University of the Western Cape

South African National Bioinformatics Institute

Robert Sobukwe Rd, Bellville, Cape Town, 7535

Telephone: + 27 21 929 3645

Email: 3648936@myuwc.ac.za

Identification of insertion-induced enhancers linked to gene drivers within non-coding DNA using a pipeline for Diffuse Large B-cell Lymphoma H3K27ac ChIP-seq data.

MSc Bioinformatics

Supervisor: Dr Hocine Bendou

Student Researcher: Wardah Jassiem

Student number: 3648936

November 2022

DECLARATION

I, Wardah Jassiem, student number 3648936, declare that 'Identification of insertion-induced enhancers linked to gene drivers within non-coding DNA using a pipeline for Diffuse Large B-cell Lymphoma H3K27ac ChIP-seq data' is my own work and that all the sources I have quoted have been indicated and acknowledged by means of complete references.

Signed 10 of November 2022 at the University of the Western Cape.

Signature: ……

ABSTRACT

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin lymphoma (NHL) and incorporates a diverse range of illnesses with varying biology, clinical manifestations, and therapeutic responses. Functional insertion mutations represent the driving mechanism behind many oncologic illnesses. Research has shown that variants associated with cancer in the non-coding portion of the genome, which is enriched with enhancer elements, is greatly underappreciated. The present study designed a bioinformatics pipeline using Nextflow DSL2 to identify insertion-induced enhancers associated with DLBCL oncogenes within the non-coding genome using H3K27ac ChIP-seq data. Gapped DLBCL reads identified by bowtie were mapped to the human reference genome with bowtie2. Non-coding insertions were identified with BEDTools and verified by pBlat. Putative enhancers located by MACS2 were intersected with the non-coding insertions. Genes linked to the identified non-coding insertion-induced enhancers were generated by BEDOPS before functional analysis was performed using DAVID. The insertion mutations were observed to target chromosomes that were gene rich, correlating to areas of the genome high in GC content and accessible to transcription. This resulted in a strong, positive correlation between enhancer rate and gene count. The enhancers were largely proximal, situated within or near the transcription start site (TSS) of their associated genes, among which were found known oncogenes relevant to several cancer types, such as *DYRK1A*, *COPB2*, *FOXP1*, *IPO11*, *PRDM2* and *PRDM15*, or specifically to DLBCL, such as *SMC3*, *MIR155HG*, *PIM1* and *NOTCH1*. Functional analysis placed the affected genes in lymphoma pathways involved with cell growth and survival, apoptosis, chemotherapy resistance, sustained angiogenesis, and metastasis. The study highlighted the non-coding genome's potential contribution to DLBCL tumorigenesis through the dysregulated effect of mutated enhancers on gene expression. The research provided a framework for further investigation of non-coding anomalies across human malignancies.

KEYWORDS: cancer, diffuse large B-cell lymphoma, enhancer, insertion, pipeline

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr Hocine Bendou, for his patience while I was learning new skills during my research conductance. I appreciate the motivation I was given to develop my abilities as well as the guidance from which I gained valuable practical knowledge. My gratitude goes to the National Research Fund for financially providing me with the opportunity to do my MSc, and to the dynamic SANBI team who were always welcoming and helpful. I would also like to express my appreciation to my parents and grandparents for their prayers throughout my studies, my brother for his kindness in providing me with a space to work, and to the rest of my family for the many generous ways in which they eased the path towards the completion of this work. Lastly, a special thank you to Edith Jacobs for her incomparable support.

ABBREVIATIONS

ABC Activated B-cell

BAM Binary Alignment/Map

BCL Base Call

BCL2 B-Cell Lymphoma 2

BCL6 B-Cell Lymphoma 6

BCR B-Cell Receptor

BED Browser Extensible Data

BLAST Basic Local Alignment Search Tool

Blat BLAST-Like Alignment Tool

BP Base Pair

BTK Bruton's Tyrosine Kinase

ChIP-Seq Chromatin Immunoprecipitation Sequencing

CNS DLBCL Central Nervous System Diffuse Large B-cell Lymphoma

COO Cell Of Origin

COPB2 COPI Coat Complex Subunit Beta 2

COSMIC Catalogue Of Somatic Mutations In Cancer

CRC Colorectal Cancer

DAVID Database for Annotation, Visualisation, and Integrated Discovery

DLBCL Diffuse Large B-cell Lymphoma

DLBCL NOS Diffuse Large B-cell Lymphoma Not Otherwise Specified

DSL Domain Scripting Language

DYRK1A Dual Specificity Tyrosine Phosphorylation Regulated Kinase 1A

DZ Dark Zone

EBV Epstein-Barr Virus

EZH1 Enhancer Of Zeste 1 Polycomb Repressive Complex 2 Subunit

EZH2 Enhancer Of Zeste 2 Polycomb Repressive Complex 2 Subunit

FOXP1 Forkhead Box P1

FTP File Transfer Protocol

GC Germinal Centre

GCB Germinal Centre B-cell

GO Gene Ontology

HPC High Performance Computing

H3K27ac Histone H3 Lysine 27 Acetylation

H3K4me1 Histone H3 Lysine 4 Monomethylation

H3K4me3 Histone H3 Lysine 4 Trimethylation

ICGC International Cancer Genome Consortium

IGV Integrative Genomics Viewer

INDEL Insertion and Deletion

IPO11 Importin 11

LINC01800 Long Intergenic Non-Protein Coding RNA 1800

LZ Light Zone

MACS Model-based Analysis of ChIP-seq

MIR155HG MIR155 Host Gene

MYC MYC Proto-Oncogene

NCBI National Centre for Biotechnology Information

NF-κB Nuclear Factor κB

NGS Next Generation Sequencing

NHL Non-Hodgkin Lymphoma

NOTCH1 NOTCH receptor 1

OS Operating System

p53 Tumor Protein p53 Gene

pBlat Parallel Blast Like Alignment Tool

PCAWG Pan-Cancer Analysis of Whole Genomes

PI3K Phosphatidylinositol 3-Kinase

PIM1 Pim-1 proto-oncogene, serine/threonine kinase

PLWH People Living With HIV

PRDM PR/SET Domain Family

PRDM2 PR/SET Domain 2

PRDM15 PR/SET Domain 15

PTEN Phosphatase and Tensin Homolog Deleted on Chromosome 10

R-CHOP Rituximab, Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone

SAM Sequence Alignment/Map

SEMA4D Semaphorin 4D

SIF Singularity Image File

SIMD Single-Instruction Multiple-Data

SMC3 Structural Maintenance of Chromosomes Protein 3

SNP Single Nucleotide Polymorphism

SNV Single-Nucleotide Variant

SRA Sequence Read Archive

SRR Sequence Read Run

TCGA The Cancer Genome Atlas

TF Transcription Factor

TRAJ35 T-cell Receptor Alpha Joining 35

TRAPPC2B Trafficking Protein Particle Complex Subunit 2B

TSS Transcription Start Site

UCSC University of California Santa Cruz

UV Ultraviolet

WES Whole Exome Sequencing

WGS Whole-Genome Sequencing

WHO World Health Organisation

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. BACKGROUND

Non-Hodgkin's lymphoma (NHL) results from the growth and accumulation of a single mature clone of lymphocytes (Gouveia, Siqueira and Pereira, 2012). Diffuse large B-cell lymphoma (DLBCL), the most common subtype of NHL, originates in the germinal centre and is characterized by a diffuse production of matured and enlarged B-cells (Frick, Dörken and Lenz, 2011). Although lasting remissions can be found in over 50% of cases, DLBCL remains a difficult clinical issue, with one-third of patients remaining uncured by standard immunochemotherapeutic regimens (Bakhshi and Georgel, 2020).

DLBCL makes up 30–40% of all NHL cases (Evrard *et al.*, 2019). In South Africa, DLBCL comprises up to 43% of NHL cases (Pather and Patel, 2022). DLBCL is the most prevalent cancer in people living with HIV (PLWH), and South Africa is home to one of the biggest HIV epidemics worldwide (Naidoo *et al.*, 2018). It is suggested that HIV may promote lymphomagenesis through direct and indirect interactions with B-cells (Re, Cattaneo and Rossi, 2019). HIV-related lymphomas are clinically distinct. In comparison to the non-HIV infected population, there is a greater predisposition to advanced stage, aggressive presentation, and extranodal activity (Pather and Patel, 2022). A clearer understanding on the molecular pathways affected in DLBCL may provide insight into the disease mechanism behind the manifestation of lymphoma in HIV infected patients that may be used as a guide in therapeutic development.

DLBCL is generally comprised of two major subgroups based on cell of origin (COO); activated B-cell (ABC), and germinal centre B-cell (GCB) DLBCL (Chettiankandy *et al.*, 2016). ABC DLBCL is the more aggressive of the two subtypes; associated with poorer outcomes, it is characterized by constitutive activation of the nuclear factor κB (NF-κB) pathway which is involved in cell immunity and regulation of cell differentiation, proliferation, and apoptosis (Nowakowski and Czuczman, 2015). ABC DLBCL develops from B-cells immediately after they have left the germinal centre (Roschewski, Phelan and Wilson, 2020). GCB DLBCL develops in the germinal-centre dark zone, where B-cells encounter antigens and somatic hypermutation

occurs (Hunter *et al.*, 2020). Each subgroup has their own molecular features indicative of differential pathogenesis. Up to 20% of DLBCL cases go unclassified (Schrader *et al.*, 2022). Refining what is known about the mutational variability on a molecular level may aid in the production of alternate approaches to disease management, including precision therapy.

Tumour genomes have numerous DNA variants that differentiates them from healthy cell genomes, but only a fraction are driver mutations involved in pathogenesis (Abraham *et al.*, 2017). Among the most poorly understood of these variants are insertions (Abraham *et al.*, 2017). Evidence has shown that somatic insertions can create binding motifs for master transcription factors, which then erroneously create super-enhancers that stimulate the overexpression of oncogenes (Mansour *et al.*, 2016). However, insertions frequently passed undetected due to limitations involving the alignment of short reads generated by sequencing technologies to human reference genomes (Mansour *et al.*, 2016).

While many variants in coding regions found in cancer cells via next generation sequencing (NGS) have been studied, there has been significantly less research into the importance of non-coding variants in cancer (Huang *et al.*, 2020). Those that have been studied were found to play critical roles in tumorigenesis suggesting non-coding mutations are underappreciated (Abraham *et al.*, 2017). Examples of functional non-coding mutations that disturb enhancers and promoters and expression of their target genes include a single-nucleotide polymorphism in the *LMO1* enhancer of neuroblastoma patients, variants in the *TERT* promoter of different cancer types, and heterozygous indels (insertions and deletions) in a super-enhancer linked to the *TAL1* promoter in T-acute lymphoblastic leukaemia patients (He *et al.*, 2020).

Non-coding DNA comprises most of the human genome and contains functional cis regulatory DNA elements like enhancers that control protein coding genes in a signal-dependent manner and are frequently dysregulated by cancerous mutations (Elliott and Larsson, 2021). Dysregulation may be brought on by trans regulatory processes, like the activation of transcription factors or epigenetic regulators that control enhancer activity, or by cis regulatory mutations that alter enhancer activity or specificity of its target gene (Elliott and Larsson, 2021). These activities produce super-enhancers that are specific to a particular type of tumour and which create a gene regulatory state that maintains the uncontrollable growth of cancer cells (Wang, Yan and Cairns, 2019). Acetylation of the histone mark H3K27

distinguishes active from inactive enhancers (Creyghton *et al.*, 2010). H3K27ac ChIP-seq allows the detection of enhancers through its ability to trace transcription factors which activate enhancers (Abraham *et al.*, 2017). Since H3K27ac sequence reads are largely produced from active regulatory regions, a more direct relationship between the variant and potential function is provided.

Bioinformatics tools arranged within a pipeline is used to analyse the data generated by ChIP-seq and convert it to meaningful information (Federico *et al.*, 2019). Workflow management systems enable bioinformatics pipelines to efficiently arrange analysis stages while processing large quantities of ChIP-seq data across varying computational environments (Ahmed *et al.*, 2021). Nextflow is a workflow framework and programming DSL that allows scalable, reproducible, and inherently parallel workflows using container technology to ensure efficient deployment (di Tommaso *et al.*, 2017).

Two steps typical of ChIP-seq analytical pipelines used in epigenetic studies involve mapping and variant calling, for which several programs have been developed. Bowtie2 allows quick and memory-efficient large-scale alignment of short, gapped sequencing reads to a reference genome (Langmead and Salzberg, 2012). This enables the detection of mutation events such as insertions in ChIP-seq data. The tool pBlat is a parallelised sequence alignment algorithm typically used for the analysis and comparison of biological sequences and can verify ChIP-seq sequences with insertions. MACS2 is used to identify transcription factor binding sites (Feng, Liu and Zhang, 2011) and, when coupled with H3K27ac ChIP-seq data, can be used to locate areas in the genome enriched with enhancer activity. The BEDTools suite offers tools for the exploration of genomic datasets through arithmetic tasks and can be used to find overlapping features between two datasets (Quinlan and Hall, 2010). Causative insertion events can therefore be linked to putative mutated enhancer activity within ChIP-seq data. The BEDOPS suite can then link a list of variants to the nearest genes based on genomic distance (Neph *et al.*, 2012), thereby suggesting biological significance.

Algorithms such as those described have been used in past studies; Abraham *et al.* (2017) successfully identified enhancer-associated small insertion variants near known oncogenes. Later research by Huang *et al.* (2020) computationally rebuilt the techniques described by Abraham *et al.* (2017) and developed a database cataloguing enhancer-associated insertion

and deletion variants for human and murine ChIP-seq data. The combination of an experimental and computational approach is therefore suggested to optimise genome-wide detection of enhancer-associated variations in tumour cells. As a result, the search for diagnostic biomarkers and therapeutic targets may turn its focus to the non-coding genome.

1.2. PROBLEM STATEMENT

DLBCL is a genetically diverse form of cancer with notable variations in manifestation and pathogenesis. Driver mutation discovery is essential to comprehending DLBCL oncogenesis and therapeutic response. The search for driver mutations in B-cell lymphoma has mostly centred on coding areas, but many tumours lack obvious driver mutations. Since non-coding sequences make approximately 98% of the human genome, it is the location of most somatic mutations in cancer genomes. Insertions were one of the main topics of earlier investigations into non-coding mutations, of which several have now been identified. Enhancers, which are non-coding regulatory sequences, are important regulators of gene expression. Enhancers and the expression of their target genes have been known to be disrupted by insertions in DLBCL. Detection of functional non-coding insertion-induced enhancers is crucial to identifying genes and pathways relevant to the onset, progression, and outcome of DLBCL pathogenesis.

1.3. RESEARCH AIM

Mutations in the non-coding genome may have an impact on the progression of tumorigenesis that is not fully realised. This study was geared to determine via a bioinformatics pipeline for H3K27ac ChIP-seq data analysis whether genomic insertions in non-coding regions can create novel enhancers associated with DLBCL oncogenes for the purpose of developing a database that future studies on DLBCL can use in South Africa.

1.4. RESEARCH OBJECTIVES

The objectives of this study were to:

a) Access DLBCL H3K27ac ChIP-seq data through public repositories and published literature.
b) Create a bioinformatics pipeline to find non-coding insertion-induced enhancers.
c) Identify genes associated with the detected novel enhancers.
d) Perform functional enrichment analysis to analyse the enhancer associated genes' involvement in biological pathways relevant to cancer.

1.5. DISSERTATION STRUCTURE

Chapter 1 describes the development of the thesis idea, its rationale and significance. It includes the main research problem that drove the investigation, and the aim and objectives the study strived to achieve to explore potential solutions to the issues described. It indicates the research framework, structure, and methodology. Chapter 2 involves deep insight into literature surrounding the study, the boundaries of the study and the research upon which the study was built. It discusses the latest technology and virtualisation of biological studies. Chapter 3 explains the research hypothesis, concepts, research instruments and processes used and the rationale behind their selection. It stipulates the specificities involving data gathering methodology, data editing and data coding. It also describes the steps taken to reduce error and validate findings. Chapter 4 describes the properties of the published data accessed together with the results drawn from the computational pipeline as well as patterns and relationships observed. Chapter 5 discusses key points, anomalies, and unexpected discoveries along with potential explanations. It underlines the importance of the study and areas that require further investigation. It also covers the study's potential practical implications. Chapter 6 summarises the key points of the investigation and areas of potential research expansion.

CHAPTER 2

LITERATURE REVIEW

2.1. INTRODUCTION

The purpose of the following chapter is to explain the reasoning behind the conception of and methodology undertaken in the present study. The chapter describes the biological theory and computational practices that gave rise to the study hypothesis and the research techniques employed to achieve the study objectives. It begins with the classification of DLBCL, the influence of genetic factors on pathogenesis, and the significance of DLBCL in South Africa. This is followed by an analysis of genomic components such as enhancers, insertions, and the non-coding genome relevant to cancer as well as DLBCL specifically. Previous literature involving similar themes are discussed and their methods and results summarised to describe the foundation upon which this study was built. Subsequently, the value of computational biology in modern science is explored along with the various bioinformatics tools that can be utilised to answer data intensive questions.

2.2. CELL OF ORIGIN CLASSIFICATION OF DLBCL

DLBCL can develop spontaneously or by transition from low grade B-cell malignancies such follicular lymphoma or chronic lymphocytic leukaemia, also known as Richter's transformation. DLBCL is caused by mature B-cells becoming cancerous after they have undergone the germinal centre response (Pasqualucci and Dalla-Favera, 2015). When B-cells are exposed to a foreign antigen, microanatomical compartments called germinal centres (GC) develop (Pasqualucci and Dalla-Favera, 2015). Clonal growth and antibody affinity maturation occurs within these compartments. The B-cells are recycled in two different regions: the dark zone (DZ), made up of proliferating cells that undergo somatic hypermutation to alter their immunoglobulin genes; and the light zone (LZ), where B-cells become either plasma cells or memory B-cells based on their affinity for the antigen (Tripodo *et al.*, 2020).

The World Health Organisation (WHO) partitioned DLBCL into distinct subgroups such as T-cell/histiocyte-rich B-cell lymphoma, primary DLBCL of the Central Nervous System (CNS DLBCL), primary cutaneous DLBCL, Epstein-Barr virus (EBV)-positive DLBCL of the elderly, and DLBCL not otherwise specified (NOS) (Beham-Schmid, 2017). DLBCL NOS is the most common category representing cases that do not fit into any specific disease subgroup and demonstrates a broad cytologic spectrum (Collares *et al.*, 2019). In DLBCL NOS, two main molecular phenotypes exist (Xie, Pittaluga and Jaffe, 2015). These phenotypes, later discussed, show the importance of the GC as a target of malignant transformation. Germinal centre B-cell (GCB) DLBCL lacks expression of early post-GC differentiation markers, whereas activated B-cell (ABC) DLBCL exhibits a transcriptional signature like that in activated B-cells or in lymphocytes poised to plasma cell differentiation (Nowakowski and Czuczman, 2015).

2.2.1. ACTIVATED B-CELL SUBTYPE

ABC DLBCL is derived from B-cells in the process of differentiating into plasma cells. Genes routinely expressed in normal germinal centre B-cells are downregulated, whilst genes routinely expressed in normal plasma cells are upregulated (Frick, Dörken and Lenz, 2011). The inactivation of the PR/SET domain family (*PRDM*) member *PRDM1*, which encodes BLIMP1 and promotes plasmacytic differentiation, suggests that a block in differentiation is a trait of ABC subtype (Xia *et al.*, 2017). The INKa/ARF tumour suppressor locus is deleted in 30% of cases and the amplification of 18q is associated with overexpression of the anti-apoptotic BCL2 protein (Frick, Dörken and Lenz, 2011). Mutations in the B-cell receptor (BCR) subunits (CD79A and CD79B) and in regulators of the NF-κB pathway (MYD88) work in unison to stimulate the activation of the NF-κB transcription factor complex via constitutive B-cell receptor signalling (Weber and Schmitz, 2022). These processes consequently promote cell survival and proliferation and impedes apoptosis. Inhibition of the NF-κB pathway is toxic to ABC but not to GCB-cell lines (Frick, Dörken and Lenz, 2011). ABC patients respond poorly to standard rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) immunochemotherapy as compared to GCB patients (Miao *et al.*, 2019). ABC DLBCL patients that carry MYD88 and CD79B mutations show a significantly better response to ibrutinib, a Bruton's tyrosine kinase (BTK) inhibitor, than patients with other ABC DLBCL tumours (Roschewski, Phelan and Wilson, 2020).

The pivotal role played by BCR signalling in the pathogenesis of ABC DLBCL validates the concept of the cell-of-origin classification.

2.2.2. GERMINAL CENTRE B-CELL SUBTYPE

GCB DLBCL arises from malignant B-cells with overexpression of genes involved in the germinal centre response including ongoing somatic hypermutation and CD10 expression. Common mutations that define GCB DLBCL include B-cell lymphoma 2 (*BCL2*) chromosomal translocations, enhancer of zeste 2 polycomb repressive complex 2 subunit (*EZH2)* oncogenic mutations, *REL* amplification and alterations in gene phosphatase and tensin homolog deleted on chromosome 10 (*PTEN)* (Roschewski, Phelan and Wilson, 2020). A t(14;18) translocation placing the *BCL2* gene and the regulatory elements of the immunoglobulin heavy chain locus close together leads to activation of the anti-apoptotic BCL2 protein (Frick, Dörken and Lenz, 2011). Loss of function of *PTEN* activates the phosphatidylinositol 3-kinase (PI3K) pathway stimulating cell proliferation and survival. EZH2 inhibits proliferation checkpoint genes and establishes bivalent chromatin domains thereby enabling a GC B-cell specific gene expression program (Weber and Schmitz, 2022). GCB DLBCL also shows amplification of *MDM2* (Miao *et al.*, 2018), a negative regulator of the tumour suppressor p53, as well as deletions of the known tumour suppressor genes *TP73* and *ING1* (Frick, Dörken and Lenz, 2011). The 2016 WHO classification demarcated a subgroup of high-grade DLBCL with a GCB phenotype based on MYC proto-oncogene (*MYC*) and *BCL2* translocations (Roschewski, Phelan and Wilson, 2020). This so-called double hit lymphoma presents in approximately 8% of DLBCL cases and has an extremely poor prognosis (Roschewski, Phelan and Wilson, 2020).

2.2.3. UNCLASSIFIED

B-cell Lymphoma unclassifiable is a group of high-grade B-cell lymphomas recognised by the WHO that cannot be classified as either Burkitt's lymphoma or DLBCL because it has features common to both groups (Chettiankandy *et al.*, 2016). Improvements in molecular characterization techniques and the introduction of new medicines that target particular DLBCL subtypes have laid the groundwork for individualized treatment of DLBCL based on molecular subtype. However, up to 20% of DLBCL NOS cases remain unclassified with biological traits largely unknown (Roschewski, Phelan and Wilson, 2020).

The cell of origin classification does not cover rare subtypes like T-cell/histiocyte-rich large B-cell lymphoma, suggesting the need for future refinement of the classification system (Roschewski, Phelan and Wilson, 2020).

2.2.4. GENETIC SUBTYPES OF DLBCL

The biology and prognosis of the DLBCL ABC and GCB subgroups continue to show significant variability. Defining genetic events have been found exclusively within and across cell of origin subtypes and can further sub-classify DLBCL (Figure 1).

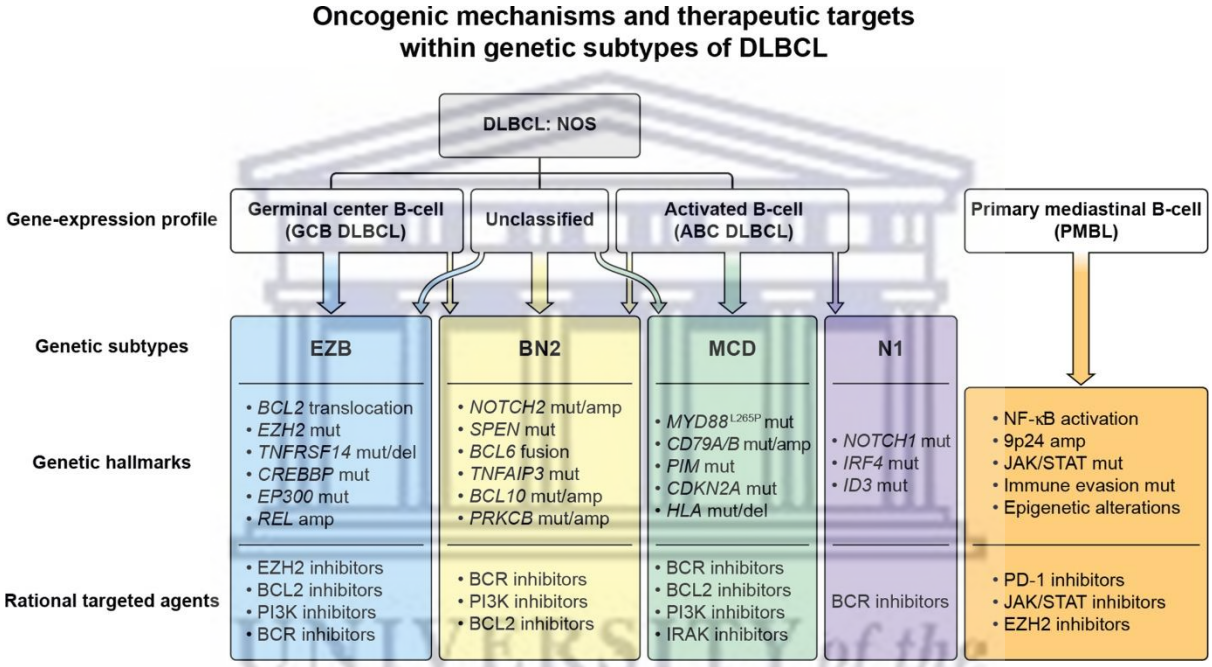


Figure 1: Genetic subtypes of DLBCL with their mutational mechanisms and therapeutic targets (Roschewski, Phelan and Wilson, 2020).

EZB is typically a GCB DLBCL with *BCL2* translocations, mutated *PTEN* and *EZH2*, and *REL* amplifications (Weber and Schmitz, 2022). *TNFRSF14*, *CREBBP*, and *EP300* are often inactivated. EZB has the best overall prognosis of DLBCL, but the worst within the GCB subgroup (Weber and Schmitz, 2022). Targeted therapy for EZB includes inhibitors of *EZH2* and *BCL2*, or of proximal BCR signalling and the PI3K signalling pathway (Roschewski, Phelan and Wilson, 2020). BN2, found across DLBCL cell of origin subtypes, is defined by B-cell lymphoma 6 (*BCL6*) fusions and alterations of the BCR and NOTCH signalling pathways (Kotlov *et al.*, 2021). It frequently shows mutations in *NOTCH2*. BN2 has the best prognosis within ABC DLBCL and inhibitors of *BCL2*, BCR or PI3K signalling are standard treatments (Roschewski, Phelan and Wilson, 2020). MCD is usually an ABC DLBCL characterized by MYD88

and CD79B mutations. Other recurrent mutations in MCD are seen in primary CNS lymphoma, including pim-1 proto-oncogene, serine/threonine kinase (*PIM1*) (Roschewski, Phelan and Wilson, 2020). Treatment of MCD includes inhibitors of BTK, PI3K, *BCL2*, and *IRAK4*. N1 is almost exclusively an ABC DLBCL; characterized by NOTCH receptor 1 (*NOTCH1*) mutations, it is associated with the worst prognosis along with MCD. N1 includes variations in *IRF4* and *ID3* controlling B-cell differentiation (Kotlov *et al.*, 2021). N1 has chronic active BCR signalling susceptible to BTK inhibitors (Roschewski, Phelan and Wilson, 2020).

Primary mediastinal B-cell lymphoma, which is distinct from DLBCL, is a post-thymic B-cell derivative (Roschewski, Phelan and Wilson, 2020). Several mutations that promote immune evasion are observed, such as *EZH2* variations that lessen the production of MHC class I and MHC class II.

More research is required to understand the therapeutic implications of these subgroups with reference to the function of targeted treatment. Precision medicine is transforming patient care by customising treatment based on genetic and phenotypic traits that distinguish between patients exhibiting comparable clinical presentations (Nowakowski *et al.*, 2019). The strategy is being increasingly used in clinical practice with improved patient response and survival. Two prime examples are trastuzumab for HER2 positive breast cancer, and vemurafenib for BRAF V600E positive melanoma (Nowakowski *et al.*, 2019). Both are novel therapies developed due to the application of molecular profiling. The study of genetic pathways provides for the identification of diagnostic and prognostic indicators, as well as for the development of precision medicine strategies focused at addressing oncogenic addictions particular to distinct DLBCL classes, e.g., proteasome inhibitors that reduce NF-κB signalling and BCR pathway inhibitors for ABC DLBCL (Roschewski, Staudt and Wilson, 2014).

2.3. GENOMIC FACTORS INFLUENCING DLBCL PATHOGENESIS

2.3.1. ONCOGENES, TUMOUR SUPPRESSOR GENES, AND TRANSCRIPTION FACTORS

The diverse and aggressive nature of diffuse large B-cell lymphoma (DLBCL), as well as its resistance and relapse after standard treatment, have spurred the exploration of the pathological mechanisms that sustain the disease. The introduction of powerful genomic technologies has enabled a deeper characterisation of the genetic and molecular landscape of DLBCL. DLBCL is maintained by the build-up of genetic abnormalities that change the structure or expression of proto-oncogenes, tumour suppressor genes, transcription factors, and other molecules of pathogenetic importance. The mutations induce dysregulation of biological functions crucial to the maintenance of healthy germinal centre B-cells.

The *BCL2* gene encodes an anti-apoptotic protein that is involved in resistance to chemotherapy (Dunleavy and Wilson, 2011). *BCL2* regulation is disrupted in DLBCL by the t(14;18) translocation, gene amplification or NF-κB signalling (Schuetz *et al.*, 2012). *BCL2* is the most mutated gene in GCB DLBCL (Schuetz *et al.*, 2012). Studies have shown a higher disease-free survival rate in DLBCL without the positively correlated expression of *BCL2* and *Cyclin D2* (Amen *et al.*, 2007). The *BCL6* gene, like the *BCL2* gene, is an anti-apoptotic factor important to B-cell development. *BCL6* is a proto-oncogene expressed in normal B-cells that blocks genes involved in cell cycle progression and response to DNA damage (Evrard *et al.*, 2019). *BCL6* expression is higher in the GCB subtype than in the ABC subtype of DLBCL (Li *et al.*, 2019). *BCL6* expression can become dysregulated because of mutations in the 5' non-coding region; in around 13% of DLBCL cases, these mutations prevent *BCL6* from negatively regulating its own expression (Bakhshi and Georgel, 2020). Approximately 40% of DLBCL instances have chromosomal translocations that result in increased *BCL6* expression (Bakhshi and Georgel, 2020). Chromosomal translocations in the *BCL6* gene come from disorders in the sequence of the promoter region of DNA. Alterations in BCL6 protein expression cause failure in cell differentiation and continuous cell proliferation with improved cell survival.

The tumour protein p53 gene (*p53*) encodes a phosphoprotein p53 that regulates DNA transcription and repair, autophagy, and apoptosis. Mutated *p53* is one of the most frequently mutated genes in GCB and ABC subtypes. Observed in 20%–25% of DLBCL cases, it is an unfavourable prognostic factor for patients undergoing R-CHOP treatment.

Patients with absent or mutated *p53* show more aggressive disease and worse prognosis and some studies suggest that *p53* may be inactivated by the *BCL6* gene during the start of lymphoma (Wang *et al.*, 2017).

The tumour suppressor p53 negatively regulates *c-Myc*, the dysregulation of which impacts the inferior prognosis in B-cell lymphoma (Yu, Yu and Young, 2019). *MYC* is a regulatory gene that is involved in cell cycle regulation, metabolism, and apoptosis (Wang *et al.*, 2017). *MYC* translocations can upregulate many growth-promoting genes. Recombination of the *MYC* gene with other genes is especially found in extranodal lymphomas and is linked to lower remission and survival rates (Richardson *et al.*, 2019). *MYC* rearrangements, usually from chromosomal translocation, is an instigator in many types of B-cell lymphoma including 10–15% of DLBCL (Wang *et al.*, 2017). Double-expressor DLBCL characterised by the overexpression of *MYC* and *BCL2*, and double-hit lymphoma characterised by the dual translocation of *MYC* with *BCL2/BCL6*, represent subgroups of DLBCL with poor prognosis (Xia and Zhang, 2020). Overexpression of anti-apoptotic protein BCL2 together with *MYC* activation induces uncontrolled cell proliferation (Xia and Zhang, 2020).

