Cancer Treatment and Research Communications 27 (2021) 100320

Contents lists available at ScienceDirect

Cancer Treatment and Research Communications

journal homepage: www.sciencedirect.com/journal/cancer-treatment-and-research-communications

Exploring the molecular subclasses and stage-specific genes of oral cancer: A bioinformatics analysis

Abdul Raheem Fathima Shafana [a,d,*], Gatamanna Arachchige Isuri Uwanthika [b,d], Thangathurai Kartheeswaran [c,d]

[a] Department of Information and Communication Technology, Faculty of Technology, South Eastern University of Sri Lanka, Oluvil, Sri Lanka
[b] Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka
[c] Department of Physical Science, Vavuniya Campus of the University of Jaffna, Vavuniya, Sri Lanka
[d] Department of Computer Science and Informatics, Faculty of Applied Sciences, Uva Wellassa University of Sri Lanka, Badulla, Sri Lanka

ARTICLE INFO

Keywords:
Oral Squamous Cell Carcinoma
Microarray
Gene co-expression network
Betel quid chewing
molecular subclasses

ABSTRACT

The rate of people getting affected by Oral cancer in Sri Lanka is growing rapidly since the root cause of such cancer, betel quid chewing is tightly coupled with the tradition of the country. The five-year survival rate of the disease is also pretty low as it is typically detected at advanced stages. This urges a comprehensive study on the marker genes of oral cancer for the successful therapeutic revisions that would potentially identify cancer in its early stages. Further, the identification of molecular subclasses can assist in individualizing the treatment for this type of fatal disease. This study uses the bioinformatics analysis on the gene expression dataset of 56 oral cancer patients from Sri Lanka and the United Kingdom to identify the differentially expressed genes where these genes are later clustered and classified into molecular subclasses. Molecular subclasses are found by clustering the genes that stratify together and the stages were identified with the use of gene co-expression networks. Five molecular subclasses of oral cancer were identified and the genes associated with each tumour stage. Out of the genes that are clustered and classified, TAGLN2, CCND2 and CCL8 were well-known tumour suppressor genes and GPX3, GRN and ITGB4 genes are involved in several carcinomas. Putative marker genes of Oral Squamous Cell Carcinoma were identified which could facilitate the medical practitioner in the early detection of oral cancer and also in the improvement of treatment methods.

1. Introduction

Cancer is one of the leading genetic diseases across the globe that results from both inherited and acquired changes in DNA mostly. In particular, Oral cancer occupies a prominent place in most common cancers in Sri Lanka in the years 2001–2008 [4]) as it is one of the commonest cancers amongst men. Further, deaths from Oral cancer accounts for nearly 12.8% of all cancers in the country. Oral squamous cell carcinoma [15]. National Cancer Control Programme of Sri Lanka has ranked oral cancer as the leading cause of death as the statistics proved that oral cancer possesses the highest crude rate of 17.0 per 100, 000 populations [15]. In Sri Lanka, the use of betel quid and smoking are considered in accounting for inclining the rate of oral cancer [2]. Despite the considerable advancements in the medical field, yet, the five-year survival rate (62%) of oral cancer is amongst the lowest of all the major cancers in humans [5].

The prime cause of this lower survival rate is that nearly 90% of the oral pharyngeal cancers are diagnosed only at the advanced stages. Further, The National Cancer Control Programme (NCCP) of Sri Lanka targeted to lower the effect of oral cancer by 15% at least by the end of the next decade. They also believed that this could be achieved mainly through the primary prevention and early detection of oral cancer. On the other hand, the therapeutic strategies of Oral Squamous Cell Carcinoma (OSCC) are also to be revised through the identification of potential molecular subclasses in oral cancer which can individualize the treatment to the OSCC patients.

The DNA microarray encompasses most of the human genome transcript. This has been a promising technology over the years [10] for successful prognosis and the unveiling of potential molecular subclasses of cancers. The higher dimensionality of such data to reveal the outcome related information has demanded the use of computational methods over manual interpretation.

* Corresponding author.
E-mail address: arfshafana@seu.ac.lk (A.R.F. Shafana).

https://doi.org/10.1016/j.ctarc.2021.100320

Available online 27 January 2021

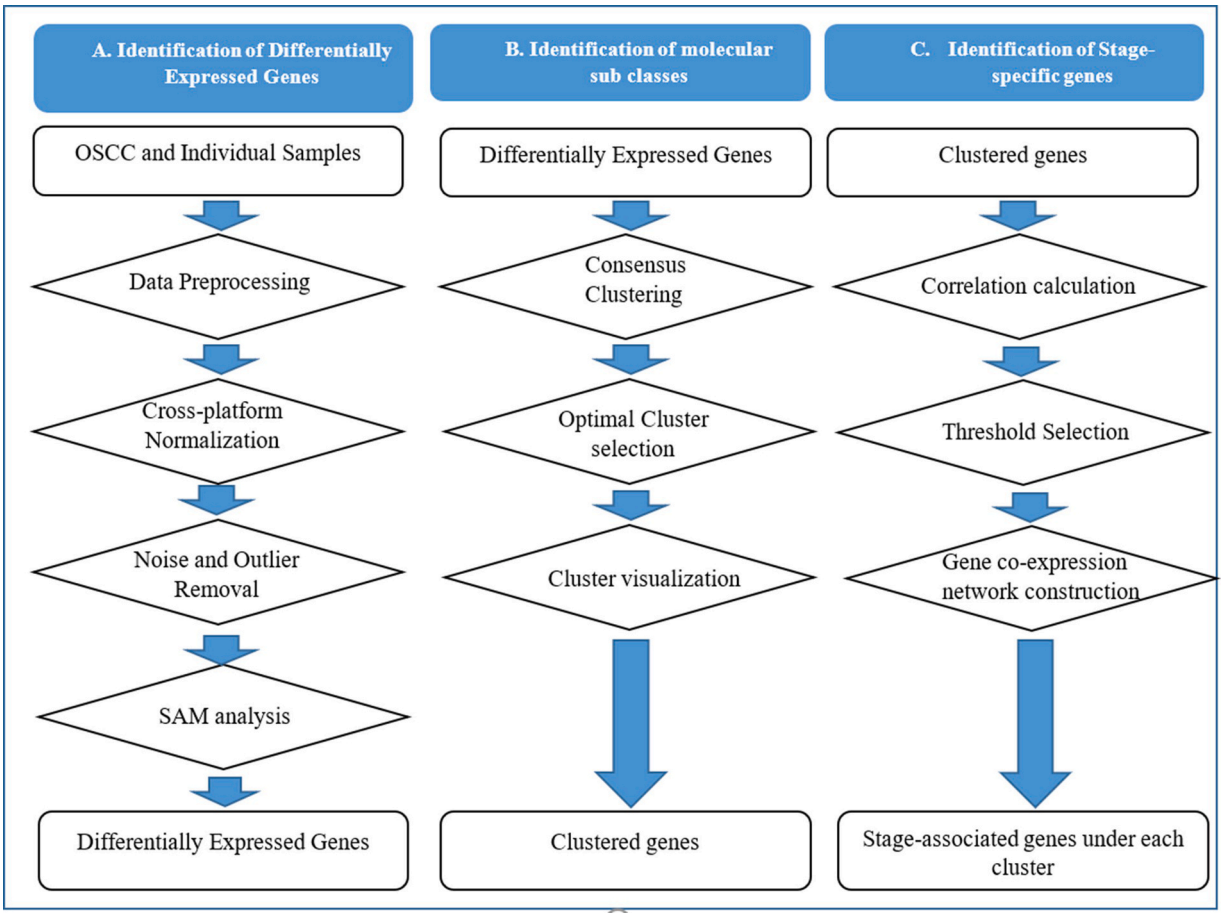


Fig. 1. Schematic representation of the applied methodology.

The present high-throughput technologies have created vast potentials for diagnostics and therapies for many cancers. Researchers were successful in using integrated network analysis and logistic regression to reveal the stage-specific genes of OSCC [18]. Further, Gene co-expression networks have been playing a vital role in many of the bioinformatics analyses of microarray gene expression data especially in revealing the association of gene expression with phenotypic traits.

There have been consistent researches that were aimed to gain insight into the subclasses of cancers as well. A meta-analysis of gene expression data has discovered six subclasses of Head and Neck Cancer Squamous Cell Carcinoma (HNSCC) that lead to new therapeutic approaches [3].

