Novel genomic biomarkers for pediatric and adult Acute Myeloid Leukemia

Nasr O M Eshibona

This dissertation submitted in fulfilment of the requirements for the degree of Doctor Philosophiae in Bioinformatics at the South African National Bioinformatics Institute,

University of the Western Cape

Supervisor: Dr Hocine Bendou

Co-supervisor: Prof Junaid Gamieldien

February 2023

DECLARATION OF AUTHORSHIP

I declare that "Novel genomic biomarkers for pediatric and adult Acute Myeloid Leukemia" is my own work, that 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**.**

Signed: **Date: Feb of 2023**

ACKNOWLEDGMENTS

First and foremost, I praise Allah, the Almighty, for providing me with this opportunity and granting me the capability to complete my PhD successfully. I could never have accomplished this without Allah granting me the courage and strength to persevere through all the obstacles and challenges I faced during the process. Through Allah, I believed in myself and pursued my dreams.

I would like to thank the valuable guidance from my supervisor, Dr Hocine Bendou, and co-supervisor Prof Junaid Gamieldien for their supervision, knowledge, assistance, encouragement, expertise, understanding, input in the research, and their time proofreading my thesis.

I thank my family for their motivation, support, and encouragement.

I would also like to thank Miss Michelle Chantel Livesey, Dr Sophia Catherine Rossouw, and Dr Abdulazeez Giwa for their assistance, advice, and valuable comment throughout the project. I would acknowledge the South African national Bioinformatics institute for providing a learning environment and collaborative team to overcome the challenge along the journey.

My gratitude goes to the South African Medical Research Council (SAMRC) and its Division of Research Capacity Development for supporting this research through the South African National Treasury's Mid-Career Scientist Programme. This work was also supported by the South African Research Chairs Initiative of the Department of Science and Innovation, the South African National Research Foundation.

DEDICATION

This thesis is dedicated to almighty Allah and my family.

ABSTRACT

Acute myeloid leukemia (AML) is a heterogeneous type of blood cancer that affects individuals of all ages. AML patients are categorized into favorable, intermediate, and adverse risks based on patients' genomic features and chromosomal abnormalities. Despite this risk stratification, the progression and outcome of the disease remain highly variable in pediatric and adult patients, which emphasizes the importance of finding more accurate genomic biomarkers studying the gene expression profiling of pediatric and adult AML patients to facilitate and improve the risk stratification of the patients. Consequently, two research aims were proposed to study the prognostic heterogeneity for pediatric and adult AML. In pediatric AML, the research project was set to identify a genetic signature related to patients with *FLT3-ITD* mutation and poor survival. While for adult AML, this study focused on establishing a genetic signature predictive of prognosis with the ability to accurately reclassify the risk of AML intermediate group.

RNA- Sequencing (RNA-Seq) count datasets for pediatric and adult AML were retrieved from the UCSC Xena browser and Gene Expression Omnibus, respectively, with their corresponding clinical information and survival data. The proposed aims were achieved by performing differential gene expression on both datasets, followed by additional bioinformatics analyses, including machine learning, Cox regression, Kaplan-Meier, receiver operating characteristics, Gene Ontology and KEGG enrichment, and statistical analyses.

High expression of *FHL1*, *SPNS3*, and *MPZL2* was associated with poor survival in patients with *FLT3-ITD* mutation and can serve as a prognostic indicator of unfavourable outcomes in AML pediatric patients. While in adult patients, alteration in expression of *CD109*, *CPNE3*, *DDIT4*, and *INPP4B* was linked to poor outcomes and had a stratification power for accurate risk classification of the intermediate-risk group.

In conclusion, identifying genetic signatures with prognostic value assist in disease management and therapeutic decisions. The genetic signatures identified can potentially improve treatment strategies for most pediatric and adult AML patients who fall into the *FLT3-ITD* and intermediate-risk stratification categories.

KEYWORDS: Acute myeloid leukemia, pediatric, adult, gene expression profile, prognostic, cytogenetic, risk classification

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
ALL	Acute Lymphocytic Leukemia
AML	Acute Myeloid Leukemia
ANOVA	Analysis Of Variance
AUC	Area Under Curve
BP	Biological Process
CC	Cellular Component
CEBPA	CCAAT/enhancer-binding Protein Alpha
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CR	Complete Remission
DEG	Differentially Expressed Genes
DGE	Differential Gene Expression
DNA	Deoxyribonucleic Acid
ELN	European LeukemiaNet
FAB	French-American-British
FHL1	Four and a Half LIM Domains Protein 1
FLT3	FMS Related Receptor Tyrosine Kinase 3
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GEP	Gene Expression Profiling
GG	Good-Good
GO	Gene Ontology
HR	Hazard Ratio
IG	Intermediate-Good
IP	Intermediate-Poor
ITD	Internal Tandem Duplication Mutations
KEGG	Kyoto Encyclopedia of Genes and Genomes
K-M	Kaplan-Meier Curve
LS	Long Survival
MF	Molecular Function
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPZL2	Myelin Protein Zero-Like 2
MRC-C	Medical Research Council Classification
NGS	Next-Generation Sequencing
NPM1	Nucleophosmin 1
OS	Overall Survival
PCA	Principal Component Analysis
PP	Poor-Poor

RFE	Recursive Feature Elimination
RMA	Robust Multi-array Average
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SPNS3	SPNS Lysolipid Transporter 3
SS	Short Survival
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
US	United States
WHO	World Health Organization
WT1	Wilms Tumor 1

Publications from this study:

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

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

Chapter 1: Introduction to thesis and research statement

1.1 Introduction

Cancer is a collection of diseases characterized by abnormal and uncontrolled cell growth caused primarily by a genetic mutation. More than 200 forms of cancer have been described, and each type can be characterized by different molecular profiles, requiring unique therapeutic strategies (Tomczak, Czerwińska and Wiznerowicz, 2015). Thus, cancer involves dynamic changes in the genome. A better understanding of which genes are most commonly mutated across all cancers and at what frequency could help prioritize genes and pathways in a manner that increases public health benefits (Martínez-Jiménez *et al.*, 2020). Cancer continues to be one of the most challenging diseases to be treated and is one of the leading causes of death around the globe. In 2020, an estimated 19.3 million new cancer cases and almost 10 million cancer deaths were recorded (Sung *et al.*, 2021). Cancers account for 13% of all deaths yearly, with cancer-related mortality expected to rise to 13.1 million by 2030 (Rashid *et al.*, 2019).

Cancer becomes a fatal disease due to late detection caused by the lack of diagnostic biomarkers, inappropriate therapy for each form of cancer, and the potential for drug resistance. These factors are the result of the accumulation of several genetic and epigenetic changes within the cell, resulting in molecular/chromosomal abnormalities and genetic instability (Yoshioka *et al.*, 2021). Individual aetiological factors are difficult to quantify. However, it can be determined that multiple risk factors contribute to cancer formation. Environmental, exogenous, and endogenous factors, as well as individual factors, including genetic predisposition, contribute to cancer development. Epidemiological research on malignant tumours has focused on environmental and genetic aspects of cancer incidence and mortality (Lewandowska *et al.*, 2019).

Acute myeloid leukemia (AML) is a type of cancer characterized by the uncontrolled proliferation of hematopoietic stem cells in the bone marrow (Nepstad *et al.*, 2020). The highly heterogeneous disease affects individuals of all ages; the prevalence generally increases with age. Therefore, elderly patients account for most newly diagnosed cases (De Kouchkovsky and Abdul-Hay, 2016). The overall survival (OS) rate in children is 60–70% and decreases gradually with age to <5% in those over 65 (Kiem Hao *et al.*, 2020). Children and adults die within five years of diagnosis due to relapse (up to 35% and 99%, respectively) and disease progression (Aung *et al.*, 2021).

Recent statistics showed that the American Cancer Society estimated 20,050 new cases and 11,540 deaths in the US alone (Siegel, Miller and Jemal, 2020). The genetic and clinical profiles of pediatric and adult AML also illustrate highly distinct diseases (Bolouri *et al.*, 2018; Liu, Spiegelman and Wang, 2022). Thus, the mutational landscape shows fewer infrequent mutations and a disproportionate prevalence of somatic structural variants in pediatrics when compared to adults (Bolouri *et al.*, 2018). Multiple chromosomal abnormalities are linked to AML, and these anomalies have comparable clinical symptoms but distinct morphologic, immunophenotypic, and cytogenetic subgroups (Maleki Behzad *et al.*, 2021). The heterogeneity of AML emphasizes the significance of studying pediatric and adult AML separately, as it can provide a unique entity and novel disease landscape specific to each age group. This, in turn, can improve survival rates and optimise treatment options.

Cytogenetics has been linked to clinical outcomes in AML, including complete remission rates, relapse risk, and OS. They served as the foundation for AML risk categorization and led to the development of a cytogenetic risk stratification system known as the original Medical Research Council classification (MRC-C) system. The MRC-C was updated in 2010 (revised MRC-C). Patients were consequently stratified into three groups, namely; favourable, intermediate and adverse, based on the cytogenetic and gene mutation profiles (Grimwade *et al.*, 1998, 2010; Grimwade and Hills, 2009). Consequently, it offered a much-needed standardized, objective and evidence-based guide for clinicians to make critical consolidation therapy decisions regarding the appropriateness, timing, and nature of stem cell transplants in a patient's first complete remission (Komanduri and Levine, 2016). The MRC-C system was further incorporated in the European LeukemiaNet (ELN) and World Health Organization (WHO) classification systems, which rely mainly on the cytogenetics in the risk classification (Harris *et al.*, 1999; Vardiman *et al.*, 2009; Döhner *et al.*, 2010; Arber *et al.*, 2016; Döhner *et al.*, 2017).

All risk stratification systems are regularly updated to incorporate new findings obtained from technological advancements, increased clinical data, and biological insightfulness of the disease. An improved individual prognosis and guide management of AML were enabled by identifying recurrent genetic mutations, such as *FLT3* internal tandem duplication (*FLT3-ITD*), *NPM1*, and *CEBPA* mutations (De Kouchkovsky and Abdul-Hay, 2016). However, the recent risk classification has undergone significant amendments. The *FLT3-ITD* allelic ratio has been excluded, and *FLT3-ITD* without NPM1 mutation is no longer classified as an adverse risk due to the incorporation of an *FLT3* inhibitor. Additionally, in-frame mutations affecting the basic leucine zipper region (bZIP) of *CEBPA*, irrespective of monoallelic or biallelic, are no longer classified as intermediate-risk but as a favourable-risk group (Döhner *et al.*, 2022). Hyperdiploid karyotypes with complex abnormalities should not simply be considered an adverse risk. However, additional assessments for specific chromosomal variations are needed to determine the risk group; for example, hyperdiploid karyotype with multiple trisomies (or polysomies) are no longer considered complex karyotypes, therefore, should not be regarded as adverse risks (Chilton *et al.*, 2014). Notably, the AML risk stratification systems continue to change or update, such as in the case of *FLT3* mutations, which emphasizes the need for a more comprehensive description and understanding of the genetic basis of the risk groups to improve AML patients' prognosis and provide more effective treatment strategies.

RNA-Sequencing (RNA-Seq) has made substantial contributions to several research fields, particularly cancer research, with the emergence of the era of precision medicine. These contributions include studies on differential gene expression analysis, cancer heterogeneity and evolution, cancer drug resistance, the cancer microenvironment and immunotherapy, neoantigens, and other topics (Hong *et al.*, 2020). Recent advancements in high-throughput sequencing enable a more comprehensive understanding of the molecular level of the genome and transcriptome. The technology further has the potential to detect early and high molecular risk mutations. Thus, it can be utilized to find novel cancer biomarkers and prospective therapeutic targets, as well as to monitor diseases and guide early treatment decisions regarding targeted therapy (Hong *et al.*, 2020). Therefore, using RNA-Seq in this study focused on AML could lead to a better understanding of the disease, including the risk classification, prognosis, and potential use as a therapeutic target.

1.2 Problem statement

AML is a distinct disease in pediatric and adult patients regarding survival and risk classification. Additionally, driver genes for high-risk pediatric and adult AML are still not fully understood (Liu, Spiegelman and Wang, 2022). Therefore, independent research on both groups must be investigated to identify gene signatures that could improve the risk classification and decipher the heterogeneity within each group.

Clinical recommendations for AML classification and risk stratification remain heavily reliant on cytogenetic findings at diagnosis, which are present in < 50% of patients (Tazi *et al.*, 2022). It should be noted that cytogenetics does not entirely account for the disease heterogeneity, despite being widely employed in the risk classification. Therefore, the current risk classification does not reflect the heterogeneity of disease survival and clinical outcome. Additionally, the recent advancement in testing abilities and mutation profiling that have become more readily available has led to rapidly changing genetic risk groups (Conneely and Stevens, 2021). Thus, it illustrates the importance of genetic signatures to obtain an accurate risk classification to facilitate the clinical management of AML.

Meanwhile, the more specific intermediate-risk group requires reclassification. Most adult AML patients are being stratified to the intermediate-risk group (an umbrella category) because they do not meet the criteria identifying specific entities of established prognostic relevance (Awada *et al.*, 2022). Therefore, due to the heterogeneity of the disease and the clinical outcome, there is still a need for more accurate prognostic biomarkers for AML.

1.3 Aim and objectives

This study aims to identify novel diagnostic and prognostic biomarkers for AML. To achieve this aim, the objectives of the study were to:

Pediatric AML

i. Retrieve and extract pediatric RNA-Seq data from the GDC TARGET database with corresponding clinical data.
ii. Identify differentially expressed genes (DEGs) between low- and high-risk patient groups and perform Principal Component Analysis (PCA).
iii. Use the features selection methods to select the genes with the highest performance in the sample segregation and apply machine learning (ML) techniques for sample classification using gene expression profiles derived from feature selection.
iv. Apply Cox regression and Kaplan Meier to identify prognostic genes.

Adult AML

i. Retrieve and extract adult RNA-Seq data from the Gene Expression Omnibus (GEO) with corresponding clinical data, and segregate samples based on survival.
ii. Apply Differential Gene Expression analysis between short- and long-survival groups using the limma R package.
iii. Perform a survival analysis on DEGs using Cox regression to identify genes implicated in prognosis.
iv. Validate the prognostic value of DEGs using Kaplan-Meier (K-M) and receiver operating characteristic (ROC).
v. Apply a one-way Analysis of Variance (ANOVA) to assess differences in the expression means of prognostic genes between the risk subcategories and OS.
vi. Perform Gene ontology (GO) and KEGG pathways analyses to illustrate the implication of the DEGs in AML.

1.4 Thesis overview

Chapter 2. Literature review.

A literature review of AML genomics, therapeutics, current knowledge of biomarker discovery, bioinformatics resources and tools.

Chapter 3. Upregulation of *FHL1*, *SPNS3*, and *MPZL2* predicts poor prognosis in pediatric acute myeloid leukemia patients with *FLT3-ITD* mutation.

Describes the use of transcriptomic data to find prognostic biomarkers in AML patients with *FLT3-ITD* and *NPM1*/*CEBPA* mutations, as the *FLT3-ITD* mutation is a factor that is responsible for poor prognosis in pediatric AML. The use of gene expression for risk prediction is also detailed in this chapter.

Chapter 4. Investigation of distinct gene expression profile patterns that can improve the classification of intermediate-risk prognosis in AML patients.

Applies different bioinformatics tools to identify predictive biomarkers associated with short survival risk groups in adult AML and can reclassify the intermediate-risk patients based on the biomarkers' expression levels.

Chapter 5. Conclusion and future recommendations.

Chapter 2: Literature Review

2.1 Leukemia

Cancers of the blood and bone marrow, collectively known as leukemias, pose a significantly high mortality risk. Leukemia is an aberrant hyper-proliferation of immature blood cells or blast cells. Depending on the affected cells, leukemia can be either myeloid or lymphoid in lineage and classified as acute or chronic. Chronic leukemias, which often have more mature cells, are uncommon in children. Contrarily, acute leukemias tend to be less mature, often affect individuals of all ages and have the potential to be lethal very quickly if not promptly treated (Juliusson and Hough, 2016; An, Fan and Xu, 2017; Bispo, Pinheiro and Kobetz, 2020). The most frequent childhood cancer is leukemia (28%), followed by brain and other nervous system tumours (26%), roughly one-third of which are benign or borderline malignant (Siegel *et al.*, 2022). According to the Surveillance, Epidemiology, and End Results (SEER) database, in 2020, leukemia accounted for around 3.4% of all newly diagnosed cancer cases and 3.8% of all cancer deaths (Lin *et al.*, 2021). There are four subtypes of leukemia; acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), and chronic myeloid leukemia (CML) (Miranda-Filho *et al.*, 2018; Du *et al.*, 2022).

ALL is the most prevalent kind of childhood cancer, and treatments offer a significant possibility of curing the disease. Adults can also develop ALL, and although the likelihood of a cure is considerably reduced, ALL is recognised by chromosomal abnormalities and genetic changes that influence the growth and proliferation of lymphoid progenitor cells (Bhojwani, Howard and Pui, 2009; Fujita *et al.*, 2021). CLL is among the most common forms of leukemia. It often affects the elderly and has a highly varied clinical course. CML is a clonal hematopoietic stem cell neoplasia defined by increased myeloid lineage cells at all differentiation stages. Specific genetic changes that interfere with the regulation of proliferation and apoptosis in clonal B-cells promote leukemic transformation (Hallek and Al-Sawaf, 2021). Approximately 15% of newly diagnosed cases of adult leukemia are myeloproliferative neoplasms, with a frequency of 1-2 cases per 100,000 persons (Alves *et al.*, 2021).

AML follows ALL as the most common form of leukemia diagnosed in patients younger than 55, and it continues to be the second in patients over the age of 65 after CLL. It is also the second cause of death in patients under 20 years and the top cause of death in those over 25 years old (Figure 2.1). AML was responsible for the majority of leukemia-related deaths among the four subtypes. Despite several years of research into the pathophysiology and molecular heterogeneity of AML, the standard treatment has remained unchanged (Lin and Levy, 2012; Du *et al.*, 2022). Investigating the gene expression profile of AML and identifying prognostic biomarkers will provide insights into the development and progression of AML, as these biomarkers are useful for physicians to monitor the disease's prognosis and could potentially be used for target therapy.

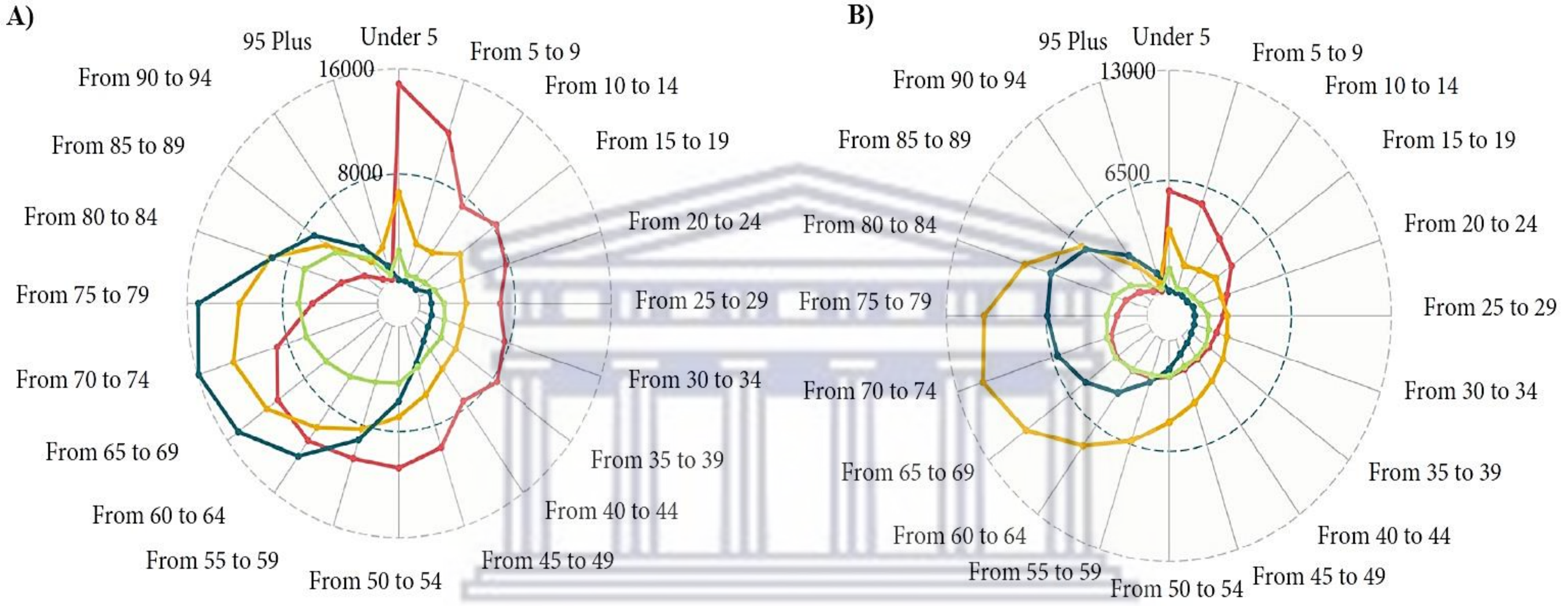

Figure 2.1: Number of cases of four leukemia subtypes AML, CML, ALL and CLL represented in lines orange, green, red, and dark blue, respectively, (A) incidence and (B) death, in different age groups. The circle starts at the top of each graph and increases by 5 years in a clockwise direction. The age ranges between under 5 and above 95 years old, modified from (Du *et al.*, 2022).

2.2 Acute myeloid leukemia

AML, also referred to as acute granulocytic leukemia, acute nonlymphocytic leukemia, acute myelogenous leukemia, and acute myeloblastic leukemia, is the most common form of acute leukemia, and its incidence increases with age. Therefore, AML is the most prevalent leukemia in adults and second most common in pediatrics (Wouters and Delwel, 2016; Bispo, Pinheiro and Kobetz, 2020; Carter *et al.*, 2020). The disease is caused by a multipotent malignant stem cell that has undergone a transformation and acquired a subsequent genetic mutation (Infante, Piris and Hernández-Rivas, 2018).

Currently, the pathogenesis of AML is unknown. However, several studies indicate that virus infection may be the primary cause (Guo, Wang and Sun, 2022). In addition, some cytotoxic drugs, such as alkylating agents and topoisomerase inhibitors, ionising radiation, benzene, and other risk factors, might cause chromosomal damage and vulnerability. They, in turn, can cause the position of oncogenes to shift and then become activated and immunological function to decrease, which is conducive to the development of leukemia (Ye *et al.*, 2019; Young *et al.*, 2019; Guo, Wang and Sun, 2022). A thorough understanding of the molecular alterations associated with chromosomal and genetic abnormalities in AML is anticipated to facilitate the design of therapies and the discovery of biomarkers (Kumar, 2011).

AML cases are diagnosed by routine blood investigations or symptomatic presentation, such as infection and bleeding. Diagnosis involves a morphological examination of bone marrow aspirate, immunophenotyping, and detection of genetic abnormalities. A morphological test of the bone marrow cells is performed to obtain the myeloid blast cell count, for which a blast cell counts of 20% or more is diagnostic of AML (Vardiman *et al.*, 2009; Vakiti and Mewawalla, 2022).

Initially, AML diagnoses were solely based on morphological evaluation. Now, AML diagnoses incorporate other approaches, such as cytomorphology, cytochemistry, immunophenotyping, cytogenetics, and molecular genetics. Combining these approaches contributes to each case's characterisation, improving AML diagnosis and treatment (Haferlach and Schmidts, 2020).

Cytotoxic chemotherapy and numerous additional medications, such as all-trans retinoic acid, are being used to treat AML. However, different cytogenetic abnormalities, somatic mutations, and alterations in the epigenome cause AML to have substantial genetic variability, making subtype categorisation and therapy challenging (Walter *et al.*, 2013; Lohse *et al.*, 2018).

2.3 Prognosis of acute myeloid leukemia

AML prognosis depends on clinical variables such as patient age, performance level, comorbidities, and leukemia-specific genetic characteristics such as cytogenetics and molecular abnormalities. Using a few molecular and clinical criteria, AML patients are classed as having a favorable, intermediate, or poor prognosis. The prognostic assignment defines treatment alternatives to optimise therapeutic efficacy and decrease recurrence. AML exhibits extensive heterogeneity and genomic complexity, depending on the presence or absence of cooperating mutations within functional categories such as epigenetic regulators, cell signalling and proliferation pathways, and master hematopoietic transcription factors (Fröhling *et al.*, 2006; Grimwade *et al.*, 2010; Dinardo and Cortes, 2016).

Cytogenetic analysis to detect prominent structural chromosomal abnormalities provided the first "genetic" prognostication schema in AML and remains the backbone of current AML genomic classification, partitioning patients based on their pre-treatment karyotype into those with favourable, intermediate, or adverse cytogenetics and correlating with 5-year OS of ~60%, 30% to 40%, and 5% to 10%, respectively (Dinardo and Cortes, 2016).

The individual prognostic information derived from the presence or absence of a specific mutation can be modified by the presence of cooperating co-mutations (i.e. *NPM1* with *FLT3*-internal tandem duplications (ITDs) or *DNMT3A* mutations), signifying that the optimal personalised prognostication requires knowledge of the complete AML genomic landscape (Dinardo and Cortes, 2016). Mutations in specific genes were closely associated with defined, prognostically distinct cytogenetic subgroups. Driver mutations typically change the prognostic consequences of specific mutations. Co-mutation patterns in NPM1-mutated AML predicted a favourable or adverse prognosis (Papaemmanuil *et al.*, 2016). Furthermore, multivariate analyses revealed that the impact of some mutations depends on patient age (Metzeler *et al.*, 2016).

Older patients have more comorbidities and poor prognostic factors. Thus, healthcare practitioners treat them less aggressively because they expect them to benefit less from intensive treatments (Oran and Weisdorf, 2012). The standard-dose or low-intensity induction regimen for elderly AML patients is controversial because of their poor prognosis (Oran and Weisdorf, 2012). All prognostic factors and risk assessments should be considered to ensure each patient receives suitable individualised treatment (Oran and Weisdorf, 2012).

The most widely accepted classification and prognostic schemes for AML include cytogenetic lesions together with *NPM1*, *FLT3-ITD*, and *CEBPA* mutations. In the near future, *TP53*, *SRSF2*, *ASXL1*, *DNMT3A*, and *IDH2* should be incorporated into prognostic guidelines because they are common and strongly influence clinical outcomes. For AML classification, evaluation of splicing-factor genes *RUNX1*, *ASXL1*, and *MLLPTD* at diagnosis would identify patients in the chromatin spliceosome mutation group, which is common in older patients and associated with poor response to induction chemotherapy (Vardiman *et al.*, 2009; Döhner *et al.*, 2010; Papaemmanuil *et al.*, 2016).

