Discovering cancer subtypes by tracking cancer progression with transcriptomic data through the multi-stage process of cancer development.

Michelle Chantel Livesey

Supervisor: Dr Hocine Bendou

Co-supervisors: Prof Alan Christoffels

South African National Bioinformatics Institute

Faculty of Natural Sciences

University of the Western Cape

This dissertation is submitted for the degree of

Doctor of Philosophy

SANBI December 2023

Discovering cancer subtypes by tracking cancer progression with transcriptomic data through the multi-stage process of cancer development.

by M.C Livesey

Keywords

Cancer progression
Transcriptomic profiles
Normalization
Cancer heterogeneity
Subtyping

Abstract

Background: The development of cancer is driven by genomic alterations, which become more heterogeneous as the disease progresses throughout the stages. Consequently, cancer patients have differential levels of sensitivity to treatment. Tumor heterogeneity thus contributes to therapeutic failure, which ultimately leads to the generally poor prognosis and poor overall survival outcome associated with cancer.

Introduction: Transcriptomic profiles can be used to track cancer progression based on gene expression changes that occur throughout the multi-stage process of cancer development. The accumulated genetic changes can be detected when gene expression levels in advanced-stage are less variable but show high variability in early-stage. Normalizing advanced-stage expression samples with early-stage and clustering of the normalized expression samples can reveal cancers with unique gene expression patterns based on cancer progression.

Aims: A computational method was employed to investigate cancer progression through RNA-Seq expression profiles across the multi-stage process of cancer development. The method was assessed in a subtype of the heterogeneous kidney cancer and enabled the discovery of in-depth cancer subtypes based on the differences in gene expression profiles.

Methods: A preliminary study was performed by downloading RNA-sequenced gene expression and associated phenotypic and survival profiles of Diffuse Large B-cell Lymphoma, Lung cancer, Liver cancer, Cervical cancer, and Testicular cancer from the UCSC Xena database. Similarly, Kidney renal clear cell carcinoma (KIRC) was downloaded as a validation dataset. Advanced-stage samples were normalized with early-stage to consider heterogeneity differences in the multi-stage cancer progression. The normalized gene expression of the preliminary cancer datasets was subjected to weighted gene co-expression network analysis. Gene modules were linked to cancer-related proteins and pathways using enrichment analyses.

Hierarchical clustering was performed to reveal clusters (subtypes) that progress differently in both the preliminary and validation datasets. Identified cancer clusters were evaluated with analysis of variance to confirm statistically significant differences. The identified KIRC clusters were subjected to two feature selection analyses: (i) differential gene expression analysis, and (ii) Recursive Feature Elimination (RFE). The optimal features were subjected to Random Forest (RF) Classifier to evaluate the cluster prediction performance. The diagnostic capacity was evaluated using Cox regression and Kaplan-Meier. Additional enrichment analyses performed included Gene Ontology and Kyoto Encyclopedia of Genes and Genomes.

Results: Normalization with early-stage revealed the true heterogeneous gene expression that accumulates across the multi-stage cancer progression. The method allowed for an in-depth clustering based on the distinct cancer types as well as clusters (subtypes) within cancer types. The validation dataset revealed three clusters that progress differently, categorized based on patients' overall survival. A total of 231 differentially expressed genes were identified between all three clusters with a pairwise comparison approach, of which RFE selected a 48-gene subset. RF Classifier revealed a 100% cluster prediction performance. Five prognostic genes were identified of which the upregulation of genes *SALL4* and *KRT15* were associated with an unfavorable prognosis, and the upregulation of genes *OSBPL11*, *SPATA18*, and *TAL2* associated with a favorable prognosis.

Conclusions: The application of the normalization method provided an increased power of differentiating cancer samples based on how they progressed from early to advanced-stages of cancer development. The enhanced accuracy of hierarchical clustering revealed cancer heterogeneity and stratified patient samples into potential new cancer subtypes based on molecular patterns that were matched to phenotypic profiles. Additional genes responsible for cancer progression were discovered that could be of great importance for the development of new targeted therapies.

Declaration

I declare that **Discovering cancer subtypes by tracking cancer progression with transcriptomic data through the multi-stage process of cancer development** is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Michelle Chantel Livesey

Signature

M. Livesey

December 2023

Acknowledgements

I would like to convey a special message of gratitude to my supervisor, Dr Hocine Bendou, for allowing me the opportunity to pursue a PhD. The effort, time, and guidance you offered to help me grow as a research student was inspirational and novel. Your deep insight helped me at various stages of my research. I am also very grateful to Prof Alan Christoffels for his co-supervision role and support, and to Prof Renette Blignaut for her help with the statistics.

Thank you to Dr Nasr O M Eshibona and Dr Catherine Rossouw, for their assistance, advice, and valuable comments throughout the project. You are a true inspiration. I would also like to acknowledge the South African National Bioinformatics Institute for providing a highly stimulating environment and collaborative team to overcome the challenges along the journey.

I wish to thank my friends and family for keeping me grounded, motivated, and inspired. My parents, JJ Livesey, and LJ Livesey, thank you for the support through the years. I consider myself the luckiest in the world to have such a supportive family, standing behind me. To PJ Aucamp, thank you for always believing in me and supporting me throughout this endeavour.

I also would like to thank the National Research Foundation of South Africa for funding me during my PhD, and the South African Medical Research Council and the Cancer Research Trust, Faculty of Health Sciences, Start-up Emerging Researcher Award from the University of Cape Town for providing financial support to undertake this research project.

Thank you to our Heavenly Father who gave me the strength to persevere and see this project through.

Table of contents

List of tables

List of figures

Appendices

List of Abbreviations

ANOVA	-	An Analysis of Variance
BH	-	Benjamini-Hochberg
BM	-	Basement membrane
BP	-	Biological Processes
cBioportal	-	Cancer Bioportal
CC	-	Cellular Component
CESC	-	Cervical squamous cell carcinoma
DEG	-	Differentially Expressed Gene
DGE	-	Differential Gene Expression
DLBCL	-	Diffuse Large B-cell Lymphoma
DNA	-	Deoxyribonucleic Acid
ECM	-	Extracellular Matrix
ENSG IDs	-	Ensembl Gene Identifiers
FDA	-	Food and Drug Administration
GDC	-	Genomic Data Commons
GEO	-	Gene Expression Omnibus
GO	-	Gene Ontology
GTEx	-	Genotype-Tissue Expression
H_0	-	Null Hypothesis
H_1	-	Alternative Hypothesis
ICGC	-	International Cancer Genome Consortium
IS	-	Intermediate Survival
K-M	-	Kaplan-Meier
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
KIRC	-	Kidney Renal Clear Cell Carcinoma
LFC	-	log2-fold change
LIHC	-	Liver Hepatocellular Carcinoma
lncRNA	-	Long Non-Coding RNA
LS	-	Long Survival
LUAD	-	Lung Adenocarcinoma
MAF	-	Mutation Annotation Format
MF	-	Molecular Function
miRNA	-	Micro Ribonucleic Acid
ML	-	Machine Learning

NCI	-	National Cancer Institute
NGS	-	Next-Generation Sequencing
NHGRI	-	National Human Genome Research Institute
NIH	-	National Institutes of Health
ORA	-	Over Representation Analysis
OS	-	Overall Survival
PCA	-	Principal Component Analysis
PCAWG	-	Pan-Cancer Analysis of Whole Genomes
QTL	-	Quantitative Trait Loci
RCC	-	Renal Cell Carcinoma
RF	-	Random Forest
RFE	-	Recursive Feature Elimination
RNA	-	Ribonucleic Acid
RNA-Seq	-	RNA Sequencing
SNP	-	Single Nucleotide Polymorphisms
SS	-	Short Survival
SVM	-	Support Vector Machine
TARGET	-	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	-	The Cancer Genome Atlas
TF	-	Transcription Factor
TGCT	-	Testicular Germ Cell Tumors
TOM	-	Topological Overlap Matrix
UCSC	-	University of California, Santa Cruz
VCF	-	Variant Calling Format
WES	-	Whole Exome Sequencing
WebGestalt	-	WEB-based GEne SeT AnaLysis Toolkit
WGCNA	-	Weighted Gene Co-Expression Network Analysis
WGS	-	Whole-Genome Sequencing

Publications from this thesis:

1. Livesey, M., Rossouw, S. C., Blignaut, R., Christoffels, A., & Bendou, H. (2023). Transforming RNA-Seq gene expression to track cancer progression in the multi-stage early to advanced-stage cancer development. *PloS one*, 18(4), e0284458. https://doi.org/10.1371/journal.pone.0284458.

2. Livesey, M., Eshibona, N., & Bendou, H. (2023). Assessment of the progression of kidney renal clear cell carcinoma using transcriptional profiles revealed new cancer subtypes with variable prognosis. *Frontiers in genetics*. 14:1291043. https://doi.org/10.3389/fgene.2023.1291043.

Chapter 1

Introduction to thesis and research statement

1.1 General Introduction

Cancer is typically described as a genetic disease driven by oncogenic mutations (Ramón *et al*., 2020). At a cellular level, cancer is viewed as a multistep process, involving mutation and the selection of cells that have progressively increasing capacities for proliferation, invasion, survival, and metastasis. The first step in the process, tumor initiation, may arise from a genetic alteration leading to the abnormal proliferation of a single cell. Thereafter, tumor progression continues as additional mutations occur within cells of the tumor population (Cooper, 2000). Numerous mechanisms based on accumulated genetic changes are thus responsible for the initiation and tumor progression, thereby modifying the biology of the cells. Certain clones may therefore be more proliferative and result in rapid clinical progression and early relapse, whereas others may be less proliferative and associated with late relapse (Morgan *et al*., 2012). These dynamic and continuous changes in tumor development and adaptation in response to external pressure are characteristics of molecular heterogeneity (Crucitta *et al*., 2022).

Malignant tumors exhibit highly diverse phenotypic and molecular characteristics both at the intra-tumor (within a tumor) and inter-tumor (tumor by tumor) levels (Jamal-Hanjani *et al*., 2015). Intra-tumor heterogeneity describes solid tumors that may contain subpopulations of cells with distinct genomic alterations within the same tumor specimen (Fisher *et al*., 2013; Jamal-Hanjani *et al*., 2015). The latter, inter-tumor heterogeneity is a term used to describe tumor variations amongst patients. It is mainly characterized by distinct genetic alterations that arise in individual tumors originating in the same organ and enables the classification of these tumors into different molecular subtypes. A cancer type can thus have several subtypes with distinct morphological and phenotypic profiles, due to the heterogeneity of cancer.

Currently, there are limited targeted therapeutic options available for multiple cancer types, in part because of the substantial intra- and inter-tumor heterogeneity, as well as an incomplete understanding of the molecular mechanisms underlying tumorigenesis (Mkrtchyan *et al*., 2022). Tumor heterogeneity is one of the major factors influencing the effectiveness of patient treatment. Consequently, it is the primary cause of drug resistance, which further contributes to therapeutic failure (Crucitta *et al*., 2022). Tumor heterogeneity has thus presented a considerable challenge to match patients with suitable treatment strategies at the appropriate time; which poses a challenge to achieving the goals of precision medicine (McGranahan *et al*., 2015; McGranahan & Swanton, 2017). As a result, tumor heterogeneity is typically associated with poor prognosis and poor overall survival (OS) outcomes in cancer patients (Jamal-Hanjani *et al*., 2015; 2017; Mroz & Rocco, 2016; Lim & Ma, 2019; Tuasha & Petros, 2020; El Khoury *et al*., 2023).

The management and treatment of cancer patients have undergone significant advances in the field of oncology, with the departure from the "one-size-fits-all" strategy and towards a personalized, alternatively, precision medicine approach based on genomic variants (Malone *et al*., 2020). Cancer precision medicine is defined as "the use of therapeutics that are expected to confer benefit to a subset of patients whose cancer displays specific molecular or cellular features (most commonly genomic changes and changes in gene or protein expression patterns)" (Yates *et al*., 2018). Therefore, it aims to identify the unique biology of an individual or group of cancer patients sharing certain characteristics, and treat them by targeting the specific oncogenic event shared by these patients (Lipinski *et al*., 2016; Russnes *et al*., 2017; Ozturk *et al*., 2018; Zhang *et al*., 2019). Consequently, next-generation sequencing (NGS) and other profiling technologies have enabled advances in tumor analysis, which has been coupled with precision medicine. The molecular profiling of tumors facilitates the identification of unique deoxyribonucleic acid (DNA) changes and gene expression patterns that are associated with specific phenotypes and prognoses. Therefore, proper analysis can also reveal groups of patients into subcategories that yield clinically relevant diagnostic, prognostic, treatment response, or other clinical features (Malone *et al*., 2020).

An NGS-based approach, RNA-Sequencing (RNA-Seq), is a rapid and affordable methodology to track transcriptomic profiles across various cells or tissues (Wang *et al*., 2009). RNA profiling allows for the measurement and comparison of genome-wide gene expression patterns at an unparalleled level (Finotello & Camillo, 2015). The technique quantifies the number of transcripts (the basic unit of a gene), which in turn enables the analysis of multiple transcripts' expression, for a specific developmental stage or in different physiological or pathological conditions. The measurement of thousands of gene expression profiles allows for the discovery of altered gene expression levels of each transcript in a single cancer type for cancer molecular

classification. Deciphering the transcriptome is vital for interpreting the functional elements of the genome, exposing the molecular components of cells and tissues, and advancing the knowledge of the development and the disease (Wang *et al*., 2009). Additionally, gene expression can be associated with tumors having complex phenotypes, thus having the potential to expand our knowledge of the relationship between the transcriptome and the phenotypic profiles of cancer patients.

Molecular classification based on gene expression profiling from RNA-Seq is driving the development of precision medicine-targeted therapies. The technique allows for the sub-classification of tumors into gene expression signatures which can be integrated into clinical decision-making to facilitate informed optimal clinical care of cancer patients (Bi & Davuluri, 2020; Malone *et al*., 2020). Therefore, the identification of cancer subtypes aims to divide patients into subgroups with distinct molecular profiles with the additional potential of associating it with clinical phenotypes such as survival time. This can also be achieved by the application of hierarchical clustering of tumor samples based on gene expression profiles from high-throughput platforms that enable the molecular stratification of cancer patients into distinct tumor subtypes for numerous cancers (Gan *et al*., 2018; Rohani & Eslahchi, 2020; Puzanov, 2022; Zhang *et al*., 2023).

For decades, molecular classification of cancer has been a major area of study as it provides a foundation for biological research and is directly related to the development of tailored therapies for distinct subtypes. A clinically relevant subtype, therefore enables the selection and administration of the most effective treatment, as different cancer subtypes may respond differently to specific treatments. Hence, the stratification of cancer patients into subtypes is crucial, however, has been recognized as a challenging step towards individualized therapy

(Sun *et al*., 2022). In addition to guiding cancer treatment, the sub-classification of cancer patients has the potential to aid early cancer diagnosis, risk assessment, improved prognosis, predict drug response, or cancer surveillance and monitoring (Sarhadi & Armengol, 2022; Park *et al*., 2023).

1.2 Research statement and rationale

In recent years, progressive profiling technologies for tissue have accumulated diverse types of data, including gene expression profiling data of bulk tumors stored in various public databases (Creighton, 2018). For any major cancer type, expression data plays an important role in the identification of molecular subtypes, diagnosis, predicting patient outcomes, and identifying markers of therapeutic response. Consequently, genome-scale molecular data readily available in public domains serves as a resource that has revolutionized the fields of biology and precision medicine. Investigating the most efficient strategy to combine the multiple profiles of data is critical to facilitate the development of a computational tool to predict cancer subtypes (Zhao *et al*., 2023). The development of such a high-throughput genome analysis technique plays an important role in the clinical treatment of various cancer types.

A computational tool that can be applied to interpret the changing molecular characteristics of aggressive, progressing, and therapy-resistant tumors remains challenging (El-Deiry *et al*., 2017). More specifically, the establishment of novel and valuable methodologies to stratify patients for personalized treatment is also still under investigation (Ying *et al*., 2020). It has been recommended that patients' survival outcomes or prognosis should be closely linked with patient stratification methods, revealing a potential clinical application (Ying *et al*., 2020).

Therefore, this study proposed the development of a computational method that captures the heterogeneity between cancerous tumors by detecting their molecular differences in progression from early to advanced-stages of tumor development using gene expression by RNA-Seq.

The method examines the continuously changing cellular transcriptome, allowing for an efficient and comprehensive description of gene expression profiles between different conditions over time. Hence, it exposes the accumulated genetic changes that occur throughout the multi-stage of cancer development. Tracking cancer progression can improve the understanding of the molecular basis of tumorigenesis and alter our clinical approach to multiple cancer types. The application of the normalization method and hierarchical clustering will result in the discovery of novel cancer subtypes (clusters) that progress differently and further find genes responsible for cancer progression. Hence, the method facilitates the sub-classification of heterogeneous cancers and will also allow for the establishment of a genotype-phenotype link to the molecularly identified clusters and thus provide insight into clinical and phenotypic patterns of patient samples within the same cancer.

1.3 Aims and objectives of the thesis research project

The aims of the project were to:

(1) Discover cancer subtypes with the implementation of a computational method that normalizes late-stage cancer samples with early-stage samples to track the progression of tumors based on transcriptomic profiles.

(2) Application and validation of the computational method and discovery of novel cancer subtypes within the kidney renal clear cell carcinoma (KIRC) subtype.

A phased approach was adopted for the project: In the first part of the project, a preliminary study was conducted using multiple cancer types. The objectives of this preliminary study were to:

i. Retrieve multiple cancers RNA-Seq data from the University of California, Santa Cruz (UCSC) Xena database browser with corresponding phenotypic and survival profiles.
ii. Implement a computational method that normalizes advanced-stage cancer RNA-Seq expression profiles with early-stage.
iii. Subject the unnormalized and normalized gene expression to Weighted Gene Co-expression Network Analysis (WGCNA) to identify groups of genes with similar expression patterns.
i. Subject both unnormalized and normalized gene expression profiles to hierarchical clustering to reveal tumors that progress differently within and between the multiple cancer types.
ii. Apply a one-way analysis of variance (ANOVA) to compare and confirm differences in the mean gene expression profiles of the identified clusters.
iii. Match associated phenotypic and survival profiles to the identified cancer clusters.
iv. Perform WikiPathways, Kyoto Encyclopedia of Genes and Genome (KEGG), and Transcription factor (TF) enrichment analyses to link gene modules to cancer related proteins and pathways.

The second part of the study focused on the validation of the normalization method using transcriptomic profiles of a subtype of heterogeneous kidney cancer.

The objectives of the validation study were to:

i. Retrieve transcriptomic profiles of KIRC from the UCSC Xena database browser with corresponding phenotypic and survival profiles.

ii. A modified normalization method was designed that focused on one cancer type.

iii. Identify new cancer subtypes that are molecularly heterogeneous and progress differently during tumor development, from early to late-stages.

iv. The genotype-phenotype relationship of the distinct molecular clusters was defined by the average OS of the KIRC patient samples.

v. Use feature selection methods; differential gene expression (DGE) analysis and Recursive Feature Elimination (RFE) to select genes with the highest performance in sample classification.

vi. Perform survival analysis on the key feature selection genes using Cox regression and Kaplan-Meier (K-M) to identify prognostic genes.

vii. Apply machine learning (ML) techniques for sample classification using gene expression profiles derived from feature selection genes.

viii. Perform Gene Ontology (GO) and KEGG pathway enrichment analyses to illustrate the implication of the key genes in KIRC.

1.4 Thesis overview

Chapter 2: Literature review.

A literature review details the genomic and transcriptomic basis of disease, as well as the heterogeneous nature of cancer, the available bioinformatics resources, and a multi-omics approach to disease.

Chapter 3: Transforming RNA-Seq gene expression to track cancer progression in the multi-stage early to advanced-stage cancer development.

A normalization method was established to track the progression of tumors, based on transcriptional profiles from early to late-stage cancer development. Thus, the method exposes the accumulated genetic changes that occur throughout the multi-stage of cancer development. This computational methodology was applied *in silico* to multiple cancer types. The clustering of the normalized gene expression allowed for in-depth segregation based on the distinct cancer types as well as clusters (subtypes) within the cancer types.

Chapter 4: Investigating the progression of kidney renal clear cell carcinoma transcriptional profiles to identify cancer subtypes.

To validate the newly developed computational method, an *in silico* analysis was performed on the heterogeneous subtype of kidney cancer, kidney renal clear cell carcinoma. A total of eighty-two KIRC transcriptomic profiles were subjected to the normalization method. Clustering of the normalized gene expression revealed three groups (subtypes) with differently evolving gene expression profiles. The genotype-phenotype relationship to the distinct clusters was defined by the average overall survival of the KIRC patient samples, categorized into short, intermediate, and long survival. The results of the study could lead to a more accurate prognosis, while the biomarkers identified could serve as targets to provide a more effective treatment strategy.

Chapter 5: Conclusions and future prospects.

Chapter 2

Literature Review

2.1 Introduction to high-throughput methodologies

Over the past few decades, many species' genomes have been mapped in an effort to gain a deeper understanding of biological processes at the molecular level. In the mid-1990's, microarray technology was introduced, which measured the abundance of a set of predetermined sequences via their hybridisation to an array of complementary probes (Schena *et al*., 1995). This allowed for a genome-wide analysis in a single experiment. High-throughput methods, like microarrays, have since advanced and gained widespread use as instruments for the investigation of numerous biological processes.

Research that focuses on genome-wide gene expression aims to identify and characterize genes involved in various processes. The goal of a healthcare application would be to identify the genes that change gene expression levels during the infection of a pathogen. This can serve as potential biomarkers that can be used for accurate diagnosis, risk stratification, improved

prognosis, an understanding of therapeutic response, or lead to a more effective therapeutic approach for a specific disease. Alternatively, a research-orientated application would be to characterize the function of the genes to build a model for a biological process.

Gene expression can be investigated either by quantifying the amount of ribonucleic acid (RNA) or the number of proteins. High-throughput methods, such as transcriptomics and proteomics, are available for both methodologies. Similar analyses are also available for studies on metabolites (metabolomics) and research that focuses on the protein's interaction with DNA. Other applications of high-throughput methodologies include the identification of mutations that may cause a disease or increase the risk of disease development. Deep sequencing (also referred to as NGS) has also been established. The development of this high-throughput technology enabled the determination of DNA- or RNA sequencing (RNA-Seq) and the RNA expression on a genome level and therefore increased the volume of information acquired in the respective experiments (Wang *et al*., 2009; D'Argenio, 2018).

RNA-Seq serves as key contemporary tool, that uses high-throughput sequencing to capture all sequences. It is a cost-effective technique that enables a comprehensive understanding of the transcriptome landscape (D'Agostino *et al*., 2022). The technique is thus the method of choice for examining tissue-level transcriptome changes. This powerful screening tool has improved transcriptome analysis in both qualitative and quantitative ways, due to its limitless dynamic range (D'Agostino *et al*., 2022). Table 2.1 below captures a comparison of the differences in RNA-Seq platforms with more details regarding to their features, specifications, and technologies.

Table 2.1: Comparison of five principal RNA-Seq platforms and technologies. The primary application of these technologies is RNA-Seq. This table includes some features and details about each platform (Jazayeri et al., 2015).

Features	454, Roche	Ion Torrent	Illumina	ABI SOLiD	Pacific Bio
Sequencing chemistry	Pyrosequencing, Chemiluminescence	Ion semiconductor	Polymerase-based sequence-by-synthesis	Sequencing by ligation	Single Molecule Real Time
Sequencing method	incorporation of normal nucleotides	measuring pH change	incorporation of fluorescent nucleotides	fluorescent short linkers	Incorporation of fluorescent Nucleotides
Sequence yield per run	0.6 -1 Gb	1 Gb	1- 60 Gb	3 Gb	0.3-0.5 Gb
Time per run	7 hours	2 hours	1-10 days	5-14 days	10 h
Read length	700 bp	400 bp	50 to 250 bp	50+35 or 50+50 bp	5,000 bp average; maximum read length ~22,000 bases
Input run type library	SE, PE, Mx	SE, PE, Mx	SE, PE, MP, Mx	SE, MP, Mx	SE

The data generated by the high-throughput techniques described above are similar in nature. These techniques enabled genome-wide analyses, as they provide information on the full set of all potential variables in an individual. As a result, a vast number of variables, ranging from hundreds to millions, are generally investigated. Variability is often introduced into the data, due to the complex experimental procedures. This variation needs to be eliminated to derive biologically meaningful insights and conclusions. Two data analysis processes are typically employed. First, pre-processing which aims to remove the technical variations and, second downstream analysis, which includes all additional analyses carried out to address the biological question, such as statistical analysis.

The general aim of these investigations is primarily to find the variables that are different between two experimental groups, such as which genes, proteins, or metabolites are different when comparing infected tissue with uninfected tissue, patients and healthy individuals, or virulent and nonvirulent bacterial strains. The direct comparison of results between experiments is facilitated by the principle of RNA-Seq. RNA-Seq enables the determination of the absolute quantity of every molecule in a cell population (Wang *et al*., 2009). Therefore, it allows researchers to measure and investigate levels of gene expression over time, to further assess the function of the genes, and find targeted treatment, or potential virulence factors. Other studies seek to categorize a disease into subtypes and find the genes that differ between the subtypes (Zhang *et al*., 2017; Chen *et al*., 2020; Ding *et al*., 2023). Generally, these studies would implement a hierarchical clustering algorithm to analyse the samples and variables. A more clinical application can be to predict diseases or strains of pathogens based on gene expression using classification methods. To achieve this, either expression from all genes or a subset of genes is employed.

In this investigation, changes in gene expression levels were assessed to understand the progression of the multi-stage cancer development using data from high-throughput RNA-Seq. A novel pre-processing computational method was developed and evaluated with respect to their performance in downstream analysis (See Chapter 3).

2.2 Cancer as a disease

Cancer is a broad category of genetic diseases that are currently classified by their primary site of origin, such as brain cancer and breast cancer (Zhao *et al*., 2019). Hence, the term "cancer" refers to over 277 different types of cancer diseases (Hassanpour & Dehghani, 2017). The disease is the most intractable medical and health challenge in the world, accounting for approximately 10 million deaths in 2020 (Ferlay *et al*., 2021). Therefore, cancer is a major problem that has an impact on the health of all human societies. The disease exhibits variability at the tissue level, which poses significant challenges for both specific diagnoses, followed by the efficacy of treatments (Meacham & Morrison, 2013; Fisher *et al*., 2013). In men, the highest percentages of cancer types occur in the prostate, lung, stomach, liver, colon, and rectum. In women, cancer prevalence is highest in the breast, lung, cervical, thyroid, non-melanoma skin, and ovary. This data indicates that the majority of cancers in men and women are prostate and breast cancers, respectively (Ferlay *et al*., 2021). For children, the most common cancers are blood cancer, cancers related to the brain, and central nervous system cancer (Wu *et al*., 2022).

Researchers have discovered different stages of cancer, suggesting that numerous gene alterations have a role in cancer pathogenesis. Cancer thus occurs by a series of continuously accumulating gene mutations that alter cell activities. These gene mutations lead to abnormal cell proliferation (Hassanpour & Dehghani, 2017). Consequently, cancer is a dynamic and

complex disease, that generally becomes more heterogeneous as the disease progresses (Meacham & Morrison, 2013; Dagogo-Jack & Shaw, 2018). As a result, different cancers may present different gene expression levels at different stages of the disease that affect the prognostic characteristics (or survival patterns) of a patient. Thus, gene expression data have similar survival-related characteristics, in which some tumours may be fast-growing and can cause mortality soon after diagnosis, while other tumors grow gradually and slowly.

2.2.1 Tumour heterogeneity

Tumour heterogeneity refers to the existence of subpopulations of cells, with unique morphological and phenotypic profiles that may harbour diverse biological behaviours within a primary tumour as well as its metastases (Fisher *et al*., 2013). This phenomenon is also referred to as intra-tumour heterogeneity. This in turn can lead to inter-tumor heterogeneity. Heterogeneity thus describes the differences among cancer cells both within tumors (intra-tumor heterogeneity) and between tumors (inter-tumor heterogeneity) (Figure 2.1) (Fisher *et al*., 2013; Proietto *et al*., 2023). Therefore, it refers to cancer cells describing variations in morphology, transcriptional profiles, metabolism, and metastatic potential.

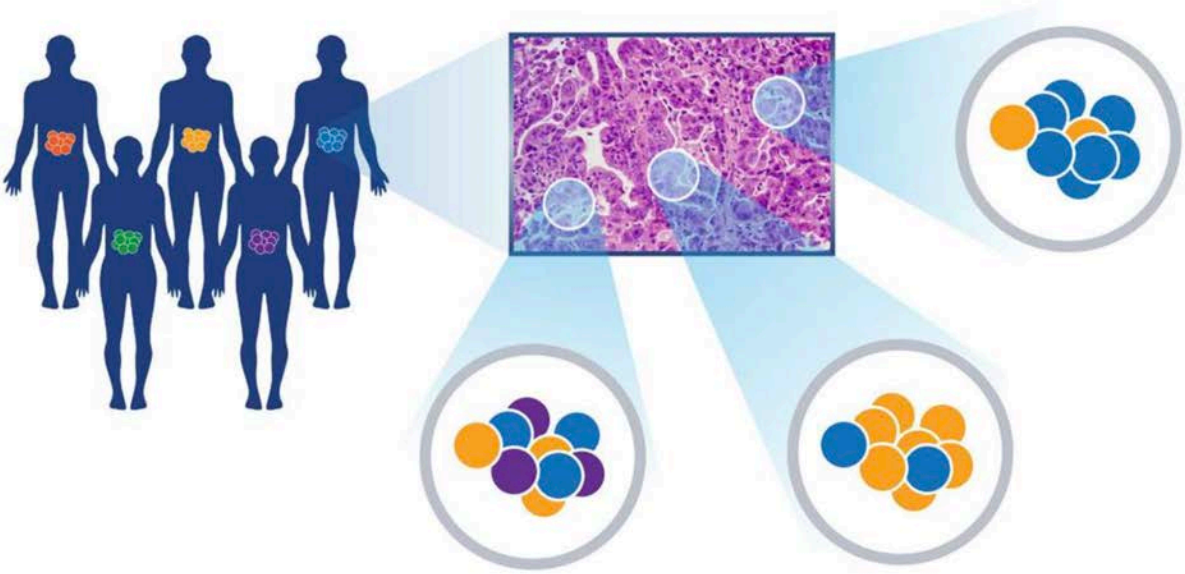

Figure 2.1: An illustration of intra-tumor and inter-tumor heterogeneity (Venkata, 2019).

One of the most challenging behaviours in cancer ecosystems is heterogeneity (Proietto *et al*., 2023), which has been discovered in the majority of tumors. This includes leukemias (Eshibona *et al*., 2023), breast (Parker *et al*., 2009), prostate (Kaffenberger & Barbieri, 2016), kidney (Zhong *et al*., 2021), colorectal (Singh *et al*., 2019), brain (Friedmann-Morvinski, 2014), esophagus (Li *et al*., 2020), head and neck (Canning *et al*., 2019), bladder (Lavallee *et al*., 2021) and gynecological carcinomas (Fujii *et al*., 2000). Heterogeneity promotes tumor resistance, more aggressive metastasis, and recurrence and is one of the major factors limiting the long-term efficacy of solid tumor therapy (Proietto *et al*., 2023). Hence, tumor heterogeneity thus provides the fuel for drug resistance (Dagogo-Jack & Shaw, 2018). However, the functional relevance of genomic heterogeneity in tumor progression and therapy resistance remains poorly understood (Marusyk *et al*., 2020).

An accurate assessment and characterization of tumor heterogeneity has the potential to advance the understanding of the causes and progression of the disease. In turn, this could serve

as guidance for the development of more advanced treatment plans that recognise the magnitude and prevalence of intra- and inter-tumor heterogeneity to yield higher efficacy.

2.2.2 Cancer Subtyping

Cancers are traditionally classified four ways: (i) primary site of origin i.e lung or liver cancer; then by (ii) histotype, and (iii) grade according to WHO classifications; and (iv) finally by spread according to the Tumor Node Metastasis system. However, this only partially captures the true heterogenic characteristics of cancer. Therefore, the World Health Organization classifications began to include molecular–genetic features of tumors, starting from the third edition in 2000 (Carbone, 2020). Molecular subtyping of cancer, as the name suggests, is a new approach to group cancers according to molecular data and classification models. For example, breast cancer is highly heterogeneous and over the years multiple molecular subtypes have evolved. Currently, four subtypes of breast cancer are widely recognized: luminal A, luminal B, HER2-positive, and triple-negative (Orrantia-Borunda *et al*., 2022). Thus, patients with different cancer subtypes often have unique groups of genomic and clinical characteristics due to the high heterogeneity and complexity of malignancies (Zhao *et al*., 2023). Molecular classifications of cancer thus rely on biomarkers and classifiers, in contrast to the traditional histological classification (Zhao *et al*., 2019). Therefore, different molecular approaches access the potency of gene expression and defective proteins, as well as the identification of novel cancer biomarkers. These discoveries can be useful to treat cancers and reduce cancer complications.

High-throughput sequencing technologies have enabled the capturing of comprehensive profiles of tumor samples at multiple levels and allow for deep phenotyping of patients. These

recent advances in technology have accelerated the increasing availability of multi-omics data for the purpose of cancer subtyping. The identification of cancer subtypes is crucial to facilitate cancer diagnosis, prognosis and selection of effective treatment. Therefore, it is vital to take advantage of the complimentary information from multi-omics data, and develop computational models that can characterize and integrate different data layers into a single framework (Zhao *et al*., 2023).

2.2.3 Cancer and Representative Signaling Pathways

Cancer-associated genetic abnormalities have been well documented since the early identification of oncogenes and tumor suppressor genes (International Human Genome Sequencing Consortium, 2004, ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Nowadays, it is widely acknowledged that signaling pathways and molecular networks play crucial roles in carrying out and regulating important pro-survival and pro-growth cellular processes. As a result, they are primarily responsible for the onset of cancer as well as potential treatments (Yip & Papa, 2021).

Several important signaling pathways have been identified as frequently genetically altered in cancer, including the RTK/RAS/MAP-Kinase pathway, PI3K/Akt signaling, amongst others (Vogelstein & Kinzler, 2004). Members of these pathways and their interactions have been captured in numerous pathway databases. The genes in key pathways are not altered at equal frequencies, with some genes frequently altered and well-known in cancer, whereas others are only rarely or never altered. Also, the alteration to specific pathways, such as RTK-RAS signaling or the cell-cycle pathway, occur at high frequency across many different tumor types, while other pathways are altered in more specific subsets of malignancies (e.g., alterations in

the oxidative stress response pathway are strongly associated with squamous histologies). Identifying the relationships of inter- and intra-pathway recurrence, co-occurrence or mutual exclusivity in various cancer types can aid in understanding the functionally relevant processes of oncogenic pathway alterations that may guide therapeutic approaches (Sanchez-Vega *et al*., 2018).

2.3 Omics Research

Advancements in technology have allowed for the collection of large quantities of molecular measurements within a tissue or cell. These technologies can be applied to a biological system of interest and reveal the underlying biology at a resolution that has never been attainable. Generally, the scientific fields associated with measuring such biological molecules in a high-throughput manner are known as omics.

Omic analysis includes different branches and categories of research. Examples include genomics (Hasin *et al*., 2017), epigenomics (Esteller, 2007), transcriptomics (Sager *et al*., 2015), proteomic (Aslam *et al*., 2017), and metabolomics (Pinu *et al*., 2019) that corresponds to the global analyses of genes, methylated DNA or modified histone proteins, RNA, proteins, and metabolites, respectively. These studies produce a large amount of data that has enabled the characterization of molecular features and provided evidence of disease diagnosis in multiple human diseases (Subramanian *et al*., 2017). However, single omics research can only provide a limited degree of understanding and thus, a combination of these studies within a suitable statistical and mathematical framework can assist in solving broader queries related to both basic and applied fields of biology.

2.3.1 Multi-omics approach to diseases

Multi-omics seeks to combine two or more omics strategies to aid in data analyses, visualization, and interpretation (Brademan *et al*., 2020, Krassowski *et al*., 2020). This method provides important insights into the flow of biological information at multiple levels and can thus reveal the mechanisms underlying the biological condition of interest (Subramanian *et al*., 2020). Therefore, multi-omics efforts have revolutionized biomedical research and are now a standard method for carrying out biological research. These integrated approaches further hold significant promise for complex diseases such as cancer. The complexity of cancer research can thus be enhanced by multi-omics research and improve the accuracy of cancer diagnosis and prognosis (Iorio *et al*., 2016, Pettini *et al*., 2021).

Multi-omics has the potential to find novel associations between biological entities, aid in biomarker discovery, and build an elaborate concept between the disease and physiology. Additionally, multi-omics helps in coherently matching genotype-to-phenotype relationships. The robust understanding of genotype-to-phenotype correlations in applied multi-omics could improve healthcare facilities by increasing the diagnostic yield for health, improving disease prognosis, and thus establishing a standard for excellence (Krassowski *et al*., 2020, Subramanian *et al*., 2020). Future research will be greatly aided by combining multi-omics resources and bioinformatics techniques to gain knowledge from existing data.

2.4 Transcriptome profiling

In the last decades, transcriptome profiling has been one of the most utilized approaches to understanding human diseases at the molecular level (Casamassimi *et al*., 2017). The term

transcriptome refers to the full range of RNA molecules expressed by a cell, tissue, or organism during a particular physiological condition or developmental stage (D'Agostino *et al*., 2022). Detailed knowledge of the transcriptome is essential for understanding genomic processes, and identifying the molecular compositions of cells, as well as the cause and progression of diseases (Wang B *et al*., 2019). The study of transcriptomics is also referred to as gene expression profiling.

2.4.1 Gene expression profiling technique

The experimental methods for obtaining gene expression profiles have rapidly advanced from measuring a small number of transcripts with microarrays, to a large number of transcripts with the more contemporary RNA-Seq technique. This approach has advantages in almost every field of life sciences and is currently being adopted for clinical purposes (Szalat *et al*., 2016; Blok *et al*., 2018; Borisov *et al*., 2020).

RNA-Seq enables the characterization of the average expression profiles for individual samples (Mortazavi *et al*., 2008; Wang *et al*., 2009; Metzker, 2010). This is achieved by expression profiling once the sequencing of a genome is completed. Essentially, gene expression profiling measures the expression level of all targeted RNA transcripts. Hence, it provides a snapshot of the transcriptional activity in a biological sample and reveals the underlying molecular processes occurring (D'Agostino *et al*., 2022). Expression profiles from various conditions can thus be compared to find expression signatures to describe a condition of interest such as a tissue type, a disease, or a treatment response. Therefore, it facilitates the discovery of molecular functions linked to genes with condition-specific differential expression.

2.4.2 Gene quantification

An RNA-Seq experiment is conducted with the extraction of a targeted RNA population from biological samples (Kukurba & Montgomery, 2015). These RNAs are fragmented into shorter sequences suitable for high-throughput sequencing platforms, transformed into cDNA, and finally ligated with sequencing adapters. The adapter-ligated fragments can then be read by a sequencer. After the fragments have been sequenced, RNA-Seq data begins to exist. These fragments must be assigned to the genomic features from which they originated, to assign a read count value per genomic feature. This process is known as quantification (Figure 2.2).

