Reconstruction of gene regulatory networks of inflammation-associated genes in different clinical stages of diffuse large B-cell lymphoma

RESPICE PROSPICE

UNIVERSITY *of the*
WESTERN CAPE

Nomlindelo Witness Mfuphi

A thesis presented in the partial fulfilment
of the requirement for the degree of
MAGISTER SCIENTIAE (M.Sc.) in Bioinformatics
at the South African National Bioinformatics Institute, University of the Western Cape

Supervisor: Dr. H. Bendou
Cosupervisor: Prof. A. Christoffels

June 2022

https://etd.uwc.ac.za/

Keywords

Diffuse large B-cell lymphoma

Next-generation sequencing

RNA-Seq

Inflammation

Differentially expressed genes

Differential gene expression

Gene regulatory networks

Functional enrichment analysis

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Abstract

Background: Diffuse large B-cell lymphoma (DLBCL) is a heterogeneous malignancy that is driven by complex gene regulatory networks (GRNs). Numerous genes exert distinct effects on the progression and therapeutic outcome of DLBCL. Previous studies have associated DLBCL with inflammation but the GRNs involved in this mechanism have not yet been explored. The objectives of this current study are to reconstruct inflammation-associated networks and to understand the effects of inflammation on the pathogenesis and progression of DLBCL in different clinical stages.

Methods and Materials: Different stages of DLBCL RNA-Seq expression data were downloaded from UCSC Xena and were subjected to differential gene expression (DGE) analyses, using edgeR and DESeq2, with Stage I as a reference group. The database for annotation, visualization, and integrated discovery (DAVID) was used for gene enrichment analysis to find insight into the sets of differentially expressed genes (DEGs) that drive inflammation in the DLBCL clinical stages. The gene expression data were used to elucidate the GRNs of inflammation-associated DEGs using gene network inference with ensemble of trees (GENIE3) and weighted gene expression co-expression network analysis (WGCNA) algorithms. The reconstructed GRNs were imported and visualized using Cytoscape. Next, survival analyses were carried out using the Receiver operating characteristic (ROC) and Kaplan–Meier (K-M) curves to determine the prognosis of inflammation-associated gene expression in different DLBCL stages.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Results: A total of 25 DEGs were found to be common between the clinical stages. The gene and KEGG pathways enrichment mapped the majority of the DEGs under inflammation. The networks inferred by GENIE3 and WGCNA showed that the interactions between these genes were similar, suggesting that the genes regulate each other and were also co-expressed. The K-M estimates indicated a significant survival impact of the high expression of most inflammation-associated genes on DLBCL patients.

Conclusion: The findings of this study indicated that the pathogenesis, as well as the progression of DLBCL is associated with inflammation and the expression of inflammation-associated genes has a significant impact on the survival of DLBCL patients.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Declaration of authorship

I declare that "Reconstruction of gene regulatory networks of inflammation-associated genes in different clinical stages of diffuse large B-cell lymphoma" is my own work and it has not been submitted for any degree or examination in any other university, and all the sources I have used or quoted have been indicated and acknowledged by complete references.

UNIVERSITY of the
WESTERN CAPE

Full name: Nomlindelo Witness Mfuphi

Date: 15 June 2022

Signed:

https://etd.uwc.ac.za/

Dedication

This dissertation is dedicated to the loving memory of my momma and daddy.

It is all because of your hard work and
the wonderful parents you were!
Continue resting in peace.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Bendou for believing in me, it was not an easy journey but your support made it durable. I owe a deep sense of gratitude to my friends and family for listening to me even when they did not understand what I was complaining about. I would like to also thank Ada Bertie Levenstein and National Research Foundation for the financial support.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

List of abbreviations

ABC DLBCL Activated-B-Cell like Diffuse Large B-cell Lymphoma

AUC Area Under Curve

DEG Differentially Expressed Genes

DGE Differential Gene Expression

DLBCL Diffuse Large B-cell Lymphoma

DLBCL NOS Diffuse Large B-cell Lymphoma Not Otherwise Specified

GCB DLBCL Germinal Center B-cell–like Diffuse Large B Cell Lymphoma

GENIE3 GEne Network Inference with Ensemble of trees

GEP Gene Expression Profiling

GO Gene Ontology

GRN Gene Regulatory Network

K-M Kaplan-Meier

KEGG Kyoto Encyclopedia of Genes and Genomes

LFC log_2 Fold Change

NF-kF Nuclear Factor Kappa

NGS Next-Generation Sequencing

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

NHL Non-Hodgkin Lymphoma

NOD Nucleotide-Binding Oligomerization

PMBCL Primary Mediastinal Diffuse Large B-cell Lymphoma

REAL Revised European-American lymphoma

RNA-Seq RNA-Sequencing

ROC Receiver Operating Characteristic

TAM Tumor-Associated Macrophages

TLR Toll-Like Receptors

TME Tumor Micro-Environment

WES Whole Exome Sequencing

WGCNA Weighted gene co-expression network analysis

WGS Whole Genome Sequencing

WF Working Formulation

WHO World Health Organization

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

List of figures

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

List of tables

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

CONTENTS

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

LIST OF APPENDICES

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 1

Introduction

1.1 Background

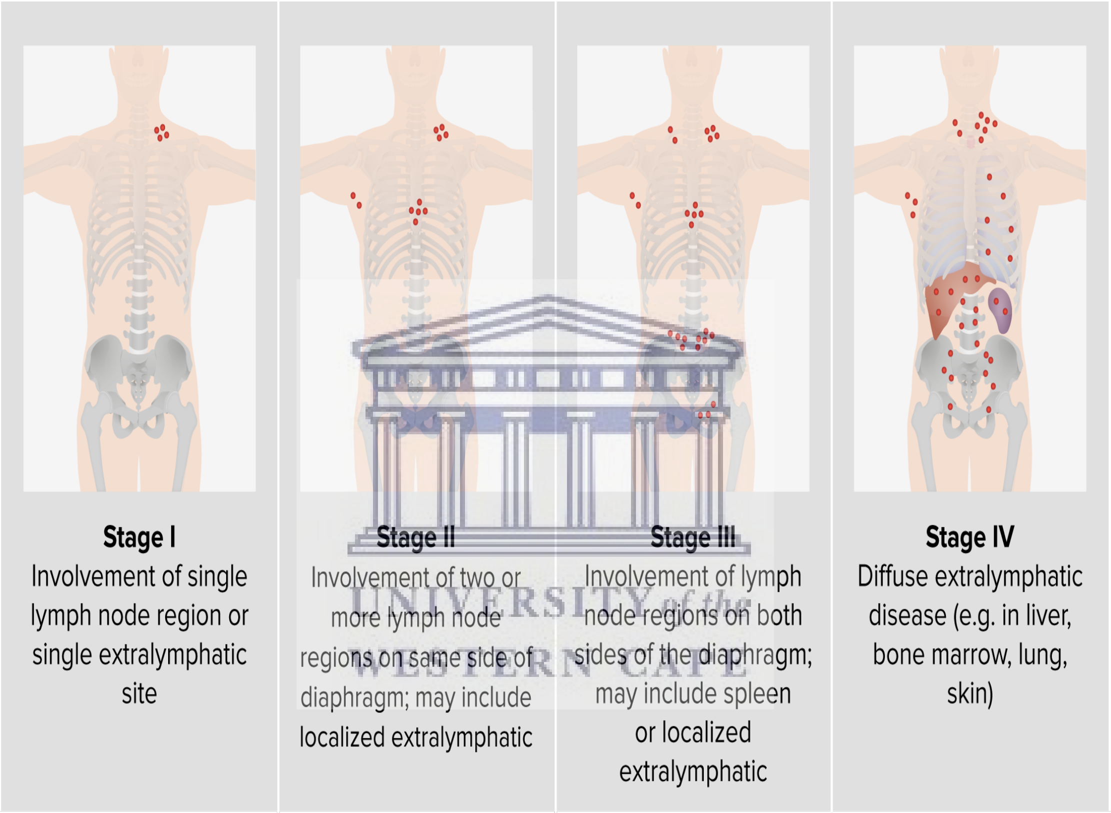
Diffuse large B-cell lymphoma (DLBCL) is the most frequently occurring non-Hodgkin lymphoma (NHL) and is characterized by the diffuse proliferation of B lymphocytes. In 2020, the estimated number of new cases of NHL was 50516 in Africa, with a slight male predominance (Global Cancer Observatory 2021). The median age of DLBCL is in the 7th decade, although other forms of DLBCL, such as primary mediastinal lymphoma, occur at a lower median age (Martelli et al. 2013; Dabrowska-Iwanicka and Walewski 2014). Most patients present with a fast-growing tumor mass encompassing one or more lymph nodes and extranodal sites (Swerdlow et al. 2016).

UNIVERSITY of the
WESTERN CAPE

DLBCL falls under a group of lymphoid malignancies that present as large lymphocytes which have a diffuse growth pattern with basophilic cytoplasm, vesicular nuclei, and nucleoli size two times or the same size as a small lymphocyte (Li et al. 2018). The Ann Arbor staging classification of Non-Hodgkin's and Hodgkin's disease and Cotswold modification of the Ann Arbor staging are accepted in staging patients with DLBCL (Lister et al. 1989). The Ann Arbor staging classification takes into account the

https://etd.uwc.ac.za/

number of sites involved based on the computational topography of the neck, chest, abdomen, pelvis, and their relation to the diaphragm. The stage is then assigned as shown in Figure 1.1. Approximately half of DLBCL patients present with Stage I-II DLBCL, whereas the other half present with Stage III-IV DLBCL (Li et al. 2018).


Stage I
Involvement of single lymph node region or single extralymphatic site
Stage II
Involvement of two or more lymph node regions on same side of diaphragm; may include localized extralymphatic
Stage III
Involvement of lymph node regions on both sides of the diaphragm; may include spleen or localized extralymphatic
Stage IV
Diffuse extralymphatic disease (e.g. in liver, bone marrow, lung, skin)

UNIVERSITY of the WESTERN CAPE

Figure 1.1: The Ann Arbor staging system of non-Hodgkin's and Hodgkin's disease. Stage I is the involvement of 1 lymph node group, Stage II is the involvement of 2 or more lymph node groups on the same side of the diaphragm, Stage III involves lymph node groups on both sides of the diaphragm and Stage IV is disseminated and a widespread disease (https://www.lecturio.com/magazine/non-hodgkins-lymphoma).

https://etd.uwc.ac.za/

DLBCL varies in terms of clinical presentation, genetic findings, therapeutic response, and prognosis (Li et al. 2018). Over the past couple of decades, next-generation sequencing (NGS) has proven to be an efficient method in the classification of DLBCL by using genomic analyses (Alizadeh et al. 2000). It has provided the first clear insight into the molecular pathogenesis of DLBCL through gene expression profiling using RNA-Seq, whole exome, and whole genome sequencing. This has not only resulted in the classification of DLBCL into different subgroups according to their genetic profiles but it also led to the generation of genetic signatures that influences mechanisms that play a role in the development and progression of DLBCL.

One of the significant challenges in biology is studying how gene expression is regulated under different conditions. Several computational methods have been identified in which transcriptional regulatory interactions can be shown at genomic levels directly from high-throughput genomic datasets. The development, as well as advancement of NGS, has provided bioscience with high throughput technologies that allow the semi-quantitative measurement of gene expression programming in great depth and on a broad genomic scale (Kwon et al. 2003). However, it is a challenge to overcome the difficulties of recognizing and evaluating relevant biological processes from vast quantities of experimental data.

UNIVERSITY of the
WESTERN CAPE

Recently, the reverse engineering of the regulatory network of genes from gene expression data has gained much attention due to emerging experimental and computational methods. The high-throughput genomic data, in particular, RNA-Seq gene expression data, can be used to computationally infer or reverse-engineer genetic interactions and graphically elucidate them as a form of regulatory interactions among genes. There is mounting evidence that supports the connection between chronic inflammation and DLBCL, as a result inflammation has been proposed to be one of the biomarkers of

https://etd.uwc.ac.za/

DLBCL (Loong et al. 2010; Didonato et al. 2012; Monti et al. 2005).

1.2 Problem Statement

DLBCL is a highly heterogeneous cancer. The complexity of achieving an accurate understanding of the pathogenesis and sub-classification criteria of all the DLBCL subtypes lies within this biological heterogeneity. The pathogenesis of DLBCL is primarily associated with oncogenes, tumor-suppressors, and genomic stable genes. However, the full spectrum of the expression of these genes contributing to the biological mechanisms of DLBCL is far from being fully known. One of the biological mechanisms frequently associated with DLBCL pathogenesis in numerous studies is inflammation. Although inflammation has been proven to be the hallmark of DLBCL, few studies have evaluated the expression of genes involved in the development and regulation of inflammation associated with DLBCL. Therefore, it is more appealing to adapt the reconstruction of GRNs to study mechanisms such as inflammation via high-throughput transcriptomic studies of DLBCL since inflammation has been associated with poor prognosis in several cancers. Such a systematic approach can provide insights into the genes commonly found to be mutated in DLBCL and are the central mediators of the GRNs that control inflammation. This may be an effective means for predicting, classifying, or targeting DLBCL.

UNIVERSITY of the
WESTERN CAPE

1.3 Aims and objectives

Most previous DLBCL studies focused more on the difference in gene expression between different molecular types of DLBCL (Arthur et al. 2018; Liu et al. 2018), but ignored the genes that are expressed in different clinical stages and the mechanisms they contribute to. Therefore, the aims and objectives of this study are as follows:

https://etd.uwc.ac.za/

1. Find differentially expressed genes (DEGs) across different DLBCL clinical stages using edgeR and DESeq2 tools.

2. Perform functional enrichment analyses using the database for annotation, visualization, and integrated discovery (DAVID) tool on the DEGs to find genes that are mainly enriched in inflammation.

3. Analyse the DEGs for inflammation-associated pathways.

4. Reconstruct the GRNs of the DEGs using GENIE3 as well as WGCNA and visualize them using Cytoscape.

5. Validate the results using statistical analyses, receiver operating characteristic and Kaplan–Meier plots to determine the impact of the expression of inflammation-associated genes on the survival of DLBCL patients under different clinical stages.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 2

Literature review

2.1 Origin and classifications of DLBCL

UNIVERSITY of the WESTERN CAPE

Rappaport described the first system of classification of lymphoma in 1966 (Rappaport 1966). This system was based on architecture and morphology, thus DLBCL was named diffuse histiocytic lymphoma. Emerging knowledge in the immunology field led to Kiel and Lukes-Collin's classification systems of lymphoma solely based on immunological concepts (Gerard-Marchant et al. 1974; Lukes and Collins 1974). In 1982, the Working Formulation for Clinical Usage classified lymphomas based on cell size, nodal pattern, and morphology (NHL-Pathological-Classification-Project 1982). At that point, the lineage and genetics of DLBCL remained largely unknown, and this lymphoma had been designated by a variety of names over the last century (Table 2.1).

https://etd.uwc.ac.za/

Table 2.1: Pathological classification history of DLBCL. Adapted from "Review in translational hematology Advances in the biology and therapy of diffuse large B-cell lymphoma: moving toward a molecularly targeted approach" by (Abramson and Shipp 2016).

Research	**The biological feature used to classify**	**Descriptor**
(Rappaport 1966)	Architecture and morphology	Diffuse histiocytic lymphoma
(Gerard-Marchant et al. 1974)	Immunology	Centroblastic lymphoma B-immunoblastic lymphoma B-large cell anaplastic lymphoma
(Lukes and Collins 1974)	Immunology	Large cleaved follicular center cell lymphoma Large noncleaved follicular center cell lymphoma B-immunoblastic lymphoma
(Rosenberg 1982)	Cell size, nodal pattern, and morphology	Diffuse mixed small and large cell lymphoma (group F) Diffuse large cell lymphoma (group G) Large cell Immunoblastic lymphoma (group H)
(Harris et al. 1994) and (Harris et al. 1999)	Genetics, immunophenotype and lymphocyte development	Diffuse Large B-cell Lymphoma

UNIVERSITY of the
WESTERN CAPE

The Revised European-American Lymphoma (REAL) classification system, which was published in 1994, comprised of immunophenotyping, lymphoid lineage, and deepened knowledge on lymphocyte development (Harris et al. 1994). This led to clearly distinguished DLBCL from other forms of aggressive NHL lymphomas such as peripheral T-cell, mantle cell, anaplastic large T-cell, follicular large cell, and Burkitt-like lym-

https://etd.uwc.ac.za/

phoma. The recent edition of the WHO classification (Swerdlow et al. 2016) subdivides DLBCL into morphological variants, molecular and immunophenotypic subgroups, and distinct disease entities (Table 2.2). The DLBCL classification has become more complex over the years due to the heterogeneity of DLBCL, therefore, many cases remain unclassified, and they are collectively termed DLBCL not otherwise specified (NOS) (Martelli et al. 2013), as shown in Table 2.2.

Table 2.2: WHO update of DLBCL Classification. Adapted from "The 2016 revision of the World Health Organization classification oflymphoid neoplasms" by (Swerdlow et al. 2016).

DLBCL subtypes and clinical entities
Diffuse large B-cell lymphoma, (not otherwise specified)
GCB versus ABC/non-GCB
MYC and BCL2 double expressor
CD5+
DLBCL subtypes
T-cell/histiocyte-rich large B-cell lymphoma
Primary DLBCL of the central nervous system
Primary cutaneous DLBCL, leg type
EBV positive DLBCL, NOS
Other lymphomas of large B-cells
Primary mediastinal (thymic) large B-cell lymphoma
Intravascular large B-cell lymphoma
DLBCL-associated with chronic inflammation
Lymphomatoid granulomatosis
ALK-positive LBCL
Plasmablastic lymphoma HHV8+ DLBCL, NOS
Primary effusion lymphoma
Borderline cases
High-grade B-cell lymphoma, with MYC and BCL2 and/or BCL6 translocations
High-grade B-cell lymphoma, NOS
B-cell lymphoma, unclassifiable, with features intermediate between DLBCL and classical Hodgkin lymphoma

UNIVERSITY of the
WESTERN CAPE

2.2 Next-generation sequencing

Although the WHO classification has had widely diagnostic valuable for some time, it lacked most of the genetic insights; this created a need for a more specific and insightful lymphoma classification system that delves deeper into the genetic make-up of DLBCL. Next-generation sequencing (NGS) collectively describes conceptual approaches such as whole genome sequencing (WGS), whole exome sequencing (WES), and RNA-Seq that make use of massively parallel sequencing of millions of deoxyribonucleic acid (DNA) templates to generate data that can be used to study genetic alterations. This has provided the first capable approach to understanding the genetic heterogeneity of DLBCL, which later led to the genomic characterization of DLBCL. Briefly discussed here, WES is the use of NGS technologies to determine the variants in gene coding regions or exons of a particular genome. Numerous studies, through WES, have redefined the genomic landscapes of DLBCL by identifying common single nucleotide variants primarily recurrent in particular subtypes of DLBCL (Pasqualucci et al. 2011; Bohers et al. 2015). WGS has made strides in deciphering genomes of tumors by uncovering somatic single-nucleotide variants, insertions/deletions (indels), structural rearrangements, and copy number alterations (Morin and Gascoyne 2013). The third NGS technique, RNA-Seq, which was used in this project, is discussed in more details in section 2.3.1.

UNIVERSITY of the
WESTERN CAPE

2.3 Gene expression profiling

Tumorigenesis of B-cell malignancies like DLBCL arises from genetic alterations in a cell that predisposes the cell to undergo further genetic alterations. Over time, cancerous cells acquire abnormalities that further promote proliferation and survival advantage above other cells. The accumulation of genetic abnormalities in a tumor

https://etd.uwc.ac.za/

results in distinct gene expression profiles, which can be determined experimentally using RNA-Seq analysis. Based on this premise, it is rational to assume that the characteristics of a tumor and clinical behavior can be predicted by profiling gene expression. Therefore, gene expression profiling (GEP) has become a powerful method for filling gaps in pathology definition by providing a molecular concept of certain cancers, wherein relatively homogenous disease entities are defined based on the common cell of origin, oncogenic mechanisms, signaling pathways, and a uniform pattern of clinical behavior (Liu et al. 2018; Miyazaki et al. 2008). This has aided in changing the definition and classification of different entities of a disease leading to the potential discovery of new, improved methods for diagnosis and treatment of a disease (Sarkozy et al. 2020).

The pathogenesis of DLBCL has been proven in numerous studies to be linked with genetic alterations that result from gene mutations. GEP studies have helped in establishing a molecular diagnosis by identifying gene expression signatures of the subtypes of this disease. Gene expression patterns revealed that gene profiles of different subgroups of DLBCL resemble normal B cells at different stages of differentiation. Hence, they are termed Germinal center B-like DLBCLs (expresses hallmark genes of normal tonsillar germinal center B cells) and Activated B cell-like DLBCLs (expresses genes that are normally activated in human blood B cells after B cell receptor stimulation) (Alizadeh et al. 2000; Eric Davis et al. 2001; Wright et al. 2003; Staudt and Dave 2005). Advancement in GEP led to the discovery of a third DLBCL sub-group namely primary mediastinal B-cell lymphoma (Savage et al. 2003) (Figure 2.1). Not only has GEP led to the segregation of molecular subtypes and given an understanding of variation in survival of DLBCL patients, but it has additionally led to deeper insights into the development and biological mechanisms of this disease, as well as the identification of rational targets for drug interventions

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

(Camicia et al. 2015).

