Molecular characterization of peripheral T-cell lymphoma through whole exome sequencing based missense variant annotation and enrichment analysis of biological pathways and disease associations

A final year project report submitted to

DEPARTMENT OF BIOINFORMATICS

BHARATHIDASAN UNIVERSITY,

TIRUCHIRAPALLI

In fulfillment of the requirements for the 5th Year of the Degree of

MASTER OF SCIENCE

(5 year Integrated)

in

BIOINFORMATICS

Submitted

by

MOHAMED THOWFEEK A

(Reg.no: MSFBI2009)

DEPARTMENT OF BIOINFORMATICS
SCHOOL OF LIFE SCIENCES
BHARATHIDASAN UNIVERSITY
TIRUCHIRAPALLI – 620024
TAMILNADU

2024-2025

DEPARTMENT OF BIOINFORMATICS
SCHOOL OF LIFE SCIENCES
BHARATHIDASAN UNIVERSITY
TIRUCHIRAPPALLI – 620 024,
TAMILNADU, INDIA.

CERTIFICATE

This is to certify that the project report entitled "**Molecular characterization of peripheral T-cell lymphoma through whole exome sequencing-based missense variant annotation and enrichment analysis of biological pathways and disease associations**" submitted to the Department of Bioinformatics, Bharathidasan University, Tiruchirappalli, in partial fulfillment of the requirements for the Fifth year of the Master of Science (Five-Year Integrated) in Bioinformatics degree, is an original work carried out by **Mr. A. MOHAMED THOWFEEK (Reg. No: MSFBI2009)** during the period 2024–2025.

Signature of the Internal Guide

Signature of the Head of Department

Submitted for the Project Evaluation and Viva-voce Held on_____

Internal Examiner

External Examiner

DECLARATION

I hereby declare that the dissertation entitled, **"Molecular characterization of peripheral T-cell lymphoma through whole exome sequencing based missense variant annotation and enrichment analysis of biological pathways and disease associations"** submitted to the Department of Bioinformatics, Bharathidasan University, Tiruchirappalli, in fulfillment of the requirements for the award of the Master of Science (Five-Year Integrated) in Bioinformatics degree, is a bonafide record of the original and independent research work carried out by me at Bionome, Bangalore**,** during the period January 2025 to April 2025 and further declare that this dissertation has not been submitted previously, either in part or full, for the award of any other degree, diploma, or fellowship from any other university or institution.

Signature of the candidate

MOHAMED THOWFEEK A
(MSFBI2009)

ACKNOWLEDGEMENT

I thank the Almighty for bestowing upon me His blessings, which made it possible for me to successfully complete this endeavour.

I express my deepest gratitude to **Mr. Sameer Sharma**, CSO and Director of **Bionome, Bangalore**, for providing me with the opportunity to work at such a prestigious company and for his continued support during my project.

I am immensely grateful to **Ms. Susha Dinesh**, Senior Bioinformatics Analyst at **Bionome**, for her invaluable guidance, encouragement, and unwavering support throughout my project. Her insights and expertise have greatly contributed to the successful completion of this research.

I am profoundly thankful to my internal guide, **Dr. K.S. Jayachandran**, Assistant Professor at Bharathidasan University, for his constructive feedback, motivation, and academic guidance, which played a crucial role in refining my research.

My sincere thanks to **Dr. P. Chellapandi**, Head of the Department, and **Dr. P. Senthilraja**, Assistant Professor, Department of Bioinformatics, Bharathidasan University, for their encouragement and for providing a conducive research environment.

I also extend my heartfelt gratitude to my family, friends, and colleagues for their unwavering support and motivation throughout this journey.

Finally, I acknowledge **Bionome** and **Bharathidasan University** for providing me with the necessary facilities and resources to conduct my research successfully.

MOHAMED THOWFEEK A

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVATIONS

PTCL	Peripheral T-Cell Lymphoma
WES	Whole Exome Sequencing
HL	Hodgkin Lymphoma
NHL	Non-Hodgkin Lymphoma
NGS	Next Generation Sequencing
INDELS	Insertion And Deletion
CNV	Copy Number Variations
SNV	Single-Nucleotide Variants
BWA	Burrows-Wheeler Aligner
SRA	Sequence Read Archive
NCBI	National Center For Biotechnology Information
BAM	Binary Alignment Map
VCF	Variant Call Format
GATK	Genome Analysis Toolkit
VEP	Variant Effect Predictor
GO	Gene Ontology
BP	Biological Processes
MF	Molecular Functions
SAM	Sequence Alignment Map

ABSTRACT

Peripheral T-cell lymphoma (PTCL) is a heterogeneous and aggressive subtype of blood cancer, classified under non-Hodgkin lymphoma and characterized by complex genetic alterations. Identifying key genetic variants in PTCL is crucial for understanding its pathogenesis and discovering potential therapeutic targets. In this study, Whole Exome Sequencing (WES) data from two PTCL samples retrieved from the NCBI SRA database were analyzed. After quality control, annotation, and filtration, missense variants were prioritized for functional enrichment analysis. Gene Ontology (Biological Process, Molecular Function), pathway analysis (Reactome, WikiPathway), disease-gene associations, and tissue-specific expression analyses were performed. The results highlighted *TP53*, *BCL6*, *BCL2*, and *AKT1* as key genes enriched across multiple biological pathways, including apoptosis, immune regulation, and signal transduction. These genes are known to play significant roles in tumor progression and immune evasion. The study provides insights into the molecular landscape of PTCL, contributing to biomarker discovery and potential therapeutic interventions. Further validation through experimental studies is necessary to confirm these findings.

CHAPTER I

INTRODUCTION

1.1. Cancer: A Global Health Burden

Cancer is a complex and multifaceted disease characterized by uncontrolled cellular proliferation, invasion, and metastasis. It arises due to genetic and epigenetic alterations that disrupt normal cellular regulatory mechanisms, including cell cycle control, apoptosis, and DNA repair pathways. The pathological hallmarks of cancer include sustained proliferative signalling, evasion of growth suppressors, resistance to cell death, replicative immortality, angiogenesis, and the ability to invade tissues and metastasize. These malignant transformations not only compromise normal physiological functions but also lead to significant metabolic dysregulation, immune evasion, and systemic effects such as cancer-associated cachexia and organ dysfunction.

Beyond its direct biological impact, cancer has profound psychosocial and economic consequences. It not only deteriorates an individual's physical health but also leads to severe psychological distress, affecting mental well-being and quality of life. The disease disrupts family dynamics and social structures, placing emotional and financial burdens on patients and caregivers alike. Cancer remains a leading cause of mortality worldwide, accounting for approximately 10 million deaths annually, underscoring its persistent global burden (Huang *et al*., 2022)

1.1.2 Epidemiology and Global Trends in Cancer

The global incidence of cancer has increased significantly over the past few decades due to demographic changes, lifestyle factors, and environmental influences. In 2015, the worldwide cancer burden reached 17.5 million new cases, with 8.7 million cancer-related deaths reported. The incidence of cancer increased by 33% between 2005 and 2015, largely due to population aging, growth, and the increasing prevalence of risk factors such as tobacco use, excessive alcohol consumption, obesity, and sedentary lifestyles (Cabral *et al*., 2018).

Cancer is currently the second leading cause of death globally, and projections suggest that the number of new cases will rise to 27.1 million by 2030. This escalation is particularly concerning in high-income and rapidly developing countries, where urbanization, dietary shifts, and exposure to carcinogens contribute to rising cancer rates. Furthermore, low- and middle-income countries (LMICs) bear a disproportionate cancer burden due to limited healthcare resources, delayed diagnoses, and inadequate access to treatment. The economic toll of cancer is staggering, with global healthcare costs and productivity losses exceeding $1 trillion per year, making it a critical public health challenge requiring urgent intervention.

By integrating advanced genomic technologies, precision medicine approaches, and public health initiatives, efforts to reduce cancer incidence and mortality are progressing. However, significant gaps remain in early detection, equitable treatment access, and effective prevention strategies, necessitating continued research and global collaboration (Bray *et al*., 2020)

1.2 BLOOD CANCER

Blood cancers, also known as hematologic malignancies, are a heterogeneous group of disorders that affect the production and function of blood cells. These malignancies typically originate in the bone marrow, where hematopoietic stem cells differentiate into three primary cell types: erythrocytes (red blood cells), leukocytes (white blood cells), and thrombocytes (platelets). In blood cancers, this tightly regulated process is disrupted by genetic mutations and epigenetic alterations, leading to the uncontrolled proliferation of abnormal blood cells. These malignant cells interfere with normal haematopoiesis, compromising immune function, oxygen transport, and haemostasis, which can result in severe complications such as immunosuppression, anaemia, and thrombocytopenia (Genovese *et al*., 2014).

1.2.1 Major Types of Blood Cancers

Hematologic malignancies can be broadly classified into three major categories: leukemia, lymphoma, and multiple myeloma. Each type originates from different cell lineages and exhibits distinct clinical and molecular characteristics.

Leukemia:

Leukemia is a malignant disorder characterized by the clonal proliferation of abnormal white blood cells in the bone marrow and peripheral blood. This uncontrolled expansion disrupts normal hematopoiesis, leading to a reduction in red blood cells, platelets, and functional leukocytes. The disease is further classified into acute and chronic subtypes based on the rate of progression and cell lineage affected:

- **Acute Lymphoblastic Leukemia (ALL)** – Rapidly progressing leukemia that arises from lymphoid progenitor cells, primarily affecting children and young adults.
- **Acute Myeloid Leukemia (AML)** – Affects myeloid precursor cells, leading to an aggressive disease course with poor prognosis in older adults.
- **Chronic Lymphocytic Leukemia (CLL)** – A slow-growing malignancy involving mature B lymphocytes, predominantly affecting elderly individuals.
- **Chronic Myeloid Leukemia (CML)** – Characterized by the presence of the Philadelphia chromosome (BCR-ABL fusion gene), which drives uncontrolled myeloid cell proliferation (Jabbour & Kantarjian, 2018).

Lymphoma

Lymphomas are a group of malignancies originating from lymphocytes, a subset of white blood cells that play a crucial role in immune function. These cancers primarily affect the lymphatic system, including lymph nodes, the spleen, and other lymphoid tissues. Lymphomas are broadly categorized into:

- **Hodgkin Lymphoma (HL)** – Characterized by the presence of Reed-Sternberg cells, HL has a relatively high cure rate with appropriate treatment.
- **Non-Hodgkin Lymphoma (NHL)** – A diverse group of lymphoid malignancies, including both B-cell and T-cell lymphomas. Among these, Peripheral T-cell Lymphoma (PTCL) represents an aggressive and poorly understood subtype.

Multiple Myeloma

Multiple myeloma is a malignancy of plasma cells, a specialized subset of B lymphocytes responsible for antibody production. In multiple myeloma, the uncontrolled proliferation of malignant plasma cells in the bone marrow leads to:

- Suppression of normal antibody production, resulting in immunodeficiency.
- Osteolytic lesions due to increased bone resorption.
- Renal impairment due to excessive production of monoclonal immunoglobulins (Rajkumar, 2020).

1.2.2 Peripheral T-Cell Lymphoma (PTCL)

Among non-Hodgkin lymphomas, Peripheral T-cell lymphoma (PTCL) is a heterogeneous and aggressive subgroup derived from mature, post-thymic T cells and natural killer (NK) cells. PTCL accounts for approximately 10–15% of all NHL cases and has a significantly poorer prognosis compared to B-cell lymphomas due to its molecular complexity and resistance to conventional therapies (Mehta-Shah *et al*., 2024).

PTCL arises due to oncogenic mutations and epigenetic dysregulation, leading to uncontrolled proliferation of T or NK cells. These malignant cells accumulate in lymphoid tissues such as the lymph nodes, spleen, and bone marrow, eventually infiltrating extra nodal sites such as the gastrointestinal tract, liver, and skin. Unlike B-cell lymphomas, which often have well-characterized molecular markers guiding targeted therapy, PTCL remains a therapeutic challenge due to its heterogeneity and lack of specific biomarkers (Savage & Macon, 2019)

1.2.3 Pathogenesis and Disease Progression in PTCL

The development of PTCL is associated with complex genetic alterations affecting key signaling pathways:

- **JAK-STAT Pathway** – Mutations in JAK1, JAK3, and STAT3 contribute to abnormal cytokine signaling and uncontrolled T-cell proliferation.
- **TCR Signaling Pathway** – Aberrations in T-cell receptor (TCR) signaling components, such as CD28 and ITK, lead to unchecked activation and survival of malignant T cells.
- **Epigenetic Dysregulation** – Mutations in epigenetic regulators such as TET2, DNMT3A, and IDH2 disrupt DNA methylation and histone modification, driving oncogenic transformation (Pizzi *et al*., 2018).

1.2.4 Clinical Manifestations of (PTCL)

Peripheral T-cell lymphomas (PTCLs) present with diverse clinical manifestations due to their ability to affect multiple organ systems, including the lymphatic system, skin, gastrointestinal tract, liver, and spleen. The heterogeneity in clinical presentation is influenced by the specific PTCL subtype and the extent of disease progression. Patients may exhibit both localized and systemic symptoms, which are crucial for early diagnosis and disease staging (Pileri *et al*., 2021)

1.2.5 General Symptoms

The most common clinical feature of PTCL is painless lymphadenopathy, which can occur in various anatomical regions, including:

- Cervical (neck)
- Axillary (armpit)
- Inguinal (groin)

In some cases, extra nodal involvement may lead to cutaneous lesions, gastrointestinal disturbances, or hepatosplenomegaly (Greer et al., 1984).

1.2.6 Systemic Symptoms and Disease Progression

PTCL frequently presents with constitutional symptoms, commonly referred to as B symptoms, which are indicative of an aggressive disease course and are associated with a poorer prognosis (Hari et al., 2008)

- Persistent or recurrent fever (>38°C)
- Drenching night sweats
- Unintentional weight loss (>10% of body weight within six months)

1.2.7 Clinical Implications

The presence of systemic symptoms, particularly B symptoms, often correlates with advanced disease and necessitates an aggressive therapeutic approach. Given the variability in clinical presentation, early and accurate diagnosis through histopathological examination, immunophenotyping, and molecular profiling is crucial for optimizing treatment strategies and improving patient outcomes (Mehta-Shah *et al*., 2024)

1.2.8 STAGING OF PTCL

Accurate staging is essential for determining the extent of disease progression, guiding treatment decisions, and predicting patient prognosis. Staging helps classify the anatomical distribution of malignancy and assess whether the lymphoma is localized or disseminated. Peripheral T-cell lymphoma (PTCL) is staged using a combination of clinical evaluation, laboratory tests, and imaging modalities (Barrington *et al*., 2014).

1.2.9 Diagnostic Tests for Staging

The staging process involves multiple diagnostic modalities to assess both nodal and extra nodal involvement, including:

- **Blood Tests:**
    - Complete blood count (CBC) to evaluate cytopenias
    - Lactate dehydrogenase (LDH) levels as a marker of tumour burden
    - Beta-2 microglobulin as a prognostic biomarker

- **Bone Marrow Aspiration and Biopsy:**
    - Determines whether the malignancy has infiltrated the bone marrow
    - Identifies abnormal lymphoid cells or fibrosis indicative of PTCL involvement
- **Imaging Techniques:**
    - **Computed Tomography (CT) Scan** – Detects lymphadenopathy, hepatosplenomegaly, and organ involvement
    - **Positron Emission Tomography (PET) Scan** – Utilized to assess metabolic activity of lymphoma cells and detect occult disease sites
    - **Magnetic Resonance Imaging (MRI)** – Occasionally used for central nervous system (CNS) involvement

1.2.10 Ann Arbor Staging System

The Ann Arbor Staging System, originally developed for Hodgkin lymphoma, is widely used for staging all subtypes of non-Hodgkin lymphoma (NHL)**,** including PTCL (Lister *et al*., 1989). This system classifies lymphoma into four stages based on the anatomical distribution of the disease

- **Stage I:**
    - Involvement of a single lymph node region or a single extra nodal site.
- **Stage II:**
    - Involvement of two or more lymph node regions on the same side of the diaphragm (either above or below).
- **Stage III:**
    - Involvement of lymph node regions on both sides of the diaphragm.
- **Stage IV:**
    - Disseminated or widespread disease with involvement of extra nodal sites such as the liver, bone marrow, or lungs.

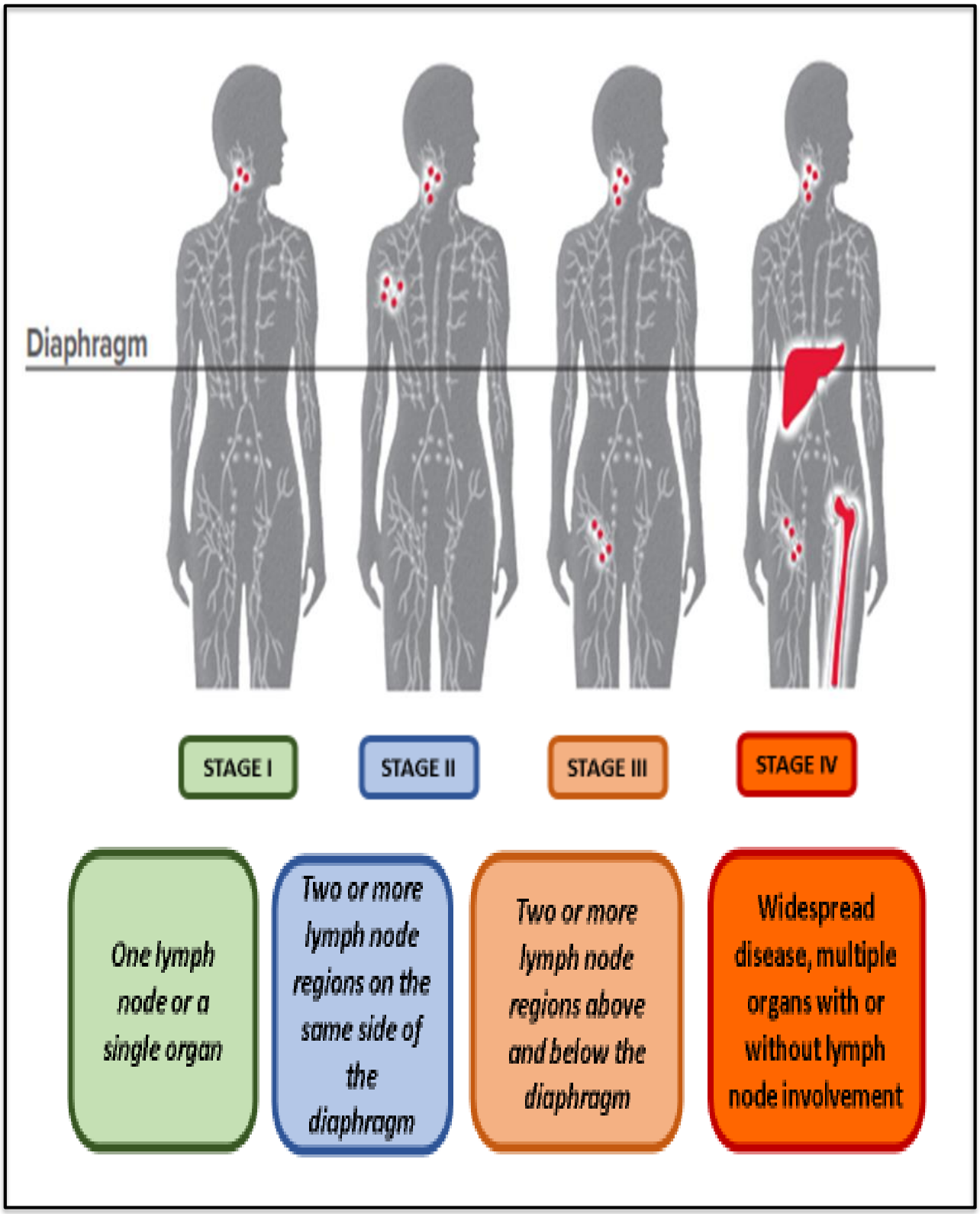

Fig 1: Stages of PTCL

1.2.11 Treatment Strategies for (PTCL)

The treatment of PTCL remains challenging due to its heterogeneity and aggressive nature. Unlike B-cell lymphomas, PTCL lacks standardized therapeutic protocols, and treatment strategies vary based on disease subtype, stage, and patient-specific factors. The primary treatment modalities include chemotherapy, radiation therapy, targeted therapy, and, in some cases, hematopoietic stem cell transplantation (HSCT) (*Stoll BTL et al., 2021*).

Chemotherapy

Chemotherapy remains the first-line treatment for most PTCL subtypes. It involves the administration of cytotoxic agents that target rapidly dividing cancer cells. The most commonly used regimen for PTCL is CHOP (cyclophosphamide, doxorubicin, vincristine, and prednisone) or its variations:

- CHOP-like regimens: Often combined with etoposide (CHOEP) in younger patients.
- Intensified chemotherapy regimens: Such as hyper-CVAD or dose-adjusted EPOCH, are considered for aggressive PTCL subtypes.

