Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by:

Ralitsa Langova, Master of Science

Born in: Sofia, Bulgaria

Oral examination:16$^{\text{th}}$ June 2023

Multi-omics of AML

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. Stefan Fröhling

Contents

ABSTRACT

Acute myeloid leukemia (AML) is one of the most aggressive hematopoietic malignancies and has been recognized as a heterogeneous disease due to a lack of unifying characteristics. It is driven by different genome aberrations, gene expression changes, and epigenomic dysregulations. Therefore a multi-omics approach is needed to unravel the complex biology of this disease. This thesis deals with the challenges of identifying driver events that account for differences in clinical phenotypes and responses to treatment. The work presented here investigates the driver events of AML and epigenetics drug response profiles. The thesis consists of three main projects. The first study identifies recurrent mutations in AML carrying t(8;16)(p11;p13), a rare abnormality. The second project is identifying prospective drivers of mutation-negative nkAML. The third project concentrates on epigenetic changes after AML drugs.

t(8;16) AML is a rare and distinguishable clinicopathological entity. Some previous reports that represented the characteristics of patients with this type of AML suggest that the t(8;16) translocation could be sufficient to induce hematopoietic cell transformation to AML without acquiring other genetic alterations. Therefore here I evaluate the frequently mutated genes and compare them with the most frequent mutated genes in AML in general and AML carrying t(8;16) translocation. FLT3 mutation was found in 3 patients of my cohort, a potential target for therapy with tyrosine kinase inhibitors. However, exciting finding was the mutations in EYS, KRTAP9-1, PSIP1, and SPTBN5 that were depicted earlier in AML.

Elucidating different layers of aberrations in normal karyotype no-driver acute myeloid leukemia provides better biology insight and may impact risk-group stratification and new potential driver events. Therefore, this study aimed to detect such anomalies in samples without known driver genetic abnormalities using multi-omic molecular profiling. Samples were analyzed using RNA sequencing (n=43), whole genome sequencing (n=43), and EPIC DNA methylation array (n=42). In 33 of 43 patients, all three layers of data were available. I developed a pipeline looking for a driver in any layer of data by connecting the information of all layers of data and utilizing public genomic, transcriptomic, and clinical data available from TCGA. Genetic alterations of somatic cells can drive malignant clone formation and promote leukemogenesis. Therefore I first built a mutation prioritization workflow that checks each patient's genomic mutation drivers. Here I use the information on the allele frequency of the specific mutation combining information from WGS and RNA sequencing data. Finally, I compared each mutation on a positional level with AML and other TCGA cancer cohorts to assess the causative genomic mutations. I found potential driver stopgain mutation in genes implicated in chromosome segregation during mitosis and some tumor suppressor genes. I found new stopgain mutations in cancer genes (NIPBL and NF1). Since fusions are increasingly acknowledged as oncology therapeutic targets, I investigated potential driver fusion events by evaluating high-confidence and in-frame cancer-related fusion findings. As a result, I found specific gene fusion patterns. Kinases activated by gene fusions define a meaningful class of oncogenes associated with hematopoietic malignancies. I identify several novel and recurrent fusions involving kinases that potentially play a role in leukemogenesis. I detected previously unreported fusions involving known cancer-related genes, such as PIM3- RAC2 and PROK2- EIF4E3. In addition, outliers, such as gene expression levels, can pinpoint potential pathogenic events. Therefore, combining my AML cohort with a healthy control group, I determined aberrant gene expression levels as possible pathogenic events using the deep learning method. Finally, I combined the data and looked for a comparison to the methylation pattern of each patient. Overall, the analysis uncovered a rich landscape of potential drivers. In different data layers, I found an altered genomic and transcriptomic signature of different GTPases, which are known to be involved in many stages of tumorigenesis. My methods and results demonstrate the power of integrating multi-omics data to study complex driver alterations in AML and point to future directions of research that aim to bridge gaps in research and clinical appli-

cations. Furthermore, I provide in vitro evidence for antileukemic cooperativity and epigenetic activity between DAC and ATRA. I performed differential methylation analysis on CpG resolution and across genomic and transposable elements regions, enhancing the results' statistical power and interpretability. I demonstrated that single-agent ATRA caused no global demethylation, nor did ATRA improve the demethylation mediated by DAC. In summary, combining multi-omics profiling is a powerful tool for studying dysregulated patterns in AML. Furthermore, multi-omics profiling performed on mutation-negative nkAML reveals several promising drivers. My findings not only go beyond augmenting my understanding of the heterogeneity landscape of AML but also may have immediate implications for new targeted therapy studies.

ZUSAMMENFASSUNG

Akute myeloische Leukämie (AML) ist eines der aggressivsten hämatopoetischen Malignome und wurde aufgrund fehlender einheitlicher Merkmale als heterogene Krankheit anerkannt. AML wird durch verschiedene Genomaberrationen, Veränderungen der Genexpression und epigenomische Dysregulationen angetrieben. Daher ist ein Multi-Omics-Ansatz erforderlich, um die komplexe Biologie dieser Krankheit zu entschlüsseln. Diese Dissertation beschäftigt sich mit den Herausforderungen, treibende Ereignisse zu identifizieren, die für Unterschiede hinsichtlich klinischer Phänotype und Ansprechen auf die Behandlung. Die hier vorgestellte Arbeit untersucht die Treiberereignisse von AML und epigenetische Arzneimittelreaktionsprofile. Die Dissertation besteht aus drei Hauptprojekten. Die erste Studie identifiziert rezidivierende Mutationen bei AML, die die die seltene Anomalie t(8;16)(p11;p13) tragen. Das zweite Projekt identifiziert potenzielle Treiber von mutationsnegativem nkAML. Das dritte Projekt konzentriert sich auf epigenetische Veränderungen nach der Behandlung von AML.

t(8;16) AML ist eine seltene und unterscheidbare klinisch-pathologische Entität. Einige frühere Berichte, die die Merkmale von Patienten mit dieser Art von AML darstellten, legen nahe, dass die t(8;16)-Translokation ausreichen könnte, um eine hämatopoetische Zelltransformation zu AML zu induzieren, ohne andere genetische Veränderungen zu erwerben. Daher werten ich hier die häufig mutierten Gene aus und vergleiche sie mit den am häufigsten mutierten Genen bei AML im Allgemeinen und AML mit t(8;16)-Translokation. Bei 3 Patienten unserer Kohorte wurde eine FLT3-Mutation gefunden, ein potenzielles Ziel für eine Therapie mit Tyrosinkinase-Inhibitoren. Ein spannendes Ergebnis waren jedoch die Mutationen in EYS, KRTAP9-1, PSIP1 und SPTBN5, die bereits früher in AML beschrieben wurden.

Die Aufklärung verschiedener Ebenen von Aberrationen bei akuter myeloischer Leukämie ohne Treibermutation mit normalem Karyotyp bietet einen besseren Einblick in die Biologie und kann sich auf die Risikogruppenstratifizierung und neue potenzielle Treiberereignisse auswirken. Daher zielte diese Studie darauf ab, solche Anomalien in Proben ohne bekannte genetische Treiberanomalien mithilfe von molekularem Multi-Omic-Profiling zu erkennen. Die Proben wurden mittels RNA-Sequenzierung (n=43), Gesamtgenomsequenzierung (n=43) und EPIC-DNA-Methylierungs-Array (n=42) analysiert. Bei 33 von 43 Patienten waren alle drei Daten-Typen verfügbar. Ich habe eine Pipeline entwickelt, die in jedem Daten-Typ nach einem Treiber sucht, indem ichdie Informationen aller Daten-Typen verbunden und öffentliche genomische, transkriptomische und klinische Daten verwendet haben, die von TCGA verfügbar sind. Genetische Veränderungen somatischer Zellen können die maligne Klonbildung vorantreiben und die Leukämogenese fördern. Daher habe ich zunächst einen Workflow zur Priorisierung von Mutationen entwickelt, der die genomischen Mutationstreiber jedes Patienten überprüft. Hier verwende ich die Informationen über die Allelhäufigkeit der spezifischen Mutation, indem ich Informationen aus WGS- und RNA-Sequenzierungsdaten kombiniere. Schließlich vergleiche ich jede Mutation auf Positionsebene mit AML- und anderen TCGA-Krebskohorten, um die ursächlichen genomischen Mutationen zu bewerten. Ich fand potenzielle Treiber-Stopgain-Mutationen in Genen, die an der Chromosomentrennung während der Mitose beteiligt sind, und in einigen Tumorsuppressorgenen. Ich fand neue Stopgain-Mutationen in Krebsgenen (NIPBL und NF1). Da Fusionen zunehmend als therapeutische Ziele in der Onkologie anerkannt werden, untersuchten ich potenzielle Treiber-Fusionsereignisse, indem ich hochgradig zuverlässige und krebsbezogene Fusionsergebnisse im Fusionsergebnisse ohne Verschiebung des Leserasters auswerten. Als Ergebnis fand ich spezifische Genfusionsmuster. Kinasen, die durch Genfusionen aktiviert werden, stellen eine bedeutende Klasse von Onkogenen dar, die mit hämatopoetischen Malignomen assoziiert sind. ich identifizierten mehrere neue und nicht beschriebene Fusionen, an denen Kinasen beteiligt sind, die möglicherweise eine Rolle bei der Leukämogenese spielen. ich entdeckten zuvor nicht gemeldete Fusionen mit bekannten krebsbezogenen Genen wie PIM3-RAC2 und PROK2-EIF4E3. Darüber hinaus können Ausreißer, wie Genexpressionsniveaus, potenzielle pathogene Ereignisse

lokalisieren. Daher haben ich durch die Kombination unserer AML-Kohorte mit einer gesunden Kontrollgruppe abweichende Genexpressionsniveaus als mögliche pathogene Ereignisse unter Verwendung der Deep-Learning-Methode bestimmt. Schließlich haben ich die Daten kombiniert und nach einem Vergleich zum Methylierungsmuster jedes Patienten gesucht. Insgesamt deckte die Analyse eine breite Landschaft potenzieller Treiber auf. In verschiedenen Datenschichten fand ich eine veränderte genomische und transkriptomische Signatur verschiedener GTPasen, von denen bekannt ist, dass sie an vielen Stadien der Tumorentstehung beteiligt sind. Unsere Methoden und Ergebnisse zeigen die Leistungsfähigkeit der Integration von Multiomics-Daten zur Untersuchung komplexer Treiber-Ereignisse bei AML und weisen auf zukünftige Forschungsrichtungen hin, die darauf abzielen, Lücken in Forschung und klinischen Anwendungen zu schließen. Darüber hinaus liefern ich in-vitro-Beweise für die antileukämische Kooperativität und epigenetische Aktivität zwischen DAC und ATRA. Ich führte eine differentielle Methylierungsanalyse zur CpG-Auflösung und über genomische und transponierbare Elementregionen hinweg durch, um die statistische Aussagekraft und Interpretierbarkeit der Ergebnisse zu verbessern. Ich habe gezeigt, dass ATRA als Einzelwirkstoff weder eine globale Demethylierung verursacht, noch die durch DAC vermittelte Demethylierung durch ATRA verbessert. Zusammenfassend lässt sich sagen, dass die Kombination von Multi-Omics-Profiling ein leistungsstarkes Werkzeug zur Untersuchung dysregulierter Muster bei AML ist. Darüber hinaus zeigt das Multi-Omics-Profiling von mutationsnegativem nkAML mehrere vielversprechende Treiber. Meine Ergebnisse erweitern nicht nur unser Verständnis der der Heterogenität von AML, sondern können auch unmittelbare Auswirkungen auf neue zielgerichtete Therapiestudien haben.

1 INTRODUCTION

1.1 Biology and classification of acute myeloid leukemia

The findings described in this dissertation are based on human and cell culture samples of Acute myeloid leukemia. It is instrumental for the reader to be introduced to the thesis focus's general characteristics, processes, and techniques.

Acute myeloid leukemia (AML) is a hematological malignancy represented by clonal expansion of abnormal and undifferentiated myeloid precursor cells. It is a highly heterogeneous disease, for which approximately 35–40% of patients younger than 60 years old can be cured and 5-15% of patients who are older than 60 years of age. However, older patients who cannot receive intensive chemotherapy or other treatment have a median survival of only 5 to 10 months. [1]. AML is a stem cell precursor malignancy starting from the myeloid lineage (red blood cells, platelets, and white blood cells). It is caused by infiltration of the myeloid precursor cells due to their clonal, abnormally, or poorly differentiation. [2]. In AML, the myeloid stem cells give rise to immature white blood cells called myeloblasts or myeloid blasts. However, they are abnormal and cannot undergo normal differentiation. Therefore, if many stem cells become abnormal, leukemia can occur. Leukemia cells can accumulate in the bone marrow and blood. That follows exhaustion and less room for healthy ones. That can follow infection, anemia, or easy bleeding as an additional effect [3]. Several factors can increase the risk of AML development. AML is observed more in elderly patients. However, it can emerge at all ages. The risk of AML has been connected to tobacco exposure and other causes. Genetic predispositions and other concurrent diseases increase AML development, such as Down syndrome, some anemias, and others. [2]. Previous cancer disease and treatment can contribute to AML progression. It is a rare but often deadly complication of myeloproliferative neoplasms (MPN). Patients who have received chemotherapy for other cancer types, such as ovarian cancer and lymphoma, have a higher risk of developing AML [4]. Based on recurrent genetic abnormalities, including chromosomal translocations, AML is classified into six categories depicted in (Table 1) [5]

AML could arise after a myeloproliferative neoplasm (MPN) complication that usually occurs years after the diagnosis. Post-MPN AML comes from the traditional French-American-British (FAB) classifications of AML. Subtypes from M0 to M5 begin in immature forms of white blood cells. M6 AML forms in very immature cells, and M7 AML forms from immature platelet cells [4].

AML and related neoplasms	AML and related neoplasms (cont'd)
AML with recurrent genetic abnormalities	Acute myelomonocytic leukemia
AML with t(8;21)(q22;q22.1); RUNX1-RUNX1T1	Acute monoblastic/monocytic leukemia
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBFB-MYH11	Pure erythroid leukemia
Acute promyelocytic leukemia with PML-RARA	Acute megakaryoblastic leukemia
AML with t(9;11)(p21.3;q23.3); MLLT3-KMT2A	Acute basophilic leukemia
AML with t(6;9)(p23;q34.1); DEK-NUP214	Acute panmyelosis with myelofibrosis
AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2,MECOM(EVI1)	Myeloid sarcoma
AML (megakaryoblastic) with t(1;22)(p13.3;q13.3); RBM15-MKL1	Myeloid proliferations related to Down syndrome
Provisional entity: AML with BCR-ABL1	Transient abnormal myelopoiesis
AML with mutated NPM1	Myeloid leukemia associated with Down syndrome
AML with biallelic mutations of CEBPA	Blastic plasmacytoid dendritic cell neoplasm
Provisional entity: AML with mutated RUNX1	Acute leukemias of ambiguous lineage
AML with myelodysplasia-related changes	Acute undifferentiated leukemia
Therapy-related myeloid neoplasms	MPAL with t(9;22)(q34.1;q11.2); BCR-ABL1
AML, NOS	MPAL with t(v;11q23.3); KMT2A rearranged
AML with minimal differentiation	MPAL, B/myeloid, NOS
AML without maturation	MPAL, T/myeloid, NOS
AML with maturation	

Table 1: WHO classification disease categories. Reprinted with permission from [5]

1.2 Driver mutations in AML

Hematopoiesis is a highly regulated process in which a primitive pluripotent stem cell generates mature blood elements after an ordered sequence of maturation and proliferation [6]. Understanding the pathophysiologic functions of key genes that disrupt hematopoiesis and exert leukemogenic potential is essential for findings for AML [7]. The next-generation sequencing technology allows us to study the heterogeneity and genomic complexity of AML, which is proved to be based on the presence of collaborating mutations within operating categories such as epigenetic modulators, cell signaling and hematopoietic transcription factors[8]. Gene mutations in AML were generally grouped into class I mutation, which correlated with activation of intracellular signals, and class II mutation, which blocks differentiation by dysregulation of specific transcription factors [9]. Although AML carrying FLT3 mutation is not part of the WHO classification as a separate entity, it is the most typically mutated gene in AML (~30% of AML), and carriers have an unfavorable prognosis. Furthermore, the internal tandem duplication of this gene (FLT3-ITD) mutations is a negative prognostic marker since they result in a constitutively active FLT3(transmembrane tyrosine kinase). Constitutively active FLT3 induces several carcinogenic pathways such as Ras/Raf/MEK/ERK and PI3K/AKT signaling. Besides, it is known to phosphorylate STAT5 through a direct mechanism independent of JAK2, which in turn contributes to the proliferation of leukemia cells [10].
Several critical signaling pathways have been identified as frequently genetically altered in AML. Their representation of individual and co-occurring manners could suggest targeted and combination therapy opportunities.

1.2.1 Ras pathway

Many cases of AMLs have mutations in NRAS, KRAS, or genes that are part or connected to Ras signaling, and therefore Ras signaling is an appealing target for AML [11]. Ras small GTPases act as molecular switches that can bind and hydrolyze guanosine triphosphate (GTP) to inactive guanine diphosphate (GDP)-bound states. They play crucial roles in different processes such as cell polarization, cell migration, membrane trafficking, cell growth, and cell differentiation [12]. Interestingly, the mutated position varies among RAS proteins in cancer types, but the cause remains unclear. Also, the mutation rates at each codon could differ between the Ras proteins [13]. While mutations of KRAS occur at codon 12, NRAS is commonly mutated at codon 61. That results in G12D and G13D substitution for KRAS and Q61R replacement for NRAS protein [14].

1.2.2 Protein phosphatase

Some tyrosine-protein phosphatases, such as PTPN11, are an essential regulator of RAS signaling and are frequently mutated in patients with AML. PTPN11 mutation is more commonly associated with the acute myelomonocytic leukemia subtype. Mutation in this gene most commonly co-occurred with NPM1 mutations and FLT3 internal tandem duplications, and its presence is associated with poor outcomes across myeloid malignancies [15][16].

1.2.3 Ubiquitin pathway

Post-translational modifications are generally enzymatic modifications of proteins following protein biosynthesis. Protein ubiquitination is the significant post-translational modification in charge of protein biological process alteration or protein degradation by the 26S proteasome. Several studies have described alterations in the ubiquitin-proteasome system (UPS) in Hematological malignancies [17][18][19].

1.2.4 Nuclear-cytoplasmic shuttling phosphoprotein

The transport across the nuclear membrane is a highly regulated process vital for cell function and survival. This transport is highly moderated by importins or exportins, depending on the path of protein transport. These proteins recreate a functional part in tumorigenesis.[20]. The nucleophosmin (NPM1) gene is predominant nucleolar localization and is the most commonly mutated gene in adult acute myeloid leukemia[21].

1.2.5 Transcription factors

Leukemic transformation is based on mutations or operational dysregulation of many transcription factors [22]. The leading role of CEBPA is in the development of granulocytes during the process of hematopoiesis. [23]. Mouse studies show that CEBPA depletion intercepts the growth from common myeloid progenitors to granulocyte monocyte progenitors, and that causes the expansion of myeloid blasts in the bone marrow[24]. Studies show AML patients with CEBPA mutations show better relapse and overall survival. However, the favorable prognosis is reversed when a patient has CEBPA mutations with the coexistence of FLT3-ITD mutations. Mutations in another transcription factor, RUNX1, can have pathological and prognostic implications. However, the frequency of RUNX1 mutation is relatively low (less than 17 %), making it hard to specify its influence on clinical outcomes [25].

1.2.6 DNA hydroxymethylation genes

IDH1 and IDH2 are isocitrate dehydrogenases that are essential to cellular metabolism. They are enzymes that transform isocitrate to ∝-ketoglutarate. Mutations in IDH1/2 can produce a gain-of-function mutation, resulting in the accumulation of an oncometabolite, the D-2-hydroxyglutarate (D-2HG)[26]. Overproduction of D-2HG interferes not only with cellular metabolism and epigenetic regulation but contributes to oncogenesis as well. Indeed, high levels of D-2HG inhibit ∝-ketoglutarate-dependent dioxygenases, including histone and DNA demethylases. That leads to histone and DNA hypermethylation and, finally, a block in cell differentiation[27]. In addition, a mutation in epigenetic regulatory enzymes such as DNA methyltransferases, histone methyltransferases, and histone deacetylases is involved in AML development and progression. TET2 is one of the three TET (Ten-Eleven Translocation) family proteins, which are evolutionarily conserved dioxygenases that catalyze DNA demethylation. TET2 protein is involved in the epigenetic regulation of myelopoeisis [9].

1.2.7 DNA methylation genes

TET2 ofter co-occur with DNMT3A (DNA methylases) despite their epistasis in the methylation-hydroxymethylation pathway. Mouse models demonstrated the contributions of TET2 and DNMT3A mutations in promoting and inhibiting hematopoietic stem cells (HSCs) differentiation [28]. Mechanistically DNMT3A itself is closely correlated with the biological characteristics of malignant tumors. For example, R882 mutant proteins interact with polycomb repressive complex 1 (PRC1), allowing him to block the differentiation of hematopoietic stem cells by down-regulating differentiation-associated genes [29].

