UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

MASTER'S DEGREE IN

MOLECULAR BIOTECHNOLOGY AND BIOINFORMATICS

A COMPARATIVE ANALYSIS OF MUTATIONAL PROFILES IN CLONAL HEMATOPOIESIS AND ACUTE MYELOID LEUKEMIA PATIENTS

External Supervisor:

Prof. Erik Ben van den Akker

Internal Supervisor:

Prof. Matteo Chiara

Student Name:

Fazel Mohammadi

ID: 984393

Academic Year 2023-2024

ACKNOWLEDGEMENT

I am deeply grateful to everyone who supported and guided me throughout the journey of completing my master's thesis. Their invaluable contributions have been instrumental in my growth as a researcher and in the successful completion of this work.

First and foremost, I would like to express my profound gratitude to the team at LUMC. I am especially indebted to Professor Akker for his exceptional supervision, which made me an independent researcher. I am equally grateful to Jeppe for his daily guidance and determined dedication, which instilled the confidence to make sound decisions. Special thanks to Ramin, whose profound influence and guidance at every step of my project were pivotal in achieving better results and shaping my research character.

I would like to express my deep appreciation for the privilege of working under the decent guidance of Professor Chaira during my time in Milan. His invaluable assistance in crafting this thesis and navigating the intricate internal processes was a beacon of support. I am sincerely grateful for his unwavering availability and kind encouragement throughout this journey.

I am eternally grateful to my family for their unwavering support throughout my life. Their constant encouragement and belief in my abilities have been my anchor, fueling my determination and inspiring me to pursue my dreams with relentless passion.

Finally, I want to extend my heartfelt thanks to my friends in both Leiden and Milan. Living and growing together through all the good and challenging times has been an enriching experience. I am grateful for the community of wonderful people around me, whose support during the toughest moments kept us from despair and inspired us to move forward.

To everyone who has been part of this journey, your contributions have been immensely appreciated. Thank you for helping me reach this milestone.

ABSTRACT

The present study aims to elucidate the unique mutational landscapes of the DNMT3A and TET2 genes in Clonal Hematopoiesis (CH) and Acute Myeloid Leukemia (AML) cohorts. Given the critical role of these epigenetic regulators, mutations within these genes may originate distinctly and perturb various cellular pathways, potentially serving as biomarkers for early detection, prognostic evaluation, and therapeutic targeting in hematologic malignancies.

The primary objectives include the comprehensive identification of DNMT3A and TET2 gene somatic mutations in CH and AML populations, the delineation of the mutation trends and hotspots, and the detailed exploration of mutation types. Notably, certain mutation sites were significantly more prevalent in AML, suggesting potential hotspots for targeted therapeutic interventions. The study systematically identifies mutations and conducts extensive statistical analyses to characterize their distribution and frequency. Furthermore, this study develops and evaluates a machine learning classifier to discriminate between CH-like and AML-like mutations, facilitating early diagnostic processes based on somatic mutation profiles.

A Generalized Linear Model (GLM) was employed to classify mutations, achieving an accuracy of 70%, indicating moderate concordance on DNMT3A mutations. Additionally, the Random Forest (RF) model, which is adept at handling non-linear relationships and providing robust feature importance measures, demonstrated superior performance with numeric features, attaining an accuracy above 70% on the test dataset of TET2 mutations. The machine learning classifier, particularly the RF model, proved effective in enhancing the accuracy of disease classification based on mutational profiles.

This research bridges critical gaps in the understanding of the genetic etiology of DNMT3A and TET2 genes in CH and AML patients. It contributes to the body of knowledge necessary for the development of early diagnostic tools and personalized treatment strategies. The integration of machine learning approaches in this context highlights their potential to advance the precision of hematologic malignancy classifications.

Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Acute Myeloid Leukemia

Acute myeloid leukemia (AML) is a cancer that originates in the precursor cells of the myeloid lineage of blood cells, leading to the rapid growth of abnormal cells in the bone marrow, peripheral blood, and other tissues. The rapid expansion of these abnormal myeloid precursor cells disrupts hematopoiesis, specifically erythropoiesis and megakaryopoiesis, resulting in immature white blood cells.

The major risk factors for AML include smoking, a high body mass index, and exposure to ionizing radiation[1]. Smoking is the most significant risk factor, with smokers having about a 20% higher risk compared to non-smokers[2]. Elderly AML patients often present with more severe diagnoses, such as lower white blood cell counts, poorer performance status, and a higher percentage of bone marrow blasts, highlighting the impact of aging on AML outcomes[3].

AML is a relatively rare cancer that primarily affects older male adults, with incidence increasing with age. In European countries like Denmark and Sweden, the incidence of AML has been reported as 5.4 per 100,000 person-years. The Italian National Health Service reported a higher incidence of AML in 2023, at 9.0 per 100,000 person-years, with 57% of cases occurring in male adults[4].

Given the severity of this cancer and its relation to aging, a longitudinal study investigating the origins and biological drivers of AML is of significant importance.

1.2 Hematopoietic Stem Cells and Hematopoiesis

Blood cells originate from hematopoietic stem cells (HSCs) in the bone marrow. HSCs play a major role in the immune system as the source of blood and immune cells with diverse morphology and function. HSCs undergo developmental differentiation in the early stages, result-

ing in blood cell precursors called hematopoietic stem cells. A classic model of hematopoiesis, based on the multipotent features of HSCs, describes a stepwise differentiation process through which different lineages branch out[5]. The first division point separates the lymphoid and myeloid lineages. HSCs can self-renew through cell division, producing a new population of HSC clones. HSCs are distinguishable by their long-term self-renewal ability throughout an individual's lifespan, making them an emerging target for cell transplantations.

The process of hematopoiesis is represented In Figure 1, beginning with self-renewing stem cells, specifically Hematopoietic Stem Cells (HSCs) and Myeloid-restricted Stem Cells (MySCs)[6]. At the top, you can see these stem cells, which are the origin of all blood cells. HSCs differentiate into Multi-Potent Progenitors (MPPs) which can further differentiate into two main types of progenitors: Lymphoid-Primed Multi-Potent Progenitors (LMPPs) and Common Lymphoid Progenitors (CLPs), shown branching off to the sides. The figure then shows how MPPs also lead to Myeloid-Erythroid Progenitors (MyEPs). These MyEPs branch into two distinct pathways. One pathway leads to the formation of Common Myeloid Progenitors (CMPs). From CMPs, the diagram illustrates further differentiation into Megakaryocyte-Erythroid Progenitors (MEPs) and directly into Megakaryocyte Progenitors (MkPs). The MEPs are depicted as giving rise to erythrocytes (red blood cells), while MkPs lead to the production of platelets. Additionally, CMPs can differentiate into other important cell types, such as neutrophils, monocytes, and T-cells, as shown by the additional branches extending from CMPs. This overview of the hematopoiesis clearly showed the main pathways of each cell type and potential differentiation pathways, indicating that these processes can have some flexibility.

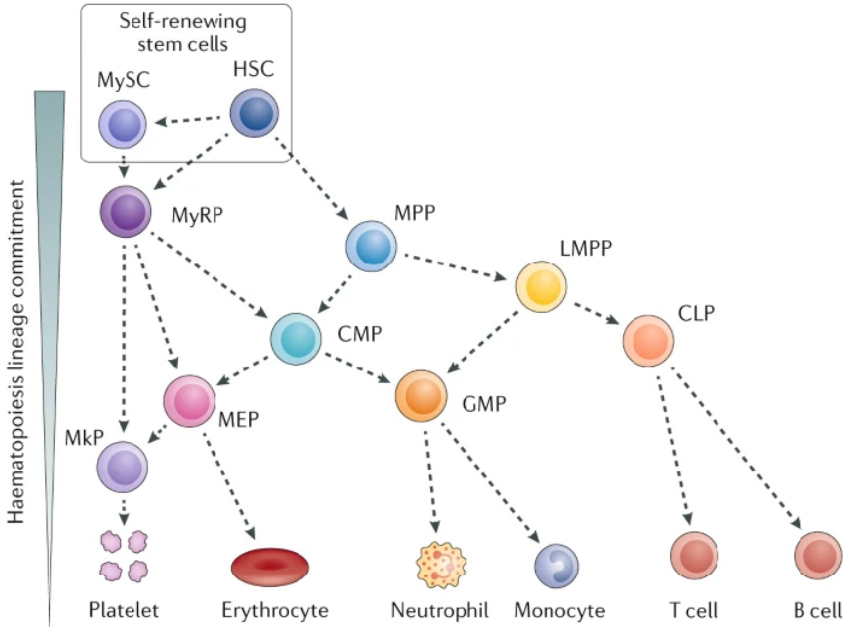


Figure 1: ***Overview of hematopoiesis process and development of blood cell formation.***
This figure shows the differentiation of Hematopoietic Stem Cells (HSCs) and Myeloid-restricted Stem Cells (MySCs) into various blood cells such as Platelets, Erythrocytes, Neutrophils, Monocytes, and T-cells. The diagram provides a visual summary of the intricate biological processes involved in blood cell formation from multipotent stem cells.[6]

1.2.1 HSCs mutation acquisition

Throughout aging, HSC clones within the bone marrow niche acquire somatic mutations in various genes. DNA damage is an inevitable part of the cell division process, while DNA damage repair deficiency plays a major role as a source of mutations in HSCs due to aging and multiple risk factors[7]. Studies have found that approximately 5% of individuals in the elderly population have somatic mutations in their hematopoietic stem cells [8]. Although most of these mutations will be repaired by the DNA damage response system, some will persist and be passed on to progeny cell clones [9].

Throughout the hematopoiesis, as a clonal and multiclonal event, mutations will be distributed within different lineages. In elderly individuals, only a few long-lived progenitors contribute to hematopoiesis. However, the pool of HSCs involved in the hematopoiesis process shifts towards

newly generated HSCs through their self-renewal ability. On the other hand, this ability also makes them prone to a higher risk of mutations. Studies revealed that HSCs through their self-renewal ability avoid DNA damage response (DDR) and showed a delay in the DNA break-end joining process by weakening the expression of DDR-related genes[10]. However, this process in some populations revealed an apoptosis-independent role for p53 enhancing the early apoptosis in HSCs[10].

Due to recent discoveries of pop-up somatic mutations in healthy older adults' blood cells, a deep investigation of mutation acquisition in HSCs is currently ongoing[11]. Although healthy individuals with somatic mutations in their blood cells may remain asymptomatic, these mutations can be detected by high-throughput sequencing methods. This state, in which HSC clones are clinically healthy but genetically mutated, is defined as Clonal Hematopoiesis (CH).

1.2.2 HSC clones evolution to malignancy state

The process of clonal evolution in hematopoietic stem cells leading to hematologic neoplasms is complex and multifaceted[12](Figure 2). Somatic mutations in blood cells and bone marrow with a variant allele frequency (VAF) of at least 5% are normally considered to be related to the CH state[13]. Any somatic mutations with a value lower than 5% VAF are more likely to be influenced by the limitations of sequencing technologies[14]. Additionally, low-frequency mutations could be sub-clonal or passenger mutations, requiring more clinical history to be considered CH-related. The mutation acquisition begins with hematopoietic stem cells, which undergo aging and early events such as mutations in specific genes like DNMT3A, TET2, and ASXL1. This condition increases the risk of hematopoietic neoplasms in peripheral blood cells, although most individuals with CH mutations will never develop blood cancers[15].

Two pathways of Clonal Hematopoiesis of Indeterminate Potential (CHIP) and Clonal Cytopenia of Unknown Significance (CCUS) are associated with increased inflammatory response and increased risk of blood cancer development, cardiovascular disease, and threatening all-cause mortality[16]. The state of having mutations with no morphological abnormalities in the blood cells and no evident clinical symptoms is defined as CHIP[17]. CHIP has been linked to in-

creased mortality in the elderly population, with a prevalence ranging from 10% to 20% in individuals over 70[18].

As the VAF of mutations in epigenetic regulatory genes, such as DNMT3A, TET2, and ASXL1 increases to a prevalence of 5-10%, the first signs of abnormalities in the hematopoietic cells may appear, followed by dysregulation in hematopoiesis. This state is defined as Myelodysplastic Syndrome (MDS) where selective pressures such as modifications in the microenvironment, and inflammation response. MDS is diagnosed clinically by decreased white blood cell production and multilineage dysplasia[19]. People with MDS undergo chemotherapy to reduce the risk of progression to acute myeloid leukemia and to improve blood cell counts, alleviating symptoms like anemia and infections[20]. It also prepares the body for potential stem cell transplantation, offering a chance for a cure[20].

The final stage shows a dispersion of cells indicating associated non-driver mutations and transformation into high-risk variants with high variant allele frequency. As the myeloid clones expand, and the VAF increases above 15%, they acquire more diverse mutations, becoming a heterogeneous hematopoietic malignancy in peripheral blood, bone marrow, and other tissues. This state is defined as Acute Myeloid Leukemia (AML). The transformation rate from MDS to AML among the elderly population is approximately 30%[21].

Throughout the stages of hematological neoplasm development, additional contributing factors like epigenetic regulators (e.g., EZH2), DNA damage response proteins (e.g., TP53), signaling molecules (e.g., SF3B1, SRSF2), and late events such as additional mutations or epigenetic changes further drive progression towards neoplasm.

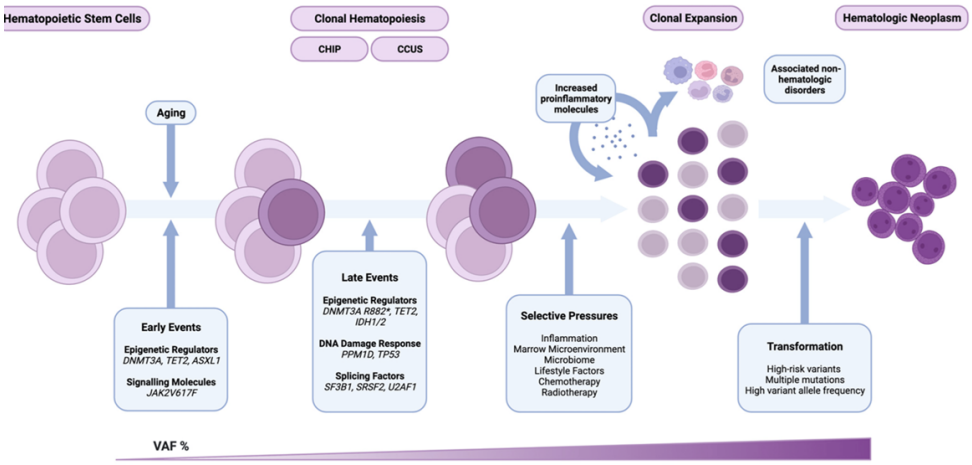


Figure 2: ***Clonal Evolution in Hematopoietic Stem Cells to Hematologic Neoplasms.***

This figure represents the process of clonal evolution in hematopoietic stem cells leading to hematologic neoplasms. It highlights the progression from normal stem cell aging to the development of cancerous conditions, emphasizing key factors and steps involved in this transformation.[12]

1.3 Clonal Hematopoiesis clinical aspect

As previously described, clonal hematopoiesis (CH) refers to a state in hematopoiesis where hematopoietic stem cells (HSCs) acquire mutations without causing symptoms in the individual. Aging is the primary mechanism of accumulation of these mutations, making CH more commonly detected in the elderly population. The initial observation of CH was made by studying the clonal origin of chronic myelocytic leukemia (CML) by Philip Fialkow et al.[22]. With advancements in next-generation sequencing (NGS), studying early clonality in the hematopoietic pathway has become more accessible through high-throughput sequencing datasets[23]. CH is a driver of hematological malignancies and is associated with nonmalignant diseases such as atherosclerotic cardiovascular disease (CVD)[24]. Recent population-wide studies have shown that CH can be a significant risk factor for new-onset type 2 diabetes based on cholesterol levels in a five-year follow-up study[25].

In addition to aging, an epidemiological study revealed that smoking accelerates the development of CH and is associated with nonhematological diseases[26]. Another study found that CH, observed in 62% of individuals over 80 years old, could also be associated with chronic

obstructive pulmonary disease (COPD)[27]. Stem cell transplantation studies have reported a common occurrence of CH among patients with poor stem cell mobilization, resulting in therapy-related myeloid neoplasms (t-MN)[28]. This study defined the criteria for the classification of stem cells before transplantation.

In conclusion, the study of CH in the elderly population has become an emerging field of research for oncologists and developmental biologists. A thorough understanding of the origin of CH and its early detection can play a pivotal role in diagnosing hematological malignancies and cardiovascular diseases. Recent studies have indicated that the detection of CH in healthy individuals can lead to feelings of anxiety[29]. Therefore, the study of CH is crucial not only for diagnosis but also for addressing the psychological impact on patients with somatic mutations.

1.4 Acute Myeloid Leukemia clinical aspect

Acute Myeloid Leukemia (AML) is defined as a cancerous malignancy in peripheral blood and bone marrow tissues in the elderly. Patients appear with a series of symptoms starting with reduced stamina, bruising and bleeding, bone pain, and gum inflation with frequent infection. Based on the European LeukemiaNet report in 2022, more than 45% of patients are at high risk of AML[30]. The mortality rate of AML increases exponentially in patients over 60 years with 53% of AML deaths in individuals over 75 years[31]. The most common diagnosis methods are blood count measurement from peripheral blood smear and bone marrow biopsy[32]. Recent advancements in molecular genetics laboratory tests allowed precise and personalized diagnosis of AML and its subtypes [33].

Considering the complexity of the biology behind AML, patient treatment remains challenging. One of the earliest treatments for patients with blood hematological malignancy was combination chemotherapy[34]. Although some patients initially showed remission, a 55-year-old patient relapsed within six months, highlighting the limitations of this method in treating AML subtypes[35]. In a comparative analysis of AML treatment Cornelissen et al., described a treatment based on hematopoietic stem cell transplantation (HSCT) involving the replacement of a patient's bone marrow with a healthy stem cell donor, which showed a reduction in the rate of

relapse and long-term mortality risk[36]. However, HSCT is prone to a high risk of infection, graft versus host disease, graft rejection, and transplant-related mortality[37].

In a more fundamental and personalized therapeutic approach, drugs are designed to target specific cancer-driven genetic mutations called targeted therapeutics[38]. In 30% of AML Patients, the mutation in tyrosine kinase receptor (FLT3) allowing tumor cell proliferation in AML clones was associated with a poor prognosis. Midostaurin (Rydapt) is designed to target FLT3 mutation and inhibit uncontrolled cell proliferation and tumor growth. Additionally, mutations in the IDH gene family affect the pattern of cell proliferation and producing oncometabolites. In a case study by DiNardo et al. Ivosidenib, a drug designed to inhibit the production of oncometabolite 2-hydroxyglutarate, showed a 30.4% rate of remission in relapsed in AML patients[39]. Monoclonal antibodies also play a role as targeted therapeutics of AML, targeting the surface of tumor cells and inhibiting cancer cell growth[40]. In a study on Gemmatuzumab, targeting CD33 in AML cells, patients receiving the medication showed a 34% reduction in the risk of relapse[41].

In conclusion, many chemo and drug-based therapies are proposed for the treatment of AML patients and each of them showed their advantages and limitations. In comparative studies of targeted therapy versus chemo- and immuno-therapy, targeted therapy based on the genetic sub-group showed more effectiveness and lower toxicity[42]. On the other hand, combination therapy is prone to the risk of overlapping toxicity highlighting the importance of focused research on genetically defined personalized therapies[42].

1.5 Somatic mutations on DNMT3A, TET2 genes

To understand the pathogenesis and progression of clonal hematopoiesis (CH), it is essential to focus on the origin and impact of the somatic mutations that define it. These genetic alterations in non-germline cells are key to understanding the development and progression of these conditions. As depicted in Figure 3, the accumulation and interaction of mutations over time can lead to clonal expansion and eventually, malignancy[43]. The process of clonal expansion is made possible by early mutations in genes such as DNMT3A, TET2, ASXL1, JAK2, PPM1D,

SF3B1, and SH2B3 (represented by green circles). As time progresses, subsequent cooperating mutations occur in genes such as FLT3, IDH1/2, NPM1, WT1, NRAS, CEBPA, U2AF1, PHF6, and STAG2 (represented by red triangles).