Chemotherapy is typically administered intravenously and may be combined with other modalities such as radiation therapy or immunotherapy to enhance treatment efficacy. However, chemotherapy is associated with systemic side effects, including bone marrow suppression, gastrointestinal toxicity, and increased risk of infections (Zain & O'Connor, 2020)

Radiation Therapy

Radiation therapy, or radiotherapy, utilizes high-energy ionizing radiation to induce DNA damage in cancer cells, leading to apoptosis. It is primarily used in:

- Localized PTCL (Stage I–II): Often combined with chemotherapy for improved disease control.
- Palliative care: To reduce tumour burden and alleviate symptoms in advanced-stage disease.

Advanced techniques such as intensity-modulated radiation therapy (IMRT) and proton beam therapy allow for precise targeting of malignant cells while minimizing damage to surrounding healthy tissue (Hoppe *et al*., 2017).

Targeted Therapy

Targeted therapies are designed to selectively inhibit oncogenic pathways and molecular aberrations driving PTCL progression. Unlike conventional chemotherapy, targeted agents provide greater specificity and reduced systemic toxicity. Key targeted therapies for PTCL include:

- **Brentuximab Vedotin (BV):** An antibody-drug conjugate targeting CD30, used in anaplastic large cell lymphoma (ALCL) and other CD30-expressing PTCLs.
- **Histone Deacetylase (HDAC) Inhibitors:** Such as romidepsin and belinosta**t**, which induce apoptosis in PTCL cells by modulating gene expression.
- **Tyrosine Kinase Inhibitors (TKIs):** Including alectinib and crizotinib, used for ALK-positive PTCL cases.

The development of novel small-molecule inhibitors and immunotherapeutic approaches continues to expand treatment options for refractory and relapsed PTCL (Marchi *et al*., 2022).

1.3 NEXT GENERATION SEQUENCING (NGS)

Next-Generation Sequencing (NGS) has revolutionized genomic research**,** enabling high-throughput sequencing with unparalleled speed, accuracy, and cost-effectiveness. This transformative technology has significantly expanded our understanding of genome architecture, gene regulation, and disease pathogenesis**,** facilitating both basic research and clinical applications

Unlike traditional Sanger sequencing**,** which relies on chain termination methods and is limited in scalability, NGS platforms can generate massive amounts of sequence data in a parallelized manner. This capability has led to significant advancements in multiple "omics" disciplines, including genomics, transcriptomics, epigenomics, and metagenomics**.** The advent of third-generation sequencing technologies**,** such as single-molecule real-time (SMRT) sequencing and Nano pore sequencing, has further enhanced read length, resolution, and throughput, enabling more comprehensive genomic analyses (Satam *et al*., 2023).

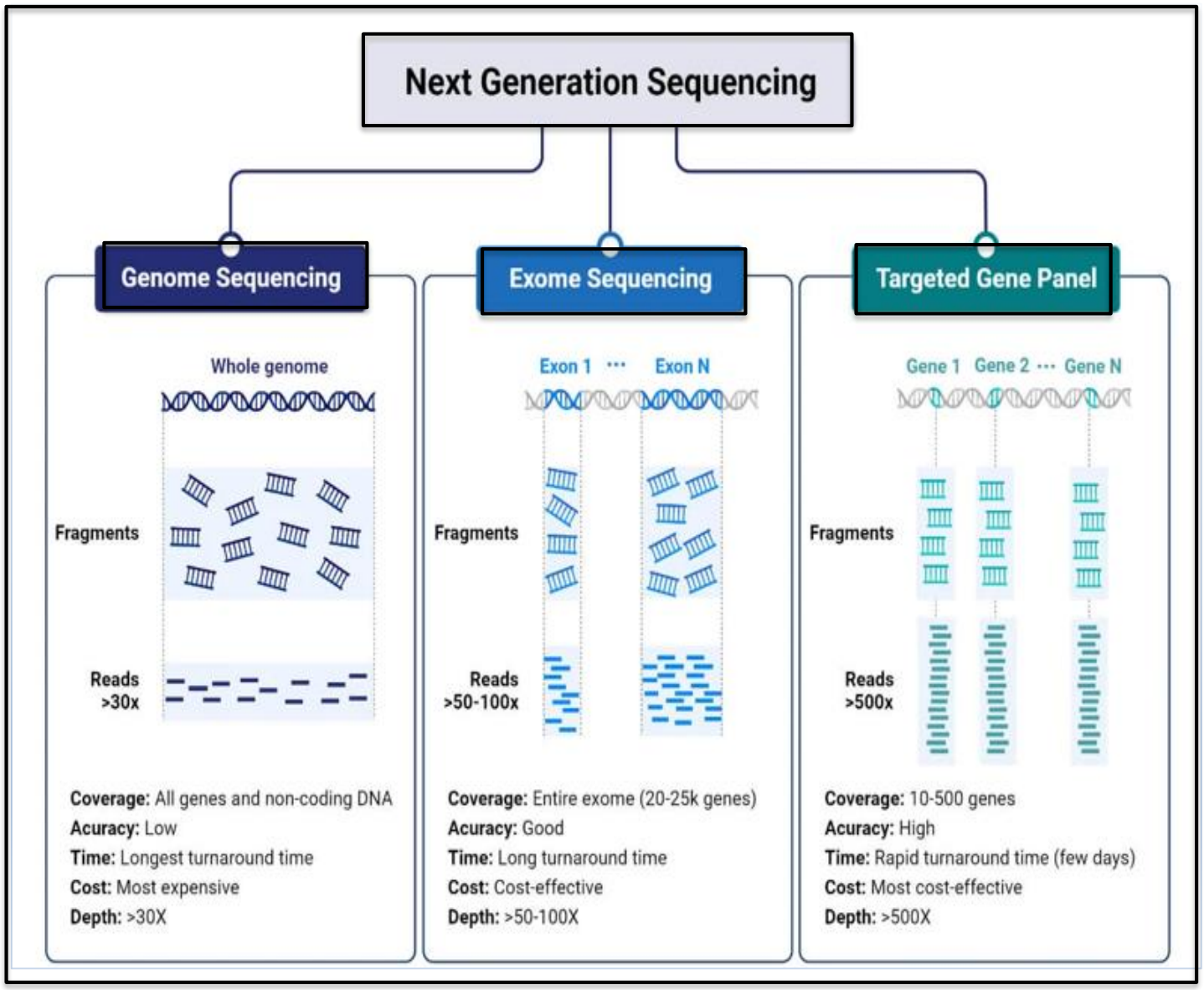

Fig 2: Overview of NGS Strategies for Genomic Analysis

1.3.1 WHOLE-EXOME SEQUENCING

Whole-exome sequencing is a widely used next-generation sequencing (NGS) method that involves sequencing the protein-coding regions of the genome. The human exome represents less than 2% of the genome, but contains ~85% of known disease-related variants, making this method a cost-effective alternative to whole-genome sequencing. Exome sequencing using exome enrichment can efficiently identify coding variants across a broad range of applications, including population genetics, genetic disease, and cancer studies. Exome sequencing is a cost-effective approach when whole-genome sequencing is not practical or necessary. Sequencing only the coding regions of the genome enables researchers to focus their resources on the genes most likely to affect phenotype, and offers an accessible combination of turnaround time and price. Exome sequencing detects variants in coding exons, with the capability to expand targeted content to include untranslated regions (UTRs) and microRNA for a more comprehensive view of gene regulation (https://www.illumina.com/).

The majority of WES computational tools are centered on the generation of a Variant Calling Format (VCF) file from raw sequencing data. Once the VCF files have been generated, further downstream analyses can be performed by other computational methods. Therefore, in this review we have classified bioinformatics methods and computational tools into Pre-VCF and Post-VCF categories. Pre-VCF workflows include tools for aligning the raw sequencing reads to a reference genome, variant detection, and annotation. Post-VCF workflows include methods for somatic mutation detection, pathway analysis, copy number alterations, INDEL identification, and driver prediction (Tetreault et al., 2015).

1.3.2 ADVANTAGES OF WES:

- Identifies variants across a wide range of applications.
- Achieves comprehensive coverage of coding regions.
- Provides a cost-effective alternative to whole-genome sequencing (4–5 Gb of sequencing per exome compared to ~90 Gb per whole human genome)
- Produces a smaller, more manageable data set for faster, easier data analysis compared to whole-genome approaches.

Table 1.1: Summary of the differences between the solution-based, exome-sequencing platforms

Solution based	NimbleGen's SeqCap EZ Exome Library	Agilent's Sure Select Human All Exon Kit	Illumina's TruSeq Exome Enrichment Kit	Illumina's NextEra Rapid Capture Exome Kit
Probe size, bp[a]	55−105	114−126	95	95
Probe type	DNA	RNA	DNA	DNA
Coverage strategy	High-density, overlapping probes	Adjacent probes	Gaps between probes	Gaps between probes
Fragmentation method	Ultra sonication	Ultra sonication	Ultra sonication	Transposomes
Target region size (human), Mb[b]	64	50	62	62
Reads remaining after filtering	66%	71.70%	54.80%	40.10%
Major Strengths	(i) High sensitivity and specificity (ii) Most uniform coverage in difficult regions	(i) Better coverage of indels (ii) High alignment rate (iii) Fewer duplicate reads than other platforms	(i)Good coverage of UTRs and miRNAs	(i)Good coverage of UTRs and miRNAs
Non-human supported	Yes	Yes	No	No

1.3.3 WHOLE GENOME SEQUENCING (WGS):

Detectable mutations by each of the genomic analysis platforms (DNA chip, target sequencing for 100 genes, WES, RNA-Seq and WGS) and their performances are summarized. WES analysis captures protein-coding exons spanning approximately 50 Mb (1%-2%) of the human genome by in-solution hybridization, microarray capturing or PCR amplification, and usually sequences approximately 100 sequence depth for each sample, which is more accurate than 30 WGS, because the accuracy of mutation calling by NGS is primarily dependent on the sequencing depth. However, some capturing bias expected is, for example due to difficulty detecting complicated or repetitive genomic regions as well as non-targeted regions. On the other hand, WGS is technically straightforward. DNA is randomly fragmented by physical shearing, and 30-50x sequence depth (90-150 Gb) of each human whole genome is usually sequenced for both cancer and normal genomes, which can cover 99% of the entire human genome.

Common NGS technology reads are 100-150 bp for both ends of a 500-600-bp DNA fragment, but WGS by NGS is still dependent on PCR, with PCR bias, indicating that GC-rich or AT-rich regions are difficult to sequence. PCR-free protocol was recently developed, which shows less GC bias and is more comprehensive than the PCR protocol, although some µg DNA is required as an input for library preparation. The largest limitation of the present 2nd NGS technology (Illumina SBS technology) is its short-read length (100-250 bp). Around 50% of the 3-Gb human genome is occupied by repetitive regions and pseudogenes in its 50%, and when short sequence reads are aligned to the redundant reference genome, alignment errors can occur around these repetitive or complicated regions, leading to mutation calling errors (Park & Kim, 2016)

1.3.4 RNA SEQUENCING:

In addition to NGS, there is third-generation sequencing, which allows for long-read sequencing of individual RNA molecules. Single-molecule RNA sequencing enables the generation of full-length cDNA transcripts without clonal amplification or transcript assembly. Thus, third-generation sequencing is free from the shortcomings generated by PCR amplification and read mapping. It can greatly reduce the false positive rate of splice sites and capture the diversity of transcript isoforms. Single-molecule sequencing platforms comprise Pacific Biosciences (Pac Bio) single-molecule real-time (SMRT) sequencing, Helicos single-molecule fluorescent sequencing and Oxford Nano pore Technologies (ONT) Nano pore sequencing.

Furthermore, RNA-seq recently evolved from bulk sequencing to single-cell sequencing. Single-cell RNA sequencing was first published in 2009 to profile the transcriptome at single-cell resolution. Drop-Seq and in drop were initially reported in 2015 by analysing mouse retina cell and embryonic stem cell transcriptomes, identifying novel cell types. Sci-RNA-seq, single-cell combinatorial indexing RNA sequencing, was developed in 2017, and SPLIT-seq (split-pool ligation-based transcriptome sequencing) was first reported in 2018 (Mingye Hong *et al*., 2020)

1.4 MUTATION:

Mutations are alterations in the genetic sequence that contribute to the genetic diversity observed among organisms. These changes can occur at various levels, from single nucleotide substitutions to large-scale chromosomal rearrangements, and they exhibit a wide range of biological consequences**.** Mutations can be heritable or somatic, depending on whether they occur in germline cells (sperm or eggs) or somatic cells. Only mutations occurring in germline cells are passed on to offspring, influencing evolutionary processes and genetic inheritance.

Mutations play a crucial role in biological evolution, as they introduce genetic variation that can be acted upon by natural selection. While some mutations confer adaptive advantages, others may be neutral or deleterious, leading to genetic disorders or cancer. The interaction between mutational events and environmental pressures is fundamental to species adaptation and evolution (Loewe, 2008).

1.4.1 TYPES OF MUTATIONS:

Mutations can be classified into several categories based on their molecular characteristics and functional consequences.

- **Point Mutations**

Point mutations involve a single nucleotide substitution in the DNA sequence. These mutations can have varying effects on protein synthesis, depending on the nature of the substitution:

- **Silent Mutation:** A nucleotide substitution that does not alter the amino acid sequence of the encoded protein. This occurs due to the redundancy of the genetic code (codon degeneracy).
  - Example: A change from GTA (valine) to GTT (valine) still codes for the same amino acid.
- **Missense Mutation:** A nucleotide substitution that leads to the incorporation of a different amino acid in the protein sequence, potentially altering protein function.
  - Example: CCC (proline) → ACC (threonine) results in a missense mutation, which may affect protein stability or activity.
- **Nonsense Mutation:** A nucleotide substitution that converts a codon into a premature stop codon, leading to early termination of protein synthesis. This typically produces a truncated, non-functional protein, particularly if the stop codon occurs far upstream of the normal termination site.

- **Insertion and Deletion (Indels):**

Insertion and deletion mutations involve the addition or removal of one or more nucleotides from the DNA sequence. These mutations can have profound effects on gene expression and protein function:

- **Frameshift Mutation:** If the insertion or deletion is not a multiple of three, it alters the reading frame of the genetic code. This disrupts the entire downstream amino acid sequence, usually resulting in a non-functional protein.
  - Example: Inserting a single nucleotide in the middle of a coding sequence changes all subsequent codons, producing an entirely different protein structure.
- **In-frame Mutation:** If the insertion or deletion involves a multiple of three nucleotides, the reading frame remains unchanged. However, the resulting protein may have extra or missing amino acids, potentially affecting its structure and function.

1.4.2 VARIANT:

Variant calling is a fundamental process in next-generation sequencing (NGS) data analysis, aimed at identifying genetic differences between a reference genome and the genome of an individual or a population. These genetic differences, collectively referred to as variants, play a crucial role in understanding genomic diversity, disease mechanisms, and evolutionary processes**.** Variant calling enables researchers to detect genetic mutations associated with inherited diseases, cancer, and complex traits, thereby facilitating precision medicine and population genetics studies (Pabinger *et al*., 2014)

1.4.3 EXONS:

Exons are the protein-coding regions of a gene, forming essential segments of messenger RNA (mRNA) after transcription. In eukaryotic genes, the initial pre-mRNA transcript contains both exons (coding regions) and introns (non-coding regions**)**. Through a process known as RNA splicing, introns are removed**,** and exons are joined together to form the mature mRNA transcript, which is subsequently translated into a protein (Corvelo & Eyras, 2008)

CHAPTER II

REVIEW OF LITERATURE

Blood cancer is a particularly aggressive form of cancer that originates from bone marrow cells, leading to the abnormal proliferation of blood cells and disrupting the production of normal blood cells. Its incidence is increasing every year due to a combination of genetic and environmental factors. While advancements in dose-intensification techniques have shown promising results in younger patients, the prognosis for older patients remains poor. Efforts to reduce cancer incidence and mortality involve not only the development of improved treatment strategies but also initiatives focused on prevention, early detection, and better access to healthcare services. Early diagnosis is crucial, as it significantly enhances treatment outcomes and survival rates (Tetreault *et al.*, 2015).

Blood cancers are broadly classified into four major types: leukemia, myeloma, Hodgkin lymphoma, and non-Hodgkin lymphoma. Collectively, these cancers account for approximately 9% of all malignancies worldwide, making them the fourth most common cancer type in both males and females (McCabe *et al*., 2015).

However, these categories include over 60 distinct subtypes, each characterized by unique immunophenotypic, genetic, and clinical features. The World Health Organization (WHO) first introduced a standardized classification system for hematologic malignancies in 2001, refining it further in 2008 to incorporate genetic abnormalities and clinical characteristics, which has helped improve targeted therapies. Among blood cancers, leukemia is influenced by various genetic, environmental, and lifestyle factors. Risk factors include smoking, exposure to harmful chemicals such as benzene, previous chemotherapy or radiation therapy, congenital conditions, pre-existing blood disorders, age, gender, and a family history of leukemia. The development of novel therapeutic strategies and targeted drugs has significantly improved overall survival rates among leukaemia patients. However, leukaemia's epidemiology continues to evolve, with variations observed across different population groups and regions. To develop effective prevention strategies, it is essential to examine global disease distribution, emerging risk factors, and evolving trends (Huang *et al*., 2022).

Peripheral T-cell lymphomas (PTCLs) are a rare and heterogeneous group of haematological malignancies with a poor prognosis across almost all subtypes. Their diverse clinicopathological features make accurate diagnosis, prognosis, and the selection of optimal treatment strategies particularly challenging. The absence of standardized treatment protocols, coupled with the reliance on therapeutic approaches extrapolated from B-cell lymphomas, further complicates disease management. However, some advancements have been made, notably with CD30 monoclonal antibody therapy, which has shown improvements in progression-free and overall survival, particularly in anaplastic large-cell lymphomas (Luminari *et al*., 2021).

Historically, significant progress in non-Hodgkin lymphoma (NHL) treatment has primarily benefited patients with B-cell lymphoma, while PTCLs have not seen comparable advancements in outcomes over the past two to three decades. In the United States alone, approximately 80,000 cases of NHL and 14,000 cases of Hodgkin lymphoma are diagnosed annually. As a subgroup of NHL, PTCLs account for only 6% to 10% of cases, making them exceptionally rare. Furthermore, the World Health Organization's 2017 classification defines 29 distinct PTCL subtypes, highlighting the intrinsic diversity of the disease. This rarity, combined with its complex molecular landscape, has hindered progress in treatment development. Most PTCL subtypes also have a significantly worse prognosis than their B-cell counterparts, largely because treatment paradigms for PTCL have been derived from B-cell neoplasms, which differ fundamentally in biology. In fact, the first drug specifically approved for PTCL treatment was introduced only a decade ago (Marchi *et al*., 2019).

PTCLs originate from post-thymic lymphocytes and exhibit heterogeneous clinic pathologic features. The 2016 WHO classification describes 27 distinct PTCL subtypes, categorizing them based on clinic pathological characteristics and immunohistochemical markers. Recent technological advancements, including gene expression profiling and whole-genome sequencing, have enhanced our understanding of PTCL by aiding in subtype differentiation and uncovering the molecular mechanisms underlying disease pathogenesis. Emerging data suggest that distinct molecular signatures may define prognostic groups, paving the way for more targeted and rational treatment strategies in the future (Zain *et al*., 2021).

Peripheral T-cell lymphoma, not otherwise specified (PTCL, NOS), is a heterogeneous and aggressive neoplasm without distinct genetic, immunological, or clinical features. Although several morphological subtypes have been described, no specific markers define this disease. Patients commonly present with symptoms such as night sweats, fever, lymphadenopathy, weight loss, splenomegaly, and skin changes. Laboratory findings often reveal anaemia, thrombocytosis, lymphocytosis, eosinophilia, hypergammaglobulinemia, or elevated lactate dehydrogenase levels. In a case study, a patient exhibited massive lymphadenopathy and right lower limb swelling over six weeks. A tissue biopsy and supporting investigations confirmed the diagnosis of PTCL, NOS (Amador *et al*., 2025).

Peripheral T-cell lymphoma (PTCL) is a rare but diverse group of haematological malignancies classified as mature T-cell non-Hodgkin's lymphomas (NHL). It is an aggressive disease associated with poor prognosis, accounting for approximately 5–10% of all NHL cases. Compared to B-cell tumours, our understanding of T-cell leukemia and lymphoma remains limited, and most patients are diagnosed at an advanced stage. The World Health Organization (WHO) classifies PTCL into nearly 30 subtypes, which are broadly categorized as nodal, extra nodal, or leukemic forms. These subtypes often manifest as aggressive diseases, with an estimated five-year survival rate of around 30%. In contrast, cutaneous PTCL generally progresses more slowly. The WHO classification of hematolymphoid tumours (WHO-HAEM5) integrates new molecular and histopathological insights, enhancing the diagnostic classification of PTCL (Luan *et al.*, 2024).