1.2.8 Histone methylation genes

Mutation in histone methylation genes has been observed in AML patient blasts. The most commonly mutated genes here are EZH2 and ASXL1. EZH2 acts as a repressor of lineage inappropriate genes and maintains cell integrity and genomic stability. Therefore it is demonstrated as a potential therapeutic target in AML [30]. Furthermore, ASXL1 mutations are found in AML and are primarily associated with signs of aggressiveness and poor clinical outcome [31].

1.2.9 Transcriptional corepressor

The most somatic mutations in the transcriptional corepressor genes could be found in BCORL1, located on the X-chromosome. Several known nonsense, splice site, and frameshift mutations were predicted to result in truncation and inactivation of its protein, suggesting that BCORL1 is a tumor suppressor gene. In AML, the BCOR gene is targeted by translocations and nonsense/splice site mutations. In addition, some studies demonstrated that the BCOR-mutated cases also carry mutations of the DNMT3A gene, suggesting that these two mutations may act cooperatively to induce AML, possibly through altering the epigenetic mechanisms [32].

1.2.10 Cohesin complex

The cohesin complex compiled of four subunits, including SMC1A, SMC3, RAD21, and either STAG1 or STAG2. That complex form a ring-shaped structure and intercede cohesion between replicated sister chromatids and segregation during mitosis and meiosis. Mutations in this gene complex have been reported in myeloid neoplasms and AML. However, patients carrying cohesin gene mutations had favorable impacts on overall survival (OS) [33].

1.2.11 Tumor suppressors

Mutations in suppressor genes usually could lead to cell proliferation, suppressing apoptosis or differentiation, or activating invasion and metastasis. The most prominent mutations in hematological malignancies are TP53 and WT1. TP53 is a key tumor suppressor gene that plays an essential role in preserving genomic balance, apoptotic pathways, metabolism functions, and DNA repair. In most cases, TP53 mutations are associated with complex karyotype, chemoresistance, and poor outcomes [34]. WT1

has long been implicated in acute myeloid leukemia. WT1 is often found either overexpressed or mutated and has given rise to the assumption that it may act depending on the context as a tumor suppressor or oncogene. [11].

1.3 DNA methylation epigenetics changes

The genetic sequence independently does not depict the complete picture of gene expression or cellular operation. Epigenetic machinery can affect gene activity without altering the DNA sequence. DNA methylation is a conversion of DNA that occurs primarily in the CpG dinucleotide region but is also observed in the non-cpg areas. It is specified by adding a methyl group to the cytosine. That and other epigenetic modifications modulate chromatin structure and gene expression. In addition, DNA methylation is critical for controlling gene regulation during embryogenesis, as well as cancer progression [35]. Therefore methylation is associated with both normal developmental processes and the changes that can be observed in oncogenesis that cause activation of pathological processes such as gene silencing of tumor suppressors or DNA repair genes.

1.3.1 Global Measures for DNA methylation measurement by microarray

The advancements in microarray and sequencing technologies create possible genome-wide profiling at a single-nucleotide level. DNA methylation analysis contains probes representing defined genomic regions. Methylation array protocols involve DNA bisulfite conversion followed by the converted DNA fragments' hybridization into arrays containing probes that differentiate between methylated and unmethylated Cs. Each probe's signal intensity sample is measured using labeled nucleotides [36]. These intensities indicate if a position has methylated, unmethylated, or hemimethylated cytosines and are utilized to quantify DNA methylation levels as β-values (or M-values). Beta-value and M-value statistics have been employed to calculate methylation levels. Beta values (β) estimate methylation level utilizing the proportion of intenseness between methylated and unmethylated alleles. β are between 0 and 1, where 0 represents unmethylated and 1 fully methylated Cs. The M-value has also been used for methylation microarray analysis. Positive M-values indicate methylation, while negative M-values indicate the contrary [37].

Wildly used and available datasets implement the Illumina Infinium HumanMethyla-tion450 BeadChip (450K), which enfolds 482,421 CpGs in the human genome. The 450K also contains 65 Single-Nucleotide Variation (SNV) probes and quality control probes [38]. The Illumina Infinium Methylation EPIC BeadChip (EPIC) is the more current platform, covers 853,307 CpGs, and significantly expands the coverage in enhancers and other regions in the genome. The 450K and the EPIC arrays contain two probe designs - type I and type II probes harness correspondently two and a single-probe design. Several methods have been developed to normalize the beta values derived from the two types of probes and make beta values from both types of probes more comparable [39].

1.3.2 Epigenetics modifications in AML

AML has one of the lowest mutational burdens compared with other cancer types, wherein in some cases, there are some samples with no apparent drivers. One plausible explanation for the visible lack of driver alterations in AML is that these could not be coded within the DNA sequence itself but act on genome organization and regulation. These events are controlled by epigenetic modifications of the genome, including DNA methylation, histone modification, and various RNA-mediated processes [37]. DNA methylation occurs almost exclusively at Cs that are followed by Gs. So cytosines are followed by guanines, called a CpG dinucleotide. The genomic methylation process is executed by DNA methyltransferases (DNMTs): DNMT1, DNMT3A, and DNMT3B. They have different functions in maintaining normal methylation DNA status. DNMT1 is needed for the maintenance of methylation. During repli-

cation, DNMT1 fixes the corresponding methylation on the daughter strand following the pattern of the parental DNA. Methylation is laid down and copied mitotically with the help of DNMT3A and DNMT3B, which can recognize hemimethylated DNA. DNMT3A and DNMT3B are de novo methyltransferases, which are in charge of setting up genomic imprints during germ cell development [40]. DNA could also be demethylated by a passive and active process. Passive DNA demethylation dilutes DNA methylation with every cell division when DNMT1 is not expressed or not in the nucleus. TET proteins TET1, TET2, and TET3 are the leading players in active DNA demethylation. The TET enzymes removed the methylated DNA base by base excision repair and replaced it with cytosine. It is known that DNA methylation is a stable epigenetic mark essential for transcriptional regulation and genome integrity. Understanding how DNA methylation orchestrates the cell functions requires attention to the methylation distribution across the genome. More than half of the genes include short (roughly 1 kb) CpG-rich regions known as CpG islands (CGIs). Many CGIs are associated with gene promoters [41]. In the normal cells, CpG islands are usually unmethylated, whereas intergenic regions and repetitive elements tend to be methylated. DNA methylation is there to maintain genomic integrity. Mutation of DNMT genes causes an acute phenotype characterized by global DNA hypomethylation and genome instability [42]. Furthermore, inappropriate methylation at these intergenic intervals could result in abnormal karyotypes, such as deletions, duplications, insertions, or translocations. It is also known that methylation plays some role in expression gene silencing. However, it is not a traditional promoter that is silenced but rather a cryptic promoter or cryptic splice site. The mechanisms of how methylation changes in cancer occur are not fully understood. However, genome-scale methylation mapping techniques allow us to study the proportion and occurrence of methylation of each element on a single CpG site. Gene promoters usually hold CpG-rich sequences (called CpG islands) in 5´ regions. In normal cells, CpG islands are generally unmethylated. Nevertheless, cancer cells frequently demonstrate abnormal methylation. Moreover, some studies have demonstrated that DNA methylation of CpG islands and epigenetic modification of some histones leads to gene silencing of tumor-suppressor genes. [43]. It is known that global hypomethylation of DNA in cancer was most closely associated with repeated DNA elements and non-protein-coding regions. DNA hypomethylation contributes to the accumulation of genetic mutations and tumor progression [44]. Furthermore, hypomethylation of some retrotransposons is considered a hallmark of malignant transformations. Repetitive elements are silenced through DNA methylation to prevent transposition and avoid transcriptional interference from strong promoters. The methylation of repeats may also prevent illegitimate recombination. Their hypomethylation could result in the reactivation of retroelements and subsequent genomic instability 1 .

Different genomic regions, like late-replicating or transposable elements, undergo DNA hypomethylation preferentially. It is also known that DNA methylation impacts genes implicated in cell-cycle checkpoints and other cancer cellular signaling [45]. Numerous genomic studies aim to establish which genes are abnormally controlled by DNA methylation in disease. However, we do not fully understand how disease-specific methylation changes affect gene expression. CpG islands have been evolutionarily conserved to enable gene expression by controlling the chromatin architecture and transcription factor direction. Methylation of CpG islands can harm transcription factor binding and stably silence gene expression [46]. Methylation of CpG islands in the promoter region of tumor suppressor genes results in gene silencing. Contrarily, hypomethylation in the promoter region of the oncogene could reactivate transcription. DNA methylation in the bodies of genes is suggested to be involved in transcription elongation and alternative splicing [41].

Pervasive modifications in the patterns of DNA methylation characterize hematological malignancies. These changes contain a global hypomethylation and the specific hypermethylation of associated tumor suppressor gene promoters. That process is associated with transcriptional gene silencing and determines distinct prognostic properties [47]. Loss of function in key DNA methylation-related enzymes is widely

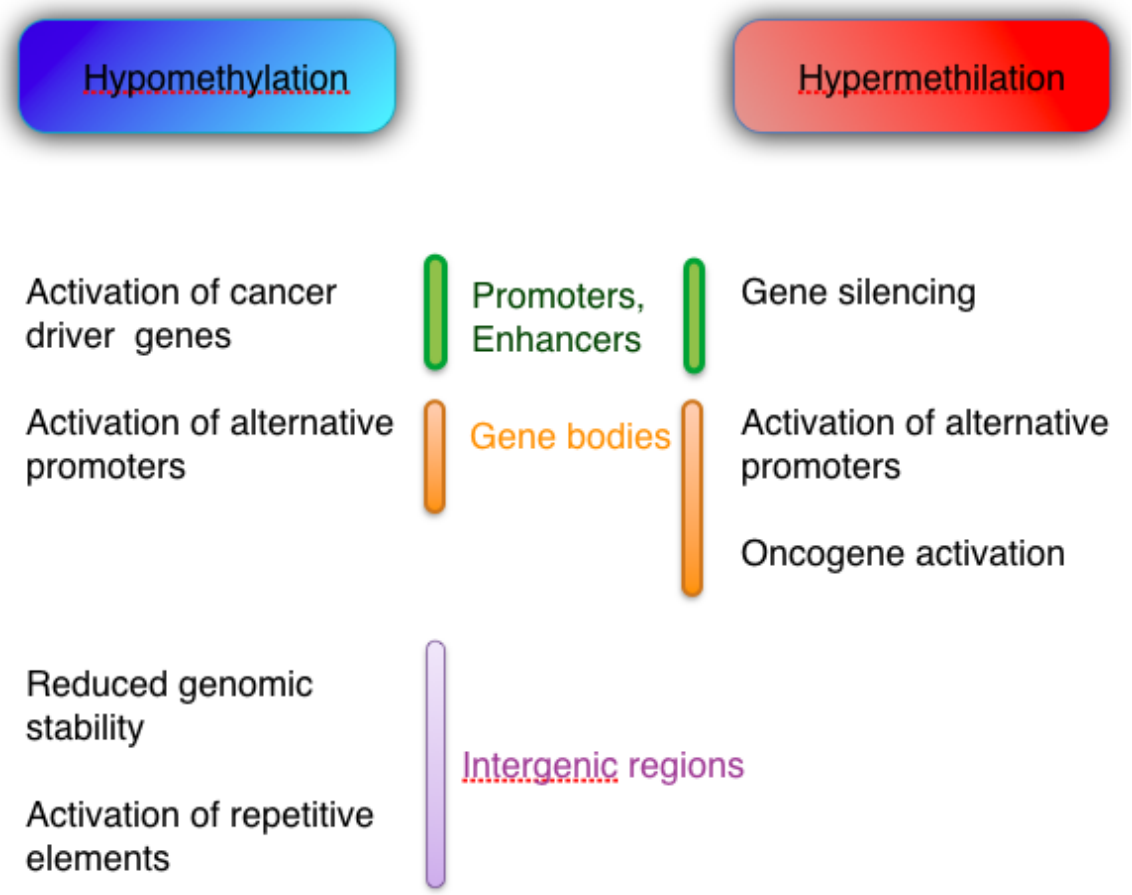


Figure 1: Overview of the role of the DNA methylation. The the figure is modified from coursera course "Epigenetic Control of Gene Expression".

observed in hematological malignancies, which results in lineage-specific aberrant DNA methylation patterns. Interestingly, TET loss or DNMT loss can lead to hypermethylation and hypomethylation of the promoters of specific target genes [48]. Although the alterations of DNA methylation in AML development are not fully comprehended, several studies have documented a substantial correlation between DNA methylation signature and patient clinical outcomes [48, 49]. For example, patients with high Cytosine methylation levels demonstrated a lower overall survival rate than the low Cytosine methylation groups [50]. Another study identified three significantly hypermethylated candidates, CDKN2A, CDKN2B, and ID4, associated with the increased risk of leukemia onset [49]. Furthermore, global DNA methylation patterns can often co-occur with the genetic portrayal of AML risk groups. One example is the CEBPA-double mutation which displayed a distinct hypermethylated profile and had a better prognosis. These studies raise the perspective that DNA methylation carries possibilities in clinical applications through diagnosis and therapeutic determination. However, there is slow progress in discovering DNA methylation biomarkers due to the lack of reproducibility in different patient cohorts. Therefore, a robust method for DNA methylation biomarkers should be contrived in the early diagnosis of personalized medicine for AML patients [51].

1.4 Fusion genes

Gene fusions have been recognized as essential drivers of cancers. Gene fusions are hybrid genes formed when two previously independent genes are joined so that they are transcribed and translated as a single molecule. The fusion can be formed from chromosomal rearrangements like translocations, deletions, or inversions. However, it could also result from transcription read-through of neighboring genes or the trans- and cis-splicing of pre-mRNAs Fig.2 [52]. Trans-splicing is when two separate pre-mRNAs are spliced together, giving rise to a single mRNA molecule. Read-through is produced after the transcription of two neighboring genes. Some literature has suggested that normal karyotype AMLs are characterized by chimeras, mainly coming from adjacent genes located on the same chromosome. That can occur through transcription-induced chimeras, which are irrelevant to chromosomal translocations. The fusion transcripts created between adjacent genes emphasize the possibility that specific such fusions could be affected in the oncological approach in AML [53] Oncogenic fusion genes and proteins produced via chromosomal rearrangements play essential roles in hematologic malignancies. These chimera proteins alter the protein's function, which could interfere with fundamental cellular properties, such as self-renewal, differentiation, and proliferation, initiating the process of leukemic transformation [54].

Especially in acute leukemia, fusion genes (FGs) are significant molecular abnormalities that play crucial tumorigenesis factors. Based on their essential role in leukemogenesis, FGs have been utilized as molecular markers for leukemia diagnosis and treatment. Therefore dozens of FGs are incorporated in the WHO classification of the hematopoietic and lymphoid tissues. Common FGs are presented in around 42% of AMLs. [55]. The fast development of sequencing technology and the decreased costs have made identifying fusion genes from either whole-genome sequencing (WGS) or RNA sequencing. That allows us to reach fusion genes produced after chromosomal rearrangements and generated through cis- or trans-splicing. The nature of equal fusion transcripts could vary in cancer and normal subjects. Two examples are fusions JAZF1-JJAZ1, and PAX3-FOXO1 are formed by chromosomal rearrangements in cancer cells but by trans-splicing in normal cells [56]. From one site, RNA-Seq sequencing only gives information about 2% of the genome that is transcribed and matured mRNA, but from another, RNA-Seq would detect intergenically spliced fusions that exclusively happen at the RNA level. That allows the detection of multiple alternative splice variants coming from the same fusion genes. The limitations of RNA-Seq are missing the fusion events involving non-transcribed or very low transcribed events [57]. Many dedicated tools for fusion detection have detected hundreds of FGs with clinical relevance [58].

1.4.1 Transcription factor gene fusions

Chromosomal translocations affecting transcription factors (TFs) are often found in leukemias, and some of them have been used as genetic markers since they have specific prognostic value. There were detected 521 transcriptions of factor-associated FGs in 149 distinct fusions. Genomic rearrangements involving TFs can produce fusion proteins that affect their activity in both directions. The newly formed chimera could have either enhanced, weakened, or lost TF activity [59]. The t(8;21) translocation is one of the most frequent chromosome abnormalities in acute myeloid leukemia, which includes a fusion of nuclear transcription factor AML1 and oncogene MTG8 [60]. The inv(16)(p13q22) and t(16;16)(p13;q22) in acute myeloid leukemia results in different CBFB-MYH11 fusion transcripts. These fusions together account for ∼25% of AML [61]. Recent studies have revealed that t(8;21) and inv(16) have typical pathogenesis and are designated the core-binding factor leukemias. These translocations follow the gene alteration that encodes the AML1/CBFβ core-binding factor transcription complex, a crucial regulator of normal hematopoiesis. Mice lacking either AML1 or CBFβ gene die in utero and fail to develop definitive hematopoiesis, proof that both genes' normal functions are critical to hematopoietic development [62].

The common translocation involved RARα gene is t(15;17) translocation, giving rise to the PML/RARα fusion protein. RARα is a nuclear hormone receptor that transmits the transcriptional response to retinoic acid (RA). Normally, physiologic concentrations of RA induce a conformational change in the RARs, driving the recruitment of transcriptional coactivator molecules and leading to the activation of gene transcription [61].

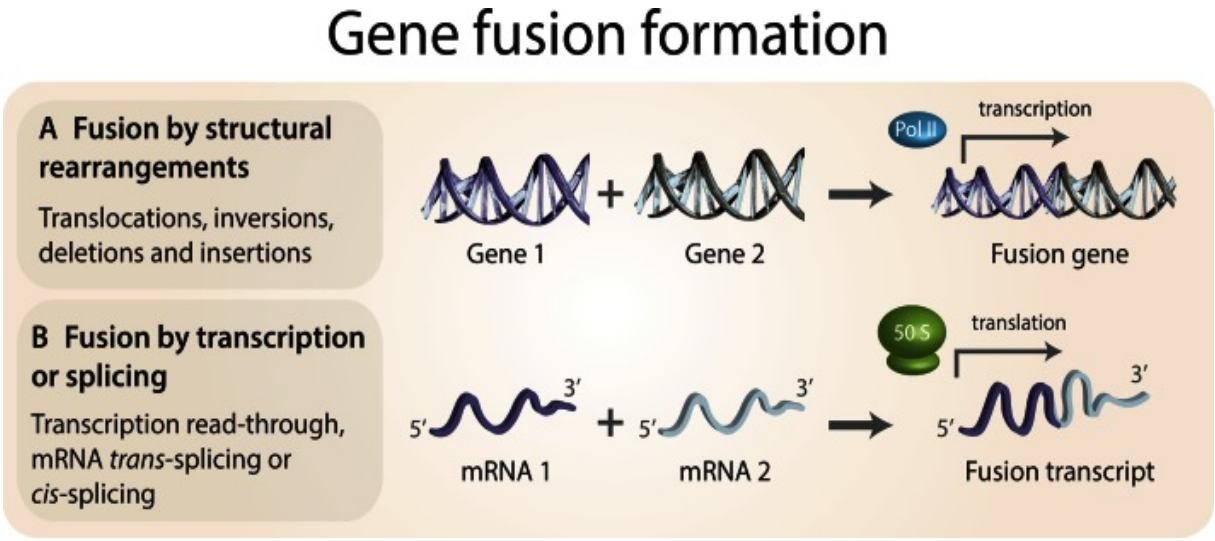


Figure 2: Formations of gene fusions. Figured reused with permission from Latysheva and Babu, 2016 , 2016

Fusions involving the MLL gene are common in patients with acute leukemia, ALL, and patients with t-AML. Rearrangements of the MLL gene involve multiple partners implicated in decision-making and therapy [63]. NUP98, TEL, and CBP/p300 gene rearrangements occur in less than 3% of adult AML [64]. The leukemia-associated fusions harboring functional similarities indicate that they may act through shared mechanisms of transcriptional dysregulation. The fusion proteins developed by these translocations often possess a DNA-binding domain, parts that target the proteins to specific DNA sites. They could also involve in chromatin remodeling and impact chromatin structure. Another well-described fusion is 11q23 Leukemias involving MLL, a huge gene spanning 90 kb containing 36 exons. The N-terminal portion of MLL holds an 'AT hooks' region that is described to bind to the AT-rich DNA segments. Contrasting the sequence-specific DNA-binding domains seen in other fusions, the AT-hooks recognize specific DNA structures rather than specific sequences. There are more than 40 leukemia-associated translocations connected to the MLL gene. The most common fusion partners are AF9, AF4, ENL, ELL, AF6, and CBP [65]. YST3 encodes a large multidomain protein that contains, besides others, a histone acetyltransferase catalytic domain. Several studies have demonstrated that MYST3 is a co-activator of diverse transcription factors related to hematopoiesiss [66]. CREBBP is a tumor suppressor gene important for hematopoiesis [67]. Chromosomal rearrangements involving the lysine methyltransferase 2A (KMT2A) gene are typical alterations in pediatric AML with an incidence of up to 26%, and carrier KMT2A gene rearrangements had a dismal prognosis [68].

