**Cluster analysis of gene expression data from microarrays of different cancer types to identify common gene expression patterns**

# **ABSTRACT**

Microarrays allow the simultaneous expression of many genes. Cancer diagnosis and prognosis utilize gene expression data from microarrays. Understanding gene expression patterns using microarray data is difficult due to the complexity of the data and its limited sample size. Clustering techniques may be used to detect comparable gene expression patterns across cancer microarray datasets. This project aimed to cluster gene expression data from 6 separate microarray studies covering 6 different cancer types. These patterns may help in explaining cancer's physiological mechanisms and in designing cancer therapeutics. Raw microarray data for 6 cancer types were downloaded from the Gene Expression Omnibus database and processed. This was followed by differential gene expression analysis to identify genes differentially expressed in cancer tissues relative to normal tissues for each cancer type (log2-fold change (LFC) -1.5 or > 1.5). GSEA was then conducted to reveal shared biological pathways among expressed gene sets. Clustering was performed to identify similar gene expression patterns. Overall, there were more downregulated (n = 3167) than upregulated (n = 1907) genes, with renal cancer having the most upregulated (n = 714) and bladder cancer the most downregulated (n = 1686). Clustering revealed 742 genes (log2-fold change = -1.5 or > 1.5 and k = 5) which were expressed in more than one cancer type. No genes were expressed in all 6 cancers. 4 genes *(ASPM, COL1A1, CXCL8, and TOP2A)* were expressed in 5 of the 6 cancer types. Clustering methods are crucial for identifying common gene expression patterns, but additional studies are needed to understand the genes' significance in tumor growth and progression.

# 

# 

[**ABSTRACT**](#_exjtunlk5sa7) **2**

[**CHAPTER ONE**](#_l8ffiro1qsi) **4**

[**1.0. INTRODUCTION**](#_x2lysama2cvt) **4**

[**CHAPTER TWO**](#_v66137nlswb9) **5**

[**2.0. METHODS**](#_i6qxdvhdv1hm) **5**

[2.1. Datasets](#_1zk5nwdn56iy) 5

[2.2. Data processing/data normalization](#_kmfys5i1g572) 6

[The data were processed by conducting background correction, normalization, and summarization in R using the Robust Multichip Averaging (RMA) method utilizing the rma() function that is provided in the affy package. Normalization was performed on the microarray data to adjust for technical differences across arrays.](#_tj1z0xlhnt4e) 6

