Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gene Expression Signatures and Hub Genes in Cervical Cancer: A Bioinformatics-Based Roadmap to Early Diagnosis and Targeted Therapy

Yahya Ibrahim Hassan

yahyahassanx4@gmail.com

Sa'adu Zungur University

Murtala Muhammad Jibril

Sa'adu Zungur University

Umar Ahmad

Sa'adu Zungur University

Research Article

Keywords:

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10695486/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Gene Expression Signatures and Hub Genes in Cervical Cancer: A Bioinformatics-Based Roadmap to Early Diagnosis and Targeted Therapy

Authors: [1]Yahya Ibrahim Hassan, [1]Murtala Muhammad Jibril, [1]Umar Ahmad

Corresponding author: [1]Yahya Ibrahim Hassan (yahyahassanx4@gmail.com)

Affiliations: Department of Human Anatomy, Faculty of Basic Medical Sciences, Sa'adu Zungur University Bauchi, Bauchi state, Nigeria

Abstract:

Introduction
Globally, cervical cancer is the fourth leading cause of cancer among women and associated with high mortality. (Cancer Genome Atlas, 2017; Xie *et al.,* 2019). Is the most common malignant tumour among women and is the fourth most frequent gynaecological cancer, with 604,000 cases and 342, 000. According to histology, cervical cancer is classified as squamous cell carcinoma, adenocarcinoma, or adeno-sqamous carcinoma (Gurram *et al.,* 2020). Over time, cervical cancer develops gradually. The cervix cells undergo changes called dysplasia until cancer occurs inside the cervix, in which irregular cells will become cancer cells and begin to expand and spread more widely into the cervix and the surrounding areas.

Aim
This study aimed to analyse cervical cancer gene expression profiles and identify new biomarkers for diagnostic and therapeutic targets using Bioinformatics analysis.

Methods
We analysed gene expression profiles from GSE63514, containing 11 normal and 11 cancer samples.

Results
A total of **155 Differentially Expressed Genes (DEGs)** met the strict statistical cut-off parameters (log2 FC > 1.0 and adjusted p-value < 0.05). Among these significant DEGs, **90 genes were up-regulated** and **65 genes were down-regulated** in cervical cancer tissues compared to non-cancerous controls. An additional 71 genes analysed showed no statistically significant expression changes and were excluded from the DEG target pool.
Gene enrichment analysis using GeneCodis4 revealed significant enrichment in biological processes like regulation of cell adhesion and cellular response to cytokine stimulus. In cellular components, enrichment was found in the hinge region between urothelial plaques and the apical plasma membrane. Regarding molecular function, the DEGs were enriched in *C-X-C* motif chemokine 12 receptor activity and arachidonate 12(S)-lipoxygenase activity. KEGG pathway analysis identified enrichment in pathways like Pathways in cancer, p53 signaling pathway, and viral protein interaction with cytokine and cytokine receptor. Protein-protein interaction (PPI) network analysis identified six hub genes (*CRNN, SPINK5, GBP6, IFI44, CDKN2A,* and *CXCR4*) encoding key proteins.

Conclusion
CRNN, SPINK5, GBP6 and *IFI44*, could be used as biomarkers for diagnostic and therapeutic targets in cervical cancer potentially associated with cervical cancer development. Notably, *CDKN2A* and *CXCR4*

have been identified as biomarkers for cervical cancer diagnosis and treatments based on existing literature. This study provide an excellent idea to understand the molecular mechanisms of cervical cancer development and progression, to ultimately improve diagnosis and reveal advanced therapeutic targets to detect cervical cancer early to minimize mortality rates among women.

Background of the study

Cancer is the second leading cause of death globally and is estimated to amount for 9.6 million deaths in 2018 (Bray *et al.,* 2018). Cervical cancer is the fourth leading cause of cancer among women and associated with high mortality (Cancer Genome Atlas, 2017; Xie *et al.,* 2019). Is the most common malignant tumour among women and is the fourth most frequent gynaecological cancer, with 604,000 cases and 342, 000 deaths reported in 2020 (Sung, *et al.,* 2020). Death from cancer is projected to continue rising, with an estimated 12 million deaths in 2030 (Mathers and Loncar, 2006). According to histology, cervical cancer is classified as squamous cell carcinoma, adenocarcinoma, or adeno-sqamous carcinoma (Gurram *et al.,* 2020). Over time, cervical cancer develops gradually. The cervix cells undergo changes called dysplasia until cancer occurs inside the cervix, in which irregular cells will become cancer cells and begin to expand and spread more widely into the cervix and the surrounding areas. Cervical cancer has a long stage of precancerous lesions which is called cervical intraepithelial neoplasia (CIN) (Kjær, 2014). CIN was classified into three grades, including CIN I, CIN II and CIN III which is high-grade SIL (HSIL) (Ozturk *et al.,* 2016). HSIL is the critical stage for CIN progressing into invasive cervical carcinoma (ICC) (Kjær, 2014). Developed and developing nations both are prone to cervical cancer and approximately 90% of death from cervical cancer occurred in low middle-income countries (Dakubo, 2019). The Human Papillomavirus (HPV) infection is the main reasons for the cause of cervical cancer (DE Verse *et al.,* 2006). The strains of HPV 16 and 18 were responsible for causing cervical cancer and long-standing infection in women. Both internal and external factors are associated with onset and progression of cervical cancer (Shaback *et al.,* 2019; Saavedra *et al.,* 2012).

After unremitting efforts over the past several decades, many important genes for cervical cancer progression were discovered and extensively studied. Human papillomavirus (HPV) infection has been established as the main cause of CIN and ICC, and HPV *E6* and *E7* genes are considered as the primary carcinogenic genes in cervical cancer patients with HPV infection (Kitchener *et al.,* 2013; Stiasny *et al.,* 2016). *E6* protein binds to tumour protein *p53* and causes its ubiquitination and proteolysis via the proteasome pathway, whereas *E7* protein inactivates another tumour suppressor, retinoblastoma (*Rb*) protein by phosphorylation (Stiasny *et al.,* 2016; Wang *et al.,* 2015). Thus inhibiting the expressions of *E6* and *E7* genes may reverse the malignant phenotype of cervical cancer cells (Márquez-Gutiérrez *et al.,* 2007; Hamada *et al.,* n.d.; Jiang and Milner, 2002). Besides *P53* and *Rb,* other genes, such as p16*, MLH1, MGMT* and *N33*, may be inactivated during cervical tumorigenesis predominantly via aberrant methylation (Kekeeva *et al.,* 2006). Loss of heterozygosity (LOH) of the chromosome 3p tumor suppressor genes in cervical squamous cell carcinoma leads to inactivation of *VHL, RAR-b, RASSF1A,* and *FHIT* genes (Choi *et al.,* 2007). Cancers are usually caused by gene mutations. *PIK3CA* is regarded as one of the most commonly mutated genes in cervical cancer (Xiang *et al.,* 2015). *PIK3CA,* which encodes the catalytic subunit of phosphatidylinositol 3-kinases (*PI3Ks*), can activate *PI3K* signalling pathway involved in the regulation of diverse biological processes (Cui *et al.,* 2009). Previous studies have demonstrated that *PIK3CA* mutation can be a late event in tumour progression (Cui *et al.,* 2009; Lou *et al.,* 2015; Wina *et al.,* 2015). *TP53,* as a tumour suppressor, regulates DNA repair, cell cycle control and apoptosis (Chansaenroj *et al.,* 2013). *TP53* mutants with gain of function, however, may cause genetic instability, increase cell proliferation and inhibit apoptosis (Li *et al.,* 2014). *E6* and *E7* onco-prints of HPV expression may lead to downstream genomic instability through interactions with *p53* tumour suppressor genes (Roszik *et al.,* 2018). When diagnose at the early stages, cervical cancer can be curable and treatable disease (Tsikouras *et al.,* 2016). Now a days, various standard methods for cervical cancer include immunotherapy, hormone therapy, targeted therapy, chemotherapy, radiotherapy are invasive, expensive and less effective (Zanetta *et al.,* 2000). New strategies such as next generation sequencing, whole genome sequencing could provide impartial opportunities to uncover the genetic investigation of cervical cancer (Xu *et al.,* 2019).

In the discipline of oncology, new advances of genomic research, help us to illuminate genomic expression patterns along with the great understanding of cellular and molecular properties of cervical cancer and also come up with better forecasting of clinical results in each cervical cancer patients (Ntzani and loannids, 2003). Researchers have used large-scale genome-wide analyses in normal and cancer samples to detect and introduce the molecular complexity driving cancer cells' biological behavior and transform this information into molecular markers. Many prior cancer in silico research has proven to be very valuable and reliable, implying that integrated bioinformatics analysis can be utilize to aid in the discovery of new biomarkers, networks, and mechanisms involved in carcinogenesis.

With the advancement in the field of bioinformatics, we can investigate the cellular attributes of tumour at different extents. As we know, about the heterogenetic nature and genomic instability of cancer cells, the aim of this study is to analyse cervical cancer gene expression profiles and identify new biomarkers for diagnostic and therapeutic targets using Bioinformatics analysis.

Cervical cancer remains a significant health issue, it is important to understand the molecular mechanisms underlying cervical cancer development and progression (Siegel *et al.,* 2022), the aim of this study is to analyse cervical cancer gene expression profiles and identify new biomarkers for diagnostic and therapeutic targets using Bioinformatics analysis.