Several studies have been undertaken to critically study the differentially expressed genes of oral cancer. However, much effort has not been pledged to identify the marker genes of OSCC that have undergone both classifications based on their molecular subclasses and their tumour stages. Thus, this study focuses to develop such a methodology. This research study focuses to contribute to the medical community by lowering the expenditures for laboratory scanning and a series of medical tests to a certain extent. Further, the research has also aided in the identification of OSCC at an early curable stage by clearly distinguishing the genes between the stages. Thus, this methodology presented here is a good initiative in such implementation especially in the context of Sri Lanka.

Table 1
Details of the dataset used.

Details	Dataset
Identifier	GSE51010
Initial Number of samples	56 (tumour: 48 + Normal: 8)
Samples left after preprocessing	56 (tumour: 48 + Normal: 8)
Affymetrix® Platform (Normal & Tumour)	Affymetrix® Human Genome U133 Plus2.0 Array & Affymetrix® Human Genome Focus Array

2. Methods and materials

The gene expression data have been analysed using R programming tools (http://www.R-project.org) and Bioconductor packages [9] to

Table 2
Patients and tumour related factors in Sri Lankan cohort of oral cancers [19].

Study No	Sex	Age	Site	Pathological Staging	Differentiation	Early Recurrence	Invasion	Smoking	Heavy Alcohol Consumption	BALT	Lymphocyte Infiltration
OCS 001 C	M	59	Buccal	T4N2bMx	Moderate	Yes	No	Smoker	No	None	Moderate
OCS 003 C	F	72	Alveolus	T4N2bMx	Poor	No	No	None	No	Oral Snuff	Dense
OCS 004 C	F	67	Alveolus	T4N0Mx	Moderate	Yes	Yes	None	No	BALT	Moderate
OCS 006 C	F	53	Tongue	T1N0Mx	Moderate	No	No	None	No	None	Dense
OCS 007 C	F	67	Palate	T4N0Mx	Well/Moderate	No	No	Smoker	No	None	Dense
OCS 008 C	F	67	Palate	T4N0Mx	Well/Moderate	No	No	Smoker	No	None	Dense
OCS 011 C	F	65	Alveolus	T4N0Mx	Moderate	Yes	No	None	No	None	Moderate
OCS 012 C	F	49	Tongue	T1N0Mx	Well/Moderate	No	No	Smoker	No	None	Moderate
OCS 013 C	F	72	Tongue	T4N2aMx	Poor	Yes	No	Smoker	Yes	None	Moderate
OCS 014 C	M	43	FOM	T1N0Mx	Moderate	No	No	Smoker	No	None	Moderate
OCS 015 C	M	46	Tongue	T4N2bMx	Poor/Moderate	No	Yes	Smoker	Yes	None	Moderate
OCS 016 C	M	51	Tongue	T4N0Mx	Moderate	Yes	Yes	None	No	None	Moderate
OCS 020 C	F	73	Alveolus	T2N0Mx	Well	No	No	None	No	Oral Snuff	Moderate
OCS 022 C	M	58	FOM	T2N0Mx	Poor/Moderate	No	No	Smoker	Yes	None	Moderate
OCS 024 C	M	70	Alveolus	T2N2bMx	Moderate	No	No	Smoker	No	None	Dense
OCS 025 C	F	68	Alveolus	T2N1Mx	Moderate	Yes	Yes	Smoker	No	None	Moderate
OCS 026 C	F	79	Buccal	T2N0Mx	Well	No	No	None	No	None	NA
OCS 027 C	M	37	FOM	T1N0Mx	Moderate	No	No	Smoker	Yes	None	Mild
OCS 029 C	M	66	FOM	T2N0Mx	Moderate	No	No	Smoker	No	None	Moderate
OCS 031 C	M	67	FOM	T4N0Mx	Moderate	Yes	No	Smoker	Yes	None	Mild
OCS 032 C	M	60	Retromolar	T2N0Mx	Moderate	No	Yes	Smoker	No	None	Moderate

identify the genes associated with the molecular subclasses of oral cancer and their respective stages. The association of Bioconductor with R statistical environment, provide a better companion for the analysis and comprehension of high-throughput genomic data. The schematic representation of the methodology applied is provided in Fig. 1. The detailed explanation of each of the steps are as follows:

2.1. Data collection and normalization

Gene Expression Profiles of OSCC concerning Sri Lankan patients were specifically obtained from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. The samples were obtained from a study [19] comprising both case and healthy control samples to identify disease-specific signals.

The dataset comprised 56 samples of which 48 were oral cancer patients and eight were healthy individuals as shown in Table 1. This data has been selected in particular, as it possessed the samples from Sri Lanka and phenotypic traits were also well defined. Out of the 48 patients, Sri Lankans were 21. Out of eight healthy individuals, Sri Lankans were three. Clinical information of the patients from Sri Lanka and the UK are provided in Table 2 and Table 3 respectively. The array datasets were of HGU133 plus 2.0 and HGU Focus Array that were from the widely used Affymetrix platforms. The studies comprised of well-defined phenotypic descriptions of cancer stages.

Bioconductor packages such as GEOquery, Biobase, affy were used to retrieve and process the data in the working environment. The researchers have obtained [19], appropriate ethical approval from the respective authorities (South Birmingham Research Ethics Committee 0769, UK and Kandy General Hospital and University of Peradeniya Ethical Committee, Sri Lanka) for English and Sri Lankan samples.

The datasets were imported and processed into R (version 3.2.5) using Bioconductor packages such as GEOquery tools and affy package

Table 3
Patients and tumour related factors in the UK cohort of oral cancers [19].

Study No	Sex	Age	Site	Pathological Staging	Differentiation	Early Recurrence	Invasion	Smoking	Heavy Alcohol Consumption	BALT	Lymphocyte Infiltration
KC01	M	72	Soft Palate	T2NxMx	Well	No	No	Yes	No	BALT	Moderate
KC02	M	50	Retromolar	T2NxMx	Moderate	Yes	No	Yes	No	BALT	Mild
KC04	M	66	Buccal	T1NxMx	Well/Moderate	No	No	No	No	BALT	Moderate
KC07	M	50	Alveolus	T2NxMx	Moderate	No	No	No	No	BALT	NA
KC09	F	54	Tongue	T1NxMx	Well	No	No	No	No	BALT	Mild
KC13	M	73	Buccal	T2NxMx	Well	No	No	No	No	BALT	NA
KC15	F	40	Tongue	T2NxMx	Moderate	No	No	No	No	No	NA
KC16	F	71	FOM	T2NxMx	Moderate	No	No	Yes	No	BALT	Mild
KC17	M	48	Alveolus	T2NxMx	Well	No	No	Yes	No	BALT	NA
KC19	M	76	Buccal	T2NxMx	Moderate	No	No	Yes	Yes	BALT	Mild
KC20	M	76	FOM	T1NxMx	Well	No	No	Yes	Yes	BALT	Moderate
KC21	F	56	Retromolar	T2NxMx	Well	No	No	No	No	BALT	Moderate
KC24	F	55	Buccal	T1NxMx	Moderate	No	No	No	No	BALT	NA
KC25	M	74	Buccal	T2NxMx	Moderate	No	No	Yes	No	BALT	NA
KC26	F	85	Alveolus	T2NxMx	Moderate	No	No	No	No	BALT	Mild
KC29	M	55	Buccal	T2NxMx	Well	No	No	Yes	Yes	BALT	Mild
KC31	M	95	Buccal	T2NxMx	Well/Moderate	No	No	No	No	BALT	Moderate
KC32	M	52	Buccal	T2NxMx	Well	No	No	Yes	Yes	BALT	Mild
KC38	M	51	Alveolus	T2NxMx	Moderate	No	No	Yes	Yes	BALT	Dense
KC39	M	64	Alveolus	T3NxMx	Moderate	No	No	Yes	Yes	BALT	Mild
KC41	M	82	Tongue	T3NxMx	Moderate	No	No	Yes	No	No	Mild
KC44	M	42	Palate	T3NxMx	Well	No	No	No	No	BALT	Mild
KC45	M	58	Tongue	T2NxMx	Moderate	No	No	Yes	Yes	BALT	Mild
KC46	M	38	Tongue	T2NxMx	Moderate	No	No	Yes	Yes	BALT	Dense
KC47	M	50	Buccal	T2NxMx	Well	No	No	No	No	BALT	NA
KC51	M	78	Buccal	T2NxMx	Moderate	No	No	No	No	BALT	NA
KC53	M	60	Buccal	T3NxMx	Well	No	No	Yes	No	BALT	NA

respectively. Then, the data were normalized using the Robust Multi-Array Average (RMA) normalization technique to increase the quality and to remove the technical variations. This process removed the local artifacts, array effects and combined probe intensities across the arrays.