B-cell differentiation and proliferation are controlled by transcription factors, two of which are the OCT-1 and OCT-2 proteins (Heckman *et al.*, 2006). The OCT-2 protein is highly expressed in mature B-cells but not in pre-B-cells, T-cells, myelomonocytic and epithelial cells (Gouveia, Siqueira and Pereira, 2012). The OCT-1 protein is highly expressed in pre-B-cells and may be involved in the early development of B-cells. OCT factors affect the survival of cells in lymphomas with the t(14; 18) translocation (the transfer of a chromosome 18 segment containing a certain gene to an IgH downstream site on chromosome 14, like that which occurs with *BCL2*) (Gouveia *et al.*, 2020). A positive correlation was found between OCT, BOB1 and BCL2 expressions. OCT-2 activates the *BCL2* gene promoter and is involved in the malignant transformation in B-lymphomas (Heckman *et al.*, 2006). Apoptosis is inversely proportional to low expression of OCT-1, OCT-2 and BOB1 (Heckman *et al.*, 2006). Gene forkhead box P1 (*FOXP1*), which encodes another transcription factor relevant to DLBCL, distinguishes the ABC subtype from the GCB subtype (Gascoyne and Banham, 2017). FOXP1 regulates the expression of ABC DLBCL signatures, such as the NF-κB and MYD88 pathways (Gascoyne and Banham, 2017). By opposing pathways specific to GCB DLBCL, such as those of the GCB regulator *BCL6*, FOXP1 increases gene expression fundamental to the transition of

the GCB cell to the plasmablast, which is the transitory B-cell stage targeted in ABC DLBCL transformation (Dekker *et al.*, 2016).

2.3.2. ENHANCERS

Enhancers are one of several cis-regulatory elements (non-coding DNA regulatory elements) that work in unison to regulate transcription by controlling cell state and cell differentiation (Wang, Yan and Cairns, 2019). Enhancers are comprised of short DNA regions that can be bound by transcription factors to activate gene expression regardless of their orientation or distance (Panigrahi and O'Malley, 2021). Single enhancers in non-coding areas outside of model genes, found by studies of elements related to disease followed by large scale comparative genomics, suggest certain cis-regulatory elements influence disease in a very big way (Creyghton *et al.*, 2010).

Enhancer dysfunction is one of the main mechanisms behind the abnormal regulation of oncogenes in cancer and is typically induced by epigenetic processes (Yao *et al.*, 2020). Super enhancers, a subclass of regulatory domains, are made up of large enhancer clusters with a stronger ability to promote gene expression than typical enhancers (Bal *et al.*, 2022). During normal cell development, super enhancers or very strong enhancers are often located close to genes that determine lineage (He, Long and Liu, 2019). During tumorigenesis, super-enhancers form de novo near oncogenes and enlist enhancer-binding proteins to activate gene expression (He, Long and Liu, 2019). This locks the growth regulation network in an activated state and promotes unchecked proliferation (Sur and Taipale, 2016). Super enhancers are often filled with H3K4me1, H3K27ac, p300, Mediator, RNA polymerase II, BRD4, CDK7, and other master transcription factors (He, Long and Liu, 2019).

Histones can undergo several post-transcriptional changes, such as acetylation, phosphorylation, methylation, and ubiquitination, which affect how histones and DNA interact, hence affecting how genes are expressed globally (Zhang *et al.*, 2020). Histone acetylation, like that of histone H3 on lys9 and lys27, is often associated with transcription that is active, while deacetylation results in transcriptional silence (Gao *et al.*, 2020). Histone acetylation entails the covalent modification of lysine with an acetyl group. Normally a positively charged amino acid, lysine has a strong binding affinity for the negatively charged DNA molecule (Ellenbroek and Youn, 2016). A more open structure that is easier for the

transcriptional machinery to access results from the inclusion of the acetyl group, which neutralizes this positive charge and lessens the interaction between histones and DNA (Ellenbroek and Youn, 2016).

Histone modifications, which are utilized as markers to distinguish putative enhancers, are necessary for enhancer activity (Zhang *et al.*, 2020). The mark for enhancer priming is histone H3 lysine 4 monomethylation (H3K4me1) (Tang *et al.*, 2020). However, the identification of enhancers focuses on the acetylation of histone H3 on lysine 27 (H3K27) to find specific cell type enhancer sites (Huang *et al.*, 2021). Active enhancers are marked with H3K4me1 and H3K27ac, with reduction of histone H3 lysine 4 trimethylation (H3K4me3); silent or inactive enhancers are marked with only H3K4me3 (Creyghton *et al.*, 2010). Thus, Histone H3K27ac can differentiate active from inactive enhancers containing H3K4me1 and is the most extensively researched histone acetylation used for the detection of enhancers and super-enhancers in published ChIP-seq studies (Gao *et al.*, 2020).

Genome-wide association studies have revealed that most cancerous mutations are found outside the exome (Huang *et al.*, 2020) in regions enriched with enhancer elements (Sur and Taipale, 2016). Mutations in regulatory elements can alter their activity, e.g., an indel in T-ALL creates a super-enhancer that drives overexpression of *TAL1* gene (Hung *et al.*, 2019). Sporadic tumours usually have enhancers with somatic mutations, like copy number changes that increase enhancer affinity and activity, structural rearrangements that direct enhancers to new targets and point mutations or insertions and deletions that create new enhancers by changing transcription factor binding sites (Sur and Taipale, 2016). Malfunctioning enhancers are the biggest contributor to heritable cancer predisposition (Sur and Taipale, 2016). Different cancers have been observed to lose or acquire super enhancers (Xu *et al.*, 2022). The expression of downstream oncogenes crucial for the development of DLBCL have been found to be stimulated by super enhancers (Chapuy *et al.*, 2013). Active super enhancers connected to the proto-oncogenes *BCL6*, *BCL2*, and *CXCR4* were found to prevent transcriptional repressors from binding to and downregulating the target gene (Bal *et al.*, 2022). DLBCL treatment with super enhancer inhibitors have been shown to reduce the expression of oncogenes impacted in this manner (Xu *et al.*, 2022).

Although transcription is a characteristic of all cells, cancer cells are dependent on increased transcription from enhancers, making them particularly vulnerable to enhancer inhibition (Sur and Taipale, 2016). Cancer therapy can be turned toward enhancer dysfunction which could pinpoint critical factors that directly contribute to pathogenesis. These can be methodically examined through procedures like ChIP-seq of histone mark H3K27ac. By focusing on mutations in reported enhancers, the search space can be reduced while statistical power is increased; indels and low frequency variants in regulatory regions can be detected (Huang *et al.*, 2020). Additionally, candidate variants can be efficiently linked to target transcripts.

2.3.3. GENOMIC INSERTIONS AND DELETIONS

The non-coding regions of tumour genomes have a lot of DNA variation, but the contribution of these variants to tumorigenesis is poorly understood (Abraham *et al.*, 2017). Somatic insertions are among the least defined due to challenges with interpreting short-read DNA sequences (Abraham *et al.*, 2017). Insertions and deletions (indels) are additions or deletions of one or more nucleotides in a DNA sequence (Gagliano *et al.*, 2019). Studies have estimated that 16% to 25% of sequence polymorphisms are indels (Chen and Guo, 2021). Indels in both coding and non-coding regions have been associated with Mendelian and complex diseases (Gagliano *et al.*, 2019). Indels are important in clinical NGS because they're the driving procedure behind many constitutional and oncologic diseases, they are also typically a mechanism of kinase activation in cancer which is a feature exploited by targeted therapy with kinase inhibitors (Sehn, 2015).

In coding regions, an indel that is not in-frame will change the reading frame resulting in a protein product different to the wild type, e.g., 40 or more CAG repeats in the first exon of the *huntingtin* gene results in Huntington's disease (Lench *et al.*, 2013). Indels that are in-frame can also result in altered proteins, e.g., a deletion in the cystic fibrosis transmembrane conductance regulator gene that leads to cystic fibrosis (Mullaney *et al.*, 2010).

A 2010 study suggested that indels are often under positive selection and can therefore be oncogenic driver mutations (Yang *et al.*, 2010). The study established a strong correlation between indels and base substitutions in cancer-related genes and observed a tendency of the indels to group at the same locus in the coding sequences of the same samples. Furthermore, a larger amount of indels were found in somatic mutations than in meiotic ones.

Indels in non-coding regions may affect chromatin structure or the affinity of a binding site for a regulatory factor, e.g., insertions in the promoter region of the *FMR1* gene can cause Fragile X syndrome (Mills *et al.*, 2006), and insertions in the promoter region of the *SNCA* gene contributes to autosomal dominant Parkinson's disease (Gagliano *et al.*, 2019). Previous work established that somatic non-coding indels in 79 lung adenocarcinoma genomes were exclusively enriched in protein genes (Imielinski, Guo and Meyerson, 2017). Another study further reported the presence of non-coding indels in different forms of lung cancer and demonstrated their clinical use as clonal markers (Nakagomi *et al.*, 2019).

Indel identification is influenced by structural features, like repeats or short interspersed elements. Variant detection methods have been NGS based for which software like SOAP and MAQ have been designed (Bennett *et al.*, 2021). The sequencing methods and bioinformatics tools used for NGS analysis influence the sensitivity and specificity of indel detection. The many NGS platforms have different error types regarding detection of substitutions and indels, comparative analyses therefore show limited agreement between identified indels. False negative rates in many NGS studies have led to about one third of indels in human genomes remaining undiscovered (Bennett *et al.*, 2021). Due to such difficulties, studies suggest that indels are underrepresented with only 55% of insertions in European and Yoruban genomes detected (Bennett *et al.*, 2021).

Many mutations linked to complex traits are found outside the exome; a study found that 88% of the single nucleotide polymorphisms (SNP) in prostate cancer fell in presumed enhancers and less than 20% of the variants were present in the coding region (Hazelett *et al.*, 2014). Therefore, it is important to examine the possible pathogenic influence of indels in non-coding areas of the genome.

2.3.4. NON-CODING DNA REGION

The central dogma of molecular biology comprises the cellular processes of replication, transcription, and translation. The Human Genome Project established that about 98.5% of the human genome does not encode proteins (Boland, 2017). Originally thought to be redundant and under no selective pressure, it is now known that 3D genomic organisation needed for gene regulation is defined by structural elements of non-coding DNA (Perenthaler *et al.*, 2019).

Studies have identified frequently mutated genes and pathways in B-cell lymphoma; however, many malignancies have no detectable driver mutations. The search for driver mutations in B-cell lymphoma has mostly been limited to coding DNA (Cornish *et al.*, 2019). Finding oncogenic mutations in the non-coding genome is problematic because of the huge search space, the challenge in determining the effect of variants that do not encode proteins, the higher mutation rates of non-coding regions due to weaker selective pressure, the analysis of a greater amount of passenger mutations to find non-coding driver variants, and a lack of understanding of the non-coding genome (Rahman and Mansour, 2019). Therefore, it is tricky to analyse the selection pressure of non-coding mutations methodically and unbiasedly, which is why few non-coding mutations have been defined and the functional contribution of non-coding mutations is underappreciated (Huang *et al.*, 2020).

There is however good reason for exploring the non-coding genome for biomarkers, therapeutic targets, and driver mutations. It is filled with cis regulatory DNA elements like enhancers that play key roles in gene expression (Rahman and Mansour, 2019). Mutations are many in the non-coding genome which is over 50 times bigger than the exome (Elliott and Larsson, 2021). New variants of regulatory potential may exist in non-coding regions and may also provide clues to finding other mutations that disrupt normal cellular development. For example, a germline deletion in the microRNA MIR17HG leads to microcephaly, and a mutation in the promoter region of MIR146A is associated with lupus; most single-nucleotide variants (SNV) identified by genome wide association studies that are linked with increased risk of complex disease are in non-coding DNA areas (Ferlaino *et al.*, 2017).

Recent studies have identified somatically acquired recurring mutations in the non-coding genome that activate protooncogene expression (Rahman and Mansour, 2019). A 2013 study found that the promoter of *TERT*, which encodes the reverse transcriptase subunit of telomerase, can somatically acquire mutations leading to its overexpression in human melanoma (Huang *et al.*, 2013). This showed that the non-coding genome can acquire driver mutations. The lesions made de novo consensus binding sites for ETS family transcription factors, allowing overexpression of *TERT* and telomere length and therefore continued cell survival (Hornshøj *et al.*, 2018). Further study showed that these *TERT* promoter mutations occur in other neoplasms too, like ovarian, follicular thyroid, and meningiomas, so similar mutations are selected during the development of other malignancies (Vinagre *et al.*, 2013).

Other comparable non-coding mutational hotspots have been found in additional cancer genomes; *FOXA1* promoter mutations (a driver of hormone-receptor positive breast cancer), recurring mutations in cis-regulatory elements that interact with the *ETV1* promoter in colorectal cancer (CRC), and recurring non-coding mutations in liver cancer (Rahman and Mansour, 2019). Recurrent mutations have also been found in the regulatory regions of the promoters of cancer related genes *WDR74*, *SDHD* and *PLEKHS1* (Gan *et al.*, 2018). These mutations change gene expression levels, transcription factor binding and are linked to poor prognosis. Analyses of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium and The Cancer Genome Atlas (TCGA) identified non-coding driver mutations in several cancer related genes such as *TP53*, *NFKBIZ*, *TOB1*, *BRD4* and *AKR1C* (Rheinbay *et al.*, 1965). Another study identified 160 significant non-coding elements, including the *TERT* promoter, as well as elements associated with known cancer related genes and regulatory genes such as *PAX5*, *TOX3*, *PCF11* and *MAPRE3* (Hornshøj *et al.*, 2018).

With all the mutational procedures at work, it is reasonable to expand the search space for driver mutations in malignancies beyond the exome.

2.4. DLBCL IN SOUTH AFRICA

DLBCL comprises approximately 43% of NHLs in South Africa (Pather and Patel, 2022). This is thought to be largely influenced by the high seroprevalence of HIV infection in the southern African region (Pather and Patel, 2022). South Africa accounts for a third of all new HIV infections in southern Africa and has the biggest HIV epidemic in the world, with 7.7 million people living with the disease (AVERT, 2020). HIV occurrence in the general population was at 20.4% as of 2020 (AVERT, 2020). HIV infection is a recognized risk factor for aggressive B-cell NHLs, which account for up to 30% of tumours in Africa (Pather and Patel, 2022). DLBCL makes up approximately 50% of all HIV-associated lymphomas worldwide (Magangane, Mohamed and Naidoo, 2020).

Studies suggest that HIV can influence B-cells and promote lymphomagenesis directly and indirectly by interacting with B lymphocyte surface molecules (de Carvalho, Leal and Soares, 2021). HIV may affect B-cells by changing how different cell types secrete cytokines. Many cytokines, like IL6, IL10, TNFα, and IFNα, that are involved in B-cell activation, differentiation,

and HIV induced modifications, are overexpressed in HIV-positive individuals (de Carvalho, Leal and Soares, 2021).

A retrospective cohort study was conducted in 2020 on patients diagnosed with de novo DLBCL NOS in Cape Town, South Africa over a 14-year period (Cassim *et al.*, 2020). The study included DLBCL patients with and without HIV comorbidity. An equal distribution of GCB and ABC subtypes was observed in the HIV-infected and HIV-uninfected groups. There is growing research indicating that DLBCL classification into GCB or ABC subtype does not predict the outcome of HIV-linked DLBCL (Wu *et al.*, 2021). The expression of antigens like *FOXP1*, *BCL2*, and *PRDM1* that indicate poor prognosis in non-AIDS DLBCL patients does not predict survival with HIV-linked DLBCL (Chadburn *et al.*, 2009). The 2020 South African study however observed no statistically significant differences in overall survival by DLBCL COO subtype, regardless of HIV status (Cassim *et al.*, 2020). Higher CD4 counts in HIV-infected patients was associated with similar survival outcomes as HIV-uninfected patients, whereas lower CD4 counts in HIV-infected patients predicted significantly poorer outcomes compared to HIV-uninfected patients (Cassim *et al.*, 2020). This was corroborated with other research that also linked worse DLBCL prognosis in PLWH to infectious complications, specifically, immunosuppression with low CD4 count (Re, Cattaneo and Rossi, 2019).

Although the rate of HIV-associated lymphoma has decreased since the introduction of highly active antiretroviral therapy (HAART), the risk of lymphoma is still higher in PLWH (Wu *et al.*, 2021). In South Africa, late establishment of anti-retroviral therapy and late diagnosis of AIDS-defining cancers remain common (Cassim *et al.*, 2020). In the setting of HIV, DLBCL is characterised by early diagnosis, later tumour staging, higher prevalence of B symptoms, and extranodal involvement, supporting accumulating evidence that indicate HIV-linked DLBCL is distinct from other forms of DLBCL (de Carvalho, Leal and Soares, 2021). HIV-linked DLBCL possesses specific molecular properties, gene expression profiles, and chromosomal rearrangements, as well as altered amounts of miRNAs (de Carvalho, Leal and Soares, 2021).

Though the survival estimates for HIV-linked DLBCL patients are consistently similar to those for immunocompetent DLBCL patients, up to 70% of PLWH are excluded from DLBCL research trials (de Carvalho, Leal and Soares, 2021). The inclusion of HIV-DLBCL patients in clinical trial

protocols may help accurately characterise DLBCL in this particularly sensitive population given the innate link between lymphoma and HIV.

2.5. NEXT GENERATION SEQUENCING

Determining the order of nucleotides in a genome or targeted region of DNA/RNA has improved due to the development of next-generation sequencing (NGS). NGS technology's success is a result of its capacity to sequence millions of DNA reads and perform multi-gene analysis with very little nucleic acid (Kanzi *et al.*, 2020). It is excellent for sequencing complicated genomes quickly and effectively, which saves time and money.

DNA NGS calls for DNA fragmentation, library preparation, massive parallel sequencing, bioinformatics analysis, and variant identification and interpretation (Nones and Patch, 2020). NGS technology has improved in reliability, sequencing chemistry, pipeline analyses, and data interpretation (Kanzi *et al.*, 2020). Additionally, it boasts an impressive degree of flexibility and is effectively used in a variety of research fields, including pharmacogenomics, molecular diagnostics of genetic disorders, infectious illnesses, and cancer (Kamps *et al.*, 2017).

Due to the ability to discover a large number of variations linked to complex pathways of oncogenesis and inter- and intra-tumour heterogeneity, NGS usage in cancer research has yielded high-quality mutation detection data, particularly for functional or rarely mutated genes, epigenetics, and transcriptomics (del Vecchio *et al.*, 2017). Molecular profiling of malignancies can offer significant insights on diagnosis, prognosis, and therapeutic response prediction, which can influence clinical decision-making.

NGS has been used in testing circulating tumour DNA and in human leukocyte antigen typing, and microbial sequencing (Thompson *et al.*, 2016; di Resta *et al.*, 2018). Targeted testing can be focused on oncogenes, like *BRCA1* and *BRCA2* genes for breast and ovarian cancer, or it can analyse a wider panel that includes genes associated with other cancers (many cancers have overlapping characteristics) (di Resta *et al.*, 2018). NGS has allowed collective efforts, like the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) project, to list the genomic environment of different cancer genomes (Meldrum, Doyle and Tothill, 2011). TCGA has generated data on 33 types of cancer (Wang *et al.*, 2018); and the Catalogue Of Somatic Mutations In Cancer (COSMIC) project has collected about six

million coding mutations and has explored other genetic mechanisms that can aid cancer like gene fusions, drug resistance mutations and non-coding mutations (Tate *et al.*, 2019). Available NGS platforms include Illumina, which uses 'sequencing by synthesis' (Illumina Custom Amplicon panels, Illumina Nextera for large amplicons, etc.); Ion Torrent, which uses fusion primers for small amplicons; Roche and Helios (Moorthie, Mattocks and Wright, 2011).

With greater than 99% accuracy, Illumina's equipment is used in a variety of fields, including transcriptomics, epigenomics, and genomics (Illumina Sequencing and array-based solutions for genetic research, 2022). The Illumina sequencing technique is based on clonal arrays paired with clonal sequencing by synthesis employing cyclic reversible termination. The process is designed to sequence both the forward and reverse strands; as a result, data from both strands are taken into account in the final analysis. During sequencing by synthesis, base calls are made for each cluster and stored for every cycle of sequencing in individual base call (BCL) files. When sequencing completes, the BCL files must be converted into sequence data in FASTQ format, the default file format for sequence reads generated from NGS technologies (FASTQ files explained, 2022). A FASTQ file is a text file that represents biological sequences and their corresponding quality scores.

NGS has been applied to whole genome sequencing (WGS), a technique designed for entire genome sequencing (del Vecchio *et al.*, 2017). It offers the most comprehensive landscape of genetic data and potential biological effects. Despite its potential, which enables the detection of undiscovered mutations at the level of both coding and non-coding areas largely engaged in the control of gene expression, it exhibits unavoidable challenges owing to the large volume of generated data and their relevance, e.g., variants with unknown significance (del Vecchio *et al.*, 2017). Whole-exome sequencing (WES), targeted sequencing, and transcriptome sequencing (RNA-Seq) have been developed to get around these problems. Targeted sequencing concentrates on particular genome regions or significant genes whose pathogenic role in disease has previously been documented (del Vecchio *et al.*, 2017). WES restricts the length of the nucleic acid under analysis to coding areas, providing details about exons in the process (del Vecchio *et al.*, 2017). RNA sequencing (RNA-Seq) offers insight into a cell's transcriptome, and the data it produces makes it easier to find new transcripts, identify alternatively spliced genes, and find allele-specific expression patterns (Conesa *et al.*, 2016).

To create profiles that may be utilized for diagnostic or prognostic reasons, it may be helpful to analyse components that play a part in these mechanisms, such as DNA-binding proteins, methylated regions, or non-coding RNAs. NGS is being increasingly applied to the analysis of epigenetic changes, most notably in the study of cancer. This is crucial for detecting alterations to the genome's regulatory components, such as transcription factor binding sites, enhancers, and insulators that control gene expression. High-throughput sequencing like ChIP-seq has enabled genome wide experimental determination of in vivo transcription factor binding regions.

2.5.1. ChIP-SEQ

Many important biological processes, like cell differentiation, gene transcription and DNA replication, etc., depend on interactions between cellular proteins and DNA. ChIP-seq comprises chromatin immunoprecipitation followed by high-throughput sequencing. Chromatin is a compound of DNA and proteins in the nucleus (Nakato and Sakata, 2021). Histones are major proteins of the chromatin; histone H1, 2A, 2B, 3 and 4, etc., (Furey, 2012). Histones and other regulatory proteins bind to DNA and preserve its 3D structure.

The chromatin packages DNA into a smaller volume to fit into the cell; it reinforces the DNA to allow mitosis and prevent damage. ChIP-seq is the standard assay for genome-wide identification of transcription factor binding sites and other DNA-binding proteins important in the understanding of cellular processes and disease (Mundade *et al.*, 2014). Applications include studies on transcriptional regulation and histone modifications (Ma and Wong, 2011).

The ChIP method was established by Gilmour and Lis while studying the involvement of RNA polymerase II with transcribed and poised genes in *Escherichia coli* and Drosophila (Gilmour and Lis, 1984). Ultraviolet (UV) irradiation was used to covalently cross-link proteins in contact with neighbouring DNA in living cells. The formaldehyde cross-link approach by Solomon and Varshavsky later replaced the UV cross-link (Solomon, Larsen and Varshavsky, 1988).

ChIP-seq begins with cross-linking DNA and DNA-bound proteins. Chromatin is then isolated from nuclei and exposed to sonication (Raha, Hong and Snyder, 2010). A specific antibody of a transcription factor or DNA-binding protein is used to immunoprecipitate specific DNA-transcription factor complexes. Purification of ChIP DNA is followed by ligation of sequencing

adapters; typically producing 30 to 35 nucleotide sequence reads. The DNA fragment sequences are mapped to a reference genome to identify binding sites. Thus, ChIP-seq ultimately reveals the binding sites for DNA-associated proteins which are then stored in FASTQ formatted files. During ChIP–seq analysis, the ChIP-seq data in FASTQ format is used to perform read mapping which enables the detection of mutations like indels by gapped alignments. The ChIP-seq reads in FASTQ format is also used to perform peak calling, which identifies regions within the genome that are enriched with reads indicating increased transcription.

Large-scale ChIP-seq data sets have been made for different transcription factors and histone modifications with potential to predict gene expression that can be used to test hypotheses about the mechanisms of gene regulation (Jiang and Mortazavi, 2018). While technology like WGS can determine the entire DNA sequence of an organism's genome, including mutations, it cannot detect which genes are functional, and with the cost of sequencing decreasing, ChIP-seq is a vital tool for studying gene regulation and epigenetics.

Due to the high sensitivity and specificity ChIP-seq has in plotting protein binding sites, it has enabled motif and target discovery and identification. Enhancers, key regulatory elements that control gene expression, are marked by specific chromatin modifications including H3K4me1 and H3K27ac. H3K27ac is particularly interesting because it differentiates active enhancers from poised ones. ChIP-seq allows the detection of enhancers through its ability to trace transcription factors which activate enhancers.

Bioinformatics is a rate limiting factor of ChIP-seq in terms of storage challenges, analysis, and data interpretation (Kulski, 2016). Bioinformatics tools and databases are required for ChIP-seq data analysis from the original raw sequencing data to functional biology.

2.6. BIOINFORMATICS STUDIES ON CANCER USING THE H3K27AC MARK

Bioinformatics is a multidisciplinary field that includes computer science, mathematics, statistics, molecular biology, and genetics (Baichoo *et al.*, 2018). Data intensive, large-scale biological problems are addressed from a computational point of view. Bioinformatics combines data generated by high throughput sequencing such as ChIP-seq to form a comprehensive picture of normal cellular activities so that researchers can investigate how these activities are changed by disease (Spjuth *et al.*, 2015). Common points of interest involve modelling biological processes at the molecular level and making inferences from the collected data (Can, 2014). The field's goal is to foster the discovery of novel biological insights and to provide a broad viewpoint from which unifying biological principles may be extracted.

ChIP-seq computational studies are becoming more thorough and complex as more experimental groups use it to elucidate transcriptional and epigenetic regulatory processes. A variety of computational and statistical methods have been designed for ChIP-seq analysis (Nakato and Sakata, 2021; Eder and Grebien, 2022).

MISREGULATION OF ONCOGENES VIA ENHANCERS FORMED BY SMALL GENOMIC INSERTIONS

A 2017 study found non-coding driver mutations by looking at sequencing reads from H3K27ac ChIP-seq (Abraham *et al.*, 2017). They developed a computational pipeline to get insertions that were present in tumour cells but not in the NCBI human reference genome to find enhancer-associated variation in cancer cells.

To detect reads without insertions as well as enhancers, the H3K27ac reads were mapped to the hg19 reference genome using bowtie. The H3K27ac reads were also used to find active enhancers using MACS with input DNA controls. H3K27ac reads that were not aligned by bowtie were mapped to the hg19 reference genome using bowtie2, which allows gaps (insertions and deletions) relative to the target genome. The reads with insertions were verified with Blat; reads with a CIGAR string containing 'I' were used as input. The BLAT output was parsed so that each accepted read hit included the entire read sequence rather than just aligning a portion of the read, it also contained only one insertion that was shorter than the read with no BLAT-called mismatches and ensured that the best hit with the highest score was kept. The CIGAR string from the bowtie hit was utilized to pinpoint the location and

nature of the insertion in BLAT hits, which were additionally filtered to have no more than a 20-bp (base pair) insertion. For a read to be kept, bowtie2 and BLAT hits had to be within 100 bp. Overlaps with the enhancers served as the basis for determining inserts in enhancers.

The study identified candidate enhancer-associated insertions ranging in size from 1 to 31 bp in the tumour samples. Small insertions were often found in enhancer DNA sequences close to known oncogenes. Further study of one insertion, somatically acquired in primary leukaemia tumour genomes, revealed that it nucleated formation of an active enhancer that drives expression of the *LMO2* oncogene. The information on enhancer-associated insertions obtained in this study contributed to the foundation for further studies to define the oncogenic impact of this type of variant.

ACTIVATION OF GENE ENHANCERS VIA MISMATCH REPAIR SIGNATURE MUTATIONS ACROSS THE EPIGENOMES OF HUMAN COLORECTAL CANCER

In 2019, a study used gains in tumour-specific enhancer activity with allele-biased mutation detection from H3K27ac ChIP-seq data to find enhancer-activating mutations in colorectal cancer (Hung *et al.*, 2019). ChIP-seq data processing, alignment, peak-calling, and identification of differentially enriched-peaks relative to normal colonic crypts was performed before the detection of enhancer mutations.

H3K27ac ChIP-seq reads were aligned to the human genome (hg19) with bowtie2 and then realigned around regions with evidence of indels. Peaks were called using MACS. For each indel, the sequence 50 bp upstream and downstream was aligned to the human genome with Blat. The reference allele was replaced with the indel allele to simulate the alignment of indel-supporting ChIP-seq reads. Indels whose second-highest alignment score was >50 indicated potential alignment error and was discarded. Indels with imbalanced read distributions favouring the indel allele were prioritized, because it suggested the enhancer signal and indel occurred on the same allele.

The analysis of CRC specimens showed that microsatellite instable (MSI) samples had a high indel rate in active enhancers which showed evidence of positive selection, upregulation of target gene expression with a recurrent subset. The indels increased affinity for FOX

transcription factors. The results suggested that mismatch-repair signature mutations activate enhancers in CRC tumour epigenomes to provide a selective advantage.

IDENTIFICATION OF TUMOUR SPECIFIC GENE EXPRESSION VIA EPIGENOME MAPPING IN PRIMARY RECTAL CANCER

ChIP-seq analyses are usually done on cell lines so data from primary tumours is limited. A study investigated the use of ChIP-seq to find tumour-specific epigenetic variations in primary rectal cancer (Flebbe *et al.*, 2019). Focus was put on H3K27ac due to its association with active gene transcription.

Tissue samples from primary rectal cancer and matched healthy mucosa was obtained. ChIP-seq for H3K27ac was performed before statistical analysis of the data. The data was mapped to the human reference genome using bowtie2. Peak calling was done using MACS2. Visualization of the ChIP-seq data was done with Integrative Genomics Viewer (IGV).

ChIP-seq for H3K27ac in primary rectal cancer and matched mucosa revealed differential binding in 44 regions. Genes with increased H3K27ac were identified; *EPHX4*, *KRT23*, *FOXQ1*, and *RIPK2*. They were also upregulated in an independent primary rectal cancer dataset. The increased expression of the four proteins was confirmed by immunohistochemistry. This study showed the viability of ChIP-seq-based epigenome mapping of primary rectal cancer and validates the value of H3K27ac to predict gene expression differences.

DBINDEL: A DATABASE OF ENHANCER-ASSSOCIATED INDEL VARIANTS BY H3K27ac ChIP-seq ANALYSIS

In 2020, dbInDel, a database cataloguing enhancer-associated indel variants for human and murine samples, was introduced (Huang *et al.*, 2020). It includes transcription factor binding motif analysis which enables the identification of upstream transcriptional regulators. The database contains enhancer-associated indels taken from H3K27ac ChIP-seq data. Survival analysis in tumour and normal samples across human cancer types and mRNA expression profiles was integrated to allow analysis of target transcripts of enhancers with indels. The method identifies the possible recruitment of transcription factors due to enhancer-associated indels, supporting the examination of the functional contributions of these non-coding variants.

To find enhancers, H3K27ac enriched regions were identified with or without a corresponding control sample using MACS. To find small indels in presumed enhancers across a range of cancers, the method described in Abraham *et al.* (2017) was computationally rebuilt and H3K27ac ChIP-seq datasets from over 250 samples of 26 types of cancer was investigated. The ChIP-seq reads was mapped to the reference genome using bowtie. To align gapped reads to the reference genome, bowtie2 was used. The SAM output files were examined for CIGAR string containing 'M' and one 'I'. Those reads were selected to align to the reference genome using pBlat (multi-threads support Blat). The insertion was confirmed if the whole read aligned with only one insertion in the pBlat result. IntersectBed was used to remove insertions in exons of hg19 refseq mRNAs from the University of California Santa Cruz (UCSC) Browser. Enhancer associated non-coding indels were therefore captured after overlapping indels with enhancers.