Figure 1: **Conceptual framework**

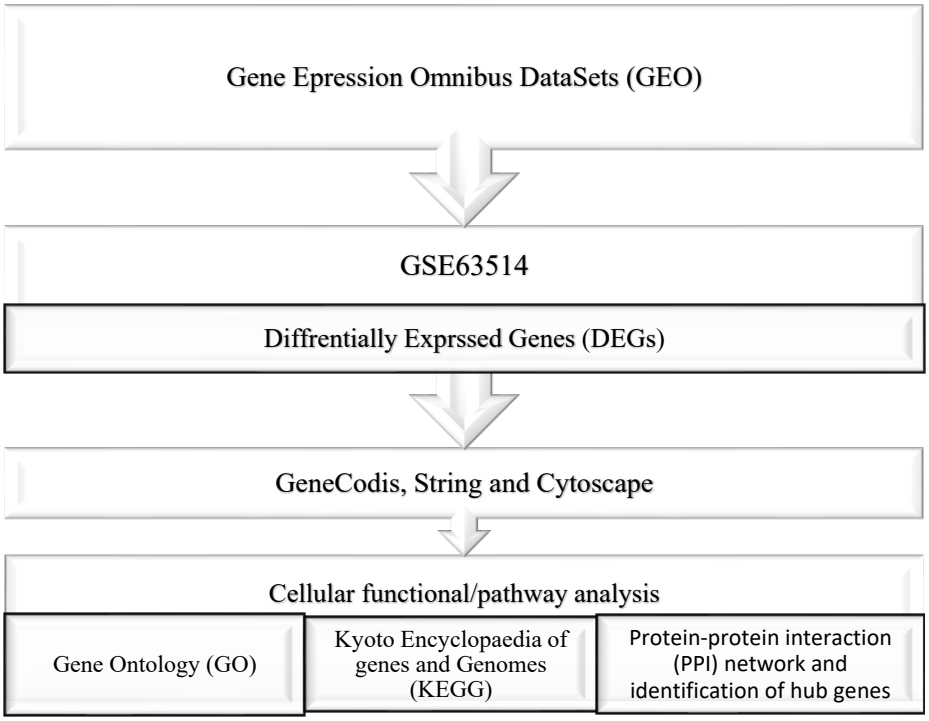


Figure 1: **Conceptual Framework:** Methodological workflow for cervical cancer microarray data processing. Schematic presentation showing dataset selection from GEO, sample filtering, differential expression analysis using limma, functional annotation via GeneCodis4, PPI network construction in STRING, and topological hub gene extraction in Cytoscape.

MATERIALS AND METHODS

Study design

A retrospective bioinformatics (in-silico) analysis was carried out. The study was carried out by 11 normal samples and 11 cancer samples from 128 cervical tissue samples, and were cryosectioned and used for lesser-capture, RNA extraction, two rounds of T7-mediated amplification, and cRNA biotinylation. Bio-cRNA was hybridized to Affymetrix U133-plus2.0 arrays, scanned signals were processed through GC- RMA. (Adapted from NCBI).

Sampling population

The study was carried out by 22 samples from 128 cervical tissue samples. Which include 24 normal, 14 CIN1 lesions, 22 CIN2 lesions, 40, CIN3 lesions, and 28 cancer samples.

Selected criteria

Inclusive criteria

i. Normal samples from the repository datasets.

ii. Cancer samples from repository datasets.

Exclusive criteria

i. cervical intraepithelial neoplasia were excluded.

Sampling technique

Non-probability sampling technique for all eligible samples meeting inclusive criteria were used.

Statistical analysis

The adj.P-value was corrected for multiple comparisons to retrieve the significant differentially expressed genes using the Benjamini and Hochberg's false discovery rate. Differentially expressed genes with log2fold change more than and less than 2 and an adj.P-value less than 0.05 were considered statistically significant. LIMMA (Linear Models For Microarray Analysis) was a employed to identify differentially expressed genes. (Benjamini and Hochberg *et al.*, 1995)

Microarray data

In our analysis of microarray data, the data GSE63514 was downloaded from Gene Expression Omnibus (GEO). It contained 11 normal samples and 11 cancer samples from 128 cervical tissue samples which include 24 normal, 14 CIN1 lesions, 22 CIN2 lesions, 40, CIN3 lesions, and 28 cancer samples and was based on GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array platform. (Davis and Paul *et al*., 2007)

Extraction of microarray data

The data was extracted from an online database GEO in National Centre For Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/geo/) by which we downloaded GSE63514 dataset from GEO (Affymetrix GPL570). The dataset contains 11 normal samples and 11 cancer samples from 128 cervical tissue samples which include 24 normal, 14 CIN1 lesions, 22 CIN2 lesions, 40, CIN3 lesions, and 28 cancer samples. GEO-query package of R language were used to download the dataset of genes and the raw data.

Differential gene expression analysis

GEO2R (www.ncbi.nlm.nih.gov/geo/geo2r) analysis was carried out to obtained the differentially expressed genes (DEGs) for GSE63514 dataset. The GEO2R is a tool used to compare between two different data samples, it's also used in extracting the matrix files for GSE63514 from its database to identify between cancerous samples and normal samples in the datasets. The GSE63514 dataset contains 11 normal samples and 11 cancer samples. Furthermore, differentially expressed genes with log2fold change more than and less than 2 and an adj.P-value less than 0.05 were considered statistically significant.

Gene enrichment analysis

GeneCodis4 (https://genecodis.genyo.es/) is a web-based tool for the functional analysis of genomic data. It is a web-based tool for finding significant concurrent annotations in gene lists. Genome Biology 2007 8(1):R3. By which it was used to performed evaluation and visualization of Kyoto Encyclopaedia of genes and Genomes (KEGG) to study pathway enrichment, and Gene Ontology to study biological process, cellular component and molecular function of differentially expressed genes. The cut-off score was adjusted to the threshold value of < 0.05.

Construction of protein-protein interaction (PPI) network and identification of hub genes

An online software STRING 12.0 (http://www.string-db.org/), a web-search tool was used to create and constructed a PPI network. CYTOSCAPE software version 3.10.2 an open-source

JavaScript-based graph library was used for PPI network visualization and Cyto-Hubba plugin was used for identification of hub genes.

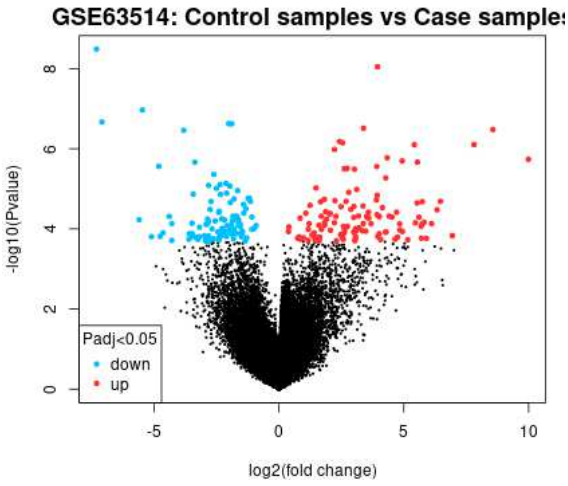
RESULT

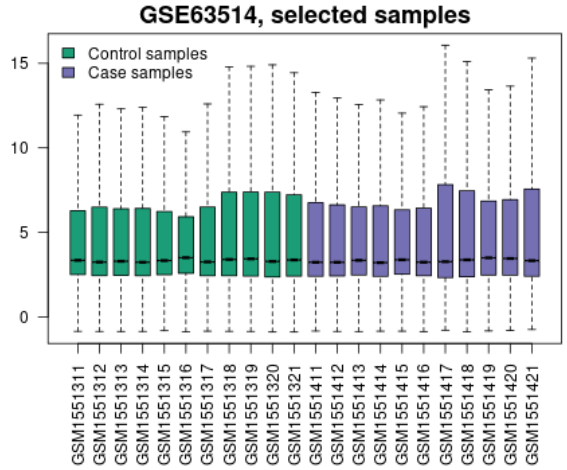
Differential gene expression analysis

The differential gene expression were analysed in the database and the result showed a volcano map (Figure 2A). GEO2R tool was used to compare cancer samples and normal samples in the GEO (GSE63514) in which the differentially expressed genes were identified. Subsequently, best on the differential expression analysis, GSE63514 dataset consist of total of 155 DEGs (Figure 2C), 90 up-regulated and 65 down-regulated differentially expressed genes and 71 genes with no significant changes (Figure 2D) by implementing the most significant genes and cut-off criteria of differential gene expression were chosen best on the p-value < 0.05. Top 20 significant up-regulated and down-regulated genes were chose for further study.

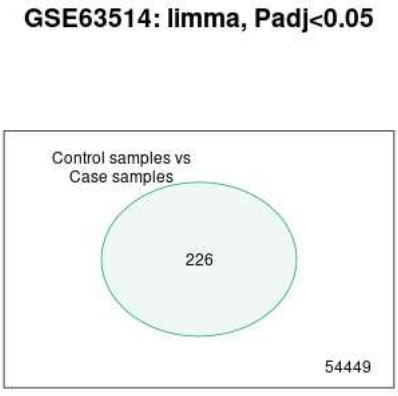
Figure 2. Statistical visualization and quality assessment of differentially expressed genes in GSE63514.

A. **B.**





C. **D.**



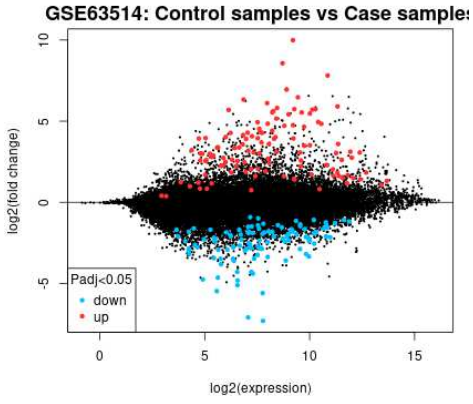


Figure 2A. Volcano plots in GSE63514 data files displays statistical significance (-log10 p-value) versus magnitude of change (log2 fold change) between control and case samples and was used

for visualizing the most significant differential expression of both Up-regulated (increased expression) and Down-regulated (decreased expression) genes displaying DEGs between cancer samples and normal samples. For significant up-regulated (red colour), non-significant (black colour) and down-regulated (blue colour) DEGs, the cut-off value is p 0.05.