2.4 Age and acute myeloid leukemia

AML affects all ages, including children and adults. It is one of the leading causes of death in children with cancer, and recurrence is the leading cause of death for children with AML. During induction and consolidation therapy, children and adults die from relapse (up to 35% and 99%, respectively) and treatment-related death (Steliarova-Foucher *et al.*, 2017; Chaudhury *et al.*, 2018). Steliarova-Foucher et al. (2017) examined the patterns of pediatric cancer incidence around the globe and discovered that leukemia accounted for 36.1% of cases in children under four and affected 15.4% of patients between the ages of 15 and 19. Although the fundamental processes of malignant transformation across all ages and the tumour spectrum are similar, childhood tumours differ considerably from adult tumours (Murphy *et al.*, 2013). The hematopoietic (40%), central nervous system (25%), and solid tumours (35%) make up the different groups of childhood cancers (Murphy *et al.*, 2013).

Genetic differences between adult and pediatric cancers are revealed by genomic sequencing of tumours. It has been established that childhood cancers have a low mutation rate compared to adult tumours (Ma *et al.*, 2018; Savary *et al.*, 2020). This may be due to environmental carcinogens, which only make a minimal contribution, and the embryonal origin of pediatric cancers (Alejandro Sweet-Cordero and Biegel, 2019). The type of genomic alteration also observed in pediatric tumours differs from adult tumours. Such alterations include copy number variations, gene fusions, and chromoplexy, which are prognostic of many pediatric cancers (Alejandro Sweet-Cordero and Biegel, 2019).

The burden and risk of potential complications of cancer treatment are more profound in children because of long-term complications than in adults, who would primarily experience short-term complications (Kattner *et al.*, 2019). However, pediatric and adult cancers would benefit from more accurate diagnosis and prognosis and the development of novel, less toxic therapies.

Soheil Meshinchi, MD, PhD, of Fred Hutchinson Cancer Research Centre, emphasised that AML in younger patients and AML in older patients are entirely distinct diseases. The senior investigator further reported, it is similar to comparing breast cancer to colon cancer. Despite the large degree of similarity between the diagnostic and therapeutic recommendations for AML in children and adults, there are significant variances in the diagnostic criteria and disease management that call for age-specific strategies (Creutzig *et al.*, 2012).

While there are numerous clinical and molecular parallels between pediatric and adult AML with a continuum across the age range, many AML features relate to disease onset. These include chromosomal abnormalities, gene mutations, and differentiation lineage. After treatment, AML cells that relapse are chemoresistant. Genetic profiling can uncover age-specific prognostic indicators and targetable molecular vulnerabilities in AML cells from adults and children pre- and post-chemotherapy (Aung *et al.*, 2021).

Comprehensive genomic profiling, which employs DNA and RNA sequencing, has improved knowledge of oncogenic mutations in AML and variations that can serve as prospective therapeutic intervention targets. Increased understanding of the variability of AML suggests that pediatric and adult AML exhibit considerable biologic variations (Tarlock *et al.*, 2018).

2.4.1 Pediatric acute myeloid leukemia

Pediatric AML is the second most common leukemia in children. Although survival rates for childhood ALL have exceeded 90%, treatment for pediatric AML has lagged. This is likely due to the heterogeneity of pediatric AML causes and the absence of innovative treatment methods until recently. Intensified conventional chemotherapy and improved supportive care have increased survival to nearly 60%. New prognostic indicators and targeted treatments may improve survival outcomes (Egan *et al.*, 2021). AML in children is diverse and requires comprehensive therapy. Over the past few decades, low-risk AML outcomes have improved, while high-risk AML remains poor. Improved molecular diagnosis, risk stratification, and supportive therapy are necessary to improve the outcomes of pediatric AML patients at high-risk (Egan *et al.*, 2021).

Significant numbers of AML cases have somatic mutations in genes already known to affect hematopoiesis, and the presence of these mutations is linked to specific clinical outcomes. Numerous mutations have been implicated in AML pathogenesis, with the number rising with discovery phase initiatives. At present, mutations in three genes (*FLT3*, *NPM1*, and *CEBPA*) have been shown to have clinical implications in childhood AML and have been incorporated in clinical trials as prognostic markers, therapeutic targets, or both (Tarlock and Meshinchi, 2015). Also, a study that was conducted on a gene expression profile pediatric dataset found that the upregulation of *FHL1*, *SPNS3*, and *MPZL2* were associated with poor outcome in a patient with *FLT3-ITD* mutation (Eshibona *et al.*, 2022).

2.4.2 Adult acute myeloid leukemia

The incidence of AML, a blood and bone marrow malignancy, rapidly increases in people aged 60 and older. As a result, the disease is the most prevalent and severe type of acute leukemia in adults and often worsens rapidly if left untreated. It has been reported that males are affected at a greater rate than females across all age groups (Heuser *et al.*, 2020). The adult leukemia statistics in the United States (US) report an age-adjusted incidence of 3.6 per 100,000 per year and a median age of 69 years at diagnosis. People with AML who are older than 75 still have relatively poor survival rates. In several population-based studies, patients over 60 years old had a 3-year survival rate of just 9-10% and a 5-year survival rate of only 3-8%, in comparison to the 5-year survival rates of up to 50% for younger patients (Oran and Weisdorf, 2012).

In AML, cytogenetic and molecular genetic aberrations frequently coexist in leukemic cells; they are not mutually exclusive. The average age of diagnosis for AML is 70, making it a disease of the elderly. The prevalence of adverse cytogenetic abnormalities increases with age, and the prognosis with standard treatment worsens with increasing age within each cytogenetic group. A high proportion of adult AML confers adverse outcomes (Kumar, 2011; Burd *et al.*, 2020) due to the presence of numerous cytogenetic and molecular abnormalities in patients tumour karyotypes. For example, loss of chromosome segments 5q, 17p, 7q, and others contribute to tumour cell survival, genomic instability, and thus poor outcome (Mrózek, 2022).

A study that aimed to investigate the relationship between *BAALC* gene expression and comprehensive molecular and clinicopathologic features in AML found that *BAALC* overexpression was associated with *CD34* positivity on leukemic blasts, the absence of *NPM1* mutation, the presence of *RUNX1* gene mutation, and poor patient outcomes, particularly in *NPM1*-wild type/*FLT3-ITD* negative adult CN-AML patients (Verma *et al.*, 2022).

Recommendations for AML patient's treatment vary depending on whether they are 60 years of age or younger (Kumar, 2011). Recent approvals of glasdegib and venetoclax, two medicines designed exclusively for the treatment of elderly people, have generated enthusiasm in the medical community. The debate remains, however, as to whether healthy older people should get combination therapy with newer drugs, given that rigorous chemotherapy is the only treatment that has proven potential to produce long-term disease-free survival (Luger, 2019).

2.5 Cytogenetics of leukemia

According to many studies, the majority of patients with AML have acquired chromosomal abnormalities (Hackl, Astanina and Wieser, 2017; Pourrajab *et al.*, 2020). In AML, numerous recurrent karyotypic abnormalities have been identified and continue to be identified, including changes in chromosomal number, rearrangements, large insertions, and large deletions. A large number of chromosomal aberrations were identified in AML, including t(8;21)(q22;q22), inv(3)(q21q26), monosomy 5/del(5q), t(6;9)(p22;q34), monosomy 7/del(7q), trisomy 8, t(15;17)(q24;q21), and complex karyotypes (Lazarevic and Johansson, 2020).

Cytogenetic research cleared the path for molecular analysis that revealed the genes involved in the leukemogenesis process. Moreover, chromosome abnormalities, whether or not they have been molecularly defined, have been demonstrated to be diagnostic and prognostic malignancy indicators (Mrózek *et al.*, 1997). Current therapeutic protocols require the detection of prognostic mutations such as *FLT3-ITD***,** *NPM1***,** and *CEBPA*, particularly in cases with normal karyotypes (Quessada *et al.*, 2021). Routine cytogenetic testing is recommended for all cases of AML, and molecular and cytogenetic studies must be integrated for risk stratification at diagnosis to improve therapeutic strategies (Gupta, Mahapatra and Saxena, 2019).

The World Health Organization (WHO) and the European Leukemia Net developed molecular classification and risk stratification schemes for AML based on the fact that most cytogenetic abnormalities do not overlap and have different links to clinical presentation, treatment response, relapse rates, and overall survival (Moarii and Papaemmanuil, 2017).

2.6 Classification of acute myeloid leukemia

Diseases classification is crucial and widely used, significantly contributing to disease management. At the time of diagnosis, it is essential to know the type or subtype of the condition, which helps physicians to use and follow the appropriate treatment approach. In the subsequent sections, we discuss the widely used classification systems for AML and their relevance to AML prognosis.

2.6.1 French-American-British classification

The French-American-British (FAB) classification divides AML patients into sub-groups based primarily on cytochemical and conventional morphological methods, with the correlation between the subgroups and laboratory findings, prognosis, and response to treatment. The classification divides AML into six sub-groups (M1-M6) defined based on the differentiation along one or more cell lines and the maturation degree of cells. Thus, M1, M2, and M3 show largely granulocytic differentiation and vary in the extent and form of granulocytic maturation; M4 demonstrates both granulocytic and monocytic differentiation; M5 mostly monocytic differentiation, and M6 predominantly erythroblastic differentiation (Bennett *et al.*, 1976). Later, immunological markers were employed to establish two additional AML subtypes, M0 and M7, both Sudan B Black negative but not lymphoblastic (Table 2.1) (Bennett *et al.*, 1985, 1991; Catovsky *et al.*, 1991; Segeren and Vantveer, 1996).

FAB classification is not outdated for use in the diagnosis of AML; however, different processes are necessary for refining the diagnosis and determining the patient prognosis using this classification. Additionally, certain AML FAB subtypes are associated with specific chromosomal abnormalities that have prognostic significance (Table 2.1). No specific chromosomal pattern is found in AML-M0 (Segeren and Vantveer, 1996).

The assignment of AML subtypes in every single case mandates an integrated approach where the following need to be considered: clinical history, history of therapy, cytogenetic studies (karyotyping and fluorescence in situ hybridisation), molecular results, next-generation sequencing (NGS) and gene panels, and examination of extramedullary tissue. Molecular genetic analysis and cytogenetic studies, including fluorescence in situ hybridisation testing, are essential in the AML diagnostic workup (Walter *et al.*, 2013).

Table 2.1: Correlation between FAB classification and chromosomal with prognostic value. Adapted from (Segeren and Vantveer, 1996).

FAB classification	Chromosomal translocation	Prognostic relevance
M2	t(8:21) (q22:q22)	Fair to good
M3	t(15;17) (q22;q21)	Fair to good
M4eo	inv(16) (p13;q22) t(16;16) (p13;q22) (p13;q22)	Good
M5	t(9;1 l) (p21;q23)	Poor
M4-M5	t(1 lq23)	Poor
M2, M4	t(6;9) (p23;q34)	Poor
M5	t(8;16) (p11;p13)	Undetermined
M4	inv(3) (q2l;q26) /t(3;33) t(1;3) (p36;q21)	Undetermined
M7	t(1;22) (p13;q13)	Undetermined

2.6.2 World Health Organisation classification

The World Health Organization (WHO) classification system (Table 2.2) aims to establish clinicopathologic entities using a combination of clinical characteristics, morphology, immunophenotype, cytogenetics, and molecular genetics. This technique was initially utilised to define disease entities (Harris *et al.*, 1999). A new WHO classification of hematologic malignancies has been under development since 1995 by the European Association of Pathologists (EAHP) and the Society for Hematopathology. Neoplasms classified under this heading include mast cell, lymphoid, myeloid, and histiocytic neoplasms (Harris *et al.*, 1999).

Within the category of AML, four main groups are recognised: (i) AML with recurrent cytogenetic translocations, (ii) AML with myelodysplasia-related features, (iii) therapy-related AML and MDS, and (iv) AML not otherwise categorised (Harris *et al.*, 1999). The WHO classification was based mostly on adult patient data and included most, but not all, pediatric age-specific cytogenetic change subgroups, and there is no evidence for the applicability in children. In the literature, there are no studies conducted using the 2016 WHO classification, while very few studies used the 2008 WHO classification (Nunes *et al.*, 2019).

One of the biggest challenges in revising the WHO classification of AML was how to include important and recently discovered genetic aberrations while adhering to the WHO principle of defining homogeneous, biologically relevant, and mutually exclusive entities based not only on prognostic value but also on morphologic, clinical, phenotypic, and other unique biological properties (Vardiman *et al.*, 2009). This was especially problematic for the most common and prognostically significant mutations currently identified in cytogenetically normal AML, namely mutant *FLT3*, *NPM1*, and *CEBPA* (Vardiman *et al.*, 2009).

Table 2.2: The WHO classification and cytogenetic abnormalities. Adapted from (Arber *et al.*, 2016).

Type of AML	Inversion and/or translocation	Gene Mutation
AML with recurrent genetic abnormalities		
AML with	t(8;21) (q22;q22.1)	*RUNX1-RUNX1T1*
AML with	inv(16)(p13.1q22)or t(16;16)(p13.1;q22)	*CBFB-MYH11*
APL (acute promyelocytic leukemia)with	t(15:17)	*PML-RARA*
AML with	t(9;11) (p21.3;q23.3)	*MLLT3-KMT2A*
AML with	t(6;9) (p23;q34.1)	*DEK-NUP214*
AML with	inv(3) (q21.3q26.2) or t(3;3)(q21.3;q26.2)	*GATA2*, *MECOM*
AML (megakaryoblastic) with	t(1;22) (p13.3;q13.3)	*RBM15-MKL1*
AML with		*NPM1*

AML with		Biallelic mutationsof *CEBPA*
AML with myelodysplasia-related changes		
Therapy-related myeloid neoplasms		
AML, not otherwise specified (NOS)		
AML with minimal differentiation		
AML without maturation		
AML with maturation		
Acute myelomonocytic leukemia		
Acute monoblastic/monocytic leukemia		
Pure erythroid leukemia		
Acute megakaryoblastic leukemia		
Acute basophilic leukemia		
Acute panmyelosis with myelofibrosis		
Myeloid sarcoma		
Myeloid proliferations related to Down syndrome		
Blastic plasmacytoid dendritic cell neoplasm		
Acute leukemias of ambiguous lineage		
Acute undifferentiated leukemia		
Mixed phenotype acute leukemia (MPAL) with	t(9;22) (q34.1;q11.2)	*BCR-ABL1*
MPAL with	t(v;11q23.3)	*KMT2A* rearranged

2.6.3 European LeukemiaNet classification

European LeukemiaNet (ELN) establishes a risk classification based on cytogenetic and molecular aberrations (Döhner *et al.*, 2010). The definitions of four AML risk groups (favorable, intermediate-I, intermediate-II, and adverse) were shown to significantly predict outcomes, mainly in patients treated with consolidation chemotherapy regimens (Grimm *et al.*, 2020). In 2017, an updated ELN risk classification was released that incorporated the most recent insights into the molecular architecture of AML and its prognostic significance in ELN2017. Comparing the ELN2010 classification to the ELN2017 classification, only three risk groups were defined: favorable, intermediate, and adverse (Döhner *et al.*, 2017)

ELN2010 separated Intermediate-I and Intermediate-II patients from one another not by prognostic factors but by genetic characteristics (Table 2.3). Although the individual patient outcome of both intermediate groups was not taken into consideration in the classification of all patients, later studies showed an association between AML genetic variation with prognosis (Röllig *et al.*, 2011; Mrózek *et al.*, 2012). Patient outcome was dependent on age, with younger patients between 18 and 60 years old showing a statistically significant difference in OS between the two groups, with Intermediate-II patients having longer OS than Intermediate-I patients. However, in older patients (over 60 years old), there was no difference in prognosis between the two groups (Döhner *et al.*, 2017). The following ELN2017 merged the two groups I and II into a single Intermediate group based on the lack of prognostic difference among patients over 60 who constitute the majority of AML cases (Döhner *et al.*, 2017). Despite the fact that younger patients showed a prognostic difference between Intermediate-I and II, the latest ELN classification overlooked this difference and placed both patients in the same risk category, which may negatively impact the therapeutic decisions, response to therapy, and therefore patient outcomes.

Additionally, a comparative study was performed between ELN2017 and ELN2022 and found that risk classification of the latter version identified a larger group of adverse-risk patients at the cost of slightly reduced prognostic accuracy compared to ELN2017. Prognostic accuracy may have decreased due to the inclusion of patients from the intermediate risk group whose OS at 5 years was higher than those initially in the adverse group (Rausch *et al.*, 2022). For the same reason, this also applies to the intermediate group with the inclusion of patients from adverse and favorable groups as of the latest changes in the ELN classification system. Thus, this points to a problem in the classification system even after the new changes that were supposed to improve prognostic accuracy, not decrease it.

Most of AML patients are categorised into an intermediate-risk group with variable prognostic outcomes. Thus, a selection of an appropriate consolidation therapy regimen remains challenging; this demonstrates the need for improved AML patient stratification (Docking *et al.*, 2021). In terms of clinical impact, adult classification systems, such as those defined by ELN, cannot be completely transferred to the classification of childhood AML because the cytogenetic (and genomic) landscapes of pediatric and adult AML and the cytogenetic risk associations are different according to age (Quessada *et al.*, 2021).

Table 2.3: The significant amendments for ELN classification of AML. Three genetic groups were used in AML risk classification, and the corresponding genetic abnormalities were reported and/or updated. Adapted from (Döhner *et al.*, 2010, 2017, 2022).

ELN classification version	**Genetic group**			
	Favorable	**Intermediate**		**Adverse**
		Intermediate-I	**Intermediate-II**	
ELN2010	t(8;21) (q22;q22); *RUNX1-RUNX1T1* inv(16) (p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* Mutated *NPM1* without *FLT3-ITD* (normal karyotype) Mutated *CEBPA* (normal karyotype)	Mutated *NPM1* and *FLT3*-ITD (normal karyotype) Wild-type *NPM1* and *FLT3*-ITD (normal karyotype) Wild-type *NPM1* without *FLT3-ITD* (normal karyotype)	t(9;11) (p22;q23); *MLLT3-MLL* Cytogenetic abnormalities not classified as favorable or adverse	inv(3) (q21q26.2) or t(3;3)(q21;q26.2); *RPN1-EVI1* t(6;9) (p23;q34); *DEK-NUP214* t(v;11) (v;q23); *MLL* rearranged -5 or del(5q); -7; abnl(17p); complex karyotype
ELN2017	t(8;21) (q22;q22.1); *RUNX1-RUNX1T1* inv(16) (p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* Mutated *NPM1* without *FLT3-ITD* or with *FLT3-ITD* low allelic ratio Biallelic mutated *CEBPA*	Mutated *NPM1* and *FLT3-ITD* high allelic ratio Wild type *NPM1* without *FLT3-ITD* or with *FLT3-ITD* low allelic ratio (without adverse risk genetic lesions) t(9;11) (p21.3;q23.3); *MLLT3-KMT2A* Cytogenetic abnormalities not classified as favorable or adverse		t(6;9) (p23;q34.1); *DEK-NUP214* t(v;11q23.3); *KMT2A* rearranged t(9;22) (q34.1;q11.2); *BCR-ABL1* inv(3) (q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2,MECOM(EVI1)* -5 or del(5q); -7; -17/abn(17p) Complex karyotype, monosomal karyotype

			Wild type *NPM1* and *FLT3-ITD* high allelic Mutated *RUNX1* Mutated *ASXL1* Mutated *TP53*
ELN2022	t(8;21) (q22;q22.1) / *RUNX1*::*RUNX1T1* inv(16) (p13.1q22) or t(16;16) (p13.1;q22)/*CBFB*::*MYH11* Mutated NPM1, without *FLT3-ITD* bZIP in-frame mutated *CEBPAk*	Mutated *NPM1*, with *FLT3-ITD* Wild-type *NPM1* with *FLT3-ITD* (without adverse-risk genetic lesions) t(9;11) (p21.3;q23.3)/*MLLT3*::*KMT2A* Cytogenetic and/or molecular abnormalities not classified as favorable or adverse	t(6;9) (p23.3;q34.1)/*DEK*::*NUP214* t(v;11q23.3)/*KMT2A*-rearranged t(9;22) (q34.1;q11.2)/*BCR*::*ABL1* t(8;16) (p11.2;p13.3)/*KAT6A*::*CREBBP* inv(3) (q21.3q26.2) or t(3;3)(q21.3;q26.2)/ *GATA2*, *MECOM*(EVI1) t(3q26.2;v)/*MECOM*(EVI1)-rearranged 25 or del(5q); 27; 217/abn(17p) Complex karyotype, monosomal karyotype Mutated *ASXL1*, *BCOR*, *EZH2*, *RUNX1*,

			SF3B1, *SRSF2*, *STAG2*, *U2AF1*, and/or *ZRSR2* Mutated TP53a

2.7 Treatment of acute myeloid leukaemia

Chemotherapy accounted for the first-line treatment of AML, depending on many factors, mainly on the genetic background of inhibitor-based therapy. Genetic defects are considered the most critical factors in determining the response to chemotherapy and its outcome. For instance, the better outcome for Gemtuzumab ozogamicin was noticed in the favorable and intermediate risk group of AML. Significant progress has been made in treating younger adults (Thol and Schlenk, 2014; Short, Rytting and Cortes, 2018). The prospects for elderly patients have remained dismal, with median survival times of only a few months. This difference is related to comorbidities associated with ageing and disease biology. Current efforts in clinical research focus on the assessment of targeted therapies. Such new approaches will probably increase the cure rate (Short, Rytting and Cortes, 2018). Most adults with acute leukemia have AML, which has a poor outcome, especially in older patients. Its cytogenetic and molecular abnormalities make it diverse. These abnormalities classify patients into prognosis categories and are therapeutic targets (Huerga-Domínguez *et al.*, 2022)

In recent years, the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have approved a number of new medications for AML as a result of breakthroughs in basic and translational research and the development of new target therapies. Primary refractory, relapsed, high-risk patients who are ineligible for allogeneic stem cell transplantation (alloSCT) still lack access to effective treatment options. Consequently, numerous experimental medications are now being studied (Huerga-Domínguez *et al.*, 2022)

For a long time, the only treatment option for "fit" AML patients who had recently been diagnosed was intensive chemotherapy based on cytarabine and anthracycline ("7+3"). Allogeneic stem cell transplantation also made a big difference in how long eligible patients with intermediate or high-risk AML survived (Huerga-Domínguez *et al.*, 2022). Midostaurin is a first-generation, orally administered, multitargeted kinase inhibitor. Combined with regular chemotherapy, it is currently the treatment of choice for patients newly diagnosed with *FLT3*-mutated AML who are "fit." (Thol, 2021).

2.8. Cancer biomarkers

Biomarkers are essential in disease diagnosis, prognosis, and predictive and treatment outcomes. While biomarkers help differentiate physiological and pathological mechanisms, they are equally important in assessing disease response to a medications therapeutic outcome, disease progression, and exploring disease mechanisms (Dhama *et al.*, 2019). A diagnostic biomarker is a biological marker used in medical diagnosis to either confirm the presence of a disease or condition or to positively identify a patient as having a certain subtype of that disease. As we enter the precision medicine age, this type of biomarker will evolve considerably. Such biomarkers may be used to identify people with a disease and redefine the disease's classification. For example, cancer detection is moving toward a molecular and imaging-based classification rather than a largely organ-based classification scheme (Califf, 2018).

Prognostic biomarkers can distinguish many stages of a disease and determine the course of therapy that should be applied to a particular patient after primary treatment. The aim of using prognostic biomarkers, which provide information on the overall cancer outcome in patients, is to facilitate cancer progression assessment, usually with no need to put invasive methods into use (Nalejska, Mączyńska and Lewandowska, 2014). Predictive biomarkers help to optimise therapy decisions, as they provide information on the likelihood of response to a given chemotherapeutic. Among the prognostic factors that identify patients with different outcome risks (e.g., recurrence of the disease), the following factors can be distinguished: somatic and germline mutations, changes in DNA methylation that lead to the enhancement or suppression of gene expression, the occurrence of elevated levels of microRNA capable of binding specific messenger RNA (mRNA) molecules, which affects gene expression, as well as the presence of circulating tumour cells in the blood, which leads to a poor prognosis for the patient (Nalejska, Mączyńska and Lewandowska, 2014). Clinical prediction models influence many medical choices across many clinical disciplines, and they are often used in oncology, for example, to evaluate the risk of getting cancer, inform cancer diagnosis, forecast cancer outcomes and prognosis, and guide treatment decisions (Dhiman *et al.*, 2022).

Molecular diagnostics of chronic myeloid leukemia, colon, breast, and lung cancer, and, more recently, melanoma use biomarkers for personalised oncology. They are successfully employed in assessing the potential benefits of targeted therapy or the toxicities of the chemotherapeutic agents used in the treatment (Nalejska, Mączyńska and Lewandowska, 2014). In medicine, mRNA transcripts are being developed as molecular biomarkers to diagnose and treat many diseases. These biomarkers offer the early and more accurate prediction and diagnosis of disease and progression and the ability to identify individuals at risk (Sunde, 2010).

2.9 Biological data

Biological data is often characterised by huge size. There are four important data generated and collected at biological sites, DNA, RNA, Protein Sequences, and Micro Array images. The first three datatypes are text data, and the last is a digital image. This large, vast, and complex amount of biological data needs to be stored, accessed, and manipulated efficiently and powerfully (Chowdhary *et al.*, 2016). Large amounts of cancer data have been collected and are available to the medical research community. So it was necessary to build databases such as sequence databases, microarray databases, genome databases, protein structure databases, and many more (Kourou *et al.*, 2015). Additionally, the complexity of the data and the amount of available data required the development of the bioinformatics field, which includes powerful tools to analyse and draw scientific insights.

2.10 Databases

Database plays an essential role in current bioinformatics research as they provide highly curated data stored in a way that makes it more accessible, reusable, and stable for the long term, such as The Cancer Genome Atlas (TCGA), protein gene atlas, National Centre for Biotechnology Information (NCBI), gene expression omnibus (GEO). There are various databases that researchers can use for research purposes and to understand the biological mechanism of any disease of interest under systematic investigation. In the following subsections, we describe databases that have been used in this study.

2.10.1 The Cancer Genome Atlas

One of the most ambitious and effective cancer genomics studies is TCGA. TCGA has produced, analysed, and made public methylation, genomic sequence, gene expression, and copy number variation data on over 11,000 people with over 30 cancer types (Wang, Jensen and Zenklusen, 2016). The TCGA researchers have so far collected a broad range of genomic data on individual cancer types, yielding a better understanding of each tumour's biology and pathology, resulting in the development of specific treatment strategies (Tomczak, Czerwińska and Wiznerowicz, 2015). With the ongoing decrease in cost for NGS and other high throughput molecular characterisation methods, many datasets are generated and provided for public access on web portals for use in the field of cancer research (Deng *et al.*, 2016).

2.10.2 Gene Expression Omnibus

Demand for a public archive for high-throughput gene expression data prompted the Gene Expression Omnibus (GEO) initiative. GEO's flexible and open architecture allows the submission, storage, and retrieval of disparate datasets from high-throughput gene expression and genomic hybridisation research (Edgar, Domrachev and Lash, 2002). The resource allows for the preservation of raw data, processed data, and indexed metadata (Barrett *et al.*, 2013). The collection of the GEO database is over 94,000 datasets and more than 2 million samples. This is a great resource that, with suitable methodology and tools, can be used to combine gene expression data for biomarker discovery applications (Toro-Domínguez *et al.*, 2019). Additionally, new investigations are being prompted by the high-throughput data made available by GEO. Moreover, data has been reanalysed and used in thousands of publications to form and test hypotheses (Clough and Barrett, 2016).