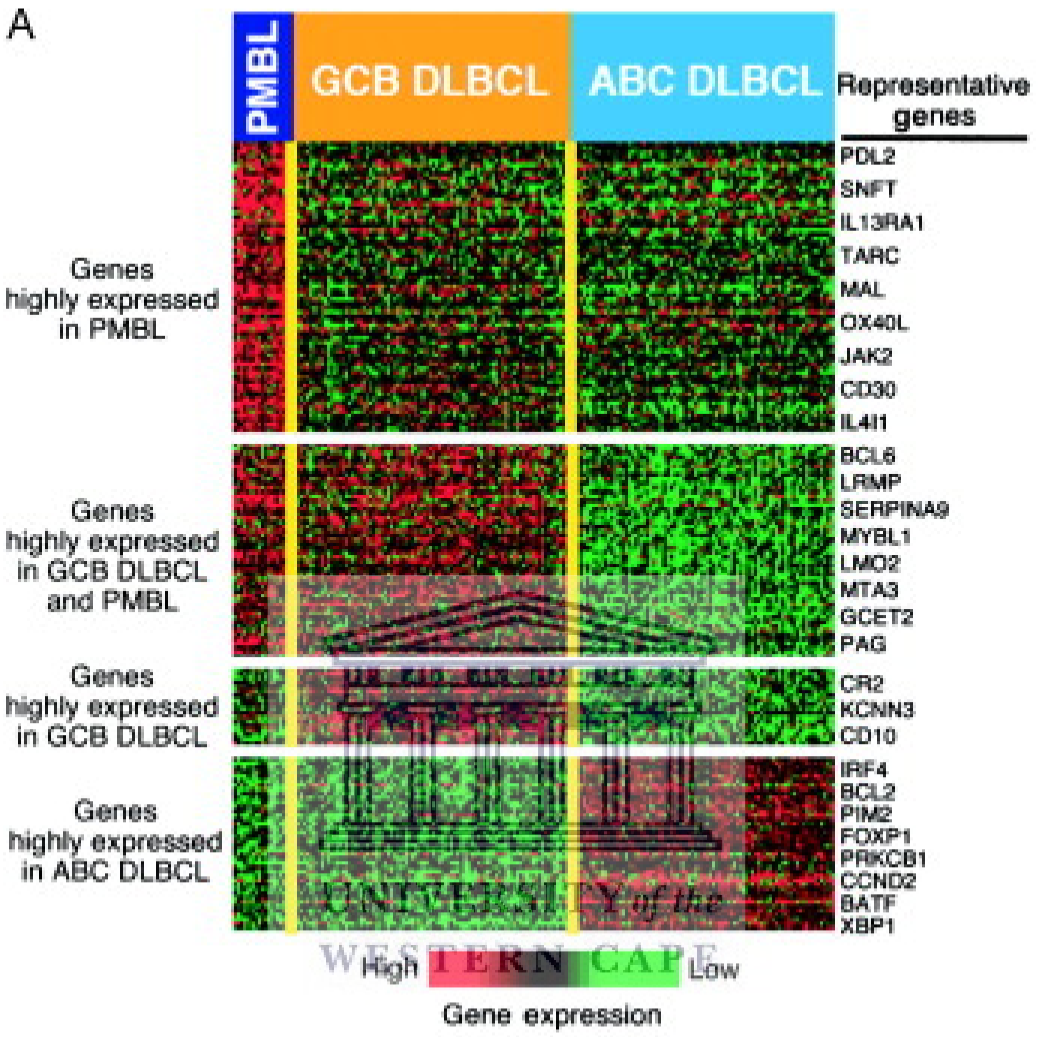


Figure 2.1: Genes characteristically expressed by three subgroups of diffuse large B-cell lymphoma (DLBCL): Primary mediastinal B-cell lymphoma (PMBL), germinal center B-cell–like (GCB) DLBCL, and the activated B-cell–like (ABC) DLBCL. Genes were grouped based on the similarities in the variation of their expression across all samples. The expression level of each gene relative to its median expression level across all samples was represented by a color, with red representing expression greater than the mean, green representing expression less than the mean, and the intensity of the color representing the magnitude of the deviation from the mean. Adapted from "Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma" by (Rosenwald et al. 2003). Open access. Source: https://doi.org/10.1084/jem.20031074.

2.3.1 Gene expression analysis using RNA-Seq

RNA-Seq uses the capabilities of high-throughput sequencing techniques to provide relevant information about the total messenger RNA, non-coding RNA, and small RNA expressed in a cell. It also provides detailed profiling of gene expression levels between conditions. RNA-Seq technology has emerged as an attractive alternative to traditional microarray platforms for conducting GEP. In comparison to Sanger sequencing and hybridization-based microarray methods, RNA-Seq provides far better coverage and clarification of the dynamic environment of the transcriptome by determining the quantity and sequences of RNA in a sample (Rapaport et al. 2013; Yang et al. 2013). In addition to quantifying gene expression, the results generated by RNA-Seq can assist researchers in discovering new transcript variants, identifying alternatively spliced genes, and also detecting allele-specific expression (Kukurba and Montgomery 2015).

RNA-Seq employs NGS techniques to sequence cDNA derived from an RNA sample, resulting in the generation of millions of short reads. These reads are generally mapped to a reference genome, and the number of reads mapping within a genomic feature of interest (such as a gene or an exon) are used as a measure of the amount of the feature present in the analyzed sample (Robinson and Oshlack 2010). The aligned reads should be greater than 40 million (Mortazavi et al. 2008), the reads are then given genes according to the common regions that they share in the alignments on the source genome. RNA expression can be calculated and subsequently compared with the amount in any other sequenced sample to quantify the expression differences during development and under different biological conditions (developmental stages, treatments, genotypes, or environments) (Mortazavi et al. 2008; Marioni et al. 2008).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

The development and existence of high-throughput data have greatly influenced the understanding and decoding of the genes that impacted the sub-grouping of DLBCL. Although not yet complete, the use of RNA-Seq on DBLCL studies has identified certain mechanisms that influence the pathogenesis of DLBCL which can be used as targets in DLBCL therapy. Table 2.3 summarizes findings from DLBCL studies using RNA-Seq.

Table 2.3: Diffuse large B-cell lymphoma (DLBCL) studies that used RNA-Seq

Sequencing methods	**Main findings**	**Cohort size**	**References**
WGS, WES, and RNA-Seq	Observed that over-expression of FCGR2B primarily in the GCB subgroup correlates with poor patient outcomes	1000	(Arthur et al. 2018)
RNA-Seq	Gene fusion between TBL1XR1 and TP63	96	(Scott et al. 2012)
RNA-Seq	Identified hub genes associated with pathogenesis	629	(Zhou et al. 2020)
RNA-Seq	Chronic active B-cell-receptor signaling in ABC DLBCL	223	(Davis et al. 2010)
RNA-Seq	Differential expression of 8 hub genes (MME, CD44, IRF4, STAT3, IL2RA, ETV6, CCND2, and CFLAR)	500	(Liu et al. 2018)
RNA-Seq	High expression of SPIB gene. FOXP1 as a potential oncogene in ABC DLBCL and mutations in PTEN gene	203	(Lenz et al. 2008)
RNA-Seq	MAFA-AS1 gene, hsa-mir-338, and hsa-mir-891a as a candidates related to the prognosis	51	(Xiao et al. 2020)
WES and RNA-Seq	Identified 313 ABC DLBCLs and 331 GCB DLBCLs, while the rest were unclassified DLBCLs	1001	(Reddy et al. 2017)

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

2.3.2 Differential gene expression using RNA-Seq

Another vital application of RNA-Seq is differential gene expression (DGE) analysis. DGE is the statistical examination of normalized read count data to uncover quantitative differences in expression levels between experimental groups. The advent of RNA-Seq technologies with reduced costs has motivated the development of statistical tools that implement approaches for the detection of differential expression of genes. The primary function of DGE analysis is to determine if genes express differently between different samples and biological conditions (Yang et al. 2013). Therefore, abnormally expressed genes are detected as differentially expressed genes (DEGs). These genes are selected using a combination of an expression shift threshold and a score cut-off, both of which are based mainly on p-values derived by statistical modeling (Rapaport et al. 2013). DEGs can be found by statistically examining if there is a significant difference between RNA-Seq read count data on different biological conditions, *i.e.,* diseased and healthy tissues or different stages of the same disease. This elucidates the essential components of the genome and reveals the molecular structure that can be targeted for the treatment of a disease (Han et al. 2015).

UNIVERSITY of the
WESTERN CAPE

Many computational tools have been developed for analyzing DGE in RNA-Seq data. This includes BBSeq (Zhou et al. 2011), DSS (Wu et al. 2013), baySeq (Hardcastle and Kelly 2010), ShinkBay (Van De Wiel et al. 2013), PoissonSeq (Li et al. 2012), limma (Mortazavi et al. 2008), edgeR (Robinson et al. 2009), and DESeq2 (Love et al. 2014). DESeq2 and edgeR are very well documented, and easy-to-use R[1] packages for DGE analysis. In recent years, edgeR and DESeq, a previous version of DESeq2, have been included in a few benchmark studies (Anders and Huber 2010; Rapaport et al. 2013) and have demonstrated great performance in replicated experiments. The

[1] R is a language and a free software environment for statistical computing and graphics.

https://etd.uwc.ac.za/

DGE analysis process consists of three major steps, namely normalization, dispersion estimation, and the test for differential expression (Varet et al. 2016). DESeq2 and edgeR normalization methods have been shown to outperform other methods, particularly when expressed genes vary across biological conditions or in the presence of highly expressed genes (Rapaport et al. 2013; Dillies et al. 2013).

2.4 Functional enrichment analyses

Following DGE analysis using RNA-Seq, the interpretation of the gene list usually requires the use of functional enrichment tools to infer gene ontology (GO) and the biological relevance of the DEGs. Functional enrichment tools help to analyze genome-scale sets of data by easing the transition from data collection to biological significance. Examples of these tools include Metascape (Zhou et al. 2019), DAVID (Dennis et al. 2003), GSEA (Subramanian et al. 2005), WebGestalt (Wang et al. 2017), Enrichr (Kuleshov et al. 2016), KOBAS (Xie et al. 2011), and g: Profiler (Reimand et al. 2007).

UNIVERSITY of the
WESTERN CAPE

Unlike other functional enrichment tools, DAVID addresses several issues that other tools have not been able to address extensively. This includes (i) integrating over 20 different types of significant gene/protein identifiers and over 40 well-known functional enrichment categories from a plethora of public databases to expand biological information coverage, (ii) possessing novel algorithms such as the DAVID Gene Functional Classification Tool, the Functional Annotation Clustering Tool, the Linear Searching Tool, the Fuzzy Gene-Term Heat Map Viewer, and others, to assist in resolving the enriched and repetitive relationships among many-genes-to-many-terms (*i.e.*, one gene may associate with several distinct, redundant terms and one term could connect with many genes), and (iii) providing the DAVID Pathway Viewer which allows the automatic graphical visualization of genes from a user's list into the most relevant

https://etd.uwc.ac.za/

Kyoto Encylopedia of Genes and Genomes (KEGG) and BioCarta enrichment pathways (Huang et al. 2007).

KEGG is a knowledge database that provides manually drawn pathway maps about genes for systematic analysis of gene functions (Wrzodek et al. 2011). KEGG is comprised of three databases namely Pathway, Genes, and Compound. The Pathway database represents higher-level functions in terms of a network of interacting molecules. The Genes database contains genomic information that is composed of gene catalogs for partially and also comprehensively sequenced genomes, as well as proteins generated by genome sequences. The Compound database is a collection of chemical compounds in a living cell, enzymes molecules, and enzymatic reactions (Ogata et al. 1999). In general, living cells' biological functions are controlled by a list of interacting genes and molecules. The function of KEGG pathways is to connect a set of genes in a genome to create a network of interacting genes or molecules in a cell to form a pathway representing a higher-order biological function (Kanehisa et al. 2002). KEGG can also be used to predict protein interaction networks and associated cellular functions by matching genes in the genome with gene products in the pathway. The KEGG databases are updated frequently and are freely accessible.

UNIVERSITY of the
WESTERN CAPE

2.4.1 Biological mechanisms of DLBCL discovered using functional enrichment analyses

In making use of cancer-specific DEGs, bioinformatics methods such as functional enrichment analyses have uncovered feasible biological mechanisms associated with different cancer development (Zhou et al. 2018). Due to the high heterogeneity of DLBCL, functional enrichment analysis has associated the genes expressed in DLBCL with numerous biological functions. Several hub and co-expressed genes are mainly

https://etd.uwc.ac.za/

enriched in immune response in general, and inflammatory response (Zhou et al. 2020). This mean that the hub genes involved in DLBCL pathogenesis play crucial roles in the development and progression processes and are DLBCL tumor-specific via affecting the immune response and inflammatory mechanisms of tumor cells. The hub genes that have been associated with DLBCL include chemokines and inflammatory cytokines (Zhou et al. 2020). Pathways enrichment tools such as KEGG have associated different cancers including DLBCL with pathways such as NF-B pathway (Didonato et al. 2012; Karin 2006; Ji et al. 2019), MAPK signaling pathway (Huang et al. 2010; Wagner and Nebreda 2009), and NOD-like receptor signaling pathway (Castaño-Rodríguez et al. 2014; Zhong et al. 2013). The activation of these pathways has been said to be a mediator of inflammation (Kaminska 2005; Franchi et al. 2009).

Inflammation in B-cell neoplasms is largely associated with the tumor microenvironment (TME) which is the interaction between cancerous B cells, inflammatory cells, stromal cells, and other immune cells. Malignant B cells and stromal cells promote the growth of the TME by releasing cytokines, chemokines, and growth factors that produce tumor-associated macrophage (TAM) (Shain et al. 2015). Cytokines and chemokines serve to lure tumor cells to the TME, in which positive cytokine and cell adhesion-mediated feedback mechanisms between neoplastic B cells and stromal cells are formed (Shain and Tao 2014). These intercellular positive feedback loops enhance not only survival and drug resistance but also the growth and proliferation of malignant B cells. TME plays a crucial role in the onset as well as the progression of lymphoma (Ruiduo et al. 2018) by either promoting the tumor progression or by using TAM to suppress the immune response for the anti-tumor response, and it is associated with poor prognosis (Kridel et al. 2012). TME assists tumor growth by further promoting inflammation, angiogenesis as well as metastasis (Shain et al. 2015).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

2.5 Gene regulatory networks

Gene regulatory networks (GRNs) are a graphical representation of the interaction of genes that is inferred from gene expression data. Reconstruction of GRNs aims to identify the interactions and expressions of subsets of genes that form the master regulatory core of biological mechanisms leading to complex conditions such as cancer. Deciphering the topology of GRNs is crucial in identifying hundreds of genes linked to the development and progression of complicated human diseases that might be used as targets for therapeutic interventions (MacNeil and Walhout 2011).

GRNs are usually denoted by directed graphs (Figure 2.2), where genes, proteins, and/or metabolites are represented by nodes, and edges that connect nodes represent molecular interactions (Kimura et al. 2005; Hecker et al. 2009). Nodes with many edges connecting to other nodes are known as hub genes and are normally the transcription factors that influence the suppression or over-expression of numerous genes. Therefore, GRNs tend to contain as much information as the high-dimensional genomic data (Wang et al. 2021).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

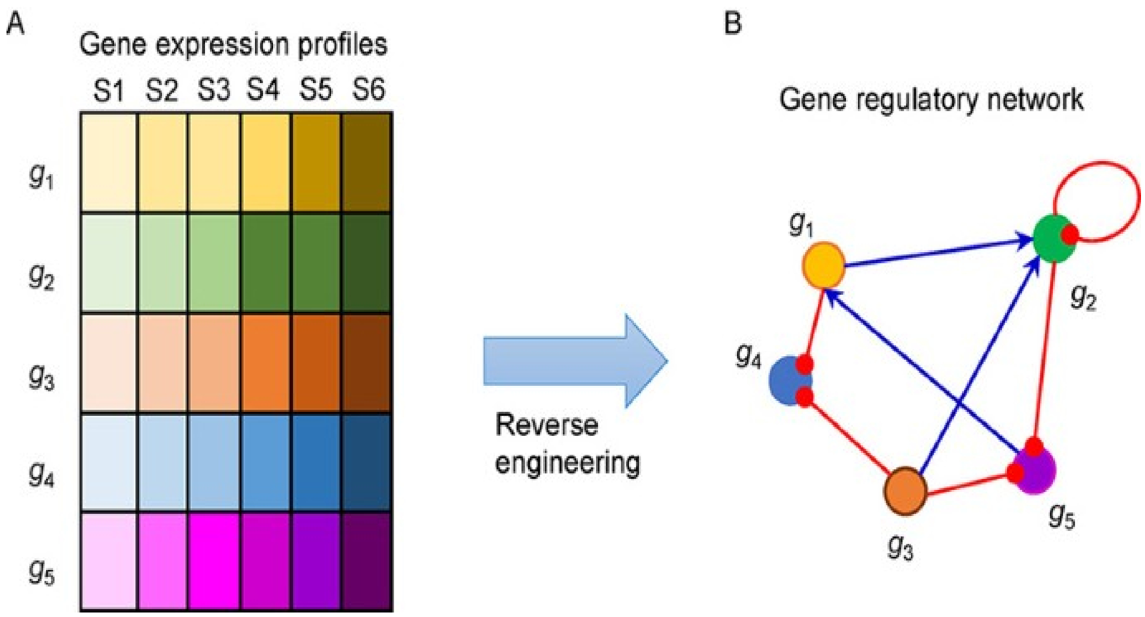


Figure 2.2: Reconstruction of gene regulatory network using expression data. Different g represent different genes and different colors represent expression levels. Adapted from "Towards precise reconstruction of gene regulatory networks by data integration." by (Liu 2018). Open access. Source: https://doi.org/10.1007/s40484-018-0139-4.

UNIVERSITY of the WESTERN CAPE

Since GRNs are constructed from gene expression data, the advances in NGS have led to the availability of enormous databases of gene expression compendia, which allows for the high-throughput and large-scale network topology inference (Ruyssinck et al. 2014). As a result, a number of computational tools for gene regulatory networks inference from gene expression data have been developed and explored, they are now employed in real-world applications (Table 2.4). Of these computation network tools are WGCNA (Langfelder and Horvath 2009) and GENIE3 (Huynh-Thu et al. 2010) which were used in this project.

2.6 Weighted Gene Co-expression Network Analysis (WGCNA)

Another powerful bioinformatics tool that has been successfully applied to high-throughput data to extract co-expressed gene networks based on their similar expression

Table 2.4: Gene regulatory network inference methods

Network inference method	**Description**	**Network inference method example**
Correlation-based network	Uses weighted correlation coefficient of two genes. Two genes are predicted to interact if their correlation coefficient is above a set threshold. Some network inference uses mutual information to infer gene regulatory interactions.	ARACNe CLR RELNET MRNET C3NET WGCNA
Boolean network	Uses binary values that define the state of a gene. Each gene (node) can take two possible values, 1 or 0. 1 represents "ON", the gene is expressed, and 0 represents "OFF", the gene is not expressed.	REVEAL
Bayesians network	The expression of each gene is considered to be a random variable following probability distributions.	BANJO networkBMA GeneNet BNT ebdbNet
Auto-regressive models	The expression of one gene is predicted from the expression of all the other genes using tree-based ensemble methods (Random Forests or Extra-Trees).	GENIE3 TIGRESS INFERELATOR
Clustering	Visualizes gene expression by grouping genes with similar expression profiles in clusters. Not used for network inference.	Hierarchical clustering

UNIVERSITY of the
WESTERN CAPE

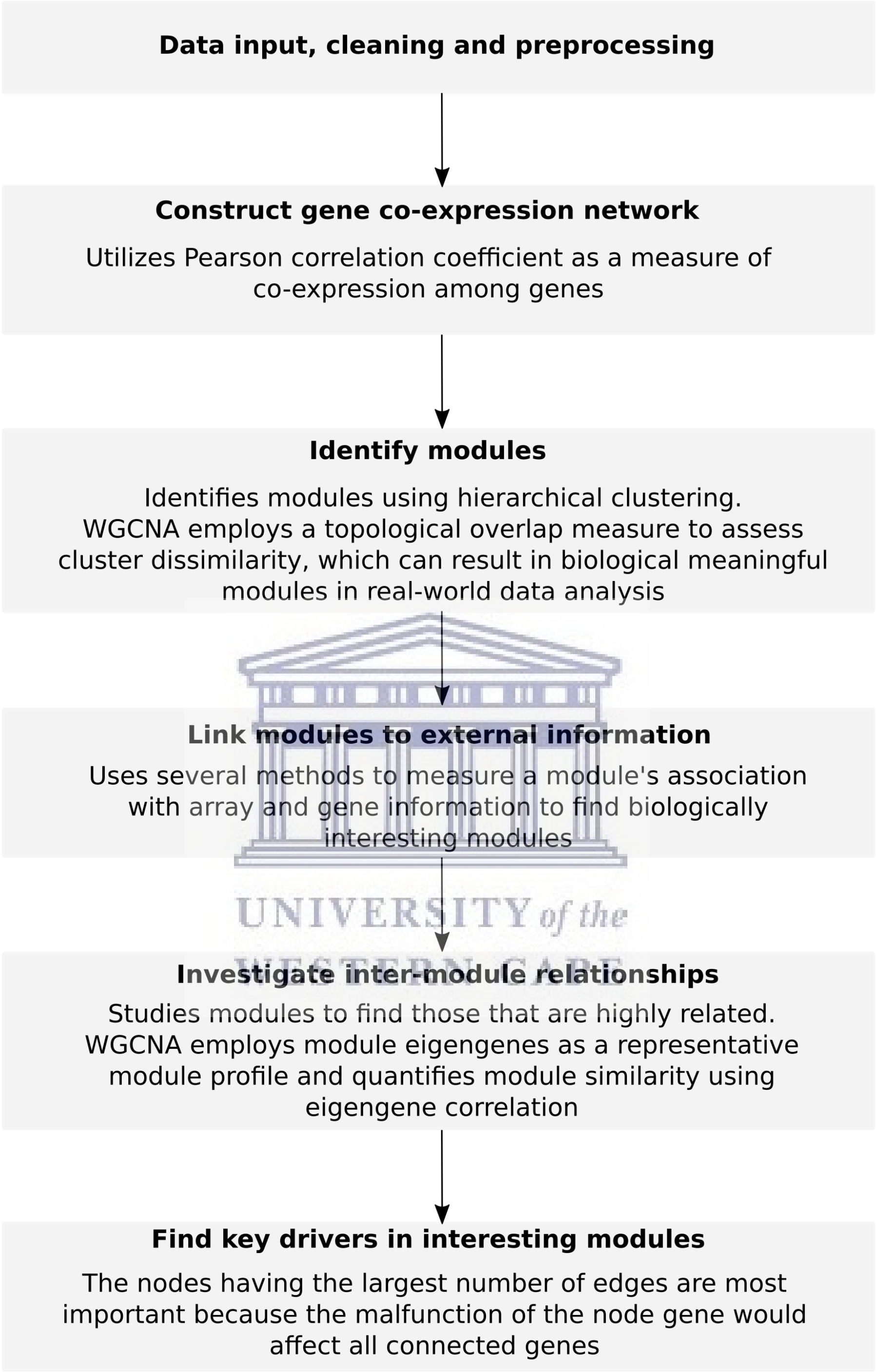
https://etd.uwc.ac.za/

profiles is WGCNA (Langfelder and Horvath 2008). WGCNA is an R package with various functions for gene networks construction, gene co-expression clusters (module) detection, gene selection, calculations of topological properties, data simulation, and interfacing with external software such as Cytoscape (Shannon et al. 2003) for the visualization of the co-expressed genes networks (Figure 2.3). One of the attractive features of WGCNA is that instead of linking thousands of genes to the physiologic trait, it focuses on the relationship between a few modules and the trait (Lu et al. 2014).