2.2. Cross-Platform normalization

The samples of cancer patients were analysed through Affymetrix Human Genome Focus Array (GPL 201) and the samples of individuals were analysed through Affymetrix Human Genome U133 Plus2.0 Array (GPL570). Since the arrays are from different Affymetrix platforms, Linear Models for Microarray Analysis (LIMMA) Bioconductor package [12] has been used to assess the availability of technical biases, artifactual differences and batch effects. Hence, a Multi-Dimensional Scaling Plot was drawn to witness the non-biological experimental variation or the batch effect as in Fig. 2.

As the samples contained batch effect, the Bioconductor package inSilicoMerging [24] was used to merge the two platforms. Though the package consists of Six merging algorithms, the COMBAT [25] algorithm was used in particular to remove the unknown batch effect. The normalization left out the dataset with 8973 genes obtained through common identifiers.

2.3. Noise removal

Since the merged dataset had a high probability of having noises and outliers, the expression sets were assessed using *biosvd* [6] Bioconductor package. The expressions with steady-state gene expression and steady-scale variances were considered to be noises. Removal of such noises [1] enables meaningful comparison of gene expression across different arrays in different experiments.

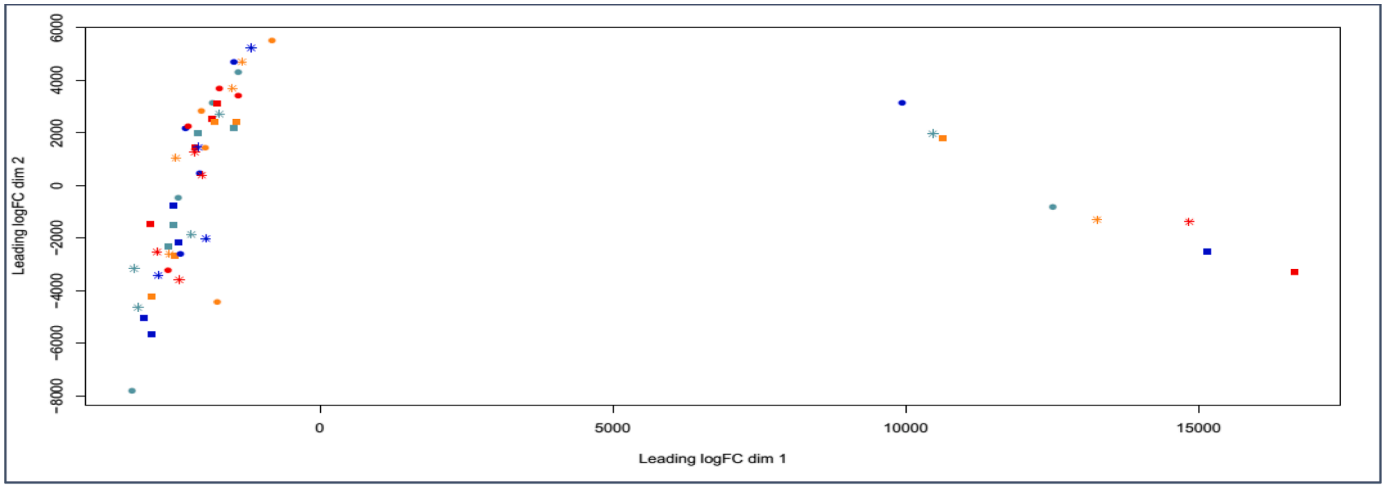
At first, the genes x sample space was transformed into a space of *eigengenes x eigenarrays* . The entropy was calculated and the bar plot for the merged expression set was plotted as shown in Fig 3.

The *allLine* graph and *eigenfeature* heatmap were also plotted for the merged dataset to visualize the available noises. it was inferred that the underlying processes were manifested by weak perturbations of the steady-state of expression from these graphs. Hence, the *exclude* function of the *biosvd* [6] package was used to filter the steady-state gene expressions as well as the steady scale variance. The comparative heatmaps before and after the noise removal are depicted in Fig. 4. a) and Fig. 4. b) respectively.

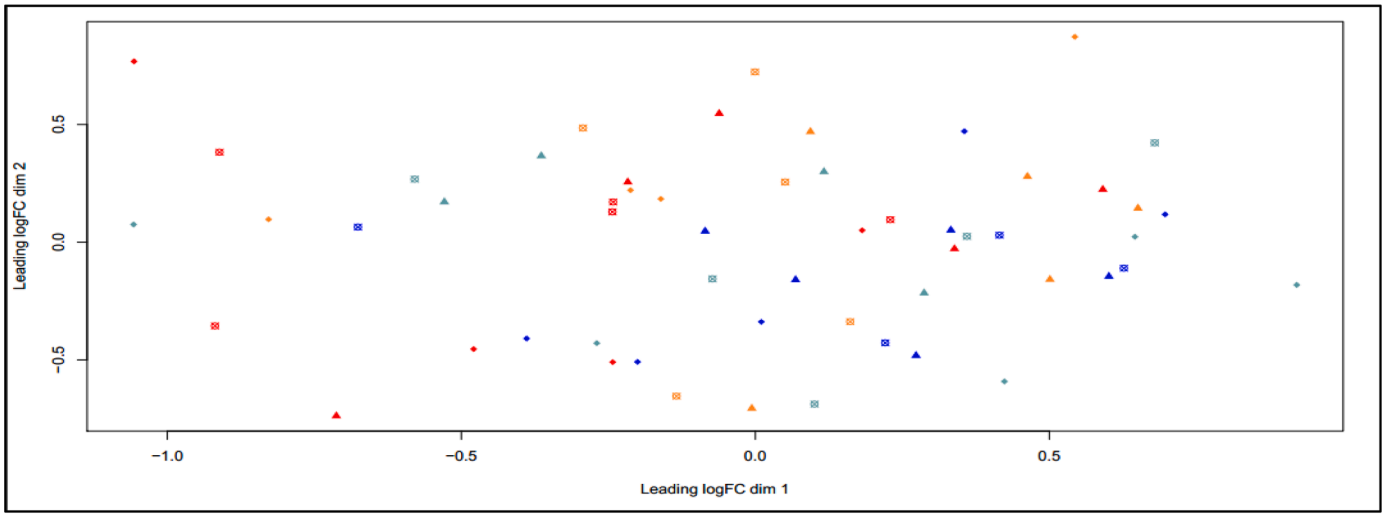
2.4. Identification of differentially expressed genes

The reliable method for the identification of differentially expressed genes is the evaluation of the log-ratio between conditions and considerations of genes that differ by more than a random cut-off value called the delta value [18]. At first, the supervised analysis was performed on the gene expression set using the *samr* (Significance Analysis of Microarrays) [26] package.

The set of differentially expressed genes were retrieved with the fold-change of ±1.5 and FDR (False Discovery Rate) of 0.5% from the list of significant genes obtained as shown in Fig. 5. Later, the identified over-expressed and lower expressed genes (together known as differentially expressed genes) were converted to gene identifiers using AnnotationDbi andhgu133plus2.db. At this stage, the control genes were also filtered using the *genefilter* [8] package.



a)



b)

Fig. 2. A multidimensional scaling (MDS) plot of the merged gene expression data. **a)** all samples are clustered by affymetrix platforms inside the MDS space without removal of the batch effect. **b)** With intra-platform batch adjustment, the samples are intermixed.