1.4.2 Epigenetic gene fusions

Not only could mutation in genes control the chromatin modifier and epigenetic programs drive human cancer, but also fusions involving epigenetic regulators. Most of them, around 79%, are histone methyltransferases- or histone acetyltransferases-related. The World Health Organization (WHO) classification of tumors of lymphoid tissues had adjoined t(9;11) to a distinct category of AML under the definition 'AML with t(9;11)(p21.3;q23.3); KMT2A-MLLT3'. KMT2A is also known as ALL-1, MLL1, or HRX [69]. KDM5A is a lysine-specific demethylase that regulates cell proliferation. When NUP98-KDM5A fusion is expressed in HSPCs drives proliferation and transforms differentiation while navigating an erythroid phenotype in vitro [70]. SWI/SNF chromatin remodeling complexes are also found to be

part of the fusion partners. BRD7 fusions alter the myogenic program and maintain the proliferative state while blocking terminal differentiation [71].

1.5 Drug targeting for acute myeloid leukemia

AML is aggressive disease that grows quickly. Therefore the treatment should begin direct after a diagnosis has been confirmed. Like other cancer, the therapy starts with chemotherapy. In case of chemotherapy resistance, an alternative option would be a bone marrow or stem cell transplant. Before transplantation, the patient will require high-dose chemotherapy and additional radiotherapy to eliminate the cells in their bone marrow. Standard care of AML for more than half of a century was treatment with cytotoxic chemotherapy. As described above, AML is a malignancy stemming from hematopoietic stem cells, portrayed by different driver gene mutations, epigenetic modifications affecting DNA methylation and chromatin structure, and microRNA dysregulation. The advance of next-generation sequencing technologies has paved the path for gene mutation-targeted and epigenetic therapies [72]. Targeted therapy can be divided into three groups: 1: agents that target oncogenic effectors of recurrent AML-associated mutations, including FLT3 and IDH inhibitors. 2: cytotoxic agents therapy, such as ADCs, and group 3: agents that act on disrupting key cell metabolic pathways. These include epigenetic modifiers and agents that directly target apoptosis [73].

1.5.1 FLT3 tyrosine kinase inhibitors

The FMS-like tyrosine kinase 3 (FLT3) gene is mutated in more than 30% of AML cases, and the FLT3 receptor is overexpressed in most acute leukemias. Alterations happen by internal tandem duplications (FLT3-ITD) or a point mutation primarily affecting the tyrosine kinase domain. FLT3-ITD is associated with a low-grade forecast of risk of relapse and overall survival (OS) [74]. Therefore, FLT3 tyrosine kinase inhibitors have been applied in recent years, such as sorafenib, lestaurtinib, gilteritinib, and others [75].

1.5.2 Antibody–drug conjugates (ADCs)

ADCs obtain a monoclonal antibody append to a cytotoxic payload by a cleavable linker. That permits highly cytotoxic mechanisms to be straight provided to leukemia cells leading to cell death and preventing extreme off-tumor toxicity. Multiple generations of ADCs have developed with enhanced stability and internalization kinetics. However, ADC development requires attention to stability, and the target antigen should be sufficiently expressed on the leukemia cell surface [76]. Even though several ADCs have been developed and tested clinically for leukemia, until now, they have accomplished only limited success [77].

1.5.3 Epigenetic Therapies

AML is a highly heterogeneous cancer with a low mutational burden and epigenetic dynamic aberrations. As described, mutations frequently hit genes, encoding DNA methylation modifications (DNMT3A, TET2, IDH1, and IDH2) and histone tails modifications (EZH2 ASXL1, and others). In addition, some of the fusion proteins could redirect the specificity of epigenetic regulators. Since AML clonal expansion was recently linked to global alterations in DNA methylation patterns, supporting the conception that epigenetic heterogeneity better explains leukemia development than genetic background [78]. If AML is frequently initially susceptible to chemotherapy, relapse possibilities stay recurring due to chemotherapy-resistant clones [79]. Therefore, more sufficient combinatorial or epigenetic treatments are needed for AML patients. The considerable part of epigenetic dysregulation in advancing leukemias pushes efforts toward developing epigenetic drugs for treating AML. The US FDA has approved agents in three

epigenetic categories (DNMT, HDAC, and EZH2 inhibitors) to treat hematological malignancies [80].

DNMT blockade is observed to be a successful procedure for the prevention of aberrant DNA hypermethylation. DNMT inhibitors could reactivate the aberrantly methylated tumor suppressor genes and cause cancer cells to reprogram into proliferation arrest and cell death. Generally, there are nucleoside and non-nucleoside inhibitors of DNMT, where both groups aim to reactivate the aberrantly methylated TSG. Nucleoside analogs comprise a modified cytosine ring and, during replication, can get incorporated into the newly synthesized DNA strands, leading to DNA methylation inhibition [81]. 5-Azacytidine was first synthesized almost 40 years ago. It was demonstrated to be an effective chemotherapeutic agent for acute myelogenous leukemia. AZA is incorporated into RNA and directs to abnormal ribosome assembly. DAC, on the other hand, is incorporated only into DNA. Low-dose DAC reduces DNA demethylation by impeding DNA methyltransferase and reactivating tumor suppressor genes [82]. Cytosine analogs perform their anti-cancer effects at low doses via targeting DNMT-dependent DNA methylation, while at high doses block the DNA synthesis and inhibit cell proliferation via incorporation into DNA [83] 4 . Furthermore, Preclinical studies have demonstrated that DAC is more effective than AZA in vivo [84]. DAC treatment is also more effective in patients with a poorer prognosis, suggesting that DAC resulted in a better outcome for patients with relapsed or refractory acute myeloid leukemia [85]. The activity of DAC combined with All-Trans Retinoic Acid (ATRA) resulted in an improved response rate and survival compared to DAC without ATRA [86]. ATRA is an intermediate vitamin A product and was first used in therapy for acute promyelocytic leukemia by the end of the nineteen-eighties. ATRA operates via signaling pathways separate from the classical nuclear receptor activation mechanism. All-trans-retinoic acid modulates the actions of molecules implicated in signal transduction in a cell-type-specific manner. ATRA has been shown to activate different multiple kinase signaling pathways via transcription factors at the end of these pathways [87] Fig 3. Complex interactions between genomic and non-genomic pathways control the activity of ATRA in the tumor. ATRA triggers genomic effects mediated through nuclear receptors RA receptors (RARs) that function as ligand-activated transcription factors. The retinoid–receptor complex modulates the transcription of differentiation-inducing sets of genes [88]. The precise pathway involved in the non-genomic effects of ATRA is still under research. However, it is evident that these influences have significant results, and they have to be regarded when estimating the overall response in the case of combinatorial treatments [89] .

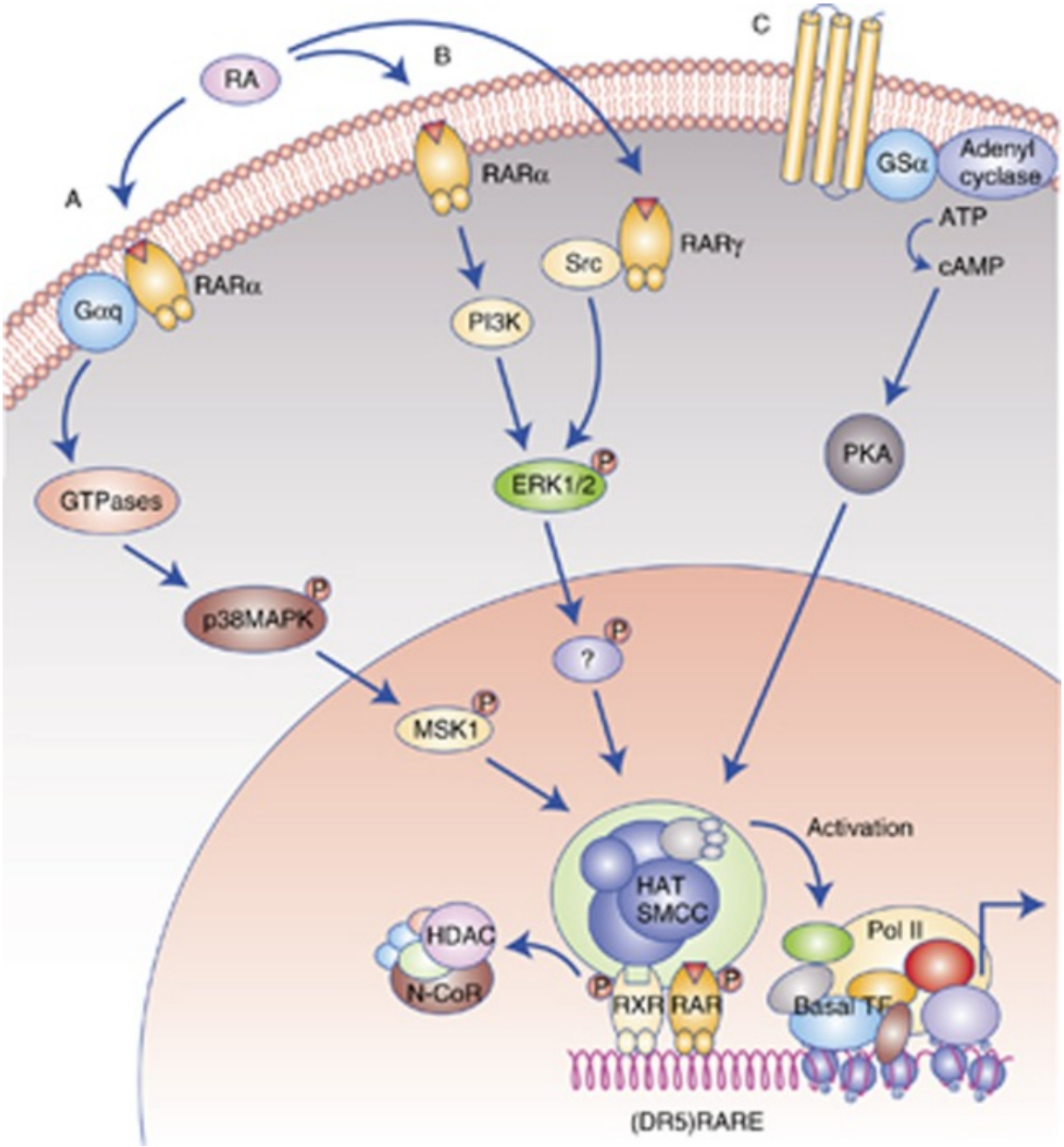


Figure 3: Non-genomic effects of ATRA. ATRA increase expression of RAR target genes and enhance differentiation. Figured reused with permission from [87]

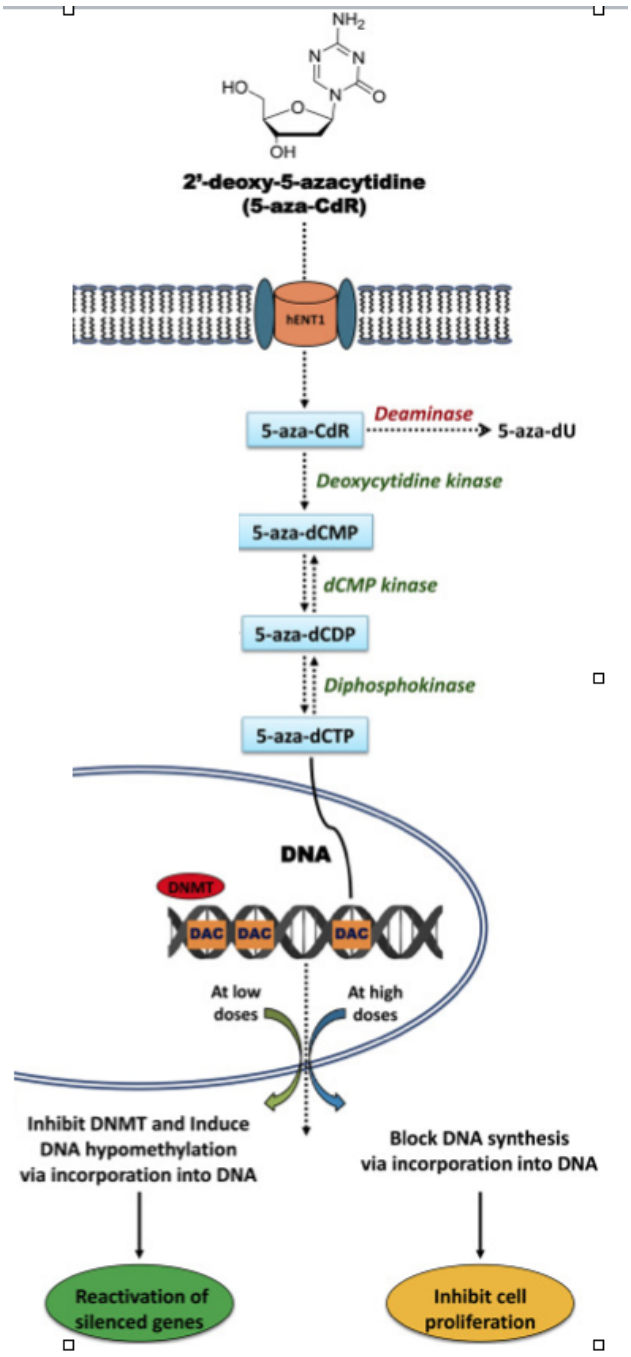


Figure 4: Mechanism of DAC. Decitabine-derived 5-aza-dCTP is incorporated into newly synthesized DNA during the replication process. Figured reused with permission from [83]

1.6 Overview of the undertaken approach

Advancements in understanding the alterations of acute myeloid leukemia (AML) have resulted in significant progress in diagnosis and provided opportunities for greater individualization of therapy. However, current approaches to genetic profiling fail to identify chromosomal abnormalities or leukemogenic driver mutations in approximately 20% of cases. Therefore It is still under progress mutations associated with distinct clinical phenotypes since the heterogeneity of the disease and high passenger rate. The reason is that AML and cancer are generally associated with the accumulation of mutations. However, only a tiny fraction of patients' recognized mutations is reliable for cancer progression. These designated driver events could help us predict clinical results for cancer carriers. One of the main difficulties in cancer study is the prioritization of alterations. The recurrence of a mutation stays one of the most dedicated markers of its driver status. However, the mutation process does not impact the genome equally. The mutability of driver mutations is not that high than that of passengers, so adapting mutation recurrence frequency is required. Even further, if one want to study these rare cases of AML where no common driver mutations are detected, one need a better approach to tackle that enormous challenge. Identifying disease cause can be anything that captures the biology of leukemia cells: from genomic mutations and rearrangements through epigenetic or gene expression profiles. Therefore in the main project, I went one step further into that part of personalized medicine and developed a pipeline to detect possible multi-omic driver events for each patient separately.

Since epigenetic machinery donated to pathogenesis and held prognostic relevance to AML, I also aimed to investigate the epigenetic mechanism of decitabine (DAC) and all-trans retinoic acid (ATRA) in cultured cells. Therefore, the project aimed to investigate the combinatorial methylation mechanism of decitabine (DAC) and all-trans retinoic acid (ATRA) in AML-cultured cells. I estimated the methylation changes corresponding to the control on the single cpg, promoter, gene body, and transposable elements levels.

In summary, the methods and output are presented and discussed in the next chapters from the following three projects: (i) Identifying recurrent mutations in AML carrying t(8;16)(p11;p13) rare abnormality, (ii) Prospective driver events in mutation-negative nkAML, (iii) The antileukemic activity of decitabine upon PML/RARA-negative AML blasts is supported by all-trans retinoic acid

2 METHODS

In this dissertation, I applied and developed several bioinformatics workflows for analyzing specific multi-omic analysis of normal karyotype no-driver AML and no-control whole-genome sequencing (WGS) of several AML and MNP samples. This chapter describes technical details involving the detection of driver alterations in AML and epigenetic drug effects. The thesis aims to find the AML driver alterations in all types all layer DNA mutation, gene fusions, expression alteration, and changes in epigenetic state. Therefore I built the workflow to prioritize the candidates for targeting, focusing on cancer-relevant mutation, highly expressed genes, and potential DNA methylation driver changes. Finally, I analyzed and compared the results with publicly available datasets. Scripts and the instructions for the reproduction of the analyses from the following chapters are internally available via the following path: /omics/groups/OE0436/data2/langova/hipo_K16R/

2.1 Patient and sample collection

The AML samples were collected from European diagnostic bone marrow aspirations after obtaining informed written consent. For a patients with acute myeloid leukaemia and t(8;16)(p11;p13) please relate to our paper [90]. In a cohort of ten patients t(8;16) AML, genomic DNA is available from bone marrow at baseline, and I performed whole-genome sequencing (WGS). The normal karyoteAML sample undergoes whole genome sequencing, and MPN samples whole exome sequencing. The study presented an analysis of omics data of AML and MPN samples. For AML, whole genome data were generated for 42 samples, transcriptomic data for 34 samples, and EPIC arrays for 41. For MPN, whole exome data were generated for 12 samples, transcriptomic data for 24 samples, and EPIC arrays for 26 5. The whole sample overview table can be found in the supplementary section. The samples are coming from different sites in Germany and Netherland. Seven of the patient underwent chemotherapy before sampling. As healthy controls, I used a buccal swap. However, in most of the samples, the DNA from the buccal swap was degraded or missing, leaving more than 95 % of samples without matching control.

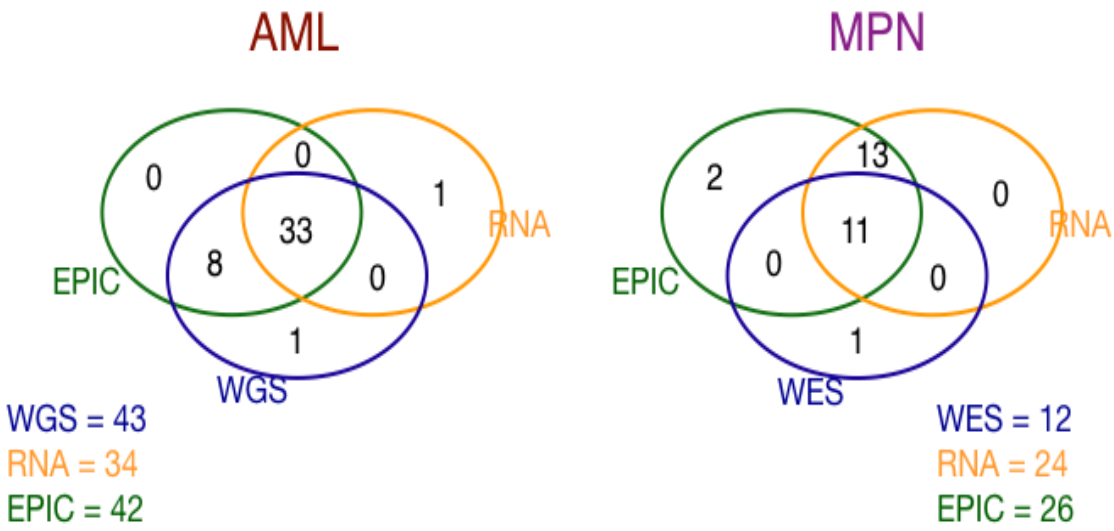


Figure 5: The number of samples ofAML and MPN patients and the overview overlap between omics data types. The number of samples representing all the layers (genome, transcriptome, and DNA methylation) is 33 in AML samples and 11 in MPNsamples.

2.2 Cell culture and drug treatment

Cell culture and drug treatment were performed by our collaborators in Freiburg. Cell lines U937 were cultured in RPMI 1640 medium. Afterward, the cells were treated in triplicates with daily pulses of 200 nM DAC for three consecutive days. After 48 hours, simultaneously with the third DAC dose, 1 µM ATRA was administered. For control, cells treated with the appropriate vehicles were used. All experiments were performed in triplicate for each condition. Samples' characteristics and experimental conditions are listed in Table 2

label	treated with	harvesting after
U937_72h_untreated_1	untreated	72 h
U937_72h_untreated_2	untreated	72 h
U937_72h_untreated_3	untreated	72 h
U937_120h_untreated_1	untreated	120 h
U937_120h_untreated_2	untreated	120 h
U937_120h_untreated_3	untreated	120 h
U937_72h_ATRA_1	ATRA after 48 h	72h
U937_72h_ATRA_2	ATRA after 48 h	72h
U937_72h_ATRA_3	ATRA after 48 h	72h
U937_72h_DAC_1	DAC after 0, 24 und 48 h	72h
U937_72h_DAC_2	DAC after 0, 24 und 48 h	72h
U937_72h_DAC_3	DAC after 0, 24 und 48 h	72h
U937_72h_DAC+ATRA_1	DAC after 0, 24 und 48 h; ATRA after 48 hours	72h
U937_72h_DAC+ATRA_2	DAC after 0, 24 und 48 h; ATRA after 48 hours	72h
U937_72h_DAC+ATRA_3	DAC after 0, 24 und 48 h; ATRA after 48 hours	72h
U937_120h_ATRA_1	ATRA after 48 hours	120h
U937_120h_ATRA_2	ATRA after 48 hours	120h
U937_120h_ATRA_3	ATRA after 48 hours	120h
U937_120h_DAC_1	DAC after 0, 24 und 48 h	120h
U937_120h_DAC_2	DAC after 0, 24 und 48 h	120h
U937_120h_DAC_13	DAC after 0, 24 und 48 h	120h
U937_120h_DAC+ATRA_1	DAC after 0, 24 und 48 h; ATRA after 48 hours	120h
U937_120h_DAC+ATRA_2	DAC after 0, 24 und 48 h; ATRA after 48 hours	120h
U937_120h_DAC+ATRA_3	DAC after 0, 24 und 48 h; ATRA after 48 hours	120h

Table 2: Table describing the conditions of sample treatment and harvesting

2.3 Priritisation of mutations

In collaboration with the Department of Internal Medicine V, Heidelberg University Hospital, Heidelberg, Germany, AML samples have been collected from patients carrying the rare translation between the 8th and 16th chromosomes. Primary tumors and matched controls from each patient were submitted to DKFZ Genomics and Proteomics Core Facility from our collaborators. In cancer mutation detection, it is essential to identify those genetic variants that occur exclusively in cancer tissue. For this goal, the non-malignant tissue of the same individual (i.e., matched control) is used to determine which variation tumor tissue-specific and driver candidates. Therefore, somatic variant calling pipelines are incorporated to call the cancer-specific somatic variants. However, not every cancer sample has matched control.