The database contains 640,432 insertions and 157,554 deletions in 593,655 presumed enhancers detected across 275 samples. Among the indels, 274,995 are unique insertions and 71,603 are unique deletions. The results indicated that many cancer drivers have unique enhancer-associated indels in the respective cancer types, e.g., *AR*, a prostate cancer oncogene, was found to have enhancer-associated indels only in prostate cancer samples. The data suggests that some of the enhancer-associated indels are under selective pressure and give advantage to certain cells because of the functions of the target genes they regulate.

2.7. BIOINFORMATICS TOOLS USED TO IDENTIFY INSERTIONS AND ENHANCERS THROUGH H3K27AC CHIP-SEQ ANALYSIS

2.7.1. ALIGNING READS TO A REFERENCE GENOME

The first step of ChIP-seq analysis involves mapping the ChIP-seq reads in FASTQ format to a reference genome to facilitate the detection of insertions and to perform peak calling which then aids the identification of putative enhancers. Aligners usually use a genome index to narrow the list of possible alignment locations. Such aligners work by looking for ways to change the read string into one that occurs in the reference. The search space is large, and many portions can be skipped without loss of sensitivity. Bowtie allows quick and memory-efficient large-scale alignment of short sequencing reads to a reference genome (Langmead, 2010). It has tools to build Burrows-Wheeler reference genome indexes with which it aligns

reads to a reference genome. More than one processor can be used at the same time to increase the speed of alignment. Bowtie can output alignments in SAM format which enables it to work with other SAM supporting tools. Bowtie is run via the command line under Solaris, Windows, Linux, and Mac OS X. Bowtie is fast when working with sets of short reads where many are of high quality with at least one valid alignment, and the number of reported alignments per read is small.

Index-aided alignment is inefficient when alignments have gaps from sequencing errors or indels. Ungapped aligners like bowtie cannot align reads spanning gaps and so these events go overlooked. Bowtie2 allows gapped alignment with a two-stage algorithm: an ungapped seed-finding stage and a gapped extension stage that benefits from single-instruction multiple-data (SIMD) parallel processing (Langmead and Salzberg, 2012). Bowtie2 takes 'seed' substrings from the read and its reverse complement; these substrings are aligned to the reference in an ungapped manner; seed alignments are prioritized and their positions in the reference genome are calculated from the index; seeds are extended into full alignments by performing SIMD-accelerated programming. The combination results in effective speed, sensitivity, and accuracy across a range of read lengths and sequencing technologies. In this way, ChIP-seq reads that contain gaps potentially caused by indels are identified and typically stored in SAM formatted files.

2.7.2. ALIGNMENT FORMATTING

Modern sequencing technologies have led to the advent of alignment tools for read mapping against reference sequences. However, the alignments produced by these tools differ in format which complicates further processing. A defined interface between alignment and further analysis is necessary to create a format that can support all types of sequences and aligners. Sequence Alignment/Map (SAM) format is an alignment format for storing read alignments to reference sequences (Li *et al.*, 2009). SAM supports short and long reads from different sequencing platforms including ChIP-seq. The format has one header section with lines starting at '@', and one alignment section. The lines are TAB delimited. Every alignment line has 11 compulsory fields with further optional fields. The sixth compulsory field contains the CIGAR string which represents spliced alignments in the SAM/BAM format.

The CIGAR string indicates which bases match with the reference, are deleted from the reference, and are insertions that are not in the reference.

Binary Alignment/Map (BAM), a binary representation of SAM compressed by the BGZF library, improves performance through quick retrieval of alignments in specific regions (Li *et al.*, 2009). An unsorted SAM/BAM file can be sorted by coordinate to streamline processing. A BAM file can be indexed by combining the UCSC binning plan and linear indexing to quickly fetch alignments that overlap a certain region. Sorting and indexing can implement genomic processing without loading the whole file into memory.

SAMtools is a software package for manipulating alignments in SAM or BAM format (Li *et al.*, 2009). SAMtools can perform jobs like converting alignment formats, sorting, viewing, and combining alignments, indexing, and variant calling. The SAM and BAM formats, together with SAMtools, provide for a modular approach to process ChIP-seq data by separating alignment from downstream analytics.

2.7.3. IDENTIFYING INSERTIONS WITHIN SEQUENCING READS

Reliable read filtering is crucial when processing ChIP-seq data in epigenetic studies. For jobs such as identifying insertion events within ChIP-seq data, some custom data processing which cannot be implemented in shell pipelines can be implemented in AWK, Bash or Python scripts. AWK is a scripting language and text manipulation toolkit for the command line (McKay, 2020). It allows the user to write small but effective statements that are programs (Spjuth *et al.*, 2015). AWK requires no compiling and allows the user to use numeric and string functions, logical operators, and variables (Spjuth *et al.*, 2015). AWK is utilized for pattern scanning and processing, manipulating, and transforming data and generating formatted reports (Aho, Kernighan and Weinberger, 1978).

Sequence aligners typically output files in SAM format. Aligned ChIP-seq sequences may contain extra bases not found in the reference or may lack bases found in the reference. The CIGAR field of SAM formatted files contains a sequence of base lengths and associated operations viz. ‘M’ for match, ‘I’ for insertion and ‘D’ for deletion (Kim *et al.*, 2017). AWK programs scan a file line by line splitting them into fields and comparing them to a provided pattern before performing an action on lines that match (Goyal and Negi, 2021).

In this way, a SAM formatted file of ChIP-seq reads that were aligned to a reference genome can be analysed for insertions using the 'I' field in the CIGAR string. Positive matches can be stored in a separate output file. ChIP-seq sequences containing insertions to the reference genome are thus filtered from sequences without insertions to the reference genome.

2.7.4. VERIFICATION OF INSERTIONS WITHIN SEQUENCING READS

Studying genomes require quick mRNA/DNA and cross-species protein alignments. To confirm ChIP-seq reads with insertions are reliably alignable, reads with a CIGAR string containing 'I' are used as input for the program Blat or pBlat. Blat (BLAST-like alignment tool) is a pairwise sequence alignment algorithm that was developed by Jim Kent to assist in the assembly and annotation of the human genome (Kent, 2002). It is used for the analysis and comparison of biological sequences to infer homology to identify the biological function of genomic sequences. Previous alignment tools could not perform such operations in a way that would allow a regular update of the human genome assembly. Blat produces alignments at the DNA level between two sequences that are of 95% or greater identity, but which may include large inserts. It searches for short matches and extends these into high-scoring pairs (Kent, 2002). Blat has an index of nonoverlapping K-mers in the genome which fits inside the RAM of inexpensive computers and must only be computed once for each genome assembly (Bhagwat, Young and Robison, 2012). Blat uses the index to find areas in the genome likely to be homologous to the query sequence. It performs an alignment between homologous regions and stitches these aligned regions (exons) together into larger alignments (genes). Blat then goes back to small internal exons possibly missed before and modifies large gap boundaries that have feasible splice sites.

Blat is typically used for gapped mapping and long sequence alignment which can't be properly done by other fast sequence mappers made for short reads (Wang and Kong, 2019). However, the number of sequences generated by high throughput sequencing projects is increasing and blat is not adept at large scale sequencing research and iterative analysis (Wang and Kong, 2019). It takes days for blat to map whole genome or transcriptome sequences to a reference genome. This is because blat was designed to be single threaded and therefore does not take advantage of modern multicore processors.

The parallel blat (pBlat) algorithm is a multithreaded program with cluster computing support that facilitates high-throughput mapping of large scale genomic and transcript sequences to reference genomes through the C programming language to implement multiple thread support and data-level parallelism (Wang and Kong, 2019). FASTA format input query files containing ChIP-seq reads with insertions are partitioned based on the number of threads specified. Each part has the same amount of query sequences. Each thread performs the blat algorithm on one part of the input sequence. The threads use the same amount of memory as blat because they share the same memory copy of the whole reference genome and the index. The number of threads used reduces the run time and the results of pBlat are identical to that of blat. The outputs of each thread are written to a temporary file and once all threads have completed their workload, the temporary output files are combined into one final output file. The order of output records therefore matches the order of query sequences in the input file no matter how many threads are used. The global variables in the original blat program are localized to ensure all the variables and subroutines are thread safe.

2.7.5. IDENTIFYING GENOMIC REGIONS ENRICHED WITH ALIGNED READS (PEAK CALLING)

Peak calling finds areas in the genome that are enriched with aligned reads due to a ChIP-seq experiment. The Model-based Analysis of ChIP-seq data (MACS) analyses short read sequencing data. It can identify transcription factor binding sites and histone modification enriched regions (Feng, Liu and Zhang, 2011). MACS can be used for the ChIP sample alone or in combination with a control sample to boost peak call specificity. MACS2 is a later version of MACS and performs several functions including duplicate filtering, peak model construction, peak identification, and multiple testing correction. It can also join close peaks together to create broad peaks (Gaspar, 2018). The application is user-friendly and gives detailed information about each peak, including genome coordinates, p-value, false discovery rate, fold enrichment, and peak centre (Gaspar, 2018).

ChIP-seq tags are used to indicate the ends of fragments in a ChIP-DNA library, and they are usually pushed towards the 3' direction to better show the protein-DNA binding site. The experimenter is unaware of the magnitude of the shift. Because both ends of ChIP-DNA fragments can be sequenced, the tag density of a real binding site should exhibit paired peaks or a bimodal enrichment pattern, with Watson strand tags enriched upstream and Crick

strand tags enriched downstream (Feng, Liu and Zhang, 2011). This bimodal pattern is used by MACS2 to anticipate the moving size and locate the exact binding points.

To develop a model, MACS2 looks at the whole dataset for substantial enriched regions to locate paired peaks (Chipster, 2021a). MACS2 slides two bandwidth windows over the genome to detect locations with tags more than mfold enriched compared to a random tag genome distribution (Chipster, 2021b). The ratio between the ChIP-seq tag count and local is reported as the fold enrichment. MACS2 selects 1,000 of these high-quality peaks at random, separates their positive and negative strand tags, and aligns them by their midpoints. The estimated fragment length is defined as 'd', the distance between the modes of the two peaks. MACS2 moves all tags by d/2 to the 3' ends, where the most likely protein-DNA interaction sites are found. By estimating the distance d and moving tags by d/2, MACS2 enhances the spatial resolution of the anticipated binding sites.

A dynamic Poisson distribution is used to show local biases in the genome which improves the prediction's validity and specificity (Feng, Liu and Zhang, 2011). This method can also be used to catch regional biases and estimate fold-enrichment in other applications such as copy number variation and digital gene expression. Because each peak is analysed separately, a multiple testing problem occurs when there are thousands of significant peaks discovered in a sample. The Benjamini-Hochberg adjustment is used to fix p-values for multiple comparisons in MACS2 (Gaspar, 2018).

Peak regions indicative of enhancer activity identified from ChIP-seq data are then filtered and any overlaps with blacklisted regions, such as genomic locations of insertion events in sequence reads, are assessed in order to infer biological cause and effect.

2.7.6. BEDTOOLS

Genomic research requires testing for connections or overlapping between sets of genomic features (aligned reads, polymorphisms, annotations, etc.). Such comparisons characterise experimental output, deduce coincidence and evaluate biological impact (Quinlan, 2014).

BEDTools is an open-source software package written in C++ and consisting of tools focused on operations for the comparison and exploration of genomic datasets through basic genome arithmetic tasks (Quinlan and Hall, 2010). BEDTools includes the UCSC Genome Browser's

genome-binning algorithm. It utilizes a hierarchical indexing plan to assign genomic features to 'bins' along the chromosome. This speeds up the search for overlapping features, since comparison of features is done between two sets that share the same bins. The most common question asked of two sets of genomic features involves feature intersection; bedtools intersect screens for overlaps between two sets of genomic features and enables the user to fine tune how the intersections are reported (Quinlan, 2014). It can be used to screen for insertion events within ChIP-seq data that do not occur within the exome, and to identify overlapping features between a set of enriched regions (peaks) and sequencing reads that contain insertion mutations, thereby connecting mutation events to sites in the genome with elevated transcription levels.

BEDTools outputs files in standard BED (Browser Extensible Data) format. BED format is typically used to store genomic data (Quinlan and Hall, 2010). Operations using genomic coordinates, nearest-element connections between feature sets, and quantitative computations across linked genomic segments are necessary for BED analyses.

2.7.7. BEDOPS

In order to infer biological context, it is customary to link called peaks from ChIP-seq data to adjacent genes, either upstream or downstream, because many cis-regulatory components like enhancers are close to the transcription start site (TSS) of their targets.

BEDOPS is a software tool for genomic analytical jobs such as set statistical operations and calculations, archiving, and conversions. Some interesting programs offered by BEDOPS include union, subset, and difference; closest features which links the nearest features (e.g., TSS or genes) between two sorted inputs based on genomic distance; and bedmap which maps source data onto genomically related target regions and generates summaries per region (Neph *et al.*, 2012). These core utilities may be combined to make pipelines while keeping efficiency and scalability with standard sorted input and output stream support. The memory overhead of the main BEDOPS utilities is unaffected by the size of the data input so BEDOPS pipelines can function with dense datasets on a variety of hardware. It has better flexibility, scalability, and execution time than most other tools (Neph *et al.*, 2012). BEDOPS compresses BED into a format that decreases access times to most data. BEDOPS only keeps the data needed to compute the next line of output, so memory use is reduced. Other tools

need more space because they load the entire file to memory and generate an index (Quinlan, 2014); this leads to longer run times too which can cause errors with big inputs.

Once gene annotations are assigned to peak calls for ChIP-seq data, biological ontologies like Gene Ontology (GO), KEGG, and Reactome can be used for functional enrichment analysis to uncover common biological themes among these genes. Functional enrichment methods that carry out over-representation analysis by querying databases holding details about gene function and relationships can be used to interpret the gene lists gained from annotation.

2.7.8. FUNCTIONAL ENRICHMENT ANALYSIS

The comprehensive quantification of DNA, RNA and proteins in biological samples has generated huge amounts of data that must be interpreted to elucidate biological functions and disease mechanisms (Reimand *et al.*, 2019). ChIP-seq data after it has undergone analysis often takes the form of extensive gene lists without structure or context, and which need an impractical amount of manual research to analyse (Tipney and Hunter, 2010). Single genes also do not accurately represent the complex operation of biological systems. Researchers can get mechanistic insight into gene lists produced by ChIP-seq studies using pathway enrichment analysis (Reimand *et al.*, 2019). This switches analysis from individual genes to biological processes by concentrating on sets of genes that share biologically significant attributes (Creixell *et al.*, 2015).

Pathway enrichment analysis identifies biological pathways that are more prevalent than would be anticipated by chance in an experimental gene list. It performs a systematic mapping of biological annotations, like GO terms, to genes and proteins, and then compares the distribution of these terms within a target gene set to the background distribution of these terms (Tipney and Hunter, 2010). Terms that are statistically overrepresented or underrepresented are thus identified and biological behaviour can be extrapolated (Tipney and Hunter, 2010). Three general processes comprise enrichment analysis: specifying a gene list from ChIP-seq studies, identifying statistically enriched pathways, and visualizing and translating the findings (Reimand *et al.*, 2019).

Functional enrichment tools like the Database for Annotation, Visualisation, and Integrated Discovery (DAVID) arrange functionally related genes and terms into a summarised number of biological modules for effective interpretation of gene lists in a network setting (Huang *et al.*, 2007). DAVID combines annotation terms from a variety of sources, e.g., InterPro for proteins, OMIM for disease associations, and KEGG and BioCarta for pathways (Reimand *et al.*, 2019). It also takes relationships between annotation terms into account and has unique visualisation methods to enable assessment of results.

Pathway analysis has been applied to cancer data sets to uncover regulators of cancer associated genetic pathways, undetected tumour subgroups characterised by repeated patterns of pathway variations, and to suggest cancer mechanisms and biomarkers (Creixell *et al.*, 2015). Pathway analysis has several advantages over analysing single genes (Creixell *et al.*, 2015). Results are simpler to understand since genetic variations are linked to well-known concepts like apoptosis. Possible causative processes can be found, e.g., by anticipating a specific transcription factor that accounts for the differential expression between tumour samples and controls. Since pathway information enables interpretation in a shared feature space, results from linked datasets become comparable. It also enhances statistical and interpretive power by enabling the assimilation of different omics inputs into a cohesive perspective of cancer biology.

2.8. CONTAINER SYSTEMS

Several bioinformatics tools and programs should typically be installed and set up before beginning a bioinformatics investigation. This requires a lot of labour, time, and the installation of software and their dependencies. The possibility that a full environment might be packaged and executed anywhere was made possible by the development of virtual machines.

A software container is used to enclose a software component and the associated dependencies (Matelsky *et al.*, 2018). It has code fragments that may be independently deployed and utilized to create and run applications. Containers share a machine's operating system (OS) kernel but do not require the overhead of associating an OS within each application (IBM, 2019). The abstraction from the host OS makes containerized applications portable and able to run uniformly and consistently across any platform or cloud.

Existing operating systems serve as the foundation for containers (Emily Mell, 2021). Because they don't include the whole guest OS, containers vary from virtual machines in that they are constructed using optimized system libraries and make use of the host OS's memory management and process controls (Matelsky *et al.*, 2018). Typically, containers are built around a single piece of software, and are made executable by creating images from them. Images are collections of files that may include an OS, software, data, and sometimes additional files for associated applications (Kurtzer, Sochat and Bauer, 2017). Two common container technologies are Docker and Singularity.

Docker packages and runs an application in a loosely isolated environment i.e., a container. Docker separates applications from infrastructure, but it was designed for virtual servers, so it tries to isolate the container (di Tommaso *et al.*, 2015). Containers have an isolated file system, so the script won't have access to the host filesystem. To run the Docker image the same way as the script, the local directory must be mounted as a volume, and the working directory changed to be the mounted volume (Mitra-Behura, Fiolka and Daetwyler, 2022). Docker also isolates user identities. The container uses a different user identity to the process that launches it. On Linux OS the output file becomes owned as root so the container must be run as a user and group identity that matches the user.

Singularity containers are frequently used on high performance computing (HPC) clusters because they do not require root access (Mitra-Behura, Fiolka and Daetwyler, 2022). It can create images from Docker definitions accessible from Docker Hub and is an excellent tool for condensing several difficult image processing operations into one. A Singularity container packages an application and all its dependencies into a single Singularity Image File (SIF). It allows you to install pre-built container images and it ensures the same software can be shared and used across Linux systems (Mitra-Behura, Fiolka and Daetwyler, 2022). A Singularity definition file contains instructions on how to build a custom container including details about the base OS to build or the base container to start from, software to install, environment variables to set at runtime, files to add from the host system, and container metadata (Mitra-Behura, Fiolka and Daetwyler, 2022).

Singularity prioritizes integration, reproducibility and security via cryptographic signatures, a fixed container image format, and in-memory decryption (Kurtzer, Sochat and Bauer, 2017). GPUs, high-speed networks, parallel filesystems, and computing mobility are used on a cluster or server. The single file SIF container format has an efficient security strategy and is simple to distribute and transfer. Singularity has direct access to the kernel so there is not a big performance penalty when using a container over installed applications. The same user rights are maintained inside a container as on the outside, and more authority is not automatically granted on the host system. Singularity does not ask for extra administrative rights for a user to run and interact with containers on a platform where it is being used.

2.9. PIPELINES AND WORKFLOW FRAMEWORKS

Genetic information gained from high throughput technologies like ChIP-seq is used to develop complex biological data models, having a mechanism to map and manage analysis step-by-step has therefore become vital. Bioinformatics analyses involve steering files through transformations, called a pipeline, that perform tasks, support reproducibility, and provide measures to reduce error (Leipzig, 2017). The components of a pipeline are linked together to form a path. Using parallel buffers, the output of one operation serves as the direct input for the next. As a result, information administration is made more efficient and human processing error is reduced.

A bioinformatics workflow usually involves collecting statistics from biological data, building a computational model, solving a computational modelling problem, and evaluating a computational algorithm (Ahmed *et al.*, 2021). Frameworks are increasingly used in bioinformatics investigations to sequence metadata.

2.9.1. NEXTFLOW

Studies typically yield millions of raw reads produced by high throughput sequencing that require computationally intensive processing tools. These tools must be easy to use and combine into stable workflows. This has led to the development of sequencing pipelines like RseqFlow and Galaxy (Federico *et al.*, 2019). However, some difficulties with these pipelines involve the limited number of computational tools and modification abilities when they are

used on existing computational resources. Other frameworks may be more flexible but usually each tool must be separately installed, which is cumbersome and hinders reproducibility.

Nextflow is a workflow framework and a programming Domain Scripting Language (DSL) for writing computational pipelines (di Tommaso *et al.*, 2017). The DSL2 syntax is an updated version of the original DSL with several enhancements, including better data flow manipulation and the introduction of module libraries which separate components to allow for flexibility and reuse (di Tommaso and Floden, 2021). Nextflow expands the Linux platform's command-line and scripting facilities for data manipulation. It uses the dataflow programming approach to create complex program interactions and a high-level parallel computing environment. It supports Docker and Singularity containers. This, along with the GitHub code sharing platform, provides for version control, the creation of self-contained pipelines, and the easy replication of previous setups.

A Nextflow process is the fundamental component used to run a user script. The process is defined by the script/command to be executed, the input to the script and the output of the script. Processes are executed independently; they are connected via their outputs and inputs to other processes and run as soon as they receive input. Data is passed between process tasks via channels which manipulate the flow of data from one process to the next. There are two types of channels (di Tommaso and Floden, 2021). A queue channel is a non-blocking, unidirectional first-in-first-out queue. A value channel is restricted to a single value and can be read limitlessly without its content is consumed. Processes can be written in any scripting language executable by Linux (Bash, Python, Perl, etc.) (di Tommaso *et al.*, 2017). Workflows are made up of chained Nextflow processes. The applications are innately parallel and can be scaled without adapting to a certain platform structure.

Nextflow operators are methods that connect channels or transform values emitted by a channel by applying customizable rules (di Tommaso *et al.*, 2017). For example, the 'join' operator generates a channel that connects two channels emitting items that have a matching key; the 'collect' operator gathers a channel's entire output and returns the resultant object as a single emission. Almost every operator produces one or more new channels, allowing operators to be chained to fit the user's needs.

The executor controls how the script is run on the target system (di Tommaso *et al.*, 2017). It allows the pipeline logic to be separate from the processing platform, i.e., the pipeline script can be written once and run on a computer, cluster or cloud depending on the executor defined in the Nextflow configuration file. Unless otherwise specified, processes are run locally, which is useful for pipeline development and testing before switching to a cluster when it must be run on production data. Executors compatible with Nextflow include SGE, Moab, LSF, SLURM, PBS/Torque, PBS Pro, NQSII, Igntie, Kubernetes, AWS Batch, Google Life Science and OAR.

In recent years, researchers have used Nextflow as a foundation with other tools to create advanced frameworks. A 2019 study presented the Pipeliner framework which used Nextflow and the Anaconda package manager to make standard computational workflows (Federico *et al.*, 2019). The study created an RNA-seq pipeline to process raw DLBCL sequencing reads from a cohort supplied by TCGA. Supplementary files were generated that could be used as a template for applying Pipeliner to publicly available datasets. A 2020 study presented GeneTEFlow, a workflow for the analysis of transposable element expression from RNA-Seq data (Liu, Bienkowska and Zhong, 2020). GeneTEFlow used Nextflow and Docker which allowed reproduceable analyses on different computing platforms without requiring separate tool installation and manual version tracking.

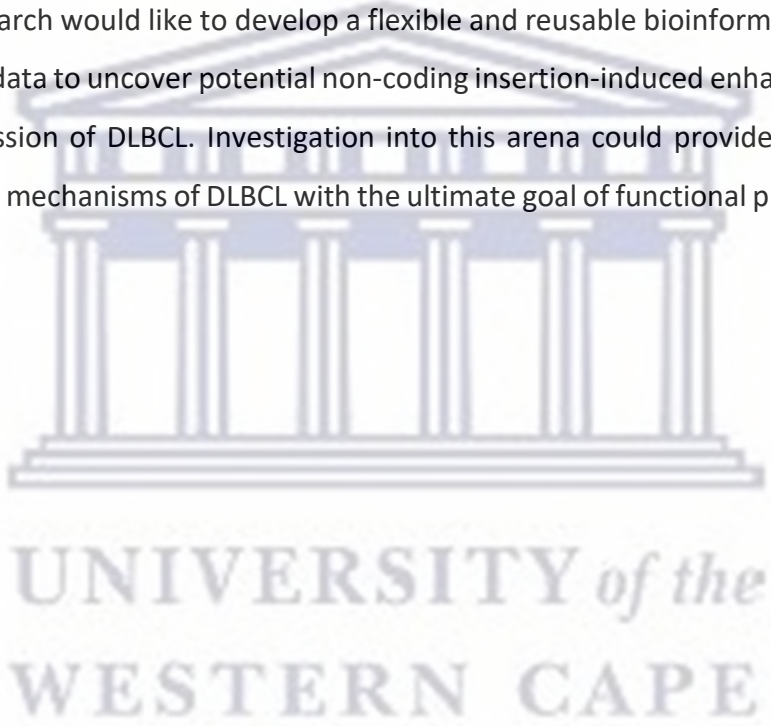
2.10. CONCLUSION

DLBCL is a hereditarily heterogenous cancer originating in the germinal centre characterized by a diffuse production of matured and enlarged B-cells. It is hypothesized to be promoted by HIV and makes up almost half all HIV-associated lymphomas in South Africa. DLBCL is divided into two phenotypic categories based on cell of origin; ABC DLBCL, defined by a post-germinal centre B-cell, and GCB DLBCL, defined by a B-cell that overexpresses genes associated with the germinal centre reaction. Each category has distinguishing genetic drivers and signalling pathways that are targetable for therapy.

Majority of the human genome does not encode proteins; the structural elements of non-coding DNA defines 3D genomic organization necessary for gene regulation. Most mutations linked to complex traits are found outside the exome, and somatic insertion mutations are among the most underrepresented and poorly defined. The non-coding genome is also filled

with cis regulatory DNA elements like enhancers that play key roles in gene expression. Enhancer activities in DLBCL lock the growth regulatory network in an activated state and can be reliably identified though the histone mark H3K27 acetylation which is examined through ChIP-seq. ChIP-seq is the standard assay for genome-wide identification of DNA-associated protein binding sites. It has enabled the detection of enhancers through its ability to trace transcription factors which activate enhancers. FASTQ files containing sequence data generated by ChIP-seq is analysed using bioinformatics tools. Bioinformatics analyses rely on frameworks to address biological problems from a computational point of view. Nextflow is a prime workflow framework and a programming DSL for writing pipelines.

The current research would like to develop a flexible and reusable bioinformatics pipeline for DLBCL ChIP-seq data to uncover potential non-coding insertion-induced enhancers associated with the progression of DLBCL. Investigation into this arena could provide insight into the different disease mechanisms of DLBCL with the ultimate goal of functional precision therapy.



CHAPTER 3

RESEARCH PROCEDURE

3.1. INTRODUCTION

This chapter's main contribution is to detail the procedures and data science methodologies employed in this investigation as well as any problems encountered throughout. The discussed tools are some of the latest and most tested for handling data problems in bioinformatics and statistical analysis. This chapter describes a unique computational pipeline that encapsulates the procedures involved in this work, from collecting and curating biological data sets to developing processes that interpret the information contained within and connecting them in flexible and reusable workflows.

3.2. DATA COLLECTION

DLBCL H3K27ac ChIP-seq data with corresponding whole cell extract controls were queried for in the National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database. Three single read Sequence Read Run (SRR) identities were selected with accession numbers SRR1020510, SRR1020512, SRR1020514. The identities had corresponding whole cell extraction controls with accession numbers SRR1020511, SRR1020513, SRR1020515. The sratoolkit.2.10.8-ubuntu64 was used to connect to the NCBI SRA database via File Transfer Protocol (FTP). The operation FASTQ dump, which formed part of the sratoolkit, was used to access DLBCL H3K27ac ChIP-seq data matching each SRR identity in FASTQ format from the NCBI SRA database. The downloaded FASTQ files were renamed so that the accession numbers of the whole cell extract control files matched the accession numbers of their corresponding ChIP-seq treatment files (Table 1). The control files were further marked with the suffix '_c' while the treatment files were marked with the suffix '_t' so that the files remained distinguishable.

Table 1: DLBCL H3K27ac ChIP-seq treatment and control data files accessed from the SRA database.

Original File Name	Reformulated File Name	Size of File (GB)
SRR1020510	SRR1020510_t	17GB
SRR1020511	SRR1020510_c	12GB
SRR1020512	SRR1020512_t	13GB
SRR1020513	SRR1020512_c	14GB
SRR1020514	SRR1020514_t	13GB
SRR1020515	SRR1020514_c	12GB

The H3K27ac ChIP-seq files were generated on the Illumina HiSeq 2000 (*Homo sapiens*) platform by a study researching mutational effects in enhancers; the study GEO accession number is GSE46663, with BioProject accession number PRJNA201426. The data is available from the GEO website with accession numbers GSM1254206, GSM1254208 and GSM1254210 for the ChIP-seq treatment files, and GSM1254207, GSM1254209, GSM1254211 for the control files.

3.3. CONTAINERISATION

Docker is a commonly used container program but running and building it requires root capabilities (Mitra-Behura, Fiolka and Daetwyler, 2022). Since most users do not have root access, this creates a problem for HPC clusters used for sophisticated image processing operations. Although rootless mode is now available, it has restrictions, like the small number of supported storage drivers (Mitra-Behura, Fiolka and Daetwyler, 2022). Singularity containers have direct access to the host system's Linux kernel and allows one to install pre-built container images while ensuring the same software can be used across Linux systems and shared in a group (Kurtzer, Sochat and Bauer, 2017). Singularity was therefore found to be most suitable for the present study.

Singularity version 3.5.3 acted as the main container management system for this study. By executing the required programs in separate containers along with their dependencies, the environment in which they were executed was better controlled. A goal for this study's bioinformatics pipeline was reproducibility, and Singularity is the container management system of choice for cases where there might be multiple users of the same script, and where exact software versions and every specific environment might be required.

A custom container's construction is made up of a header and a body in the Singularity definition file. Details on the OS that had to be made or the base container to start from were included, along with instructions for installing software, configuring metadata and environment variables, and adding files from the host system. Existing images from Docker Hub for the tools' bowtie, bowtie2, MACS2, sambamba, SAMtools, BEDTools, BEDOPS and pBlat were used as a base for creating new Singularity images using the 'docker' bootstrap agent. Once the Singularity definition file for each tool was complete, the 'build' command was used to create fixed images of the pre-existing containers in the SIF format. Processes that made use of these tools and therefore required access to the images were specified in the nextflow.config file.

3.4. WORKFLOW FRAMEWORK AND SUPPORTING SYSTEMS

The aim of this study was to create a bioinformatics pipeline for DLBCL H3K27ac ChIP-seq data that could identify non-coding insertion-induced enhancers linked to DLBCL gene drivers. The bioinformatics pipeline was based on the computational techniques written in shell by Abraham *et al.* (2017); select perl scripts used in this investigation were downloaded from the link provided by the study. Nextflow version 21.04.0-edge (di Tommaso *et al.*, 2017) was used to design the script for the bioinformatics pipeline. Using Singularity software containers for which Nextflow has built-in support, scalable and repeatable operations were customised in standard scripting languages. The DSL2 syntax was the default setting and allowed parallel and modulated operations. The Nextflow pipeline was executed using SLURM on the South African National Bioinformatics Institute (SANBI) Dell HPC cluster with 232 CPU cores and 1952GB of RAM. The cluster used the operating system Linux version 5.4.0-121-generic (Ubuntu 20.04.4 LTS).

The source code of the developed bioinformatics pipeline is shared and saved in the following GitHub repository: https://github.com/wardahjassiem/enhancerAssociatedInsertions.

3.5. WORKFLOW STRUCTURE

The implicit workflow, which served as the entry point for the DLBCL H3K27ac ChIP-seq data in Table 1, was made up of three sub workflows incorporated as separate modules each with their own objective and invoked as functions with input channels passing as parameters.