2.10.3 Kyoto Encyclopedia of Genes and Genomes pathway maps

As part of the Japanese Human Genome Program, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database project was founded in 1995. Anticipating the need for a reference knowledge base for biological interpretation of genome sequence data and biological pathways, the primary objective of the KEGG pathway was to find connections between collective sets of genes in the genome, and high-level cell and organism functions widely used (Kanehisa *et al.*, 2016). The networks of KEGG Orthology nodes that make up the KEGG pathway maps, BRITE hierarchies, and KEGG modules represent the high-level functions of the cell and the organism. The KEGG GENES database currently contains annotations for over 4000 complete genomes with KEGG Orthology, which can be used as a reference data set for assigning KEGG Orthology and subsequently reconstructing KEGG pathways and other molecular networks (Kanehisa *et al.*, 2016). The KEGG database contains information on the progression of various types of cancer as signalling pathway combinations. Signalling pathways are the molecular interactions and reactions that transport signals from the outside to the cell's nucleus, where transcriptional regulation occurs. Signalling pathways such as MAPK, Wnt, and TGF-beta signalling have been extensively studied in the context of cell proliferation. Cancer is the only human disease for which pathway information of disease progression is available, i.e. from normal tissue to the advanced tumor phase in KEGG. KEGG cancer pathways differ in including members of multiple signalling pathways within a single pathway. This is because they contain information on various stages of cancer, and each stage involves different signalling pathways (Dalkic *et al.*, 2009).

2.11 Bioinformatics

The subject of bioinformatics, or systems biology, which is the integration of computer and biological scientific disciplines, has become a crucial instrument for the organisation and analysis of a large quantity of biological data. The primary objective of bioinformatics is to uncover vital biological information hidden within a mass of raw data to detect significant trends or patterns, ultimately leading to the identification of new biomarkers for diagnostic and therapeutic reasons (Bayat, 2002; Jiang *et al.*, 2022). Modern bioinformatics emerged recently to assist with NGS data analysis. De novo sequence assembly, biological sequence databases, and substitution models laid the groundwork for bioinformatics in the early 1960s. Later, DNA analysis emerged owing to simultaneous breakthroughs in (i) molecular biology procedures, which made DNA manipulation and sequencing simpler, and (ii) computer science, which produced smaller, more powerful computers and bioinformatics-specific software. Sequencing technology advances and cost reductions caused exponential data growth in the 1990s and 2000s (Gauthier *et al.*, 2019).

Bioinformatics is thus applied in cancer research to understand metabolisms, signalling, communication, and proliferation in cancer incidence. Emerging from the intersection of clinical informatics, bioinformatics, medical informatics, computing, mathematics, and omics research, the field of clinical bioinformatics can aid in the disease identification, effective treatment, and prognostication of cancer patients (Wu, Rice and Wang, 2012).

2.12 Bioinformatics analysis tools

Bioinformatics provide numerous packages, tools, and algorithms based on mathematical models created in R, Python and other programming languages to analyse and draw scientific conclusions from massive amounts of biological data.

2.12.1 Differential Gene Expression

Correctly identifying differentially expressed genes (DEGs) between specific conditions is critical to understanding phenotypic variation (Figure 2.2). High-throughput transcriptome sequencing, such as (RNA-Seq) has become the primary option for these studies. Thus, the number of methods and software for differential expression analysis from RNA-Seq data has also increased rapidly (Costa-Silva, Domingues and Lopes, 2017). Different packages were developed in the R language, such as limma, edgeR, and DESeq2 (Robinson, McCarthy and Smyth, 2009; Love, Huber and Anders, 2014; Ritchie *et al.*, 2015).

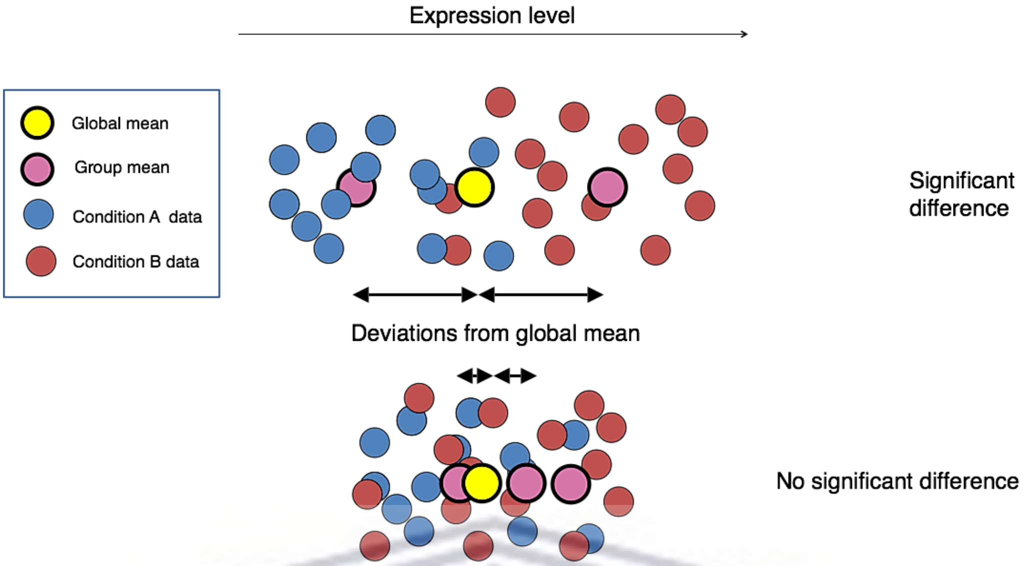

Figure 2.2: Illustrate significant and insignificant differences in expression levels between two conditions, A in blue and B in red. The bottom image shows one cluster with samples from both conditions. There is no significant difference in gene expression between conditions A and B. The top image shows two clusters mainly composed of samples from one condition. Conditions A and B have different expression patterns (https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html).

Modern high-throughput sequencing technologies are increasingly replacing traditional methods to quantify RNA expression levels (RNA-Seq). Due to advances in rapid sequencing technology and declining prices, whole profiling of gene expression levels is now possible, with repercussions throughout the life sciences. This information is already being embraced for clinical usage (Rapaport *et al.*, 2013).

In a biological system, there is a need for advancements in identifying genes related to a trait to understand complex conditions better. This can be achieved by enhancing our knowledge about gene expression through statistical models to perform statistical analysis of gene expression profiles to quantify gene expression and the sequencing reads aligned to a known reference genome sequence (e.g. Tophat and Star aligners). The proportion of reads matching a given transcript is used as quantification of its expression level (e.g. Salmon tool), followed by statistical testing of differences in quantification values between samples (e.g. DESeq2 and edgeR) (Dobin *et al.*, 2013; Kim *et al.*, 2013; Anjum *et al.*, 2016; Patro *et al.*, 2017).

DGE analysis is widely applied for biomarker discovery for different types of cancer. In breast cancer, DGE was utilised to identify gene signatures associated with a worse prognosis (Pan *et al.*, 2017). High expression of four genes was linked to the early stage of papillary thyroid carcinoma (Han *et al.*, 2018). Furthermore, microarrays and NGS have generated molecular signatures for prostate cancer that differentiate between malignant and non-malignant states and are considered promising prostate cancer biomarkers (Myers *et al.*, 2015).

Several studies have used meta-analysis approaches to discover DEGs between cancer patients and controls using microarray data. These techniques may be used to establish gene expression signatures in a single cancer type or to search for common expression patterns across several cancer types (Kais *et al.*, 2022). In 2004, Rhodes and colleagues evaluated 40 published cancer microarray datasets, including 38 million gene expression values from over 3,700 cancer samples. This led to the identified meta-signature of neoplastic transformation by integrating microarray data and analysis from a variety of cancer types. The aforementioned defined a transcriptional program that is almost always activated in cancer, irrespective of the origin of the cell (Rhodes *et al.*, 2004; Kais *et al.*, 2022).

2.12.2 Machine learning

The machine learning (ML) approach has become popular among medical researchers. These techniques can discover and identify patterns and relationships between them from complex datasets. At the same time, they can effectively predict future outcomes of a cancer type (Kourou *et al.*, 2015). ML is often described as providing more flexible modelling, the capacity to analyse a vast amount of data, non-linear and high-dimensional data, and the ability to simulate complicated clinical events (Dhiman *et al.*, 2022). ML is used in various disciplines, including disease diagnosis in health care. Many academics and practitioners demonstrate the promise of machine learning-based disease diagnosis, which is cost-efficient and time saving. The advancement of biomedical and translational research and the use of sophisticated statistical analysis and ML approaches are driving factors in the advancement of prognostic cancer prediction (W. Zhu *et al.*, 2020; Ahsan, Luna and Siddique, 2022).

2.12.3 Survival analysis

Survival analysis is a set of statistical processes for analysing data in which the outcome variable of interest is the time until an event happens. A part of the survival periods of interest is often unknown due to censoring, which is the no observation of the event of interest after a follow-up period. It is considered that patients who are censored have the same chances of surviving as those who continue to be monitored; hence, censoring is presumed to be non-informative (Clark *et al.*, 2003).

The survivor and hazard functions are often used to describe and model survival data. The survivor function depicts the chance that an individual lives from the moment of origin to some point after time t. It directly characterises the survival experience of a study cohort and is often estimated using the Kaplan-Meier curves. The log-rank test may be used to compare the survival curves of different groups, such as treatment arms. Given survival up to that point, the hazard function provides the immediate probability of experiencing an event (Clark *et al.*, 2003). In numerous cancer studies, the time to an event of interest is the primary outcome being evaluated. Survival time is the generic term for the period, which can refer to the time 'survived' from complete remission to relapse or progression as well as the time from diagnosis to death (Clark *et al.*, 2003). Many studies use survival analysis to find a prognostic biomarker for different cancers; in breast cancer, high expression of six genes was associated with poor prognosis in younger patients (Ingebriktsen *et al.*, 2022). Survival analysis combined with differential gene expression profiling has been performed and found that overexpression of *STAT6* and *SOX2* genes impacts the survival rate in prostate cancer (Mohammad *et al.*, 2022).

2.13 Summary

Intensive studies have been conducted on AML to understand the biological nature of the disease, and improved knowledge has been gained over time. The significant improvement was relating the cytogenetic, molecular abnormalities, morphology, and immunophenotype to the prognosis of AML, which led to the establishment of the AML risk classification. The existing classification systems are used for the clinical management of AML patients. However, all the established classifications fall short of providing a particular group of patients with an appropriate prognostic value due to the heterogeneity of AML. This has a negative impact on the accurate assessment of overall survival and therapeutic options for AML patients. Furthermore, these classification systems underwent multiple justifications and updates, as mentioned in section 2.6.

Notably, the prognosis of AML varies depending on the patient's age and genetic abnormalities, which further affects the therapeutic options. Moreover, the response to conventional treatment differs depending on the risk group. The advancement of genomic studies has enabled numerous researchers and organisations to overcome the challenge of risk classification and prognostic prediction. Similarly, this study exploits gene expression profiles to establish genetic signatures that could be used for prognosis and accurate risk classification in AML patients.

Chapter 3: Upregulation of FHL1, SPNS3, and MPZL2 predicts poor prognosis in pediatric acute myeloid leukemia patients with *FLT3-ITD* mutation

This is an original manuscript of an article published by Taylor & Francis in Leukemia & Lymphoma on March 2022, available at: https://doi.org/10.1080/10428194.2022.2045594.

Abstract

Chromosomal translocations and gene mutations are characteristics of the genomic profile of acute myeloid leukemia (AML). We aim to identify a gene signature associated with poor prognosis in AML patients with *FLT3-ITD* compared to AML patients with *NPM1/CEBPA* mutations. RNA-sequencing (RNA-Seq) count data were downloaded from the UCSC Xena browser. Samples were grouped by their mutation status into high and low-risk groups. Differential gene expression (DGE), machine learning (ML) and survival analyses were performed. A total of 471 differentially expressed genes (DEGs) were identified, of which 16 DEGs were used as features for the prediction of mutation status. An accuracy of 92% was obtained from the ML model. *FHL1*, *SPNS3*, and *MPZL2* were found to be associated with overall survival in *FLT3-ITD* samples. *FLT3-ITD* mutation confers an indicative gene expression profile different from *NPM1/CEBPA* mutation, and the expression of *FHL1*, *SPSN3*, and *MPZL2* can serve as prognostic indicators of unfavorable disease.

3.1 Introduction

Acute myeloid leukemia (AML) makes up about 20% of acute leukemia in pediatric patients (de Rooij, Zwaan and van den Heuvel-Eibrink, 2015). Common clinical symptoms of AML include leukocytosis, anemia, and thrombocytopenia. It is a very heterogeneous disease accounting for more than half of the deaths from leukemia (Szalontay and Shad, 2014; Chaudhury *et al.*, 2018). Deciphering disease heterogeneity at the molecular level is crucial for accurate diagnosis, treatment and prognosis, and identifying possible gene therapeutic targets requires deciphering the genetic patterns underlying the etiology of the disease. AML clonal expansion results from abnormal genetic and epigenetic changes in hematopoietic stem and progenitor cells that cause changes or impairment of important physiological processes, such as self-renewal, proliferation, and differentiation (Saultz and Garzon, 2016; Siveen, Uddin and Mohammad, 2017). Consequently, AML results in the insufficient generation of normal mature blood cells. In addition, AML is associated with multiple chromosomal translocations and mutations that are responsible for the disease pathology and influence disease prognosis. Three gene mutations are proven to be prognostically significant in AML, namely, *NPM1*, *CEBPA*, and *FLT3-ITD* (Torrebadell *et al.*, 2018). Mutations in *FLT3* are associated with a higher rate of relapse and unfavorability (Meshinchi *et al.*, 2006), while mutations in *NPM1* and *CEBPA* are associated with favorable survival prognoses (Fröhling *et al.*, 2004; Hollink *et al.*, 2009).

Of great interest to many research studies are the genome-wide detection of differentially expressed genes (DEGs) from two or more conditions of interest (Law *et al.*, 2016). Currently, one of the main techniques of choice used for gene expression profiling (GEP) is RNA-sequencing (RNA-Seq). GEP can be used to find many systematic differences between cancer and normal conditions, thereby defining new clinically relevant disease subtypes (Gerstung *et al.*, 2015). The European LeukemiaNet (ELN-2017) provides guidelines to stratify AML patients into good, intermediate, and adverse risk groups based on cytogenetics and mutation status of some genes, including *ASXL1*, *CEBPA*, *FLT3*, *NPM1*, *RUNX1*, and *TP53* (Döhner *et al.*, 2017).

Based on the results of a preliminary data analysis of high- and low-risk AML samples from the GDC TARGET AML dataset, the PCA analysis showed that samples with *NPM1*, *CEBPA*, and *FLT3-ITD* mutations were clustered together (Figure 3.1), although they have different survival rates. In addition, samples with t(8;21) and inv(16) variations were well separated into two other independent clusters (Figure 3.1). In line with these results, a 36-gene expression signature enabled the accurate classification of AML samples with t(8;21) and inv(16) mutations (Handschuh and Lonetti, 2019). However, segregation between AML samples with *NPM1* and *CEBPA* using GEP is less accurate, revealing that the expression pattern between the two types of samples is likely similar (Verhaak *et al.*, 2009). Therefore, the fact that *FLT3-ITD* samples are grouped with *NPM1* and *CEBPA* samples of favorable prognosis raises the question of why the survival of AML patients with *FLT3-ITD* is low? Thus, it is crucial to look for DEGs between AML *FLT3-ITD* on the one hand and AML *NPM1*/*CEBPA* on the other hand, which are potentially the cause of poor survival in *FLT3-ITD* samples.

This study aims to identify a gene expression signature that can differentiate the AML *FLT3-ITD* mutation from the AML *NPM1* and AML *CEBPA* mutations, which are implicated in poor survival. In addition, we examined the predictive value of this signature in risk classification. These novel gene expression signatures may assist clinicians and pathologists in patient diagnosis and assessment, thereby ensuring more accurate and individualized treatment options.

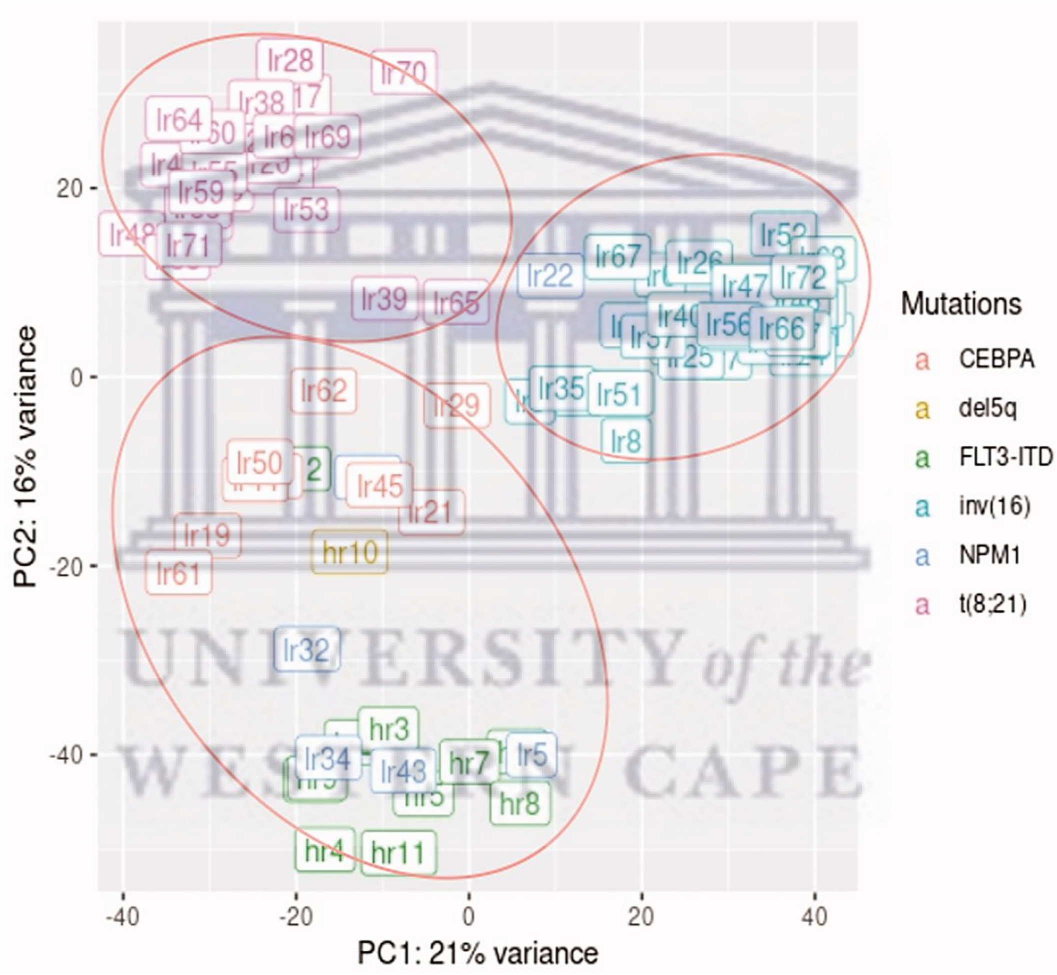

Figure 3.1: Clustering of the high- and low-risk AML samples based on their cytogenetics and genetic aberrations using principal component analysis (PCA).

3.2 Materials and Methods

The workflow describing the steps and methods undertaken in this study is illustrated in (Figure 3.2). It includes six essential steps: dataset retrieval, differential gene expression (DGE), PCA and clustering, machine learning (ML), Cox regression, and Kaplan–Meier's (K-M) analyses.

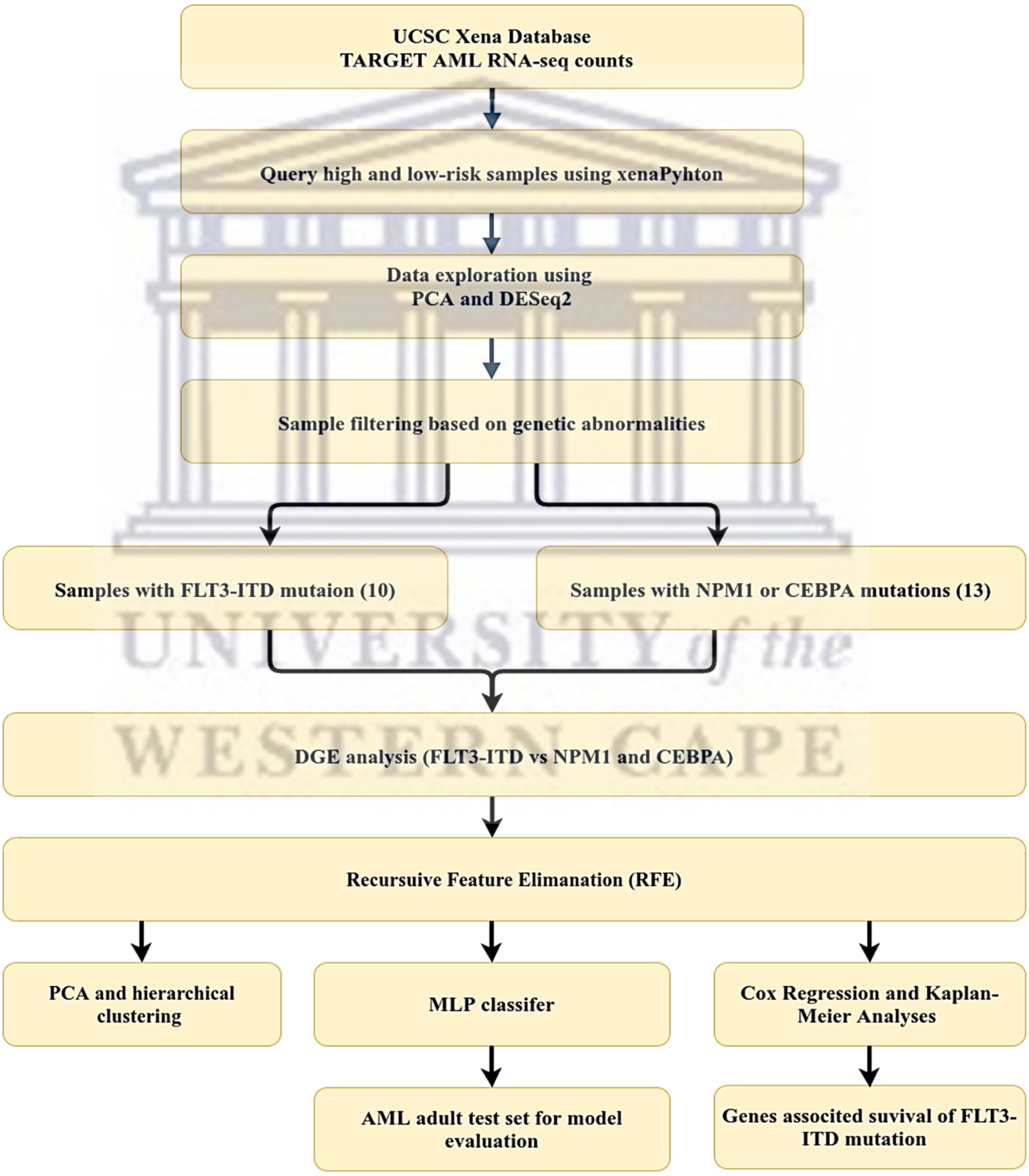
UCSC Xena Database
TARGET AML RNA-seq counts
Query high and low-risk samples using xenaPyhton
Data exploration using
PCA and DESeq2
Sample filtering based on genetic abnormalities
Samples with FLT3-ITD mutaion (10)
Samples with NPM1 or CEBPA mutations (13)
DGE analysis (FLT3-ITD vs NPM1 and CEBPA)
Recursuive Feature Elimanation (RFE)
PCA and hierarchical clustering
MLP classifer
Cox Regression and Kaplan-Meier Analyses
AML adult test set for model evaluation
Genes associted suvival of FLT3-ITD mutation

Figure 3.2: Workflow depicting the steps and methods used for the identification of genetic signature associated with poor prognosis in AML samples with *FLT3-ITD* mutation. AML: acute myeloid leukemia; *CEBPA*: CCAAT enhancer binding protein alpha; DGE: differential gene expression; *FLT3-ITD*: FMS-like tyrosine kinase 3-internal tandem duplications; MLP: multi-layer perceptron; *NPM1*: nucleophosmin-1; PCA: principal component analysis; TARGET: therapeutically applicable research to generate effective treatment.

3.2.1 Datasets

The therapeutically applicable research to generate effective treatment (TARGET) project employed a multi-omic strategy to molecularly characterize hard-to-treat pediatric cancers, including AML. TARGET data are accessible through the TARGET data matrix and the UCSC Xena Browser, a web-based visual integration and exploration tool for multi-omic data. The UCSC Xena Browser is a high-performance visualization, exploration, and analysis tool for multi-omic data of large public repositories and private datasets (Goldman *et al.*, 2018).

The TARGET AML dataset in Xena consists of a total of 187 samples with their accompanying clinical data (Table 3.1). Xena Python, a Python package implementing APIs to query and download data from Xena, was used to obtain gene expression in log2 transformed RNA-Seq counts format of the TARGET AML dataset (dataset ID: TARGET-AML.htseq_counts.tsv) from the GDC hub. The criteria used to query the dataset were the risk group (low and high) into which the samples were classified. The data samples were then filtered based on their cytogenetic abnormalities and genetic mutations. AML data samples with *FLT3-ITD*, *NPM1*, and *CEBPA* mutations were used for downstream analysis. Samples with multiple gene mutations were not considered, as multiple mutations may impact AML prognosis (Stölzel *et al.*, 2016). There is no precise risk classification of patients with *FLT3-ITD* and *NPM1* double mutation. The National Comprehensive Cancer Network considers them to have a favorable or intermediate risk. However, an unfavorable prognosis was manifested in this category of patients in recent studies (Huang *et al.*, 2019). Similarly, *CEBPA*-mutated patients with *WT1* mutations were reported to have an unfavorable risk (Wang *et al.*, 2022). Due to these uncertainties and changes in the risk classification of AML patients with multiple mutations, two samples, one with *FLT3-ITD* and *NPM1* double mutation and the other with both *CEBPA* and *WT1* mutations, were not considered.

Table 3.1: Relevant clinical and mutational variables of 187 pediatric samples from the TARGET database and their distribution by prognosis.