[2.3. Differential gene expression](#_gaxqlmdityd7) 6

[**2.4. Gene set enrichment analysis (GSEA)**](#_hb8q0t4199gs) **6**

[2.5. Clustering of gene expression data](#_veirz0ldq1xm) 6

[**CHAPTER THREE**](#_eswl67d3y31a) **7**

[**3.0. RESULTS**](#_86fipv39121f) **7**

[**3.1. Differential gene expression**](#_uix3rqzgz7wj) **7**

[**3.2. Gene set enrichment analysis**](#_l8gtzcyap968) **9**

[**3.3. Clustering of gene expression data**](#_rp5lipa2k3p5) **14**

[**CHAPTER FOUR**](#_olvvs9egytj3) **15**

[4.1. Discussion](#_n678gtx5o6zl) 15

[4.2. Conclusion](#_6p6ezv8aip65) 16

[**REFERENCES**](#_ubi6qrhpe0ef) **17**

[**SUPPLEMENTARY DATA**](#_rmahhqxtnz31) **18**

[Supplementary data 1](#_fz2uufwvgwln) 18

[Supplementary data 2](#_c52rv0eqfc3f) 18

[Supplementary data 3](#_bsdhy4rr3z2v) 18

[Supplementary data 4](#_jmrj7zqor5zv) 18

# **CHAPTER ONE**

# **1.0. INTRODUCTION**

A microarray is a laboratory tool used to determine the expression of a large number of genes simultaneously. Microarrays are microscope slides called gene chips because there are thousands of spots on the slides each containing a gene (known DNA sequence). The genes act as probes [(Bumgarner, 2013)](https://www.zotero.org/google-docs/?K8Kbii) to detect gene expression or the amount of messenger RNA (mRNA) transcripts expressed by a collection of genes ("microarray | Learn Science at Scitable", 2022). During microarray analysis, mRNA molecules are isolated from both the experimental sample and the control sample. For example, the control sample may be obtained from a healthy individual. In contrast, the experimental sample may be obtained from a patient with a disease such as cancer ("microarray | Learn Science at Scitable", 2022).

Cancer diagnosis and prognosis rely on the analysis of gene expression data from tools like microarrays. Understanding common patterns of gene expression from microarray data is difficult due to the data's complexity, multidimensionality, and limited sample size [(Pawar et al., 2020)](https://www.zotero.org/google-docs/?sxYUX0). Techniques such as gene expression data clustering can be used to overcome this issue by identifying similar gene expression patterns across several cancer microarray experiments. Clustering gene expression data also facilitates the identification of gene functions, cellular processes, and cell subtypes, the extraction of meaningful information from noisy data, and the interpretation of gene regulation [(Oyelade et al., 2016)](https://www.zotero.org/google-docs/?Y6RhxN). Common clustering approaches include hierarchical clustering and the use of hybridized K-means.

This project aims to perform hierarchical clustering on gene expression data from 6 different microarray studies involving 6 different cancer types so as to obtain common gene expression patterns. These patterns can be potentially useful in understanding the major physiological processes associated with cancers and in designing cancer therapies.

# **CHAPTER TWO**

# **2.0. METHODS**

## **2.1. Datasets**

In this project, raw microarray data (Affymetrix CEL data files) from six distinct cancer studies, each concentrating on a particular cancer type, were selected and downloaded from the Gene Expression Omnibus (GEO) database (Supplementary data 1) which is managed by the National Center for Biotechnology Information (NCBI). The 6 studies considered were on bladder, cervical, colorectal, esophageal, pancreatic, and, renal cancer respectively. For each form of cancer, 10 malignant and 10 normal tissue samples were chosen and imported into Rstudio for downstream analysis using R programming tools.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cancer type** | **Organism** | **Malignant samples** | **Normal samples** |
| Bladder cancer | Homo sapiens | 10 | 10 |
| Cervical cancer | Homo sapiens | 10 | 10 |
| Colorectal | Homo sapiens | 10 | 10 |
| Esophageal cancer | Homo sapiens | 10 | 10 |
| Pancreatic cancer | Homo sapiens | 10 | 10 |
| Renal cancer | Homo sapiens | 10 | 10 |

## **2.2. Data processing/data normalization**

## The data were processed by conducting background correction, normalization, and summarization in R using the Robust Multichip Averaging (RMA) method utilizing the *rma()* function that is provided in the affy package. Normalization was performed on the microarray data to adjust for technical differences across arrays.

## **2.3. Differential gene expression**

Using the normalized microarray data, differential gene expression analysis was performed to identify the genes that are differentially expressed in cancer tissues compared to normal tissues for each kind of cancer. R's limma package was used to conduct differential gene expression. Only genes that were differentially expressed with a log2-fold (LFC) change < -1.5 or > 1.5 were considered for downstream analysis.

# **2.4. Gene set enrichment analysis (GSEA)**

GSEA was performed to reveal the common biological pathways present among the different sets of genes. GSEA was performed using the clusterprofile package in R.

## **2.5. Clustering of gene expression data**

Hierarchical clustering was performed on the gene expression data to group genes with similar expression patterns together based on a log2-fold change (LFC) of -1.5 or > 1.5 and a maximum of 5 clusters (k = 5). This aided in identifying common expression patterns of genes across the different cancers. Clustering was performed using the *hclust()* function from the R stats package. Expression patterns were visualized using heatmaps. The pheatmap package in R was used to generate the heatmaps, and the log2-fold change was modified for each cancer type such that each heatmap could include no more than 50 genes.

