doi.org/10.36721/PJPS.2023.36.1.SP.301-306.1

Bioinformatics analyses of microarray data reveal novel biomarker associated with cutaneous squamous cell carcinoma

Ayesha Rashid[1], Raza Rasool[2], Muhammad Zeshan Zaib[3], Muhammad Hamza Khan[4], Saeedah Musaed Almutairi[5], Dina S. Hussein[6], Amna Mumtaz[7], Zain Ul Abedien[8] and Ihteshamul Haq[9]*

[1]Basic Health Unit Phullan Toli, Depalpur, Okara, Pakistan
[2]Department of Surgery, DHQ Hospital, Khanewal, Pakistan
[3]Department of Medicine, DHQ Hospital, Khanewal, Pakistan
[4]Peshawar Medical College (Riphah international university Islamabad)
[5]Department of Botany and Microbiology, College of Science, King Saud University, Riyadh, Saudi Arabia
[6]Department of Chemistry, College of Sciences and Health, Cleveland State University, Cleveland, USA
[7]Department of Clinical Laboratory Sciences, Women University, Swabi, Pakistan
[8]Institute of Microbiology, University of Agriculture, Faisalabad, Pakistan
[9]Department of Biotechnology and Genetic Engineering Hazara University, Mansehra, Pakistan

Abstract: Global incidence of cutaneous squamous cell carcinoma is rising. This study investigated the molecular mechanism of CSCC and screened genes that could serve as biomarkers, providing a theoretical framework for pathogenesis and drug development research. We examined differentially expressed genes (DEGs) between normal tissue specimens and CSCC patient tissues using microarray data analysis.. Also, signal pathway and functional enrichment analysis were employed for target genes to rule out the specific genes that display close association with CSCC with the potential to serve as unique CSCC biomarkers. Two datasets GSE66359 and GSE45216 were used in the differential expression analysis. Results revealed the upregulation of potential candidate genes like P13, MMP1, A100A12, and KRT16 while downregulation of rf132, EGR3, PAMR1, LRP4, and KRT2. Verifying these genes demonstrated that KRT16 and EGR3 had obvious differential expression patterns. Further analyses showed that EGR3 was the only gene that exhibited a similar differential pattern in CSCC. Thus, we infer that EGR3 is closely related to the occurrence and development of CSCC, and it has the potential to function as the candidate biomarker gene against CSCC and facilitate subsequent diagnosis and treatment in clinical management.

Keywords: Cutaneous squamous cell carcinoma, DEGs, microarray data, bioinformatical analysis, biomarker, EGR3.

INTRODUCTION

Cutaneous squamous cell carcinoma (CSCC) is one of the most common human cancers in the world, second to no melanoma cancer (Parekh and Seykora, 2017). Recently, a growing number of people are being affected by SCC including patients with white skin and suppressed immune systems. Multiple factors are known to be associated with cutaneous squamous cell carcinoma. In addition, the clinical and pathological diversity of CSCC makes it complicated due to escape or misdiagnosis (Wei *et al.,* 2018). At present, the most effective means to enhance the survival rate of chronic skin diseases include early diagnosis/prediction and timely treatment thus preventing disease deterioration. This disease affects middle-aged and elderly people more than younger ones. CSCC biomarker research is scarce, hence it's important to find them to improve clinical care. In clinical research, DNA microarray data is a commonly used approach because it can be used in clinical research since it can screen the expression levels of thousands of genes simultaneously. Besides, it has become a novel and widely used tool to identify and compare the DEGs between normal and cancer samples (Zhang *et al.,* 2020). Moreover, this approach offers a complete, systematic, and reliable comparison to analyze gene expression among different types of specimens (Fang *et al.,* 2022; Zou *et al.,* 2021). Hence, the investigation of DEGs provides clues and evidence associated with distinct signaling pathways and biological processes involved in tumor development. Lately, several studies have reported that analysis of DEGs can provide efficient and rapid diagnosis using DNA microarray data, as in the case of prostate cancer (Chen *et al.,* 2013) and human gliomas (Chen *et al*., 2013), etc. The specific objective of this study was to investigate the molecular mechanism underlying CSCC and ultimately provide a theoretical framework to facilitate the research on disease diagnostics medicine development.