Variable	**Good,N =72**	**Intermediate, N = 93**	**Poor, N = 12**	**Unknown , N = 10**
Gender				
Female	37 (51%)	51 (55%)	5 (42%)	3 (30%)
Male	35 (49%)	42 (45%)	7 (58%)	7 (70%)
Vital status				
Alive	50 (69%)	37 (40%)	2 (17%)	2 (20%)
Dead	22 (31%)	56 (60%)	10 (83%)	8 (80%)
FLT3 Mutation	2 (2.8%)	4 (4.3%)	11 (92%)	0 (0%)
CEBPA Mutation				
Neg	63 (88%)	91 (98%)	12 (100%)	10 (100%)
Unknown	0 (0%)	2 (2.2%)	0 (0%)	0 (0%)

Pos	9 (13%)	0 (0%)	0 (0%)	0 (0%)
NPM1 Mutation				
Neg	66 (92%)	87 (94%)	9 (75%)	10 (100%)
Unknown	0 (0%)	6 (6.5%)	2 (17%)	0 (0%)
Pos	6 (8.3%)	0 (0%)	1 (8.3%)	0 (0%)
Overall Survival	2,080 (828, 2,626)	917 (462, 2,205)	398 (351, 772)	637 (444, 700)

3.2.2 Differential Gene Expression analysis

The gene expression values from the AML samples with *FLT3-ITD*, *NPM1*, and *CEBPA* were converted to raw counts using the mathematical expression $R=2^L - 1$, where R is the raw count, and L is the log2 normalized value. Filtering to remove low expressed genes was done using edgeR's filterByExpr function (Robinson, McCarthy and Smyth, 2009). Statistically, eliminating low expressed genes enables the mean-variance association in the data to be measured more accurately and reduces the number of computational checks conducted in downstream differentially expressed tests (Law *et al.*, 2016). The DGE analysis between *FLT3-ITD* and *NPM1/CEBPA* samples was performed using the DESeq2 package in R (Love, Huber and Anders, 2014). DESeq2 uses shrinkage estimators for dispersion and fold change for comparative DGE estimation (Love, Huber and Anders, 2014). Genes that met the criteria of an adjusted p value <.01 were considered significant and therefore differentially expressed.

3.2.3 Machine learning

To determine if the DEGs identified could serve as risk classification biomarkers in AML, a multi-layer perceptron (MLP) classifier was constructed using the pediatric DEGs and tested on adult AML samples, as explained below. MLP is a feed-forward artificial neural network (ANN) algorithm. It is implemented in Python under the Scikit-learn library (Pedregosa *et al.*, 2011) and imported using the MLPClassifer class. Prior to the classification task, a linear support vector machine and repeated stratified 10-fold cross-validation were used for recursive feature elimination (RFE). The RFE was applied to the DEGs to obtain the essential features for optimum model performance. The selected genes were then used as features for classification. Also, principal component analysis (PCA) and clustering analysis were applied to the expression values of these selected genes to assess their ability to discriminate the samples of the two AML groups. The PCA plot and clustering heatmap were generated using the ggplot2 and pheatmap R packages, respectively.

The features of the training set were extracted from the log2-transformed normalized counts of the pediatric AML samples. The test set was constructed from the adult AML dataset (dataset ID: TCGA-LAML.htseq_counts.tsv) obtained from the GDC hub in the UCSC Xena Browser. Construction of the test set followed the same process used to construct the training set, i.e. consider AML samples with *FLT3-ITD* mutation and AML samples with *NPM1*/*CEBPA* mutation. The features (genes) not present in the test dataset were removed from the training set and were not used in the ML classification. In addition, the gene expression values of the training and test sets were scaled using Scikit-learn’s StandardScaler function. The MLP model was then applied to predict the classification of the samples in the test set. The evaluation metric for the model performance was the accuracy of classification.

3.2.4 Cox’s Regression Analysis and Kaplan–Meier’s estimates

To determine among the genes from ML those that best correlated with patient's survival, we used a Cox regression model based on the Lasso algorithm of the glmnet R package (Friedman, Hastie and Tibshirani, 2010; Simon *et al.*, 2011; Tibshirani *et al.*, 2012). The model assigns each gene a regression coefficient value. Genes with a zero coefficient were eliminated, having no effect on survival. Prognostic genes with positive coefficients suggest that their upregulation signifies low survival in *FLT3-ITD* patients. For each gene with positive coefficient, a score value is calculated for each patient as a product of the expression value of the gene and its corresponding coefficient obtained from the Cox regression model. A median value was inferred from the patient scores. Each score was then compared to the median, and patients were assigned a status value of 1 or 0 depending on whether the score was above or below the median. According to patient status information, K-M estimates were then calculated for overall survival (OS). K-M curves were generated using the *ggsurvplot* function from the survminer R package.

3.3 Results

3.3.1 Differential gene expression analysis

Querying UCSC Xena Browser for TARGET AML samples using risk group as search criteria returned 12 high-risk samples and 72 low-risk samples. The distribution and clustering of samples by cytogenetic and mutational aberrations are shown in (Table 3.2 and Figure 3.1), respectively. Samples with *FLT3-ITD*, *NPM1*, and *CEBPA* mutations were selected for downstream analysis, comprising of 10 samples of *FLT3-ITD* mutation and 13 *NPM1*/*CEBPA* samples. All were clustered together in the exploratory analysis (Figure 3.1). Filtering of the gene expression data eliminated 38,980 low expressed genes and retained 21,503 genes for downstream analysis. The DGE analysis using DESeq2 identified 471 DEGs between the AML *FLT3-ITD* and AML *NPM1*/*CEBPA* groups. Of these, 208 genes were upregulated, while 263 genes were downregulated.

Table 3.2: Type of mutation, number of samples and risk category of TARGET high- and low-risk samples.

Type of mutation	Number of samples	Risk classification
CEBPA	9	Favorable
NPM1	6	Favorable
Del(5q)	1	Unfavorable
FLT3-ITD	11	Unfavorable
inv(16)	32	Favorable
t(8; 21)	25	Favorable

3.3.2 Machine learning

The RFE selected 22 genes out of 471 DEGs as the most critical features for classifying samples according to their mutational status (Table 3.3). These 22 DEGs accurately discriminated and clustered the TARGET samples into their respective mutation groups (Figure 3.4). However, some of these selected genes were not found in the adult test dataset and were not included in the pediatric training set used to create the MLP model. Sixteen DEGs (features) were thus used for the ML classification, including *FHL1*, *VWF*, *MPZL2*, *SPNS3*, *LINC00515*, *TCEA3*, *DCN*, *MACC1*, *ADD2*, *CCDC152*, *KCNA6*, *ADAMTS3*, *KCTD15*, *KCNMB4*, *HESX1*, and *IL3RA*. The training set was constructed based on the gene expression values of the TARGET samples from the *FLT3-ITD* group with 10 samples and the *NPM1/CEBPA* group with 13 samples. The test set was 25 samples, including five *FLT3-ITD* samples and 20 *NPM1/CEBPA* samples. Evaluation of the MLP classifier on the test set yielded 92% accuracy of samples correctly assigned to their corresponding mutational status. All *FLT3-ITD* samples were correctly predicted, while 18 of the 20 *NPM1* and *CEBPA* samples were correctly predicted.

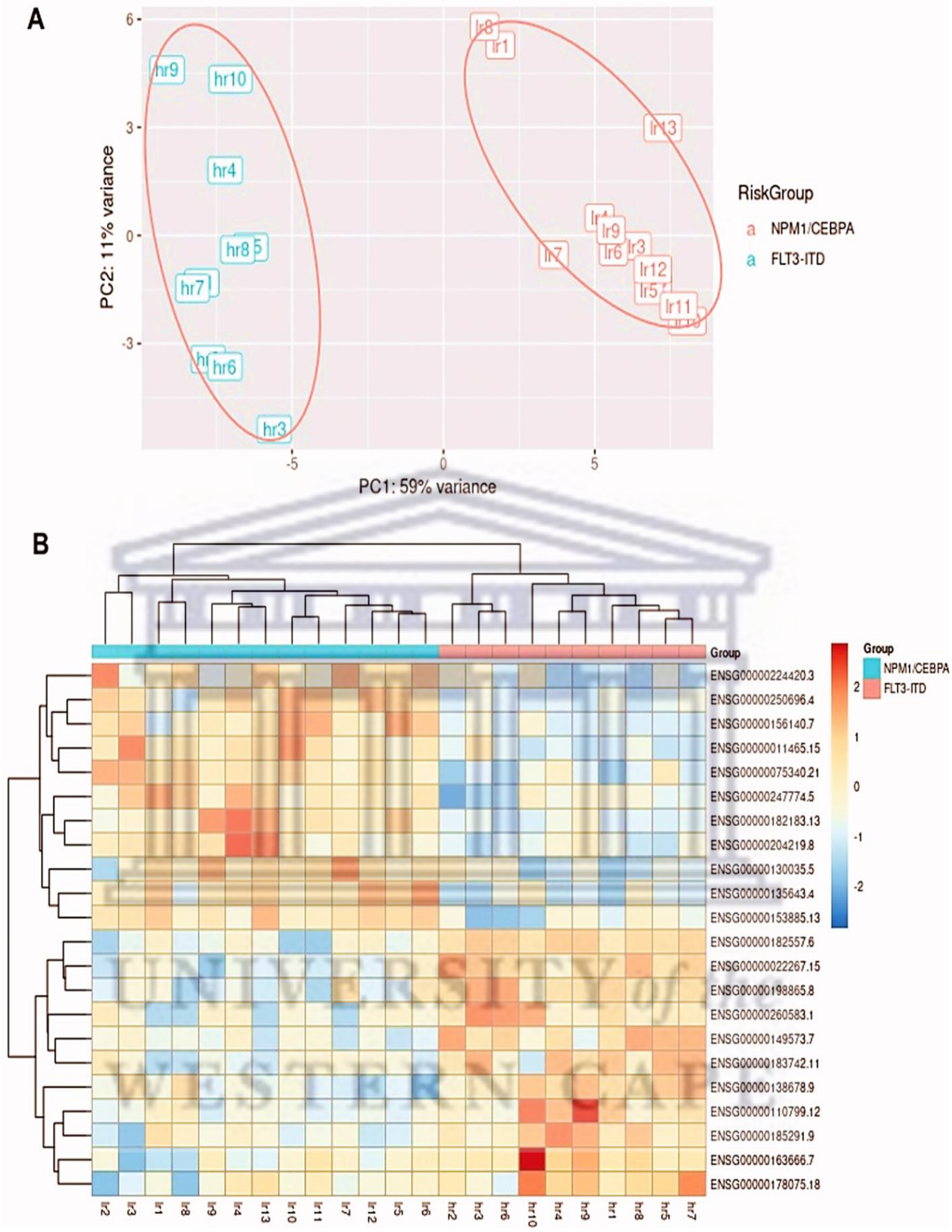

Figure 3.3: Clustering of the AML samples with *FLT3-ITD* and *NPM1*/*CEBPA* mutations using the 22 DEGs selected by RFE using (A) principal component analysis (PCA) and (B) hierarchical clustering using pheatmap R package.

Table 3.3: List of the 22 DEGs selected by RFE genes with * was used in the ML model.

Ensembl_Id	**Symbol**	**log2FoldChange**	**padj**
ENSG00000250696.4	*LOC105377267*	–7.31571544058504	5.57786607011193E-16
ENSG00000022267.15	*FHL1**	3.3904649981173	1.74302159426198E-11
ENSG00000110799.12	*VWF**	5.57344738786003	7.63061866000091E-11
ENSG00000182557.6	*SPNS3**	2.34741592806669	1.73370534268721E-08
ENSG00000149573.7	*MPZL2**	4.11373735379172	2.24605771346956E-07
ENSG00000182183.13	*SHISAL2A*	–2.80284553156498	3.31597387364724E-07
ENSG00000011465.15	*DCN**	–4.93071112156197	3.74939407438753E-07
ENSG00000183742.11	*MACC1**	3.11711209042932	4.62361610488009E-06
ENSG00000156140.7	*ADAMTS3**	–6.20242014449732	5.12366228681635E-06
ENSG00000075340.21	*ADD2**	–3.56747714605576	6.92781902701875E-06
ENSG00000135643.4	*KCNMB4**	–2.96743485475082	1.20611585085678E-05
ENSG00000153885.13	*KCTD15**	–2.73614037690847	2.01801229683281E-05
ENSG00000247774.5	*PCED1B-AS1*	–1.90010539713958	0.000355270349374
ENSG00000224420.3	*ADM5*	–1.99554499874056	0.000408776507021
ENSG00000204219.8	*TCEA3**	–1.99082117539315	0.000651667363783
ENSG00000163666.7	*HESX1**	1.59282492017001	0.001841360690536
ENSG00000185291.9	*IL3RA**	1.36386204250051	0.002217439812688
ENSG00000178075.18	*GRAMD1C*	1.93515803839245	0.003109807383799
ENSG00000130035.5	*KCNA6**	–2.40840491942698	0.003146881886037
ENSG00000260583.1	*LINC00515**	2.48087500485109	0.003149939931692
ENSG00000198865.8	*CCDC152**	2.70119814011179	0.003498603901287
ENSG00000138678.9	*GPAT3*	2.22603438762192	0.004203371885436

3.3.3 Cox's Regression analysis and Kaplan–Meier's estimates

The Cox regression model associated five of the 22 DEGs obtained from RFE with patient survival in the two AML groups. The model created using the LASSO algorithm assigned non-zero, positive or negative coefficients to the five genes

(Table 3.4). The genes with a positive coefficient (*FHL1*, *SPNS3*, and *MPZL2*) had high expression in the *FLT3-ITD* group with the over fivefold increase, while those with a negative coefficient (*KCNMB4* and *ADD2*) had high expression in the *NPM1*/*CEBPA* group with over sevenfold increase. Kaplan–Meier's estimates for OS based on patient statuses of each gene with a positive coefficient were derived and presented in (Figure 3.4). Expression of the three genes in the *FLT3-ITD* group correlated with low survival (*p*<.0001). Patients with a score above the median were predominantly *FLT3-ITD*, and those below mainly were *NPM1*/*CEBPA*.

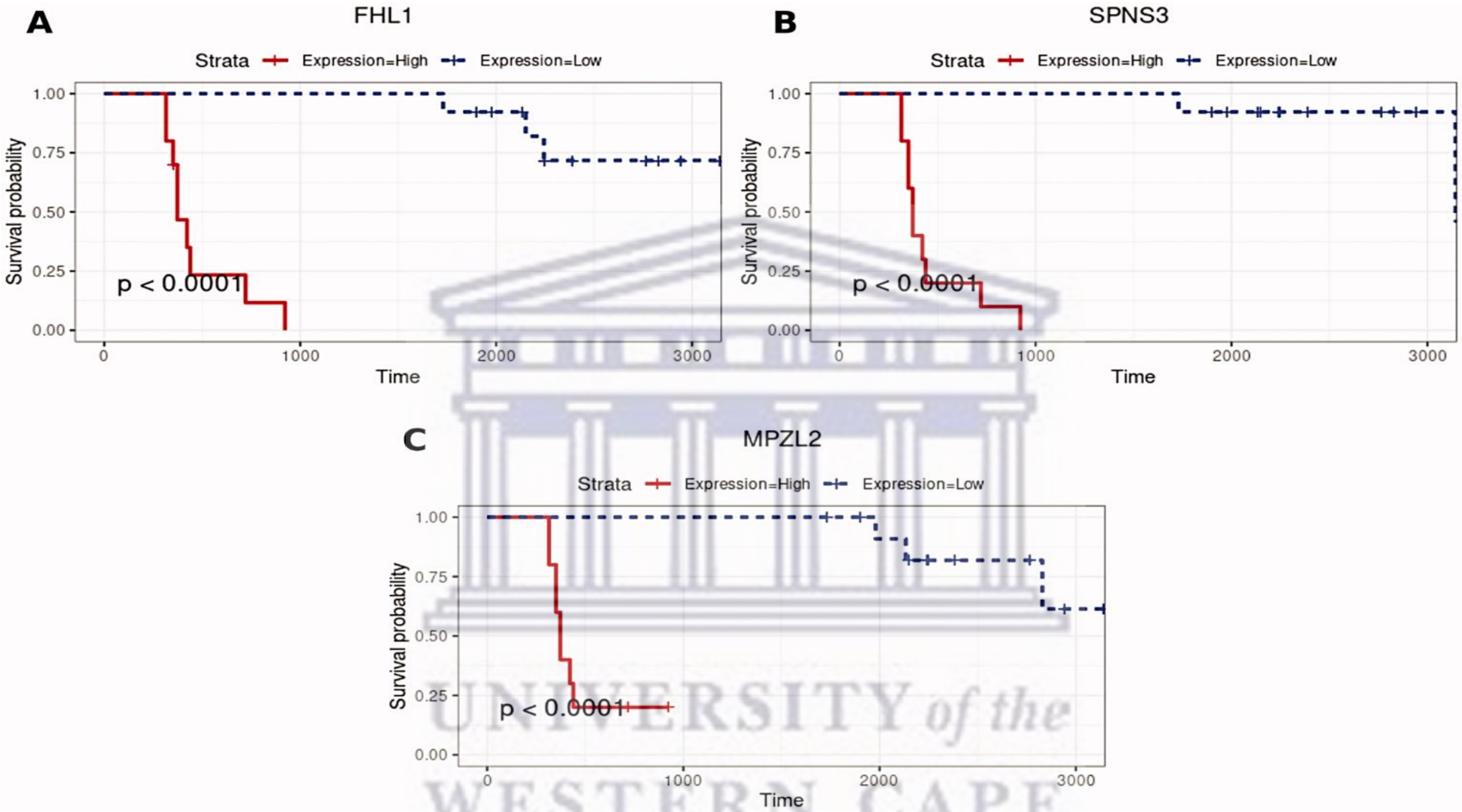

Figure 3.4: Kaplan–Meier's plots of overall survival of patients with *FLT3-ITD* and *NPM1*/*CEBPA* mutations for survival scores above the median (continuous line), corresponding to high expression, and below the median (dotted line), corresponding to low expression, for the genes with positive Cox coefficients (A) *FHL1*, (B) *SPNS3*, and (C) *MPZL2*.

Table 3.4: The list of the DEGs related to overall survival and their corresponding coefficient values from the Cox regression model.

Gene name	Coefficient
FHL1	0.077996384008932
SPNS3	0.050799306065861
MPZL2	0.08555275154584
ADD2	–0.000713971400328
KCNMB4	–0.03112122883124

3.4 Discussion

Proper risk stratification of AML patients at diagnosis is essential for making optimal therapeutic decisions. The results of the DGE analysis demonstrated that the gene expression profile of *FLT3-ITD* AML is different from that of the other mutation groups mentioned above. *FLT3-ITD* mutation is associated with unfavorable AML disease outcomes (Meshinchi *et al.*, 2006). Similar prognostic driver genes have been cataloged in several cancers. For example, in neuroblastoma, another pediatric cancer, *MYCN* amplification is a predictor of poor prognosis (Seeger *et al.*, 1985). Some of the DEGs identified here have been associated with AML and several other cancers (Stein *et al.*, 2009; Lee *et al.*, 2020; Zhou and Chen, 2021). The downstream analysis identified DEGs associated with survival (Table 3.4) and possible responsibility for the poor survival in AML *FLT3-ITD* as opposed to AML *NPM1*/*CEBPA*, which clustered together in the exploratory analysis (Figure 3.1).

3.4.1 DEGs in the expression signature

Some of the genes in the identified signature have been reported to have tumor-suppressive roles and downregulated in several cancers. *TCEA3*, a gene with apoptosis-promoting functions (Liao *et al.*, 2016), was downregulated in the *FLT3-ITD* group in our study. It has been reported downregulated as well in rhabdomyosarcoma (Kazim *et al.*, 2020), gastric cancer (Li *et al.*, 2015), and ovarian cancer (Cha *et al.*, 2013). *DCN*, another downregulated gene in the *FLT3-ITD* group, is well known for its oncosuppressive function (Baghy *et al.*, 2020). *DCN* expression is decreased in different cancer types (Bozoky *et al.*, 2014; Shi *et al.*, 2015; Neill, Schaefer and Iozzo, 2016). In contrast, some upregulated genes have tumor-promoting roles. *VWF*, upregulated in the *FLT3-ITD* group, is a glycoprotein involved in hemostasis whose expression has been associated with lymph node metastasis in prostate cancer patients (Kong *et al.*, 2020). It was also demonstrated to be related to metastatic activities in glioma and osteosarcoma cells (Mojiri *et al.*, 2017).

MACC1 regulates the HGF-MET pathway, which is involved in cellular activities like growth, motility, angiogenesis, invasiveness, epithelial–mesenchymal transition, and metastasis (Birchmeier *et al.*, 2003; Mazzone and Comoglio, 2006). As expected, its high expression has therefore been reported in some cancers, such as retinoblastoma (Nair *et al.*, 2020), gastric cancer (Tong *et al.*, 2019), and colon cancer (Stein *et al.*, 2009). *MACC1* is an independent prognostic indicator for metastasis in colon cancer (Stein *et al.*, 2009). *LINC00515* has also been reported upregulated in glioma (Wu and Lin, 2019) and multiple myeloma (Lu *et al.*, 2018).

IL3RA expression was also correlated with *FLT3-ITD*-mutated AML, although it is also expressed in some *NPM1*-mutated AML cases (Rollins-Raval *et al.*, 2013). High expression of *IL3RA* also seemed to be associated with worse clinical outcomes in AML (Jiang *et al.*, 2020). Arai et al, (2019) reported *IL3RA* expression associated with chemotherapy response failure and poor survival in *de novo* AML patients. It, therefore, suggests that *FLT3-ITD* mutation causes the downregulation of protective tumor-suppressing genes while upregulating genes with oncogenic function.

3.4.2 Machine learning

The high classification accuracy of 92% obtained indicates that the 16-gene expression signature can be used as prognostic markers of *FLT3-ITD* mutation and unfavorable disease outcome in AML and could be potential therapeutic targets. This high accuracy was achieved despite using an adult AML dataset as the testing set. Differences in molecular mechanisms between pediatric and adult AML have been reported (Jeha *et al.*, 2002; Bolouri *et al.*, 2018; Chaudhury *et al.*, 2018). Our classification results suggest that pediatric and adult AML have likely similar disease mechanisms for the mutations investigated in our study. Different gene expression signatures for prognosis in AML have been proposed before. Using expression microarray, Bullinger et al. (Bullinger *et al.*, 2008) found a 20-gene expression signature predictive of *FLT3-ITD* mutation status using prediction analysis of microarray method (Tibshirani *et al.*, 2002) and Kaplan–Meier's analysis. They obtained a prediction accuracy of 81% for identifying cases with *FLT3-ITD* mutation. Our study used an RNA-Seq dataset for analysis, performed

DGE analysis, and applied machine-learning methods to find and validate a 16-gene expression signature predictive of *FLT3-ITD* mutation with 92% higher accuracy achieved. RNA-Seq has become the primary technology used for GEP (Law *et al.*, 2016). Our expression signature of a lesser number of genes should be implementable and potentially cost-effective.

Similarly, Zhu et al. (2020) found six immune-related gene signatures to predict AML prognosis using Cox and LASSO regression analysis. However, their classifier is based on immune-related genes, and the genes identified may not provide a larger picture of the mechanisms at play in the disease pathogenesis (R. Zhu *et al.*, 2020). As a limitation, the sample size of our test dataset was small. Therefore, validation in much larger sample cohorts may still be essential. However, we expect that further studies of the genes identified in our study will shed light on the underlying pathogenesis of AML with *FLT3-ITD* mutation.

3.4.3 Cox's Regression Analysis and Kaplan–Meier's estimates

By reducing the list of genes obtained from the ML analysis, the Cox regression model selected five genes related to patient OS including three genes with positive Cox coefficients. The K-M curves (Figure 3.4) show the ability of these three genes to distinguish between shorter and longer OS, i.e. *FLT3-ITD* and *NPM1*/*CEBPA* groups, respectively (p<.0001). The results of the survival analysis (Figure 3.4) validated the poor prognosis in the *FLT3-ITD* group. Since a patient score above the median is associated with low OS, higher expression of *FHL1*, *SPNS3*, and *MPZL2* with positive Cox coefficients is a prognostic indicator of poor outcome in AML with *FLT3-ITD* mutation. The roles of the genes with negative Cox coefficients, *ADD2* and *KCNMB4*, in the survival of AML patients with *NPM1*/*CEBPA* require further investigation.

FHL1 belongs to a family of genes that play a role in focal adhesion and differentiation. High expression of *FHL1* has been described as a powerful prognostic indicator of worse survival and poor outcome, independent of existing genetic factors for the prognosis of AML. Its overexpression was further found to be related to chemotherapy-resistance and relapse in AML, and *FHL1* knockdown enhanced the sensitivity of AML cells response to treatment (Fu *et al.*, 2020). *SPNS3 i*s a transmembrane transporter whose overexpression has been associated with poor prognosis in AML patients receiving chemotherapy or allogeneic hematopoietic stem cell transplantation (Huang *et al.*, 2020). Huang et al. (2020) suggested that over-expression of *SPNS3* may regulate and control proliferation and differentiation of AML by autophagy. *MPZL2* expressed in the lymphoid organ, thymus, and other epithelial structures maintains stemness in glioblastoma (Ohtsu *et al.*, 2016). Its increased expression has been reported in AML and hepatocellular carcinoma, which is associated with poor prognosis and recurrence (Ni *et al.*, 2020; Yu *et al.*, 2020).

Notably, the high expression of these genes in *FLT3-ITD* patients promotes poor outcomes, and the implication of *FHL1* and *SPNS3* in chemotherapy resistance could be the reason for treatment failure, which may therefore explain the low survival in this category of patients. Functional studies such as knockdown experiments to decipher the precise roles of these genes in *FLT3-ITD* AML would be illuminative.

3.5 Conclusions

In this study, we differentiated AML *FLT3-ITD* and AML *NPM1*/*CEBPA* using GEP. We also identified and validated a 16-gene expression signature for risk classification in AML, which has diagnostic and prognostic value. The upregulation of *FHL1*, *SPNS3*, and *MPZL2* was found to be associated with poor survival in AML *FLT3-ITD*. These genes could therefore be potential therapeutic targets in pediatric and adult AML.

Chapter 4: Investigation of distinct gene expression profile patterns that can improve the classification of intermediate-risk prognosis in AML patients.

This is an original manuscript of an article published in *Frontiers in Genetics* on February 2023, available at: https://doi.org/10.3389/fgene.2023.1131159.

Abstract

Background: Acute myeloid leukemia (AML) is a heterogeneous type of blood cancer that generally affects the elderly. AML patients are categorized with favorable-, intermediate-, and adverse-risks based on an individual's genomic features and chromosomal abnormalities. Despite the risk stratification, the progression and outcome of the disease remain highly variable. To facilitate and improve the risk stratification of AML patients, the study focused on gene expression profiling of AML patients within various risk categories. Therefore, the study aims to establish gene signatures that can predict the prognosis of AML patients and find correlations in gene expression profile patterns that are associated with risk groups.

Methods: Microarray data were obtained from Gene Expression Omnibus (GSE6891). The patients were stratified into four subgroups based on risk and overall survival. Limma was applied to screen for differentially expressed genes (DEGs) between short-survival (SS) and long-survival (LS). DEGs strongly related to general survival were discovered using Cox regression and LASSO analysis. To assess the model's accuracy, Kaplan-Meier (K-M) and receiver operating characteristics (ROC) were used. A one-way ANOVA was performed to assess for differences in the mean gene expression profiles of the identified prognostic genes between the risk subcategories and survival. GO and KEGG enrichment analyses were performed on DEGs.