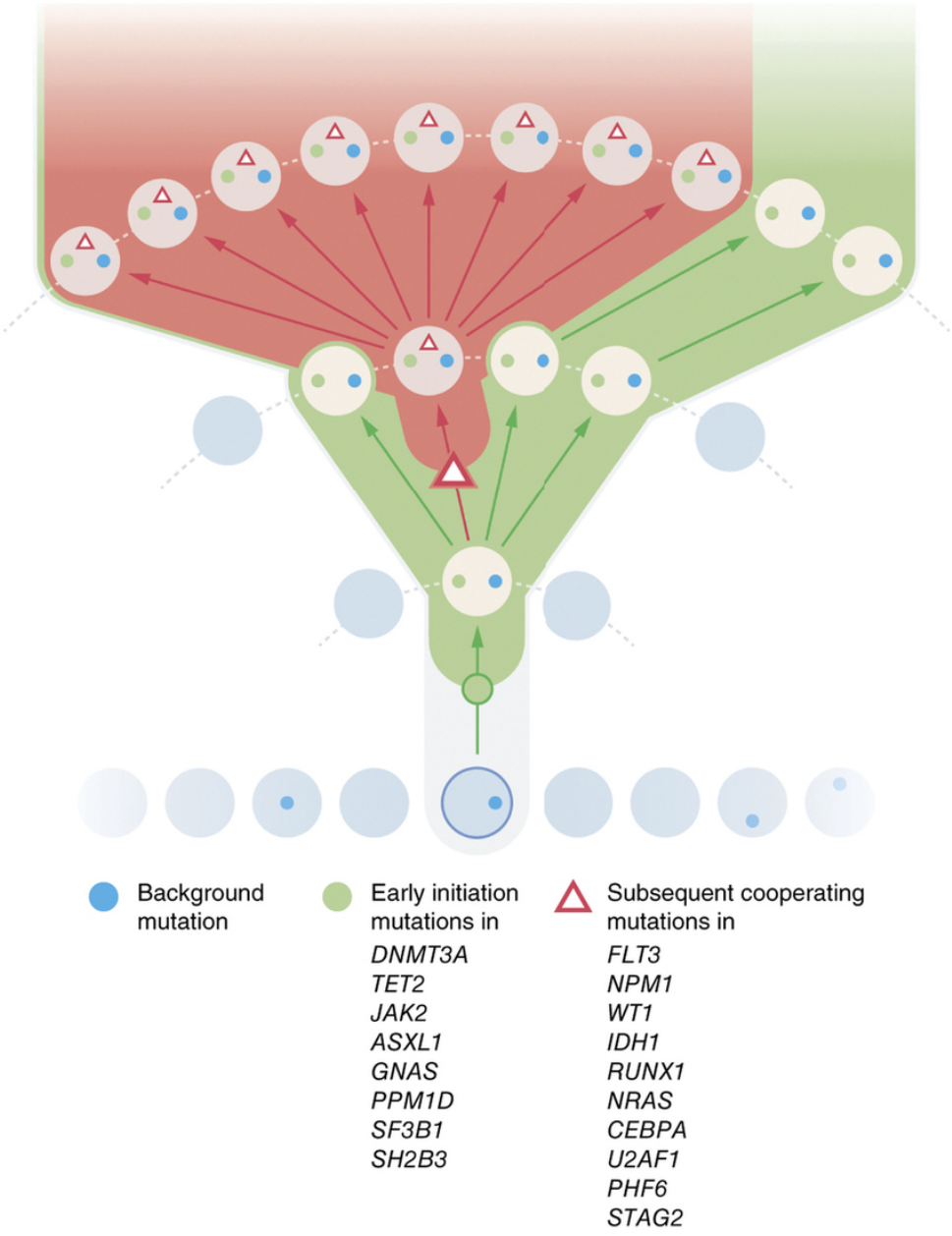


Figure 3: ***Visual Representation of the Evolution of Mutations in Clonal Hematopoiesis Progressing to Acute Myeloid Leukemia.***

The figure illustrates the accumulation of diverse mutations over time, starting with background mutations (blue circles), followed by early initiation mutations (green circles), and subsequent cooperating mutations (red triangles). This progression highlights the complexity and diversity of the mutational landscape in the transition from CHIP to MDS to AML.[43]

The most commonly mutated genes identified in CH are those related to epigenetic regulation, such as DNMT3A, TET2, and ASXL1, accounting for nearly 90% of all mutations[11]. In AML patients with a wider range of mutated genes, the frequency of DNMT3A mutations is around 20%, and TET2 mutations are around 12%[44]. Considering the prominent presence of mutations in DNMT3A and TET2 genes across CH, MDS, and AML stages, these genes are considered pre-clinical AML markers[45]. These somatic mutations can be synonymous or non-synonymous, including truncating and missense mutations.

The DNMT3A gene family encodes methyltransferase enzymes that regulate the methylation of CpG islands upstream of genes. Mutations in DNMT3A can lead to hypermethylation of tumor suppressor gene promoters[46]. The most common mutations are R882H, R882C, R882P, and R882S, representing over 50% of all missense mutations. Individuals with CH and heterozygous R882 mutations, even in normal karyotype AML patients, have also shown DNA hypomethylation at specific locations[47].

TET2 encodes an enzyme that helps oxidize methylated cytosines, facilitating the demethylation of genomic DNA. TET2 is involved in various physiological and pathological processes, including cell fate determination and cancer development. It plays a particularly important role in the hematopoietic hierarchy, and loss-of-function mutations in TET2 are known to be major drivers of myeloid malignancies in humans. Studies have shown that TET2 loss can provide a proliferative advantage to HSPCs compared to wild-type HSCs after exposure to TNFα and IFNγ[48].

1.6 NGS-based methods on the early detection of CH mutations

Next-generation sequencing (NGS)-based methodologies push the boundaries of molecular biology studies applied in disease diagnosis. Two main principal approaches, whole genome sequencing (WGS) and whole exome sequencing (WES), play the streamlined role of CH detection in peripheral blood and bone marrow tissue[49].

The exome, which covers about 1-2% of the genome, is the target of WES, focusing on those

regions encoding proteins that are most critical for detecting disease-causing mutations[50]. This technique allows us to explore genetically defined diseases effectively and is economical and efficient. Conversely, WGS examines the entire genome including coding and non-coding sections enabling a full view of genetic composition[51]. Such additional capacity makes it possible to identify mutations or other genomic changes that may be missed by WES, especially in non-coding regions that can also drive clonal expansion and disease progression. Figure 4 shows a comparison between WES and WGS in terms of their respective genome coverage as well as related costs. Despite its limited scope because it focuses on coding regions and has lower costs compared to WES, the wider coverage of WGS allows for more comprehensive genetic analysis although at increased cost and data complexity[52].

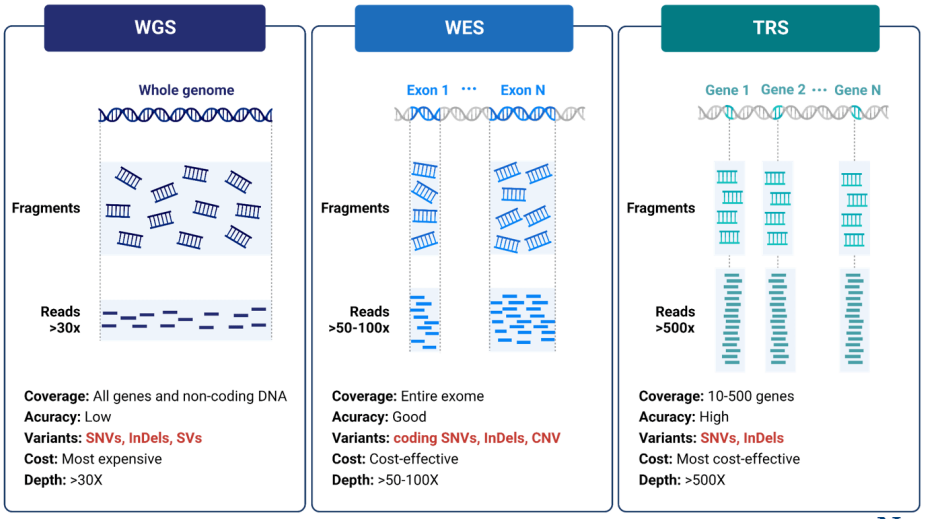


Figure 4: ***Comparative overview of Genomic Sequencing Techniques.***

Whole Genome Sequencing (WGS), and Whole Exome Sequencing (WES) - A Study on Coverage, Accuracy, Variants, and Cost[52].

According to recent studies, about 25% of patients with early detection of CH can develop subsequent hematological malignancies[53]. Further, patients with early detected CH showed not only hematological cancers but also an accelerated risk of heart failure[54]. Due to these facts, early detection of CH state is associated with an early diagnosis and targeted therapies for patients with the risk of malignancy improving their survival[55].

Low coverage and detection sensitivity of WGS and WES restrict early detection of CH-related

mutations and introduce false discoveries. Additionally, due to the low rate of mutation frequency, the discovery of true CH-related mutations is challenging because of sequencing artifacts and germline variants[56].

Further, investigation of AML clones evolved from CH requires a deep understanding and pattern detection of CH mutational profiles. On the other hand, more effective and focused methodologies such as targeted and amplicon-based sequencing, while advantageous for their precision, face budget constraints when applied to large-scale studies[57]. However, despite their effectiveness, these methods still do not achieve full coverage of mutations associated with clonal hematopoiesis. Considering effective sequencing methods, however, full coverage of CH-related mutations is still limited[58].

1.6.1 Evolutionary tumor growth studies

The advent of NGS-based technologies opened the way to the study of tumor evolution and growth dynamics. Starting with the detection of genetic mutational patterns, researchers can study tumor development in diverse sub-populations within a single tumor. Considering the mutational profile of patients, evolutionary tumor evolution studies facilitate a major mutation classification into driver or passenger. This step is crucial in identifying the main mutations in tumor progression or mutations as a byproduct of uncontrolled cell growth[59]. In tumor evolution studies, the nature of endogenous and exogenous genome instability processes can also be investigated as tumor subclones[60].

1.7 Machine Learning Approaches and Application in Omics Studies

Machine learning (ML) is a branch of artificial intelligence (AI) that is based on the development of algorithms that allow computers to learn from and make predictions or decisions about data. Unlike traditional programming where specific instructions are written for every task, machine learning systems are trained using data to recognize patterns and make decisions with

little human intervention[61].

Integrating machine learning methods into omics studies has altered the biomedical research world regarding unveiling meaningful patterns from high-dimensional datasets. Genomics, transcriptomics, proteomics, and metabolomics are among the technological tools in omics that produce massive data sets that come with their challenges and opportunities in terms of analysis. Machine learning provides some of the most powerful tools for analyzing these intricate data sets making it possible to identify biomarkers, disease subtypes as well as therapeutic targets[62]. The significance of ML approaches in omics studies will be explored in this section.

1.7.1 Machine learning techniques applied in bioinformatics

Broadly speaking, machine learning can be divided into three primary types: supervised learning, unsupervised learning, and reinforcement learning (Figure 5). Each has its unique features, applications, and methodologies.

One of the most fundamental techniques in machine learning is training a predictive model based on known results. In supervised machine learning, the model is trained based on parameters and features, called independent variables, to predict a desired label known as the dependent variable. This algorithm training aims to learn unseen correlations and map from independent to dependent variables. Supervised learning is divided into two major sections: classification, where the dependent variables are distinct labels, and regression, which defines the dependent variable as a continuous value. As an example of supervised learning in the field of omics, classification can be illustrated through RNA-seq data analysis, where an algorithm is trained on gene expression patterns to classify new genes such as upregulated or downregulated [62]. In the same dataset, when a model is trained to predict the expression level of a gene based on the expression patterns of other genes within the dataset, this is defined as regression.

On the other hand, in unsupervised learning, the algorithm is trained based on independent variables and a set of features within the dataset without labeling or dependent variables. The main aim of this approach is to discover underlying patterns or structures within the data. Con-

sidering pattern recognition within data, unsupervised learning is divided into two main approaches: clustering and dimensionality reduction. Clustering techniques refer to categorizing and grouping data points sharing common or close inherent characteristics. Dimensionality reduction, a prevalent method in the omics field, reduces the feature complexity of the dataset while preserving essential characteristics to improve the interpretation and visualization of the data. Both dimensionality reduction and clustering are fundamental steps in the analysis of single-cell RNA sequencing (RNA-seq) to reduce the complexity of a large number of cell gene expressions and group cells based on their apparent expression profiles[62].

In the field of omics, obtaining labeled data can be challenging and expensive. Semi-supervised learning (SSL) and self-supervised learning (Self-SL) are powerful techniques that leverage unlabeled data alongside labeled ones. In SSL, the model's performance is improved by using both labeled and unlabeled data (Figure 5). For example, in cancer genomics, cancer-associated genes have been identified using SSL by merging multiple multi-omics datasets where only a few parts of the data are marked with known gene roles. By exploiting the vast amount of unlabeled data, this method results in more accurate and robust gene discovery[63]. Additionally, self-SL generates supervisory signals from within a given dataset. This approach is ideal for omics data analysis when there is limited labeled information[63]. Self-supervised approaches on protein sequences, for example, provide training labels through techniques like masked language modeling, which lead to highly accurate predictions about protein structures[55]. These models learn general feature representations that can be fine-tuned for specific tasks such as predicting protein function or mapping interactions.

Reinforcement learning (RL) is the third paradigm of machine learning approaches. Unlike supervised learning, the algorithm is not given the correct dependent-independent variable pairs but learns from the consequences of actions in the environment of the dataset. In this method, the algorithm is trained to make decisions based on the interaction of a defined agent and a rewarding system for its actions. The learning or decision-making in reinforcement learning is developed based on the summation of rewards achieved by the agent and the states fed into the agent at each iteration of action. Despite its potential, RL faces several challenges in computational resources and accurate modeling of complex biological systems, besides the critical

hurdle of accepting decisions from the RL model, especially in clinical settings[64].

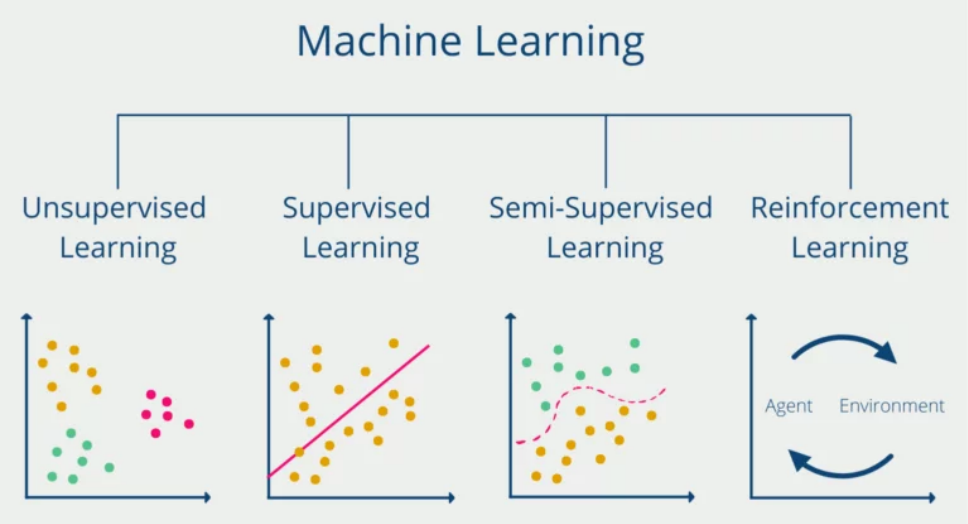


Figure 5: ***Overview of the Four Main Types of Machine Learning.***

Unsupervised Learning is represented by a scatter plot with two distinct clusters of dots in different colors, signifying the algorithm's ability to identify patterns and group unlabeled data. Supervised Learning is depicted by a scatter plot with a clear decision boundary separating two groups of colored dots. Semi-supervised learning on a scatter plot includes additional uncolored dots representing unlabeled data, indicating the algorithm's ability to learn from a mix of labeled and unlabeled data. Reinforcement Learning is illustrated by an interaction loop between an 'Agent' and 'Environment', depicted by arrows forming a circular flow[65].

1.7.2 Machine Learning Applications in Omics: Case Studies

The advantages of Machine Learning (ML)-based approaches in the Bioinformatics area are undeniable as records demonstrated tremendous advancement in ML paradigms applied in omics data analysis[66]. Besides the fundamental application in bioinformatics pipelines, training a predictive model based on large-scale omics and multi-omics datasets became emerging.

With regard to all types of cancer, a new study on ML models and pan-cancer classification of multi-omics data presents a complete picture of the latest developments in oncology. Despite the problems brought by growing amounts of information stored in multiple formats and unanticipated diversity inside tumors with identical histogenesis, integration has demonstrated its potential for advancing our perception of cancer from multi-omics perspectives[67].

Another study also suggested an ML framework of multi-omic integration to predict cancer-related LncRNAs. As the true association mechanism between lncRNAs and complex diseases remains challenging, this method captures essential information and enhances the performance of predicting cancer-related lncRNA[68].

In evolutionary tumor research, a curated omics data analysis using multiple approaches provides gene expression profiles across various cancers. This study aims to investigate possible frequently affected molecular mechanisms underlying tumorigenesis by systems biology approach along with ML-based analysis. Despite its inherent heterogeneity in cancer biology, this work examines microarray and RNA-seq data from diverse angles using an integrative tool kit including machine learning as well as system biology tools toward identifying primary expressions[69].

1.8 Objective of this study

The primary objective of this project is to investigate and analyze the unique mutations in the DNMT3A and TET2 genes within CH and AML populations. Each unique mutation in epigenetics regulatory genes can derive from a distinct origin and disrupt different pathways in clonal expansion. Besides, each of them has the potential to play the role of biomarker in the early detection of clonal state, prognosis, and therapeutic targets in hematologic malignancies. Existing studies focused on overall mutation analysis within the population which remains a lack of a deeper study of genetics behind CH.

The first objective of the study is to systematically identify and document specific mutations unique to DNMT3A and TET2 genes in CH and AML populations and perform a statistical analysis comparing mutation sites on the DNMT3A and TET2 genes in AML as well as CH populations to reveal any discernible trends or hotspots that are peculiar to each condition increasing our understanding of the dispersion of the mutations.

Additionally, the project explores such types of mutations as point mutations, insertions, and deletions among both CH and AML patients. Consequently, this research effort highlights mechanisms of molecular clonal expansion as well as leukemogenesis by identifying differences in mutation patterns between the two conditions. In other words, that is not only type analysis but also an investigation into how these mutations appeared first and the ways they spread within populations. This takes into account studying possible environmental, lifestyle, and genetic factors that may lead to mutation occurrence in DNMT3A and TET2 genes.

The other primary aim is to create and test a classifier based on machine learning that can differentiate between CH-like and AML-like mutations, to facilitate the early diagnosis based on the observed somatic mutational profiles, and eventually discriminate mutations detected in CH and AML populations. Then, this thesis will concentrate on testing its ability to accurately sort out novel changes emerging in either population.

These endeavors are aimed at bridging the existing knowledge gap about what causes CH and

AML to facilitate early diagnosis, guide therapeutic choices, and contribute to global understanding of hematological malignancies.

2 RESULTS

2.1 Mutational Spectrum

2.1.1 Overall comparison of mutations in CH and AML

The CH somatic mutation data were collected from the high-throughput genome and exome sequencing studies conducted by Miller et al., Uddin et al., and Zink et al. [70, 58, 43]. The AML somatic mutation data were obtained from public cohorts (TCGA, TARGET, BEAT, Leucegene) and in-house cohorts of AML patients[71, 72, 73, 74, 75]. Reannotation and converting genome assembly references of mutations in Clonal Hematopoiesis (CH) and Acute Myeloid Leukemia (AML) were performed to integrate and harmonize all the datasets. The dataset for CH somatic mutations is collected from comprehensive studies We found 3032 variants in the CH datasets and 9442 variants in the AML datasets representing somatic mutations observed across all patients in the selected studies. After selecting mutations with unique features among all the datasets, we have 1875 somatic mutations in the CH dataset and 6364 somatic mutations in the AML dataset. The analysis found that AML samples exhibited a greater number of mutations, more than 5000 mutations, in genes that were not mutated in CH samples. We found 1665 mutations in the CH dataset and 1775 mutations in the AML dataset on the genes commonly mutated in both conditions by filtering out mutations on genes exclusively mutated in one state such as NMP1.

The analysis of distinct mutations in commonly mutated genes in CH and AML samples reveals several notable patterns in Figure 6. In CH samples, DNMT3A, TET2, and ASXL1 were found to be genes with 684, 459, and 173 unique somatic mutations respectively. AML samples, on the other hand, have 270, 266, and 53 unique somatic mutations respectively in the DNMT3A, TET2, and ASXL1 genes. This finding confirms that not only is the frequency of unique mutations in DNMT3A, TET2, and ASXL1 higher but also more diverse than mutations found in AML samples[76]. The genetic alterations observed in AML samples are broader in scope and more complex than those observed in CH samples. These results highlight a potential difference in the mutational landscape between CH and AML.