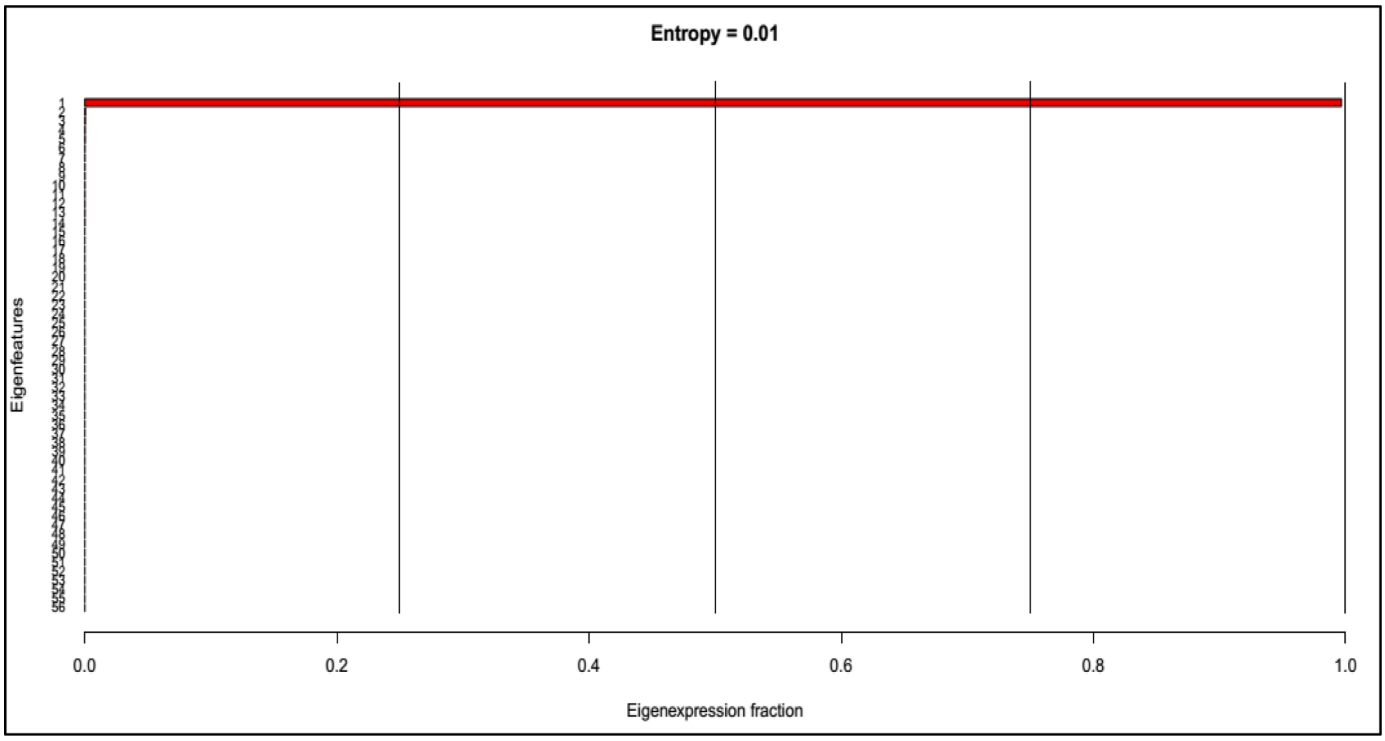


Fig. 3. Bar plot with all eigenfeatures.

2.5. Identification of molecular subclasses

Identification of molecular subclasses can substantially assist the oncologist to revise the therapeutic strategies to individualize the treatments. Consensus Clustering has gained wide popularity in cancer genomics as it could uncover the potential molecular subclasses of cancers [20]. *ConsensusClusterPlus* package was used to cluster the significant genes to identify the molecular subclasses of oral cancer.

The number of clusters starting from two was gradually increased by one until an optimal range was identified. The respective consensus matrix is presented in Fig. 6. As there was no significant difference after Six, the range was limited to Six. Clusters ranging from Two to Six were plotted and from the consensus Cumulative Distribution Function (CDF) curve analysis obtained, the optimal cluster was chosen when $k = 5$. The delta area, consensus CDF and the tracking plot are illustrated in Fig. 7.

2.6. Identification of stage-specific genes

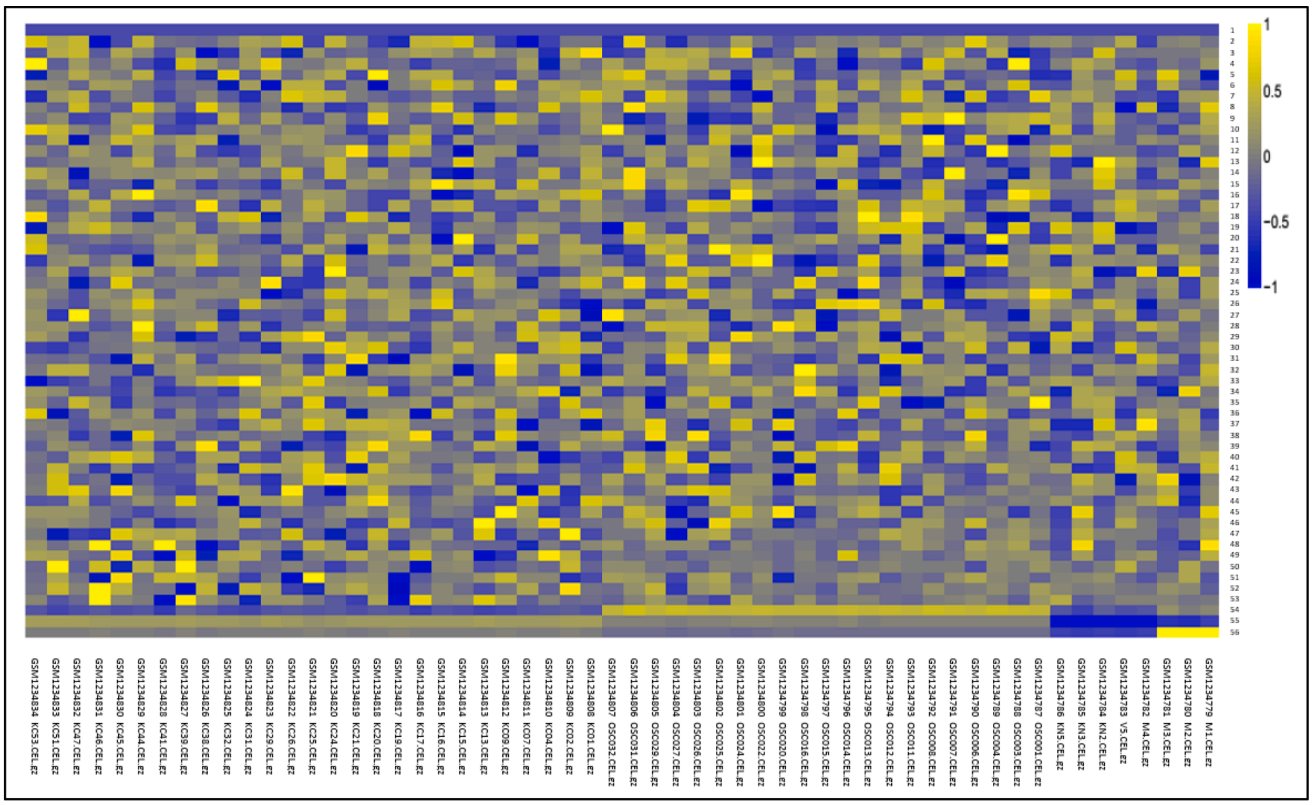
Gene co-expression networks have been playing a vital role in many of the bioinformatics analysis of microarray gene expression data. A gene co-expression network is typically built based on the similarity matrix between gene expression profiles of many genes over a set of samples or experimental conditions. The set of genes showing a high correlation i.e. having a similar pattern of expression tend to have similar functionalities and their transcript levels rise and fall together across samples [17]. This feature in particular is useful for revealing the association of gene expression with phenotypic traits.

The differentially expressed genes and their relative expressions were separated from other genes. Since the study is aimed to categorize between the early stage-specific genes and later stage-specific genes, the datasets were preliminarily categorized into two based on their tumour stages available in the clinical data. The correlation between the genes was calculated using Pearson's Correlation Coefficient separately. The results are presented in Tables 4 and 5.