Results: A total of 87 DEGs were identified between the SS and LS groups. The Cox regression model selected nine genes *CD109*, *CPNE3*, *DDIT4*, *INPP4B*, *LSP1*, *CPNE8*, *PLXNC1*, *SLC40A1*, and *SPINK2* that are associated with AML survival. K-M illustrated that the high expression of the nine-prognostic genes is associated with poor prognosis in AML. ROC further provided high diagnostic efficacy of the prognostic genes. ANOVA also validated the difference in gene expression profiles of the nine genes between the survival groups and highlighted four prognostic genes to provide novel insight into risk subcategories poor and intermediate-poor, as well as good and intermediate-good that displayed similar expression patterns.

Conclusion: Prognostic genes can provide more accurate risk stratification in AML. *CD109*, *CPNE3*, *DDIT4*, and *INPP4B* provided novel targets for better intermediate-risk stratification. This could enhance treatment strategies for this group, which constitutes the majority of adult AML patients.

4.1 Introduction

Acute myeloid leukemia (AML) is hematologic cancer characterized by clonal proliferation and the accumulation of immature myeloid progenitors (Arber *et al.*, 2016). AML is the most prevalent leukemia subtype in adults. The disease is highly heterogeneous, with a variable prognosis and a high mortality rate (Gregory *et al.*, 2009; Vakiti and Mewawalla, 2022). Recent intensive research in genomics, novel treatments, and prognostic markers have substantially improved our understanding of many of the biological aspects of this complex disease (Green and Konig, 2020). However, the global outcome of AML patients remains poor (Wheatley *et al.*, 2009).

The revised European LeukemiaNet (ELN) risk classification system categorizes newly diagnosed AML patients into favorable-, intermediate-, and adverse-risk groups based on cytogenetic and molecular profiles, which serves as a guideline to establish treatment strategies (Döhner *et al.*, 2022). However, it has been noted that this classification system does not completely reflect the heterogeneity within each subgroup. In particular, the intermediate-risk group exhibits significantly diverse biology and prognosis (Hu *et al.*, 2021).

A poorly defined intermediate-risk group results in the majority of AML patients being stratified to an intermediate-risk category (an umbrella category) because they do not meet the criteria that identify specific entities of established prognostic relevance (Awada *et al.*, 2022). Intermediate-risk AML patients feature heterogeneous clinical outcomes, and it further remains a challenge to assign a suitable consolidation of therapy (Döhner, Weisdorf and Bloomfield, 2015; Hu *et al.*, 2021). This emphasizes the need for a more comprehensive description and understanding of the genetic basis of the intermediate-risk group to improve AML patients' prognosis and provide more effective treatment strategies.

The original aim of the ELN genetic categories was to standardize reporting of genetic abnormalities, particularly for correlations with clinical characteristics and outcomes. However, significant modifications to the risk classification for AML from 2017 (Döhner *et al.*, 2017) to 2022 revision (Döhner *et al.*, 2022), which excluded the *FLT3-ITD* mutation, shows that the diagnosis and management of the intermediate-risk group, in particular, remain inexact. Generally, the AML classification and prognostic criteria are based on cytogenetic and molecular features at the time of diagnosis, and thus studies tend to exclude prognostic stratification and base the distinction between the intermediate-I and intermediate-II categories solely on genetic characteristics (Döhner *et al.*, 2010). Meanwhile, a subsequent study demonstrated longer OS in the intermediate-I group than in the intermediate-II group. However, the two groups were prognostically indistinguishable in older patients, who constitute most AML cases (Mrózek *et al.*, 2012).

The purpose of this study is to facilitate improved intermediate-risk stratification of AML and also focus on prognostication. The gene expression profiles of AML patients were investigated to identify gene signatures that differentiate between short- and long-term survival for patients categorized as good- or poor-risk as well as the intermediate-risk group. Therefore, the benefit of this study was twofold (i) the study enabled the segregation of intermediate-risk patients into good and poor-prognosis based on distinct gene expression profiles, (ii) significant prognostic gene signature was identified to differentiate AML patients with good and poor-prognosis. The identified gene signatures associated with survival in AML patients have the potential to serve as prognostic biomarkers that can aid in the prognosis and monitoring of AML. All contribute to a better understanding of the genetic basis of the disease.

4.2 Materials and Methods

4.2.1 Microarray data

The microarray expression profiles of 537 samples and accompanied clinical data were extracted from the Gene Expression Omnibus (GEO) database under the accession number GSE6891 (Verhaak *et al.*, 2009) by the *getGEO* function in the GEOquery R package (version 2.64.2) (Davis and Meltzer, 2007). The patients' survival data were provisioned by the authors (Verhaak *et al.*, 2009), and samples without clinical data and survival information were excluded from subsequent analyses and 447 samples remained (Table 4.1). A complete illustration of the workflow employed in this study is shown in (Figure 4.1).

Table 4.1: Relevant clinical and mutational variables of 447 adult AML samples from the GSE6891 dataset and their distribution by prognosis.

Variable	**Good**, N = 97	**Intermediate**, N = 259	**Poor**, N = 91
Gender			
Female	49 (51%)	130 (50%)	45 (49%)
Male	48 (49%)	129 (50%)	46 (51%)
Vital status			

Alive	57 (59%)	92 (36%)	17 (19%)
Dead	40 (41%)	167 (64%)	74 (81%)
FLT3 ITD Mutation			
Neg	85 (88%)	160 (62%)	81 (89%)
Pos	12 (12%)	99 (38%)	10 (11%)
CEBPA Mutation			
Double Mutation	0 (0%)	23 (8.9%)	1 (1.1%)
Single Mutant	0 (0%)	9 (3.5%)	1 (1.1%)
Wild Type	97 (100%)	227 (88%)	89 (98%)
NPM1 Mutation			
Neg	97 (100%)	131 (51%)	86 (95%)
Pos	0 (0%)	128 (49%)	5 (5.5%)
Overall Survival	2,972 (419, 4,143)	652 (260, 2,606)	358 (145, 751)

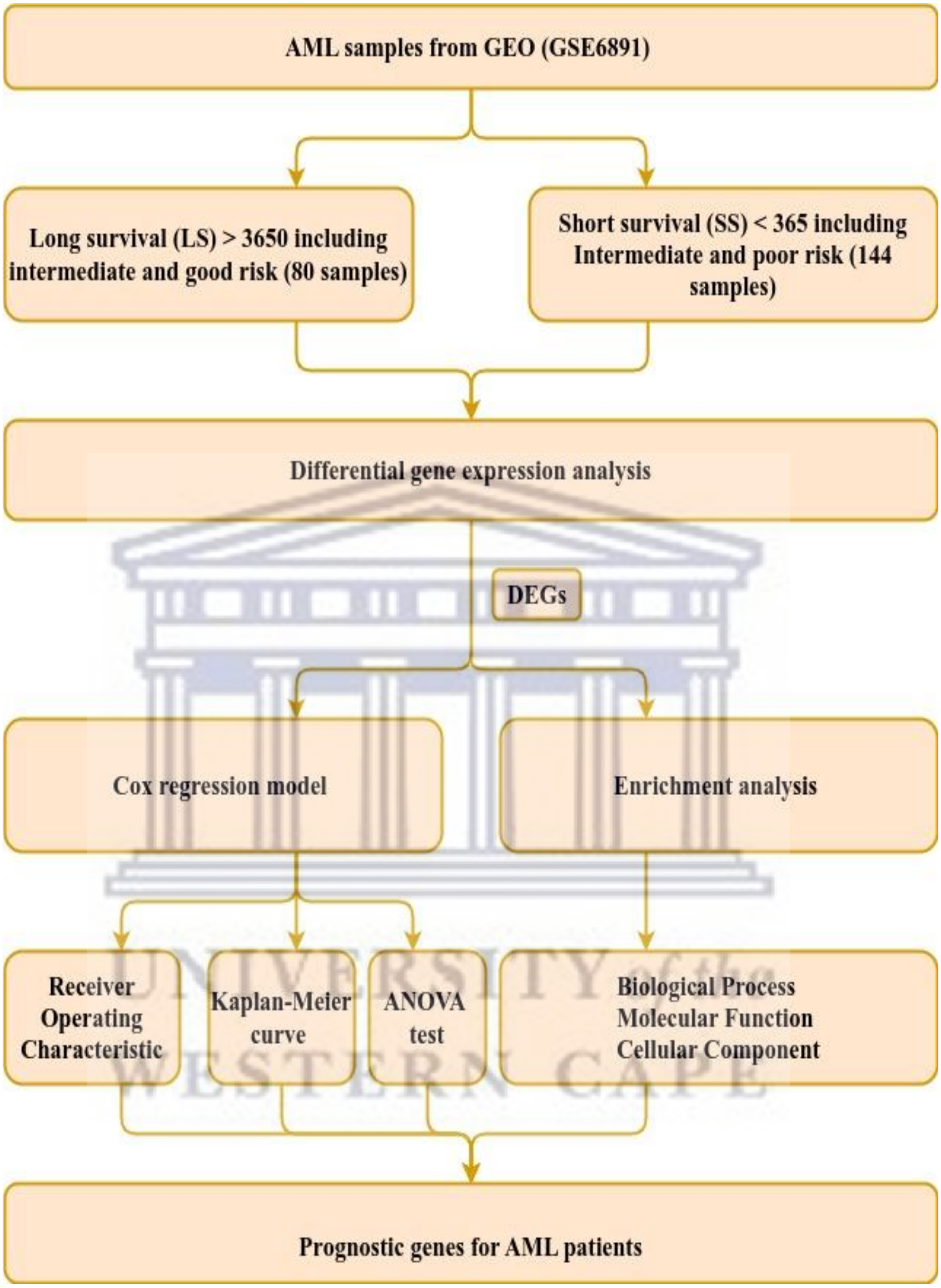
AML samples from GEO (GSE6891)
Long survival (LS) > 3650 including intermediate and good risk (80 samples)
Short survival (SS) < 365 including Intermediate and poor risk (144 samples)
Differential gene expression analysis
DEGs
Cox regression model
Enrichment analysis
Receiver Operating Characteristic
Kaplan-Meier curve
ANOVA test
Biological Process
Molecular Function
Cellular Component
Prognostic genes for AML patients

Figure 4.1: Study workflow. Steps used to identify genetic signature in AML patients with intermediate-risk. The steps comprise data extraction, sample grouping, differential gene expression, survival and functional enrichment analyses. DEGs refer to differentially expressed genes.

4.2.2 Samples selection based on risk profile

The patient samples were divided based on OS into short survival (SS) and long survival (LS) (Table 4.2). The SS includes patients with a survival of less than 365 days, while LS contains patients with a survival of greater than 3,650 days. The two groups were composed by further evaluating the cytogenetic risk classes of the samples in the clinical file, which were categorized into poor-, intermediate-, and good-risk samples. The SS was further stratified into two risk subcategories: poor (PP) and intermediate-poor (IP) risk, while LS was divided into two risk subcategories: good (GG) and intermediate-good (IG) risk. This additional filtering based on OS and cytogenic risk yielded 224 samples for downstream analysis.

Table 4.2: The number of samples stratified by survival time and risk subcategory. Good-Good (GG), Intermediate-Good (IG), Intermediate-Poor (IP), and Poor-Poor (PP) risk of AML sample. Long Survival (LS) and short survival (SS) terms.

Survival time	Risk subcategory	Sample
LS	GG	38
	IG	42
SS	PP	47
	IP	97
Total		224

4.2.3 Data preprocessing

Raw expression data from the 224 selected samples were subjected to background correction, quantile normalization, and $\log_2$ transformation through the RMA algorithm from the affy R package (version 1.74.0). A filtering operation was applied to reduce the probes that exhibited low variation and a consistently low signal across samples. The median expression of the dataset was calculated and returned a median value of 7.2. Thus, a probe was kept if the probe expression is above the median in more than 10 samples. The probe identification numbers were then transformed into official gene symbols and duplicate probes were deleted.

4.2.4 Differential gene expression analysis

The normalized gene expression of 224 samples and 31,140 genes were analyzed to identify differentially expressed genes (DEGs) between the two survival groups (SS and LS). The limma R package (Ritchie *et al.*, 2015) performs differential gene expression (DGE) analysis and experimental design through linear modeling. The limma package was applied to screen for DEGs that differentiate SS from LS. The DEGs were identified with the parameters of the filter set to $| \log_2$ fold change $|> 1$ and adjusted *p*-value < 0.01.

4.2.5 Identification of gene signatures correlated with prognosis

The identified DEGs were subjected to a Cox regression model based on the Lasso algorithm of the glmnet R package (version 4.1-3), to determine which genes were best correlated with patient survival (Friedman, Hastie and Tibshirani, 2010; Simon *et al.*, 2011; Tibshirani *et al.*, 2012). The model reduces the number of candidate genes and selected the most significant genes for a patient's survival, assigning a regression coefficient value to each gene. Genes with a zero coefficient did not affect survival and were discarded. The product of the coefficient value and the corresponding gene's expression value resulted in a prognostic risk score for each patient in the complete dataset (GSE6891) that provided a survival time. The patient scores were used to calculate a median risk score. A status value of 1 or 0 was assigned to each patient based on whether the patient's score was greater than or less than the median risk score.

Using Kaplan-Meier (K-M) survival analysis, the prognostic difference between the short- and long-term survival groups was calculated. The K-M curves were created using the *ggsurvplot* function from the survminer R package (version 3.4-0). Additionally, the predicting power (sensitivity and specificity) of the prognostic gene signatures was calculated using the receiver operating characteristic (ROC) curve analysis (Florkowski, 2008). The ROC curves with the observing AUC values were created in Python by applying the metrics.roc_curve function from sklearn using logistic regression algorithms. The results of the Cox regression model were subjected to a validation step using an independent dataset (GSE37642). This test dataset comprises 11 favorable, 78 intermediate, and 35 adverse cytogenetic risk samples. The survival data were inquired and provided by the authors (Herold *et al.*, 2018). K-M curves and Hazard Ratio (HR) of the prognostic genes were generated for the test dataset.

4.2.6 One-way ANOVA

The statistical analysis was performed using the stats R package (version 4.2.1). The statistics were conducted to evaluate for differences in the mean expression profiles of the prognostic genes identified by Cox regression analysis between the survival groups (SS and LS) and risk subcategories. One-way analysis of variance (ANOVA) was applied, followed by Tukey's post-hoc test for pairwise comparisons (Tukey, 1949). The null hypothesis (H_0) of equal mean between the risk subcategories and survival groups was accepted if the *p*-value > 0.05; H_0: there is no significant difference among the group means.

4.2.7 Functional enrichment analyses

A list of DEGs was subjected to functional annotations of Gene ontology (GO) (Ashburner *et al.*, 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses, the *EnrichGO* and *EnrichKEGG* functions were used, respectively, in the clusterProfiler R package (version 4.4.4) (Yu *et al.*, 2012). *P*-value < 0.05 was determined as a cut-off criterion for significant enrichment.

4.3 Results

4.3.1 Data extraction and DGE analysis

The selected dataset was composed of 144 SS and 80 LS based on the criteria of survival time split set out in section 3.2 (Table 4.2) as input for DGE analysis. In the DGE, a total of 31,140 genes were screened for DEGs to differentiate between SS and LS. A total of 87 DEGs were identified, where 69 genes were up-regulated and 18 genes were down-regulated (Supplementary Table 4.1).

4.3.2 Identification of prognostic genes

By performing univariate Cox regression analysis between the 87 candidate DEGs and patient survival data of (GSE6891), nine prognostic genes were detected and associated with AML patient survival. The prognostic genes were identified using the LASSO algorithm, which assigns non-zero, positive, or negative coefficients. All nine genes had a positive coefficient (Table 4.3).

Table 4.3: Nine prognostic genes with positive coefficient value.

Gene name	Coefficient value
CD109	0.0875676482
CPNE3	0.0755063783
CPNE8	0.0585824601
INPP4B	0.0554178086
SPINK2	0.0544528326
PLXNC1	0.0483530691
LSP1	0.0344057691
DDIT4	0.0216117829
SLC40A1	0.0009147425

Kaplan–Meier's estimates for OS based on patient statuses of each gene with a positive coefficient were derived and presented in (Figure 4.2). All prognostic genes show that a high gene expression level has a poor survival outcome compared to patients with a low gene expression level (Figure 4.2). The estimates, HR and P-value, of the Cox regression model for the prognostic genes were all significant, which confirms the involvement of the alteration in the expression of these genes in the survival of AML patients (Table 4.4). Additionally, same significant results for K-M and HR were obtained for the validation dataset (GSE37642) (Figure 4.3 and Table 4.5).

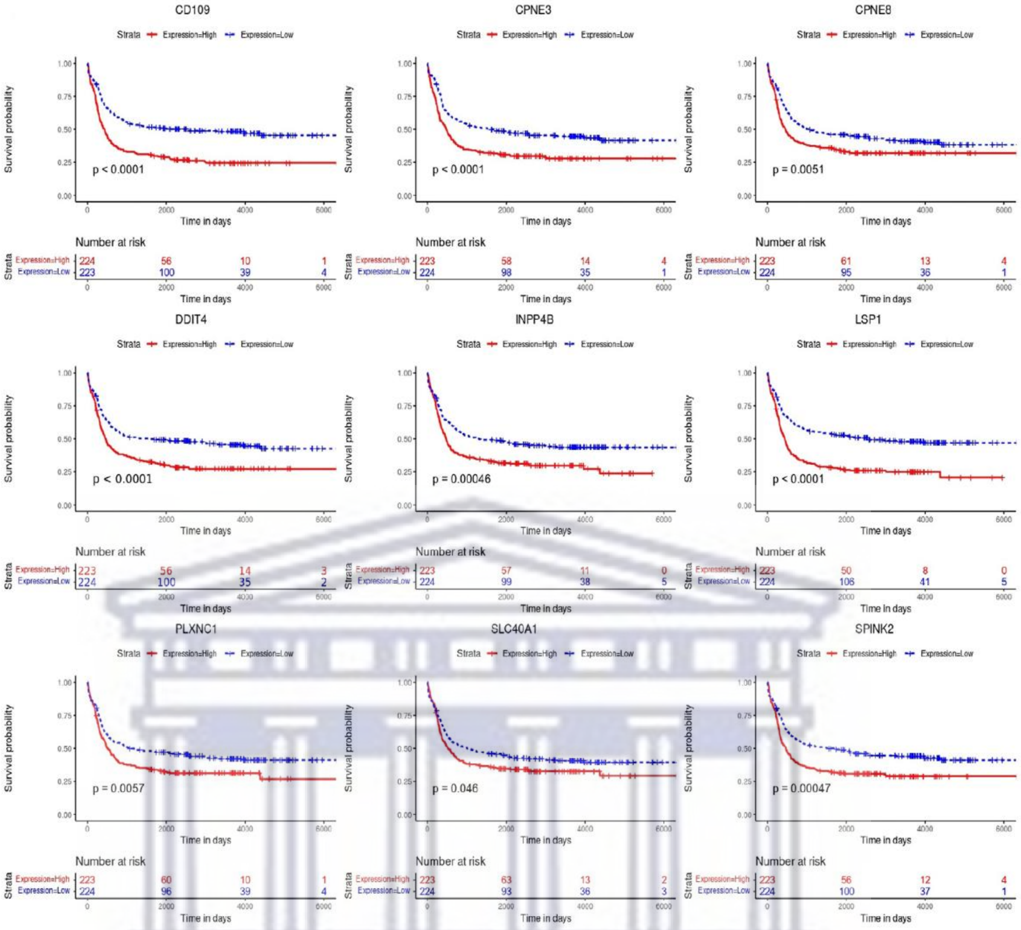

Figure 4.2: Kaplan-Meier (K-M) survival curves. Analysis revealed the survival prediction associated with high and low gene expression profiles of the prognostic genes in AML patients.

Table 4.4: The estimated hazard ratio of each prognostic gene included in the Cox regression for the GSE6891 dataset.

Prognostic Genes	**HR**	**P value**
CD109	0.5322	<0.0001
CPNE3	0.6183	<0.0001
CPNE8	0.7158	0.0051
DDIT4	0.616	<0.0001
INPP4B	0.6572	<0.0001
LSP1	0.5227	<0.0001
PLXNC1	0.7184	0.0062
SLC40A1	0.7875	0.0462
SPINK2	0.6578	0.0005

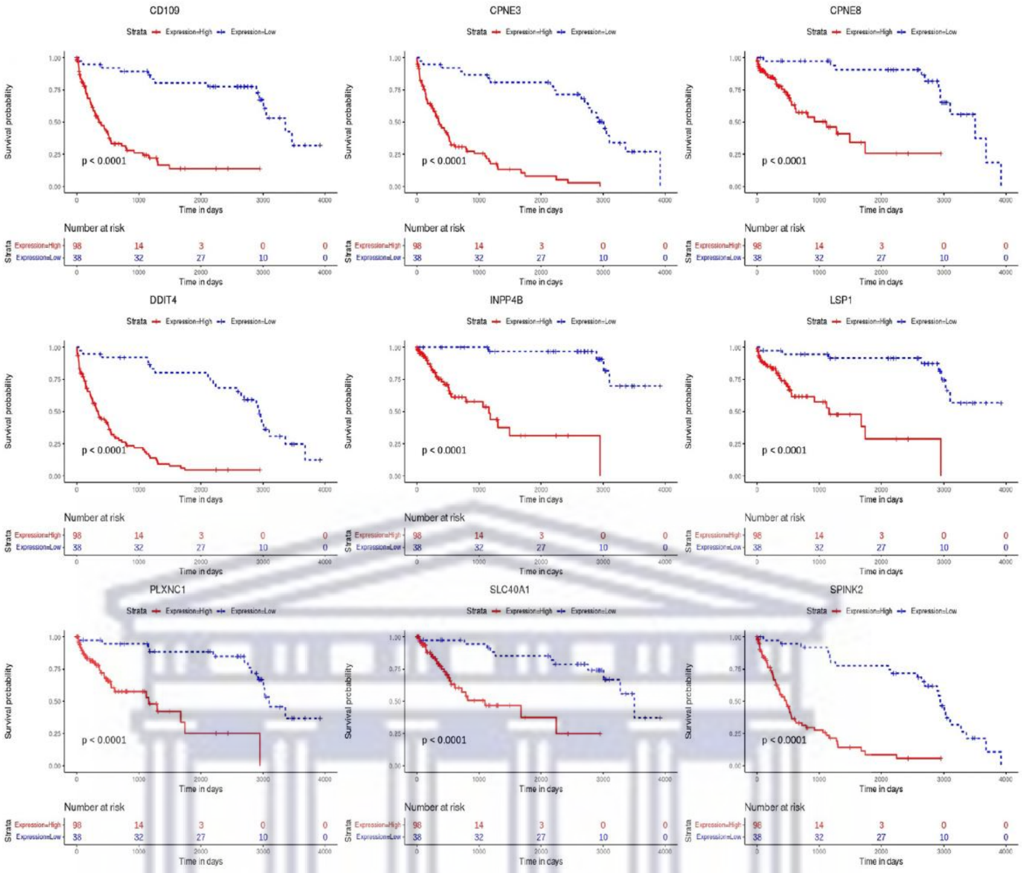

Figure 4.3: Kaplan-Meier (K-M) survival curves. Analysis on the prognostic genes in the validation dataset (GSE37642).

Table 4.5: The estimated hazard ratio of each prognostic gene included in the Cox regression for the independent test dataset (GSE37642).

Gene Name	HR	P value
CPNE3	0.12	< 0.0001
DDIT4	0.13	< 0.0001
LSP1	0.10	< 0.0001
SPINK2	0.15	< 0.0001
PLXNC1	0.14	< 0.0001
SLC40A1	0.17	< 0.0001
CD109	0.13	< 0.0001
CPNE8	0.13	< 0.0001
INPP4B	0.04	< 0.0001

4.3.3 Efficiency evaluation of prognostic gene signatures

The prognostic difference between the high and low gene expression profiles of identified prognostic genes in AML patients was also evaluated using ROC curves. ROC analysis evaluated the accuracy of the aforementioned nine-genes model for survival prediction in AML patients. The ROC curve showed the best performance for the area under the curve (AUC) for *CD109* of 0.84. This followed by AUC > 0.81 for *CPNE3*, *CPNE8*, *PLXNC1*, and *SPINK2* (Figure 4.4). Genes *LSP1*, *DDIT4*, and *INPP4B* were 0.74 ≤ AUC ≤ 0.79, with the lowest AUC of *SLC40A1*, was 0.69 (Figure 4.4).

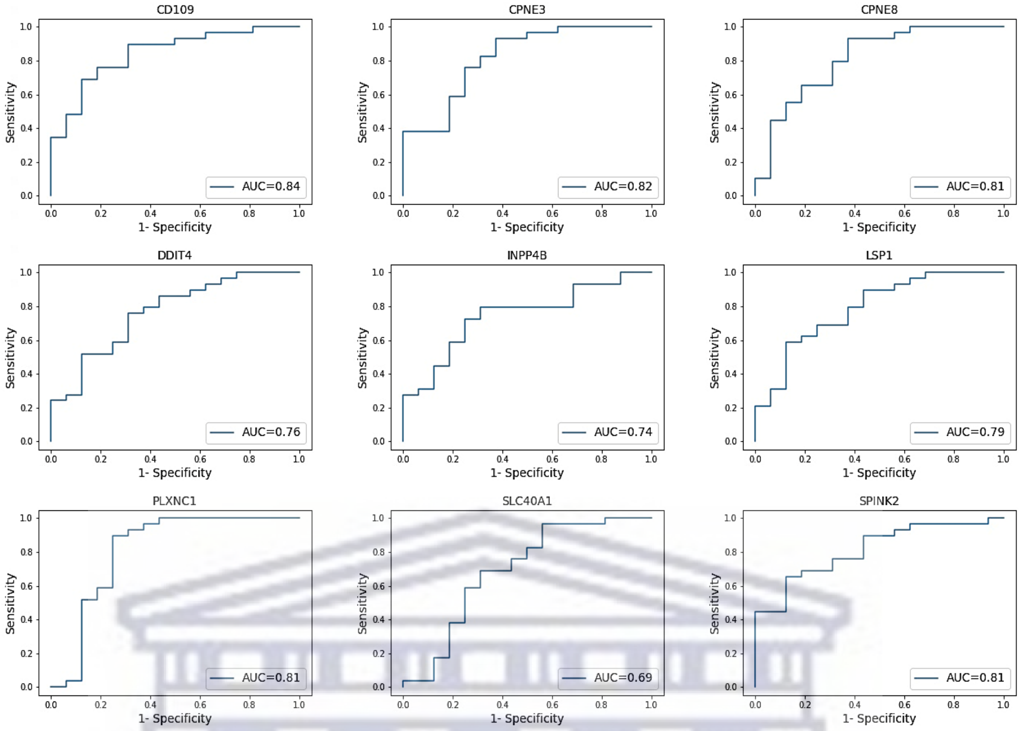

Figure 4.4: Receiver operating characteristic (ROC) curves. Evaluating the accuracy of high and low gene expression profiles of the nine-genes model in AML patients. *AUC = area under curve

4.3.4 Gene expression patterns between risk categories

One-way ANOVA was used to evaluate for differences in the mean gene expression profile of each prognostic gene identified between the survival groups and risk subcategories. This include the difference between the short- (PP and IP) and long-term survival (GG and IG) (Figure 4.5). ANOVA result confirmed that short- and long-term survival for all prognostic genes are statistically different in gene expression profiles (*p*-value $\leq 1.3 \times 10^{-8}$) (Figure 4.5).