WGS no-control pipeline for SNV, Indel, and copy number variation is harmonized with our in-house workflow management systems, Roddy (https://roddy-documentation.readthedocs.io/en/latest/) and the One Touch Pipeline (OTP) [91]. The pipelines are optimized to detect the most significant and reliable alterations without requiring a matched control by deducting common variants (in some publicly available common variants databases) from a germline variant calling output. The method is further described in our paper [90].

2.3.1 Filtering the germline variants

The germline variant deduction calling pipeline uses three public databases, ExAC [92], dbSNP [93], and EVS (http://evs.gs.washington.edu/EVS/), and an in-house control dataset (containing 280 controls. I used ANNOVAR [94] to distinguish regional and functional types for each genomic mutation found on a precise chromosome position. Finally, the remaining variants are reported as a variant calling format

Table 1: Gene Regions Assessed by the TruSight Myeloid Sequencing Panel

Gene	Target Region (exon)	Gene	Target Region (exon)	Gene	Target Region (exon)	Gene	Target Region (exon)
ABL1	4–6	*DNMT3A*	full	*KDM6A*	full	*RAD21*	full
ASXL1	12	*ETV6/TEL*	full	*KIT*	2, 8–11, 13 + 17	*RUNX1*	full
ATRX	8–10 and 17–31	*EZH2*	full	*KRAS*	2 + 3	*SETBP1*	4 (partial)
BCOR	full	*FBXW7*	9 + 10 + 11	*MLL*	5–8	*SF3B1*	13–16
BCORL1	full	*FLT3*	14 + 15 + 20	*MPL*	10	*SMC1A*	2, 11, 16 + 17
BRAF	15	*GATA1*	2	*MYD88*	3–5	*SMC3*	10, 13, 19, 23, 25 + 28
CALR	9	*GATA2*	2–6	*NOTCH1*	26–28 + 34	*SRSF2*	1
CBL	8 + 9	*GNAS*	8 + 9	*NPM1*	12	*STAG2*	full
CBLB	9, 10	*HRAS*	2 + 3	*NRAS*	2 + 3	*TET2*	3–11
CBLC	9, 10	*IDH1*	4	*PDGFRA*	12, 14, 18	*TP53*	2–11
CDKN2A	full	*IDH2*	4	*PHF6*	full	*U2AF1*	2 + 6
CEBPA	full	*IKZF1*	full	*PTEN*	5 + 7	*WT1*	7 + 9
CSF3R	14–17	*JAK2*	12 + 14	*PTPN11*	3 + 13	*ZRSR2*	full
CUX1	full	*JAK3*	13				

Figure 6: Gene Regions Assessed by the TruSight Myeloid Sequencing Panel. Obtainaned from http://www.biosystems.com.ar/archivos/folletos/217/datasheet-trusight-myeloid.pdf

(VCF) file. After subtracting the common variants, the SNVs found in coding regions, called functional SNVs, were deducted to 250 - 300. Then the allele frequency was calculated by dividing the number of reads of the alternative allele by the total number of copies of all the alleles at that particular genetic position. The criteria for identifying a driver mutation is SNV with variant allele frequency > 0.3.

The exonic regions were defined and only considered in mine study. The next step implicated removing poorly-covered variants potentially driven by technical anomalies. Here, sufficient coverage is defined as bases with 9 or more reads at a given location. I also addressed complex indel events (SNVs in indel regions) to assure there is no deduplication of variants. For the no-driver AML project, I received samples that first went through The myeloid sequencing panel to exclude samples carrying myeloid driver 54 genes (tumor suppressor genes and oncogenic hotspots) in one test Figure 6 . Targeted genes contain clinical relevance genes, delivering a broad picture of the MDS and AML disease and their progression. The panel concentrates on ~141 kb of genomic range, consisting of 568 amplicons of ~250 bp in length created against the human hg19 reference genome. The oligo pool targets 15 entire genes (exons only) plus exonic hotspots of a further 39 genes, supplying nearly 100% coverage of all targeted regions. Samples carrying any hotspots were discarded from further analysis, and only no-driver samples were kept.

2.3.2 cancer related genes with prognostic value

I built an integrated analysis workflow to reveal key genes with prognostic value using our data's molecular profiles of gene mutation, methylation, and expression characteristics and utilize the data from publicly available cancer databases. Firstly, each SNV position and exact nucleotide exchange is compared to the Catalogue Of Somatic Mutations In Cancer (COSMIC) which includes almost 6 million coding mutations across 1.4 million tumor samples [95]. The Cancer Genome Atlas (TCGA) is another to

catalog discovering significant cancer-driving alterations in order to construct a broad "atlas" of cancer genomes. TCGA researchers have investigated extensive cohorts of over 30 human tumors via large-scale genome sequencing and integrated multidimensional investigations. The cBioPortal for Cancer Genomics (http://cbioportal.org) supplies help for the visualization of multifaceted cancer data. It could query by gene and show us how many patients independent of cancer type carried a mutation at a specific position. From here, I considered a mutation that occurs in more patients independent of the cancer type as a potential driver. Furthermore, in the pipeline, OncoVar output (https://oncovar.org/) is implemented, which utilized published bioinformatics algorithms and incorporated known driver events to identify driver mutations. Higher priority is given to the functional mutant alleles defined in hotspot package [96].

2.3.3 Allele dosage

A critical effort when conducting multi-omics is to establish that all assays executed on samples received from the same individual correspond. Therefore, I designed a dataflow to match the variants derived from DNA and RNA sequencing. I performed this by comparing VCF files from DNA sequencing with the BAM files from RNA-seq at predefined SNV positions. Furthermore, the impact of SNVs depends on whether the mutated allele is transcribed to the RNA. Therefore in parallel, the pipeline has implemented a method to check if mutated tumor alleles are presented and expressed according to their DNA frequency by assessing the allele asymmetries in RNA and DNA datasets using VAF. Here I established a samttols-based method to compute RNA-seq read coverage for the broad type and mutated allele and then evaluate heterozygosity DNA, and RNA mutation allele fractions were defined by analyzing the sequence of the reads coinciding with each SNVs Figure 7. Mutations were considered if at least eight DNA and eight RNA reads overlapped the genomic position. This minimum cutoff of eight was chosen to improve the accuracy when defining mutation allele percentages. The outcomes received when using higher or lower cutoffs are very comparable. Effectively, frequencies from samples with more reads were given more weight. Of 736 SNVs, only 578 are located in generally expressed genes. The next step is to find if a specific SNV position is presented in the transcript. Around 50 % of SNVs are expressed (281 mutation locations are presented in the RNA reads, and 297 are absent). For each SNV position, count tables for all nucleotides (A/C/G/T) were created from the BAM files of WGS and of RNA seq data. The mutation allele fraction at DNA and RNA levels was estimated as the number of mutation-containing reads divided by all reads overlapping the SNV. Differences between VAF (DNA) and VAF (RNA) values show if the mutated allele is generally expressed and where one of the alleles is preferentially transcribed over the other.

2.3.4 co-occurence of mutations

The co-occurrence of multiple mutations has gained increasing attention in identifying cooperating mutations or pathways that contribute to cancer. First, as a standard procedure, I dropped the synonymous mutations from the data because of their insignificant impact on protein sequences. Then, co-mutation was performed at gene and position levels. At the position level, the two exact co-occurring genomic positions were assigned as a co-positional pair. At the gene level, two genes were considered a co-mutation pair if cross-gene simultaneous mutations occurred. However, here the precise number of co-occurring gene positions is not considered. For instance, if two patients harbor one mutation in gene A and two mutations in gene B, I assigned only one co-mutation pair co-mutation pairs. Throughout this work, I only analyzed co-mutation pairs SNVs with allele frequency > 0.3 considered a possible driver. Be-

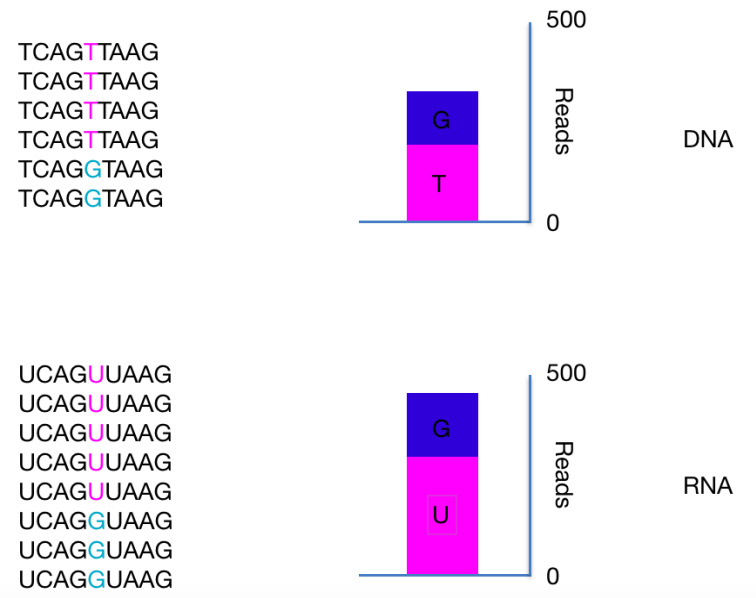


Figure 7: **DNA and RNA mutation allele frequency** The example depicts the sequence content and normalized read counts from the DNA reads where T > G.

cause mutations can be divided into stop gain SNPs and nonsynonymous SNPs set, I examined three types of co-mutation gene pairs: stop gain SNPs: stop gain SNPs, stop gain SNPs: nonsynonymous SNPs, and nonsynonymous SNPs: nonsynonymous SNPs. Since gene length was essential for analysis, I estimated the span between the transcription start and end site. I also implemented the FrequentLy mutAted GeneS (FLAGS) database that represents rare protein-coding mutations to exclude less likely disease-associated genes [97].

2.3.5 Protein domain-based approaches for prioritization of actionable cancer variants

In this step, I applied a protein domains-based computational approach to uncover the functional consequences of somatic mutations. I downloaded a list of protein domains from Prot2HG database [98]. This library contained a map for the known protein domain to a chromosomal hg19 genomic location. I parsed out all 281 expressed variants to obtain their protein domain coordinates. This mapping procedure yielded a total of 84 regions. That range suggests that the non-mapped mutations are not in the protein domain location. The database provides the protein domain name to assess if some modifications share distinct functional and/or structural unit alterations. Among them, there were 14 stopgain mutations and 70 nonsynonymous mutations. For the two samples (F2Q2AA and HS4N8A), I have a submitted control sample from the same patient. Whole genome sequencing of buccal swab DNA confirmed that the mutation found was somatic in origin as it was not found in a matched normal tissue, proving there are real mutations. The final mutations are evaluated for their cancer specificity and functionality from cBioPortal. Genetic aberrations in all these genes are visualized with the Oncoprint from complex heatmaps package [99].

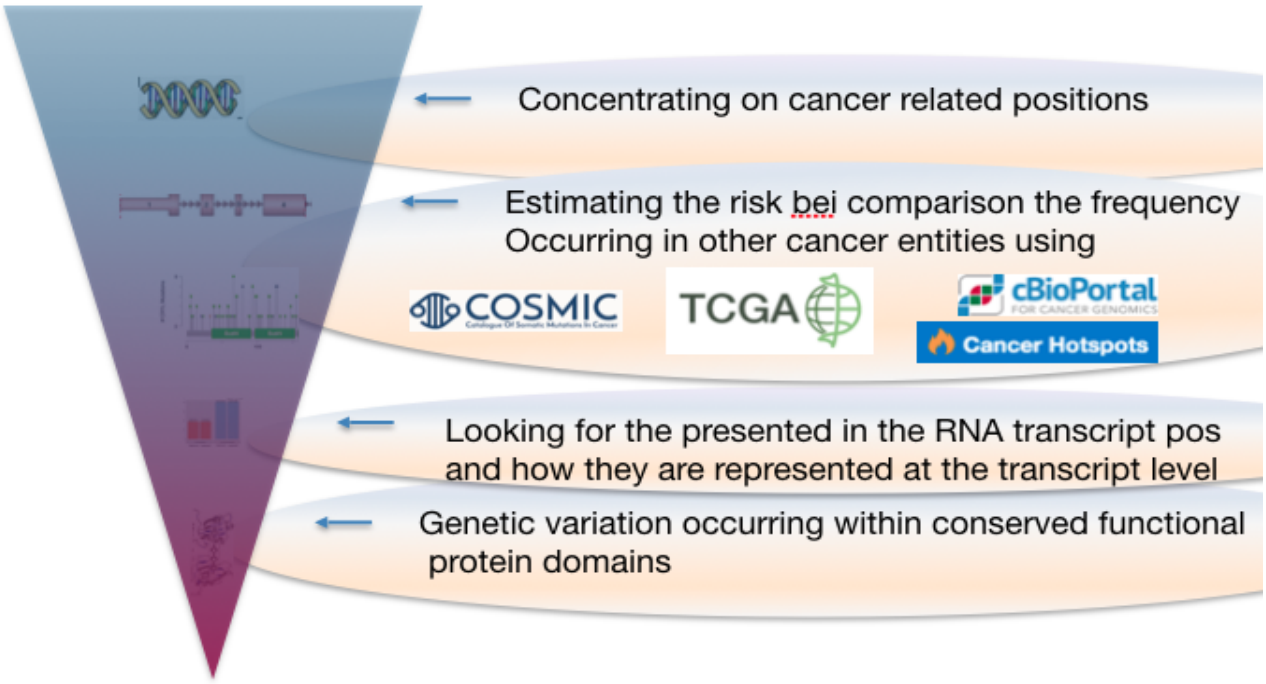


Figure 8: Overview of the mutation prioritization process

2.4 Methods for epigenetic project

2.4.1 Cell culture and drug treatment

Cell culture and drug treatment were performed by our collaborators in Freiburg. Cell lines U937 were cultured in RPMI 1640 medium. Afterward, the cells were treated in triplicates with daily pulses of 200 nM DAC for three consecutive days. After 48 hours, simultaneously with the third DAC dose, 1 µM ATRA was administered. For control, cells treated with the appropriate vehicles were used. All experiments were performed in triplicate for each condition. Samples' characteristics and experimental conditions are listed in Table **??**

2.4.2 Analysis of genome-wide DNA methylation data

- Data processing and statistics

After acquiring the raw data IDAT files, I performed quality control, preprocessing and basic analysis using R/Bioconductor with the RnBeads package [100].

Quality control plots and statistics for the methylation data were prepared by us using the **rnb.run.qc** command of the RnBeads package. This analysis assures the correct execution of the experimental procedures by checking all analyzed samples' bisulfite conversion, hybridization, and intensity distributions. Overall hybridization execution was evaluated using synthetic reference targets present in the hybridization buffer at three concentrations (low, medium, and high), resulting in signals with nicely separable intensity intervals, as desired. Further, the normalization method is well described in our paper described later the results. Beta values were conducted on-site and region level according to the sample groups specified in the analysis. RnBSet objects already contain summarized methylation levels for genes, promoters, and cpgislands. Since I would like to track the DNA methylation changes in transposable

elements, I incorporated into the workflow the database from different transposable elements (LINE, SINE, L1). The implementation was successful, and the command **summarized.regions** recognized the default and introduced by us new features by name. The output is represented as interpretable analysis reports in Html format, including the following analysis modules: Quality Control, Preprocessing (assess the effects of the normalization procedure), Exploratory analysis (Dimension reduction, statistical association tests, unsupervised learning techniques (clustering), and visualization techniques are applied to discover associations of the methylation data and sample groups), Differential Methylation Analysis (differentially methylated sites and genomic regions between different treatments are identified and visualized). The entire DNA methylation microarray dataset has been uploaded to the NCBI Gene Expression Omnibus (GEO) database and consists of the processed data matrix and a metadata table describing the samples and their treatment procedures.

Similarly, the method mval is used to extract M values. The annotation function obtains coordinates and additional annotation for sites or regions contained in the results. Differential methylation is to be computed for single cpg sites, genes, promoters, cpgislands, and newly introduced different transposable elements. Differential methylation for each selected annotation column is performed in a treatment group manner, defined in the condition column in the annotation table. ggplot package code is used to create a volcano plot comparing size and significance effects for all by introducing regions across all treatments and comparisons. A table for a given comparison and region type that contains comparison information on differential methylation regions was produced. Furthermore, different models are used to look into these regions. Finally, a density distribution plot from the ggplot package is used to demonstrate the dose-effect of the drugs. Futher method description could be found in out accepted paper “add citation”

About the patient, the DNA methylation is measured with EPIC array. Raw intensity data were obtained as IDAT files. DNA methylation data quality control and analysis were performed using the Bioconductor package RnBeads (version 1.10.8). Sites that overlapped with SNPs and cross-reactive probes were filtered, resulting in the removal of 17371 sites. Furthermore, probes giving unreliable measurements, 2410 determined by the Greedycut algorithm, were excluded from the analysis. Finally, the data from the remaining 847114 were subjected to background subtraction using the methylumi package (method "enmix.oob") [1] and beta-mixture quantile normalization (BMIQ) [2]. Hierarchical clustering analysis was performed in R using the Manhattan distance metric and complete linkage criteria.

Previous analysis had shown that standard normalization by RnBeads or openSesame did result in a highly unequal distribution of beta values. Therefore, a new undocumented normalization method from Plass lab DKFZ was used. That adds to the analysis extra steps to generate newly normalized beta values and some exploratory downstream analysis. The new normalization is done with the functions: sesame.noobsb (for background correction) and scaling.internal (for between-array normalization). The next step is calculating PCA on a subset of 3000 most variable probes, overlaying this with the annotation information to identify potential batch effects. For example, plotting gender as a factor reviews a strong SEX effect, probably due to probes on chromosome X. I will remove those. In addition, there seem to be moderate batch effects for SITE, likely due to the submission date. I use the removeBatchEffect method from limma for correction based on linear models. I also transform to M values instead of beta values because the distribution is much closer to normal.

12 BM-MSC samples from healthy decor controls contained in dataset GSE79695 were downloaded from the GEO database [101]. It includes 450K array data that I would like to compare with theAML EPIC data. Combining methylation values from 450K and EPIC is essential to regard the correlation at each CpG site. Therefore, I integrated only overlying CpG sites on both platforms. Before probe filtering, there were 452,567 overlapping probes. These fractions of removed and retained values, as well as the distributions of the removed methylation β values and the retained ones, are depicted in Figure 9. To perform differential expression between EPIC array and 450K array healthy control samples, I

combine the RNB set from overlapping CpGs using rnb.combine.arrays command. I defined differentially methylated regions on sites, CpG islands, promoters, gene bodies, and transposable elements. Tracking data pre-processing, differential analyses were conducted to estimate the differences between the AML and healthy groups. First, δ-β was computed as the discrepancy in the mean β-values for each CpG site using the RnBeads package. Then, the threshold for DMRs was established as a total value of $p<0.05$; δ-β >0.2 was described as hypermethylated sites, and δ-β <-0.2 was evaluated to display hypomethylated sites.