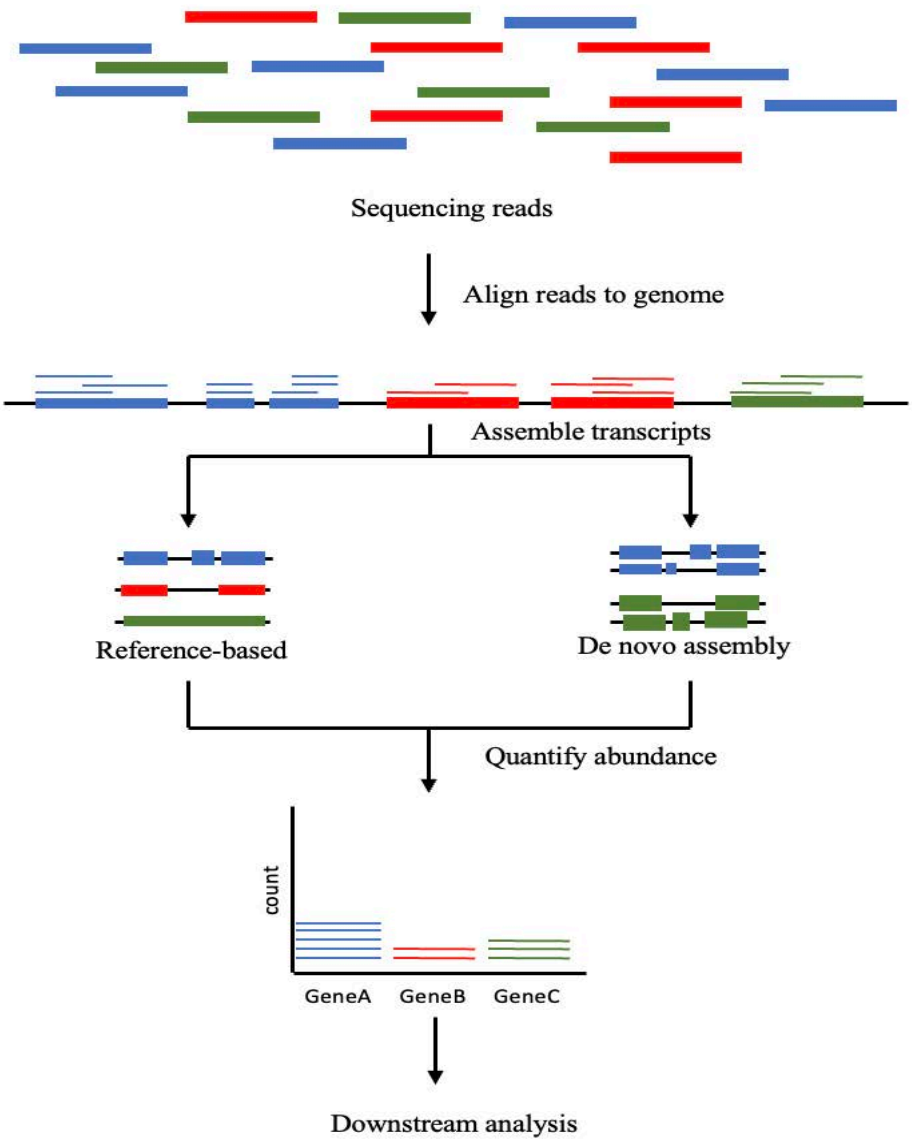

Figure 2.2: Quantifying transcription levels. In a typical RNA-Seq investigation, the reads are first aligned to a reference genome, after which the reads may be assembled into transcripts by either using reference transcript annotations or de novo assembly methods. The expression level of a single gene is determined by counting the number of reads that align to an exon or full-length transcript. Next, downstream analysis with RNA-Seq data can be performed (Adapted from Kukurba & Montgomery, 2015).

The expression level of each RNA unit is based on the number of sequenced fragments mapped to the gene or transcript, which in turn is expected to directly correlate with its abundance level (Rapaport et al., 2013). Once quantified, the expression profiles from individual assays are merged to create a matrix X with features as rows and samples as columns. Also referred to as a gene-by-sample matrix. Each element of the matrix Xij represents the raw read count of feature i in sample j.

Subsequently, changes in the gene expression profiles between samples can be established. Therefore, the purpose of a gene quantification investigation is to recognize the changes that occur under various experimental conditions, in disease states, and response to medical treatments. Therefore, RNA-Seq expression profiles have the potential to result in expression signatures that aid in the understanding of disease mechanisms and the development of clinically relevant biomarkers (Goossens et al., 2015; Bhowmick et al, 2019), or machine learning models that enhance the quality of data and medical care.

2.5 Normalization

Unintentional experimental errors are frequently introduced into RNA-Seq data by sequencing technology. To counteract this, a mathematical adjustment known as normalization is

standardly used to reduce the non-biologically derived variability present in transcriptome measurements. Therefore, normalization corrects for systemic biases introduced during sample processing and data generation and makes gene expressions directly comparable within and between samples.

The method involves adjusting data from one domain to another so that the results are relatively normally distributed. When applied to numerical data, normalization converts the numbers to a common scale without distorting the underlying differences (Chawade *et al*., 2014). Other methods include min-max, z-score, TPM (transcripts per million), RPKM (reads per kilobase million), and quantile, among others (Bolstad *et al*., 2003, Baumgartner *et al*., 2011, Roy *et al*., 2019, Quackenbush, 2002, Anders & Huber, 2010, Oshlack & Wakefield, 2009, Wagner *et al*., 2012).

The above normalization methods can be computed with R code, while DESeq and TMM (Trimmed Mean of M-values) normalization is implemented in the DESeq and edgeR Bioconductor packages, respectively (Robinson *et al*., 2010, Anders & Huber, 2010; Love *et al*., 2014). The two methods use a combination of mathematically based and biologically based normalizing strategies and are most frequently used for differential gene expression analysis. The selected method depends on the following factors: (i) the type of genomic data, (ii) the platform originally used to collect the data (iii) the scale of the data, and (iv) the intended downstream analyses. Normalization is critical to accurately interpret the results of genomic and transcriptomic investigations (Abrams *et al*., 2019).

2.6 RNA-sequencing: Application in cancer research

The in-depth analysis of RNA-Seq and comprehension of gene expression have facilitated the interpretation of diseases and their genetic causes at the molecular level. Therefore, it allows for the identification of different cancer types as well as rare diseases. The method further offers a tool that can identify the genetic and epigenetic cause of cancer, and thus aid in better therapy by identifying resistant genes and defining gene mutations as cancer biomarkers (Hong *et al*., 2020).

The method is thus important for accurate cancer diagnosis, shedding light on the development of more effective treatments and more specifically offering targeted therapy (Ergin *et al*., 2022). Additionally, normal tissues and cells can be compared to abnormal conditions to track and reveal the cause of various diseases and identify metabolic abnormalities or alterations occurring at the molecular and cellular levels (Ergin *et al*., 2022).

2.6.1 Transforming RNA-Seq to track cancer progression

The use of RNA-Seq data for disease assessment is growing, and normalization (*section 2.5*) is generally accepted to be a necessary step in order to generate comparable samples. However, a study found that raw data may perform better in capturing more original transcriptome patterns in different pathological conditions (Han & Men, 2018). Therefore, this study developed a computation method to adjust or transform raw count RNA-Seq gene expression profiles to provide more meaningful biological information. The computational method was developed to track gene expression changes that occur throughout the multi-stage development

of cancer. The rationale of this approach can be illustrated by a bar graph of a single raw count gene expression profile from the same cancer type (Figure 2.3).

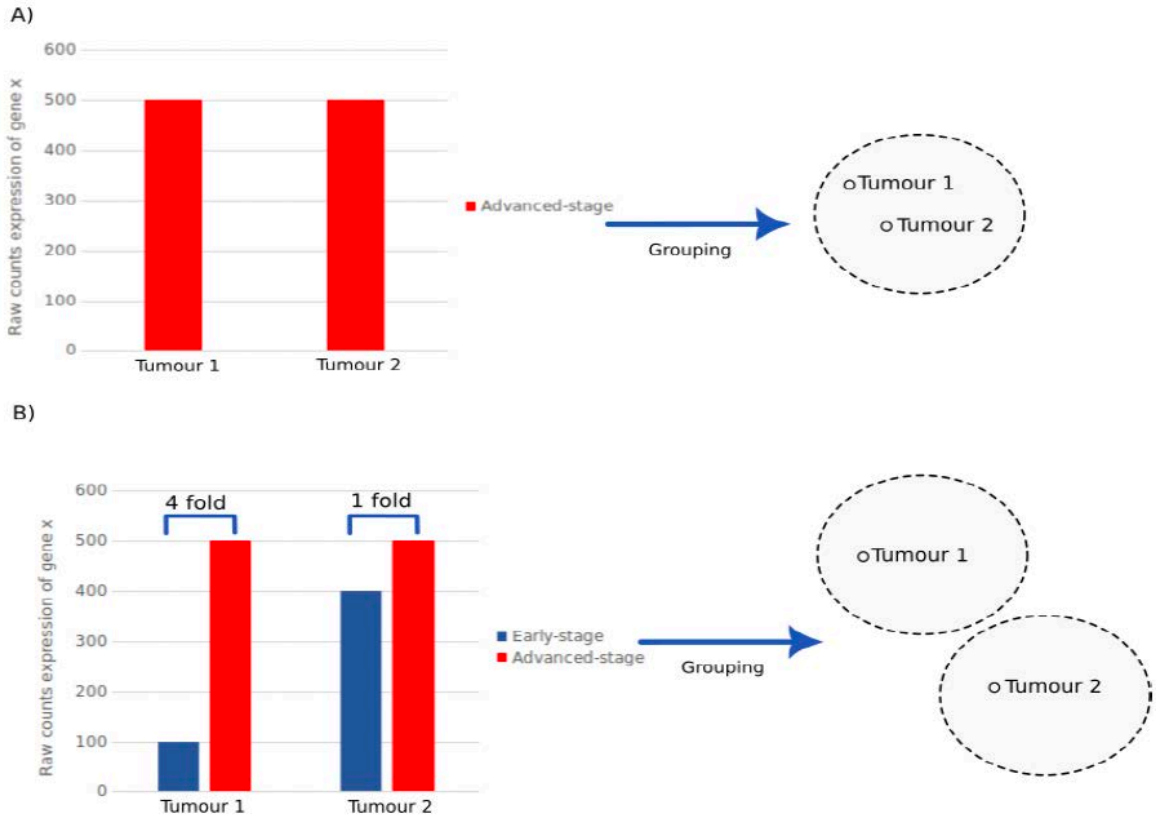

Figure 2.3: Raw RNA-Seq data of advanced-stage and early-stage gene expression of gene *x* in two tumor types. Tumour 1 and tumour type 2 show a gene expression fold increase of 4 and 1, respectively, from early to advanced-stage cancer. (Adapted from Livesey *et al*., 2023).

For gene *x*, the red and blue bars represent advanced-stage and early-stage cancer gene expression profiles, respectively (Figure 2.3). It can be noted that gene *x* reveals identical advanced-stage expression profiles in both distinct tumour types. Therefore, these tumour types will group together based on transcriptional profiles (Figure 2.3A). However, when considering the early-stage gene expression profiles in both tumour types, it can be noted that there is a significant difference in the expression levels between advanced-stage and early-stage. Tumour type 1 illustrates a greater difference than the expression levels in tumour type 2. Therefore, a

computational method that corrects for genes that display less expression variability in advanced-stage cancer samples but display a high variability in early-stage cancer samples, and grouping of the normalized output will allow for the segregation of the heterogeneous tumour types (Figure 2.3B).

The computation method in this study detects the accumulated genetic changes when gene expression levels in advanced-stage are less variable but display high variability in early-stage, by calculating the quotient of cancerous samples (dividend) and early-stage samples (divisor) (Livesey *et al*., 2023). The method produces 'normalized' differential RNA gene expression within a specific condition, therefore representing the continuously changing cellular transcriptome in which two distinct tumour types or subtypes can be differentiated based on the differences in the progression of gene expression profiles in the multi-stage cancer development. This enables a more efficient and comprehensive description of heterogeneous gene expression profiles.

2.7 Hierarchical Clustering

The molecular patterns can be further explored with hierarchical clustering analysis to reveal unique gene expression patterns. An algorithm referred to as hierarchical clustering organizes similar objects into groups called clusters. The sole concept of hierarchical clustering lies in the creation and evaluation of a dendrogram output. A dendrogram is a tree-like structure that shows the hierarchical relationship among all the data points. As a result, the endpoint is a set of clusters, where each cluster is distinct from other clusters, and the objects within each cluster are generally similar to each other. Numerous research studies have focused on clustering cancer patient samples based on gene expression profiles (Alon *et a*l., 1999; Ma *et al*., 2009;

De Souto *et al*., 2008; Yu *et al*., 2017, Cao *et al*., 2021; Xing *et al*., 2022). Consequently, the application of clustering analysis has successfully been used to identify novel cancer subtypes based on high-dimensional RNA-sequencing data from samples taken from cancer patients (Vidman *et al*., 2019).

2.8 Biomarkers

The emergence of genomics and advances in molecular biology have allowed for a promising era of biomarker research. The Food and Drug Administration (FDA) in collaboration with the National Institutes of Health (NIH) Joint Leadership Council described a biomarker as "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention" (FDA-NIH Biomarker Working Group, 2016). The most common experimental approach for identifying biomarkers is to compare diseased samples with control samples

2.8.1 Transcriptional Biomarkers in Cancer

A tumor may consist of a diverse collection of cells, each with its own unique molecular signatures, due to the high degree of genetic heterogeneity. A cancer biomarker is thus any quantifiable molecular indicator of cancer risk, occurrence of cancer, or patient outcome (Sarhadi & Armengol, 2022). This process involves the profiling of tumors to detect changes in DNA, RNA, proteins, or other biomolecules. Cancer biomarkers have a wide range of useful healthcare applications, including cancer risk assessment, screening of disease and early detection, cancer diagnosis, patient prognosis, the prediction of response to therapy including the safety and toxicity of therapeutic regimen, and cancer monitoring (Sarhadi & Armengol,

2022). Their ultimate goal is to achieve precision medicine to enhance the prevention, screening, and treatment approaches of cancer.

Biomarkers are classified into seven categories; susceptibility/risk, diagnostic, predictive, prognostic, monitoring, pharmacodynamic/response, and safety. A susceptibility/risk biomarker can indicate the potential for developing a disease in an individual who does not currently have clinically apparent disease, while a diagnostic biomarker detect the presence of a disease or identify a subtype of the disease (Califf, 2018). Predictive biomarkers provide information about clinical outcomes based on treatment decisions, while prognostic biomarkers provide information about the probable course of the disease, including its recurrence, progression, and patient's OS, irrespective of the treatment (Ballman, 2015; Sarhadi & Armengol, 2022). A monitoring biomarker can be frequently measured to assess disease status or for evidence of exposure to a medical product or an environmental agent. They are thus useful for measuring the pharmocodynamic effects, to detect early evidence of a therapeutic response and detect complications of a disease or therapy. Conversely, a pharmacodynamic biomarkers are those whose levels alter in response to exposure to medical products or environmental agents. Lastly, a safety biomarker is measured before or after exposure to medical intervention or an environmental agent to determine the likelihood, presence, or extent of a toxicity as an adverse event (Califf, 2018). Therefore, biomarker discovery is advancing the understanding of disease pathogenesis, providing novel targets for disease characterization, and early diagnosis, and improving targeted therapy to facilitate personalized treatment that benefits a patient based on their unique profile (Novelli *et al*., 2008).

A cancer biomarker is a characteristic that is measured as an indicator of cancer risk, cancer occurrence, or patient prognosis. These characteristics can be either molecular, cellular,

physiologic, or imaging-based. Cancer biomarkers that are frequently research include *AK2* gene mutation, which aid in the diagnosis of certain types of leukemia, whereas *BRCA1* and *BRCA2* gene mutation help in the treatment of ovarian and breast cancers. A *DPD* gene mutation helps predict the risk of a toxic reaction to 5-fluorouracil therapy in breast, colorectal cancer, gastric, and pancreatic cancer. Meanwhile, the *HE4* biomarker helps with ovarian cancer therapy planning, disease progression assessment, and recurrence monitoring (Sarhadi & Armengol, 2022). Therefore, the identification of novel molecular biomarkers has the potential to improve personalized disease prevention and management, therefore, resulting in a more precise, safe, and cost-effective healthcare outcome, ultimately improving patient health outcomes. Accordingly, a new era of precision and personalized cancer therapeutics has been brought about as biomarker discovery has led to the development of drugs targeting tumor-specific biomarkers in a subgroup of patients (Moore & Guinigundo, 2023). Continuous advances in precision oncology are needed for the development of novel cancer biomarkers with increased sensitivity, specificity, and positive predictive value.

2.9 Multi-omic resouces

Massively parallel sequencing technology has generated an increasing amount of complex cancer genomic data, providing a need for large repositories and databases to store this data. The cancer recourse community further requires user-friendly data-centric tools for data visualization and interpretation.

Numerous resources are available to explore for integrated multi-omics research. Most of these resources are publicly available and can be queried, utilized, and studied without restrictions for the purpose of reproducibility, discovery, and validating results (Yang *et al*., 2015;

Pavlopoulou *et al*., 2015). Globally, several data types are being curated in bioinformatics resources. These data types are stored in different file formats and can be retrieved from relevant cancer data repositories such as UCSC Xena browser (Goldman *et al*., 2020), Cancer Bioportal (cBioportal) (Gao *et al*., 2013), The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and Gene Expression Omnibus (GEO), among others.

2.9.1 UCSC Xena Browser

UCSC Xena (UCSC Xena; http://xena.ucsc.edu) is a high-performance resource for visualizing and exploring multi-omic data from large public repositories and private datasets (Goldman *et al*., 2020). The platform comprises of two components: the front-end Xena Browser and the back-end Xena Hubs. The web-based Xena Browser (UCSC Xena Browser; https://xenabrowser.net) empowers biologists to easily explore data across multiple open-public Xena Hubs, while Xena Hubs securely hosts genomics data from laptops, public servers, or the cloud (Goldman *et al*., 2013, 2015, 2020).

Xena focuses on cancer genomics and showcases more than 1600 datasets across 50 types of cancer. Significant cancer genomics datasets include TCGA (Chin *et al*., 2011), International Cancer Genome Consortium (ICGC), TCGA Pan-Cancer Atlas (Hoadley *et al*., 2018), Pan-Cancer Analysis of Whole Genomes (PCAWG) (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), Genomic Data Commons (GDC) (Grossman *et al*., 2016), and more. The platform also hosts results from the UCSC Toil RNA-Seq Recompute Compendium which uniformly realign gene and transcript expression data from all TCGA, TARGET, and Genotype-Tissue Expression (GTEx) (GTEx Consortium, 2017) samples to

enable users to compare gene and transcript expression from these datasets (Vivian *et al*., 2017).

Additional analyses and visualization tools available through Xena Browser include dynamic K-M survival analysis, powerful filtering and subgrouping, box plots, scatter plots, and statistical tests (Goldman *et al*., 2020). It further supports a wide range of data types, including clinical data such as phenotypic and survival information (Goldman *et al*., 2020). Other data types include somatic and germline single nucleotide polymorphisms (SNPs), indels, large structural variants, copy-number variation, gene, transcript, exon, protein or micro RNA expression, DNA methylation, and ATAC-seq peak signals (Cieślik & Chinnaiyan, 2018, Langmead & Nellore, 2018).

The many unique features, broad data type support, high performance, easy and secure view, and open access to public and private data differentiate Xena from other genomic tools (Goldman *et al*., 2020). Several recent bioinformatics studies have used data (Zhu *et al*., 2019; Giwa *et al*., 2020; Hu *et al*., 2021; Eshibona *et al*., 2022; Song *et al*., 2022), published data (Kang *et al*., 2020), and made visualizations from or on the Xena browser (Chen *et al*., 2019; Zheng & Fu, 2020; Zhang *et al*., 2020; Wang *et al*., 2021; Jin *et al*., 2021).

2.9.2 The Genotype-Tissue Expression portal

The GTEx (GTEx; https://gtexportal.org) project is an ongoing effort to create a comprehensive open-access resource for the scientific community. Building the GTEx project, the initiative aimed to establish a molecular and data analysis resource, and a tissue bank to study human gene expression and regulation, and its relationship to genetic variation. Tissue-specific

regulation of gene expression levels is obtained from multiple healthy reference neonatal, pediatric, and adolescent tissues. The data types include gene expression levels across numerous 'normal' (non-diseased) human tissues, quantitative trait loci (QTL), and histology images (GTEx consortium, 2015 and 2017).

The GTEx database presents a sample collection from 54 non-diseased tissue across nearly 1000 individuals, primarily from molecular assays which include Whole-Genome Sequencing (WGS), Whole Exome Sequencing (WES), and RNA-Seq. The platform allows controlled access to de-identified individual-level genotype, expression, and clinical data, and users are able to browse and download computed expression QTL results. The associated tissue repository is also a source for numerous other types of analysis.

The project enables research on the relationship among genetic variation, gene expression, and other molecular phenotypes among a diverse set of human body tissues, many of which are not easily accessible (GTEx consortium, 2013, 2015). Correlations between genotype and tissue-specific gene expression levels will aid in the identification of regions of the genome that affect whether and how much a gene is expressed. GTEx will also aid researchers in understanding the inheritance of disease susceptibility.

2.9.3 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/) is one of the most significant and successful cancer genomics programs (Wang *et al*., 2016). The project began in 2006 as a collaborative effort led by the National Cancer Institute (NCI) and the National

Human Genome Research Institute (NHGRI), both of which are components of the National Institutes of Health, U.S. Department of Health and Human Services.

The project aims to use large-scale genome sequencing applications to accelerate the understanding of the molecular characteristics of cancer. For this purpose, this initiative generates rich molecular and genetic profiles from primary tumor samples of various cancers and their subtypes (Cancer Genome Atlas Research Network *et al*., 2013). TCGA thus houses one of the largest collections of multi-omics datasets of more than 20,000 individual tumor samples, representing 33 types of cancers.

Data from TCGA projects are organized into two tiers: open and controlled access. Controlled access requires an application and approval for access, while open access TCGA data is available through the GDC Data Portal (GDC Data Portal; https://portal.gdc.cancer.gov/). The data type is broadly categorized into biospecimen and clinical data, molecular analysis (genomic characterization) data, and analysis metadata. The platform also allows for web-based analysis and visualization tools (Gao *et al*., 2019). Every data file can be classified as either metadata (alternatively, level 0) or one of three data levels. Level 1 is equivalent to raw data, where examples include an Affymetrix CEL file. While level 2 and level 3 refer to processed and segmented or interpreted files, respectively. Examples of the files can be variant calling format (VCF) or mutation annotation format (MAF) files (Wang *et al*., 2016).

Other TCGA data access methods include using software packages such as the R packages, TCGAbiolinks (Colaprico *et al*., 2016), TCGA2STAT (Wan *et al*., 2016), TCGAIntegrator, and xenaPython python package.

2.9.4. Pathway databases

The interpretation of molecular signatures that are generally yielded by genome-scale investigations is often supported by pathway-centric techniques through which mechanistic insights can be gained by pointing at a collection of biological processes. WikiPathways (WikiPathways; https://wikipathways.org), and Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.kegg.jp/), among other pathway databases, present a curated resource, in a machine-readable form (Ogata *et al*., 1999; Kelder *et al*., 2009; Kanehisa *et al*., 2017; Slenter *et al*., 2018).

WikiPathways is a biological pathway database, founded in 2007 (Pico *et al*., 2008) and currently consists of 242 pathways, and 9014 genes and proteins in the human pathway collection (Martens *et al*., 2021). KEGG was originally developed in 1995 as a comprehensive database resource for the biological interpretation of completely sequenced genomes. Currently, the database consists of fifteen manually curated databases (Kanehisa *et al*., 2017). These databases facilitate the biological interpretations of large-scale molecular datasets. Pre-assembled and ready-to-use menus of pathways and networks from several open sources can be obtained through publicly available web-based applications. Also, software for pathway analysis is available in the form of desktop programs, or packages written in languages such as R and Python.

2.10 Access publicly available RNA-sequenced datasets

RNA-Seq datasets that are publicly available in repositories can be downloaded as either raw sequencing data (.fastq sequencing files) and/or pre-processed files. Generally, the pre-

processed files are stored as tabular formatted files containing matrices with sequenced read counts after trimming and alignment to a reference genome (Sanchis *et al*., 2021). The gene-by-sample matrix comprises columns that are all replicates of the same experiment, and the rows contain the gene names, most frequently corresponding Ensembl identifiers (also known as ENSG IDs). The Ensembl gene IDs are stable identifiers that serve as a method for databases to label features, such as genes, transcripts, exons, or proteins (Aken *et al*., 2016).

The analysis of these large datasets can be incredibly powerful and can reveal many novel findings, however, requires substantial analysis to be interpreted. Thus the demand for bioinformatics expertise is rapidly expanding as a result of the increased popularity of RNA-Seq.

2.11 Bioinformatics

The vast volume of biological data stored in the aforementioned repositories, demands analysis and interpretation, tasks that are being managed by the evolving science of bioinformatics (Bayat, 2002). Within the fields of genetics and genomics, bioinformatics is a scientific subdiscipline that uses computer technology to collect, store, analyse, and distribute biological data and information, including sequences of amino acid and DNA or annotations related to those sequences (Paszkiewicz & Giezen, 2011). Hence, it combines several fields of study, including computer sciences, molecular biology, biotechnology, and statistics. Bioinformatics aims to organize large volumes of molecular data, develop tools that facilitate the analysis of such data, and uncover vital biological information hidden in a large amount of unprocessed data to identify significant trends or patterns (Jiang *et al*., 2022).

Bioinformatics is currently applied in numerous fields, including microbial genome applications, personalized medicine, evolutionary studies, and biotechnology, among others. Cancer bioinformatics is focused on bioinformatics methodologies linked to disease specificity, proliferation, communication and signaling in cancer. While, clinical bioinformatics is an emerging science that combines mathematics, medical informatics, and clinical informatics (Beg & Parveen, 2021). Clinical bioinformatics seeks to comprehend the potential application of biological and medical informatics in the development of personalized healthcare, medication, and therapies.

2.12 Bioinformatics tools and methods

The interdisciplinary field of bioinformatics provides a wide range of packages, tools, and algorithms based on mathematical models developed in R, Python, and other programming languages to analyse and draw scientific findings from the vast volumes of biological data.

2.12.1 Weighted gene co-expression network analysis

WGCNA is an algorithm widely used in cancer research. This method addresses the drawback of most studies that focus on differential genes when screening for differences and ignores the correlations between genes. Therefore, this novel biological method is employed to identify highly correlated gene clusters referred to as modules and key genes based on gene expression data (Langfelder & Horvath, 2008; Langfelder & Horvath, 2012). WGCNA simplifies the interpretation of thousands of genes and builds a co-expression network based on similarities in expression profiles among samples (Niemira *et al*., 2019). Hence, the genes that are clustered into a module have similar expression patterns. Therefore, these genes have the potential to be

involved in the same biological processes or signaling pathways (Liu *et al*., 2017; Kakati *et al*., 2019). Additionally, these gene modules can also be associated with clinical features.

2.12.2 Differential Gene Expression

The most frequent use of transcriptome profiling is to compare one experimental group to another group (or more) to identify which genes change significantly between the conditions (Figure 2.4). The method applied, is known as DGE analysis. The aim of DGE analysis is to perform a statistical analysis that evaluates for differences or changes in the expression level of gene transcripts between experimental groups (Conesa *et al*., 2016). The genes that exhibit differences in expression level between conditions or in other ways are linked to specific predictors or responses are referred to as differentially expressed genes (DEGs), and are critical to advance the understanding of phenotypic variation.

The number of methodologies and tools available for analysing DEGs has rapidly increased (Costa-Silva *et al*., 2017). Among the R language packages developed are limma (Ritchie *et al*., 2015), DESeq2 (Love *et al*., 2014), Cuffdiff (Trapnell *et al*., 2012), NOISeq (Tarazona *et al*., 2011), and edgeR (Robinson *et al*., 2010). Similar results are produced by these techniques, which mostly focus on the interpretation of the log2 fold change value, *p*-value, and *p*-adjusted value. These techniques may be applied to identify gene expression signatures in a single cancer type or to search for shared expression patterns across several cancer types (Kais & Hamdi, 2022). A DGE analysis thus results in a list of genes having significant differences in the gene expression levels between the comparative experimental groups.

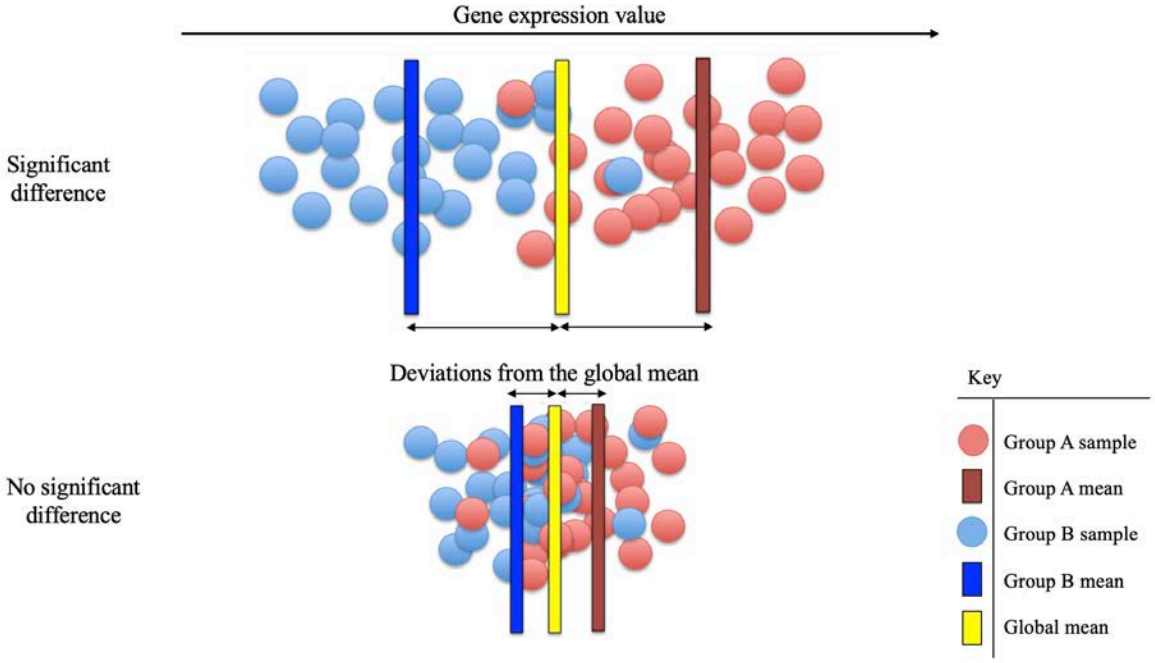

Figure 2.4: Two experimental groups; group A in red and group B in blue, illustrate significant and non-significant differences in gene expression levels. One cluster with samples from both groups shows no significant difference in gene expression (below). While two segregated clusters composed of samples from each group, respectively (top). Hence, groups A and B exhibit different gene expression levels. (Adapted from: https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html).

DGE analysis is widely used to find biomarkers for various cancer types. Numerous studies have employed meta-analysis techniques to identify DEGs between cancer patients and controls using gene expression profiles. Several methods could further be applied to the DGE analysis outputs for validation and prediction studies, as well as machine learning applications.

2.12.3 Machine Learning

Machine learning algorithms are mathematical model mapping techniques that are used to recognize or find underlying patterns and relationships between them from complex data. It comprises a collection of computational algorithms that can classify, adapt, predict, and learn

from existing data (training set) (DeGregory *et al*., 2018). Therefore, many ML tasks aim to optimize the performance of models built on independent test datasets (Zou *et al*., 2019). The three types of ML are (i) supervised learning, which implements labelled data, to develop predictive capabilities, (ii) unsupervised learning, which is a discovering technique, that involves unlabelled data to find hidden information, while (iii) semi-supervised combines both unsupervised and supervised learning (Sarker, 2021) (Figure 2.5).

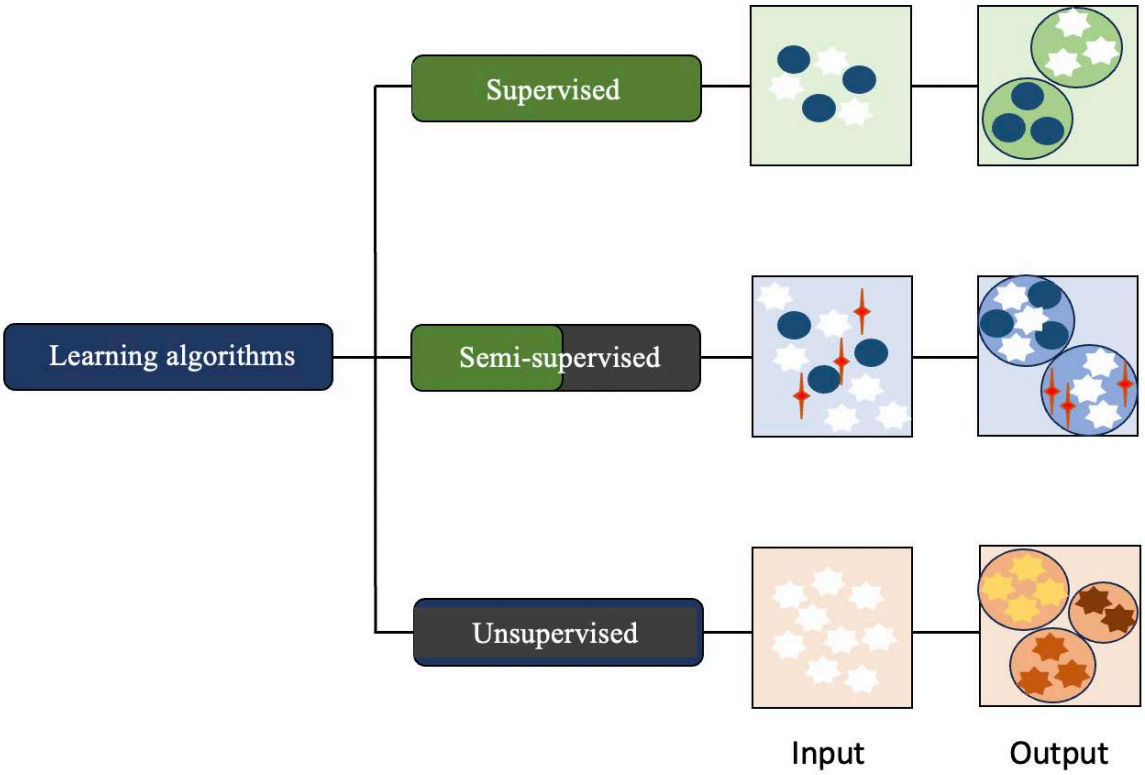

Figure 2.5: ML approaches. The main approaches of machine learning include: (i) Supervised, which relies on labelled input data. (ii) Unsupervised, processes unlabelled data, and (iii) semi-supervised uses both labelled and unlabelled data simultaneously to improve learning accuracy (Adapted from: Rafique *et al*., 2021).

In recent years, some advances have been made through the collaborations between ML and multi-omics data analysis of cancer with the primary intent to provide a broad view of the complexities of the patterns involved in the cancer process (de Anda-Jáuregui & Hernández-

Lemus, 2020). Generally, the application of ML in cancer is used to find and validate potential pathology-based biomarkers that may be useful for diagnosis, improved prognosis, and disease monitoring (Kourou *et al.*, 2014; Yamada *et al.*, 2019; Matek *et al.*, 2019; Ahsan *et al.*, 2022). The prediction in healthcare is vital considering 41u ekcations of delayed diagnosis and treatment.

2.12.4 Survival analysis

Survival analysis, also referred to as time-to-event analysis, is a branch of statistics that investigates the amount of time it takes until a specific event of interest occurs (Schober & Vetter, 2018). Generally, this time is also referred to as "survival time". In numerous cancer studies, the time to an event of interest is the primary outcome being evaluated. In medical studies, an example of an event of interest is the time from diagnosis to death. However, it can also refer to the time 'survived' from complete remission to relapse or progression (Clark *et al.*, 2003). A specific challenge arises if only some individuals have experienced the event of interest. The survival time will thus be unknown for a subset of the study group; this phenomenon is known as censoring (Clark *et al.*, 2003). Censoring is presumed to be non-informative since patients who are censored are considered to have the same probability of surviving as those who continue to be monitored (Clark *et al.*, 2003).

Generally, two related probabilities, survival and hazard are used to describe and model survival data. First, the survival probability (also referred to as the survival function) is the likelihood that an individual will live from the time of origin to a given time in the future. These statistics provide a clear description of a study cohort's survival experience and are often estimated using the K-M curves. Second, the hazard probability (or hazard function) provides

the immediate probability of experiencing an event, given survival up to that time (Clark *et al*., 2003). In short, the survivor function, focuses on not having an event, whereas the hazard function focuses on the event 42u eking. In summary, the hazard relates to the incident (current) event rate, while survival reflects the cumulative non-occurrence.

2.13 Summary

Remarkable efforts have been made to characterize the molecular changes underlying the development and progression of a broad range of complex human diseases, including cancer, due to the recent advancements in omics technology. As a result, multi-omics analyses have been proposed as the key to advancing precision medicine. Several important mechanisms in cancer development, treatment resistance, and recurrence risk have been revealed in the field of precision oncology through genomics approaches. These findings have been applied in clinical oncology to help guide treatment decisions. However, the lack of widespread use of truly integrated multi-omics analysis has limited future advancements in precision medicine. Additional efforts are required to develop an assessment model to accurately generate, evaluate, and annotate multi-omics data to facilitate precision medicine-based decision-making.

Cancer is a major malignant and heterogeneous lethal genetic disease that present significant challenges in both research and clinical treatment. RNA-Seq has served as an essential tool used in numerous aspects of cancer research and therapy, such as the identification of biomarkers, characterization of cancer heterogeneity and evolution, and drug resistance, among others. Therefore, in this study, a computational method was developed that tracks cancer progression through the multi-stages of cancer progression based on RNA-Seq gene expression

profiles. The method normalizes advanced-stage cancer samples with early-stage samples to consider the heterogeneity differences. Therefore subjecting heterogeneous cancer types to the method will allow for the detection of differences in the transcriptional profiles from early to advanced-stage of cancer development.

New cancer clusters (subtypes) that progressed differently in gene expression patterns may be discovered by using hierarchical clustering to the normalized gene expression. As a result, this method's application can recognize molecular heterogeneity and establish a genotype-phenotype relationship with the molecularly identified subtypes. The study thus advances knowledge of the transcriptional landscape of multiple cancer patients with an emphasis on cancer progression. Additionally, the identification of new cancer subtypes has the potential to improve prognosis, identify druggable aberrations in various cancer types, and enable more effective therapeutic strategies. Consequently, the research output will contribute to an advanced understanding of cancer heterogeneity to inform strategies for improving health for cancer patients.

Chapter 3

Transforming RNA-Seq gene expression to track cancer progression in the multi-stage early to advanced-stage cancer development

This is an original manuscript of an article published in PloS ONE on April 2023, available at: https://doi.org/10.1371/journal.pone.0284458.

3.1 Abstract

Background: Cancer progression can be tracked by gene expression changes that occur throughout early-stage to advanced-stage cancer development. The accumulated genetic changes can be detected when gene expression levels in advanced-stage are less variable but show high variability in early-stage. Normalizing advanced-stage expression samples with early-stage and clustering of the normalized expression samples can reveal cancers with similar

or different progression and provide insight into clinical and phenotypic patterns of patient samples within the same cancer.

Objective: This study aims to investigate cancer progression through RNA-Seq expression profiles across the multi-stage process of cancer development.