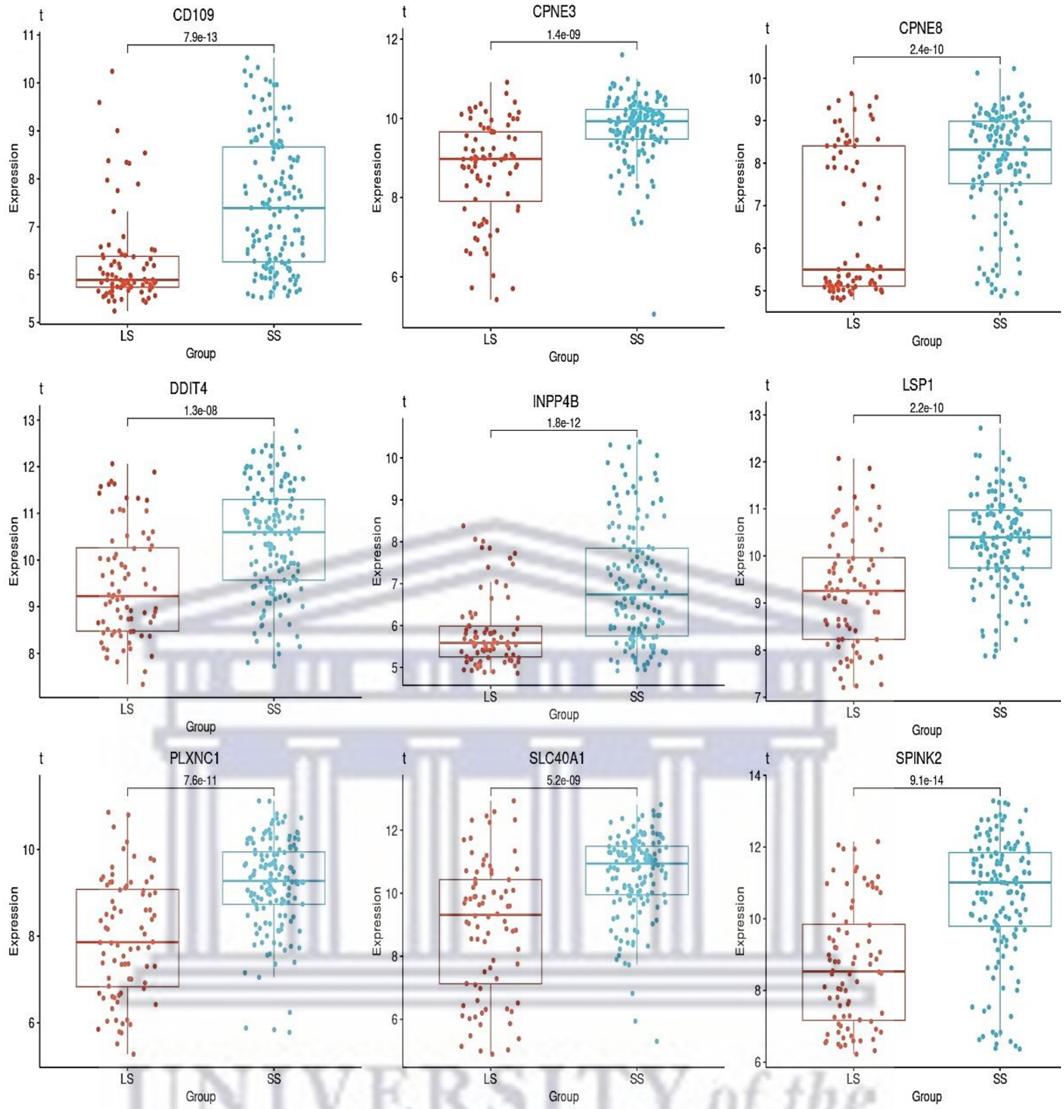

Figure 4.5: Boxplots based on the survival times of the prognostic genes in AML patients. A boxplot was constructed with the gene expression profile of each prognostic gene in all the samples that were categorized as short- and long-term survival.

The samples that were categorized into PP, IP, IG, and GG-risk groups, respectively, were investigated for each of the nine-genes models that were identified with prognostic significance. Each risk group was composed of a set of samples in which the gene expression profile of a specific prognostic gene was extracted to construct a boxplot (Figure 4.6). The differences in the mean gene expression profiles of each prognostic gene identified between PP and IP-risk groups, as well as the GG and IG-risk groups. Also, the difference between the two intermediate-risk groups was evaluated with the IP and IG-risk groups (Figure 4.6).

All prognostic genes showed a statistically significant difference between the two intermediate-risk groups, i.e. IG and IP-risk (p-value $\leq 5.5 \times 10^{-5}$). Also, the ANOVA results between the risk subcategories showed that the mean gene expression profiles of genes *CD109*, *CPNE3*, *DDIT4*, and *INPP4B* showed no statistically significant difference between the PP and IP-risk groups (p-value $\geq$ 0.16). The same was found for the IG and GG-risk groups (p-value $\geq$ 0.54) (Figure 4.6).

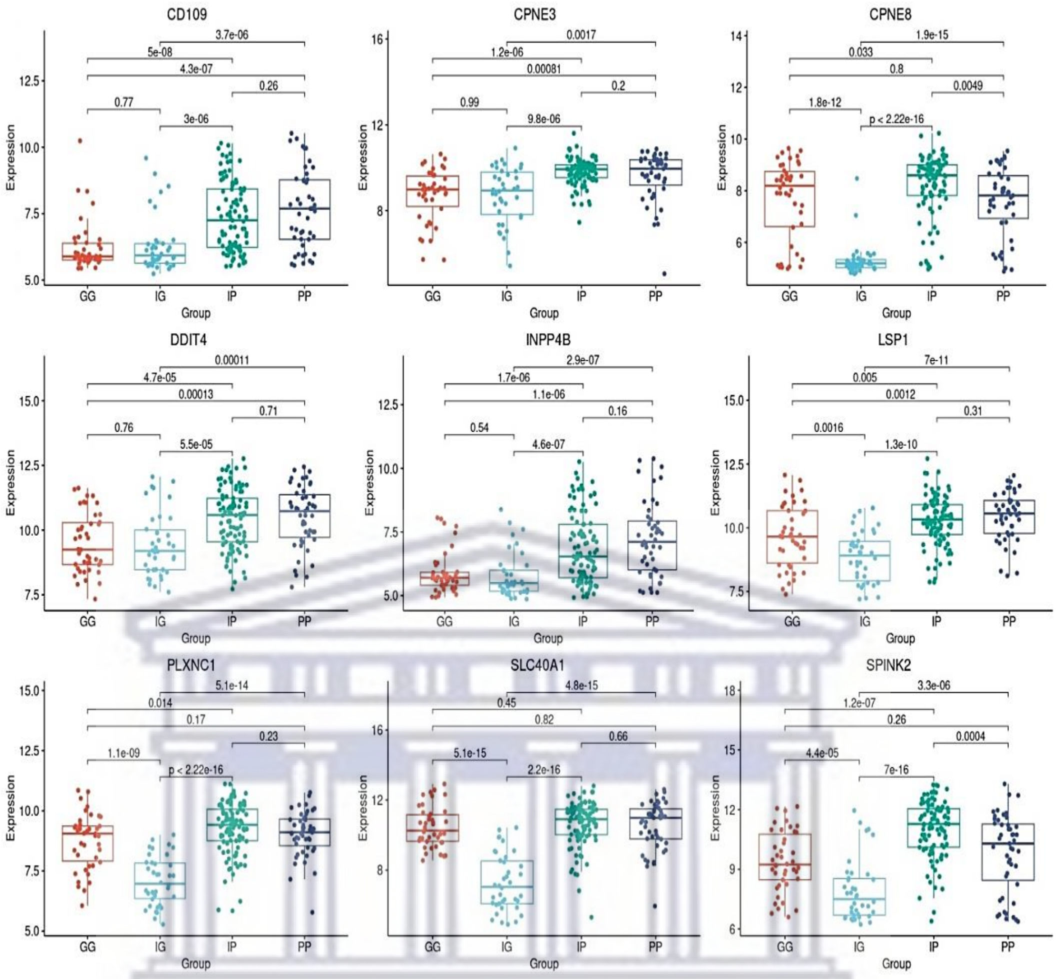

Figure 4.6: Boxplots based on risk subcategories of the nine prognostic genes in AML patients. A boxplot was constructed with the gene expression profile of each prognostic gene in all the samples that were categorized into the Good-Good (GG), Intermediate-Good (IG), Intermediate-Poor (IP), and Poor-Poor (PP) risk categories.

4.3.5 Enrichment analysis

The GO enrichment analysis showed that AML DEGs were significantly enriched in functional items, such as DNA-binding transcription activator activity, RNA and polymerase II-specific and DNA-binding transcription activator activity, and so on of the biological process (BP). In terms of molecular function (MF), AML DEGs were significantly enriched in functional items such as negative regulation of cytokine production, myeloid cell differentiation, and pattern specification process, among other terms (Figure 4.7). In terms of the cellular component (CC), AML DEGs were significantly enriched in functional items such as secretory granule lumen, cytoplasmic vesicle lumen, and vesicle lumen (Figure 4.7). The KEGG analysis indicated significant differences in the transcriptional mis-regulation in the cancer pathway, PI3K-Akt signalling pathway, and Rap1 signalling pathway (Figure 4.7).

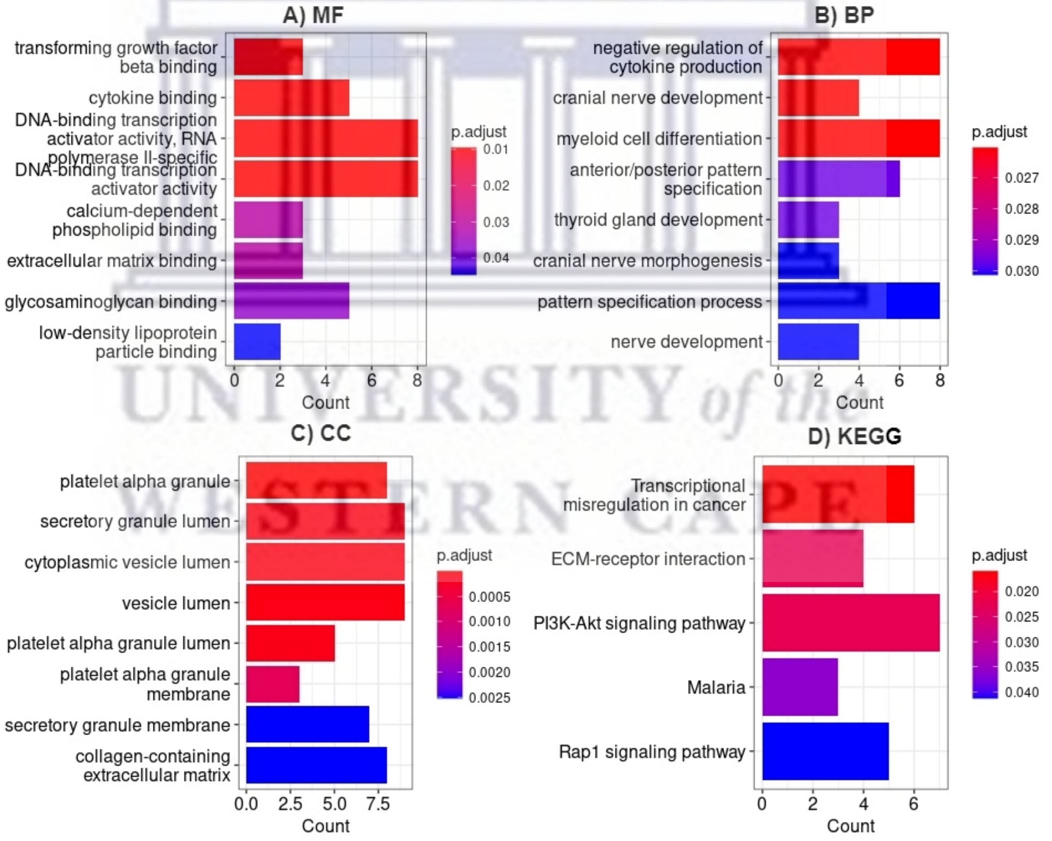

Figure 4.7: AML DEGs were enriched in Gene Ontology and KEGG pathways. (A) Molecular function, (B) Biological process, (C) Cellular component, (D) Kyoto Encyclopedia of Genes and Genomes. The horizontal axis represents the number of

enriched genes, and the vertical axis represents the gene ontology project and KEGG pathways, respectively.

4.4 Discussion

The risk stratification of AML patients into favorable-, intermediate- and adverse-risk groups is crucial to determine an effective therapy strategy and medical care. However, AML patients continue to feature heterogeneous clinical outcomes, and it remains a challenge to assign a suitable consolidation of therapy. Therefore, it is vital to investigate new leukemogenesis-related characteristics. This study aimed to investigate gene expression profiles in AML patients with long and short survival to decipher the heterogeneity in outcomes of intermediate-risk patients and propose a genetic signature that accurately predicts survival of intermediate-risk patients.

The study screened DEGs through the gene expression profiles between short- and long-term survival of AML samples. GO terms and KEGG pathways enrichment analyses was carried out on a total of 87 DEGs to explore the function of the DEGs. GO enrichment analysis illustrated that the DEGs of AML were significantly enriched in functional items such as DNA-binding transcription activator activity, myeloid cell differentiation, secretory granule lumen, cytoplasmic vesicle lumen, and vesicle lumen which was similarly found in studies that focused on predicting disease prognosis for AML (Chen *et al.*, 2020, 2021; Kuang *et al.*, 2021). Interestingly, the prognostic gene *CD109* enriched for all three types of GO terms (BP, MF, and CC). Additionally, the *CD109* gene was enriched in the functional item myeloid cell differentiation, which suggests significant involvement in the development of AML disease.

The KEGG pathway analysis revealed that AML DEGs were enriched in the transcriptional misregulation in cancer, Rap1 signalling pathway, and PI3K-Akt signalling pathway. Consistent with previous studies, the aforementioned pathways have been reported to have an impact on the pathogenesis and prognosis of AML (Martelli *et al.*, 2006; Bertacchini *et al.*, 2015; Yin *et al.*, 2018; Chen *et al.*, 2020). The prognostic *DDIT4* gene enriched in the PI3K-Akt signalling pathway may play a crucial role in the activation of cancer. Therefore, the GO enrichment analysis and KEGG pathway enrichment results showed that the identified DEGs may be important pathogenic genes of AML, contributing to the occurrence and progression of the disease.

This study identified a nine-genes model as potential prognostic biomarkers and therapeutic targets for AML (Table 4.3). Cox regression and Kaplan-Meier analyses validated the prognostic biomarkers and illustrated that high gene expression of all nine genes has a poor prognosis, whereas a low gene expression is associated with a good prognosis in AML. Therefore, both Kaplan-Meier and high AUC values confirmed that the nine-genes model has good diagnostic efficacy in predicting prognosis for AML. Previous studies supported the findings and reported that the higher expression of the genes is associated with poor prognosis in AML (Woolley, Dzneladze and Salmena, 2015; Fu *et al.*, 2017; Gasparetto *et al.*, 2019; Lebedev *et al.*, 2019; Xue *et al.*, 2019; Cheng *et al.*, 2020; Ding *et al.*, 2021; Zhao, Li and Wu, 2018). A recent study (Deepak Shyl *et al.*, 2022) revealed the potential of *CD109* as a biomarker with diagnostic capabilities in AML, and this study further aligns with this finding, in which CD109 was also found with the highest specificity and sensitivity with AUC (Figure 4.4).

The difference in mean gene expression profiles of the prognostic genes were evaluated with ANOVA to determine if there is a difference in gene expression profiles between short- and long-term survival samples. ANOVA confirmed a statistically significant difference between the short- and long-term survival in the nine-genes model and therefore confirms the prognostic significance of the nine prognostic genes identified in this study. The intermediate-risk category was further investigated to improve the risk category in which the majority of AML patients are classified. It is noteworthy that all nine prognostic biomarkers displayed a statistically significant difference between the gene expression profiles in the intermediate-good and intermediate-poor risk categories (p-value ≤ 5.5×10^{-5}). The nine prognostic genes are therefore essential in intermediate-risk group classification as AML patients categorized into this risk group could be provided with an improved prognosis.

A crucial finding was made between the gene expression profiles of good-risk compared to intermediate good-risk. It was found that the prognostic biomarkers *CD109*, *CPNE3*, *DDIT4*, and *INPP4B* found in this study displayed the same pattern of gene expression in both GG and IG-risk categories. Hence, GG and IG-risk categories gene expression was not significantly different in the four genes (p-value ≥ 0.54) (Figure 4.6). The same observation was made when comparing the gene expression profiles of poor-risk and intermediate poor-risk. The same four genes displayed the same pattern of gene expression in both PP and IP-risk categories (p-value ≥ 0.16) (Figure 4.6). Therefore, the four genes may enable a reclassification of the intermediate-risk category in AML patients into either good- or poor-risk based on the gene expression levels of the four genes. Hence, this finding is important as it could predict the outcome of intermediate risk patients as it is directly associated with survival. This discovery provides a more comprehensive description and understanding of the genetic basis of the intermediate-risk group and therefore has the potential to improve AML patients' prognosis and provide more effective treatment strategies.

4.5 Conclusion

In this study, we found correlations between risk categories and gene signatures that differentiate short- and long-term survival using gene expression profile data from an AML GEO dataset. The gene expression profiles of nine prognostic genes, including *CD109*, *CPNE3*, *DDIT4*, *INPP4B*, *LSP1*, *CPNE8*, *PLXNC1*, *SLC40A1*, and *SPINK2*, showed that high gene expression is associated with poor prognosis. Therefore, the nine genes have the prognostic ability and successfully predict the prognosis of AML patients. Also, the prognostic biomarkers were able to segregate intermediate-risk into poor- and good-risk categories that improve the risk classification by adding prognostic significance to the particular risk category. The prognostic biomarkers *CD109*, *CPNE3*, *DDIT4*, and *INPP4B* provided novel insights as the gene expression pattern were similar between poor and intermediate-poor as well as good and intermediate-good. Therefore, these biomarkers provide targets that can enhance prognosis and provide a more effective treatment strategy for AML patients categorized into the intermediate-risk group. Hence, these biomarkers could serve as potential therapeutic targets in adult AML.

Chapter 5: Conclusion and future recommendations

5.1 Conclusion

AML is a heterogeneous disease in various aspects, including the fact that the disease affects patients of all ages, adult AML patients have a greater mortality rate than pediatric AML patients, and pediatric AML patients have a higher survival rate. As a result, AML prognosis is still challenging to predict. Many risk classifications that have been established for decades are still frequently updated due to an increased understanding of the disease's molecular profile and other mechanisms. This research project comprises GEP datasets for AML that are publically available from TARGET and accessed through the Xena database portal and gene expression omnibus (GEO) for pediatric and adult patients, respectively.

Recent advancements in pediatric AML have improved survival rates in pediatric AML. However, pediatric patients with *FLT3-ITD* have an unfavorable outcome, and *FLT3-ITD* has been reported as the most prevalent AML mutation (Wu *et al.*, 2016). Therefore, this study aimed to use DGE analysis to identify gene signatures for pediatric AML with *FLT3-ITD* as a prognostic biomarker. A total of 471 DEGs were identified, and downstream analysis revealed 16 genes that could classify and segregate between *FLT3-ITD* and *NPM1*/*CEBPA* patient samples with an accuracy of 92%. Additionally, high expression of genes *FHL1*, *SPNS3*, and *MPZL2* was associated with poor outcomes in AML *FLT3-ITD* patients.

The incidence of AML increases with age and is thus more prevalent in adults, with a lower survival rate in comparison to pediatric. The majority of adult AML patients are categorized as intermediate-risk, and reports of variable outcomes for this category have been made. The GEP profiles of adult AML patients were investigated to find DEGs between long- and short-survival and their ability to reclassify the intermediate-risk group. A total of 87 DEGS were found between short- and long-survival, and the Cox regression model revealed that only nine prognostic genes were linked to short survival, of which four genes, namely (*CD109*, *DDIT4*, *CPNE3*, and *INPP4B*) significantly distinguish between the short- and long-survival within the intermediate-risk group as demonstrated by ANOVA.

This study's key discovery was that pediatric AML patients with high gene expression of *FHL1*, *SPNS3*, and *MPZL2* were identified as prognostic biomarkers and associated with poor outcomes. This was similarly found in genes *CD109*, *DDIT4*, *CPNE3*, *INPP4B*, *CPNE8*, *LSP1*, *PLXNC1*, *SLC40A1*, and *SPINK2* for adult AML. The first four prognostic genes were able to reclassify the intermediate-risk group in adult AML further. Therefore, this study demonstrated that using the gene expression profile of AML patients was beneficial for biomarker discovery.

5.2 Clinical importance

Genomic profiling enhances the understanding of disease progression. In this study, the identified prognostic biomarkers have the potential to be applied in clinical practices and improve the course of the disease. From the time of diagnosis, a clinician can be guided and able to predict the outcome for administering a more effective therapy based on the gene expression level of the identified prognostic biomarkers and the possibility of being used as target therapy. Furthermore, the discovery of the four genes that enabled a reclassification of adult AML intermediate-risk could facilitate the choice of the best therapy option for this specific risk group.

5.3 Future recommendations

Molecular biomarkers are important for disease diagnosis, prognosis, and outcome prediction. The result obtained from this research is highly recommended for clinical applications, such as developing prognostic panels that include the three prognostic genes for pediatric AML patients, specifically with the *FLT3-ITD* mutation. The same principle can be applied to the nine prognostic genes in adult AML. This will contribute to AML patient status monitoring and management, as well as the development of therapeutic targets for prognostic genes whose expression is linked to the prognosis of AML. The genes with risk reclassification properties should further be considered for reclassification of adult AML patients categorized as intermediate-risk or incorporated with existing classification systems.

References

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

Alejandro Sweet-Cordero, E. and Biegel, J.A. (2019) 'The genomic landscape of pediatric cancers: Implications for diagnosis and treatment', *Science*, 363(6432), pp. 1170–1175. Available at: https://doi.org/10.1126/science.aaw3535.

Alves, R. *et al.* (2021) 'Resistance to Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia—From Molecular Mechanisms to Clinical Relevance', *Cancers*, 13(19), p. 4820. Available at: https://doi.org/10.3390/cancers13194820.

An, Q., Fan, C.-H. and Xu, S.-M. (2017) 'Recent perspectives of pediatric leukemia - an update', *European Review for Medical and Pharmacological Sciences*, 21(4 Suppl), pp. 31–36.

Anjum, A. *et al.* (2016) 'Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach', *Journal of Computational Biology*, 23(4), pp. 239–247. Available at: https://doi.org/10.1089/cmb.2015.0205.

Arai, N. *et al.* (2019) 'Impact of CD123 expression, analyzed by immunohistochemistry, on clinical outcomes in patients with acute myeloid leukemia', *International Journal of Hematology*, 109(5), pp. 539–544. Available at: https://doi.org/10.1007/s12185-019-02616-y.

Arber, D.A. *et al.* (2016) 'The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia', *Blood*, 127(20), pp. 2391–2405. Available at: https://doi.org/10.1182/BLOOD-2016-03-643544.

Ashburner, M. *et al.* (2000) 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nature Genetics*, 25(1), pp. 25–29. Available at: https://doi.org/10.1038/75556.

Aung, M.M.K. *et al.* (2021) 'Insights into the molecular profiles of adult and paediatric acute myeloid leukaemia', *Molecular Oncology*, 15(9), pp. 2253–2272. Available at: https://doi.org/10.1002/1878-0261.12899.

Awada, Hassan *et al.* (2022) 'A Focus on Intermediate-Risk Acute Myeloid Leukemia: Sub-Classification Updates and Therapeutic Challenges', *Cancers*, 14(17), pp. 1–20. Available at: https://doi.org/10.3390/cancers14174166.

Baghy, K. *et al.* (2020) 'Decorin in the Tumor Microenvironment', in *Advances in Experimental Medicine and Biology*, pp. 17–38. Available at: https://doi.org/10.1007/978-3-030-48457-6_2.

Barrett, T. *et al.* (2013) 'NCBI GEO: archive for functional genomics data sets—update', *Nucleic Acids Research*, 41(D1), pp. D991–D995. Available at: https://doi.org/10.1093/nar/gks1193.

Bayat, A. (2002) 'Science, medicine, and the future: Bioinformatics', *BMJ*, 324(7344), pp. 1018–1022. Available at: https://doi.org/10.1136/bmj.324.7344.1018.

Bennett, J.M. *et al.* (1976) 'Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group', *British Journal of Haematology*, 33(4), pp. 451–458. Available at: https://doi.org/10.1111/j.1365-2141.1976.tb03563.x.

Bennett, J.M. *et al.* (1985) 'Criteria for the diagnosis of acute leukemia of megakaryocyte lineage (M7). A report of the French-American-British Cooperative Group', *Annals of Internal Medicine*, 103(3), pp. 460–462. Available at: https://doi.org/10.7326/0003-4819-103-3-460.

Bennett, J.M. *et al.* (1991) 'Proposal for the recognition of minimally differentiated acute myeloid leukaemia (AML-MO)', *British Journal of Haematology*, 78(3), pp. 325–329. Available at: https://doi.org/10.1111/j.1365-2141.1991.tb04444.x.

Bertacchini, J. *et al.* (2015) 'Targeting PI3K/AKT/mTOR network for treatment of leukemia', *Cellular and Molecular Life Sciences*, 72(12), pp. 2337–2347. Available at: https://doi.org/10.1007/s00018-015-1867-5.

Bhojwani, D., Howard, S.C. and Pui, C.-H. (2009) 'High-Risk Childhood Acute Lymphoblastic Leukemia', *Clinical lymphoma & myeloma*, 9(Suppl 3), p. S222. Available at: https://doi.org/10.3816/CLM.2009.s.016.

Birchmeier, C. *et al.* (2003) 'Met, metastasis, motility and more', *Nature Reviews Molecular Cell Biology*, 4(12), pp. 915–925. Available at: https://doi.org/10.1038/nrm1261.

Bispo, J.A.B., Pinheiro, P.S. and Kobetz, E.K. (2020) 'Epidemiology and Etiology of Leukemia and Lymphoma', *Cold Spring Harbor Perspectives in Medicine*, 10(6), p. a034819. Available at: https://doi.org/10.1101/cshperspect.a034819.

Bolouri, H. *et al.* (2018) 'The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions', *Nature Medicine*, 24(1), pp. 103–112. Available at: https://doi.org/10.1038/nm.4439.