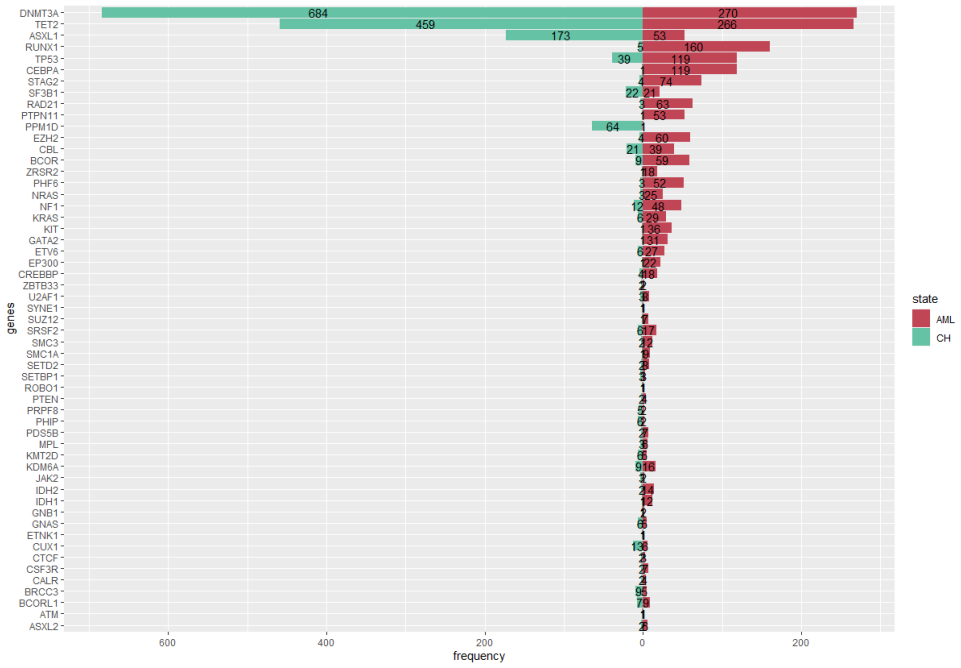


Figure 6: ***The mutation frequency plot of the common mutated genes within the CH and AML populations.***

The plot represents the number of unique mutations on each gene and the frequency of their states as a representation of the library space.

To evaluate the quality of the mutation data in CH and AML conditions, we compared the average Variant Allele frequency (VAF) values of the mutations on DNMT3A, TET2, and ASXL1 (Figure 7). VAF value is a metric of the proportion of sequencing reads supporting a specific genetic locus within each sample measured through variant calling of NGS data. A VAF of about 15% for DNMT3A, TET2, and ASXL1 mutations in CH samples indicates that these mutations are present in a subset of cells but do not dominate the hematopoietic cell population. In bone marrow and peripheral blood, a small number of mutated cells coexist with non-mutated cells according to the concept of CH[77].

AML has a VAF of around 40% for mutations in DNMT3A, TET2, and ASXL1, indicating a higher prevalence of these mutations and their clonal dominance in leukemic cells (Figure 7). As a result of AML, malignant clones are accumulated and expanded, including those affecting epigenetic regulators such as DNMT3A, TET2, and ASXL1. Based on the higher VAF in AML samples, it appears that these mutations have a more significant impact on the leukemic cells[72].

It is important to keep in mind that the VAF values mentioned are averages and may vary among the individual samples within each group. Furthermore, mutations within DNMT3A, TET2, and ASXL1 genes may also affect VAF values. A slightly higher average VAF value (0.43) in DNMT3A could be due to clonal dominance, and earlier mutation acquisition in the leukemogenesis process. This step was taken to ensure the quality of the data collection by analyzing all unique mutations in the sample population.

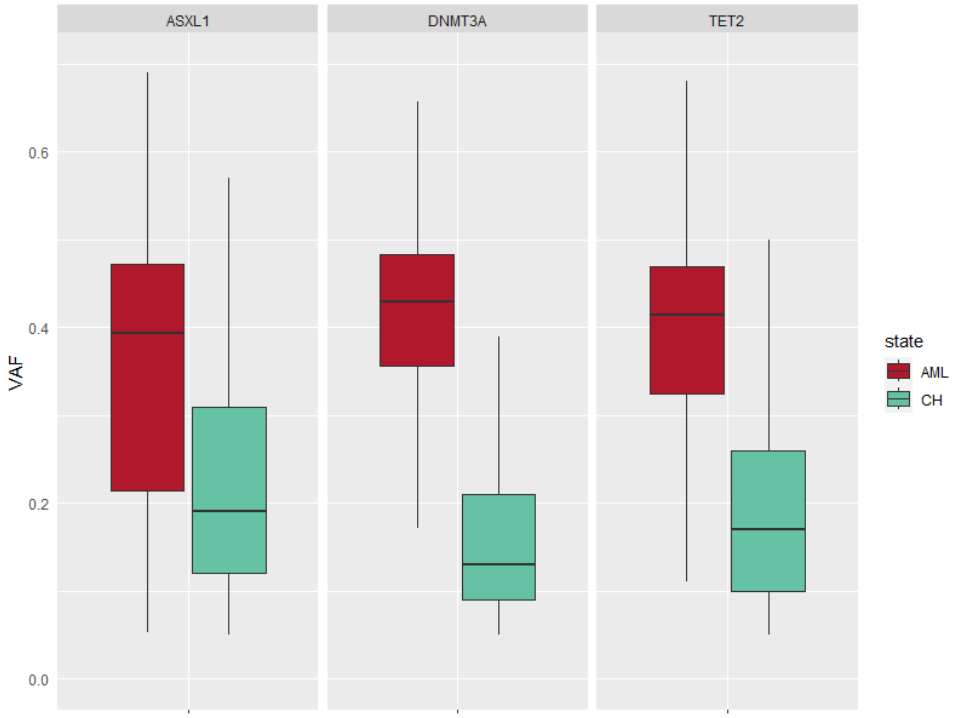


Figure 7: ***VAF box plot showing the average VAF values for the top three CH-related mutated genes within the AML and CH populations of patients.***

2.1.2 Pairwise comparison of DNMT3A mutation spectrum

To investigate similarities and differences in the CH and AML mutational profiles, we performed a pairwise comparison of mutations on the top three most mutated genes in CH. We have found critical similarities and differences between the CH and AML mutations spectrum.

The lollipop and ridge plot showed that DNMT3A mutations are more diverse and distributed in

CH samples compared to AML samples (Figure 8A). Not only the hotspot region, R882 in the MTase domain but also the ADD (Amine-Terminal Domain) domain has a considerable density of missense mutations in CH samples. On the other hand, we have seen a peak of Frameshift insertion mutation between ADD and MTase domain in AML samples. We assume this peak of a frameshift mutation in the middle of the amino acid sequence would lead the DNMT3A protein to an immature translation and nonfunctional protein. The loss of DNMT3A function can disrupt normal gene regulation, contributing to the pathogenesis of AML by promoting clonal expansion of mutated cells[78].

In contrast, the CH population, which does not exhibit a high peak of frameshift mutations in DNMT3A, suggests that mutations within this group are predominantly passenger mutations. These passenger mutations do not significantly disrupt cellular functions or confer a selective advantage, hence, they do not lead to malignant transformation[11].

Through these mutation spaces in CH and AML, we have found a considerable number of shared mutations between both conditions (Figure 8B). However, a decent number of mutations happen in CH and AML which could be explained as clonal expansion and mutation perseverance. In addition, only a small part of the mutation in the AML samples is found, particularly within the AML population. However, most of the DNMT3A mutations are preserved from the CH to AML state. The particular subsets of DNMT3A mutation would help us in defining labels in the supervised classification of the mutations.

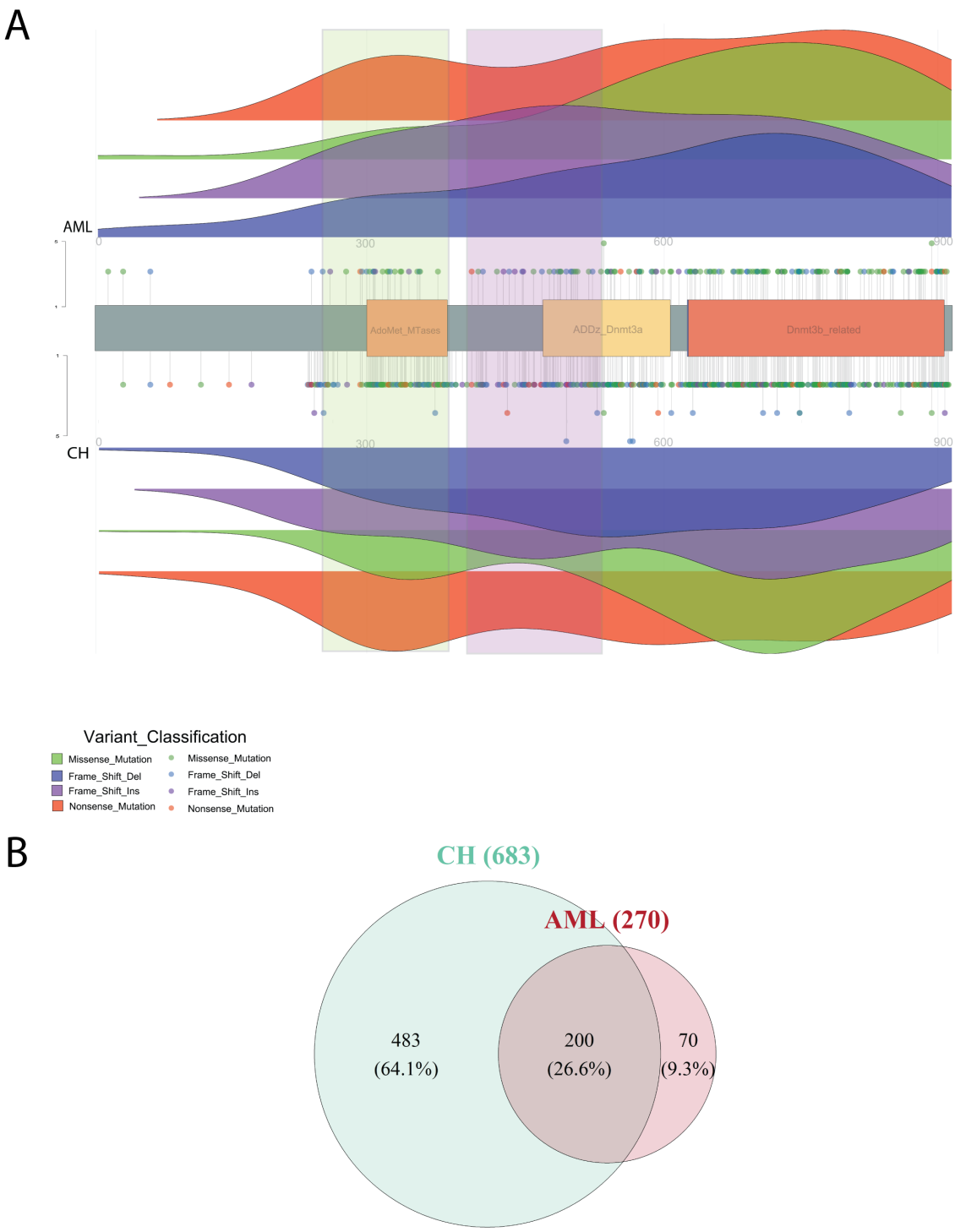


Figure 8: ***DNMT3A mutational spectrum in CH and AML samples.***

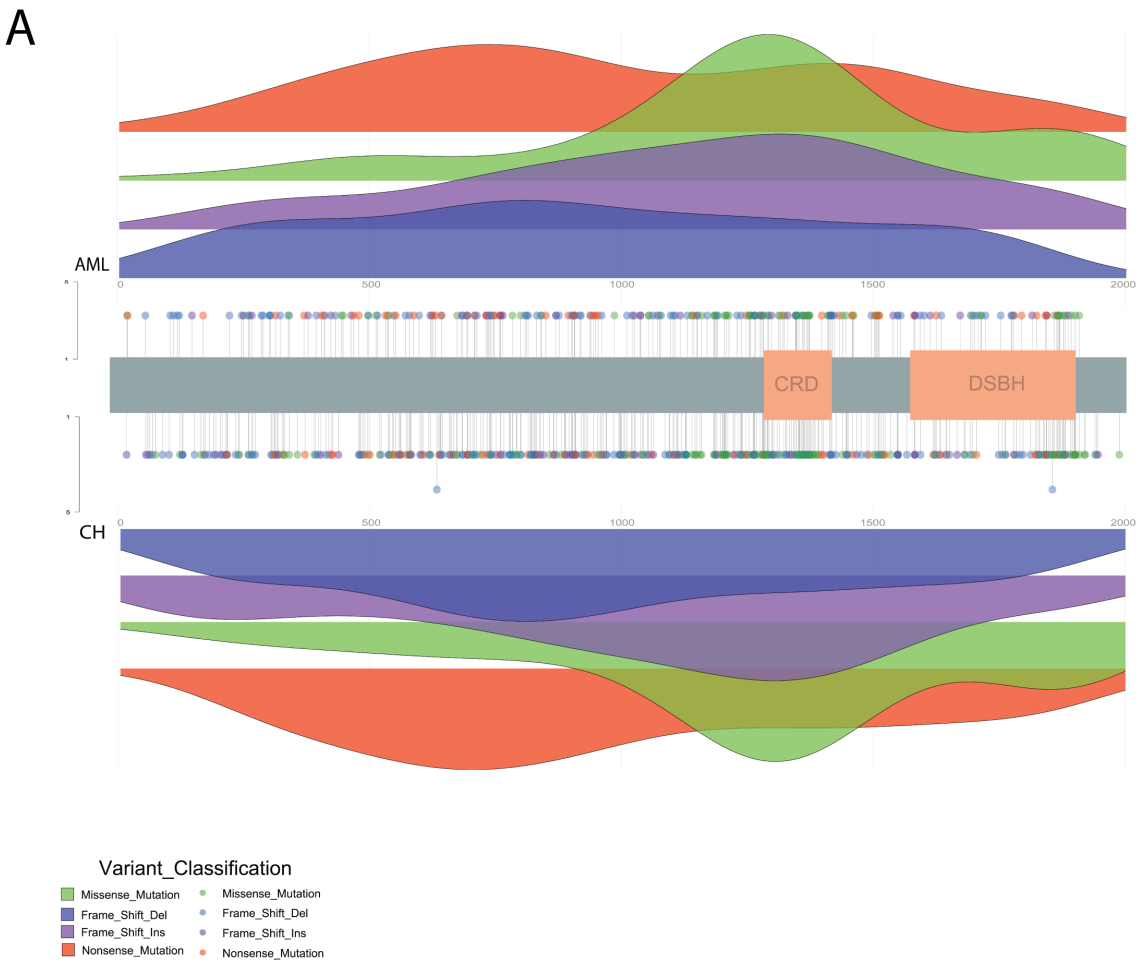
The figure presents a comprehensive comparison of DNMT3A somatic mutations in AML and CH samples. ***Panel A*** *displays lollipop and ridge plots for both conditions, illustrating the position and density of mutations along the DNMT3A protein sequence. The top half of Panel A represents AML samples, while the bottom half represents CH samples.* ***Panel B*** *features a Venn diagram that quantifies the unique and shared mutations between AML and CH samples.*

2.1.3 Pairwise comparison of TET2 mutation spectrum

According to the mutational spectrum of TET2 mutations in CH and AML samples, although TET2 has a high frequency of mutations in the conditions, the mutation distribution patterns are consistent (Figure 9A). In addition, the pattern of mutation impact on different regions of the protein is the same in CH and AML mutations with a slightly small difference in Frame Shift Insertion and Missense mutations.

We have found a small subset of mutations in TET2 which is shared between CH and AML populations (Figure 9B). The small proportion (13%) of shared mutations between CH and AML suggests that while there is some overlap in the mutation patterns, most mutations in CH are likely passenger mutations that do not significantly drive disease progression. The shared mutations might represent early driver mutations that provide a growth advantage and could be necessary for the initial clonal expansion in both CH and AML. In CH, mutations may accumulate gradually with age and do not necessarily lead to malignant transformation. In contrast, AML likely requires a combination of specific mutations, including those unique to AML, to drive the aggressive, malignant phenotype.

Although the similar pattern of mutation spectrum in TET2 for CH and AML populations, the huge difference in the subset of CH-only shared, and AML-only mutations would help us in training the classifier.



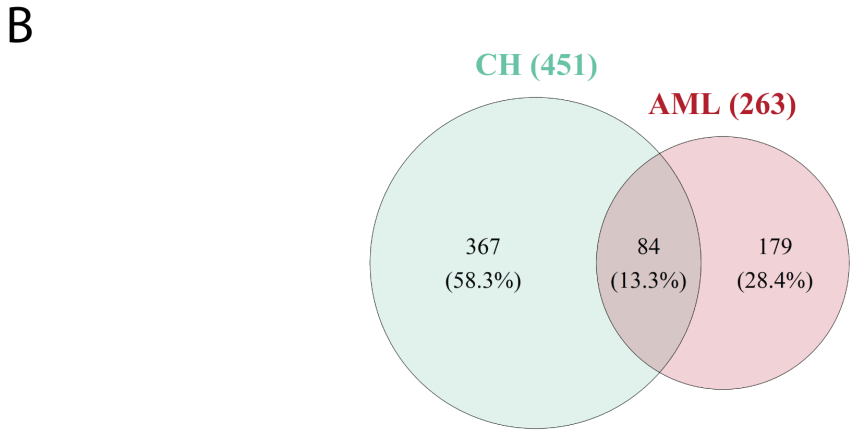


Figure 9: ***TET2 mutational spectrum in CH and AML samples.***

The figure presents a comprehensive comparison of TET2 somatic mutations in AML and CH samples. ***Panel A*** *displays lollipop and ridge plots for both conditions, illustrating the position and density of mutations along the TET2 protein sequence. The top half of Panel A represents AML samples, while the bottom half represents CH samples.* ***Panel B*** *features a Venn diagram that quantifies the unique and shared mutations between AML and CH samples.*

2.1.4 Pairwise comparison of ASXL1 mutation spectrum

According to the mutational spectrum in CH and AML, ASXL1 mutations in CH samples targeted the ASXH (N-terminal ASX Homology) domain of the protein. Compared to AML, in the CH samples, we have a more distributed pattern of mutations (Figure 10A). The mutations in CH and AML follow the same pattern in the impacted regions by different classes of mutation with minor differences in CH samples which is the consequence of a higher number of mutations in the CH samples. Peaks of frameshift insertion and missense mutations close to the PHD domain in the CH population suggest that a part of mutations in this category are passenger mutations and would not lead to the next clones.

On the other hand, only a small portion of mutations in AML are shared with the mutations in the CH condition (Figure 10B). The Venn diagram showed a biased subset of mutations in the CH population, in which 70% of mutations occurred only in the CH. This finding suggests a low-quality standard for training a classifier due to the high similarity in the pattern of mutations and biased subset. However, we decided to continue our study focusing on DNMT3A and TET2 mutation profiles.

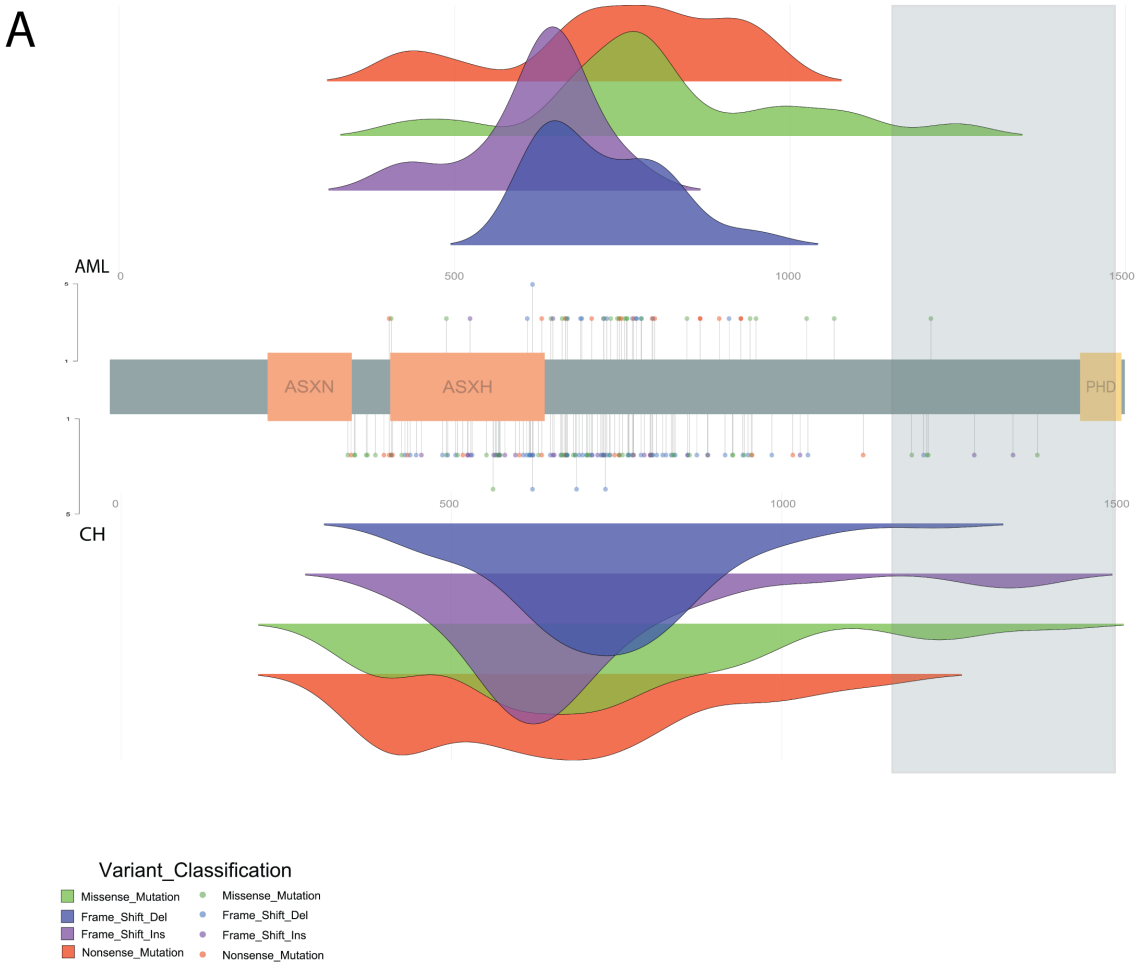


Figure 10: ***ASXL1 mutational spectrum in CH and AML samples.***

The figure presents a comprehensive comparison of ASXL1 somatic mutations in AML and CH samples. ***Panel A*** *displays lollipop and ridge plots for both conditions, illustrating the position and density of mutations along the ASXL1 protein sequence. The top half of Panel A represents AML samples, while the bottom half represents CH samples.* ***Panel B*** *features a Venn diagram that quantifies the unique and shared mutations between AML and CH samples.*

2.1.5 DNMT3A, TET2, ASXL1 mutations within AML subtypes

AML subtyping analysis is done based on the transcriptomic profile of AML samples by the team[79]. Using mRNA-seq data from 1337 patients, this study presents a comprehensive survey of transcriptomic subtypes in AML. We mapped our mutation dataset on their clusters to study the state of mutations in different subtypes of AML patients (Figure 11). DNMT3A mutations across AML subtype clusters showed a small portion that happened only in AML samples. However, most of their mutations are considered shared mutations, which also happen in the CH state.