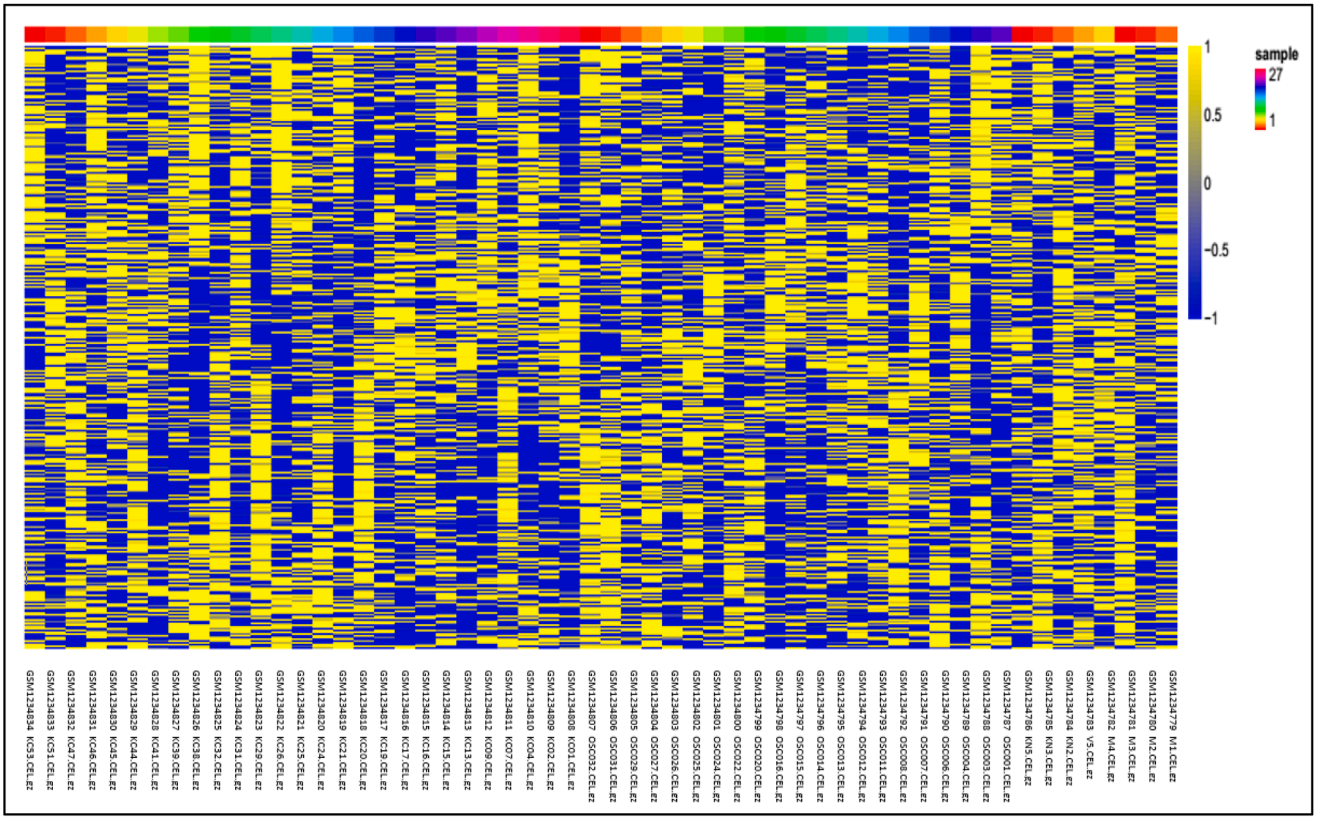
CoExpress software [16] has been used to obtain the number of subnetworks for varying threshold values of the similarity matrix. The similarity matrix was based on Pearson's Correlation Coefficient. This was plotted into a bar chart as shown in Fig. 8. Since the number of subnetworks decreased when proceeded to pass 0.8, the cut-off was set to 0.8. The plugin *ExpressionCorrelation* of Cytoscape software [22] computes a similarity network from either the genes or conditions in an expression matrix and it has been used to obtain the gene co-expression network in this study. The gene co-expression networks were built separately for early stage-specific genes and later stage-specific genes based on the significant correlation amongst them as indicated in Figs. 9 and 10 respectively.

2.7. Pathway analysis

Pathway analysis is one of the crucial steps in a bioinformatics analysis since it has the potential to identify important proteins in one pathway. The popular DAVID (The Database for Annotation, Visualization, and Integrated Discovery) [11] tool has been used to identify the most affected molecular and cellular functions, diseases and disorders, canonical pathways, and the transcriptional regulators [19] from amongst the genes that have been classified based on the subclasses and



a)



b)

Fig. 4. Comparative heatmaps before and after the noise removal. **a)** This heatmap shows that the availability of steady-state gene expressions across samples 1, 54, 55, and 56. **b)** The sorted heatmap after removing the steady state gene-expression and steady- scale variance.

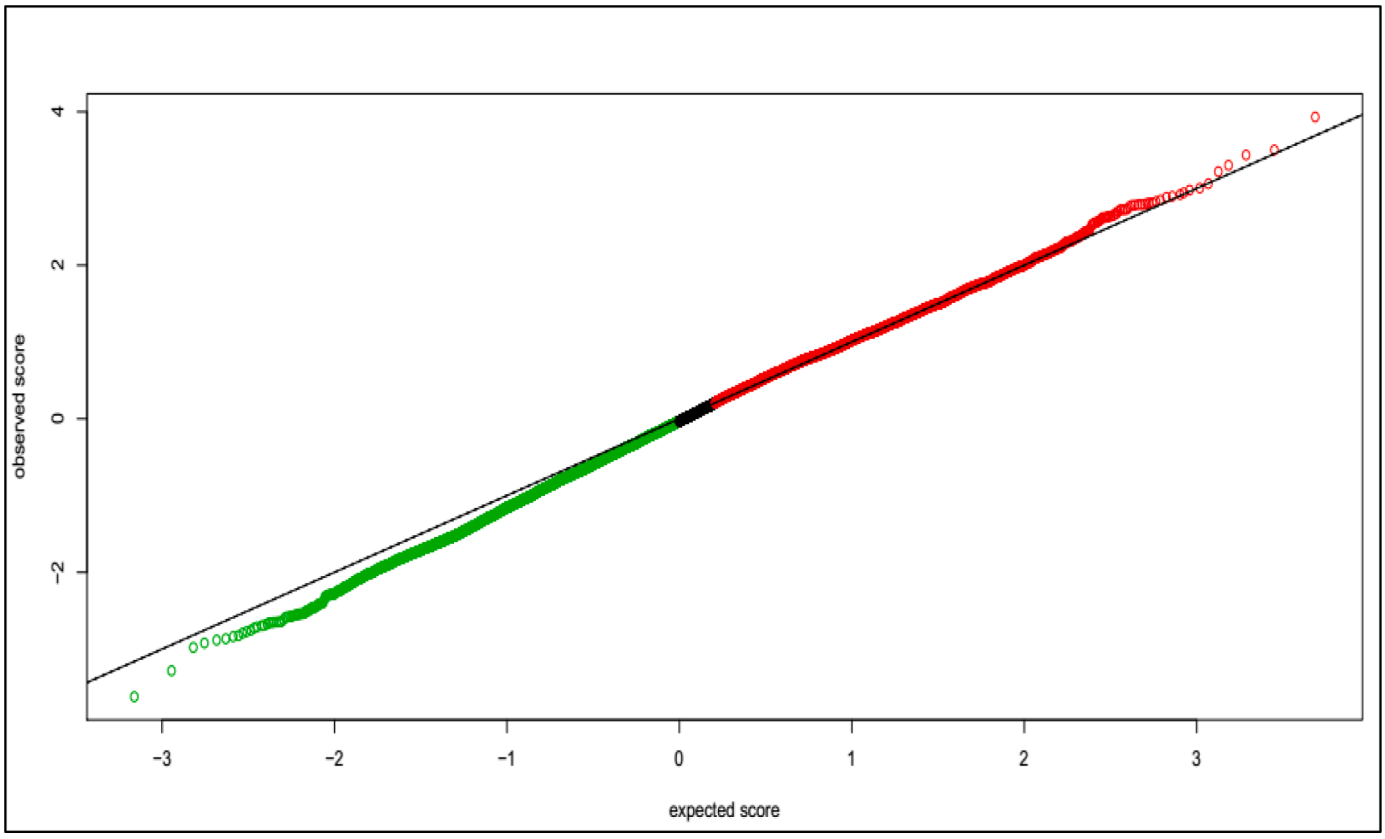


Fig. 5. The plot of differentially expressed genes for the selected delta value.

stages respectively in the previous steps.

The identified genes were tested against databases and the following interpretations were made. A previous research [21] revealed that there is a significant impact on the weight of the newborn baby born to the women who chew betel during pregnancy. This particular research investigated this hypothesis using 310 pregnant women and revealed that there is birth weight reduction in those who chew betel nut during pregnancy. This may be accredited to the SPINT1 gene as it plays a major role in placenta development, embryonic organ development.

According to our study, Integrin (ITGB4) has been identified as one of the under-expressed genes in OSCC patients. The principal role of Integrin in the human body is its contribution to focal adhesion that is mandated for the transmission of regulatory signals and mechanical forces between the extra-cellular matrix and the interacting cell. However, the arecoline present in the areca nut, chewed with the betel, deteriorates the stability of such matrix and result in the accumulation of extracellular matrix causing a barrier in transmission [23]. Thus, the cellular signals for White Blood Cells (WBC) to repair the damaged tissues are also not transmitted and leads to the lack of immunity in the human body.

On the other hand, Serglycin (SRGN) is a down-regulated gene for patients with OSCC. Serglycin gene is one of the important genes that have the cellular function of apoptosis, the programmed cell death [14]. Eugenol is a compound that triggers apoptosis in oral cancer patients and it is one of the essential components present in betel leaf. Thus, the general cyclic process of the development of cells is disturbed and the dysfunctional cells also exist.