2.5 RNA data

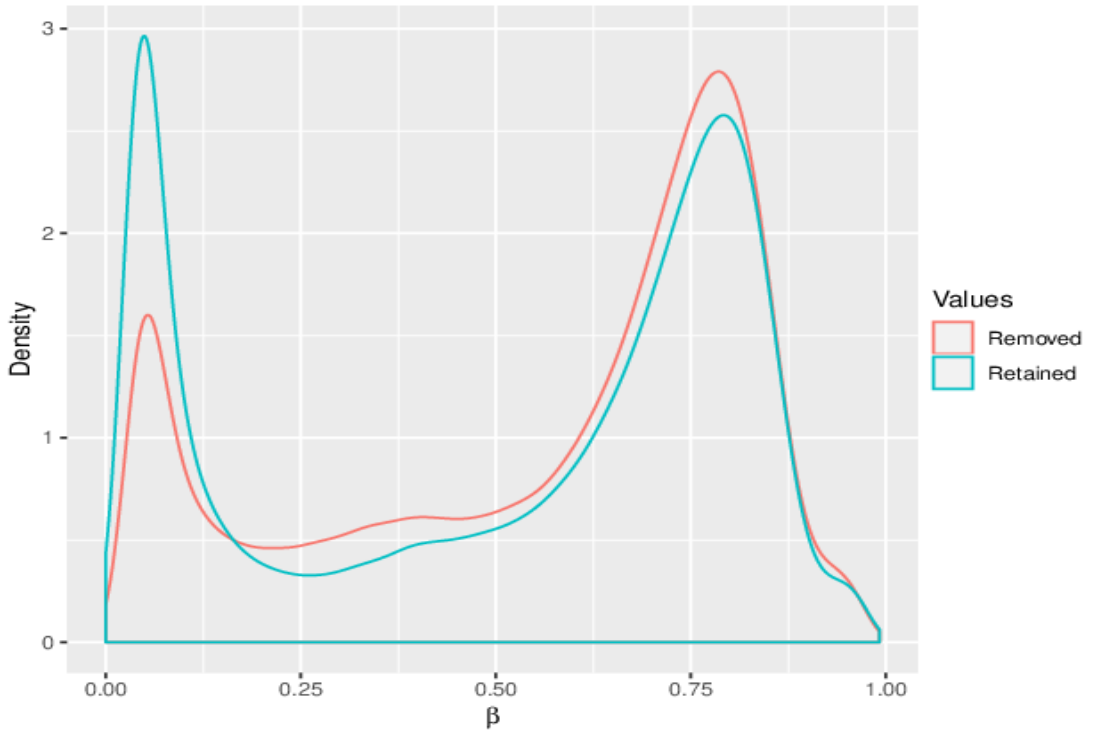
RNA was extracted from our collaborators with Manual RNAeasy Lipid Tissue Mini Kit. Libraries were sequenced on an Illumina HiSeq 2000 platform using 150-bp paired-end sequencing. An in-house RNA-seq pipeline was used to map and align the sequenced data. The filtered reads were mapped to the human genome (version hg19) using STAR aligner (version 2.5.2b). Raw reads were held in the fastq format and subjected to routine quality control (QC) criteria to withdraw the unfitted according to the following parameters: Reads aligned to adaptors or primers. Reads with >10% unidentified bases (N bases). Reads with >50% of low-quality bases in a single read. Normalization is a critical step in gene expression studies for RNA-seq data. TPM was utilized to estimate the expression level. The TPM method could eliminate the impact of different gene lengths and sequencing discrepancies on the calculation of gene expression levels.

For comparison with healthy individuals, the raw RNA sequencing data from normal bone marrow samples from the Human Protein Atlas were downloaded with the following accession codes: ERR315333, ERR315395, ERR315396, ERR315404, ERR315406, ERR315425, ERR315469, ERR315486. Differentially expressed genes between healthy controls and AML samples were identified using the Bioconductor package DESeq2 [102]. Based on the concurrent mutations, patients were separated on The non-parameter Wilcoxon rank-sum test was used to examining the relationship of continuous variables between each mutation group. In order to identify aberrant gene expression levels as potential pathogenic events, I used OUTRIDER (Outlier in RNA-Seq Finder), an algorithm that uses an autoencoder to model read-count expectations according to the gene covariation [103]. The package was run with AML and healthy samples to check if the algorithm would distinguish and extract only deceased outlier expressions. The Z-scores on the log-transformed counts were utilized for visualization and sample hierarchy. In order to view the dispersal of P-values on a single sample, volcano plots are demonstrated with the plotVolcano function. In addition, the quantile-quantile plot using plotQQ function shows if the fit converged nicely.

2.6 combination of RNA and DMA methylation

Indeed, DNA methylation holds the beta values for each methylated site, so each sample has several methylated sites belonging to a given gene promoter or gene body. Therefore, for combining methylation quantity at the gene/promoter level, I evaluate the sum of the beta values as a measurement of the general intensity of the methylation on each gene and promoter in separate matrices. When assessing RNA sequencing, the rows represent the samples and the genes' columns, and the matrix items contain the TPN gene expression values for each gene for the same gene set as for DNA methylation data. Canonical correlation analysis(CCA) is a well-known strategy for studying the associations between two groups of variables [104]. I peroformed CCA analysis on the promother methylation and gene expresion data, as well as on the gene bodies methylation and gene expresion data.

A.



B.

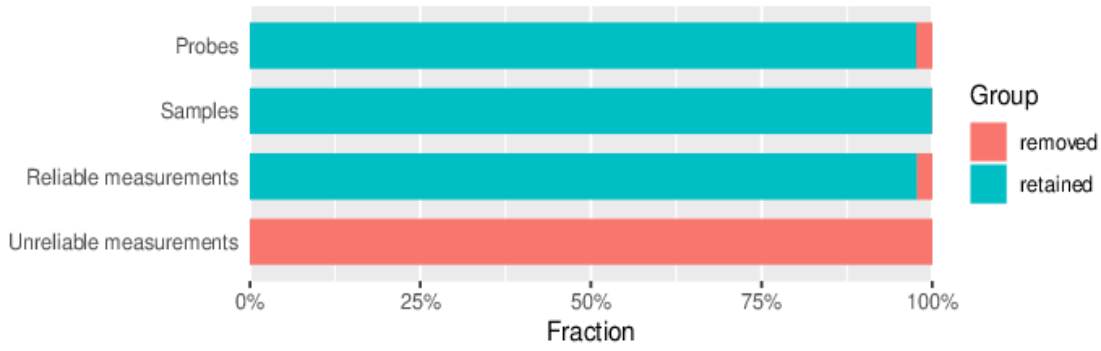


Figure 9: A. The distributions of the removed methylation β values and of the retained ones. The distribution of retained betas is evaluated by randomly sampling 1000000 values., B. Fractions of removed values in the dataset after applying filtering procedures.

2.7 fusion genes

For fusion detection from RNA-seq data I used fusion detection algorithm [105]. Arriba takes the main output BAM file, and the script download_references.sh is used to download the hg19 assembly. Arriba can be ordered to be particularly sensitive toward events between specific gene pairs by providing a list of gene pairs (parameter -k). Here I downloaded all known frequent fusions in all hematological malignancies from COSMIC dataset and integrated them into the Ariba. In the output fusions.tsv (as specified by the parameter -o) holds highly enriched accurate predictions fusions that pass all Arriba's filters. The predictions are listed as high, middle, and lowest confidence. Since the initial output gives us many recurrent fusion genes, I applied and built additional filters and strategies to filter possible artifacts. Initially, built-in filters of the callers were applied, and provided blacklist helped filter artifact fusions. I removed fusions for which both partner genes belong to the same gene family, as that could be the suggestion about the potentially high sequence similarity. Furthermore, low read coverage of a fusion event proximate to the read coverage of its partner genes suggests an artifact. The number of supporting reads is a valuable attribute in distinguishing artifacts from actual events. Accurate predictions usually have a high number of supporting reads. Therefore I examine that >6 is sufficient for the study. It is prevalent for fusions to be covered by only a part of the expression of a gene. Some real translocations affect only a few supporting reads, although the gene is highly expressed. One example is when one allele of a tumor suppressor is inactive due to a fusion and expressed at a low level, and the other allele is favorably expressed but also impaired as a consequence of a mutation. In that case, even output with just 3-6 reads can be valid. Here I combined SNV and indel data with the Ariba fusion output and checked if I found some mutations in any of the fusion genes. In the case of transchromosomal fusions, I compare the results with WGS data to determine if the finding is a real fusion or artifact. However, Ariba's initial output provided us with several frequently fused genes analog to blacklist genes (mostly hemoglobin genes) which brings us to the next step of individual filtering. Even though Ariba provided us with a fusion score (high, medium, and low), I filtered and considered only fusion with more than six supporting reads. The number of supporting reads most influences the ranking. However, additional features are also taken into accounts, such as the closeness of the breakpoints (intragenic vs. read-through vs. distal), multiple transcript variants between the same pair of genes caused by alternative splicing, or the level of background noise in a given gene. Since the tool reports alternatively spliced transcript variants involved in the fusion, I rely on this information and put more weight on fusions from different spicing forms. Each high confidence fusion breakpoint was double-checked using IGV to examine the reading frames of the exons adjacent to the breakpoints.

I assessed gene expression employing the same RNA-Seq dataset. I normalized the data using the common approach from the DESeq2 package [102]. I depicted log-scaled normalized expression values to reach the partner gene expression of samples with fusions to other instances. Because I only had rather small numbers of fusions, I conducted these tests on all high-confidence fusion genes. I estimated differential expression employing two-sided Student's t-tests subject to normalized expression values.

2.8 Kaplan-Meier survival analysis

I collected patients' mRNA expression profiles and clinical annotation data from the TCGA website [106]. for all transcriptionally dysregulated genes and mutations. For a subset of mutated, overexpressed, and downregulated genes, I studied their relation to relapse and survivor from TCGA AML data carrying the same artifact. I computed Kaplan–Meier curves and P-values from the log-rank test using the survival package (version 2.38).

3 RESULTS

3.1 Identifying recurrent mutations in AML carrying t(8;16)(p11;p13) rare abnormality

In this project, I studied the application of WES in tumor mutation detection, particularly for AML-related genes from samples carrying t(8;16)(p11;p13) aberration. Using the no-control pipeline, I extracted mutations that could be candidates for targeted therapy in AML-t(8;16)(p11;p13) patients without a matching control.

The translocation, t(8;16)(p11;p13), results in the fusion of MYST3 (located on chromosome 8p11) and CREB-binding protein (CREBBP) (located on chromosome 16p13). Both proteins maintain histone acetyltransferase activity, and in addition, they are implicated in transcriptional regulation and cell cycle control [107] . MYST domain with histone acetyltransferase (HAT) activity stays intact in all the t(8;16) translocations described thus far. It regulates the transcription of distinct target genes by coactivating the RUNX1 transcription factor complex [108]. MYST3-CREBBP chimeric protein may disrupt several hematopoietic pathways due to the disease's essential mechanisms.

This type of AML is found predominantly in female patients and can occur in children (median age two years) and adults (median age 60 years) [109]. Therefore not surprisingly, 6 out of 10 patients are female.

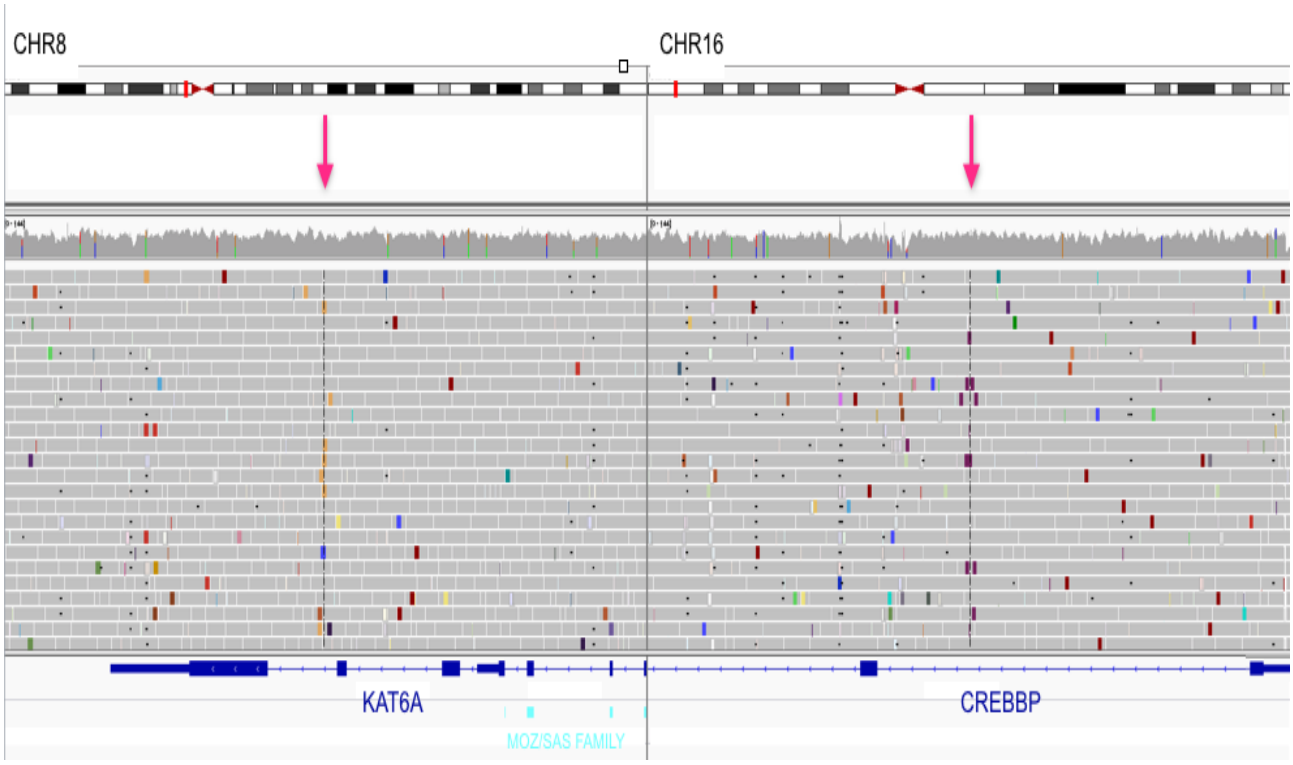


Figure 10: IGV view from a patient WGS with typical MYST3-CREBBP fusion. The pink arrow highlights the exact breakpoints of the fusion between these two genes. The WGS fusion gene detection pipeline determines the fusion.

To further examine t(8;16) AML, I have calculated mutational signatures from 10 patients, 7 of which with cytogenetically demonstrated t(8;16) MYST3-CREBBP, and 3 with an atypical t(8;16)(p11;p13) fusion not resulting in MYST3-CREBBP fusion. Since all the samples came without matching control, I utilized an in-house designed no-control pipeline. Therefore, the initial variant calling liberally must collect the sum of all possible variants to obtain high sensitivity. Consequently, I first characterized sequencing quality filtering criteria to filter variants on read depth ($\geq$6). Then, I observed that filtering

using dbSNP [93] and Exome Variant Server (EVS) [92] dramatically reduces the potential germline mutations by excluding common SNPs. In addition, our in-house control dataset (>2%, among 280 controls) is used to reduce the false positive cases further, as expected, by enriching the probabilities that any single normal control would have a rare SNP that could be named a mutation.

Here I found several mutations accompanying t(8;16) fusion. I detected non-synonymous mutations in the following genes: ASXL1 (in 3 patients), BCORL1 (in 1 patient), BRD3 (in 3 patients), CBL (in 1 patient), EYS (in 4 patients), FLT3 (in 3 patients), IDH1(in 1 patient), KRTAP9-1 (in 4 patients), MLH1 (in 3 patients), NUP98 (in 1 patient), PSIP1 (in 1 patient), POLG (in 3 patients), RUNX1 (in 3 patients), SAMD4B (in 3 patients), SETD2 (in 1 patient), SPTBN5 (in 4 patients), TET2 (in 2 patients), TP53 (in 3 patients) and WT1 (in 1 patient) (Figure11). FLT3, a receptor tyrosine kinase, is recurrently altered in acute myeloid leukemia and other hematologic malignancies. Mutations in FLT3 are very typical for AML, specified in roughly one-third of the cases. In addition, FLT3 internal tandem duplication mutations (FLT3-ITD) are correlated with increased relapse and low overall survival [110]. In 3 of the typical MYST3-CREBBP fusion patients, there are FLT3 mutations. Two are most likely oncogenic since they are found in several TCGA AML cases. However, those mutation has not been functionally or clinically validated. The FLT3 D835Y alteration in AML6 has been identified as a statistically significant hotspot and is likely to be oncogenic since it is recognized as a recurrent hotspot in samples of various cancer types, especially AML, using a method based in part on [111] ASXL1 mutations are found in myeloproliferative neoplasms (MPN) and acute myeloid leukemia (AML). They are commonly associated with indications of aggressiveness and poor clinical outcome [112]. I observe ASXL1 mutations in 4 out of 10 patients, where 3 have a typical t(8;16) breakpoint, resulting in MYST3-CREBBP fusion. The ASXL1 mutations are found in myeloproliferative neoplasms (MPN) and acute myeloid leukemia (AML) [31]. They are commonly associated with indications of aggressiveness and poor clinical outcome. The cohort has four patients with a mutation in this gene. In AML cell lines, POLG plays an essential role in supporting cell viability and mitochondrial structure; inhibition of POLG promotes differentiation of myeloblasts. That suggests that POLG mutations with resultant mitochondrial dysfunction in solid malignancies contribute to bad disease biology and poorer outcomes [113].

I found that half of the samples possess one or more mutational hotspots from a gene with prominent hematological cell pathways such as epigenetic machinery, mismatch repair, and tumor suppressor genes. Even with the missing controls, heterogeneous nature of the disease, and a small cohort of patients, I could distinguish possible drivers in all ten patients and go deeper in understanding the genetics of this rare t(8;16)(p11;p13) AML abnormality.

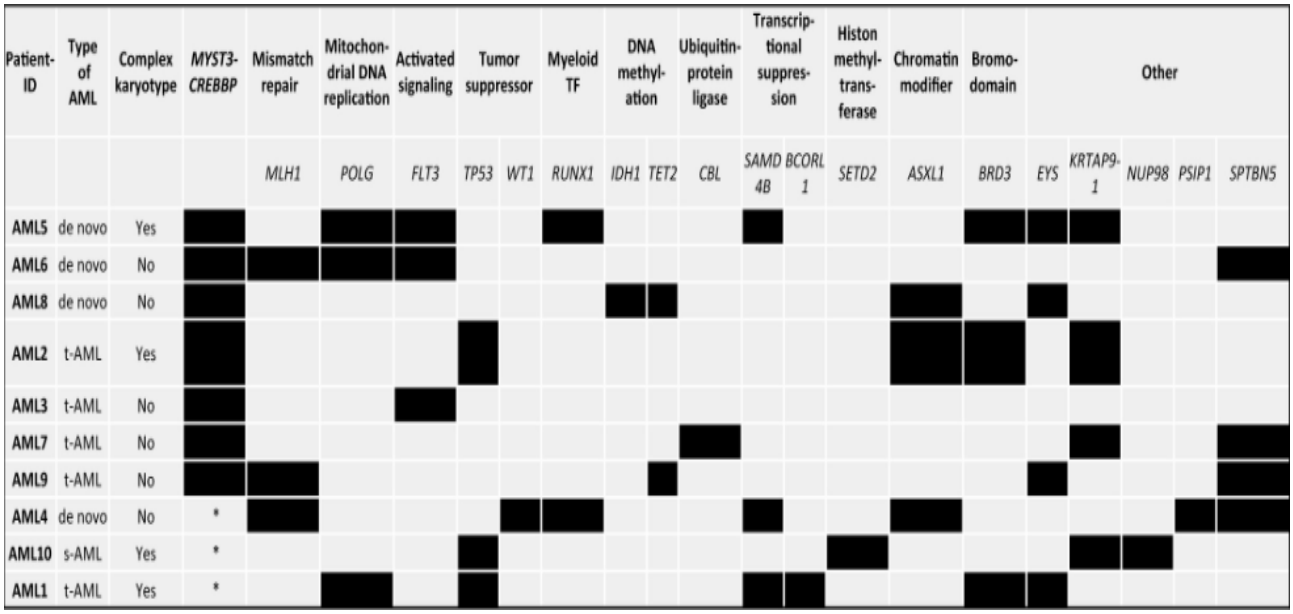


Figure 11: Oncoprint depicts common alterations in all ten patients. s-AML stands for secondary AML after myelodysplastic syndrome; t-AML stands for therapy-related AML. In the patient, AML1, AML4, and AML10 are atypical t(8;16) breakpoints observed, not resulting in MYST3-CREBBP fusion protein. The Figure is published in our paper [90]

3.2 The antileukemic activity of decitabine upon PML/RARA-negative AML blasts is supported by all-trans retinoic acid

In this project, I studied the methylation change of U937 cell lines after their treatment with DAC, ATRA, and their combination. The effective treatment of AML is very demanding because of the disease heterogeneity. In addition, most patients with AML are older and exhibit a poor prognosis even after intensive therapy. This is because AML survival mechanisms are involved in reprogramming metabolism, impairing differentiation and drug resistance. Therefore, when combining drugs or treatments, more processes in the cell can be targeted together to increase the response in a broader group of patients. [114]. Therefore, inducing differentiation and apoptosis of leukemic blasts by DNA-hypomethylating agents, like decitabine (DAC), represent well-tolerated alternative treatment approaches [115]. Another differentiation therapy with all-trans retinoic acid (ATRA) has strongly improved outcomes in different types of AML [116]. Several extensive clinical studies have handled the role of ATRA when presented in a mixture with high-dose chemotherapy. In addition, some studies show that adding ATRA to induction treatment in AMl patients older than 60 years resulted in a survival advantage [117]. However, other studies do not present this additional effect of ATRA, and the data suggest no difference in the early death rate with or without ATRA [118][119]. In the study with elderly patients with oligoblastic AML unsuitable for induction chemotherapy, the addition of ATRA to DAC improved the overall survival of patients [120] . Furthermore, similar studies described that AML cell lines treated in vitro with DAC in combination with ATRA showed enhanced in vitro differentiation [121], providing additional motivation for further investigation of combining ATRA with the DAC. Therefore, in the present study, I tested if the ATRA cooperates with DAC on epigenetic levels triggering antileukemic activity, including dysregulation of transposable elements. To assess DNA methylation, the HumanMethylation450 BeadChip array was performed. After quality control and normalization steps, I obtained DNA methylation β-values (from 0.0 (unmethylated) to 1.0 (fully methylated)). I used methylation data measured with the Illumina Infinium HumanMethylation450 BeadChip array from triplicates for each condition: 72h untreated, 120h untreated, 72h DAC, 120h DAC, 72h ATRA, 120h ATRA, 72h DAC+ATRA, 120h DAC+ATRA.