The sub workflows were designed for three tasks respectively:

a) The identification of non-coding insertions.
b) The identification of areas in the genome enriched with aligned reads, i.e., peak calling.
c) The identification of enhancers and their associated genes linked to DLBCL.

Module scripts were written to define the various processes that were included and executed in each workflow. Parameters, executors, and other configuration specifications were defined in the nextflow.config file. Figure 2 depicts the structure of the investigative pipeline which will be further elaborated upon in the following sections.

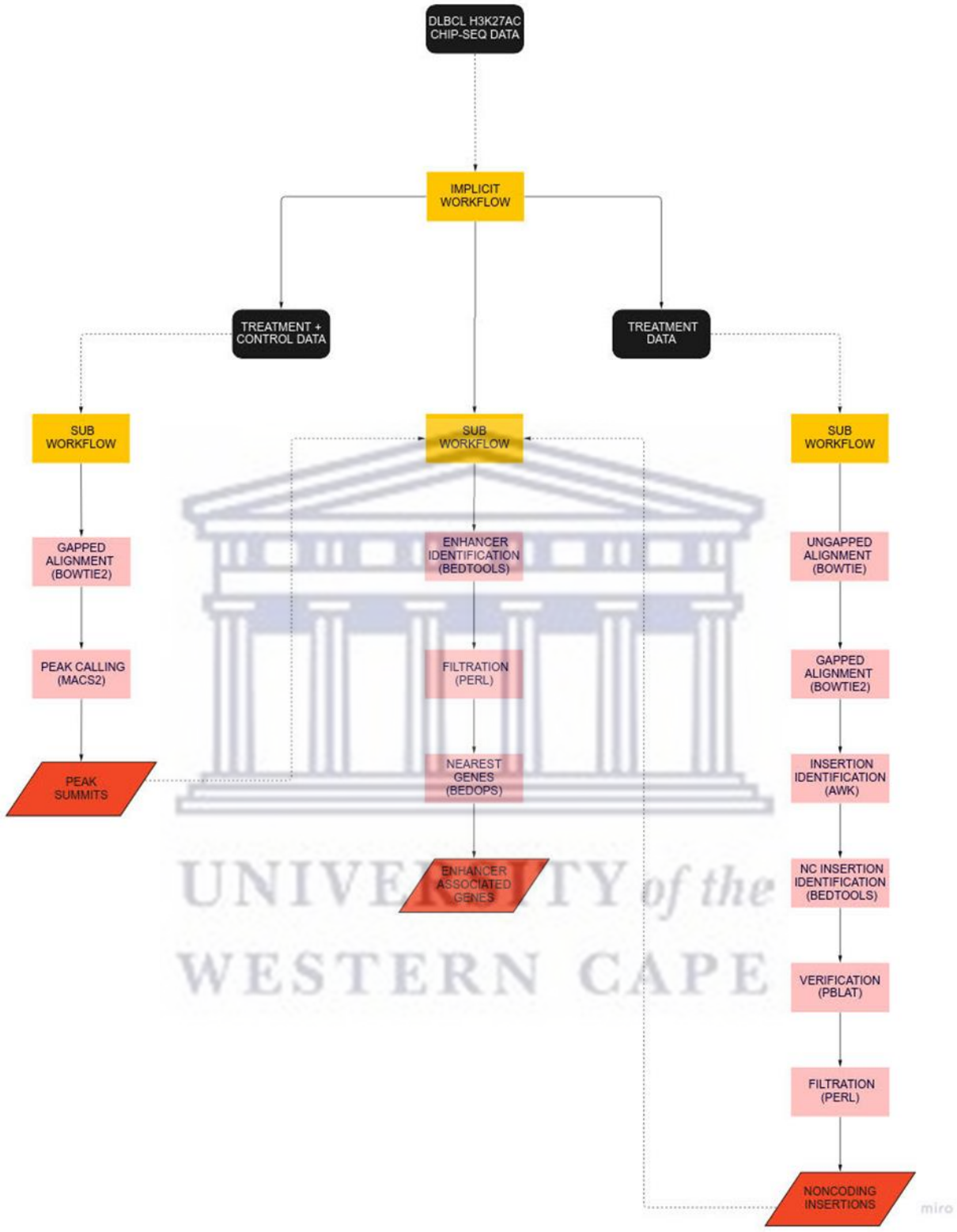


Figure 2: Diagram depicting the workflow frame for the bioinformatics pipeline.

The DLBCL ChIP-seq data was directed into the implicit workflow which then channelled the treatment data into a sub workflow for the identification of non-coding insertions, and the treatment and control data into a sub workflow for the identification of regions in the genome enriched with ChIP-seq reads. The output of these two sub workflows were channelled as input for the third and final sub workflow; the identification of insertion-induced enhancer associated genes. Abbreviations: nc, non-coding.

3.6. SUB WORKFLOW: IDENTIFICATION OF NON-CODING DLBCL INSERTIONS

The purpose of this sub workflow was to identify and verify insertions located in the non-coding genome using the DLBCL H3K27ac ChIP-seq treatment data files shown in Table 1.

The sub workflow declared six input channels from those specified in the implicit workflow. Three channels emitted the DLBCL H3K27ac treatment files and indices used for sequence mapping with bowtie and bowtie2, the fourth channel emitted the NCBI hg19 RefSeq genes used to extract coding sequences which was then emitted through a fifth channel to identify non-coding insertions. The final channel emitted hg19 human reference chromosomes used by pBlat to verify the identified non-coding insertion sequences.

The sub workflow declared a single, final output channel which emitted aligned, filtered, and sorted non-coding DLBCL insertions. The following subsections elaborate on the components defined in module (process) scripts that were imported into the sub workflow script to achieve the final output.

3.6.1. SEQUENCE ALIGNMENT

3.6.1.1. UNGAPPED READ ALIGNMENT

A Nextflow process was created to align the DLBCL H3K27ac ChIP-seq treatment data in FASTQ format to the hg19 human genome index (*H.sapiens*, UCSC hg19) using bowtie version 1.3.1 (Langmead, 2010), a fast and efficient short read aligner (Figure 3). The purpose was to identify ChIP-seq reads that contained gaps from potential insertions. Since bowtie cannot align reads spanning gaps, these events go overlooked. Therefore, reads that could be aligned successfully to the reference genome with bowtie were separated from reads that could not be aligned due to gaps.

```
bowtie --best --strata -m 1 -n 2 -p 24 -S --un ${fastq.baseName}-un.fastq -x
$bowtie1Ind --max /dev/null $fastq > ${fastq.baseName}_ungapped.sam
```

Figure 3: Command line for aligning DLBCL H3K27ac ChIP-seq reads in FASTQ format to the human reference genome index using bowtie.

The number of parallel search threads was set to 24 and only 2 mismatches were allowed in the seed. If valid reportable alignments were found in many alignment strata, only the alignments that fell into the best was reported. Alignments for a read was suppressed if there was more than 1 valid alignment for it, all reads with more than 1 valid alignment was put into a separate file. Only alignments with the lowest number of mismatches in the seed was reported along with the quality at the mismatch position. Aligned reads were saved in SAM format. All reads that could not be aligned were saved to a separate file in FASTQ format.

3.6.1.2. GAPPED READ ALIGNMENT

A recent work evaluated 12 pipelines for their ability to detect single nucleotide variations (Kisakol *et al.*, 2021). Pipelines that used aligners like Novalign identified a greater number of variants, but the precision rate was around 65% whereas pipelines that used bowtie2 had a precision rate of 90%. The FASTQ reads that could not be aligned with bowtie were therefore channelled into a process that ran bowtie2 version 2.4.5 (Langmead and Salzberg, 2012), an alignment tool that can be used for gapped reads (Figure 4). The purpose was to map DLBCL ChIP-seq reads that contained gaps from potential insertions to the hg19 human reference genome index (*H.sapiens*, UCSC hg19).


```
bowtie2 -k 1 -p 24 -q -x $bowtie2Ind -U $fastq -S
${fastq.baseName}_mapped.sam
```

Figure 4: Command line for aligning DLBCL H3K27ac ChIP-seq gapped reads to the human reference genome index using bowtie2.

The input DLBCL ChIP-seq reads were in FASTQ format. Bowtie2 was set to search for 1 valid alignment for each read before stopping the search. The number of parallel search threads was set to 24 and the aligned reads were saved in SAM format.

3.6.2. INSERTION IDENTIFICATION

A process was designed to filter the aligned DLBCL ChIP-seq reads from bowtie2 in SAM format according to query sequences that had insertions to the hg19 reference genome. To accomplish this, an AWK program defined a text pattern to be searched for in each line of each input SAM file and the action to be taken when a match was found. Each alignment line in SAM format represents the linear alignment of a segment. Each line has 11 or more TAB delimited fields. The sixth field is the CIGAR string used to show base matches, deletions, and insertions. In this case, the program searched for insertions in the sixth column (CIGAR string) of the SAM files and printed the entire alignment line when a match was found. All Nextflow processes in this study employing AWK in the command line used a shell block definition to allow the script to have Bash and Nextflow variables without needing to escape the first.

The files containing the identified insertion alignment lines did not have SAM headers necessary for complete SAM formatted files. To address this, aligned ChIP-seq reads from bowtie2 were fed into a process that extracted the SAM headings of each file using the head command. The SAM headings were then concatenated with the files containing insertion sequences.

The SAM files were fed into a process that made use of the view method from the SAMtools package version 1.15 (Li *et al.*, 2009) to convert the SAM formatted files to BAM format. The BAM files were then coordinate-sorted and indexed with sambamba (Tarasov *et al.*, 2015).

3.6.3. NON-CODING INSERTION IDENTIFICATION

3.6.3.1. EXON ACQUIREMENT

A process was designed to extract exons from the NCBI RefSeq Genes composite track, which shows human protein-coding and non-protein-coding genes from the NCBI RNA reference sequences collection. The exons would later be used in a downstream process to identify non-coding DLBCL insertions.

A perl script detailed code for the extraction of the chromosome, start and end frame of the exon, and the gene name for each sequence by specifying the necessary columns of information from the human RefSeq file and saving them in tabulated BED[1] format for compatibility between programs.

3.6.3.2. OVERLAPPING FEATURE IDENTIFICATION

The human exons in BED format and the sorted DLBCL non-coding insertions in BAM format were used as input for bedtools intersect, an operation that formed part of the BEDTools package version 2.30.0 (Quinlan, 2014). Bedtools intersect was used to screen for overlapping features between the insertions and the exons, i.e., to determine whether any of the insertion sequences were not found in the coding regions of the human genome. Regions in each chromosome was intersected with the region of each insertion. The bedtools intersect option -v dictated that the operation would output features that did not overlap, i.e., to report insertion sequences that were in the non-coding regions of the human genome. The non-coding DLBCL insertions were emitted in SAM format.

[1] BED format has a minimum of three columns and nine optional columns. The first three contain the chromosome, and the start and end coordinates of the sequences. The fourth column contains the name of the sequence.

3.6.4. NON-CODING INSERTION SEQUENCE FILTRATION

The non-coding DLBCL insertions in SAM format were fed into two individual processes that used AWK to extract certain columns of information stored in the alignment section of the SAM formatted files to create FASTQ and FASTA formatted files for downstream processing of the non-coding insertions. Both types of formatted files contained the same sequence information, however, the FASTA files were required for non-coding insertion alignment to the human reference genome, and the FASTQ files were required for non-coding insertion sequence filtration.

3.6.5. NON-CODING INSERTION VERIFICATION

The non-coding insertions in FASTA format were verified with pBlat (Wang and Kong, 2019) by locating the positions of the insertions in the human genome (Figure 5). The program pBlat facilitates the high-throughput mapping of large-scale sequences to reference genomes with speed and accuracy. The prepared FASTA files containing the insertion sequences and the hg19 human reference chromosomes (chromosomes 1-22 as well as chromosomes X, Y and M) from the UCSC Browser were used as input for the pBlat program.

```
for x in $chrpath
  do
    export chr=`basename \$x .fa`
      pBlat -threads=23 -minScore=0 -stepSize=1 \$x $fasta \$chr.${fasta.baseName}.psl
      tail -n +6 \$chr.${fasta.baseName}.psl > \$chr.${fasta.baseName}.blat.psl
done
```

Figure 5: Command line for insertion sequence alignment using pBlat.

The number of parallel threads was set to 23, the stepSize was reduced from default 11 to 1 with the minimum score at 0 (the number of matches minus the number of mismatches minus a gap penalty). The output was in the default TAB separated PSL[2] format.

[2] In PSL format, each alignment is represented by a line with 21 necessary fields. The format contains information about the alignments (insertions, deletions, matches, mismatches) but not the sequences themselves.

The process attempted to align each FASTA file, containing non-coding insertion sequences, to each hg19 human reference chromosome. This resulted in 25 output PSL files for each of the 3 FASTA files. The aligned non-coding insertion sequence PSL files were concatenated by file base name into a single PSL file for each of the 3 SRR identities.

3.6.6. ALIGNED NON-CODING INSERTION SEQUENCE FILTRATION

A process was designed that executed a perl script to identify non-coding insertions aligned by pBlat that occurred once and select the best hit with the highest score amongst those that occurred multiple times or were PCR duplicates.

The pBlat output was parsed so that each accepted read hit included the entire read sequence. Given that the whole read was aligned, and the insertion was smaller than the read size, if a hit occurred more than once, it was examined to determine whether it was a PCR duplicate. The hit along with the query name containing the chromosome position in the CIGAR string was then printed. If the hit was not identical to the original read line, it was taken that the insertion was present in multiple chromosomes. The program then printed "has multi hits" along with the read line. If the hit was identified as occurring for the first time, the program then printed "has first hit". In the case of multiple hits, the best hit was selected. A read was accepted upon the following conditions: there were no pBlat-called mismatches and there was only one insertion of less than 20bp; the reference chromosome name from the bowtie2 process was the same as the target sequence name from the pBlat process; and the start positions of the target sequences from bowtie2 and pBlat were less than 100bp apart.

The output files were in SAM format and included the query name of the non-coding insertions, the bowtie2 reference chromosome names, the alignment start positions of the pBlat target sequences, the CIGAR strings, and the observed template lengths.

Within the Nextflow script of the process, the base name of the pBlat aligned insertion files was used as a key for the Nextflow tuple[3] qualifier. This would enable downstream processes to receive tuples of values as input that had to be handled individually.

[3] The Nextflow tuple qualifier allows multiple parameters to be grouped as one.

3.6.7. NON-CODING INSERTION AND ALIGNED NON-CODING INSERTION INTERSECTION

A process was designed that used a perl script to identify the matched reads between the aligned and filtered non-coding insertions from pBlat and the unaligned non-coding insertions in FASTQ format. The output files were in SAM format and included the length of the sequence but not the sequence quality. The SAM files were channelled into a second process that made use of the same perl script to identify overlaps once again with the unaligned non-coding insertions in FASTQ format. The output of the second filtration process was also in SAM format but included the sequence length as well as the sequence quality.

The Nextflow scripts of the processes invoked the tuple qualifier using the key previously defined which specified the SRR identity names. The Nextflow join() operator created a channel that joined the two input channels for each of these processes by the defined key. This enabled the use of paired files; each filtered PSL file aligned by pBlat was processed with its matching unaligned read file in FASTQ format.

The SAM files containing the filtered non-coding DLBCL insertion sequences were concatenated with the SAM headings extracted from the bowtie2 aligned reads and converted to BAM format. The BAM alignments were then sorted according to the leftmost coordinates and indexed.

3.7. SUB WORKFLOW: IDENTIFICATION OF DLBCL PEAK REGIONS

The sub workflow used both the treatment and control DLBCL H3K27ac ChIP-seq data files depicted in Table 1 to identify transcription factor binding sites and locate places in the genome that were enriched with aligned reads, i.e., peak calling. Processes defined in modules were imported into the sub workflow script.

The sub workflow called for two input channels defined in the implicit workflow. The first emitted the previously downloaded DLBCL H3K27ac ChIP-seq data in FASTQ format. The treatment and control data were emitted in corresponding pairs according to their reformulated names as per Table 1. The sub workflow processes used the Nextflow tuple qualifier so that values would be grouped by file base name but handled individually in a single parameter definition. Each treatment file in the tuple was therefore processed with its

corresponding control file as per tuple definition. The second input channel emitted the hg19 index used for read mapping with bowtie2.

The peaks[4] generated by MACS2 in BED format served as the final output of the sub workflow. The following subsections elaborate on the components defined in module scripts that were imported into the sub workflow script to achieve the final output.

3.7.1. GAPPED READ ALIGNMENT

It is highly recommended that mapped reads from treatment and control samples/input DNA are used during peak calling, which is a procedure that outputs a set of regions representative of transcription factor binding locations. The reads were aligned to the hg19 human reference genome index (*H.sapiens*, UCSC hg19) using bowtie2 (Figure 6). Bowtie2 works best for reads that are at least 50 bp and has a local alignment mode which performs soft clipping to remove poor quality bases or adapters from untrimmed reads (Langmead and Salzberg, 2012).

```
bowtie2 -p 5 -q --local -x $bowtie2Ind -U ${fastqs.find{it.toString().contains('_t.fastq')}}
-S ${fq1_name}.sam

bowtie2 -p 5 -q --local -x $bowtie2Ind -U ${fastqs.find{it.toString().contains('_c.fastq')}}
-S ${fq2_name}.sam
```

Figure 6: Command line for aligning DLBCL H3K27ac ChIP-seq reads and whole cell extract controls in FASTQ format to the human reference genome using bowtie2.

The input ChIP-seq reads were in FASTQ format. The number of parallel search threads was set to 5, and the --local mode was activated to allow soft clipping at the read ends to get the best alignment scores. Bowtie2 was run on the treatment and control data in parallel. Files that contained the control marker ('_c') and files that contained the treatment marker ('_t') were processed separately by bowtie2. The aligned reads were saved in SAM format.

The aligned reads from bowtie2 in SAM formatted files were converted to BAM format. The BAM files were then coordinate-sorted and indexed with sambamba.

[4] Regions of the genome where many reads align that are suggestive of enhancer activity.

3.7.2. UNIQUELY MAPPED READS

It is recommended to use uniquely mapped reads for peak calling to improve specificity since ChIP-seq data tends to contain duplicates and much redundancy. Sambamba was used to filter the sorted BAM files to keep only uniquely mapped ChIP-seq reads (Figure 7).

```
sambamba view -h -t 2 -f bam -F "[XS] == null and not unmapped and not duplicate" ${ctl.baseName}.bam > ${ctl.baseName}_unique.bam

sambamba view -h -t 2 -f bam -F "[XS] == null and not unmapped and not duplicate" ${trt.baseName}.bam > ${trt.baseName}_unique.bam
```

Figure 7: Command line to filter uniquely mapped DLBCL reads using sambamba.

The parameters -t, -h, and -f specified that the program was to use 2 threads, print the SAM headers before the reads and produce all output files in BAM format. The -F parameter then described the filters implemented; unmapped reads were removed by specifying 'not unmapped', and duplicates were removed with 'not duplicate'. Multimappers were removed from among the aligned reads by specifying '[XS] == null'. The bowtie2 'XS' tag provides an alignment score for the second-best alignment and is only present if the read has more than one alignment. Each treatment file with its corresponding control file was run in parallel but separately.

3.7.3. PEAK CALLING

Peak calling identifies transcription factor binding sites; regions of the genome enriched with aligned data from a ChIP-seq experiment. To determine the characteristics of strategies that enable some to perform better than others, a 2017 study evaluated six peaking calling methods (Thomas *et al.*, 2017). In terms of sensitivity, accuracy, and F-score metrics for low, medium, and high noise levels, MACS2 excelled. Methods that rate their candidate peaks using a Poisson test, like MACS2, rather than a Binomial test are more effective for statistical testing of candidate peaks. The best operating features on simulated transcription factor binding data was found in MACS2. The information provided direction and justification for the peak caller chosen in this investigation.

The aligned and sorted DLBCL ChIP-seq data in BAM format was channelled into process that used MACS2 (Feng, Liu and Zhang, 2011) to identify peak regions, i.e., transcription binding sites indicative of enhancer activity (Figure 8). The whole cell extracts for each treatment file were used as controls to increase the robustness of called peaks.

```
macs2 callpeak -t ${trt} -c ${ctl} -g hs --bdg -n ${baseName} -f BAM -p 1e-9
```

Figure 8: Peak calling performed on DLBCL ChIP-seq treatment and control data using MACS2.

MACS2 has seven functions available as sub-commands. Callpeak is the main function and was invoked with 'macs2 callpeak'. A --p value cut off of 1e-9 for peak detection was used. The option -f was set to specify that the input file format would be BAM. The process emitted the peak summits in BED format. Peak summits are useful for finding motifs at binding sites.

3.8. SUB WORKFLOW: IDENTIFICATION OF ENHANCERS AND ASSOCIATED GENES

The sub workflow was used to identify putative non-coding insertion-induced enhancers and associated genes that may act as drivers in DLBCL. The non-coding insertions in BAM format and the peak summits in BED format outputted by the previous two sub workflows were intersected to obtain non-coding insertion-induced enhancers. The hg19 human reference genes (UCSC Table Browser) was used by BEDOPS to locate genes closest to the identified enhancers.

The final output of the workflow was a list of potential noncoding insertion-induced enhancer associated genes that may play a role in DLBCL pathogenesis. The following subsections elaborate on the components defined in module scripts that were imported into the sub workflow script to achieve the final output.

3.8.1. INSERTION-INDUCED ENHANCER IDENTIFICATION

The sorted and filtered non-coding insertions in BAM format and the peak summits in BED format were directed into the operation bedtools intersect from the BEDTools package (Quinlan and Hall, 2010). The sub workflow script used the Nextflow join() operator to combine the elements emitted by the two input channels based on the matching predefined key (SRR identity). The process script used the Nextflow tuple qualifier to associate the elements of the two parameters based on the tuple definition while still allowing them to be handled separately. Bedtools intersect was used to screen for overlapping features between the insertions and the peak summits, i.e., to determine which insertion sequences were found in enriched regions in the genome. The non-coding insertion-induced enhancers in BAM format was converted to SAM format.

3.8.2. POTENTIAL ENHANCER FILTRATION

The non-coding insertion-induced enhancers in SAM format was channelled into a process that filtered the sequences and converted the files to BED format using a perl script.

The CIGAR string of the SAM files was used to check if an insertion was present in each SAM alignment line and then to determine its length. The script went on to determine how many unique letters were present within each insertion. The downstream position of the insertion was determined using the number of matching bases after the insertion, starting from position 0 to the insert length. Similarly, the upstream position of the insertion was determined using the number of matching bases before the insertion. The start position was the length of the matching bases less the length of the insertion, and the end point was the length of the insertion. The start position of each insertion was determined by adding the length of matching bases before the insertion to the 1-based leftmost mapping position of the read.

The output files emitted in BED format contained the reference chromosome name, the start and end positions of the insertions, and the insertion sequences, and was used in downstream analysis of the putative insertion-induced enhancers.

3.8.3. ENHANCER ASSOCIATED GENES

BEDOPS version 2.4.41 (Neph *et al.*, 2012) was used to locate the genes nearest to the sequence positions of the insertion-induced enhancers in BED format based on genomic distance.

The start and end coordinates of the non-coding insertions in BED format were identical because they described zero length features. BED format is defined as half-open, so this is what is required by BEDOPS tools (Neph *et al.*, 2012). The reads in BED format were fed into a process that used AWK to modify the end coordinates of the insertions by adding the 1 integer. The sort-bed operation from the BEDOPS package version 2.4.41 sorted the resultant BED files first by lexicographic chromosome order, then by ascending integer start coordinate order, and finally by ascending integer end coordinate order. This allowed downstream BEDOPS tools to work properly and quickly without software modifications. The BED files were piped to the linux uniq command which filtered out repeated lines.

The operation closest-features from the BEDOPS package was used to identify the genes nearest to the non-coding insertion-induced enhancers. The hg19 human reference genes was downloaded from the UCSC Table Browser with group 'Genes and Gene Predictions', track 'GENCODE V41lift37', and table 'Comprehensive (wgEncodeGencodeCompV41lift37'). Each reference gene was compared to each enhancer region in the BED files. The closest gene to each noncoding insertion-induced enhancer was outputted in BED format along with the distance between them. The files were delimited by tabulation to make further user specific processing easier.

3.9. FUNCTIONAL ANNOTATION ANALYSIS

A bioinformatics resource system called the Database for Annotation, Visualisation, and Integrated Discovery (DAVID) for gene ontology and pathway analyses was used to perform functional enrichment analysis on the DLBCL non-coding insertion-induced enhancer associated gene list detected by BEDOPS outside of the pipeline as a validation step (Huang *et al.*, 2007).

CHAPTER 4

RESULTS

4.1. INTRODUCTION

This chapter establishes the findings produced using the bioinformatics methodologies and technology previously described. It explores the identification of potential novel enhancers involved in the tumorigenesis of DLBCL and the impact of insertions located outside of the exome as driver mutations. It also analyses the flexibility and sensitivity of the pipeline developed to answer the research questions put forth by the study. All experimental gene symbols with their associated names can be found in the appendix.

4.2. MUTATIONAL EVENTS DETECTED

On a high-performance computing cluster with 4 CPUs, 1 node, 1 thread per core, and a memory of 90GB per task, the pipeline took 10 days, 9 hours, and 34 minutes to complete. The tool pBlat monopolised most of this time but was still significantly more time efficient than Blat due to its inherent parallel processing.

The results from the analysis and filtration of the H3K27ac ChIP-seq data obtained from DLBCL tissue samples were summarised in Table 2. The first sub workflow of the investigative pipeline identified a total of 104,628 insertions not located in the exome and therefore deduced to be non-coding events. Peak calling, which was the objective of the second sub workflow, was done on the DLBCL treatment data along with the corresponding whole-cell extracts which contained sequence reads from chromatin lysates prepared for ChIP-seq but without antibody selection. The results of the pipeline's second sub workflow showed 127,407 areas in the genome enriched with aligned reads from ChIP-seq indicative of transcription factor binding. The third and final sub workflow of the pipeline intersected the peak summits with the non-coding insertions and identified a total of 1,437 potential non-coding insertion induced DLBCL enhancers. The number of mutational events found in each SRR file did not correspond with the size of the file, i.e., mutational activity was not seen to increase along with file size (number of ChIP-seq reads per file in GB). The mutational activity referred to includes the number of non-coding insertions, peak regions, and enhancers per DLBCL H3K27ac ChIP-seq file.

Table 2: Summary of the results from the analysis of DLBCL H3K27ac ChIP-seq using the designed bioinformatics pipeline.

SRR File	Size of SRR File (GB)	Non-Coding Insertions	Number of Peak Regions	Insertion-Induced Enhancers
SRR1020510	17GB	20,194	26,224	104
SRR1020512	13GB	48,754	42,934	964
SRR1020514	13GB	35,680	58,249	369
Total	43GB	104,628	127,407	1,437

It was expected that the number of mutational events would increase as the size of the file increased due to growing amounts of DLBCL H3K27ac ChIP-seq data. This, however, was not observed. File SRR1020510 was the largest at 17GB, followed by files SRR1020512 and SRR1020514 at 13GB each. The largest file consistently showed significantly fewer non-coding insertions, enriched genomic regions, and enhancers than the other two files. Conversely, one of the smaller files (SRR1020512) yielded the most insertions and enhancers. Although files SRR1020512 and SRR1020514 were the same in size at 13GB, analysis revealed vastly different results for both files. The rate of enhancers also did not consistently correspond with the number of enriched regions (peaks); while file SRR1020514 contained more peak regions than file SRR1020512, it showed fewer insertions and correspondingly fewer enhancers. The rate of enhancer activity did, however, correspond with the number of insertions found in each file, i.e., as more insertions were located in a file, the number of enhancers also increased. Clinical factors, such as disease stage, may have influenced the discrepancy in the mutational density of the DLBCL ChIP-seq data files.

The highest percentages of non-coding insertion-induced enhancers were identified in chromosomes 1, 19, 2, 3, 6 and 17, while the lowest percentages of enhancers were recorded in chromosomes 13, 21, and 18 (Figure 9). It was necessary to take chromosome length into consideration.

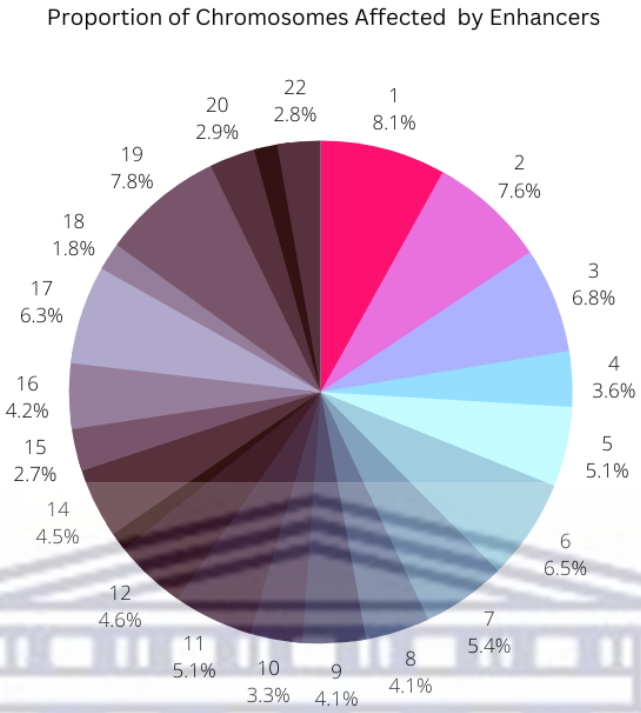


Figure 9: Graph depicting the percentage contribution of non-coding insertion-induced enhancers per chromosome.

Each slice is representative of the percentage of enhancers identified in each somatic chromosome. Chromosomes 1 and 19 contributed the greatest share of enhancers at percentages of 8.1 and 7.8 respectively. Chromosomes 13 and 21 had slices of less than 1.5% which could not be displayed.

The chromosome lengths referred to in this investigation were taken in centimetres from a published study (Piovesan *et al.*, 2019). Chromosomes 1-3 are typically the longest chromosomes at 8.14cm, 7.92cm, and 6.48cm respectively and so it was expected that more enhancers would be identified in these areas, with the proportion of enhancers increasing with chromosome length. This was the general trend observed, the most notable exceptions being in chromosomes 13, 17 and 19.

Chromosome 19 is relatively small in size (1.92cm) compared to chromosomes 1-3, yet the data showed that it contributed a percentage of enhancers close to that of chromosome 1 (Figure 9). Similar results were found for chromosome 17 with a length of 2.72cm. The trend of enhancer activity increasing with chromosome length would suggest that the lowest number of enhancers among somatic chromosomes should be found in the shortest chromosome, 21. However, chromosome 13, at 3.74cm in length, contributed the fewest enhancers of all the somatic human chromosomes at a percentage of 1.18. This was followed by chromosomes 21 and 18, at percentages of 1.41 and 1.8 respectively (Figure 9). The former is 1.53cm in length while the latter is 2.63cm in length.

The relationship between the rate of enhancer activity and chromosome length was graphically displayed in Figure 10. Enhancer activity was seen to increase along with chromosome length, except in the case of chromosomes 19 and 17 where steep upward inclines were observed which were not in keeping with the trend observed regarding the lengths of the chromosomes. Another anomaly observed was in chromosome 13, which was greater in length than chromosomes 21 or 18, yet it showed the lowest rate of enhancer activity among all the somatic chromosomes. Chromatin accessibility and other epigenetic factors may play a role in this observation.