The standard treatment for lymphoma involves chemotherapy, either alone or combined with radiotherapy, while radiotherapy alone is not recommended due to its toxicity. Long-term complications from radiotherapy include an increased risk of secondary cancers, such as breast or lung cancer. Additionally, chemotherapy-treated patients may develop malignancies like melanoma, acute myeloid leukemia, or other secondary cancers. Older patients, particularly those over 60, generally have poorer prognoses regardless of disease stage. The National Comprehensive Cancer Network (NCCN) advises against using certain chemotherapeutic agents in patients above 60 years due to increased toxicity risks. Physicians should prioritize shared decision-making, especially for elderly patients, ensuring they understand their treatment options and potential outcomes before pursuing therapy (Lewis *et al*., 2020).

Whole-exome sequencing (WES) is a targeted sequencing approach that focuses on capturing and sequencing the protein-coding regions of the genome, collectively known as the exome. Although the exome constitutes only about 1–2% of the entire genome, it harbors the majority of known disease-causing variants. By sequencing these regions, WES enables the identification of genetic variations, including single-nucleotide variants (SNVs), insertions, deletions, and copy number variations (CNVs) within protein-coding genes. It is a cost-effective alternative to whole-genome sequencing (WGS) for diagnosing rare clinical diseases with distinct symptom clusters and for studying genetic variants in population and cancer genetics. WES relies on enrichment techniques such as hybrid capture or target-specific amplification, followed by high-throughput sequencing. Several exome capture assays from NimbleGen, Agilent, Illumina, Twist, and IDT are available, all of which are compatible with the Illumina NGS platform. The bioinformatics pipeline used for WES data analysis closely follows that of WGS since WES is essentially a subset of WGS (Satam *et al.,* 2023).

Whole-exome sequencing (WES) data analysis requires multiple bioinformatics tools and pipelines to process raw sequencing data, identify variants, and interpret their significance. The analysis typically involves several key steps, including quality control, read alignment, variant calling, annotation, and functional interpretation**.** Read alignment is often performed using the Burrows-Wheeler Aligner (BWA) or Bowtie2, followed by variant calling using tools such as GATK, freebayes, or samtools/bcftools. The called variants are then annotated with functional information using annotation tools like ANNOVAR, snpeff, or VEP (Variant Effect Predictor)**.** For interpretation, various databases are used to classify and assess the significance of identified variants. Clinvar provides information on the clinical relevance of variants based on expert curation and reported cases. COSMIC (Catalogue of Somatic Mutations in Cancer) is a widely used resource for identifying cancer-related mutations, while dbsnp serves as a reference database for common single nucleotide polymorphisms (snps). These bioinformatics pipelines allow researchers to filter out benign polymorphisms and focus on pathogenic or likely pathogenic variants relevant to disease studies, including PTCL (Iqbal *et al.*, 2019).

The interpretation of PTCL-related variants is a crucial step in WES analysis, distinguishing between pathogenic and benign mutations**.** Pathogenic variants are genetic alterations that have been established to contribute to disease progression, whereas benign mutations are typically neutral with no known clinical consequences. The classification of these variants follows guidelines from the American College of Medical Genetics and Genomics (ACMG), which considers factors such as population frequency, functional studies, and computational predictions. Certain driver mutations in PTCL have been linked to oncogenic pathways, including mutations in *TET2, RHOA, DNMT3A, IDH2, TP53*, and *STAT3***,** which contribute to altered epigenetic regulation and aberrant signalling in T-cell lymphomas. The presence of such mutations can influence prognosis, with some alterations correlating with poor survival outcomes. For instance, *TP53* mutations have been associated with increased disease aggressiveness, while *RHOA* mutations play a key role in angioimmunoblastic T-cell lymphoma (AITL) (Palomero *et al*., 2014).

Whole-exome sequencing has rapidly emerged as an effective and accurate tool for analysing the protein-coding regions of the genome, with numerous studies demonstrating its impact on rare disease diagnosis and clinical applications. Despite its effectiveness, WES identifies a causative variant in only about 25% of cases. While this diagnostic rate may seem low, it must be considered in the context of challenges such as limited power to detect multigenic effects and the potential influence of non-exonic variants, including deep intronic or regulatory mutations that may contribute to disease. Notably, exome studies investigating both rare and more common disorders, such as autism and sporadic schizophrenia, have revealed a remarkably high rate of de novo mutations, highlighting the complexity of genetic contributions to these conditions (Arjen *et al*., n.d)

Peripheral T-cell lymphoma (PTCL) is an aggressive and highly malignant form of lymphoma characterized by its poor prognosis and high mortality rate. It is distinguished by its strong specificity and primarily develops after the maturation of NK and T cells in the thymus. The disease's violent nature and clinical complexity make it a challenging malignancy to diagnose and treat, necessitating further research into its genetic and molecular underpinnings (Vasmatzis *et al*., 2019).

A study was conducted to refine and adapt previous preliminary bioinformatics algorithms to detect de novo chromosomal rearrangements in peripheral T-cell lymphoma (PTCL). Out of 21 PTCL patients analysed, 13 were classified with recurrent chromosomal rearrangements. Among these, five genetic modifications affected p53-related genes, including *ANKRD11, WWOX, CDKN2A, TP63,* and *TP53*. The rearrangement involving *TP63* is particularly significant, as it results in a fusion protein containing a truncated form of p63 (Np63), which closely resembles the normal Np63 subtype. This fusion protein exhibits strong carcinogenic properties and inhibits the *p53* tumour suppressor pathway through a dominant-negative mechanism. Approximately 5.8% of patients exhibited these rearrangements, which were associated with poor survival outcomes, particularly in cases with ALK-negative mutations. The *TP63* rearrangement was specifically correlated with worse prognosis in PTCL patients. Unlike other malignant neoplasms, PTCL rarely results from *TP53* mutations. Instead, findings suggest that in PTCL, the tumour-suppressive function of *p53* may be inactivated through alternative genetic aberrations. In one study, sequencing was performed on patient samples from angioimmunoblastic T-cell lymphoma (AITL, n=30), anaplastic large cell lymphoma (ALCL, n=21), and PTCL-not otherwise specified (PTCL-NOS, n=12). The key signalling proteins *pSTAT3*, *pMAPK*, and p*AKT* were analysed through amplification and sequencing, along with immunohistochemically staining to assess their expression patterns (Zhu *et al*., 2021).

WES has significantly influenced personalized medicine by guiding targeted therapy decisions for PTCL patients. By identifying actionable mutations, WES enables the selection of therapies that specifically target altered molecular pathways. For instance, PTCL cases with mutations in the JAK-STAT pathway may benefit from JAK inhibitors, while patients with alterations in epigenetic regulators such as *TET2* and *DNMT3A* might be candidates for hypo ethylating agents**.**

Additionally, WES plays a role in novel drug discovery by uncovering new oncogenic drivers that can be exploited for therapeutic development. For example, studies have identified recurrent RHOA mutations in AITL, suggesting that inhibitors targeting downstream signalling could be potential treatment options. Furthermore, immune checkpoint inhibitors, such as PD-1/PD-L1 inhibitors, are being explored in PTCL cases where immune evasion mechanisms are evident through WES findings (Vasmatzis *et al*., 2019).

Initial studies comparing the mutational burden between liquid and tissue biopsies in peripheral T-cell lymphoma (PTCL) utilized whole-exome sequencing (WES) of synchronous peripheral blood mononuclear cells (PBMNCs) and lymph node tissue from a single patient with angioimmunoblastic T-cell lymphoma (AITL) (PTCL_02). Mutations in genes previously associated with PTCL, including *TET2*, *IDH2*, and *VAV1*, were identified in the lymph node tissue. Additionally, a novel mutation in *IL12RB1* exon 10, CA1068T (pArg356Ser), was detected. This mutation affects a cytokine receptor critical for T-cell function. Interestingly, the PBMNCs did not exhibit mutations in *TET2*, *IDH2*, or *VAV1*; however, they did reveal a *RHOA* mutation (c.73A>G), predicted to result in a p.Phe25Leu amino acid change. This specific mutation has not been previously reported, although a similar substitution, p.Phe25Val, has been described in lung cancer. Phenylalanine at residue 25 is part of an α-helix that interacts with phenylalanine at position 171. The replacement with bulky aliphatic side chains, such as valine or leucine, is predicted to disrupt this interaction, potentially altering the function of *RHOA* (Zhu, L., *et al*, 2021)

This case demonstrated presumptive evidence for peripheral blood lymphoma cells. However, previous studies have suggested that circulating tumour DNA is present in 98% of diffuse large B-cell lymphoma (DLBCL) patients, whereas circulating lymphoma cells are detectable in only approximately 50% of peripheral T-cell lymphoma (PTCL) patients. Therefore, in this study, DNA was chosen over peripheral blood mononuclear cells (PBMNCs) because only a fraction of patients have detectable circulating lymphoma cells, whereas DNA is likely to be present in nearly all cases and has been shown to reflect global tissue mutational heterogeneity in haematological malignancies (Barbara Ottolini, 2020).

Empirical priors were constructed to determine the distribution of variant frequencies for each sample. High-credibility intervals (posterior probability $\geq 1–10^{-5}$) were obtained to assess frequency changes between tumour and normal samples using the SAVI (Statistical Algorithm for Variant Identification) algorithm developed at Columbia University. The number of germline SNPs in the coding regions was approximately 18,000, which was comparable to previous reports. Most candidate germline SNPs (approximately 16,000, or 90%) were reported in the dbSNP database. Candidate somatic variants were identified using the following criteria: total variant depth in both tumour and normal samples between 10× and 300×, variant frequency above 15% in tumour and below 3% in normal samples, and at least a 1% frequency change between normal and tumour samples with high posterior probability (≥

$1–10^{-5}$). To further eliminate systematic errors, all variants detected in any normal cases were excluded (Teresa Palomero *et al*., 2014).

There are limited studies on mutation profiling for Peripheral T-cell lymphomas (PTCL) in the Chinese population. In this study, we retrospectively analyzed the clinical and genetic landscape of 66 newly diagnosed Chinese patients. Targeted next-generation sequencing (NGS) was performed on tissue samples from these patients. At least one mutation was detected in 60 patients (90.9%), with a median of three mutations per patient (range: 0–7), and 32 cases (48.5%) harbored more than four mutations. The most frequently mutated genes included *TET2, RHOA, DNMT3A, IDH2, TP53, STAT3,* and *KMT2D.* Functionally, the most prevalent mutations were associated with epigenetic regulation and signal transduction pathways.

Clinical factors such as an International Prognostic Index (IPI) score ≥2, Prognostic Index for T-cell lymphoma (PIT) score ≥2, and failure to achieve partial remission (PR) were associated with inferior progression-free survival (PFS) and overall survival (OS). Multivariate analysis identified *TP53* mutations as an adverse factor for PFS (HR = 3.523; 95% CI, 1.262–9.835; $p = 0.016$) and KMT2D mutations as an adverse factor for OS (HR = 10.097; 95% CI, 1.000–101.953; $p = 0.048$). Mutation profiling could help distinguish distinct PTCL subtypes and serve as a valuable tool for guiding treatment decisions and prognostic assessments (Lingling Wan *et al*., 2024).

The phenotypic presentation of individuals with multiple primary tumours is often heterogeneous, complicating the establishment of a genetic diagnosis. The absence of a genetic diagnosis may lead to inappropriate surveillance recommendations and suboptimal treatment choices. This study aimed to determine whether whole-exome sequencing (WES) and variant prioritization across cancer predisposition genes could identify pathogenic variants that explain the phenotypes of individuals who developed multiple primary tumours. We analyzed exome-based cancer predisposition gene testing in 72 individuals who developed multiple primary tumours (both malignant and benign) before the age of 65 years**.** A germline pathogenic variant (gPV) in a cancer-predisposing gene was identified in 9.7% of individuals **(***CHEK2, FANCM, NF1, POT1, and PTEN***),** while 4.2% of individuals harbored a candidate variant ***(****HOXB13, MAX,* and *RECQL4***).** Additionally, by analyzing variants in genes involved in cancer-associated pathways, we identified *RECQL5* as a novel candidate gene for further study. In

conclusion, this study highlights that exome-based cancer predisposition gene testing may aid in the identification of pathogenic variants in individuals who develop multiple primary tumours, providing insights for genetic counselling, risk assessment, and clinical management (Snezana Hinic *et al*., 2019).

Despite its advantages, WES faces several technical and biological challenges in PTCL research. One major limitation is the depth and accuracy of sequencing**;** as certain low-frequency variants may be missed due to insufficient coverage. Additionally, errors in read alignment and variant calling can lead to false-positive or false-negative results, requiring validation through orthogonal methods such as Sanger sequencing or droplet digital PCR (ddPCR)**.**

From a biological standpoint, tumour heterogeneity presents a significant challenge in PTCL. The presence of multiple sub clones within a single tumour can complicate variant detection, as different regions of the tumour may harbor distinct mutations. Moreover, circulating tumour DNA (ctDNA) analysis has shown that not all PTCL cases shed detectable DNA into the bloodstream, limiting the utility of liquid biopsies in some patients (Ottolini, 2020).

Another key limitation is the inability of WES to capture non-coding regulatory variants that may influence gene expression. Since WES focuses on protein-coding regions, it excludes deep intronic mutations, enhancers, and other non-coding elements that may contribute to PTCL pathogenesis. Whole-genome sequencing (WGS) offers a more comprehensive approach but is significantly more expensive and computationally demanding (Hinic *et al*., 2019).

CHAPTER -III

3.AIM AND OBJECTIVES

3.1 AIM:

The goal of the study aims to characterize the molecular landscape of PTCL using whole exome sequencing by integrating variant identification, functional annotation, pathway enrichment, disease association, and tissue-specific expression analysis. Through a systematic bioinformatics approach, this research will provide novel insights into the genetic mechanisms underlying PTCL and identify potential biomarkers for targeted therapeutic interventions.

3.2 OBJECTIVES:

- To perform whole exome sequencing-based molecular characterization of peripheral T-cell lymphoma by identifying and analyzing missense variants.
- To systematically annotate the identified missense variants using bioinformatics tools to determine their potential biological significance.
- To assess the functional impact of missense variants through gene ontology enrichment analysis, focusing on biological processes and molecular functions.
- To identify pathways disrupted by the missense variants using pathway enrichment analysis with Reactome and WikiPathways databases.
- To explore the association of mutated genes with disease association databases such as DISEASES from Jensen Lab.
- To investigate tissue-specific expression patterns of genes harboring missense variants to determine their relevance in lymphoid tissues and PTCL pathogenesis

CHAPTER -IV

METHODOLOGY

4. SYSTEM CONFIGURATION:

The Whole Exome Sequencing (WES) analysis was performed on a Lenovo ThinkPad T560 running Ubuntu, featuring an Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz, 16GB RAM, and a 64-bit architecture. This setup ensured efficient processing of sequencing data, including alignment, variant calling, and annotation

4.1 COMPUTATIONAL ENVIRONMENT SETUP

To ensure reproducibility and efficient package management, Miniconda was used to create an isolated environment for exome sequencing data analysis. Miniconda, a lightweight version of Anaconda, allowed for efficient dependency management and package installation.

4.1.1 COMMANDS

Command for creating a Conda environment

```
➢ conda activate exome_seq_env
```

4.2 INSTALLING COMMANDS FOR PACKAGES

PACKAGES	INSTALLING COMMANDS
MINICONDA	▪ *wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh* ▪ *bash Miniconda3-latest-Linux-x86_64.sh*
FASTQC	• *conda install -c bioconda fastqc*
FASTP	• *conda install -c bioconda fastp*
BWA	▪ *conda install -c bioconda bwa*
SAM TOOLS	▪ *conda install -c bioconda samtools*
PICARD	▪ *conda install -c bioconda picard*
FREEBAYES	▪ *conda install -c bioconda free bayes*
GATK4	▪ *conda install -c bioconda gatk4*

4.3 PURPOSE OF PACKAGES

PACKAGES	PURPOSE
Miniconda	Used to create an isolated environment for exome sequencing data analysis.
FASTQC	Used to quality control checks on raw sequence data coming from high throughput sequencing pipeline.
FASTP	Used for quality control of FASTQ data, performing functions like quality profiling, adapter trimming, read filtering, and base correction.
BWA	BWA (Burrows-Wheeler Aligner) is a fast and efficient tool for aligning short DNA/RNA reads to a reference genome. It accurately handles mismatches and gaps, making it essential for NGS analysis
SAM TOOLS	SAM tools is a software package for processing SAM/BAM files, enabling format conversion, sorting, merging, and PCR duplicate removal. It is essential for NGS data analysis and alignment handling
PICARD	Picard is performs tasks like duplicate marking, sorting, merging, and quality score recalibration and aiding in variant calling
FREEBAYES	FreeBayes is a Bayesian haplotype-based variant caller that identifies genetic variations (SNPs, indels, MNPs, etc.) by analyzing short-read alignments (BAM files) against a reference genome, determining the most probable genotypes at each position.
GATK4	GATK4 used to filter low-quality or false-positive variants from VCF files based on criteria like depth **(**DP), quality score (QUAL), and strand bias (FS, MQ**)**. It ensures high-confidence variant calls by removing sequencing artifacts and errors. This step is essential for accurate downstream analysis in WES

4.4 DATA COLLECTION

The Sequence Read Archive (SRA) is a database maintained by the National Center for Biotechnology Information (NCBI) that stores sequence data generated by next-generation sequencing. SRA is the world's largest library of high-throughput sequencing data that is open to the public. The SRA is a large database of experimental DNA and RNA sequences that exhibit genomic diversity across the tree of life. Researchers can use this database to find sequence reads for additional analysis by searching metadata for those sequences. Through the STRIDES Initiative, SRA data is available on Google Cloud Platform and Amazon Web Services clouds. These cloud services enable access to and computation for any publicly available, unassembled read data as well as authorized-access to human data. To improve reproducibility and make new discoveries easier, SRA saves raw sequencing data and alignment metadata. The whole-exome sequencing datasets of Blood cancer were collected from the SRA database. The retrieved datasets were sequenced using Illumina HiSeq 4000 and it is a paired-end sequence. The retrieved data was submitted by Weill Cornell Medicine.

SRA ID: 1) *SRR31851783*

2)*SRR31851784*

4.5 DATA ANALYSIS

4.5.1 QUALITY CONTROL

FastQC is a tool that analyses high-throughput sequencing files for potential issues. It performs a series of analysis on one or more raw sequence files in fastq or bam format, and then generates a report summarizing the results. FastQC will highlight any areas where this library appears to be out of the ordinary and where you should investigate more. The tool is not restricted to a single type of sequencing technique and can be used to examine libraries generated by a wide range of experiments (Genomic Sequencing, ChIP-Seq, RNA-Seq, BS-Seq etc.). The analysis of exome sequencing data involves assessing the quality of the raw sequencing reads. The raw data, typically in FASTQ format, contains millions of reads that may include sequencing errors, adapter contamination, and low-quality bases. To evaluate the quality of the raw data, FastQC was employed.

FastQC provides a comprehensive report on various quality metrics, including per-base sequence quality, per-sequence quality scores, GC content, sequence length distribution, and the presence of adapter sequences.

4.5.2 COMMANDS

Running FastQC

- *fastqc SRR31851783_1.fastq.gz SRR31851783_2.fastq.gz*

4.6 TRIMMING

4.6.1 FASTP

Fastp is a fast and versatile tool that automates quality control and adapter trimming for high-throughput sequencing data. It supports both single-end and paired-end sequencing reads and offers additional capabilities such as quality filtering, base correction, and polyG tail trimming. Fastp is optimized for speed and efficiency, utilizing multi-threading to process large datasets quickly. For paired-end data, Fastp ensures synchronization and can perform slight trimming at the 3' end to prevent alignment issues, especially with Bowtie2 or BWA. It accepts both standard and gzip-compressed FASTQ files and generates detailed HTML and JSON reports for quality assessment. Additionally, FastQC-style quality control can be run automatically after trimming, providing a comprehensive overview of sequencing quality. Following quality assessment, adapter sequences and low-quality reads were trimmed using f Fastp, a fast and efficient tool designed for pre-processing high-throughput sequencing data. Fastp performs adapter trimming, quality filtering, and length filtering in a single step, making it highly efficient for large datasets. The tool uses a sliding window approach to trim low-quality bases from the 5' and 3' ends of reads, ensuring that only high-quality reads are retained for alignment.