3.2.1 Global methylation level change

To characterize the effect of all types of treatments, I performed different unsupervised analyses with different tools. Dimension reduction is used to visually examine the dataset for a strong signal in the methylation values related to sample treatment. I first performed a Principal component analysis (PCA) and applied hierarchical clustering based on all methylation values to identify and visualize potential clusters of DNA methylation profiles (Fig 12). PCA is a mathematical conversion of correlated variables into multiple uncorrelated variables named principal components. The consequent components from this transformation are depicted in such a way that the first principal component has resulted in the highest variance and accounts for most of the variability in the data. I plotted the PC1 and PC2, revealing how the sample treatments will cluster within these dimensions. The hierarchical clustering dendrogram uses correlation distance, $1-\rho$, where ρ is the Pearson correlation coefficient between two pairs of samples. I observed that ATRA alone does not cause a change in global DNA methylation level. As expected, DAC reduces the methylation level drastically on both 72h and 120h DAC treatment. Despite the evidence that ATRA targets the PML-RARA (promyelocytic leukemia/retinoic acid receptor-alpha)/DNA methyltransferase (DNMT) [122], DAC and DAC + ATRA treatments showed similar demethylation patterns. This approach shows no evidence that ATRA demethylates DNA. When comparing DAC and DAC+ATRA treatment, I observed that samples cluster based on time, not the treatment difference (Fig 12).

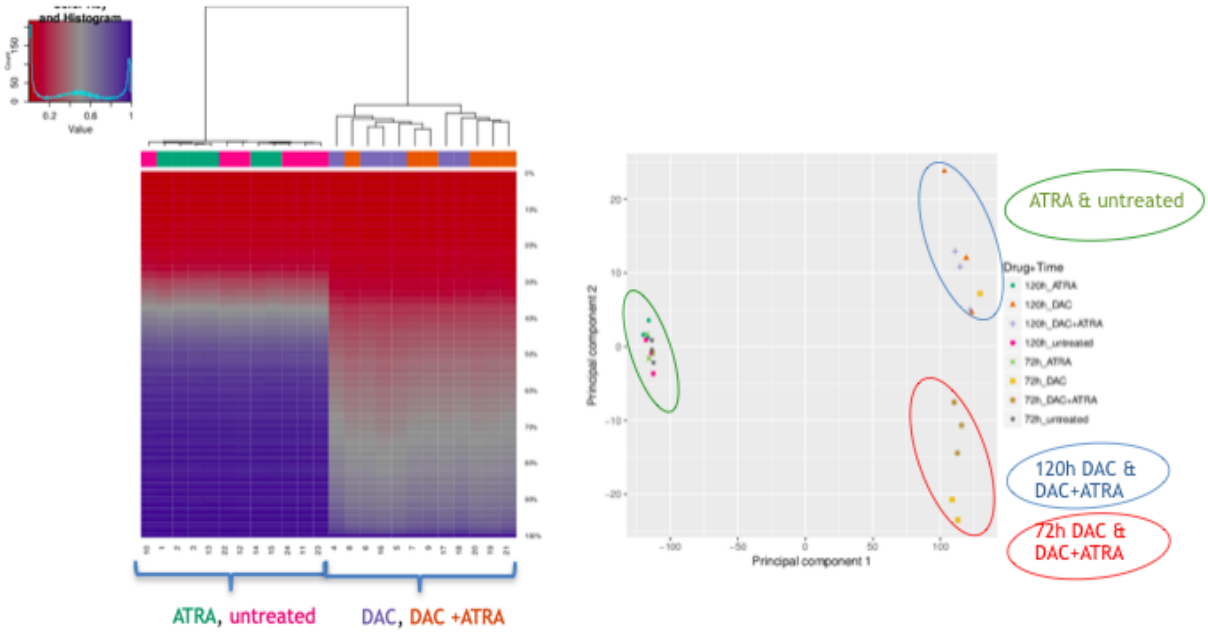


Figure 12: The hierarchical clustering of samples (left) is based on all methylation values. Therefore, the heatmap depicts methylation percentiles per sample. The scatter plot (right) displays the samples and different groups of treatment coordinates on principal components.

3.2.2 transposable elements coverage

A big part of the human genome is repetitive sequences derived from transposable elements (TEs). TEs are DNA sequences that can change their position within a genome. Thus, it is likely, not surprising that TEs are randomly dispersed in the genome [123]. TEs interact with pluripotency factors such as NANOG and OCT4 during early development, which supports a possible link between TEs and the dedifferentiation of tumor cells. Furthermore, TEs may regulate the expression of oncogenes and promote oncogenesis. TEs are usually transcriptionally repressed, but epigenetic changes in tumors may enable their expression and allow them to retrotranspose [124]. Therefore, in the study, I tracked the methylation changes on promoter and gene levels and different transposable elements. It is known that 450K technology covers 485,577 CpG sites and 99% of RefSeq genes [125]. I measured the array coverage of TEs. Long interspersed nuclear (LINE) elements, there is a 7 % overlap. L1 elements are 5% of them. The array covers 10 % of all short interspersed nuclear (SINE) elements.

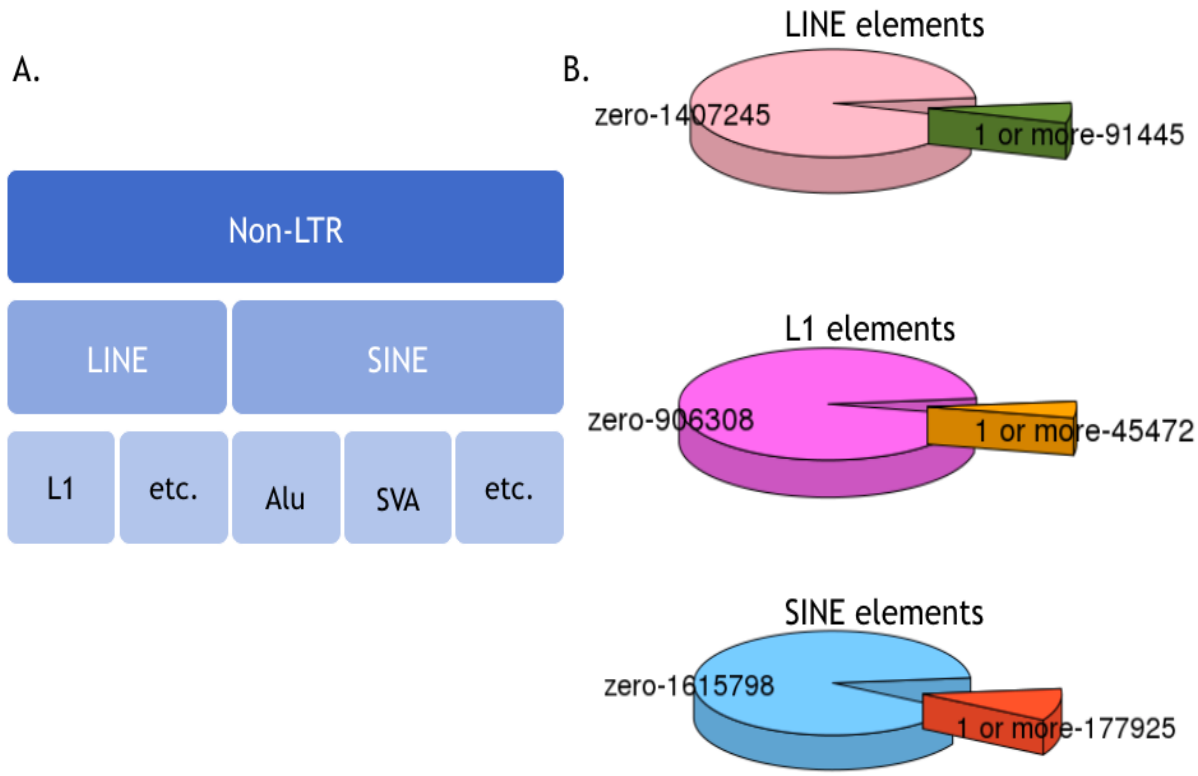


Figure 13: A. Schema describing the subclassification of non-LTR retrotransposons are elements lacking long terminal repeats. B. Pie chart presenting the coverage of 450K array. LINE elements without any probes in the array are 1407245. One and more probes cover 91445 LINE elements on the array, which make up 7 % of all Line elements, from 5 %, 45472 are L1 elements. SINE elements have 10 % coverage in the array, where 177925 SINES are represented.

3.2.3 Differentially methylated regions

I investigated which regions are affected by different drug treatments and their combination. My analysis tested the hypothesis that ATRA cooperates with DAC, initiating transcriptional programs activating antileukemic processes, including reactivating transposable elements (LINE, SINE). Repetitive elements are involved in several cellular processes, and due to their innate qualities, they can be the origin of genome instability. Thus, numerous repetitive elements are usually methylated to preserve a heterochromatic, repressed form. However, growing evidence shows repetitive elements are often hypomethylated in various cancers, especially in AML [126].

Since DNA methylation changes over time, a reference is used for each condition as an untreated sample from the same time point. Therefore, I compared the triplicates of the same treatment and time (72h ATRA, 72h DAC, 72h DAC+ATRA, 120h ATRA, 120h DAC, 120hDAC+ATRA) with samples with corresponding matched time controls (72h untreated and 120h untreated). Since DNA methylation changes over time, a reference is used for each condition as an untreated sample from the same time point. To determine whether differences in DNA methylation differ at the CpG level between untreated samples, I compared the CpG level methylation variation between the untreated samples. I observe that for all CpG, the deviation of the triplicates is less than 0.035, which promises comparable results (Fig 14A). I also checked if there is a significant variance in the methylation values between the two

untreated conditions (72h untreated and 120h untreated). There are no differential methylation CpG regions between 72h untreated and 120h untreated (Fig 14B). Most data fell on the diagonal, representing similar methylation levels in the two untreated timepoints. An overall significant difference between the 72h and 120h untreated interval (Kruskal-Wallis, p = 0.9991) was not observed: Fig 14B . All these results speak of good reproducibility and interpretability.

I performed differential methylation analysis on single CpG resolution and across all genomic regions of interest (promoters, gene bodies, CpG islands, LINE, L1, SINE), enhancing the results' statistical power and interpretability. Differential methylation for each selected region is performed from treated(name of the treatment and time) versus untreated (same time). For each region level (promoters, gene bodies, CpG islands, LINE, L1, SINE), differential methylation for the individual comparisons is facilitated through aggregating p-values of all sites in the selected region using a generalization of Fisher's method [127]. Differential methylation for each selected annotation column is performed from treated(name of the treatment and time) versus untreated (same time) combined p-value using a generalization of Fisher's method. My study identified more than 10 000 CpG hypomethylated sites in both DAC+ATRA conditions(72h and 120h). Moreover, these extreme hypomethylation events could be seen on single CpG sites and functional genomic regions such as CpG islands, genes, and promoters. In addition, I uncovered that the major repetitive elements, such as short interspersed elements (SINE) and long interspersed elements (LINE), have a high percentage of differentially methylated regions in all DAC-related treatments (Fig14C, D).

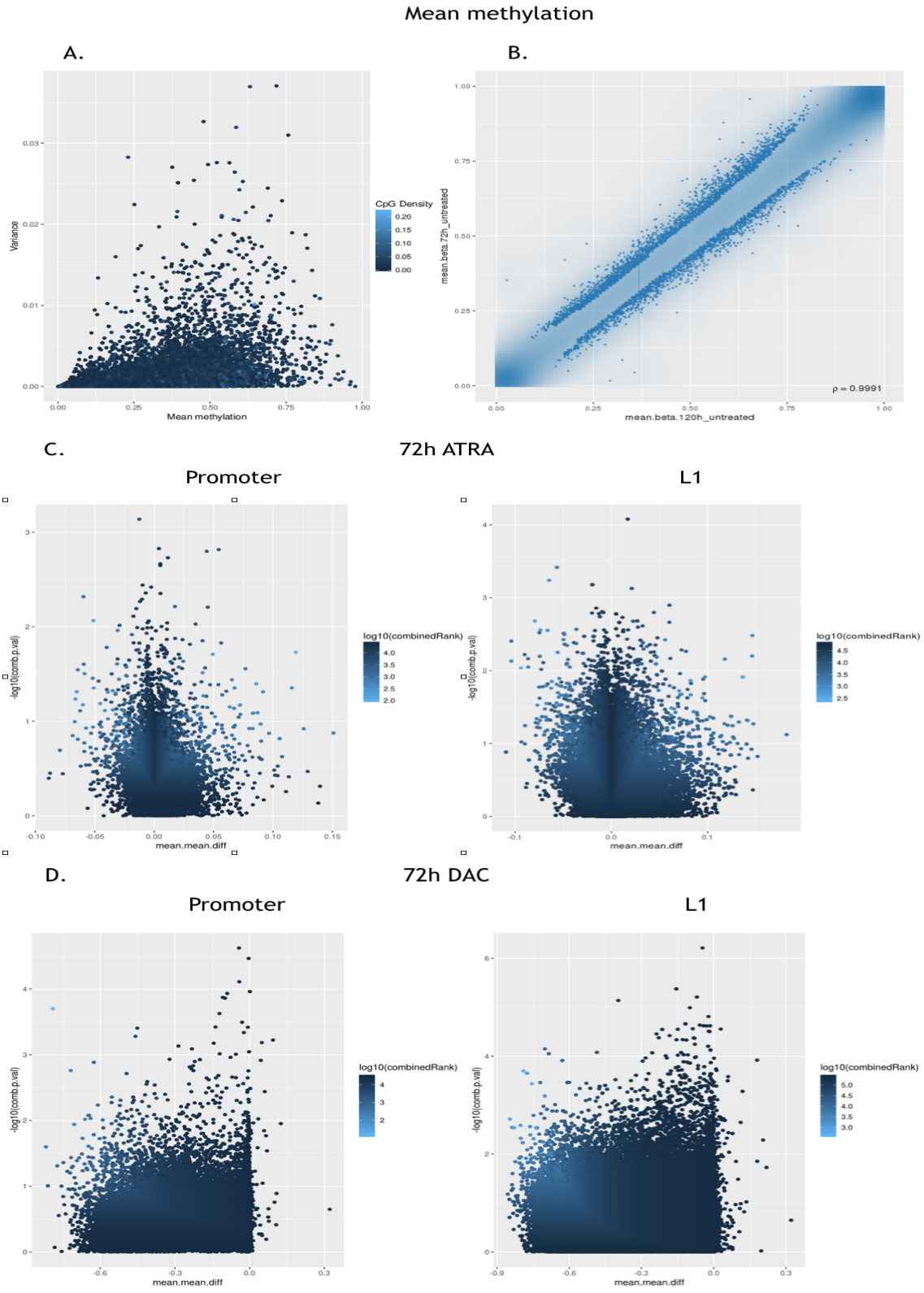


Figure 14: Control untreated sample-wise methylation variability, B. Scatterplot for differential methylation (sites) between 72h untreated and 120h untreated. C. Volcano plot of gene methylation changes after 72h treatment with ATRA. D. Volcano plot of gene methylation changes after 72h treatment with DAC.

c

3.3 Prospective driver events in mutation-negative nkAML

Genetic abnormalities contribute to AML. It develops due to a sequence of genetic changes in a hematopoietic precursor cell. These alterations transform normal hematopoietic to growth and stop differentiate differentiation, resulting in an expansion of considerable numbers of immature and abnormal myeloid cells in the bone marrow and peripheral blood [128]. Advances in understanding the genetic basis of this disease have resulted in significant improvements in diagnosis and prognostication and provide opportunities for greater individualization of therapy. However, about half of the AMLs do not possess known genetic abnormalities [129]. These patients are categorized into a normal karyotype AML (nkAML) subgroup since they do not harbor clear markers for additional classification. The deviation of therapy response and prediction of normal karyotype AML indicates that this subgroup of AML is heterogeneous with different genetic abnormalities. A study including 393 patients with normal karyotype AML has identified driver mutations across 40 genes, with one or more driver mutations found in more than 95 % of patients. Those are mutated chromatin or RNA-splicing genes, TP53, CEBPA, tumor suppressors, cohesin complex, myeloid transcription factors, activators in signaling pathways, and others [130]. The most shadowy are patients with nkAML with no detected driver mutation since they don't harbor an apparent driver gene. It is presently unidentified whether patients with cytogenetically normal, non-driver mutation AML harbor genetic and/or epigenetic alterations that might be associated with distinct clinical phenotypes since comprehensive multi-omic analyses focusing on this specific type of AML have not yet been conducted. Nevertheless, uncovering abnormalities on the genomic, epigenomic, and transcriptomic levels that drive non-driver nkAML development could serve as therapeutic targets.

In the study, I performed extensive analyses of mutation, gene expression, and DNA methylation data from AML and MPN patient data to uncover coding or non-coding genetic abnormalities and epigenomic alterations that drive nkAML. I thus investigated multi-omic patterns along the nkAML following the workflow depicted in Figure 15 I present a multi-omics-based framework to identify cancer driver events by combining multi-omics nkAML and MPN data. To generate driver gene representations, I propose an interactive workflow to combine multiple omic-specific information with public genomic, transcriptomic, and clinical data from TCGA. For convincing candidate genes, I examined the sequencing data from WGS and RNA-seq and array data from EPIC methylation data. Each patient's cancer is a distinct and heterogenous entity and can be investigated from various angles. Therefore, different approaches to data analysis with many data types are carried out, including detecting SNVs, indels from DNA, gene expression, DNA methylation, and fusion detection from RNA. Firstly, I built a mutation prioritization workflow described below that narrows each patient's genomic drivers. Here I used information on the allele dosage of the specific mutation combining information from WGS and RNA sequencing data. Furthermore, I evaluate the features and impacts of the gene mutations. I compared them on a positional level with AML and other cancer cohorts funding to assess the harmful genomic mutations. From RNA seq data, I could get asses if one mutation is expressed or not and its dosage representation. I further applied different methods for RNA expression of outlier and fusion genes. I downloaded and analyzed raw RNA data from healthy bone marrows for comparison and pipeline guidance. The methylation array data analysis was the most challenging since I observed the most batch effects there. After normalization, I could track the methylation changes within the sample. The pipeline monitor which events would coincide and compare with publicly available data from TCGA and evaluates the output with clinical implications.

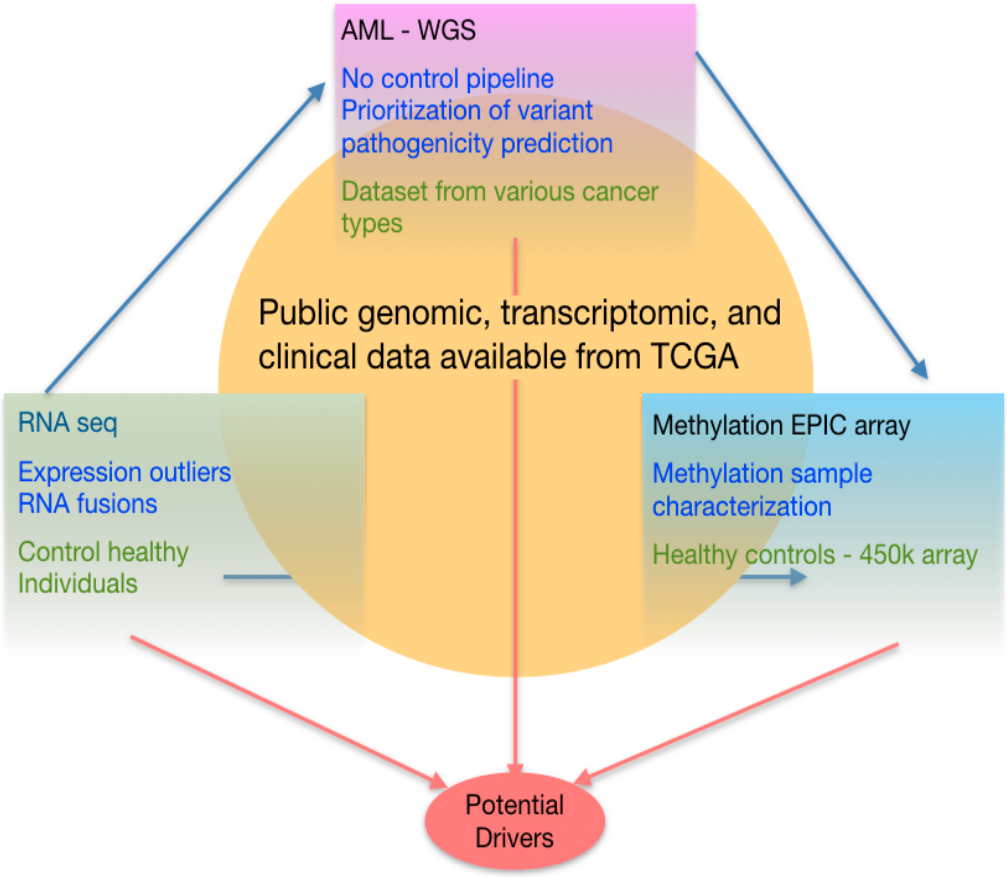


Figure 15: Overview of multi-omics workflow for identifying driver mutation profiles.