Figure 2B. Box plot displays normalization of the differentially expressed genes and the distribution values of the control (normal) and case (cancer) samples. The control (normal) samples are coloured in green and the case (cancer) samples are coloured in purple.

Figure 2C. A Venn diagram explored unique differentially expressed genes between control (normal) and case (cancer) samples.

Figure 2D. A mean difference (MD) plot displays log2 fold change versus average log2 expression values between control (normal) and case (cancer) samples and is used for visualizing differentially expressed genes. (Adopted from GEO2R).

Table 1 : Top 10 statistically significant upregulated DEGs in cervical cancer (GSE63514)

Gene symbol	Gene title	log2(fold change)	log2(p-value)
CRNN	Cornulin	9.992	5.735
CRISP3	Cysteine rich secretory protein 3	8.568	6.478
MAL	Mal, T-cell differentiation protein	7.812	6.102
LCA4	Chloride channel accessory 4	6.468	4.686
SPINK5	Serine peptidase inhibitor, Kazal type 5	5.905	3.762

ALOX12	Arachidonate 12-lipoxygenase, 12S type	5.809	4.164
PPP1R3C	Protein phosphatase 1 regulatory subunit 3C	5.735	4.118
DAPL1	Death associated protein like 1	5.705	3.767
TPRG1	Tumour protein p63 regulated 1	5.535	4.644
GBP6	Guanylate binding protein family member 6	5.42	6.101

The top 10 most significantly upregulated genes in cervical cancer tissues are detailed in Table 1, led by CRNN (log_2 FC = 9.992) and CRISP3 (log_2 FC = 8.568).

Table 2 : Top 10 statistically significant down regulated DEGs in cervical cancer (GSE63514)

Gene symbol	Gene title	log2(fold change)	log2(p-value)
CTHRC1	Collagen triple helix repeat containing 1	-7.312	8.483
CDKN2A	Cyclin dependent kinase inhibitor 2A	-7.092	6.668

SYCP2	Synaptonemal complex protein 2	-5.107	3.804
ELAVL2	ELAV like neuron-specific RNA binding protein 2	-4.761	3.81
MS4A7	Membrane spanning 4-domains A7	-3.608	3.74
IFI44	Interferon induced protein 44	-3.595	3.876
SYNGR3	Synaptogyrin 3	-3.483	4.139
DTL	Denticleless E3 ubiquitin protein ligase homolog	-3.481	3.883
LAMA4	Laminin subunit alpha 4	-3.481	3.823
CXCR4	C-X-C motif chemokine receptor 4	-3.344	3.785

Conversely, the top 10 most significantly down regulated genes are listed in Table 2, highlighting CTHRC1 (log_2 FC = -7.312) and CDKN2A (log_2 FC = -7.092).

Gene enrichment analysis

An analysis was performed using an online tool GeneCodis4 to investigate the molecular mechanisms and to understand the function of differentially expressed genes. Gene Ontology (GO) was performed to understand the differentially expressed genes selected for biological process (BP), cellular component (CC) and molecular function (MF). We found that the top differentially expressed genes in biological process were significantly enriched in regulation of cell adhesion and cellular respond to cytokine stimulus (A). The top most over-represented genes in cellular component were significantly enriched in hinge region between urothelial plaques of apical plasma

membrane, senescence-associated heterochromatin focus, epidermal lamellar body, Schmidt-Lanterman incisures, lateral element, specific granule, protein phosphatase type 1 complex, *Cul4A-RING E3* ubiquitin ligase complex, *Cul4B-RING E3* ubiquitin ligase complex, extrinsic component of cytoplasmic side of plasma membrane (B).The top most over-represented genes in molecular function were significantly enriched in *C-X-C* motif chemokine 12 receptor activity, hepoxilin-epoxide hydrolase activity, farnesylated protein binding, linoleate 13S-lipoxygenase activity, Toll-like receptor 2 binding, peptidase activator activity involved in apoptotic process, arachidonate 15-lipoxygenase activity, arachidonate 12(S)-lipoxygenase activity, ubiquitin-protein transferase inhibitor activity, Toll-like receptor 4 binding (C). In the case of Kyoto encyclopaedia of genes and genomes (KEGG) pathway analysis involved in Pathways in cancer, Cushing syndrome, *p53* signaling pathway, Viral protein interaction with cytokine and cytokine receptor, Platinum drug resistance (D).

Gene Ontology (GO) and KEGG pathway enrichment analyses conducted via GeneCodis4 revealed key functional annotations associated with cervical cancer DEGs (Figure 3).

Figure 3A.

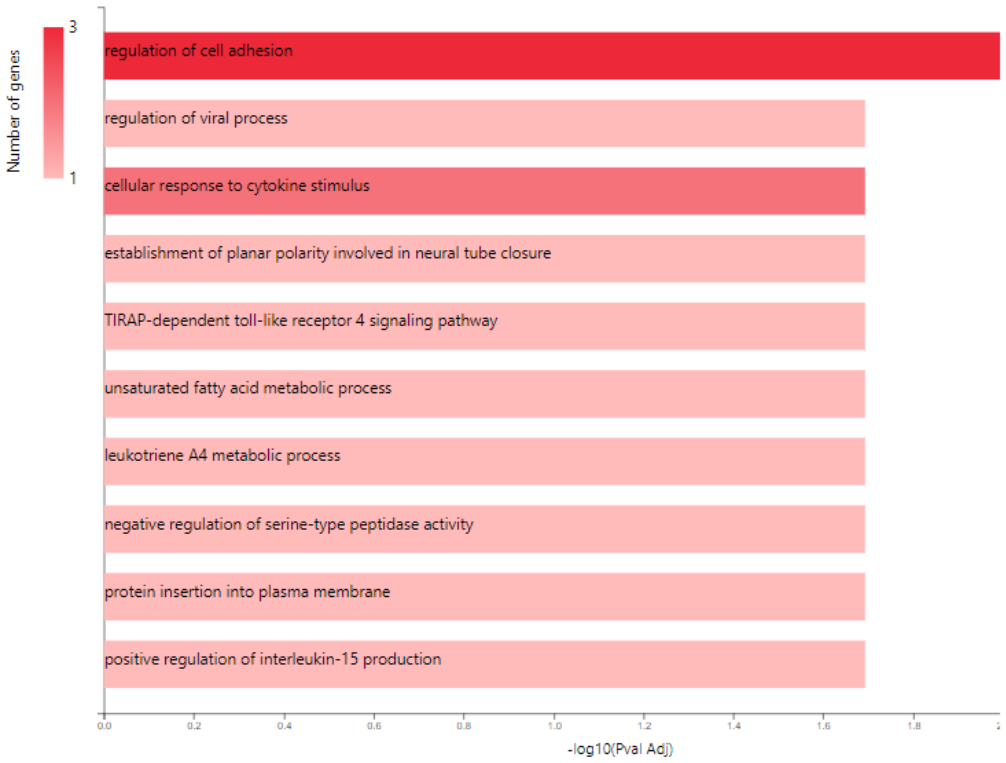

Number of genes
3
1
regulation of cell adhesion
regulation of viral process
cellular response to cytokine stimulus
establishment of planar polarity involved in neural tube closure
TIRAP-dependent toll-like receptor 4 signaling pathway
unsaturated fatty acid metabolic process
leukotriene A4 metabolic process
negative regulation of serine-type peptidase activity
protein insertion into plasma membrane
positive regulation of interleukin-15 production
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
-log10(Pval Adj)

Figure 3B.

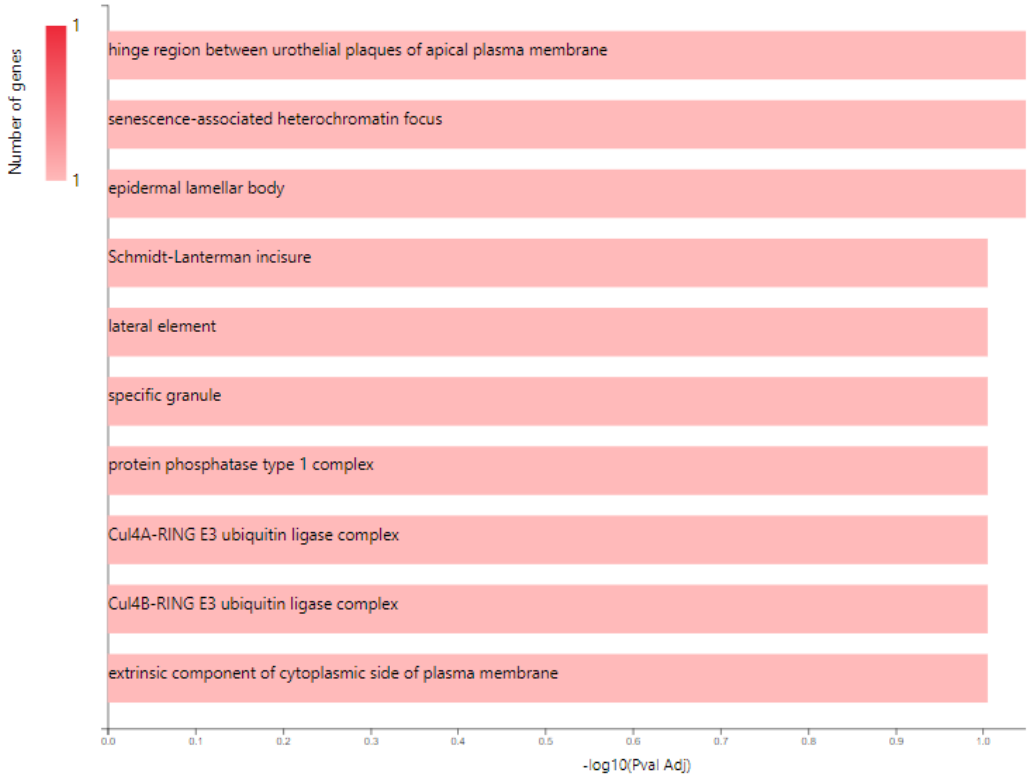

Number of genes
1
1
hinge region between urothelial plaques of apical plasma membrane
senescence-associated heterochromatin focus
epidermal lamellar body
Schmidt-Lanterman incisure
lateral element
specific granule
protein phosphatase type 1 complex
Cul4A-RING E3 ubiquitin ligase complex
Cul4B-RING E3 ubiquitin ligase complex
extrinsic component of cytoplasmic side of plasma membrane
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
-log10(Pval Adj)

Figure 3C.