Methods: RNA-sequenced gene expression of Diffuse Large B-cell Lymphoma, Lung cancer, Liver cancer, Cervical cancer, and Testicular cancer were downloaded from the UCSC Xena database. Advanced-stage samples were normalized with early-stage samples to consider heterogeneity differences in the multi-stage cancer progression. WGCNA was used to build a gene network and categorized normalized genes into different modules. A gene set enrichment analysis selected key gene modules related to cancer. The diagnostic capacity of the modules was evaluated after hierarchical clustering.

Results: Unnormalized RNA-Seq gene expression failed to segregate advanced-stage samples based on selected cancer cohorts. Normalization with early-stage revealed the true heterogeneous gene expression that accumulates across the multi-stage cancer progression, this resulted in well segregated cancer samples. Cancer-specific pathways were enriched in the normalized WGCNA modules. The normalization method was further able to stratify patient samples based on phenotypic and clinical information. Additionally, the method allowed for patient survival analysis, with the Cox regression model selecting gene *MAP4K1* in cervical cancer and K-M confirming that upregulation is favourable.

Conclusion: The application of the normalization method further enhanced the accuracy of clustering of cancer samples based on how they progressed. Additionally, genes responsible for cancer progression were discovered.

3.2 Introduction

Cancer is an ever-changing disease that generally becomes more heterogeneous as the disease progresses (Dagogo-Jack & Shaw, 2018). Different cancers progress and evolve in different ways. Some cancers are fast-growing and can cause mortality soon after diagnosis, while other cancers can be successfully treated (Natrajan *et al*., 2016). One way of tracking cancer progression is to assess gene expression differences across the multi-stage process of cancer development. To our knowledge, limited research has focused on the progression of cancer in relation to gene expression. The numerous genetic changes that accrue over the course of early-stage to advanced-stage cancer development can be traced by RNA-Seq.

RNA-Seq is a high-throughput sequencing technology with computational methods to determine the quantity of RNA present in a biological sample. The method examines the continuously changing cellular transcriptome, allowing for an efficient and comprehensive description of gene expression profiles between different conditions over time (Wang *et al*., 2009). RNA-Seq data is often in the format of a gene-by-sample count matrix, with genes in rows, and samples along the columns. The elements in the matrix report for each sample, the number of reads that could be uniquely aligned to a particular gene. The raw read counts have to be adjusted or "transformed" to aid our understanding of cancer progression.

To demonstrate our approach to investigating RNA-Seq cancer progression over the course of early-stage to advanced-stage cancer, we illustrate a bar graph of a single raw count gene expression profile in two cancer types (Figure 3.1). The dark blue and light blue bars represent advanced-stage and early-stage cancer gene expression, respectively, for gene x. In advanced-stage, gene x shows an identical expression profile in cancer types 1 and 2. Based on the same

raw expression value, both cancer types will group together. However, when considering the early-stage gene expression profiles in both cancer types, it's worth noting that the difference in expression between advanced-stage and early-stage cancer gene expression in cancer type 1 is greater than the difference in cancer type 2.

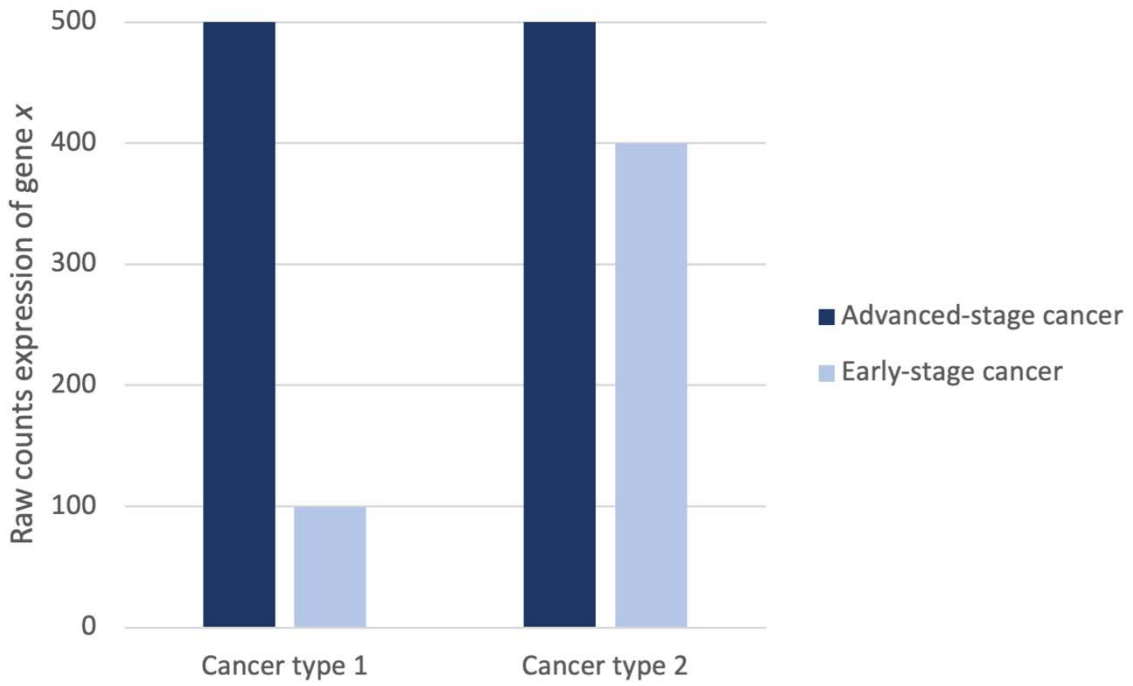

Figure 3.1. Raw RNA-Seq data of advanced-stage and early-stage gene expression of gene x in two cancer types. Cancer type 1 and cancer type 2 show a gene expression fold increase of 4 and 1, respectively, from early to advanced-stage cancer.

The present study aims to normalize advanced-stage with early-stage RNA-Seq data to investigate cancer progression in relation to gene expression. The normalization method corrects for genes that display less expression variability in advanced-stage cancer samples but display a high variability in early-stage cancer samples. As a result, more meaningful information is available in which the two distinct cancer types can be differentiated based on the differences in gene expression profiles, or cancer progression, from early-stage to advanced-stage cancer. The development of such high-throughput genome analysis techniques

for research on cancer has a significant impact on clinical treatment, as the discovery of cancers that differentiate in gene expression profiles (subtypes) is useful for guiding clinical treatment of multiple cancer (Berger & Mardis, 2018).

The normalization method evaluated was performed by Frost and colleagues (Frost *et al*., 2020). This method involves calculating the quotient of cancerous samples (dividend) and normal/non-cancerous samples (divisor), thereby producing normalized differential RNA expression profiles within a specific condition. However, many RNA-Seq research projects do not generate normal sequenced samples. Accordingly, we propose that early-stage cancer samples be used. We further hypothesize that using early-stage cancer samples will provide a more accurate genetic landscape of the multi-stage cancer progression.

3.3 Materials and methods

3.3.1 Data acquisition and processing

Cancer progression was investigated in early-stage and advanced-stage cancer. The datasets examined were selected based on cancers known to have an increased survival risk among patients due to associated autoimmune diseases, prevalent in South Africa and in the African continent in general. This includes five cancers; Diffuse Large B-cell Lymphoma (DLBCL) (Mörth *et al*., 2019; de Carvalho *et al*., 2021), Lung Cancer (Shiels *et al*., 2009), Cervical cancer (Dugué *et al*., 2013), Liver cancer (Clifford *et al*., 2008; Lleo *et al*., 2019), and Testicular cancer (Goedert *et al*., 2007).

RNA-sequenced gene expression profiles for both early- and advanced-stage cancer were downloaded from the UCSC Xena database using cancer-specific data from The Cancer Genome Atlas cohort, from the Genomic Data Commons (GDC-TCGA) (Goldman *et al*., 2020) (Table 3.1). Each patient's expression profile was organized in a gene-by-sample genomic matrix. Additional metadata includes the associated phenotypic and survival profiles of each patient. This data is publicly accessible from the UCSC Xena data browser (https://xenabrowser.net) from individual cancer cohorts. (Appendix A, Table A1, A2).

Table 3.1. Cancer datasets. The cancer cohorts were limited according to clinical or tumor stage and the primary site involved in each cancer. Patient samples were categorized in early-stage and advanced-stage, as well as the primary sites.

		Number of samples	
Cancer cohort	Primary site	Early-stage	Advanced-stage
Diffuse Large B-cell Lymphoma	Lymph Node	4	8
Lung Adenocarcinoma	Bronchus and Lung	28	28
Cervical Cancer	Cervix uteri	8	22
Liver Cancer	Liver and intrahepatic bile ducts	20	6
Testicular Cancer	Testis	15	15

The cancer datasets were made up of 60,483 unique Ensembl identifiers, which included transcript-non-specific expression data for all coding genes plus long non-coding RNA (lncRNA), pseudogenes, and multiple forms of non-coding transcripts (Aken *et al*., 2016). The datasets quantified gene expression as $\log_2(x+1)$ with x referring to the count of reads mapped to a specific genetic region in the human reference genome (GRCh38.p2, gencode release 22).

Genes having ENSG identifiers annotated with a protein-coding biotype were extracted using Ensembl BioMart (GRCh38.p13, Ensembl 104, May 2021) (Smedley *et al*., 2015). This eliminated 40,927 (67,7%) non-coding entries leaving 19,556 protein-coding entries. The gene expression of the 19,556 protein-coding genes as $\log_2(x+1)$ was converted to raw counts for further analysis, as it was found that raw RNA-Seq data may perform better for capturing more original transcriptome patterns in different disease conditions (Han & Men, 2018).

3.3.2 Data normalization

The normalization method involves calculating the quotient of advanced-stage gene expression and early-stage gene expression (GitHub code: https://github.com/3270006/trackingcancer-progression). We followed the same calculations established by Frost et al. (2020).

3.3.2.1 Gene and tissue correction

The gene-by-sample matrices from each cancer cohort in Table 3.1 were used to assemble early-stage I and advanced-stage (A) gene expression matrices. This included:

A, s x q matrix for advanced-stage gene expression and,

E, s x r matrix for early-stage gene expression.

Where q and r represent the number of advanced-stage and early-stage cancer samples, respectively, and s the number of protein-coding genes. Two binary primary site classification matrices were created for each gene expression matrix. This included:

P^A, t x q matrix for advanced-stage cancer primary sites and,

P^E, t x r matrix for early-stage cancer primary sites.

Where q and r represent the number of advanced-stage and early-stage cancer samples, respectively, and t the number of primary sites. The advanced-stage cancer expression vector of gene *I* in matrix A was multiplied by the binary classification vector for primary site *I* in matrix P^A as shown in Eq 1, resulting in a vector of tissue-specific advanced-stage cancer gene expression X_i.

$$X_i = P^A{}_I \odot A_i \quad (1)$$

The early-stage expression vector of gene *I* in matrix E was multiplied by the binary classification vector for primary site *I* in matrix P^E as shown in Eq 2, resulting in a vector of tissue-specific early-stage gene expression Y_i.

$$Y_i = P^E{}_I \odot E_i \quad (2)$$

X_i and Y_i, were computed based on the series of vectors of all primary sites and all protein-coding genes to build three-dimensional matrices for X (advanced-stage cancer) and Y (early-stage cancer). The $X_{i,j,I}$ three-dimensional matrix represents the raw count gene expression value for gene *I* in advanced-stage cancer j of primary site *I*. While, the three-dimensional matrix of $Y_{i,k,I}$ represents the raw count gene expression value for gene *I* in early-stage cancer k of primary site *I*.

The initial phase of calculating for the normalized dataset (subsequently called 'Tissue-corrected'), involved creating a mean normalized expression G^{tissue} for gene *I* at each primary

site *I*, as given in Eq 3. To summarize, the sum of early-stage gene *I* within each primary site *I* was calculated.

$$G_{i,I}^{tissue} = \frac{1}{m_I} \sum_{k=1}^{r} Y_{i,k,I} \tag{3}$$

Where r is the number of early-stage cancer samples in primary site *I*. The calculation to determine for m_I are shown in Eq 4, where the sum of a given primary site in the binary matrix P^E were calculated for all early-stage samples.

$$m_I = \sum_{k=1}^{r} P_{k,I}^{E} \tag{4}$$

Finally, the tissue-corrected gene expression matrix L^{tissue} was calculated as shown in Eq 5.

$$L_{i,j,I}^{tissue} = ln\left(\frac{X_{i,j,I}}{G_{i,I}^{tissue}}\right) \tag{5}$$

3.3.3 Weighted gene co-expression network analysis

Both the advanced-stage cancer gene expression as raw count (uncorrected) and the normalized tissue-corrected datasets were analysed. The 19,556 protein-coding genes were subjected to Weighted Gene Co-expression Network Analysis (v. 1.70–3) (WGCNA) R package (Langfelder & Horvath, 2008; Zhao *et al*., 2010).

3.3.3.1 Data pre-processing.

The uncorrected matrix was filtered of genes that had a count of less than 10 in more than 90% of samples as recommended by the WGCNA authors, resulting in 17,436 protein-coding genes.

The tissue-corrected matrix was filtered by removing all genes that had a row sum of zero, resulting in 19,350 protein-coding genes.

3.3.3.2 Gene co-expression network construction

To construct a weighted network, a correlation matrix between each pair of genes across all samples was calculated. A soft threshold power β was calculated to amplify the correlation between genes. The optimal power value was selected based on a scale-free topology criterion ($R^2 > 0.8$). Based on this, an adjacency matrix was constructed, followed by the generation of a topological overlap matrix (TOM), and computation of the corresponding dissimilarity (1-TOM) values (Zhang & Horvath, 2005; Yip & Horvath, 2007).

To group the protein-coding genes, an average linkage hierarchical clustering based on the *hclust* function in conjunction with the dissimilarity TOM was used, resulting in a gene hierarchical clustering tree (tree graph). A novel dynamicTreeCut algorithm (v. 1.63–1) was employed to identify the clusters, in which branches of the dendrogram were sliced to determine the modules. Modules represent the partitioning of protein-coding genes into distinct groups based on expression values co-correlated and variable across the cancer cohorts. Modules were named using the default WGCNA settings, which assign a colour to each module.

3.3.4 Pathways and transcription factor enrichment analyses

A popular gene set enrichment analysis tool, WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) was used to extract biological information from genes of interest (Zhang *et al*.,

2005). The over representation analysis (ORA) in the WebGestaltR package (v. 0.4.4) was used to characterize the genes of interest that were grouped inside each module found by WGCNA (Wang *et al*., 2013; 2017; Liao *et al*., 2019). The ORA used all protein-coding genes as a reference set, the WikiPathways (Kelder *et al*., 2009; Slenter *et al*., 2018) and KEGG (Kanehisa *et al*., 2017) databases for functional annotations, and the Benjamini-Hochberg (BH) method for multiple testing correction (Benjamini & Hochberg, 1995).

Transcription factor enrichment analysis was performed on the genes of interest that were grouped inside each module found by WGCNA using the ChEA3 database webserver application (Keenan *et al*., 2019). To estimate the TF-target enrichment, the ARCHS4 resource were selected as it uses a co-expression method to compile a list of genes that are controlled by each TF.

3.3.5 Clustering by transcript profiling

The clustering of cancer samples is the most basic and exploratory analysis to find groups of samples sharing similar gene expression patterns, which can lead to the discovery of new cancer subtypes. Therefore, gene expression profiles will be subjected to clustering analysis to investigate the grouping of cancer samples. Accordingly, the computation model was used to predict cancer clusters (subtypes) that progressed differently and/or similarly. The cosine distance between the expression profiles of the genes included in the modules and Ward's method for agglomeration were used to create clusters of similar cancers established by hierarchical clustering (Ward, 1963; Jaskowiak *et al*., 2014). The number of clusters was identified using the *find_k* function, which estimates k using maximal average silhouette widths

(Rousseeuw, 1987). This function forms part of the dendextend (v. 1.15.2) R package. Finally, the dendrograms were split into *k* groups to assign samples to a cluster.

3.3.6 Survival analysis

The genes categorized in each module by WGCNA across the clusters were subjected to a Cox regression model based on the Lasso algorithm of the glmnet R package (v. 4.1–3) (Friedman *et al*., 2010; Simon *et al*., 2011; Tibshirani *et al*., 2012). The model reduces the number of candidate genes and selects the most significant genes for a patient's survival, assigning a regression coefficient value to each gene. The product of the coefficient value and the corresponding gene's expression value resulted in a prognostic risk score for each patient. The patient scores were used to calculate a median risk score. A status value of 1 or 0 was assigned to each patient based on whether the patient's score was above or below the median risk score. Kaplan-Meier estimates for overall survival were generated according to the patient status information. The K–M curves were created using the *ggsurvplot* function from the survminer R package (v. 0.4.9).

3.3.7 Statistics

The statistical analysis was performed using the car (v. 3.0–11), DescTools (v. 0.99.43), and agricolae (v. 1.3–5) R packages. The statistics were conducted to evaluate for different gene expression in each module and primary sites across the clusters.

The differences in the gene expression were first evaluated for normality and equal variance using Shapiro-Wilk test of normality (Shapiro & Wilk, 1965) and Levene's test of homogeneity

(Levene *et al*.,1960), respectively. If the Shapiro-Wilk null hypothesis (H_0) was not rejected ($P \geq 0.05$; H_0: normal distribution) and Levene's test null hypothesis were not rejected ($P \geq 0.05$; H_0: equal variance across groups), an analysis of variance (ANOVA) (Fisher, 1921) was employed. If the ANOVA null hypothesis of equal mean gene expression in each module and primary site was rejected by chance ($P \leq 0.05$), a Tukey's post-hoc test was used for pairwise comparisons (Tukey, 1949).

In the event that Levene's test null hypothesis was rejected ($P \leq 0.05$; H_1: difference in variances between groups) and Shapiro-Wilk test resulted in either normal ($P \geq 0.05$) or not normal distribution ($P \leq 0.05$), then the Kruskal-Wallis test (Kruskal & Wallis, 1952) was used to evaluate for differences in the gene expression in each module and primary site across clusters. If the Kruskal-Wallis was rejected, it can be concluded that equal median gene expression across groups was rejected, a post-hoc analysis was performed using Dunn's test (Dunn, 1964).

3.4 Results and Discussion

Both the uncorrected and tissue-corrected matrices were evaluated to determine if the normalization method represents differences in the true gene expression. The normalization method is considered effective if the normalized gene expression has an increased power in differentiating samples based on cancer type and clinical and phenotypic information.

3.4.1 Uncorrected RNA-Seq

The uncorrected protein-coding genes were inserted into WGCNA. The soft-thresholding power was defined as 20, with a scale-free topological index of above 0.8. This resulted in a gene tree and corresponding module colours. Similar modules were merged using the associated adjacency heatmap. The merged modules and the number of genes in each module was used for further analysis (Appendix A, Figure A1).

A total of 3175 genes were categorized into 32 modules using WGCNA. Of those, only 10 modules were enriched for functional pathway annotations with WikiPathways: brown, cyan, grey60, magenta, purple, dark green, dark grey, light cyan, light steel blue 1, and tan. The first five modules were enriched for tissue-specific processes (ORA, $P \leq 0.047$). The latter five modules were enriched for cancer-relevant processes (ORA, $P \leq 0.045$).

It was found that the tan module had the highest total genes detected in biological pathways. It was also noteworthy that a repetition of the same pathway description appeared in several different modules. The same behaviour was noted with KEGG pathway analysis (Appendix A, Figure A2) This indicates that the uncorrected dataset, which did not undergo normalization, did not efficiently depict gene expression differences.

The hierarchical clustering of cancer samples using the 3175 genes resulted in two cancer clusters (Figure 3.2). The primary site composition of each cluster was evaluated to determine if each primary site corresponded to the cluster assignment. Both clusters were primary sites heterogeneous. Cluster 1 was composed of samples of DLBCL (13.2%), lung (35.8%), liver (5.7%), cervical (22.6%), and testicular cancer (22.6%). While cluster 2 was composed of

DLBCL (3.8%), lung (34.6%), liver (11.5%), cervical (38.5%), and testicular cancer (11.5%). The uncorrected dataset failed to correctly segregate the cancer samples in different clusters (Figure 3.2).

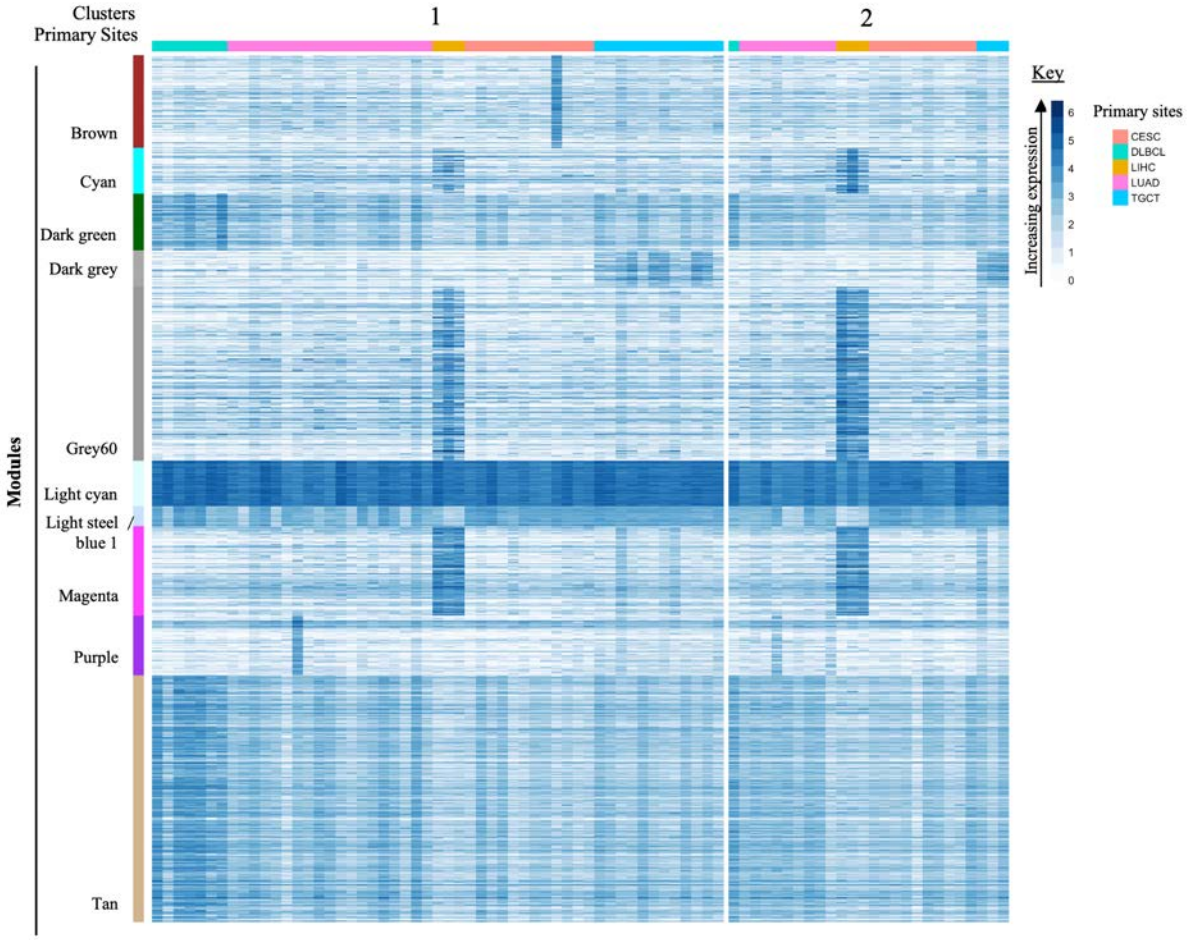

Figure 3.2. Heatmap of uncorrected RNA-Seq data illustrating module expression within cancer clusters. The colour bar on the left shows modules identified by WGCNA and enriched for functional pathway annotations. The rows are further composed of protein-coding genes with raw count values. Clusters of similar cancer cohorts are indicated across the top, and cancer cohorts are displayed by the colour bar along the top with the key on the right. *Primary sites abbreviations: CESC = Cervical squamous cell carcinoma; DLBCL = Diffuse Large B-cell Lymphoma; LIHC = Liver Hepatocellular Carcinoma; LUAD = Lung Adenocarcinoma; TGCT = Testicular Germ Cell Tumors.

The statistical analysis outlined in the methods section was performed to compare each module across the cancer clusters. From the 10 enriched modules, seven modules; cyan, dark green, dark grey, grey60, light cyan, light steel blue 1, and tan were characterized by significantly different expressions (Kruskal-Wallis $P \leq 0.0008$) across cancer clusters. While the magenta, purple (ANOVA, $P \geq 0.08$) and brown modules (Kruskal-Wallis, $P = 0.31$) did not show differential expression across clusters. That is, WGCNA selected genes with less differential power, because of non-normalization, resulting in heterogeneous clusters composed of samples from different primary sites (Figure 3.2).

The same statistical analysis was performed to compare each primary site in Cluster 1 to the equivalent primary site in Cluster 2 for each module. This computation was performed to determine if the segregation of primary sites into Clusters 1 and 2 was based on changes in the gene expression. The statistical test showed no significant difference between sample groups of the same primary sites from the two different clusters. It can be said that the clustering of the uncorrected dataset failed to segregate the primary sites based on different gene expression. Evidently, the unnormalized genes failed to show differentiation.

3.4.2 Tissue-corrected RNA-Seq data

The tissue-corrected protein-coding genes were inserted into WGCNA. A soft threshold selection of the lowest β value that leads to $R^2 > 0.8$ was selected as 21. This resulted in a gene tree and corresponding module colours. Similar modules were merged using the associated adjacency heatmap. The merged modules and the number of genes in each module was used for further analysis (Appendix A, Figure A3).

WGCNA identified 617 genes distributed into seven modules. The module that composed the most and least genes was the brown and pink modules, respectively. Of the seven modules, KEGG analysis enriched five modules (Appendix A, Figure A4), while a total of four modules were found to be enriched for functional pathway annotations with WikiPathways. This included the black, brown, magenta, and turquoise modules (Figure 3.3), of which all four modules were enriched for cancer-related processes (ORA, $P \leq 0.038$). The pathway descriptions identified in the four modules are indicated in the bar chart in Figure 3.3. Each colour bar represents the module colour and shows the number of genes that were enriched for that module. Analysing the degree of enrichment and terms further signifies the difference of each module.

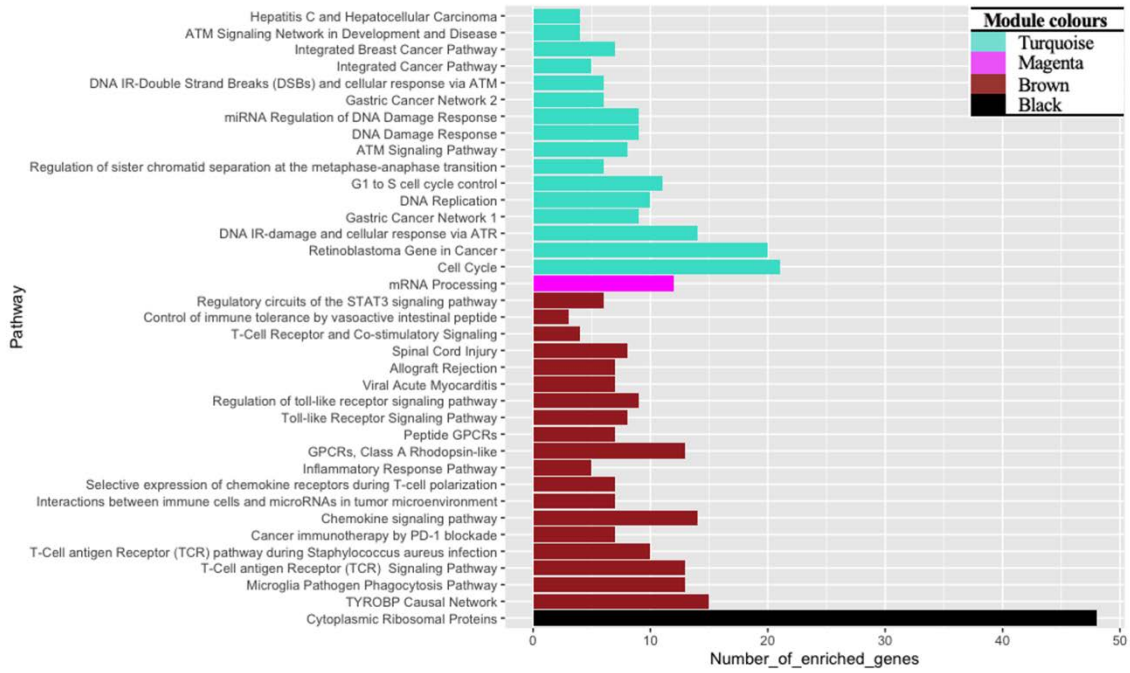

Figure 3.3. WikiPathways enrichment of gene modules detected by WGCNA from the tissue-corrected dataset using the ORA, WebGestalt.

The black module was enriched for cytoplasmic ribosomal proteins (ORA, $P < 0.001$). The brown module was enriched for NK cell, T cell or inflammatory signalling (ORA, $P \leq 0.021$). It was also found that the brown module has the highest total genes detected in biological

pathways. The magenta module enriched for mRNA processing (ORA, $P < 0.0001$). Meanwhile, processes relevant to the cell cycle progression were enriched in the turquoise module (Figure 3.3). The turquoise module was the largest module comprising 139 genes and also identified pathways that were related to other cancers such as breast cancer, gastric cancer, and retinoblastoma. Gastric adenocarcinoma has been reported to be correlated to the investigated cancers including liver carcinoma and lung cancer through specific genes (Salarikia *et al*., 2022). It was noted that some genes were shared between the detected cancer pathways, this included the *AURKA* gene, which was involved in the gastric and breast cancer pathways. An increased gene expression of the *AURKA* gene has been previously identified in the liver and lung cancer (Miralaei *et al*., 2021). Gastric and the retinoblastoma pathways further shared the *MCM4*, *TOP2A*, and *RFC4* genes, that have been reported in the studied cancers, where *MCM4* is overexpressed in liver cancer (Zheng *et al*., 2021), *TOP2A* promotes lung cancer (Kou *et al*., 2020), and *RFC4* has a high expression in liver, lung, and cervical cancer (Li *et al*., 2018a).

Moreover, cancer progression and the retinoblastoma pathway are closely connected (Du & Searle, 2009; Marshall *et al*., 2019). It was found that the retinoblastoma and the breast cancer pathways shared the *CHEK1* gene, a gene that has been reported in the development of human malignant tumors, such as lung and cervical cancers (Wu *et al*., 2019). Therefore, the enriched module genes detected in the studied cancers could suggest that they play a role in cancer development and thus could also be relevant to other cancer types.

The WGCNA module genes were further subjected to TF enrichment analysis, to gain evidence for potential mechanistic connection of transcriptome changes to specific TFs. ChEA3 TF analysis revealed associations between the observed gene expression changes and involved

TFs. The top 5 prioritized TFs for each module are presented in Appendix A, Table A3, with documented information about their biological involvement in the context of cancer (Appendix A, Table A3). The analysis confirms, with supported literature, several TF relationships with the multiple cancers evaluated in this study.

Hierarchical clustering of the 617 genes in WGCNA modules detected eight clusters characterized by distinct expression of the four enriched modules (Kruskal-Wallis Test, $P < 0.0001$) (Figure 3.4). Post hoc analysis by Dunn's Test to assess pairwise differences across clusters in each module showed differential expression for 21 of 28 cluster comparisons for the black module, 25 of 28 comparisons for the brown module, 24 of 28 comparisons for the magenta module, and 27 out of 28 comparisons for the turquoise module. The high proportion of pairwise cluster comparisons with significant differences highlights the distinctive expression patterns in each module across clusters.

The primary site composition of each cluster was evaluated to determine if the cancer primary site corresponded to the cluster assignment. Cluster 1 was primary site homogenous, composed of only DLBCL samples, while Cluster 2 was primary site heterogeneous, composed of DLBCL and liver samples. Clusters 3 and 4 were primary site homogenous, however shows a segregation of lung samples. The same was observed in Clusters 5 and 6 with cervical samples and Clusters 7 and 8 composed of testicular samples (Figure 3.4).

The associated metadata of the cancer samples were investigated to determine if distinct phenotypes could have caused similar cancer cohorts to partition into separate clusters in Figure 3.4. The DLBCL samples present in Cluster 1 show gene profiles that are more upregulated in comparison to the Cluster 2 DLBCL samples. In addition, it was noted that DLBCL samples

in Cluster 1 showed a higher number of extranodal sites involvement (≥ 2), while those in Cluster 2 showed no or low number of extranodal sites involvement (≤ 2). Common sites of extranodal spread are lung, liver, kidney, and bone marrow (Jamil & Mukkamalla, 2022). It has also been reported that DLBCL can be involved in virtually any organ (Beham-Schmid, 2017). Therefore, the DLBCL Cluster 2 found grouped with liver samples is an interesting finding, given the high prevalence of secondary liver involvement by lymphoma including DLBCL and indicates advanced disease (Rajesh *et al*., 2015).

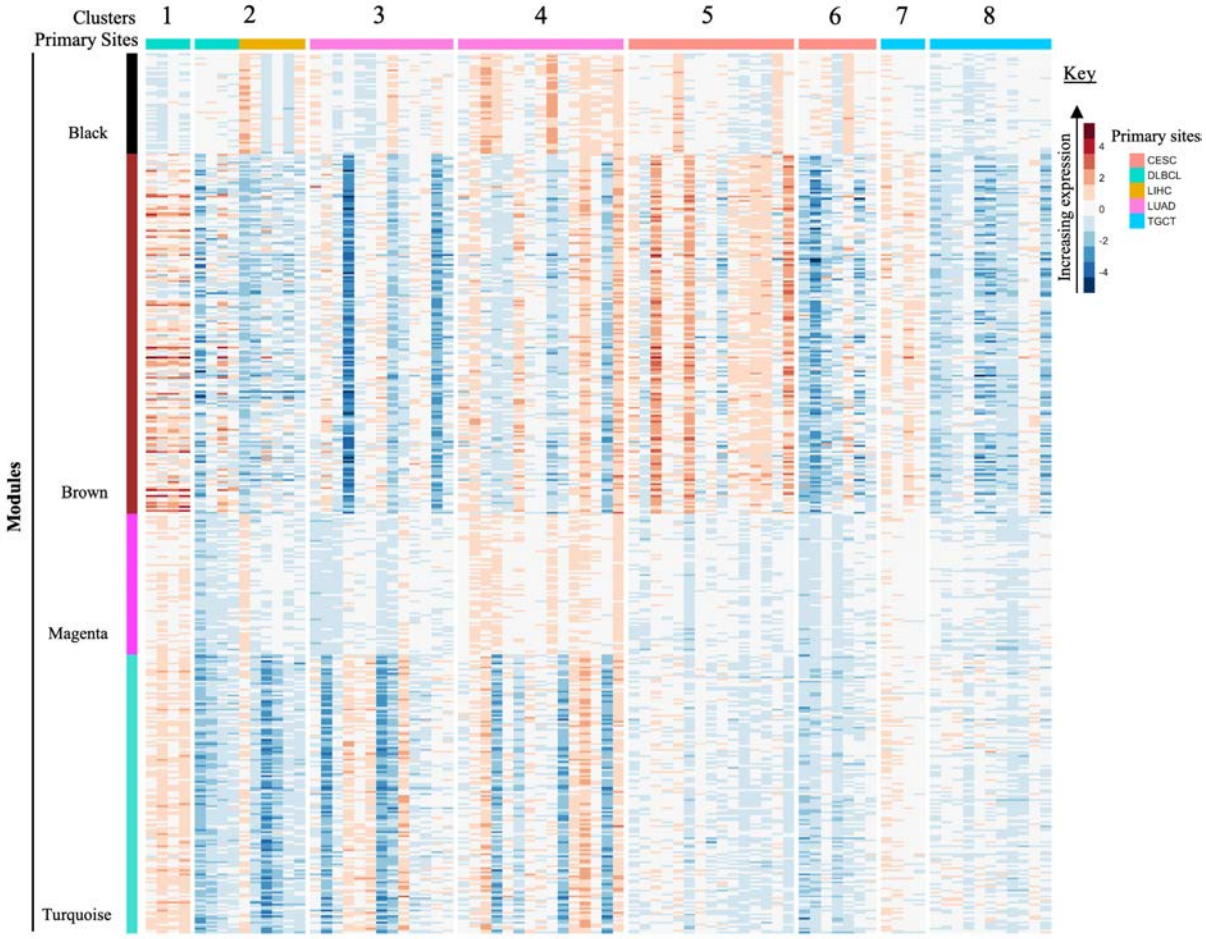

Figure 3.4. Heatmap of tissue-corrected RNA-Seq data illustrating module expression within cancer clusters. The colour bar on the left shows modules identified by WGCNA and enriched for functional pathway annotations. The rows are further composed of protein-coding genes with expression values obtained after data normalization. Clusters of similar cancer cohorts are indicated across the top and the cancer cohort are displayed by the colour bar along

the top with the key on the right. *Primary sites abbreviations: CESC = Cervical squamous cell carcinoma; DLBCL = Diffuse Large B-cell Lymphoma; LIHC = Liver Hepatocellular Carcinoma; LUAD = Lung Adenocarcinoma; TGCT = Testicular Germ Cell Tumors.

However, this information of secondary liver involvement in the metadata associated to DLBCL is unavailable, and requires further investigation to support the claim that DLBCL patients have liver infection, as well as the use of a higher sample number, which was not possible for this study since the public data was not available. The phenotypic data for lung samples in Clusters 4 and 5 did not provide a clear reason for the segregation of the cancer cohort as some clinical information on the samples were incomplete.

It was discovered that the average overall survival of patients with cervical cancer represented in Cluster 5 were greater than the average overall survival of cervical cancer patients in Cluster 6. This led to a survival analysis in which the Cox regression model selected *MAP4K1* (ENSG00000104814) categorized in the brown module as a prognostic gene. The upregulation of the *MAP4K1* gene has been found to be favourable in cervical cancer (Uhlen *et al*., 2017; The human protein atlas; 2022). According to K-M results in a recent study, the high expression of the *MAP4K1* gene was beneficial to cervical cancer patients (Kannan *et al*., 2021). Their research focussed on *PDCD1*, a gene that is most typically related to its expression on tumor-infiltrating lymphocytes. Moreover, they showed that *PDCD1* significantly co-expressed with the following 15 genes, whose high expression is beneficial for cervical cancer patients; *MAP4K1, ACAP1, CST7, CXCR6, GPR171, GZMH, GZMK, P2RY10, RASAL3, SH2D1A, TBC1D10C, ZNF831, GZMM, JAKMIP1*, and *PSTPIP1* (Kannan *et al*., 2021). We compared their finding to the results of our normalization method and discovered the *PDCD1* gene as well as the first 12 of the 15 genes were co-expressed within the brown module. This finding validates the normalization method in this study, as upregulation is observed in the brown

module for Cluster 5, whereas the brown module in Cluster 6 mainly illustrates downregulation (Figure 3.4). The normalized gene expression of *MAP4K1* in cervical patient samples from Clusters 5 and 6 were extracted from the brown module and shown in Figure 3.5.