Bozoky, B. *et al.* (2014) 'Decreased decorin expression in the tumor microenvironment', *Cancer medicine*, 3(3), pp. 485–491. Available at: https://doi.org/10.1002/cam4.231.

Bullinger, L. *et al.* (2008) ‘An FLT3 gene-expression signature predicts clinical outcome in normal karyotype AML’, *Blood*, 111(9), pp. 4490–4495. Available at: https://doi.org/10.1182/blood-2007-09-115055.

Burd, A. *et al.* (2020) ‘Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial’, *Nature medicine*, 26(12), pp. 1852–1858. Available at: https://doi.org/10.1038/s41591-020-1089-8.

Califf, R.M. (2018) ‘Biomarker definitions and their applications’, *Experimental Biology and Medicine*, 243(3), pp. 213–221. Available at: https://doi.org/10.1177/1535370217750088.

Carter, J.L. *et al.* (2020) ‘Targeting multiple signaling pathways: the new approach to acute myeloid leukemia therapy’, *Signal Transduction and Targeted Therapy*, 5(1), p. 288. Available at: https://doi.org/10.1038/s41392-020-00361-x.

Catovsky, D. *et al.* (1991) ‘A classification of acute leukaemia for the 1990s’, *Annals of Hematology*, 62(1), pp. 16–21. Available at: https://doi.org/10.1007/BF01714978.

Cha, Y. *et al.* (2013) ‘TCEA3 binds to TGF-beta receptor I and induces Smad-independent, JNK-dependent apoptosis in ovarian cancer cells’, *Cellular Signalling*, 25(5), pp. 1245–1251. Available at: https://doi.org/10.1016/j.cellsig.2013.01.016.

Chaudhury, S. *et al.* (2018) ‘Age-specific biological and molecular profiling distinguishes paediatric from adult acute myeloid leukaemias’, *Nature Communications*, 9(1), p. 5280. Available at: https://doi.org/10.1038/s41467-018-07584-1.

Chen, M. *et al.* (2021) ‘The Prognostic Value and Function of HOXB5 in Acute Myeloid Leukemia’, *Frontiers in Genetics*, 12. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2021.678368 (Accessed: 22 December 2022).

Chen, S. *et al.* (2020) ‘Bioinformatics Analysis Identifies Key Genes and Pathways in Acute Myeloid Leukemia Associated with DNMT3A Mutation’, *BioMed Research International*, 2020, p. e9321630. Available at: https://doi.org/10.1155/2020/9321630.

Cheng, Z. *et al.* (2020) ‘Up-regulation of DDIT4 predicts poor prognosis in acute myeloid leukaemia’, *Journal of Cellular and Molecular Medicine*, 24(1), pp. 1067–1075. Available at: https://doi.org/10.1111/jcmm.14831.

Chilton, L. *et al.* (2014) ‘Hyperdiploidy with 49–65 chromosomes represents a heterogeneous cytogenetic subgroup of acute myeloid leukemia with differential outcome', *Leukemia*, 28(2), pp. 321–328. Available at: https://doi.org/10.1038/leu.2013.198.

Chowdhary, M. *et al.* (2016) 'Bioinformatics: an overview for cancer research', *Journal of Drug Delivery and Therapeutics*, 6(4), pp. 69–72. Available at: https://doi.org/10.22270/jddt.v6i4.1290.

Clark, T.G. *et al.* (2003) 'Survival Analysis Part I: Basic concepts and first analyses', *British Journal of Cancer*, 89(2), pp. 232–238. Available at: https://doi.org/10.1038/sj.bjc.6601118.

Clough, E. and Barrett, T. (2016) 'The Gene Expression Omnibus database', *Methods in molecular biology (Clifton, N.J.)*, 1418, pp. 93–110. Available at: https://doi.org/10.1007/978-1-4939-3578-9_5.

Conneely, S.E. and Stevens, A.M. (2021) 'Acute Myeloid Leukemia in Children: Emerging Paradigms in Genetics and New Approaches to Therapy', *Current Oncology Reports*, 23(2), p. 16. Available at: https://doi.org/10.1007/s11912-020-01009-3.

Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017) 'RNA-Seq differential expression analysis: An extended review and a software tool', *PLoS ONE*, 12(12), pp. 1–18. Available at: https://doi.org/10.1371/journal.pone.0190152.

Creutzig, U. *et al.* (2012) 'Diagnosis and management of acute myeloid leukemia in children and adolescents: recommendations from an international expert panel', *Blood*, 120(16), pp. 3187–3205. Available at: https://doi.org/10.1182/blood-2012-03-362608.

Dalkic, E. *et al.* (2009) 'Integrative Analysis of Cancer Pathway Progression and Coherence', *Proteomics. Clinical applications*, 3(4), pp. 473–485. Available at: https://doi.org/10.1002/prca.200800074.

Davis, S. and Meltzer, P.S. (2007) 'GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor', *Bioinformatics*, 23(14), pp. 1846–1847. Available at: https://doi.org/10.1093/bioinformatics/btm254.

De Kouchkovsky, I. and Abdul-Hay, M. (2016) '"Acute myeloid leukemia: a comprehensive review and 2016 update".', *Blood cancer journal*, 6(7), p. e441. Available at: https://doi.org/10.1038/bcj.2016.50.

Deepak Shyl, E.S. *et al.* (2022) 'Mining of transcriptome identifies CD109 and LRP12 as possible biomarkers and deregulation mechanism of T cell receptor pathway in Acute Myeloid Leukemia', *Heliyon*, 8(10), p. e11123. Available at: https://doi.org/10.1016/j.heliyon.2022.e11123.

Deng, M. *et al.* (2016) ‘Web-TCGA: An online platform for integrated analysis of molecular cancer data sets’, *BMC Bioinformatics*, 17(1), pp. 1–7. Available at: https://doi.org/10.1186/s12859-016-0917-9.

Dhama, K. *et al.* (2019) ‘Biomarkers in stress related diseases/disorders: Diagnostic, prognostic, and therapeutic values’, *Frontiers in Molecular Biosciences*, 6(October). Available at: https://doi.org/10.3389/fmolb.2019.00091.

Dhiman, P. *et al.* (2022) ‘Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review’, *BMC Medical Research Methodology*, 22(1), p. 101. Available at: https://doi.org/10.1186/s12874-022-01577-x.

Dinardo, C.D. and Cortes, J.E. (2016) ‘Mutations in AML: prognostic and therapeutic implications’, *Hematology Am Soc Hematol Educ Program*, (1), pp. 348–355. Available at: https://doi.org/10.1182/asheducation-2016.1.348.

Ding, F. *et al.* (2021) ‘A review of the mechanism of DDIT4 serve as a mitochondrial related protein in tumor regulation’, *Science Progress*, 104(1), pp. 1–16. Available at: https://doi.org/10.1177/00368504211997273.

Dobin, A. *et al.* (2013) ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics*, 29(1), pp. 15–21. Available at: https://doi.org/10.1093/bioinformatics/bts635.

Docking, T.R. *et al.* (2021) ‘A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia’, *Nature Communications*, 12(1), p. 2474. Available at: https://doi.org/10.1038/s41467-021-22625-y.

Döhner, H. *et al.* (2010) ‘Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet’, *Blood*, 115(3), pp. 453–474. Available at: https://doi.org/10.1182/blood-2009-07-235358.

Döhner, H. *et al.* (2017) ‘Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel’, *Blood*, 129(4), pp. 424–447. Available at: https://doi.org/10.1182/blood-2016-08-733196.

Döhner, H. *et al.* (2022) ‘Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN’, *Blood*, 140(12), pp. 1345–1377. Available at: https://doi.org/10.1182/blood.2022016867.

Döhner, H., Weisdorf, D.J. and Bloomfield, C.D. (2015) ‘Acute Myeloid Leukemia’, *New England Journal of Medicine*. Edited by D.L. Longo, 373(12), pp. 1136–1152. Available at: https://doi.org/10.1056/NEJMra1406184.

Du, M. *et al.* (2022) ‘The Global Burden of Leukemia and Its Attributable Factors in 204 Countries and Territories: Findings from the Global Burden of Disease 2019 Study and Projections to 2030’, *Journal of Oncology*. Edited by I. Ilic, 2022, pp. 1–14. Available at: https://doi.org/10.1155/2022/1612702.

Edgar, R., Domrachev, M. and Lash, A.E. (2002) ‘Gene Expression Omnibus: NCBI gene expression and hybridization array data repository’, *Nucleic Acids Research*, 30(1), pp. 207–210. Available at: https://doi.org/10.1093/nar/30.1.207.

Egan, G. *et al.* (2021) ‘Treatment of acute myeloid leukemia in children: A practical perspective’, *Pediatric Blood & Cancer*, 68(7). Available at: https://doi.org/10.1002/pbc.28979.

Eshibona, N. *et al.* (2022) ‘Upregulation of FHL1, SPNS3, and MPZL2 predicts poor prognosis in pediatric acute myeloid leukemia patients with FLT3-ITD mutation’, *Leukemia and Lymphoma*, 0(0), pp. 1–10. Available at: https://doi.org/10.1080/10428194.2022.2045594.

Florkowski, C.M. (2008) ‘Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests’, *The Clinical Biochemist Reviews*, 29(Suppl 1), pp. S83–S87. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556590/ (Accessed: 22 December 2022).

Friedman, J., Hastie, T. and Tibshirani, R. (2010) ‘Regularization paths for generalized linear models via coordinate descent’, *Journal of Statistical Software*, 33(1), pp. 1–22. Available at: https://doi.org/10.18637/jss.v033.i01.

Fröhling, S. *et al.* (2004) ‘CEBPA mutations in younger adults with acute myeloid leukemia and normal cytogenetics: Prognostic relevance and analysis of cooperating mutations’, *Journal of Clinical Oncology*, 22(4), pp. 624–633. Available at: https://doi.org/10.1200/JCO.2004.06.060.

Fröhling, S. *et al.* (2006) ‘Cytogenetics and age are major determinants of outcome in intensively treated acute myeloid leukemia patients older than 60 years: results from AMLSG trial AML HD98-B’, *Blood*, 108(10), pp. 3280–3288. Available at: https://doi.org/10.1182/blood-2006-04-014324.

Fu, L. *et al.* (2017) ‘High expression of CPNE3 predicts adverse prognosis in acute myeloid leukemia’, *Cancer Science*, 108(9), pp. 1850–1857. Available at: https://doi.org/10.1111/cas.13311.

Fu, Y. *et al.* (2020) ‘Genome-wide identification of FHL1 as a powerful prognostic candidate and potential therapeutic target in acute myeloid leukaemia’, *EBioMedicine*, 52, p. 102664. Available at: https://doi.org/10.1016/j.ebiom.2020.102664.

Fujita, T.C. *et al.* (2021) ‘Acute lymphoid leukemia etiopathogenesis’, *Molecular Biology Reports*, 48(1), pp. 817–822. Available at: https://doi.org/10.1007/s11033-020-06073-3.

Gasparetto, M. *et al.* (2019) ‘Low ferroportin expression in AML is correlated with good risk cytogenetics, improved outcomes and increased sensitivity to chemotherapy’, *Leukemia Research*, 80(February), pp. 1–10. Available at: https://doi.org/10.1016/j.leukres.2019.02.011.

Gauthier, J. *et al.* (2019) ‘A brief history of bioinformatics’, *Briefings in Bioinformatics*, 20(6), pp. 1981–1996. Available at: https://doi.org/10.1093/bib/bby063.

Gerstung, M. *et al.* (2015) ‘Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes’, *Nature Communications*, 6, pp. 1–11. Available at: https://doi.org/10.1038/ncomms6901.

Goldman, M. *et al.* (2018) ‘The UCSC Xena platform for public and private cancer genomics data visualization and interpretation’, *bioRxiv*, p. 326470. Available at: https://doi.org/10.1101/326470.

Green, S.D. and Konig, H. (2020) ‘Treatment of Acute Myeloid Leukemia in the Era of Genomics-Achievements and Persisting Challenges’, *Frontiers in Genetics*, 11, p. 480. Available at: https://doi.org/10.3389/fgene.2020.00480.

Gregory, T.K. *et al.* (2009) ‘Molecular prognostic markers for adult acute myeloid leukemia with normal cytogenetics’, *Journal of Hematology & Oncology*, 2, p. 23. Available at: https://doi.org/10.1186/1756-8722-2-23.

Grimm, J. *et al.* (2020) ‘Prognostic impact of the ELN2017 risk classification in patients with AML receiving allogeneic transplantation.’, *Blood advances*, 4(16), pp. 3864–3874. Available at: https://doi.org/10.1182/bloodadvances.2020001904.

Grimwade, D. *et al.* (1998) ‘The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children’s Leukaemia Working Parties’, *Blood*, 92(7), pp. 2322–2333.

Grimwade, D. *et al.* (2010) ‘Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials’, *Blood*, 116(3), pp. 354–365. Available at: https://doi.org/10.1182/blood-2009-11-254441.

Grimwade, D. and Hills, R.K. (2009) ‘Independent prognostic factors for AML outcome’, *Hematology. American Society of Hematology. Education Program*, pp. 385–395. Available at: https://doi.org/10.1182/asheducation-2009.1.385.

Guo, Y., Wang, W. and Sun, H. (2022) ‘A systematic review and meta-analysis on the risk factors of acute myeloid leukemia’, *Translational Cancer Research*, 11(4), pp. 796–804. Available at: https://doi.org/10.21037/tcr-22-27.

Gupta, M., Mahapatra, M. and Saxena, R. (2019) ‘Cytogenetics’ impact on the prognosis of acute myeloid leukemia’, *Journal of Laboratory Physicians*, 11(2), pp. 133–137. Available at: https://doi.org/10.4103/JLP.JLP_164_18.

Hackl, H., Astanina, K. and Wieser, R. (2017) ‘Molecular and genetic alterations associated with therapy resistance and relapse of acute myeloid leukemia’, *Journal of Hematology & Oncology*, 10(1), p. 51. Available at: https://doi.org/10.1186/s13045-017-0416-0.

Haferlach, T. and Schmidts, I. (2020) ‘The power and potential of integrated diagnostics in acute myeloid leukaemia.’, *British journal of haematology*, 188(1), pp. 36–48. Available at: https://doi.org/10.1111/bjh.16360.

Han, J. *et al.* (2018) ‘Identification of Biomarkers Based on Differentially Expressed Genes in Papillary Thyroid Carcinoma’, *Scientific Reports*, 8(1), p. 9912. Available at: https://doi.org/10.1038/s41598-018-28299-9.

Handschuh, L. and Lonetti, A. (2019) ‘Not only Mutations Matter: Molecular Picture of Acute Myeloid Leukemia Emerging from Transcriptome Studies’, *Journal of Oncology*, 2019. Available at: https://doi.org/10.1155/2019/7239206.

Harris, N.L. *et al.* (1999) ‘The World Health Organization Classification of Neoplastic Diseases of the Hematopoietic and Lymphoid Tissues’, *Annals of Oncology*, 10(12), pp. 1419–1432. Available at: https://doi.org/10.1023/A:1008375931236.

Herold, T. *et al.* (2018) ‘A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia’, *Haematologica*, 103(3), pp. 456–465. Available at: https://doi.org/10.3324/haematol.2017.178442.

Heuser, M. *et al.* (2020) ‘Acute myeloid leukaemia in adult patients: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†’, *Annals of Oncology*, 31(6), pp. 697–712. Available at: https://doi.org/10.1016/j.annonc.2020.02.018.

Hollink, I. *et al.* (2009) ‘Favorable prognostic impact of NPM1 gene mutations in childhood acute myeloid leukemia, with emphasis on cytogenetically normal AML’, *Leukemia*, 23, pp. 262–270. Available at: https://doi.org/10.1038/leu.2008.313.

Hong, M. *et al.* (2020) ‘RNA sequencing: new technologies and applications in cancer research’, *Journal of Hematology and Oncology*, 13(1), pp. 1–16. Available at: https://doi.org/10.1186/s13045-020-01005-x.

Hu, X. *et al.* (2021) ‘A clinical prediction model identifies a subgroup with inferior survival within intermediate risk acute myeloid leukemia’, *Journal of Cancer*, 12(16), pp. 4912–4923. Available at: https://doi.org/10.7150/JCA.57231.

Huang, W. *et al.* (2020) ‘Prognostic significance of Spinster homolog gene family in acute myeloid leukemia’, *Journal of Cancer*, 11(15), pp. 4581–4588. Available at: https://doi.org/10.7150/jca.44766.

Huang, Y. *et al.* (2019) ‘Acute myeloid leukemia patient with FLT3-ITD and NPM1 double mutation should undergo allogeneic hematopoietic stem cell transplantation in CR1 for better prognosis’, *Cancer Management and Research*, 11, pp. 4129–4142. Available at: https://doi.org/10.2147/CMAR.S194523.

Huerga-Domínguez, S. *et al.* (2022) ‘Updates on the Management of Acute Myeloid Leukemia’, *Cancers*, 14(19), p. 4756. Available at: https://doi.org/10.3390/cancers14194756.

Infante, M.S., Piris, M.Á. and Hernández-Rivas, J.Á. (2018) ‘Molecular alterations in acute myeloid leukemia and their clinical and therapeutical implications’, *Medicina Clínica (English Edition)*, 151(9), pp. 362–367. Available at: https://doi.org/10.1016/j.medcle.2018.05.044.

Ingebriktsen, L.M. *et al.* (2022) ‘A novel age-related gene expression signature associates with proliferation and disease progression in breast cancer’, *British Journal of Cancer*, 127(10), pp. 1865–1875. Available at: https://doi.org/10.1038/s41416-022-01953-w.

Jeha, S. *et al.* (2002) ‘Comparison between pediatric acute myeloid leukemia (AML) and adult AML in VEGF and KDR (VEGF-R2) protein levels’, *Leukemia Research*, 26(4), pp. 399–402. Available at: https://doi.org/10.1016/S0145-2126(01)00149-7.

Jiang, G. *et al.* (2020) ‘Prognostic relevance of CD123 expression in adult AML with normal karyotype’, *British Journal of Haematology*, 188(1), pp. 181–184. Available at: https://doi.org/10.1111/bjh.16307.

Jiang, P. *et al.* (2022) ‘Big data in basic and translational cancer research’, *Nature Reviews Cancer*, 22(11), pp. 625–639. Available at: https://doi.org/10.1038/s41568-022-00502-0.

Juliusson, G. and Hough, R. (2016) ‘Leukemia’, *Progress in Tumor Research*, 43, pp. 87–100. Available at: https://doi.org/10.1159/000447076.

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

Kanehisa, M. *et al.* (2016) ‘KEGG as a reference resource for gene and protein annotation’, *Nucleic Acids Research*, 44(D1), pp. D457–D462. Available at: https://doi.org/10.1093/nar/gkv1070.

Kattner, P. *et al.* (2019) ‘Compare and contrast: pediatric cancer versus adult malignancies’, *Cancer and Metastasis Reviews*, 38(4), pp. 673–682. Available at: https://doi.org/10.1007/s10555-019-09836-y.

Kazim, N. *et al.* (2020) ‘The transcription elongation factor TCEA3 induces apoptosis in rhabdomyosarcoma.’, *Cell death & disease*, 11(1), p. 67. Available at: https://doi.org/10.1038/s41419-020-2258-x.

Kiem Hao, T. *et al.* (2020) ‘Long-term outcome of childhood acute myeloid leukemia: A 10-year retrospective cohort study’, *Pediatric Reports*, 12(1), p. 8486. Available at: https://doi.org/10.4081/pr.2020.8486.

Kim, D. *et al.* (2013) ‘TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions’, *Genome Biology*, 14(4), p. R36. Available at: https://doi.org/10.1186/gb-2013-14-4-r36.

Komanduri, K.V. and Levine, R.L. (2016) ‘Diagnosis and Therapy of Acute Myeloid Leukemia in the Era of Molecular Risk Stratification’, *Annual Review of Medicine*, 67, pp. 59–72. Available at: https://doi.org/10.1146/annurev-med-051914-021329.

Kong, Q.F. *et al.* (2020) ‘Association of von Willebrand factor (vWF) expression with lymph node metastasis and hemodynamics in papillary thyroid carcinoma’, *European Review for Medical and Pharmacological Sciences*, 24(5), pp. 2564–2571. Available at: https://doi.org/10.26355/eurrev_202003_20525.

Kourou, K. *et al.* (2015) ‘Machine learning applications in cancer prognosis and prediction’, *Computational and Structural Biotechnology Journal*, 13, pp. 8–17. Available at: https://doi.org/10.1016/j.csbj.2014.11.005.

Kuang, Y. *et al.* (2021) ‘New prognostic factors and scoring system for patients with acute myeloid leukemia’, *Oncology Letters*, 22(6), p. 823. Available at: https://doi.org/10.3892/ol.2021.13084.

Kumar, C.C. (2011) ‘Genetic Abnormalities and Challenges in the Treatment of Acute Myeloid Leukemia’, *Genes & Cancer*, 2(2), pp. 95–107. Available at: https://doi.org/10.1177/1947601911408076.

Law, C.W. *et al.* (2016) ‘RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR’, *F1000Research*, 5(1), p. 1408. Available at: https://doi.org/10.12688/f1000research.9005.1.

Lazarevic, V.L. and Johansson, B. (2020) 'Why classical cytogenetics still matters in acute myeloid leukemia', *Expert Review of Hematology*, 13(2), pp. 95–97. Available at: https://doi.org/10.1080/17474086.2020.1711733.

Lebedev, T.D. *et al.* (2019) 'Two receptors, two isoforms, two cancers: Comprehensive analysis of kit and trka expression in neuroblastoma and acute myeloid leukemia', *Frontiers in Oncology*, 9(OCT). Available at: https://doi.org/10.3389/fonc.2019.01046.

Lee, K.Y. *et al.* (2020) 'Elevation of CD109 promotes metastasis and drug resistance in lung cancer via activation of EGFR-AKT-mTOR signaling', *Cancer Science*, 111(5), pp. 1652–1662. Available at: https://doi.org/10.1111/cas.14373.

Lewandowska, A.M. *et al.* (2019) 'Environmental risk factors for cancer - review paper', *Annals of Agricultural and Environmental Medicine*, 26(1), pp. 1–7. Available at: https://doi.org/10.26444/aaem/94299.

Li, J. *et al.* (2015) 'TCEA3 attenuates gastric cancer growth by apoptosis induction', *Medical Science Monitor*, 21, pp. 3241–3246. Available at: https://doi.org/10.12659/MSM.895860.

Liao, J.M. *et al.* (2016) 'TFIIS.h, a new target of p53, regulates transcription efficiency of pro-apoptotic bax gene', *Scientific Reports*, 6(1), pp. 1–10. Available at: https://doi.org/10.1038/srep23542.

Lin, T.L. and Levy, M.Y. (2012) 'Acute myeloid leukemia: focus on novel therapeutic strategies', *Clinical Medicine Insights. Oncology*, 6, pp. 205–217. Available at: https://doi.org/10.4137/CMO.S7244.

Lin, X. *et al.* (2021) 'Global, regional, and national burdens of leukemia from 1990 to 2017: a systematic analysis of the global burden of disease 2017 study', *Aging*, 13(7), pp. 10468–10489. Available at: https://doi.org/10.18632/aging.202809.

Liu, Z., Spiegelman, V.S. and Wang, H.-G. (2022) 'Distinct noncoding RNAs and RNA binding proteins associated with high-risk pediatric and adult acute myeloid leukemias detected by regulatory network analysis', *Cancer Reports*, 5(10), p. e1592. Available at: https://doi.org/10.1002/cnr2.1592.

Lohse, I. *et al.* (2018) 'Precision medicine in the treatment stratification of AML patients: challenges and progress', *Oncotarget*, 9(102), pp. 37790–37797. Available at: https://doi.org/10.18632/oncotarget.26492.

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

Lu, D. *et al.* (2018) 'Knockdown of Linc00515 Inhibits Multiple Myeloma Autophagy and Chemoresistance by Upregulating miR-140-5p and

Downregulating ATG14', *Cellular Physiology and Biochemistry*, 48(6), pp. 2517–2527. Available at: https://doi.org/10.1159/000492690.

Luger, S.M. (2019) 'Acute myeloid leukemia: How to treat the fit patient over age 75?', *Best Practice and Research: Clinical Haematology*, 32(4), p. 101105. Available at: https://doi.org/10.1016/j.beha.2019.101105.

Ma, X. *et al.* (2018) 'Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours.', *Nature*, 555(7696), pp. 371–376. Available at: https://doi.org/10.1038/nature25795.

Maleki Behzad, M. *et al.* (2021) 'Effects of Lifestyle and Environmental Factors on the Risk of Acute Myeloid Leukemia: Result of a Hospital-based Case-Control Study', *Journal of Research in Health Sciences*, 21(3), pp. e00525–e00525. Available at: https://doi.org/10.34172/jrhs.2021.58.

Martelli, A.M. *et al.* (2006) 'Phosphoinositide 3-kinase/Akt signaling pathway and its therapeutical implications for human acute myeloid leukemia', *Leukemia*, 20(6), pp. 911–928. Available at: https://doi.org/10.1038/sj.leu.2404245.

Martínez-Jiménez, F. *et al.* (2020) 'A compendium of mutational cancer driver genes', *Nature Reviews Cancer*, 20(10), pp. 555–572. Available at: https://doi.org/10.1038/s41568-020-0290-x.

Mazzone, M. and Comoglio, P.M. (2006) 'The Met pathway: master switch and drug target in cancer progression', *The FASEB Journal*, 20(10), pp. 1611–1621. Available at: https://doi.org/10.1096/fj.06-5947rev.

Meshinchi, S. *et al.* (2006) 'Clinical implications of FLT3 mutations in pediatric AML', *Blood*, 108(12), pp. 3654–3661. Available at: https://doi.org/10.1182/blood-2006-03-009233.

Metzeler, K.H. *et al.* (2016) 'Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia', *Blood*, 128(5), pp. 686–698. Available at: https://doi.org/10.1182/blood-2016-01-693879.

Miranda-Filho, A. *et al.* (2018) 'Epidemiological patterns of leukaemia in 184 countries: a population-based study', *The Lancet Haematology*, 5(1), pp. e14–e24. Available at: https://doi.org/10.1016/S2352-3026(17)30232-6.

Moarii, M. and Papaemmanuil, E. (2017) 'Classification and risk assessment in AML: integrating cytogenetics and molecular profiling', *Hematology: the American Society of Hematology Education Program*, 2017(1), pp. 37–44. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6142605/ (Accessed: 14 February 2023).

Mohammad, T. *et al.* (2022) 'Differential Gene Expression and Weighted Correlation Network Dynamics in High-Throughput Datasets of Prostate Cancer', *Frontiers in Oncology*, 12, p. 881246. Available at: https://doi.org/10.3389/fonc.2022.881246.

Mojiri, A. *et al.* (2017) ‘Functional assessment of von Willebrand factor expression by cancer cells of non-endothelial origin’, *Oncotarget*, 8(8), pp. 13015–13029. Available at: https://doi.org/10.18632/oncotarget.14273.