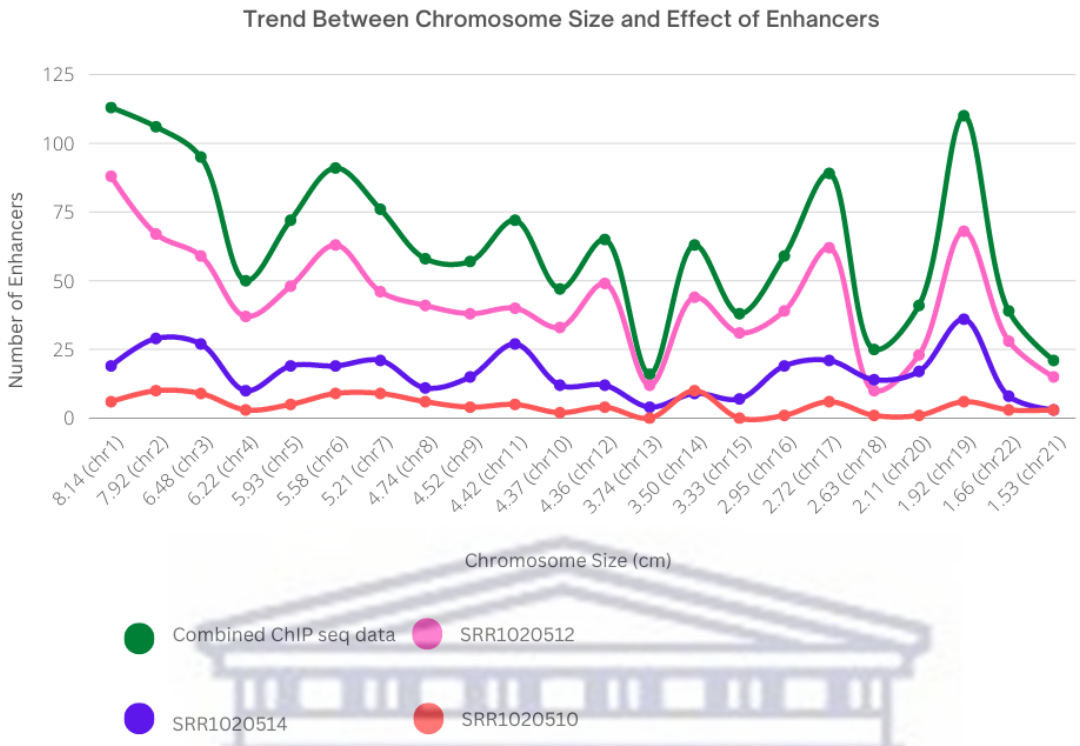
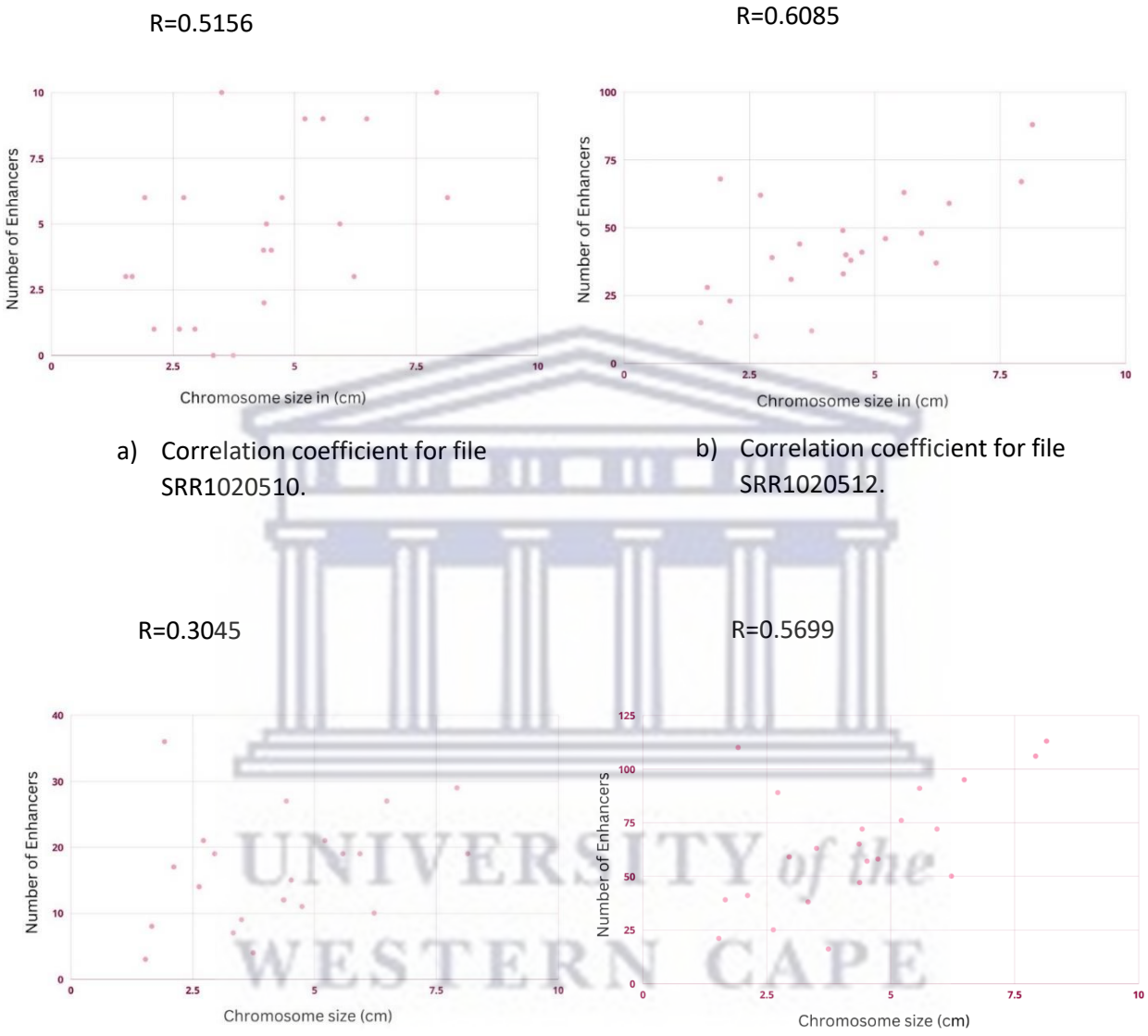


Figure 10: Graph depicting enhancer occurrence in relation to somatic chromosome size (cm) on the combined DLBCL data and individual SRR ChIP-seq files.

The chromosome lengths were taken in centimetres from a published study (Piovesan et al., 2019). The pink line representing ChIP-seq file SRR1020512 had notably more enhancer activity across all somatic chromosomes. The red line representing ChIP-seq file SRR1020510 had the lowest rate of enhancer activity across all chromosomes, despite containing the most DLBCL reads. The blue line representing ChIP-seq file SRR1020514 had a lower enhancer rate than that of file SRR1020512, despite being the same size in GB, but was still higher than that of file SRR1020510. The green line represented the enhancer rate of all the DLBCL ChIP-seq data in total.

Enhancer activity generally increasing along with chromosome length indicated a positive correlation between the variables which was confirmed by statistical analysis. The Pearson correlation coefficient of the combined ChIP-seq data indicated a moderate and positive relationship between enhancer occurrence and chromosome length, i.e., as the chromosome length increased, the number of enhancers in that chromosome generally also tended to increase (Figure 11). The exception was the ChIP-seq data from file SRR1020514, which did not statistically display a strong relationship between enhancers and chromosomes in terms of length. Table 3 showed that the p-value was estimated to be significant at $p<0.05$ for each SRR file except file SRR1020514. Since the p-value for the combined ChIP-seq data was

significant at $p<0.05$, the consensus was that the probability of enhancer activity being consistently proportional to chromosome length was likely.



a) Correlation coefficient for file SRR1020510.

b) Correlation coefficient for file SRR1020512.

c) Correlation coefficient for file SRR1020514.

d) Correlation coefficient for combined ChIP-seq data.

Figure 11: Correlation coefficient for the DLBCL ChIP-seq data in relation to chromosome length.

Each SRR file of DLBCL ChIP-seq data showed a positive relationship between enhancer activity and chromosome length, file SRR1020514 however only had a weak linear association between the variables.

Table 3: Statistical scoring of the DLBCL ChIP-seq data in relation to chromosome size.

File	Correlation coefficient (R)	Coefficient of determination (R2)	p-value (exact)
SRR1020510	0.5156	0.2658	0.014048
SRR1020512	0.6085	0.3703	0.002656
SRR1020514	0.3045	0.0927	0.0168252
Combined ChIP-seq data	0.5699	0.3248	0.005624

By reorganising the enhancer rate according to the number of genes within each chromosome, a similar pattern to that seen with chromosome length was observed; the number of enhancers increased as the number of genes increased within each chromosome (Figure 12), which translated into the number of non-coding insertion mutations increasing as the number of genes increased.

As previously stated, the highest percentages of enhancers were identified in chromosomes 1 and 19. Chromosome 1 contains the highest number of genes at 2,100, followed by chromosome 19 at 1500 genes. The lowest percentages of enhancers were recorded in chromosomes 21 and 13, which contain 300 and 400 genes respectively, two of the lowest chromosomal gene counts. Chromosome length is not necessarily an indicator of the number of genes within a chromosome. Chromosomes 19 and 17 are among the smaller chromosomes in terms of length, but they are the most gene dense. Gene density, which is the ratio of the number genes per number of base pairs, may be a factor influencing the aggregation of enhancers to certain chromosomes. The study identified chromosome 18 had the third lowest rate of enhancer activity at a percentage of 1.8 (Figure 9). Both chromosome 21 and 18 contain approximately 300 genes, however the former is 1.53cm in length while the latter is 2.63cm in length. Chromosome 18 has the lowest gene density (Figure 12).

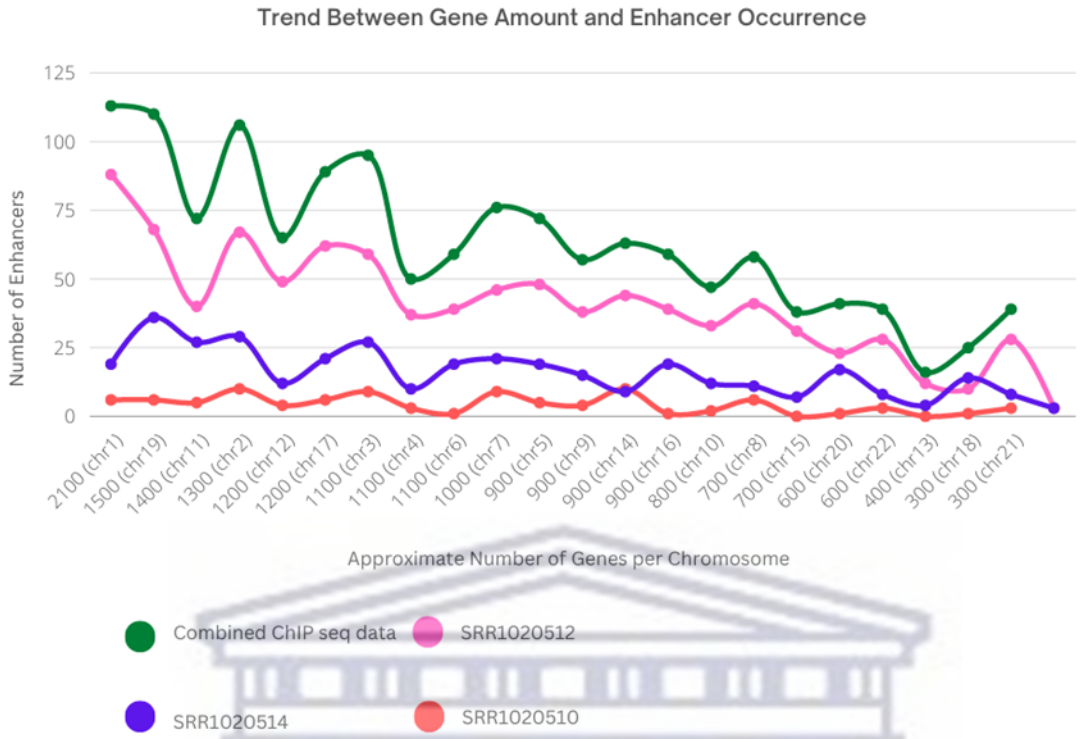
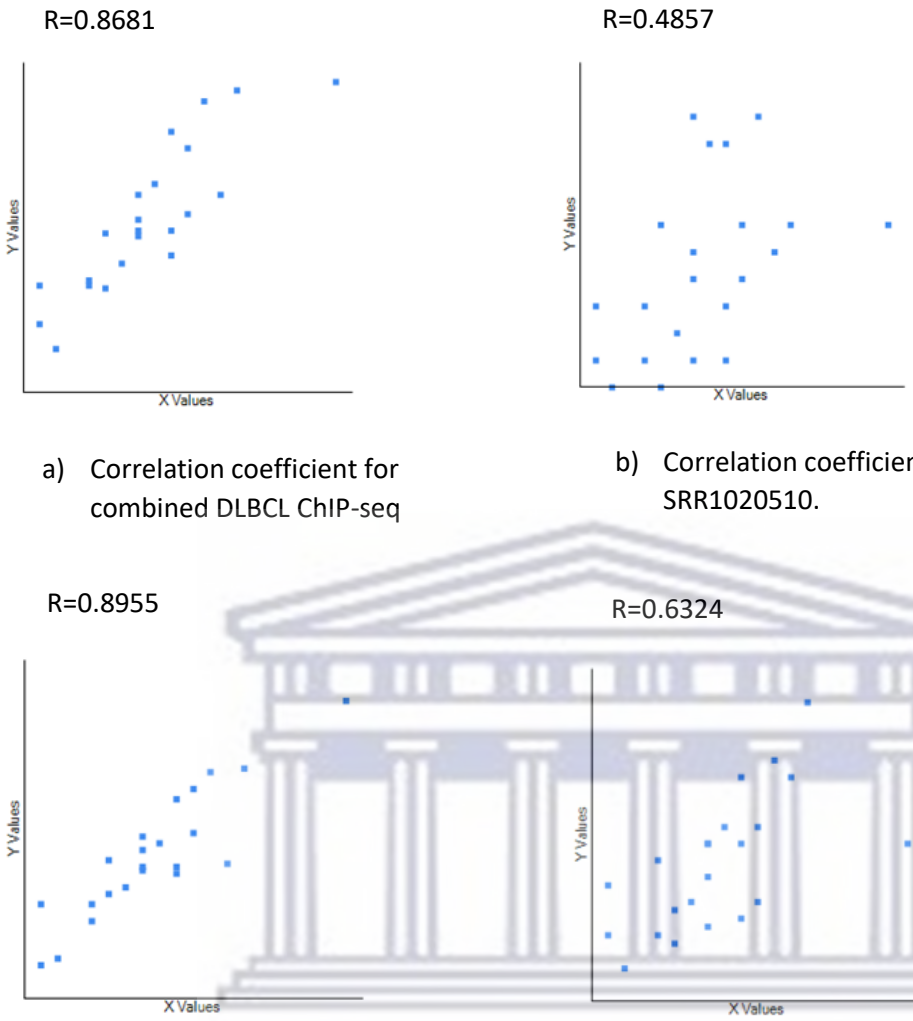


Figure 12: Graph depicting enhancer occurrence in relation to the number of genes per somatic chromosome for the combined DLBCL data and the individual SRR ChIP-seq files.

The approximate number of genes were taken from a medical online site (MedlinePlus: Chromosomes & mtDNA, 2021). *File SRR1020510 showed the lowest enhancer activity while file SRR1020512 showed the highest enhancer activity. Each data line in the graph curved in a general upward slope indicating that the number of enhancers was directly proportional to the number genes.*

The Pearson correlation coefficient of the DLBCL ChIP-seq data confirmed a very strong and positive relationship between enhancer occurrence and number of genes (Figure 13). Table 4 shows that the p-value was estimated to be significant at $p<0.05$ for each group of DLBCL ChIP-seq data.



a) Correlation coefficient for combined DLBCL ChIP-seq

b) Correlation coefficient for file SRR1020510.

c) Correlation coefficient for file SRR1020512.

d) Correlation coefficient for file SRR1020514.

Figure 13: Correlation coefficient for the DLBCL ChIP-seq data in relation to the number of genes.

Each group of DLBCL ChIP-seq data showed a positive relationship between enhancer activity and number of genes, with a strong linear association between the variables.

Table 4: Statistical scoring of the DLBCL ChIP-seq data in relation to the number of genes.

File	Correlation coefficient (R)	Coefficient of determination (R2)	p-value (exact)
SRR1020510	0.4857	0.2359	.021928
SRR1020512	0.8955	0.8019	< .00001
SRR1020514	0.6324	0.3999	.001589
Combined ChIP-seq data	0.8681	0.7536	< .00001

The bioinformatics pipeline did not identify any non-coding insertion-induced enhancer activity in chromosome Y, and only one enhancer event was detected in chromosome M (mitochondrial DNA) (Figure 14). Chromosome Y is the second smallest chromosome at 1.87cm and has the fewest genes (70-200) amongst all the human chromosomes. All of the enhancers found among the sex chromosomes came from chromosome X (Figure 15). The chromosome with the highest rate of enhancer activity for each SRR file was chromosome 2 for SRR1020510, chromosome 1 for SRR1020512, and chromosome 19 for SRR1020514.

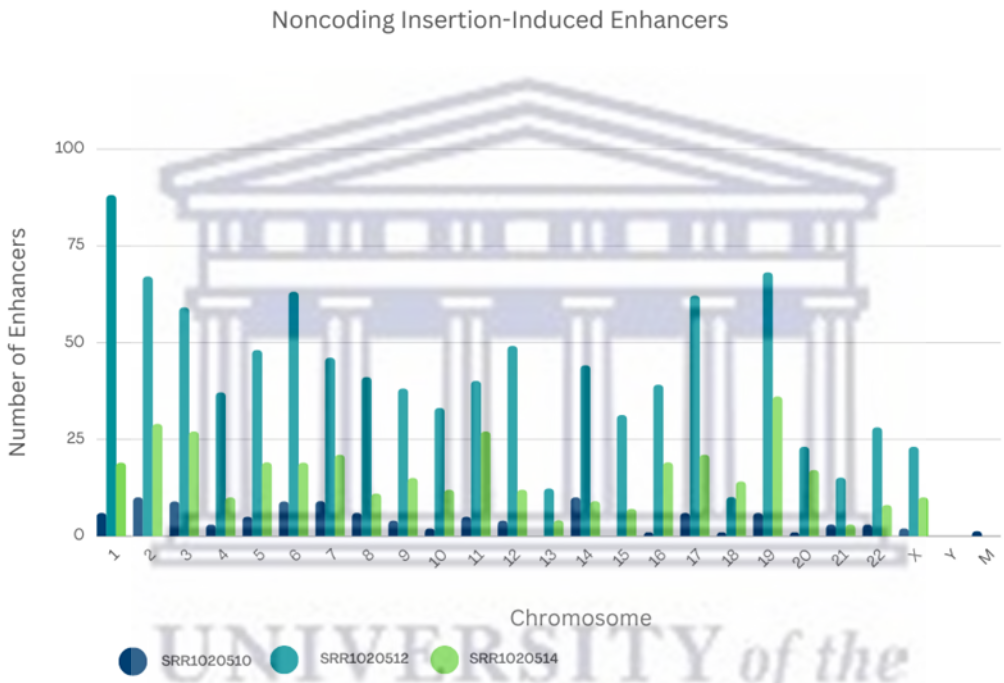


Figure 14: Graph depicting the contribution of non-coding insertion-induced enhancers per chromosome for each DLBCL H3K27ac ChIP-seq file.

File SRR1020510 contained the least enhancers across all chromosomes even though it was the largest file, for chromosomes 13 and 15 no enhancers were found. However, SRR1020510 was the only file in which an enhancer event was detected for chromosome M. Files SRR1020512 and SRR1020514 contained the same amount of ChIP-seq data, yet the rate of enhancer activity in file SRR1020512 far outstripped that found in either of the other files.

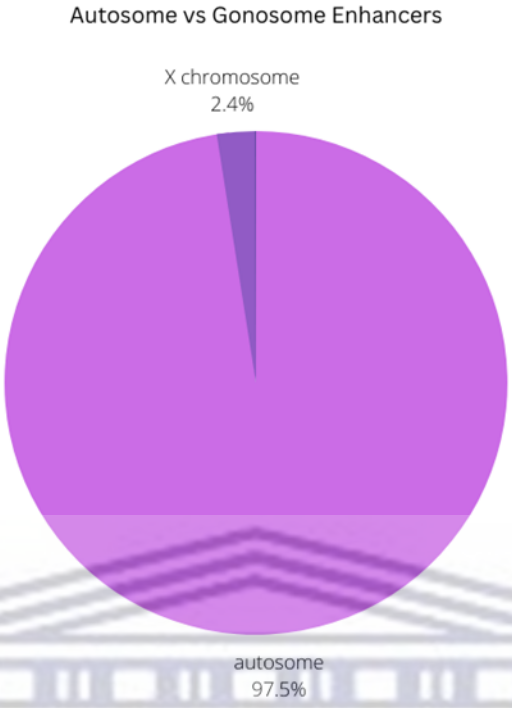


Figure 15: Graph depicting proportion of enhancers in autosomal chromosomes as compared to enhancers in non-autosomal chromosomes.

Majority of the non-coding insertion-induced enhancers were identified in somatic chromosomes/autosomes. Non-autosomal/gonosomal enhancer activity was observed to be relatively low in comparison and was mainly restricted to chromosome X.

A total of 6 enhancer associated genes were found to be common to all 3 SRR files; two of the genes were well-known oncogenes, dual specificity tyrosine phosphorylation regulated kinase 1A (*DYRK1A*) and COPI coat complex subunit beta 2 (*COPB2*) (Table 5). *DYRK1A* is located in chromosome 21, one of the smallest chromosomes in size and lowest in gene count. Of the 6 genes mentioned, 4 genes were in chromosomes 1-3 and 19, which were observed as being hotspots for enhancer events detected in this investigation. The distance between the DLBCL peak region and the TSS of the enhancer associated gene was at 0 in most cases of this investigation.

Table 5: Enhancer associated genes commonly identified in ChIP-seq files SRR1020510, SRR1020512, and SRR1020514.

Chromosome	Start position	End position	Gene	Distance to TSS
chr1	148556094	148577660	*NBPF15*	0
chr14	22975628	22975687	*TRAJ35*	283
chr19	57874934	57876677	*TRAPPC2B*	0
chr2	65073263	65090760	*LINC01800*	-1644
chr21	38739402	38885075	*DYRK1A*	0
chr3	139098852	139108488	*COPB2*	0

The relatively few (only 6) genes commonly affected by insertion-induced enhancers in all 3 ChIP-seq data files was thought to be due to the low rate of activity in file SRR1020510- the few genes affected in this file lowered the number of commonly affected genes amongst all 3 files. Files SRR1020512 and SRR1020514 had a comparatively greater rate of enhancer activity than file SRR1020512, with 24 genes identified to be commonly affected (Table 6). Instances where a gene was reported more than once was not representative of duplicates, rather they indicated the presence of different insertion events within the same locus of the gene, i.e., one base insertion, two base insertions, etc.

Table 6: Enhancer associated genes commonly identified in ChIP-seq files SRR1020512 and SRR1020514.

Chromosome	Start position	End position	Gene	Distance to TSS
chr1	148556094	148577660	*NBPF15*	0
chr10	112327484	112350844	*SMC3*	0
chr14	22975628	22975687	*TRAJ35*	283
chr16	9056562	9060847	*USP7-AS1*	0
chr16	9056562	9060847	*USP7-AS1*	0
chr17	10600932	10609245	*ADPRM*	0
chr17	20059401	20140492	*SPECC1*	0
chr19	20736597	20844389	*ENSG00000269110*	0
chr19	57874934	57876677	*TRAPPC2B*	0
chr2	64415648	64479736	*ENSG00000225889*	0
chr2	65073263	65090760	*LINC01800*	-1644
chr20	5556545	5591570	*GPCPD1*	0
chr21	38739402	38885075	*DYRK1A*	0
chr3	8543573	8609450	*LMCD1*	0
chr3	139098852	139108488	*COPB2*	0
chr4	25863451	25931167	*SMIM20*	0
chr4	25863451	25931167	*SMIM20*	0
chr5	178152376	178157660	*ZNF354A*	0
chr6	20321691	20333424	*ENSG00000286590*	1587
chr6	32485129	32498064	*HLA-DRB5*	0
chr6	36853727	36896740	*C6orf89*	0
chr7	130736623	130792687	*LINC-PINT*	0
chr7	154735399	154794834	*PAXIP1*	-133
chrX	153178663	153200452	*ARHGAP4*	0

The cut off for increased mutational activity reported for file SRR1020510 was at 2 enhancer events because, in keeping with the trend of low non-coding insertional activity in this file, no genes were affected by more than 2 enhancer events. A total of 7 genes were affected by more than 1 enhancer (Table 7).

Table 7: Genes affected by more than 1 enhancer event in ChIP-seq file SRR1020510.

Chromosome	Gene	Number of Enhancers
chr6	*ENSG00000285064*	2
chr2	*LINC01825*	2
chr21	*MIR155HG*	2
chr3	*RAB7A*	2
chr11	*SLC22A18*	2
chr14	*TRAC*	2
chr19	*TRAPPC2B*	2

The cut off for increased mutational activity reported for files SRR1020512 and SRR1020514 was at 3 enhancer events because of the increased rate of enhancer activity in these files. The genes most affected by multiple enhancer events (3 or more enhancers) were located in files SRR10205012 and SRR10205014 (Table 8).

Table 8: Genes affected by more than 2 enhancer events in ChIP-seq files SRR1020512 and SRR1020514.

SRR1020512			SRR1020514		
Chromosome	Gene	No. of Enhancers	Chromosome	Gene	No. of Enhances
chr6	*ATXN1*	4	chr19	*TRAPPC2B*	5
chr9	*SEMA4D*	4	chr19	*CCDC106*	3
chr5	*CDC42SE2*	3	chr20	*ENSG00000270299*	3
chr17	*CYTH1*	3	chr3	*NUP210*	3
chr20	*GPCPD1*	3			
chr3	*IQSEC1*	3			
chr2	*UBXN4*	3			

The trafficking protein particle complex subunit 2B (*TRAPPC2B*) in chromosome 19 and *GPCPD1* in chromosome 20 showed the most enhancer activity across the DLBCL data, at 8 and 5 enhancers respectively (Table 9). *TRAPPC2B* was also common to each set of enhancer associated genes identified among the 3 SRR files. The trend was observed again whereby chromosomes 1, 2 and 3, known to be large in length and gene rich, not only contributed greatly to the amount of enhancer associated genes identified but also to the genes most affected by enhancer activity (Table 9). Chromosomes 17 and 19, with two of the highest gene counts, also featured in the list of chromosomes housing genes most affected by enhancer activity.

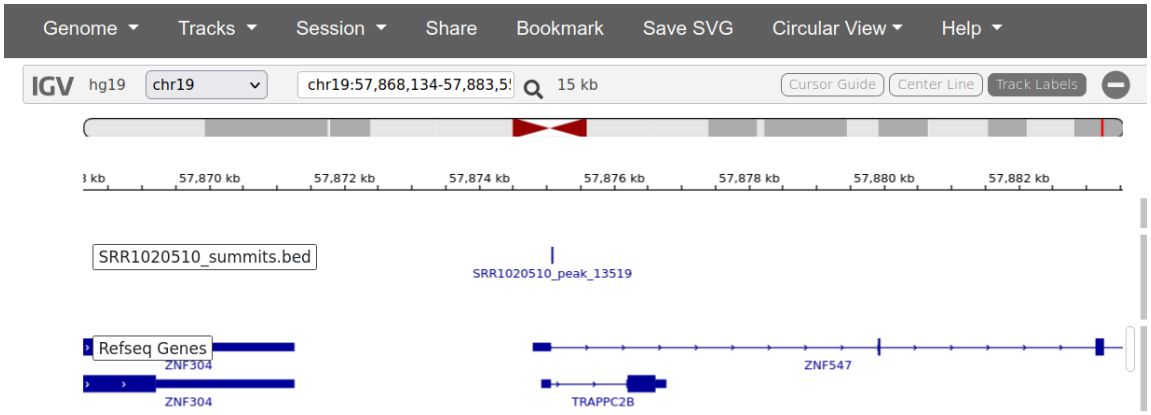
Some chromosomes that were among the smallest in length, such as chromosomes 20 and 21, housed individual genes with multiple enhancers, even though the overall rate of enhancer activity in those chromosomes were relatively low compared to that of other chromosomes. The number of genes impacted as well as the degree of impact was focused upon in this study. The results showed that potentially significant genes affected in DLBCL by non-coding insertion-induced enhancers can be found regardless of chromosome length or gene count.

Table 9: Genes most affected by enhancer activity across all DLBCL ChIP-seq files.

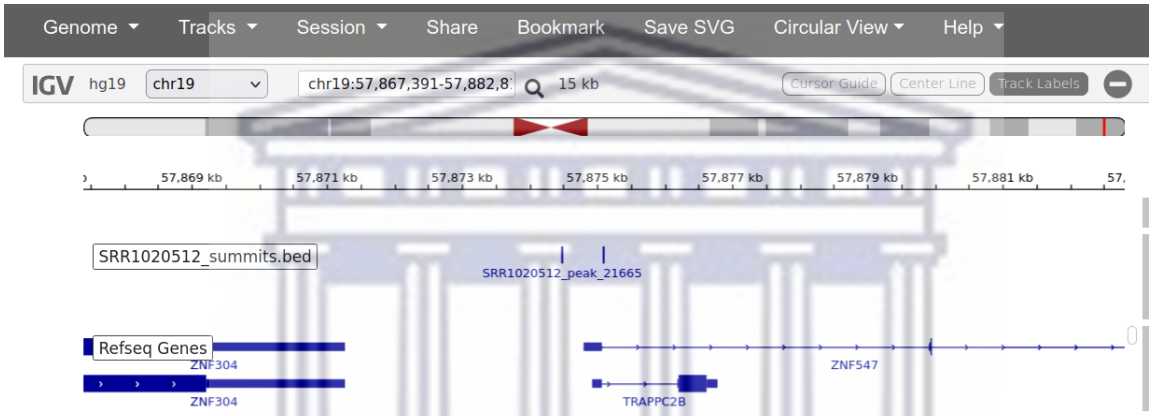
Chromosome	Gene	Total Number of Enhancers
chr19	*TRAPPC2B*	8
chr20	*GPCPD1*	5
chr6	*ATXN1*	4
chr5	*CDC42SE2*	4
chr17	*CYTH1*	4
chr2	*LINC01800*	4
chr9	*SEMA4D*	4
chr14	*TRAC*	4
chr14	*TRAJ35*	4
chr7	*TTYH3*	4
chr16	*USP7-AS1*	4
chrX	*ARHGAP4*	3
chr11	*CADM1*	3
chr19	*CCDC106*	3
chr3	*COPB2*	3
chr21	*DYRK1A*	3
chr2	*ENSG00000225889*	3
chr20	*ENSG00000270299*	3
chr11	*ENSG00000279491*	3
chr6	*ENSG00000285064*	3
chr9	*FNBP1*	3
chr3	*IQSEC1*	3
chr3	*LMCD1*	3
chr21	*MIR155HG*	3
chr17	*MSI2*	3
chr1	*NBPF15*	3
chr18	*NDUFV2*	3

chr6	*NEDD9*	3
chr3	*NUP210*	3
chr11	*SLC22A18*	3
chr4	*SMIM20*	3
chr4	*TNIP3*	3
chr2	*UBXN4*	3
chr19	*ZC3H4*	3

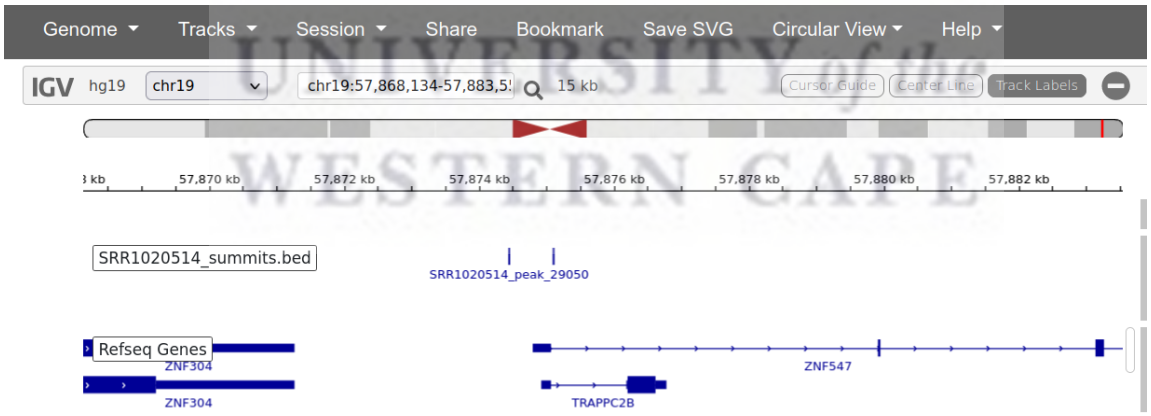
The BED files containing the DLBCL peak summits obtained from MACS2 were visualised using IGV (online application). The peak regions were later intersected with non-coding DLBCL insertions to identify genes associated with the enhancers. The images displayed in Figures 16 and 17 exemplify the potential non-coding insertion-induced enhancer locations of two genes found within DLBCL peak regions.



a) Gene *TRAPPC2B* in file SRR1020510.



b) Gene *TRAPPC2B* in file SRR1020512.



c) Gene *TRAPPC2B* in file SRR1020514.