A gene co-expression network is inferred from a weighted adjacency matrix indicating the degree of connections across gene pairs. Connection strength is calculated by adjusting the pairwise correlations among gene expression profiles across samples with a power-law function that down-weights poorer correlation such that the gene co-expression network approaches a scale-free topology. This is used to estimate the topological similarity between genes, then adjacency is modified for the proportion of shared connections, and the topological overlap-based dissimilarity matrix is subjected to hierarchical clustering. Finally, WGCNA arranges genes into co-expressed gene sets by cutting the cluster dendrogram at a height that maximizes intra-connectedness within a cluster of genes (Greenfest-Allen et al. 2017; Langfelder and Horvath 2008).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/



Data input, cleaning and preprocessing

Construct gene co-expression network

Utilizes Pearson correlation coefficient as a measure of co-expression among genes

Identify modules

Identifies modules using hierarchical clustering. WGCNA employs a topological overlap measure to assess cluster dissimilarity, which can result in biological meaningful modules in real-world data analysis

Link modules to external information

Uses several methods to measure a module's association with array and gene information to find biologically interesting modules

Investigate inter-module relationships

Studies modules to find those that are highly related. WGCNA employs module eigengenes as a representative module profile and quantifies module similarity using eigengene correlation

Find key drivers in interesting modules

The nodes having the largest number of edges are most important because the malfunction of the node gene would affect all connected genes

UNIVERSITY of the
WESTERN CAPE

Figure 2.3: A flow diagram representing main steps of weighted gene co-expression network analysis. This flowchart was adapted from ”WGCNA: An R package for weighted correlation” by (Langfelder and Horvath 2008). Open access. Source: https://doi.org/10.1186/1471-2105-9-559

https://etd.uwc.ac.za/

2.7 GEne Network Inference with Ensemble of trees (GENIE3)

GENIE3 has been proven to be an overall top performer when it comes to topology inference of GRNs from high-throughput data in the DREAM4 Multifactorial Network challenge and the DREAM5 Network Inference challenge (Ruyssinck et al. 2014). This algorithm has also been analyzed and compared to other algorithms in several independent studies (Bellot et al. 2015; Feizi et al. 2013; Maetschke et al. 2014; Omranian et al. 2016), usually exhibiting competitive performance results in most cases. The key advantage of GENIE3 over other techniques is that it makes very few assumptions about the existing relationship between variables, and can possibly capture high-order conditional interconnections between expression patterns. It also generates a directed graph of regulatory interactions and normally enables feedback loops to exist in the network. Simultaneously, it remains intuitive, algorithmically tractable, and simple to implement (Huynh-Thu et al. 2010).

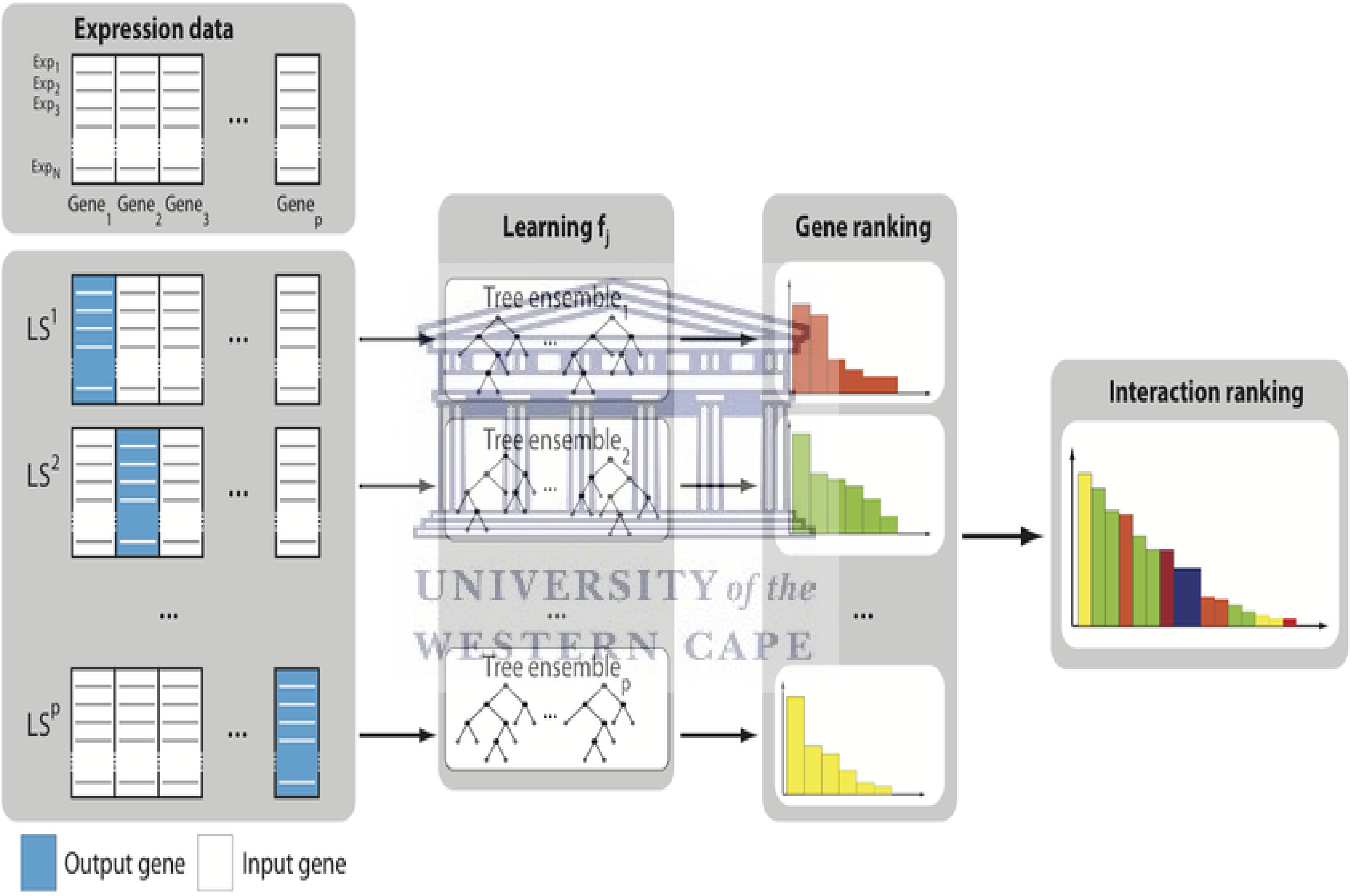
The GENIE3 algorithm predicts the gene regulatory network by using a tree-based ensemble Random forest from steady-state expression data. This algorithm decomposes the network inference task into separate regression problems for each gene in the network in which the expression values of a target gene are predicted using all other genes as possible predictors. Tree-based ensemble methods are used to calculate how important a predictor gene is to the target gene. With greater importance signifying a likely interaction or regulatory link between both genes (Breiman 2001). GENIE3 then provides a ranking of the regulators of the target gene by deriving a weight for each regulator based on an ensemble of the tree (Figure 2.4).

GENIE3 first generates a learning sample where the expression profile of gene j is the output and the expression of all other genes in the sample is the input. Genes that

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

are strong predictors for gene j expression profiles are considered the gene regulators. A decision tree is constructed for each gene j thousand times (using bootstrapped samples) where the root of each tree contains all observations, which are split into subsets that are more similar than those in the parent node, as shown in Figure 2.4. These trees are averaged in order to get the most likely genes regulating gene j (Huynh-Thu et al. 2010).



UNIVERSITY of the WESTERN CAPE

Figure 2.4: GENIE3 procedure. A learning sample (LS^1...LS^j) is generated for each gene ($Gene_1$,...,$Gene_p$,) with expression levels of gene j as output values and expression levels of all other genes as input values. Adapted from “Inferring regulatory networks from expression data using tree-based methods” by (Huynh-Thu et al. 2010). Open access. Source: https://doi.org/10.1371/journal.pone.0012776.

https://etd.uwc.ac.za/

2.8 Cytoscape

Cytoscape (Shannon et al. 2003) is implemented as an easily accessible, open-source software package with a programmable application programming interface written in Java[2], which integrates biological network analysis and visualization. The structural basis of Cytoscape is a network model with genes, proteins, metabolites, cells, or patients represented as nodes, and the interactions as edges between nodes. Attributes, which map nodes or edges to specific data values such as gene expression levels or protein functions, are used to integrate data with the network. Attribute values can be used to modify the appearance of nodes and edges (such as shape, color, and size), and to execute complicated network queries, filtering procedures, and other analyses (Smoot et al. 2011).

Function annotations, pathways, and expression profiles of genes can be imported and mapped into Cytoscape networks. Apart from these fundamental features, Cytoscape stands out from other network visualization tools by allowing and promoting re-architecture by active third-party development of add-on visualization and analysis applications hence providing performance and versatility. The Cytoscape App Store (Lotia et al. 2013) has a significant variety of biological network plugins (Saito et al. 2012) which provide additional functionalities such as data import from other sources, functional annotation and identification, module detection, literature search, network layouts, and network filtering.

[2] Java is a multi-platform, object-oriented, and network-centric programming language that is designed to have as few dependencies as possible.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

2.9 GRNs overview

The use of GRNs has resulted in the functional classification of genes that are mainly responsible for the development and progression of particular diseases (Khan et al. 2020). GRNs have been used for the detection of the most important genes in several cancer studies, and the inferred networks are used to further find the functional relevance of the networks. Through reconstruction and analysis of canine DLBCL GRNs, Zamani-Ahmadmahmudi *et al.,* (2015) found critical canine DLBCL hub genes associated with biological processes such as cell activation, cell cycle phase, immune effector process, immune system development, immune system process, integrin-mediated signaling pathway, intracellular protein kinase cascade, intracellular signal transduction, leucocyte activation and differentiation, lymphocyte activation and differentiation (Zamani-Ahmadmahmudi et al. 2015).

Using the reconstructed breast cancer GRNs, Emmert-Streib *et al.,* (2014) found cellular processes contributing to breast cancer, including cell cycling, cell adhesion, translation, organelle fission, immune response, and mitosis (Emmert-Streib et al. 2014). GRNs are a considerable endeavor to improve the diagnosis, prediction, and prognosis of different cancers. Agnelli *et al.,* (2011) reconstructed multiple myeloma GRNs and identified the most important genes associated with poor prognosis in patients with multiple myeloma (Agnelli et al. 2011). Reverse engineering of chronic lymphocytic leukemia GRNs shed light on genes that are associated with the poor prognosis of this leukemia (Yepes et al. 2015).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

2.10 Inferential statistical analyses

A common problem in practical statistical analyses is determining whether multiple samples should be considered to come from the same population. Almost always, the samples differ, and the question is whether the differences reflect population differences or are merely chance variations to be expected among random samples from the same population. When this problem arises, it is common to assume that the populations are roughly the same, in the sense that if they differ, it is due to a shift or translation. Levene's test (Levene 1960) is used to determine whether or not k samples have equal variances. The presence of equal variances across samples is referred to as variance homogeneity. Some statistical tests, such as analysis of variance, make the assumption that variances are equal across groups or samples. That assumption can be validated using the Levene's test. If the resulting p-value of Levene's test is below some threshold of significance (generally < 0.05), the observed differences in sample variances are unlikely to have occurred through random sampling from a population with equal variances. As a result, the null hypothesis of equal variances is rejected, and it is concluded that the variances in the population differ.

UNIVERSITY of the
WESTERN CAPE

The Kruskal Wallis test (Kruskal and Wallis 1952) is the non-parametric counterpart to the One Way ANOVA (Fisher 1921). The term non-parametric refers to a test that does not assume the data is coming from a specific distribution. When the assumptions for ANOVA are not met, the H test is used (like the assumption of normality). It is also known as the one-way ANOVA on ranks because the test uses the ranks of the data values rather than the actual data points. The test evaluates whether two or more groups' medians differ. One can compute a test statistic and compare it to a distribution cut-off point. The Kruskal Wallis determines whether or not there is a statistically significant difference between groups. The H statistic is

https://etd.uwc.ac.za/

the test statistic used in this test. The test hypotheses are:

$$\begin{cases} H_0 : \text{population medians are equal} \\ H_1 : \text{population medians are not equal} \end{cases} \tag{2.1}$$

2.11 Survival analyses

Survival analyses in biomedical studies are concerned with uncovering new prognosis biomarkers capable of differentiating between high- and low-risk patients. Performing survival analyses across gene expression databases determines the best performing genes in the occurrence of a disease and the minimal hazard rates to achieve clinically robust significance (Győrffy 2021). The best performing genes can be seen as predictive, prognosis, and/or diagnostic biomarkers in diseases and are capable of predicting the expected survival of patients. This can also be used in future genetic and transcriptomic studies. For the biomarkers to be useful in clinical decision-making about patient therapy and follow-up, it is common to identify and classify the diagnostic accuracy of these markers when comparing different groups (Pina et al. 1999).

UNIVERSITY of the
WESTERN CAPE

The receiver operating characteristics, or the ROC curve, is a graphical plot showing the accuracy of a biomarker for differentiating between two different groups (Fluss et al. 2005). In recent years, the ROC curve has been increasingly used in biomedical practice to investigate the effectiveness of a diagnostic marker between healthy and diseased individuals (Xiao et al. 2019; Wang et al. 2019). The ROC curve is created by plotting values of the true positive rate (sensitivity, y-axis) versus the false positive rate (1-specificity, x-axis) for a given cut-off value. The sensitivity is the actual number of true positive decisions and the specificity is the number of actual negative cases, this is well summarized in the confusion matrix shown in Figure 2.5.

https://etd.uwc.ac.za/

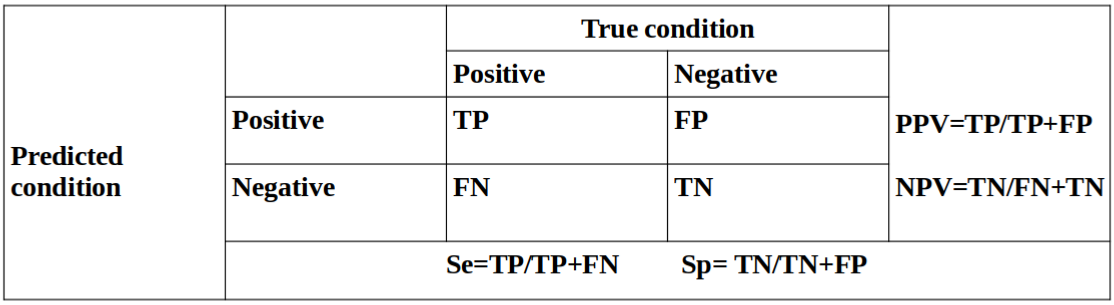
In medical research, ROC sensitivity is described as the probability of a diseased individual being predicted to have the disease, while ROC specificity is defined as the probability of a non-diseased individual being predicted to not have the disease (Kamarudin et al. 2017). This method has proven to be a well-established and reliable statistical tool for assessing the efficacy and accuracy of biomarkers as well as prognosis (Kamarudin et al. 2017).

The ROC curve is associated with numerous summary indices. The area under the ROC curve (AUC) is one of the frequently used indices. The AUC can be used to determine the discrimination power of tumor markers between different patient groups (Weiss et al. 2003). AUC is the average measure of sensitivity for all possible values of specificity and it measures the overall performance of the biomarker (Obuchowski 2003). It ranges from 0 to 1, the closer it is to 1 the better the overall performance of the test (Park et al. 2004). The practical lower limit for an AUC is 0.5, $(0.5<\text{AUC}\leq 0.7)$ the performance of the test is considered acceptable, $(0.7<\text{AUC}\leq 0.9)$ is considered excellent, and the ideal or perfect test AUC falls between 0.9 and 1 (Mandrekar 2010). This means that if the AUC is greater than 0.7 then it has a good ability to distinguish between different subjects. The ROC AUC is an excellent summary measure of classification performance because it is unaffected by disease prevalence or the cut-off points used to form the curve (Obuchowski 2003).

UNIVERSITY of the
WESTERN CAPE

The next step that follows the identification of the biomarkers is the ranking and establishment of threshold values, which aid in the rapid identification and filtering of genes associated with the disease and that can be used in the development of treatments that targets the biomarkers. The Youden index (J) is a frequently used

https://etd.uwc.ac.za/



		True condition		
		Positive	Negative	
Predicted condition	Positive	TP	FP	PPV=TP/TP+FP
	Negative	FN	TN	NPV=TN/FN+TN
		Se=TP/TP+FN	Sp= TN/TN+FP	

Figure 2.5: Confusion matrix. Se= Sensitivity, Sp= Specificity, TN= True Negative, TP= True Positive, FP= False Positive, FN= False Negative, PPV= Positive Predictive Value, and NPV= Negative Predictive Value.

threshold measure of the ROC curve (Fluss et al. 2005). It assesses the efficacy of a diagnostic marker while also allowing for the classification of an ideal cut-off point for the diagnostic marker. J can be calculated as follows:

$$J = (sensitivity + specificity - 1) \tag{2.2}$$

UNIVERSITY of the
WESTERN CAPE

and it ranges from 0 to 1. The optimal outcome is when the sensitivity is equal to 1 and at the same time, the specificity is also 1, *i.e.*, the false positive rate $(1 - specificity)$ is 0. When $J = 1$ it means the distributions of marker values for the different populations are completely separated, whereas $J = 0$ means they completely overlap. The closer J is to 1 the clearer the separation is between the different subjects. The J has an appealing feature that the AUC does not have. J specifies a criterion for selecting the optimal threshold value, *i.e.*, the threshold value for which formula 2.2 is maximized (Greiner et al. 2000).

One of the main focuses of survival analysis is investigating the time between entry to a study and a subsequent event like death. The Kaplan-Meier (K-M) is one of the methods for survival analysis which shows the probability of surviving a disease after a given time while taking into consideration time in small intervals (Goel et al.

https://etd.uwc.ac.za/

2010). The K-M method entail calculating survival probability at a given point in time. The survival probability in each interval is calculated as the number of subjects surviving divided by the number of patients at risk. The total probability of survival until that time interval is calculated by multiplying all the probabilities of survival at all-time intervals preceding that time to get the final estimate.

K-M can also be used to compare two different groups of a subject, *i.e.,* patients in different stages of the same disease. The log-rank testing in K-M is used to assess if the difference in survival times between two groups is statistically significant by providing a p-value, but it does not allow for the effect of other independent variables, *i.e.,* confidence intervals to be tested (Stel et al. 2011). In the log-rank test, the expected number of events in each group, E1 and E2 is calculated, while O1 and O2 represent the total number of observed events in each group. The K-M survival analysis is also capable of assessing and comparing the survival of patients treated with different treatments, and has been used in DLBCL in several studies (Han et al. 2019; Adams et al. 2009; Huang et al. 2012).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 3

Methods and Materials

3.1 Dataset retrieval

The UCSC Xena is a web-based visualization and exploration tool that hosts open access and easy to retrieve cancer genomics datasets derived from public hubs (TCGA, Pan-Cancer Atlas, PCAWG, ICGC, TARGET, and the GDC) (Goldman et al. 2020). With more than 1500 datasets across 129 cohorts, UCSC Xena allows users to explore for connections between genomic and phenotypic features in functional genomic datasets. The datasets found in the cohorts include SNPs, INDELs, large structural variants, CNV, RNA-Seq gene expression, DNA methylation, clinical and phenotypic annotations.

UNIVERSITY of the
WESTERN CAPE

XenaPython, a Python[1] package implementing application programming interfaces (APIs), was used to connect to the GDC hub in the UCSC Xena and download TCGA DLBCL RNA-Seq gene expression dataset (dataset ID: TCGA-DLBC.htseq_counts.tsv). First, the XenaPython *dataset_samples()* function was used to retrieve the DLBCL sample names using the link address to the GDC hub and the dataset ID as

[1] Python is a dynamically structured, interpreted, object-oriented high-level programming language used for software and development, mathematics, and system scripting.

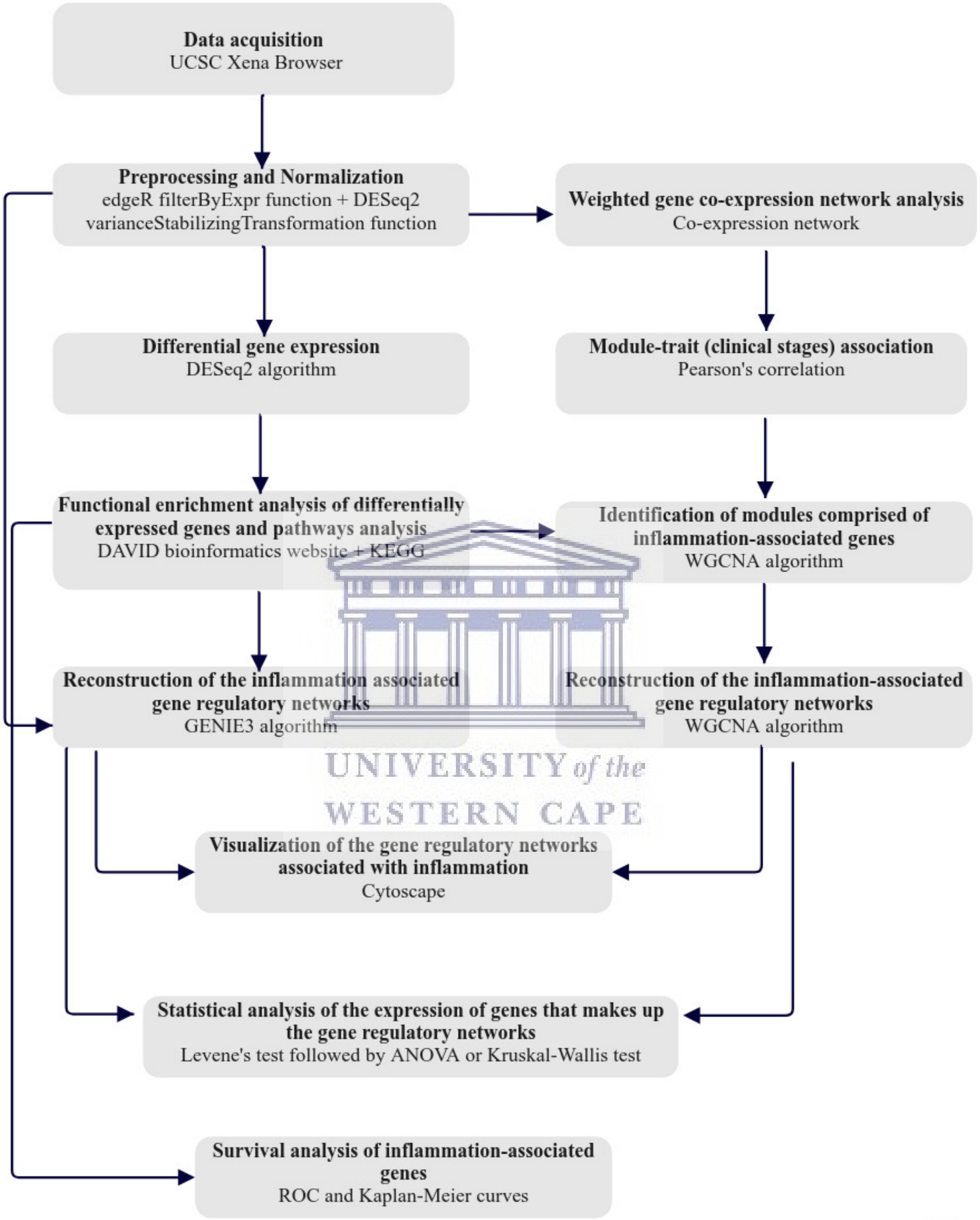
https://etd.uwc.ac.za/

arguments. The samples were filtered based on the clinical stages information from the DLBCL phenotype file downloaded from Xena. This resulted in the removal of samples with no clinical information. Next, the XenaPython *dataset_field()* function was used to retrieve probe (gene) names using the link address to the GDC hub and the dataset ID as arguments. This was followed by the use of the Xenapython *dataset_fetch()* function to retrieve the gene expression counts of each probe of all samples. This query yielded a Python data frame of 42 DLBCL samples as column names (Stage I = 8, Stage II = 17, Stage III= 5, and Stage IV = 12) and 60,489 probes as row names with RNA-Seq normalized counts in $log_2(x+1)$ as content, where x represents the raw count value of a probe.