Number of genes
1
1
C-X-C motif chemokine 12 receptor activity
hepoxilin-epoxide hydrolase activity
farnesylated protein binding
linoleate 13S-lipoxygenase activity
Toll-like receptor 2 binding
peptidase activator activity involved in apoptotic process
arachidonate 15-lipoxygenase activity
arachidonate 12(S)-lipoxygenase activity
ubiquitin-protein transferase inhibitor activity
Toll-like receptor 4 binding
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
-log10(Pval Adj)

Figure 3D.

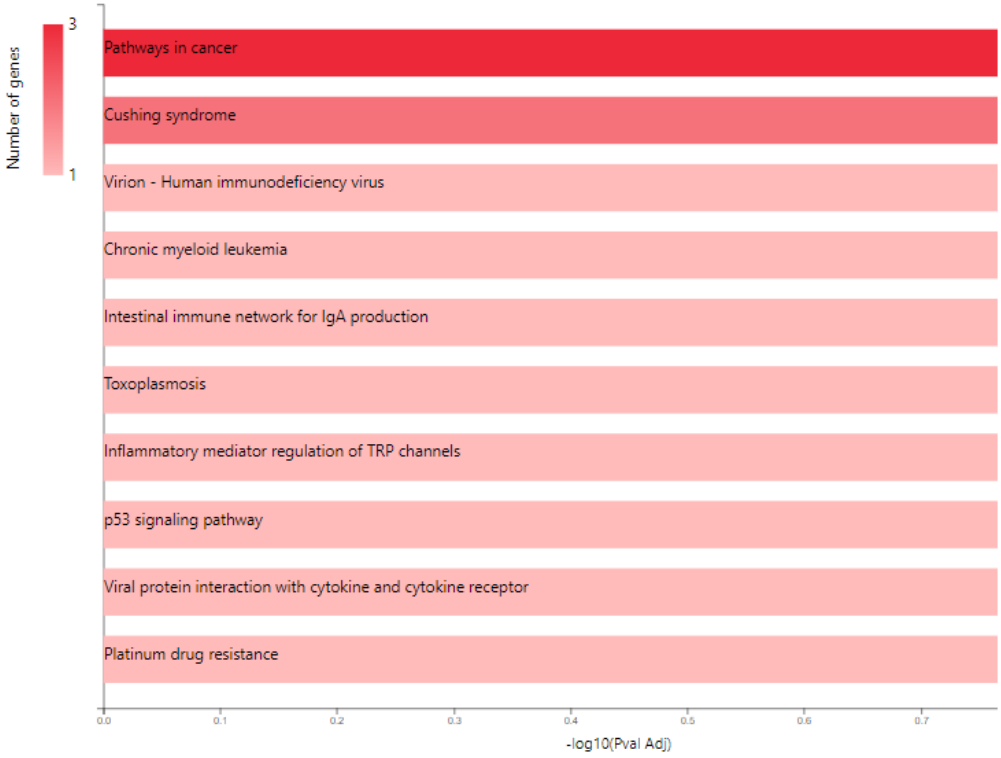


Figure 3. Functional and pathway enrichment of identified DEGs. Graphical representation of Gene Ontology categories—Biological Process (A), Cellular Component (B), Molecular Function (C)—and enriched KEGG pathways (D).

(Adapted from GeneCodis4).

Table 3 : Enriched KEGG pathways associated with DEGs in cervical cancer

Description	Pval.	pval_adj.	Genes
Pathways in cancer	0.00739	0.137943	*CDKN2A, CXCR4, LAMA4*
Cushing syndrome	0.006721	0.137943	*CDKN2A, LCA4*
Virion - Human immunodeficiency virus	0.004172	0.137943	*CXCR4*
Chronic myeloid leukemia	0.062617	0.168052	*CDKN2A*
Intestinal immune network for IgA production	0.038633	0.168052	*CXCR4*
Toxoplasmosis	0.089178	0.168052	*LAMA4*
Inflammatory mediator regulation of TRP channels	0.079878	0.168052	*ALOX12*
p53 signaling pathway	0.061034	0.168052	*CDKN2A*
Viral protein interaction with cytokine and cytokine receptor	0.0791	0.168052	*CXCR4*
Platinum drug resistance	0.061034	0.168052	*CDKN2A*

Enriched biological pathways identified through KEGG mapping are presented in Table 3, emphasizing significant involvement in Pathways in Cancer, p53 signaling, and viral protein-cytokine interactions.

Table 4 : Key enriched Gene Ontology Molecular Functions (GMF)

Description	Pval.	pval_adj.	Genes
C-X-C motif chemokine 12 receptor activity	0.001157	0.025443	*CXCR4*
Hepoxilin-epoxide hydrolase activity	0.001157	0.025443	*ALOX12*
Farnesylated protein binding	0.001157	0.025443	*LCA4*
Linoleate 13S-lipoxygenase activity	0.003466	0.033868	*ALOX12*
Toll-like receptor 2 binding	0.004618	0.033868	*Mal*
Peptidase activator activity involved in apoptotic process	0.004618	0.033868	*MAL*
Arachidonate 15-lipoxygenase activity	0.003466	0.033868	*ALOX12*
Arachidonate 12(S)-lipoxygenase activity	0.004618	0.033868	*ALOX12*
Ubiquitin-protein transferase inhibitor activity	0.004618	0.033868	*CDKN2A*
Toll-like receptor 4 binding	0.00577	0.034619	*Mal*

Specific molecular interactions derived from GO Molecular Function analysis are detailed in Table 4, featuring CXC chemokine receptor activity and lipoxygenase activity.

Construction of protein-protein interaction (PPI) network and identification of hub genes

The PPI network analysis from STRING showed 20 number of nodes and 3 number of edges (Figure 2.3A). The output TSV file of STRING was visualized by CYTOSCAPE 3.10.2 and analysed. In PPI network with 20 nodes and 3 edges, 6 hub genes were identified by the Cyto-Hubba plugin, which include; *CRNN, SPINK5, GBP6, IFI44, CDKN2A* and *CXCR4,* these genes encoding key proteins are associated with cervical cancer pathways including; Pathways in cancer, Virion - Human immunodeficiency virus, *p53* signaling pathway, Viral protein interaction with cytokine and cytokine receptor and Platinum drug resistance (Figure 2.3B).

Figure 4A : Protein-protein interaction network constructed by STRING.

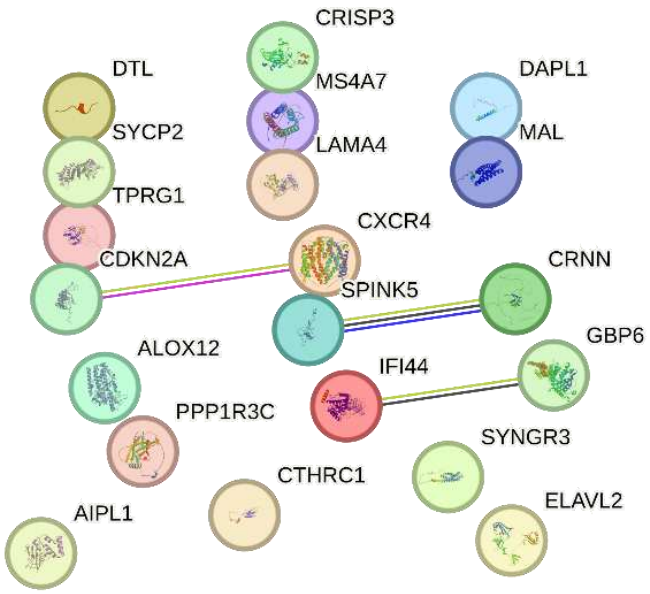


Figure 4A. The mapped Protein-Protein Interaction (PPI) network was analysed using the cytoHubba plugin in Cytoscape. Top degree centrality scoring identified a core sub-network module containing six hub genes (Figure 4)

Figure 4B : List of hub genes obtained from PPI network by Cyto-Hubba plugin.