On the other hand, TET2 mutations across AML clusters are uniquely observed in AML samples. Only a small fraction of TET2 mutations in each cluster are considered as shared mutations with CH samples. In NPM1 clusters which are under precise investigation by our team, we have found that most of the TET2 mutations are shared-like, meaning that they may be raised from the CH state of the clones. This finding provided insight for the team's main research and consideration for deep sequencing of TET2 genes within AML patients for further investigation of early clone development.

Figure 11: ***The fraction of the presence of DNMT3A, TET2, and ASXL1 AML-only mutations across AML subtype clusters was found based on their transcriptional profile.***

In this figure, we represented the percentage of AML-like mutation in AML subtype clusters based on the transcriptional profile found in the team. The higher the value in this heatmap shows the higher AML-like somatic mutation within the clusters.

2.2 Mutational Signatures

2.2.1 Mutational signatures of DNMT3A and TET2 and all genes

Mutational signatures were used to compare and analyze distinct mutational profiles. COSMIC generates a comprehensive mutational catalog, which captures the mutational profile of some reference samples of cancer patients[80]. The mutational signatures are defined based on the single nucleotide variants (SNVs) and the neighboring sequences representing SNV patterns in cancer genome samples. The signature patterns are defined as single-base substitutions (SBS).

The pattern of mutation signatures is statistically calculated based on the type of substitution (C>T, C>G, C>A, T>A, T>C, T>G) and one nucleotide upstream and downstream of SNVs. Currently, 96 SBS classes are defined on the COSMIC website. The alternative substitutions are translated based on the complementary DNA strands. Using these patterns and clinical data of reference samples, a catalog is developed to address the mutational patterns to the origin of mutations.

Comparing signatures of unique mutations in AML and CH conditions, we observed differences in the patterns of each substitution in all and specific genes (Figure 12). In all the signatures we mostly have a higher frequency in C>T substitution and various patterns in other substitutions. The higher frequency of mutations in C>T substitution could be assumed due to the CpG islands mutation in chromatin regulatory genes which included the majority of somatic mutations in both CH and AML conditions[81]. Signatures on DNMT3A and TET2 are the same in both conditions and it showed a conservation of the mutations through the development from CH to AML.DNMT3A specifically showed slightly higher differences in C>A and T>C substitutions. On the other hand in the TET2 profile, we observed a more conserved pattern of signatures with a slightly higher frequency in AML conditions. These findings suggest a very low probability of difference in the origin of mutations in both conditions.

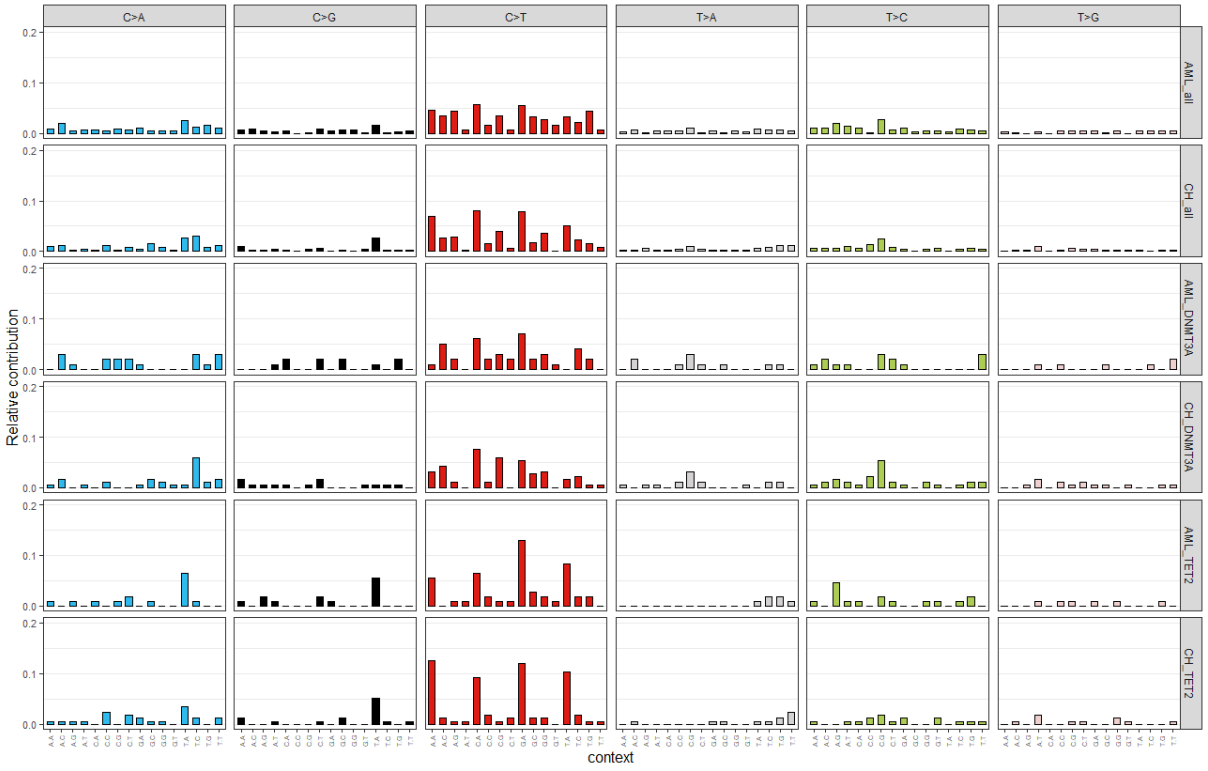


Figure 12: ***Mutational signatures of AML-related and CH-related mutation profiles.***

The first and second panels are mutational signatures of somatic mutations on all commonly mutated genes in the AML and CH samples. The third and fourth panels are mutational signatures of DNMT3A mutations in AML and CH conditions. The fifth and sixth panels are TET2 mutational signatures in AML and CH conditions.

2.2.2 Deconvolution analysis of signature profiles

To extract the origin of the mutations, mathematical decomposition techniques were employed, including non-negative matrix factorization (NMF). Based on the results (Figure 13), we can see that all the mutation profiles have a similarity with SBS30 which is due to a deficiency in base excision repair due to inactivating mutations in NTHL1. This similarity is because of the higher frequency in C>T substitutions. On the other hand, we also see differences between AML and CH conditions in their relative contribution to the references. Specifically, in DNMT3A we have an enormous difference in their similarity pattern. However, we need to grab top similar signatures in each profile to be able to compare profiles.

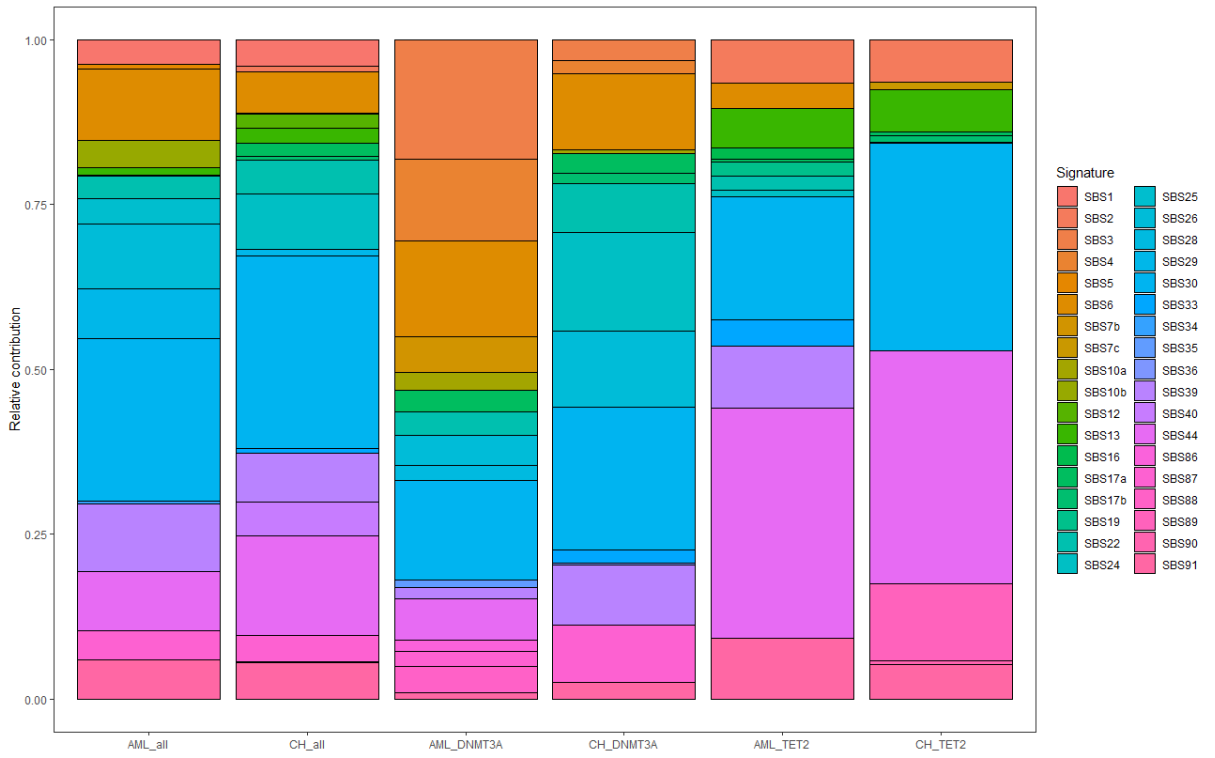


Figure 13: ***Contribution plot of each mutational profile with the known COSMIC mutational signatures.***

This figure represents the percentage of the similarity of the mutational profile with the reference signature within each subset of mutations.

We selected the top five signatures based on their similarity score for each mutation profile (Table 1). Comparing profiles of all genes in two conditions, SBS30 has the highest similarity in both conditions. In the second rank, the AML profile has a similarity with SBS6, and the CH profile has a similarity with SBS44. SBS6 is associated with defective DNA mismatch repair and is found in tumors with microsatellite instability[80]. Among the seven mutational signatures associated with defective DNA mismatch repair and microsatellite instability (MSI), SBS6 is often found in the same samples as SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44. According to their similarity with signatures, we can assume that CH and AML mutations share the same origin.

DNMT3A profiles showed a more significant difference In particular, SBS30 is in the top three ranks of these profiles. Ignoring SBS30 due to its presence in all profiles, we have a high similarity of DNMT3A profile in AML condition with SBS3. The SBS3 signature represents

defective homologous recombination-based DNA damage repair which manifests primarily as small indels and genome rearrangements because of abnormal double-strand break repair but can also take the form of base substitutions[82]. SBS3 has been suggested as a predictor of homologous recombination-based repair failure and response to therapies exploiting this defect. On the other hand, the DNMT3A profile in the CH condition shows a similarity with SBS24. SBS24 represents exposure to aflatoxin. As a result of aflatoxin exposure, SBS24 has been observed in cancer samples with mutation patterns that are consistent with those that have been observed in experimental systems exposed to aflatoxin, and these mutation patterns correlate with those observed in cancer samples with known aflatoxin exposure.

TET2 profiles in both conditions are similar except for one signature on the third rank. TET2 profile in AML condition shows a similarity with SBS39 and in CH condition a similarity with SBS89. Both SBS39 and SBS89 are annotated as unknown signatures on the COSMIC website.

Table 1: *Top five COSMIC mutational signatures reconstructed in the mutation profiles.*

AML_ALL	CH_ALL	AML_DNMT3A	CH_DNMT3A	AML_TET2	CH_TET2
SBS30	SBS30	SBS3	SBS30	SBS44	SBS44
SBS6	SBS44	SBS30	SBS24	SBS30	SBS30
SBS39	SBS24	SBS6	SBS26	SBS39	SBS89
SBS26	SBS39	SBS4	SBS6	SBS91	SBS2
SBS44	SBS6	SBS44	SBS39	SBS2	SBS13

Note. By solving the nonnegative least-squares constraints problem, mutation signatures are deconvolved, and the mutation matrix is reconstructed.

2.3 Model Training

We have chosen the top two most mutated genes within both conditions, DNMT3A and TET2 to train the predictive model based on their mutations. These two genes have been identified through our research as having the highest mutation rates across both conditions, meaning that they are most likely to be affected by certain treatments. We will use these two genes to create a predictive algorithm that accurately identifies the state of mutations for available datasets as well as novel pop-up mutations on these genes. We believe that the predictive model will help

us gain a better understanding of the gene regions that are more prone to a mutation in CH and AML.

We have identified sets of mutations for each gene that comes from CH and AML samples and a subset of mutations that are shared and commonly occur in both conditions, we have three subsets of mutations. Intending to have a broader inspection and training experiment, we approached the training with two assumptions. The first assumption is based on the actual subsets of mutations for which we have three labels: the mutations that only happened in CH samples as *CH_only*, the mutations that were observed only in AML samples as *AML_only*, and the mutations that are shared between two conditions as shared. Another assumption is that considering shared subset and *AML_only* as one subset because the shared mutations will develop into an AML situation one day. However, in this assumption, we have two classes: the mutations that only happened in CH labeled as *CH_only* and the rest of the mutations labeled as *shared+AML_only*.

Due to the controversy on the Amino Acid (AA) position feature which could be considered categorical or numeric, we trained models with both numeric and categorical conditions. AA position features have a categorical spirit that represents the specific location of the mutation occurrence. To give the model more flexibility on prediction, we can also consider this AA position feature as numeric, and it could represent the distance between the mutations and the regional impact of the mutations on the protein sequence. Therefore, we ended up with four distinctive model training models for DNMT3A as well as the TET2 mutations dataset.

2.3.1 Feature importance analysis using ROC/AUC

To compare the feature importance of the mutations, we perform an analysis of the ROC metric of models trained on each feature. ROC refers to receiver operating characteristics, representing the performance of classification problems[83]. In a default classification problem, the threshold is 0.5; every data point classified with a probability higher than 0.5 will be classified into one class, while a data point with a lower probability than 0.5 will be classified into the other class[83]. In this method, the operator calculates the classification performance at different

thresholds based on the true positive rate and false positive rate and creates a curve plot for each model. The true positive rate, also known as sensitivity, is the probability of positive test results being correctly labeled as positive[83]. On the other hand, the false positive rate is defined as 1-specificity, which is the probability of incorrectly classifying negative data as positive. Using the ROC curve and the area under the curve (AUC), we can estimate the contribution of each feature set to the classifier and its performance in accurately identifying the correct class of data based on that feature.

As a result of the DNMT3A dataset, we observed that the AA position is of the highest importance for the model, with 0.97 AUC (Figure 14). Other numeric features such as molecular weight differences and Hydrophobicity differences also have an AUC value higher than 0.5. Although most of the categorical features have a low AUC value which shows that the model cannot rely only on these features, a combination of all features shows the highest AUC value even more than the AA position. However, our model can be trained using these features.

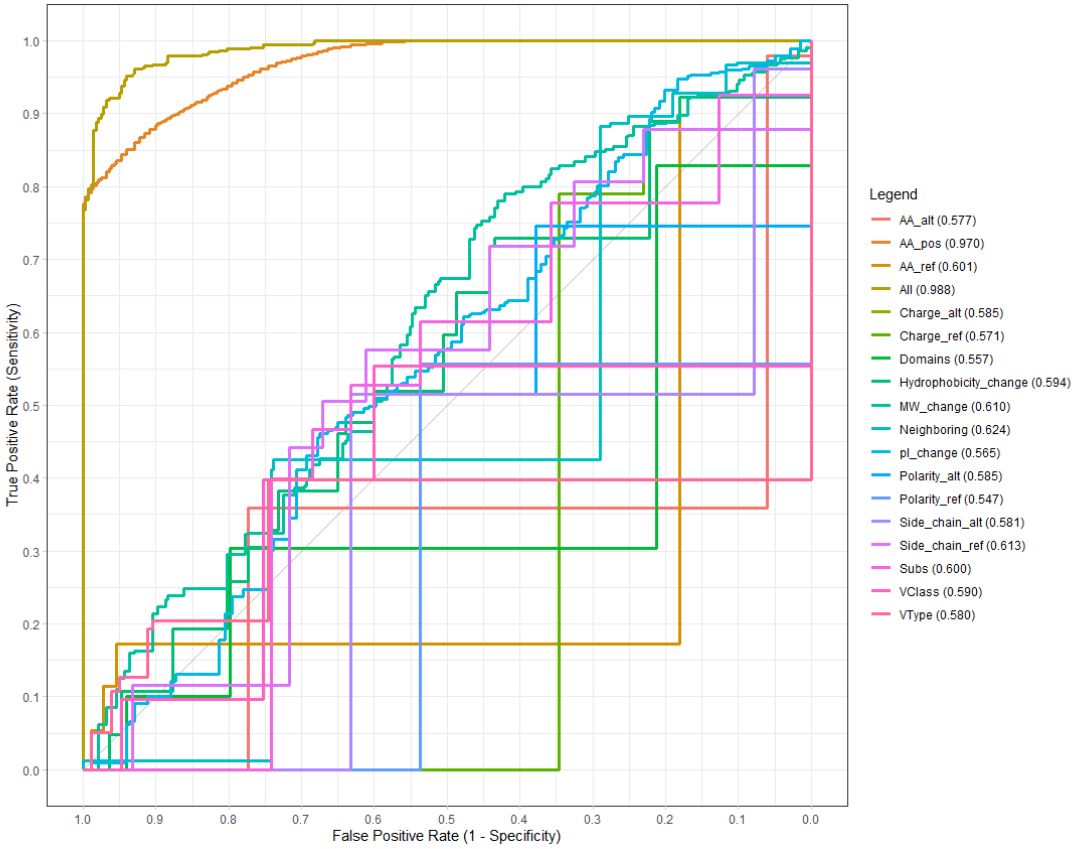


Figure 14: ***ROC plot of each feature of DNMT3A mutations.***

This plot shows the false positive and false negative rates of a simple model trained on each feature to show the contribution of the feature to the training process.

The ROC results of TET2 mutations' features also show the same pattern as DNMT3A features (Figure 15). AA position feature has the highest AUC value (0.99) among all other features. Categorical features have higher importance for the model compared to the DNMT3A mutations' features. The combination of all the features together gives a significant ability to model in training as it is obvious with the AUC value of all features (0.99).

Therefore, the AA position feature showed the best contribution to the model training among all other features of the mutations while a combination of all features showed a better performance in the model training.