In the present study, we conclude that the EGR3 gene displays a close association with the occurrence of CSCC and it can function as a potential biomarker of CSCC. Thus, monitoring of EGR3 expressions would be critical to clarify this disease's clinical and histopathological

**Corresponding author:* e-mail: ihteshamulhaq384@gmail.com

diversity. Consequently, it will lead to improved diagnostic and prognostic approaches for the clinical management of CSCC.

MATERIALS AND METHODS

Microarray data collection

Microarray information has been downloaded from a website of GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/). For the study, two distinct group data sets have been selected with the accession number GSE45216 (Lambert *et al*., 2014) and GSE45164 (Brooks *et al.,* 2014). In the GSE45216 dataset, 30 tumor samples were present and in the GSE42677 data collection, 15 samples from the same history and test platform were chosen. In the screening of essential genes GSE45164 and GSE45216 were used. There were five natural keratinocyte cells in GSE45164 and eight cutaneous squamous cell carcinoma cells, as well as three normal specimens and 10 tumor samples, were reported. Affymetrix human genome (HG) U133A 2.0 Array and Affymetrix HG U133 Plus2.0, both test platforms were used. The framework was Affymetrix Human Genome U133 Plus 2-Array used to monitor gene expression from accession number GSE66359.

Data quality control and preliminary analysis

For the study of data quality based on the linear model at the microarray level AffyPLM package (Bolstad *et al.,* 2013) was used. Box RLE (Relative Log Expression) and NUSE (Normalized Unscaled Standard Errors) figures were painted to measure the pattern according to the testing data. Furthermore, the AffyRNAdeg feature checked the degradation situation of the RNA. Finally, the downstream study included selected high-quality RNA datasets with the same tendency.

To ensure the integrity and comparability of the data collection, Gcrma package (Wu *et al.,* 2012) was used to normalize and context correction of microarray data. An error was found in the micro-array and among the genes which were checked more than once, and an average value was determined. A significant indicator for evaluating experimental reliability and rationality for sample selection was the analysis of the gene expression level among samples. The global and primary analysis of components was then performed in the test samples and correlation and distribution data were measured with the Pearson correlation coefficient.

Identification of DEG

In identifying differentially expressed genes between control group samples and processed samples, Limma packages (Ritchie *et al.,* 2015) were used. Finally, various genes with log 2 values greater than 1 were screened, with a p-value less than 0.05.

Pathways and functional enrichment analysis of DEGs

These differentially expressed genes had pathways and functional enrichment. DAVID functional annotation and GO/KEGG enrichment analysis were employed. The database also annotated functions. DAVID measured the P-value before and after (benjamini correction or FDR adjustment) and set the threshold at 0.05.

Validation of the study

Microarray data regarding control and skin cancer was taken from (Nindl *et al*., 2006) using 3 different biopsies including 5 skin tissues, 5 immunosuppressed organ transplanted audiences, and five SCC invasive were determined. The phylogenetic analysis was made to determine the gene expression profile of the selected genes under normal skin cells and SCC invasiveness. Furthermore, the study was validated with thirteen genes using qPCR to determine the expression profiling of control and SCC genes.

STATISTICAL ANALYSIS

DAVID software (v3.10.0) was used for functional enhancement analysis on these genes.

RESULTS

Data normalization

After consistency screening, the Gcrma package was used to standardize samples. The results obtained after data normalization are given in fig. 1. The product of the density of the expression curve and the box plot showed the expression of 0 to 15 in the two groups; with minor variation that matched the actual situation. GSE45216 and GSE66359 expression values were adjusted at 3 following normalizations. The two sets of microarray datasets displayed identical patterns of expression.

The uniform manifold approximation and projection (UMAP) analysis

There were two groups found from UMAP analysis in the dataset and 5 groups were made in the GSE45216 data set (fig. 3). In addition, the candidate genes were tested for reliability in the data set of GSE66359 (Supplementary 1). EGR3, KRT16, and PI3 are key genes for cutaneous squamous cell carcinoma, yet they have opposite patents for expression in GSE42677 and GSE45164 (Wei *et al*., 2018).