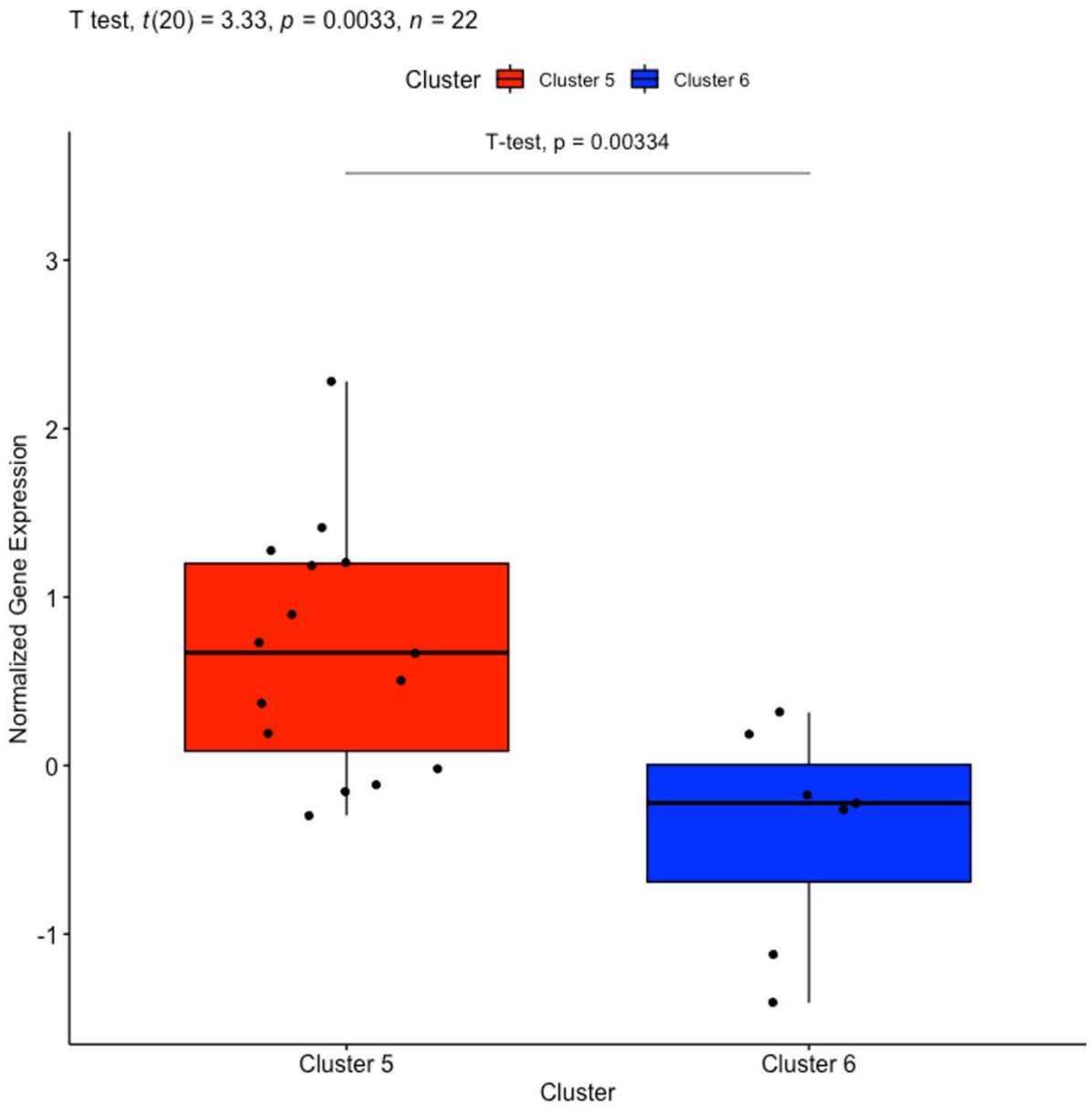

Figure 3.5. Boxplot of gene MAP4K1 from cervical cancer samples categorized in the brown module by WGCNA. The red box plot, constructed with Cluster 5 samples, shows upregulation of gene MAP4K1, while the blue box plot, constructed with Cluster 6 samples, shows downregulation of MAP4K1 gene.

We corroborate the previous findings (Uhlen *et al*., 2017; The human protein atlas; 2022; Kannan *et al*., 2021) in that the upregulation of gene *MAP4K1*, in Cluster 5, is favourable in cervical cancer patients as shown by the K-M curve, in Figure 3.6. Cluster 5 presents a longer life expectancy than the patient samples in Cluster 6.

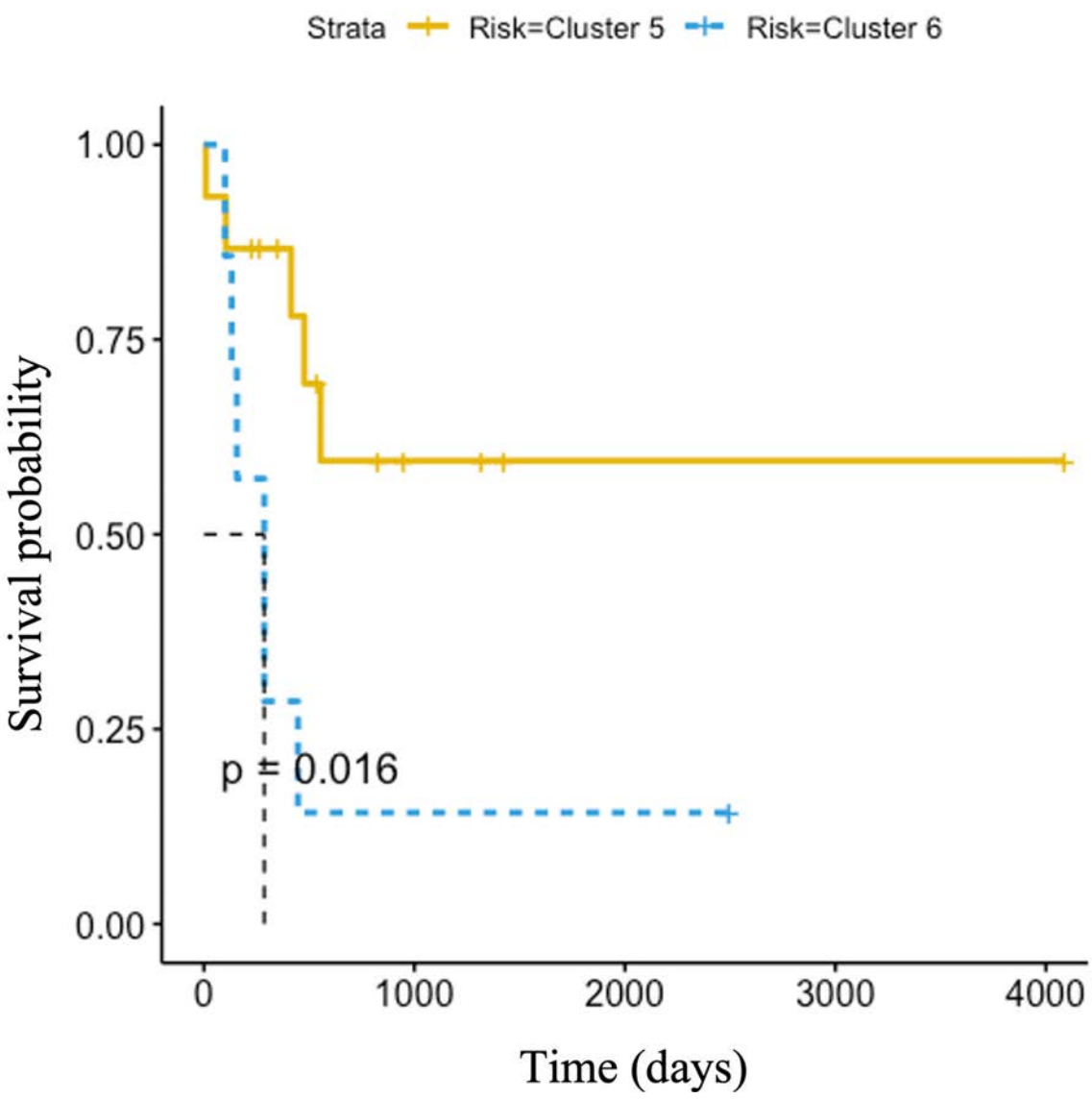

Figure 3.6. Kaplan-Meier of *MAP4K1* gene in cervical cancer patients. Analysis shows the correlation between normalized gene expression level and patient survival in days. Patients were divided as detected in Clusters 5 and 6 after clustering according to transcript profiling.

The brown module was further subjected to TF enrichment analysis using an established computational tool to offer a better understanding of the associations between the observed gene expression changes and TFs in the context of cervical cancer. The TFs that were associated with the *MAP4K1* gene in which the TF was found to effect cervical cancer survival was extracted and documented (Appendix A, Table A4). Several co-expressed genes that also play a role in cervical cancer survival identified in (Kannan *et al*., 2021) were also linked to the TFs and highlighted (Appendix A, Table A4).

Lastly, the phenotypic data of testicular cancer, divided in Clusters 7 and 8, showed that the primary diagnosis of the patients in Cluster 7 was seminomas, while Cluster 8 were made up of patient samples that were primarily diagnosed with type embryonal carcinoma testicular cancer, mixed germ cell tumor or Teratoma malignant.

To further demonstrate the significance of late-stage cancer samples normalized with early-stage cancer samples, an investigation was carried out with a normal tissue expression dataset from the GTEx Portal (GTEx Consortium, 2017). Normalized gene expression profiles using normal tissue samples were clustered and allowed for the segregation between distinct cancer types (Appendix A, Figure A5). However, it failed to provide in-depth clustering based on subtypes within cancer types. As a result, the variations in gene expression, such as in cervical cancer that was associated with survival, could not be stratified by normalizing late-stage cancer samples with normal tissue. The results obtained with our method by normalizing late-stage with early-stage cancer samples demonstrate the ability of the method to cluster samples by cancer progression, rather than simply by cancer type as with the use of normal samples.

3.5 Conclusion

The RNA-Seq read count before normalization showed discrepancies in comparison to normalized gene expression. The goal of our normalization method was achieved, since it shows that advanced-stage cancer gene expression data can be normalized using early-stage cancer gene expression data. WGCNA analysis validated the results of the tissue-corrected matrix as the correct relationships between normalized gene expression were presented. It was further illustrated that the biological information was preserved and allowed more meaningful comparisons of each cancer cohort, including survival analyses.

The benefit of the normalization method used in the present study was twofold; (i) it was able to segregate tumor samples with different and similar progression, (ii) and it could cluster samples from distinct cancer types as well as samples within the same cancer type. A significant result of the latter was in the case of cervical cancer, in which gene *MAP4K1* was segregated according to the disease prognosis. This discovery demonstrated that the normalization method can be used in conjunction with cancer clustering to identify areas of higher cancer risk as well as the cause of the increased risk.

The value of this method thus aids with hypotheses that seek to explore various novel cancer subtypes that segregate by different gene expression profiles and further investigate the biological association, clinical, or prognostic features linked to the cancer subtypes (clusters). Additionally, hypotheses that investigate cancer progression and identify cancer subtypes with different progression. New users can further use this method to find new subtypes in their data and associate it with the clinical data that they have.

Chapter 4

Assessment of the progression of kidney renal clear cell carcinoma using transcriptional profiles revealed new cancer subtypes with variable prognosis

This is an original manuscript of an article published in Frontiers in Genetics on November 2023, available at: https://doi.org/10.3389/fgene.2023.1291043.

4.1 Abstract

Background: Kidney renal clear cell carcinoma is the most prevalent subtype of renal cell carcinoma encompassing a heterogeneous group of malignancies. Accurate subtype identification and an understanding of the variables influencing prognosis are critical for personalized treatment, but currently limited. To facilitate the sub-classification of KIRC patients and improve prognosis, this study implemented a normalization method to track cancer

progression by detecting the accumulation of genetic changes that occur throughout the multi-stage of cancer development.

Objective: To reveal KIRC patients with different progression based on gene expression profiles using a normalization method. The aim is to refine molecular subtyping of KIRC patients associated with survival outcomes.

Methods: RNA-sequenced gene expression of eighty-two KIRC patients were downloaded from UCSC Xena database. Advanced-stage samples were normalized with early-stage to account for differences in the multi-stage cancer progression's heterogeneity. Hierarchical clustering was performed to reveal clusters that progress differently. Two techniques were applied to screen for significant genes within the clusters. First, DEGs were discovered by Limma, thereafter, an optimal gene subset was selected using RFE. The gene subset was subjected to Random Forest (RF) Classifier to evaluate the cluster prediction performance. Genes strongly associated with survival were identified utilizing Cox regression analysis. The model's accuracy was assessed with K-M. Finally, a Gene ontology and Kyoto Encyclopedia of Genes and Genomes enrichment analyses were performed.

Results: Three clusters were revealed and categorized based on patients' overall survival into short, intermediate, and long. A total of 231 DEGs were discovered of which RFE selected 48 genes. RF Classifier revealed a 100% cluster prediction performance of the genes. Five genes were identified with significant diagnostic capacity. The downregulation of genes *SALL4* and *KRT15* and upregulation of genes *OSBPL11*, *SPATA18*, and *TAL2* were associated with favorable prognosis.

Conclusion: The normalization method based on tumour progression from early to late stages of cancer development revealed the heterogeneity of KIRC and identified three potential new subtypes with different prognoses. This could be of great importance for the development of new targeted therapies for each subtype.

4.2 Introduction

Multiple different forms of kidney tumors make up the complex disease known as kidney cancer (Hu *et al*., 2019). Renal cell carcinoma (RCC) is a heterogeneous group of kidney parenchyma tumors that can be further divided into histologically defined subtypes (Znaor *et al*., 2015; Casuscelli *et al*., 2017; Xiong *et al*., 2022). The different subtypes have undergone multiple revisions in the past two decades, due to advancements in the morphological as well as molecular characterization of renal tumors (Kovacs *et al*., 1997; Lopez-Beltran *et al*., 2006; Srigley *et al*., 2013; Moch *et al*., 2016; Udager & Mehra, 2016).

The recent discoveries in renal tumor transcriptome profiling studies have had a substantial influence in the field of genomics as a category for "molecularly defined renal carcinomas" has been introduced by the World Health Organization 2022 classification of urinary and male genital tumors (5th edition) (Trpkov *et al*., 2021a; 2021b; Mohanty *et al*., 2023). These studies have significantly improved our understanding of RCC, however, effective diagnostic and therapeutic approaches have yet to be achieved (Caliskan *et al*., 2020). Additionally, these studies revealed the high molecular heterogeneity of these tumors, necessitating further sub-classification.

In this study, the most prevalent and aggressive subtype Kidney renal clear cell carcinoma was investigated as it accounts for 80%–90% of the total number of RCC patients (Wang Q. *et al*., 2019). Patients with KIRC are associated with a high mortality rate and poor clinical outcomes (Gray & Harris, 2019; Puzanov, 2022). Also, there are limited therapeutic options available; surgery is the primary option since KIRC is resistant to radiotherapy and chemotherapy (Yin *et al*., 2019). The resistance to treatment may be due to the heterogeneity of these tumors.

Therefore, an accurate assessment of the heterogeneity of these tumors is crucial to identify subtypes of patients that can benefit from targeted therapy. This can be achieved by investigating the underlying molecular mechanisms and progression of KIRC, which are currently not fully understood (You *et al*., 2021).

To track cancer progression, we implemented a recently established normalization method, which also has the potential to facilitate the sub-classification of KIRC (Livesey *et al*., 2023). The normalized gene expression reveals how cancer progresses by detecting the accumulated genetic changes that emerge from early-stages of cancer development to advanced-stages. The application of the normalization method and hierarchical clustering will allow for the identification of clusters (subtypes) that progress differently.

This study aims to reveal KIRC patients with different progression (subtypes) and establish a genotype-phenotype link to the identified clusters. In this study, the genotype-phenotype relationship to the distinct clusters was defined by the average OS of the KIRC patient samples. Prognostic gene signatures were identified that differentiate between the different survival clusters and have the potential to function as prognostic biomarkers that can facilitate the prognosis and monitoring of KIRC. Therefore, the study advances knowledge of the transcriptional landscape of KIRC patients with an emphasis on cancer progression.

4.3 Materials and methods

4.3.1 Data acquisition and processing

The RNA-Sequencing (RNA-Seq) gene expression profiles of KIRC were downloaded from the UCSC Xena database using cancer-specific data from The Cancer Genome Atlas cohort, from the Genomic Data Commons (GDC-TCGA) (Goldman *et al*., 2020). A total of eighty-two advanced-stage cancer samples, along with a matched number of randomly selected early-stage samples were extracted. The accompanying metadata included the corresponding patient phenotypic and survival profiles. The gene expression profile of each patient was organized in a gene-by-sample genomic matrix. The cancer datasets consisted of 60,483 unique Ensembl identifiers (ENSG) (Aken *et al*., 2016), quantified as log2(x+1), where x represents the count of reads mapped to a specific genomic location in the human reference genome (GRCh38.p2, gencode release 22). Ensembl BioMart (GRCh38.p13, Ensembl 104 May 2021) (Smedley *et al*., 2015) was utilized to retrieve a total of 19,556 ENSG identifiers that were annotated with a protein-coding biotype. Hence, 40,927 (67,7%) non-coding entries were eliminated. For further analysis, the 19,556 protein-coding gene expressions were converted to counts. The source code for the implementation of reproducibility of the analyses for the study is available in GitHub: https://github.com/LiveseyM/KIRC_Subtyping.git.

4.3.2 Data normalization

The normalization method that tracked cancer progression and corrected for multiple cancers (Livesey *et al*., 2023) was modified to investigate a cancer type. The normalization method

involves calculating the quotient of advanced-stage gene expression and early-stage gene expression.

4.3.2.1 Tracking cancer progression

A normalization method was implemented to capture the heterogeneity between cancerous tumors by detecting their molecular differences in progression from early to late-stages of tumor development using gene expression by RNA-Seq. As a result, the method exposes the accumulated genetic changes that occur throughout the multi-stage of cancer development. To track the development of cancer, the gene expression profiles of both early-stage and late-stage cancer samples were required. Thus, the gene-by-sample matrix of KIRC was used to create two distinct matrices; early-stage I and advanced-stage (A) gene expression as follows:

E, s x r matrix for early-stage gene expression and,

A, s x q matrix for advanced-stage gene expression.

The early-stage and advanced-stage gene expression matrices are represented by E and A, respectively. Where r and q correspond to the number of cancer samples in early-stage and advanced-stage, and s the number of protein-coding genes represented with raw count gene expression value.

The early-stage patient profiles do not match the same patient profiles in the late-stages. Thus, the initial approach to calculating the normalized dataset involves generating a mean normalized expression, or "m_i", for gene *I* in the early-stage dataset. The sum of early-stage gene *I* for all early-stage cancer k samples was calculated, as shown in Eq 1. The average early-

stage expression vector of gene I produced by this equation offers a more accurate representation of the early-stage expression of a particular gene.

$$m_i = \frac{1}{r}\sum_{k=1}^{r} E_{i,k} \quad \text{(eq 1)}$$

$$L_i = ln\left(\frac{A}{m_i}\right) \quad \text{(eq 2)}$$

Finally, the gene expression matrix that represents cancer progression, L was calculated as demonstrated in Eq 2. Matrix L contains normalized counts of the quotients of advanced-stage (dividend) and the mean gene expression of early-stage cancer samples (divisor). Therefore, the normalized gene expression represents the continuously changing cellular transcriptome, allowing for an efficient and comprehensive description of gene expression profiles.

4.3.3 Hierarchical clustering

The clustering of cancer samples is the most fundamental strategy to identify groups of samples that progressed differently in gene expression patterns. This approach may result in the identification of novel cancer clusters (subtypes) within a cancer type. Therefore, the normalized gene expression profiles of the KIRC cancer samples were subjected to hierarchical clustering analysis, to reveal the grouping of cancer samples.

The clusters of cancer samples were created by hierarchical clustering, using the cosine distance between the gene expression profiles and Ward's method for agglomeration (Ward, 1963; Jaskowiak *et al*., 2014). The optimal number of clusters was determined using the *find_k* function as part of the dendextend R package (version 1.17.1), which calculates k using

maximal average silhouette widths (Rousseeuw, 1987). Finally, the dendrograms were split into k groups to assign samples to a cluster.

4.3.4 Feature analysis

4.3.4.1 Differential gene expression

Limma package in R (version 3.54.2) (Ritchie *et al*., 2015) was used to screen for DEGs, by applying an empirical Bayesian approach to evaluate for differences in gene expression profiles between the identified clusters. The *decideTests* (Law *et al*., 2016) function assigned binary values (0: not detected, 1: upregulated, and −1: downregulated) to the genes, to identify and extract genes that differentiate between the altered (up or down) gene expression. Significant DEGs were defined as those with a BH adjusted p-value <0.05 and log2-fold change (LFC) ≥ 0.5 or ≤ -0.5.

4.3.4.2 Marker gene selection using machine learning

Recursive Feature Elimination algorithm was implemented to identify key genes playing a role in the classification of the identified KIRC clusters (subtypes), using the Scikit-learn python package (Pedregosa *et al*., 2011). RFE with a linear kernel support vector machine (SVM) was utilized to find optimal genes that predict the cancer clusters. The k-fold cross-validation procedure, with a value of K set to 10, was repeated 3 times.

The model was built with all identified DEGs and In several iterations eliminates a single gene deemed least important for segregating the identified clusters (Guyon *et al*., 2002). The model

is rebuilt, and the new gene subset are evaluated based on their classification performance. Hence, the genes are ranked according to their relevance. In this study, the final gene subset was selected based on the highest classification accuracy by linear SVM with *C* set to 5. The final gene subset was further subjected to principal component analysis (PCA) using the R packages FactoMineR (version 2.8) (Lê *et al*., 2008) and factoextra (version 1.0.7) (Kassambara & Mundt, 2020).

4.3.5 Predictive and validation of marker genes

The performance of the RFE selected gene subset was validated using RF classifier with a "test-train split ()" class to split the data into train and test sets with a ratio of 75: 25. The performance of the RF classifier was measured using accuracy, precision, and recall score as the performance metrics. All machine learning implementations were run in Anaconda environment based on python programming language and Scikit-learn package (Pedregosa *et al*., 2011).

4.3.6 Survival analysis

The gene subset selected by RFE was subjected to a Cox regression model based on the Lasso algorithm of the glmnet R package (version 4.1-7), to further understand the relative importance of the gene subset (Friedman *et al*., 2010; Simon *et al*., 2011; Tibshirani *et al*., 2012). The model reduces the total number of the gene subset and identifies the genes with the most significant impact on a patient's survival. This step assigned a regression coefficient value to the given gene that is multiplied by the corresponding gene's expression and results in a prognostic risk score for each patient. The patient scores were used to calculate a median risk

score. Each patient was assigned a status value of 0 or 1 based on whether the patient's score was higher or lower than the median risk score. The patient status information was used to generate K-M estimates for OS. The K–M curves were constructed using the *ggsurvplot* function from the survminer R package (version 0.4.9).

4.3.7 One-way ANOVA

A one-way analysis of variance (ANOVA) was performed to compare the mean gene expression of the prognostic genes discovered by Cox regression analysis between the identified clusters. Statistical analysis was conducted with the stats R package (version 4.2.2). Following the application of ANOVA, Tukey's *post hoc* test for pairwise comparisons was applied (Tukey, 1949). The null hypothesis of equal mean between the clusters was rejected if the p-value < 0.05; H_1: the cluster means are significantly different from one another.

4.3.8 Enrichment

The list of DEGs were subjected to functional annotations of GO (Ashburner *et al*., 2000), with an adjusted p-value < 0.05 determined as a cut-off criterion for significant enrichment. Additionally, the 48 RFE gene subset were subjected to KEGG pathways enrichment, with the threshold for significant enrichment established as p-value <0.05. The enrichment analysis was performed utilizing the clusterProfiler R package (version 4.6.2) (Yu *et al*., 2012).

4.4 Results

4.4.1 Cancer clusters detection with normalized expression

The gene expression profiles of eighty-two advanced-stage KIRC samples were normalized with early-stage cancer samples to consider the heterogeneity differences that occur in the multistage cancer progression.

In this study, all 19,556 normalized protein-coding genes were subjected to clustering. The clusters are visually represented in a hierarchical tree called a dendrogram. The clustering of all eighty-two KIRC samples revealed three unique KIRC progression patterns based on gene expression profiles (Figure 4.1).

Three unique cancer clusters (subtypes) as Clusters 1, 2, and 3 were identified and encompass a total of 42, 24, and 16 KIRC patient samples, respectively. These three molecularly identified clusters were further correlated with the patients' average overall survival to reflect its genotype-phenotype relationship. Cluster 1 showed the lowest average OS of 864.43 days, Cluster 2 displayed an average OS of 1076.38, and Cluster 3 had the highest average OS of 1522.31 days. Therefore, these Clusters were categorized as Short (SS), Intermediate (IS), and Long Survival (LS) (Table 4.1).

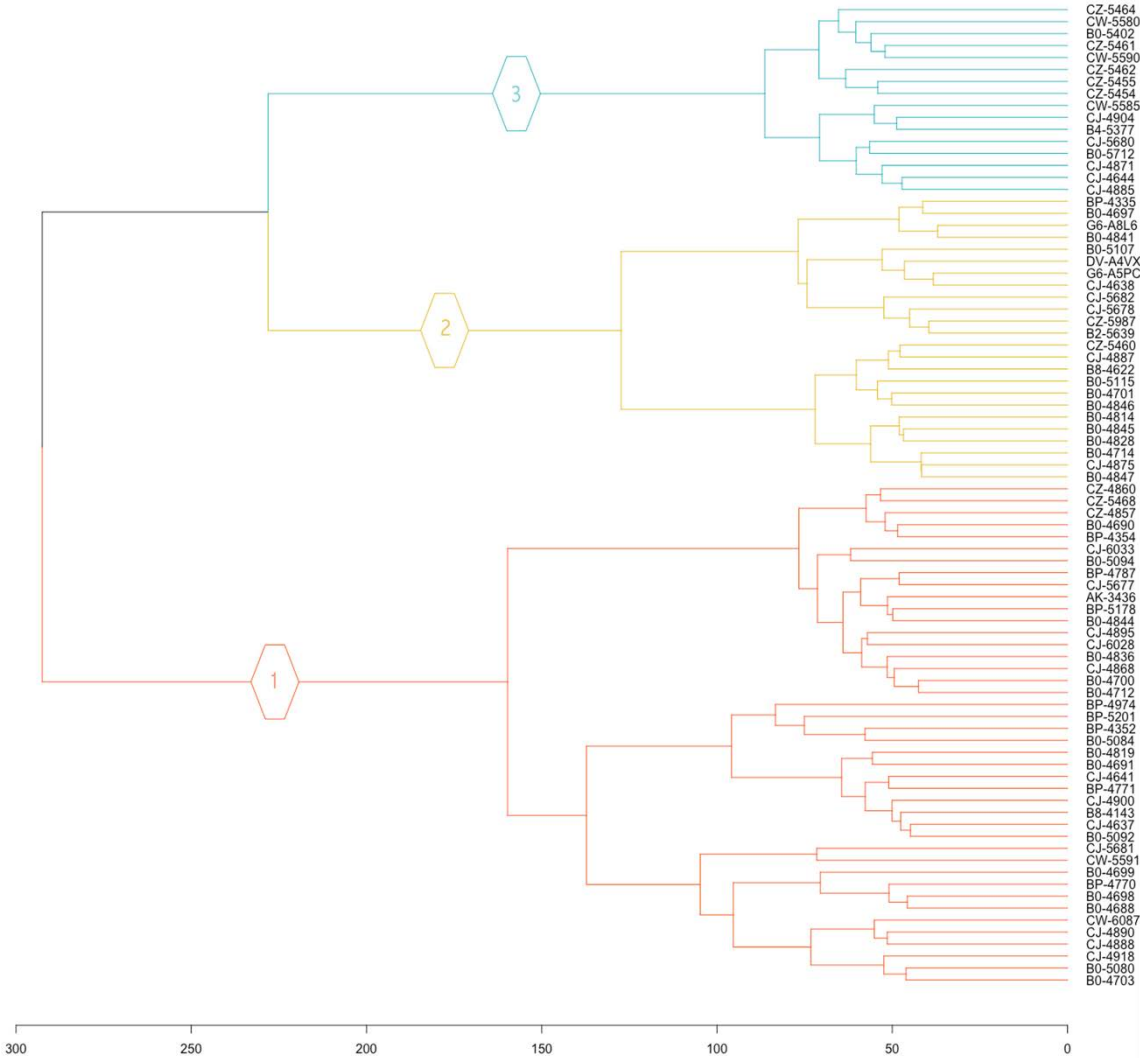

Figure 4.1. Hierarchical clustering dendrogram of KIRC patient. The 19,556 normalized gene expression profiles of the eighty-two KIRC cancer samples were subjected to clustering analysis, to reveal the grouping of cancer samples.

Table 4.1: The number of patient samples stratified by hierarchical clustering. The average overall survival of all patients within a cluster was calculated and further categorized into Short, Intermediate, and Long Survival.

Cluster	Average survival (days)	Survival time	Risk subcategory	Samples
1	864.43	Short	SS	42
2	1076.38	Intermediate	IS	24
3	1522.31	Long	LS	16
		Total		82

4.4.2 Differential gene expression analysis

In the DGE analysis, a total of 19,556 protein-coding genes were evaluated for DEGs to distinguish between SS, IS, and LS. A pairwise comparison approach between the gene expression profiles of IS and SS, LS and SS, and LS and IS were used, and only the genes with an adjusted *p*-value <0.05 and LFC ≥ 0.5 or ≤ −0.5 between all three pairwise comparisons were used for further analysis. Thus, a total of 231 DEGs were discovered.

Considering only the DEGs that were significant between all three pairwise comparisons, a total of 47 genes were identified as upregulated, when IS was compared to SS, whereas 184 genes were found to be downregulated. While 159 genes were upregulated, and 72 genes were downregulated in the comparison of LS and SS. Finally, the comparison of LS and IS, identified 221 and 10 genes as upregulated and downregulated, respectively.

4.4.3 Selection of optimal gene subset

All 231 DEGs identified between SS, IS, and LS KIRC patients were screened by the RFE algorithm. The optimal gene subset is defined by the best combination of genes that has candidate characteristics of classification and prognosis. This also refers to the performance of the RFE and is quantified by the feature importance score. In this study, the optimal gene subset of 48 genes (Appendix B, Table B1) with the highest performance score of 0.963 was selected for further analysis (Figure 4.2A).

4.4.3.1 Validation of optimal RFE gene subset

An RF classifier model was constructed to evaluate the classification power of the 48 RFE gene subset for SS, IS, and LS. A tenfold cross-validation on a forest model in the training phase (75% of the samples) and testing phase (25% of the samples) was computed. The RF classification yielded an accuracy score of 100%, a precision of 100%, and a recall of 100%.

A confusion matrix that defines the performance of the classification algorithm is presented in Figure 4.2B. The importance of each gene for risk subcategory prediction to the RF classifier model is presented in Figure 4.2C.

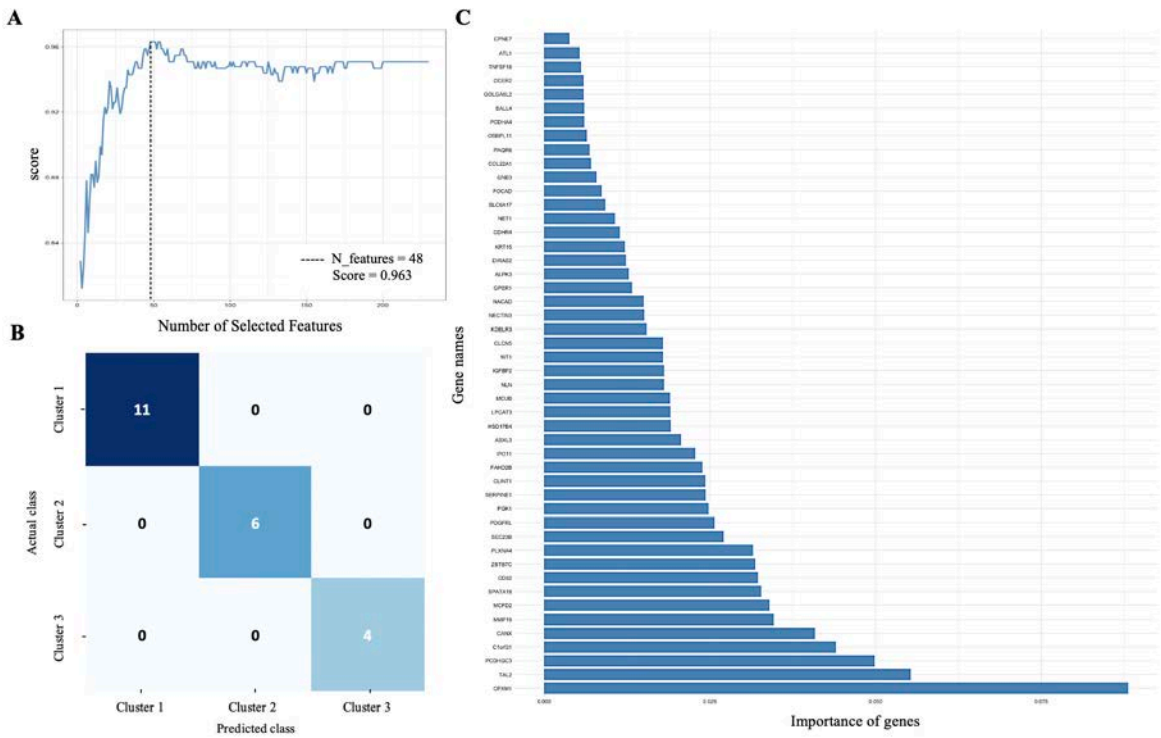

Figure 4.2. Supervised machine learning. (A) Recursive feature elimination selected 48 genes with the highest performance score of 0.963. (B) Confusion matrix that defines the performance of RF classifier. Each row and columns represent the instances in an actual and predicted class, respectively. (C) The importance of each gene for RF classifier prediction.

A PCA model was built to determine the heterogeneity in gene expression between the SS, IS, and LS risk subcategories. The PCA assessed and identified the key sources of variance, allowing samples to be grouped based on similar and different gene expression profiles. Dim 1 represented 29.8% of the overall variance, whereas Dim 2 represented 23.6% (Figure 4.3). A clear segregation between KIRC patient samples can be observed to distinguish between the three risk subcategories.

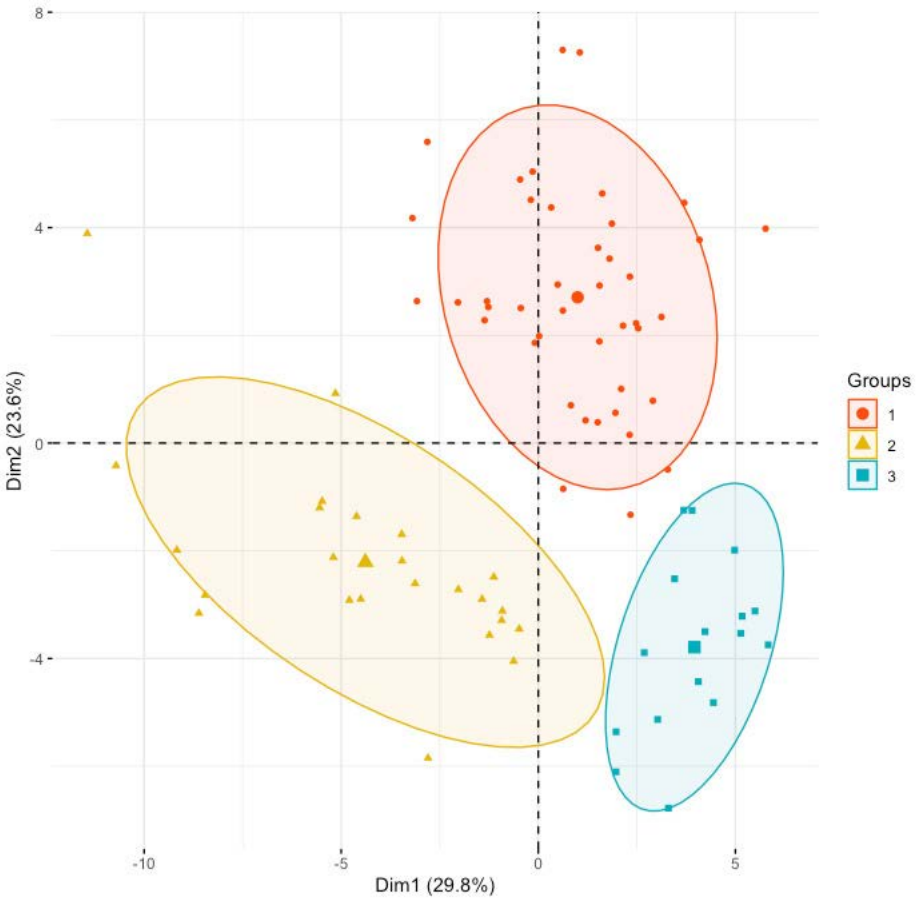

Figure 4.3. Principal component analysis using the normalized gene expression profiles of the 48 RFE gene subset. KIRC samples were stratified according to the initial hierarchical clustering analysis.

To further compare the initial clustering analysis of protein-coding genes to the clustering of the selected 48 RFE gene subset, a hierarchical clustering was performed with the normalized

gene expression of the 48 RFE gene subset of the eighty-two KIRC cancer samples. The correspondence between the two hierarchical clusters is represented by a tanglegram (Figure 4.4). It can be observed that only four samples were assigned to a different cluster (risk subcategory) with the reduced gene subset (Figure 4.4).

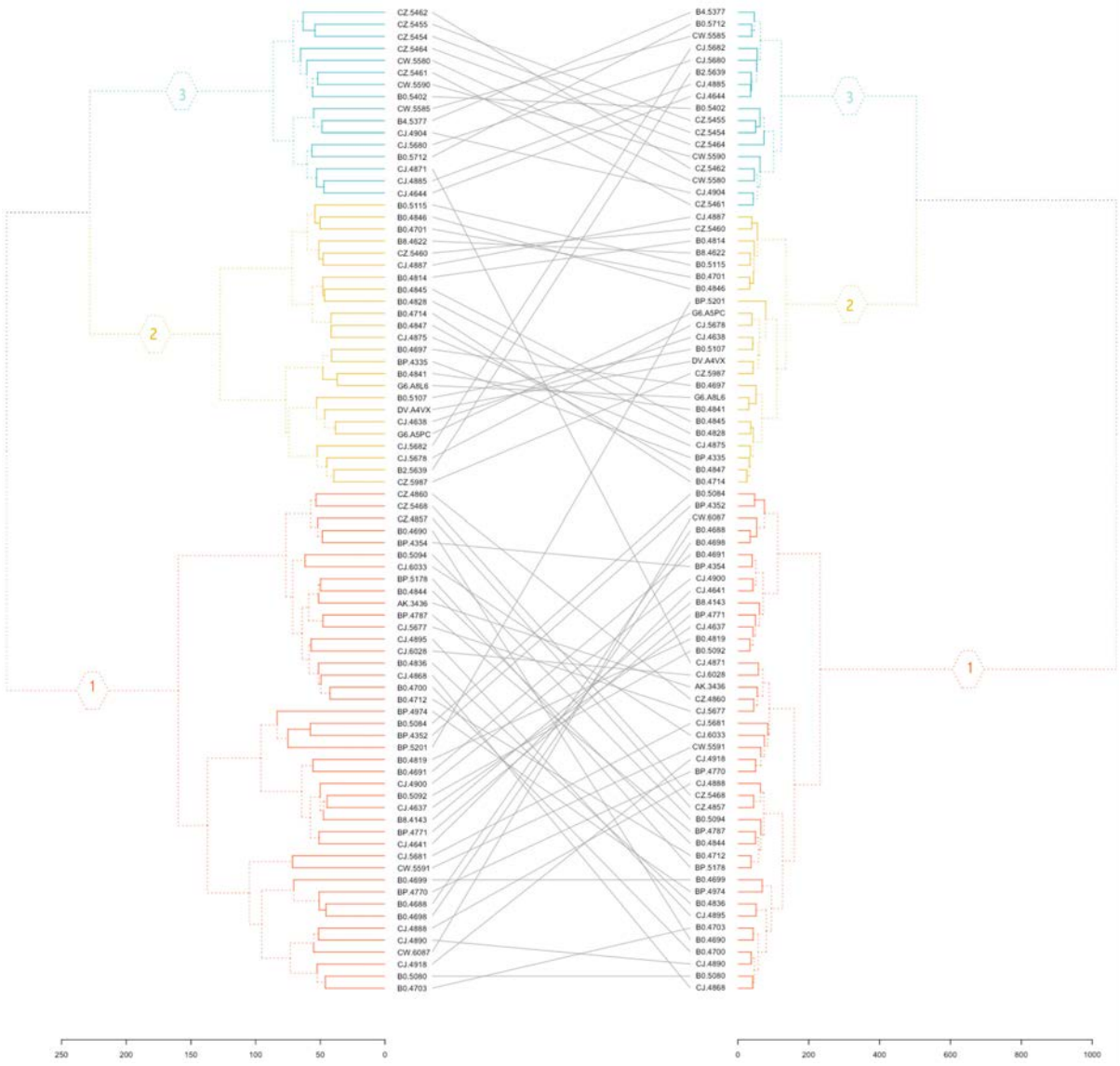

Figure 4.4. Tanglegram. The initial hierarchical clustering of 19,556 protein-coding genes (left) and clustering analysis of the 48 RFE gene subset (right).