Mrózek, K. *et al.* (1997) ‘Clinical significance of cytogenetics in acute myeloid leukemia’, *Seminars in oncology*, 24(1), pp. 17–31.

Mrózek, K. *et al.* (2012) ‘Prognostic Significance of the European LeukemiaNet Standardized System for Reporting Cytogenetic and Molecular Alterations in Adults With Acute Myeloid Leukemia’, *Journal of Clinical Oncology*, 30(36), pp. 4515–4523. Available at: https://doi.org/10.1200/JCO.2012.43.4738.

Mrózek, K. (2022) ‘Molecular cytogenetics in acute myeloid leukemia in adult patients: practical implications’, *Polish Archives of Internal Medicine* [Preprint]. Available at: https://doi.org/10.20452/pamw.16300.

Murphy, M.F.G. *et al.* (2013) ‘Childhood and adult cancers: Contrasts and commonalities’, *Maturitas*, 76(1), pp. 95–98. Available at: https://doi.org/10.1016/j.maturitas.2013.05.017.

Myers, J.S. *et al.* (2015) ‘Differentially Expressed Genes and Signature Pathways of Human Prostate Cancer’, *PLoS ONE*, 10(12), p. e0145322. Available at: https://doi.org/10.1371/journal.pone.0145322.

Nair, R.M. *et al.* (2020) ‘Overexpression of metastasis-associated in colon cancer 1 in retinoblastoma’, *Tumor Biology*, 42(11). Available at: https://doi.org/10.1177/1010428320975973.

Nalejska, E., Mączyńska, E. and Lewandowska, M.A. (2014) ‘Prognostic and predictive biomarkers: Tools in personalized oncology’, *Molecular Diagnosis and Therapy*, 18(3), pp. 273–284. Available at: https://doi.org/10.1007/s40291-013-0077-9.

Neill, T., Schaefer, L. and Iozzo, R.V. (2016) ‘Decorin as a multivalent therapeutic agent against cancer’, *Advanced Drug Delivery Reviews*, 97, pp. 174–185. Available at: https://doi.org/10.1016/j.addr.2015.10.016.

Nepstad, I. *et al.* (2020) ‘The PI3K-Akt-mTOR Signaling Pathway in Human Acute Myeloid Leukemia (AML) Cells’, *International Journal of Molecular Sciences*, 21(8), p. 2907. Available at: https://doi.org/10.3390/ijms21082907.

Ni, Q.Z. *et al.* (2020) ‘Epithelial V-like antigen 1 promotes hepatocellular carcinoma growth and metastasis via the ERBB-PI3K-AKT pathway’, *Cancer Science*, 111(5), pp. 1500–1513. Available at: https://doi.org/10.1111/cas.14331.

Nunes, A. de L. *et al.* (2019) 'Cytogenetic abnormalities, WHO classification, and evolution of children and adolescents with acute myeloid leukemia', *Hematology, Transfusion and Cell Therapy*, 41(3), pp. 236–243. Available at: https://doi.org/10.1016/j.htct.2018.09.007.

Ohtsu, N. *et al.* (2016) 'Tumor and Stem Cell Biology Eva1 Maintains the Stem-like Character of Glioblastoma-Initiating Cells by Activating the Noncanonical NF-kB Signaling Pathway', *Cancer Res*, 76(1), pp. 171–181. Available at: https://doi.org/10.1158/0008-5472.CAN-15-0884.

Oran, B. and Weisdorf, D.J. (2012) 'Survival for older patients with acute myeloid leukemia: A population-based study', *Haematologica* [Preprint]. Available at: https://doi.org/10.3324/haematol.2012.066100.

Pan, Y. *et al.* (2017) 'Analysis of differential gene expression profile identifies novel biomarkers for breast cancer', *Oncotarget*, 8(70), pp. 114613–114625. Available at: https://doi.org/10.18632/oncotarget.23061.

Papaemmanuil, E. *et al.* (2016) 'Genomic Classification and Prognosis in Acute Myeloid Leukemia', *The New England Journal of Medicine*, 374(23), pp. 2209–2221. Available at: https://doi.org/10.1056/NEJMoa1516192.

Patro, R. *et al.* (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods*, 14(4), pp. 417–419. Available at: https://doi.org/10.1038/nmeth.4197.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python.', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: http://scikit-learn.sourceforge.net. (Accessed: 12 August 2021).

Pourrajab, F. *et al.* (2020) 'Genetic characterization and risk stratification of acute myeloid leukemia', *Cancer Management and Research*, 12, pp. 2231–2253. Available at: https://doi.org/10.2147/CMAR.S242479.

Quessada, J. *et al.* (2021) 'Cytogenetics of Pediatric Acute Myeloid Leukemia: A Review of the Current Knowledge', *Genes*, 12(6), p. 924. Available at: https://doi.org/10.3390/genes12060924.

Rapaport, F. *et al.* (2013) 'Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data', *Genome Biology*, 14(9). Available at: https://doi.org/10.1186/gb-2013-14-9-r95.

Rashid, H.U. *et al.* (2019) 'Research advances on anticancer activities of matrine and its derivatives: An updated overview', *European Journal of Medicinal Chemistry*, 161, pp. 205–238. Available at: https://doi.org/10.1016/j.ejmech.2018.10.037.

Rausch, C. *et al.* (2022) ‘Validation of the 2022 European Leukemianet Genetic Risk Stratification of Acute Myeloid Leukemia’, *Blood*, 140(Supplement 1), pp. 3408–3409. Available at: https://doi.org/10.1182/blood-2022-167022.

Rhodes, D.R. *et al.* (2004) ‘Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression’, *Proceedings of the National Academy of Sciences of the United States of America*, 101(25), pp. 9309–9314. Available at: https://doi.org/10.1073/pnas.0401994101.

Ritchie, M.E. *et al.* (2015) ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic Acids Research*, 43(7), p. e47. Available at: https://doi.org/10.1093/nar/gkv007.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009) ‘edgeR: A Bioconductor package for differential expression analysis of digital gene expression data’, *Bioinformatics*, 26(1), pp. 139–140. Available at: https://doi.org/10.1093/bioinformatics/btp616.

Röllig, C. *et al.* (2011) ‘Long-term prognosis of acute myeloid leukemia according to the new genetic risk classification of the European leukemianet recommendations: Evaluation of the proposed reporting system’, *Journal of Clinical Oncology*, 29(20), pp. 2758–2765. Available at: https://doi.org/10.1200/JCO.2010.32.8500.

Rollins-Raval, M. *et al.* (2013) ‘CD123 Immunohistochemical expression in acute myeloid leukemia is associated with underlying FLT3-ITD and NPM1 mutations’, *Applied Immunohistochemistry and Molecular Morphology*, 21(3), pp. 212–217. Available at: https://doi.org/10.1097/PAI.0b013e318261a342.

de Rooij, J.D.E., Zwaan, C.M. and van den Heuvel-Eibrink, M. (2015) ‘Pediatric AML: From Biology to Clinical Management’, *Journal of Clinical Medicine*, 4(1), pp. 127–149. Available at: https://doi.org/10.3390/jcm4010127.

Saultz, J. and Garzon, R. (2016) ‘Acute Myeloid Leukemia: A Concise Review’, *Journal of Clinical Medicine*, 5(3), p. 33. Available at: https://doi.org/10.3390/jcm5030033.

Savary, C. *et al.* (2020) ‘Depicting the genetic architecture of pediatric cancers through an integrative gene network approach’, *Scientific Reports*, 10(1), pp. 1–15. Available at: https://doi.org/10.1038/s41598-020-58179-0.

Seeger, R.C. *et al.* (1985) ‘Association of Multiple Copies of the N- myc Oncogene with Rapid Progression of Neuroblastomas’, *New England Journal of Medicine*, 313(18), pp. 1111–1116. Available at: https://doi.org/10.1056/nejm198510313131802.

Segeren, C. and Vantveer, M. (1996) ‘The FAB classification for acute myeloid leukaemia?is it outdated?’, *The Netherlands Journal of Medicine*, 49(3), pp. 126–131. Available at: https://doi.org/10.1016/0300-2977(96)00024-1.

Shi, X. *et al.* (2015) ‘Decorin is responsible for progression of non-small-cell lung cancer by promoting cell proliferation and metastasis’, *Tumor Biology*, 36(5), pp. 3345–3354. Available at: https://doi.org/10.1007/s13277-014-2968-8.

Short, N.J., Rytting, M.E. and Cortes, J.E. (2018) ‘Acute myeloid leukaemia’, *The Lancet*, 392(10147), pp. 593–606. Available at: https://doi.org/10.1016/S0140-6736(18)31041-9.

Siegel, R.L. *et al.* (2022) ‘Cancer statistics, 2022’, *CA: A Cancer Journal for Clinicians*, 72(1), pp. 7–33. Available at: https://doi.org/10.3322/caac.21708.

Siegel, R.L., Miller, K.D. and Jemal, A. (2020) ‘Cancer statistics, 2020’, *CA: A Cancer Journal for Clinicians*, 70(1), pp. 7–30. Available at: https://doi.org/10.3322/caac.21590.

Simon, N. *et al.* (2011) ‘Regularization paths for Cox’s proportional hazards model via coordinate descent’, *Journal of Statistical Software*, 39(5), pp. 1–13. Available at: https://doi.org/10.18637/jss.v039.i05.

Siveen, K.S., Uddin, S. and Mohammad, R.M. (2017) ‘Targeting acute myeloid leukemia stem cell signaling by natural products.’, *Molecular cancer*, 16(1), pp. 1–12. Available at: https://doi.org/10.1186/s12943-016-0571-x.

Stein, U. *et al.* (2009) ‘MACC1, a newly identified key regulator of HGF-MET signaling, predicts colon cancer metastasis’, *NATURE MEDICINE VOLUME*, 15(1), pp. 59–67. Available at: https://doi.org/10.1038/nm.1889.

Steliarova-Foucher, E. *et al.* (2017) ‘International incidence of childhood cancer, 2001–10: a population-based registry study’, *The Lancet Oncology*, 18(6), pp. 719–731. Available at: https://doi.org/10.1016/S1470-2045(17)30186-9.

Stölzel, F. *et al.* (2016) ‘Karyotype complexity and prognosis in acute myeloid leukemia’, *Blood Cancer Journal*, 6(1), pp. 7–10. Available at: https://doi.org/10.1038/bcj.2015.114.

Sunde, R.A. (2010) ‘mRNA transcripts as molecular biomarkers in medicine and nutrition☆’, *The Journal of Nutritional Biochemistry*, 21(8), pp. 665–670. Available at: https://doi.org/10.1016/j.jnutbio.2009.11.012.

Sung, H. *et al.* (2021) ‘Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.’, *CA: a cancer journal for clinicians*, 71(3), pp. 209–249. Available at: https://doi.org/10.3322/caac.21660.

Szalontay, L. and Shad, A.T. (2014) ‘Pediatric Acute Myeloid Leukemia: How to Improve Outcome?’, *Current Pediatrics Reports*, 2(1), pp. 26–37. Available at: https://doi.org/10.1007/s40124-013-0036-2.

Tarlock, K. *et al.* (2018) ‘Distinct age-associated molecular profiles in acute myeloid leukemia defined by comprehensive clinical genomic profiling’, *Oncotarget*, 9(41), pp. 26417–26430. Available at: https://doi.org/10.18632/oncotarget.25443.

Tarlock, K. and Meshinchi, S. (2015) ‘Pediatric acute myeloid leukemia: Biology and therapeutic implications of genomic variants’, *Pediatric Clinics of North America*, 62(1), pp. 75–93. Available at: https://doi.org/10.1016/j.pcl.2014.09.007.

Tazi, Y. *et al.* (2022) ‘Unified classification and risk-stratification in Acute Myeloid Leukemia’, *Nature Communications*, 13(1). Available at: https://doi.org/10.1038/s41467-022-32103-8.

Thol, F. (2021) ‘What to use to treat AML: the role of emerging therapies’, *Hematology: the American Society of Hematology Education Program*, 2021(1), pp. 16–23. Available at: https://doi.org/10.1182/hematology.2021000309.

Thol, F. and Schlenk, R.F. (2014) ‘Gemtuzumab ozogamicin in acute myeloid leukemia revisited’, *Expert Opinion on Biological Therapy*, 14(8), pp. 1185–1195. Available at: https://doi.org/10.1517/14712598.2014.922534.

Tibshirani, R. *et al.* (2002) ‘Diagnosis of multiple cancer types by shrunken centroids of gene expression’, *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp. 6567–6572. Available at: https://doi.org/10.1073/pnas.082099299.

Tibshirani, R. *et al.* (2012) ‘Strong rules for discarding predictors in lasso-type problems’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(2), pp. 245–266. Available at: https://doi.org/10.1111/j.1467-9868.2011.01004.x.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) ‘The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge’, *Wspolczesna Onkologia*, 1A, pp. A68–A77. Available at: https://doi.org/10.5114/wo.2014.47136 .

Tong, G. *et al.* (2019) ‘MACC1 regulates PDL1 expression and tumor immunity through the c-Met/AKT/mTOR pathway in gastric cancer cells’, *Cancer Medicine*, 8(16), pp. 7044–7054. Available at: https://doi.org/10.1002/cam4.2542 .

Toro-Domínguez, D. *et al.* (2019) ‘ImaGEO: integrative gene expression meta-analysis from GEO database’, *Bioinformatics*, 35(5), pp. 880–882. Available at: https://doi.org/10.1093/bioinformatics/bty721 .

Torrebadell, M. *et al.* (2018) 'A 4-gene expression prognostic signature might guide post-remission therapy in patients with intermediate-risk cytogenetic acute myeloid leukemia', *Leukemia and Lymphoma*, 59(10), pp. 2394–2404. Available at: https://doi.org/10.1080/10428194.2017.1422859.

Tukey, J.W. (1949) 'Comparing Individual Means in the Analysis of Variance Author (s): John W . Tukey Published by : International Biometric Society Stable . *International Biometric Society*, 5(2), pp. 99–114. Available at: http://www.jstor.org/stable/3001913.

Vakiti, A. and Mewawalla, P. (2022) 'Acute Myeloid Leukemia', in *StatPearls*. Treasure Island (FL): StatPearls Publishing. Available at: http://www.ncbi.nlm.nih.gov/books/NBK507875/ (Accessed: 12 December 2022).

Vardiman, J.W. *et al.* (2009) 'The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: Rationale and important changes', *Blood*, 114(5), pp. 937–951. Available at: https://doi.org/10.1182/blood-2009-03-209262.

Verhaak, R.G.W.W. *et al.* (2009) 'Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling', *Haematologica*, 94(1), pp. 131–134. Available at: https://doi.org/10.3324/haematol.13299.

Verma, D. *et al.* (2022) 'BAALC gene expression tells a serious patient outcome tale in NPM1-wild type/FLT3-ITD negative cytogenetically normal-acute myeloid leukemia in adults', *Blood Cells, Molecules, and Diseases*, 95, p. 102662. Available at: https://doi.org/10.1016/j.bcmd.2022.102662.

Walter, R.B. *et al.* (2013) 'Significance of FAB subclassification of "acute myeloid leukemia, NOS" in the 2008 WHO classification: Analysis of 5848 newly diagnosed patients', *Blood*, 121(13), pp. 2424–2431. Available at: https://doi.org/10.1182/blood-2012-10-462440.

Wang, T. *et al.* (2022) 'Frequency and clinical impact of WT1 mutations in the context of CEBPA-mutated acute myeloid leukemia', *Hematology*, 27(1), pp. 994–1002 .

Wang, Z., Jensen, M.A. and Zenklusen, J.C. (2016) 'A practical guide to The Cancer Genome Atlas (TCGA)', in *Methods in Molecular Biology*, pp. 111–141. Available at: https://doi.org/10.1007/978-1-4939-3578-9_6.

Wheatley, K. *et al.* (2009) 'Prognostic factor analysis of the survival of elderly patients with AML in the MRC AML11 and LRF AML14 trials', *British Journal of Haematology*, 145(5), pp. 598–605. Available at: https://doi.org/10.1111/j.1365-2141.2009.07663.x.

Woolley, J.F., Dzneladze, I. and Salmena, L. (2015) ‘Phosphoinositide signaling in cancer: INPP4B Akt(s) out’, *Trends in Molecular Medicine*, 21(9), pp. 530–532. Available at: https://doi.org/10.1016/j.molmed.2015.06.006.

Wouters, B.J. and Delwel, R. (2016) ‘Epigenetics and approaches to targeted epigenetic therapy in acute myeloid leukemia’, *Blood*, 127(1), pp. 42–52. Available at: https://doi.org/10.1182/blood-2015-07-604512.

Wu, D., Rice, C.M. and Wang, X. (2012) ‘Cancer bioinformatics: A new approach to systems clinical medicine’, *BMC Bioinformatics*, 13(1), pp. 13–16. Available at: https://doi.org/10.1186/1471-2105-13-71.

Wu, X. *et al.* (2016) ‘Prognostic significance of FLT3-ITD in pediatric acute myeloid leukemia: a meta-analysis of cohort studies’, *Molecular and Cellular Biochemistry*, 420(1–2), pp. 121–128. Available at: https://doi.org/10.1007/s11010-016-2775-1.

Wu, Z. and Lin, Y. (2019) ‘Long noncoding RNA LINC00515 promotes cell proliferation and inhibits apoptosis by sponging mir-16 and activating PRMT5 expression in human glioma’, *OncoTargets and Therapy*, 12, pp. 2595–2604. Available at: https://doi.org/10.2147/OTT.S198087.

Xue, C. *et al.* (2019) ‘Elevated SPINK2 gene expression is a predictor of poor prognosis in acute myeloid leukemia’, *Oncology Letters*, 18(3), pp. 2877–2884. Available at: https://doi.org/10.3892/ol.2019.10665.

Ye, X. *et al.* (2019) ‘The incidence, risk factors, and survival of acute myeloid leukemia secondary to myelodysplastic syndrome: A population-based study’, *Hematological Oncology*, 37(4), pp. 438–446. Available at: https://doi.org/10.1002/hon.2660.

Yin, X. *et al.* (2018) ‘Identification of long non-coding RNA competing interactions and biological pathways associated with prognosis in pediatric and adolescent cytogenetically normal acute myeloid leukemia’, *Cancer Cell International*, 18, p. 122. Available at: https://doi.org/10.1186/s12935-018-0621-0.

Yoshioka, K.-I. *et al.* (2021) ‘Genomic Instability and Cancer Risk Associated with Erroneous DNA Repair’, *International Journal of Molecular Sciences*, 22(22), p. 12254. Available at: https://doi.org/10.3390/ijms222212254.

Young, A.L. *et al.* (2019) ‘Clonal hematopoiesis and risk of acute myeloid leukemia’, *Haematologica*, 104(12), pp. 2410–2417. Available at: https://doi.org/10.3324/haematol.2018.215269.

Yu, G. *et al.* (2012) ‘ClusterProfiler: An R package for comparing biological themes among gene clusters’, *OMICS A Journal of Integrative Biology*, 16(5), pp. 284–287. Available at: https://doi.org/10.1089/omi.2011.0118.

Yu, P. *et al.* (2020) 'High Expression of the SH3TC2-DT / SH3TC2 Gene Pair Associated With FLT3 Mutation and Poor Survival in Acute Myeloid Leukemia : An Integrated TCGA Analysis', 10(June), pp. 1–15. Available at: https://doi.org/10.3389/fonc.2020.00829.

Zhao, X., Li, Y. and Wu, H. (2018) 'A novel scoring system for acute myeloid leukemia risk assessment based on the expression levels of six genes', *International Journal of Molecular Medicine*, 42(3), pp. 1495–1507. Available at: https://doi.org/10.3892/ijmm.2018.3739.

Zhou, F. and Chen, B. (2021) 'Prognostic significance of ferroptosis-related genes and their methylation in AML', *Hematology (United Kingdom)*, 26(1), pp. 919–930. Available at: https://doi.org/10.1080/16078454.2021.1996055.

Zhu, R. *et al.* (2020) 'Identification of an Immune-Related Gene Signature Based on Immunogenomic Landscape Analysis to Predict the Prognosis of Adult Acute Myeloid Leukemia Patients', *Frontiers in Oncology*, 10, p. 1. Available at: https://doi.org/10.3389/fonc.2020.574939.

Zhu, W. *et al.* (2020) 'The Application of Deep Learning in Cancer Prognosis Prediction', *Cancers*, 12(3), p. 603. Available at: https://doi.org/10.3390/cancers12030603.

Supplementary Table

Supplementary Table 4.1: List of significant DEGs between SS and LS. Up-regulation ($\log_2$ fold change > 1 and adjusted *p*-value < 0.01) and down-regulation ($\log_2$ fold change < -1 and adjusted *p*-value < 0.01)

PROBEID	SYMBOL	adj.P.Val	logFC
203948_s_at	MPO	6.66433722721692E-07	-1.72692340999298
206940_s_at	POU4F1	2.99304871047971E-07	-1.62030038280376
211341_at	POU4F1	4.09282639057784E-09	-1.50074305365523
206622_at	TRH	8.08283334949699E-13	-1.44338266419843
203949_at	MPO	1.95633487577802E-06	-1.44279703513265
228827_at	RUNX1T1	2.34940598929572E-08	-1.44068393399573
205529_s_at	RUNX1T1	1.43188160413641E-07	-1.37944834127139
210755_at	HGF	4.61931627940309E-07	-1.37844488835543
1556395_at	NA	1.40712909401244E-08	-1.27328264275696
205528_s_at	RUNX1T1	4.88149475962196E-09	-1.22807447972022
219890_at	CLEC5A	2.18394017821693E-06	-1.21314230417013
209960_at	HGF	6.37438244373354E-07	-1.19307588474501
204885_s_at	MSLN	1.4285543854801E-06	-1.09885463573598
206871_at	ELANE	0.00289341802260763	-1.09007669098783
210997_at	HGF	2.10185979492379E-07	-1.08334261066224
202760_s_at	PALM2AKAP2	3.20749222852502E-06	-1.03662595092478
206135_at	ST18	3.38929160340253E-05	-1.03656585877682
226694_at	PALM2AKAP2	1.87652495899579E-05	-1.01188591826101
209392_at	ENPP2	8.01597460509493E-06	1.01999968557689
205608_s_at	ANGPT1	3.37073197990753E-05	1.02209939714152
212070_at	ADGRG1	2.71179359337343E-09	1.02369614629028
205237_at	FCN1	0.000964305606661832	1.02891401701537
228708_at	RAB27B	1.6645107596387E-06	1.03103241488127
201110_s_at	THBS1	0.000912064539401371	1.03253023834489
209555_s_at	CD36	0.00142336729432303	1.03692773653932
203523_at	LSP1	1.058099465274E-09	1.04078578752424
228766_at	CD36	0.00179909758565153	1.04481849237915
206471_s_at	PLXNC1	3.47107235490379E-11	1.04513420467289
227856_at	FAM241A	4.43984044278956E-07	1.0486654746159
205453_at	HOXB2	3.12126323737221E-06	1.05116104965351
228372_at	TMEM273	2.40792271690105E-06	1.05375659972598
213056_at	FRMD4B	1.51837648810166E-08	1.05925003724532
202118_s_at	CPNE3	6.51205436049427E-11	1.06091498007244
203741_s_at	ADCY7	1.73936867121803E-10	1.07017594808137
212386_at	TCF4	1.95633487577802E-06	1.07400194125269
211597_s_at	HOPX	1.99957085724885E-06	1.07732077565928

217853_at	TNS3	1.58136813802767E-07	1.08247881453307
205898_at	CX3CR1	0.000802424997408272	1.08294487642028
205844_at	VNN1	5.81081268921827E-05	1.08317054423163
224596_at	SLC44A1	2.20400412506535E-06	1.08563302518562
208792_s_at	CLU	4.1985776275758E-06	1.08637095112643
205767_at	EREG	0.00426328570026295	1.08937130708731
217800_s_at	NDFIP1	2.07429129794218E-07	1.10340104048834
201669_s_at	MARCKS	0.000549944485851676	1.11140092285802
235046_at	INPP4B	1.5632243995606E-07	1.11322417808221
202890_at	MAP7	2.08059065739214E-09	1.11534179028208
213110_s_at	COL4A5	0.000536846674299879	1.11775284448327
208791_at	CLU	7.75291783876769E-06	1.11997003006859
225512_at	ZBTB38	4.34525531513468E-10	1.12388026880831
210145_at	PLA2G4A	5.34531176012095E-10	1.12875711386747
206494_s_at	ITGA2B	3.34364882016399E-08	1.14020163802072
202887_s_at	DDIT4	6.25131153334214E-09	1.14159491900488
202119_s_at	CPNE3	1.31325682609609E-11	1.14602334980558
1559477_s_at	MEIS1	7.0390264882506E-08	1.15178101301953
227236_at	TSPAN2	9.18351435683882E-08	1.17541076621842
226545_at	CD109	1.427419373503E-08	1.2035099909681
204082_at	PBX3	4.21444253796353E-06	1.20649096251638
215646_s_at	VCAN	0.00770273391925119	1.2156232701604
208029_s_at	LAPTM4B	2.17321366540848E-06	1.22899530426182
223204_at	GASK1B	0.000169080411871333	1.25167532851227
212192_at	KCTD12	0.000632539040536299	1.26913166723962
238778_at	MPP7	1.00181902752795E-09	1.27507027869064
205609_at	ANGPT1	8.97389734416107E-06	1.28605125089355
1554679_a_at	LAPTM4B	1.37248941990576E-06	1.2952538339301
203373_at	SOCS2	3.10398572171745E-06	1.30015085913583
205612_at	MMRN1	2.14281709311068E-08	1.33372158044422
213241_at	PLXNC1	4.3539160907754E-12	1.39637739218553
212314_at	SEL1L3	2.00896864576196E-10	1.39639885594187
1553808_a_at	NKX2-3	1.39927546742931E-06	1.40024148342575
236738_at	C3orf80	1.57316988799411E-07	1.40586020285082
222717_at	CAVIN2	6.093405431701E-10	1.42325738753958
203680_at	PRKAR2B	2.32288150005616E-08	1.42460371713351
203372_s_at	SOCS2	2.82759196320902E-06	1.44935848946594
235521_at	HOXA3	8.06457710725795E-08	1.48780858453495
217963_s_at	BEX3	2.17540302441161E-10	1.54509735615073
204069_at	MEIS1	1.77246642396798E-09	1.58822863308414
217975_at	TCEAL9	1.15738767302774E-10	1.59228413982703
206478_at	FAM30A	4.74940257995711E-12	1.73105003266287
213844_at	HOXA5	2.36042250744854E-08	1.76682399643656
228365_at	CPNE8	1.41224811931834E-13	1.77427718942994

214039_s_at	LAPTM4B	8.20184976721563E-08	1.79418792260646
228904_at	HOXB3	5.28947169890202E-07	1.82206709508076
201427_s_at	SELENOP	4.54091536457947E-06	1.82471579449115
223044_at	SLC40A1	8.08283334949699E-13	1.96820903194534
213150_at	HOXA10	8.08283334949699E-13	2.02113737641849
214146_s_at	PPBP	3.4666938660632E-08	2.03848293040479
206310_at	SPINK2	1.41224811931834E-13	2.18370800351311