4.6.2 COMMANDS

Trimming FASTQ file

- *fastp -i SRR31851783_1.fastq.gz –I SRR31851783_2.fastq.gz –o SRR31851783_1_trimmed.fastq.gz –O SRR31851783 _2_trimmed.fastq.gz*

4.7 ALIGNMENT

4.7.1 BURROWS-WHEELER ALIGNER (BWA)

BWA is a tool used to map low-divergent sequences against the reference genome. BWA-backtrack, BWA-SW, and BWA-MEM are the three algorithms included. The first algorithm is optimised for Illumina sequence reads of up to 100bp, while the other two are optimised for larger sequences of 70bp to 1Mbp. Long read support and split alignment are shared by BWA-MEM and BWA-SW, although BWA MEM, the most recent, is often favoured for high-quality queries since it is faster and more accurate. For 70-100 bp Illumina readings, BWA-MEM also outperforms BWA-backtrack.

By default, BWA finds an alignment within edit distance 2 to the query sequence, except for disallowing gaps close to the end of the query. It can be tuned to find a fraction of longer gaps at the cost of speed and of more false alignments. BWA is known for its speed. In 20 minutes, 2 million high-quality 35bp short reads may be mapped against the human genome. Typically, speed is obtained at the expense of a large amount of memory, allowing gaps and/or imposing strict limits on maximum read length and maximum mismatches. BWA is not one of them. It still has a small memory footprint (2.3GB for human alignment), does gapped alignment, and has no hard limits on read length or maximum mismatches. Once the raw reads were pre-processed, the next step were aligned them to a reference genome. For this study, the human reference genome (GRCh38) was used. Alignment was performed using Burrows-Wheeler Aligner (BWA-MEM), a widely used tool for mapping sequencing reads to large reference genomes. The trimmed reads were aligned to the human reference genome (GRCh38) using Burrows-Wheeler Aligner (BWA-MEM), a widely used and accurate alignment tool.

4.7.2 BWA-INDEX

BWA-MEM is well-suited for aligning high-throughput sequencing data due to its ability to handle long reads and its sensitivity in detecting both small and large variants. Prior to alignment, the reference genome was indexed using bwa index to facilitate efficient mapping. BWA-MEM was chosen for its ability to handle large datasets and its sensitivity in detecting both small and large variants. BWA first needs to construct the FM-index for the reference genome (the index command). First the index file for the reference genome is created using bwa-index and along with the index file GRCh38.fasta.fai other files like GRCh38.fasta.pac, GRCh38.fasta.sa, GRCh38fasta.bwt, GRCh38.fasta.ann, GRCh38.fasta.amb are created for the alignment process. BWA was used for the alignment of both reads using GRChg38 as the reference genome. The trimmed FASTQ files of Blood cancer were aligned to the GRCh38 FASTA file as the reference. The output is in SAM file format.

4.7.3 COMMAND

creating index for reference

- *bwa index GRCh38.fasta.fna*

- *bwa mem -t8 GRCh38.fasta.fna SRR31851783_1_trimmed.fastq.gz SRR31851783_2_trimmed.fastq.gz > aligned.Sam*

4.8 SAM TOOLS

SAM tools was employed to perform various post-alignment processing steps, ensuring that the alignment data was correctly formatted and indexed. The first step involved converting the SAM file to a BAM format. BAM files are more compact and efficient for downstream processing. The alignment process produced Sequence Alignment Map (SAM) files, which contain information about the alignment of each read to the reference genome.

These SAM files were converted to Binary Alignment Map (BAM) format using SAM tools, a suite of utilities designed for manipulating alignment files. BAM files are more compact and efficient for storage and processing compared to SAM files. Additionally, the BAM files were sorted and indexed using SAM tools to facilitate rapid access to specific genomic regions during downstream analysis.

4.8.1 COMMAND

SAM to BAM:

➢ *samtools view -bS SRR31851783_aligned.sam > SRR31851783_aligned.bam*

Sorting and Indexing the BAM:

➢ *samtools sort -o SRR31851783_sorted.bam SRR31851783_aligned.bam*

➢ *samtools index SRR31851783_sorted.bam*

4.9 MARKING DUPLICATES

4.9.1 PICARD TOOL

Picard Tools was utilized to mark duplicate reads in the BAM file, which helps in identifying PCR artifacts and redundant sequences arising from the library preparation. The Mark Duplicates function detects duplicate reads based on identical mapping positions and marks them without removal, ensuring that downstream analyses, such as variant calling, are not biased by overrepresented sequences.

The process generates a deduplicated BAM file along with a metrics file that provides detailed statistics on the number of duplicates, estimated library complexity, and duplication rates. This step is crucial for improving the accuracy of variant detection and ensuring reliable downstream interpretations.

4.9.2 COMMAND

Removing Mark Duplicates

➢ *picard markduplicates i = srr31851783_sorted.bam o = dedup.bam m = metrics.txt*

4.10 VARIANT CALLING

4.10.1 FREE BAYES

Free Bayes, a Bayesian-based variant caller, was used to identify single nucleotide variants (SNVs) and small insertions/deletions (INDELs) from the aligned and pre-processed BAM file. Unlike heuristic-based variant callers, FreeBayes performs haplotype-based variant detection, considering read alignments, sequencing errors, and population-level polymorphisms to improve variant accuracy. The output is a Variant Call Format (VCF) file containing detailed information about the identified variants, including their genomic positions, reference and alternate alleles, quality scores, and depth of coverage. This step is crucial for downstream analyses, such as functional annotation and disease association studies.

4.10.2 COMMAND

Running variant calling

➢ *freebayes -f GRCh38.36.fasta.fna dedup.bam > raw_variants.vcf*

4.11 VARIANT FILTRATION

4.11.1 GATK4

GATK4 (Genome Analysis Toolkit version 4) is a comprehensive and widely used framework for variant discovery and analysis in high-throughput sequencing data. Developed by the Broad Institute, GATK4 employs advanced algorithms to identify single nucleotide variants (SNVs), insertions/deletions (INDELs), and structural variants from aligned sequencing data (BAM files)**.** The primary output of GATK4 is a Variant Call Format (VCF) file, which contains detailed information about the identified variants, including their genomic coordinates, allele frequencies, quality scores, and annotations. These results can be further filtered and annotated using GATK4's built-in tools or integrated with downstream analysis pipelines for functional interpretation. Overall, GATK4 provides researchers with a robust and scalable solution for variant analysis, combining high accuracy, flexibility, and reproducibility, which are essential for advancing genomic research and precision medicine.

4.11.2 COMMAND

Creating a dict file:

➢ *gatk createsequence Dictionary -R GRCh38.36.fasta.fna*

Filtering variants:

➢ *gatk Variant Filtration -R GRCh38.36.fasta.fna -V raw_variants.vcf --filter -name "LowQual "-- filter -expression "QUAL < 30.0 | | DP < 10 "- O filtered_output.vcf*

4.12 ANNOTATION

4.12.1 ENSEMBL VEP

Variant annotation was performed using the Ensembl VEP (Variant Effect Predictor) https://www.ensembl.org/vep web interface with the *GRCh38* reference genome. Variant annotation is a critical step in understanding the biological and clinical significance of genetic mutations. In this study, the missense variants identified from the variant calling pipeline were annotated using the Ensembl Variant Effect Predictor (VEP) webserver.

VEP is a widely used tool that predicts the impact of genetic variants on genes, transcripts, and protein function. It integrates data from multiple sources, including the Ensembl genome database, ClinVar, dbSNP, and UniProt, to provide comprehensive variant annotations. The annotation process involved submitting variant call format (VCF) files to the VEP server, which then mapped the variants to their respective genomic locations and provided information such as consequence type, affected gene, protein impact, conservation scores, and pathogenicity predictions.

4.13 FUNCTIONAL ENRICHMENT OF MISSENSE VARIANTS

Functional enrichment analysis was performed using STRING (https://string-db.org/) to investigate the biological significance of genes affected by missense variants in Peripheral T-cell Lymphoma (PTCL). To understand the broader biological significance of the identified missense variants, we performed Gene Ontology (GO) analysis, which provides insights into the functional roles of genes at the cellular and molecular levels. GO analysis is structured into three categories: biological processes (BP), molecular functions (MF), and cellular components (CC). In this study, we specifically focused on biological process and molecular function analysis using the eg:Profiler tool, which is widely used for functional enrichment analysis. This tool maps input gene lists to predefined ontologies and determines statistically significant enrichments using hypergeometric testing. The biological process analysis allowed us to identify key pathways that the mutated genes were involved in, while molecular function analysis helped determine the types of biochemical activities affected by these missense variants. Benjamini-Hochberg correction was applied to control the false discovery rate (FDR), ensuring statistical robustness. The results provided a functional context for the missense variants, highlighting disrupted biological pathways and molecular activities that may contribute to peripheral T-cell lymphoma progression. The enriched gene ontology terms were visualized using ebar plots for better interpretation of the results.

4.14 PATHWAY ENRICHMENT OF MISSENSE VARIANTS

To further explore the mechanistic implications of the identified missense variants, we performed pathway enrichment analysis using two widely recognized tools: Reactome and WikiPathways. Pathway enrichment analysis helps identify biochemical pathways that are significantly affected by genetic mutations, providing insight into the underlying molecular mechanisms of disease progression. Reactome is a manually curated open-source database that maps genes to biological pathways, ensuring accurate representation of human molecular processes. Similarly, WikiPathways is a community-driven database that provides detailed pathway annotations across various biological functions. The genes harboring missense variants were submitted to these tools, and enrichment analysis was performed to identify pathways with a significant representation of altered genes. Statistical enrichment was determined using Fisher's exact test with FDR correction to account for multiple testing. The identified pathways were ranked based on their enrichment scores, and only those with an adjusted p-value < 0.05 were considered significant.

4.15 DISEASE ASSOCIATION STUDIES

To explore the potential clinical relevance of the identified missense variants, we conducted disease association studies using the DISEASES database from the Jensen Lab. The DISEASES database integrates information from manually curated literature, experimental datasets, and text-mining approaches to establish gene-disease associations. This step is essential for identifying potential links between the mutated genes and known diseases, including haematological malignancies and immune disorders. We input the list of mutated genes into the DISEASES database, which then ranked the associated diseases based on confidence scores derived from co-occurrence frequencies in scientific literature and experimental validation studies. A higher confidence score indicated a stronger association between the gene and a specific disease.

4.16 TISSUE EXPRESSION STUDIES

To determine the tissue-specific expression patterns of genes harboring missense variants, we performed tissue expression analysis using the TISSUES database from the Jensen Lab. This database integrates expression data from multiple sources, including RNA sequencing, microarrays, and immunohistochemistry, to provide a comprehensive overview of gene expression across different human tissues. The goal of this analysis was to identify whether the mutated genes are predominantly expressed in lymphoid tissues or other organ systems, which could provide insight into their potential involvement in peripheral T-cell lymphoma pathogenesis. The gene list was queried against the TISSUES database, and expression levels were retrieved for each gene in various tissues.

CHAPTER - V

RESULT

5. FASTQC:

The FASTQC tool is used to assess the sequence quality. The various analyses performed for the FASTQC report are summarized on the left side of the screen for each FASTQ file used for the analysis.

5.1 BASIC STATISTICS:

Basic Statistics module provides an overview of the sequencing data, including key information such as file name, format, encoding, total sequences, GC content, and sequence length distribution.

Basic Statistics

a)

Measure	Value
Filename	SRR31851783_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	51030629
Total Bases	5.1 Gbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	47

b)

Measure	Value
Filename	SRR31851784_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	53826805
Total Bases	5.4Gbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	47

Fig 5.1: Basic Statistics of FASTQC result

a) Sample (1) - *SRR31851783* **b)** Sample (2) - *SRR31851784*

The sequencing file passed this module, indicating that the reads have expected properties. The base composition is balanced, and there are no immediate concerns regarding the integrity of the dataset. These metrics confirm that the sequencing data is correctly formatted and ready for downstream analysis.

5.2 PER BASE SEQUENCE QUALITY

A box plot module evaluates the quality of each nucleotide position across all reads using Phred scores. High scores indicate accurate base calls, while lower scores suggest potential sequencing errors. Quality usually remains high at the start but may decrease towards the end of reads.

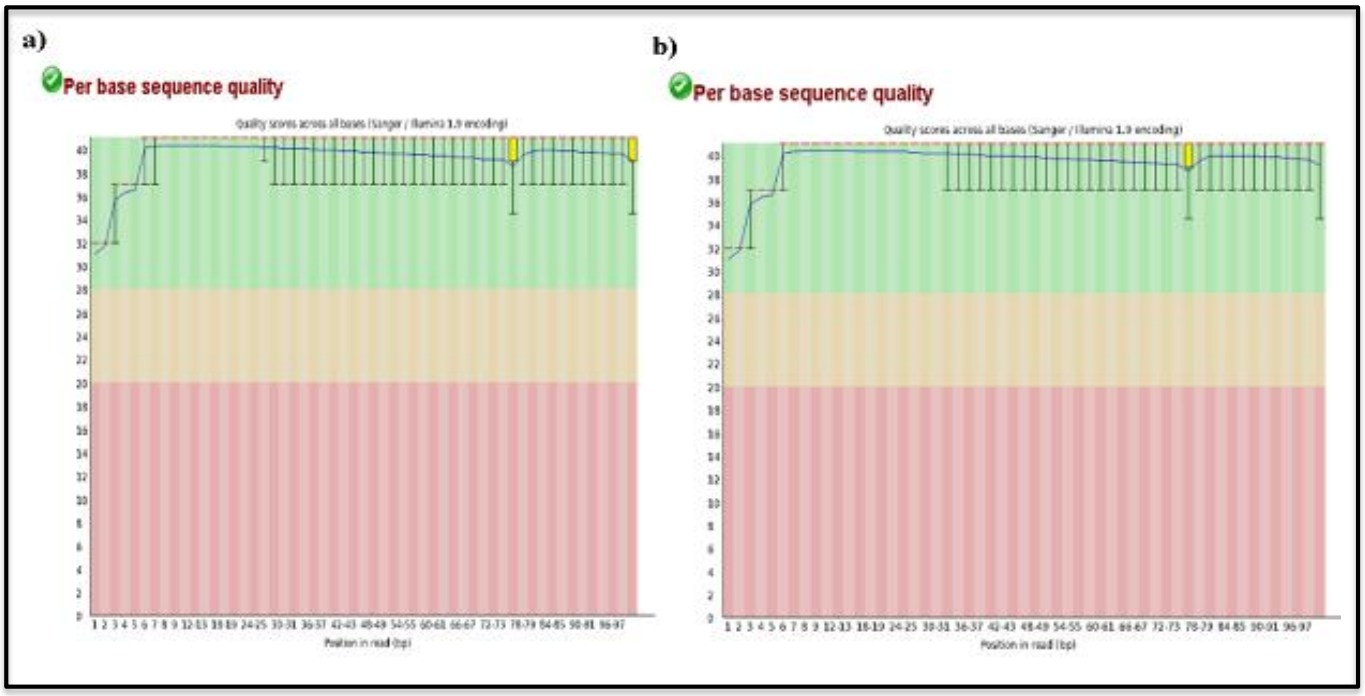

Fig 5.2: Per base sequence quality of FASTQC result

a) Sample-1 b) Sample-2

The results indicate high quality across all bases, with only minor reductions towards the end of some reads. The median quality scores remain well above the recommended threshold, suggesting reliable sequencing. No significant degradation or systematic errors were detected.

5.3 FASTP

As an all-in-one FASTQ pre-processor, fastp provides functions including quality profiling, adapter trimming, read filtering and base correction. It supports both single-end and paired-end short read data and also provides basic support for long-read data, which are typically generated by Pac Bio and Nanopore sequencers.

General

Fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
Sequencing:	paired end (101 cycles + 101 cycles)
Mean length before filtering:	101bp, 101bp
Mean length after filtering:	100bp, 100bp
Duplication rate:	35.460069%
Insert size peak:	142

Filtering result

Reads passed filters:	101.381466 m (99.333937%)
Reads with low quality:	641.804000 k (0.628842%)
Reads with too many N:	37.988000 K (0.037221%)
Reads too short:	0 (0.000000%)

Before filtering

Total reads:	102.061258 m
Total bases:	10.308187 g
Q20 bases:	10.151399 G (98.478995%)
Q30 bases:	9.884706 G (95.891803%)
GC content:	47.026293%

After filtering

Total reads:	101.381466 M
Total bases:	10.218861 G
Q20 bases:	10.083198 G (98.672427%)
Q30 bases:	9.829700 G (96.191737%)
GC content:	47.004370%

Fig 5.3: Summary of Fastp result sample -1

General

Fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
Sequencing:	paired end (101 cycles + 101 cycles)
Mean length before filtering:	101bp, 101bp
Mean length after filtering:	100bp, 100bp
Duplication rate:	35.460069%
Insert size peak:	142

Filtering result

Reads passed filters:	101.381466 M (99.333937%)
Reads with low quality:	641.804000 K (0.628842%)
Reads with too many N:	37.988000 K (0.037221%)
Reads too short:	0 (0.000000%)

Before filtering

Total reads:	102.061258 m
Total bases:	10.308187 g
Q20 bases:	10.151399 G (98.478995%)
Q30 bases:	9.884706 G (95.891803%)
GC content:	47.026293%

After filtering

Total reads:	101.381466 m
Total bases:	10.218861 g
Q20 bases:	10.083198 G (98.672427%)
Q30 bases:	9.829700 G (96.191737%)
GC content:	47.004370%