Gene set enrichment analysis

After normalization, the Limma package conducted differentially expressed gene (DEGs) identification of the microarray results. Comparing the 30 tumor tissues and 10 normal tissues in the data set GSE45216. The GSE45216 data set was categorized into poorly differentiated vs unknown, moderately differentiated vs moderately poorly differentiated (103), well-differentiated

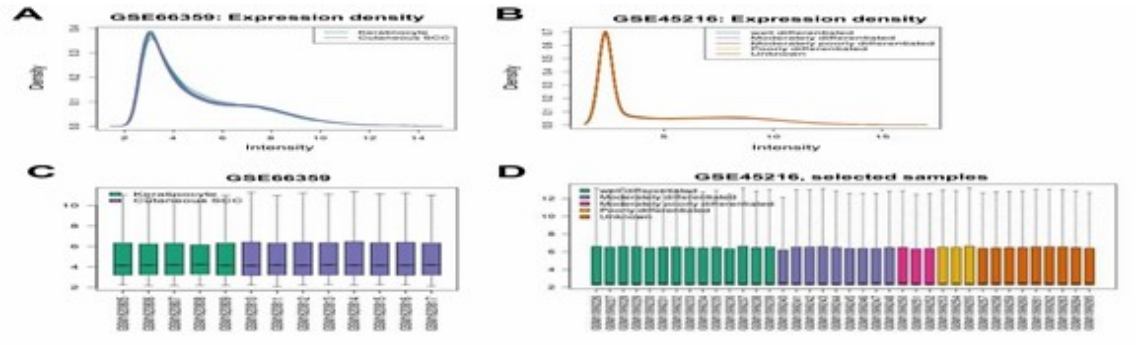

Fig. 1: Distribution figure after normalization of two data set. figs. 1A and 1B represent the expression density after data normalization, respectively. However, Figures 1C and 1D represent the box chart (gene expression) after the normalization of GSE45216 and GSE66359 respectively.

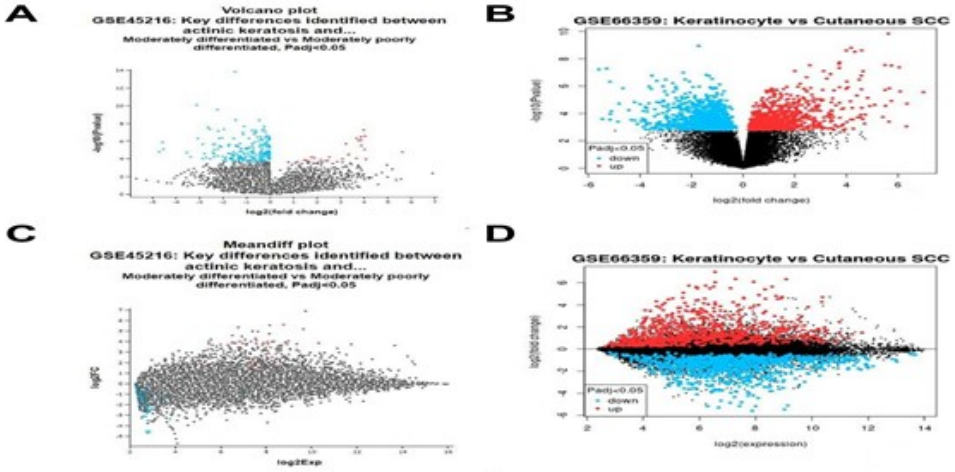

Fig. 2: Key differences were identified between actinic keratosis and Cutaneous SCC samples; fig. 2 a &c represents the volcano plot and mean difference plot in the GSE45216 data set while fig. 2b & d represents the volcano plot and mean difference plot respectively for the GSE66359 data set.

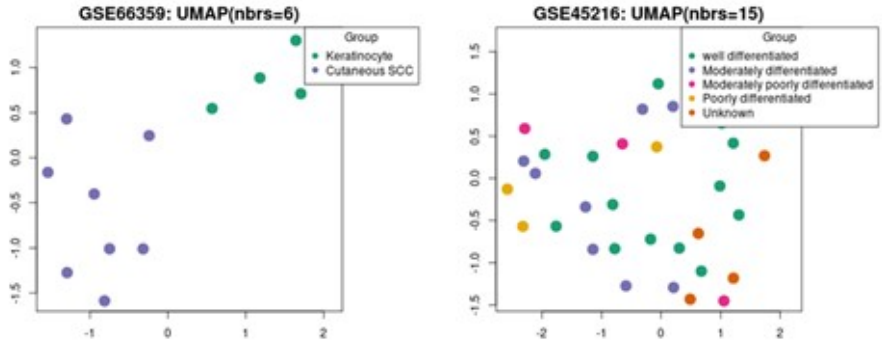

Fig. 1: Uniform manifold approximation and projection (UMAP) of the GSE66359 and GSE45216 data respectively