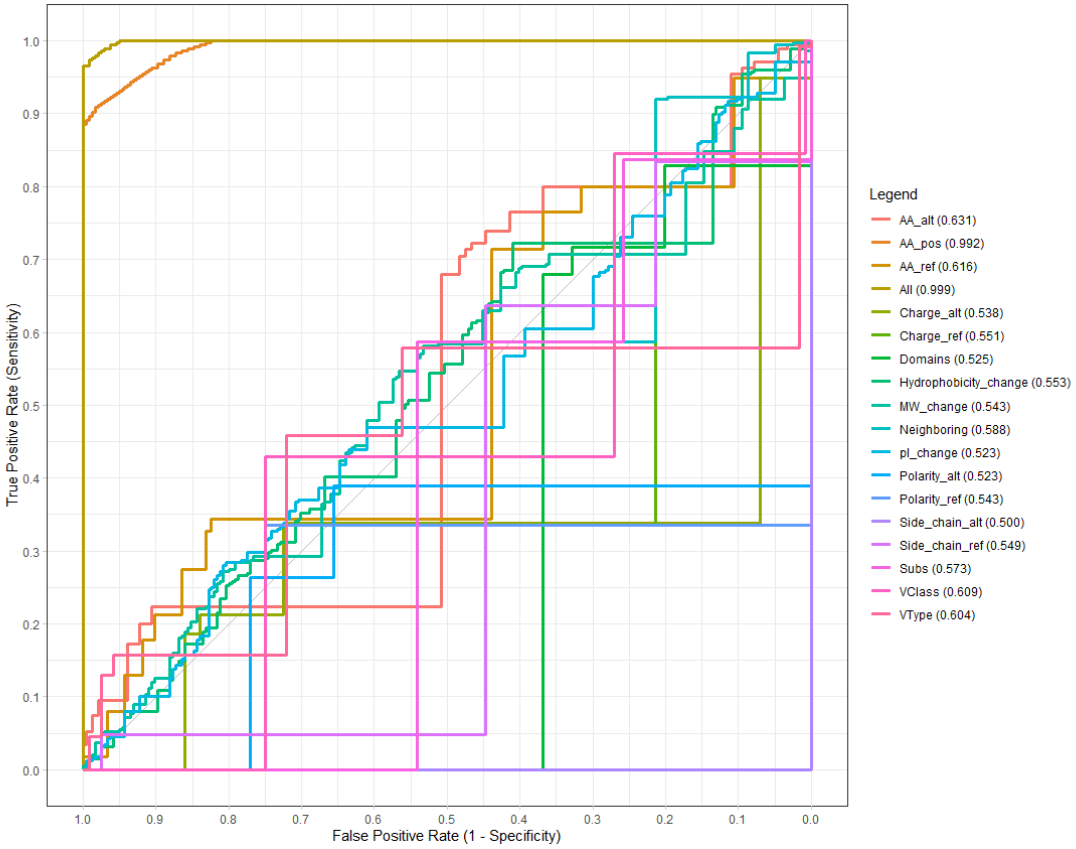


Figure 15: ***ROC plot of each feature of TET2 mutations.***

This plot shows the false positive and false negative rates of a simple model trained on each feature to show the contribution of the feature to the training process.

2.3.2 DNMT3A models result

After training the models with both assumptions, we have received metrics values, Accuracy, Kappa, and their standard deviation directly from the model which are the average of cross-validation folds (Figure 16). We can see that the best model among all the trained models is the *GLM_OH*, which is the model trained using the Generalized Linear Model (GLM) algorithm on the dataset with a categorical AA position. With an accuracy of 0.7, the model correctly predicted 70% of the instances in the dataset along with the cross-validation folds. As a result, the model has a Kappa value of 0.41, which indicates moderate agreement between predicted and actual classes. According to the standard deviation of the accuracy scores, the accuracy results are consistent across different runs or folds.

Additionally, Kappa values have a standard deviation of 0.05, indicating that Kappa results are stable and consistent across different runs. We evaluated the model based on Kappa value as well as accuracy considering the imbalance dataset. Kappa value has more complex calculations of agreement which calculate the chance of performance of each agreement in the minority class as well as the majority class[84].

In the assumption by considering AA position as a numeric feature to capture the distances, we can see that the RF model can perform better. As a result of their capability to handle non-linearity, provide feature importance measures, robustness to outliers, and natural handling of categorical features without explicit encoding, Random Forest models can perform better with numeric features than categorical features in comparison to GLM models. The GLM model assumes linearity and may not be able to capture complex relationships effectively. The models' performances are the same with two assumptions of two classes and three classes showing the flexibility of the training process with binary and non-binary classification.

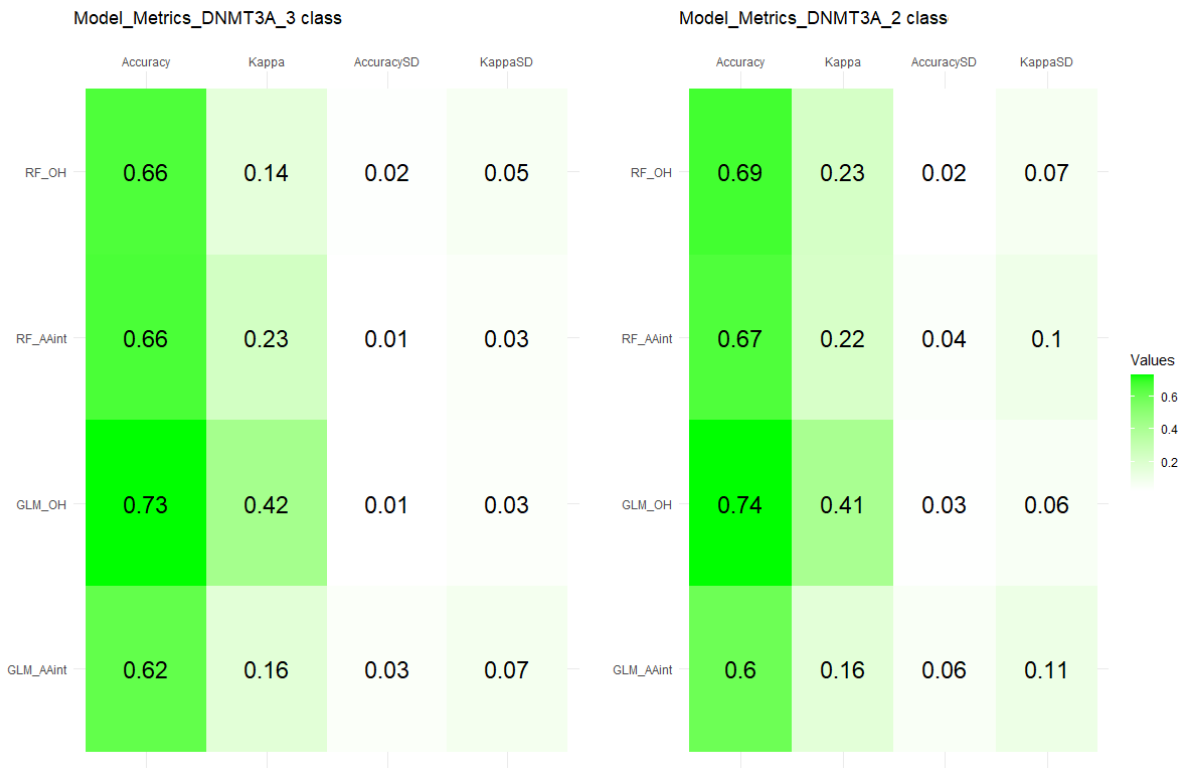


Figure 16: ***Model metrics output of training on DNMT3A mutations.***

We also fitted the models with an unseen test set to evaluate the prediction performance of the model. Based on the results of the *two-class* assumption in Table 2, with an accuracy of 0.77, the model correctly predicted the class labels for approximately 77% of the instances in the test set. It is important to remember that accuracy alone may not provide a complete picture of model performance, especially if the classes are imbalanced. A kappa coefficient of 0.49 indicates a prominent level of agreement between your model's predictions and the actual class labels, taking into consideration the likelihood of such an agreement occurring by chance. There is minor agreement beyond chance when the kappa value is less than 0.5. A sensitivity score of 0.64 indicates the proportion of positive instances that the model correctly identified. Positive instances are more likely to be detected when the sensitivity value is higher. The model correctly identified 84 percent of actual negative instances with a specificity score of 0.84 which is assigned to the larger class *(CH_only)*. The higher the specificity value, the better the ability to identify negative incidents. It measures the proportion of predicted positive instances that were positive. The precision value of 0.69 measures the accuracy of the prediction. By labeling instances as positive, it indicates the reliability of the model.

According to the negative predictive value of 0.81, the proportion of predicted negative instances that occurred is 0.81. This indicates that the model is capable of correctly classifying instances as negative. With an F1-score of 0.67, precision and recall are combined into one metric. Providing a balance between the two is useful when class distribution is uneven. In the test set, the prevalence of 0.35 represents the proportion of positive examples. This is the actual occurrence of the positive class within the dataset. A detection rate of 0.23 indicates the proportion of positive instances correctly identified by the model. The model predicts that 0.33 instances will be positive, indicating a detection prevalence of 0.33. The balanced accuracy of 0.74 is the average of the sensitivity and specificity. By considering both true positives and true negatives, it provides a balanced picture of the model's performance.

Table 2: *Metric values of the model's performance on the test set in two-class assumptions on DNMT3A mutations.*

Model	Accuracy	Kappa	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence
RF_OH	0.70	0.25	0.32	0.90	0.64	0.71	0.64	0.32	0.43	0.35
RF_AA	0.66	0.19	0.36	0.82	0.53	0.70	0.53	0.36	0.43	0.35
GLM_OH	0.77	0.49	0.64	0.84	0.69	0.81	0.69	0.64	0.67	0.35
GLM_AA	0.58	0.16	0.61	0.57	0.44	0.73	0.44	0.61	0.51	0.35

This analysis reveals that the model performs well in identifying instances labeled as *CH* (true positives) and instances labeled as *AML_only+shared* (true negatives). Despite this, it is unable to distinguish between *CH* and *AML_only+shared*, resulting in misclassifications in both directions (false positives and false negatives). In the confusion matrices of two class assumptions, we can observe the actual number of data truly labeled as their actual reference labels (Figure 17).

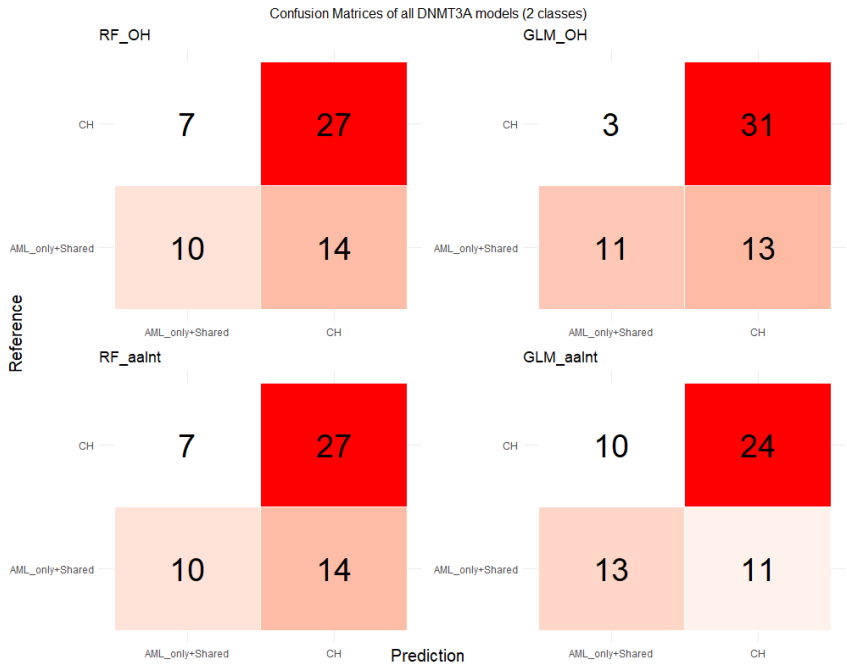


Figure 17: ***Confusion matrices of all DNMT3A models fitted with test data based on two classes assumption.***

On the other hand, in the three classes assumption, we can also see higher performance in the *GLM_OH* model (Table 3). An accuracy of 0.74 indicates that approximately 74% of the

instances were correctly classified by the multi-class classification model. While the kappa coefficient is low at 0.37, it indicates that there is a significant disagreement that is beyond chance. The model has difficulty identifying instances that are labeled as *AML_only*, resulting in a notable discrepancy (Table 4). It is evident from the metrics that this issue exists, such as the absence of true positives (sensitivity of 0.00), the absence of instances predicted as *AML_only* (positive predictive value N/A), and the low detection rate of 0.00. These results suggest a potential mismatch between the model and the data, indicating that further investigation into the characteristics of the *AML_only* class is necessary, as well as improvements to data quality, feature selection, or model architecture to resolve this issue. Multi-class classification models would be more balanced and robust if they were able to accurately identify *AML_only* instances.

Table 3: *Overall values of the evaluation metrics of the models' performance of three-classes assumption on DNMT3A mutations.*

Model	**Accuracy**	**Kappa**
GLM_OH	0.74	0.37
GLM_AAint	0.66	0.10
RF_OH	0.69	0.21
RF_AAint	0.71	0.33

Table 4: *Metric values of the model's performance on the test set in three-class assumption on DNMT3A mutations*

Model	Class	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
RF_OH	AML_only	0.00	1.00	NA	0.92	NA	0.00	NA
	CH_only	0.94	0.22	0.70	0.67	0.70	0.94	0.80
	shared	0.29	0.95	0.67	0.78	0.67	0.29	0.40
GLM_OH	AML_only	0.00	1.00	NA	0.92	NA	0.00	NA
	CH_only	0.94	0.37	0.74	0.77	0.74	0.94	0.83
	shared	0.48	0.95	0.77	0.83	0.77	0.48	0.59
RF_AAint	AML_only	0.00	1.00	NA	0.92	NA	0.00	NA
	CH_only	0.90	0.41	0.74	0.69	0.74	0.90	0.81
	shared	0.48	0.89	0.63	0.82	0.63	0.48	0.54
GLM_AAint	AML_only	0.00	1.00	NA	0.92	NA	0.00	NA
	CH_only	0.96	0.11	0.67	0.60	0.67	0.96	0.79
	shared	0.14	0.96	0.60	0.75	0.60	0.14	0.23

The confusion matrices of different models in three classes assumption showed the model performs perfectly when predicting the *CH_only* classes in the configuration (Figure 18). However, models struggle a bit when these configurations are low data space. This could suggest that the three-class configuration is more prone to misclassification and confusion for this particular dataset.

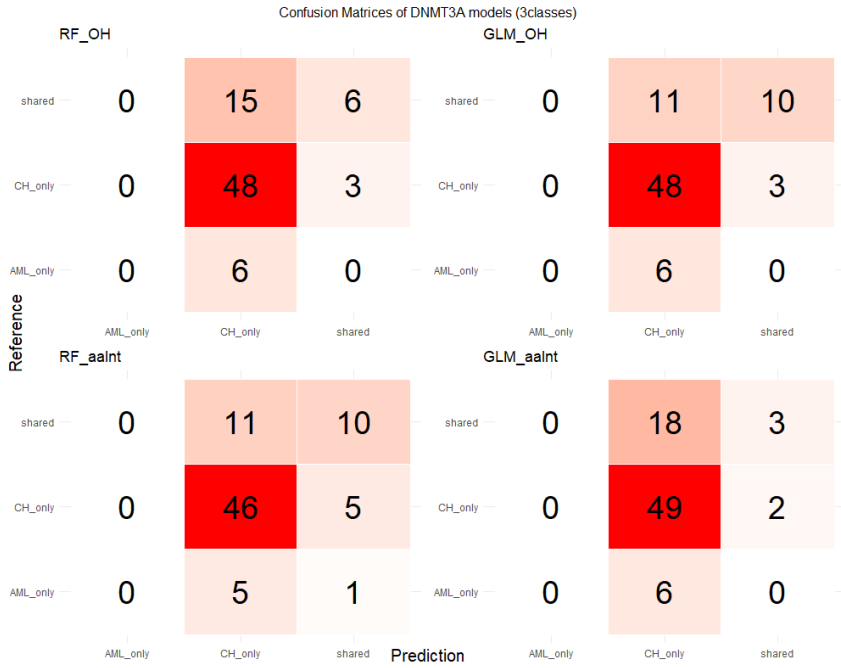


Figure 18: ***Confusion matrices of all DNMT3A models fitted with test data based on three classes assumption.***

2.3.3 TET2 models result

The models with the same setting as models that trained on DNMT3A were performed on the TET2 mutations dataset. The results of model training indicate that the TET2 model performance follows the same trend as the model performance on the DNMT3A dataset.(Figure 19). We can see that the *GLM_OH* model has 0.65 accuracy which is not a significant performance. Also, the kappa value is 0.3 which is not a significant value to deal with the different agreement of the data and this is the result of prediction by chance. In both assumptions, the results are the same and it shows that something related to the feature sets would be the problem of these predictions.

Figure 19: ***Model metrics output of training on TET2 mutations.***

After fitting the unseen test set to the models, we observe a better performance in the *GLM_OH* model (Table 5). As indicated by the accuracy of 0.72, the model correctly predicted the class labels for approximately 72% of the instances in the test set. A kappa coefficient of 0.39 indicates a reasonable level of agreement beyond chance. According to the model's sensitivity (recall) score of 0.46, 46% of the instances belonging to the *CH_only* class were correctly identified. A specificity score of 0.91 indicates that the model is capable of correctly identifying instances outside the *CH_only* class. The positive predictive value (precision) of 0.79 indicates that 79% of the time, the model is correct when predicting that an instance is *CH_only*. It is estimated that 70 percent of the time, the model is correct when predicting an instance to be *AML_only+shared*. F1 scores of 0.58 combine precision and recall into a single metric, providing a balance between the two. In terms of identifying *CH_only* instances correctly, it indicates a moderate performance. The prevalence of 0.41 indicates the proportion of instances in the test set that belong to the *CH_only* class. According to the model, 0.19 percent of *CH_only* instances are correctly identified.

Table 5: *Metric values of the model's performance on the test set in two-class assumptions on TET2 mutations.*

Model	Accuracy	Kappa	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
RF_OH	0.64	0.22	0.42	0.79	0.59	0.66	0.59	0.42	0.49
RF_AA	0.64	0.22	0.42	0.79	0.59	0.66	0.59	0.42	0.49
GLM_OH	0.72	0.39	0.46	0.91	0.79	0.70	0.79	0.46	0.58
GLM_AA	0.64	0.25	0.54	0.71	0.57	0.69	0.57	0.54	0.55

The confusion matrices of the fitted model on the test set showed the actual number of data points truly and falsely predicted by the models in two class assumptions (Figure 20). According to confusion matrices, the *GLM_OH* model showed a better performance also in the TET2 gene mutations dataset.

Figure 20: ***Confusion matrices of all TET2 models fitted with test data based on two classes assumption.***

Although the *GLM_OH* model has an accuracy of 0.64 and kappa of 0.30, a certain prediction is not guaranteed for the model in the three-classes assumption (Table 6). The corresponding confusion matrices showed the same problem with the three-classes assumption as DNMT3A models (Figure 21). By creating smaller classes the model will be more confused and the three-classes assumption is not a proper approach to increase the quality of model training.

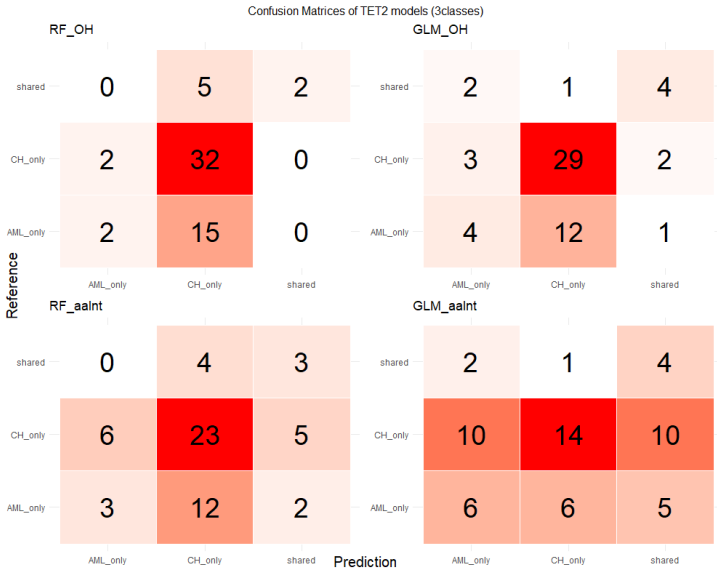


Figure 21: ***Confusion matrices of all TET2 models fitted with test data based on three classes assumption.***

With more detail on the *GLM_OH* model performance, in Table 7 we have more information on various metrics specified in each class. Among the instances belonging to the *AML_only* class, the model correctly identified only 24% of them. According to the specificity score of 0.88, instances in classes other than *AML_only* can be correctly identified with a high degree of accuracy. A positive predictive value (precision) of 0.44 indicates that the model is accurate approximately 44% of the time when predicting an instance to be *AML_only*. An F1-score of 0.31 indicates a moderate balance between precision and recall for the *AML_only* class.

85% of the instances belonging to the *CH_only* class were correctly identified by the model, based on the sensitivity score of 0.85. In addition, the specificity score of 0.46 indicates that the model is moderately capable of identifying instances in classes other than *CH_only*. A positive predictive value (precision) of 0.69 indicates that when the model predicts an instance to be *CH_only*, it is correct approximately 69% of the time. An F1-score of 0.76 indicates a good balance between precision and recall for the *CH_only* class.