Fig 5.4: Summary of Fastp Result Sample-2

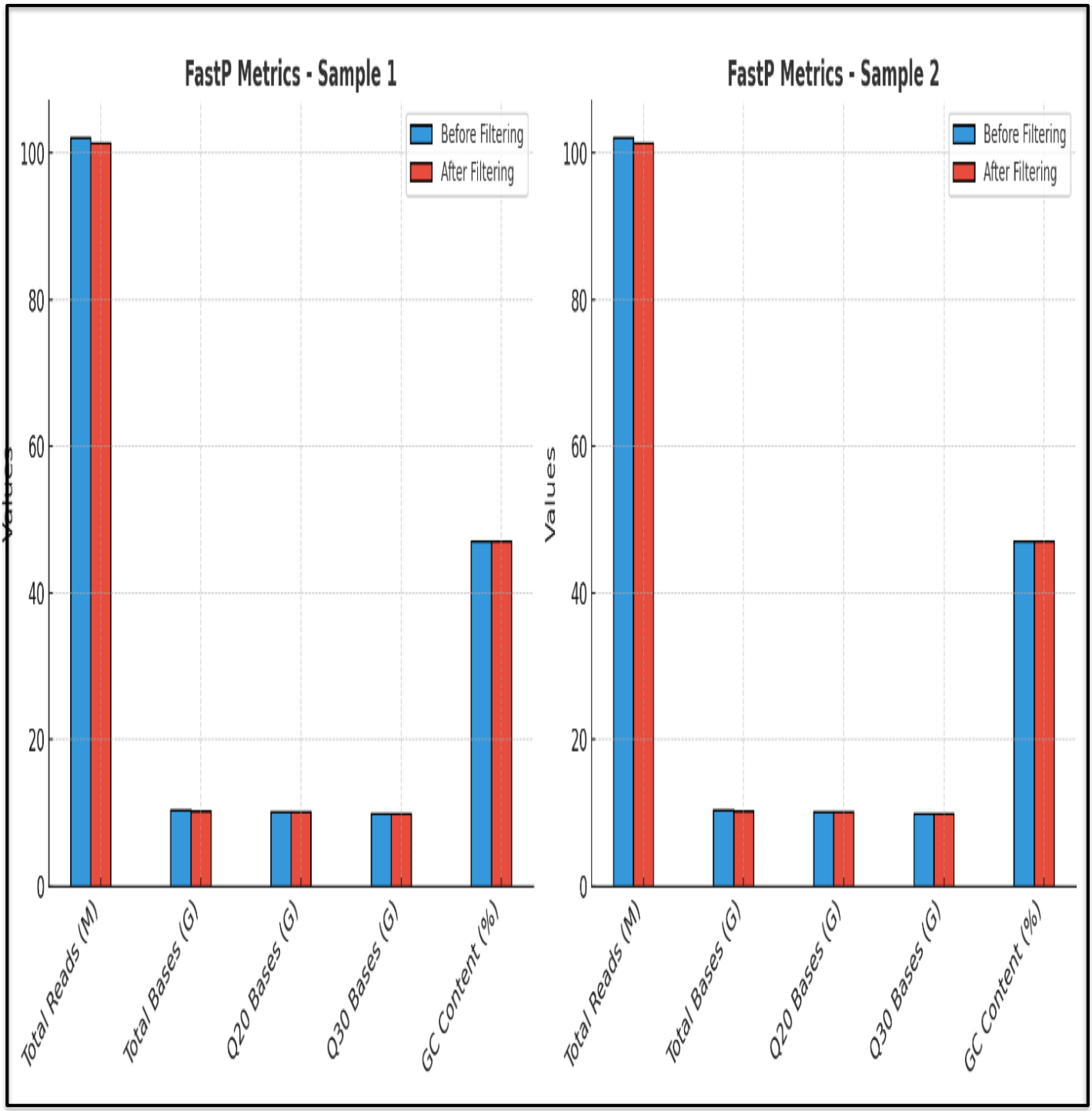

Fig 5.5: Fastp Metrics for Both Sample

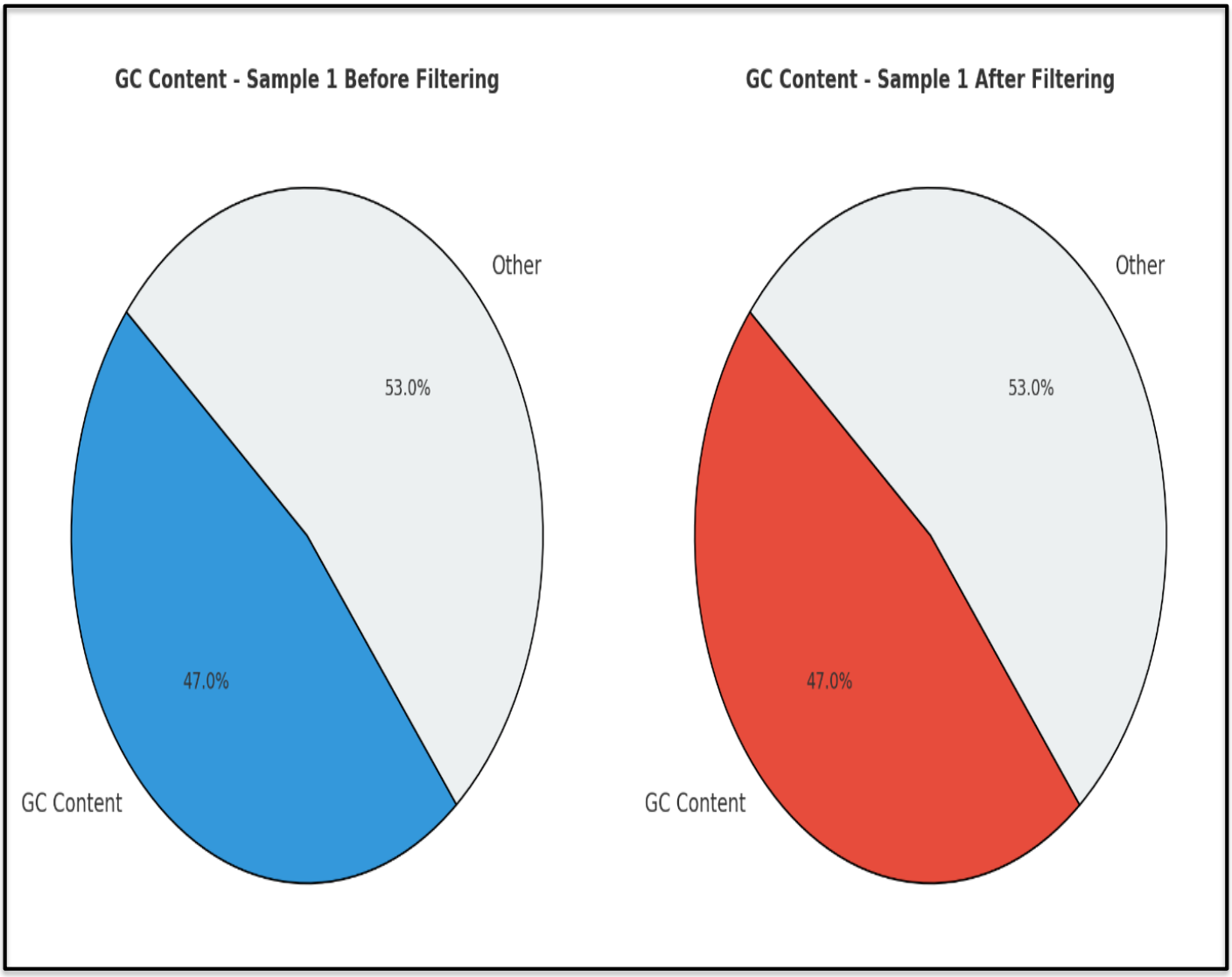

Fig 5.6: GC Content of Before & After Filtering

The Fastp results for both Sample-1 and Sample-2 show that over 99.33% of reads passed the filters, with less than 0.63% classified as low quality. The mean read length remains stable (101bp before and 100bp after filtering), and duplication rates are slightly above 30%. High sequencing quality is maintained, with Q20 and Q30 base percentages at approximately 98.47% and 96.19%, respectively. The GC content (~47.02%) remains consistent, confirming the data's reliability for downstream analysis.

5.4 BWA

The sequencing reads from Whole Exome Sequencing (WES) were aligned to the reference genome using BWA-MEM, generating a SAM file, which was converted to a BAM file for efficient processing. Alignment statistics confirmed a high mapping rate, ensuring reliable variant detection for downstream analysis.

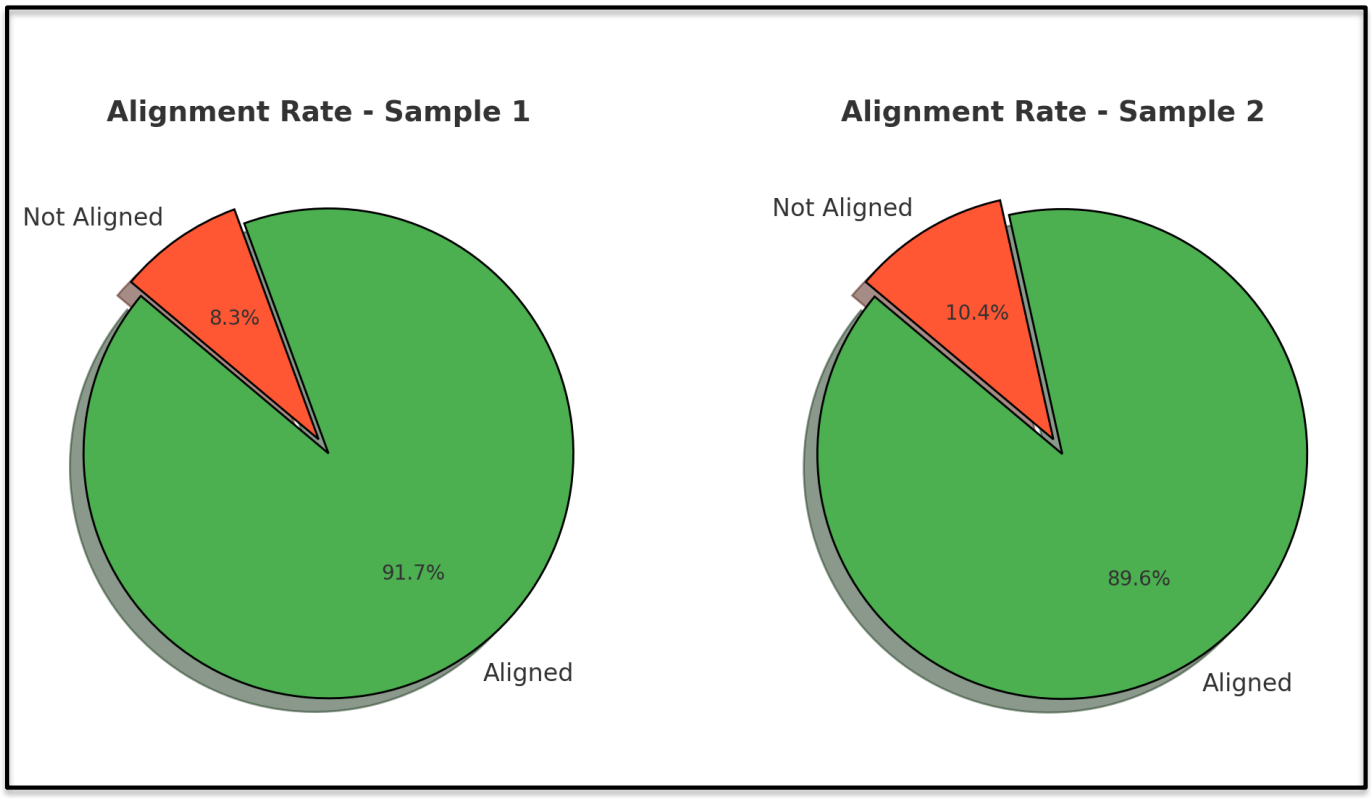

Fig 5.7: Alignment Mapping Rate for Both Sample

The alignment rates for Whole Exome Sequencing (WES) samples were analyzed using BWA-MEM. Sample 1 showed a 91.7% alignment rate, while Sample 2 had 89.6%. The remaining 8.3% and 10.4% of reads, respectively, were unmapped. High alignment rates indicate effective read mapping for variant detection. Unaligned reads may result from sequencing errors or genomic variations. These results confirm data quality for downstream analysis.

5.5 PICARD

Picard tool was employed to identify and remove duplicate reads, which commonly arise from PCR amplification during library preparation. These duplicates can introduce biases, affecting variant calling accuracy and overall data quality. By marking duplicates, Picard helps reduce false-positive variant calls, ensuring a more reliable and unbiased analysis

Table 5.1: Summary of Duplicate Read Metrics from Picard

METRIC	Sample 1	SAMPLE 2
LIBRARY	Unknown Library	Unknown Library
UNPAIRED_READS_EXAMINED	1168995	30973
READ_PAIRS_EXAMINED	45067424	53458846
SECONDARY_OR_SUPPLEMENTARY_RDS	148249	446469
UNMAPPED_READS	10077623	225047
UNPAIRED_READ_DUPLICATES	828798	16836
READ_PAIR_DUPLICATES	20293896	28454820
PERCENT_DUPLICATION	0.453613	0.532279
ESTIMATED_LIBRARY_SIZE	33497489	30100316

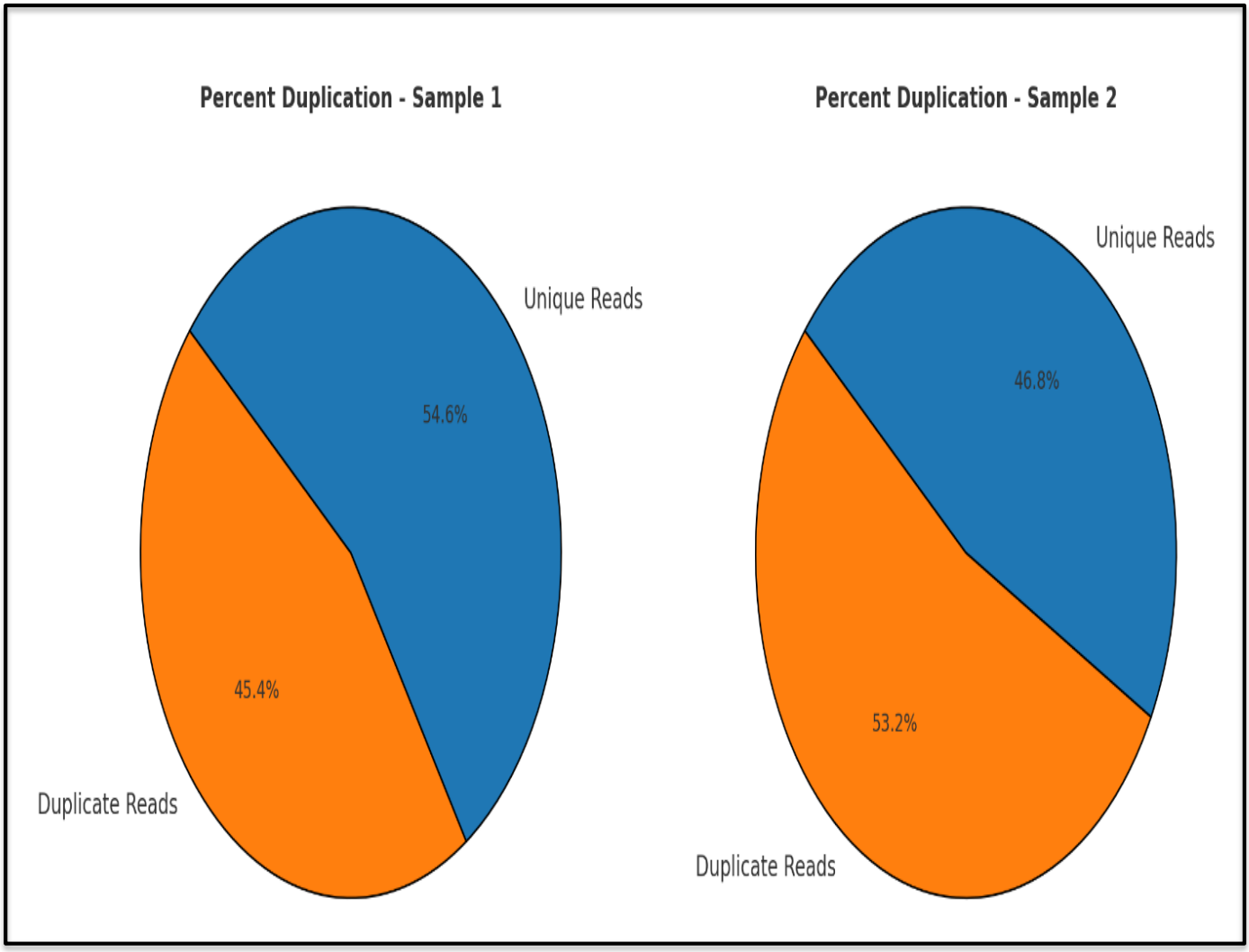

Fig 5.8: Percent Duplication for Both Samples

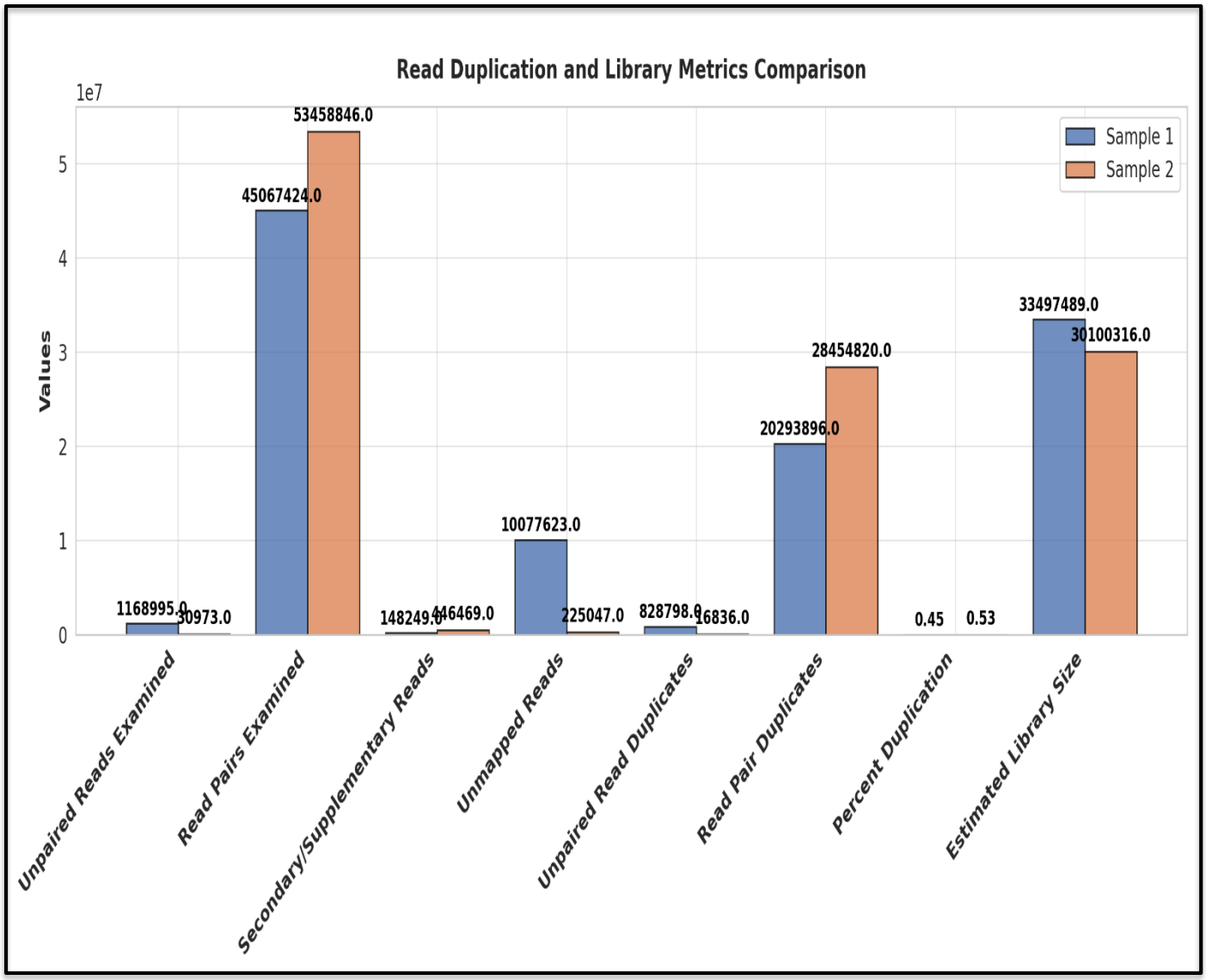

Fig 5.9: Read Duplication & Library Metrics Comparisons

The Picard metrics for Sample-1 and Sample-2 indicate differences in duplication rates and library sizes. Sample-2 has a higher percent duplication (53.23%) compared to Sample-1 (45.36%), suggesting more redundant reads. The estimated library size is slightly smaller in Sample-2 (30.1M) than in Sample-1 (33.5M). Both samples show significant read duplication, with Sample-2 having more read-pair duplicates (28.45M) than Sample-1 (20.29M). These results highlight variations in sequencing complexity and duplication levels, which may impact downstream variant analysis.

5.6 ENSEMBL VEP

Ensembl Variant Effect Predictor (VEP) was used for variant annotation, providing gene and transcript information, functional consequences, and protein impact predictions. It identified missense and other variant effects, along with pathogenicity scores from SIFT and PolyPhen. Additionally, regulatory region effects and population allele frequencies from gnomAD and 1000 Genomes were assessed. These annotations aid in understanding the functional and clinical significance of detected variants

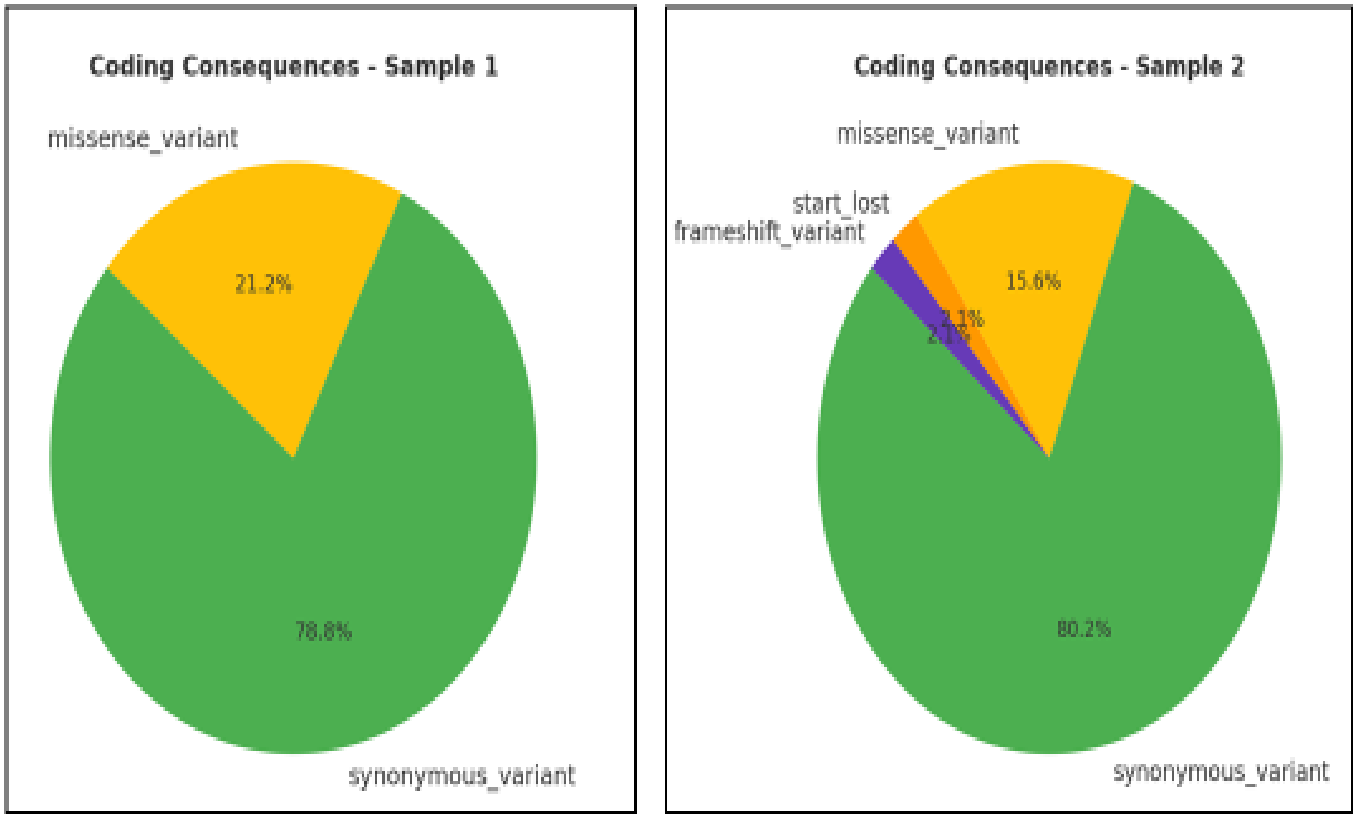

Fig 5.10: Coding Consequences for Both Sample

The functional impact of coding variants was analyzed for both samples. In Sample 1, 78.8% of variants were synonymous, causing no amino acid change, while 21.2% were missense, leading to amino acid substitutions. Similarly, Sample 2 had 80.2% synonymous variants and 15.6% missense variants. Additionally, Sample 2 contained a small proportion of frameshift (2.1%) and start-lost (2.1%) variants, which may have significant functional consequences