WGCNA documentation suggests that the combined number of samples used for analysis should be greater than 15, otherwise the network would simply be too noisy for the network to be biologically meaningful (Langfelder and Horvath 2017). While considering Stage I (8 samples) as a reference group, the number of samples in Stage III is not sufficient to reach a statistically significant network for the WGCNA analysis. Therefore, Stage III samples were not used in the downstream analyses in this study. The schematic diagram of the study methods is shown in Figure 3.1. The R and Python scripts used for data preparation and analyses can be found in the publicly accessible repository available on https://github.com/Nomlindelow/DLBCL_Inflammation_GRNs.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/


Data acquisition
UCSC Xena Browser
Preprocessing and Normalization
edgeR filterByExpr function + DESeq2
varianceStabilizingTransformation function
Weighted gene co-expression network analysis
Co-expression network
Differential gene expression
DESeq2 algorithm
Module-trait (clinical stages) association
Pearson's correlation
Functional enrichment analysis of differentially
expressed genes and pathways analysis
DAVID bioinformatics website + KEGG
Identification of modules comprised of
inflammation-associated genes
WGCNA algorithm
Reconstruction of the inflammation associated
gene regulatory networks
GENIE3 algorithm
Reconstruction of the inflammation-associated
gene regulatory networks
WGCNA algorithm
Visualization of the gene regulatory networks
associated with inflammation
Cytoscape
Statistical analysis of the expression of genes that makes up
the gene regulatory networks
Levene's test followed by ANOVA or Kruskal-Wallis test
Survival analysis of inflammation-associated
genes
ROC and Kaplan-Meier curves

UNIVERSITY of the
WESTERN CAPE

Figure 3.1: Schematic representation of the study methodology. This figure summarizes the methods employed in the current project. This includes dataset retrieval, differential gene expression, gene ontology enrichment, gene regulatory network inference, and statistical analyses.

https://etd.uwc.ac.za/

3.2 Differential gene expression

3.2.1 Data preprocessing

Prior to DGE analyses, the normalized counts from section 3.1 were converted to raw counts by raising 2 to the power of $log_2(x+1)$ value minus one.

$$2^{log_2(x+1)} - 1 \tag{3.1}$$

The DGE analyses were proceeded using the R packages edgeR and DESeq2. The edgeR *filterByExpr()* function was used to filter out low expressed genes from the raw counts' expression data of the different clinical stages (Robinson et al. 2009). This function was used to remove rows with no or nearly no reads since low-expressed features tend to reflect noise leading to the increased memory size of the data object and decreased speed of the transformation and testing functions conducted downstream.

UNIVERSITY of the
WESTERN CAPE

3.2.2 Differential gene expression analyses

Using DESeq2, DGE analysis was performed in Stage II and Stage IV using Stage I as the reference group for both stages (Love et al. 2014). DESeq2 uses shrinkage estimators for dispersion and fold change for comparative DGE estimation (Love et al. 2014). Genes with a p-adjusted value of < 0.05 were considered DEGs. Venn diagram analysis was performed using Venny 2.1 to identify and display the common DEGs across the different clinical stages.

https://etd.uwc.ac.za/

3.3 Gene ontology and functional enrichment analyses

GO and KEGG pathway enrichment analyses were performed on Stage II and Stage IV DEGs separately using DAVID. The DEGs' Ensembl IDs file for each stage was uploaded onto DAVID and the ENSEMBL_GENE_ID was chosen as the identifier, then the list was submitted. The functional annotation tool was used to analyze and perform GO and KEGG pathway enrichment on the submitted gene list. The group of genes that were considered in this study were the DEGs that were mapped by DAVID under inflammation and immune response, and have a statistically significant p-value < 0.05.

3.4 Inference of gene regulatory and co-expression networks

3.4.1 Data preprocessing

(Langfelder and Horvath 2017) suggest the use of DESeq2 varianceStabilizingTransformation function in R for data normalization. Therefore, the gene raw counts were normalized using this function prior to performing WGCNA and GENIE3 analyses. The function computes a variance stabilizing transformation (VST) from the fitted dispersion-mean relations, then normalizes the raw count data by dividing the counts by the size factors resulting in a matrix of values that have a constant variance along with the range of mean values (Love et al. 2014).

UNIVERSITY of the WESTERN CAPE

3.4.2 Inference of gene co-expression networks using WGCNA

WGCNA package in R (Langfelder and Horvath 2008) was used to create weighted gene co-expression networks of Stage II and IV DEGs. To begin, the interaction patterns among genes were calculated using the absolute value of the Pearson correlation in

https://etd.uwc.ac.za/

order to generate a gene expression similarity matrix S =(s_{ij}).

$$s_{ij} = \left[\frac{1 + corr(x_i, x_j)}{2}\right] \tag{3.2}$$

Where s_{ij} is the Pearson correlation coefficient between gene i and gene j. x_i and x_j represent the gene expression values of gene i and gene j, respectively.

A gradient test (power value ranging from 1 to 10) was used to determine scale independence and average connectivity of modules with various power values. The adjacency matrix $A = (a_{ij})$ was then created from the gene expression similarity matrix.

$$a_{ij} = |s_{ij}|^{\beta} \tag{3.3}$$

Where s_{ij} is the Pearson correlation coefficient between gene i and gene j from matrix S and β is the soft-power threshold.

UNIVERSITY of the
WESTERN CAPE

Since the present study was performing network analysis on a dataset with a large number of genes, a block-wise network construction and module detection method was used. Firstly, the genes were pre-clustered into blocks of size close to and not exceeding a maximum of 5000. This was followed by a full network analysis in each block separately. The adjacent matrix A was transformed into a topological overlap matrix (TOM) in order to identify modules of highly co-expressed genes based on adjacency a_{ij}.

$$TOM_{ij} = \left[\frac{\sum_{\mu \neq, i,j} a_{i\mu} a_{\mu j} + a_{ij}}{min(\sum_{\mu} a_{i\mu}, \sum_{\mu} a_{\mu j}) + 1 - a_{ij}}\right] \tag{3.4}$$

Where $a_{i\mu}$ represents the adjacency coefficient between gene i and gene μ when μ is

https://etd.uwc.ac.za/

different from i and j.

Next, hierarchical clustering was applied to the dissimilarity topological overlap matrix to generate a dendrogram. Co-expressed genes were given relevant modules via dynamic minimum tree cutting, thereafter, modules with similar expression patterns greater than 75% similarity were merged into one module. Thus, WGCNA identified genes that have similar co-expression and hence similar biological functions. The modules of interest that were used for the inference of gene co-expression networks were the ones that included DEGs associated with inflammation according to DAVID.

The WGCNA *exportNetworkToCytoscape()* function was used to export the pairwise gene correlation into a file for visualization. A weight threshold of 0.02 was applied to Stage II and Stage IV networks in order to select highly correlated genes. The edge and node list file was imported and visualized using Cytoscape (Shannon et al. 2003).

UNIVERSITY of the
WESTERN CAPE

3.4.3 Inference of gene regulatory networks using GENIE3

The reconstruction of the gene regulatory networks (GRNs) was done using the GENIE3 algorithm on the normalized raw counts from section 3.4.1. The GENIE3 algorithm predicts the GRNs by using a tree-based ensemble of random forests from steady-state expression data (Huynh-Thu et al. 2010). In GENIE3, by default, all the genes in expression data are used as candidate regulators and target genes. The list of candidate targets and regulators can, however, be restricted to a subset of genes. In this study, only DEGs of each stage was set to be target genes in that stage GRNs so as to only find the interactions between DEGs. Different clinical stages in this study had different numbers of DEGs. Therefore, different weight thresholds were

https://etd.uwc.ac.za/

used in GENIE3 to only consider gene connections with high regulation weights. The thresholds were 0.0016 and 0.00211 for Stage II and Stage IV, respectively. This was then visualized using Cytoscape software (Shannon et al. 2003) enabling the detection of DEGs that closely regulate one another.

3.5 Inferential statistical and survival analyses

Inferential statistical analyses were incorporated in this project to assess the significant difference in the expression means of the genes composing the inflammation-associated GRNs in Stage II and Stage IV compared to the reference stage, Stage I. These analyses were processed via R scripting language[2]. Levene's test of homogeneity of variance across groups (Stage II vs Stage I and Stage IV vs Stage I) was carried out to investigate equal variance across the WGCNA and GENIE3 networks. In the event that the Levene's test null hypothesis was rejected, *i.e.,* there was no equal variance across groups (p-value < 0.05), Kruskal-Wallis algorithm of non-parametric analysis of variance (Kruskal and Wallis 1952) was used to evaluate the variances in the expression medians across stages. Otherwise, the ANOVA test (Fisher 1921) was used. The test statistic values are included in all statistical reports, including the F ratio, chi-square ($\chi 2$), and p-value.

UNIVERSITY of the
WESTERN CAPE

ROC curves (Fluss et al. 2005) were used to validate the effect of the difference in expression of the inflammatory DEGs across groups in order to assess their diagnostic capacity. The AUC of the genes were calculated using their expression data and was evaluated to check their significance. Subsequently, the maximum Youden index (Youden 1950) of each gene based on ROC analysis was calculated to choose the optimum cut-off value that maximizes both sensitivity and specificity.

[2] R version 4.1.2 ("Bird Hippie")

https://etd.uwc.ac.za/

To visualize the correlation between the gene expressions and overall survival (OS) of DLBCL patients, the Kaplan-Meier (K-M) curves were constructed. The Youden indices of DEGs with AUC greater than 0.7 were utilized in specifying the cut-off gene expression values. The patient status was classified as 1 if the DEG expression value exceeds the Youden Index cut-off value (high-expression), otherwise, the patient's status was classified as 0 (low-expression). Furthermore, the statistical significance was calculated using log-rank tests in K-M to select survival-related DEGs. Statistical significance was established at a p-value < 0.05. The OS was measured from the date of diagnosis to the last follow-up.

UNIVERSITY of the
WESTERN CAPE

Chapter 4

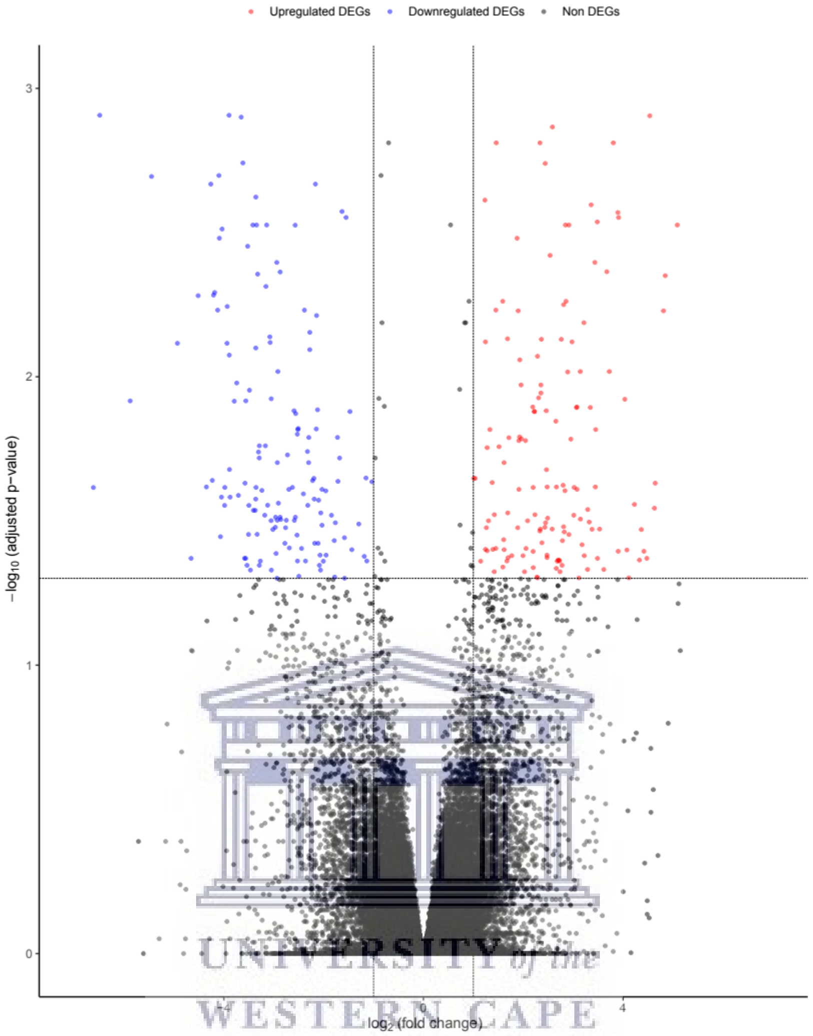
Results

4.1 Selection of DEGs

Querying the dataset ID: TCGA-DLBCL.htseq_counts.tsv from the GDC hub returned a total of 48 DLBCL samples and 60,489 genes. Based on the gene expression in these samples, the edgeR *filterByExpr()* function removed low expressed genes and kept 20,022 genes for downstream analysis. The dataset was queried using clinical-stage criteria, yielding 37 samples (Stage I = 8, Stage II = 17, and Stage IV = 12). Based on DESeq2, applying p-adjusted value < 0.05, absolute value $|LFC| > 0$ as cut-off and Stage I as the reference group, a total of 53 genes (19 down-regulated and 34 up-regulated genes) in Stage II and 207 (102 down-regulated and 105 up-regulated genes when compared to the reference stage, Stage I) in Stage IV were found to be differentially expressed. A volcano plot was created to visualize the distribution of the expression of DEGs across the LFC and p-adjusted value parameters (Figure 4.1).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/



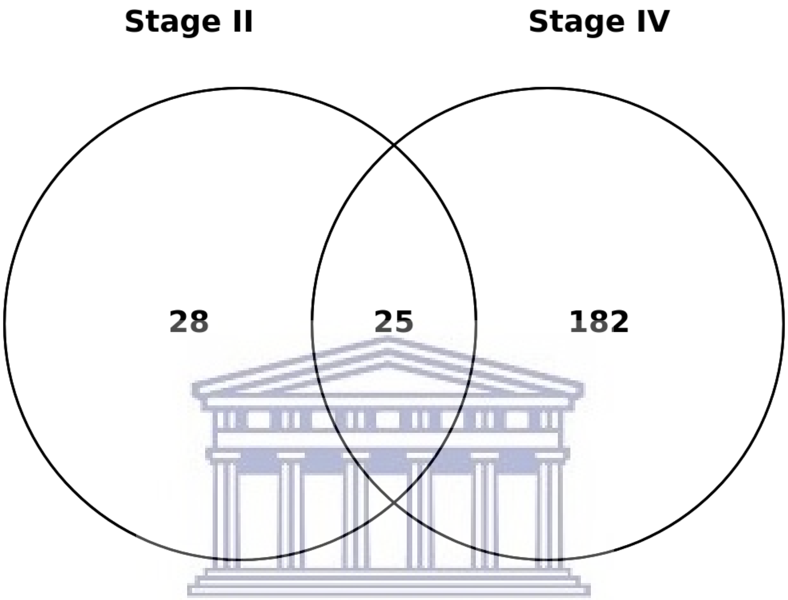
UNIVERSITY of the
WESTERN CAPE

Figure 4.1: Volcano plot portraying expressed genes in DLBCL samples. Red and blue dots indicate differentially expressed up- and down-regulated genes respectively (p-adjusted value < 0.05) respectively. While the black dots represent non-significant genes with no difference in expression between stages.

The Venn diagram of common and unique DEGs between Stage II and IV is shown in Figure 4.2. Twenty-five DEGs are common between the two stages, *i.e.,* the difference in expression of these DEGs compared to Stage I is maintained throughout Stage II to Stage IV of DLBCL. Therefore, these genes can be considered as possible drivers of DLBCL (Table 4.1). The level of changes in LFC was maintained throughout DLBCL cancer progression, and up-regulated genes in Stage II remained up-regulated

https://etd.uwc.ac.za/

even in Stage IV when compared to the reference stage, Stage I. The same behavior was observed for the down-regulated genes (Table 4.1)



UNIVERSITY of the
WESTERN CAPE

Figure 4.2: Venn diagram plot displaying the number of common and unique differentially expressed genes (DEGs) of DLBCL Stage II and IV. The Venn diagram was plotted using https://bioinfogp.cnb.csic.es/.

https://etd.uwc.ac.za/

Table 4.1: Log_2 Fold Change (LFC) and p-adjusted values of the common DEGs in different clinical stages of DLBCL

	p-adjusted value		**LFC**	
Genename	**Stage II**	**Stage IV**	**Stage II**	**Stage IV**
PTBP1P	3.122E-05	1.63E-06	4.039	4.634
TSPAN1	4.338E-05	0.0002	-3.427	-3.557
C11orf53	0.0001	0.0352	-4.575	-3.505
CTSE	0.0002	0.008	-4.978	-4.93
SPTBN2	0.0002	0.00015	4.248	3.364
BCL2L10	0.0003	0.001	5.005	4.695
RP4-781K5.6	0.0005	0.001	4.594	4.535
MYOCD	0.0006	1.2E-09	3.994	6.670
BARX2	0.001	0.027	-3.651	-3.361
CACNA1E	0.0026	0.0015	3.896	3.806
ADAMTS6	0.0074	0.032	2.364	2.087
VWCE	0.0074	0.00	2.76	2.600
RP11-830F9.7	0.0095	2.1E-06	3.143	3.129
LINC01415	0.0095	0.001	3.726	4.36
CASP5	0.0127	0.001	3.344	4.166
SOX7	0.01	0.011	1.963	1.956
RASGRP4	0.0210	0.011	2.467	2.35
TMEM132B	0.0233	0.006	4.645	4.811
MEFV	0.024	0.01	2.8	2.891
SHANK1	0.027	0.003	-2.906	-3.417
KCNQ1OT1	0.042	2.8E-06	1.697	2.203
CCL8	0.043	0.006	2.717	3.216
FFAR2	0.043	0.004	2.697	3.675
CABLES1	0.046	0.005	2.202	2.858
CXCL2	0.046	0.013	2.663	2.191

UNIVERSITY of the
WESTERN CAPE

4.2 Functional enrichment analysis

The GO functions of Stage II and Stage IV DEGs were evaluated using DAVID. The Stage II and Stage IV DEGs were mainly enriched in inflammation, immune response, positive regulation of cell proliferation, negative regulation of cell proliferation, cell adhesion, calcium ion transport, cell migration, regulation of ion transmembrane transport, cytoskeleton organization, and cell chemotaxis. The subset of DEGs that were investigated further in this study by creating their GRNs were those enriched

https://etd.uwc.ac.za/

in inflammation and immune response, as shown in Table 4.2.

Table 4.2: Results of the gene ontology (GO) analysis executed on Stage II and Stage IV DEGs

Stages	**Description**	**DEGs**	**p-value**
Stage II	chemokine-mediated signaling pathway	*CCL8* *CXCL2* *CXCL5*	1.1E-2
	Immune response	*CCL8* *CXCL2* *CXCL5* *IGHV4-34* *CSF3*	1.3E-2
Stage IV	Inflammatory response	*CXCL2* *CCL8* *MEFV* *NFKBIZ* *NLRP4* *TNFRSF10B* *ALOX15* *HMGB1P1* *FPR2* *PTX3* *SIGLEC1* *ZC3H12A*	3.6E-4
	Immune response	*CCL8* *CXCL2* *TNFRSF10B* *CLNK* *ENPP3* *IL7* *PKHD1L1* *TNFSF15* *IGHV2-70* *IGKV4-1*	9.2E-3

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

KEGG pathways analysis found 2 significant inflammation-related pathways in Stage II and none in Stage IV (Table 4.3).

Table 4.3: Significantly enriched KEGG pathways for DEGs in DLBCL

clinical stage	KEGG pathway	p-value	DEGs
Stage II	NOD-like receptor signaling pathway	0.0056	*CXCL2, MEFV, CASP5*
	Cytokine-cytokine receptor interaction	0.012	*CCL8, CXCL2, CXCL5, CSF3*

4.3 Co-expression networks inferred using WGCNA

The network topology analysis was applied on the normalized gene expression raw counts to determine the appropriate soft-thresholding power (β). Normally, the scale-free topology fit index is set to a value ≥ 0.80, indicating that the network complies with the requirements of non-scale distribution. Therefore, β was set to 5 and 7 for Stage II and Stage IV networks respectively (Figure 4.3A and B). Next, the genes were clustered into modules according to their correlation similarities. The correlation among modules was calculated, and modules with a strong correlation, *i.e.,* with correlation dissimilarity of less than 25%, were merged into one module (Figure 4.3C and D).