vs moderately differentiated (158), and unknown vs well differentiated (68) (fig. 4). Furthermore, the other data set recorded 1828 genes keratinocytes vs cutaneous SCC. GSE45216 and GSE45164 had 833 common genes, with 680 genes with the same tendency, 465 genes up-regulated and 215 genes down-regulated (fig. 2).

Functional enrichment analysis

There were 250 top genes used for the analysis of functional enrichment analysis screened from NCBI (https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE66359). In the GO functional enrichment analysis, 202 were mainly enriched. The annotation summary results were categorized into disease, functional categories, gene ontology, and pathways. The disease section includes (OMIM disease. The number of biological processes to be enriched was as high as 130 and molecular functions were to be enriched by a minimum of 35 molecular functions. The functional categories include COG_ontology, up keywords, and UP_SEQ_feature. The UP keywords

category includes threonine protease, proteasome, hydrolase, host-virus interaction, immunity, acetylation, protease, cytoplasm, cell cycle, innate immunity, SH2 domain, antiviral defense, tumor suppressor, protease inhibitor, Intermediate filament, nucleus, golgi apparatus, GTP-binding, a serine protease, zymogen, chromatin regulator, ichthyosis, DNA replication, alternative splicing, phosphoprotein, tyrosine-protein kinase andMHC_1. The 20 UP_seq_feature was recorded during the DAVID analysis. It includes SH2, mutagenesis site, UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2, N-acetyllactosaminide beta-1,3-N- acetylglucosaminyltransferase, Coil 2, linker 12, Coil 1B, linker 1, Coil 1A, Rod, head, nucleophile, charge relay system, proton acceptor, Cys-rich, MCM, EGF-like; calcium -binding, nucleotide phosphate-binding region: GTP, hazel-like 2 and Peptidase S1. The top three functional enrichment (GOTERM_BP_DIRECT; GOTERM_CC_DIRECT; GOTERM_MF_DIRECT) GO terms include different biological pathways such as proteasome core complex, nucleoplasm, proteasome complex, extracellular exosome, cytoplasm, cytosol, chromatin, Golgi apparatus, epidermal lamellar body, Golgi membrane, nuclear chromosome, cytoplasmic mRNA processing body, sarcolemma, intracellular membrane-bounded organelle, proteasome core complex, alpha-subunit complex, cornified envelope, MCM complex, and membrane. Furthermore, the 6 important KEGG pathway include proteasome, cell cycle, pertussis, DNA replication, Parkinson's disease, and non-alcoholic fatty liver disease. The most important protein domains INTERPRO; PIR_SUPERFAMILY and SMART were also reported (fig. 5).

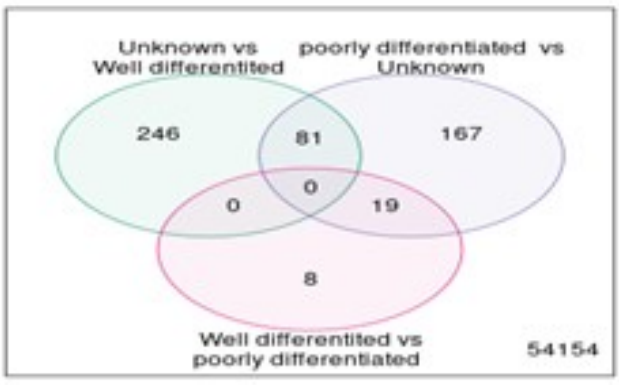

Fig. 2: Differential expression gene analysis from GSE45216 data set expressed in volcano fig.

The identification of biomarkers for CSCC by *in vitro* from GSE45216 and were verified from GSE66359, avoidance of background overlapping, and the reliability of the findings. The most important genes were up-regulated i.e., KRT16, MMP1, S100A12, and P13 while downregulated genes KRT2, EGR3, rf132, PAMR1, and LRP4 were recognized as potential genes or biomarkers to validate it for future use (Supplementary 2). Moreover, it was also used to determine cutaneous squamous cell carcinoma (CSCC). Further detailed analysis results suggested that EGR3 and KRT16 were found to be more enriched while EGR3 has a similar expression pattern from the GSE45164 data set. Here, we can infer that EGR3 is more closely related to the occurrence and development of CSCC and functions as the biomarker against cutaneous squamous cell carcinoma.