3.3.1 personalized prioritization of driver mutations

Methods for sequencing analysis have undergone continued, rapid improvement. However, detecting somatic mutations in cancer samples has been challenging due to inherited germline variants, sample heterogeneity, and genomic instability [131] . Typically, a tumor and a "normal" sample from the same patient are sequenced to be compared. Variants observed only in the tumor are evaluated as somatic mutations. However, this approach requires two samples for each individual. Therefore, I created a workflow downstream of our no-control pipeline to detect somatic mutations in patients for which only tumor samples are available.

Each healthy or diseased person carries around 88 million genetic variants and polymorphisms. Most of them cannot be identified as the single cause for a given phenotype [132]. Therefore, several strategies were applied to identify the germline variants to improve somatic mutation calling.Since I are interested in functional alteration, I concentrated the analysis on the exome part of the genome. The exome comprises <2% of the human genome but holds ∼85% of known disease-causing variants [133]. Since all the samples came without matching control, as in the previous project see 1.1, I utilized and adapted a no-control pipeline previously developed in the lab to exclude the most common SNPs. After running this pipeline, the observed patient mutations were over 200, which is too much for a disease such as AML with a low mutational burden (add citation). Therefore the task was to develop a pipeline to filter the possible germline variants, estimate the most probable cancerous from driver genes and prioritize functional driver mutations. Furthermore, I look into different angles into the genomic, epigenomic, and transcriptomic patient signature to prioritize driver alterations. Since the cohort consisted of 34 AML samples, a relatively small number of samples, I relied not only on the cohort's statistics but also took advantage of different publicly available datasets.

The search for cancer drivers has quickly advanced with systematic exome-sequencing studies. The primary purpose of these analyses is to determine signals of positive selection and differentiate them from passenger mutations. In the case of samples without a matching control, there are both issues to solve; if this SNP is a germline variant or a mutation but a passenger that does not cause any phenotypical change [132]. Since AML is a complex disease with heterogeneous genetic causes and clinical outcomes, large-scale analyses have thoroughly specified somatic mutations across multiple AML types, providing datasets for comparing patients based on genomic alterations. However, one challenge with this study is that modifications are rarely shared across patients. I propose multi-omics-based stratification approaches to overcome this challenge and are used to find drivers based on genome and transcriptome levels. Meanwhile, at the decision-making process in the pipeline, the publicly available data from AML (and cancer in general) and healthy datasets are used to assess the interpretability of the results. First, I studied individual variant characteristics and sequential NGS assays to identify potential germline variants by analyzing mutational variant allelic frequency (VAF) implications. Assessing tumor heterogeneity is an essential factor of cancer genomics since comprehending the subclonal arrangement can indicate important clues for how a tumor has developed and what targeted treatments could be adequate [134]. Based on the tumor cellularity, I expect driver mutations to present an allelic frequency of at least 0.3 on average (average cellularity divided by half). Since I proved I are dealing with a flat genome, the genome is absent from copy number gains, and I also do not expect an increase of any observed allelic frequency. We, therefore, used a cutoff of 0.3 to split high from low allelic frequency as a marker of early clonal mutations versus late, subclonal mutations. Here not surprisingly, the most mutated genes in the study had VAF between 0.3 and 0.6, with one sample exception (F2Q2AA) with more than 800 subclonal variants with VAF less than 0.3 (Fig. 16). That is also the only sample that got a treatment with chemotherapy medication Daunorubicine before sample extraction. The number of variants with allele frequency larger than 0.6 is stably low in all samples that are most probably germline snips. In the best scenario, germline

variants have VAF $\sim$ 0.5 if heterozygous and VAF $\sim$ 1.0 if homozygous. However, the VAF must be decoded comparable to germline mosaicism, loss of heterozygosity, and sequencing artifacts, including statistical oscillation, especially with shallow sequencing depths. Therefore, a detailed understanding of NGS data from clinical samples is required. The patient from the cohort passed the sequencing panel specifically targets 54 genes known to be frequently mutated in hematologic malignancies, concentrating on leukemia and myeloproliferative disorders [135]. In addition, this preliminary screening approach includes determining the status of some driver genes gene, purity, and ploidy of the sample to verify that I are dealing only with no-driver AML samples. However, I identified some samples with known AML drivers that slipped through the first-panel screening. From RNA seq data, I found some samples having interchromosomal translocations involving a switch of a chromosomal segment(s) between chromosomes. F2Q2AA carry t(11;12)(p15;p13) NUP98-KDM5A translocation and therefore cannot be classified as normal karyotype AML. The same issues have three more samples, B2PHB9, 52XXGW, and GP28EK, which have the following interchromosomal fusions: RUNX1-USP42, KMT2A-MLLT4, and again, KMT2A-MLLT4. The panel did not detect some very prominent hotspot driver AML mutations in some samples, such as JAK2V617F and IDH2R140Q, all with VAF around 0.4. The reason why these samples are handed to us as no-driver is unknown. Even if those samples are normal karyotypes, they do not pass the condition of non-driver variants. Therefore are analyzed separately from the real normal karyotype and non-driver samples. After separation, the number of non-driver samples is 28. Since all 34 of the provided samples have a flat genome, the samples that do not cover all the criteria of non-drivers are not directly excluded but rather parallel analyzed with all other samples. I do so because the information they harbor is still beneficial for us. For example, they can help solve batch effect issues from one hand and, from the other, direct us to different driver alterations that a no-driver sample may also have. To further prioritize mutations, I took advantage of publicly available data sets. It is known that some genes are more cancerogenic than others. For example, some gene mutations, like those in TP53, typically arise as somatic alleles.

Big cancer sequencing projects, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), supply exceptional possibilities to pinpoint alterations that are the probable cause of human cancers [136]. Therefore I filtered all cancer-related poison to concentrate only on previously oncogenic loci, which reduces the number of functional mutations in the whole cohort from 9 223 to 1 239. However, the prevalence of somatic missense mutations does not have an apparent effect on the disease progression and development [137].

Here I implemented the FLAGS dataset [97] that revealed significantly longer protein-coding sequences and genes that display less evolutionarily selective pressure than expected. For instance, the TTN gene is the largest known human protein [138]. Titin is a vital sarcomeric protein because it is implicated in the development of muscles and defines sarcomeric architecture. However, catching rare missense variants in the TTN gene in samples is reasonably common and does not signify the involvement of this gene in the cancer phenotype. Furthermore, some of the mutations appeared only in one patient, and their oncogenic activity is questioned. To see if one mutation is a rare cancerogenic mutation, I evaluated frequently mutated genes at the position level and performed a comprehensive analysis of co-mutations to identify cooperating mechanisms of leukemogenesis. I compared the most commonly mutated genes with cBioportal and saw that most of them are not oncogenic since they are mostly found in one sample unrelated to hematological malignancies and cancer type and are not described in OncoKB [139](Fig). One of the most commonly mutated positions was splicing factor SRSF2 (P95R). Splicing factor (SF) mutations are frequently mutated in hematological malignancies, and their relevance in the risk classification of acute myeloid leukemia needs further investigation. Furthermore, I performed a co-occurrence analysis. The co-occurrence of multiple mutations in cancer studies has earned growing awareness in identifying cooperating mutations or pathways that contribute to cancer. However, the

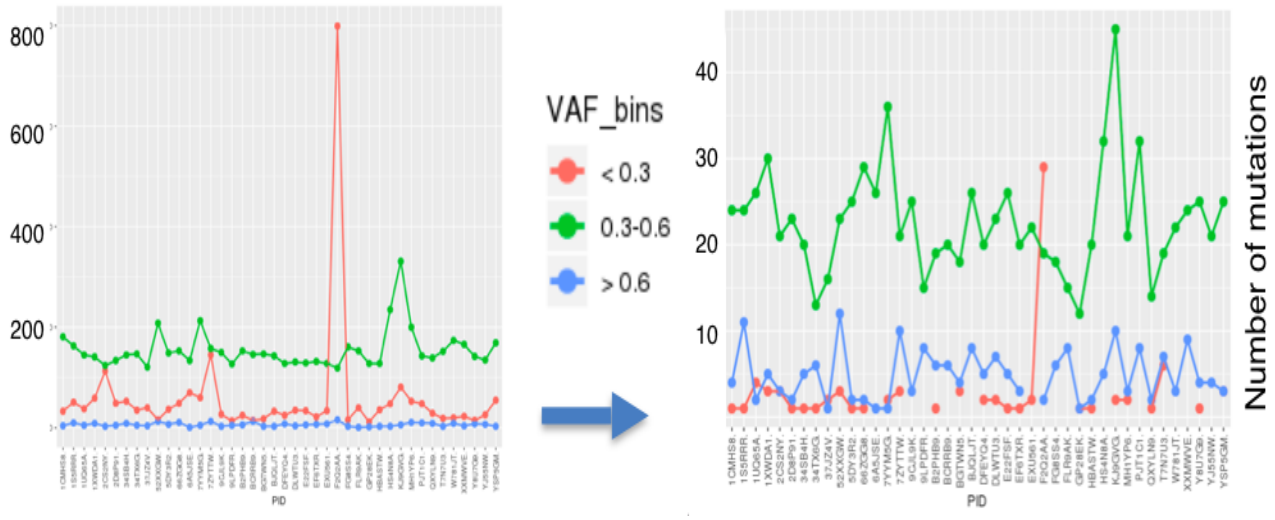


Figure 16: Number of mutations for each VAF after the no-control output (left) and after filtering only cancer-related positions (right)

data is heterogeneous, as with all other AML cohorts. Therefore the co-occurrence events from filtered cancer-related genes are not frequent. SRSF2 co-coocur in 3 patients with ZMI22, ZAN, SH3BP14, and ZNF66. However, the gene mutation found is not cancerogenic according to the cBioportal or dbSNIP database; therefore, I mark them as possible passenger mutations. The exciting funding is that I found SRSF2 and TET2 mutated in 4 patients, which directs to the idea that DNA methylation and alternative splicing join forces in leukemogenic development.

To prioritize the oncogenic capability of somatic mutations, I further go deeper into genes to rank actionable driver alterations. I used Oncovar, where an AI-Driver score $\geq$0.95 s was used, and mutations that appeared in a minimum of two patients were characterized as a driver, reducing the mutations to 736. Various methods have been designed to indicate whether a mutation is deleteriously founded on ideas including evolutionarily conserved regions, structural regulations, and the nature of amino acids. Nevertheless, cancer driver mutations characterized by these techniques lack consistency and tend to be false positives [140].

3.3.2 Expression of tumor alleles according to their DNA frequency

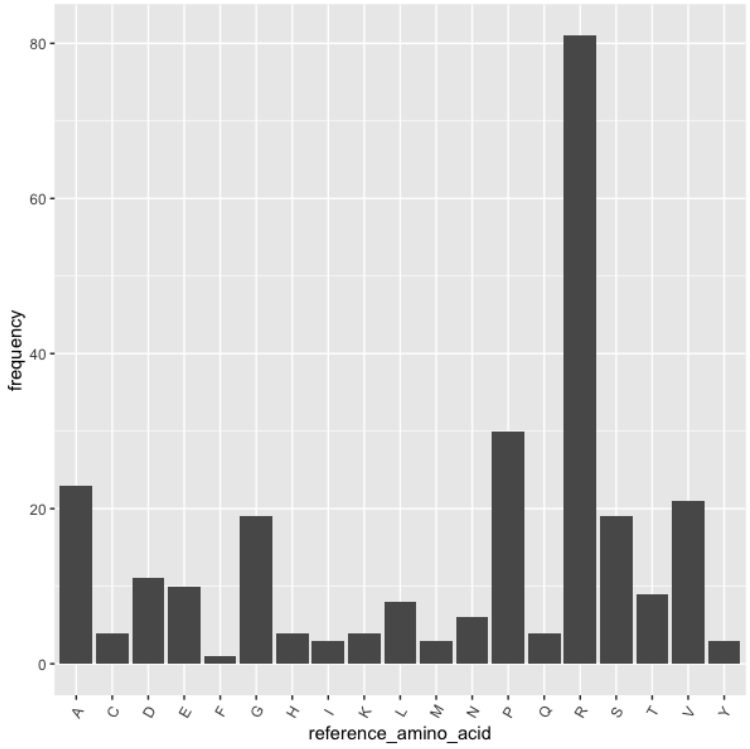
Moreover, although many methods predict cancer-driving deleterious mutations based only on public databases, surprisingly, a genome-wide study of combining publicly datasets and the relationship between DNA and RNA mutation allele frequency in tumor samples in tumor samples in mutation exploring has not been done. I presented a workflow combining the prioritization method that not only combines association and the biological relevance of genes but also considers the allele frequency and the expression relevance of a specific position. Unlike somatic mutation detection, which demands an inter-sample comparison (tumor and normal tissue), the finding of mutation allele frequencies is a self-normalizing intra-sample comparison and robust investigation of mutations.

For example, transcribing tumor mutations from DNA into RNA has significant clinical relevance. However, it is unclear if mutations are generally transcribed and, if so, at what proportion to the wild-type allele. I computed and matched DNA-RNA variants to control for sample mismatches, allowing us to correctly match the RNA-seq samples to their corresponding DNA samples. Therefore I studied the correlation between DNA mutation and RNA mutation allele frequency. I estimated the DNA

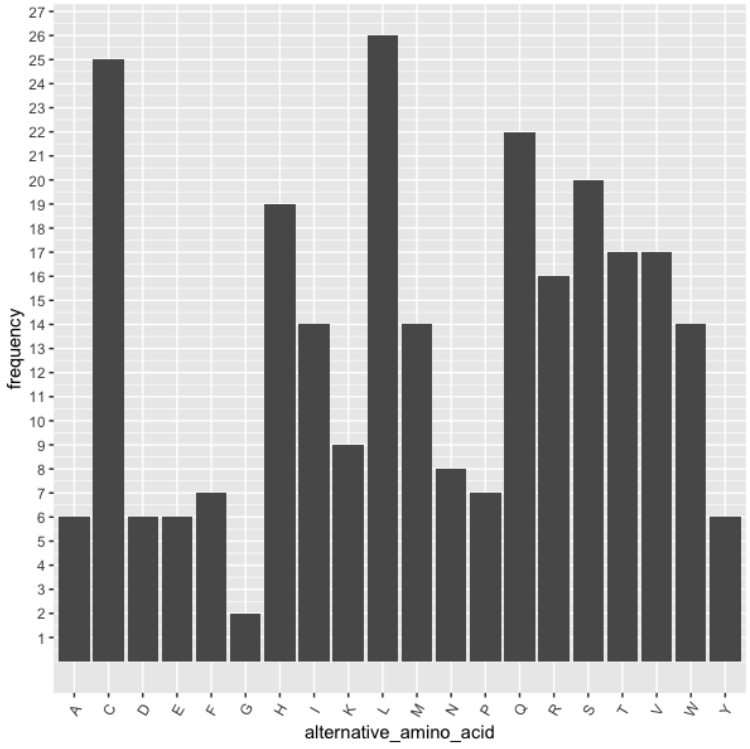
mutation allele frequency using the sequence range of the reads overlapping each SNV. Next, I defined the RNA mutation allele frequency using the NGS RNA reads covering the specified position. I next analyzed the concurrence between genome and transcriptome data. Of 736 SNVs, only 578 are located in generally expressed genes. Next, I uncovered that around 50 % of the DNA mutations in expressed genes are expressed as RNA (281 mutation locations are presented in the RNA reads, and 297 are absent). Therefore, the mutation allele fraction at DNA and RNA levels was estimated as the number of mutation-containing reads divided by all reads overlapping the SNV.

3.3.3 base substitution

Base pair modifications come in two different categories. Transition refers to a point mutation that changes a purine nucleotide to another purine (A to G or reverse) or a pyrimidine nucleotide to another pyrimidine (C to T or reverse). Transversions are single mutations when (two rings) purine (A or G) is changed for a (one ring) pyrimidine (T or C), or vice versa[141]. In other words, transitions are DNA mutations that preserve the same number of rings in the nucleotide base, particularly exchanging a one-ring pyrimidine with another pyrimidine or a purine for another purine. Alexandrov and colleagues lately reported a thorough analysis of mutational signatures, studying nearly 5 million somatic mutations from over 7,000 tumors describing 30 different cancer types. Transitions are more frequent than transversions, represented mainly by C>T mutations. In the cohort, most mutations are C to T transitions (102 mutations), followed by G to A transitions (93 mutations). Curiously the vice versa transitions are represented by just several mutations (A to G are 10 and T to C are 13). Cytosine to thymine transitions is the most abundant single-base change because cytosine is vulnerable to deamination, cytosines in CpG dinucleotides are often methylated, and deamination of 5-methylcytosine (5mC) produces thymidine. DNA methylation-mediated mutagenic events have intensely influenced vertebrate genome evolution since most CpG dinucleotide sequences have been mislaid. Therefore, CpGs are present only at about one-fifth of their expected random frequency in mammalian genomes, and only about 1% of all DNA bases are 5-methylcytosine pfe[142]. In addition, CpG dinucleotides are mutation hotspots in some cancer types [143]. Therefore not surprisingly, in the cohort, in each PID, around 40% of the reference base is cytosine. Therefore, I investigated the cohort's mutations that overlap with CpG sites. I mapped C/A, C/G, and C/T mutations to CpG sites regardless of their methylation levels. From 102 C to T transition mutations, 79 are located in CpG dinucleotides. Additionally, CpGs influence the mutation rates of neighboring non-CpG DNA [144]. Interestingly the 93 G to A transitions, 68 of them are located on the opposite strain of CpG dinucleotides. Totally 147 somatic exon mutations observed in my cohort could lead to loss of CpG sites. After analyzing methylation data, which will be discussed later in this thesis, I analyzed if there are some methylation changes in mutated CpGs. Since methylated CpGs are highly mutable, I would expect to observe that CpGs with high methylation levels are more likely to mutate. Indeed, I found that mutations overlap mostly with methylated CpGs. I also observed that a sample carrying the mutation has a methylation difference in a specific CpG position.



A.



B.

Figure 17: A. Frequencies of loss of amino acids as a result of mutations in the analyzed subset in 26 patients. B. Frequencies of gain of amino acids as a result of mutations in the analyzed subset in 26 patients.

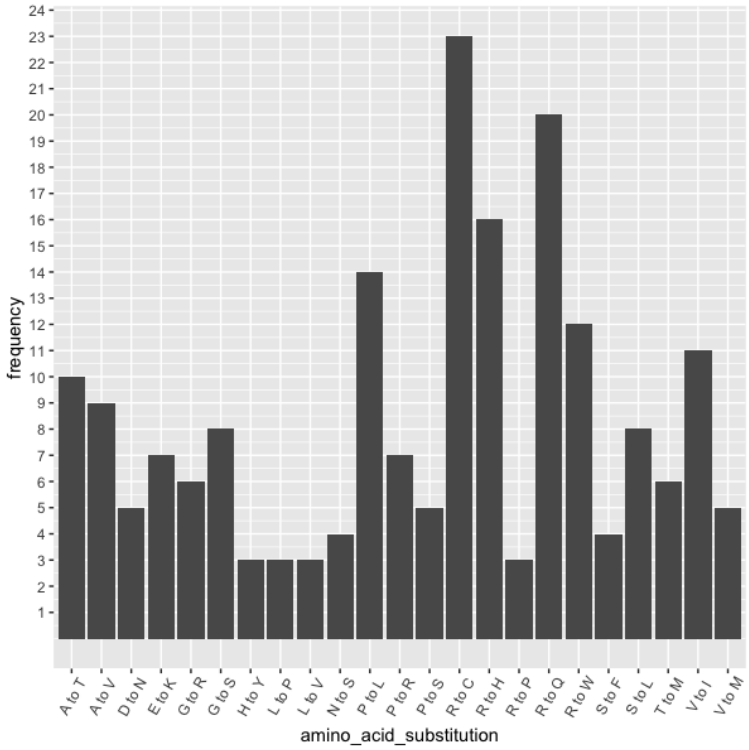


Figure 18: Amino acid substitution due to mutations in the analyzed subset in 26 patients.

3.3.4 protein domains

Mutations can be mapped onto protein structures, revealing how mutations distant apart on the linear DNA sequence could impact the same operational units in three-dimensional protein space. To test the functional impact of a mutation, I created a computational protocol that extracts cancer- and protein domain-associated variants and examined their possible outcome.