UNIVERSITY of the
WESTERN CAPE

The generated WGCNA modules consist of highly co-expressed genes. The modules significantly enriched by DAVID GO in inflammation and immune response were the light-yellow and dodgerblue for Stage II and the tan for Stage IV. The co-expression networks associated with the above modules are shown in Figure 4.4A and B for Stage II, and Figure 4.5A for Stage IV. In Stage II, WGCNA identified 2 networks, one with 5 genes related to inflammation (Figure 4.4A) while the other network is composed of 3 genes related to immune response (Figure 4.4B). In regards to

https://etd.uwc.ac.za/

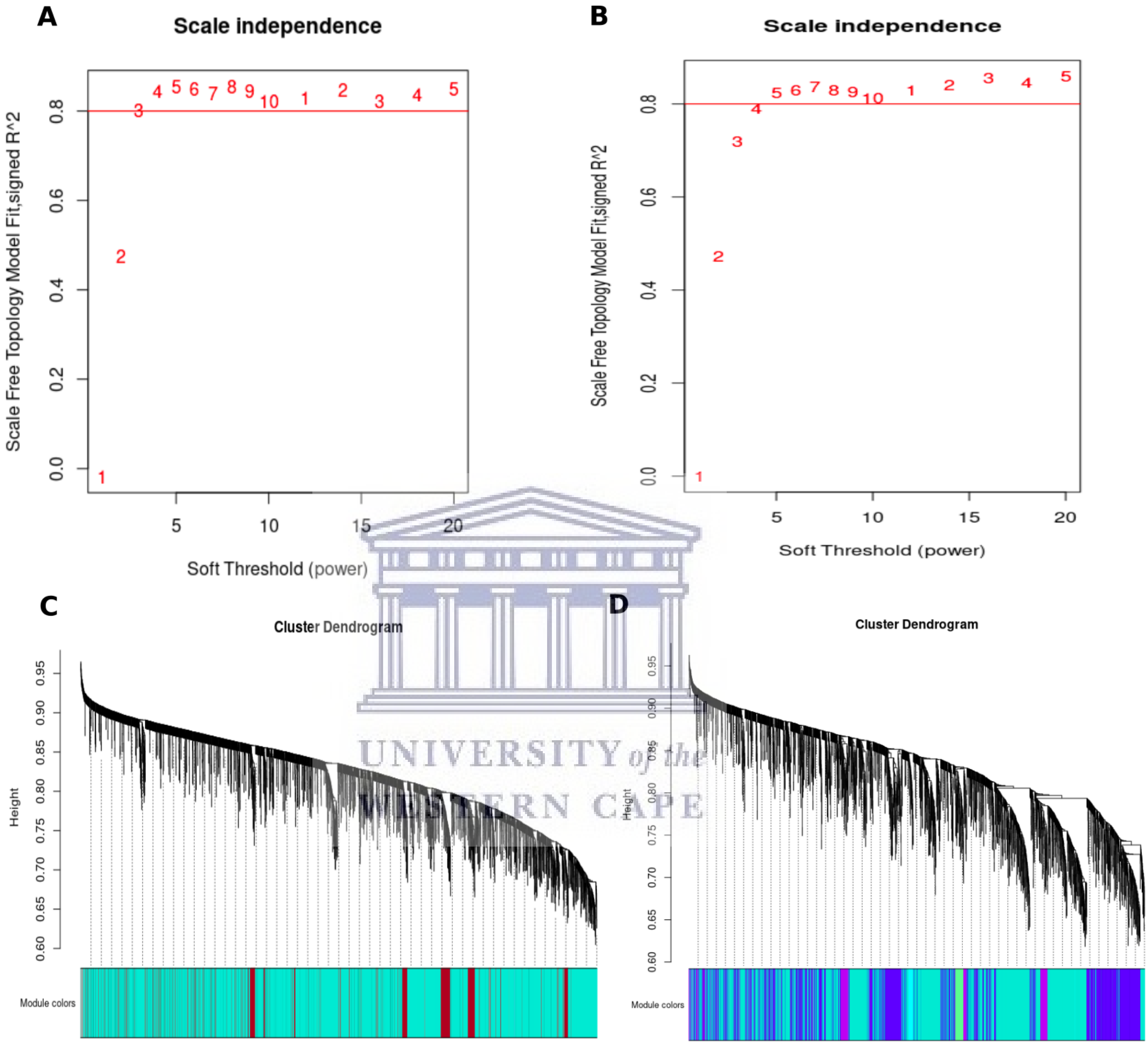


Figure 4.3: Network topology analysis for different soft-thresholding powers in a scale free fit index function; (A) for Stage II and (B) for Stage IV. Cluster dendrograms were yielded by the average linkage hierarchical clustering (C) for Stage II and (D) for Stage IV. The colors beneath the dendrograms represent the module assignments as determined by the dynamic tree-cutting algorithm.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

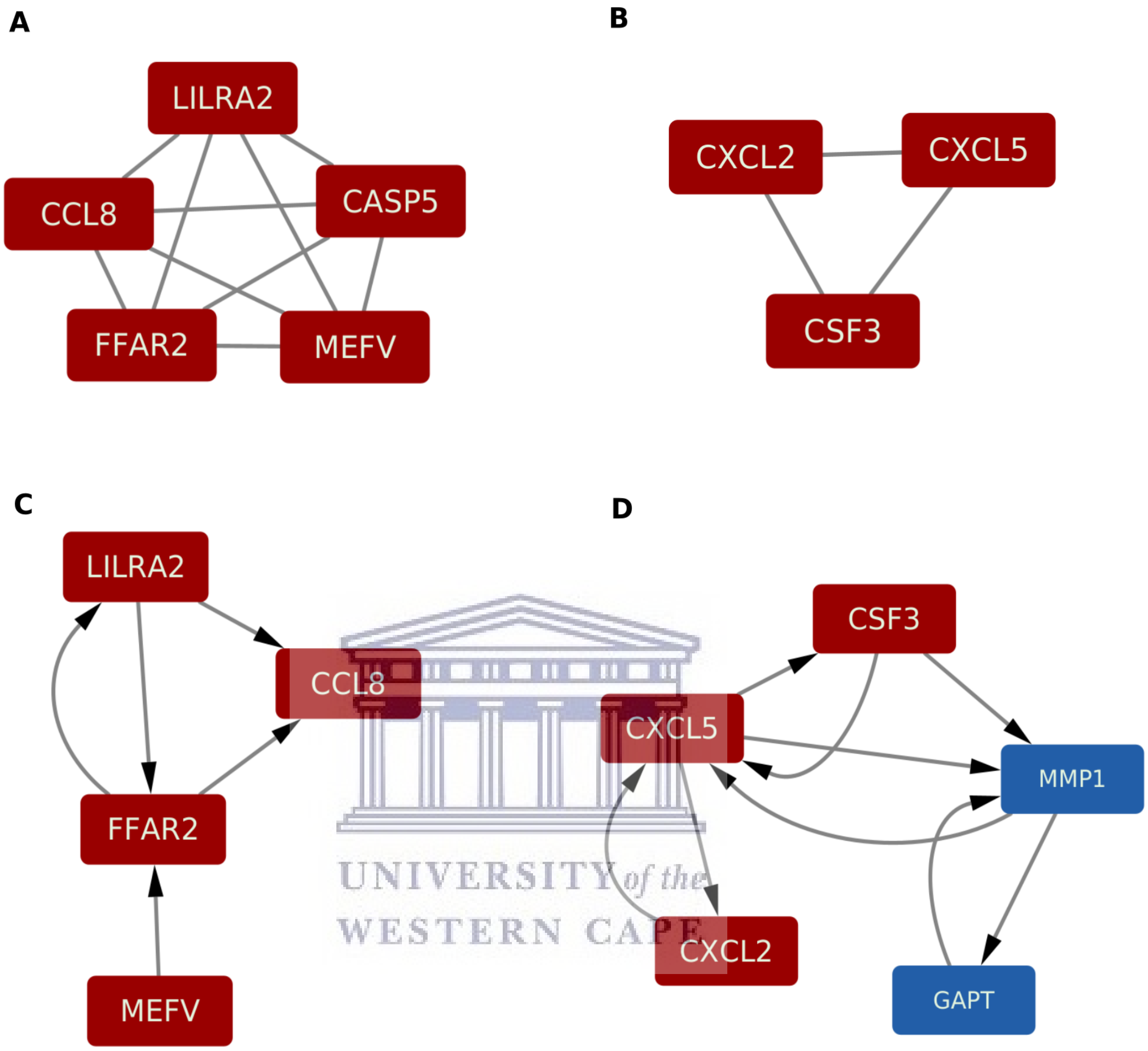
Stage IV, WGCNA identified a large network of 19 genes related to inflammation and immune response (Figure 4.5A). It is noteworthy that the majority of the genes constituting the WGCNA networks are mainly up-regulated genes, and the progression in clinical stages leads to an increase in the number of genes involved in inflammation and immune response.

4.4 Gene regulatory networks inferred using GENIE3

Similar to WGCNA, GENIE3 identified that DEGs related to inflammation and immune response mutually regulate one another. In Stage II, GENIE3 generated two sub-networks for inflammation and immune response composed of 4 and 5 DEGs, respectively (Figure 4.4C and D). In Stage IV, a larger network was identified consisting of 18 DEGs enriched in both inflammation and immune response (Figure 4.5B). Upon reconstruction of the GENIE3 networks, it was noted that many identified DEG connections were also reported by the WGCNA algorithm. Out of the 8 DEGs in WGCNA Stage II networks, 7 DEGs were present in the GENIE3 Stage II networks (Figure 4.4). In Stage IV, 12 DEGs were present in both WGCNA and GENIE3 networks (Figure 4.5).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/



UNIVERSITY of the
WESTERN CAPE

Figure 4.4: A&B represent the co-expression networks of Stage II DEGs related to inflammation and immune response generated using WGCNA; C&D represent the gene regulatory networks constructed using GENIE3. The red rectangles represent up-regulated genes while the blue rectangles represent down-regulated genes.

https://etd.uwc.ac.za/

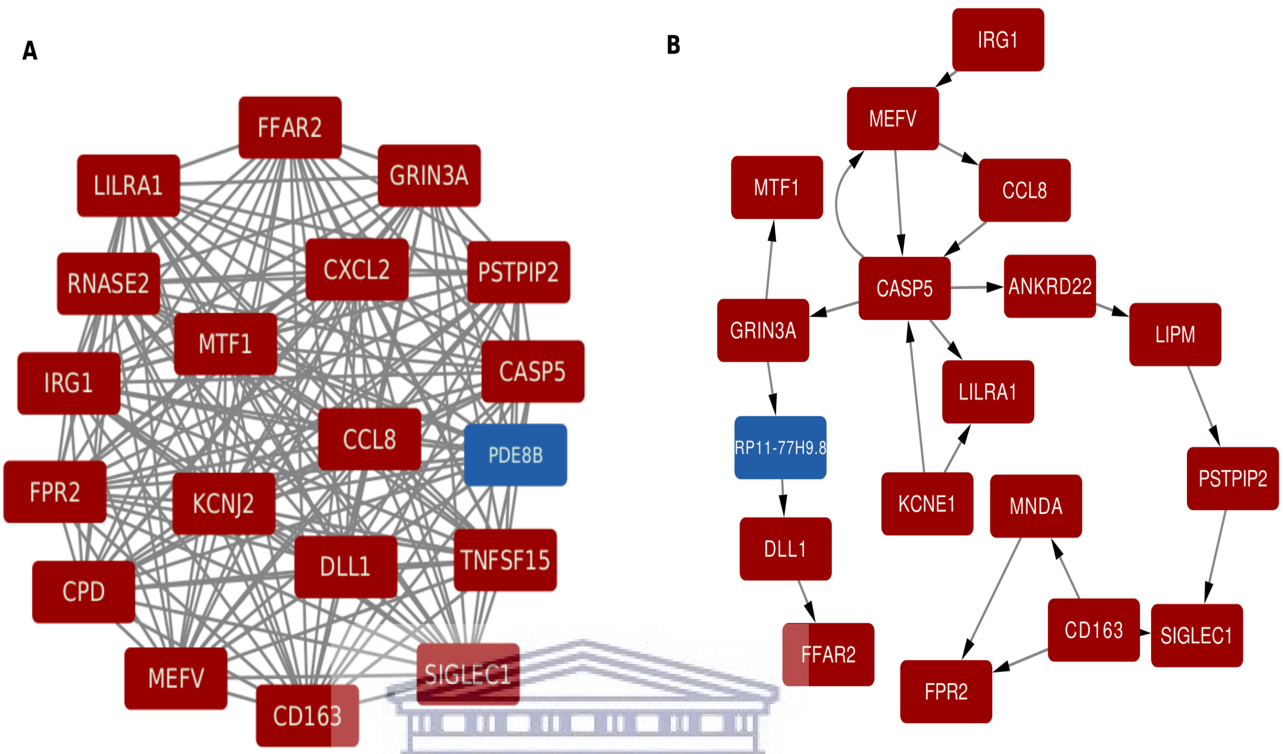


Figure 4.5: A represents the co-expression network of Stage IV DEGs related to inflammation and immune response constructed using WGCNA and B represents the gene regulatory network constructed using GENIE3. The red rectangles represent up-regulated genes, while the blue rectangles represent down-regulated genes.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

4.5 Inferential statistical and survival analyses

The statistical analyses were carried out to investigate variances in the mean/median expression of Stage II and IV GRNs using Stage I as reference. In all of the comparisons of equal variance across groups, the Kruskal-Wallis and ANOVA algorithms reported significant difference in expression between the clinical groups (Table 4.4). This was expected since the networks were composed of DEGs. Additionally, a statistical analysis was performed to verify whether the expression of the 25 common DEGs (including the immune and inflammatory DEGs) was maintained throughout cancer progression from Stage II to Stage IV (Stage IV vs II). The Levene's test for homogeneity of variances assumption was satisfied, *i.e.,* the null hypothesis was not rejected (p-value = 0.17). Therefore, the ANOVA test was used, giving a p-value of 0.48 ensuring equal means of expression of the 25 common DEGs between stages II and IV.

From the results of the ROC analyses, represented here by the AUC prediction probabilities, the cut-off values of the Youden index (Formula 2.2) were deduced for the immune and inflammatory DEGs and are shown in Table 4.5. Among the 5 Stage II DEGs classified by DAVID under inflammation (Table 4.2), 3 genes (*CXCL2, CXCL5* and *CCL8*) have an AUC > 0.7 indicating that their expression values have a strong prediction for clinical stages I and II. This number increased at Stage IV with 11 out of 19 genes (*CXCL2, CCL8, MEFV, NFKBIZ, TNFRSF10B, FPR2, PTX3, SIGLEC1, ZC3H12A, IL7*, and *TNFSF15*) having an AUC > 0.7 indicating a significant difference in gene expression between these clinical stages (Table 4.5).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Table 4.4: Statistical tests of equal expression variances in Stage II and IV networks relative to the reference stage (Stage I)

Data Groups	Statistical Tests	Conclusions
Stage II vs I (WGCNA)	1. Levene's Test: F value = 17.90 p-value = 3.56e-05 2. Kruskal-Wallis: $\chi 2$ = 26.46 p-value = 2.69e-07	The homogeneity of variance assumption across groups was rejected. Therefore, the test used for equal variances was Kruskal-Wallis algorithm. The null hypothesis which states that there is no difference in means was rejected. Therefore, the expression of the genes in the WGCNA Stage II GRNs was significantly different compared to Stage I.
Stage II vs I (GENIE3)	1. Levene's Test: F value = 2.16 p-value = 0.14 2. Anova: F value = 8.99 p-value = 0.003	The homogeneity of variance assumption across groups was accepted. Therefore, the test used for equal mean variances was ANOVA. The null hypothesis which states that there is no difference in means was rejected. Therefore, the expression of the genes in the GENIE3 Stage II GRNs was significantly different compared to Stage I.
Stage IV vs I (WGCNA)	1. Levene's Test: F value = 6.72 p-value = 0.01 2. Kruskal-Wallis: $\chi 2$ = 20.21 p-value = 6.94e-06	The homogeneity of variance assumption across groups was rejected. Therefore, the test used for equal variances was Kruskal-Wallis algorithm. The null hypothesis which states that there is no difference in means was rejected. Therefore, the expression of the genes in the WGCNA Stage IV GRNs was significantly different compared to Stage I.
Stage IV vs I (GENIE3)	1. Levene's Test: F value = 18.22 p-value = 2.35e-05 2. Kruskal-Wallis: $\chi 2$ = 10.04 p-value = 0.001	The homogeneity of variance assumption across groups was rejected. Therefore, the test used for equal variance was Kruskal-Wallis algorithm. The null hypothesis which states that there is no difference in means was rejected. Therefore, the expression of the genes in the GENIE3 Stage IV GRNs was significantly different compared to Stage I.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Table 4.5: Area under the curve (AUC) and Youden Index (YI) cut-off values of the immune and inflammatory DEGs in Stage II and IV of DLBCL. Values in bold represent models with AUC $>$ 0.7

	Genes	**AUC**	**YI Cut-off**
Stage II	CCL8	**0.779**	6.285
	CXCL2	**0.783**	6.00
	CXCL5	**0.732**	0.00
	IGHV4-34	0.614	12.378
	CSF3	0.676	6.700
Stage IV	*CXCL2*	**0.865**	5.492
	CCL8	**0.719**	6.833
	MEFV	**0.729**	5.954
	NFKBIZ	**0.781**	9.255
	NLRP4	0.214	0.0
	TNFRSF10B	**0.906**	11.037
	ALOX15	0.292	3.17
	HMGB1P1	0.198	2.0
	FPR2	**0.766**	8.109
	PTX3	**0.854**	4.524
	SIGLEC1	**0.75**	10.500
	ZC3H12A	**0.875**	11.046
	CLNK	0.328	1.585
	ENPP3	0.156	2.807
	IL7	**0.781**	10.151
	PKHD1L1	0.115	2.807
	TNFSF15	**0.854**	7.34
	IGHV2-70	0.271	1.00
	IGKV4-1	0.229	2.807

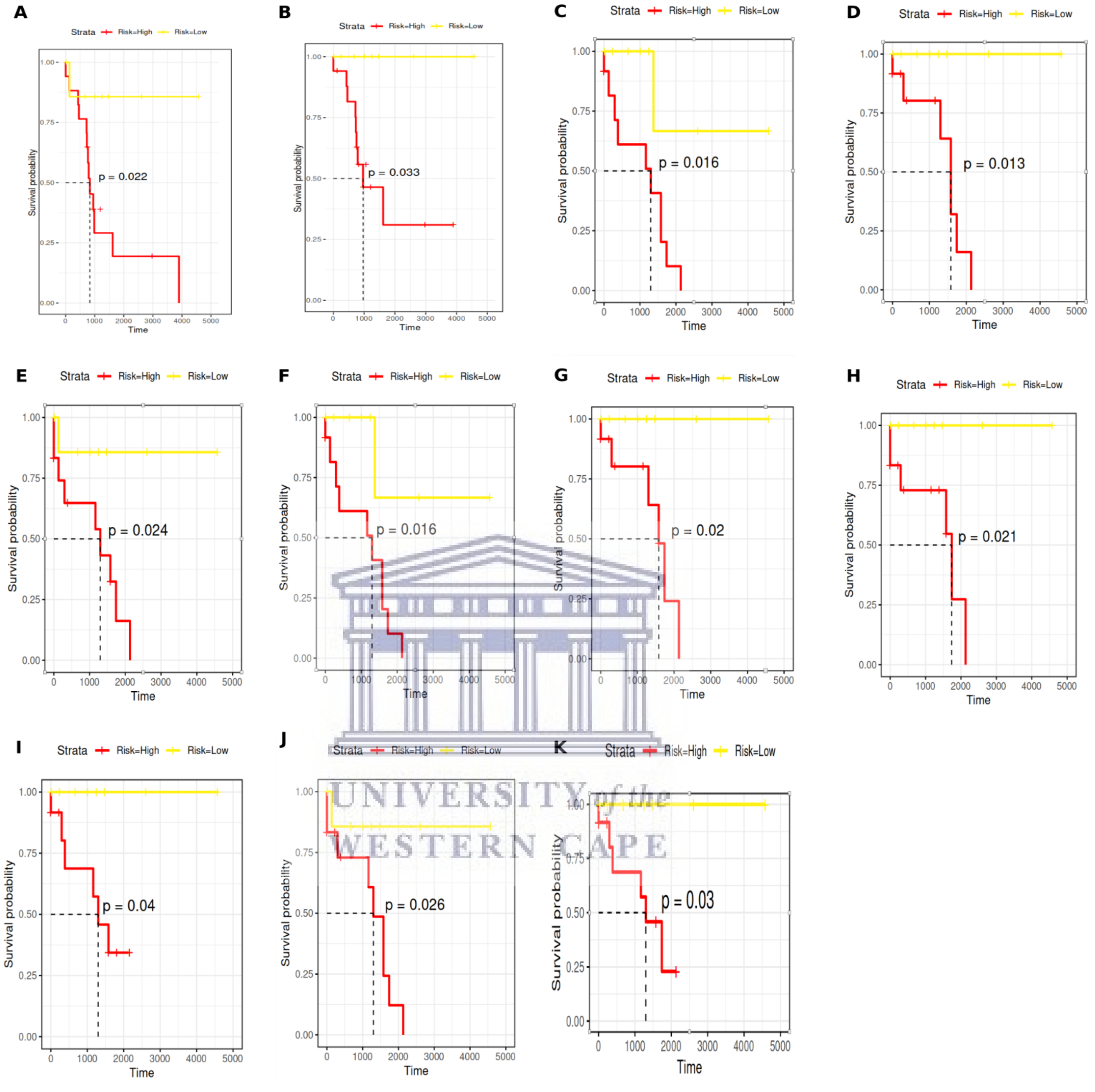
UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

K-M estimates were calculated for the genes with AUC $>$ 0.7 (Table 4.5) to determine whether the regulation of genes impacted patients' survival. Genes with log-rank test p-value $>$ 0.05 were not reported and their K-M plots are not shown (Figure 4.6). Patients with a high expression *CCL8* and *CXCL2* genes had shorter OS in DLBCL Stage II and I. Similarly, high expression of *CXCL2, MEFV, TNFRSF10B, FPR2, PTX3, SIGLEC1, ZCH312A, IL7,* and *TNFSF15* genes in Stage IV and I patients had poorer clinical outcomes (Figure 4.6). Although *CXCL5* and *NFKBIZ* have AUC $>$ 0.7 in Stage II and IV respectively, their high expression had no prognostic impact on OS (p-value $>$ 0.05).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/



UNIVERSITY of the
WESTERN CAPE

Figure 4.6: Kaplan-Meier (K-M) plots displaying the relationship between gene expression and patients' overall survival (OS). A and B represent K-M plots of *CCL8* and*CXCL2* expression relative to Stage II and I patients' OS. C-K represent K-M plots of *TNFSF15, MEFV, CXCL2, TNFRSF10B, FPR2, SIGLEC1, ZC3H12A, PTX3, IL7* expression relative to Stage IV and I patients' OS.

Chapter 5

Discussion

Integrated bioinformatics tools were used to analyze the DLBCL dataset to explore the hub genes and essential functions associated with DLBCL DEGs. Despite its clinical significance, the genes expression mechanism that contribute to inflammation and immune response in DLBCL is poorly understood. The findings showed 25 genes that are commonly expressed throughout Stage II and IV of this lymphoma (Table 4.1). Amongst these genes *CCL8, CASP5,* and *CXCL2* were commonly enriched in inflammation and immune response in both stages (Table 4.2 and 4.3).