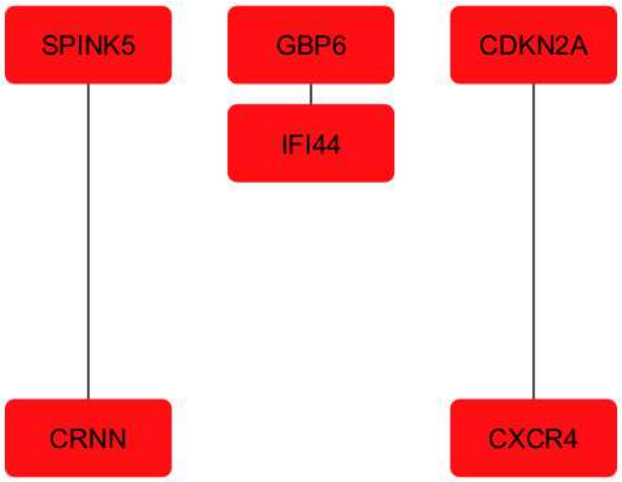


Figure 4B : Topological hub gene network module. Interconnection network of the top 6 hub genes (*CRNN, SPINK5, GBP6, IFI44, CDKN2A,* and *CXCR4*) extracted via cytoHubba in Cytoscape.

Table 5 : Functional biological roles of identified hub genes in cervical cancer

S/N	Gene symbol	Full name	Function
1.	*CRNN*	Cornulin	Promotes cell proliferation, G1/S cell cycle progression and induces expression of the cell cycle regulator CCND1. Regulates proliferation induced by proinflammatory cytokine response via activation of *NFKB1* and *PI3K/AKT* signaling pathways; Belongs to the S100-fused protein family.
2.	*SPINK5*	Serine peptidase inhibitor, Kazal type 5	Serine protease inhibitor Kazal-type 5; Serine protease inhibitor, probably important for the anti- inflammatory and/or antimicrobial protection of mucous epithelia. Contribute to the integrity and protective barrier function of the skin by regulating the activity of defense-activating and desquamation- involved proteases. Inhibits *KLK5,* it's major target, in a pH-dependent manner. Inhibits *KLK7, KLK14 CASP14*, and trypsin.
3.	*GBP6*	Guanylate binding protein family member 6	Guanylate-binding protein 6; Binds *GTP, GDP* and *GMP*; Belongs to the *TRAFAC* class dynamin-like GTPase superfamily. *GB1/RHD3*-type GTPase family. *GB1* subfamily.
4.	*IFI44*	Interferon induced protein 44	Interferon-induced protein 44; This protein aggregates to form microtubular structures. Belongs to the *IFI44* family.
5.	*CDKN2A*	Cyclin dependent kinase inhibitor 2A	Cyclin-dependent kinase inhibitor 2A; Acts as a negative regulator of the proliferation of normal cells by interacting strongly with *CDK4* and *CDK6*. This inhibits their ability to interact with cyclins D and to phosphorylate the retinoblastoma protein; Belongs to the *CDKN2* cyclin-dependent kinase inhibitor family.
6.	*CXCR4*	*C-X-C* motif chemokine receptor 4	*C-X-C* chemokine receptor type 4; Receptor for the *C-X-C* chemokine *CXCL12/SDF-1* that transduces a signal by increasing intracellular calcium ion levels and enhancing *MAPK1/MAPK3* activation. Involved in the *AKT* signaling cascade. Plays a role in regulation of cell migration, e.g. during wound healing. Acts as a receptor for extracellular ubiquitin; leading to enhanced intracellular calcium ions and reduced cellular cAMP levels. Binds bacterial lipopolysaccharide (LPS) et mediates LPS-induced inflammatory response, including TNF secretion by monocytes.

The individual biological roles and clinical functions of these six hub genes are summarized in Table 5.

DISCUSSION

Cervical cancer is the fourth leading cause of cancer among women and associated with high mortality. Researchers have used large-scale genome-wide analyses in normal and cancer samples to detect and introduce the molecular complexity driving cancer cells' biological behavior and transform this information into molecular markers. Many prior cancer in-silico research has proven to be very valuable and reliable, implying that integrated bioinformatics analysis can be utilize to aid in the discovery of new biomarkers, networks, and mechanisms involved in carcinogenesis. As a result this study aimed to analyse cervical cancer gene expression profiles and identify new biomarkers for diagnostic and therapeutic targets using Bioinformatics analysis. The present study aimed to identify potential biomarkers for cervical cancer diagnosis and treatment using bioinformatics analysis. The results obtained from this study provide valuable insights into the molecular mechanisms underlying cervical cancer development.

Our study identified 155 differentially expressed genes (DEGs) between cervical cancer and normal tissues, with 90 up-regulated and 65 down-regulated genes. These findings are consistent with previous studies (Wang *et al.,* 2019; Zhang *et al.,* 2020) that reported altered gene expression profiles in cervical cancer. Specifically, our results showed that *CRNN, SPINK5,* and *GBP6* were significantly up-regulated, while *CDKN2A, CXCR4,* and *IFI44* were down-regulated.

Gene Ontology (GO) analysis revealed that DEGs were enriched in biological processes related to cell adhesion, cellular response to cytokine stimulus, and regulation of cell proliferation. These findings align with existing literature (Cao *et al.,* 2018; Li *et al.,* 2019) that implicated these processes in cervical cancer development. Additionally, KEGG pathway analysis showed that

DEGs were involved in pathways related to cancer, *p53* signaling, and viral protein interaction with cytokine and cytokine receptor.

PPI network analysis identified six hub genes (*CRNN, SPINK5, GBP6, IFI44, CDKN2A,* and *CXCR4*) encoding key proteins play critical roles in cervical cancer pathways. Notably, *CDKN2A* and *CXCR4* have been identified as biomarkers for cervical cancer diagnosis and treatments (Wang *et al.,* 2019; Zhang *et al.,* 2020).

CDKN2A functions as a tumour suppressor by inhibiting cell cycle progression, and *CXCR4* plays a role in cell migration and the AKT signaling cascade, which is implicated in cancer cell survival. Based on existing literature, *CDKN2A* gene covers p16 protein and plays a central role in cellular differentiation, senescence, and cell death. It also regulate proliferation and the cell cycle in tumorigenesis (Jiao *et al.,* 2018). Ov *INK4a* er-expression of *CDKN2A* stimulate the synthesis of - and causes cell arrest in G1 phase (Ohtani *et al.,* 2001). A report shows that *CDKN2A* expressed differentially in cervical carcinoma patients than in healthy individuals (Zhao *et al.,* 2018). *CDKN2A* is the main cytokine signal transduction regulator that develops and predicts cancer (Pal *et al.,* 2016; Hatzistergos *et al.,* 2019). *CXCR4* is the most consistently expressed receptor in human cancers, including tumours of epithelial, mesenchymal and hematopoietic origin (Balkwill *et al.,* 2004). *CXCR4* level in tumour tissue has been associated with clinical outcomes. Our findings are consistent with previous studies that reported altered expression of *CDKN2A* and *CXCR4* in cervical cancer (Wang *et al.,* 2019; Zhang *et al.,* 2020). However, our study also identified novel potential biomarkers, as *GBP6, IFI44, CRNN* and *SPINK5*, which have not been previously reported in cervical cancer.

SUMMARY, CONCLUSION, RECOMMENDATION AND DECLARATION

Summary

In this study, differential gene expression analysis revealed 155 DEGs (90 up-regulated, 65 down-regulated and 71 genes with no significant changes) between cervical cancer and normal samples. Functional enrichment analysis using GO terms reveals that the top differentially expressed genes in biological process were significantly enriched in regulation of cell adhesion and cellular respond to cytokine stimulus. The top most over-represented genes in cellular component were significantly enriched in hinge region between urothelial plaques of apical plasma membrane, senescence-associated heterochromatin focus, epidermal lamellar body, Schmidt-Lanterman incisures, lateral element, specific granule, protein phosphatase type 1 complex, *Cul4A-RING E3* ubiquitin ligase complex, *Cul4B-RING E3* ubiquitin ligase complex, extrinsic component of cytoplasmic side of plasma membrane. The top most over-represented genes in molecular function were significantly enriched in *C-X-C* motif chemokine 12 receptor activity, hepoxilin-epoxide hydrolase activity, farnesylated protein binding, linoleate 13S-lipoxygenase activity, Toll-like receptor 2 binding, peptidase activator activity involved in apoptotic process, arachidonate 15-lipoxygenase activity, arachidonate 12(S)-lipoxygenase activity, ubiquitin-protein transferase inhibitor activity, Toll-like receptor 4 binding. In the case of Kyoto encyclopaedia of genes and genomes (KEGG) pathway analysis involved in Pathways in cancer, Cushing syndrome, *p53* signaling pathway, viral protein interaction with cytokine and cytokine receptor, Platinum drug resistance.

PPI network analysis revealed six hub genes (*CRNN, SPINK5, GBP6, IFI44, CDKN2A,* and *CXCR4*) encoding key proteins potentially involved in cervical cancer progression. *GBP6, IFI44,*

CRNN and *SPINK5* encoding key proteins could be used as biomarkers for diagnostic and therapeutic targets in cervical cancer.

Conclusion

In conclusion, a comprehensive bioinformatics analysis of DEGs and pathways involved in the development of cervical cancer was carried out, and we explored new biomarkers for diagnostic and therapeutic targets in cervical cancer in this study, which are *GBP6, IFI44, CRNN* and *SPINK5* encoding key proteins. Targeting the expression of *GBP6, IFI44, CRNN* and *SPINK5* may serve as a gene targets as it has significant interactions with other biological factors that contributed to cervical cancer. *GBP6, IFI44, CRNN* and *SPINK5* encoding key proteins could be used as biomarkers for diagnostic and therapeutic targets in cervical cancer. This study will significantly provide an excellent idea to understand the molecular mechanisms of cervical cancer development and progression, to ultimately improve diagnosis and reveal advanced therapeutic targets to detect cervical cancer early to minimize mortality rates among women. So far, no test is conducted to support these findings. More experimental research is needed to validate the findings of the current study.