DISCUSSION

Epidermoid or squamous cell tumors are frequent in appendages or epidermal cells (Losquadro, 2017). Keratosis varies in tumor cells. CSCC is more likely to harm squamous epithelium-covered organs such skin, lips, mouth, vagina, cervix, and esophagus. Squamous metaplasia can also occur in the bladder, renal pelvis, and bronchus without squamous epithelium. Recently, the SCC incidence has dramatically risen on a global scale (Arnold *et al.,* 2015). Thus, to avoid tumor incidence and development, early detection, and effective treatment are the most efficient ways for disease management.

Little has been known regarding the biomarker-related research on SCC which imposes an imperative need for SCC biomarker identification to further facilitate the early prediction and prognostic treatment of malignant tumors in clinical research. In the present study, a total of 805 genes exhibiting co-expression were identified and further evaluation of differential gene expression revealed a total of 250 top differentially expressed genes in two datasets, i.e., GSE66359 and GSE45216. Consequently, a total of 521 genes were observed to show differential expression, including 292 upregulated and 229 downregulated genes. Additionally, the verification of the potential candidate gene was done in 4 groups, making the results more reliable. The differentially expressed analysis was performed using a group comprising 40 samples in total: 30 tumor samples in dataset GSE45216 and 10 normal samples in dataset GSE66359. Consequently, 675 genes in total were observed to have similar expression patterns, including 463 upregulated and 212 downregulated genes.

Next, the candidate biomarker genes were verified in the 13 genes from the microarray dataset and were checked to enhance the reliability of the results (Nindl *et al.,* 2006). It was noticed that three genes, P13, KRT16, and EGR3 displayed a distinct differential expression pattern in CSCC; EGR3 was the only biomarker candidate gene because P13 and KRT16, the most differentially expressed genes in cutaneous squamous cell carcinoma, expressed oppositely.

Early growth response-3 (EGR3) is associated with the EGR gene family along with 3 other members i.e., EGR1,

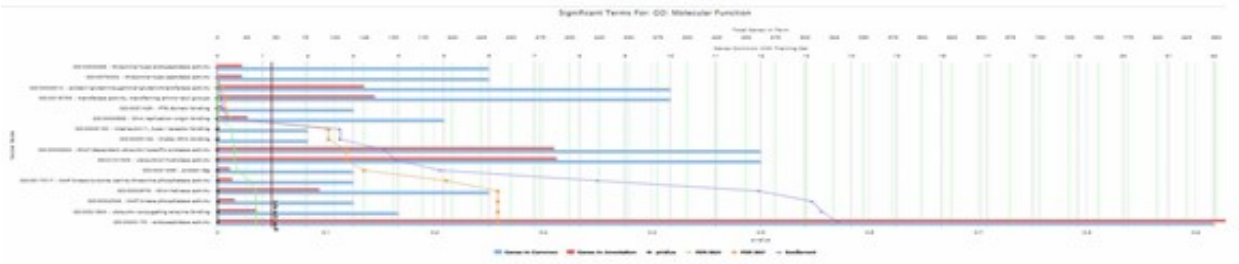

Fig. 5: GO and KEGG pathway for top differentially expressed terms by GO enrichment analysis.