UNIVERSITY of the
WESTERN CAPE

Caspases are important regulators of apoptosis, cell differentiation, inflammation, and innate immunity (Lan et al. 2009). The enzymes found in *CASP5* positively regulate inflammation (Widmann and Numbers 2007). There is growing evidence-supported belief that caspases contribute to lymphomagenesis. In this study, the gene ontology analysis mapped *CASP5* in inflammation-associated function. This is in agreement with a study done by Lan *et al.,* (2009) where caspases were found to play a crucial role in the regulation of apoptosis, inflammation, and innate immunity in several NHL, including DLBCL (Lan et al. 2009). Moreover, another study found *CASP5* to be among caspases that contribute to biological processes important in lymphomagenesis, such as promoting inflammation by playing a pivotal role

https://etd.uwc.ac.za/

in maturation and activation of pro-inflammatory cytokines such as inflammasome (Ghayur et al. 1996; Martinon et al. 2002).

The molecular heterogeneity of DLBCL poses a major obstacle to the accurate classification of DLBCL as well as the development of effective treatments for all DLBCL patients (Kuang et al. 2021). Three other studies have published reports on the association of the up-regulation of *CCL8* with immunoregulatory and inflammatory processes in DLBCL (Nakayama et al. 2021; Tamma et al. 2020; Xu-Monette et al. 2016). Based on these results, it may be inferred that even though DLBCL is a heterogeneous malignancy, *CCL8* forms part of inflammatory biomarkers in patients with DLBCL. However, the exact mechanisms of *CCL8* in inflammatory processes are yet to be reported in DLBCL.

The functional analysis showed that *CCL8, CXCL2, CXCL5, IL7, CSF3, TNFSF15*, and *TNFRSF10B* genes were mainly enriched in inflammation-related functions (Table 4.2). These genes are commonly referred to as inflammatory chemokines which are inflammatory prognostic biomarkers in cancer (Ansell et al. 2012; Hong et al. 2017). Inflammatory chemokines are small cytokines that play a role in recruiting immune cells such as TAM into a site of inflammation (Tamma et al. 2020). Chronic inflammation has been linked with sustained and excessive production of chemokines in advanced tumors (Pahwa and Jialal 2019). This may be the case in Stage IV since more genes are involved in inflammation, as seen in Figure 4.5. The high levels of chemokines and cell migration are key triggers for promoting the malignancy and metastasis of tumor cells (Mantovani et al. 2008). Interestingly, chronic inflammation has been associated with DLBCL in previous studies (Loong et al. 2010; Monti et al. 2005), which supports the results demonstrating chronic inflammation in this study. In the transformation of follicular lymphoma to aggressive DLBCL, nearly two-thirds

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

of the most discriminative genes were related to cellular immune response and inflammatory processes (Glas et al. 2005; Kiaii et al. 2013). Therefore, chemokines are potential targets for preventing cancer cell dissemination from primary tumors to the circulation. Another mechanism that has been observed to lead to tumor metastasis in different studies in oncology is the recruitment of TAM to TME by inflammatory chemokines (Argyle and Kitamura 2018). The up-regulation of these chemokines, as observed in this study, and the presence of immune cells such as TAM in TME have been demonstrated to be involved in tumor growth, invasion, metastasis, angiogenesis, and immunosuppression (Fridlender and Albelda 2012; Zhang et al. 2012).

The findings of this study are consistent with reports that associate high expression of *CSF3* in DLBCL with inflammatory response (Zhao et al. 2016). Saunders *et al.*, (2021) demonstrated that *CSF3* signaling is associated with changes in the immune infiltrate within the TME and has a strong correlation with gene signatures associated with pro-tumor immune responses such as *CXCL5*, *CXCL6*, and *CXCL8* in colon cancer (Saunders et al. 2021). Furthermore, Kuang *et al.*, (2021) reported *CSF3* as one of the 25 cachexia-inducing factors found in the DLBCL immune microenvironment that played a non-negligible role in immune regulation in the TME (Kuang et al. 2021). It becomes obvious that the high expression of *CSF3* plays an important role in the TME and may offer a potential therapeutic target and should be further explored.

Tumor necrosis is spontaneous cell death that is usually associated with necrosis-induced inflammation and controlled by the up-regulation of inflammatory chemokines such as *TNFSF15* and *TNFRSF10B* which were found to be enriched in inflammatory responses in this study (Table 4.2). These genes form the tumor necrosis factor (TNF) family and have been proven to possess both pro-tumor (via nuclear factor-B (NF-B) pathway) and anti-tumor properties (via activation of c-Jun N-terminal kinase

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

(JNK) pathway) (Wang and Lin 2008). However, a study has shown that TNFs act as pro-tumors in NHL in general, specifically in DLBCL (Song et al. 2017). TNFs act as a double-edged sword due to their ability to induce tumor necrosis to shrink the tumor which results in necrosis-induced inflammation leading to more aggressive patterns of DLBCL growth (Wang and Lin 2008). TNFs also promote tumor growth by stimulating proliferation, survival, migration, and angiogenesis in most cancer cells that are resistant to TNF-induced cytotoxicity (Zhou et al. 2018).

Among the genes that remained up-regulated throughout the stages was *FFAR2*. *FFAR2/FFA2*, also known as *GPR43*, are free fatty acid receptors that have been proven in numerous studies to have a role in a variety of physiological processes, including the regulation of inflammatory mediators (Bindels. et al. 2013). *FFAR2* down-regulation led to reduced colonic inflammation which usually leads to colon cancer (Bindels. et al. 2013). Therefore, the up-regulation of *FFAR2* in this study promotes inflammation in DLBCL. In addition, FFARs were found to be among inflammatory defining genes in a gene expression profiling of gray zone lymphoma (Sarkozy et al. 2020).

UNIVERSITY of the
WESTERN CAPE

KEGG enrichment analysis suggested the involvement of *CXCL2* in two immune-related pathways in Stage II and *CXCL5* in one (Table 4.3). Of these, *CXCL2* participates in nucleotide-binding oligomerization domain (NOD)-like receptor signaling pathway which has been associated with the production of inflammasomes and several cancer types with inflamed tissue (Castaño-Rodríguez et al. 2014). Activation of NOD-like receptor signaling leads to the production of numerous pro- and/or anti-inflammatory signals including interferons, tumor necrosis factors, and cytokines (Zhong et al. 2013). Another study by Zhou and co-workers also associated NOD-like receptor pathway with host response DLBCL (Zhou et al. 2020).

https://etd.uwc.ac.za/

The inferential statistical analyses showed the existence of variance in the mean/median expression of stage II and IV inflammatory GRNs compared to the reference group, Stage I. Therefore, inflammation is a hallmark of DLBCL that is likely maintained by the TME and kept active throughout the DLBCL stages leading to tumors with chronic inflammation. Additionally, the ANOVA test reported no significant difference in expression of the common DEGs, including DEGs related to inflammation and immune response, between Stage II and IV. That is, the high expression of the inflammation-associated DEGs is maintained throughout the stages, demonstrating possible chronic inflammation in this study. Furthermore, as shown in Figure 4.5, Stage IV involves more inflammation-associated DEGs, further validating the concept of chronic inflammation in DLBCL that begins at an early-stage and worsens with the disease progression. Targeting the set of early-stage connections (Figure 4.4) by knocking down the genes involved is essential to assess its impact on inflammation at the advanced stage.

UNIVERSITY of the
WESTERN CAPE

The ROC analyses reported that the expression of certain Stage II and IV inflammatory DEGs have good discriminating power ($AUC > 0.7$) with respect to the reference stage, Stage I. Moreover, the expression YI cut-off values inferred from these genes associated them with patients' OS. Selected cut-off values for chemokines (*CCL8, CXCL2,* and *CXCL5*) appear to have good discriminating power to differentiate between clinical stages. This was also true for the TNF family of genes (*TNFRSF10B* and *TNFSF15*) in Stage IV. In fact, in the advanced stage better discrimination due to higher AUC was observed for the TNF genes compared to other chemokines from early-stage, indicating a trend of new biomarkers controlling tumor growth.

https://etd.uwc.ac.za/

Survival analyses of the DEGs that were classified by DAVID to be related to inflammation showed that the expression of a larger proportion of the DEGs had an impact on the survival of DLBCL patients. The construction of the overall K-M plots revealed the association of many inflammatory DEGs with patient survival, and their elevated expression is linked to poor outcomes. For example, high expression of the chemokine *CCL8* has been associated with poor outcomes in patients with early-stage DLBCL (Figure 4.6) placing it as well as those interacting with it (Figure 4.4) as primary drivers of inflammation and targets for drug discovery.

The statistical analyses of this study emphasized the effects of TNF on the clinical outcomes of DLBCL patients. Several other studies have demonstrated that the overexpression of TNF negatively affects survival in both aggressive and indolent lymphoma (Pedersen et al. 2005; Pedersen and Sørensen 2003; Seymour et al. 2019; Warzocha et al. 1997). In this study, it was validated that high expression of *PTX3, MEFV, FPR2, SIGLEC1, ZC3H12A*, and *IL7* have been shown to have a significantly poor outcome in advanced-stage patients. A study previously showed that high Pentraxin 3 (*PTX3*) expression is associated with a poor prognosis in DLBCL (Carreras et al. 2022). To date, there have been no published work assessing the prognostic value of *MEFV, FPR2, SIGLEC1, ZC3H12A*, and *IL7* in DLBCL patients survival. Nevertheless, previous studies reported significant association up-regulation of *FPR2* with immune response, inflammation, and host defense (Xu-Monette et al. 2016), *IL7* with immune response (Charbonneau et al. 2012), and *ZC3H12A* with genetics and pathogenesis of DLBCL (Schmitz et al. 2018).

UNIVERSITY of the
WESTERN CAPE

The inflammatory response of the reported genes in a DLBCL cell could be due to stress from the microenvironment (Monti et al. 2005), and that inflammation can be harmful to cell life. Therefore, studying the microenvironment surrounding tumor

https://etd.uwc.ac.za/

cells is vital to understanding the causes of inflammation in DLBCL. Targeting the inflammatory genes can possibly turn off the expression of anti-apoptotic genes and allow DLBCL death.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 6

Limitations

The networks inferred in this study constituted of up-regulated genes, but the effect of the down-regulation of *SHANK1, CTSE, TSPAN1, BARX2,* and *C11orf53* as the pathogenesis of DLBCL progresses remains elusive and require further and deepened exploration.

It would have been more interesting to perform analyses on all DLBCL stages, however, in order for an analysis to be significant, WGCNA requires a total number of samples used between groups to be greater than 15. This was not the case when combining Stage I and Stage III in this study since Stage III had only 5 available samples. Therefore, Stage III network would simply be too noisy to be considered biologically significant, and hence it was not inferred and analyzed.

Even though this study proved DLBCL is associated with chronic inflammation, functional studies are required to validate these results in a larger DLBCL sample cohort.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 7

Conclusion

The purpose of this study, as discussed in section 1.3, was to reconstruct the inflammation-associated GRNs of DLBCL across clinical stages, and identify genes in these networks that are linked to the pathogenesis and progression of DLBCL. This was made possible by the use of RNA-Seq expression data and the application of different bioinformatics and statistical methods. This resulted in the discovery of network-centric genes that are responsible for inflammation and short survival in patients with DLBCL.

UNIVERSITY of the
WESTERN CAPE

This project demonstrated the existence of chronic inflammation in DLBCL, as well as increasing levels of inflammation as cancer progression continues, as evidenced by the increased size of the inflammatory network in the advanced stage. The exhibited networks revealed the involvement of chemokines such as *CXCL2, CCL8, IL7, TNFSF15*, and *TNFRSF10B* and genes from other families, such as caspases, in the pathogenesis and progression of DLBCL. These genes were also found to have a significant prognosis value on the overall survival of DLBCL patients. Further work can be done to test the effectiveness of DLBCL treatment strategies targeting the above-discussed genes.

References

J. S. Abramson and M. A. Shipp. Review in translational hematology Advances in the biology and therapy of diffuse large B-cell lymphoma: moving toward a molecularly targeted approach. *Blood*, 106(4):1164–1175, 2016.

H. Adams, A. Tzankov, A. Lugli, and I. Zlobec. New time-dependent approach to analyse the prognostic significance of immunohistochemical biomarkers in colon cancer and diffuse large b-cell lymphoma. *Journal of clinical pathology*, 62(11): 986–997, 2009.

L. Agnelli, M. Forcato, F. Ferrari, G. Tuana, K. Todoerti, B. A. Walker, G. J. Morgan, L. Lombardi, S. Bicciato, and A. Neri. The reconstruction of transcriptional networks reveals critical genes with implications for clinical outcome of multiple myeloma. *Clinical Cancer Research*, 17(23):7402–7412, 2011.

A. A. Alizadeh, M. B. Elsen, R. E. Davis, C. L. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marü, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Nature Precedings*, 11(10):1–1, 2010.

S. M. Ansell, M. J. Maurer, S. C. Ziesmer, S. L. Slager, T. M. Habermann, B. K. Link, T. E. Witzig, W. R. Macon, A. Dogan, J. R. Cerhan, and A. J. Novak. Elevated pretreatment serum levels of interferon-inducible protein-10 (CXCL10) predict disease relapse and prognosis in diffuse large B-cell lymphoma patients. *American Journal of Hematology*, 87(9):865–869, 2012.

D. Argyle and T. Kitamura. Targeting macrophage-recruiting chemokines as a novel therapeutic strategy to prevent the progression of solid tumors. *Frontiers in Immunology*, 9(NOV):1–15, 2018.

S. E. Arthur, A. Jiang, B. M. Grande, M. Alcaide, R. Cojocaru, C. K. Rushton, A. Mottok, L. K. Hilton, P. K. Lat, E. Y. Zhao, L. Culibrk, D. Ennishi, S. Jessa, L. Chong, N. Thomas, P. Pararajalingam, B. Meisnber, M. Boyle, J. Davidson, K. R. Bushell, D. Lai, P. Farinha, G. W. Slack, G. B. Morin, S. Shah, D. Sen, S. J. Jones, A. J. Mungall, R. D. Gascoyne, T. E. Audas, P. Unrau, M. A. Marra, J. M. Connors, C. Steidl, D. W. Scott, and R. D. Morin. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nature Communications*, 9(1), 2018.

P. Bellot, C. Olsen, P. Salembier, A. Oliveras-Vergés, and P. E. Meyer. NetBenchmark: A bioconductor package for reproducible benchmarks of gene regulatory network inference. *BMC Bioinformatics*, 16(1):1–15, 2015.

Bindels., E. M. Dewulf, and N. M. Delzenne. GPR43/FFA2: Physiopathological relevance and therapeutic prospects. *Trends in Pharmacological Sciences*, 34(4): 226–232, 2013.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

E. Bohers, S. Mareschal, P. Bertrand, P. J. Viailly, S. Dubois, C. Maingonnat, P. Ruminy, H. Tilly, and F. Jardin. Activating somatic mutations in diffuse large B-cell lymphomas: Lessons from next generation sequencing and key elements in the precision medicine era. *Leukemia and Lymphoma*, 56(5):1213–1222, 2015.

L. Breiman. Random Forests. *Machine Learning 2001 45:1*, 45(1):5–32, oct 2001.

R. Camicia, H. C. Winkler, and P. O. Hassa. Novel drug targets for personalized precision medicine in relapsed/refractory diffuse large b-cell lymphoma: a comprehensive review. *Molecular cancer*, 14(1):1–62, 2015.

J. Carreras, Y. Yukie, K. Shinichiro, H. Masashi, S. Tomita, H. Ikoma, A. Ito, Y. Kondo, J. Itoh, G. Roncador, A. Martinez, L. Colomo, R. Hamoudi, K. Ando, and N. Nakamura. High PTX3 expression is associated with a poor prognosis in diffuse large B-cell lymphoma. (June 2021):334–348, 2022.

N. Castaño-Rodríguez, N. O. Kaakoush, K. L. Goh, K. M. Fock, and H. M. Mitchell. The NOD-like receptor signalling pathway in Helicobacter pylori infection and related gastric cancer: A case-control study and gene expression analyses. *PLoS ONE*, 9 (6):117870, 2014.

B. Charbonneau, M. J. Maurer, S. M. Ansell, S. L. Slager, Z. S. Fredericksen, S. C. Ziesmer, W. R. Macon, T. M. Habermann, T. E. Witzig, and B. K. Link. Pretreatment circulating serum cytokines associated with follicular and diffuse large b-cell lymphoma: a clinic-based case-control study. *Cytokine*, 60(3):882–889, 2012.

A. Dabrowska-Iwanicka and J. A. Walewski. Primary mediastinal large B-cell lymphoma. *Current Hematologic Malignancy Reports*, 9(3):273–283, 2014.

R. E. Davis, V. N. Ngo, G. Lenz, P. Tolar, R. M. Young, P. B. Romesser, H. Kohlhammer, L. Lamy, H. Zhao, Y. Yang, W. Xu, A. L. Shaffer, G. Wright,

UNIVERSITY of the
WESTERN CAPE

W. Xiao, J. Powell, J. K. Jiang, C. J. Thomas, A. Rosenwald, G. Ott, H. K. Muller-Hermelink, R. D. Gascoyne, J. M. Connors, N. A. Johnson, L. M. Rimsza, E. Campo, E. S. Jaffe, W. H. Wilson, J. Delabie, E. B. Smeland, R. I. Fisher, R. M. Braziel, R. R. Tubbs, J. R. Cook, D. D. Weisenburger, W. C. Chan, S. K. Pierce, and L. M. Staudt. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature*, 463(7277):88–92, 2010.

G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for annotation, visualization, and integrated discovery. *Genome biology*, 4(5):1–11, 2003.

J. A. Didonato, F. Mercurio, and M. Karin. NF-κB and the link between inflammation and cancer. *Immunological Reviews*, 246(1):379–400, 2012.

M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, and J. Estelle. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

F. Emmert-Streib, R. d. M. Simoes, P. Mullan, B. Haibe-Kains, and M. Dehmer. The gene regulatory network for breast cancer: Integrated regulatory landscape of cancer hallmarks. *Frontiers in Genetics*, 5(FEB):1–12, 2014.

R. Eric Davis, K. D. Brown, U. Siebenlist, and L. M. Staudt. Constitutive nuclear factor κB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. *Journal of Experimental Medicine*, 194(12):1861–1874, 2001.

S. Feizi, D. Marbach, M. Médard, and M. Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnology*, 31(8): 726–733, 2013.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

R. Fisher. Theoretical foundations of mathematical statistics. *Phil. Trans. Roy. Soc. London, Series A*, 222:309–368, 1921.

R. Fluss, D. Faraggi, and B. Reiser. Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal*, 47(4):458–472, 2005.

L. Franchi, N. Warner, K. Viani, and G. Nuñez. Function of nod-like receptors in microbial recognition and host defense. *Immunological reviews*, 227(1):106–128, 2009.

Z. G. Fridlender and S. M. Albelda. Tumor-associated neutrophils: Friend or foe? *Carcinogenesis*, 33(5):949–955, 2012.

R. Gerard-Marchant, I. Hamlin, K. Lennert, F. Rilke, A. Stansfeld, and J. Von Unnik. A. m.: Letter to the editor. *Lancet*, pages 406–408, 1974.

T. Ghayur, S. Banerjee, M. Hugunin, D. Butler, L. Herzog, A. Carter, L. Quintal, L. Sekut, R. Talanian, M. Paskind, W. Wong, R. Kamen, D. Tracey, and H. Allen. Caspase-1 processes ifn--inducing factor and regulates lps-induced ifn- production. *letters to nature 25.*, 389(983):619–623, 1996.

A. M. Glas, M. J. Kersten, L. J. Delahaye, A. T. Witteveen, R. E. Kibbelaar, A. Velds, L. F. Wessels, P. Joosten, R. M. Kerkhoven, R. Bernards, J. H. Van Krieken, P. M. Kluin, L. J. Van'T Veer, and D. De Jong. Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. *Blood*, 105(1):301–307, 2005.

M. K. Goel, P. Khanna, and J. Kishore. Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*, 1(4):274, 2010.

M. J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A. N. Brooks, J. Zhu, and D. Haussler. Visualizing and

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38 (6):675–678, 2020.

E. Greenfest-Allen, J.-P. Cartailler, M. A. Magnuson, and C. J. Stoeckert. iterativewgcna: iterative refinement to improve module detection from wgcna co-expression networks. *bioRxiv*, page 234062, 2017.

M. Greiner, D. Pfeiffer, and R. Smith. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive veterinary medicine*, 45(1-2):23–41, 2000.

B. Győrffy. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Computational and Structural Biotechnology Journal*, 19:4101–4109, 2021.

Y. Han, S. Gao, K. Muegge, W. Zhang, and B. Zhou. Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*, 9:29–46, 2015.

Y. Han, J. Yang, P. Liu, X. He, C. Zhang, S. Zhou, L. Zhou, Y. Qin, Y. Song, Y. Sun, and Y. Shi. Prognostic nomogram for overall survival in patients with diffuse large B-cell lymphoma. *The Oncologist*, 24(11), 2019.

T. J. Hardcastle and K. A. Kelly. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(August), 2010.

N. L. Harris, E. S. Jaffe, H. Stein, P. M. Banks, J. K. Chan, M. L. Cleary, G. Delsol, B. Falini, and K. Gatter. A revised european-american classification of lymphoid neoplasms: a proposal from the international lymphoma study group. *blood*, 84(5): 1361–1392, 1994.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

N. L. Harris, E. S. Jaffe, J. Diebold, G. Flandrin, H. K. Muller-Hermelink, J. Vardiman, T. A. Lister, and C. D. Bloomfield. The world health organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the clinical advisory committee meeting, airlie house, virginia, november, 1997. *Annals of Oncology*, 10(12):1419–1432, 1999.

M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems*, 96 (1):86–103, 2009.

J. Y. Hong, K. J. Ryu, J. Y. Lee, C. Park, Y. H. Ko, W. S. Kim, and S. J. Kim. Serum level of CXCL10 is associated with inflammatory prognostic biomarkers in patients with diffuse large B-cell lymphoma. *Hematological Oncology*, 35(4):480–486, 2017.

D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(2):169–175, 2007.

P. Huang, J. Han, and L. Hui. MAPK signaling in inflammation-associated cancer development. *Protein and Cell*, 1(3):218–226, 2010.

Y. Huang, S. Ye, Y. Cao, Z. Li, J. Huang, H. Huang, M. Cai, R. Luo, and T. Lin. Outcome of R-CHOP or CHOP regimen for germinal center and nongerminal center subtypes of diffuse large B-cell lymphoma of chinese patients. *The Scientific World Journal*, 2012(December 2013), 2012.