4.4.4 Identification of prognostic genes

Five prognostic genes were identified and linked with KIRC patient survival by univariate Cox regression analysis between the 48 RFE gene subset and patient survival data. The prognostic genes were detected utilizing the LASSO algorithm, which assigns non-zero, positive, or negative coefficients. Two of the five genes had positive coefficients, while three genes had negative coefficients (Table 4.2).

Table 4.2: Five prognostic genes. The coefficient value obtained by LASSO algorithm.

Gene name	Coefficient value
SALL4	0.06613418699953
KRT15	0.0296694189909953
OSBPL11	-0.121246995833747
SPATA18	-0.0770127595245775
TAL2	-0.18919349247905

Based on patient statuses, the K-M estimations for overall survival were derived and presented below. The K-M curves illustrate low, intermediated, and high gene expression in blue, green, and red colors, respectively. The K-M curves of genes *SALL4* and *KRT15* with positive coefficient values are presented in Figure 4.5.

The K-M curves for the three genes *OSBPL11*, *SPATA18*, and *TAL2* with negative coefficient values are presented in Figure 4.6.

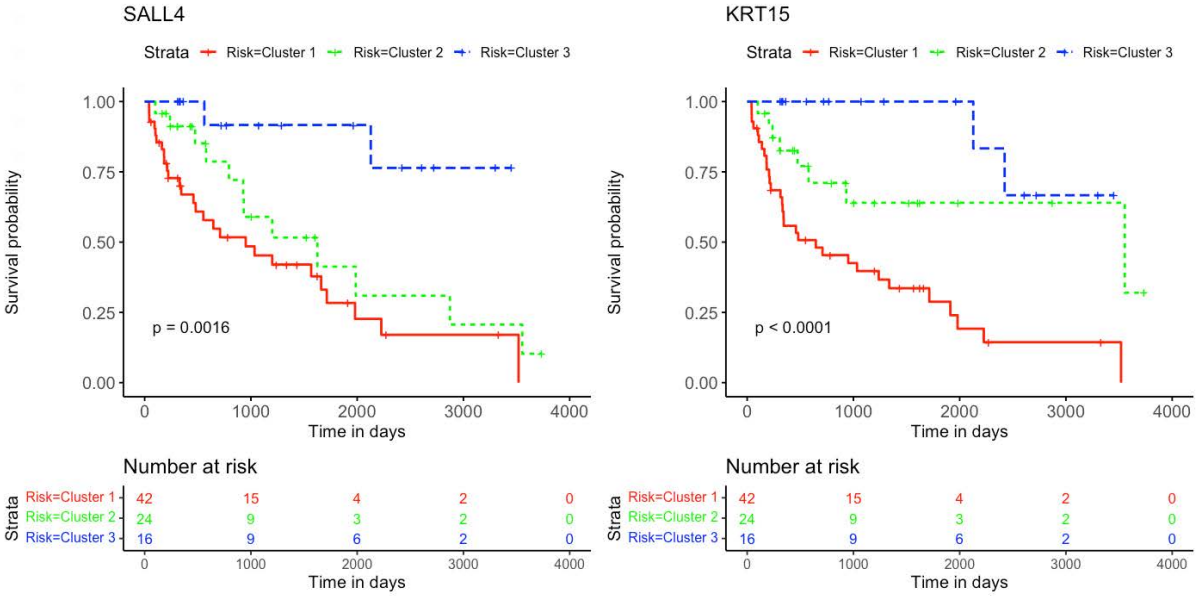

Figure 4.5. Kaplan-Meier survival curves of *SALL4* and *KRT15*. Analysis revealed the survival prediction associated with high and low gene expression profiles of *SALL4* and *KRT15* prognostic genes in KIRC patients.

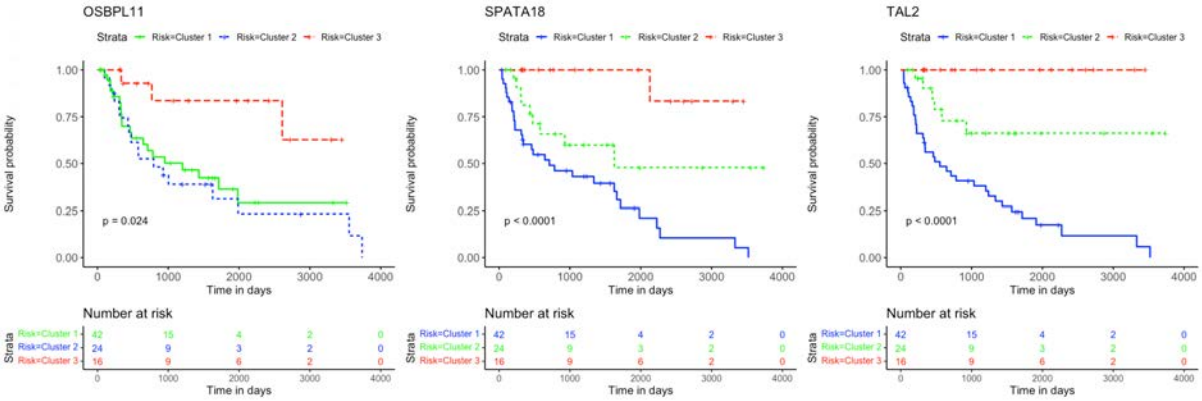

Figure 4.6. Kaplan-Meier survival curves of *OSBPL11*, *SPATA18*, and *TAL2*. Analysis revealed the survival prediction associated with high and low gene expression profiles of *OSBPL11*, *SPATA18*, and *TAL2* prognostic genes in KIRC patients.

The five prognostic genes' estimations and *p*-values in the Cox regression model were all significant, which demonstrates that the altered expression of these genes affects KIRC survival.

4.4.5 Gene expression patterns between risk subcategories

One-way ANOVA was performed to assess for differences in the mean normalized gene expression profiles of each of the prognostic genes detected between the risk subcategories. This evaluation included the differences between SS and IS, IS and LS, and SS and LS. Each survival group consisted of a set of samples that make up that risk subcategory, from which a boxplot was created using the normalized gene expression profile of a specific prognostic gene (Figure 4.7).

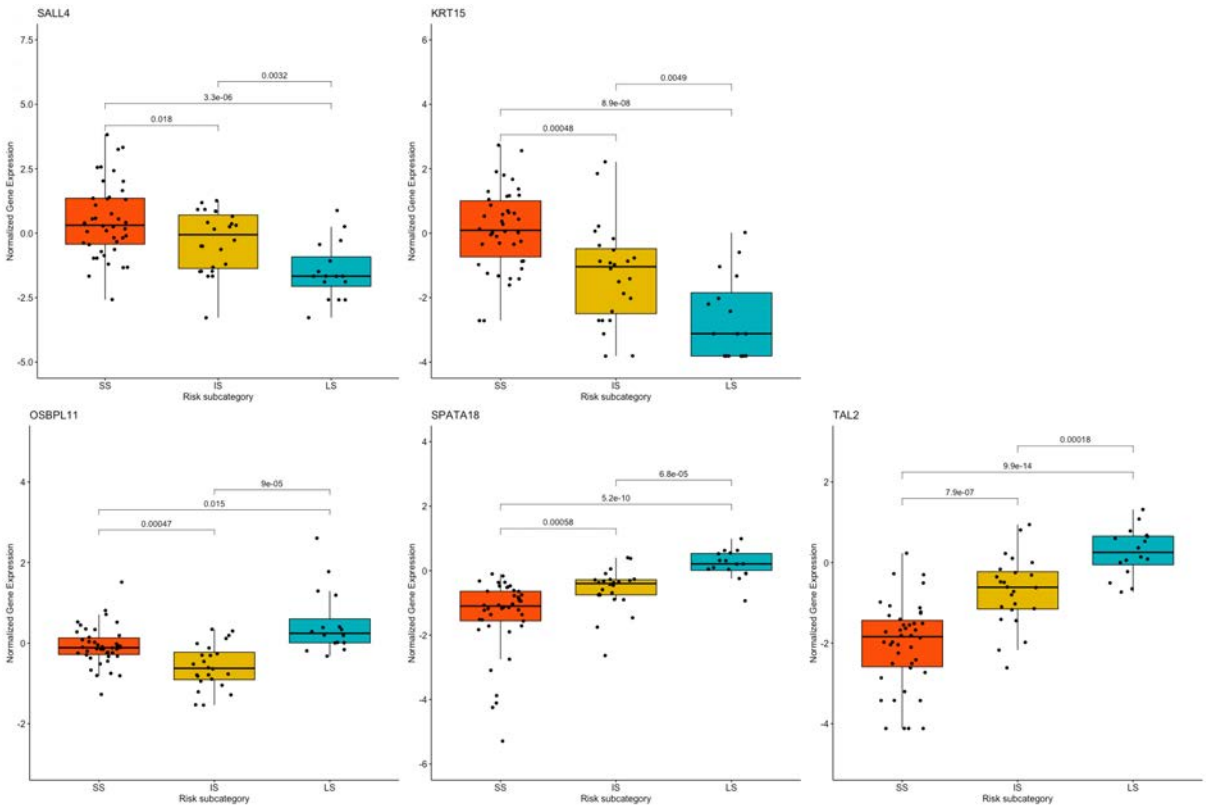

Figure 4.7. Boxplots based on risk subcategories of the five prognostic genes in KIRC patients. A boxplot was constructed with the normalized gene expression profile of each prognostic gene in all the samples that were categorized into the SS, IS, and LS categories.

All prognostic genes showed a statistically significant difference between SS and LS (*p*-value ≤ 0.015). It is further noteworthy that ANOVA resulted in a statistical difference in the normalized gene expression between IS and LS (*p*-value ≤ 0.0032) as well as between survival IS and SS (*p*-value ≤ 0.018) (Figure 4.7).

4.4.6 Enrichment analysis

The GO enrichment analysis illustrated that KIRC DEGs were significantly enriched in biological processes (BP), including extracellular matrix (ECM) organization, extracellular structure organization, and external encapsulating structure organization (Figure 4.8). In terms of cellular component (CC), collagen-containing ECM, cell leading edge, and cell projection membrane, among other terms were significantly enriched in KIRC DEGs (Figure 4.8). Lastly, the molecular function (MF), were significantly enriched in ECM structural constituent, growth factor binding, and hormone binding (Figure 4.8). The KEGG analysis revealed that the 48 gene subset significantly enriched for the p53 signaling pathway, HIF-1 signaling pathway, and estrogen signaling pathway (Figure 4.9).

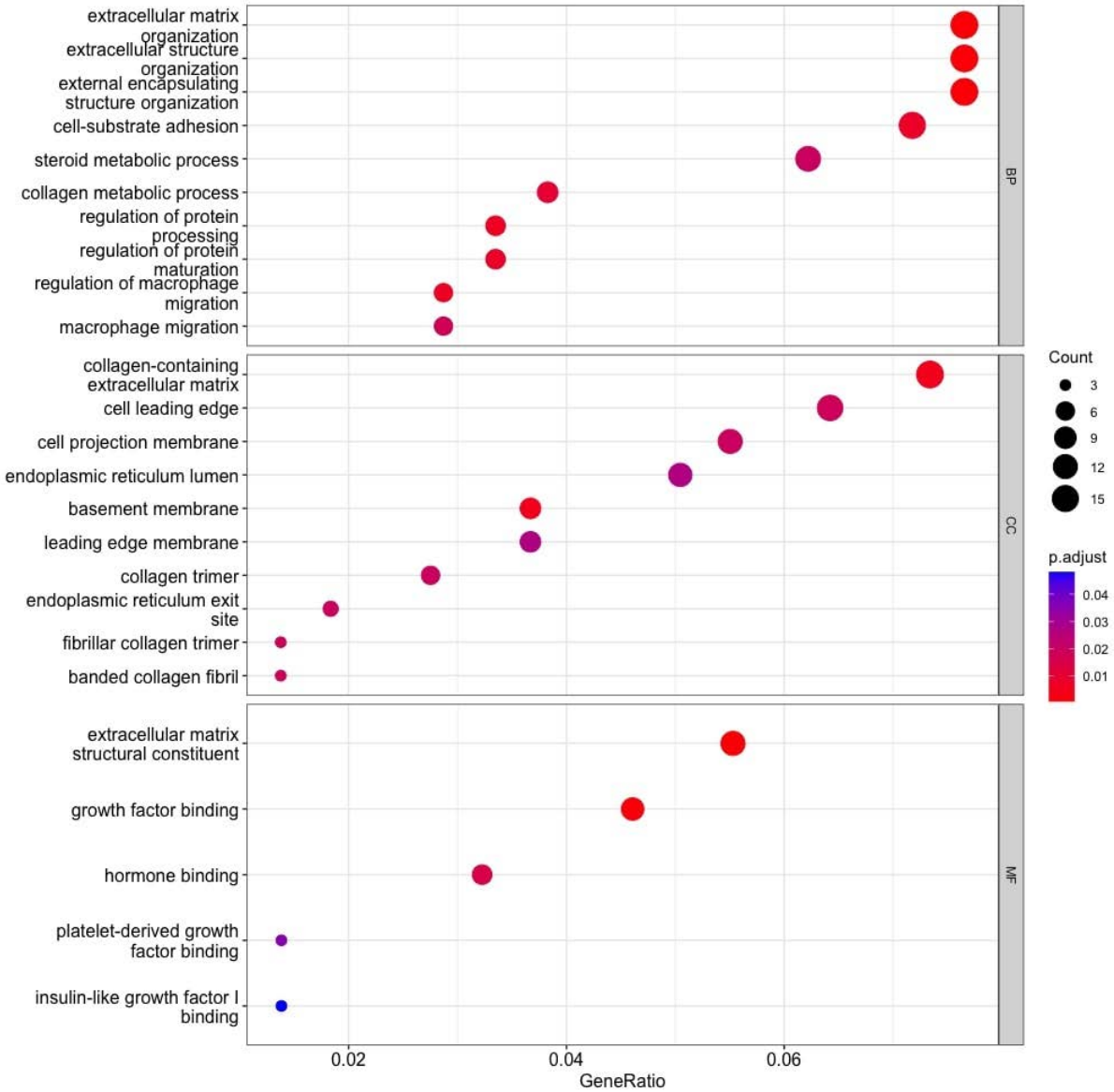

Figure 4.8. Gene Ontology enrichment analysis. Top 10 functional items of KIRC DEGs based on clusterProfiler. *Functional databases: BP, Biological process; CC, Cellular component; and MF, Molecular function.

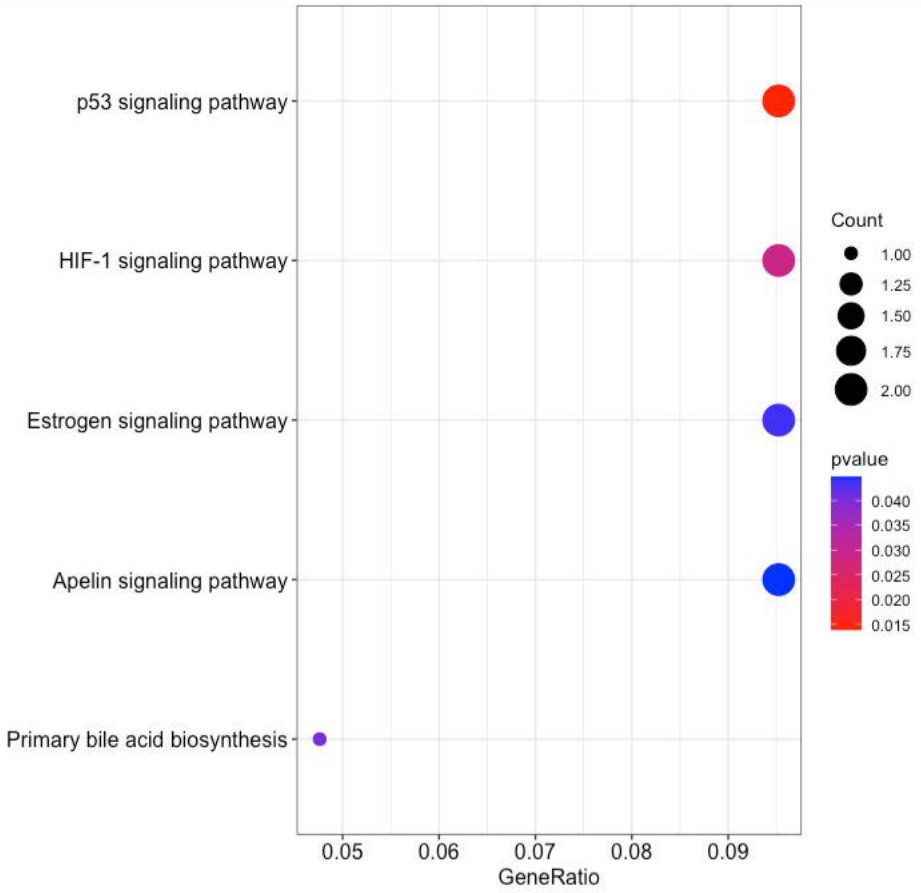

Figure 4.9. The results of KEGG pathways enrichment analysis of the 48 RFE gene subset based on clusterProfiler.

4.5 Discussion

The high molecular heterogeneity of RCC necessitates further sub-classification to establish a successful treatment strategy and medical care. Therefore, this study focussed on KIRC as it represents the majority of RCC diagnoses. The study aims to identify subtypes that reflect a genotype-phenotype relationship for KIRC patients that provide a more accurate prognosis, with an emphasis on cancer progression.

This study implemented a normalization method in which the gene expression profiles of eighty-two advanced-stage KIRC samples were normalized with early-stage cancer samples to

consider heterogeneity differences in the multi-stage cancer progression. The normalization method corrects for genes that present with high expression variability in early-stage samples but less expression variability in advanced-stage cancer samples. This leads to the availability of more meaningful information to track the cancer progression from early- to advanced-stage, based on the differences in the gene expression profiles.

The normalized gene expression was subjected to a hierarchical clustering method, to detect cancer samples that progress differently in gene expression patterns. The approach allows for the grouping, alternatively, clustering of cancer samples to identify samples within a group/cluster that are similar to each other and different from samples in other groups. This popular method revealed three cancer clusters (subtypes) for KIRC cancer. The three molecularly defined clusters were correlated with the patients' average OS. It can be noted that patients in Cluster 3 lived on average 657.88 days longer than patients in Cluster 1. Meanwhile patients in Cluster 2 and Cluster 3 live on average 211.95 days and 445.93 days longer than patients in Cluster 1 and Cluster 2, respectively. Thus, the obtained three clusters by the use of our normalization method illustrate different KIRC tumors that progressed differently from early-stage to late-stage cancer development (Figure 4.3). Consequently, these clusters have different prognoses and can be considered as different subtypes. The results of the hierarchical clustering analysis were subjected to a validation step using an independent GEO dataset (Appendix B, Table B2 & Figure B1). This test dataset includes sixty-five KIRC samples, and the normalization method also identified three clusters in the GEO KIRC dataset (Appendix B, Figure B2).

The 48 genes identified through the Machine Learning analysis have the capacity to accurately classify and predict the KIRC subtypes to an extent similar to the use of the 19,556 protein-

coding genes. This demonstrates the existence of genetic heterogeneity within KIRC tumors and the ability of our normalization method to recognize this heterogeneity and associate it with prognosis and OS. The gene set contains genes that were reported to play a critical role in the aggressiveness of renal tumors, and our study revealed their involvement in the heterogeneity of the most prevalent and aggressive subtype in renal cancer, KIRC.

Analysis of GO enrichment illustrates the involvement of DEGs in the biological processes that promote tumor aggressiveness. It has been reported that ECM regulates fundamental properties of tumors, such as growth and invasion. The most prevalent genetic mutations in KIRC inactivate the *VHL* gene, which plays a direct role in ECM organization. Therefore, therapeutic approaches to control ECM are currently being investigated and an advanced understanding of KIRC ECM will determine if ECM-modifying drugs are appropriate for KIRC (Oxburgh, 2022). An additional BP enrichment was macrophages that are highly enriched in RCC, and the RCC survival rate is strongly correlated with the inflammatory cytokines secreted by macrophages (Xie *et al*., 2022).

In terms of the cellular component, KIRC DEGs were significantly enriched in functional elements such as basement membrane (BM). According to a recent study, KIRC is associated with unique BM gene expression patterns, and the characterization of the BM has the potential to guide clinical therapy (Xiong *et al*., 2022). Cellular component, collagen trimer has been similarly found in studies focused on renal cancer progression (Wang A. *et al*., 2019), along with molecular function enriched extracellular matrix structural constituent and platelet-derived growth factor binding (Wang A. *et al*., 2019; van Roeyen *et al*., 2019). Lastly, MF is significantly enriched for hormone binding, and hormones plays a role in RCC etiology.

Hormone receptor expression in RCC cells has been demonstrated to be aberrant (Czarnecka *et al*., 2016).

Analysis of KEGG pathways revealed signalling pathways that promote cancer progression and resistance to therapies. The *SERPINE1* gene was enriched in the p53 signaling pathway, HIF-1 signaling pathway, and apelin signaling pathway. The interaction between P53 and HIF signaling can promote cancer progression (Zhang *et al*., 2021a). While apelin signaling has also been linked to the development of cancer and its progression (Liu *et al*., 2021). It is thus noteworthy, that the survival analysis of *SERPINE1* expression in TCGA found a correlation between shorter survival, and the increased tumor grade, lymph node metastasis, and tumor stage (Guo *et al*., 2023). Therefore, *SERPINE1* plays a crucial role in the progression of KIRC. KIRC patients categorized as SS revealed high levels of *SERPINE1* gene expression, whereas LS displayed low levels of gene expression. Hence, the method tracked the progression of KIRC and further indicated the potential of *SERPINE1* as a therapeutic target for KIRC patients.

Together with *SERPINE1*, the *PGK1* gene was also enriched for HIF-1 signaling pathway. HIF-1 is known to modulate a number of signaling pathways, having a significant impact on the cancer's response to radiotherapy (Huang & Zhou, 2020). Therefore, a viable approach for sensitization of KIRC to radiotherapy is to target *SERPINE1* and *PGK1*. Also, *PGK1* has been linked to several roles in the development of cancer, tumor progression, and drug resistance. The gene is known to promote sorafenib resistance, which is a first-line treatment for KIRC patients as a tyrosine kinase inhibitor. However, resistance to sorafenib significantly reduces the effectiveness of therapy (He *et al*., 2022). Therefore, the large patient group (n = 42),

accounting for about half of the KIRC patients investigated in this study encompassed in SS, may be affected by this resistance to therapy.

Genes *KRT15* and *GPER1* enriched for estrogen signaling pathways can also serve as treatment targets for KIRC patients. Estrogen is known to inhibit the proliferation, migration, and infiltration of RCC cells as well as increase RCC apoptosis (Yu *et al*., 2013). This study illustrated that the downregulation of *KRT15* had favorable prognostic outcomes for KIRC patients for Cluster 2 and 3 (Figures 4.5, 4.7), whereas the downregulation of *GPER1* was linked to unfavorable prognosis in Cluster 1. Therefore, the two genes may serve as valuable prognostic markers for KIRC and a novel developmental approach for enhancing KIRC therapeutics.

This study further identified five prognostic genes as promising prognostic biomarkers and treatment targets for KIRC patients (Table 4.2). Cox regression together with K-M analyses confirmed the prognostic biomarkers and showed that patients with high levels of *SALL4* and *KRT15* gene expression have a poor survival outcome than patients with low levels of gene expression (Figure 4.5). While the high gene expression level of *OSBPL11*, *SPATA18*, and *TAL2* has a favorable survival outcome than patients with a low level of gene expression (Figure 4.6). Therefore, K-M confirmed that the five genes are effective at diagnosing KIRC patients and predicting prognosis.

The results are supported by previous research, Ih indicated that the high gene expression level of *SALL4* has a poor survival outcome in comparison to KIRC patients with a low gene expression level (Che *et al*., 2020). Also, data from Sun et al. (2020) showed that the downregulation of *SALL4* reduces KIRC tumor growth, metastasis, and angiogenesis.

Therefore, it is noteworthy that Cluster 2 with intermediate survival followed a similar trend in cumulative survival probabilities as Cluster 1 with short survival (Figure 4.5). Furthermore, the high gene expression of *KRT15* has also been reported to correlate with a poor prognosis for RCC (Zhang *et al*., 2023). This study was able to detect *KRT15* as a prognostic gene in the KIRC subtype. The levels of gene expression correspond with the SS, IS, and LS (Figure 4.7). Previous studies have also reported higher levels of *SPATA18* gene expression associated with favorable OS in the KIRC subtype (Lingui *et al*., 2023) aswell as in RCC (The human protein atlas, 2023a). High expression of *TAL2* has been reported with a favorable OS in RCC (The human protein atlas, 2023b). This is the first article to our knowledge to report *OSBPL11* as a prognostic biomarker. A similar observation as with the *SALL4* K-M curve is observed with the *OSBPL11* gene. The K-M curve of Cluster 2 followed a similar trend in cumulative survival probabilities as Cluster 1 (Figure 4.6). Therefore, the upregulation of *OSBPL11* could reduce KIRC progression.

ANOVA was used to assess the heterogeneity in the prognostic genes' mean gene expression profiles, to establish whether SS, IS, and LS samples' gene expression profiles differ from one another. The prognostic value of the five prognostic genes found was confirmed by ANOVA, which also indicated a statistically significant difference in gene expression between short- and long-term survival. A crucial discovery was made between the gene expression profiles in the intermediate- and long survival as well as intermediate- and short survival. ANOVA showed statistically significant differences between the gene expression profiles of both IS and LS, and IS and SS. This further validates the finding of an intermediate-survival group. The unique gene expression pattern of each of the five prognostic genes were further subjected to a validation step using the independent GEO dataset (Appendix B, Table B2 & Figure B1). This test dataset verified prognostic genes *OSBPL11* and *TAL2* in the GEO dataset illustrated a

similar gene expression pattern for cluster 1 (short survival) and cluster 3 (long survival). The remaining three prognostic genes, *SALL4, KRT15*, and *SPATA18* showed similar gene expression patterns for all three clusters (Appendix B, Figure B3). The five prognostic genes are therefore essential as they may enable an improved KIRC patient prognosis based on the gene expression level of the five genes. Hence, this discovery is important as it is directly correlated with survival and could aid in predicting the outcome of KIRC patients.

The investigation detected molecular mechanisms that allowed for the segregation of three unique cancer clusters (subtypes) that progress differently in gene expression profiles and correlate with KIRC patient survival. Therefore, the normalization method was successfully implemented in this study and hierarchical clustering was able to provide an accurate assessment of the heterogeneity of KIRC. The cellular functions detected by GO enrichment along with the pathogenic genes detected by KEGG pathway analysis further confirmed the contribution to the progression of the disease. Additionally, the heterogeneity of KIRC served as a fuel for therapy resistance and emphasized the urgent need to expand the clinical subtypes for KIRC patients. As a result, this investigation facilitated and contributed to the current KIRC cancer classification with in-depth patient subtyping. The discovery of the five prognostic genes, combined with the biomarkers detected in pathway analysis, can provide a more accurate prognosis, and serve as targets to provide a more effective therapeutic approach for KIRC patients.

4.6 Conclusion

The implemented normalization method has the potential to reveal cancer patients that progress differently (subtypes) and establish a genotype-phenotype relationship between the identified

subtypes and the patient's OS. In this study, correlations between the risk subcategories and gene signatures differentiated short, intermediate, and long survival in KIRC patients. The prognostic capacity of the prognostic genes can successfully classify and predict the prognosis of KIRC patients. Moreover, the prognostic genes were able to segregate patients into additional survival subcategories and thus provide targets that can enhance patient prognosis and aid in the development of individualized treatment approaches.

Chapter 5

Conclusion and future recommendations

5.1 Conclusion

Cancer is a complex and dynamic genetic disease. During the multi-stage of cancer development, the disease generally becomes more heterogeneous. As a result of this heterogeneity, the tumor may consist of a diverse collection of cells harboring unique molecular signatures with differential levels of sensitivity to treatment. Consequently, this may be the cause of the poor overall survival associated with cancer. Therefore, this study focused on the discovery of cancer subtypes with the implementation and validation of a normalization method. In this study, the method captures the heterogeneity between cancerous tumors by detecting their molecular differences in progression between early- and advanced-stages of tumor development using gene expression by RNA-Seq.

The method examines the continuously changing cellular transcriptome, allowing for an efficient and comprehensive description of gene expression profiles between different conditions over time. The method calculates the quotient of cancerous samples (dividend) and

early-stage samples (divisor), thereby producing normalized differential RNA expression profiles within a specific condition. Therefore, it corrects for genes that display less expression variability in advanced-stage cancer samples but display a high variability in early-stage cancer samples. The method exposed the accumulated genetic changes that occur throughout the multi-stage of cancer development. Therefore, the application of the normalization method and hierarchical clustering allowed for the identification of cancer subtypes (clusters) that progressed differently. Therefore, the method facilitated the sub-classification of heterogeneous diseases.

Tracking of cancer progression demonstrated its potential to enhance the understanding of the molecular basis of carcinogenesis. The approach further demonstrated its potential to explore clinical relevance to the identified molecular subtypes that will enabled altered clinical approaches to heterogeneous diseases. Knowing the attributes of heterogeneity and their magnitude in carcinogenesis further allowed for the identification of biomarkers that can facilitate the screening and identification of individuals who are at risk of developing specific diseases, improve prognosis, or predict the response to treatment. The findings can further support the design of clinical trials for targeted therapies and stratification of heterogeneous cancer patients with differential therapeutic efficacy and prognosis.

5.2 Study limitations

The main limitation of this study was the number of cancer samples that were available to subject to the normalization method and downstream analyses in Chapter 3. A larger group of patients is recommended to validate the findings of this research. This will also render the five prognostic biomarkers identified in KIRC highly recommended for use in clinical applications.

An additional drawback was the lack of clinical information available in the phenotypic data for the lung samples to provide a reason for the segregation of Clusters 4 and 5 (Chapter 3). It would have also been of interest to validate the DEGs and RFE gene subset found in KIRC to the independent microarray GEO dataset (Chapter 4). However, the analysis of RNA-Seq has a higher sensitivity and specificity than microarray analysis.

5.3 Clinical importance

The discovery of cancer subtypes will have a significant impact on the field of cancer biology and precision medicine research. The approach outlined in this study allows for the accurate assessment of cancer heterogeneity and enables the tracking of cancer progression. The method facilitated the sub-classification of heterogeneous cancers and also allowed for the establishment of a genotype-phenotype link to the molecularly identified subtypes (clusters) and thus provided insight into clinical and phenotypic patterns of patient samples. This knowledge can be integrated into future clinal practices and research efforts to optimise patient care and clinical outcome. Additionally, the discovery of potential predictive biomarkers can also be implemented into clinical practices and improve the course of the disease. Therefore, the sub-classification of heterogeneous cancer allows for improved prognosis and the development of more effective targeted treatment strategies that aid in patients' welfare.

5.4 Future recommendations

This novel avenue for genome-based classification of heterogeneous cancers that focuses on the transcriptional landscape of tumor sequencing can be applied to numerous diseases to investigate the progression of the disease. The method can aid hypotheses that aim to

investigate new cancer subtypes that segregate by different gene expression profiles and also find the biological relationship, clinical characteristic, or prognostic features associated with the molecularly defined subtype. Also, the application can contribute to a better understanding of molecular heterogeneity linked to cancer.

The molecular biomarkers found in the study are vital for disease prognosis, treatment strategies, and outcome prediction. For clinical applications, it is highly recommended that the results obtained from the validation study be verified in a larger group of patients. The optimal RFE selected gene subset can further be used to accurately classify patients into subtypes with enhanced prognosis. The findings can further contribute to patient status monitoring and management to identify patients with short-, intermediate or long survival, as well as the development of targeted therapeutic strategies for the prognostic genes whose expression is associated with KIRC prognosis.

6 References

Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O., & Coombes, K. (2019). A protocol to evaluate RNA sequencing normalization methods. *BMC bioinformatics*, *20*(Suppl 24), 679. https://doi.org/10.1186/s12859-019-3247-x.

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare (Basel, Switzerland)*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. (2016). The Ensembl gene annotation system. *Database : the journal of biological databases and curation*, *2016*, baw093. https://doi.org/10.1093/database/baw093.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(12), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1), 25–29. https://doi.org/10.1038/75556.

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and Their Applications. *Journal of chromatographic science*, *55*(2), 182–196. https://doi.org/10.1093/chromsci/bmw167.

Bach, D. H., Long, N. P., Luu, T. T., Anh, N. H., Kwon, S. W., & Lee, S. K. (2018). The Dominant Role of Forkhead Box Proteins in Cancer. International journal of molecular sciences, 19(10), 3279. https://doi.org/10.3390/ijms19103279.

Ballman K. V. (2015). Biomarker: Predictive or Prognostic?. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 33(33), 3968–3971. https://doi.org/10.1200/JCO.2015.63.3651.

Baumgartner, C., Osl, M., Netzer, M., & Baumgartner, D. (2011). Bioinformatic-driven search for metabolic biomarkers in disease. *Journal of clinical bioinformatics*, *1*(1), 2. https://doi.org/10.1186/2043-9113-1-2.

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. BMJ (Clinical research ed.), 324(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018.

Beg, A., & Parveen, R. (2021). Chapter 11—role of bioinformatics in cancer research and drug development. In Raza K, Dey NBT-TB in H and M (eds) Advances in ubiquitous sensing applications for healthcare, vol 13. Academic Press, Cambridge, pp 141–148. https://doi.org/10.1016/B978-0-323-89824-9.00011-2.

Beham-Schmid C. (2017). Aggressive lymphoma 2016: revision of the WHO classification. *Memo*, *10*(4), 248–254. https://doi.org/10.1007/s12254-017-0367-8.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 57(1):289–300.

Berger, M. F., & Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature reviews. Clinical oncology*, *15*(6), 353–365. https://doi.org/10.1038/s41571-018-0002-6.

Bhowmick, S. S., Bhattacharjee, D., & Rato, L. (2019). Identification of tissue-specific tumor biomarker using different optimization algorithms. *Genes & genomics*, *41*(4), 431–443. https://doi.org/10.1007/s13258-018-0773-2.

Bi, Y., & Davuluri, R.V. (2020). Platform-Independent Gene-Expression Based Classification-System for Molecular Sub-typing of Cancer. In: Adam, T., Aliferis, C. (eds) Personalized and

Precision Medicine Informatics. Health Informatics. Springer, Cham. https://doi.org/10.1007/978-3-030-18626-5_10.

Blok, E. J., Bastiaannet, E., van den Hout, W. B., Liefers, G. J., Smit, V. T. H. B. M., Kroep, J. R., & van de Velde, C. J. H. (2018). Systematic review of the clinical and economic value of gene expression profiles for invasive early breast cancer available in Europe. *Cancer treatment reviews*, *62*, 74–90. https://doi.org/10.1016/j.ctrv.2017.10.012.

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, *19*(2), 185–193. https://doi.org/10.1093/bioinformatics/19.2.185.

Borisov, N., Sorokin, M., Tkachev, V., Garazha, A., & Buzdin, A. (2020). Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments. *BMC medical genomics*, *13*(Suppl 8), 111. https://doi.org/10.1186/s12920-020-00759-0.

Brademan, D. R., Miller, I. J., Kwiecien, N. W., Pagliarini, D. J., Westphall, M. S., Coon, J. J., & Shishkova, E. (2020). Argonaut: A Web Platform for Collaborative Multi-omic Data Visualization and Exploration. *Patterns (New York, N.Y.)*, *1*(7), 100122. https://doi.org/10.1016/j.patter.2020.100122.

Califf R. M. (2018). Biomarker definitions and their applications. Experimental biology and medicine (Maywood, N.J.), 243(3), 213–221. https://doi.org/10.1177/1535370217750088.

Caliskan, A., Andac, A. C., & Arga, K. Y. (2020). Novel molecular signatures and potential therapeutics in renal cell carcinomas: Insights from a comparative analysis of subtypes. Genomics, 112(5), 3166–3178. https://doi.org/10.1016/j.ygeno.2020.06.003.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764.

Canning, M., Guo, G., Yu, M., Myint, C., Groves, M. W., Byrd, J. K., & Cui, Y. (2019). Heterogeneity of the Head and Neck Squamous Cell Carcinoma Immune Landscape and Its Impact on Immunotherapy. *Frontiers in cell and developmental biology*, *7*, 52. https://doi.org/10.3389/fcell.2019.00052.

Cao, J., Gong, J., Li, X., Hu, Z., Xu, Y., Shi, H., et al. (2021). Unsupervised Hierarchical Clustering Identifies Immune Gene Subtypes in Gastric Cancer. *Frontiers in pharmacology*, *12*, 692454. https://doi.org/10.3389/fphar.2021.692454.

Carbone A. (2020). Cancer Classification at the Crossroads. Cancers, 12(4), 980. https://doi.org/10.3390/cancers12040980.

Casamassimi, A., Federico, A., Rienzo, M., Esposito, S., & Ciccodicola, A. (2017). Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *International journal of molecular sciences*, *18*(8), 1652. https://doi.org/10.3390/ijms18081652.

Casuscelli, J., Vano, Y. A., Fridman, W. H., & Hsieh, J. J. (2017). Molecular Classification of Renal Cell Carcinoma and Its Implication in Future Clinical Practice. *Kidney cancer (Clifton, Va.)*, *1*(1), 3–13. https://doi.org/10.3233/KCA-170008.

Chawade, A., Alexandersson, E., & Levander, F. (2014). Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *Journal of proteome research*, *13*(6), 3114–3120. https://doi.org/10.1021/pr401264n.

Che, J., Wu, P., Wang, G., Yao, X., Zheng, J., & Guo, C. (2020). Expression and clinical value of *SALL4* in renal cell carcinomas. *Molecular medicine reports*, *22*(2), 819–827. https://doi.org/10.3892/mmr.2020.11170.