According to the sensitivity score of 0.57, 57% of the instances belonging to the shared class were correctly identified by the model. The specificity score of 0.94 suggests that instances in

non-shared classes can be correctly identified. A positive predictive value (precision) of 0.57 indicates that the model is correct approximately 57% of the time when it predicts that an instance will be shared. F1-score of 0.57 indicates a moderate balance between precision and recall.

Table 6: *Overall values of the evaluation metrics of the models' performance of three-classes assumption on TET2 mutations.*

Model	**Accuracy**	**Kappa**
GLM_OH	0.64	0.30
GLM_AA	0.41	0.11
RF_OH	0.62	0.16
RF_AA	0.50	0.07

Table 7: *Metric values of the model's performance on the test set in three-class assumption on TET2 mutations*

Model	**Class**	**Sensitivity**	**Specificity**	**Pos Pred Value**	**Neg Pred Value**	**Precision**	**Recall**	**F1**	**Prevalence**
RF_OH	AML_only	0.12	0.95	0.50	0.72	0.50	0.12	0.19	0.29
	CH_only	0.94	0.17	0.62	0.67	0.62	0.94	0.74	0.59
	shared	0.29	1.00	1.00	0.91	1.00	0.29	0.44	0.12
GLM_OH	AML_only	0.24	0.88	0.44	0.73	0.44	0.24	0.31	0.29
	CH_only	0.85	0.46	0.69	0.69	0.69	0.85	0.76	0.59
	shared	0.57	0.94	0.57	0.94	0.57	0.57	0.57	0.12
RF_AAint	AML_only	0.18	0.85	0.33	0.71	0.33	0.18	0.23	0.29
	CH_only	0.68	0.33	0.59	0.42	0.59	0.68	0.63	0.59
	shared	0.43	0.86	0.30	0.92	0.30	0.43	0.35	0.12
GLM_AAint	AML_only	0.35	0.71	0.33	0.73	0.33	0.35	0.34	0.29
	CH_only	0.41	0.71	0.67	0.46	0.67	0.41	0.51	0.59
	shared	0.57	0.71	0.21	0.92	0.21	0.57	0.31	0.12

3 DISCUSSION AND FUTURE PERSPECTIVES

The analysis of the somatic mutation landscape in clonal hematopoiesis (CH) and acute myeloid leukemia (AML) samples yields critical insights into the mutational spectrum and their regional impact, revealing significant underpinnings of these hematologic states and their implications for early disease progression. This study demonstrates distinct mutational profiles between CH and AML, underscoring the pivotal roles of specific genes and their mutations in the pathogenesis and progression of these hematologic conditions.

Key findings indicate a higher frequency of DNMT3A, TET2, and ASXL1 mutations in AML samples compared to CH samples. These genes are integral to epigenetic regulation and hematopoietic differentiation, with mutations in these loci frequently associated with leukemic transformation and disease progression. The clonal dominance of these mutations in AML suggests their involvement in the proliferative advantage of leukemic cells, corroborating previous studies that highlight the adverse clinical outcomes associated with DNMT3A, TET2, and ASXL1 mutations in AML[78].

Contrastingly, CH samples exhibit a broader diversity of mutations within DNMT3A, TET2, and ASXL1, which are present in a subset of hematopoietic cells without clonal dominance. This observation suggests that these mutations alone may be insufficient for leukemogenesis, requiring additional genetic alterations to drive AML development. The presence of these mutations in CH implies a selective advantage conferred upon affected cells, promoting clonal expansion and elevating the risk of AML.

Specifically, mutations in the ADD domain of DNMT3A, which is crucial for recruiting DNMT3A to specific genomic regions for DNA methylation[85], may disrupt protein-protein interactions and regulatory complex formation, leading to altered gene expression and cellular dysregulation. Similarly, ASXL1 mutations targeting the ASXH domain, a region critical for protein-protein interactions and epigenetic regulation[86], further substantiate the role of these mutations in clonal hematopoiesis.

While the mutation profiles of TET2 in CH and AML exhibit similar regional impacts and distributions, shared mutations between these states are infrequent. TET2 mutations either arise during CH and later diminish or emerge concurrently with IDH1/2 mutations during AML progression[87]. These findings emphasize the complex interplay of genetic alterations in the transition from CH to AML.

TET2 mutations are uniquely prevalent in AML clusters, with minimal overlap with CH samples. Within NPM1 clusters, a significant portion of TET2 mutations are shared-like, suggesting they may originate from the CH state of clones. This finding indicates that some TET2 mutations present in AML patients might have originated during an earlier, pre-leukemic phase, highlighting the role of the CH state in AML development.

These insights have led our team to prioritize deep sequencing of TET2 genes in AML patients to better understand their role in early clone development. By investigating these mutations at a granular level, we aim to elucidate the mechanisms driving the transition from CH to AML. This research could inform the development of targeted therapies to intercept the progression of pre-leukemic clones, ultimately improving patient outcomes and contributing to the broader field of hematologic malignancies.

Comparative analysis reveals that AML samples possess a broader mutation spectrum across various genes than CH samples, with a higher prevalence of mutations in DNMT3A, TET2, and ASXL1. This suggests that AML development is influenced by a more extensive array of genetic mutations, potentially affecting critical signaling pathways, DNA repair mechanisms, and other cellular processes involved in leukemogenesis. The identification of shared mutations between CH and AML samples indicates clonal expansion and preservation of specific mutations during disease progression, highlighting the significant impact of DNMT3A mutations from CH to AML and their potential as therapeutic targets. This supports the hypothesis that NPM1 mutations drive the evolution of DNMT3A-mutant CH to AML, accelerating disease progression through extended CH latency and clonal expansion[88].

Our analysis of mutation signatures in AML and CH revealed a predominant C>T substitution

pattern, likely influenced by CpG island mutations in chromatin regulatory genes. DNMT3A and TET2 mutation signatures were largely conserved between CH and AML, suggesting these mutations persist through disease progression. DNMT3A showed minor variations in C>A and T>C substitutions, while TET2 displayed a stable pattern with a slightly higher frequency in AML. These findings imply a continuity of mutation origin from CH to AML and underscore the potential of these mutations as early biomarkers for disease progression and targeted therapy development in AML.

Furthermore, the application of machine learning techniques to classify CH-related mutations in DNMT3A and TET2 illustrates the potential for enhancing mutation classification and disease prediction. The machine learning models, trained on known mutation patterns and their disease associations, demonstrate that mutational impact features are more influential than other features in the classification process. The preservation of specific mutations from CH to AML suggests their utility as predictive markers for disease progression, necessitating further investigation and refinement of the models to improve predictive accuracy and reliability.

Nevertheless, several limitations must be acknowledged. The study's focus on a subset of genes warrants further research to explore the mutational landscapes of additional genes and their contributions to CH and AML. Validation of the machine learning methods in larger cohorts is essential to assess their accuracy and generalizability. The imbalance between CH and AML samples could introduce biases, impacting the robustness of the findings due to variations in sample size and mutation frequencies. Addressing these challenges through comprehensive studies and balanced datasets will enhance our understanding of the molecular mechanisms underlying CH and AML, thereby improving risk prediction, prognostic assessment, and the development of targeted therapies.

In conclusion, this study provides valuable insights into the somatic mutation dynamics of DNMT3A, TET2, and ASXL1 in CH and AML, highlighting their implications for disease progression and therapeutic targeting. Continued exploration of imbalanced data analysis and refinement of machine learning approaches will further optimize diagnostic and prognostic studies, ultimately applied in personalized treatment strategies for hematologic malignancies.

4 CONCLUSION

This study comprehensively examined the somatic mutation landscape in clonal hematopoiesis (CH) and acute myeloid leukemia (AML) samples, providing novel insights into the mutational impact on the DNMT3A, TET2, and ASXL1 genes. The analysis revealed a higher prevalence and diversity of DNMT3A, TET2, and ASXL1 mutations in AML samples compared to CH samples, indicating a significant mutational burden in AML. This suggests that additional genetic alterations beyond DNMT3A, TET2, and ASXL1 contribute to AML development and progression.

Notably, certain mutations were found to persist from CH to AML, underscoring their potential role in disease progression. DNMT3A mutations were particularly notable for their preservation from CH to AML, suggesting their critical impact on leukemic transformation. Similarly, the conservation of TET2 mutations from CH to AML highlights their relevance throughout disease development.

The study also identified distinct differences in mutation profiles and impacted regions between CH and AML. In CH samples, DNMT3A mutations exhibited a broad distribution, whereas AML samples showed a pronounced hotspot in the MTase domain. ASXL1 mutations were widely distributed in CH but showed limited overlap with AML mutations. Both CH and AML exhibited similar TET2 mutation patterns with minor differences in specific mutation classes. Moreover, the analysis of AML subtypes revealed that DNMT3A mutations were shared between AML and CH, while TET2 mutations were more specific to AML. These findings emphasize the pivotal role of DNMT3A mutations in disease progression and suggest a differential role for TET2 mutations in CH and AML.

Analysis of mutational signatures reveals distinct patterns in CH and AML, with a predominant C>T substitution pattern influenced by CpG island mutations in chromatin regulatory genes. The conserved mutation signatures of DNMT3A and TET2 across disease stages suggest their potential as early biomarkers and therapeutic targets in AML.

The machine learning approach employed in this study demonstrated the potential for classifying mutations and predicting disease states. Models trained on DNMT3A and TET2 mutations exhibited varying performance, with the Random Forest model outperforming others when utilizing numeric features. However, further refinement, including feature selection and model optimization, is necessary to enhance predictive power. Future research should focus on validating these models and exploring additional features to improve accuracy. Testing these models on larger datasets is also crucial for obtaining more reliable results.

In conclusion, this study provides valuable insights into the mutational landscape of CH and AML and their implications for disease progression and prognosis. The findings suggest that further research and validation in larger cohorts are essential to refine risk prediction models and identify potential therapeutic targets. Future studies should encompass a broader range of genes and mutations to elucidate additional genetic changes and their interactions. This will enhance our understanding of disease progression and AML development. The identified mutations and their associated patterns may also inform the development of diagnostic tools for CH and AML. Integrating these findings with other clinical and molecular features, such as gene expression profiles and epigenetic modifications, could lead to the development of comprehensive risk stratification and treatment selection models, ultimately improving patient outcomes in hematologic malignancies.

5 MATERIALS AND METHODS

5.1 Data Collection

We collected variant data for CH state from comprehensive studies with 2426 samples in total[43, 58, 70]. CH studies included samples from their in-house patients and public resources such as the UK Biobank[89]. AML mutations were collected from public cohorts and mRNA data resources, TCGA[71](n = 151), TARGET[72](n = 187), BEAT[73](n = 450), Leucegene[74](n = 449), LUMC[75](n =100), and from Papaemmanuil et al.[90] (n= 1540). A dataset of somatic mutations discovered in the studies mentioned above was generated for our investigation.

5.1.1 UK BIOBANK

The UK Biobank is a biomedical database and research resource established to facilitate the study of a wide array of diseases and health conditions. Launched in 2006, it has amassed a wealth of genetic and health-related data from approximately 500,000 participants across the United Kingdom, making it one of the most comprehensive biobanks in the world. This substantial and meticulously curated resource is invaluable for researchers to enhance our understanding of the genetic, environmental, and lifestyle factors contributing to human health and disease. Enriched resources of WGS and WES resources from a large-scale participant provided the potential for longevity analysis and age-related diseases.

5.1.2 TCGA

The Cancer Genome Atlas (TCGA) is a groundbreaking initiative in cancer genomics, providing a comprehensive and openly accessible repository of molecular and clinical data for various types of cancer. Launched in 2006 through a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), TCGA has significantly advanced our understanding of cancer's genetic and molecular underpinnings. By

meticulously documenting genetic mutations, changes in gene expression, and epigenetic modifications in over 11,000 cancer patients across more than 30 different cancer types, TCGA has become a crucial resource for cancer research and personalized medicine.

5.1.3 BEAT AML

The Beat AML project is an innovative endeavor in the field of hematologic oncology, particularly focusing on AML. Launched by the Leukemia and Lymphoma Society (LLS) in 2016, the Beat AML initiative aims to transform the treatment approach to AML through a comprehensive precision medicine strategy. By utilizing advanced genomic and molecular profiling technologies, Beat AML aims to unravel the complex molecular foundations of the disease, identify actionable genetic changes, and support the development of targeted treatments. The primary objective is to enhance patient outcomes by providing personalized treatment plans that are more effective and less harmful than traditional therapies.

5.1.4 LEUCEGENE

The Leucegene initiative encompasses diverse cohorts focused on comprehensive genetic and molecular characterization within AML research. These cohorts consist of newly diagnosed patients, individuals with relapsed/refractory disease, and groups categorized by treatment response. The initiative aims to discover biomarkers that can predict prognosis, treatment response, and relapse risk while identifying new therapeutic targets through in-depth genomic and transcriptomic analyses. Leucegene utilizes extensive resources such as high-resolution genomic sequencing, comprehensive transcriptomic profiles, and integrated clinical data, supported by advanced bioinformatics tools. This collaborative framework encourages interdisciplinary research efforts aimed at translating genomic insights into personalized treatment strategies, ultimately improving outcomes for AML patients.

5.1.5 TARGET

The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative is a crucial undertaking led by the National Cancer Institute (NCI) to improve our knowledge and management of pediatric cancers using extensive data sources. Since its inception in 2006, TARGET has aimed to utilize cutting-edge genomic and transcriptomic technologies to decipher the intricate molecular characteristics of different pediatric cancers, such as AML, neuroblastoma, and osteosarcoma, among others.

5.2 Harmonizing and Filtration

To harmonize the mutation datasets, we performed reannotation and conversion of the genome coordination reference of somatic mutations from different studies. We applied the *"Lift Genome Annotation"* tool of the UCSC Genome Browser[91]. In addition, we used the *Variant Effect Predictor (VEP)*[92] to adjust somatic mutation annotations from different studies. We used a 5% VAF threshold to filter low-impact mutations. To analyze unique mutations within the genes, we compared the mutation dataset and removed any duplicate mutations from the list.

5.2.1 File Formats

The mutations from samples were collected in the Variant Call Format (VCF). The VCF is a standardized file format used in bioinformatics to store genetic variants detected in genomic sequencing data. The VCF file structure is designed to comprehensively capture and describe genetic variants identified in genomic sequencing data (Figure 22). It begins with header lines that provide essential metadata, including information about the reference genome, software version, and definitions of fields used within the file. Following the headers, each row corresponds to a specific variant detected in the genome, with columns detailing its genomic location (‘CHROM‘ and ‘POS‘), identifier (‘ID‘), reference allele (‘REF‘), alternate alleles (‘ALT‘), variant quality (‘QUAL‘), and filter status (‘FILTER‘). Additional information about the variant is stored in the ‘INFO‘ field, encompassing diverse annotations like allele frequency and

functional impact. The ‘FORMAT‘ field defines the structure of genotype information for each sample, including genotype calls (‘GT‘), genotype quality (‘GQ‘), and allelic depths (‘AD‘). This structured tabular format enables efficient storage, sharing, and analysis of genetic variation data across populations and studies, facilitating insights into the genetic basis of diseases and traits in genomic research and clinical applications.

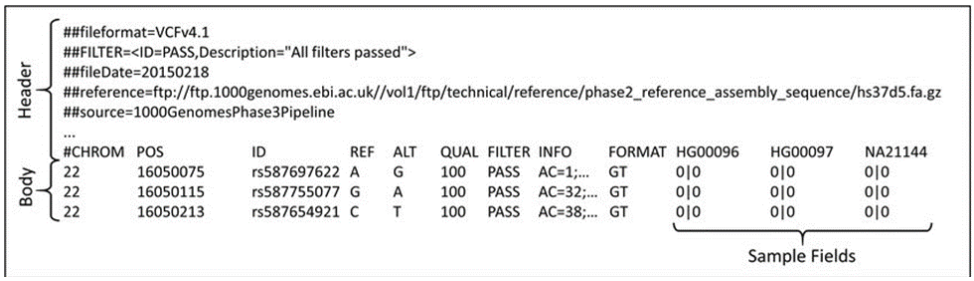


Figure 22: ***An overview of VCF file structure.***

This file includes a header line describing the overall information of the sample data. The body part includes the list of mutations observed in the sample with their specific location on the chromosomes and features in the quality of the mutation calling process. [93]

Further steps regarding the comparative analysis and training Machine Learning classifier have been done using the Mutation Annotation Format (MAF). MAF is a specialized file format used extensively in cancer genomics to catalog and describe somatic mutations detected in tumor samples. Unlike the Variant Call Format (VCF), which is versatile and used broadly across genetics research, MAF files focus specifically on somatic mutations in cancer genomes. We converted VCF files to MAF files for integration and harmonization using the *Funcotator* (*FUNCtional annOTATOR*)[92].

The structure of the MAF file typically includes mandatory and optional columns such as genomic coordinates, allele information, mutation type, sample identifiers, and additional annotations like protein changes and mutation classification (Figure 23). These columns provide detailed insights into the genomic alterations present in tumor samples, aiding in the study of cancer biology, treatment response, and the identification of potential therapeutic targets. While VCF files are more general-purpose and capture a broader range of genetic variants across genomes, MAF files are tailored for the specific needs of cancer researchers, emphasizing somatic mutations and their implications in oncology. This specialized format ensures efficient

storage, analysis, and exchange of critical mutation data within the cancer research community.

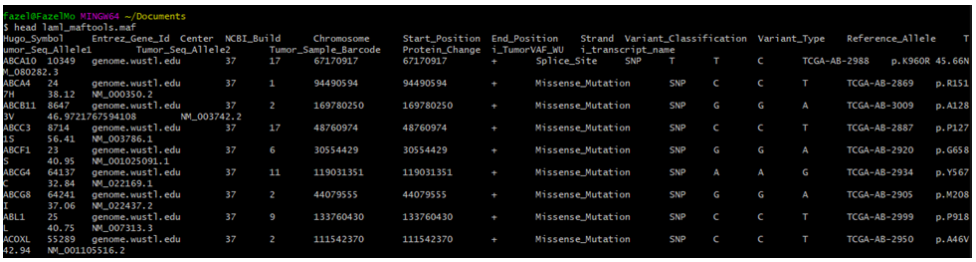


Figure 23: ***An overview of MAF file structure.***

TA Detailed View of an example Mutation Annotation File (MAF) in Action. The main feature of the interface is a table that displays various detailed annotations about genomic variants.

5.3 Somatic Variant Calling

Variant calling was performed using the *Mutect2* tool from the Genome Analysis Toolkit (GATK)[94], which is widely recognized for its accuracy and sensitivity in detecting somatic mutations. The analysis was conducted on patient samples from Leiden University Medical Center (LUMC) and the LEUCEGENE project by the team. *Mutect2* was chosen due to its robust performance in identifying low-frequency variants in heterogeneous tumor samples. The workflow involved aligning the raw sequencing reads to the reference genome using the Burrows-Wheeler Aligner (BWA) and preprocessing the alignments with GATK tools to mark duplicates, realign around indels, and recalibrate base quality scores. Subsequently, *Mutect2* was employed to call somatic mutations by comparing tumor samples against matched normal samples, allowing for precise differentiation between somatic and germline variants.

5.4 Programming Languages

5.4.1 R Programming

R is a powerful statistical programming language widely used for data analysis, statistical modeling, and visualization. In this study, R *version 4.3.1*[95] was used to perform all statistical analyses and data visualizations. The decision to use R was based on its extensive library of

packages dedicated to mutational analysis, which offers a wide range of functionalities that are essential for advanced data analysis. The use of R facilitated the efficient handling of data, execution of complex machine learning models, and creation of high-quality plots to illustrate the findings.

5.4.2 Unix Shell

Unix shell Bash scripting is a powerful tool widely used in bioinformatics for automating and managing complex workflows. Bash, the Bourne Again Shell, is the default command-line interpreter for most Unix-based systems, providing a versatile environment for executing commands, running scripts, and automating repetitive tasks.

In this study, Unix shell Bash scripting was employed to handle large bioinformatics datasets and direct the execution of various bioinformatics tools. The choice of Bash scripting was driven by its robustness, efficiency, and widespread adoption in the bioinformatics community. Bash scripts facilitate the automation of data processing pipelines, ensuring reproducibility and scalability of analyses. Large genomic datasets were processed using a series of custom Bash scripts, which streamlined tasks such as data downloading, preprocessing, and format conversion. The modular nature of Bash scripts allowed for the flexible combination of these tools, facilitating complex data analyses.

5.5 Mutational spectrum

5.5.1 Maftools package

To find the differences between impacted regions of the genes in each condition and the type of mutational consequences on the protein, we performed a pairwise analysis of unique mutations in CH and AML on the top three mutated genes, DNMT3A, TET2, and ASXL1. In this comparison, we used the R package *"Maftools"* (version 2.10.05) to perform an overall analysis of the mutations in each state[96]. We used the *'lollipopplot'* function to visually represent mutations,

highlighting specific mutations or along a reference sequence for easy analysis and comparison. We used a ridge density plot to represent the impacted regions of the protein sequence for each class of mutations in both AML and CH. Using a Venn diagram, we represent the intersection of the mutations in CH and AML samples.