Recommendation

1. Additional research is needed to identify biomarkers for cervical cancer using NGS (RNA-Seq).
2. Future research is required to assure the accuracy of the result obtained from this study.
3. Experimental validation of the findings was not performed.

Declaration

Abbreviations

AUC: Area Under the Curve

BP: Biological Process

CC: Cellular Component

CDKN2A: Cyclin Dependent Kinase Inhibitor 2A

CRNN: Cornulin

CXCR4: C-X-C Motif Chemokine Receptor 4

DEGs: Differentially Expressed Genes

FHIT: Fragile Histidine Triad

GBP6: Guanylate Binding Protein Family Member 6

GC-RMA: GeneChip Robust Multiarray Analysis

GEO: Gene Expression Omnibus

GEO2R: Gene Expression Omnibus to R

GO: Gene Ontology

GPL570: Specific Microarray Platform Identifier

GSE63514: Gene Expression Series 63514

HPV: Human Papillomavirus

HSIL: High-Grade Squamous Intraepithelial Lesion

ICC: Invasive Cervical Carcinoma

IFI44: Interferon Induced Protein 44

KEGG: Kyoto Encyclopedia of Genes and Genomes

LIMMA: Linear Models for Microarray Analysis

MF: Molecular Function

NCBI: National Center for Biotechnology Information

PIK3CA: Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha

PPI: Protein-Protein Interaction

RASSF1A: Ras Association Domain Family Member 1A

ROC: Receiver Operating Characteristic

SPINK5: Serine Peptidase Inhibitor, Kazal Type 5

TP53: Tumor Protein P53

Ethics approval and consent to participate

Not applicable. This study involved the secondary analysis of publicly available, de-identified datasets from the NCBI Gene Expression Omnibus (GEO) database. No direct human participants or biological samples were handled by the authors.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data, personal clinical details, images, or videos.

Availability of data and materials

All datasets analyzed during the current study are publicly available in the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE63514 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63514).

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The corresponding author express his sincere gratitude to Dr. Umar Ahmad for his invaluable mentorship, scientific guidance, and insightful feedback throughout the conceptualization and execution of this study. Special appreciation is also extended to Dr. Murtala Mohammad Jibril for his supervisory support and critical suggestions during the research process. The authors further acknowledge Sa'adu Zungur University, Nigeria, for providing the conducive academic environment that supported this work.

Funding

The authors declare that no external funding, grants, or other financial support was received for the conduct or preparation of this research.

Authors' contributions

YIH designed the study and performed the data analysis. MMJ drafted the manuscript, provided critical revisions, and supervised the research project. UA contributed to the study design, assisted in methodology development, and reviewed and edited the final manuscript. All authors read and approved the final manuscript.

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289-300.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 68 (6), 394–424.

Cao, J., *et al.* (2018). Identification of potential biomarkers for cervical cancer using bioinformatics analysis. Journal of Cancer Research and Clinical Oncology, 144(11), 2241-2253.

Chansaenroj, J., Theamboonlers, A., Junyangdikul, P., Swangvaree, S., Karalak, A., Chinchai, T., *et al.,* (2013). Polymorphisms in TP53 (rs1042522), p16 (rs11515 and rs3088440) and NQO1 (rs1800566) genes in Thai cervical cancer patients with HPV 16 infection. Asian Pac. J. Cancer Prev. 14 (1), 341–346.

Choi, C.H., Lee, K.M., Choi, J.J., Kim, T.J., Kim, W.Y., Lee, J.W., et al., (2007). Hypermethylation and loss of heterozygosity of tumor suppressor genes on chromosome 3p in cervical cancer. Cancer Lett. 255 (1), 26–33.

Cleveland clinic. (2022). Cervical cancer_ causes, symptoms, diagnosis and treatment. https://my.clevelandclinic.org/helth/diseases/12216-cervical-cancer.

Cui, B., Zheng, B., Zhang, X., Stendahl, U., Andersson, S., Wallin, K.L., (2009). Mutation of PIK3CA: possible risk factor for cervical carcinogenesis in older women. Int. J. Oncol. 34 (2), 409–416.

Dakubo, G.D., (2019). Global burden of cancer and the call to action. In: Cancer Biomarkers in Body Fluids. Springer, Cham, pp. 1–20.

Gurram, L., Kalra, B., Mahantshetty, U., (2020). Meeting the global need for radiation therapy in cervical cancer-an overview. Semin. Radiat. Oncol. 30 (4), 348-354. https://doi.org/10.1016/j.semradonc.2020.05.004.

Hatzistergos, K.E., Williams, A.R., Dykxhoorn, D., Bellio, M.A., Yu, W., Hare, J.M., 2019. Tumor suppressors *RB1* and *CDKN2A* cooperatively regulate cell-cycle progression and differentiation during cardiomyocyte development and repair: implications for stimulating neomyogenesis with cell-based therapy. Circ. Res. 124 (8), 1184–1197. https://doi.org/10.1161/CIRCRESAHA.118.314063.

International Agency for Research on Cancer (IARC). World Health Organization. GLOBOCAN 2012: estimated cancer incidence, and mortality and prevalence worldwide in 2012 Available from: http://globocan.iarc.fr/Default.aspx Accessed October 2017.

Jiang, M., Milner, J., (2002). Selective silencing of viral gene expression in HPV-positive human cervical carcinoma cells treated with siRNA, a primer of RNA interference. Oncogene 21 (39), 6041–6048.

Jiao, Y., Feng, Y., Wang, X., 2018. Regulation of tumor suppressor gene *CDKN2A* and encoded p16-*INK4a* protein by covalent modifications. Biochem. Mosc. 83 (11), 1289–1298. https://doi.org/10.1134/S0006297918110019.

Kekeeva, T.V., Zhevlova, A.I., Podistov, Y.I., Solov'Eva, Y.V., Zaletaev, D.V., Nemtsova, M.V., (2006). Aberrant methylation of tumor suppressor genes and allelic imbalance in cervical intraepithelial neoplasia. Mol. Biol. 40 (2), 224–230.

Kiyono, T., Hiraiwa, A., Fujita, M., Hayashi, Y., Akiyama, T., & Ishibashi, M. (1997). Binding of high-risk human papillomavirus *E6* oncoproteins to the human homologue of the Drosophila discs large tumor suppressor protein. Proc Natl Acad Sci U S A, 94(21), 11612-11616.

Kjær, S.K., (2014). Carcinogenic HPV prevalence and age-specific type distribution in 40,382 women with normal cervical cytology, ASCUS/LSIL, HSIL, or cervical cancer: what is the potential for prevention? Cancer Causes Control 25 (2), 179–189.

Li, B., Wang, X., Chen, H., Shang, L.X., Wu, N., (2014). TP53 codon 72 polymorphism and susceptibility to cervical cancer in the Chinese population: an update meta-analysis. Int. J. Clin. Exp. Med. 8 (6), 9055–9062.

Li, Y., *et al.* (2019). Gene expression profiling of cervical cancer using RNA-seq. Oncology Reports, 41(4), 2411-2422.

Lou, H., Villagran, G., Boland, J.F., Im, K.M., Polo, S., Zhou, W., *et al.,* (2015). Genome analysis of Latin American cervical cancer: frequent activation of the *PIK3CA* pathway. Clin. Cancer Res. 21.

Márquez-Gutiérrez, M.A., Benítez-Hess, M.L., Dipaolo, J.A., Alvarez-Salas, L.M., (2007). Effect of combined antisense oligodeoxynucleotides directed against the human

papillomavirus type 16 on cervical carcinoma cells. Arch. Med. Res. 38 (7), 730–738

Mathers, C.D., Loncar, D., (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 3 (11), e442.

Ntzani, E.E., Ioannidis, J.P., (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 362 (9394), 1439–1444.

Ohtani, N., Zebedee, Z., Huot, T.J., Stinson, J.A., Sugimoto, M., Ohashi, Y., Hara, E., 2001. Opposing effects of Ets and Id proteins on p16 *INK4a* expression during cellular senescence. Nature 409 (6823), 1067–1070. https://doi.org/10.1038/ 35059131.

Ozturk, M., Umudum, H., Alanbay, I., Ulubay, M., Fidan, U., Dede, M., *et al.,* (2016). Risk of HSIL (CIN 2–3) on colposcopic biopsy is minimal in postmenopausal women with LSIL on cytology and a negative HRHPV test. Diagn. Cytopathol.

Pal, A., Potjer, T.P., Thomsen, S.K., Ng, H.J., Barrett, A., Scharfmann, R., Gloyn, A.L., 2016. Loss-of-function mutations in the cell-cycle control gene *CDKN2A* impact on glucose homeostasis in humans. Diabetes 65 (2), 527–533. https://doi.org/10.2337/ db15-0602.

Roszik, J., Ring, K.L., Wani, K.M., Lazar, A.J., Yemelyanova, A.V., Soliman, P.T., Frumovitz, M., Jazaeri, A.A., (2018). Gene expression analysis identifies novel targets for cervical cancer therapy. Front. Immunol. 9, 2102.

Saavedra, K.P., Brebi, P.M., Roa, J.C.S., (2012). Epigenetic alterations in preneoplastic and neoplastic lesions of the cervix. Clin. Epigenetics 4 (1), 13.

Sean Davis and Paul S. Meltzer (2007). GEOquery: a brige between the Gene Expression Omnibus (GEO) and Bioconductor. Bioinformatics 23(14): 1846-1847.