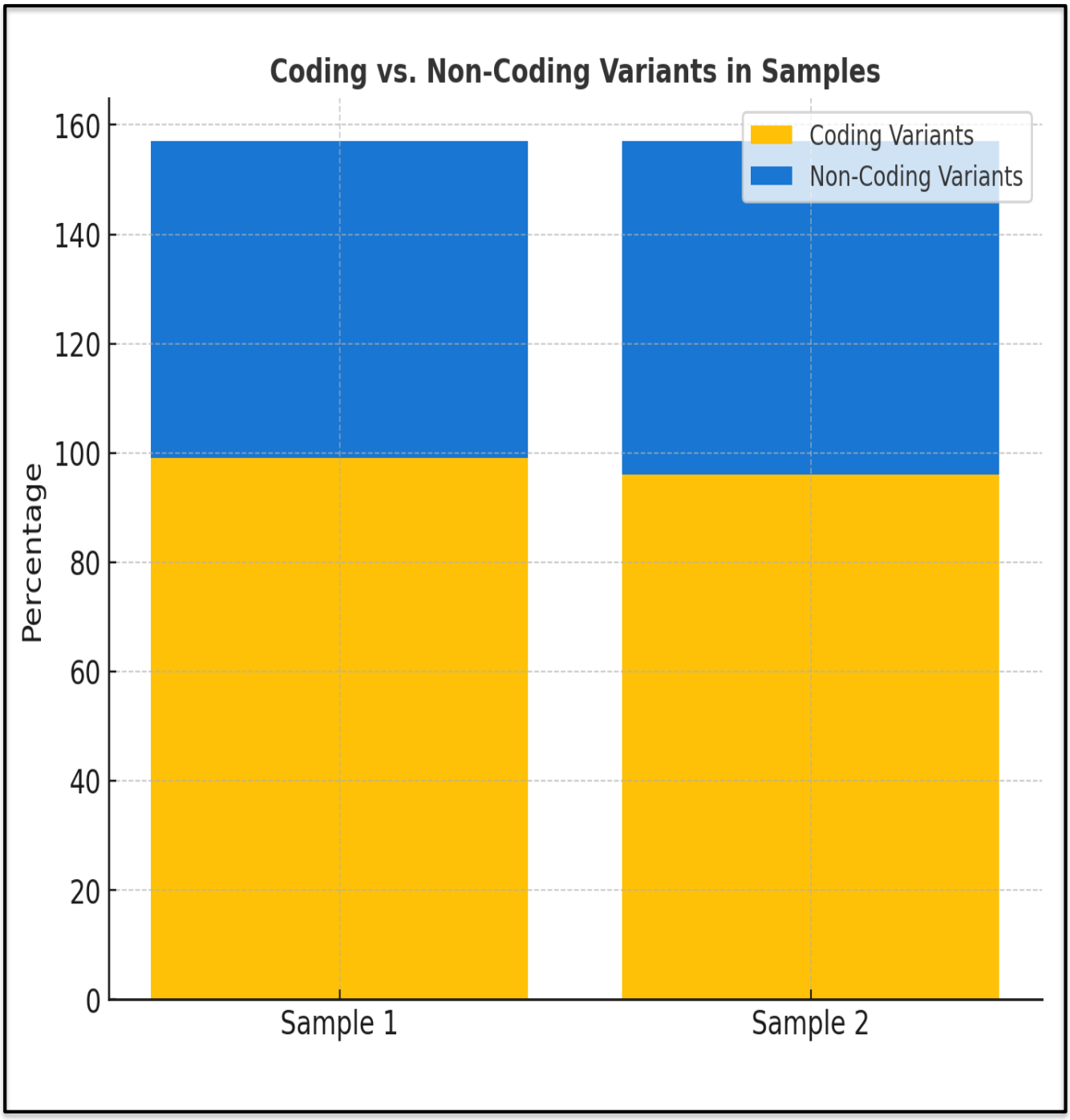

Fig 5.11: Coding (vs) Non-Coding Variants Samples

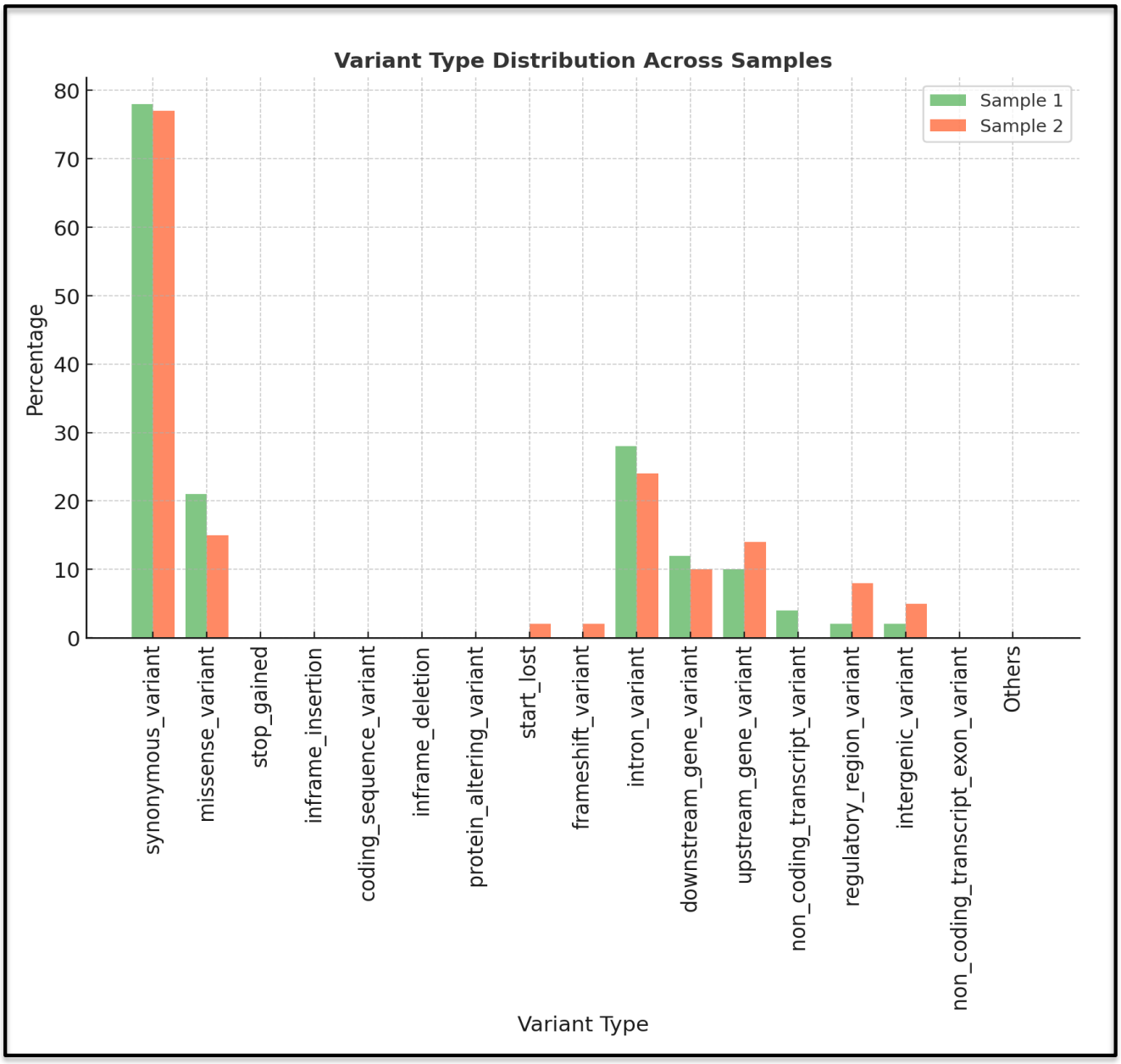

Fig 5.12: Comparison of Variant Type Distribution Between Samples

The analysis of variant types revealed that synonymous variants were the most abundant in both samples, followed by missense variants. Sample 1 exhibited a slightly higher percentage of synonymous and intronic variants, whereas Sample 2 had a greater proportion of stop-gained and regulatory region variants. Frameshift and start-lost mutations, which can have significant functional impacts, were present in lower frequencies. The distribution patterns suggest differences in mutation effects between samples, which may influence gene function and regulatory mechanisms.

5.7 FUNCTIONAL ENRICHMENTS

The STRING-based functional enrichment analysis identified key biological processes and molecular functions associated with the filtered missense variants. Additionally, pathway enrichment analysis using Wiki and Reactome provided insights into the molecular mechanisms underlying these variations. Disease association and tissue expression analysis further highlighted the potential clinical relevance of the identified genes.

Table 5.2: Summary of Protein-Protein Interaction (PPI) Network Statistics

Network Stats	
number of nodes:	206
number of edges:	795
average node degree:	7.72
avg. local clustering coefficient:	0.441
expected number of edges:	366
PPI enrichment p-value:	< 1.0e-16

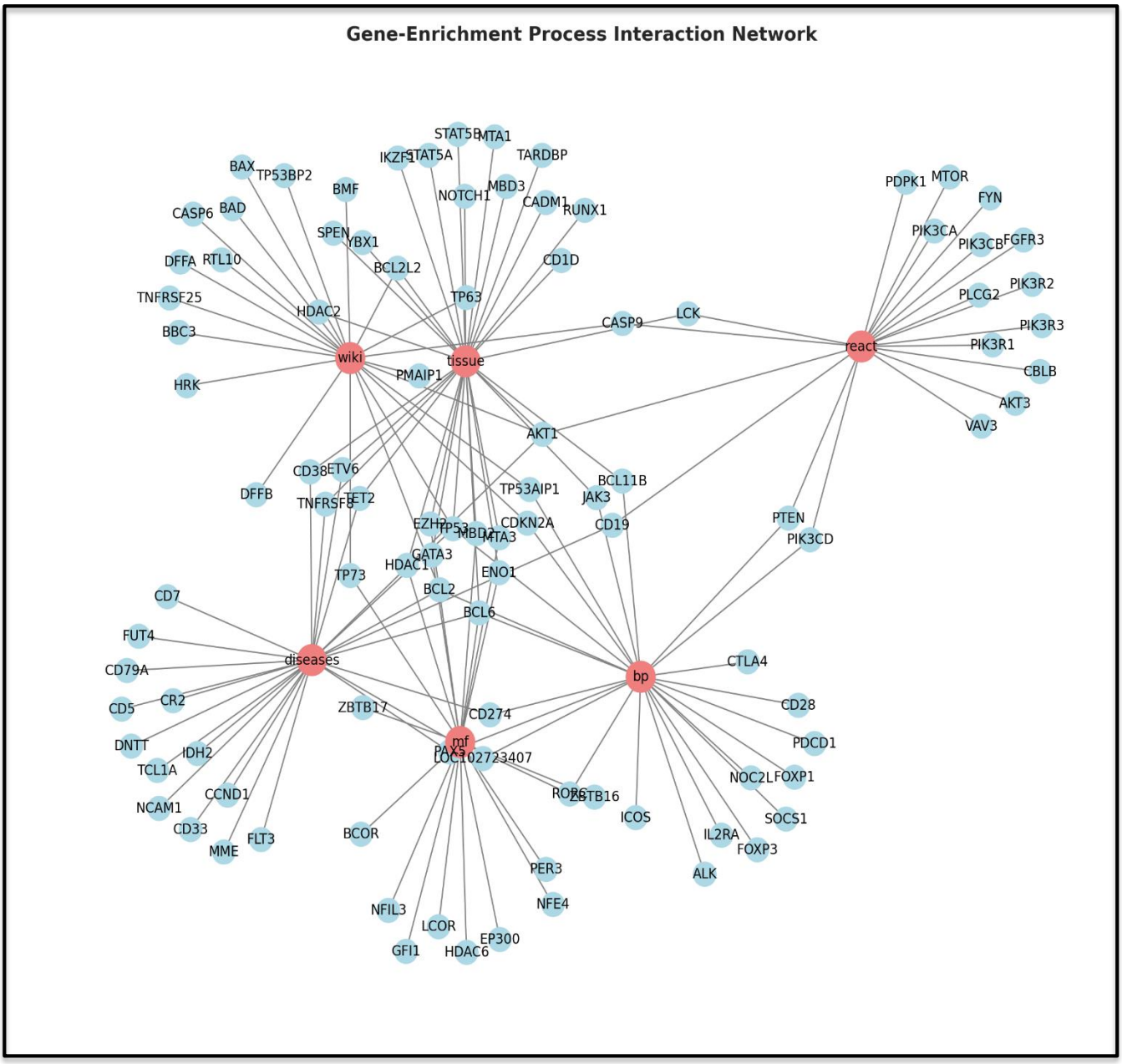

Fig 5.13: Gene-Enrichment Process Interaction Network

The network visualization represents the interaction of genes involved in biological processes, diseases, and pathways. The red nodes correspond to key functional categories such as biological processes (bp), diseases, tissue-specific interactions, and pathways (react, wiki)**,** while the blue nodes represent individual genes. The edges indicate interactions between these genes and functional categories. A higher number of edges suggest strong associations and involvement in multiple biological pathways. The clustering of nodes around specific functional categories highlights key regulatory genes potentially playing crucial roles in disease mechanisms or cellular processes

5.7 GENE ONTOLOGY

Gene Ontology (GO) is a standardized framework used to classify and describe the functions of genes and their products across different species. It provides a systematic way to categorize gene functions based on their roles in biological systems.

5.7.1 BIOLOGICAL PROCESS (BP)

Biological Process (BP) refers to a series of molecular activities that work together to achieve a specific biological goal. These processes include cell cycle regulation, immune response, apoptosis, signal transduction, and metabolism. GO enrichment analysis of BP helps researchers determine how genetic variations influence cellular mechanisms and disease progression.

5.7.2 MOLECULAR FUNCTION (MF)

Molecular Function (MF) refers to the specific biochemical activities carried out by a gene product at the molecular level. This includes enzymatic activities such as kinase and protease functions**,** ligand interactions like ATP or DNA binding**,** and roles in signal transduction**.** MF annotations provide insights into how proteins interact with other molecules, facilitating essential cellular processes. Understanding these functions helps determine the biological impact of genetic variations and their potential role in disease mechanisms.

Table 5.3: Significant GO Biological Processes Identified in PTCL Analysis

GO Term	Description	Count	Strength	Signal	False discovery rate
GO:2000106	Regulation of leukocyte apoptotic process	13 of 95	1.12	1.48	2.43e-07
GO:0070228	Regulation of lymphocyte apoptotic process	11 of 64	1.22	1.47	6.01e-07
GO:0030098	Lymphocyte differentiation	20 of 279	0.84	1.27	1.46E-07
GO:0070229	Negative regulation of lymphocyte apoptotic process	8 of 38	1.3	1.19	1.86E-05
GO:0046649	Lymphocyte activation	25 of 457	0.72	1.15	1.46E-07
GO:0002521	Leukocyte differentiation	22 of 396	0.73	1.08	7.93e-07
GO:0071887	Leukocyte apoptotic process	8 of 45	1.23	1.08	4.66e-05
GO:0030217	T-cell differentiation	14 of 171	0.89	1.07	9.39e-06
GO:0070227	Lymphocyte apoptotic process	7 of 34	1.29	1.02	0.00010
GO:0046632	Alpha-beta T cell differentiation	9 of 70	1.09	0.99	7.59e-05

Table 5.4: Enriched Molecular Function Terms in PTCL Analysis

GO Term	Description	Count	Strength	Signal	False discovery rate
GO:0097718	Disordered domain specific binding	5 of 34	1.15	0.0345	0.0345
GO:0097371	MDM2/MDM4 family protein binding	4 of 12	1.5	0.0151	0.0151
GO:0042802	Identical protein binding	45 of 2144	0.3	0.0080	0.0080
GO:0019904	Protein domain specific binding	23 of 695	0.5	0.0058	0.0058
GO:0008134	Transcription factor binding	20 of 587	0.51	0.0080	0.0080
GO:0001222	Transcription corepressor binding	7 of 46	1.16	0.0058	0.0058
GO:0001221	Transcription coregulatory binding	9 of 114	0.88	0.0080	0.0080

5.8 REACTOME PATHWAY

The DISEASES database integrates gene-disease associations from curated sources, text mining, and experimental studies to establish links between genetic variants and diseases. It provides insights into how specific genes contribute to disease mechanisms, helping identify potential biomarkers and therapeutic targets. The results offer a comprehensive view of genetic factors involved in various conditions, aiding in precision medicine and biomedical research.

Table 5.5: Enriched Reactome Pathways in PTCL Analysis

Pathway	Description	Count	Strength	Signal	False discovery rate
HSA-2219528	PI3K/AKT Signalling in Cancer	9 of 105	0.92	0.54	0.0064
HSA-9020558	Interleukin-2 signalling	4 of 12	1.51	0.54	0.0101
HSA-6804759	Regulation of TP53 Activity through Association with Co-factors	4 of 14	1.44	0.54	0.0101
HSA-6803207	TP53 Regulates Transcription of Caspase Activators and Caspases	4 of 12	1.51	0.54	0.0101
HSA-6803205	TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	4 of 14	1.44	0.54	0.0101
HSA-201556	Signalling by ALK	5 of 27	1.25	0.53	0.0101
HSA-114452	Activation of BH3-only proteins	5 of 29	1.22	0.53	0.0101
HSA-114604	GPVI-mediated activation cascade	5 of 35	1.14	0.52	0.0101
HSA-109606	Intrinsic Pathway for Apoptosis	6 of 52	1.04	0.51	0.0101
HSA-6803204	TP53 Regulates Transcription of Genes Involved in Cytochrome C Release	4 of 19	1.31	0.51	0.0126