Chen, W. X., Yang, L. G., Xu, L. Y., Cheng, L., Qian, Q., Sun, L., & Zhu, Y. L. (2019). Bioinformatics analysis revealing prognostic significance of *RRM2* gene in breast cancer. *Bioscience reports*, *39*(4), BSR20182062. https://doi.org/10.1042/BSR20182062.

Chen, R., Yang, L., Goodison, S., & Sun, Y. (2020). Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics (Oxford, England), 36(5), 1476–1483. https://doi.org/10.1093/bioinformatics/btz769.

Chin, L., Hahn, W. C., Getz, G., & Meyerson, M. (2011). Making sense of cancer genomic data. *Genes & development*, *25*(6), 534–555. https://doi.org/10.1101/gad.2017311.

Cieślik, M., & Chinnaiyan, A. M. (2018). Cancer transcriptome profiling at the juncture of clinical translation. *Nature reviews. Genetics*, *19*(2), 93–109. https://doi.org/10.1038/nrg.2017.96.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *British journal of cancer*, *89*(2), 232–238. https://doi.org/10.1038/sj.bjc.6601118.

Clifford, G. M., Rickenbach, M., Polesel, J., Dal Maso, L., Steffen, I., Ledergerber, B., et al. (2008). Influence of HIV-related immunodeficiency on the risk of hepatocellular carcinoma. *AIDS (London, England)*, *22*(16), 2135–2141. https://doi.org/10.1097/QAD.0b013e32831103ad.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, *44*(8), e71. https://doi.org/10.1093/nar/gkv1507.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8.

Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates. (2000). The Development and Causes of Cancer. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9963/.

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, *12*(12), e0190152. https://doi.org/10.1371/journal.pone.0190152.

Creighton C. J. (2018). Making Use of Cancer Genomic Databases. *Current protocols in molecular biology*, *121*, 19.14.1–19.14.13. https://doi.org/10.1002/cpmb.49.

Crucitta, S., Cucchiara, F., Mathijssen, R., Mateo, J., Jager, A., Joosse, A., et al. (2022). Treatment-driven tumour heterogeneity and drug resistance: Lessons from solid tumours. *Cancer treatment reviews*, *104*, 102340. https://doi.org/10.1016/j.ctrv.2022.102340.

Czarnecka, A. M., Niedzwiedzka, M., Porta, C., & Szczylik, C. (2016). Hormone signaling pathways as treatment targets in renal cell cancer (Review). *International journal of oncology*, *48*(6), 2221–2235. https://doi.org/10.3892/ijo.2016.3460.

D'Agostino, N., Li, W., & Wang, D. (2022). High-throughput transcriptomics. *Scientific reports*, *12*(1), 20313. https://doi.org/10.1038/s41598-022-23985-1.

D'Argenio V. (2018). The High-Throughput Analyses Era: Are We Ready for the Data Struggle?. High-throughput, 7(1), 8. https://doi.org/10.3390/ht7010008.

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature reviews. Clinical oncology*, *15*(2), 81–94. https://doi.org/10.1038/nrclinonc.2017.166.

de Anda-Jáuregui, G., & Hernández-Lemus, E. (2020). Computational Oncology in the Multi-Omics Era: State of the Art. *Frontiers in oncology*, *10*, 423. https://doi.org/10.3389/fonc.2020.00423.

de Carvalho, P. S., Leal, F. E., & Soares, M. A. (2021). Clinical and Molecular Properties of Human Immunodeficiency Virus-Related Diffuse Large B-Cell Lymphoma. *Frontiers in oncology*, *11*, 675353. https://doi.org/10.3389/fonc.2021.675353.

de Haas, T., Oussoren, E., Grajkowska, W., Perek-Polnik, M., Popovic, M., Zadravec-Zaletel, L., et al. (2006). OTX1 and OTX2 expression correlates with the clinicopathologic classification of medulloblastomas. Journal of neuropathology and experimental neurology, 65(2), 176–186. https://doi.org/10.1097/01.jnen.0000199576.70923.8a.

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC bioinformatics*, *9*, 497. https://doi.org/10.1186/1471-2105-9-497.

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., et al. (2018). A review of machine learning in obesity. *Obesity reviews : an official journal of the International Association for the Study of Obesity*, *19*(5), 668–685. https://doi.org/10.1111/obr.12667.

Deng, Z. M., Dai, F. F., Zhou, Q., & Cheng, Y. X. (2021). I_circ_0000301 facilitates the progression of cervical cancer by targeting miR-1228-3p/IRF4 Axis. *BMC cancer*, *21*(1), 583. https://doi.org/10.1186/s12885-021-08331-4.

Ding, R., Liu, Q., Yu, J., Wang, Y., Gao, H., Kan, H., & Yang, Y. (2023). Identification of Breast Cancer Subtypes by Integrating Genomic Analysis with the Immune Microenvironment. ACS omega, 8(13), 12217–12231. https://doi.org/10.1021/acsomega.2c08227.

Du, W., & Searle, J. S. (2009). The rb pathway and cancer therapeutics. *Current drug targets*, *10*(7), 581–589. https://doi.org/10.2174/138945009788680392.

Dugué, P. A., Rebolj, M., Garred, P., & Lynge, E. (2013). Immunosuppression and risk of cervical cancer. *Expert review of anticancer therapy*, *13*(1), 29–42. https://doi.org/10.1586/era.12.159.

Dunn, O.J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, 6:241–52. https://doi.org/10.1080/00401706.1964.10490181.

El-Deiry, W. S., Taylor, B., & Neal, J. W. (2017). Tumor Evolution, Heterogeneity, and Therapy for Our Patients With Advanced Cancer: How Far Have We Come?. *American Society*

of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting, *37*, e8–e15. https://doi.org/10.1200/EDBK_175524.

El Khoury, L. Y., Pan, X., Hlady, R. A., Wagner, R. T., Shaikh, S., Wang, L., et al. (2023). Extensive intratumor regional epigenetic heterogeneity in clear cell renal cell carcinoma targets kidney enhancers and is associated with poor outcome. *Clinical epigenetics*, *15*(1), 71. https://doi.org/10.1186/s13148-023-01471-3.

Ergin, S., Kherad, N., & Alagoz, M. (2022). RNA sequencing and its applications in cancer and rare diseases. *Molecular biology reports*, *49*(3), 2325–2333. https://doi.org/10.1007/s11033-021-06963-0.

Eshibona, N., Giwa, A., Rossouw, S. C., Gamieldien, J., Christoffels, A., & Bendou, H. (2022). Upregulation of *FHL1*, *SPNS3*, and *MPZL2* predicts poor prognosis in pediatric acute myeloid leukemia patients with *FLT3-ITD* mutation. *Leukemia & lymphoma*, *63*(8), 1897–1906. https://doi.org/10.1080/10428194.2022.2045594.

Eshibona, N., Livesey, M., Christoffels, A., & Bendou, H. (2023). Investigation of distinct gene expression profile patterns that can improve the classification of intermediate-risk prognosis in AML patients. *Frontiers in genetics*, *14*, 1131159. https://doi.org/10.3389/fgene.2023.1131159.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews. Genetics*, *8*(4), 286–298. https://doi.org/10.1038/nrg2005.

Farnie, G., Clarke, R. B., Spence, K., Pinnock, N., Brennan, K., Anderson, N. G., & Bundred, N. J. (2007). Novel cell culture technique for primary ductal carcinoma in situ: role of Notch and epidermal growth factor receptor signaling pathways. Journal of the National Cancer Institute, 99(8), 616–627. https://doi.org/10.1093/jnci/djk133.

FDA-NIH Biomarker Working Group. (2016). BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK326791/Co-published by National Institutes of Health (US), Bethesda (MD).

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 10.1002/ijc.33588.

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in functional genomics*, *14*(2), 130–142. https://doi.org/10.1093/bfgp/elu035.

Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*. 1:3–32.

Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, *108*(3), 479–485. https://doi.org/10.1038/bjc.2012.581.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, *33*(1), 1–22.

Friedmann-Morvinski, D. (2014). Glioblastoma heterogeneity and cancer cell plasticity. *Critical reviews in oncogenesis*, *19*(5), 327–336. https://doi.org/10.1615/critrevoncog.2014011777.

Frost, F. G., Cherukuri, P. F., Milanovich, S., & Boerkoel, C. F. (2020). Pan-cancer RNA-seq data stratifies tumours by some hallmarks of cancer. *Journal of cellular and molecular medicine*, *24*(1), 418–430. https://doi.org/10.1111/jcmm.14746.

Fujii, H., Yoshida, M., Gong, Z. X., Matsumoto, T., Hamano, Y., Fukunaga, M., et al. (2000). Frequent genetic heterogeneity in the clonal evolution of gynecological carcinosarcoma and its influence on phenotypic diversity. *Cancer research*, *60*(1), 114–120.

Fujimoto, J., Aoki, I., Toyoki, H., Khatun, S., & Tamaya, T. (2002). Clinical implications of expression of ETS-1 related to angiogenesis in uterine cervical cancers. *Annals of oncology : official journal of the European Society for Medical Oncology*, *13*(10), 1598–1604. https://doi.org/10.1093/annonc/mdf248.

Gan, Y., Li, N., Zou, G., Xin, Y., & Guan, J. (2018). Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method. *BMC medical genomics*, *11*(Suppl 6), 117. https://doi.org/10.1186/s12920-018-0433-z.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, *6*(269), pl1. https://doi.org/10.1126/scisignal.2004088.

Gao, G. F., Parker, J. S., Reynolds, S. M., Silva, T. C., Wang, L. B., Zhou, W., et al. (2019). Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell systems*, *9*(1), 24–34.e10. https://doi.org/10.1016/j.cels.2019.06.006.

Giwa, A., Fatai, A., Gamieldien, J., Christoffels, A., & Bendou, H. (2020). Identification of novel prognostic markers of survival time in high-risk neuroblastoma using gene expression profiles. *Oncotarget*, *11*(46), 4293–4305. https://doi.org/10.18632/oncotarget.27808.

Goedert, J. J., Purdue, M. P., McNeel, T. S., McGlynn, K. A., & Engels, E. A. (2007). Risk of germ cell tumors among men with HIV/acquired immunodeficiency syndrome. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *16*(6), 1266–1269. https://doi.org/10.1158/1055-9965.EPI-07-0042.

Goldman, M., Craft, B., Swatloski, T., Ellrott, K., Cline, M., Diekhans, M., et al. (2013). The UCSC Cancer Genomics Browser: update 2013. *Nucleic acids research*, *41*(Database issue), D949–D954. https://doi.org/10.1093/nar/gks1008.

Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., et al. (2015). The UCSC Cancer Genomics Browser: update 2015. *Nucleic acids research*, *43*(Database issue), D812–D817. https://doi.org/10.1093/nar/gku1073.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature biotechnology*, *38*(6), 675–678. https://doi.org/10.1038/s41587-020-0546-8.

Goossens, N., Nakagawa, S., Sun, X., & Hoshida, Y. (2015). Cancer biomarker discovery and validation. *Translational cancer research*, *4*(3), 256–269. https://doi.org/10.3978/j.issn.2218-676X.2015.06.04

Gray, R. E., and Harris, G. T. (2019). Renal cell carcinoma: diagnosis and management. Am. Fam. Physician 99 (3), 179–184.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *The New England journal of medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591.

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653.

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110.

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277.

Guo, L., An, T., Wan, Z., Huang, Z., & Chong, T. (2023). SERPINE1 and its co-expressed genes are associated with the progression of clear cell renal cell carcinoma. BMC urology, 23(1), 43. https://doi.org/10.1186/s12894-023-01217-6.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1–3): 389–422. https://doi.org/10.1023/A:1012487302797.

Hammerich-Hille, S., Bardout, V. J., Hilsenbeck, S. G., Osborne, C. K., & Oesterreich, S. (2010). Low SAFB levels are associated with worse outcome in breast cancer patients. Breast cancer research and treatment, 121(2), 503–509. https://doi.org/10.1007/s10549-008-0297-6.

Han, H., & Men, K. (2018). How does normalization impact RNA-seq disease diagnosis? *J Biomed Inform*. 85: 80-92. https://doi.org/10.1016/j.jbi.2018.07.016.

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1.

Hassanpour, H., & Dehghani, M. (2017). Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice*, 4(1), 127-129. https://doi.org/10.1016/j.jcrpr.2017.07.001.

He, Y., Wang, X., Lu, W., Zhang, D., Huang, L., Luo, Y., et al. (2022). PGK1 contributes to tumorigenesis and sorafenib resistance of renal clear cell carcinoma via activating CXCR4/ERK signaling pathway and accelerating glycolysis. Cell death & disease, 13(2), 118. https://doi.org/10.1038/s41419-022-04576-4.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, *173*(2), 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

Hong, M., Tao, S., Zhang, L., Diao, L. T., Huang, X., Huang, S., et al. (2020). RNA sequencing: new technologies and applications in cancer research. Journal of hematology & oncology, 13(1), 166. https://doi.org/10.1186/s13045-020-01005-x.

Hu, F., Zeng, W., & Liu, X. (2019). A Gene Signature of Survival Prediction for Kidney Renal Cell Carcinoma by Multi-Omic Data Analysis. *International journal of molecular sciences*, *20*(22), 5720. https://doi.org/10.3390/ijms20225720.

Hu, C., Liu, C., Li, J., Yu, T., Dong, J., Chen, B., et al. (2021). Construction of Two Alternative Polyadenylation Signatures to Predict the Prognosis of Sarcoma Patients. *Frontiers in cell and developmental biology*, *9*, 595331. https://doi.org/10.3389/fcell.2021.595331.

Huang, R., & Zhou, P.K. (2020). HIF-1 signaling: a key orchestrator of cancer radio resistance. *Radiation Medicine and Protection*. 1 (1), 7–14. https://doi.org/10.1016/j.radmp.2020.01.006.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature, 578(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature, 431(7011), 931–945. https://doi.org/10.1038/nature03001.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, *166*(3), 740–754. https://doi.org/10.1016/j.cell.2016.06.017.

Jamal-Hanjani, M., Quezada, S. A., Larkin, J., & Swanton, C. (2015). Translational implications of tumor heterogeneity. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *21*(6), 1258–1266. https://doi.org/10.1158/1078-0432.CCR-14-1429.

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B. K., Veeriah, S., et al. (2017). Tracking the Evolution of Non-Small-Cell Lung Cancer. *The New England journal of medicine*, *376*(22), 2109–2121. https://doi.org/10.1056/NEJMoa1616288.

Jamil, A., & Mukkamalla, S.K.R. (2022). Lymphoma. In: StatPearls. Treasure Island (FL): StatPearls Publishing.

Jaskowiak, P. A., Campello, R. J., & Costa, I. G. (2014). On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics*, *15 Suppl 2*(Suppl 2), S2. https://doi.org/10.1186/1471-2105-15-S2-S2.

Jensen, J., Pedersen, E. E., Galante, P., Hald, J., Heller, R. S., Ishibashi, M., et al. (2000). Control of endodermal endocrine development by Hes-1. Nature genetics, 24(1), 36–44. https://doi.org/10.1038/71657.

Jia, C., Ma, Y., Wang, M., Liu, W., Tang, F., & Chen, J. (2022). Evidence of Omics, Immune Infiltration, and Pharmacogenomics for BATF in a Pan-Cancer Cohort. Frontiers in molecular biosciences, 9, 844721. https://doi.org/10.3389/fmolb.2022.844721.

Jiang, P., Sinha, S., Aldape, K., Hannenhalli, S., Sahinalp, C., & Ruppin, E. (2022). Big data in basic and translational cancer research. Nature reviews. Cancer, 22(11), 625–639. https://doi.org/10.1038/s41568-022-00502-0.

Jin, T. Y., Saindane, M., Park, K. S., Kim, S., Nam, S., Yoo, Y., et al. (2021). LEP as a potential biomarker in prognosis of breast cancer: Systemic review and meta analyses (PRISMA). *Medicine*, *100*(33), e26896. https://doi.org/10.1097/MD.0000000000026896.

Jazayeri, S. M., Melgarejo, L. M., & Romero, H. M. (2015). RNA-Seq: a glance at technologies and methodologies. Acta Biologica Colombiana. 20. 10.15446/abc.v20n2.43639.

Kaffenberger, S. D., & Barbieri, C. E. (2016). Molecular subtyping of prostate cancer. *Current opinion in urology*, *26*(3), 213–218. https://doi.org/10.1097/MOU.0000000000000285.

Kais, G., & Hamdi, Y. (2022). Introductory Chapter: Application of Bioinformatics Tools in Cancer Prevention, Screening, and Diagnosis, Cancer Bioinformatics. *IntechOpen*. Available at: https://doi.org/10.5772/intechopen.104794.

Kakati, T., Bhattacharyya, D. K., Barah, P., & Kalita, J. K. (2019). Comparison of Methods for Differential Co-expression Analysis for Disease Biomarker Prediction. *Computers in biology and medicine*, *113*, 103380. https://doi.org/10.1016/j.compbiomed.2019.103380.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research, 45(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092.

Kang, T., Yau, C., Wong, C. K., Sanborn, J. Z., Newton, Y., Vaske, C., et al. (2020). A risk-associated Active transcriptome phenotype expressed by histologically normal human breast tissue and linked to a pro-tumorigenic adipocyte population. *Breast cancer research : BCR*, *22*(1), 81. https://doi.org/10.1186/s13058-020-01322-6.

Kannan, S., O'Connor, G. M., & Bakker, E. Y. (2021). Molecular Mechanisms of PD-1 and PD-L1 Activity on a Pan-Cancer Basis: A Bioinformatic Exploratory Study. *International journal of molecular sciences*, *22*(11), 5478. https://doi.org/10.3390/ijms22115478.

Kassambara, A., & Mundt, F. (2020). Factoextra: extract and visualize the results of multivariate data analyses. R Package Version. 1.0.7 https://CRAN.R-project.org/package=factoextra.

Katoh, M., & Katoh, M. (2007). Integrative genomic analyses on HES/HEY family: Notch-independent HES1, HES3 transcription in undifferentiated ES cells, and Notch-dependent HES1, HES5, HEY1, HEY2, HEYL transcription in fetal tissues, adult tissues, or cancer. International journal of oncology, 31(2), 461–466. https://doi.org/10.3892/ijo.31.2.461.

Keenan, A. B., Torre, D., Lachmann, A., Leong, A. K., Wojciechowicz, M. L., Utti, V., et al. (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic acids research*, *47*(W1), W212–W224. https://doi.org/10.1093/nar/gkz446.

Kelder, T., Pico, A. R., Hanspers, K., van Iersel, M. P., Evelo, C., & Conklin, B. R. (2009). Mining biological pathways using WikiPathways web services. *PloS one*, *4*(7), e6447. https://doi.org/10.1371/journal.pone.0006447.

Kim, L. K., Park, S. A., Eoh, K. J., Heo, T. H., Kim, Y. T., & Kim, H. J. (2020). E2F8 regulates the proliferation and invasion through epithelial-mesenchymal transition in cervical cancer. International journal of biological sciences, 16(2), 320–329. https://doi.org/10.7150/ijbs.37686.

Kou, F., Sun, H., Wu, L., Li, B., Zhang, B., Wang, X., & Yang, L. (2020). TOP2A Promotes Lung Adenocarcinoma Cells' Malignant Progression and Predicts Poor Prognosis in Lung Adenocarcinoma. *Journal of Cancer*, *11*(9), 2496–2508. https://doi.org/10.7150/jca.41415.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

Kovacs, G., Akhtar, M., Beckwith, B. J., Bugert, P., Cooper, C. S., Delahunt, B., et al. (1997). The Heidelberg classification of renal cell tumours. *The Journal of pathology*, *183*(2), 131–133. https://doi.org/10.1002/(SICI)1096-9896(199710)183:2<131::AID-PATH931>3.0.CO;2-G.

Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers in genetics*, *11*, 610798. https://doi.org/10.3389/fgene.2020.610798.

Kruskal, W.H., & Wallis, W.A. (1952) Use of Ranks in One-Criterion Variance Analysis. *J Am Stat Assoc*, 47 (260):583–621. https://doi.org/10.1080/01621459.1952.10483441.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970.

Kwon, W., Choi, S. K., Kim, D., Kim, H. G., Park, J. K., Han, J. E., et al. (2021). ZNF507 affects TGF-β signaling via TGFBR1 and MAP3K8 activation in the progression of prostate cancer to an aggressive state. Journal of experimental & clinical cancer research : CR, 40(1), 291. https://doi.org/10.1186/s13046-021-02094-3.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*, 559. https://doi.org/10.1186/1471-2105-9-559.

Langfelder, P., & Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of statistical software*, *46*(11), i11.

Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature reviews. Genetics*, *19*(4), 208–219. https://doi.org/10.1038/nrg.2017.113.

Lavallee, E., Sfakianos, J. P., & Mulholland, D. J. (2021). Tumor Heterogeneity and Consequences for Bladder Cancer Treatment. *Cancers*, *13*(21), 5297. https://doi.org/10.3390/cancers13215297.

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*, *5*, ISCB Comm J-1408. https://doi.org/10.12688/f1000research.9005.3.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1–18. https://doi.org/10.18637/jss.v025.i01.

Lee, S. U., & Maeda, T. (2012). POK/ZBTB proteins: an emerging family of proteins that regulate lymphoid development and function. Immunological reviews, 247(1), 107–119. https://doi.org/10.1111/j.1600-065X.2012.01116.x.

Levene, H., Olkin, I.I., & Hotelling, H. (1960). Robust Tests for Equality of Variances. Contributions to Probability and Statistics; Essays in Honor of Harold Hotelling, 78–92.

Li, Y., Gan, S., Ren, L., Yuan, L., Liu, J., Wang, W., et al. (2018a). Multifaceted regulation and functions of replication factor C family in human cancers. *American journal of cancer research*, *8*(8), 1343–1355.

Li, Z., Fan, P., Deng, M., & Zeng, C. (2018b). The roles of RUNX3 in cervical cancer cells *in vitro*. *Oncology letters*, *15*(6), 8729–8734. https://doi.org/10.3892/ol.2018.8419.

Li, R., Li, P., Xing, W., & Qiu, H. (2020). Heterogeneous genomic aberrations in esophageal squamous cell carcinoma: a review. *American journal of translational research*, *12*(5), 1553–1568.

Liang, R., Xiao, G., Wang, M., Li, X., Li, Y., Hui, Z., Sun, X., Qin, S., Zhang, B., Du, N., Liu, D., & Ren, H. (2018). SNHG6 functions as a competing endogenous RNA to regulate E2F7 expression by sponging miR-26a-5p in lung adenocarcinoma. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie, 107, 1434–1446. https://doi.org/10.1016/j.biopha.2018.08.099.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped Uis and APIs. *Nucleic acids research*, *47*(W1), W199–W205. https://doi.org/10.1093/nar/gkz401.

Lim, Z. F., & Ma, P. C. (2019). Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *Journal of hematology & oncology*, *12*(1), 134. https://doi.org/10.1186/s13045-019-0818-2.

Lin, Z. W., Wu, L. X., Xie, Y., Ou, X., Tian, P. K., Liu, X. P., et al. (2015). The expression levels of transcription factors T-bet, GATA-3, RORγt and FOXP3 in peripheral blood lymphocyte (PBL) of patients with liver cancer and their significance. International journal of medical sciences, 12(1), 7–16. https://doi.org/10.7150/ijms.8352.

Lingui, X., Weifeng, L., Yufei, W., & Yibin, Z. (2023). High SPATA18 expression and its diagnostic and prognostic value in clear cell renal cell carcinoma. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*. 29, e938474. https://doi.org/10.12659/msm.938474.

Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., & Gerlinger, M. (2016). Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in cancer*, *2*(1), 49–63. https://doi.org/10.1016/j.trecan.2015.11.003.

Liu, W., Li, L., Ye, H., & Tu, W. (2017). Weighted gene co-expression network analysis in biomedicine research. Sheng Wu Gong Cheng Xue Bao, 33(11),1791-1801. https://doi.org/10.13345/j.cjb.170006.

Liu, L., Yi, X., Lu, C., Wang, Y., Xiao, Q., Zhang, L., et al. (2021). Study Progression of Apelin/APJ Signaling and Apela in Different Types of Cancer. *Frontiers in oncology*, *11*, 658253. https://doi.org/10.3389/fonc.2021.658253.

Livesey, M., Rossouw, S. C., Blignaut, R., Christoffels, A., & Bendou, H. (2023). Transforming RNA-Seq gene expression to track cancer progression in the multi-stage early to

advanced-stage cancer development. *PloS one*, 18(4), e0284458. https://doi.org/10.1371/journal.pone.0284458.

Lleo, A., de Boer, Y. S., Liberal, R., & Colombo, M. (2019). The risk of liver cancer in autoimmune liver diseases. *Therapeutic advances in medical oncology*, *11*, 1758835919861914. https://doi.org/10.1177/1758835919861914.

Lopez-Beltran, A., Scarpelli, M., Montironi, R., & Kirkali, Z. (2006). 2004 WHO classification of the renal tumors of the adults. *European urology*, *49*(5), 798–805. https://doi.org/10.1016/j.eururo.2005.11.035.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8.

Ma, S., Huang, J., & Shen, S. (2009). Identification of cancer-associated gene clusters and genes via clustering penalization. *Statistics and its interface*, *2*(1), 1–11. https://doi.org/10.4310/sii.2009.v2.n1.a1.

Ma, Y. S., Lv, Z. W., Yu, F., Chang, Z. Y., Cong, X. L., Zhong, X. M., et al. (2018). MicroRNA-302a/d inhibits the self-renewal capability and cell cycle entry of liver cancer stem cells by targeting the E2F7/AKT axis. Journal of experimental & clinical cancer research : CR, 37(1), 252. https://doi.org/10.1186/s13046-018-0927-8.

Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L., & Siu, L. L. (2020). Molecular profiling for precision cancer therapies. *Genome medicine*, *12*(1), 8. https://doi.org/10.1186/s13073-019-0703-1.

Marshall, A. D., Bailey, C. G., Champ, K., Vellozzi, M., O'Young, P., Metierre, C., Feng, Y., Thoeng, A., Richards, A. M., Schmitz, U., Biro, M., Jayasinghe, R., Ding, L., Anderson, L., Mardis, E. R., & Rasko, J. E. J. (2017). CTCF genetic alterations in endometrial carcinoma are pro-tumorigenic. Oncogene, 36(29), 4100–4110. https://doi.org/10.1038/onc.2017.25.

Marshall, A. E., Roes, M. V., Passos, D. T., DeWeerd, M. C., Chaikovsky, A. C., Sage, J., et al. (2019). *RB1* Deletion in Retinoblastoma Protein Pathway-Disrupted Cells Results in DNA Damage and Cancer Progression. *Molecular and cellular biology*, *39*(16), e00105-19. https://doi.org/10.1128/MCB.00105-19.

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., et al. (2021). WikiPathways: connecting communities. Nucleic acids research, 49(D1), D613–D621. https://doi.org/10.1093/nar/gkaa1024.

Marusyk, A., Janiszewska, M., & Polyak, K. (2020). Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer cell*, *37*(4), 471–484. https://doi.org/10.1016/j.ccell.2020.03.007.

Matek, C., Schwarz, S., Spiekermann, K., & Marr, C. (2019). Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell*, 1:538–544. https://doi.org/10.1038/s42256-019-0101-9.

McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science translational medicine*, *7*(283), 283ra54. https://doi.org/10.1126/scitranslmed.aaa1408.

McGranahan, N., & Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, *168*(4), 613–628. https://doi.org/10.1016/j.cell.2017.01.018.

Meacham, C. E., & Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, *501*(7467), 328–337. https://doi.org/10.1038/nature12624.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nature reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626.

Miralaei, N., Majd, A., Ghaedi, K., Peymani, M., & Safaei, M. (2021). Integrated pan-cancer of AURKA expression and drug sensitivity analysis reveals increased expression of AURKA

is responsible for drug resistance. *Cancer medicine*, *10*(18), 6428–6441. https://doi.org/10.1002/cam4.4161.

Mkrtchyan, G. V., Veviorskiy, A., Izumchenko, E., Shneyderman, A., Pun, F. W., Ozerov, I. V., et al. (2022). High-confidence cancer patient stratification through multiomics investigation of DNA repair disorders. *Cell death & disease*, *13*(11), 999. https://doi.org/10.1038/s41419-022-05437-w.

Moch, H., Humphrey, P. A., Ulbright, T. M., and Reuter, V. E. (2016). WHO classification of tumours of the urinary system and male genital organs. 4th ed. Lyon (France): International Agency for Research on Cancer.

Mohanty, S. K., Lobo, A., & Cheng, L. (2023). The 2022 revision of the World Health Organization classification of tumors of the urinary system and male genital organs: advances and challenges. *Human pathology*, *136*, 123–143. https://doi.org/10.1016/j.humpath.2022.08.006.

Moore, D. C., & Guinigundo, A. S. (2023). Biomarker-Driven Oncology Clinical Trials: Novel Designs in the Era of Precision Medicine. *Journal of the advanced practitioner in oncology*, *14*(Suppl 1), 9–13. https://doi.org/10.6004/jadpro.2023.14.3.16.

Morgan, G. J., Walker, B. A., & Davies, F. E. (2012). The genetic architecture of multiple myeloma. *Nature reviews. Cancer*, *12*(5), 335–348. https://doi.org/10.1038/nrc3257.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226.

Mörth, C., Valachis, A., Abu Sabaa, A., Marshall, K., Hedström, G., Flogegård, M., et al. (2019). Autoimmune disease in patients with diffuse large B-cell lymphoma: occurrence and impact on outcome. *Acta oncologica (Stockholm, Sweden)*, *58*(8), 1170–1177. https://doi.org/10.1080/0284186X.2019.1619936.

Mroz, E. A., & Rocco, J. W. (2016). Intra-tumor heterogeneity in head and neck cancer and its clinical implications. *World journal of otorhinolaryngology – head and neck surgery*, *2*(2), 60–67. https://doi.org/10.1016/j.wjorl.2016.05.007.

Natrajan, R., Sailem, H., Mardakheh, F. K., Arias Garcia, M., Tape, C. J., Dowsett, M., et al. (2016). Microenvironmental Heterogeneity Parallels Breast Cancer Progression: A Histology-Genomic Integration Analysis. *PloS medicine*, *13*(2), e1001961. https://doi.org/10.1371/journal.pmed.1001961.

Niemira, M., Collin, F., Szalkowska, A., Bielska, A., Chwialkowska, K., Reszec, J., et al. (2019). Molecular Signature of Subtypes of Non-Small-Cell Lung Cancer by Large-Scale Transcriptional Profiling: Identification of Key Modules and Genes by Weighted Gene Co-Expression Network Analysis (WGCNA). *Cancers*, *12*(1), 37. https://doi.org/10.3390/cancers12010037.

Novelli, G., Ciccacci, C., Borgiani, P., Papaluca Amati, M., & Abadie, E. (2008). Genetic tests and genomic biomarkers: regulation, qualification and validation. *Clinical cases in mineral and bone metabolism : the official journal of the Italian Society of Osteoporosis, Mineral Metabolism, and Skeletal Diseases*, *5*(2), 149–154.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1), 29–34, https://doi.org/10.1093/nar/27.1.29.

Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L.E., Gómez-Valleset, F. O., & Ramírez-Valdespino C. A. (2022) Subtypes of Breast Cancer. In: Mayrovitz HN, editor. Breast Cancer [Internet]. Brisbane (AU): Exon Publications; Chapter 3. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583808/ doi: 10.36255/exon-publications-breast-cancer-subtypes.

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, *4*, 14. https://doi.org/10.1186/1745-6150-4-14.

Oxburgh L. (2022). The Extracellular Matrix Environment of Clear Cell Renal Cell Carcinoma. *Cancers*, *14*(17), 4072. https://doi.org/10.3390/cancers14174072.

Ozturk, K., Dow, M., Carlin, D. E., Bejar, R., & Carter, H. (2018). The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. *Journal of molecular biology*, *430*(18 Pt A), 2875–2899. https://doi.org/10.1016/j.jmb.2018.06.016.

Parajuli, G., Tekguc, M., Wing, J. B., Hashimoto, A., Okuzaki, D., Hirata, T., et al. (2021). Arid5a Promotes Immune Evasion by Augmenting Tryptophan Metabolism and Chemokine Expression. Cancer immunology research, 9(8), 862–876. https://doi.org/10.1158/2326-6066.CIR-21-0014.

Park, J., Lee, J. W., & Park, M. (2023). Comparison of cancer subtype identification methods combined with feature selection methods in omics data analysis. *BioData mining*, *16*(1), 18. https://doi.org/10.1186/s13040-023-00334-0.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370.

Paszkiewicz, K. H., & Giezen, M. V. D. (2011). Omics, Bioinformatics, and Infectious Disease Research. Genetics and Evolution of Infectious Disease, 523–539. https://doi.org/10.1016/B978-0-12-384890-1.00018-2.

Pavlopoulou, A., Spandidos, D. A., & Michalopoulos, I. (2015). Human cancer databases (review). *Oncology reports*, *33*(1), 3–18. https://doi.org/10.3892/or.2014.3579.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit- learn Machine learning in python. *Journal of Machine Learning Research*, *12*: 2825–2830.

Pettini, F., Visibelli, A., Cicaloni, V., Iovinelli, D., & Spiga, O. (2021). Multi-Omics Model Applied to Cancer Genetics. *International journal of molecular sciences*, *22*(11), 5751. https://doi.org/10.3390/ijms22115751.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., & Evelo, C. (2008). WikiPathways: pathway editing for the people. PloS biology, 6(7), e184. https://doi.org/10.1371/journal.pbio.0060184.

Pinu, F. R., Goldansaz, S. A., & Jaine, J. (2019). Translational Metabolomics: Current Challenges and Future Opportunities. *Metabolites*, *9*(6), 108. https://doi.org/10.3390/metabo9060108.

Proietto, M., Crippa, M., Damiani, C., Pasquale, V., Sacco, E., Vanoni, M., & Gilardi, M. (2023). Tumor heterogeneity: preclinical models, emerging technologies, and future applications. *Frontiers in oncology*, *13*, 1164535. https://doi.org/10.3389/fonc.2023.1164535.

Puzanov G.A. (2022). Identification of key genes of the ccRCC subtype with poor prognosis. *Scientific reports*, *12*(1), 14588. https://doi.org/10.1038/s41598-022-18620-y.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, *32 Suppl*, 496–501. https://doi.org/10.1038/ng1032.

Rafique, R., Islam, S. M. R., & Kazi, J. U. (2021). Machine learning in the prediction of cancer therapy. Computational and structural biotechnology journal, 19, 4003–4017. https://doi.org/10.1016/j.csbj.2021.07.003.

Rajesh, S., Bansal, K., Sureka, B., Patidar, Y., Bihari, C., & Arora, A. (2015). The imaging conundrum of hepatic lymphoma revisited. *Insights into imaging*, *6*(6), 679–692. https://doi.org/10.1007/s13244-015-0437-6.

Ramón Y Cajal, S., Sesé, M., Capdevila, C., Aasen, T., De Mattos-Arruda, L., Diaz-Cano, S. J., et al. (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of molecular medicine (Berlin, Germany)*, *98*(2), 161–177. https://doi.org/10.1007/s00109-020-01874-2.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **14**, 3158. https://doi.org/10.1186/gb-2013-14-9-r95.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616.

Rohani, N., & Eslahchi, C. (2020). Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach. *Frontiers in genetics*, *11*, 553587. https://doi.org/10.3389/fgene.2020.553587.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Roy, S., Sharma, P., Nath, K., Bhattacharyya, D.K., & Kalita, J.K. (2019). Pre-Processing: A Data Preparation Step. In: Guenther R, Steel D. (eds.). Encyclopedia of Bioinformatics and Computational Biology, 1:463–471. Oxford:Elsevier.

Russnes, H. G., Lingjærde, O. C., Børresen-Dale, A. L., & Caldas, C. (2017). Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *The American journal of pathology*, *187*(10), 2152–2162. https://doi.org/10.1016/j.ajpath.2017.04.022.

Sager, M., Yeat, N. C., Pajaro-Van der Stadt, S., Lin, C., Ren, Q., & Lin, J. (2015). Transcriptomics in cancer diagnostics: developments in technology, clinical research and commercialization. *Expert review of molecular diagnostics*, *15*(12), 1589–1603. https://doi.org/10.1586/14737159.2015.1105133.

Salarikia, S. R., Kashkooli, M., Taghipour, M. J., Malekpour, M., & Negahdaripour, M. (2022). Identification of hub pathways and drug candidates in gastric cancer through systems biology. *Scientific reports*, *12*(1), 9099. https://doi.org/10.1038/s41598-022-13052-0.

Saleh, A. D., Cheng, H., Martin, S. E., Si, H., Ormanoglu, P., Carlson, S., et al. (2019). Integrated Genomic and Functional microRNA Analysis Identifies miR-30-5p as a Tumor Suppressor and Potential Therapeutic Nanomedicine in Head and Neck Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research, 25(9), 2860–2873. https://doi.org/10.1158/1078-0432.CCR-18-0716.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell, 173(2), 321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035.

Sanchis, P., Lavignolle, R., Abbate, M., Lage-Vickers, S., Vazquez, E., Cotignola, J., Bizzotto, J., & Gueron, G. (2021). Analysis workflow of publicly available RNA-sequencing datasets. *STAR protocols*, *2*(2), 100478. https://doi.org/10.1016/j.xpro.2021.100478.

Sarhadi, V. K., & Armengol, G. (2022). Molecular Biomarkers in Cancer. *Biomolecules*, *12*(8), 1021. https://doi.org/10.3390/biom12081021.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN computer science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x.

Sarode, P., Zheng, X., Giotopoulou, G. A., Weigert, A., Kuenne, C., Günther, S., et al. (2020). Reprogramming of tumor-associated macrophages by targeting β-catenin/FOSL2/ARID5A signaling: A potential treatment of lung cancer. Science advances, 6(23), eaaz6105. https://doi.org/10.1126/sciadv.aaz6105.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467.

Schober, P., & Vetter, T. R. (2018). Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesthesia and analgesia, 127(3), 792–798. https://doi.org/10.1213/ANE.0000000000003653.

Shapiro, S.S., & Wilk, M.B. (1965). An Analysis of Variance Test for Normality (complete Samples). *Biometrika*, 52:591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Shen, S., & Wang, Y. (2021). Expression and Prognostic Role of E2F2 in Hepatocellular Carcinoma. International journal of general medicine, 14, 8463–8472. https://doi.org/10.2147/IJGM.S334033.

Shiels, M. S., Cole, S. R., Kirk, G. D., & Poole, C. (2009). A meta-analysis of the incidence of non-AIDS cancers in HIV-infected individuals. *Journal of acquired immune deficiency syndromes (1999)*, *52*(5), 611–622. https://doi.org/10.1097/QAI.0b013e3181b327ca.

Shin, S. H., Kim, B. H., Jang, J. J., Suh, K. S., & Kang, G. H. (2010). Identification of novel methylation markers in hepatocellular carcinoma using a methylation array. Journal of Korean medical science, 25(8), 1152–1159. https://doi.org/10.3346/jkms.2010.25.8.1152.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of statistical software*, *39*(5), 1–13. https://doi.org/10.18637/jss.v039.i05.

Singh, M. P., Rai, S., Pandey, A., Singh, N. K., & Srivastava, S. (2019). Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. *Genes & diseases*, *8*(2), 133–145. https://doi.org/10.1016/j.gendis.2019.10.013.

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic acids research, 46(D1), D661–D667. https://doi.org/10.1093/nar/gkx1064.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data

repositories. *Nucleic acids research*, *43*(W1), W589–W598. https://doi.org/10.1093/nar/gkv350.