The cytotoxicity in the human body can be inhibited by glutathione metabolism, which is the key function of Glutathione peroxidase 2 (GPX2). The particular gene is a master anti-oxidant and assists in the detoxification process. However, the up-regulation of this particular gene lowers the immune function and also makes people susceptible to infection. The root cause that induces the cytotoxicity in humans is again the arecoline of the areca nut.

2.8. *Validation and testing*

The obtained results were tested against the previously identified list of marker genes from the *Catalogue of Somatic Mutations in Cancer* (COSMIC) [7] database and the NCBI database. Most of the identified genes were related to OSCC and certain other carcinomas as well. It has been found that most of the genes coincided with the previously identified marker genes of carcinomas.

Out of the genes classified based on subclasses and the tumour stages, *TAGLN2, CCND2 and CCL8* were well-known tumour suppressor genes. Further, *GPX3, GRN, and ITGB4* like genes are involved in several

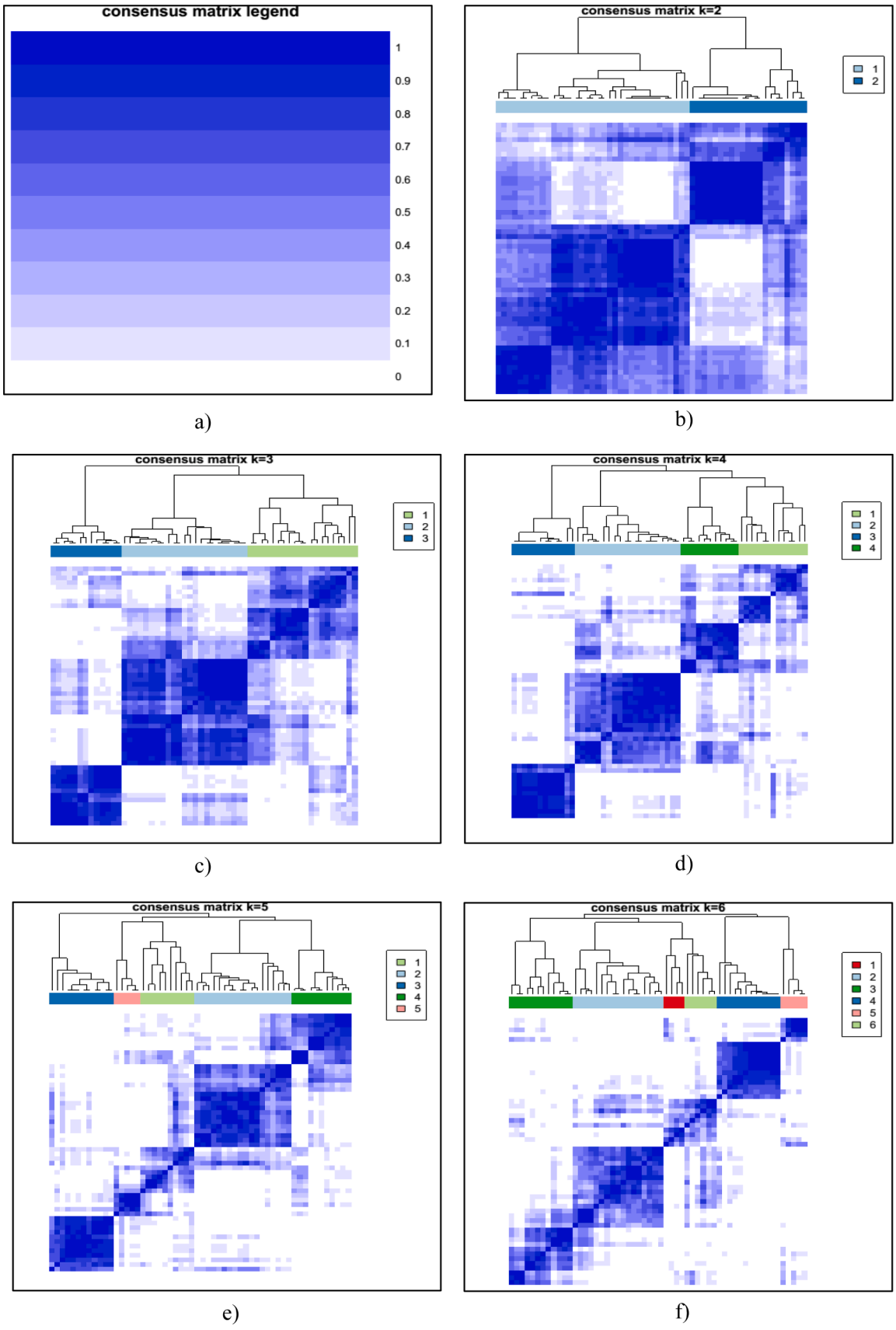
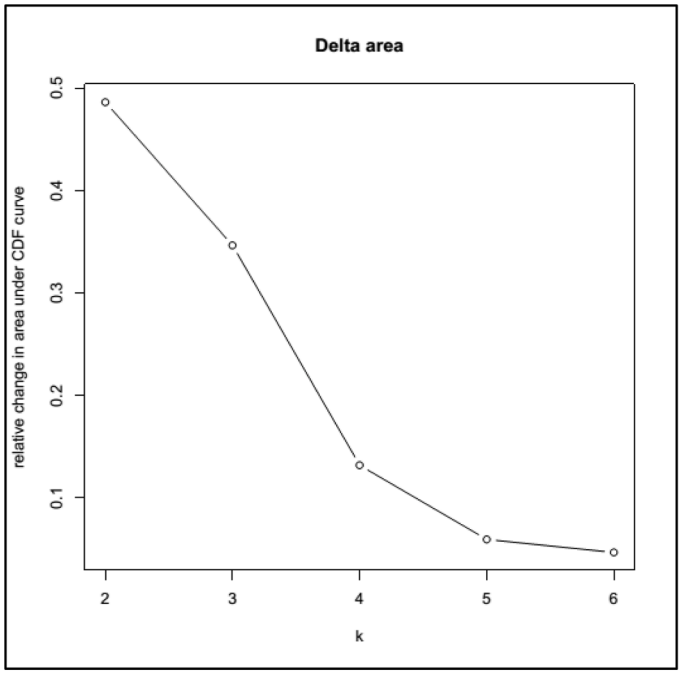
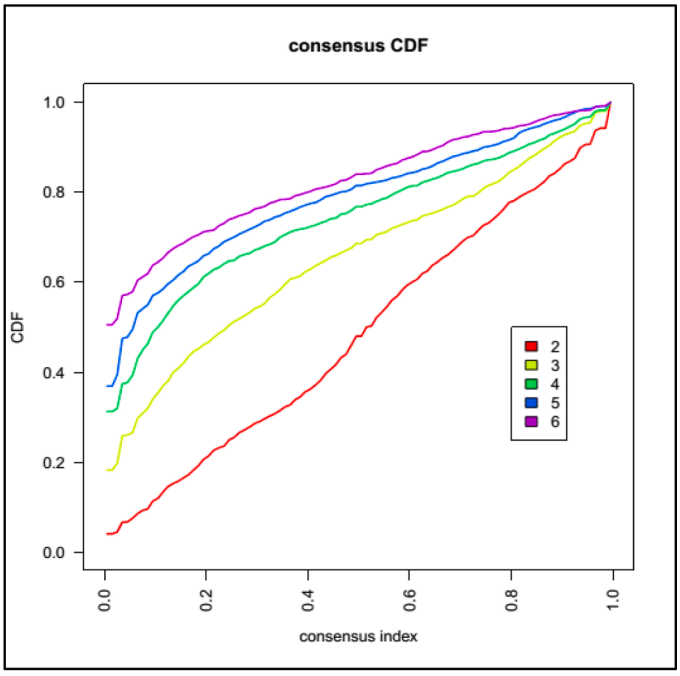


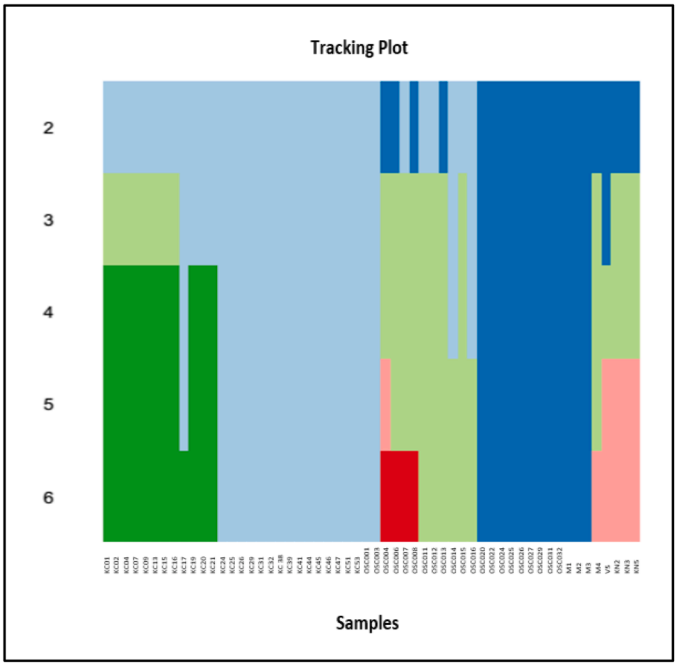
Fig. 6. Consensus matrix of the clusters **a)** Consensus matrix legend **b)** when $k = 2$ **c)** when $k = 3$ **d)** when $k = 4$ **e)** when $k = 5$ **f)** when $k = 6$.



a)



b)



c)

Fig. 7. Identification of Clusters **a)** The plot representing the delta area of varying number of clusters **b)** Their respective Cumulative Distribution Function and **c)** The tracking plot of clusters.