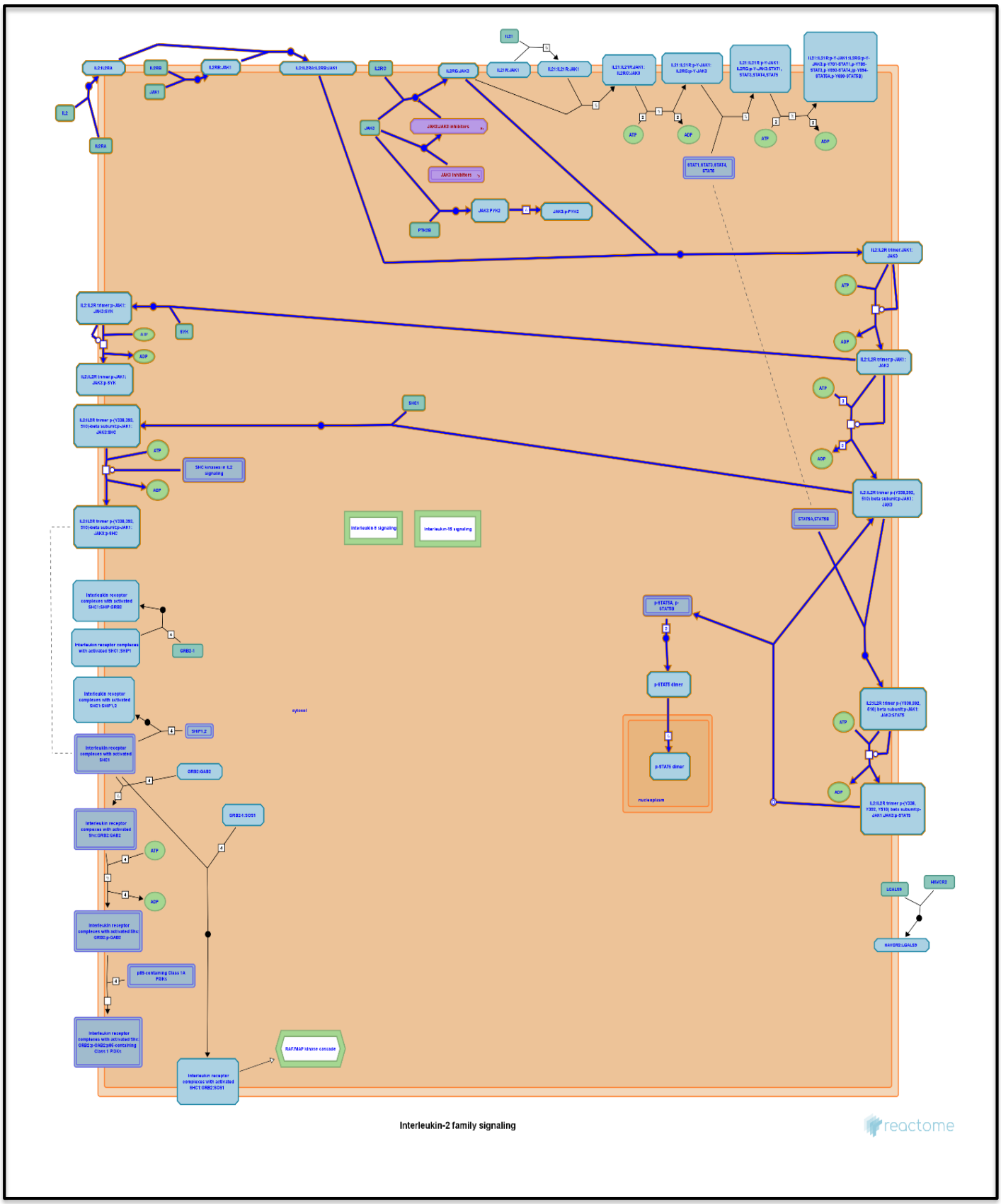

Fig 5 .14: Reactome Pathway Representation of Interleukin-2 Signalling in PTCL

5.9 WIKI PATHWAY

WikiPathways is a community-driven, open-access database that provides manually curated biological pathways. It includes metabolic, signalling, and disease-related pathways, allowing researchers to visualize molecular interactions and functional networks. The results help in understanding pathway dysregulation in diseases, facilitating hypothesis generation and drug target identification

Table 5.6: Enriched WikiPathways in PTCL Analysis

Pathway	Description	Count	Strength	Signal	False discovery rate
WP254	Apoptosis	11 of 83	1.11	1.28	2.70E-06
WP710	DNA Damage Response (ATM Dependent)	12 of 109	1.02	1.23	2.70E-06
WP1742	TP53 Network	6 of 19	1.48	1.15	3.92E-05
WP4674	Head and Neck Squamous Cell Carcinoma	9 of 72	1.08	1.04	3.92E-05
WP3584	MECP2 and Associated Rett Syndrome	9 of 72	1.08	1.04	3.92E-05
WP49	IL-2 Signaling Pathway	7 of 42	1.2	1.03	7.37E-05
WP2261	Glioblastoma Signaling Pathways	9 of 82	1.02	1	5.38E-05
WP5098	T-cell Activation SARS-CoV-2	9 of 88	0.99	0.96	7.37E-05
WP5218	Extra follicular and Follicular B cell Activation by SARS-CoV-2	8 of 72	1.03	0.91	0.00015
WP4255	Non-small Cell Lung Cancer	8 of 72	1.03	0.91	0.00015

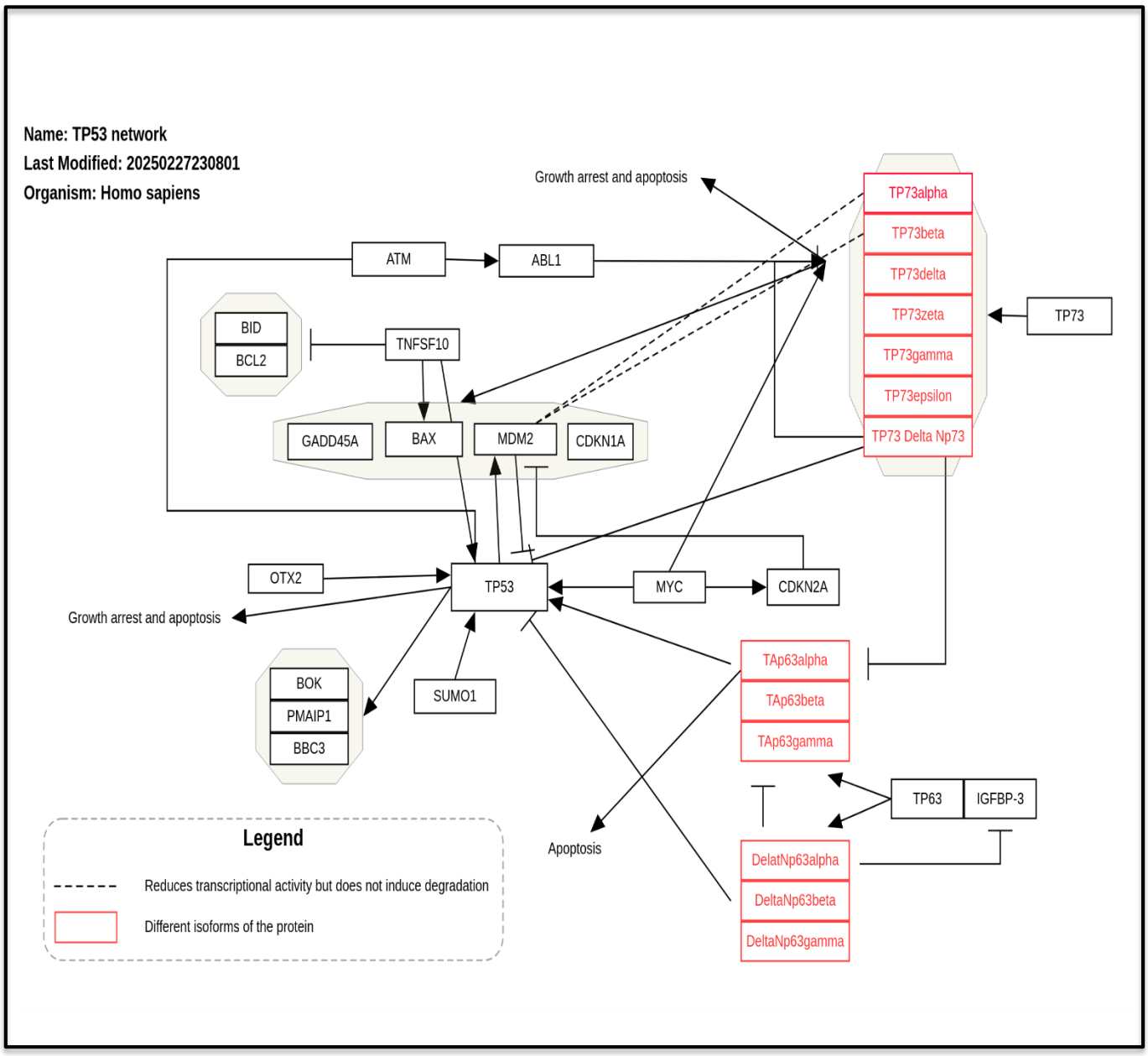

Fig 5.15: TP53 Signalling Network in Peripheral T-cell Lymphoma (PTCL)

5.10 Disease-Gene Associations (DISEASES)

The DISEASES module provides gene-disease associations by integrating curated datasets, text mining, and experimental evidence. It identifies genes linked to specific diseases, helping in understanding disease mechanisms and prioritizing candidate genes for further study. The results offer insights into known pathogenic mutations and potential therapeutic targets.

Table 5.7: Disease-Gene Associations in PTCL Analysis

Disease	Description	Count	Strength	Signal	False discovery rate
DOID:0060060	Non-Hodgkin Lymphoma	12 of 61	1.28	1.73	4.44E-08
DOID:707	B-cell Lymphoma	9 of 34	1.41	1.64	2.54E-07
DOID:0060058	Lymphoma	14 of 105	1.11	1.6	4.44E-08
DOID:0060073	Lymphatic System Cancer	15 of 120	1.08	1.58	4.44E-08
DOID:0050745	Diffuse Large B-cell Lymphoma	7 of 17	1.6	1.55	1.17E-06
DOID:3996	Urinary System Cancer	12 of 89	1.11	1.47	2.54E-07
DOID:75	Lymphatic System Disease	16 of 174	0.95	1.41	8.74E-08
DOID:11934	Head and Neck Cancer	6 of 12	1.68	1.4	5.69E-06
DOID:2531	Hematologic Cancer	16 of 190	0.91	1.31	2.54E-07
DOID:3908	Lung Non-Small Cell Carcinoma	7 of 26	1.41	1.27	1.08E-05

5.11 Tissue Expression (TISSUES)

The TISSUES module analyzes gene expression across different tissues using data from transcriptomics, proteomics, and immunohistochemistry. It highlights tissue-specific gene activity, revealing functional roles and disease relevance. The results help in identifying genes predominantly expressed in disease-affected tissues, supporting biomarker discovery and targeted research.

Table 5.8: Tissue Expression Analysis in PTCL

Tissue	Description	Count	Strength	Signal	False discovery rate
BTO:0000744	Lymphocytic leukemia cell	20 of 419	0.66	0.87	2.02E-05
BTO:0000583	Bone marrow cancer cell	21 of 442	0.66	0.87	2.02E-05
BTO:0001546	Chronic lymphocytic leukemia cell	14 of 222	0.78	0.86	8.16E-05
BTO:0001130	Prostate gland cancer cell	4 of 7	1.74	0.8	0.0012
BTO:0001271	Leukemia cell	34 of 1067	0.49	0.74	2.02E-05
BTO:0000580	Blood cancer cell	37 of 1234	0.46	0.72	2.02E-05
BTO:0000093	MCF-7 cell	4 of 11	1.54	0.67	0.0033
BTO:0000426	Erythroleukemia cell	13 of 244	0.71	0.64	0.0011
BTO:0001879	H9c2 cell	3 of 4	1.86	0.59	0.0074
BTO:0000256	Myoblast cell line	4 of 15	1.41	0.57	0.0074
BTO:0000744	Lymphocytic leukemia cell	20 of 419	0.66	0.87	2.02E-05

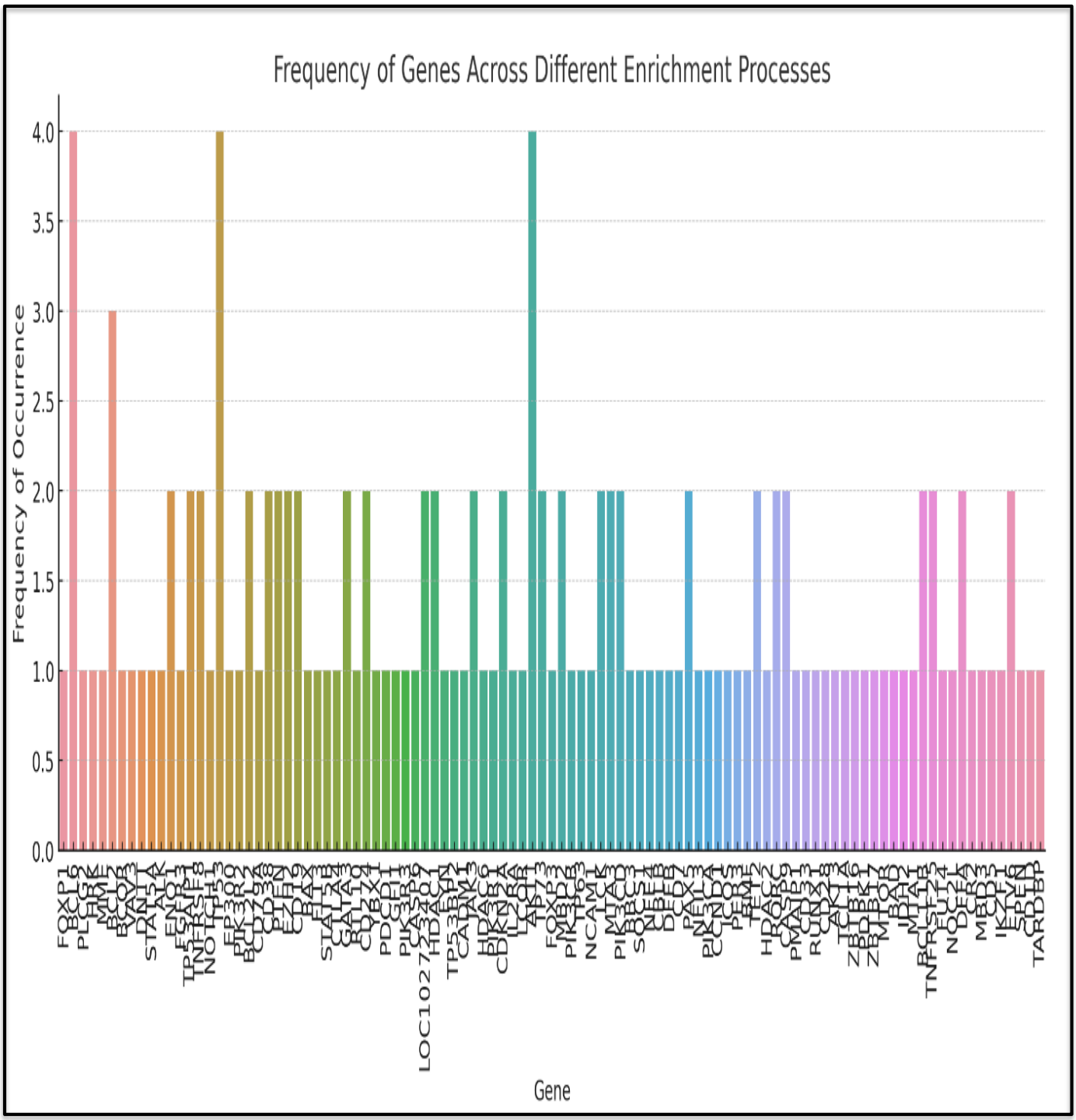

Fig 5.16: Gene Frequency Distribution Across Enrichment Processes

This bar chart illustrates the frequency of occurrence of different genes in various enrichment processes. The x-axis represents individual genes, while the y-axis indicates how many times each gene appears in different enriched biological processes.

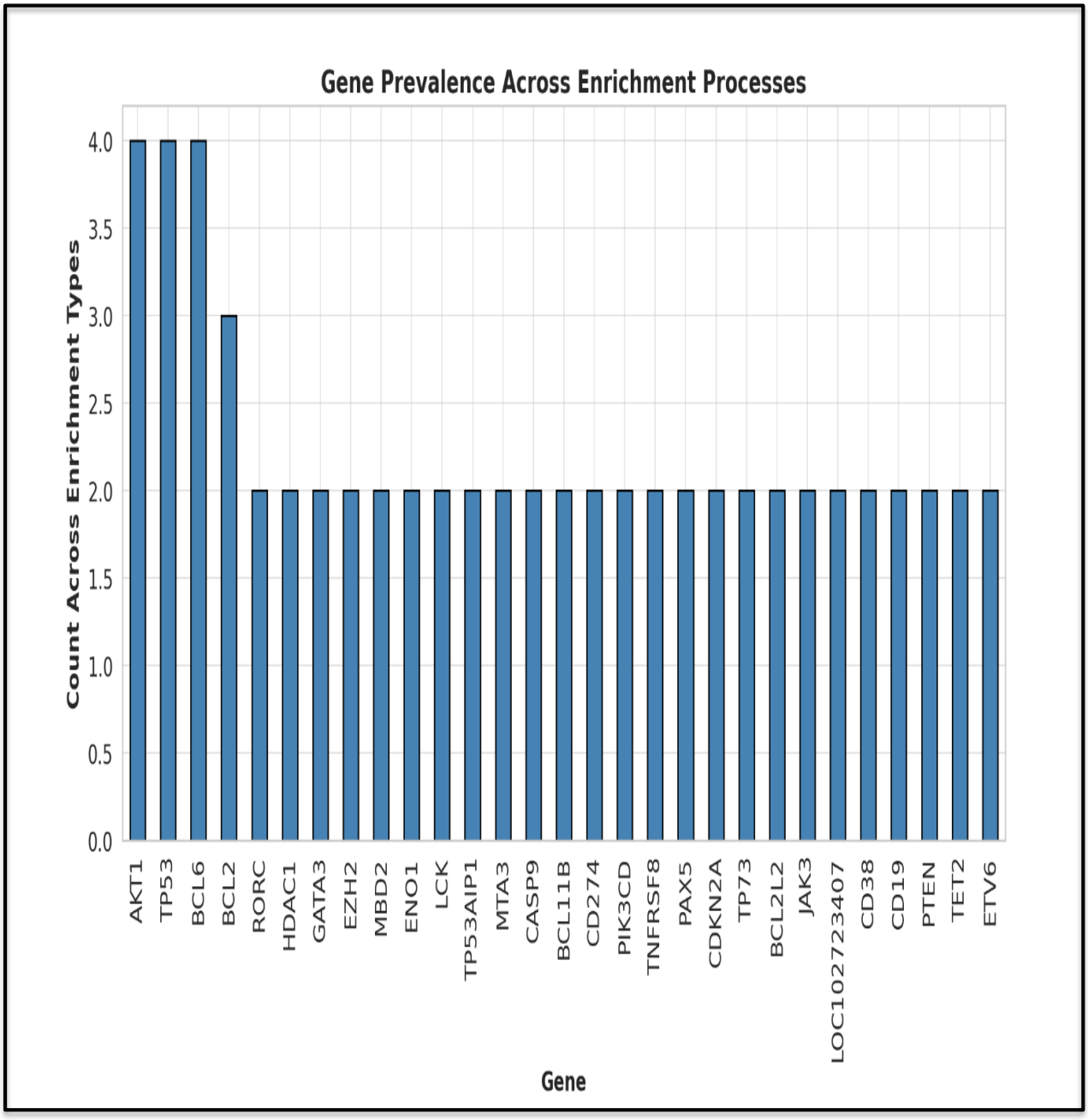

Fig 5.17: Prevalence of Genes Across Enrichment Processes

The bar chart illustrates the frequency of specific genes across different enrichment processes. Genes such as ***AKT1, TP53,*** and ***BCL6*** exhibit the highest prevalence, appearing in four enrichment types, indicating their significant role in multiple biological pathways. Other genes, including ***BCL2, RORC, HDAC1,*** and ***GATA3***, are also highly enriched, appearing in two or more processes. These findings suggest that certain genes play a central role in the biological functions under investigation, potentially serving as key regulators or biomarkers in the context of the study