V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

networks from expression data using tree-based methods. *PLoS ONE*, 5(9):1–10, 2010.

Z. Ji, L. He, A. Regev, and K. Struhl. Inflammatory regulatory network mediated by the joint action of NF-kB, STAT3, and AP-1 factors is involved in many human cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 116(19):9453–9462, 2019.

A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):1–19, 2017.

B. Kaminska. Mapk signalling pathways as molecular targets for anti-inflammatory therapy—from molecular mechanisms to therapeutic benefits. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1754(1-2):253–262, 2005.

M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The kegg databases at genomenet. *Nucleic acids research*, 30(1):42–46, 2002.

M. Karin. Nuclear factor-κB in cancer development and progression. *Nature*, 441 (7092):431–436, 2006.

A. Khan, G. Saha, and R. K. Pal. Modified half-system based method for reverse engineering of gene regulatory networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(4):1303–1316, 2020.

S. Kiaii, A. J. Clear, A. G. Ramsay, D. Davies, A. Sangaralingam, A. Lee, M. Calaminici, D. S. Neuberg, and J. G. Gribben. Follicular lymphoma cells induce changes in T-cell gene expression and function: Potential impact on survival and risk of transformation. *Journal of Clinical Oncology*, 31(21):2654–2661, 2013.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

S. Kimura, K. Ide, A. Kashihara, M. Kano, M. Hatakeyama, R. Masui, N. Nakagawa, S. Yokoyama, S. Kuramitsu, and A. Konagaya. Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics*, 21 (7):1154–1163, 2005.

R. Kridel, L. H. Sehn, and R. D. Gascoyne. Review series Pathogenesis of follicular lymphoma. *the Journal of Clinical Investigation*, 122(10):3424–3431, 2012.

W. H. Kruskal and W. A. Wallis. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

Z. Kuang, X. Li, R. Liu, S. Chen, and J. Tu. Comprehensive Characterization of Cachexia-Inducing Factors in Diffuse Large B-Cell Lymphoma Reveals a Molecular Subtype and a Prognosis-Related Signature. *Frontiers in Cell and Developmental Biology*, 9(May):1–14, 2021.

K. R. Kukurba and S. B. Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):951–969, 2015.

M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(1):90–97, 2016.

A. T. Kwon, H. H. Hoos, and R. Ng. Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, 19(8):905–912, 2003.

Q. Lan, L. M. Morton, B. Armstrong, P. Hartge, I. Menashe, T. Zheng, M. P. Purdue, J. R. Cerhan, Y. Zhang, A. Grulich, W. Cozen, M. Yeager, T. R. Holford, C. M. Vajdic, S. Davis, B. Leaderer, A. Kricker, M. Schenk, S. H. Zahm, N. Chatterjee,

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

S. J. Chanock, N. Rothman, and S. S. Wang. Genetic variation in caspase genes and risk of non-Hodgkin lymphoma: A pooled analysis of 3 population-based case-control studies. *Blood*, 114(2):264–267, 2009.

P. Langfelder and S. Horvath. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 2008.

P. Langfelder and S. Horvath. 2.a Automatic network construction and module detection. *Genetics*, pages 1–5, 2009.

P. Langfelder and S. Horvath. WGCNA package FAQ. pages 1–20, 2017.

G. Lenz, G. W. Wright, N. C. Emre, H. Kohlhammer, S. S. Dave, R. E. Davis, S. Carty, L. T. Lam, A. L. Shaffer, W. Xiao, J. Powell, A. Rosenwald, G. Ott, H. K. Muller-Hermelink, R. D. Gascoyne, J. M. Connors, E. Campo, E. S. Jaffe, J. Delabie, E. B. Smeland, L. M. Rimsza, R. I. Fisher, D. D. Weisenburger, W. C. Chan, and L. M. Staudt. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13520–13525, 2008.

H. Levene. Robust tests for equality of variances. *Korean journal of radiology*, 1(1): 278–292, 1960.

J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

S. Li, K. H. Young, and L. J. Medeiros. Diffuse large B-cell lymphoma. *Pathology*, 50(1):74–87, 2018.

T. A. Lister, D. Crowther, S. B. Sutcliffe, E. Glatstein, G. P. Canellos, R. C. Young,

UNIVERSITY of the
WESTERN CAPE

S. A. Rosenberg, C. A. Coltman, and M. Tubiana. Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. *Journal of Clinical Oncology*, 7(11):1630–1636, 1989.

Z. Liu, J. Meng, X. Li, F. Zhu, T. Liu, G. Wu, and L. Zhang. Identification of hub genes and key pathways associated with two subtypes of diffuse large B-cell lymphoma based on gene expression profiling via integrated bioinformatics. *BioMed Research International*, 2018, 2018.

Z.-P. Liu. Towards precise reconstruction of gene regulatory networks by data integration. *Quantitative Biology*, 6(2):113–128, 2018.

F. Loong, A. C. Chan, B. Ho, Y.-P. Chau, H.-Y. Lee, W. Cheuk, W.-K. Yuen, W.-S. Ng, H.-L. Cheung, and J. K. Chan. Diffuse large b-cell lymphoma associated with chronic inflammation as an incidental finding and new clinical scenarios. *Modern Pathology*, 23(4):493–501, 2010.

S. Lotia, J. Montojo, Y. Dong, G. D. Bader, and A. R. Pico. Cytoscape app store. *Bioinformatics*, 29(10):1350–1351, 2013.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

X. Lu, Y. Deng, L. Huang, B. Feng, and B. Liao. A co-expression modules based gene selection for cancer recognition. *Journal of Theoretical Biology*, 362:75–82, 2014.

R. Lukes and R. Collins. A functional approach to the classification of malignant lymphoma. In *Diagnosis and therapy of malignant lymphoma*, pages 18–30. Springer, 1974.

L. T. MacNeil and A. J. Walhout. Gene regulatory networks and the role of robustness

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

and stochasticity in the control of gene expression. *Genome research*, 21(5):645–657, 2011.

S. R. Maetschke, P. B. Madhamshettiwar, M. J. Davis, and M. A. Ragan. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, 15(2):195–211, 2014.

J. N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

A. Mantovani, P. Allavena, A. Sica, and F. Balkwill. Cancer-related inflammation. *Nature*, 454(7203):436–444, 2008.

J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

M. Martelli, A. J. Ferreri, C. Agostinelli, A. Di Rocco, M. Pfreundschuh, and S. A. Pileri. Diffuse large B-cell lymphoma. *Critical Reviews in Oncology/Hematology*, 87 (2):146–171, 2013.

F. Martinon, K. Burns, and J. Tschopp. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proil-β. *Molecular cell*, 10(2):417–426, 2002.

K. Miyazaki, M. Yamaguchi, M. Suguro, W. Choi, Y. Ji, L. Xiao, W. Zhang, S. Ogawa, N. Katayama, H. Shiku, and T. Kobayashi. Gene expression profiling of diffuse large B-cell lymphoma supervised by CD21 expression. *British Journal of Haematology*, 142(4):562–570, 2008.

S. Monti, K. J. Savage, J. L. Kutok, F. Feuerhake, P. Kurtin, M. Mihm, B. Wu,

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

L. Pasqualucci, D. Neuberg, R. C. Aguiar, P. D. Cin, C. Ladd, G. S. Pinkus, G. Salles, N. L. Harris, R. Dalla-Favera, T. M. Habermann, J. C. Aster, T. R. Golub, and M. A. Shipp. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105(5):1851–1861, 2005.

R. D. Morin and R. D. Gascoyne. Newly identified mechanisms in B-cell non-hodgkin lymphomas uncovered by next-generation sequencing. *Seminars in Hematology*, 50 (4):303–313, 2013.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.

Y. Nakayama, H. Chiu, R. K. Narla, A. Shakya, J. Gamez, I. Wall, D. Papazoglou, B. Apollonio, A. G. Ramsay, P. Hagner, and A. K. Gandhi. Differential effects of iberdomide versus revlimid on leukocyte trafficking, immune activation and DLBCL tumor cell killing. *Blood*, 138(Supplement 1):718, nov 2021.

NHL-Pathological-Classification-Project. National cancer institute sponsored study of classifications of non-hodgkin's lymphomas: summary and description of a working formulation for clinical usage. *Cancer*, 49(10):2112–2135, 1982.

N. A. Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8, 2003.

H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34, 1999.

N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski. Gene

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

regulatory network inference using fused LASSO on multiple data sets. *Scientific Reports*, 6:1–14, 2016.

R. Pahwa and I. Jialal. Chronic inflammation—statpearls—ncbi bookshelf. *Stat Pearls*, 1, 2019.

S. H. Park, J. M. Goo, and C.-H. Jo. Receiver operating characteristic (roc) curve: practical review for radiologists. *Korean journal of radiology*, 5(1):11–18, 2004.

L. Pasqualucci, D. Dominguez-sola, A. Chiarenza, G. Fabbri, A. Grunn, V. Trifonov, L. H. Kasper, S. Lerach, H. Tang, J. Ma, D. Rossi, A. Chadburn, V. V. Murty, C. G. Mullighan, G. Gaidano, R. Rabadan, P. K. Brindle, and R. Dalla-favera. Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature*, 471 (7337):189–195, 2011.

L. M. Pedersen and P. G. Sørensen. Mediators of inflammation correlate with microalbuminuria in patients with non-Hodgkin's lymphoma. *British Journal of Haematology*, 121(2):275–279, 2003.

L. M. Pedersen, G. W. Jürgensen, and H. E. Johnsen. Serum levels of inflammatory cytokines at diagnosis correlate to the bcl-6 and CD10 defined germinal centre (GC) phenotype and bcl-2 expression in patients with diffuse large B-cell lymphoma. *British Journal of Haematology*, 128(6):813–819, 2005.

T. C. Pina, I. T. Zapata, J. B. López, J. L. Pérez, P. P. Paricio, and P. M. Hernández. Tumor markers in lung cancer: does the method of obtaining the cut-off point and reference population influence diagnostic yield? *Clinical biochemistry*, 32(6):467–472, 1999.

F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason,

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

N. D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9), 2013.

H. Rappaport. Tumors of the hematopoietic system. *In Atlas of Tumor Pathology. Washington, DC: Armed Forces Institute of Pathology*, 3, 1966.

A. Reddy, J. Zhang, N. S. Davis, A. B. Moffitt, C. L. Love, A. Waldrop, S. Leppa, A. Pasanen, L. Meriranta, M. L. Karjalainen-Lindsberg, P. Nørgaard, M. Pedersen, A. O. Gang, E. Høgdall, T. B. Heavican, W. Lone, J. Iqbal, Q. Qin, G. Li, S. Y. Kim, J. Healy, K. L. Richards, Y. Fedoriw, L. Bernal-Mizrachi, J. L. Koff, A. D. Staton, C. R. Flowers, O. Paltiel, N. Goldschmidt, M. Calaminici, A. Clear, J. Gribben, E. Nguyen, M. B. Czader, S. L. Ondrejka, A. Collie, E. D. Hsi, E. Tse, R. K. Au-Yeung, Y. L. Kwong, G. Srivastava, W. W. Choi, A. M. Evens, M. Pilichowska, M. Sengar, N. Reddy, S. Li, A. Chadburn, L. I. Gordon, E. S. Jaffe, S. Levy, R. Rempel, T. Tzeng, L. E. Happ, T. Dave, D. Rajagopalan, J. Datta, D. B. Dunson, and S. S. Dave. Genetic and functional drivers of diffuse large B cell lymphoma. *Cell*, 171(2):481–494, 2017.

J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(2):193–200, 2007.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

S. Rosenberg. National-cancer-institute sponsored study of classifications of non-hodgkins

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

lymphomas-summary and description of a working formulation for clinical usage. *Cancer*, 49(10):2112–2135, 1982.

A. Rosenwald, G. Wright, K. Leroy, X. Yu, P. Gaulard, R. D. Gascoyne, W. C. Chan, T. Zhao, C. Haioun, T. C. Greiner, D. D. Weisenburger, J. C. Lynch, J. Vose, J. O. Armitage, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, E. Campo, E. Montserrat, A. Lopez-Guillermo, G. Ott, H. K. Muller-Hermelink, J. M. Connors, R. Braziel, T. M. Grogan, R. I. Fisher, T. P. Miller, M. LeBlanc, M. Chiorazzi, H. Zhao, L. Yang, J. Powell, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, and L. M. Staudt. Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma. *Journal of Experimental Medicine*, 198(6):851–862, 2003.

C. Ruiduo, D. Ying, and W. Qiwei. CXCL9 promotes the progression of diffuse large B-cell lymphoma through up-regulating β-catenin. *Biomedicine and Pharmacotherapy*, 107(April):689–695, 2018.

J. Ruyssinck, V. A. Huynh-Thu, P. Geurts, T. Dhaene, P. Demeester, and Y. Saeys. NIMEFI: Gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS ONE*, 9(3):1–13, 2014.

R. Saito, M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, S. Lotia, A. R. Pico, G. D. Bader, and T. Ideker. A travel guide to cytoscape plugins. *Nature methods*, 9(11):1069–1076, 2012.

C. Sarkozy, L. Chong, K. Takata, E. A. Chavez, T. Miyata-Takata, G. Duns, A. Telenius, M. Boyle, G. W. Slack, C. Laurent, P. Farinha, T. J. Molina, C. Copie-Bergman, D. Damotte, G. A. Salles, A. Mottok, K. J. Savage, D. W. Scott, A. Traverse-Glehen, and C. Steidl. Gene expression profiling of gray zone lymphoma. *Blood Advances*, 4(11):2523–2535, 2020.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

A. S. Saunders, D. E. Bender, A. L. Ray, X. Wu, and K. T. Morris. Colony-stimulating factor 3 signaling in colon and rectal cancers: Immune response and CMS classification in TCGA data. *PLoS ONE*, 16(2 February):1–21, 2021.

K. J. Savage, S. Monti, J. L. Kutok, G. Cattoretti, D. Neuberg, L. De Leval, P. Kurtin, P. Dal Cin, C. Ladd, F. Feuerhake, R. C. Aguiar, S. Li, G. Salles, F. Berger, W. Jing, G. S. Pinkus, T. Habermann, R. Dalla-Favera, N. L. Harris, J. C. Aster, T. R. Golub, and M. A. Shipp. The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma. *Blood*, 102(12):3871–3879, 2003.

R. Schmitz, G. W. Wright, D. W. Huang, C. A. Johnson, J. D. Phelan, J. Q. Wang, S. Roulland, M. Kasbekar, R. M. Young, A. L. Shaffer, D. J. Hodson, W. Xiao, X. Yu, Y. Yang, H. Zhao, W. Xu, X. Liu, B. Zhou, W. Du, W. C. Chan, E. S. Jaffe, R. D. Gascoyne, J. M. Connors, E. Campo, A. Lopez-Guillermo, A. Rosenwald, G. Ott, J. Delabie, L. M. Rimsza, K. Tay Kuang Wei, A. D. Zelenetz, J. P. Leonard, N. L. Bartlett, B. Tran, J. Shetty, Y. Zhao, D. R. Soppet, S. Pittaluga, W. H. Wilson, and L. M. Staudt. Genetics and pathogenesis of diffuse large B-cell lymphoma. *New England Journal of Medicine*, 378(15):1396–1407, 2018.

D. W. Scott, K. L. Mungall, S. Ben-Neriah, S. Rogic, R. D. Morin, G. W. Slack, K. L. Tan, F. C. Chan, R. S. Lim, J. M. Connors, M. A. Marra, A. J. Mungall, C. Steidl, and R. D. Gascoyne. TBL1XR1/TP63: A novel recurrent gene fusion in B-cell non-Hodgkin lymphoma. *Blood*, 119(21):4949–4952, 2012.

B. J. F. Seymour, M. Talpaz, F. Cabanillas, M. Wetzler, and R. Kurzrock. INTERLEUKIN-6. 13(3):575–582, 2019.

K. H. Shain and J. Tao. The B-cell receptor orchestrates environment-mediated

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

lymphoma survival and drug resistance in B-cell malignancies. *Oncogene*, 33(32): 4107–4113, 2014.

K. H. Shain, W. S. Dalton, and J. Tao. The tumor microenvironment shapes hallmarks of mature B-cell malignancies. *Oncogene*, 34(36):4673–4682, 2015.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3): 431–432, 2011.

M. K. Song, J. S. Chung, D. Y. Shin, S. N. Lim, G. won Lee, J. C. Choi, W. Y. Park, and S. Y. Oh. Tumor necrosis could reflect advanced disease status in patients with diffuse large B cell lymphoma treated with R-CHOP therapy. *Annals of Hematology*, 96(1):17–23, 2017.

L. M. Staudt and S. Dave. The biology of human lymphoid malignancies revealed by gene expression profiling. *Advances in Immunology*, 87(05):163–208, 2005.

V. S. Stel, F. W. Dekker, G. Tripepi, C. Zoccali, and K. J. Jager. Survival analysis I: The Kaplan-Meier method. *Nephron - Clinical Practice*, 119(1):83–88, 2011.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

S. H. Swerdlow, E. Campo, S. A. Pileri, N. Lee Harris, H. Stein, R. Siebert, R. Advani, M. Ghielmini, G. A. Salles, A. D. Zelenetz, and E. S. Jaffe. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, 127(20):2375–2390, 2016.

R. Tamma, G. Ranieri, G. Ingravallo, T. Annese, A. Oranger, F. Gaudio, P. Musto, G. Specchia, and D. Ribatti. Inflammatory cells in diffuse large B-cell lymphoma. *Journal of Clinical Medicine*, 9(8):2418, 2020.

M. A. Van De Wiel, G. G. Leday, L. Pardo, H. Rue, A. W. Van Der Vaart, and W. N. Van Wieringen. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128, 2013.

H. Varet, L. Brillet-Guéguen, J. Y. Coppée, and M. A. Dillies. SARTools: A DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PLoS ONE*, 11(6):1–8, 2016.

E. F. Wagner and Á. R. Nebreda. Signal integration by JNK and p38 MAPK pathways in cancer development. *Nature Reviews Cancer*, 9(8):537–549, 2009.

J. Wang, S. Vasaikar, Z. Shi, M. Greer, and B. Zhang. WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, 45(1):130–137, 2017.

J. F. Wang, Z. Y. He, S. H. Huang, H. Chen, W. Z. Wang, and F. Pourpanah. Fuzzy measure with regularization for gene selection and cancer prediction. *International Journal of Machine Learning and Cybernetics*, 12(8):2389–2405, 2021.

X. Wang and Y. Lin. Tumor necrosis factor and cancer, buddies or foes? *Acta Pharmacologica Sinica*, 29(11):1275–1288, 2008.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Y. Wang, L. Chen, G. Wang, S. Cheng, K. Qian, X. Liu, C. L. Wu, Y. Xiao, and X. Wang. Fifteen hub genes associated with progression and prognosis of clear cell renal cell carcinoma identified by coexpression analysis. *Journal of Cellular Physiology*, 234(7):10225–10237, 2019.

K. Warzocha, G. Salles, J. Bienvenu, Y. Bastion, C. Dumontet, N. Renard, E. M. Neidhardt-Berard, and B. Coiffier. Tumor necrosis factor ligand-receptor system can predict treatment outcome in lymphoma patients. *Journal of Clinical Oncology*, 15 (2):499–508, 1997.

H. L. Weiss, S. Niwas, W. E. Grizzle, and C. Piyathilake. Receiver operating characteristic (ROC) to determine cut-off points of biomarkers in lung cancer patients. *Disease Markers*, 19(6):273–278, 2003.

C. Widmann and C. Numbers. Caspases. 35:9–11, 2007.

G. Wright, B. Tan, A. Rosenwald, E. H. Hurt, A. Wiestner, and L. M. Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9991–9996, 2003.

C. Wrzodek, A. Dräger, and A. Zell. Keggtranslator: visualizing and converting the kegg pathway database to various formats. *Bioinformatics*, 27(16):2314–2315, 2011.

H. Wu, C. Wang, and Z. Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243, 2013.

J. Xiao, X. Wang, and H. Bai. Clinical features and prognostic impact of coexpression modules constructed by WGCNA for diffuse large B-cell lymphoma. *BioMed Research International*, 2020, 2020.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

X. B. Xiao, Y. Gu, D. L. Sun, L. Y. Ding, X. G. Yuan, H. W. Jiang, and Z. X. Wu. Effect of rituximab combined with chemotherapy on the expression of serum exosome miR-451a in patients with diffuse large b-cell lymphoma. *European Review for Medical and Pharmacological Sciences*, 23(4):1620–1625, 2019.

C. Xie, X. Mao, J. Huang, Y. Ding, J. Wu, S. Dong, L. Kong, G. Gao, C. Y. Li, and L. Wei. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, 39(SUPPL. 2):316–322, 2011.

Z. Y. Xu-Monette, L. Li, J. C. Byrd, K. J. Jabbar, G. C. Manyam, C. M. De Winde, M. Van Den Brand, A. Tzankov, C. Visco, J. Wang, K. Dybkaer, A. Chiu, A. Orazi, Y. Zu, G. Bhagat, K. L. Richards, E. D. Hsi, W. W. Choi, J. Huh, M. Ponzoni, A. J. Ferreri, M. B. Møller, B. M. Parsons, J. N. Winter, M. Wang, F. B. Hagemeister, M. A. Piris, J. H. Van Krieken, L. J. Medeiros, Y. Li, A. B. Van Spriel, and K. H. Young. Assessment of CD37 B-cell antigen and cell of origin significantly improves risk prediction in diffuse large B-cell lymphoma. *Blood*, 128 (26):3083–3100, 2016.

E. W. Yang, T. Girke, and T. Jiang. Differential gene expression analysis using coexpression and RNA-Seq data. *Bioinformatics*, 29(17):2153–2161, 2013.

S. Yepes, M. M. Torres, and L. López-Kleine. Regulatory network reconstruction reveals genes with prognostic value for chronic lymphocytic leukemia. *BMC Genomics*, 16 (1):1–12, 2015.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

M. Zamani-Ahmadmahmudi, A. Najafi, and S. M. Nassiri. Reconstruction of canine diffuse large B-cell lymphoma gene regulatory network: Detection of functional modules and hub genes. *Journal of Comparative Pathology*, 152(2-3):119–130, 2015.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

L. Zhang, J. Yang, J. Qian, H. Li, J. E. Romaguera, L. W. Kwak, M. Wang, and Q. Yi. Role of the microenvironment in mantle cell lymphoma: IL-6 is an important survival factor for the tumor cells. *Blood*, 120(18):3783–3792, 2012.

S. Zhao, N. Bai, J. Cui, R. Xiang, and N. Li. Prediction of survival of diffuse large B-cell lymphoma patients via the expression of three inflammatory genes. *Cancer Medicine*, 5(8):1950–1961, 2016.