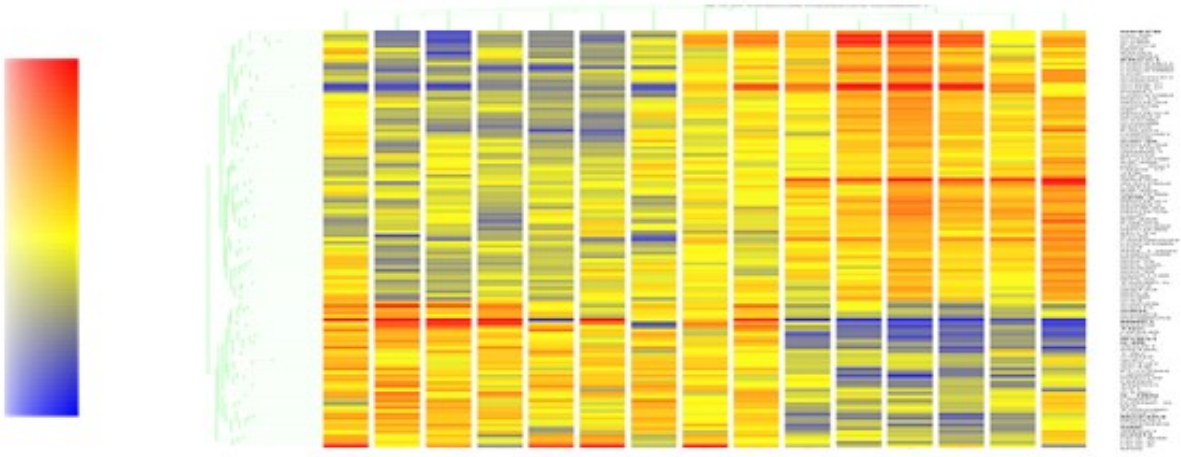

Fig. 3: Expression analysis of 118 Genes using Microarray prediction analysis under normal vs SCC patients' samples. The 13 bold SCC genes were verified using qPCR. normalized gene expression of each gene in each group to expression levels from dark blue (minimum) to dark red (maximum).

EGR2, and EGR4. The EGR family encoded proteins contain a highly conserved zinc finger (DNA-binding) domain; enabling them to bind with other proteins depending on the zinc finger reactive domain. EGR3-encoded proteins are Cys2His2-related zinc finger transcription factors that respond to Kary mitosis-driven early growth (Chen *et al.,* 2012; Kothinti *et al.,* 2010). Early studies found that EGR2 and EGR3 genes help breast adipose fibroblasts (TNF) stimulate MAPK and NFB pathways (To *et al.,* 2013). It exhibits critical functions in the fibrotic responses and up-regulates during scleroderma disease (Artlett, 2018). Oral squamous cell carcinoma tissues and several cell lines were downregulated or missing EGR family genes. Oral squamous cell carcinoma tissues and several cell lines lacked EGR family genes. EGR3 gene expression and its role in CSCC remain unclear.

Moreover, a study reports that the EGR3 expression pattern has a crucial function during the differentiation, proliferation, progression, and metastasis of gastric tumor cells (Wang *et al.,* 2018). While another study describes that EGR3 functions as an intracellular mediator of estrogen signaling mechanism in breast cancer. Bother studies provide vital evidence on the association of EGR3 with cancer (Li *et al.,* 2020). In the current study, we found that EGR3 was a highly enriched gene and exhibited a similar expression pattern in both datasets i.e., GSE66359 and GSE45216. In contrast to other methods, the present study identified 3 candidate biomarker genes (P13, KRT16, and EGR3), and further validation revealed that only EGR3 was the most enriched gene and had a similar expression pattern in the analyzed datasets. Additionally, several previous studies report the well-known association of EGR family genes with SCC (Padam *et al.,* 2022). Thus, considering previous reports, the results of our study encourage us to infer that EGR3 is a novel CSCC-related biomarker gene with featured functions that are vital for early diagnosis and efficient treatment of cancer in clinical disease management.

In vitro validation of SCC genes

Microarray data regarding control and skin cancer was taken from (Nindl *et al.,* 2006) using 3 different biopsies including 5 skin tissues, 5 immunosuppressed organ transplanted audiences, and five SCC invasive were determined. The phylogenetic analysis was made to determine the gene expression profile of the selected genes under normal skin cells and SCC invasiveness. Furthermore, the study was validated with thirteen genes using qPCR to determine the expression profiling of control and SCC genes. The results suggested that the genes i.e., MMP1, TNC, GRN, RAB31, IL-1RN, NMI, IL-4R, and CDH1 were upregulated while four genes (NKEFB, ERCC1, CGI-39, APR3) were down-regulated under normal and SCC invasiveness study (fig. 6).

CONCLUSION

In conclusion, this study provides a theoretical framework for understanding the pathogenesis of CSCC and developing new drugs for its treatment. The identification of EGR3 as a potential biomarker gene for CSCC is a significant finding that may aid in early diagnosis and effective treatment of CSCC.