Song, C., Lee, Y., & Kim, S. (2022). Bioinformatic Analysis for the Prognostic Implication of Genes Encoding Epithelial Sodium Channel in Cervical Cancer. *International journal of general medicine*, *15*, 1777–1787. https://doi.org/10.2147/IJGM.S346222.

Srigley, J. R., Delahunt, B., Eble, J. N., Egevad, L., Epstein, J. I., Grignon, D., et al. (2013). The International Society of Urological Pathology (ISUP) Vancouver Classification of Renal Neoplasia. *The American journal of surgical pathology*, *37*(10), 1469–1489. https://doi.org/10.1097/PAS.0b013e318299f2d1.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, *171*(6), 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and biology insights*, *14*, 1177932219899051. https://doi.org/10.1177/1177932219899051.

Sun, J., Tang, Q., Gao, Y., Zhang, W., Zhao, Z., Yang, F., et al. (2020). VHL mutation-mediated SALL4 overexpression promotes tumorigenesis and vascularization of clear cell renal cell carcinoma via Akt/GSK-3β signaling. *Journal of experimental & clinical cancer research : CR*, *39*(1), 104. https://doi.org/10.1186/s13046-020-01609-8.

Sun, P., Wu, Y., Yin, C., Jiang, H., Xu, Y., & Sun, H. (2022). Molecular Subtyping of Cancer Based on Distinguishing Co-Expression Modules and Machine Learning. *Frontiers in genetics*, *13*, 866005. https://doi.org/10.3389/fgene.2022.866005.

Szalat, R., Avet-Loiseau, H., & Munshi, N. C. (2016). Gene Expression Profiles in Myeloma: Ready for the Real World?. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *22*(22), 5434–5442. https://doi.org/10.1158/1078-0432.CCR-16-0867.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome research*, *21*(12), 2213–2223. https://doi.org/10.1101/gr.124321.111.

Terrinoni, A., Pagani, I. S., Zucchi, I., Chiaravalli, A. M., Serra, V., Rovera, F., et al. (2011). OTX1 expression in breast cancer is regulated by p53. Oncogene, 30(27), 3096–3103. https://doi.org/10.1038/onc.2011.31.

The human protein atlas. (2022). Human pathology atlas: MAP4K1 gene. Available From: https://www.proteinatlas.org/ENSG00000104814-MAP4K1/pathology (Accessed June 7, 2022).

The human protein atlas. (2023a). Human pathology atlas. SPATA18 gene. Available From: https://www.proteinatlas.org/ENSG00000163071-SPATA18/pathology/renal+cancer (Accessed August 20, 2023).

The human protein atlas. (2023b). Human pathology atlas. TAL2 gene. Available From: https://www.proteinatlas.org/ENSG00000186051-TAL2/pathology/renal+cancer (Accessed August 20, 2023).

The human protein atlas (2023c). Human Pathology Atlas. CHCHD3 gene. Available from: https://www.proteinatlas.org/ENSG00000106554-CHCHD3/pathology/liver+cancer. (Accessed Jan 9, 2023).

The human protein atlas. (2023d). Human Pathology Atlas. ZNF581 gene. Available from: https://www.proteinatlas.org/ENSG00000171425-ZNF581/pathology/liver+cancer. (Accessed Jan 9, 2023).

The human protein atlas (2023e). Human Pathology Atlas. ZNF207 gene. Available from: https://www.proteinatlas.org/ENSG00000010244-ZNF207/pathology/liver+cancer. (Accessed Jan 9, 2023).

The human protein atlas (2023f). Human Pathology Atlas. CENPA gene. Available from: https://www.proteinatlas.org/ENSG00000115163-CENPA/pathology/liver+cancer. (Accessed Jan 9, 2023).

The human protein atlas. (2023g). Human Pathology Atlas. CENPA gene. Available from: https://www.proteinatlas.org/ENSG00000115163-CENPA/pathology/lung+cancer. (Accessed Jan 9, 2023).

The human protein atlas. (2023h). Human Pathology Atlas. SNAI3 gene. Available from: https://www.proteinatlas.org/ENSG00000185669-SNAI3/pathology/cervical+cancer (Accessed Jan 10, 2023).

The human protein atlas. (2023i). Human Pathology Atlas. IKZF1 gene. Available from: https://www.proteinatlas.org/ENSG00000185811-IKZF1/pathology/cervical+cancer (Accessed Jan 10, 2023).

The human protein atlas. (2023j). Human Pathology Atlas. IKZF3 gene. Available from: https://www.proteinatlas.org/ENSG00000161405-IKZF3/pathology/cervical+cancer (Accessed Jan 10, 2023).

The human protein atlas. (2023k). Human Pathology Atlas. FOXP3 gene. Available from: https://www.proteinatlas.org/ENSG00000049768-FOXP3/pathology/cervical+cancer (Accessed Jan 10, 2023).

The human protein atlas. (2023l). Human Pathology Atlas. ZNF266 gene. Available from: https://www.proteinatlas.org/ENSG00000174652-ZNF266/pathology/cervical+cancer (Accessed Jan 10, 2023).

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *74*(2), 245–266. https://doi.org/10.1111/j.1467-9868.2011.01004.x.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7:562-578. https://doi.org/10.1038/nprot.2012.016.

Trpkov, K., Williamson, S. R., Gill, A. J., Adeniran, A. J., Agaimy, A., Alaghehbandan, R., et al. (2021a). Novel, emerging and provisional renal entities: The Genitourinary Pathology Society (GUPS) update on renal neoplasia. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, *34*(6), 1167–1184. https://doi.org/10.1038/s41379-021-00737-6.

Trpkov, K., Hes, O., Williamson, S. R., Adeniran, A. J., Agaimy, A., Alaghehbandan, R., et al. (2021b). New developments in existing WHO entities and evolving molecular concepts: The Genitourinary Pathology Society (GUPS) update on renal neoplasia. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, *34*(7), 1392–1424. https://doi.org/10.1038/s41379-021-00779-w.

Tuasha, N., & Petros, B. (2020). Heterogeneity of Tumors in Breast Cancer: Implications and Prospects for Prognosis and Therapeutics. *Scientifica*, 4736091. https://doi.org/10.1155/2020/4736091.

Tukey J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99–114. https://doi.org/10.2307/3001913.

Udager, A. M., & Mehra, R. (2016). Morphologic, Molecular, and Taxonomic Evolution of Renal Cell Carcinoma: A Conceptual Perspective With Emphasis on Updates to the 2016 World Health Organization Classification. *Archives of pathology & laboratory medicine*, *140*(10), 1026–1037. https://doi.org/10.5858/arpa.2016-0218-RA.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science (New York, N.Y.)*, *357*(6352), eaan2507. https://doi.org/10.1126/science.aan2507.

van Roeyen, C. R. C., Martin, I. V., Drescher, A., Schuett, K. A., Hermert, D., Raffetseder, U., et al. (2019). Identification of platelet-derived growth factor C as a mediator of both renal

fibrosis and hypertension. *Kidney international*, *95*(5), 1103–1119. https://doi.org/10.1016/j.kint.2018.11.031.

Venkata, C. (2019). Tumor Heterogeneity in Preclinical Oncology Models – Crown Bioscience. https://blog.crownbio.com/pdx-tumor-heterogeneity.

Vidman, L., Källberg, D., & Rydén, P. (2019). Cluster analysis on high dimensional RNA-seq data with applications to cancer research – An evaluation study. *PloS one*, *14*(12), e0219102. https://doi.org/10.1371/journal.pone.0219102.

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature biotechnology*, *35*(4), 314–316. https://doi.org/10.1038/nbt.3772.

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nature medicine, 10(8), 789–799. https://doi.org/10.1038/nm1087.

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften*, *131*(4), 281–285. https://doi.org/10.1007/s12064-012-0162-3.

Wan, Y. W., Allen, G. I., & Liu, Z. (2016). TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics (Oxford, England)*, *32*(6), 952–954. https://doi.org/10.1093/bioinformatics/btv677.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484.

Wang, X., Fu, Y., Chen, X., Ye, J., Lü, B., Ye, F., et al. (2010). The expressions of bHLH gene HES1 and HES5 in advanced ovarian serous adenocarcinomas and their prognostic significance: a retrospective clinical study. Journal of cancer research and clinical oncology, 136(7), 989–996. https://doi.org/10.1007/s00432-009-0744-8.

Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based Gene SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research*, *41*(Web Server issue), W77–W83. https://doi.org/10.1093/nar/gkt439.

Wang, Z., Jensen, M. A., & Zenklusen, J. C. (2016). A Practical Guide to The Cancer Genome Atlas (TCGA). *Methods in molecular biology (Clifton, N.J.)*, *1418*, 111–141. https://doi.org/10.1007/978-1-4939-3578-9_6.

Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research*, *45*(W1), W130–W137. https://doi.org/10.1093/nar/gkx356.

Wang, B., Kumar, V., Olson, A., & Ware, D. (2019). Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing. *Frontiers in genetics*, *10*, 384. https://doi.org/10.3389/fgene.2019.00384.

Wang, A., Chen, M., Wang, H., Huang, J., Bao, Y., Gan, X., et al. (2019). Cell Adhesion-Related Molecules Play a Key Role in Renal Cancer Progression by Multinetwork Analysis. *BioMed research international*, *2019*, 2325765. https://doi.org/10.1155/2019/2325765.

Wang, Q., Zhang, H., Chen, Q., Wan, Z., Gao, X., & Qian, W. (2019). Identification of METTL14 in Kidney Renal Clear Cell Carcinoma Using Bioinformatics Analysis. *Disease markers*, *2019*, 5648783. https://doi.org/10.1155/2019/5648783.

Wang, J., Sun, Z., Wang, J., Tian, Q., Huang, R., Wang, H., et al. (2021). Expression and prognostic potential of PLEK2 in head and neck squamous cell carcinoma based on bioinformatics analysis. *Cancer medicine*, *10*(18), 6515–6533. https://doi.org/10.1002/cam4.4163.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58:236–244. https://doi.org/10.1080/01621459.1963.10500845.

Wu, M., Pang, J. S., Sun, Q., Huang, Y., Hou, J. Y., Chen, G., et al. (2019). The clinical significance of CHEK1 in breast cancer: a high-throughput data analysis and immunohistochemical study. *International journal of clinical and experimental pathology*, *12*(1), 1–20.

Wu, Y., Deng, Y., Wei, B., Xiang, D., Hu, J., Zhao, P., et al. (2022). Global, regional, and national childhood cancer burden, 1990-2019: An analysis based on the Global Burden of Disease Study 2019. Journal of advanced research, 40, 233–247. https://doi.org/10.1016/j.jare.2022.06.001.

Xie, D., Mao, Y., Du, N., Ji, H., & Li, J. (2022). Macrophages promote growth, migration and epithelial-mesenchymal transition of renal cell carcinoma by regulating GSDMD/IL-1β axis. *Cytokine*, *159*, 156021. https://doi.org/10.1016/j.cyto.2022.156021.

Xing, B., Shi, L., Bao, Z., Liang, Y., Liu, B., & Liu, R. (2022). Molecular clustering based on gene set expression and its relationship with prognosis in patients with lung adenocarcinoma. *Journal of thoracic disease*, *14*(5), 1638–1650. https://doi.org/10.21037/jtd-22-557.

Xiong, X., Chen, C., Yang, J., Ma, L., Wang, X., Zhang, W., et al. (2022). Characterization of the basement membrane in kidney renal clear cell carcinoma to guide clinical therapy. *Frontiers in oncology*, *12*, 1024956. https://doi.org/10.3389/fonc.2022.1024956.

Xu, F. L., Li, Y. L., Wang, Z. D., & Feng, Y. J. (2003). *Zhongguo yi 135u eke xue yuan xue bao. Acta Academiae Medicinae Sinicae*, *25*(4), 396–400.

Yamada, M., Saito, Y., Imaoka, H., Saiko, M., Yamada, S., Kondo, H., et al. (2019). Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Scientific reports*, *9*(1), 14465. https://doi.org/10.1038/s41598-019-50567-5.

Yamashita, M., Toyota, M., Suzuki, H., Nojima, M., Yamamoto, E., Kamimae, S., et al. (2010). DNA methylation of interferon regulatory factors in gastric cancer and noncancerous gastric

mucosae. Cancer science, 101(7), 1708–1716. https://doi.org/10.1111/j.1349-7006.2010.01581.x.

Yang, Y., Dong, X., Xie, B., Ding, N., Chen, J., Li, Y., et al. (2015). Databases and web tools for cancer genomics study. *Genomics, proteomics & bioinformatics*, *13*(1), 46–50. https://doi.org/10.1016/j.gpb.2015.01.005.

Yao, Y., Zhang, T., Qi, L., Liu, R., Liu, G., Li, J., & Sun, C. (2021). Identification of Four Genes as Prognosis Signatures in Lung Adenocarcinoma Microenvironment. Pharmacogenomics and personalized medicine, 14, 15–26. https://doi.org/10.2147/PGPM.S283414.

Yates, L.R., Seoane, J., Le Tourneau, C., Siu, L.L., Marais, R., Michiels S, et al. (2018). The European Society for Medical Oncology (ESMO) precision medicine glossary. *Ann Oncol,* **29**:30–35. Doi: 10.1093/annonc/mdx707.

Yin, L., Li, W., Wang, G., Shi, H., Wang, K., Yang, H., & Peng, B. (2019). NR1B2 suppress kidney renal clear cell carcinoma (KIRC) progression by regulation of LATS ½-YAP signaling. *Journal of experimental & clinical cancer research : CR*, *38*(1), 343. https://doi.org/10.1186/s13046-019-1344-3.

Ying, L., Yan, F., & Xu, D. (2020). Cancer patient stratification based on the tumor microenvironment. *Journal of thoracic disease*, *12*(8), 4522–4526. https://doi.org/10.21037/jtd.2020.03.77.

Yip, A. M., & Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, *8*, 22. https://doi.org/10.1186/1471-2105-8-22.

Yip, H. Y. K., & Papa, A. (2021). Signaling Pathways in Cancer: Therapeutic Targets, Combinatorial Treatments, and New Developments. Cells, 10(3), 659. https://doi.org/10.3390/cells10030659.

You, Y., Ren, Y., Liu, J., & Qu, J. (2021). Promising Epigenetic Biomarkers Associated With Cancer-Associated-Fibroblasts for Progression of Kidney Renal Clear Cell Carcinoma. *Frontiers in genetics*, *12*, 736156. https://doi.org/10.3389/fgene.2021.736156.

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118.

Yu, C. P., Ho, J. Y., Huang, Y. T., Cha, T. L., Sun, G. H., Yu, D. S., et al. (2013). Estrogen inhibits renal cell carcinoma cell progression through estrogen receptor-β activation. *PloS one*, *8*(2), e56667. https://doi.org/10.1371/journal.pone.0056667.

Yu, H., Yang, J., Jiao, S., Li, Y., Zhang, W., & Wang, J. (2014). T-box transcription factor 21 expression in breast cancer and its relationship with prognosis. International journal of clinical and experimental pathology, 7(10), 6906–6913.

Yu, X., Yu, G., & Wang, J. (2017). Clustering cancer gene expression data by projective clustering ensemble. *PloS one*, *12*(2), e0171429. https://doi.org/10.1371/journal.pone.0171429.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*, Article17. https://doi.org/10.2202/1544-6115.1128.

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, *33*(Web Server issue), W741–W748. https://doi.org/10.1093/nar/gki475.

Zhang, Y. H., Huang, T., Chen, L., Xu, Y., Hu, Y., Hu, L. D., Cai, Y., & Kong, X. (2017). Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. Oncotarget, 8(50), 87494–87511. https://doi.org/10.18632/oncotarget.20903.

Zhang, X., Yang, H., & Zhang, R. (2019). Challenges and future of precision medicine strategies for breast cancer based on a database on drug reactions. *Bioscience reports*, *39*(9), BSR20190230. https://doi.org/10.1042/BSR20190230.

Zhang, M., Chen, H., Wang, M., Bai, F., & Wu, K. (2020). Bioinformatics analysis of prognostic significance of COL10A1 in breast cancer. *Bioscience reports*, *40*(2), BSR20193286. https://doi.org/10.1042/BSR20193286.

Zhang, C., Liu, J., Wang, J., Zhang, T., Xu, D., Hu, W., & Feng, Z. (2021a). The Interplay Between Tumor Suppressor p53 and Hypoxia Signaling Pathways in Cancer. *Frontiers in cell and developmental biology*, *9*, 648808. https://doi.org/10.3389/fcell.2021.648808.

Zhang, J., Hou, S., You, Z., Li, G., Xu, S., Li, X., et al. (2021b). Expression and prognostic values of ARID family members in breast cancer. Aging, 13(4), 5621–5637. https://doi.org/10.18632/aging.202489.

Zhang, J. Z., & Wang, C. (2023). A comparative study of clustering methods on gene expression data for lung cancer prognosis. *BMC research notes*, *16*(1), 319. https://doi.org/10.1186/s13104-023-06604-8.

Zhang, W., Chen, P., Li, Z., Zhang, R., & Zhang, J. (2023). Clinical Implication of Keratin-15 Quantification for Renal Cell Carcinoma Management: Its Dysregulation and Association with Clinicopathologic Characteristics and Prognostication. *The Tohoku journal of experimental medicine*, *260*(2), 99–107. https://doi.org/10.1620/tjem.2023.J017.

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., & Hovarth, S. (2010). Weighted gene coexpression network analysis: state of the art. *Journal of biopharmaceutical statistics*, *20*(2), 281–300. https://doi.org/10.1080/10543400903572753.

Zhao, S., Shen, W., Yu, J., & Wang, L. (2018). TBX21 predicts prognosis of patients and drives cancer stem cell maintenance via the TBX21-IL-4 pathway in lung adenocarcinoma. Stem cell research & therapy, 9(1), 89. https://doi.org/10.1186/s13287-018-0820-6.

Zhao, L., Lee, V.H.F., Ng, M.K., Yan, H., Bijlsma, M.F. (2019). Molecular subtyping of cancer: current status and moving toward clinical applications, *Briefings in Bioinformatics*. 20(2), 572–584, https://doi.org/10.1093/bib/bby026.

Zhao, J., Zhao, B., Song, X., Lyu, C., Chen, W., Xiong, Y., & Wei, D. Q. (2023). Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data. *Briefings in bioinformatics*, *24*(2), bbad025. https://doi.org/10.1093/bib/bbad025.

Zheng, S., & Fu, Y. (2020). Age-related copy number variations and expression levels of F-box protein FBXL20 predict ovarian cancer prognosis. *Translational oncology*, *13*(12), 100863. https://doi.org/10.1016/j.tranon.2020.100863.

Zheng, R., Lai, G., Li, R., Hao, Y., Cai, L., & Jia, J. (2021). Increased expression of *MCM4* is associated with poor prognosis in patients with hepatocellular carcinoma. *Journal of gastrointestinal oncology*, *12*(1), 153–173. https://doi.org/10.21037/jgo-20-574.

Zhong, W., Li, Y., Yuan, Y., Zhong, H., Huang, C., Huang, J., et al. (2021). Characterization of Molecular Heterogeneity Associated With Tumor Microenvironment in Clear Cell Renal Cell Carcinoma to Aid Immunotherapy. *Frontiers in cell and developmental biology*, *9*, 736540. https://doi.org/10.3389/fcell.2021.736540.

Zhong, P-Q., Yan, X-X, Wang, W-J., Hong, M., Chen, P., & Liu, M. (2022). Identification and Validation of LYZ and CCL19 as Prognostic Genes in the Cervical Cancer Micro-Environment. *Clin. Exp. Obstet. Gynecol,* 49(6), 144. https://doi.org/10.31083/j.ceog4906144.

Zhou, Q., Zhou, J., & Fan, J. (2021). Expression and Prognostic Value of ARID5A and its Correlation With Tumor-Infiltrating Immune Cells in Glioma. Frontiers in oncology, 11, 638803. https://doi.org/10.3389/fonc.2021.638803.

Zhu, Q., Yang, H., Cheng, P., & Han, Q. (2019). Bioinformatic analysis of the prognostic value of the lncRNAs encoding snoRNAs in hepatocellular carcinoma. *BioFactors (Oxford, England)*, *45*(2), 244–252. https://doi.org/10.1002/biof.1478.

Znaor, A., Lortet-Tieulent, J., Laversanne, M., Jemal, A., & Bray, F. (2015). International variations and trends in renal cell carcinoma incidence and mortality. *European urology*, *67*(3), 519–530. https://doi.org/10.1016/j.eururo.2014.10.002.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature genetics*, *51*(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5.

Appendices

Appendix A

Tables

Table A1: Datasets used in uncorrected and tissue-corrected analysis. The RNA-Seq gene expression and curated clinical public datasets that have been used: Large B-cell Lymphoma (DLBC), Liver Cancer (LIHC), Lung Adenocarcinoma (LUAD), Cervical Cancer (CESC), and Testicular Cancer (TGCT).

Cancer Name	Dataset ID	Dataset	Phenotypes
Large B-cell Lymphoma (DLBC)	TCGA-DLBC	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-DLBC.htseq_counts.tsv.gz.	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-DLBC.GDC_phenotype.tsv.gz.
Lung Adenocarcinoma (LUAD)	TCGA-LUAD	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-LUAD.htseq_counts.tsv.gz.	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-LUAD.GDC_phenotype.tsv.gz.
Liver Cancer (LIHC)	TCGA-LIHC	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-LIHC.htseq_counts.tsv.gz.	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-LIHC.GDC_phenotype.tsv.gz.
Cervical Cancer (CESC)	TCGA-CESC	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-CESC.htseq_counts.tsv.gz.	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-CESC.GDC_phenotype.tsv.gz.
Testicular Cancer (TGCT)	TCGA-TGCT	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-TGCT.htseq_counts.tsv.gz.	https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-TGCT.GDC_phenotype.tsv.gz.

Table A2: Normal tissue expression dataset was obtained from the GTEx Portal. Dataset from the primary sites were extracted to match the individual cancer cohorts (last column).

Primary sites	Dataset ID	Dataset	Phenotypes	Matched Cancer
Whole Blood	GTEX	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/gtex_gene_expected_count.gz	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/GTEX_phenotype.gz	Large B-cell Lymphoma
Lung	GTEX	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/gtex_gene_expected_count.gz	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/GTEX_phenotype.gz	Lung Adenocarcinoma
Liver	GTEX	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/gtex_gene_expected_count.gz	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/GTEX_phenotype.gz	Liver Cancer
Cervix	GTEX	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/gtex_gene_expected_count.gz	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/GTEX_phenotype.gz	Cervical Cancer
Testis	GTEX	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/gtex_gene_expected_count.gz	https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/GTEX_phenotype.gz	Testicular Cancer

Table A3: Top 5 TFs derived from the ChEA3 enrichment analysis of each tissue-corrected WGCNA module. The biological role indicates the role of the identified TF in cancer according to literature.

Module	Biological Role	TF	Overlapping genes	FDR
Black	*FOX* proteins are significantly implicated in cancer (Bach *et al*., 2018).	*FOXB1*	*RPL5,RPL30,RPL32,RPL31,RPL34,RPLP0,RPL9,RPL7,RPS14,RPLP2,RPS10,RPL39,RPS13,RPL21,RPL23,RPS3A,RPL37A,RPL36A,RPS15A,RPS3,RPL15,RPL23A,RPS25,RPS27,RPS29,RPS20,RPS24,RPS23*	1.61E-28
Black	Prognostic marker, high expression is unfavorable in liver cancer (The human protein atlas, 2023c).	*CHCHD3*	*RPL5,RPL30,RPL32,RPL31,RPL34,RPLP0,RPL10A,RPL9,RPL7,RPS14,RPLP2,RPS10,RPL39,RPS13,RPL21,RPL23,RPS3A,RPL37A,RPL36A,RPS3,RPL15,RPL23A,RPS25,RPS27,RPS29,RPS24,RPS23*	2.14E-27
Black	Prognostic marker, high expression is unfavorable in liver cancer (The human protein atlas, 2023d).	*ZNF581*	*RPL5,RPL3,RPL32,RPLP0,RPL8,RPL10A,EEF1B2,RPL7A,RPS14,RPS18,RPLP2,RPS10,RPS13,RPS8,RPS6,RPSA,RPL27,RPL29,RPL12,RPS3,RPL14,RPL15,RPS23*	5.41E-22
Black	Over expressed in various cancers, including hepatocellular carcinoma (de Haas *et al*., 2006; Terrinoni *et al*., 2011)	*OTX1*	*RPL30,RPL31,RPL34,RPL9,RPL7,RPS14,RACK1,RPS10,RPL39,RPL21,RPL37A,RPL36A,RPS15A,RPL13,RPL15,RPS25,RPS27,RPS29,RPS24*	9.13E-17
Black	Overexpression has been associated with the development of pancreatic (Jensen *et al*., 2000; Katoh & Katoh, 2007), breast (Farnie *et al*., 2007) and ovarian (Wang *et al*., 2010) cancers.	*HES1*	*RPL30,RPL31,RPL34,RPL9,RPL7,RPS14,RPL39,RPL21,RPL23,RPL37A,RPL36A,RPL15,RPL23A,RPS27,RPS29,RPS20,RPS24,RPS23*	1.2E-15
Brown	Down-regulated expression in hepatocellular carcinoma and gastric cancer (Shin *et al*., 2010; Yamashita *et al*., 2010).	*IRF5*	*CD86,CD84,SPI1,CD80,LST1,ICAM3,CMKLR1,RNASE6,CYBB,MPEG1,OSCAR,TYROBP,BTK,CSF1R,IGSF6,FPR3,CORO1A,PIK3R5,SLAMF7,NCKAP1L,CD14,SLAMF1,CCR1,CD163,FAM78A,LY86,PILRA,ARHGAP30,FERMT3,SIGLEC9,ITGAM,PLEK,ITGB2,SIRPB2,SPN,HK3,FCGR3A,CD37,CCR5,CD53,FCER1G,NFAM1,FGR,HCK,MS4A6A,TLR8,LCP2,LCP1,PLEKHO2,DOCK2,SASH3,LILRA6,C1QA,WAS,LILRA2,*	1.44E-67

			AIF1,FGD2,CYTH4,LAIR1,LRRC25,IL10RA,LAPTM5,LILRB1,LILRB2,LILRB3,LILRB4,CD4,SIGLEC1,MYO1F,C1QC	
Brown	Abnormal *BATF* expression in tumors predicted survival times of patients (Jia *et al*., 2022). *BATF* expression could also predict immunotherapeutic and chemotherapy responses in cancer (Jia *et al*., 2022).	*BATF*	*TRAF3IP3,CD80,SLA,IKZF1,SIT1,GPR171,TBC1D10C,CD96,TYROBP,ACAP1,CD8A,SP140,RASAL3,CORO1A,SAMSN1,PIK3R5,SLAMF7,LPXN,ICOS,SLAMF1,SH2D1A,NKG7,PILRA,PTPRC,ARHGAP30,CD27,FERMT3,ITK,PLEK,CD3G,CD3E,CD3D,GNGT2,CD37,CYTIP,CCR5,MAP4K1,CD53,IL16,APBB1IP,ZAP70,LCP2,LCP1,SASH3,CST7,CXCR3,CCL5,IL21R,TIGIT,S1PR4,P2RY10,TRAT1,IL10RA,LAPTM5,LILRB1,LILRB2,LILRB4,CD2,CD4,CD6,ABI3,CD5,IL2RB,CD7,SIGLEC1,PTPN7,CD247*	1.03E-63
Brown	An increased incidence of *TBX21* has been linked to cancer development (Yu *et al*., 2014; Lin *et al*., 2015). *TBX21* has been associated with poor prognosis in patients with lung adenocarcinoma (Zhao *et al*., 2018).	*TBX21*	*TRAF3IP3,SPI1,ICAM3,IKZF1,IL18RAP,SIT1,TBC1D10C,CD300A,DOK2,ACAP1,CD8A,RASAL3,CORO1A,SLAMF6,FAM78A,PYHIN1,NKG7,ARHGAP25,PTPRC,ARHGAP30,ITK,ITGAM,ITGB2,SIRPB2,ITGAL,CD3E,FCGR3A,TAGAP,CD37,TNFAIP8L2,CYTIP,MAP4K1,CD53,NFAM1,IL16,FGR,APBB1IP,HCK,ZAP70,TLR8,LCP2,LCP1,DOCK2,SASH3,LILRA6,WAS,AOAH,CST7,LILRA2,CYTH4,CCL5,S1PR4,LRRC25,P2RY13,IL10RA,LAPTM5,LILRB3,CD2,CD6,ABI3,IL2RB,CD7,MNDA,CD247,EVI2B,MYO1F*	1.27E-62
Brown	Recent studies have reported significant functions of *Arid5a* in numerous types of cancer, including lung cancers (Sarode *et al*., 2020; Zhou *et al*., 2021; Parajuli *et al*., 2021; Zhang *et al*., 2021b)	*ARID5A*	*CD86,SPI1,ICAM3,IKZF1,C3AR1,TBC1D10C,OSCAR,TYROBP,ACAP1,CSF1R,RASAL3,CORO1A,SAMSN1,PIK3R5,SLAMF7,LPXN,CD14,SLAMF1,CCR1,PILRA,ARHGAP25,ARHGAP30,FERMT3,SIGLEC9,PLEK,ITGB2,ITGAL,CD3E,TAGAP,CD37,CYTIP,MAP4K1,CD53,FCER1G,IL16,FGR,APBB1IP,HCK,ZAP70,LCP2,LCP1,PLEKHO2,DOCK2,SASH3,C1QA,WAS,CYTH4,IL21R,FCGR1A,S1PR4,IL10RA,LAPTM5,LILRB1,*	2.81E-61

			LILRB2,LILRB3,LILRB4,CD4,CD6,CD5,IL2RB,CD7, CD247,EVI2B,MYO1F,C1QC	
Brown	Significantly associated with cervical cancer prognosis (Zhong *et al*., 2022).	*SCML4*	*TRAF3IP3,GPR65,ICAM3,SLA,IKZF1,SLA2,SIT1,TBC1D10C,CD96,ACAP1,CD8A,ZNF831,RASAL3,CORO1A,LY9,TESPA1,LPXN,SLAMF6,ICOS,CD300LF,FAM78A,PYHIN1,SH2D1A,NKG7,ARHGAP25,PTPRC, ARHGAP30,CD27,ITK,CD3G,ITGAL,CD3E,CD3D,TNFSF13B,TAGAP,CD37,CYTIP,MAP4K1,CD53,IL16, APBB1IP,ZAP70,LCP2,LCP1,SASH3,WAS,CYTH4,CXCR3,CCL5,IL12RB1,S1PR4,P2RY10,TRAT1,IL10RA, SNX20,CD2,CD6,CD5,IL2RB,CD7,PTPN7,CD247,EVI2B,MYO1F*	4.09E-60
Magenta	Affect TGF-β signaling to promote prostate cancer (Kwon *et al*., 2021).	*ZNF507*	*HNRNPU,XPO1,KPNB1,DHX9,IREB2,SRSF1,LARP1, TRA2B,NSD1,TRPM7,HNRNPA3,MGA,TJP1,HNRNPL*	5.11E-9
Magenta	Prognostic marker, high expression is unfavorable in liver cancer (The human protein atlas, 2023e).	*ZNF207*	*TCERG1,ICE2,HNRNPU,XPO1,SRSF2,SRSF3,DHX9, IREB2,SRSF1,TRA2B,NSD1,TRPM7,NONO,MGA*	5.11E-9
Magenta	*SAFB* protein levels predict poor prognosis of breast cancer patients (Hammerich-Hille *et al*., 2010)	*SAFB*	*HNRNPU,RBM14,NCL,HNRNPH3,KPNB1,SF1,DHX9,SRSF1,GANAB,LARP1,HNRNPA3,NONO,HNRNPL, HNRNPD*	5.11E-9
Magenta	These proteins have critical roles in development, differentiation, and tumorigenesis (Lee & Maeda, 2012).	*ZBTB39*	*HNRNPU,XPO1,GEMIN5,KPNB1,SF1,DHX9,SRSF1, LARP1,TRA2B,NSD1,HNRNPA3,NONO,HNRNPL,HNRNPD*	5.11E-9
Magenta	*CTCF* has been identified as a putative driver gene in several cancer types (Marshall *et al*., 2017).	*CTCF*	*DDX46,HNRNPU,XPO1,NCL,KPNB1,SF1,DHX9,SRSF1,LARP1,NSD1,HNRNPA3,NONO,HNRNPL,HNRNPD*	5.11E-9
Turquoise	*E2F8* is correlated with the progression of cervical cancer (Kim *et al*., 2020)	*E2F8*	*DSCC1,CCNF,HJURP,BUB1B,MKI67,CDC20,CHEK1,NUSAP1,OIP5,GTSE1,ESCO2,CDC25C,HASPIN,WDR76,CDC25A,SGO1,DEPDC1B,MELK,TIMELESS, KIF20A,CDCA2,PARPBP,CDCA3,TROAP,CDCA5,NCAPG,CDCA8,HMMR,PKMYT1,SKA3,IQGAP3,NCA*	1.35E-143

			PH,RAD51AP1,CCNB2,CCNB1,ORC1,RACGAP1,CLSPN,FAM83D,FANCI,PLK4,STIL,PLK1,CDC6,NDC80,ZWINT,ANLN,TPX2,KIF18A,KIF18B,UBE2T,KIF4A,CDK1,TOP2A,ARHGAP11A,FEN1,NCAPG2,KIF14,MCM10,BRCA1,KIF11,FOXM1,LMNB1,KIF15,EXO1,NUF2,PBK,MYBL2,SPDL1,DLGAP5,CEP55,RFC4,CKAP2L,KIF23,CIP2A,CCNA2,ASPM,ESPL1,INCENP,KIFC1,DEPDC1,BIRC5,MCM4,KIF2C,MCM6,MTFR2,DTL,FAM72B,FAM72A,UHRF1,PRIM1,TTK,TYMS,AURKB,AURKA,CDC45,E2F2,RAD54L,BUB1,E2F7,GINS1,POLQ,CENPU,RRM2,SPAG5,SHCBP1,TICRR,CENPE,CENPF,CENPI,PRC1,TRIP13,CDKN3,MAD2L1	
Turquoise	Prognostic marker, high expression is unfavorable in liver cancer and lung (The human protein atlas, 2023f; 2023g)	*CENPA*	*DSCC1,CCNF,HJURP,BUB1B,CDC20,CHEK1,NUSAP1,OIP5,NEK2,KPNA2,GTSE1,CDC25C,HASPIN,KNSTRN,CDC25A,SGO1,DEPDC1B,MELK,TIMELESS,KIF20A,PRR11,PIF1,CDCA2,PARPBP,CDCA3,TROAP,CDCA5,NCAPG,CDCA8,HMMR,PKMYT1,SKA3,IQGAP3,NCAPH,RAD51AP1,CCNB2,CCNB1,RACGAP1,FAM83D,FANCI,PLK4,STIL,UBE2C,PLK1,CDC6,NDC80,ZWINT,ANLN,TPX2,KIF18A,KIF18B,UBE2T,KIF4A,CDK1,TOP2A,ARHGAP11A,FEN1,NCAPG2,KIF14,MCM10,KIF11,FOXM1,LMNB1,KIF15,EXO1,NUF2,PBK,MYBL2,SPDL1,DLGAP5,CEP55,CKAP2L,KIF23,CIP2A,CCNA2,ASPM,ESPL1,INCENP,KIFC1,DEPDC1,BIRC5,MCM4,KIF2C,MTFR2,DTL,FAM72B,FAM72A,UHRF1,TTK,TYMS,AURKB,AURKA,CDC45,RAD54L,BUB1,GINS1,CENPU,RRM2,SPAG5,SHCBP1,TICRR,CENPE,CENPF,RAD51,PRC1,TRIP13,CDKN3,MAD2L1*	7.77E-134
Turquoise	*E2F7* promotes cell proliferation and metastasis in lung adenocarcinoma, liver	*E2F7*	*DSCC1,CCNF,HJURP,BUB1B,MKI67,CDC20,NUSAP1,GTSE1,HASPIN,CDC25A,MELK,TIMELESS,KIF2*	2.21E-103

	cancer and head and neck cancer (Liang *et al*., 2018; Ma *et al*., 2018; Saleh *et al*., 2019).		*0A,CDCA2,CDCA5,NCAPG,CDCA8,HMMR,PKMYT1,SKA3,IQGAP3,NCAPH,RAD51AP1,CCNB1,ORC1,RACGAP1,CLSPN,FANCI,PLK4,STIL,PLK1,CDC6,NDC80,ZWINT,ANLN,TPX2,KIF18B,UBE2T,KIF4A,CDK1,TOP2A,ARHGAP11A,FEN1,NCAPG2,KIF14,MCM10,KIF11,FOXM1,LMNB1,KIF15,EXO1,NUF2,PBK,MYBL2,SPDL1,DLGAP5,CEP55,CKAP2L,KIF23,CIP2A,CCNA2,ASPM,ESPL1,INCENP,KIFC1,DEPDC1,BIRC5,MCM4,KIF2C,DTL,UHRF1,TTK,TYMS,AURKB,AURKA,RAD54L,BUB1,GINS1,POLQ,CENPU,RRM2,SPAG5,SHCBP1,TICRR,CENPE,CENPF,PRC1,TRIP13,MAD2L1*	
Turquoise	*E2F2* plays a significant role in tumor progression (Shen & Wang, 2021).	*E2F2*	*CCNF,HJURP,MKI67,CHEK1,NUSAP1,GTSE1,ESCO2,HASPIN,WDR76,CDC25A,TIMELESS,TROAP,CDCA5,NCAPG,CDCA8,PKMYT1,NCAPH,SKA1,ORC1,CLSPN,FANCI,PLK4,STIL,CDC6,NDC80,ZWINT,KIF18B,ARHGAP11A,FEN1,NCAPG2,KIF14,MCM10,BRCA1,KIF11,FOXM1,LMNB1,KIF15,CHAF1B,EXO1,MYBL2,CKAP2L,CCNA2,ASPM,ESPL1,INCENP,KIFC1,MCM4,KIF2C,DTL,UHRF1,TYMS,CDC45,RAD54L,GINS1,POLQ,CENPU,RRM2,SPAG5,SHCBP1,TICRR,PRC1,CENPK,SPC24*	2.49E-64
Turquoise	*FOXN4* can be used as candidate prognostic biomarkers for lung adenocarcinoma (Yao *et al*., 2021).	*FOXN4*	*CCNF,HJURP,BUB1B,CDC20,NUSAP1,NEK2,KPNA2,GTSE1,CDC25C,KNSTRN,CDC25A,SGO1,DEPDC1B,MELK,KIF20A,PIF1,CDCA2,CDCA3,TROAP,NCAPG,CDCA8,IQGAP3,NCAPH,CCNB2,CCNB1,RACGAP1,PLK4,UBE2C,PLK1,NDC80,TPX2,KIF18A,UBE2T,KIF4A,CDK1,TOP2A,KIF14,MCM10,BRCA1,KIF11,KIF15,NUF2,PBK,DLGAP5,CKAP2L,KIF23,CIP2A,CCNA2,ASPM,ESPL1,KIFC1,BIRC5,KIF2C,FAM72B,TTK,AURKB,BUB1,E2F7,SPAG5,CENPE,CENPF,PRC1,CDKN3*	2.49E-64

Table A4: Transcription factors enrichment analysis of tissue-corrected WGCNA brown module. A list of TFs and their corresponding rank according to ARCHS4 co-expression, with documented information about their biological function associated with survival in the context of cervical cancer. The genes in bold were previously found (Kannan *et al*., 2021) to play a role in cervical cancer survival.