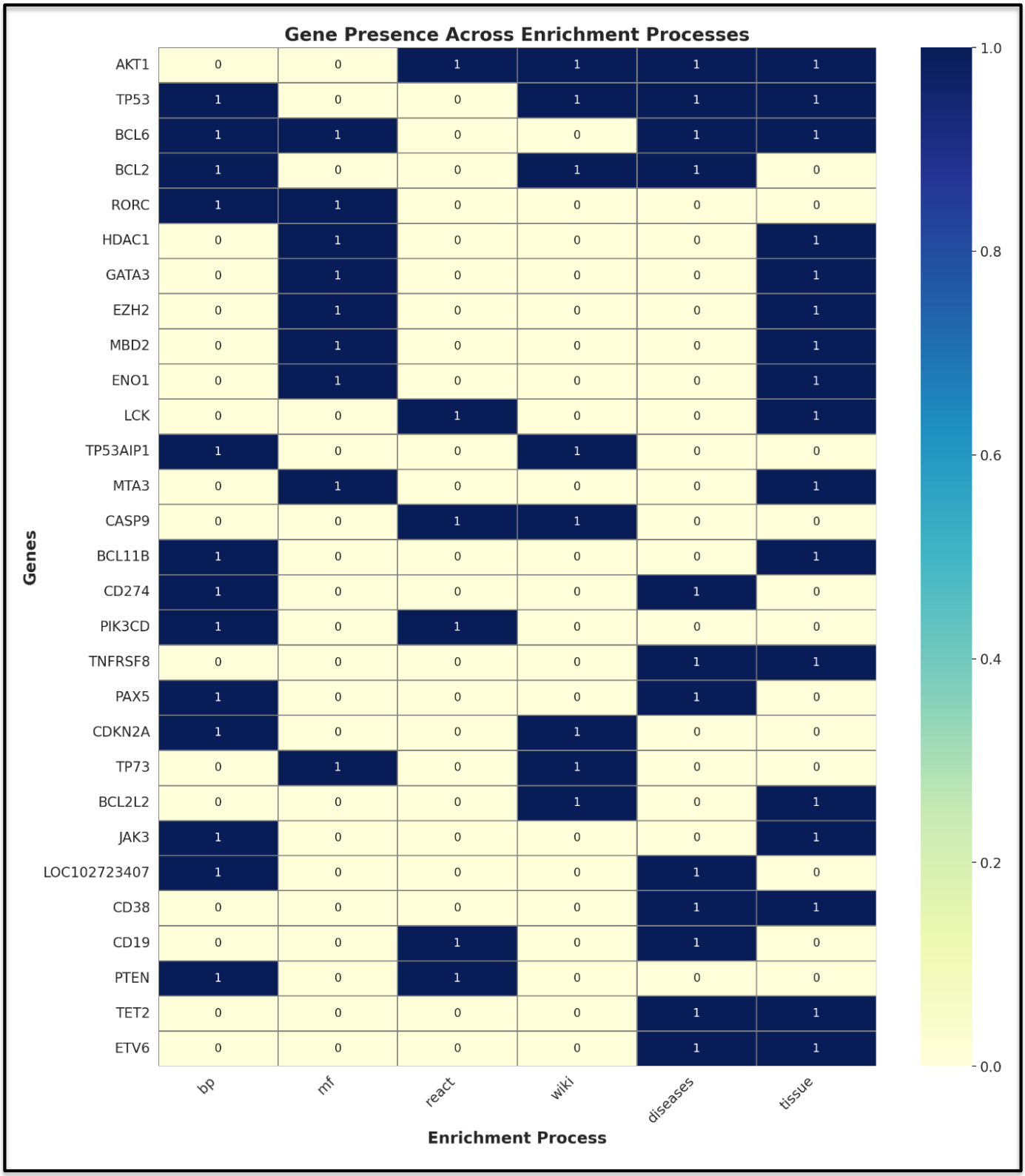

Fig 5.18: Heat map of Gene Presence Across Different Enrichment Processes

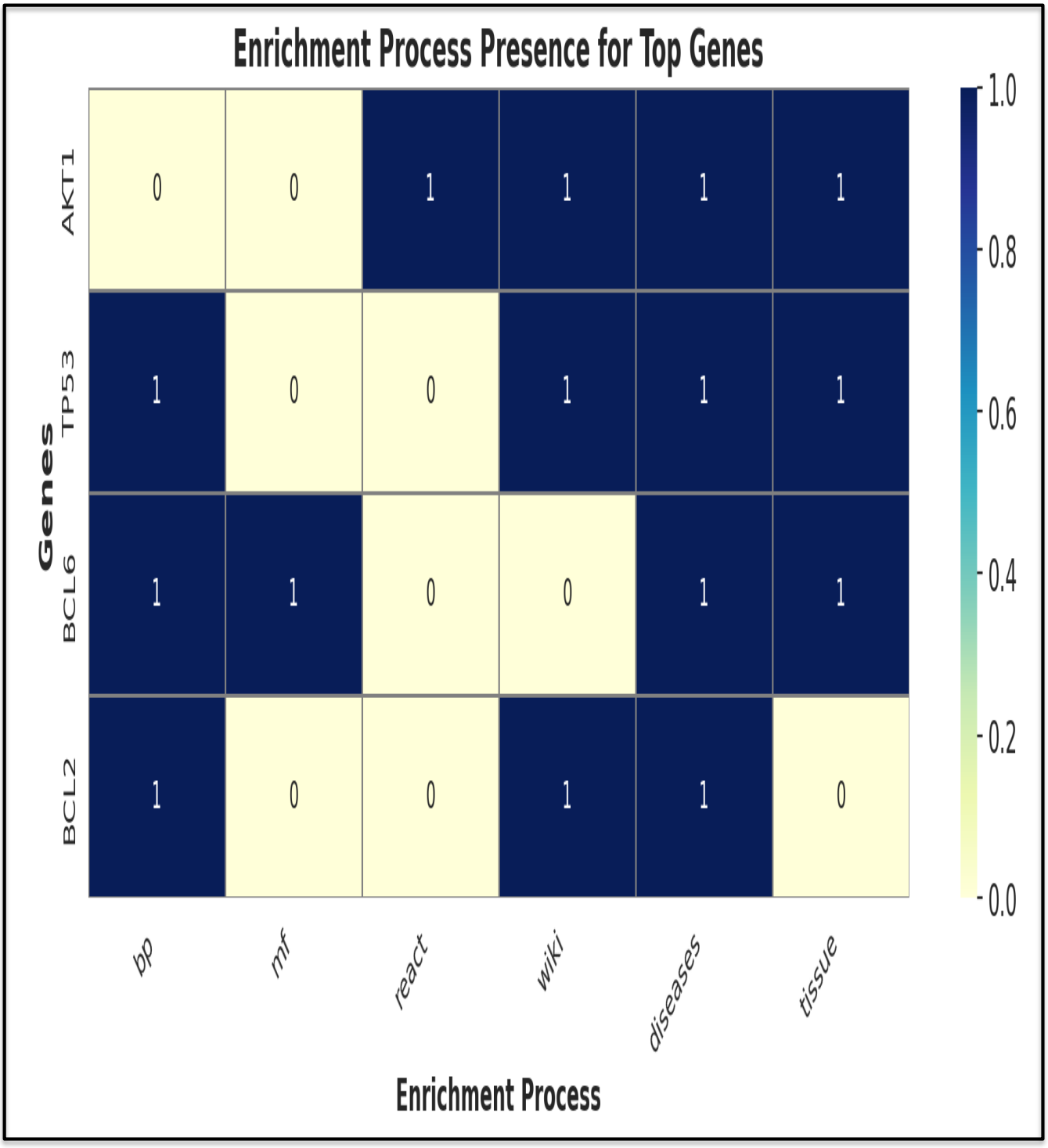

Fig 5.19: Heat map of Enrichment Process Presence for Top Genes

The heat map analysis identified ***TP53, BCL6, BCL2,*** and ***AKT1*** as key genes recurrent across multiple enrichment categories, highlighting their role in PTCL. ***TP53*** was linked to disease-associated pathways and tumor suppression, while ***BCL6*** and ***BCL2*** were prominent in apoptosis and lymphocyte differentiation. ***AKT1***, crucial for cell survival and proliferation, was consistently enriched in pathway-based analysis.

DISCUSSION

The Whole Exome Sequencing (WES) analysis of two PTCL samples identified several missense variants, which were filtered and annotated to prioritize functionally relevant mutations. Functional enrichment analysis revealed key genes associated with multiple biological processes, pathways, and disease networks. Among the enriched genes, ***TP53, BCL6, BCL2,*** and ***AKT1*** emerged as highly relevant across different biological categories. TP53 and BCL2 were prominently associated with apoptosis regulation and tumor suppression pathways, indicating their potential role in PTCL progression. BCL6, a well-known oncogene, was enriched in transcriptional regulation and immune response pathways, suggesting its involvement in lymphomagenesis. Additionally, AKT1, a crucial regulator of cell survival and proliferation, was linked to multiple signalling pathways, emphasizing its significance in cancer development.

Heat map analysis illustrated the presence of these key genes across multiple functional categories, with *TP53*, *BCL6*, and *BCL2* showing enrichment in disease association, pathway enrichment, and tissue expression analyses. Furthermore, frequency distribution analysis highlighted the recurrence of these genes across various enrichment processes, reinforcing their importance in PTCL biology. The overlap of these genes in different biological pathways suggests their potential as therapeutic targets and biomarkers. These findings contribute to a deeper understanding of the genetic landscape of PTCL and provide insights into the molecular mechanisms underlying the disease, paving the way for future research and targeted therapeutic strategies.

CHAPTER – VI

CONCLUSION

The Whole Exome Sequencing (WES) analysis of PTCL samples enabled the identification and functional characterization of key missense variants. The enrichment analysis revealed the significance of genes such as TP53, BCL6, BCL2, and AKT1, which are involved in critical biological processes, including cancer progression, apoptosis, and immune regulation. The integration of functional enrichment, pathway analysis, and gene-disease association studies provided deeper insights into the molecular mechanisms underlying PTCL, which may contribute to future biomarker discovery and therapeutic advancements.

Through this project, I gained hands-on experience in Linux, bioinformatics pipelines, and data visualization using Python and Matplotlib. This strengthened my ability to perform NGS data analysis, variant annotation, and functional interpretation, equipping me with essential skills for computational biology and translational research

REFERENCE

1. Advancements in next-generation sequencing, SE Levy, RM Myers - Annual review of genomics and human. (n.d.).Amador, C., Weisenburger, D. D., Gomez, A., Bouska, A., Alshomrani, A., Sharma, S., Shah, A. R., Greiner, T. C., Vega, F., Rosenwald, A., Ott, G., Feldman, A. L., Jaffe, E. S., Ozkaya, N., Ondrejka, S. L., Cook, J. R., Raess, P. W., Savage, K. J., Slack, G. W.

2. Leukemia and Lymphoma Molecular Profiling Project Consortium. (2025). Refining diagnostic subtypes of peripheral T-cell lymphoma using a multipara meter approach. Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc, 38(2), 100646. https://doi.org/10.1016/j.modpat.2024.100646

3. Exome-based cancer predisposition gene testing can provide a genetic diagnosis for individuals with heterogeneous tumor phenotypes Snežana Hinić. Barrington, S. F., Mikhaeel, N. G., Kostakoglu, L., Meignan, M., Hutchings, M., Müeller, S. P., Schwartz, L. H., Zucca, E., Fisher, R. I., Trotman, J., Hoekstra, O. S., Hicks, R. J., O'Doherty, M. J., Hustinx, R., Biggi, A., & Cheson, B. D. (2014).

4. Role of imaging in the staging and response assessment of lymphoma: consensus of the International Conference on Malignant Lymphomas Imaging Working Group. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 32(27), 3048–3058. https://doi.org/10.1200/JCO.2013.53.5229 ,Cabral, B. P., da Graça Derengowski Fonseca, M., & Mota, F. B. (2018).

5. The recent landscape of cancer research worldwide: a bibliometric and network analysis. Oncotarget, 9(55), 30474–30484. https://doi.org/10.18632/oncotarget.25730 Corvelo, A., & Eyras, E. (2008).

6. Exon creation and establishment in human genes. Genome Biology, 9(9), R141. https://doi.org/10.1186/gb-2008-9-9-r141 ,Genovese, G., Kähler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., Chambert, K., Mick, E., Neale, B. M., Fromer, M., Purcell, S. M., Svantesson, O., Landén, M., Höglund, M., Lehmann, S., Gabriel, S. B., Moran, J. L., Lander, E. S., Sullivan, P. F., McCarroll, S. A. (2014).

7. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. The New England Journal of Medicine, 371(26), 2477–2487. https://doi.org/10.1056/NEJMoa1409405 Greer, J. P., York, J. C., Cousar, J. B., Mitchell, R. T., Flexner, J. M., Collins, R. D., & Stein, R. S. (1984).

8. Peripheral T-cell lymphoma: a clinic pathologic study of 42 cases. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 2(7), 788–798. https://doi.org/10.1200/JCO.1984.2.7.788 Hari, P., Carreras, J., Zhang, M.-J., Gale, R. P., Bolwell, B. J., Bredeson, C. N., Burns, L. J., Cairo, M. S., Freytes, C. O., Goldstein, S. C., Hale, G. A., Inwards, D. J., Lemaistre, C. F., Maharaj, D., Marks, D. I., Schouten, H. C., Slavin, S., Vose, J. M., Lazarus, H. M., & van Besien, K. (2008).

9. Allogeneic transplants in follicular lymphoma: higher risk of disease progression after reduced-intensity compared to myeloablative conditioning. Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation, 14(2), 236–245. https://doi.org/10.1016/j.bbmt.2007.11.004 Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., & Zhang, H. (2020).

10. RNA sequencing: new technologies and applications in cancer research. Journal of Hematology & Oncology, 13(1), 166. https://doi.org/10.1186/s13045-020-01005-x Huang, J., Chan, S. C., Ngai, C. H., Lok, V., Zhang, L., Lucero-Prisno, D. E., 3rd, Xu, W., Zheng, Z.-J., Elcarte, E., Withers, M., & Wong, M. C. S. (2022)

11. Disease burden, risk factors, and trends of leukaemia: A global analysis. Frontiers in Oncology, 12, 904292. https://doi.org/10.3389/fonc.2022.904292 Huang, J., Chan, W. C., Ngai, C. H., Lok, V., Zhang, L., Lucero-Prisno, D. E., 3rd, Xu, W., Zheng, Z.-J., Elcarte, E., Withers, M., Wong, M. C. S., & On Behalf Of Ncd Global Health Research Group Of Association Of Pacific Rim Universities Apru. (2022).
12. Worldwide burden, risk factors, and temporal trends of ovarian cancer: A global study. Cancers, 14(9), 2230. https://doi.org/10.3390/cancers14092230 Iqbal, J., Amador, C., McKeithan, T. W., & Chan, W. C. (2019).

13. Molecular and genomic landscape of peripheral T-cell lymphoma. Cancer Treatment and Research, 176, 31–68. https://doi.org/10.1007/978-3-319-99716-2_2 Jochum, D. T., Bouska, A., Sharma, S., Vishwakarma, S., Amador, C., Lone, W. G., Shah, A. R., Mir, A. R., Soma, M. A., Mehmood, S. M., Nasser, Z., Feldman, A. L., Greiner, T. C., Vose, J. M., Cook, J. R., Ondrejka, S. L., Jaffe, E. S., Rosenwald, A., Ott, G., Chan, W. C. (2024).

14. An llmpp study: Genomic characterization of novel PTCL- biological subtypes reveal distinctive therapeutic vulnerabilities. Blood, 144(Supplement 1), 456–456. https://doi.org/10.1182/blood-2024-209396 Jordhøy, M. S., Jaeger, S., & Hammerstrøm, J. (1995).

15. Diagnosis and treatment of lymphomas. Retrospective quality assessment of a 10-year material. Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke, 115(26), 3243–3248. Loewe, L., & Hillston, J. (2008).

16. The distribution of mutational effects on fitness in a simple circadian clock. In Computational Methods in Systems Biology (pp. 156–175). Springer Berlin Heidelberg. Lorenzoni, V., Chaturvedi, A. K., Vignat, J., Laversanne, M., Bray, F., & Vaccarella, S. (2022)

17. . The current burden of oropharyngeal cancer: A global assessment based on GLOBOCAN 2020. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 31(11), 2054–2062. https://doi.org/10.1158/1055-9965.EPI-22-0642 Lymphoma, T.-C., Luan, Y., Li, X., Luan, Y., Luo, J., Dong, Q., Ye, S., Li, Y., Li, Y., & Jia, L. (n.d.). Therapeutic challenges in peripheral T-cell lymphoma, Yunpeng Luan.Ma, H., Marchi, E., O'Connor, O. A., & Lue, J. K. (2023).

18. Mature T-cell and NK-cell lymphoma involvement of the central nervous system: a single center experience. Leukemia & Lymphoma, 64(12), 1964–1970. https://doi.org/10.1080/10428194.2023.2245513 McCabe, B., Liberante, F., & Mills, K. I. (2015).

19. Repurposing medicinal compounds for blood cancer treatment. Annals of Hematology, 94(8), 1267–1276. https://doi.org/10.1007/s00277-015-2412-1 Nizamuddin, I. A., & Mehta-Shah, N. (2024).

20. BV and beyond: how to incorporate novel agents into PTCL management. Hematology, 2024(1), 54–61. https://doi.org/10.1182/hematology.2024000530 Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014).

21. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics, 15(2), 256–278. https://doi.org/10.1093/bib/bbs086 Palomero, T., Couronné, L., Khiabanian, H., Kim, M.-Y., Ambesi-Impiombato, A., Perez-Garcia, A., Carpenter, Z., Abate, F., Allegretta, M., Haydu, J. E., Jiang, X., Lossos, I. S., Nicolas, C., Balbin, M., Bastard, C., Bhagat, G., Piris, M. A., Campo, E., Bernard, O. A., Ferrando, A. A. (2014)
.

22. Recurrent mutations in epigenetic regulators, RHOA and FYN kinase in peripheral T cell lymphomas. Nature Genetics, 46(2), 166–170. https://doi.org/10.1038/ng.2873 Park, S. T., & Kim, J. (2016).

23. Trends in next-generation sequencing and a New Era for whole genome sequencing. International Neurourology Journal, 20(Suppl 2), S76-83. https://doi.org/10.5213/inj.1632742.371

24. Peripheral, T.-C. (n.d.). Peripheral T-cell lymphoma, not otherwise specified , Kunal Kishor Jha 1, ✉, Suresh K Gupta 2, Harpreet Saluja 3. Kunal Kishor Jha 1,✉, Suresh K Gupta, 2.Pileri, S. A., Tabanelli, V., Fiori, S., Calleri, A., Melle, F., Motta, G., Lorenzini, D., Tarella, C., & Derenzini, E. (2021).

25. Peripheral T-cell lymphoma, not otherwise specified: Clinical manifestations, diagnosis, and future treatment. Cancers, 13(18), 4535. https://doi.org/10.3390/cancers13184535 Pizzi, M., Margolskee, E., & Inghirami, G. (2018).

26. Pathogenesis of peripheral T cell lymphoma. Annual Review of Pathology, 13(1), 293–320. https://doi.org/10.1146/annurev-pathol-020117-043821 Sandell, R. F., Boddicker, R. L., & Feldman, A. L. (2017).

27. Genetic landscape and classification of peripheral T cell lymphomas. Current Oncology Reports, 19(4), 28. https://doi.org/10.1007/s11912-017-0582-9 Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2024).

28. Correction: Satam et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology 2023, 12, 997. Biology, 13(5). https://doi.org/10.3390/biology13050286 Skrypets, T., Chauvie, S., Manni, M., Fallanca, F., Racca, M., Hitz, F., Advani, R., Ramos, C. D., Miranda, E., Tomuleasa, C., Minoia, C., Marino, D., Noyan-Atalay, F., Stepanishyna, Y., De Maggi, A., Marcheselli, L., Chang, C., Federico, M., & Luminari, S. (2023).

29. Baseline PET Total Metabolic Tumor Volume has a prognostic role in PTCLs—Data from International Prospective T-Cell Project 2.0. Hematological Oncology, 41(S2), 349–349. https://doi.org/10.1002/hon.3164_250 Stoll BTL, J., Pulitzer, M., Moskowitz, A., Horwitz, S., Myskowski, P. L., & Noor, S. J. (2021)

30. Lymphomatoid papulosis: A retrospective review of clinical characteristics, symptomatology, treatments, and associated malignancies at a single institution. Journal of the American Academy of Dermatology, 85(3), AB29. https://doi.org/10.1016/j.jaad.2021.06.143 Tetreault, M., Bareke, E., Nadaf, J., Alirezaie, N., & Majewski, J. (2015).

31. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. Expert Review of Molecular Diagnostics, 15(6), 749–760. https://doi.org/10.1586/14737159.2015.1039516 Timmins, M. A., Wagner, S. D., & Ahearne, M. J. (2020).

32. The new biology of PTCL-NOS and AITL: current status and future clinical impact. British Journal of Haematology, 189(1), 54–66. https://doi.org/10.1111/bjh.16428 Tripti, R. K., Bharathi, G., & Aditi, J. (2025).

33. Comparison of Implementation in Blood Cancer Causes and Diseases. International Journal of Trend in Scientific Research and Development, 9(1), 503–512. Updating the Definition of Cancer, Joel S Brown 1, *, Sarah R Amend 2, Robert H Austin 3, Robert A Gatenby 1. (n.d.). Emma U Hammarlund, 4. Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015).

34. Exome sequencing: Current and future perspectives. G3 (Bethesda, Md.), 5(8), 1543–1550. https://doi.org/10.1534/g3.115.018564 ,2022

35. Cagirici, H. B., Akpinar, B. A., Sen, T. Z., & Budak, H. (2021). Multiple variant calling pipelines in wheat whole exome sequencing. International Journal of Molecular Sciences, 22(19), 10400. https://doi.org/10.3390/ijms221910400

36. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England), 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

37. de Sena Brandine, G., & Smith, A. D. (2019). Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Research, 8, 1874. https://doi.org/10.12688/f1000research.21142.2

38. Karcı, H., & Kafkas, S. (2024). Impact of Mark duplicate reads during variant calling in next generation sequencing (NGS) data of Pistacia vera L. *Ereğli Tarım Bilimleri Dergisi*. https://doi.org/10.54498/etbd.2024.29

39. Peters, D., Qiu, K., Liang, P., Kotsireas, I., Melnik, R., & West, B. (2011). Faster short DNA sequence alignment with parallel BWA.

40. Riccio, C., Jansen, M. L., Guo, L., & Ziegler, A. (2024). Variant effect predictors: a systematic review and practical guide. Human Genetics, 143(5), 625–634. https://doi.org/10.1007/s00439-024-02670-5
41. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 51(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000
42. Ulintz, P. J., Wu, W., & Gates, C. M. (2019). Bioinformatics analysis of whole exome sequencing data. Methods in Molecular Biology (Clifton, N.J.), 1881, 277–318. https://doi.org/10.1007/978-1-4939-8876-1_21