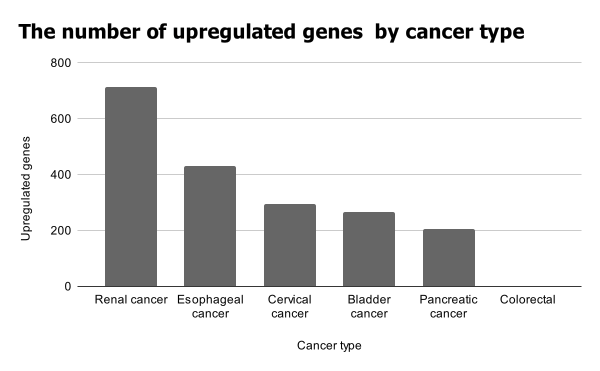
# **CHAPTER THREE**

# **3.0. RESULTS**

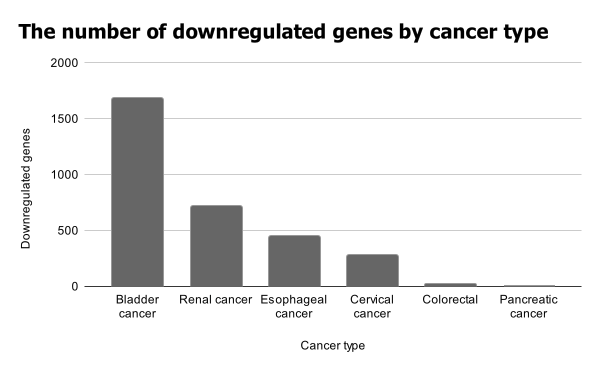
# **3.1. Differential gene expression**

When filtered using a log2-fold change (LFC) -1.5 or > 1.5, the total number of differentially expressed genes differed across cancer types. Overall, more genes were downregulated than were upregulated (Supplementary data 2).

|  |  |  |
| --- | --- | --- |
| **Cancer type** | **Upregulated genes** | **Downregulated genes** |
| Bladder cancer | 265 | 1686 |
| Cervical cancer | 294 | 283 |
| Colorectal | 0 | 24 |
| Esophageal cancer | 430 | 453 |
| Pancreatic cancer | 204 | 5 |
| Renal cancer | 714 | 716 |
| **Total** | **1907** | **3167** |

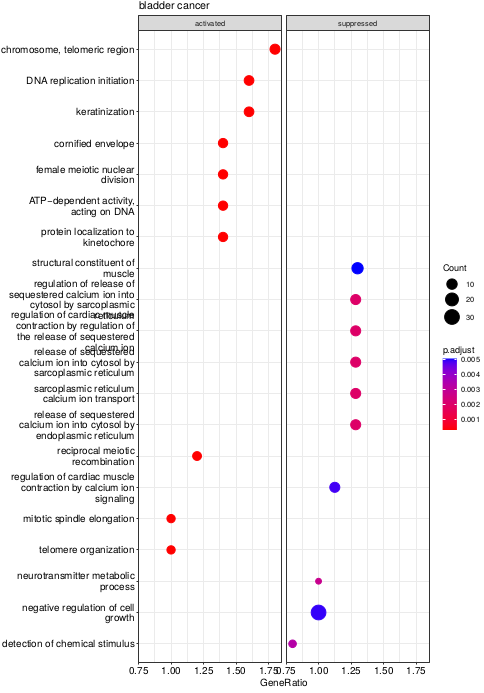
Renal cancer had the most upregulated genes, followed by esophageal, cervical, bladder, pancreatic, and colorectal cancer, which had no upregulated genes.