Figure 16: IGV images of peak summits identified in gene TRAPPC2B of chromosome 19 in each SRR file.

The images displayed show the peak summits indicative of enhancer activity identified in the gene TRAPPC2B of chromosome 19, which was common to all 3 SRR files. The vertical bar above the peak name indicates the position of the peak region in relation to the chromosome as well as the gene to which it is nearest.

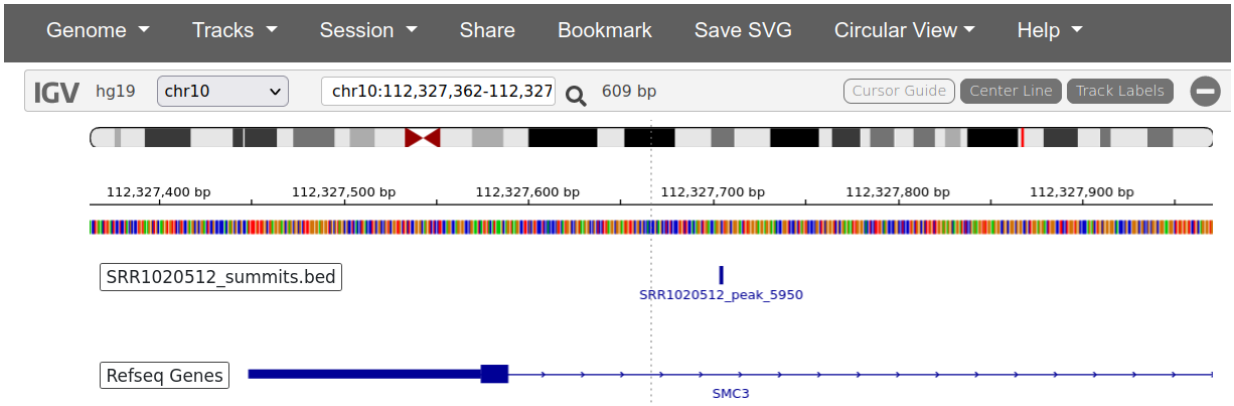


Figure 17: IGV image of peak summit identified in gene SMC3 of chromosome 10 in ChIP-seq file SRR1020512.

The structural maintenance of chromosomes 3 (SMC3) gene is known to play a role in B-cell lymphomagenesis. The investigation revealed the DLBCL peak region to be at 112,327,679 bp in chromosome 10, directly within the SMC3 gene region (112,327,484bp-112,350,844bp).

4.3. FUNCTIONALLY ENRICHED PATHWAYS

Functional analysis was performed using DAVID. Regarding disease analysis, the enhancer associated gene list identified from the SRR file SRR1020510 was not directly linked to any specific type of cancer. The gene list generated from SRR file SRR1020512 was found to contain proto-oncogenes within the Uniprot database and elements associated with somatic prostate cancer in the OMIM database. Some genes produced by SRR file SRR1020514 matched those within the Uniprot database that translated into tumour suppressors.

Scanning of the enhancer associated genes through KEGG databases identified their inclusion among proteoglycans in cancer pathways (Table 10). Proteoglycans produced by tumours aid in their growth, invasion, and maintenance (Elgundi *et al.*, 2020). The investigation discovered several genes affected by DLBCL non-coding insertion-induced enhancers among the heparan sulphate proteoglycans, e.g., AKT, Moesin, and Actin linked to cell growth and survival and SHP-1 linked to angiogenesis. By boosting the affinity of adhesion molecules to their receptors, the heparan sulphate proteoglycans can function as a co-receptor of growth factors and extracellular matrix proteins (Elgundi *et al.*, 2020). They interact with signalling pathways that influence proliferation, adhesion, invasion, and angiogenesis. Enhancer affected genes among the hyaluronan proteoglycans were CDC42, filamin, and F-actin linked to cell migration and invasion. Hyaluronic acid interacts with cell surface receptor CD44 which indirectly activates Rho, MAPK and PI3K signalling cascades to promote cell survival, growth,

proliferation, migration and invasion and transcription of pro-cancer genes (Price, Lokman and Ricciardelli, 2018). Among the keratan sulphate proteoglycans affected by enhancers was Fas, involved in growth suppression. Keratan sulphate is a glycosaminoglycan which bedecks proteoglycan core proteins (Wei *et al.*, 2020). Proteoglycans carrying keratan sulphate epitopes in cancer are highly associated with advanced tumour grade and poor prognosis.

The enhancer associated genes were seen to be involved in B-cell lymphoma transcriptional dysregulation (Table 10). Chemotherapy resistance was linked to gene *CDKN1B*; *Zeb1* was linked to cell migration and invasion; H3 was linked to cell cycle progression. The gene *EWSR1* was associated with processes linked to the escape from growth inhibition, senescence, and apoptosis, tumour growth and survival, proliferation, and angiogenesis.

An enhancer associated gene identified to be affected by non-coding insertions was *NOTCH1*, which formed part of the NOTCH signalling pathway (Table 10). The NOTCH pathway controls cell division, differentiation, proliferation, and death (Bray, 2006). NOTCH is a cell-surface receptor that translates short-range signals by connecting with transmembrane ligands on nearby cells. Multiple pathways in neoplastic B cells cooperate to activate the NOTCH pathway, which is shown by mutations amplifying positive signals or impairing negative regulators.

Table 10: Enhancer associated genes identified by DAVID to be involved in cancer pathways.

Proteoglycans involved in DLBCL	Transcriptional dysregulation	BCR signalling pathway	NOTCH signalling pathway
AKT	*CDKN1B*	*CD22*	*NOTCH1*
Moesin	*Zeb1*	*SHP-1*	*MAML*
CDC42	H3	*SYK*	*APH-1*
Actin	*EWSR1*	*PKCB*	*HATs*
filamin	*ENL*	*VAV*	*Numb*
SHP-1	*LYL1*	*GRB2*	*ATXN1L*
Fas	*PLZF*	*AKT*	
F-actin	*TEL*		

Other enhancer associated genes like *CASP3*, *ITGA*, *PKB*/*Akt*, *PKC*, *PIM1*, *HSP*, *Survivin*, *TRAFs*, *p27*, *Cyclin D*, *VEGF*, *SMC3*, and *Max,* were among those also identified by DAVID to be involved in various cancerous pathways including apoptosis evasion, cell proliferation, tumour invasion and metastasis, genomic instability, insensitivity to anti-growth signals, genomic damage, resistance to chemotherapy, cell immortality and block of cell differentiation.

Genes *CD22*, *SHP-1*, *Syk*, *VAV*, *AKT*, *GRB2*, which form part of the BCR signalling pathway, were detected by the bioinformatics pipeline to be affected by non-coding insertion-induced enhancer activity in DLBCL (Figure 18). B-cell survival, development, and antibody production in both normal and pathological situations depend on signalling via the B cell receptor (Young *et al.*, 2015).

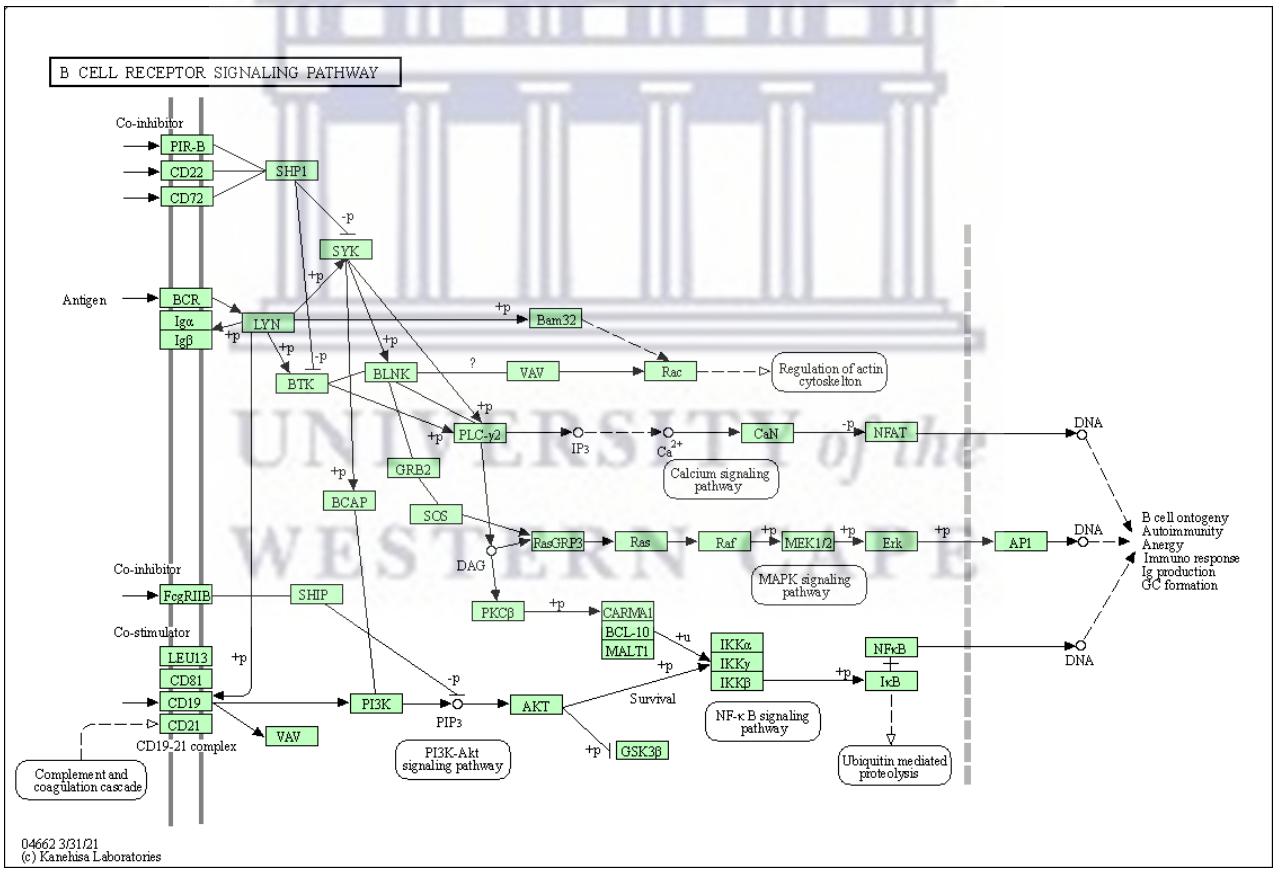


Figure 18: Diagram depicting enhancer associated genes involved in the BCR signalling pathway.

While tonic BCR signalling is necessary for B cell survival and development through molecular mechanisms, chronic active BCR signalling promotes B-cell lymphoma growth (Havranek *et al.*, 2017).

Additionally, there were genes involved in DLBCL pathogenesis identified in this investigation that were shown by DAVID to be involved in HIV-infection, such as *IRAK4*, *STING*, *RIP1*, *TNFR1*, *Fas*, *AP-1*, *CASP3*, *PYK2*, *PKC*, and *AKT*. The viral life cycle of HIV-1 was also found to be impacted by certain DLBCL enhancer associated genes such as *PIN1*, *ELL2*, *INI1*, *CycT1*, *ENL*, *EP300* and *SERINC*.

Disease mechanisms are complex and tend to rely upon the mutated effect of multiple pathways and genes working in unison. Focusing on groups of identified genes that have a similar function, chromosomal location and regulatory role may be beneficial to creating a more inclusive and precise picture of the landscape of DLBCL pathogenicity.

4.4. PIPELINE APPLICABILITY, ADAPTABILITY AND SENSITIVITY

The pipeline structure is easily adaptable depending on the needs of the user. Nextflow's built-in set up for Singularity allowed multiple processes to make use of the same tools without the need for repeated lines of code and made it convenient to change and add software. The container system can be swapped out for Docker or Conda, two of the most frequently used container and environment management systems, for which there are also in-built setups. The modularity of DSL2 allowed tested processes to be reused with different inputs and required outputs which, along with parallel processing, greatly sped up development and increased the robustness of the pipeline. The tools employed by the pipeline also reduced computational run time; pBlat as compared to Blat decreased the run time considerably with parallel threads, and one of the main advantages of using BEDOPS over BEDTools to find enhancer associated genes was that BEDOPS only kept the data needed to compute the next line of output, so memory use was reduced.

The output files must be filtered, like with other variant calling applications, before additional investigation. The necessary input and output file formats of the pipeline are commonly used in variant detection procedures, making it simple to incorporate them in more extensive pipelines. To produce the appropriate custom formats and file associations, complicated processing of the input files was automated.

The pipeline was designed to enable parallel computations for numerous samples, allowing faster analysis of corresponding input and control data sets. Output from each workflow was

standardised for convenient comparison to results from different pipelines, and to enable output viewing with common visualisation tools; the sorted BAM file in which the non-coding insertions were stored can be used for visual verifications, the enhancers and their associated genes were both sorted and stored in BED format which can be redirected or piped directly to other utilities. Depending on their requirements, users can choose to use and adapt one or more sub workflows or their outputs.

Several genes known for their tumorigenic roles in DLBCL and other malignancies, like *FOXP1*, *NOTCH1, IPO11 and PRDM2* were identified by the pipeline to be affected by non-coding insertion-induced enhancers. The genes common to each file was searched for in the dbInDel database (Huang *et al.*, 2020) under cancer type DLBCL, and each gene other than T-cell receptor alpha joining 35 (*TRAJ35*) was located within samples for DLBCL. Genes *TRAPPC2B* and *GPCPD1* with the most aberrations were also found in the dbIndel database. Gene long intergenic non-protein coding RNA 1800 (*LINC01800)* in chromosome 2 was found at a distance of -1641 upstream (towards the 5' end) to the TSS in the dbInDel database, which was close to the distance of -1644 upstream identified in the study, and further supported the accuracy of the bioinformatics pipeline.

4.5. SUMMARY

Using the bioinformatics tools described in the previous chapter, this chapter aimed to present and elucidate upon the DLBCL non-coding insertions, peak regions, and insertion-induced enhancers yielded by the computational pipeline. The aggregation of non-coding insertion mutations and enhancer activity was selective with a direct relationship to chromosomal gene count. Based on the very low rate of mutational activity per chromosome in the biggest DLBCL ChIP-seq sample file and the surge of mutational activity per chromosome in one of the smaller sample files, analysis seemed to suggest the influence of clinical factors specific to the DLBCL ChIP-seq data samples over the research findings. The pipeline produced results that were taken to be accurate based on comparisons to published data. The modularity and parallel processing provided by the programming DSL along with the standardised results optimised the reproducibility and adaptability of the computational pipeline.

CHAPTER 5

DISCUSSION

5.1. INTRODUCTION

Non-coding mutations is an area of great interest in epigenetics. Understanding the functional relevance of these variations is crucial for cancer researchers. Specific transcriptional programs that are characteristic of cancer are dysregulated by genomic abnormalities in DNA regulatory regions. Numerous mutations have been discovered in cis regulatory elements like enhancers. This study provided a robust but flexible computational method of identifying enhancer-associated non-coding mutations in the genome of DLBCL to make it easier for cancer researchers to conduct their studies. The non-coding insertion-induced enhancers and associated genes can be further investigated to determine functional relevance. Important mutations might be found consequently, and the development of innovative therapies that target the non-coding genome might be aided.

5.2. CHROMOSOMAL IMPACT ON DLBCL ENHANCER ACTIVITY

It was assumed that the number of non-coding insertion-induced enhancers would increase with chromosome length, which was the general trend observed upon analysis of the ChIP-seq data (Figure 10). There were, however, exceptions to this trend, most notably in chromosomes 17 and 19, in which more enhancer activity was observed than expected, and chromosomes 13, in which less enhancer activity was observed than expected. Chromosome 19 had the second highest non-coding insertion-induced enhancer rate of all somatic chromosomes, rivalling that of chromosome 1 which is almost 4 times its length (Figures 9 and 10). Chromosome 17 had the fourth highest enhancer rate, rivalling that of chromosome 6 which is almost double its length. Conversely, chromosome 13 had the lowest enhancer rate, even lower than that of chromosome 21 which is the smallest somatic chromosome (Figures 9 and 10). While there was a positive correlation between the rate of non-coding insertion-induced enhancers and chromosome length (Table 3), the strength of the relationship between the variables was weaker than expected at a barely moderate correlation coefficient of 0.56 for the total DLBCL ChIP-seq data (Figure 11), which was likely influenced by the discrepancies in chromosomes 13, 17, and 19.

The rate of enhancers was expected to increase with chromosome length because it was assumed that the number of genes (targets for mutational activity) would increase the larger the chromosome. Since chromosome length alone was unable to satisfactorily justify the observed trend in the non-coding insertion mutational pattern, the enhancer rate was measured according to chromosomal gene count without making assumptions about chromosome length.

The rate of enhancers was found to increase with the number of genes in each chromosome (Figure 12). A strong, positive association was found between enhancer rate and gene count at a correlation coefficient of 0.86 for the total DLBCL ChIP-seq data (Figure 13), which was more significant than what was observed between enhancer rate and chromosome length. Enhancer activity in relation to chromosomal gene count and density was thus able to elucidate upon the discrepancies observed in chromosomes 13, 17, and 19.

The present study identified the second highest rate of insertion-induced enhancers within chromosome 19 which was attributed to the fact that it had the second highest number of genes among all chromosomes (Figure 12), with a gene density more than two times the genome average (Grimwood *et al.*, 2004). Chromosome 19 has large, clustered gene families with a high GC content, many CpG islands, and is packed with repetitive DNA (55% vs. the genome average of 44.8%), which is significant in terms of both biology and evolution (Grimwood *et al.*, 2004). Studies on lung cancer subtypes found chromosome 19 contained the most frequently occurring aberrations (Jinesh *et al.*, 2021). The SPIB locus on chromosome 19 was also found to be a target of mutations in ABC DLBCL (Pasqualucci and Dalla-Favera, 2018).

Chromosome 17 is the fifth smallest somatic chromosome in length, yet it houses the fifth highest number of genes, along with chromosome 12 (Figure 12). Chromosome 17 contains the second-highest gene density and is enriched in segmental duplications (Zody *et al.*, 2006). Numerous studies have revealed a connection between chromosome 17 genes and the development, progression, and response to cancer treatment, especially breast cancer (Zody *et al.*, 2006). Treatment development might be made easier with a better knowledge of chromosome 17 anomalies since it houses several well-known tumour suppressors, like *TAU* and *TOP2A*, and oncogenes, like *p53* and *BRCA1* (Zhang and Yu, 2011).

The upscale observed in the number of insertion-induced enhancers found in chromosomes 17 and 19 was therefore consistent with the literature on what is known about the chromosomes regarding gene density. On the other hand, although chromosome 13 has a length that is average (Figure 10), it has the third lowest number of genes among somatic chromosomes, and the second lowest gene density, after chromosome 18 (Figure 12) (Dunham *et al.*, 2004). When taken in this context, the low enhancer rate in this chromosome was in keeping with the trend in which enhancer activity increased along with the number of genes and vice versa.

The non-coding insertion mutations were not random but selective; they targeted chromosomes that were gene rich. Chromosomes that are gene rich have an increased GC content, as mentioned with chromosome 19, which serve as structural markers for transcription (Han and Zhao, 2009). GC-rich regions are home to CpG islands, clusters of CpG dinucleotides that contribute to approximately 30% of the genome-wide variability in indel rates (Makova and Hardison, 2015). An increased indel rate is found at high GC content because of the increased frequency of CpG nucleotides, which become mutation hotspots when methylated and therefore have higher mutation rates (Makova and Hardison, 2015). Furthermore, while the DLBCL insertions were non-coding mutations, they induced enhancers close to or within the TSS of the genes nearest to them, and increased GC content of sequences is known to be found at and around the TSS of genes (Koudritsky and Domany, 2008). Higher GC content equates to stronger binding of DNA regulatory elements, and possibly a higher density of binding sites (Koudritsky and Domany, 2008).

Further investigation was made into the underlying mechanisms behind the non-coding insertions targeting regions of the genome rich in GC content. GC rich genes have 100-fold greater transcription rates than GC poor genes (Khuu *et al.*, 2007). High GC content has been correlated with elevated levels of gene transcription, whereas low GC content has been correlated with chromatin condensation (Khuu *et al.*, 2007).

5.3. IMPACT OF CHROMATIN ACCESSIBILITY ON DLBCL ENHANCER ACTIVITY

The distribution pattern of the non-coding insertion mutations could be explained by a particular chromatin conformation. Chromatin accessibility is a hallmark of transcription factor binding and regulatory elements like enhancers (Ocsenas and Reimand, 2022). The mutations aggregated to regions that were rich in genes, high in GC content, and therefore likely to be more accessible and less compact, collectively such genomic regions are referred to as the euchromatin (Gilbert *et al.*, 2004). Regional mutation rates are highly influenced by chromatin organization; greater indel mutation densities are seen in early-replicating, transcriptionally active areas of open chromatin (euchromatin) (Makova and Hardison, 2015). Accordingly, low non-coding insertion-induced enhancer rates were observed in chromosomes that were gene poor, which are typically the locations of constitutive heterochromatin that maintains a condensed and transcriptionally inert chromatin conformation (Marsano and Dimitri, 2022).

The DNA regulators that majority of the non-coding DLBCL insertions were associated with were putative proximal enhancers located within the TSS of the genes nearest to them. The enhancers induced by the non-coding DLBCL insertions may have been exposed to less downstream regulation if the chromatin was in an active state since this is what has been observed by previous cancer studies on response to gene regulation at proximal enhancers in rapidly dividing cells (Sanghi *et al.*, 2021).

A slight increase in enhancer activity was observed in chromosome 21 (Figure 12), which is gene poor and smallest in length of all somatic chromosomes. It is the location of known DLBCL oncogenes *DYRK1A*, *PRDM15*, and *MIR155HG* found to be affected by non-coding insertion-induced enhancers. Previous research identified the presence of particularly fragile regions of the chromatin in chromosomes that contain oncogenes and tumour suppressor genes, e.g., a region with one of the highest accessibility scores contains the *MYC* gene, a well-known oncogene in DLBCL; another susceptible region contains the *PRDM1* gene, a master regulator of pan-immune response (Liu, 2020). Fragile regions comparable to those mentioned may have made chromosome 21 slightly more susceptible to the non-coding insertion mutations, due to the oncogenes it contains which possibly have high accessibility scores.

Majority of the detected enhancers were found among the somatic chromosomes, enhancers found among the sex chromosomes came solely from chromosome X (Figure 15). Chromosome X at 5.10cm (Piovesan *et al.*, 2019) is similar in length to chromosome 7 (Figure 10), with approximately 1400 genes (MedlinePlus: Chromosomes & mtDNA, 2021), it ties with chromosome 11 for third most gene rich chromosome. However, it had an enhancer rate closer to that observed in chromosome 22 (Figure 14), which is less than half its length (Figure 10). It is understood from published literature that indels target regions of the genome that is less compact and accessible to transcription (Makova and Hardison, 2015), the silent X chromosome in women is an example of facultative heterochromatin (Wutz, 2011), which describes a condensed and inactive environment not observed to be targeted by the non-coding insertions.

Chromatin accessibility has important clinical implications in cancer and the data gained by this study provided an additional perspective in mutational targets of DLBCL tumours.

5.4. IMPACT OF CLINICAL FACTORS ON ENHANCER ACTIVITY

There were significant differences in the quantity of non-coding insertion-induced enhancers detected among the DLBCL ChIP-seq data files. The mutational rate observed in the largest ChIP-seq file, SRR1020510, was outstandingly low across all chromosomes. The mutational rate observed in the smaller ChIP-seq file, SRR1020512, was exceedingly high in comparison, while the mutational rate of ChIP-seq file SRR1020514 averaged between that of the other two files, higher than the former but lower than the latter. It was hypothesized that clinical factors defining the sources of the DLBCL samples might have played a role in these observations.

Most patients are at an advanced disease stage when DLBCL is diagnosed, with poor prognosis, numerous extranodal involvement and a high percentage of the double expressor subtype (Zhu *et al.*, 2022). The ChIP-seq data files reflected mutational rates from 3 DLBCL patients in potentially varying stages of disease (Figure 14). The data in file SRR1020510 described a patient in an early stage of disease development, the cancer genome would still have been close to a non-diseased state, and so very low numbers of non-coding insertion mutations were observed. The data in file SRR1020512 described a patient in an advanced stage of the disease with an acute clinical presentation which would have been facilitated by

multiple dysregulated genes and pathways working in unison, hence the particularly high rate of mutations. Lastly, the data in file SRR1020514 described a patient in a disease stage that had progressed further than that of the first patient, thus an increased rate in mutations was found in comparison but was still not as advanced as the second patient, thus fewer mutations were detected in comparison.

Despite the difference in mutational rates, enhancer associated genes were identified by the bioinformatics pipeline among all the DLBCL ChIP-seq files. Several were known to play key roles in various cancer types, including lymphomagenesis, and more specifically, DLBCL.

5.5. NON-CODING INSERTION-INDUCED ENHANCER ASSOCIATED GENES

This section provided insight on a selection of enhancer associated genes identified by the bioinformatics pipeline. Among the genes discussed were some of those common to each DLBCL file of ChIP-seq data as well as some of those most affected by enhancer activity and known from literature to be involved in cancer and, more specifically, DLBCL (*SEMA4D*, *PRDM15*, *MIR155HG*, *PRDM2*, *UBX4N*, *IPO11*, *EZH1*, *FOXP1*, *NOTCH1* and *PIM1*). Previous studies found that enhancers located within genes were predictive of correlated RNA and protein expression (Sanghi *et al.*, 2021). The non-coding insertion-induced enhancers therefore likely had an impact on the expression of the genes within which they were located. A description of the enhancer associated genes' roles in tumorigenesis was provided to substantiate their significance as prime featural candidates for research and further validate the bioinformatics pipeline through which they were identified.

5.5.1. ENHANCER ASSOCIATED GENES COMMON TO EACH DLBCL DATA FILE

The genes *COPB2*, *LINC01800*, *TRAPPC2B* and *DYRK1A* were among the genes commonly found in each ChIP-seq file (Table 5) and most affected by non-coding insertion-induced enhancer activity (Table 9).

Several studies support the relevance *COPB2* as a known oncogene, potential therapeutic target, and biomarker (Feng *et al.*, 2021). Breast cancer tissue is dependent on the overexpression of *COPB2* gene (Bhandari *et al.*, 2019). Additionally, CRC cell proliferation and development are significantly influenced by *COPB2* and may be inhibited by *COPB2* silencing (Feng *et al.*, 2021). *COPB2* was found to encourage the growth of lung cancer cells through

YAP1, which is a gene highly expressed in DLBCL (Feng *et al.*, 2021). The *COPB2* gene was affected by 3 DLBCL enhancers, the same number of enhancers that affected the *DYRK1A* gene (Table 9), which is thought to be both a tumour suppressor and an oncogene (Hurtz *et al.*, 2021). *DYRK1A* is downregulated in cancers of the colon, oesophagus, kidney, liver, stomach, thyroid, and uterus (Rammohan *et al.*, 2022). However, it is upregulated in glioblastoma multiforme, lung cancer, and pancreatic ductal adenocarcinoma (Rammohan *et al.*, 2022). A study found that by inhibiting *DYRK1A*, cancer cells become sensitive to *BCL2* inhibition through the hyperactivation and hyperphosphorylation of *MYC* and *ERK* (Hurtz *et al.*, 2021). *BCL2*, is an overexpressed gene linked to poor prognosis in DLBCL. *DYRK1A* inhibition could enhance the effect of *BCL2* inhibitors currently available for DLBCL.

The gene *LINC01800*, affected by 4 DLBCL enhancers, is part of the lncRNA class which plays a significant role in controlling oncogenic genes and signalling pathways in DLBCL through epigenetic regulatory mechanisms (Huang, Qian and Ye, 2020). LncRNAs with high specificity and accuracy, like *HOTAIR* and *MALAT-1* which play critical prognostic roles in DLBCL (Huang, Qian and Ye, 2020), are excellent candidates for use as biomarkers or therapeutic targets due to the expression patterns they exhibit (Karstensen *et al.*, 2021).

The *TRAPPC2B* gene was the most affected by non-coding insertions, with 8 resident DLBCL enhancers. It is located within chromosome 19, the chromosome with the second highest number of genes and correspondingly, the chromosome with the second highest enhancer rate. This gene is involved in transcriptional repression and induction of cell death (UniProt, 2022). Its role in cancer, and furthermore in DLBCL, has yet to be properly explored.

5.5.2. SELECTION OF ENHANCER ASSOCIATED GENES FOUND THROUGHOUT THE STUDY

The genes UBX domain protein 4 (*UBXN4)* and semaphorin 4D (*SEMA4D*) were among those most affected by non-coding insertion-induced enhancers (Table 9). *UBXN4* was found to be downregulated in the EZB DBCL genetic subtype and unclassified DLBCL cases by analysis for *MYC*/*BCL2* double-high expression (Derenzini *et al.*, 2021). *SEMA4D*, located in chromosome 19, is commonly dysregulated in cancer, and linked to invasive characteristics and a poor prognosis (Ch'ng and Kumanogoh, 2010). It is also a well-known immune regulator, supporting the significance of dysregulated *SEMA4D* in cancer cells' immunological evasion (Li *et al.*, 2018).

Enhancer activity in the gene *SMC3* was identified in both the SRR1020512 and SRR1020514 ChIP-seq files (Table 6). In GC-derived DLBCL patients, *SMC3* haploinsufficiency is known to aid the cancerous transformation of GC B-cells by disrupting connectivity of enhancers regulating tumour suppressor genes and induces lymphomagenesis with increased expression of *BCL6* (Rivas *et al.*, 2021). Loss of *SMC3* has been linked to decreased gene stability which leads to poor prognosis and a lower survival rate in DLBCL patients (Rivas *et al.*, 2021).

The MIR155 host gene (*MIR155HG*) in chromosome 21 was one of the genes most affected by enhancer activity in the SRR1020510 ChIP-seq file (Table 7). Resistance to the antimitotic drug vincristine, which is crucial to the efficacy of the multiagent chemotherapy regimen R-CHOP, is thought to be brought on by *MIR155HG* suppression and deletion (Due *et al.*, 2019). A clinical cohort of DLBCL patients who received R-CHOP treatment showed improved survival for the GCB subtype when *MIR155HG* expression levels were high. Overexpression of the PR/SET domain 15 (*PRDM15*) gene, which was also affected by enhancer activity and located in chromosome 21, fuels B-cell lymphomagenesis (Mzoughi *et al.*, 2020). It was found that by genetically reducing *PRDM15* levels, B-cell lymphoma lines were killed both in vitro and in vivo. *PRDM15* regulates transcriptional programs that maintains NOTCH signalling-related genes (Mzoughi *et al.*, 2020).

Genes *PRDM2, IPO11, PIM1, FOXP1*, and *EZH1* are housed in some of the largest, most gene rich chromosomes (1, 5, 6, 3, and 17 respectively) and were identified to be affected by non-coding insertion-induced enhancers. The genes are linked to tumorigenesis, genes *FOXP1* and *PIM1* specifically are relevant to ABC DLBCL while gene enhancer of zeste 1 polycomb repressive complex 2 subunit (*EZH1*) is relevant to GCB DLBCL.