Y. Zhong, A. Kinio, and M. Saleh. *Functions of NOD-Like Receptors in Human Diseases*, volume 4. 2013. ISBN 5143982065.

L. Zhou, L. Ding, Y. Gong, J. Zhao, G. Xin, R. Zhou, and W. Zhang. Identification of hub genes associated with the pathogenesis of diffuse large B-cell lymphoma subtype one characterized by host response via integrated bioinformatic analyses. *PeerJ*, 8, 2020.

X. A. Zhou, A. Louissaint, A. Wenzel, J. Yang, M. E. Martinez-Escala, A. P. Moy, E. A. Morgan, C. N. Paxton, B. Hong, E. F. Andersen, J. Guitart, A. Behdad, L. Cerroni, D. M. Weinstock, and J. Choi. Genomic Analyses Identify Recurrent Alterations in Immune Evasion Genes in Diffuse Large B-Cell Lymphoma, Leg Type. *Journal of Investigative Dermatology*, 138(11):2365–2376, 2018.

Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, 10(1), 2019.

Y. H. Zhou, K. Xia, and F. A. Wright. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19):2672–2678, 2011.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

APPENDIX A

Differentially expressed genes between DLBCL Stage II and I

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Table A.1: Differentially expressed genes between DLBCL Stage II and I

ensembl	gene	p-adjusted value	LFC
ENSG00000043039	BARX2	0.00125857448277	-3.65171682963786
ENSG00000049192	ADAMTS6	0.007414940545398	2.36437088606225
ENSG00000064655	EYA2	0.000270365339038	-3.89105223724715
ENSG00000079393	DUSP13	0.04659140285072	3.28971247310608
ENSG00000081041	CXCL2	0.046183059182225	2.66323357962152
ENSG00000087494	PTHLH	0.024106589314239	2.66637501474068
ENSG00000103313	MEFV	0.023778588158931	2.79991301764833
ENSG00000108342	CSF3	0.040281528774407	4.41860645591265
ENSG00000108700	CCL8	0.043210890508962	2.71727762890334
ENSG00000117472	TSPAN1	4.33875415758483e-05	-3.42709090123723
ENSG00000124466	LYPD3	0.016234903859388	-1.72215604378848
ENSG00000125618	PAX8	0.016234903859388	-2.29759814520272
ENSG00000125775	SDCBP2	0.039263836766313	-0.909366361862117
ENSG00000126262	FFAR2	0.043210890508962	2.69735067674894
ENSG00000134508	CABLES1	0.045890809553687	2.2018677880535
ENSG00000135480	KRT7	0.009591267324751	-2.9168005497789
ENSG00000137757	CASP5	0.012777090110659	3.34474584568217
ENSG00000137875	BCL2L10	0.000326993611187	5.00503350494086
ENSG00000139364	TMEM132B	0.023384052958253	4.64498851403681
ENSG00000141052	MYOCD	0.000635124921833	3.99471394324698
ENSG00000150750	C11orf53	0.000161056593341	-4.57498126228384
ENSG00000159261	CLDN14	0.04659140285072	3.61482713422454
ENSG00000161681	SHANK1	0.027437008553771	-2.90619448646591
ENSG00000162366	PDZK1IP1	0.002147801028592	-4.2610792719897
ENSG00000163331	DAPL1	0.023384052958253	-3.57567328157941
ENSG00000163735	CXCL5	0.043210890508962	4.32917095014284
ENSG00000164619	BMPER	0.037005802118134	2.78618208633511
ENSG00000167992	VWCE	0.007414940545398	2.7626863676361
ENSG00000169583	CLIC3	0.024863420216287	-1.95659691043809
ENSG00000171056	SOX7	0.0164175524956	1.96397542036314
ENSG00000171777	RASGRP4	0.021000588788647	2.46714758316757
ENSG00000173898	SPTBN2	0.000185096641237	4.2481837893917
ENSG00000175857	GAPT	0.015817387628694	-2.52675374526054
ENSG00000178773	CPNE7	0.032718314115826	-2.03268037243123
ENSG00000182013	PNMAL1	0.004462138272189	4.84661659454258
ENSG00000185933	CALHM1	0.008411562920912	-3.89027586889268
ENSG00000196188	CTSE	0.000161056593341	-4.97762473082687
ENSG00000196611	MMP1	0.039863508333555	4.07776835045208
ENSG00000196943	NOP9	0.002975355817788	0.544129789675922
ENSG00000198216	CACNA1E	0.002695005238933	3.89667768120171
ENSG00000198744	RP5-857K21.11	0.031896754943518	3.32327726022926
ENSG00000211956	IGHV4-34	0.040281528774407	3.90292667217729
ENSG00000230628	RP4-781K5.6	0.000530309604068	4.59431318466852
ENSG00000239998	LILRA2	0.034256580809665	2.33298283812688
ENSG00000247516	MIR4458HG	0.002147801028592	-2.16443843306188
ENSG00000256916	RP11-817J15.2	0.020967029697476	-3.8871059840873
ENSG00000259078	PTBP1P	3.12227873767283e-05	4.03893401535294
ENSG00000261226	RP11-830F9.7	0.009591267324751	3.14390453054503
ENSG00000267325	LINC01415	0.009591267324751	3.72689153449699
ENSG00000269416	LINC01224	0.027708488686373	4.22884339630951
ENSG00000269821	KCNQ1OT1	0.042358266154057	1.69702729559704
ENSG00000270885	RASL10B	0.042358266154057	2.47725290463822
ENSG00000272825	LL21NC02-1C16.2	0.006134241955645	-2.14166495747368

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

APPENDIX B

Differentially expressed genes between DLBCL Stage IV and I

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Table B.1: Differentially expressed genes between DLBCL Stage IV and I

ensembl	gene	p-adjusted value	LFC
ENSG00000012171	SEMA3B	0.0080551972793531	-2.27814079296751
ENSG00000013588	GPRC5A	0.0182332373595419	-3.30570387375667
ENSG00000019102	VSIG2	0.0225040204616075	-2.19468207122905
ENSG00000043039	BARX2	0.0265506786478361	-3.36118002593695
ENSG00000048540	LMO3	0.000197755869877	-4.16709415065189
ENSG00000049192	ADAMTS6	0.0315746385255138	2.08739268913258
ENSG00000050767	COL23A1	0.0415459264841189	-1.95073414821732
ENSG00000058056	USP13	0.0054740008920711	0.913490145654227
ENSG00000064393	HIPK2	0.0054740008920711	1.58936644168583
ENSG00000064787	BCAS1	0.0306857909135208	-2.96261427665607
ENSG00000070601	FRMPD1	0.012727215188364	3.07552933298849
ENSG00000072121	ZFYVE26	0.0326697519983108	0.734122889077704
ENSG00000075643	MOCOS	0.0025344600509646	3.36075623964939
ENSG00000078114	NEBL	0.0043317802905685	-2.87063662527267
ENSG00000081041	CXCL2	0.012727215188364	2.19057027210369
ENSG00000088827	SIGLEC1	0.0311451362065017	3.14774063482562
ENSG00000089685	BIRC5	0.0231191793550127	-1.03251738163298
ENSG00000095932	SMIM24	0.0435967145857963	-2.00543399011126
ENSG00000099810	MTAP	0.0315746385255138	1.304250298861
ENSG00000100154	TTC28	0.039312593856124	0.948712630543442
ENSG00000100426	ZBED4	0.0347137947810801	0.985313021333457
ENSG00000102760	RGCC	0.0191304058646445	-1.68013246448044
ENSG00000102794	IRG1	0.0029774651578346	5.08756487100984
ENSG00000103313	MEFV	0.0096225265946299	2.89086857650538
ENSG00000104432	IL7	0.0198489583280891	1.61589864451849
ENSG00000104974	LILRA1	0.0403249312079825	2.19785030489159
ENSG00000105492	SIGLEC6	0.028557664659224	4.62949762550641
ENSG00000107018	RLN1	0.0070146015480038	-2.27738564417865
ENSG00000108582	CPD	0.0454524748020012	0.983956376163854
ENSG00000108700	CCL8	0.0064959244402167	3.21586607539304
ENSG00000108797	CNTNAP1	0.0029774651578346	2.84669555226769
ENSG00000109684	CLNK	0.0254669975901011	-2.40507017905329
ENSG00000110400	PVRL1	0.0002709134715679	2.51403144843464
ENSG00000111319	SCNN1A	0.0044073777106215	-3.32285570609485
ENSG00000113231	PDE8B	0.0245746176828392	-2.03226816446198
ENSG00000117115	PADI2	0.0152420163902181	3.45254096748188
ENSG00000117472	TSPAN1	0.0002149122559547	-3.55674686188998
ENSG00000118971	CCND2	0.0131181316493276	2.46213184612533
ENSG00000119523	ALG2	0.0064959244402167	0.847655628934464
ENSG00000120889	TNFRSF10B	0.0064959244402167	0.824605539626906
ENSG00000122035	RASL11A	0.0152420163902181	-2.51627864985847
ENSG00000123700	KCNJ2	5.69465658323163e-05	3.41006470024504
ENSG00000124097	HMGB1P1	0.0450592699104166	-2.08489592064968
ENSG00000126262	FFAR2	0.0043233455334087	3.67518249656705
ENSG00000128052	KDR	0.0396860625745521	1.39129016916697
ENSG00000128438	TBC1D27	0.0009518427639386	3.84063415145706
ENSG00000128534	LSM8	0.0191304058646445	-0.96566321560695
ENSG00000132938	MTUS2	0.0186644528991332	-2.72404307915044
ENSG00000134250	NOTCH2	0.0435967145857963	1.14671468718466
ENSG00000134508	CABLES1	0.0054740008920711	2.85810378526979
ENSG00000136014	USP44	0.0435351003878504	2.72754712343898
ENSG00000136160	EDNRB	0.023273397592036	1.3787869111215

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Table B.3: Differentially expressed genes of DLBCL Stage IV vs I (continued)

ensembl	gene	p-adjusted value	LFC
ENSG00000170142	UBE2E1	0.002005129632108	-0.850357255255202
ENSG00000170409	CTA-313A17.2	0.0315349432599015	-3.05220971714422
ENSG00000170421	KRT8	0.0151162782901656	-2.50881172404271
ENSG00000170776	AKAP13	0.0110628076343316	0.726382898860182
ENSG00000170786	SDR16C5	0.000389724740537	-4.30890862913366
ENSG00000170791	CHCHD7	0.0126668093177323	-0.783223817948022
ENSG00000171049	FPR2	0.0127886547672688	3.06820727028513
ENSG00000171056	SOX7	0.010675976543345	1.95627421973198
ENSG00000171345	KRT19	0.0076553783740579	-3.93613403066927
ENSG00000171657	GPR82	0.0409997939206971	-2.95296891142415
ENSG00000171777	RASGRP4	0.010675976543345	2.3549413677555
ENSG00000172201	ID4	0.0415459264841189	-2.07222663191088
ENSG00000172493	AFF1	0.0396860625745521	1.24282350507689
ENSG00000172752	COL6A5	0.0491828042747732	-2.49648799991144
ENSG00000172799	ZBTB8OSP2	0.0435351003878504	-2.52390379342413
ENSG00000173239	LIPM	0.0474257626788713	2.73636622084053
ENSG00000173406	DAB1	0.011922203844216	-3.04991747267122
ENSG00000173898	SPTBN2	0.00015729771068	3.36468869662649
ENSG00000174500	GCSAM	0.0302106271999736	-1.84312562505499
ENSG00000175449	RFESD	0.0419638909135114	-1.18467153520287
ENSG00000176049	JAKMIP2	0.0377908315023455	-2.179124792326
ENSG00000176438	SYNE3	0.0075688622507166	1.2437951976913
ENSG00000177575	CD163	0.0245746176828392	3.07778798084948
ENSG00000178662	CSRNP3	0.0105100384473284	-3.74185612621851
ENSG00000179178	TMEM125	0.0002149122559547	-4.15855909936942
ENSG00000180509	KCNE1	0.0302106271999736	3.01360960533757
ENSG00000181143	MUC16	0.0004835338796713	-5.52437564779146
ENSG00000181634	TNFSF15	0.0059068143881358	1.89939644835946
ENSG00000182158	CREB3L2	0.0174814305744348	1.51517779151068
ENSG00000183117	CSMD1	0.0198489583280891	-2.92729471386845
ENSG00000183248	PRR36	0.0302106271999736	-3.18051075902029
ENSG00000183833	MAATS1	0.000471152127032	3.14971585469776
ENSG00000184292	TACSTD2	0.0361466694422472	-1.58286451091648
ENSG00000185565	LSAMP	0.0306857909135208	-2.56615436753942
ENSG00000185668	POU3F1	0.0300834992674709	2.23433069825331
ENSG00000186594	MIR22HG	0.0225040204616075	1.03501900711919
ENSG00000187045	TMPRSS6	0.0075688622507166	2.9773490411564
ENSG00000187486	KCNJ11	0.0265506786478361	-2.18817836872782
ENSG00000187595	ZNF385C	0.0040137168771352	3.4353245788761
ENSG00000188786	MTF1	0.0225040204616075	1.00734715509738
ENSG00000189043	NDUFA4	0.0435967145857963	-0.774609400888395
ENSG00000196188	CTSE	0.0076553783740579	-4.92981446877325
ENSG00000197081	IGF2R	0.0474284046192058	1.41970666710442
ENSG00000198125	MB	0.0392183912326489	-2.68725304039078
ENSG00000198216	CACNA1E	0.0015451070225996	3.80648050748213
ENSG00000198626	RYR2	0.0048625831392111	-3.15563711884461
ENSG00000198719	DLL1	0.0162697536487791	1.7123556330258
ENSG00000198785	GRIN3A	0.0132012976831993	2.22665846800334
ENSG00000203386	LINC01317	0.0450592699104166	-3.534902547627
ENSG00000203685	C1orf95	0.0498156675355045	2.28387953642016
ENSG00000169385	RNASE2	0.0450592699104166	2.77791968688495
ENSG00000169554	ZEB2	0.0333721184581284	1.26240187395594

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Table B.2: Differentially expressed genes between DLBCL Stage IV and I (continued)

ensembl	gene	p-adjusted value	LFC
ENSG00000136205	TNS3	0.0015451070225996	1.4592692184579
ENSG00000136929	HEMGN	0.0131181316493276	-2.59165324003488
ENSG00000137558	PI15	0.0382263711497532	3.17875221474347
ENSG00000137757	CASP5	0.0006056658454371	4.16627964314157
ENSG00000137875	BCL2L10	0.0009518427639386	4.69518862896482
ENSG00000138119	MYOF	0.0333721184581284	1.88966837941456
ENSG00000138185	ENTPD1	0.0152420163902181	1.33129921949752
ENSG00000138639	ARHGAP24	0.0397726715679785	1.61689979623957
ENSG00000138678	AGPAT9	0.0320810071922897	2.4485668401467
ENSG00000138771	SHROOM3	0.0018140932137202	-3.61847450397684
ENSG00000139269	INHBE	0.026792236137796	-2.10003529133923
ENSG00000139364	TMEM132B	0.0059068143881358	4.81116695690986
ENSG00000141052	MYOCD	1.2170403843048502e-09	6.67066924622333
ENSG00000143816	WNT9A	9.18577349012687e-05	3.45747605281194
ENSG00000144791	LIMD1	0.0058739710970645	1.45148981236211
ENSG00000144802	NFKBIZ	0.0113839088283033	2.35482297735021
ENSG00000145721	LIX1	0.0290592133222244	-3.40719679589278
ENSG00000150750	C11orf53	0.0352139757497171	-3.50517357662055
ENSG00000151835	SACS	0.0087373804161567	1.93106497121009
ENSG00000151876	FBXO4	0.0015451070225996	-0.698081363741058
ENSG00000152229	PSTPIP2	0.0029774651578346	2.91614841499783
ENSG00000152766	ANKRD22	0.0352300870469837	2.91626411109311
ENSG00000154118	JPH3	0.0415459264841189	2.47585474360955
ENSG00000154258	ABCA9	0.0003713355416425	2.96565857558523
ENSG00000154269	ENPP3	0.0111353690485616	-3.4865653621399
ENSG00000154556	SORBS2	0.013016166522899	-2.12187961720396
ENSG00000157303	SUSD3	0.0499457645937296	-1.58023873773983
ENSG00000157833	GAREML	0.0300834992674709	1.42249156782948
ENSG00000158966	CACHD1	0.0401156968666733	1.27725674481363
ENSG00000160447	PKN3	0.0166489799717187	2.04617641226922
ENSG00000160505	NLRP4	0.0009518427639386	-5.72328200561015
ENSG00000161681	SHANK1	0.0029774651578346	-3.41739623101971
ENSG00000161888	SPC24	0.0324059607864822	-1.29540542699127
ENSG00000161905	ALOX15	0.0009518427639386	-3.41916461040407
ENSG00000163563	MNDA	0.0248396705290312	2.83616975897988
ENSG00000163618	CADPS	0.0279107730229619	-2.31760501156194
ENSG00000163661	PTX3	0.0393847823258287	2.05609013281012
ENSG00000163874	ZC3H12A	0.0024413277815428	1.23237758802742
ENSG00000163888	CAMK2N2	0.0314847187798884	-2.58637484594453
ENSG00000164047	CAMP	1.62994604928191e-06	-4.54157902066579
ENSG00000164611	PTTG1	0.0026724138900659	-1.6311508493504
ENSG00000164695	CHMP4C	0.0393847823258287	-2.44262989933113
ENSG00000164742	ADCY1	0.0435351003878504	1.6729987634669
ENSG00000164758	MED30	0.0409997939206971	-0.849264235382671
ENSG00000165078	CPA6	0.0002149122559547	-5.03376557953131
ENSG00000165238	WNK2	0.0306857909135208	-2.87301511168366
ENSG00000166446	CDYL2	0.0450592699104166	0.951811163820251
ENSG00000167779	IGFBP6	0.0320810071922897	-1.76469610255372
ENSG00000167900	TK1	0.0435351003878504	-1.1343650936714
ENSG00000167992	VWCE	0.0002284941017711	2.60034437668822
ENSG00000168594	ADAM29	0.0058739710970645	-2.38457807214895
ENSG00000168952	STXBP6	0.0450592699104166	-3.28886828297119

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Table B.4: Differentially expressed genes of DLBCL Stage IV vs I (continued)

ensembl	gene	p-adjusted value	LFC
ENSG00000204271	SPIN3	0.0474284046192058	2.02644938999801
ENSG00000205038	PKHD1L1	0.0198489583280891	-2.30123750831337
ENSG00000205302	SNX2	0.0064959244402167	-0.830404776449015
ENSG00000205358	MT1H	0.0338585355418436	4.3666793303662
ENSG00000205517	RGL3	0.0181654587591541	-2.32404064420159
ENSG00000205593	DENND6B	0.0377908315023455	-2.11103714959896
ENSG00000211598	IGKV4-1	0.0030743528532098	-4.03488465781606
ENSG00000211829	TRDC	0.0377908315023455	-2.77647208054978
ENSG00000211972	IGHV3-66	0.0058739710970645	-4.12377829781765
ENSG00000212232	SNORD17	0.0393825272504419	1.45064693306643
ENSG00000213073	RP11-288H12.3	0.0166489799717187	1.9121877704805
ENSG00000213231	TCL1B	0.0331189173294702	2.80955658707461
ENSG00000222009	BTBD19	0.0018210446022504	2.44003576968628
ENSG00000224184	AC096559.1	0.0040137168771352	-2.93843448074832
ENSG00000225313	RP11-415J8.3	0.0245746176828392	1.91332932452802
ENSG00000227070	RP11-191G24.1	0.0354006809348203	-2.53292684957212
ENSG00000229666	MAST4-AS1	0.0134179188821547	-2.56070045795315
ENSG00000229921	KIF25-AS1	0.0191236217887054	-3.29037123788946
ENSG00000230061	TRPM2-AS	0.0118318015355181	2.30898792332622
ENSG00000230628	RP4-781K5.6	0.0012449969951687	4.53564559046853
ENSG00000231106	LINC01436	0.0257826223148095	-3.70790744879889
ENSG00000231345	BEND3P1	0.0498156675355045	3.11667246775749
ENSG00000232043	RP4-530I15.9	0.0345212982202113	2.82307062658556
ENSG00000232884	AF127936.3	0.0004690778031771	-4.27783004399558
ENSG00000234184	RP5-887A10.1	0.0333721184581284	-2.76281125685593
ENSG00000237973	RP5-857K21.6	0.0073842309427695	1.68524750588099
ENSG00000241294	IGKV2-24	0.0435351003878504	-3.26776095660522
ENSG00000242612	DECR2	0.0161584912590203	1.92695936779911
ENSG00000243147	MRPL33	0.0118924588375511	-0.893197975505742
ENSG00000243440	AF165138.7	0.0290592133222244	-3.36988570081188
ENSG00000244345	RP11-654C22.2	0.0029774651578346	-3.14176804390245
ENSG00000247675	LRP4-AS1	0.0076077731294433	-3.07142493370033
ENSG00000250138	RP11-848G14.5	0.0435967145857963	2.68210989492626
ENSG00000256751	PLBD1-AS1	0.0013611440973839	2.58547505736682
ENSG00000259078	PTBP1P	1.62994604928191e-06	4.6342226350428
ENSG00000260177	RP4-536B24.3	0.0312469054304134	-2.86094093879842
ENSG00000260979	RP11-77H9.8	0.0029774651578346	-3.34419623334892
ENSG00000261226	RP11-830F9.7	2.16205108485397e-06	3.12939380962182
ENSG00000261618	RP11-79H23.3	0.0142600182364689	2.65417524142108
ENSG00000262636	CTD-3088G3.4	0.0230471653687201	-1.71285135788748
ENSG00000267325	LINC01415	0.0007749306313066	4.3692857617222
ENSG00000269821	KCNQ1OT1	2.82147861150487e-06	2.20380760826682
ENSG00000270344	RP11-734K2.4	0.0224407126671891	-1.15109450538211
ENSG00000272720	CTA-228A9.3	0.0369899440385719	-1.78653802452368
ENSG00000272734	ADIRF-AS1	0.0015451070225996	2.33487034974766
ENSG00000273102	AP000569.9	0.0245746176828392	-2.69705385051449
ENSG00000274443	C8orf89	0.0072719074444676	-3.07776012188047
ENSG00000274576	IGHV2-70	0.0435351003878504	-2.3631776587471
ENSG00000275418	RP11-126O1.6	0.0151672645914482	-2.14704409799634
ENSG00000278970	HEIH	0.0491828042747732	-0.97579203001652
ENSG00000280206	CTB-193M12.5	0.0397726715679785	-1.43506445975165

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/