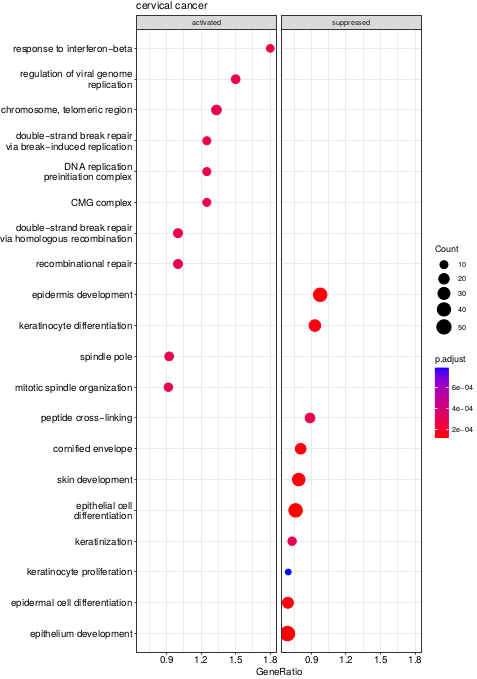
The highest number of downregulated genes was observed in bladder cancer, followed by renal, esophageal, cervical, colorectal, and pancreatic cancers, in that order.



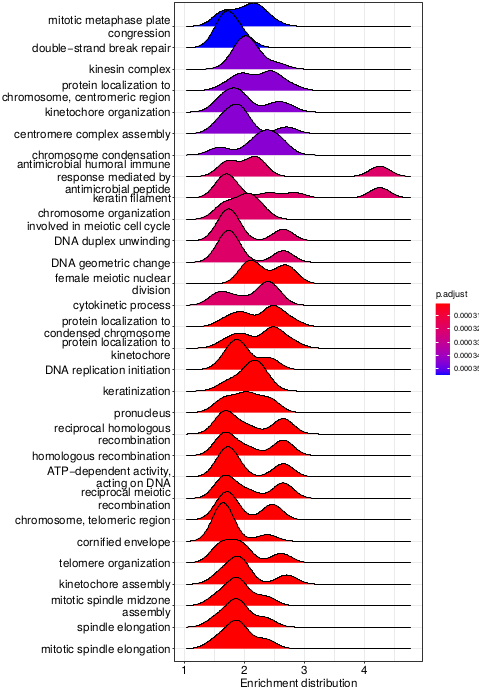
# **3.2. Gene set enrichment analysis**

GSEA revealed the number of pathways that are activated or suppressed in each type of cancer. The activated and suppressed pathways differed by cancer type. The dot plots below depict the top 10 activated and suppressed pathways in bladder and cervical cancers. Plots for all cancer types have been included in supplementary data 3 files.





The GSEA results also categorized upregulated and downregulated pathways according to gene sets. As seen in the ridge plots below for bladder and cervical cancer, these pathways also differed across various cancer types.



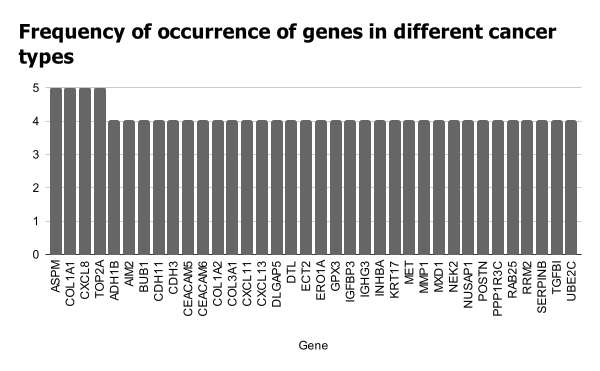
*Bladder cancer*

# 

*Cervical cancer*

# **3.3. Clustering of gene expression data**

A total of 742 genes were expressed in more than one cancer type when clustering was performed using a log2-fold change (LFC) of -1.5 or > 1.5 and a maximum of 5 clusters for each cancer type (k = 5). No genes were expressed across all 6 types of cancer. However, 4 genes were expressed in 5 of the 6 cancer types, and 32 genes were expressed in 4 of the 6 cancer types. Clustered genes and corresponding heatmaps by cancer type were included in supplementary data 4 files