5.5.2 Adobe Illustrator

In this study, the graphical representation of the mutational spectrum was constructed using *Adobe Illustrator 2024*. Adobe Illustrator is a leading vector graphics editor used extensively in both professional and academic settings for creating precise and scalable graphical illustrations. Its advanced features, such as customizable vector paths, and robust layering, make it particularly suitable for detailed scientific visualizations. The mutational spectrum plots were initially generated using specialized bioinformatics tools and subsequently imported into Adobe Illustrator. Here, they were meticulously combined and refined to produce a cohesive and visually appealing representation. This approach not only enhanced the clarity and precision of the visual data but also ensured that the graphical elements were easily modifiable and scalable for inclusion in the thesis.

5.6 Mutational signatures

For the creation of mutational signatures, we utilized the *"MutationalPattern" (version 3.10.0)* in the R programming language[97]. In the *"mutationalPattern"* package, statistical algorithms and visualization tools are integrated to facilitate the efficient and accurate analysis of mutational data, enabling the identification of mutational processes and patterns within the dataset or sample that is being studied. The *'fit_to_signatures'* function finds the linear combination of mutation signatures that most closely reconstructs the mutation matrix by solving the nonnegative least-squares constraints problem.

5.7 Model Training

5.7.1 Caret Package

All the steps regarding the model training and performed using the *"Caret" (version 6.0-94)*[98] and relative packages in the R programming language. The Caret package in R was utilized for data preprocessing, model training, and evaluation. Caret provides a consistent interface for performing these tasks across a wide variety of machine-learning algorithms.

5.7.2 Feature Engineering

To create and engineer features of our dataset, we considered 17 features that could explain each mutation (Table 8). These features were derived from a combination of intrinsic properties of the mutations and external annotations. Intrinsic properties included genomic coordinates, mutation type (single nucleotide variant, insertion, deletion), and sequence context. External annotations were sourced from public databases and included information such as gene function, physicochemical properties of Amino Acids, and mutation neighboring sequence. The *"dplyr"* package from the R programming language was utilized to facilitate the efficient manipulation and transformation of the dataset.

Table 8: *Table of features used in the model training*

Feature Name	**Type**	**Description**
Domain	Categorical	Protein domain of the mutation
Amino Acid Position	Numeric	Position of mutated amino acid in the protein sequence
Variant Class	Categorical	Class of mutation consequence (e.g. Frameshifting, Missense)
Variant type	Categorical	Type of mutation (e.g. Frameshift, SNV)
Substitution	Categorical	Type of the substitution in SNV mutations
Trimers	Categorical	Mutated nucleotides and neighbors
AA reference	Categorical	The mutated amino acid
AA alternative	Categorical	The altered amino acid in Missense mutations
Polarity reference	Categorical	The polarity of the reference amino acid (e.g., polar, nonpolar)
Charge reference	Categorical	Charge of the reference amino acid (e.g., positive, negative, neutral)
Side chain reference	Categorical	Sidechain classification of the reference amino acid (e.g., aliphatic, aromatic)
pI difference	Numeric	The difference in isoelectric point (pI) between the reference and alternative amino acids
Hydrophobicity difference	Numeric	The difference in hydrophobicity score of the reference and alternative amino acid
Molecular Weight difference	Numeric	The difference in molecular weight between the reference and alternative amino acids
Polarity alternative	Categorical	The polarity of the alternative amino acid (e.g., polar, nonpolar)
Charge alternative	Categorical	Charge of the reference amino acid (e.g., positive, negative, neutral)
Side chain alternative	Categorical	Sidechain classification of the alternative amino acid (e.g., aliphatic, aromatic)

5.7.3 Data Encoding

The categorical data in this dataset was encoded using one-hot encoding as represented in the scheme figure (figure 24). One-hot encoding is a method used to convert categorical data into a numerical format that can be provided to machine learning algorithms. This technique transforms each category value into a new binary column and assigns a 1 or 0 (True/False) to indicate the presence of the category. For instance, if we have a mutation-type categorical variable with three categories: *"Nonsense," "Missense," and "Frameshift,"* one-hot encoding will convert this into three separate binary columns: *"Nonsense," "Missense," and "Frameshift."* Each row will

have a 1 in the column corresponding to its category and a 0 in the other columns. This approach helps machine learning models interpret categorical data without assuming any ordinal relationship between the categories.

Mutation class	LABEL
Missense	AML
Nonsense	CH
Frameshift	AML

Mutation class Missense	Mutation class Nonsense	Mutation class Frameshift	LABEL
1	0	0	AML
0	1	0	CH
0	0	1	AML

Figure 24: ***Schematic representation of the One hot encoding method on a single feature.***

5.7.4 Nested Cross-Validation

To avoid overfitting and to obtain an unbiased estimate of the model's performance, we perform Nested Cross-Validation (NCV). In machine learning, nested cross-validation is used to estimate the performance of models and select hyperparameters in an unbiased and robust manner[99]. As part of this method, the dataset is divided into multiple folds, with each fold being used for both testing and training. This method consists of an inner loop that is used to select the best hyperparameters for the model and an outer loop that is used to estimate the performance of the model based on the hyperparameters selected[100].

As it is shown in Figure 25, we divided the dataset into the train (90%) and test (10%) sets, and then NCV is performed on the train set, and the test set is kept for the final evaluation of the model. In the outer loop of NCV, our training data is split into train and test sets into five

folds to estimate model performance. Using each set of hyperparameters, the outer loop trains a model on the entire training dataset and evaluates its performance on the corresponding test dataset. To obtain a more accurate estimate of the model's performance, the performance estimates from each fold of the outer loop are averaged. In the inner loop of NCV which is a train set of each fold of the outer loop, data is split into train and evaluation sets to select the best hyperparameter for the model. It trains the model using different hyperparameter values based on the training data from each fold of the outer loop. It then evaluates the model's performance on the validation set, which is a separate subset of the training data. It is the inner loop that selects the hyperparameters that result in the best performance on the validation set. For each fold of the outer loop, this process is repeated, resulting in multiple sets of hyperparameters.

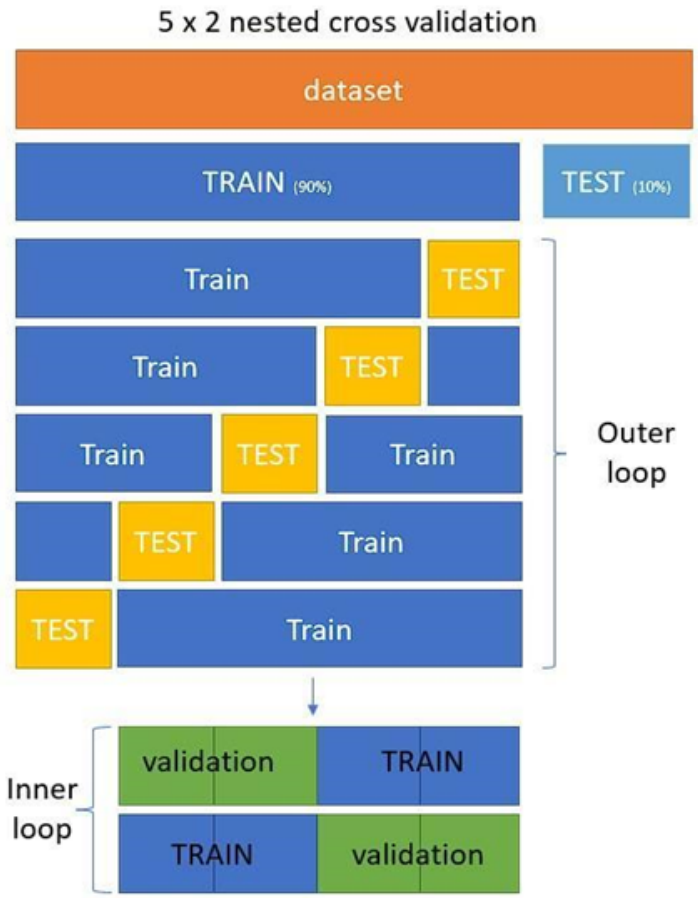


Figure 25: ***Schematic representation of Nested five*two cross-validation.***

5.7.5 Model Algorithms

The dataset in this study was trained and analyzed using Random Forest algorithms and Generalized Linear Models (GLM) to deal with the linearity or non-linearity of the data.

Generalized Linear Models (GLMs) are flexible statistical modeling techniques that generalize linear regression models by incorporating several types of response variables and error distributions. The GLM framework facilitates the modeling of relationships between dependent variables and explanatory variables[101]. In GLMs, the key hyperparameters are the link function and regularization parameter (alpha). To prevent overfitting, the link function specifies how the linear predictor relates to the response variable, and the regularization parameter controls the amount of regularization applied to the model[100]. Iteratively training models with different hyperparameter values, evaluating their performance on validation sets, and selecting the most appropriate hyperparameter configuration is accomplished using approaches such as cross-validation.

Random Forest (RF) algorithms combine multiple decision trees to produce a potent predictive model. For training decision trees, a subset of features and data instances are randomly selected. Based on the aggregation of the predictions from all the trees, the final prediction is determined[102]. For Random Forest algorithms, important hyperparameters include the number of trees, maximum depth, minimum samples for leaf nodes, maximum features, and bootstrap sampling. Tuning methods for Random Forests include grid search, random search, and Bayesian optimization[103]. These methods help explore the hyperparameter space and identify optimal configurations by evaluating models on various combinations of hyperparameters.

Due to their ensemble of decision trees, Random Forests are well suited for handling high-dimensional datasets with many features, as well as capturing nonlinear relationships among features. Furthermore, they are capable of automatically selecting features and are resistant to outliers and noise. Therefore, Random Forests are more computationally intensive and less interpretable than GLMs, as the ensemble nature of the trees makes it difficult to interpret their

predictions. In contrast, GLMs provide interpretability and statistical inference, allowing a better understanding of the relationship between the predictor variables and the response variables. Additionally, GLMs can handle a wide range of data types by using a variety of error distributions and link functions. GLMs, however, may not be as flexible in capturing complex interactions and nonlinearities as Random Forests.

These two algorithms were chosen because Random Forests captures complex interactions and handles high-dimensional datasets, which is particularly useful when dealing with large numbers of predictors. In contrast, GLMs enable statistical inference and interpretation of model parameters, thereby providing insight into the relationships between the predictor and response variables. In combination with both algorithms, we aim to obtain a comprehensive understanding of the data as well as a predictive performance by leveraging the strengths of both approaches.

5.7.6 Data Imbalance Issue

A class weight approach was employed to address the issue of class imbalance in the dataset. Machine learning commonly uses class weighting to mitigate the impact of imbalanced class distributions. The algorithm can adjust its learning process in response to the disproportionate representation of minorities by assigning different weights to each class[104]. To implement class weighting, the weights were calculated using the inverse of the class frequencies in the training data. Therefore, the weight assigned to a particular class is inversely proportional to its frequency, aiming to increase the importance of the minority class during the training process. As a result, the model becomes more sensitive to the minority class and reduces the possibility of misclassification or biased predictions because of the dominance of the majority class.

5.7.7 Evaluation Methods

To assess the importance of each feature in our models, we used the Receiver Operating Characteristic (ROC) curve analysis along with the Area Under the Curve (AUC) metric as part of

a comprehensive strategy. It was possible to evaluate the discriminative power and predictive performance of individual features using this approach[105].

We created multiple models, each exclusively incorporating a single feature, while keeping the remaining model parameters constant, to isolate the effect of each feature. Through this systematic approach, the evaluation was able to focus on the unique contribution of each feature. AUC and ROC curves were then calculated for each trained model. An insightful visual representation of the model's performance is provided by the ROC curve which plots the true positive rate against the false positive rate at various classification thresholds. AUC is a metric that summarizes the area under the ROC curve to evaluate the overall performance of a model. In general, a higher AUC value indicates superior discriminative ability and predictive accuracy.

To comprehensively assess the importance of each feature, we plotted the resulting ROC curves and AUC values together, enabling a direct comparison of their contributions. This visualization allowed us to examine the varying impact of each feature on the overall predictive power of our models.

A comprehensive set of evaluation metrics was utilized to evaluate the performance of our RF and GLM models, including Accuracy, Kappa, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, Precision, Recall, F1-score, Prevalence, Detection Rate, Detection Prevalence, and Balanced Accuracy under both two-class assumptions and three-class assumptions.

The accuracy of the models is a fundamental measure of their ability to make accurate predictions. In contrast, Kappa measures the level of agreement among the raters, which is particularly important when dealing with imbalanced datasets. The model calculates the observed agreement and compares it to the expected agreement. However, a Kappa value greater than 0.6 indicates substantial agreement beyond chance, whereas a value below 0.2 indicates poor agreement. When classes are imbalanced, Kappa helps to account for random agreement, making it a valuable metric for assessing prediction reliability and consistency.

Sensitivity, also known as the True Positive Rate, is a measure of the model's ability to correctly identify positive instances. Specificity, or True Negative Rate, is a measure of a model's ability to identify negative instances correctly. By using these metrics, we can determine whether the models can minimize false negatives and false positives, respectively.

Recall, also known as True Positive Rate or Sensitivity, indicates whether the model is capable of correctly identifying positive instances. An F1-score provides a balanced measure of the accuracy of the models, as it is the harmonic mean of precision and recall.

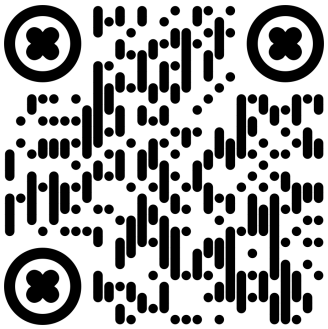
As a measure of the base rate of the target variable, prevalence represents the proportion of positive instances in the dataset. Detection Rate represents the proportion of correctly predicted positive instances, while Detection Prevalence represents the prevalence of these instances.

An overall assessment of model performance, based on the average of Sensitivity and Specificity, is presented by Balanced Accuracy, a measure combining both positive and negative results. Our study employs this comprehensive set of evaluation metrics to thoroughly assess the predictive capabilities of the RF and GLM models. We can make informed comparisons between the two approaches by using these metrics collectively to provide a holistic understanding of the models' performance.

5.8 Code Availability and Supplementary Material

The code used for the analysis and simulations in this thesis, along with supplementary materials including datasets, additional figures, and tables, are available on GitHub. These resources are provided to support the findings of this thesis. The repository includes detailed instructions on how to set up and use the code, as well as access to all supplementary materials.
Access the repository through the following link or by scanning the QR code:

"https://github.com/Fazelmohammadii/MasterThesisProject.git"



6 BIBLIOGRAPHY

[1] Ming Yi et al. “The global burden and attributable risk factor analysis of acute myeloid leukemia in 195 countries and territories from 1990 to 2017: estimates based on the global burden of disease study 2017”. In: *Journal of Hematology & Oncology* 13 (2020). DOI: 10.1186/s13045-020-00908-z.

[2] S. Fircanis, P. Merriam, Naushaba Khan, and J. Castillo. “The relation between cigarette smoking and risk of acute myeloid leukemia: An updated meta-analysis of epidemiological studies”. In: *American Journal of Hematology* 89 (2014). DOI: 10.1002/ajh.23744.

[3] F. Appelbaum et al. “Age and acute myeloid leukemia.” In: *Blood* 107 9 (2006), pp. 3481–5. DOI: 10.1182/BLOOD-2005-09-3724.

[4] Silvia Calabria et al. “Acute myeloid leukemia: Incidence, transplantation and survival through Italian administrative healthcare data”. eng. In: *Tumori* 109.5 (Oct. 2023), pp. 496–503. ISSN: 2038-2529. DOI: 10.1177/03008916231153698.

[5] Simon Haas, Andreas Trumpp, and Michael D. Milsom. “Causes and Consequences of Hematopoietic Stem Cell Heterogeneity”. In: *Cell Stem Cell* 22.5 (May 2018), pp. 627–638. ISSN: 1934-5909. DOI: 10.1016/j.stem.2018.04.003.

[6] Adam C. Wilkinson, Kyomi J. Igarashi, and Hiromitsu Nakauchi. “Haematopoietic stem cell self-renewal in vivo and ex vivo”. en. In: *Nature Reviews Genetics* 21.9 (Sept. 2020). Publisher: Nature Publishing Group, pp. 541–554. ISSN: 1471-0064. DOI: 10.1038/s41576-020-0241-0.

[7] Oriol Pich, Iker Reyes-Salazar, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. “Discovering the drivers of clonal hematopoiesis”. In: *Nature Communications* 13 (July 2022), p. 4267. ISSN: 2041-1723. DOI: 10.1038/s41467-022-31878-0.

[8] A. Bick et al. “Inherited Causes of Clonal Hematopoiesis in 97,691 TOPMed Whole Genomes”. In: *Nature* (2020). DOI: 10.1038/s41586-020-2819-2.

[9] Eva Mejia-Ramirez and Maria Carolina Florian. “Understanding intrinsic hematopoietic stem cell aging”. In: *Haematologica* 105.1 (Jan. 2020), pp. 22–37. ISSN: 0390-6078. DOI: 10.3324/haematol.2018.211342.

[10] Shahar Biechonski et al. “Attenuated DNA damage responses and increased apoptosis characterize human hematopoietic stem cells exposed to irradiation”. en. In: *Scientific Reports* 8.1 (Apr. 2018). Publisher: Nature Publishing Group, p. 6071. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24440-w.

[11] Siddhartha Jaiswal et al. “Age-related clonal hematopoiesis associated with adverse outcomes”. eng. In: *The New England Journal of Medicine* 371.26 (Dec. 2014), pp. 2488–2498. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1408617.

[12] Charles Gaulin, Katalin Kelemen, and Cecilia Arana Yi. “Molecular Pathways in Clonal Hematopoiesis: From the Acquisition of Somatic Mutations to Transformation into Hematologic Neoplasm”. en. In: *Life* 12.8 (Aug. 2022). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 1135. ISSN: 2075-1729. DOI: 10.3390/life12081135.

[13] Julia T. Warren and Daniel C. Link. “Clonal hematopoiesis and risk for hematologic malignancy”. eng. In: *Blood* 136.14 (Oct. 2020), pp. 1599–1605. ISSN: 1528-0020. DOI: 10.1182/blood.2019000991.

[14] Tyler S. Alioto et al. “A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing”. en. In: *Nature Communications* 6.1 (Dec. 2015). Publisher: Nature Publishing Group, p. 10001. ISSN: 2041-1723. DOI: 10.1038/ncomms10001.

[15] D. Steensma. “Clinical Implications of Clonal Hematopoiesis”. In: *Mayo Clinic Proceedings* 93 (2018). DOI: 10.1016/j.mayocp.2018.04.002.

[16] S. Reed, Sarah Croessmann, and B. H. Park. “CHIP Happens: Clonal hematopoiesis of indeterminate potential and its relationship to solid tumors.” In: *Clinical cancer research : an official journal of the American Association for Cancer Research* (2022). DOI: 10.1158/1078-0432.CCR-22-2598.

[17] Christopher S. Marnell, Alexander Bick, and Pradeep Natarajan. “Clonal Hematopoiesis of Indeterminate Potential (CHIP): Linking Somatic Mutations, Hematopoiesis, Chronic Inflammation and Cardiovascular Disease”. In: *Journal of molecular and cellular cardiology* 161 (Dec. 2021), pp. 98–105. ISSN: 0022-2828. DOI: 10.1016/j.yjmcc.2021.07.004.

[18] M. Rossi et al. “Clinical Relevance of Clonal Hematopoiesis in the Oldest-Old Population: Analysis of the "Health and Anemia" Study”. In: *Blood* (2018). DOI: 10.1182/BLOOD-2018-99-114717.

[19] Camilla Bertuzzo Veiga, Erin M. Lawrence, Andrew J. Murphy, Marco J. Herold, and Dragana Dragoljevic. “Myelodysplasia Syndrome, Clonal Hematopoiesis and Cardiovascular Disease”. In: *Cancers* 13.8 (Apr. 2021), p. 1968. ISSN: 2072-6694. DOI: 10.3390/cancers13081968.

[20] B. Deschler, T. D. de Witte, R. Mertelsmann, and M. Lübbert. “Treatment decision-making for older patients with high-risk myelodysplastic syndrome or acute myeloid leukemia: problems and approaches.” In: *Haematologica* 91 11 (2006), pp. 1513–22.

[21] Xiao-Qian Xu et al. “Characteristics of acute myeloid leukemia with myelodysplasia-related changes: A retrospective analysis in a cohort of Chinese patients”. eng. In: *American Journal of Hematology* 89.9 (Sept. 2014), pp. 874–881. ISSN: 1096-8652. DOI: 10.1002/ajh.23772.

[22] P J Fialkow, S M Gartler, and A Yoshida. “Clonal origin of chronic myelocytic leukemia in man.” In: *Proceedings of the National Academy of Sciences of the United States of America* 58.4 (Oct. 1967), pp. 1468–1471. ISSN: 0027-8424.