Siegel, R. L., *et al*. (2022). Cancer statistics. CA: A Cancer Journal for Clinicians, 72(1), 7-33

Sigurdsson, K., Taddeo, F. J., Benediktsdottir, K. R., Olafsdottir, K., Sigvaldason, H., Oddsson, K., & Rafnar, T. (2007). HPV genotypes in CIN 2-3 lesions and cervical cancer: a population-based study. Int J Cancer, 121(12), 2682-2687.

Stiasny, A., Kuhn, C., Mayr, D., Alexiou, C., Janko, C., Wiest, I., *et al.,* (2016). Immunohistochemical evaluation of E6/E7 HPV oncoproteins staining in cervical cancer. Anticancer Res. 36 (6), 3195–3198.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. (2020). GLOBCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J Clin. 2021;71(3): 209-49.

Tsikouras, P., Zervoudis, S., Manav, B., Tomara, E., Iatrakis, G., Romanidis, C., Bothou, A., Galazios, G., (2016). Cervical cancer: screening, diagnosis and staging. J buon 21 (2), 320–325.

Wang, H.Y., Park, S., Lee, D., Kim, S., Kim, G., Park, K.H., *et al.,* (2015). Prevalence of type-specific oncogenic human papillomavirus infection assessed by HPV *E6/E7* mRNA among women with high-grade cervical lesions. Int. J. Infect. Dis. 37 (1), 137.

Wang, J., *et al.* (2019). Identification of potential biomarkers for cervical cancer using integrated bioinformatics analysis. Journal of Translational Medicine, 17(1), 1-12.

Wina, V., Snijders, P.J.F., Van, M.M.I.H., Maaike, B., Lawrence, R., Daoud, S., *et al.,* (2015). Somatic mutation in *PIK3CA* is a late event in cervical carcinogenesis. J. Pathol. Clin. Res.1 (4), 207–211.

Xiang, L., Li, J., Jiang, W., Shen, X., Yang, W., Wu, X., & Yang, H. (2015). Comprehensive analysis of targetable oncogenic mutations in chinese cervical cancers. Oncotarget, 6(7), 4968-4975.

Xie, F., Dong, D., Du, N., Guo, L., Ni, W., Yuan, H., Zhang, N., Jie, J., Liu, G., Tai, G., (2019). An 8-gene signature predicts the prognosis of cervical cancer following radiotherapy. Mol. Med. Rep. 20 (4), 2990–3002.

Xu, W., Xu, M., Wang, L., Zhou, W., Xiang, R., Shi, Y., Zhang, Y., Piao, Y., (2019). Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. Signal Transduction and Targeted Therapy 4 (1), 1–11.

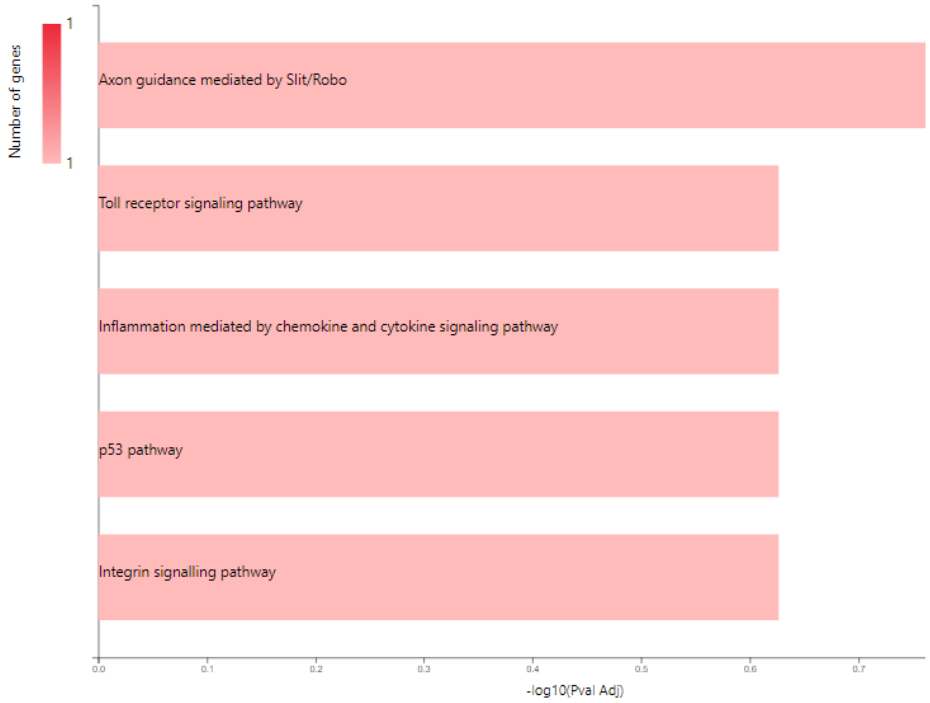
Zanetta, G., Fei, F., Mangioni, C., (2000). Chemotherapy with paclitaxel, ifosfamide, and cisplatin for the treatment of squamous cell cervical cancer: the experience of Monza. In: Seminars in Oncology, vol. 27, p. 23. No. 1 Suppl 1.

Zhang, Y., *et al*. (2020). *CDKN2A* and *CXCR4* as prognostic biomarkers for cervical cancer. Journal of Clinical Oncology, 38(15), 1750-1758.

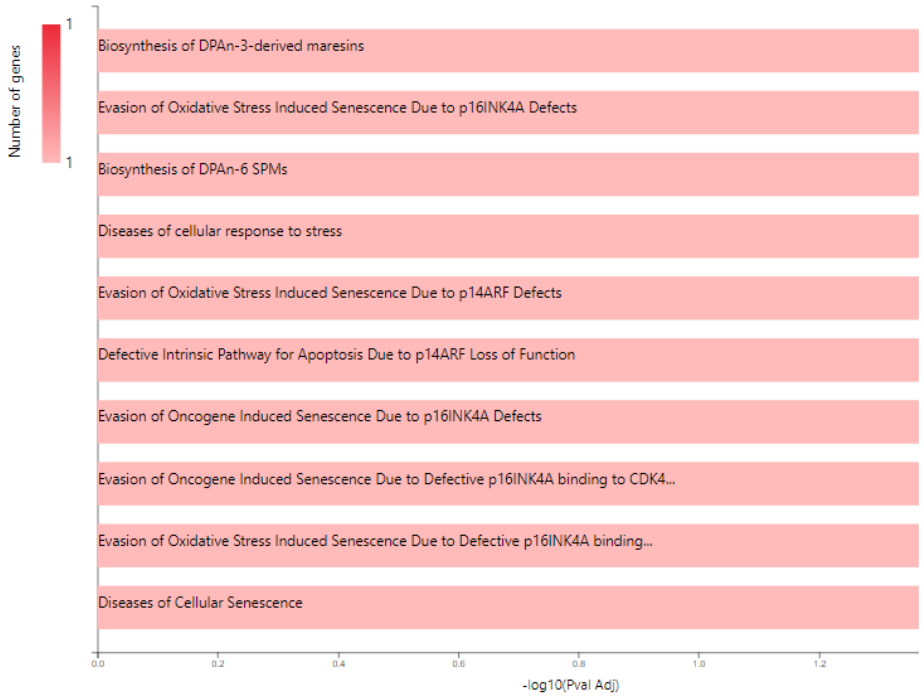
Zhao, L., Zhang, Z., Lou, H., Liang, J., Yan, X., Li, W., Ou, R., 2018. Exploration of the molecular mechanisms of cervical cancer based on mRNA expression profiles and predicted microRNA interactions. Oncol. Lett. 15 (6), 8965–8972. https://doi.org/10.3892/ol.2018.8494.