# 

# **CHAPTER FOUR**

## **4.1. Discussion**

Variations in the number of up-regulated and down-regulated genes for the different cancer types may be linked to the many biological pathways and molecular functions that are affected by the genetic alterations that result in the creation of malignant cells. Normal cell proliferation-related genes are often inhibited in malignant cells, while aberrant cell proliferation-related genes are typically activated. Consequently, discrepancies in the numbers of activated, deactivated, upregulated, and downregulated pathways were observed.

The 4 genes that were expressed in 5 of the 6 cancer types included the Abnormal spindle-like microcephaly-associated *(ASPM)* gene which is important for mitotic spindle function during cell replication. Its overexpression has been linked to cancer aggressiveness. The collagen 1A1 *(COL1A1)* gene is also expressed in 5 of the 6 cancer types and is involved in the manufacture of collagen to strengthen and support body tissues. Altered expression levels of this gene have been implicated in tumor invasion and progression. The CXC motif chemokine ligand 8 *(CXCL8)* gene which mediates inflammatory response, was also expressed in 5 of the 6 cancer types. Altered expression of this gene had been linked to tumor progression and metastasis by regulating cancer stem cell proliferation and self-renewal. DNA Topoisomerase II Alpha *(TOP2A)* gene was also expressed in 5 of the six cancer types. This gene controls the topological state of DNA during transcription. Its altered expression has also been associated with tumor development and progression.

## 

## 

## **4.2. Conclusion**

It is vital to cluster gene expression data from different cancer research in order to identify common gene expression patterns. These patterns may aid in the comprehension of the key physiological processes connected with tumors and the development of cancer therapeutics.

# **REFERENCES**

[Bumgarner, R. (2013). DNA microarrays: Types, Applications, and their future. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *0 22*, Unit-22.1. https://doi.org/10.1002/0471142727.mb2201s101](https://www.zotero.org/google-docs/?8dRmBd)

[Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., & Adebiyi, E. (2016). Clustering Algorithms: Their Application to Gene Expression Data. *Bioinformatics and Biology Insights*, *10*, 237–253. https://doi.org/10.4137/BBI.S38316](https://www.zotero.org/google-docs/?8dRmBd)

[Pawar, S., Stanam, A., & Lahiri, C. (2020). Clustering Reveals Common Check-Point and Growth Factor Receptor Genes Expressed in Six Different Cancer Types. In I. Rojas, O. Valenzuela, F. Rojas, L. J. Herrera, & F. Ortuño (Eds.), *Bioinformatics and Biomedical Engineering* (pp. 581–589). Springer International Publishing. https://doi.org/10.1007/978-3-030-45385-5\_52](https://www.zotero.org/google-docs/?8dRmBd)

*microarray | Learn Science at Scitable*. Nature.com. (2022). Retrieved 24 July 2022, from <https://www.nature.com/scitable/definition/microarray-202/>.

# **SUPPLEMENTARY DATA**

## Supplementary data 1

Raw microarray data (Affymetrix CEL data files)

<https://drive.google.com/drive/folders/1XA4ePd2cIKp5RCuv3mfuZlPjwMFyQIw_?usp=sharing>

## Supplementary data 2

Differentially expressed genes by cancer type

<https://drive.google.com/drive/folders/1Pl_PpotnPdBcuKvyBB2veSQlGJDlQJb_?usp=sharing>

## Supplementary data 3

Gene set enrichment analysis

<https://drive.google.com/drive/folders/13UOjgKe_nVuV5ALQbH4UaK5SrjMQ_qyR?usp=sharing>

## Supplementary data 4

Clustering results

<https://drive.google.com/drive/folders/1FXcAh95jFqICQpINRgt_x0_HAyecGtP4?usp=sharing>