[23] Thomas Köhnke and Ravindra Majeti. “Clonal hematopoiesis: from mechanisms to clinical intervention”. In: *Cancer discovery* 11.12 (Dec. 2021), pp. 2987–2997. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-21-0901.

[24] Herra Ahmad, Nikolaus Jahn, and Siddhartha Jaiswal. “Clonal Hematopoiesis and Its Impact on Human Health”. In: *Annual Review of Medicine* 74.1 (2023), pp. 249–260. DOI: 10.1146/annurev-med-042921-112347.

[25] Min Joo Kim et al. “Clonal hematopoiesis as a novel risk factor for type 2 diabetes mellitus in patients with hypercholesterolemia”. eng. In: *Frontiers in Public Health* 11 (2023), p. 1181879. ISSN: 2296-2565. DOI: 10.3389/fpubh.2023.1181879.

[26] Simon N. Stacey et al. “Genetics and epidemiology of mutational barcode-defined clonal hematopoiesis”. en. In: *Nature Genetics* 55.12 (Dec. 2023). Publisher: Nature Publishing Group, pp. 2149–2159. ISSN: 1546-1718. DOI: 10.1038/s41588-023-01555-z.

[27] Isabelle A. van Zeventer et al. “Evolutionary landscape of clonal hematopoiesis in 3,359 individuals from the general population”. en. In: *Cancer Cell* (May 2023). ISSN: 1535-6108. DOI: 10.1016/j.ccell.2023.04.006.

[28] Carin L. E. Hazenberg et al. “Clonal hematopoiesis in patients with stem cell mobilization failure: a nested case-control study”. In: *Blood Advances* 7.7 (Mar. 2023), pp. 1269–1278. ISSN: 2473-9529. DOI: 10.1182/bloodadvances.2022007497.

[29] David P. Steensma and Kelly L. Bolton. “What to tell your patient with clonal hematopoiesis and why: insights from 2 specialized clinics”. In: *Blood* 136.14 (Oct. 2020), pp. 1623–1631. ISSN: 0006-4971. DOI: 10.1182/blood.2019004291.

[30] Mihee Kim et al. “Prognostic Analysis According to European Leukemianet 2022 Risk Stratification for Elderly Patients with Acute Myeloid Leukemia Treated with Decitabine”. In: *Blood* 140.Supplement 1 (Nov. 2022), pp. 6094–6095. ISSN: 0006-4971. DOI: 10.1182/blood-2022-164916.

[31] Krzysztof Mrózek et al. “Outcome prediction by the 2022 European LeukemiaNet genetic-risk classification for adults with acute myeloid leukemia: an Alliance study”. eng. In: *Leukemia* 37.4 (Apr. 2023), pp. 788–798. ISSN: 1476-5551. DOI: 10.1038/s41375-023-01846-8.

[32] Stephen A. Strickland and Norbert Vey. “Diagnosis and treatment of therapy-related acute myeloid leukemia”. In: *Critical Reviews in Oncology/Hematology* 171 (Mar. 2022), p. 103607. ISSN: 1040-8428. DOI: 10.1016/j.critrevonc.2022.103607.

[33] Dahui Qin. “Molecular testing for acute myeloid leukemia”. In: *Cancer Biology & Medicine* 19.1 (Jan. 2022), pp. 4–13. ISSN: 2095-3941. DOI: 10.20892/j.issn.2095-3941.2020.0734.

[34] H. Keiser, D. Goldstein, James L. Wade, F. Douglas, and S. Averbuch. “Treatment of Malignant Pheochromocytoma with Combination Chemotherapy”. In: *Hypertension* 7 (1985). DOI: 10.1161/01.HYP.7.3_PT_2.I18.

[35] Kenny Tang, Andre C. Schuh, and Karen Wl Yee. “3+7 Combined Chemotherapy for Acute Myeloid Leukemia: Is It Time to Say Goodbye?” eng. In: *Current Oncology Reports* 23.10 (Aug. 2021), p. 120. ISSN: 1534-6269. DOI: 10.1007/s11912-021-01108-9.

[36] Jan J. Cornelissen et al. “Results of a HOVON/SAKK donor versus no-donor analysis of myeloablative HLA-identical sibling stem cell transplantation in first remission acute myeloid leukemia in young and middle-aged adults: benefits for whom?” eng. In: *Blood* 109.9 (May 2007), pp. 3658–3666. ISSN: 0006-4971. DOI: 10.1182/blood-2006-06-025627.

[37] Karam Khaddour, Caroline K. Hana, and Prerna Mewawalla. “Hematopoietic Stem Cell Transplantation”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[38] N. Daver, R. Schlenk, N. Russell, and M. Levis. “Targeting FLT3 mutations in AML: review of current knowledge and evidence”. In: *Leukemia* 33 (2019), pp. 299–312. DOI: 10.1038/s41375-018-0357-9.

[39] Courtney D. DiNardo et al. “Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML”. eng. In: *The New England Journal of Medicine* 378.25 (June 2018), pp. 2386–2398. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1716984.

[40] Yasmin Abaza and Amir T. Fathi. “Monoclonal Antibodies in Acute Myeloid Leukemia-Are We There Yet?” eng. In: *Cancer Journal (Sudbury, Mass.)* 28.1 (Feb. 2022), pp. 37–42. ISSN: 1540-336X. DOI: 10.1097/PPO.0000000000000577.

[41] Sylvie Castaigne et al. “Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): a randomised, open-label, phase 3 study”. eng. In: *Lancet (London, England)* 379.9825 (Apr. 2012), pp. 1508–1516. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(12)60485-1.

[42] Kieran D Sahasrabudhe et al. “Effect of High Intensity Chemotherapy Vs Targeted Therapy on Survival in AML Patients Aged 60-75”. In: *Blood* 138.Supplement 1 (Nov. 2021), p. 4125. ISSN: 0006-4971. DOI: 10.1182/blood-2021-148676.

[43] Mingchao Xie et al. “Age-related mutations associated with clonal hematopoietic expansion and malignancies”. eng. In: *Nature Medicine* 20.12 (Dec. 2014), pp. 1472–1478. ISSN: 1546-170X. DOI: 10.1038/nm.3733.

[44] Salah Aref et al. “Clinical Implication of DNMT3A and TET2 Genes Mutations in Cytogenetically Normal Acute Myeloid Leukemia”. In: *Asian Pacific Journal of Cancer Prevention : APJCP* 23.12 (Dec. 2022), pp. 4299–4305. ISSN: 1513-7368. DOI: 10.31557/APJCP.2022.23.12.4299.

[45] Jifeng Yu, Yingmei Li, Danfeng Zhang, Dingming Wan, and Zhongxing Jiang. “Clinical implications of recurrent gene mutations in acute myeloid leukemia”. In: *Experimental Hematology & Oncology* 9 (Mar. 2020), p. 4. ISSN: 2162-3619. DOI: 10.1186/s40164-020-00161-7.

[46] Ley Timothy J. et al. “DNMT3A Mutations in Acute Myeloid Leukemia”. In: *New England Journal of Medicine* 363.25 (2010), pp. 2424–2433. DOI: 10.1056/NEJMoa1005143.

[47] Quanyi Lu, Yamei Chen, Hang Wang, and Zhipeng Li. “DNMT3A mutations and clinical features in Chinese patients with acute myeloid leukemia”. In: *Cancer Cell International* (2013). DOI: 10.1186/1475-2867-13-1.

[48] Samuel Ojo Abegunde and Michael J. Rauh. “*Tet2*-Deficient Bone Marrow Progenitors Have a Proliferative Advantage in the Presence of TNF-Alpha and IFN-Gamma: Implications for Clonal Dominance in Inflammaging and MDS”. In: *Blood* 126.23 (Dec. 2015), p. 2850. ISSN: 0006-4971. DOI: 10.1182/blood.V126.23.2850.2850.

[49] Brooke Snetsinger et al. “Myeloid-Derived Suppressor Cell (MDSC) Dynamics In FVIII-Exposed Hemophilia A Mice: Novel Therapeutic Implications”. In: *Blood* 122.21 (Nov. 2013), p. 3569. ISSN: 0006-4971. DOI: 10.1182/blood.V122.21.3569.3569.

[50] Michael J. Bamshad et al. “Exome sequencing as a tool for Mendelian disease gene discovery”. eng. In: *Nature Reviews. Genetics* 12.11 (Sept. 2011), pp. 745–755. ISSN: 1471-0064. DOI: 10.1038/nrg3031.

[51] Matthew Meyerson, Stacey Gabriel, and Gad Getz. “Advances in understanding cancer genomes through second-generation sequencing”. eng. In: *Nature Reviews. Genetics* 11.10 (Oct. 2010), pp. 685–696. ISSN: 1471-0064. DOI: 10.1038/nrg2841.

[52] *A Beginner’s Guide to DNA Sequencing*.

[53] Catherine C. Coombs et al. “Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes”. eng. In: *Cell Stem Cell* 21.3 (Sept. 2017), 374–382.e4. ISSN: 1875-9777. DOI: 10.1016/j.stem.2017.07.010.

[54] Domingo Pascual-Figal et al. “Sacubitril-Valsartan, Clinical Benefits and Related Mechanisms of Action in Heart Failure With Reduced Ejection Fraction. A Review”. eng. In: *Frontiers in Cardiovascular Medicine* 8 (2021), p. 754499. ISSN: 2297-055X. DOI: 10.3389/fcvm.2021.754499.

[55] Sejin Park, Jihee Soh, and Hyunju Lee. “Super.FELT: supervised feature extraction learning using triplet loss for drug response prediction with multi-omics data”. In: *BMC Bioinformatics* 22.1 (May 2021), p. 269. ISSN: 1471-2105. DOI: 10.1186/s12859-021-04146-z.

[56] Caitlyn Vlasschaert et al. *A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic datasets*. en. Pages: 2022.10.21.22281368. Oct. 2022. DOI: 10.1101/2022.10.21.22281368.

[57] J. Fresnedo-Ramírez et al. “Computational Analysis of AmpSeq Data for Targeted, High-Throughput Genotyping of Amplicons”. In: *Frontiers in Plant Science* 10 (2019). DOI: 10.3389/fpls.2019.00599.

[58] Md Mesbah Uddin et al. *Cost effective sequencing enables longitudinal profiling of clonal hematopoiesis*. en. Pages: 2022.01.31.22270028. Feb. 2022. DOI: 10.1101/2022.01.31.22270028.

[59] Daniele Raimondi, Antoine Passemiers, Piero Fariselli, and Yves Moreau. “Current cancer driver variant predictors learn to recognize driver genes instead of functional variants”. In: *BMC Biology* 19.1 (Jan. 2021), p. 3. ISSN: 1741-7007. DOI: 10.1186/s12915-020-00930-0.

[60] Theodoros Rampias. “Exploring the Eco-Evolutionary Dynamics of Tumor Subclones”. en. In: *Cancers* 12.11 (Nov. 2020). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 3436. ISSN: 2072-6694. DOI: 10.3390/cancers12113436.

[61] W. Schneider and Hua Guo. “Machine Learning.” In: *The journal of physical chemistry. B* 122 4 (2018). DOI: 10.1021/acs.jpcb.8b00035.

[62] Maxwell W. Libbrecht and William Stafford Noble. “Machine learning applications in genetics and genomics”. en. In: *Nature Reviews Genetics* 16.6 (June 2015). Publisher: Nature Publishing Group, pp. 321–332. ISSN: 1471-0064. DOI: 10.1038/nrg3920.

[63] Mitja Briscik, Gabriele Tazza, László Vidács, Marie-Agnes Dillies, and Sébastien Dejean. “Supervised Multiple Kernel Learning approaches for multi-omics data integration”. Mar. 2024. DOI: 10.48550/arXiv.2403.18355.

[64] Debabrata Acharya and Anirban Mukhopadhyay. “A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology”. In: *Briefings in Functional Genomics* (Apr. 2024), elae013. ISSN: 2041-2657. DOI: 10.1093/bfgp/elae013.

[65] *What is a Generative Adversarial Network? | Data Basecamp*. en-US. Section: Machine Learning. Apr. 2022.

[66] Nofe Alganmi. “A Comprehensive Review of the Impact of Machine Learning and Omics on Rare Neurological Diseases”. en. In: *BioMedInformatics* 4.2 (June 2024). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, pp. 1329–1347. ISSN: 2673-7426. DOI: 10.3390/biomedinformatics4020073.

[67] Claudia Cava, Soudabeh Sabetian, Christian Salvatore, and Isabella Castiglioni. “Pan-cancer classification of multi-omics data based on machine learning models”. en. In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 13.1 (Feb. 2024), p. 6. ISSN: 2192-6670. DOI: 10.1007/s13721-024-00441-w.

[68] Lin Yuan, Jing Zhao, Tao Sun, and Zhen Shen. “A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs”. In: *BMC Bioinformatics* 22 (June 2021). DOI: 10.1186/s12859-021-04256-8.

[69] Bruno César Feltes, Joice de Faria Poloni, Itamar José Guimarães Nunes, Sara Socorro Faria, and Marcio Dorn. "Multi-Approach Bioinformatics Analysis of Curated Omics Data Provides a Gene Expression Panorama for Multiple Cancer Types". English. In: *Frontiers in Genetics* 11 (Nov. 2020). Publisher: Frontiers. ISSN: 1664-8021. DOI: 10.3389/fgene.2020.586602.

[70] Florian Zink et al. "Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly". In: *Blood* 130.6 (Aug. 2017), pp. 742–752. ISSN: 0006-4971. DOI: 10.1182/blood-2017-02-769869.

[71] null null null. "Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia". In: *New England Journal of Medicine* 368.22 (May 2013), pp. 2059–2074. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1301689.

[72] Jason E. Farrar et al. "Genomic Profiling of Pediatric Acute Myeloid Leukemia Reveals a Changing Mutational Landscape from Disease Diagnosis to Relapse". eng. In: *Cancer Research* 76.8 (Apr. 2016), pp. 2197–2205. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-15-1015.

[73] Jeffrey W. Tyner et al. "Functional genomic landscape of acute myeloid leukaemia". en. In: *Nature* 562.7728 (Oct. 2018). Number: 7728 Publisher: Nature Publishing Group, pp. 526–531. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0623-z.

[74] Tara Macrae et al. "RNA-Seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells". eng. In: *PloS One* 8.9 (2013), e72884. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0072884.

[75] Wibowo Arindrarto et al. "Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing". eng. In: *Leukemia* 35.1 (Jan. 2021), pp. 47–61. ISSN: 1476-5551. DOI: 10.1038/s41375-020-0762-8.

[76] Andrew L. Young, R. Spencer Tong, Brenda M. Birmann, and Todd E. Druley. "Clonal hematopoiesis and risk of acute myeloid leukemia". en. In: *Haematologica* 104.12 (Dec. 2019). Number: 12, pp. 2410–2417. ISSN: 1592-8721. DOI: 10.3324/haematol.2018.215269.

[77] C. J. Watson et al. “The evolutionary dynamics and fitness landscape of clonal hematopoiesis”. In: *Science* (2020). DOI: 10.1126/science.aay9333.

[78] T. Ley et al. “DNMT3A mutations in acute myeloid leukemia.” In: *The New England journal of medicine* 363 25 (2010). DOI: 10.1056/NEJMoa1005143.

[79] Jeppe F. Severens et al. *Mapping AML heterogeneity – multi-cohort transcriptomic analysis identifies novel clusters and divergent ex-vivo drug responses*. en. ISSN: 2328-7896 Pages: 2023.03.29.23287896. Dec. 2023. DOI: 10.1101/2023.03.29.23287896.

[80] Ludmil B. Alexandrov et al. “The repertoire of mutational signatures in human cancer”. en. In: *Nature* 578.7793 (Feb. 2020). Number: 7793 Publisher: Nature Publishing Group, pp. 94–101. ISSN: 1476-4687. DOI: 10.1038/s41586-020-1943-3.

[81] David H. Spencer et al. “CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression”. In: *Cell* 168.5 (Feb. 2017), 801–816.e13. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.01.021.

[82] Serena Nik-Zainal et al. “Mutational Processes Molding the Genomes of 21 Breast Cancers”. In: *Cell* 149.5 (May 2012), pp. 979–993. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.04.024.

[83] Mark R. J. Junge and Joseph R. Dettori. “ROC Solid: Receiver Operator Characteristic (ROC) Curves as a Foundation for Better Diagnostic Tests”. In: *Global Spine Journal* 8.4 (June 2018), pp. 424–429. ISSN: 2192-5682. DOI: 10.1177/2192568218778294.

[84] Hadi Emami and Mostafa Emami. “Local Influence in Constrained General Linear Models”. en. In: *Journal of Data Science* 12.4 (Aug. 2022). Publisher: School of Statistics, Renmin University of China, pp. 717–726. ISSN: 1680-743X, 1683-8602. DOI: 10.6339/JDS.201410_12(4).0008.

[85] H. Takeshima et al. “Distinct DNA methylation activity of Dnmt3a and Dnmt3b towards naked and nucleosomal DNA.” In: *Journal of biochemistry* (2006). DOI: 10.1093/JB/MVJ044.

[86] M Katoh. “Functional and cancer genomics of ASXL family members”. In: *British Journal of Cancer* 109.2 (July 2013), pp. 299–306. ISSN: 0007-0920. DOI: 10.1038/bjc.2013.281.

[87] M. Figueroa et al. “Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation.” In: *Cancer cell* 18 6 (2010). DOI: 10.1016/j.ccr.2010.11.015.

[88] M. Loberg et al. “Sequentially inducible mouse models reveal that Npm1 mutation causes malignant transformation of Dnmt3a-mutant clonal hematopoiesis”. In: *Leukemia* 33 (2019). DOI: 10.1038/s41375-018-0368-6.

[89] Cathie Sudlow et al. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLoS Medicine* 12.3 (Mar. 2015), e1001779. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.1001779.

[90] Elli Papaemmanuil et al. “Genomic Classification and Prognosis in Acute Myeloid Leukemia”. In: *New England Journal of Medicine* 374.23 (June 2016), pp. 2209–2221. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1516192.

[91] W. James Kent et al. “The Human Genome Browser at UCSC”. en. In: *Genome Research* 12.6 (June 2002). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 996–1006. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.229102.

[92] William McLaren et al. “The Ensembl Variant Effect Predictor”. In: *Genome Biology* 17.1 (June 2016), p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

[93] Tariq Abdullah and Ahmed Ahmet. “Extracting Insights: A Data Centre Architecture Approach in Million Genome Era”. In: Nov. 2020, pp. 1–31. ISBN: 978-3-662-62385-5. DOI: 10.1007/978-3-662-62386-2_1.

[94] Geraldine A. Van der Auwera et al. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline”. en. In: *Current Protocols in Bioinformatics* 43.1 (2013), pp. 11.10.1–11.10.33. ISSN: 1934-340X. DOI: 10.1002/0471250953.bi1110s43.

[95] R Core Team. “R: A Language and Environment for Statistical Computing”. In: *R Foundation for Statistical Computing, Vienna, Austria* (2023).

[96] Anand Mayakonda, De-Chen Lin, Yassen Assenov, Christoph Plass, and H. Phillip Koeffler. "Maftools: efficient and comprehensive analysis of somatic variants in cancer". en. In: *Genome Research* 28.11 (Nov. 2018). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1747–1756. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.239244.118.

[97] Francis Blokzijl, Roel Janssen, Ruben van Boxtel, and Edwin Cuppen. "MutationalPatterns: comprehensive genome-wide analysis of mutational processes". In: *Genome Medicine* 10.1 (Apr. 2018), p. 33. ISSN: 1756-994X. DOI: 10.1186/s13073-018-0539-0.

[98] Max Kuhn. "Building Predictive Models in R Using the caret Package". en. In: *Journal of Statistical Software* 28 (Nov. 2008), pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05.

[99] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics Surveys* 4.none (Jan. 2010). Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada, pp. 40–79. ISSN: 1935-7516. DOI: 10.1214/09-SS054.

[100] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7.1 (Feb. 2006), p. 91. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-91.

[101] P. McCullagh. *Generalized Linear Models*. 2nd ed. New York: Routledge, Jan. 2019. ISBN: 978-0-203-75373-6. DOI: 10.1201/9780203753736.

[102] Leo Breiman. "Random Forests". en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[103] Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix. "Hyperparameters and Tuning Strategies for Random Forest". In: *WIREs Data Mining and Knowledge Discovery* 9.3 (May 2019). arXiv:1804.03515 [cs, stat]. ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1301.

[104] Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. *Class-Weighted Classification: Trade-offs and Robust Approaches*. arXiv:2005.12914 [cs, stat]. May 2020. DOI: 10.48550/arXiv.2005.12914.

[105] Lei Sun, Jun Wang, and Jinmao Wei. "AVC: Selecting discriminative features on basis of AUC by maximizing variable complementarity". In: *BMC Bioinformatics* 18.3 (Mar. 2017), p. 50. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1468-4.