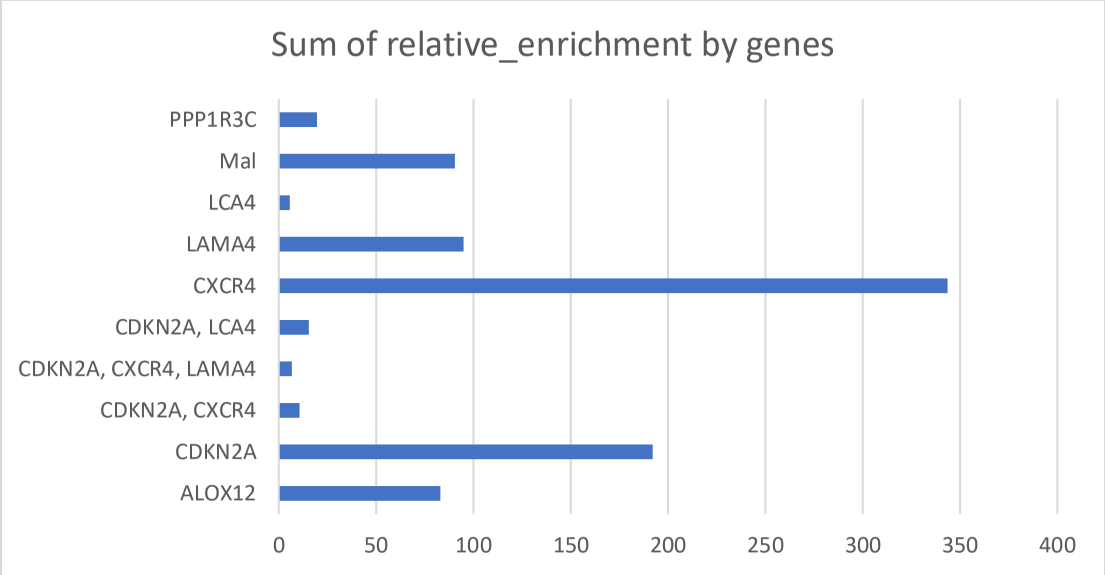
APPENDICES

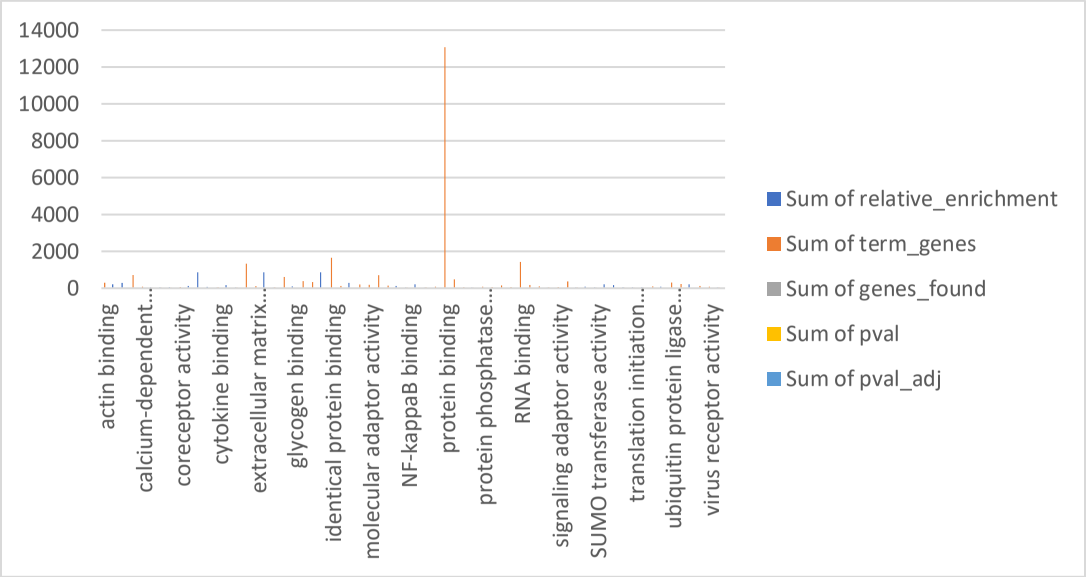
Panther pathways adopted from GeneCodis4

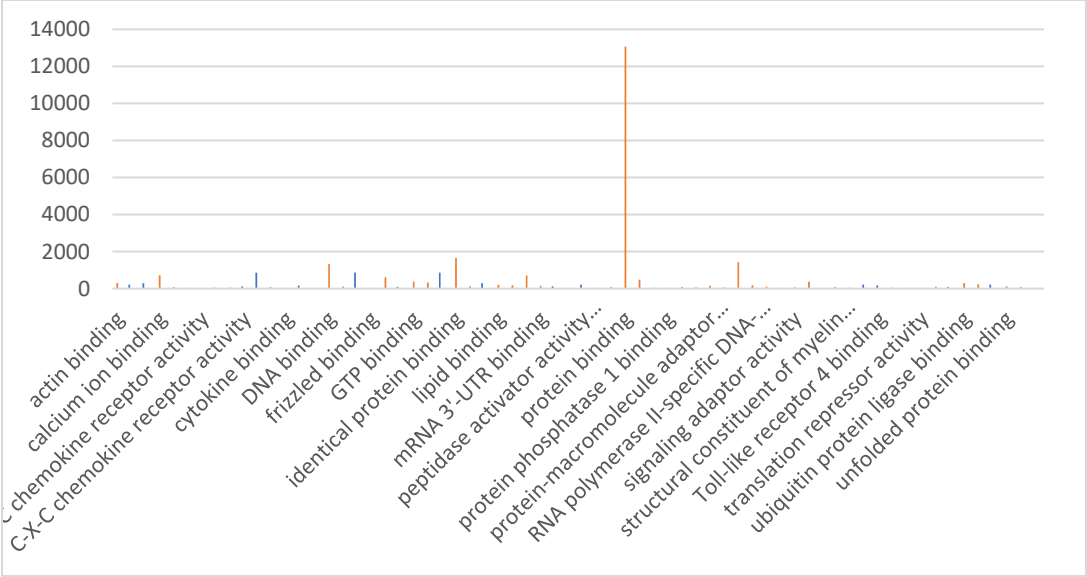


Reactome pathways adopted from GeneCodis4




Sum of relative_enrichment by genes
PPP1R3C
Mal
LCA4
LAMA4
CXCR4
CDKN2A, LCA4
CDKN2A, CXCR4, LAMA4
CDKN2A, CXCR4
CDKN2A
ALOX12
0
50
100
150
200
250
300
350
400


14000
12000
10000
8000
6000
4000
2000
0
actin binding
calcium-dependent...
coreceptor activity
cytokine binding
extracellular matrix...
glycogen binding
identical protein binding
molecular adaptor activity
NF-kappaB binding
protein binding
protein phosphatase...
RNA binding
signaling adaptor activity
SUMO transferase activity
translation initiation...
ubiquitin protein ligase...
virus receptor activity
Sum of relative_enrichment
Sum of term_genes
Sum of genes_found
Sum of pval
Sum of pval_adj



R script from GEO2R

```
# Version info: R 4.2.2, Biobase 2.58.0, GEOquery 2.66.0, limma 3.54.0
################################################################
#   Differential expression analysis with limma
library(GEOquery)
library(limma)
library(umap)

# load series and platform data from GEO

gset <- getGEO("GSE63514", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx
<- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group membership for all samples
gsms <- paste0("00000000000XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX",
        "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX",
        "11111111111XXXXXXXXXXXXXXXXX")
sml <- strsplit(gsms, split="")[[1]]
```

```
# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]

# log2 transformation
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
          (qx[6]-qx[1] > 50 && qx[2] > 0)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex) }

# assign samples to groups and set up design matrix
gs <- factor(sml)
groups <- make.names(c("Control Samples","Case Samples"))
levels(gs) <- groups
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)

gset <- gset[complete.cases(exprs(gset)), ] # skip missing values

fit <- lmFit(gset, design)  # fit linear model

# set up contrasts of interest and recalculate model coefficients
cts <- paste(groups[1], groups[2], sep="-")
cont.matrix <- makeContrasts(contrasts=cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)

# compute statistics and table of top significant genes
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)

tT                              <-                              subset(tT,
select=c("ID","adj.P.Val","P.Value","t","B","logFC","GB_ACC","SPOT_ID",
"Gene.Symbol","Gene.symbol","Gene.title"))
```

```
write.table(tT, file=stdout(), row.names=F, sep="\t")

# Visualize and quality control test results.
# Build histogram of P-values for all genes. Normal test
# assumption is that most genes are not differentially expressed.
tT2 <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
hist(tT2$adj.P.Val, col = "grey", border = "white", xlab = "P-adj",
  ylab = "Number of genes", main = "P-adj value distribution")

# summarize test results as "up", "down" or "not expressed"
dT <- decideTests(fit2, adjust.method="fdr", p.value=0.05, lfc=0)

# Venn diagram of results
vennDiagram(dT, circle.col=palette())

# create Q-Q plot for t-statistic
t.good <- which(!is.na(fit2$F)) # filter out bad probes
qqt(fit2$t[t.good], fit2$df.total[t.good], main="Moderated t statistic")

# volcano plot (log P-value vs log fold change)
colnames(fit2) # list contrast names
ct <- 1        # choose contrast of interest
# Please note that the code provided to generate graphs serves as a
guidance to
# the users. It does not replicate the exact GEO2R web display due to
multitude
# of graphical options.
#
# The following will produce basic volcano plot using limma function:
volcanoplot(fit2, coef=ct, main=colnames(fit2)[ct], pch=20,
  highlight=length(which(dT[,ct]!=0)), names=rep('+', nrow(fit2)))

# MD plot (log fold change vs mean log expression)
# highlight statistically significant (p-adj < 0.05) probes
plotMD(fit2, column=ct, status=dT[,ct], legend=F, pch=20, cex=1)
abline(h=0)

################################################################
```

```
# General expression data analysis
ex <- exprs(gset)

# box-and-whisker plot
ord <- order(gs)  # order samples by group
palette(c("#1B9E77", "#7570B3", "#E7298A", "#E6AB02", "#D95F02",
          "#66A61E", "#A6761D", "#B32424", "#B324B3", "#666666"))
par(mar=c(7,4,2,1))
title <- paste ("GSE63514", "/", annotation(gset), sep ="")
boxplot(ex[,ord], boxwex=0.6, notch=T, main=title, outline=FALSE, las=2,
col=gs[ord])
legend("topleft", groups, fill=palette(), bty="n")

# expression value distribution
par(mar=c(4,4,2,1))
title <- paste ("GSE63514", "/", annotation(gset), " value distribution",
sep ="")
plotDensities(ex, group=gs, main=title, legend ="topright")

# UMAP plot (dimensionality reduction)
ex <- na.omit(ex) # eliminate rows with NAs
ex <- ex[!duplicated(ex), ]  # remove duplicates
ump <- umap(t(ex), n_neighbors = 9, random_state = 123)
par(mar=c(3,3,2,6), xpd=TRUE)
plot(ump$layout, main="UMAP plot, nbrs=9", xlab="", ylab="", col=gs,
pch=20, cex=1.5)
legend("topright", inset=c(-0.15,0), legend=levels(gs), pch=20,
col=1:nlevels(gs), title="Group", pt.cex=1.5)
library("maptools")  # point labels without overlaps
pointLabel(ump$layout, labels = rownames(ump$layout), method="SANN",
cex=0.6)

# mean-variance trend, helps to see if precision weights are needed
plotSA(fit2, main="Mean variance trend, GSE63514")
```