Survival associated with TF	Rank	TF	Overlapping genes	FDR
Significantly associated with cervical cancer prognosis (Zhong *et al*., 2022).	5	*SCML4*	***MAP4K1***,*TRAF3IP3,GPR65,ICAM3,SLA,IKZF1,SLA2,SIT1,****TBC1D10C***,*CD96,****ACAP1***,*CD8A,****ZNF831***,***RASAL3***,*CORO1A,LY9,TESPA1,LPXN,SLAMF6,ICOS,CD300LF,FAM78A,PYHIN1,****SH2D1A***,*NKG7,ARHGAP25,PTPRC,ARHGAP30,CD27,ITK,CD3G,ITGAL,CD3E,CD3D,TNFSF13B,TAGAP,CD37,CYTIP,CD53,IL16,APBB1IP,ZAP70,LCP2,LCP1,SASH3,WAS,CYTH4,CXCR3,CCL5,IL12RB1,S1PR4,****P2RY10***,*TRAT1,IL10RA,SNX20,CD2,CD6,CD5,IL2RB,CD7,PTPN7,CD247,EVI2B,MYO1F*	4.09E-60
Prognostic marker, high expression is favorable in cervical cancer (The human protein atlas, 2023h).	9	*SNAI3*	***MAP4K1***,*TRAF3IP3,SPI1,LST1,ICAM3,****TBC1D10C***,*CD300A,OSCAR,TYROBP,****ACAP1***,*IGSF6,****RASAL3***,*CORO1A,PIK3R5,CCR1,FAM78A,PILRA,ARHGAP25,ARHGAP30,FERMT3,SIGLEC9,ITGAM,ITGB2,SIRPB2,ITGAL,CD3E,HK3,FCGR3A,CD37,TNFAIP8L2,CD53,NFAM1,IL16,FGR,APBB1IP,HCK,TLR8,LCP2,LCP1,PLEKHO2,SASH3,LILRA6,WAS,LILRA1,AOAH,LILRA2,CYTH4,S1PR4,LRRC25,P2RY13,IL10RA,LAPTM5,LILRB2,LILRB3,CD4,ABI3,CD7,MNDA,CD247,EVI2B,MYO1F*	3.1E-56
Prognostic marker, high expression is favorable in cervical cancer (The human protein atlas, 2023i).	14	*IKZF1*	***MAP4K1***,*TRAF3IP3,ICAM3,GPR174,****TBC1D10C***,*MPEG1,****ACAP1***,***RASAL3***,*CORO1A,PIK3R5,NCKAP1L,FAM78A,ARHGAP25,PTPRC,ARHGAP30,FERMT3,ITK,ITGB2,SIRPB2,ITGAL,CD3E,PIK3CG,SPN,TAGAP,CD37,CYTIP,CCR2,CD53,NFAM1,IL16,FGR,APBB1IP,HCK,ZAP70,TLR8,LCP2,LCP1,DOCK2,SASH3,WAS,LILRA1,AOAH,CYTH4,S1PR4,LRRC25,P2RY13,IL10RA,LAPTM5,CD4,CD6,CD5,IL2RB,CD7,MNDA,CD247,EVI2B,MYO1F*	3.13E-51
Prognostic marker, high expression is favorable in cervical cancer (The human protein atlas, 2023j).	24	*IKZF3*	***MAP4K1***,*TRAF3IP3,ICAM3,IKZF1,GPR174,SIT1,****GPR171***,***TBC1D10C***,*PRKCB,****ACAP1***,*CD8A,SP140,****RASAL3***,*CORO1A,LY9,NCKAP1L,SLAMF6,FAM78A,PYHIN1,NKG7,ARHGAP25,PTPRC,ARHGAP30,ITK,CD3G,ITGAL,CD3E,CD3D,SPN,TAGAP,CD37,CYTIP,CD53,FCRL3,IL16,APBB1IP,ZAP70,LCP2,LCP1,DOCK2,SASH3,IL21R,IL12RB1,TIGIT,*	1.41E-47

			P2RY10,*IL10RA,LAPTM5,SCIMP,CD6,IL2RB,CD7,**PDCD1**,CD247,EVI2B*	
Prognostic marker, high expression is favorable in cervical cancer (The human protein atlas, 2023k).	34	*FOXP3*	***MAP4K1***,*TRAF3IP3,ICAM3,SLA,IKZF1,GPR174,SIT1,**TBC1D10C**,UBASH3A,**ACAP1**,SP140,**RASAL3**,CORO1A,LPXN,ICOS,FAM78A,ARHGAP25,PTPRC,ARHGAP30,CD27,ITK,ITGAL,CD3E,CD3D,TAGAP,CD37,CYTIP,CD53,IL16,APBB1IP,ZAP70,LCP2,LCP1,DOCK2,SASH3,WAS,CYTH4,CXCR3,TIGIT,S1PR4,**P2RY10**,TRAT1,IL10RA,CD2,CD4,CD6,CD5,IL2RB,CD7,CD247*	1.19E-42
May serve as a tumor suppressor gene in cervical cancer (Li *et al*., 2018b)	49	*RUNX3*	***MAP4K1***,*TRAF3IP3,ICAM3,IKZF1,IL18RAP,**TBC1D10C**,**ACAP1**,**RASAL3**,CORO1A,PIK3R5,NCKAP1L,SLAMF1,FAM78A,ARHGAP25,PTPRC,ARHGAP30,ITGB2,ITGAL,CD3E,SPN,CD37,CYTIP,CD53,IL16,FGR,APBB1IP,ZAP70,IFNG,LCP2,LCP1,DOCK2,SASH3,WAS,CYTH4,TBX21,IL21R,S1PR4,**P2RY10**,IL10RA,LAPTM5,CD6,IL2RB,CD7,CD247,EVI2B,MYO1F*	8.23E-38
High *ETS1* levels exhibit a poorer prognosis than those with low *ETS1* levels in cervical cancer (Xu *et al*., 2003, Fujimoto *et al*., 2002).	59	*ETS1*	***MAP4K1***,*TRAF3IP3,ICAM3,IKZF1,GPR174,**TBC1D10C**,**ACAP1**,**ZNF831**,**RASAL3**,CORO1A,NCKAP1L,SLAMF6,FAM78A,ARHGAP25,PTPRC,ARHGAP30,ITK,ITGB2,ITGAL,CD3E,TAGAP,CD37,CYTIP,CD53,IL16,APBB1IP,ZAP70,LCP2,LCP1,DOCK2,SASH3,WAS,CYTH4,S1PR4,**P2RY10**,IL10RA,LAPTM5,CD6,CD5,IL2RB,CD7,CD247,EVI2B*	2.44E-34
Low expression is associated with poor prognosis in cervical cancer (Deng *et al*., 2021).	136	*IRF4*	***MAP4K1***,*CD80,IKZF1,**GPR171**,SP140,**RASAL3**,LPXN,ICOS,SLAMF1,ARHGAP30,ITK,SPN,CYTIP,APBB1IP,IFNG,LCP1,DOCK2,SASH3,IL21R,**P2RY10**,LILRB1,SCIMP,CD6,IL2RB,CD7*	4.34E-15
Prognostic marker, high expression is favorable in cervical cancer The human protein atlas, 2023l).	157	*ZNF266*	*MAP4K1,IKZF1,**TBC1D10C**,**ACAP1**,RASAL3,LY9,ICOS,PYHIN1,PTPRC,CD3G,CD37,CYTIP,FCRL3,IL16,APBB1IP,ZAP70,IL12RB1,CD6,EVI2B*	9.94E-10

Figures

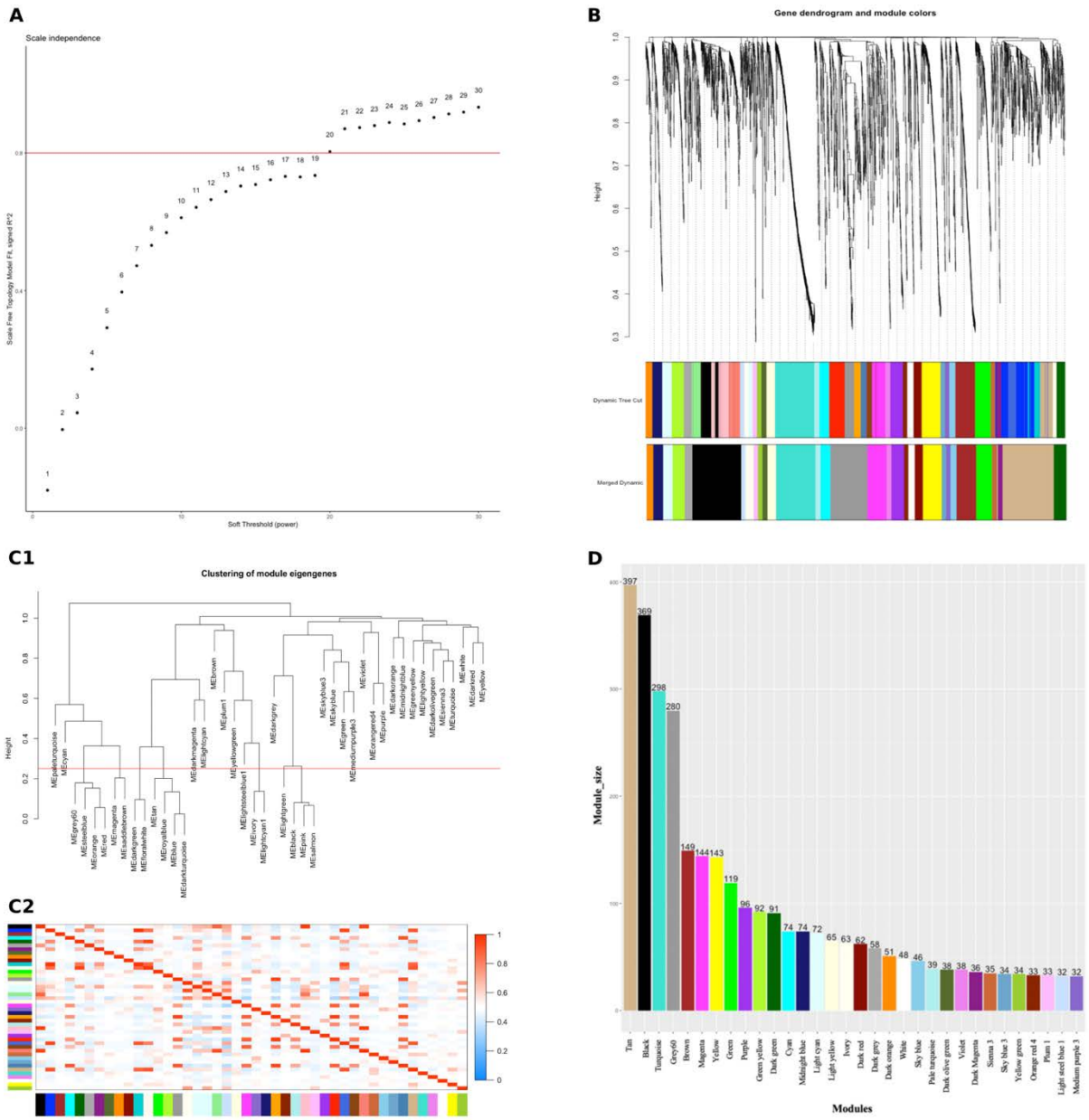

Figure A1: Uncorrected RNA-Seq data were inserted into WGCNA to identify gene modules. (A) Soft threshold power. (B) Gene clustering tree. Each colour underneath the dendrogram shows the module assignment, and branches above represent the genes. The dynamic tree cut shows the initial module detection and merged dynamic indicates the modules divided according to their similarity. (C1) Module eigengene dendrogram identified groups of correlated modules. The red line indicates the module eigengene threshold of 0.25 and (C2) Eigengene adjacency heatmap of different gene co-expression modules. In the heatmap, the blue colour represents low adjacency, while the red represents high adjacency. (D) Barplot of 32 co-expression modules constructed after similar modules were merged with module size at the top of each bar.

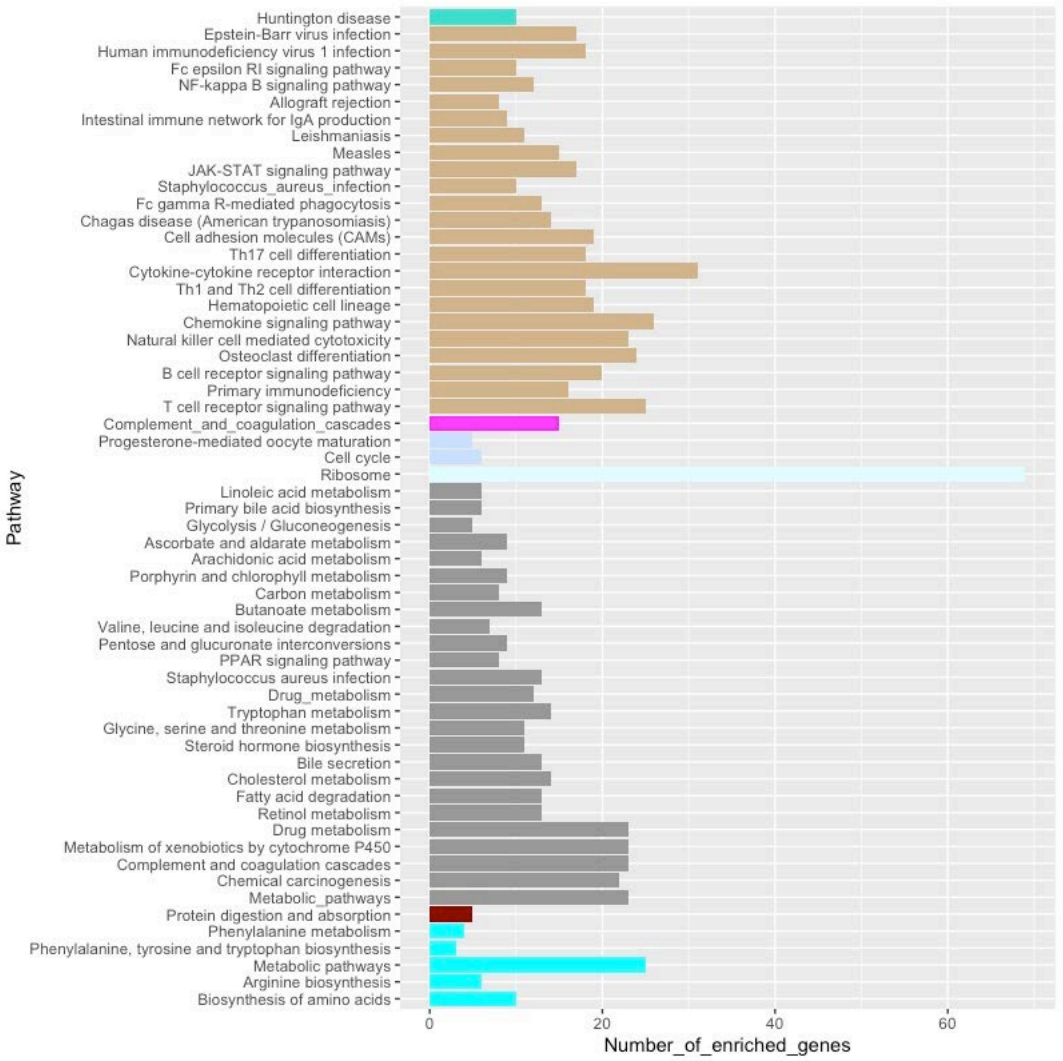

Figure A2: KEGG enrichment of gene modules detected by WGCNA from the uncorrected RNA dataset using the ORA, WebGestalt.

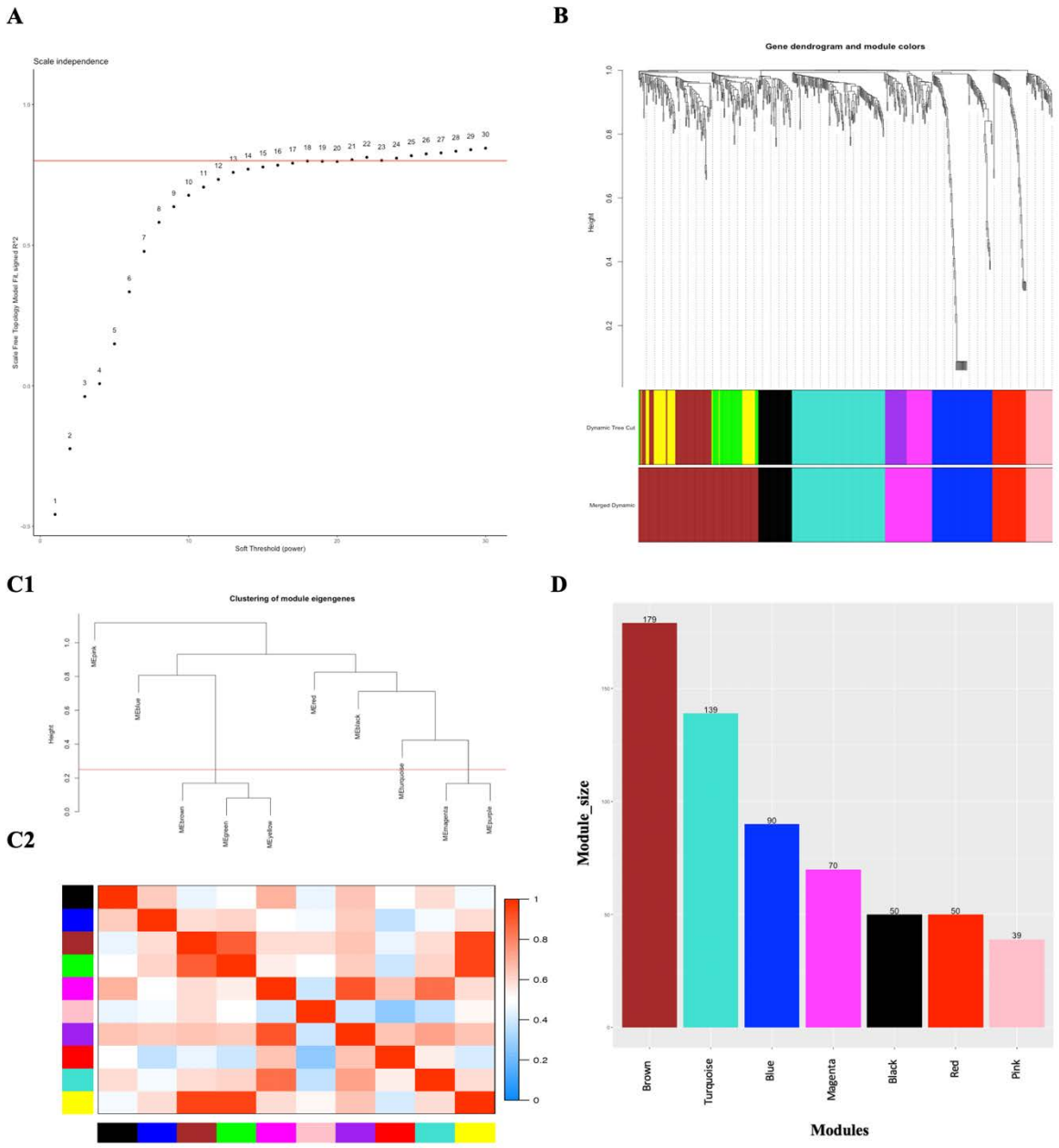

Figure A3: Tissue-corrected dataset were inserted into WGCNA to identify gene modules. (A) Soft threshold power. (B) Gene clustering tree. Each colour underneath the dendrogram shows the module assignment, and branches above represent the genes. The dynamic tree cut shows the initial module detection and merged dynamic indicates the modules divided according to their similarity. (C1) Module eigengene dendrogram identified groups of correlated modules. The red line indicates the module eigengene threshold of 0.25 and (C2) Eigengene adjacency heatmap of different gene co-expression modules. In the heatmap, the blue colour represents low adjacency, while the red represents high adjacency. (D) Barplot of seven co-expression modules constructed after merged modules with module size at the top of each bar.

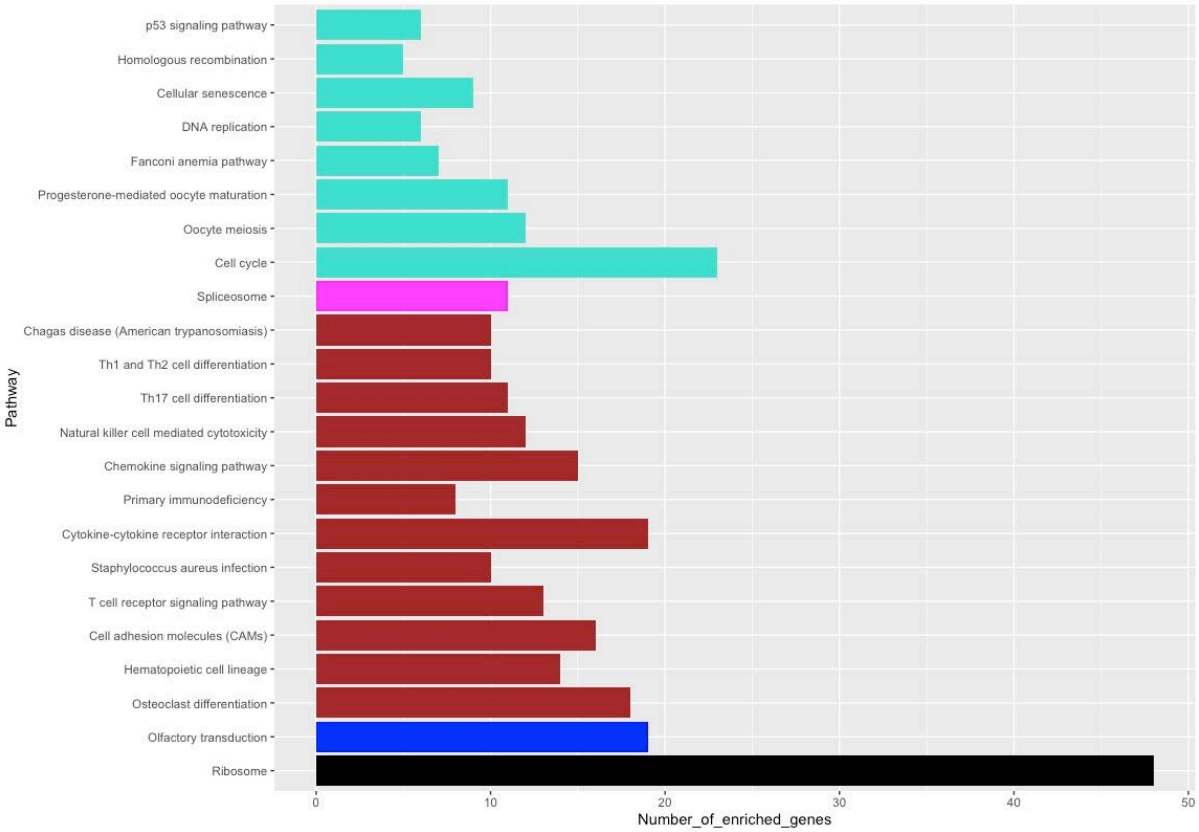

Figure A4: KEGG enrichment of gene modules detected by WGCNA from the tissue-corrected RNA dataset using the ORA, WebGestalt.

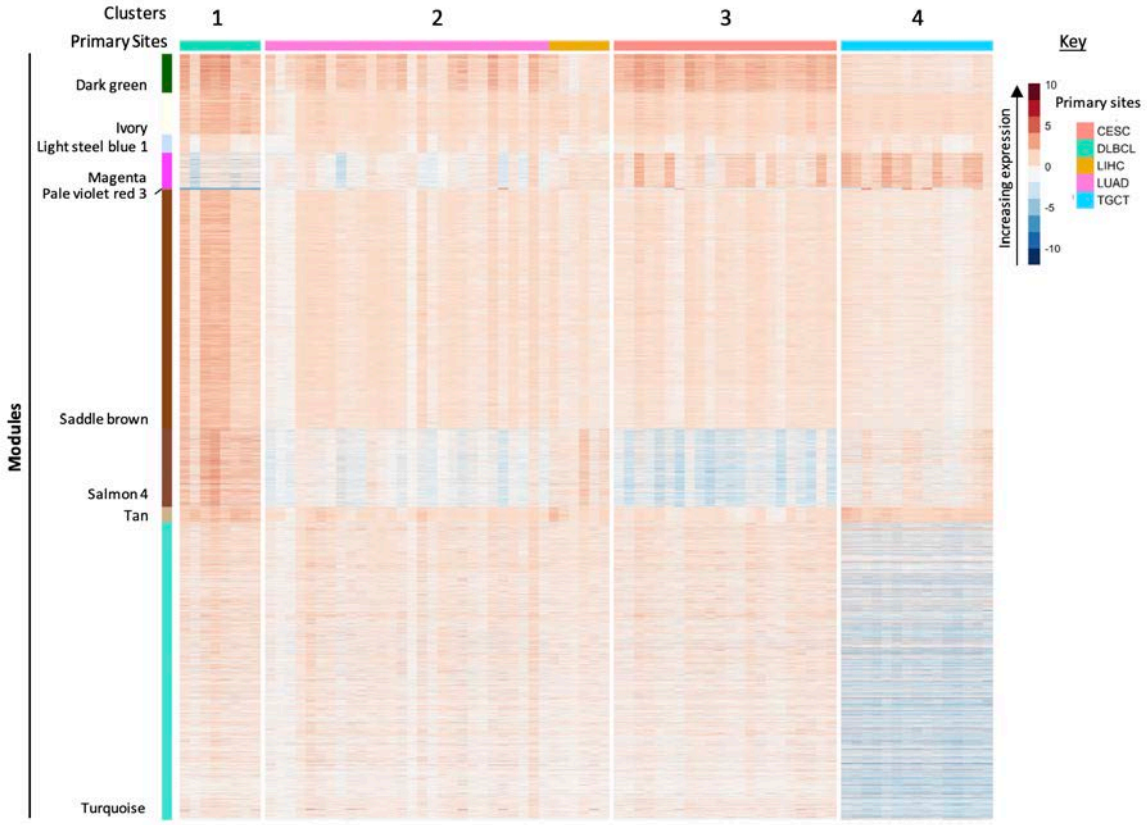

Figure A5. Heatmap of tissue-corrected RNA-Seq data of late-stage cancer samples normalized with normal tissue samples, illustrating module expression within cancer clusters. Normal tissue expression dataset was obtained from the GTEx Portal. To match the number of male/female ratios as in the late-stage cancer samples, the same number normal tissue samples of male/female ratios were randomly selected, except for cervical cancer, which only had 10 normal tissue samples. The colour bar on the left shows modules identified by WGCNA and enriched for functional pathway annotations. The rows are further composed of protein-coding genes with expression values obtained after data normalization. Clusters of similar cancer cohorts are indicated across the top and the cancer cohort are displayed by the colour bar along the top with the key on the right. *Primary sites abbreviations: CESC = Cervical squamous cell carcinoma; DLBCL = Diffuse Large B-cell Lymphoma; LIHC = Liver Hepatocellular Carcinoma; LUAD = Lung Adenocarcinoma; TGCT = Testicular Germ Cell Tumors.

Appendix B

Tables

Table B1: List of 48 gene subset selected by RFE.

Gene name
MCUB
CD82
IPO11
CPXM1
KDELR3
SALL4
SEC23B
PGK1
PDGFRL
SERPINE1
GNB3
LPCAT3
CLINT1
IGFBP2
C1orf21
TNFSF18
NLN
MMP19
CANX
HSD17B4
NACAD
ALPK3
ASXL3
FAHD2B
OSBPL11
PAQR6
SPATA18
GPER1
DIRAS2
COL22A1
KRT15
CLCN5
NET1
GOLGA6L2
NECTIN3
CPNE7

MCFD2
ZBTB7C
WT1
TAL2
CDHR4
FOCAD
SLC6A17
ATL1
PCDHA4
PLXNA4
PCDHGC3
CCER2

Table B2: GEO datasets used to verify the results obtained. An independent test dataset was created from three KIRC-specific GEO datasets; GSE73731, GSE53757, and GSE36895, which includes a total of 70 early-stage and 65 late-stage raw CEL files, which were robust multi-array average (RMA) normalized.

GEO datasets	Early-stage	Late-stage
GSE73731	41	44
GSE53757	24	15
GSE36895	5	6
	70	65

Figures

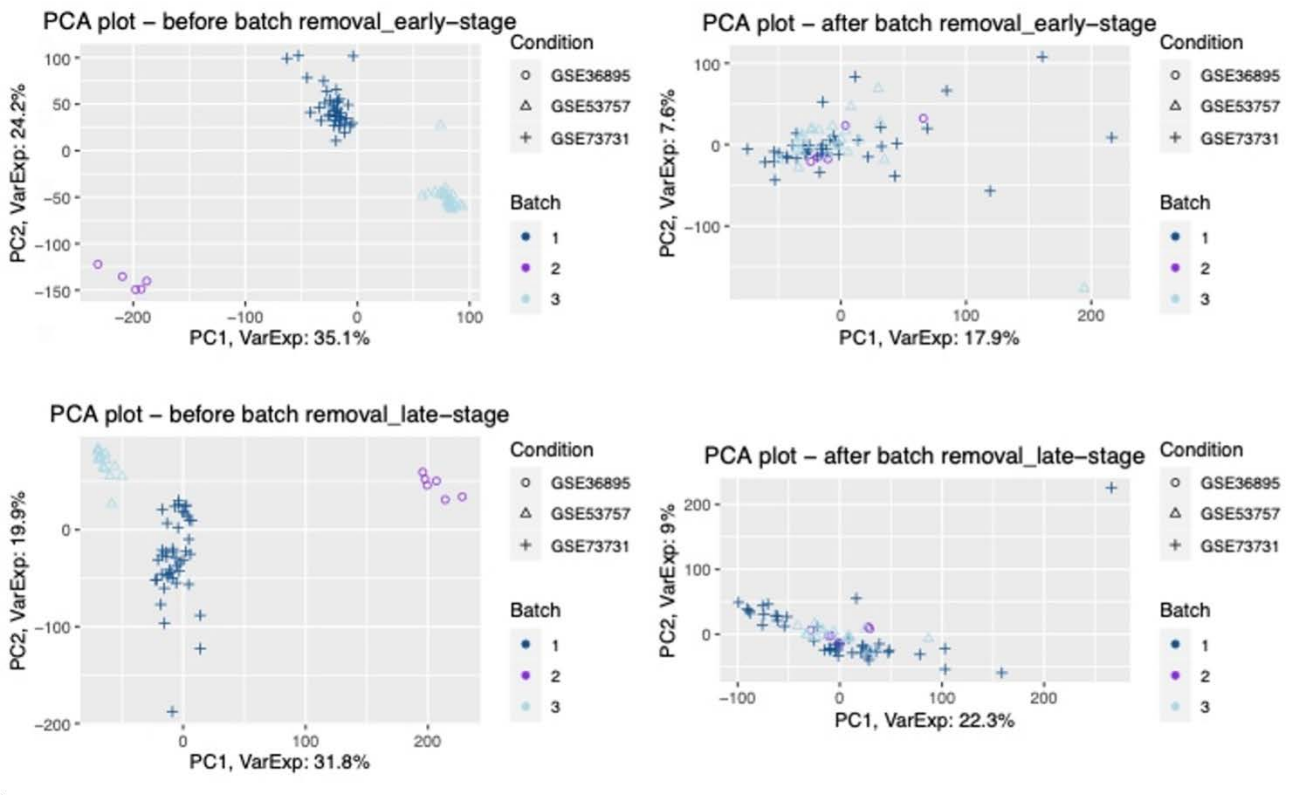

Figure B1: PCA plots before and after batch effect removal. The three GEO datasets were subjected to batch effect removal using ComBat. The GEO expression dataset after batch effect removal were used for further analysis.

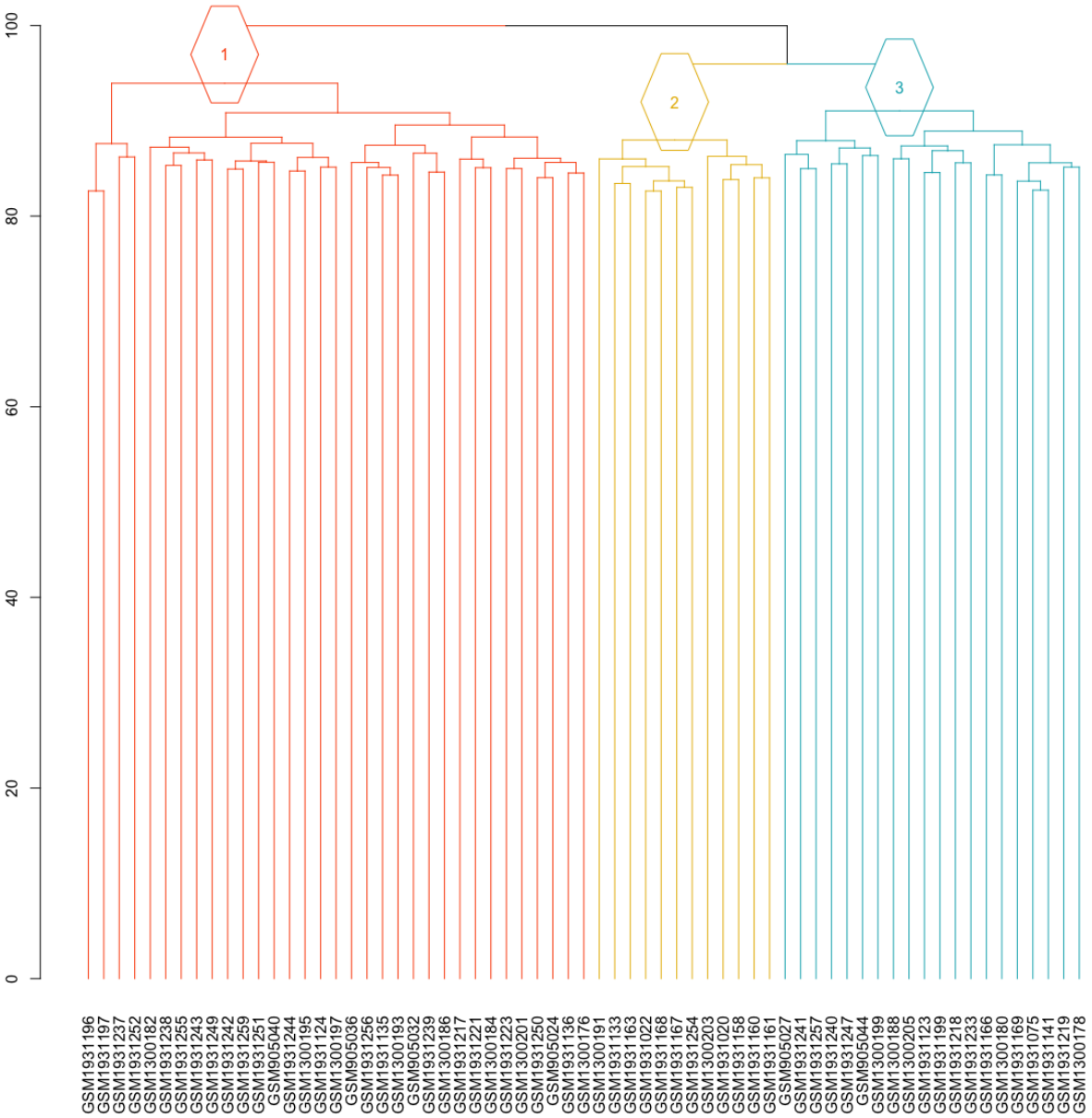

Figure B2: Hierarchical clustering dendrogram of KIRC patients in GEO dataset. The normalized gene expression of the sixty-five KIRC cancer samples were subjected to clustering analysis, to reveal the grouping of cancer samples. The GEO dataset verified the three KIRC subtypes.

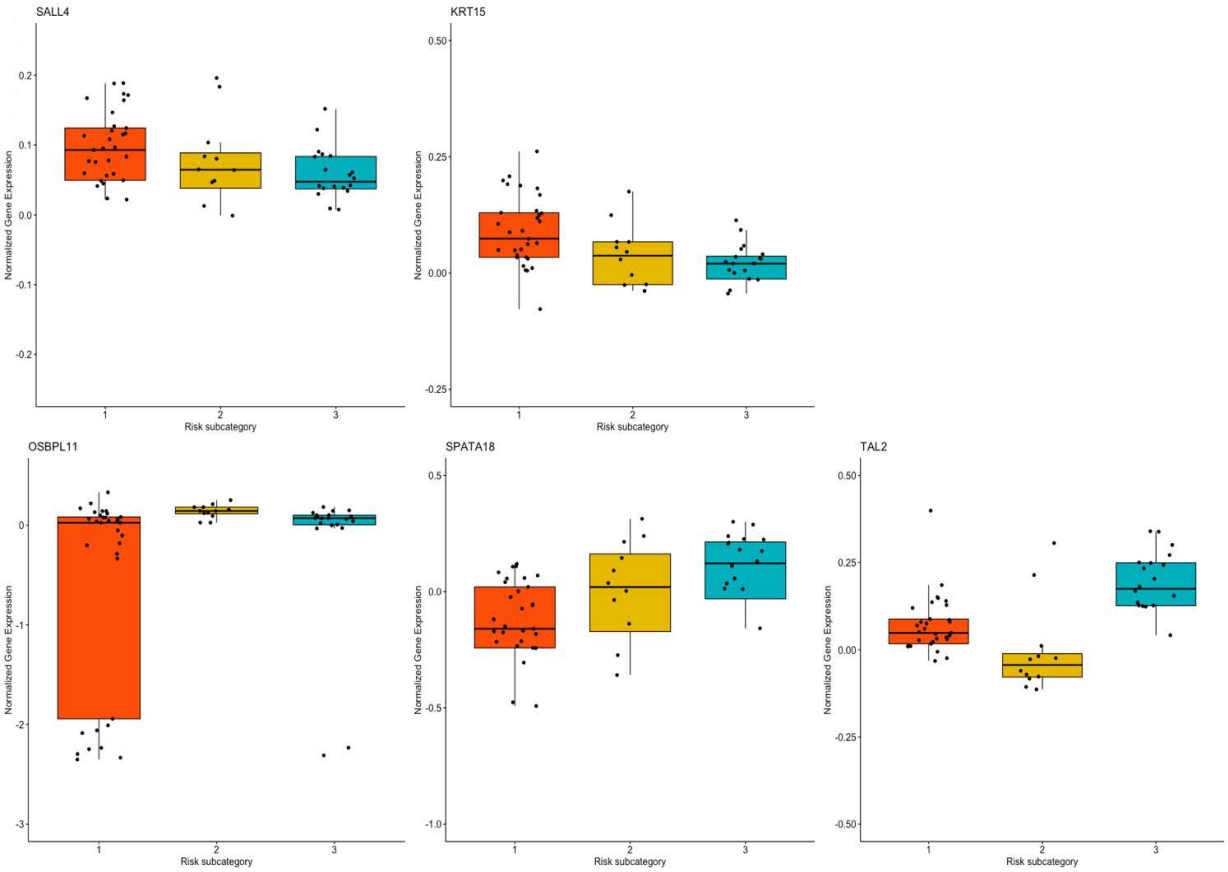

Figure B3: Boxplots were constructed of the five prognostic genes identified by the TCGA dataset. The normalized gene expression profiles of the five prognostic genes in all the samples that were categorized into clusters was extracted from the GEO dataset. Genes *OSBPL11* and *TAL2* in the GEO dataset illustrated a similar gene expression pattern to the TCGA dataset for cluster 1 (short survival) and cluster 3 (long survival). The remaining three prognostic genes, *SALL4*, *KRT15*, and *SPATA18* showed similar gene expression patterns for all three clusters in the TCGA and GEO datasets.