In ABC DLBCL, the *FOXP1* gene is highly expressed and regulates pathways that suppress apoptosis and GCB-cell identity, and influence plasmablast identity and NF-κB signalling (Gascoyne and Banham, 2017). It is associated with poor outcomes and therapeutic resistance (Gascoyne and Banham, 2017). Furthermore, *PIM1* mutations in ABC DLBCL decrease susceptibility to ibrutinib, a BTK inhibitor, by stabilising the protein and improving NF-κB signalling (Kuo *et al.*, 2016). Research indicates that ibrutinib coupled with pan-PIM inhibitors may overcome treatment resistance in DLBCL (Szydłowski *et al.*, 2021). *PIM1* mutations have

been used to categorise a distinct group of DLBCL, central nerve system diffuse large B-cell lymphoma (CNS DLBCL) (Zhou *et al.*, 2022). DLBCL patients with *PIM1*-mutant conditions have adverse characteristics such as advanced stage, non-GCB, and poor survival (Kuo *et al.*, 2016). Both *PIM1* and *FOXP1* were identified exclusively in the ChIP-seq file SRR1020512, in which the most enhancer activity was identified which was suspected to be due to the patient having been in an advanced disease stage (Figure 14). Both genes are associated with ABC DLBCL, the more aggressive of the two DLBCL COO subtypes, and furthermore, *PIM1* is linked to the MCD genetic subtype of DLBCL which is associated with one of the worst prognoses.

EZH1, or its close homolog *EZH2,* in chromosome 17 acts as a catalytic subunit that inhibits gene expression (Wassef *et al.*, 2019). *EZH1*/2 dual inhibitors demonstrate anticancer efficacy in vitro and in vivo against DLBCL cells carrying EZH2 gain-of-function mutations. About 22% of GCB-DLBCL show *EZH2* gene mutations that are not found in ABC subtype (Honma *et al.*, 2017). *EZH2* is associated with the EZB genetic subtype that has the worst prognosis in GCB DLBCL (Honma *et al.*, 2017). *EZH1* was found only within ChIP-seq file SRR1020514, which is suspected to contain DLBCL data from a patient with a moderate clinical manifestation (Figure 14).

The gene PR/SET domain 2 (*PRDM2*) in chromosome 1 is commonly deleted or altered in cancer, *PRDM2* deficiency has been found to lead to DLBCL development in mice (Xia *et al.*, 2017). Gene importin 11 (*IPO11*) in chromosome 5 is the transport receptor of PTEN, upon which tumour suppression in DLBCL is dependent (Chen *et al.*, 2017). PTEN degradation is constrained by the *IPO11* cargo UBE2E1 (Chen *et al.*, 2017). Loss of *IPO11* was found to lead to PTEN degradation in lung cancer.

Functional enrichment analysis summarised the enhancer associated genes identified by the bioinformatics pipeline and linked them to specific biological processes pertinent to DLBCL. Associating the individual genes with biological terms was done to better appreciate the complex nature of DLBCL biological processes.

5.6. SIGNALLING PATHWAYS IMPACTED BY DLBCL ENHANCERS

The genes *CD22*, *Syk* and *SHP-1* (Figure 18) were affected by non-coding insertion-induced enhancers and are involved in BCR signalling. BCR signalling with mutated CD79B, to which ABC DLBCL is addicted (Havranek *et al.*, 2017), is ill suited to activate Lyn kinase which works to inhibit BCR signalling through CD22 phosphorylation and recruitment of the phosphatase SHP-1, augmenting chronically active BCR signalling (Young *et al.*, 2015).

A subgroup of GCB DLBCL is reliant on induction of the PI3K pathway through *Syk*, indicating the relevance of chronic BCR signalling in GCB DLBCL (Young *et al.*, 2015). Research has suggested the activation of the BCR pathway in GCB DLBCL is induced by a phosphatase SHP-1 deficiency (Sasi *et al.*, 2018). *SHP-1*, a negative regulator of the BCR pathway, is downregulated in 40% of primary DLBCL tumours and was highlighted as a predictive marker for therapeutic response to a venetoclax/BCR inhibitor combination (Sasi *et al.*, 2018). The clinical success of drugs that target the BCR pathway emphasizes the significance of knowing the metabolic workings of BCR signalling in DLBCL.

Non-coding insertion-induced enhancers were identified within the NOTCH signalling pathway, most notably within the *NOTCH1* gene. *NOTCH1* mutations in DLBCL are suggested to be oncogenic because they are linked to poor prognosis and survival in patients (Fabbri *et al.*, 2011). *NOTCH1* activation defines the N1 DLBCL genetic subtype, which has the worst prognosis along with the MCD genetic subtype (Kotlov *et al.*, 2021), and promotes tumour development that evades the host immune system (Shanmugam *et al.*, 2021). *NOTCH1* mutations were only found within ChIP-seq file SRR1020512, which was suspected to define a patient with severe clinical manifestations based on the rate of enhancer activity (Figure 14). *NOTCH1* mutations have predictive value as they were found to be adversely linked with full remission of patients treated with R-CHOP chemotherapy (Li *et al.*, 2021). In other studies, chronic lymphocytic leukaemia susceptible to Richter transformation into DLBCL was shown to have much higher rates of *NOTCH1* activation (Fabbri *et al.*, 2011). Medications are being developed that target *NOTCH1* and some are already on the market, such as those that prevent its enzymatic conversion to an active transcription factor (Li *et al.*, 2021) .

Functional analysis supported the role of the identified genes in processes involved with DLBCL proliferation, metastasis, apoptosis evasion, and resistance to therapy. The bioinformatics pipeline also identified genes that played a role in both DLBCL tumorigenesis and HIV pathogenesis. Clinical studies related to lymphoma typically exclude PLWH. The pipeline should be employed using samples of HIV DLBCL ChIP-seq data, the insertion-induced enhancers identified can then undergo comparative studies with those identified in DLBCL samples without the HIV comorbidity. The inclusion of HIV-DLBCL may increase our understanding of DLBCL induction given the inherent link between lymphoma and HIV.

5.7. VALIDITY OF NON-CODING INSERTION-INDUCED ENHANCERS

The bioinformatics pipeline results were validated by the mutational trend the non-coding insertions followed where they selectively accumulated in genomic regions that were gene rich and would therefore be high in GC content indicating elevated levels of transcription and open chromatin, which was supported by literature describing indels' euchromatic regional preferences. The mutated enhancers were found to be largely proximal, within or very close to the TSS of their associated genes. Since proximal genes are known to be subject to less downstream regulation, the chance of their dysregulated effect going unchecked was increased, enhancing a diseased genomic state. Furthermore, many of the enhancer associated genes were known oncogenes or tumour suppressor genes involved in numerous oncological processes.

As an extra validative step, some of the non-coding insertion-induced enhancers were searched for in the dbInDel database, which curates non-coding somatic indels and their associated cis-regulatory elements in human malignancies (Huang *et al.*, 2020). Majority of the enhancer associated genes identified by the present study were found in the database under cancer type DLBCL, the similarity in TSS between the same genes further supported the validity of the enhancers identified. An exception was *TRAJ35*, a non-functional gene found within the present study but not within the dbInDel database. *TRAJ35* was found to be affected by DLBCL enhancers in each ChIP-seq file (Table 5) and was also one of the genes most impacted by enhancers (Table 9). The effect of mutated *TRAJ35* in DLBCL will require further investigation, although present information suggests that its protein product is not prognostic in cancer (TRAJ35 protein expression summary- The Human Protein Atlas, 2022).

Enhancers were expected to be found within *MYC*, *BCL6*, *CREBBP, MLL2* and *BCL2,* some of the most significantly mutated genes in DLBCL (Evrard *et al.*, 2019). However, these genes were not identified by the bioinformatics pipeline to be affected by non-coding insertion-induced enhancers although their presence in the dbInDel database indicates insertions are present in enhancer regions within these genes. Similarly, enhancer activity was not detected by the study in the Y chromosome, but that did not mean there were none to be found. The *P2RY8* gene located in the Y chromosome is frequently mutated in GCB-DLBCL (Lau, 2020). It may be that the sample of ChIP-seq data analysed in the study was too small to provide a representative selection of genes involved in DLBCL tumorigenesis.

According to Huang *et al.* (2020) some insertions and deletions connected with specific enhancers are found in known cancer type-specific drivers, e.g., the prostate cancer oncogene *AR* is shown to exhibit enhancer-associated insertions and deletions only in samples of prostate cancer. The fact that non-coding insertion-induced enhancers associated with non-DLBCL cancer type-specific drivers were not detected by the bioinformatics pipeline supports the idea that some enhancer-associated insertions are under selective pressure and offer growth advantages to particular cell types due to the properties of their targeted genes.

5.8. BIOINFORMATICS PIPELINE MECHANICS

Studies on non-coding mutations such as insertions in enhancer elements have been conducted in the past with promising results. The present study designed a bioinformatics pipeline based on a source script written in shell (Abraham *et al.*, 2017). However, some of the methods used were outdated. Abraham *et al.* (2017) used Blat to verify the insertion sequences. Blat is single threaded and can take days to finish when used to map whole genome sequences to reference genomes. The present study used pBlat instead, a parallelized blat algorithm with multithread and cluster computing support which reduces the run time. It uses the same amount of memory and generates the same results as Blat. Abraham *et al.* (2017) also used MACS for peak calling whereas as the present study used the latest version of the tool, MACS2. The underlying algorithm for peak calling is the same but MACS2 comes with enhancements in the form of twelve functions serving as sub-commands. The main function callpeaks was used to identify DLBCL peak regions from alignment results.

Lastly, Abraham *et al.* (2017) used SAMtools to sort and index data files, however this involved a two-step process with two separate commands, one for sorting and one for indexing. The present study used sambamba instead, which has dual functionality in that it sorts and generates an index for a file in one step.

Furthermore, the workflows by Abraham *et al.* (2017) were limited in terms of their management systems. The computational procedures were not assembled in an exact pipeline, tasks were compiled in single workflow scripts that had to be run individually within the same directory and required the user to wait for completion of one script before commencing the next. The present study made use of Nextflow as a workflow framework. Nextflow allowed sub workflows to be run in a single step as part of a main workflow. DSL2 provided parallel processing, and modularity, which is the ability to define reusable processes or sub workflows that can be included and invoked as a function from another script within a separate workflow. DSL2 also allowed for fewer lines of coding, making the scripts neater and more succinct. By simply specifying it in the nextflow.config file, the workflow was executed using SLURM. The processes were linked through channels and were also isolated, which made it easy to exchange tools and manage shell script problems that can be difficult to trace, or that stem from missing dependencies and lack of resources.

Many intermediary files are generated when using shell that must be manually removed after each script has run. Intermediary files were managed by Nextflow and automatically stored in a separate 'work' directory. The publishDir directive allowed process output files to be exported to a specified folder that could be manually accessed, allowing the user to selectively view the results of each process. In this way, debugging was made easier.

With shell, a link must be provided for each data file used in the script, if another user wants to make use of the same script, these links must be removed, and new links provided separately. Nextflow allowed the specification of a single link to multiple files which greatly eased reproducibility. Additionally, Singularity is difficult to integrate into shell, Singularity images must be imported individually, whereas Nextflow has built in support that renders importation of Singularity images a single step process.

Whereas as studies like that of Bal *et al.* (2022) explored the effect of structural variants and focused on known DLBCL oncogenes and how they were connected to identified enhancers, the present study sought to identify novel enhancers formed by insertions through a pipeline that can be manipulated to highlight genes of interest across cancer types. The investigation built off previous published works and designed a pipeline using state of the art infrastructure to setup, execute, and monitor computational workflows, incorporating the latest bioinformatics tools to achieve accurate and reproduceable results.

5.9. STUDY LIMITATIONS

This study made use of only three SRR files of DLBCL H3K27ac ChIP-seq data, and all were from the same research study. The pipeline was not tested on a larger sample size from a combination of different studies. The present study was tested on single reads generated by Illumina sequencing. In a scenario where data from different studies are being tested, the reads might be single or paired and have been generated by different sequencing platforms which have varying parameters for reporting errors and variants. Therefore, there may be a need for troubleshooting in such circumstances. Such expansion would, however, likely lead to better representation and increased validity of identified non-coding insertion-induced enhancers. The sub workflow for peak calling was designed to operate on only treatment ChIP-seq data with corresponding control data. However, some ChIP-seq treatment data do not come with corresponding controls, yet they may still hold valuable biological information, and so provision should be made for the accommodation of samples such as these. Due to time constraints, it was not possible to recreate the perl scripts used in this study as python scripts. Python scripting may be desirable over perl scripting because of its extensive library support and basic syntax that requires fewer lines of coding for larger programs.

CHAPTER 6

CONCLUSION

This study set out to uncover potential non-coding insertion-induced enhancers associated with the progression of DLBCL to develop a database that future studies on DLBCL can use in South Africa. The pipeline detected several enhancer-associated genes known for their role in DLBCL tumorigenesis and other cancer types, as well as genes jointly involved in lymphomagenesis and HIV, between which there is an innate association. The identification of known oncogenes and tumour suppressor genes indicated the accuracy of the pipeline and encouraged confidence in those genes identified for which little to no data has been recorded as potential new targets for research on the different disease mechanisms of DLBCL with the goal of functional precision therapy.

The study indicated the necessity of considering clinical factors when selecting DLBCL data to be analysed due to the hypothesized but unconfirmed influence that they may have on mutational rates. Sample heterogeneity should be considered to optimize the algorithm. Future studies on DLBCL should categorize their data according to criteria like age, gender, disease subtype, disease stage, and other patient dependent information to confirm and track associated patterns. A greater selection of data should also be incorporated to verify whether the data trends hold true and to gain a better picture of the effect of insertion-induced enhancers on the genomic landscape.

The research supported the growing data on the impact of the non-coding environment on gene expression and disease development. The promising results suggested that it might be worthwhile to expand the investigation into deletions in the non-coding regions and its effect on DLBCL. Nextflow might be one of the most developed workflow management systems to date, it is a complete system combining workflow language and execution engine. The coding involved was simple and the workflow framework provided for desirable properties like readability, compactness, portability, and provenance tracking.

REFERENCES

Abraham, B.J. *et al.* (2017) 'Corrigendum: Small genomic insertions form enhancers that misregulate oncogenes', *Nature communications*, 8, p. 15797. Available at: https://doi.org/10.1038/ncomms15797.

Ahmed, A.E. *et al.* (2021) 'Design considerations for workflow management systems use in production genomics research and the clinic', *Scientific Reports 2021 11:1*, 11(1), pp. 1–18. Available at: https://doi.org/10.1038/s41598-021-99288-8.

Aho, A. v, Kernighan, B.W. and Weinberger, P.J. (1978) 'Awk A Pattern Scanning and Processing Language (Second Edition)'.

Amen, F. *et al.* (2007) 'Absence of cyclin-D2 and Bcl-2 expression within the germinal centre type of diffuse large B-cell lymphoma identifies a very good prognostic subgroup of patients', *Histopathology*, 51(1), pp. 70–79. Available at: https://doi.org/10.1111/j.1365-2559.2007.02721.x.

AVERT (2020) *HIV and AIDS in South Africa | Avert*, *HIV and AIDS in South Africa*. Available at: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa (Accessed: 19 May 2020).

Baichoo, S. *et al.* (2018) 'Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics', *BMC Bioinformatics*, 19(1). Available at: https://doi.org/10.1186/S12859-018-2446-1.

Bakhshi, T.J. and Georgel, P.T. (2020) 'Genetic and epigenetic determinants of diffuse large B-cell lymphoma', *Bakhshi and Georgel Blood Cancer Journal*, 10, p. 123. Available at: https://doi.org/10.1038/s41408-020-00389-w.

Bal, E. *et al.* (2022) 'Super-enhancer hypermutation alters oncogene expression in B cell lymphoma', *Nature*, 607(7920), pp. 808–815. Available at: https://doi.org/10.1038/S41586-022-04906-8.

Beham-Schmid, C. (2017) 'Aggressive lymphoma 2016: revision of the WHO classification', *Memo*, 10(4), p. 248. Available at: https://doi.org/10.1007/S12254-017-0367-8.

Bennett, E.P. *et al.* (2021) 'INDEL detection, the "Achilles heel" of precise genome editing: A survey of methods for accurate profiling of gene editing induced indels', *Nucleic Acids Research*, 48(21), pp. 11958–11981. Available at: https://doi.org/10.1093/nar/gkaa975.

Bhagwat, M., Young, L. and Robison, R.R. (2012) 'Using BLAT to find sequence similarity in closely related genomes', *Current Protocols in Bioinformatics*, 0 10(SUPPL.37), p. Unit10.8. Available at: https://doi.org/10.1002/0471250953.bi1008s37.

Bhandari, A. *et al.* (2019) 'COPB2 is up-regulated in breast cancer and plays a vital role in the metastasis via N-cadherin and Vimentin', *Journal of Cellular and Molecular Medicine*, 23(8), p. 5235. Available at: https://doi.org/10.1111/JCMM.14398.

Boland, C. (2017) 'Non-coding RNA: It's Not Junk', Digestive Diseases and Sciences. Springer New York LLC, pp. 1107–1109. Available at: https://doi.org/10.1007/s10620-017-4506-1.

Bray, S.J. (2006) 'Notch signalling: a simple pathway becomes complex', *Nature Reviews Molecular Cell Biology 2006 7:9*, 7(9), pp. 678–689. Available at: https://doi.org/10.1038/nrm2009.

Can, T. (2014) 'Introduction to bioinformatics', *Methods in Molecular Biology*, 1107, pp. 51–71. Available at: https://doi.org/10.1007/978-1-62703-748-8_4.

de Carvalho, P.S., Leal, F.E. and Soares, M.A. (2021) 'Clinical and Molecular Properties of Human Immunodeficiency Virus-Related Diffuse Large B-Cell Lymphoma', *Frontiers in Oncology*, 11, p. 1550. Available at: https://doi.org/10.3389/FONC.2021.675353/BIBTEX.

Cassim, S. *et al.* (2020) 'Diffuse large B-cell lymphoma in a South African cohort with a high HIV prevalence: an analysis by cell-of-origin, Epstein–Barr virus infection and survival', *Pathology*, 52(4), p. 453. Available at: https://doi.org/10.1016/J.PATHOL.2020.02.007.

Chadburn, A. *et al.* (2009) 'Immunophenotypic Analysis of AIDS-Related Diffuse Large B-Cell Lymphoma and Clinical Implications in Patients From AIDS Malignancies Consortium Clinical Trials 010 and 034', *Journal of Clinical Oncology*, 27(30), p. 5039. Available at: https://doi.org/10.1200/JCO.2008.20.5450.

Chapuy, B. *et al.* (2013) 'Discovery and Characterization of Super-Enhancer Associated Dependencies in Diffuse Large B-Cell Lymphoma', *Cancer cell*, 24(6), p. 777. Available at: https://doi.org/10.1016/J.CCR.2013.11.003.

Chen, J. and Guo, J.-T. (2021) 'Structural and functional analysis of somatic coding and UTR indels in breast and lung cancer genomes', *Scientific Reports |*, 11, p. 21178. Available at: https://doi.org/10.1038/s41598-021-00583-1.

Chen, M. *et al.* (2017) 'The nuclear transport receptor Importin-11 is a tumor suppressor that maintains PTEN protein', *The Journal of Cell Biology*, 216(3), p. 641. Available at: https://doi.org/10.1083/JCB.201604025.

Chettiankandy, T.J. *et al.* (2016) 'B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and classical Burkitt's lymphoma: A case report and review', *Journal of Oral and Maxillofacial Pathology*. Medknow Publications, p. 333. Available at: https://doi.org/10.4103/0973-029X.185936.

Chipster (2021a) *macs2 manual*. Available at: https://chipster.csc.fi/manual/macs2.html (Accessed: 28 December 2021).

Chipster (2021b) *Peak calling with MACS2 | Introduction to ChIP-Seq using high-performance computing*. Available at: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html (Accessed: 28 December 2021).

Ch'ng, E.S. and Kumanogoh, A. (2010) 'Roles of Sema4D and Plexin-B1 in tumor progression', *Molecular Cancer*, 9(1), pp. 1–9. Available at: https://doi.org/10.1186/1476-4598-9-251/TABLES/1.

Collares, D. *et al.* (2019) 'Diffuse Large B-Cell Lymphoma, NOS (Not Otherwise Specified)', pp. 1–9. Available at: https://doi.org/10.1007/978-3-319-28845-1_3887-1.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, 17(1), pp. 1–19. Available at: https://doi.org/10.1186/s13059-016-0881-8.

Containerization Explained - South Africa | IBM (2019) *IBM*. Available at: https://www.ibm.com/za-en/cloud/learn/containerization (Accessed: 11 November 2022).

Cornish, A.J. *et al.* (2019) 'Identification of recurrent noncoding mutations in B-cell lymphoma using capture Hi-C', *Blood Advances*, 3(1), pp. 21–32. Available at: https://doi.org/10.1182/bloodadvances.2018026419.

Creixell, P. *et al.* (2015) 'Pathway and Network Analysis of Cancer Genomes', *Nature methods*, 12(7), p. 615. Available at: https://doi.org/10.1038/NMETH.3440.

Creyghton, M.P. *et al.* (2010) 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), pp. 21931–21936. Available at: https://doi.org/10.1073/pnas.1016071107.

Dekker, J.D. *et al.* (2016) 'Subtype-specific addiction of the activated B-cell subset of diffuse large B-cell lymphoma to FOXP1', *Proceedings of the National Academy of Sciences of the United States of America*, 113(5), pp. E577–E586. Available at: https://doi.org/10.1073/PNAS.1524677113/SUPPL_FILE/PNAS.1524677113.SAPP.PDF.

Derenzini, E. *et al.* (2021) 'A three-gene signature based on MYC, BCL-2 and NFKBIA improves risk stratification in diffuse large B-cell lymphoma', *Haematologica*, 106(9), p. 2405. Available at: https://doi.org/10.3324/HAEMATOL.2019.236455.

Due, H. *et al.* (2019) 'MicroRNA-155 controls vincristine sensitivity and predicts superior clinical outcome in diffuse large B-cell lymphoma', *Blood Advances*, 3(7), p. 1185. Available at: https://doi.org/10.1182/BLOODADVANCES.2018029660.

Dunham, A. *et al.* (2004) 'The DNA sequence and analysis of human chromosome 13 Europe PMC Funders Group', *Nature*, 428(6982), pp. 522–528. Available at: https://doi.org/10.1038/nature02379.

Dunleavy, K. and Wilson, W.H. (2011) 'The differential role of BCL-2 within molecular subtypes of DLBCL', *Clinical cancer research : an official journal of the American Association for Cancer Research*, 17(24), p. 7505. Available at: https://doi.org/10.1158/1078-0432.CCR-11-2372.

Eder, T. and Grebien, F. (2022) 'Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection', *Genome Biology*, 23(1). Available at: https://doi.org/10.1186/s13059-022-02686-y.

Elgundi, Z. *et al.* (2020) 'Cancer Metastasis: The Role of the Extracellular Matrix and the Heparan Sulfate Proteoglycan Perlecan', *Frontiers in Oncology*, 9, p. 1482. Available at: https://doi.org/10.3389/FONC.2019.01482/BIBTEX.

Ellenbroek, B. and Youn, J. (2016) 'Environment Challenges and the Brain', *Gene-Environment Interactions in Psychiatry*, pp. 107–139. Available at: https://doi.org/10.1016/B978-0-12-801657-2.00005-7.

Elliott, K. and Larsson, E. (2021) 'Non-coding driver mutations in human cancer', *Nature Reviews Cancer 2021 21:8*, 21(8), pp. 500–509. Available at: https://doi.org/10.1038/s41568-021-00371-z.

Emily Mell (2021) *What Is Container Management and Why Is It Important?*, *TechTarget*. Available at: https://www.techtarget.com/searchitoperations/definition/container-management-software (Accessed: 11 November 2022).

Evrard, S.M. *et al.* (2019) 'Targeted next generation sequencing reveals high mutation frequency of CREBBP, BCL2 and KMT2D in high-grade b-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements', *Haematologica*. Ferrata Storti Foundation, pp. e154–e157. Available at: https://doi.org/10.3324/haematol.2018.198572.

Fabbri, G. *et al.* (2011) 'Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation', *The Journal of Experimental Medicine*, 208(7), p. 1389. Available at: https://doi.org/10.1084/JEM.20110921.

FASTQ files explained (2022). Available at: https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html (Accessed: 10 November 2022).

Federico, A. *et al.* (2019) 'Pipeliner: A Nextflow-Based Framework for the Definition of Sequencing Data Processing Pipelines', *Frontiers in Genetics*, 10(JUN). Available at: https://doi.org/10.3389/FGENE.2019.00614.

Feng, J., Liu, T. and Zhang, Y. (2011) 'Using MACS to Identify Peaks from ChIP-Seq Data', *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, CHAPTER(SUPPL. 34), p. Unit2.14. Available at: https://doi.org/10.1002/0471250953.BI0214S34.

Feng, Y. *et al.* (2021) 'COPB2: a transport protein with multifaceted roles in cancer development and progression', *Clinical & Translational Oncology*, 23(11), p. 2195. Available at: https://doi.org/10.1007/S12094-021-02630-9.

Ferlaino, M. *et al.* (2017) 'An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome', *BMC Bioinformatics*, 18(1), p. 442. Available at: https://doi.org/10.1186/s12859-017-1862-y.

Flebbe, H. *et al.* (2019) 'Epigenome mapping identifies tumor-specific gene expression in primary rectal cancer', *Cancers*, 11(8). Available at: https://doi.org/10.3390/cancers11081142.

Frick, M., Dörken, B. and Lenz, G. (2011) 'The molecular biology of diffuse large B-cell lymphoma', *Therapeutic Advances in Hematology*, 2(6), p. 369. Available at: https://doi.org/10.1177/2040620711419001.

Furey, T.S. (2012) 'ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions', *Nature Reviews Genetics*. NIH Public Access, pp. 840–852. Available at: https://doi.org/10.1038/nrg3306.

Gagliano, S.A. *et al.* (2019) 'Relative impact of indels versus SNPs on complex disease', *Genetic Epidemiology*, 43(1), pp. 112–117. Available at: https://doi.org/10.1002/gepi.22175.

Gan, K.A. *et al.* (2018) 'Identification of single nucleotide non-coding driver mutations in cancer', *Frontiers in Genetics*. Frontiers Media S.A. Available at: https://doi.org/10.3389/fgene.2018.00016.

Gao, Y. *et al.* (2020) 'Acetylation of histone H3K27 signals the transcriptional elongation for estrogen receptor alpha', *Communications Biology 2020 3:1*, 3(1), pp. 1–10. Available at: https://doi.org/10.1038/s42003-020-0898-0.

Gascoyne, D.M. and Banham, A.H. (2017) 'The significance of FOXP1 in diffuse large B-cell lymphoma', *Leukemia & lymphoma*, 58(5), pp. 1037–1051. Available at: https://doi.org/10.1080/10428194.2016.1228932.

Gaspar, J.M. (2018) 'Improved peak-calling with MACS2', *bioRxiv*, p. 496521. Available at: https://doi.org/10.1101/496521.

Gilbert, N. *et al.* (2004) 'Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers', *Cell*, 118(5), pp. 555–566. Available at: https://doi.org/10.1016/j.cell.2004.08.011.

Gilmour, D.S. and Lis, J.T. (1984) 'Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes', *Proceedings of the National Academy of Sciences of the United States of America*, 81(14 I), pp. 4275–4279. Available at: https://doi.org/10.1073/pnas.81.14.4275.

Gouveia, G.R. *et al.* (2020) 'Overexpression of OCT-1 gene is a biomarker of adverse prognosis for diffuse large B-cell lymphoma (DLBCL): data from a retrospective cohort of 77 Brazilian patients', *BMC cancer*, 20(1). Available at: https://doi.org/10.1186/S12885-020-07553-2.

Gouveia, G.R., Siqueira, S.A.C. and Pereira, J. (2012) 'Pathophysiology and molecular aspects of diffuse large B-cell lymphoma', *Revista Brasileira de Hematologia e Hemoterapia*, 34(6), pp. 447–451. Available at: https://doi.org/10.5581/1516-8484.20120111.

Goyal, A. and Negi, P. (2021) *AWK command in Unix/Linux with examples - GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/awk-command-unixlinux-examples/ (Accessed: 2 September 2021).

Grimwood, J. *et al.* (2004) 'The DNA sequence and biology of human chromosome 19', *Nature 2004 428:6982*, 428(6982), pp. 529–535. Available at: https://doi.org/10.1038/nature02399.

Han, L. and Zhao, Z. (2009) 'CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome?', *BMC Bioinformatics*, 10, p. 65. Available at: https://doi.org/10.1186/1471-2105-10-65.

Havranek, O. *et al.* (2017) 'Tonic B-cell receptor signaling in diffuse large B-cell lymphoma', *Blood*, 130(8), pp. 995–1006. Available at: https://doi.org/10.1182/BLOOD-2016-10-747303.

Hazelett, D.J. *et al.* (2014) 'Comprehensive functional annotation of 77 prostate cancer risk loci', *PLoS genetics*, 10(1). Available at: https://doi.org/10.1371/JOURNAL.PGEN.1004102.

He, B. *et al.* (2020) 'Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers', *Science Advances*, 6(30). Available at: https://doi.org/10.1126/SCIADV.ABA3064.

He, Y., Long, W. and Liu, Q. (2019) 'Targeting super-enhancers as a therapeutic strategy for cancer treatment', *Frontiers in Pharmacology*, 10(APR), p. 361. Available at: https://doi.org/10.3389/FPHAR.2019.00361/BIBTEX.

Heckman, C.A. *et al.* (2006) 'Oct transcription factors mediate t(14;18) lymphoma cell survival by directly regulating bcl-2 expression', *Oncogene*, 25(6), pp. 888–898. Available at: https://doi.org/10.1038/sj.onc.1209127.

Honma, D. *et al.* (2017) 'Novel orally bioavailable EZH1/2 dual inhibitors with greater antitumor efficacy than an EZH2 selective inhibitor', *Cancer science*, 108(10), pp. 2069–2078. Available at: https://doi.org/10.1111/CAS.13326.

Hornshøj, H. *et al.* (2018) 'Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival /631/67/69 /631/114 article', *npj Genomic Medicine*, 3(1). Available at: https://doi.org/10.1038/s41525-017-0040-5.

Huang, D.W. *et al.* (2007) 'The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists', *Genome Biology*, 8(9), p. R183. Available at: https://doi.org/10.1186/GB-2007-8-9-R183.

Huang, F.W. *et al.* (2013) 'Highly recurrent TERT promoter mutations in human melanoma', *Science*, 339(6122), pp. 957–959. Available at: https://doi.org/10.1126/science.1229259.

Huang, H. *et al.* (2021) 'Defining super-enhancer landscape in triple-negative breast cancer by multiomic profiling', *Nature Communications*, 12(2242). Available at: https://doi.org/10.1038/s41467-021-22445-0.

Huang, M. *et al.* (2020) 'dbInDel: a database of enhancer-associated insertion and deletion variants by analysis of H3K27ac ChIP-Seq | Bioinformatics | Oxford Academic', *Bioinformatics*, 36(5), pp. 1649–1651. Available at: https://academic.oup.com/bioinformatics/article/36/5/1649/5585749 (Accessed: 13 May 2020).

Huang, X., Qian, W. and Ye, X. (2020) 'Long Noncoding RNAs in Diffuse Large B-Cell Lymphoma: Current Advances and Perspectives', *OncoTargets and therapy*, 13, p. 4295. Available at: https://doi.org/10.2147/OTT.S253330.

Hung, S. *et al.* (2019) 'Mismatch repair-signature mutations activate gene enhancers across human colorectal cancer epigenomes', *eLife*, 8, pp. 1–18. Available at: https://doi.org/10.7554/eLife.40760.

Hunter, E. *et al.* (2020) 'Comparative molecular cell-of-origin classification of diffuse large B-cell lymphoma based on liquid and tissue biopsies', *Translational Medicine Communications*, 5(1), p. 5. Available at: https://doi.org/10.1186/s41231-020-00054-1.

Hurtz, C. *et al.* (2021) 'Pharmacologic Inhibition of DYRK1A Results in Hyperactivation and Hyperphosphorylation of MYC and ERK Rendering KMT2A-R ALL Cells Sensitive to BCL2 Inhibition', *Blood*, 138(Supplement 1), p. 506. Available at: https://doi.org/10.1182/BLOOD-2021-144476.

Illumina | Sequencing and array-based solutions for genetic research (2022). Available at: https://www.illumina.com/ (Accessed: 10 November 2022).