ACKNOWLEDGMENT

The authors extend their appreciation to the Researchers supporting project number (RSP2023R470) at King Saud University, Riyadh, Saudi Arabia.

REFERENCES

Arnold M, Soerjomataram I, Ferlay J and Forman D (2015). Global incidence of oesophageal cancer by histological subtype in 2012. *Gut,* **64**(3): 381-387.

Artlett CM (2018). The IL-1 family of cytokines. Do they have a role in scleroderma fibrosis? *Immunol Lett*, **195** (1): 30-37.

Bolstad B, Bolstad MB, BiocGenerics I, bioc Views Microarray O and Preprocessing Q (2013). Package 'affyPLM', pp.185-193.

Chen J, Wang Y, Shen B and Zhang D (2013). Molecular signature of cancer at gene level or pathway level? Case studies of colorectal cancer and prostate cancer microarray data. *Comp. Math. Methods Med*, **2013**: 909525.

Chen Y, Bates DL, Dey R, Chen P-H, Machado ACD, Laird-Offringa IA, Rohs R and Chen L (2012). DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Reports*, **2**(5): 1197-1206.

Fang S, Wu Y, Zhang H, Zeng Q, Wang P, Zhang L, Yan G, Zhang G and Wang X (2022). Molecular characterization of gene expression changes in murine cutaneous squamous cell carcinoma after 5-aminolevulinic acid photodynamic therapy. *Photodiagn Photodyn Ther*, **39**(3): 102907.

Kothinti R, Blodgett A, Tabatabai NM and Petering DH (2010). Zinc finger transcription factor Zn3-Sp1 reactions with Cd^{2+}. *Chem Res Toxicol*, **23**(2): 405-412.

Li J, Cao H, Feng H, Xue Q, Zhang A and Fu J (2020). Evaluation of the estrogenic/antiestrogenic activities of perfluoroalkyl substances and their interactions with the human estrogen receptor by combining *in vitro* assays and in silico modeling. *Environ. Sci. Technol.*, **54**(22): 14514-14524.

Losquadro WD (2017). Anatomy of the skin and the pathogenesis of nonmelanoma skin cancer. *Facial Plast. Surg. Clin. North Am.,* **25**(3): 283-289.

Nindl I, Dang C, Forschner T, Kuban RJ, Meyer T, Sterry W and Stockfleth E (2006). Identification of differentially expressed genes in cutaneous squamous cell carcinoma by microarray expression profiling. *Mol. Cancer*, **5**(1): 1-17.

Padam KSR, Morgan R, Hunter K, Chakrabarty S, Kumar NA and Radhakrishnan R (2022). Identification of HOX signatures contributing to the oral cancer phenotype. *Sci. Rep.,* **12**(1): 1-12.

Parekh V and Seykora JT (2017). Cutaneous squamous cell carcinoma. *Clinics Lab. Med.*, **37** (3): 503-525.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nuc. Acids Res.*, **43**(7): e47-e47.

To SQ, Knower KC and Clyne CD (2013). NFκB and MAPK signaling pathways mediate TNFα-induced Early Growth Response gene transcription leading to aromatase expression. *Biochem. Biophys. Res. Comm.*, **433**(1): 96-101.

Wang N, Tan HY, Feng YG, Zhang C, Chen F, and Feng Y (2018). Micro RNA-23a in human cancer: Its roles, mechanisms and therapeutic relevance. *Cancers*, **11**(1): 7.

Wei W, Chen Y, Xu J, Zhou Y, Bai X, Yang M and Zhu J (2018). Identification of biomarker for cutaneous squamous cell carcinoma using microarray data analysis. *J. Cancer*, **9**(2): 400.

Wu J, Irizarry R, MacDonald J and Gentry J (2012). Gcrma: Background adjustment using sequence information. *R Package Ver,* **2200**: 3-10.

Zhang Z, Lin E, Zhuang H, Xie L, Feng X, Liu J and Yu Y (2020). Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma. *Cancer Cell Int.*, **20**(1): 1-18.

Zou DD, Xu D, Deng YY, Wu WJ, Zhang J, Huang L and He L (2021). Identification of key genes in cutaneous squamous cell carcinoma: a transcriptome sequencing and bioinformatics profiling study. *Ann. Trans. Medi*, **9**(19): 1497.