Table 4
Comparative analysis of differentially expressed genes with previously published results.

	Genes that comply with previously published results	Genes that do not comply with previously published results
Down-regulated genes	DDX1 IFIH1 CCT6A UCHL3 NAMPT ERAP2 MPZL2 RASGRP1 F3 IFI44L PTPRC CLEC2B CD24	GCH1 NMI IFIT1 DDX58 LAP3 GBP1 SAMSN1 RGS2 IFI44 CCND2 CCL8 IVNS1ABP IFITM1 SRGN CXCL8 CXCL9 WARS TPD52
Up-regulated genes	CAPNS1 GRN CTSD GPX3 SPINT1 ACTA2 H1FX CRIP1 AP2M1 SH3BGRL3 CST3 TAGLN2 CYP1B1 KRT76 APOD HSPB1 DES P4HB CLIC3	ITGB4 GPX2 TYMP

carcinomas.

3. Results

The differentially expressed genes were first categorized according to the molecular subclasses and later, they were classified against their tumour stages such as early stage and later stage. The differentially expressed genes obtained from the study were tested against the results obtained from the previous research [19] which is also based on the same microarray data. Many of the differentially expressed genes identified were similar and this study has identified few unique marker genes as well. Table 4 provides a tabular view of up-regulated and down-regulated genes identified.

Further, Consensus Clustering revealed that there could be five molecular subclasses in the gene expression data used for this study. Hence, the differentially expressed genes were categorized into five molecular subclasses. The genes were also classified based on their tumour stages. Table 5 provides a concise view of the genes that have undergone classification based on both factors.

Table 5
List of Clustered and Classified genes.

Cluster number	Early Stage	Later Stage
Cluster 1	GBP1 DDX1 CCT6A TPD52 CCL8 SRGN RGS2 PTPRC GGH AIM2 WARS	CCND2 WARS
Cluster 2	GPX2 ITGB4	GRN ITGB4
Cluster 3	ACTA2 CST3 CTSD CRIP1 IFIH1 ERAP2	CYP1B1 CRIP1
Cluster 4	TAGLN2	
Cluster 5	P4HB	

4. Discussion

There are plenty of conventional therapeutic approaches available for the treatment of oral cancer such as altered resection, chemotherapies, and radiotherapies. However, the impacts that these treatments imprint on the patients are yet destroying the wellbeing and quality of their life [13]. Hence, the necessity in revising or modifying the existing approaches is paramount to improve health outcomes as well as survival. Since the primary reason for the low survival rate is often claimed as the detection of oral cancer at a later stage, this study proposes a methodology that could expedite the detection of oral cancer relatively earlier by reducing the time as well as the money spent in many screening tests done for detection, thus leads to the detection at an early-curable stage individually. Further, the classification of oral cancer into five molecular subclasses could help to revise the therapeutic measures such that the treatments could be individualized. This approach is promising since the co-expression networks proved that most of the genes are differentially expressed at an early stage rather than later stage.

The sensitivity to drugs could be tested using the Genomics Drug Sensitivity Project to the putative genes identified through this research, such that it can be used to identify potential associations with drug sensitivity. The successful implementation could output a more robust molecularly defined subtype and stage classified genes of OSCC which can improve patient selection and pave the way to the development of appropriate therapeutic strategies for OSCC.

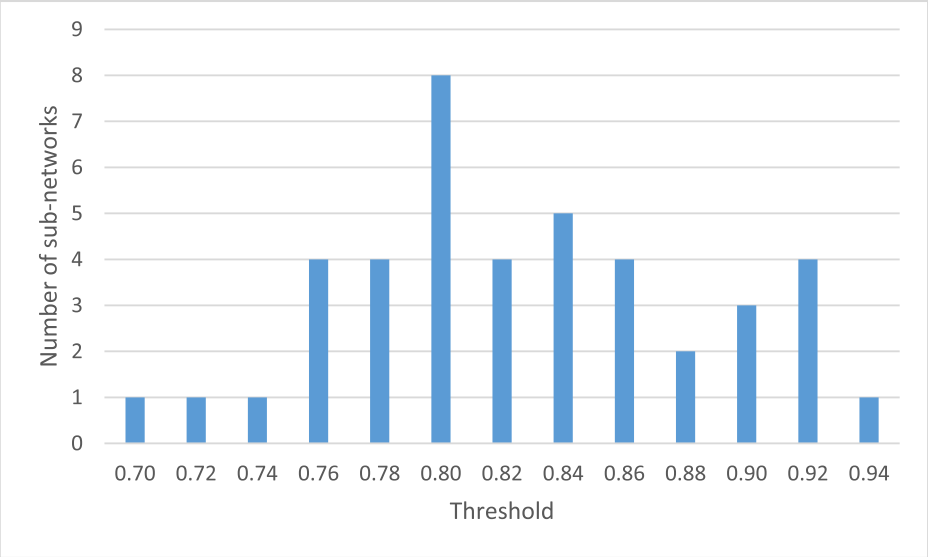


Fig. 8. Bar chart depicting the variability of number of clusters for varying threshold.

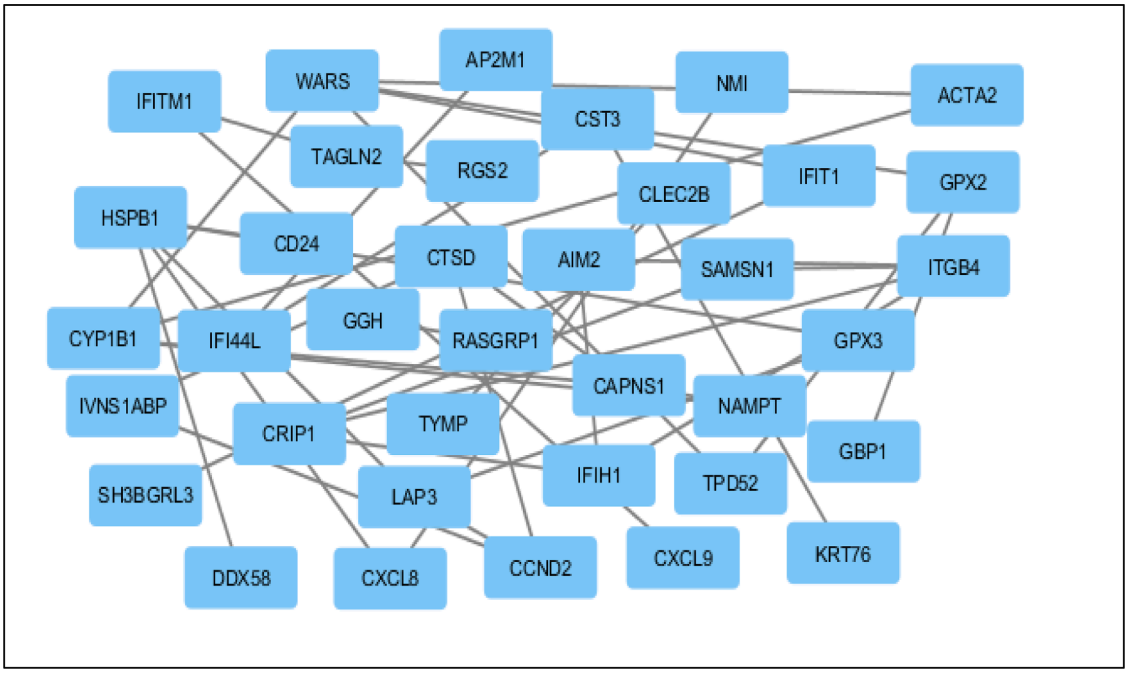


Fig. 9. Gene co-expression networks built amongst the early stage specific-genes.