Imielinski, M., Guo, G. and Meyerson, M. (2017) 'Insertions and Deletions Target Lineage-Defining Genes in Human Cancers', *Cell*, 168(3), pp. 460-472.e14. Available at: https://doi.org/10.1016/J.CELL.2016.12.025.

Jiang, S. and Mortazavi, A. (2018) 'Integrating ChIP-seq with other functional genomics data', *Briefings in Functional Genomics*, 17(2), pp. 104–115. Available at: https://doi.org/10.1093/bfgp/ely002.

Jinesh, G.G. *et al.* (2021) 'Mutant p53s and chromosome 19 microRNA cluster overexpression regulate cancer testis antigen expression and cellular transformation in hepatocellular carcinoma', *Scientific Reports 2021 11:1*, 11(1), pp. 1–13. Available at: https://doi.org/10.1038/s41598-021-91924-7.

Kamps, R. *et al.* (2017) 'Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification', *International Journal of Molecular Sciences*. MDPI AG. Available at: https://doi.org/10.3390/ijms18020308.

Kanzi, A.M. *et al.* (2020) 'Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance', *Frontiers in Genetics*, 11, p. 1250. Available at: https://doi.org/10.3389/FGENE.2020.544162/BIBTEX.

Karstensen, K.T. *et al.* (2021) 'Long Non-Coding RNAs in Diffuse Large B-Cell Lymphoma', *Non-Coding RNA*, 7(1), pp. 1–13. Available at: https://doi.org/10.3390/NCRNA7010001.

Kent, W.J. (2002) 'BLAT---The BLAST-Like Alignment Tool', *Genome Research*, 12(4), pp. 656–664. Available at: https://doi.org/10.1101/gr.229202.

Khuu, P. *et al.* (2007) 'Phylogenomic analysis of the emergence of GC-rich transcription elements', *Proceedings of the National Academy of Sciences of the United States of America*, 104(42), pp. 16528–16533. Available at: https://doi.org/10.1073/PNAS.0707203104.

Kim, B.Y. *et al.* (2017) 'Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data', *PloS one*, 12(8), p. e0182272. Available at: https://doi.org/10.1371/journal.pone.0182272.

Kisakol, B. *et al.* (2021) 'Detailed evaluation of cancer sequencing pipelines in different microenvironments and heterogeneity levels', *Turkish Journal of Biology*, 45(2), p. 114. Available at: https://doi.org/10.3906/BIY-2008-8.

Kotlov, N. *et al.* (2021) 'Clinical and biological subtypes of b-cell lymphoma revealed by microenvironmental signatures', *Cancer Discovery*, 11(6), pp. 1468–1489. Available at: https://doi.org/10.1158/2159-8290.CD-20-0839/333553/AM/CLINICAL-AND-BIOLOGICAL-SUBTYPES-OF-B-CELL.

Koudritsky, M. and Domany, E. (2008) 'Positional distribution of human transcription factor binding sites', *Nucleic Acids Research*, 36(21), pp. 6795–6805. Available at: https://doi.org/10.1093/NAR/GKN752.

Kulski, J.K. (2016) 'Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications', in *Next Generation Sequencing - Advances, Applications and Challenges*. InTech. Available at: https://doi.org/10.5772/61964.

Kuo, H.P. *et al.* (2016) 'The role of PIM1 in the ibrutinib-resistant ABC subtype of diffuse large B-cell lymphoma', *American Journal of Cancer Research*, 6(11), p. 2489. Available at: https://doi.org/10.1182/blood.v126.23.699.699.

Kurtzer, G.M., Sochat, V. and Bauer, M.W. (2017) 'Singularity: Scientific containers for mobility of compute', *PLoS ONE*, 12(5). Available at: https://doi.org/10.1371/JOURNAL.PONE.0177459.

Langmead, B. (2010) 'Aligning short sequencing reads with Bowtie', *Current protocols in bioinformatics*, Chapter 11(SUPP.32). Available at: https://doi.org/10.1002/0471250953.BI1107S32.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Lau, Y.F.C. (2020) 'Y chromosome in health and diseases', *Cell & Bioscience*, 10(1). Available at: https://doi.org/10.1186/S13578-020-00452-W.

Leipzig, J. (2017) 'A review of bioinformatic pipeline frameworks', *Briefings in Bioinformatics*, 18(3), pp. 530–536. Available at: https://doi.org/10.1093/bib/bbw020.

Lench, N. *et al.* (2013) 'The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made', *Prenatal Diagnosis*, 33(6), pp. 555–562. Available at: https://doi.org/10.1002/pd.4124.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), p. 2078. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTP352.

Li, H. *et al.* (2018) 'Promotion of Sema4D expression by tumor-associated macrophages: Significance in gastric carcinoma', *World Journal of Gastroenterology*, 24(5), p. 593. Available at: https://doi.org/10.3748/WJG.V24.I5.593.

Li, S. *et al.* (2019) 'BCL6 rearrangement indicates poor prognosis in diffuse large B-cell lymphoma patients: A meta-analysis of cohort studies', *Journal of Cancer*, 10(2), pp. 530–538. Available at: https://doi.org/10.7150/jca.25732.

Li, Z. *et al.* (2021) 'Clinical Features and Prognostic Significance of NOTCH1 Mutations in Diffuse Large B-Cell Lymphoma', *Frontiers in Oncology*, 11, p. 5084. Available at: https://doi.org/10.3389/FONC.2021.746577/BIBTEX.

Liu, X., Bienkowska, J.R. and Zhong, W. (2020) 'GeneTEFlow: A Nextflow-based pipeline for analysing gene and transposable elements expression from RNA-Seq data', *PLOS ONE*, 15(8), p. e0232994. Available at: https://doi.org/10.1371/JOURNAL.PONE.0232994.

Liu, Y. (2020) 'Clinical implications of chromatin accessibility in human cancers', *Oncotarget*, 11(18), p. 1666. Available at: https://doi.org/10.18632/ONCOTARGET.27584.

Ma, W. and Wong, W.H. (2011) 'The Analysis of ChIP-Seq Data', in *Methods in enzymology*. Methods Enzymol, pp. 51–73. Available at: https://doi.org/10.1016/b978-0-12-385075-1.00003-2.

Magangane, P.S., Mohamed, Z. and Naidoo, R. (2020) 'Diffuse large B-cell lymphoma in a high human immunodeficiency virus (HIV) prevalence, low-resource setting', *South African Journal of Oncology*, 4(0), p. 7. Available at: https://doi.org/10.4102/SAJO.V4I0.104.

Makova, K.D. and Hardison, R.C. (2015) 'The effects of chromatin organization on variation in mutation rates in the genome'. Available at: https://doi.org/10.1038/nrg3890.

Mansour, M.R. *et al.* (2016) 'Mutation of a Noncoding Intergenic Element', *Science*, 346(6215), pp. 1373–1377. Available at: https://doi.org/10.1126/science.1259037.An.

Marsano, R.M. and Dimitri, P. (2022) 'Constitutive Heterochromatin in Eukaryotic Genomes: A Mine of Transposable Elements', *Cells*, 11(5). Available at: https://doi.org/10.3390/CELLS11050761.

Matelsky, J. *et al.* (2018) 'Container-Based Clinical Solutions for Portable and Reproducible Image Analysis', *Journal of Digital Imaging*, 31(3), p. 315. Available at: https://doi.org/10.1007/S10278-018-0089-4.

McKay, D. (2020) *How to Use the awk Command on Linux*. Available at: https://www.howtogeek.com/562941/how-to-use-the-awk-command-on-linux/ (Accessed: 2 September 2021).

MedlinePlus: Chromosomes & mtDNA (2021). Available at: https://medlineplus.gov/genetics/chromosome/ (Accessed: 25 November 2022).

Meldrum, C., Doyle, M.A. and Tothill, R.W. (2011) 'Next-generation sequencing for cancer diagnostics: A practical perspective', *Clinical Biochemist Reviews*. The Australian Association of Clinical Biochemists, pp. 177–195.

Miao, Y. *et al.* (2018) 'Diffuse large B-cell lymphoma with molecular variations more than ABC and GCB classification', *Precision Cancer Medicine*, 1(0). Available at: https://doi.org/10.21037/PCM.2018.06.03.

Miao, Y. *et al.* (2019) 'Dysregulation of cell survival in diffuse large B cell lymphoma: Mechanisms and therapeutic targets', *Frontiers in Oncology*. Frontiers Media S.A., p. 107. Available at: https://doi.org/10.3389/fonc.2019.00107.

Mills, R.E. *et al.* (2006) 'An initial map of insertion and deletion (INDEL) variation in the human genome', *Genome Research*, 16(9), pp. 1182–1190. Available at: https://doi.org/10.1101/gr.4565806.

Mitra-Behura, S., Fiolka, R.P. and Daetwyler, S. (2022) 'Singularity Containers Improve Reproducibility and Ease of Use in Computational Image Analysis Workflows', *Frontiers in Bioinformatics*, 0, p. 61. Available at: https://doi.org/10.3389/FBINF.2021.757291.

Moorthie, S., Mattocks, C.J. and Wright, C.F. (2011) 'Review of massively parallel DNA sequencing technologies', *HUGO Journal*, 5(1–4), pp. 1–12. Available at: https://doi.org/10.1007/s11568-011-9156-3.

Mullaney, J.M. *et al.* (2010) 'Small insertions and deletions (INDELs) in human genomes', *Human Molecular Genetics*, 19(2), pp. 131–136. Available at: https://doi.org/10.1093/hmg/ddq400.

Mundade, R. *et al.* (2014) 'Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond', *Cell Cycle*. Landes Bioscience, pp. 2847–2852. Available at: https://doi.org/10.4161/15384101.2014.949201.

Mzoughi, S. *et al.* (2020) 'PRDM15 is a key regulator of metabolism critical to sustain B-cell lymphomagenesis', *Nature Communications 2020 11:1*, 11(1), pp. 1–14. Available at: https://doi.org/10.1038/s41467-020-17064-0.

Naidoo, N. *et al.* (2018) 'Incidence of hodgkin lymphoma in HIV-positive and HIV-negative patients at a tertiary hospital in South Africa (2005-2016) and comparison with other African countries', *South African Medical Journal*, 108(7), pp. 563–567. Available at: https://doi.org/10.7196/SAMJ.2018.v108i7.12844.

Nakagomi, T. *et al.* (2019) 'Clinical Implications of Noncoding Indels in the Surfactant-Encoding Genes in Lung Cancer', *Cancers*, 11(4). Available at: https://doi.org/10.3390/CANCERS11040552.

Nakato, R. and Sakata, T. (2021) 'Methods for ChIP-seq analysis: A practical workflow and advanced applications', *Methods*, 187, pp. 44–53. Available at: https://doi.org/10.1016/J.YMETH.2020.03.005.

Neph, S. *et al.* (2012) 'BEDOPS: high-performance genomic feature operations', *Bioinformatics*, 28(14), p. 1919. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTS277.

Nones, K. and Patch, A.M. (2020) 'The Impact of Next Generation Sequencing in Cancer Research', *Cancers*, 12(10), pp. 1–4. Available at: https://doi.org/10.3390/CANCERS12102928.

Nowakowski, G.S. *et al.* (2019) 'Integrating precision medicine through evaluation of cell of origin in treatment planning for diffuse large B-cell lymphoma', *Blood Cancer Journal 2019 9:6*, 9(6), pp. 1–10. Available at: https://doi.org/10.1038/s41408-019-0208-6.

Nowakowski, G.S. and Czuczman, M.S. (2015) 'ABC, GCB, and Double-Hit Diffuse Large B-Cell Lymphoma: Does Subtype Make a Difference in Therapy Selection?', *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, (35), pp. e449–e457. Available at: https://doi.org/10.14694/EDBOOK_AM.2015.35.E449.

Ocsenas, O. and Reimand, J. (2022) 'Chromatin accessibility of primary human cancers ties regional mutational processes and signatures with tissues of origin', *PLOS Computational Biology*, 18(8), p. e1010393. Available at: https://doi.org/10.1371/JOURNAL.PCBI.1010393.

Panigrahi, A. and O'Malley, B.W. (2021) 'Mechanisms of enhancer action: the known and the unknown', *Genome Biology 2021 22:1*, 22(1), pp. 1–30. Available at: https://doi.org/10.1186/S13059-021-02322-1.

Pasqualucci, L. and Dalla-Favera, R. (2015) 'The Genetic Landscape of Diffuse Large B-Cell Lymphoma', *Seminars in Hematology*, 52(2), pp. 67–76. Available at: https://doi.org/10.1053/j.seminhematol.2015.01.005.

Pasqualucci, L. and Dalla-Favera, R. (2018) 'Genetics of diffuse large B-cell lymphoma', *Blood*, 131(21), p. 2307. Available at: https://doi.org/10.1182/BLOOD-2017-11-764332.

Pather, S. and Patel, M. (2022) 'HIV-associated DLBCL: Clinicopathological factors including dual-colour chromogenic in situ hybridisation to assess MYC gene copies', *Annals of Diagnostic Pathology*, 58, p. 151913. Available at: https://doi.org/10.1016/J.ANNDIAGPATH.2022.151913.

Perenthaler, E. *et al.* (2019) 'Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development', *Frontiers in Cellular Neuroscience*. Frontiers Media S.A. Available at: https://doi.org/10.3389/fncel.2019.00352.

Piovesan, A. *et al.* (2019) 'On the length, weight and GC content of the human genome', *BMC Research Notes*, 12(1). Available at: https://doi.org/10.1186/S13104-019-4137-Z.

Price, Z.K., Lokman, N.A. and Ricciardelli, C. (2018) 'Differing Roles of Hyaluronan Molecular Weight on Cancer Cell Behavior and Chemotherapy Resistance', *Cancers*, 10(12). Available at: https://doi.org/10.3390/CANCERS10120482.

Quinlan, A.R. (2014) 'BEDTools: the Swiss-army tool for genome feature analysis', *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 47, p. 11.12.1. Available at: https://doi.org/10.1002/0471250953.BI1112S47.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTQ033.

Raha, D., Hong, M. and Snyder, M. (2010) 'ChIP-Seq: A method for global identification of regulatory elements in the genome', *Current Protocols in Molecular Biology*. Curr Protoc Mol Biol. Available at: https://doi.org/10.1002/0471142727.mb2119s91.

Rahman, S. and Mansour, M.R. (2019) 'The role of noncoding mutations in blood cancers', *DMM Disease Models and Mechanisms*. Company of Biologists Ltd. Available at: https://doi.org/10.1242/dmm.041988.

Rammohan, M. *et al.* (2022) 'The chromosome 21 kinase DYRK1A: emerging roles in cancer biology and potential as a therapeutic target', *Oncogene 2022 41:14*, 41(14), pp. 2003–2011. Available at: https://doi.org/10.1038/s41388-022-02245-6.

Re, A., Cattaneo, C. and Rossi, G. (2019) 'HIV and lymphoma: From epidemiology to clinical management', *Mediterranean Journal of Hematology and Infectious Diseases*, 11(1), pp. 1–17. Available at: https://doi.org/10.4084/mjhid.2019.004.

Reimand, J. *et al.* (2019) 'Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap', *Nature Protocols 2019 14:2*, 14(2), pp. 482–517. Available at: https://doi.org/10.1038/s41596-018-0103-9.

di Resta, C. *et al.* (2018) 'Integration of multigene panels for the diagnosis of hereditary retinal disorders using Next Generation Sequencing and bioinformatics approaches.', *Ejifcc*, 29(1), pp. 15–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29765283%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5949615.

Rheinbay, E. *et al.* (1965) 'Analyses of non-coding somatic drivers in 2,658 cancer whole genomes, PCAWG Drivers and Functional Interpretation Working Group 68 , PCAWG Structural Variation Working Group', *Nature*, 578(7), p. 67. Available at: https://doi.org/10.1038/s41586-020-1965-x.

Richardson, A.I. *et al.* (2019) 'p53 expression in large B-cell lymphomas with MYC extra copies and CD99 expression in large B-cell lymphomas in relation to MYC status', *Human Pathology*, 86, pp. 21–31. Available at: https://doi.org/10.1016/j.humpath.2018.11.015.

Rivas, M.A. *et al.* (2021) 'Smc3 dosage regulates B cell transit through germinal centers and restricts their malignant transformation', *Nature immunology*, 22(2), p. 240. Available at: https://doi.org/10.1038/S41590-020-00827-8.

Roschewski, M., Phelan, J.D. and Wilson, W.H. (2020) 'Molecular Classification and Treatment of Diffuse Large B-cell Lymphoma and Primary Mediastinal B-cell Lymphoma',

Cancer journal (Sudbury, Mass.), 26(3), p. 195. Available at: https://doi.org/10.1097/PPO.0000000000000450.

Roschewski, M., Staudt, L.M. and Wilson, W.H. (2014) 'Diffuse large B-cell lymphoma—treatment approaches in the molecular era', *Nature reviews. Clinical oncology*, 11(1), p. 12. Available at: https://doi.org/10.1038/NRCLINONC.2013.197.

Sanghi, A. *et al.* (2021) 'Chromatin accessibility associates with protein-RNA correlation in human cancer'. Available at: https://doi.org/10.1038/s41467-021-25872-1.

Sasi, B.K. *et al.* (2018) 'SHP1 Deficiency Is Responsible for the Constitutive Activation of the BCR Pathway in GCB DLBCL', *Blood*, 132(Supplement 1), pp. 2860–2860. Available at: https://doi.org/10.1182/BLOOD-2018-99-118120.

Schrader, A.M.R. *et al.* (2022) 'Cell-of-origin classification using the Hans and Lymph2Cx algorithms in primary cutaneous large B-cell lymphomas', *Virchows Archiv*, 480(3), pp. 667–675. Available at: https://doi.org/10.1007/S00428-021-03265-5/TABLES/1.

Schuetz, J.M. *et al.* (2012) 'BCL2 mutations in diffuse large B-cell lymphoma', *Leukemia*, 26(6), pp. 1383–1390. Available at: https://doi.org/10.1038/LEU.2011.378.

Sehn, J.K. (2015) 'Insertions and Deletions (Indels)', *Clinical Genomics*, pp. 129–150. Available at: https://doi.org/10.1016/B978-0-12-404748-8.00009-5.

Shanmugam, V. *et al.* (2021) 'Notch activation is pervasive in SMZL and uncommon in DLBCL: implications for Notch signaling in B-cell tumors', *Blood Advances*, 5(1), p. 71. Available at: https://doi.org/10.1182/BLOODADVANCES.2020002995.

Solomon, M.J., Larsen, P.L. and Varshavsky, A. (1988) 'Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene', *Cell*, 53(6), pp. 937–947. Available at: https://doi.org/10.1016/S0092-8674(88)90469-2.

Spjuth, O. *et al.* (2015) 'Experiences with workflows for automating data-intensive bioinformatics', *Biology Direct*, 10(1), pp. 1–12. Available at: https://doi.org/10.1186/s13062-015-0071-8.

Sur, I. and Taipale, J. (2016) 'The role of enhancers in cancer', *Nature Reviews Cancer*. Nature Publishing Group, pp. 483–493. Available at: https://doi.org/10.1038/nrc.2016.62.

Szydłowski, M. *et al.* (2021) 'Inhibition of PIM Kinases in DLBCL Targets MYC Transcriptional Program and Augments the Efficacy of Anti-CD20 Antibodies', *Cancer research*, 81(23), pp. 6029–6043. Available at: https://doi.org/10.1158/0008-5472.CAN-21-1023.

Tang, F. *et al.* (2020) 'Super-enhancer function and its application in cancer targeted therapy', *Precision Oncology*, 4(2). Available at: https://doi.org/10.1038/s41698-020-0108-z.

Tarasov, A. *et al.* (2015) 'Sambamba: fast processing of NGS alignment formats', *Bioinformatics*, 31(12), p. 2032. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTV098.

Tate, J.G. *et al.* (2019) 'COSMIC: The Catalogue Of Somatic Mutations In Cancer', *Nucleic Acids Research*, 47(D1), pp. D941–D947. Available at: https://doi.org/10.1093/nar/gky1015.

Thomas, R. *et al.* (2017) 'Features that define the best ChIP-seq peak calling algorithms', *Briefings in Bioinformatics*, 18(3), p. 441. Available at: https://doi.org/10.1093/BIB/BBW035.

Thompson, J.C. *et al.* (2016) 'Detection of therapeutically targetable driver and resistance mutations in lung cancer patients by next-generation sequencing of cell-free circulating tumor DNA', *Clinical Cancer Research*, 22(23), pp. 5772–5782. Available at: https://doi.org/10.1158/1078-0432.CCR-16-1231.

Tipney, H. and Hunter, L. (2010) 'An introduction to effective use of enrichment analysis software', *Human Genomics*, 4(3), pp. 202–206. Available at: https://doi.org/10.1186/1479-7364-4-3-202/METRICS.

di Tommaso, P. *et al.* (2015) 'The impact of Docker containers on the performance of genomic pipelines', *PeerJ*, 2015(9). Available at: https://doi.org/10.7717/PEERJ.1273/SUPP-1.

di Tommaso, P. *et al.* (2017) 'Nextflow, an efficient tool to improve computation numerical stability in genomic analysis', *Biologie aujourd'hui*, 211(3), pp. 233–237. Available at: https://doi.org/10.1051/JBIO/2017029.

di Tommaso, P. and Floden, E. (2021) *Nextflow - A DSL for parallel and scalable computational pipelines*, *Seqera Labs*. Available at: https://www.nextflow.io/ (Accessed: 3 October 2021).

TRAJ35 protein expression summary - The Human Protein Atlas (2022). Available at: https://www.proteinatlas.org/ENSG00000211854-TRAJ35 (Accessed: 26 September 2022).

TRAPPC2B - Trafficking protein particle complex subunit 2B - Homo sapiens (Human) | UniProtKB | UniProt (2022). Available at: https://www.uniprot.org/uniprotkb/P0DI82/entry (Accessed: 4 November 2022).

Tripodo, C. *et al.* (2020) 'A Spatially Resolved Dark- versus Light-Zone Microenvironment Signature Subdivides Germinal Center-Related Aggressive B Cell Lymphomas', *iScience*, 23(10). Available at: https://doi.org/10.1016/J.ISCI.2020.101562.

del Vecchio, F. *et al.* (2017) 'Next-generation sequencing: recent applications to the analysis of colorectal cancer', *Journal of Translational Medicine 2017 15:1*, 15(1), pp. 1–19. Available at: https://doi.org/10.1186/S12967-017-1353-Y.

Vinagre, J. *et al.* (2013) 'Frequency of TERT promoter mutations in human cancers', *Nature Communications*, 4. Available at: https://doi.org/10.1038/ncomms3185.

Wang, M. and Kong, L. (2019) 'pblat: a multithread blat algorithm speeding up aligning sequences to genomes', *BMC Bioinformatics 2019 20:1*, 20(1), pp. 1–4. Available at: https://doi.org/10.1186/S12859-019-2597-8.

Wang, X. *et al.* (2018) 'Characteristics of The Cancer Genome Atlas cases relative to U.S. general population cancer cases', *British Journal of Cancer*, 119(7), pp. 885–892. Available at: https://doi.org/10.1038/s41416-018-0140-8.

Wang, X., Yan, J. and Cairns, M.J. (2019) 'Super-enhancers in transcriptional regulation and genome organization | Nucleic Acids Research | Oxford Academic', *Nucleic Acids Research*, 47(22), pp. 11481–11496. Available at: https://academic.oup.com/nar/article/47/22/11481/5625531 (Accessed: 13 May 2020).

Wang, X.J. *et al.* (2017) 'P53 expression correlates with poorer survival and augments the negative prognostic effect of MYC rearrangement, expression or concurrent MYC/BCL2 expression in diffuse large B-cell lymphoma', *Modern Pathology*, 30(2), pp. 194–203. Available at: https://doi.org/10.1038/modpathol.2016.178.

Wassef, M. *et al.* (2019) 'EZH1/2 function mostly within canonical PRC2 and exhibit proliferation-dependent redundancy that shapes mutational signatures in cancer', *Proceedings of the National Academy of Sciences of the United States of America*, 116(13), pp. 6075–6080. Available at: https://doi.org/10.1073/PNAS.1814634116/SUPPL_FILE/PNAS.1814634116.SD01.XLSX.

Weber, T. and Schmitz, R. (2022) 'Molecular Subgroups of Diffuse Large B Cell Lymphoma: Biology and Implications for Clinical Practice', *Current Oncology Reports*, 24(1), pp. 13–21. Available at: https://doi.org/10.1007/S11912-021-01155-2/TABLES/1.

Wei, J. *et al.* (2020) 'Molecular Sciences Roles of Proteoglycans and Glycosaminoglycans in Cancer Development and Progression'. Available at: https://doi.org/10.3390/ijms21175983.

Wu, J. *et al.* (2021) 'Clinical characteristics and outcomes in HIV-associated diffuse large B-cell lymphoma in China: A retrospective single-center study', *Journal of Cancer*, 12(10), p. 2903. Available at: https://doi.org/10.7150/JCA.51027.

Wutz, A. (2011) 'Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation', *Nature reviews. Genetics*, 12(8), pp. 542–553. Available at: https://doi.org/10.1038/NRG3035.

Xia, Y. *et al.* (2017) 'Loss of PRDM1/BLIMP-1 function contributes to poor prognosis of activated B-cell-like diffuse large B-cell lymphoma', *Leukemia*, 31(3), pp. 625–636. Available at: https://doi.org/10.1038/leu.2016.243.

Xia, Y. and Zhang, X. (2020) 'The Spectrum of MYC Alterations in Diffuse Large B-Cell Lymphoma', *Acta haematologica*, 143(6), pp. 520–528. Available at: https://doi.org/10.1159/000505892.

Xie, Y., Pittaluga, S. and Jaffe, E.S. (2015) 'The Histological Classification of Diffuse Large B-cell Lymphomas', *Seminars in hematology*, 52(2), p. 57. Available at: https://doi.org/10.1053/J.SEMINHEMATOL.2015.01.006.

Xu, H. *et al.* (2022) 'A Novel Defined Super-Enhancer Associated Gene Signature to Predict Prognosis in Patients With Diffuse Large B-Cell Lymphoma', *Frontiers in Genetics*, 13. Available at: https://doi.org/10.3389/FGENE.2022.827840/.

Yang, H. *et al.* (2010) 'Important role of indels in somatic mutations of human cancer genes', *BMC Medical Genetics*, 11(1), p. 128. Available at: https://doi.org/10.1186/1471-2350-11-128.

Yao, J. *et al.* (2020) 'Epigenetic plasticity of enhancers in cancer', *Transcription*, 11(1), p. 26. Available at: https://doi.org/10.1080/21541264.2020.1713682.

Young, R.M. *et al.* (2015) 'B-cell receptor signaling in diffuse large B-cell lymphoma', *Seminars in hematology*, 52(2), pp. 77–85. Available at: https://doi.org/10.1053/J.SEMINHEMATOL.2015.01.008.

Yu, L., Yu, T. and Young, K.H. (2019) 'Cross-talk between Myc and p53 in B-cell lymphomas', *Chronic Diseases and Translational Medicine*, 5(3), pp. 139–154. Available at: https://doi.org/10.1016/J.CDTM.2019.08.001.

Zhang, T. *et al.* (2020) 'Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells', *Genome Biology*, 21(1), pp. 1–7. Available at: https://doi.org/10.1186/S13059-020-01957-W/FIGURES/2.

Zhang, W. and Yu, Y. (2011) 'The important molecular markers on chromosome 17 and their clinical impact in breast cancer', *International journal of molecular sciences*, 12(9), pp. 5672–5683. Available at: https://doi.org/10.3390/IJMS12095672.

Zhou, J. *et al.* (2022) 'PIM1 and CD79B Mutation Status Impacts the Outcome of Primary Diffuse Large B-Cell Lymphoma of the CNS', *Frontiers in Oncology*, 12, p. 234. Available at: https://doi.org/10.3389/FONC.2022.824632/BIBTEX.

Zhu, Y. *et al.* (2022) 'Oncogenic Mutations and Tumor Microenvironment Alterations of Older Patients With Diffuse Large B-Cell Lymphoma', *Frontiers in Immunology*, 13, p. 842439. Available at: https://doi.org/10.3389/FIMMU.2022.842439/FULL.

Zody, M.C. *et al.* (2006) 'DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage', *Nature*, 440(7087), p. 1045. Available at: https://doi.org/10.1038/NATURE04689.

APPENDIX

Table 11: Experimental gene symbols and associated gene names.

GENE SYMBOL	GENE NAME
AP-1	Activator protein 1
ADPRM	ADP-ribose/CDP-alcohol diphosphatase, manganese dependent
AKT	AKT serine/threonine kinase
APH-1	Anterior pharynx-defective 1
ATXN1	Ataxin 1
ATXN1L	Ataxin 1 like
Survivin	Baculoviral inhibitor of apoptosis repeat-containing 5
CASP3	Caspase 3
CD22	CD22 Molecule
CDC42SE2	CDC42 small effector 2
CADM1	Cell adhesion molecule 1
CDC42	Cell division cycle 42
C6orf89	Chromosome 6 open reading frame 89
CCDC106	Coiled-coil domain containing 106
CDKN1B	Cyclin dependent kinase inhibitor 1B
CycT1	Cyclin T1
CYTH1	Cytohesin 1
EP300	E1A binding protein p300
ELL2	Elongation factor for RNA polymerase II 2
EZH1	Enhancer of zeste 1 polycomb repressive complex 2 subunit
TEL	ETS variant transcription factor 6
EWSR1	EWS RNA binding protein 1
Fas	Fas cell surface death receptor
FNBP1	Formin binding protein 1
GPCPD1	Glycerophosphocholine phosphodiesterase 1
GRB2	Growth factor receptor bound protein 2
HSP	Heat shock genes
HATs	Histone acetyltransferases
INI1	Integrase interactor 1
ITGA	Integrin subunit alpha 1
IRAK4	Interleukin 1 receptor associated kinase 4
IQSEC1	IQ motif and SEC7 domain-containing protein 1
LMCD1	LIM and cysteine-rich domains 1
LINC01825	Long intergenic non-protein coding RNA 1825
LINC-PINT	Long intergenic non-protein coding RNA, p53 induced transcript
LYL1	LYL1 basic helix-loop-helix family member
HLA-DRB5	Major histocompatibility complex, class II, DR beta 5
MAML	Mastermind like transcriptional coactivator
MIR155HG	MIR155 host gene
MAPK	Mitogen-activated protein kinases

ENL	MLLT1 super elongation complex subunit
MSI2	Musashi RNA binding protein 2
Max	MYC associated factor X
NDUFV2	NADH:ubiquinone oxidoreductase core subunit V2
NEDD9	Neural precursor cell expressed, developmentally down-regulated 9
NBPF15	Neuroblastoma breakpoint family, member 15
NUP210	Nucleoporin 210
Numb	NUMB endocytic adaptor protein
P2RY8	P2Y receptor family member 8
PAXIP1	PAX interacting protein 1
PIN1	Peptidylprolyl cis/trans isomerase, NIMA-interacting 1
PKC	Proline rich transmembrane protein 2
PKB	Protein kinase B
PKCB	Protein kinase C beta
p27	Protein nb
PYK2	Protein tyrosine kinase 2 beta
SHP-1	Protein tyrosine phosphatase non-receptor type 6
RAB7A	RAB7A, member RAS oncogene family
RIP1	Receptor-interacting serine/threonine-protein kinase 1
ARHGAP4	Rho GTPase activating protein 4
Rho	RHO Family GTPases
SERINC	Serine incorporator
SMIM20	Small integral membrane protein 20
SLC22A18	Solute carrier family 22 member 18
SPECC1	Sperm antigen with calponin homology and coiled-coil domains 1
SYK	Spleen associated tyrosine kinase
STING	Stimulator of interferon genes
TRAC	T-cell receptor alpha constant
TNFR1	TNF receptor superfamily member 1A
TNIP3	TNFAIP3 interacting protein 3
TRAFs	Tumor necrosis factor receptor-associated factors
TTYH3	Tweety family member 3
USP7-AS1	Ubiquitin specific peptidase 7 antisense RNA 1
VEGF	Vascular endothelial growth factor A
VAV	VAV guanine nucleotide exchange factor 1
PLZF	Zinc finger and BTB domain containing 16
ZC3H4	Zinc finger CCCH-type containing 4
Zeb1	Zinc finger E-box binding homeobox 1
ZNF354A	Zinc finger protein 354A