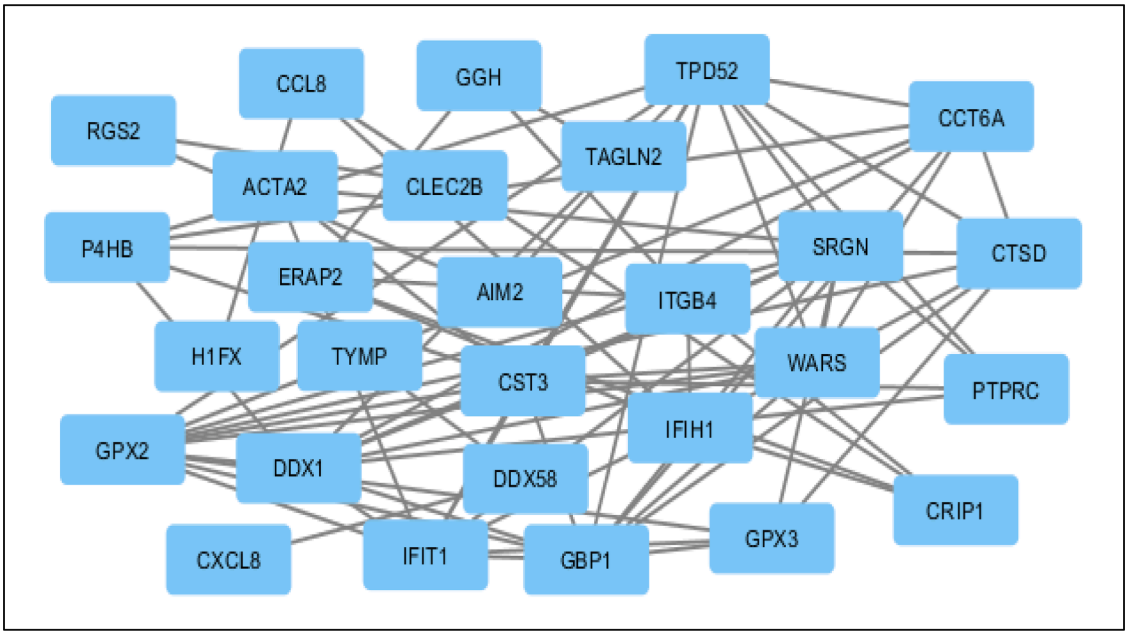


Fig. 10. Gene co-expression networks built amongst the later stage specific-genes.

5. Conclusion

The candidate marker genes of OSCC were identified and classified based on molecular subclasses and the stages OSCC. Five molecular subclasses of oral cancer have been identified and the results were further clustered as genes associated in early-stage and later stage of OSCC. This would be helpful in revising the therapeutic strategies based on further analysis of impact level and survival rate of each of the subtypes identified. Hence, a clinician might be able to design and implement a targeted drug therapy which would be more effective and practically feasible. The revision of therapeutic measures based on this classification could improve the wellbeing and quality of patients' life. Also, it has been discovered that most of the genes are differentially expressed at the early stages. Thus, it could be concluded that further study on this could help to detect oral cancer at an early curable stage as well as to design and develop a suitable treatment protocol

Credit author statement

Abdul Raheem Fathima Shafana: Methodology, Software, Formal analysis, Investigation, Data curation, Writing, Review & editing

Gatamanna Arachchige Isuri Uwanthika: Methodology, Software, Validation, Formal analysis, Investigation, Visualization

Thangathurai Kartheeswaran: Conceptualization, Methodology, Formal analysis, Resources, Review & Editing, Visualization, Supervision, Project administration

6. Funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank the Department of Computer Science and Informatics of Uva Wellassa University that helped to undertake this study successfully. The research was conducted in this institute and authors would like to acknowledge here. Funding was not provided by the particular institute.

References

[1] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. 97 (101) (2000).
[2] Anura Ariyawardana, Saman Warnakulasuriya, Declining oral cancer rates in Sri Lanka: are we winning the war after being at the top of the cancer league table? Oral Dis. 17 (2011) 636–641, 10.1111/j.1601-0825.2011.01809. x.
[3] L.D. Cecco, M. Nicolau, M. Giannoccaro, M.G. Daidone, P. Bossi, L. Locati, L. Licitra, S. Canevari, Head and neck cancer sub classes with biological and clinical relevance: meta-analysis of gene-expression data, Oncotarget 6 (11) (2015).
[4] Centre for Research in Oral Cancer | University of Peradeniya 2020., [Online] Available at: http://www.pdn.ac.lk/centers/croc/ocfacts.php [Accessed 11 November 2020].
[5] Y.L. Cheng, T. Rees, J. Wright, A review of research on salivary biomarkers for oral cancer detection, Clin. Transl. Med. 3 (3) (2014).
[6] Daemen, A. & Brauer, M., 2013. 'biosvd: package for high-throughput data processing, outlier detection, noise removal and dynamic modeling'.
[7] S.A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P.A. Futreal, M.R. Stratton, The Catalogue of Somatic Mutations in Cancer (COSMIC), Curr. Protocol. Hum. Genetic. (2008). Chapter 10, Unit–10.11. 10.1002/0471142905.hg1011s57.
[8] Gentleman R., Carey V., Huber W. and Hahne F. (2016). genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1.56.0.
[9] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol 5 (2004) R80.
[10] O. Gevaert, B. De Moore, Prediction of cancer outcome using DNA microarray technology: past, present and future, Expert Opin. Med. Diagn 3 (2) (2009) 157–165.
[11] D.W. Huang, B.T. Sharman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (2008) 44–57.
[12] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (1) (2007) 118–127.
[13] F. Ketabat, M. Pundir, F. Mohabatpour, L. Lobanova, S. Koutsopoulos, L. Hadjiiski, X. Chen, P. Papagerakis, S. Papagerakis, Controlled drug delivery systems for oral cancer treatment—current status and future perspectives, Pharmaceutics 11 (2019) 302.
[14] Korpetinou, A., Skandalis, S.S., Labropoulou, V.T., Smirlaki, G., Noulas, A., Karamanos, N.K., & Theocharis, A.D. 2014. Serglycin: at the crossroad of inflammation and malignancy. In *Frontiers in Oncology*. 10.3389/fonc.2013.00327.

[15] National Cancer Control Programme Sri Lanka, Cancer Incidence Data: Sri Lanka Year, NCCP, Colombo, 2014, 2014.
[16] Okamura, COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems, Nucl. Acids Res. 43 (2015) D82–D86.
[17] A.D. Perkins, M.A. Langston, Threshold selection in gene co-expression networks using spectral graph theory techniques, BMC Bioinformat. 10 (2009) (Suppl. 11): S4.
[18] V. Randhawa, V. Acharya, Integrated network analysis and logistic regression modeling identify stage-specific genes in Oral Squamous Cell Carcinoma, BMC Med. Genomic. 39 (8) (2015).
[19] A.A. Saeed, A.H. Sims, S.S. Prime, I. Paterson, P.G. Murray, V.R. Lopes, Gene expression profiling reveals biological pathways responsible for phenotypic heterogeneity between UK and Sri Lankan oral squamous cell carcinomas, Oral. Oncol. 3 (2015) 237–246.
[20] Yasin Şenbabaoğlu, George Michailidis, Jun. Li, Critical limitations of consensus clustering in class discovery, Sci. Rep. 4 (2014) 6207, 10.1038/srep06207.
[21] M. Senn, F. Baiwog, J. Winmai, I. Mueller, S. Rogerson, N. Senn, Betel nut chewing during pregnancy, Madang province, Papua New Guinea (2009) viewed 08 October 2019https://www.ncbi.nlm.nih.gov/pubmed/19665325.
[22] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.
[23] Yu-Tzu Shih, Po Chen, Chi-Han Wu, Yu-Ting Tseng, Yang-Chang Wu, Yi-Ching Lo, Arecoline, a major alkaloid of the areca nut, causes neurotoxicity through enhancement of oxidative stress and suppression of the antioxidant protective system, Free Radic. Biol. Med. 49 (2010) 1471–1479, https://doi.org/10.1016/j.freeradbiomed.2010.07.017.
[24] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (3) (2004).
[25] J. Taminau, S. Meganck, C. Lazar, D. Steenhoff, A. Coletta, C. Molter, et al., Unlocking the potential of publicly available microarray data using. inSilicoDb and inSilicoMerging R/Bioconductor packages, BMC Bioinformat. 13 (335) (2012).
[26] V. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, Proc. Natl. Acad. Sci. USA. 98 (2001), 5116–5.