I analyzed data on amino acid substitutions in my expressed mutation cohort that resulted from 281 nonsynonymous single nucleotide substitutions in 252 genes. I further validated the results by comparing amino acid substitutions in the same set of protein positions in the COSMIC and TCGA database. Mutation in different sequence positions influences protein conformation and its function to different extents. Sometimes even a single amino acid substitution of the protein sequence can influence the entire protein structure and function [145]. The most familiar damaging substitutions are R > H, R > W, R > C and E > K. The high commonness of arginine can be clarified by its six codons, four of which have CpG dinucleotide, a common mutation hotspot [146]. A high percentage of cancer mutations convert the arginine (R) to glutamine (W) [147]. However, in my cohort, the most mutated amino acids are arginine (R) to Cysteine (C) in 10 different proteins. C > T transitions are caused by the deamination of CpG sites in cancer cells [148]. Mutations that arise during the lifetime of a cell, also called "Clock-like" mutational signatures, appear to be a significant creator of C > T changes in cancer cells. C > T transitions are caused by the deamination of CpG sites in cancer cells. Since methylation of CpG islands and areas within 2 kb of islands, called CpG island shores, it has been hypothesized that the elevated CpG mutation frequency is caused by increased DNA methylation of islands and shores, pursued by deamination of the methylated CpGs.

Single-nucleotide changes cause the vast majority of human genetic mutations. Consequently, single-

nucleotide mutations in amino acid codons cause amino transition of acid within the protein. I generated the amino acid reference and benign spectra to characterize the observed cohort scopes [149]. Firstly, I establish the exact aminoacid exchange caused by all nonsynonymous expressed mutations by mapping genomic coordinates from every genomic expressed mutation to within-protein sequence coordinates using EnsDb databases [150]. The donation of mutations at various amino acids to the cancer scope is highly heterogeneous. Interestingly, mutations at Arg residues contribute to almost 15% of the modifications. That is an effect of the well-known high mutability of Arg (as a consequence of deamination of 5'-CpG dinucleotides in Arg codons) and the relatively high commonness of Arg in human proteins (<4%) [147]. Furthermore, arginine is subject to several post-translational transformations, including methylation, acetylation, and ubiquitylation, that affect a broad spectrum of cellular functions such as epigenetic mechanisms and DNA damage response [151]. In my cohort, highly pronounced reference amino acid arginine is substituted in 81 proteins (Fig. 17). Arginine is a positively charged amino acid. Thus, it favors substituting for the other positively charged amino acid Lysine, though in some cases, it will also accept a change to different polar amino acids. As one could find in the TCGA mutation dataset [152], I found that nonsynonymous mutation results in a gain of cysteine, tryptophan, and histidine at the cost of a loss of arginine (Fig. 18). The loss of arginine may be attributed to the composition of its codons and the implication of arginine in protein function.

Missense mutations can cause proteins to be nonfunctional, especially when they fall into protein domains. Protein domains are operating units of proteins. They are responsible for a particular function donating to protein's overall role. Because of this vital role, most genetic variants occur in the domains [153]. A protein domain or region is part of a protein with a specific function and usually has a length of around 100 amino acids. In addition, distinct protein regions can adopt a particular 3D structure, for instance, zinc fingers, WD40 repeats, and leucine-rich repeats. To test the functional impact of a mutation, I created a computational protocol that extracts cancer- and protein domain-associated variants and examined their possible outcome. Therefore I implemented Prot2HG, a protein domains database, in my prioritization process. The resulting output reduced the number of mutations from 200, where I started, to just a couple of mutations pro samples. By comparing the output from my prioritization pipeline and control pipeline, I demonstrated exemplary performance in my prioritization process. Exact mutations in both patients from the HPS1 gene were found in both pipelines. Furthermore, SNX33 exact mutation was found in the control sample of HS4N8A.

I obtained the function details for mutated protein domains, which review the heterogeneity of my data again. In most samples, each mutated domain has a different function. Only in 6 samples I have three mutated genes. Since each gene is mutated in the same protein domain, I investigate their function. First, HPS1 is often found in Inherited thrombocytopenia and other bleeding disorders [154]. HPS1 is mainly involved in trafficking cargo proteins to recently formed cytoplasmic organelles [155]. The first point mutation in the HPS1 gene is at the R158H protein position and is also found within the same position and amino acid exchange in Monoclonal B-Cell Lymphocytosis, Chronic Lymphocytic Leukemia, and Small Lymphocytic Lymphoma. The second P259L is also tumor-related and found in Cutaneous Melanoma using the cBioportal dataset.

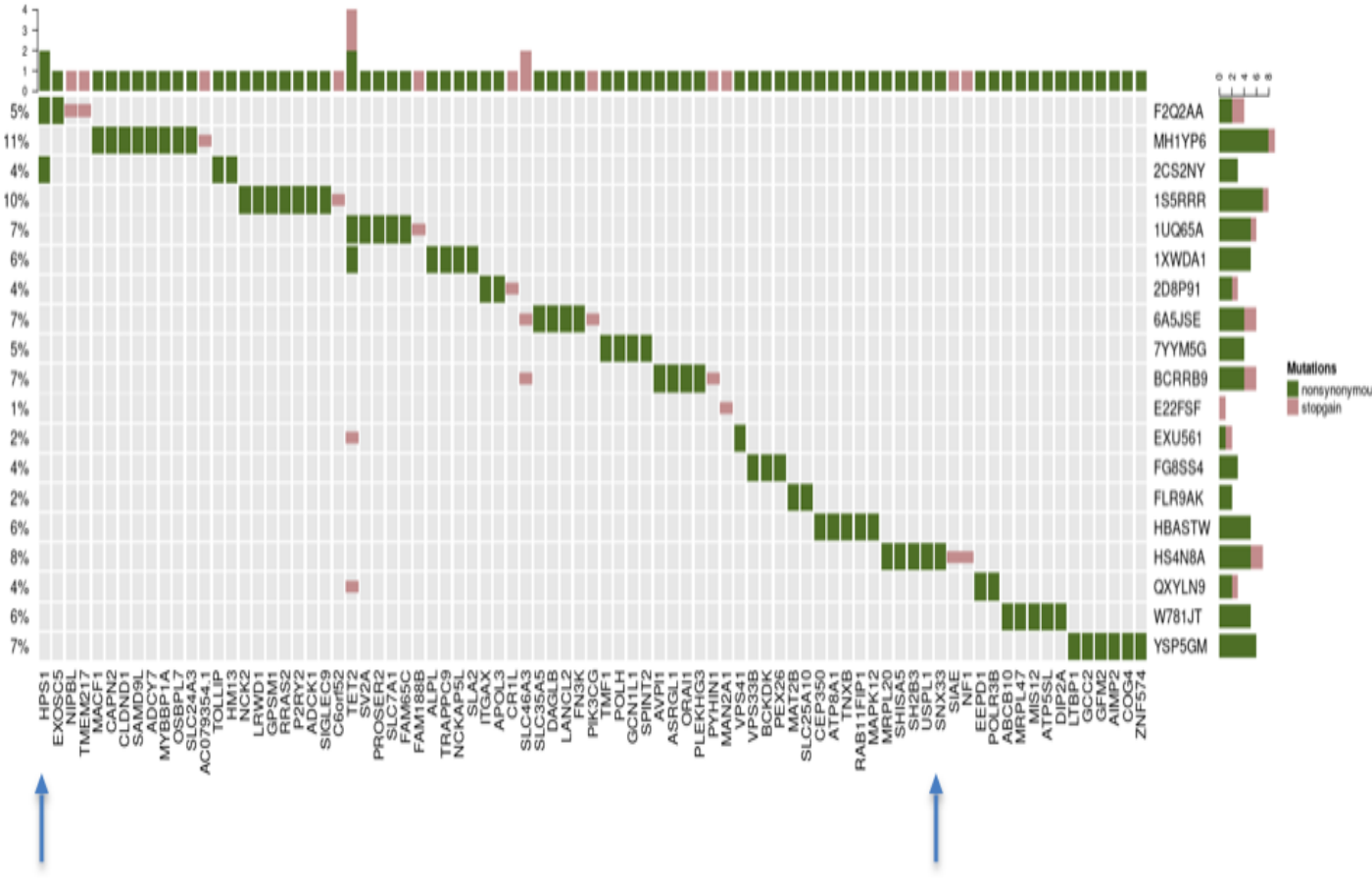


Figure 19: Oncoprint of mutations found in patients from my cohort with de novo acute myeloid leukemia who have mutations within protein domains. The blue arrow indicates the genes that are confirmed by the control pipeline. The color indicates the type of the mutation (green for nonsynonymous mutation and pink for stopgain mutation).

3.3.5 Fusion genes

Fusions are a type of somatic alteration showing in many cancer types and are associated with up to 20% of cancer morbidity[156]. Even in evolution, new genes assembled by gene fusion can occur between 2 neighboring genes and occasionally guide the change of novel complex domain structures [157]. For example, oncogenic fusion genes could lead to a gene product with a new or diverse function from the two fusion partners. Furthermore, the fused gene product can enhance the protein interaction networks in cancer [158].Gene fusions commonly exercise their oncogenic impact by either altering one of the involved genes (e.g. by fusing to a proto-oncogene), creating a fusion protein with oncogenic functionality (e.g. by inducing a constitutive activation of a tyrosine kinase) or causing a loss of function (e.g. by trimming a tumor suppressor gene). One typical example of a fusion is between promyelocytic leukemia (PML) gene and the transcription factor, retinoic acid receptor alpha (RARA), which is described in acute promyelocytic leukemia (APL) patients. The PML-RARA fusion protein drives lowered transcriptional activation and inhibition of myeloid differentiation [159].

WGS is powerful for finding possible driver genes, but the picture is different about fusion genes. Since fusion genes can arise from structural rearrangements like translocations and deletions and trans-splicing or read-through of neighboring genes, detecting fusion genes sometimes requires RNA-seq data [160]. Furthermore, RNA-seq yields only the transcriptionally active regions of the fusion, which can reflect on functionally relevant modifications in the cancer genome. Indeed, a substantial number of fusion genes have been detected in AML [161] [162] [71]. Because of biological significance and tumor-specific expression, some fusion genes are appealing diagnostic tools and therapeutic targets. One example is Tyrosine kinase inhibitors (TKI), which target BCR-ABL1 fusions [162]. Furthermore, the combination of ATRA and ATO targeting PML-RARA fusions has been demonstrated to be efficacious in acute promyelocytic leukemia (APL) [163][162]. Therefore, fusion gene studies may benefit leukemia patients by providing more future diagnostic markers and therapies.

I utilized the Arriba [105] algorithm to detect fusion transcripts in RNA-seq data. The predictions in Ariba output are listed as high, middle, and lowest confidence. Since the initial output gives us many recurrent fusion genes, I applied and built additional filters and strategies to filter possible artifacts. Next, I computed several characteristics associated with the fusion based on these alignments and can help evaluate them. First, I used the number of reads exclusively supporting each fusion as a proxy for the expression of the fusion transcript. Second, I computed the allelic fusion ratio for the fusion concerning each (5' or 3') partner transcript (5'-FAR and 3'-FAR) as the proportion of mutually exclusive reads supporting the fusion against each unfused partner gene. As a result, I could concentrate on high trust fusion genes. However, within, there are several highly confident Hemoglobin genes. Indeed in 23 of my samples, I have a fusion that the Hemoglobin (HBB) gene is involved (see supplementary). That is an artifact, and the reason is that shattered genomes often induce many fusions, most of which are passenger aberrations, and only a few are relevant to disease. However, passenger events will still be highly confident because they fulfill the above criteria. Therefore I filter out some fusion genes that are part of the "blacklist".

Since I assume that all of my samples are normal karyotype AML, I first check from my RNA-seq combined with WGS if there are some transchromosomal rearmaments. Indeed in PID B2PHB9, I found the rare but recurrent RUNX1-USP42 fusion gene is the outcome of a t(7;21)(p22;q22) chromosomal translocation, which has already been described in AML [164]. Likewise, in F2Q2AA, I found high-risk AML subtype fusion: NUP98-KDM5A. In 2 patients (52XXGW, GP28EK), I found KMT2A-MLLT4 gene fusion caused by an unbalanced translocation between chromosomes 6 and 11.In all other samples, there is no apparent transchromosomal translocation.

The predictions in Ariba output are listed as high, middle, and lowest confidence. Since the initial output

gives us many recurrent fusion genes, I applied and built additional filters and strategies to filter possible artifacts. As a result, I have several highly confident Hemoglobin genes. Indeed in 23 of my samples, I have a fusion that the Hemoglobin (HBB) gene is involved (see supplementary). That is an artifact, and the reason is that shattered genomes often induce many fusions, most of which are passenger aberrations, and only a few are relevant to disease. However, passenger events will still be highly confident because they fulfill the above criteria. Therefore I filter out some fusion genes that are part of the "blacklist" or not most probably decease causing. I further focused on decease-causing genes and concentrated on in-frame fusions that may retain protein domains essential for oncogenic function. This post-filtering process allows us to interpret the results better and point to the correct conclusions. For example, from the high confidence Ariba output, there are several solute carrier (SLC) membrane transport proteins (SLC19A1, SLC11A1, SLC25A37, SLC7A5) (see supplementary). Different SLC proteins are involved in fusion in several cancer types playing crucial roles in cancer transformation.[165][166][167].

I indicated open reading frames for these fusions and separated them into three categories regarding the frame of the fusion genes: inframe, frameshift, and no frame information (by breakpoint at UTR, intron, or non-coding RNA). I further focused on decease-causing genes and concentrated on in-frame fusions that may retain protein domains essential for oncogenic function. I found 16 high-confidence fusions, all between neighboring genes. They arise from small deletions, read-through, and inversions. From my filtered high-confidence fusions, one or both fusion partners have been described to be part of the cancer fusions according to Fusion Gene annotation DataBase (FusionGDB) [168]. I found 16 fusions that include genes involved in carcinogenic processes (Fig 20 A.)

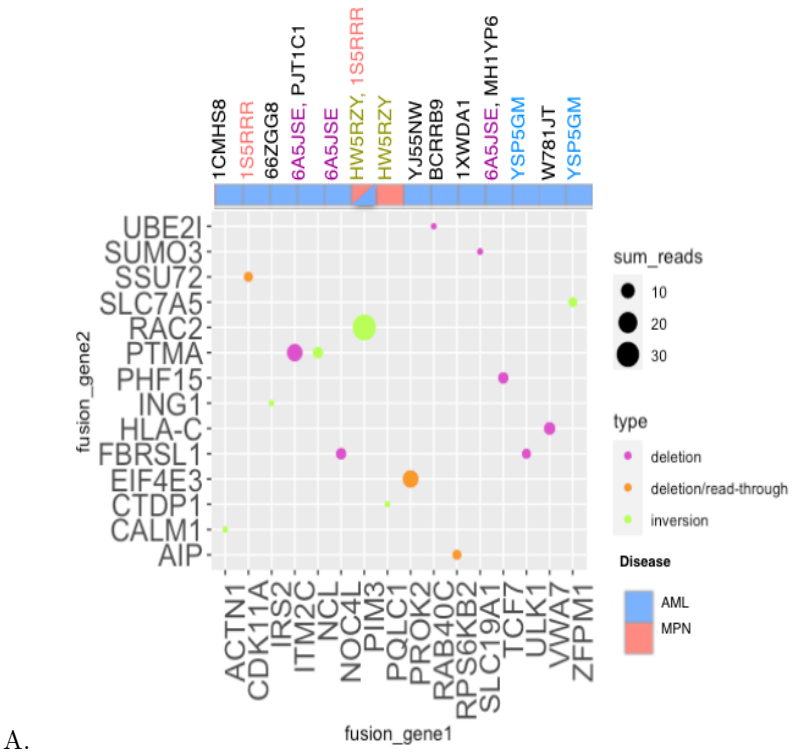
Modifications in tumor suppressor genes can lead to tumor genesis or the uncontrolled growth of cells [169]. Fusion could disrupt tumor suppressor genes and the resulting pathogenic effects. I have two tumor suppressor genes in 2 patients, ING1 and AIP. The ING tumor suppressors function on the histone epigenetic code, affecting DNA repair, chromatin remodeling, cell cycle regulation, and apoptosis [170]. ING is fused to cyclin-dependent kinase (cdk11a,) a protein kinase that regulates RNA transcription, splicing, and mitosis [171]. Aryl hydrocarbon receptor interacting protein AIP encodes a co-chaperone protein with tumor suppressor properties and predisposes tumorigenesis [172]. Similarly, in Patient Id YJ55NW, one can find the PROK2-EIF4E3 fusion, where PROK2 reviews oncogenic function in many cancers [173] and eIF4E3 acts as a tumor suppressor by utilizing a methyl-7-guanosine cap recognition [174].

I found several fusions through small inversions: ACTN1-CALM1, PIM3-RAC2, IRS2-ING1, NCL-PTMA, PQLC1-CTDP1, and ZFPM1-SLC7A5. In fusion is between IRS2 and ING1. Insulin receptor substrate 2 (IRS2) is a candidate driver oncogene frequently amplified or fused to another gene in several cancer types [175], whereas ING1 is a well-known tumor suppressor gene [176]. Another exciting fusion is NCL-PTMA fusion, which has already been described in [177]. NCL regulates mRNA translation and stability of several tumor progression genes [178], and PTMA, is a cell survival and proliferation protein, which has often been described as a fusion protein in AML [177]. Interestingly in my cohort, PTMA is fused with the cancer-related ITM2C gene.

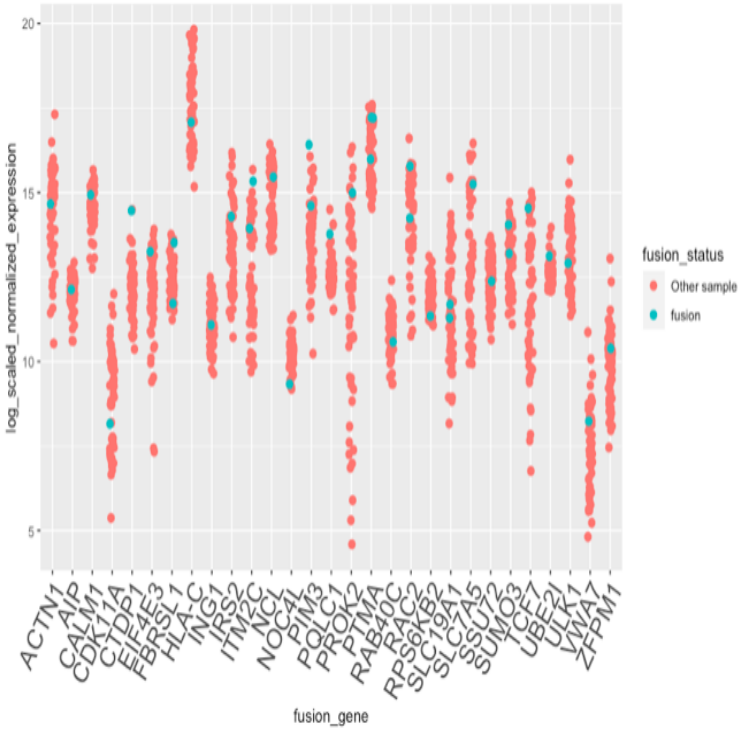
Some oncogenic kinase fusions are sensitive to kinase inhibitors, implying that additional therapeutic candidates might be discovered by studying fusion transcripts affecting protein kinase genes. The PI3K pathway mediates different cellular processes, including cell survival, migration, and proliferation.In 2 patients (one AML and one MPN) PIM3 gene, which is Ser/Thr protein kinase family, is fused with a small GTPase, Rac2, a gene controlling tumor growth and metastasis [179]. The Unc-51-like autophagy activating kinase 1 (ULK1) plays the leading role in the initiation step of autophagy [180]. ULK1 is fused with a function unknown FBRSL1 gene [181]. The retention of critical functional domains (FDs) is essential in assessing whether it plays an oncogenic role and has clinical relevance. Driver fusion genes typically retain functional domains (e.g., DNA-binding domains or kinase domains) [59].

Most in-frame fusions genes retain major functional domains. of kinase Ssu72 regulates sister chromatid cohesion and the split of duplicated chromosomes during the cell cycle. Furthermore, Ssu72 is involved in transcription by interacting with transcription initiation and termination complexes. Ssu72 is needed for 3' end cleavage of pre-mRNA but is dispensable for poly(A) expansion [182]. Cyclin-dependent kinase 11A (CDK11A) is a member of the serine/threonine protein kinase family that play multiple roles in cell cycle progression and cytokinesis [183]. The CDK11A protein location is in the nucleus and cytoplasm, whereas the Ssu72 cytoplasmic protein. Since this fusion is not before described there would be interesting to examine its function.

I explored if the fusions regulate gene expression because fusion events can be associated with one or both fusion partners' altered expressions. Therefore, I systematically integrated gene expression and fusion annotations to test for associations between gene expression and fusion status. I specified whether that sample was an expression outlier for the selected gene for each fusion. Fig 20 B. shows that no fusions displayed outlier overexpression of the fusion partner.



A.



B.

Figure 20: Fusion genes found in my cohort. A. The dot plot indicates the gene partners for each fusion. The dot size displays the number of reads containing the fusion point. The color shows the type of fusion. . B. Expression of high confidence fusion. Blue dots indicate samples for which the gene is involved in a fusion, and red dots indicate the expression value for the other samples. Generally, the expression of these genes does not have a significantly higher expression when fused than not (two-sided adjusted Student's t-test, P-value > 0.01). Only CTDP1 shows a separation of fused samples versus others.

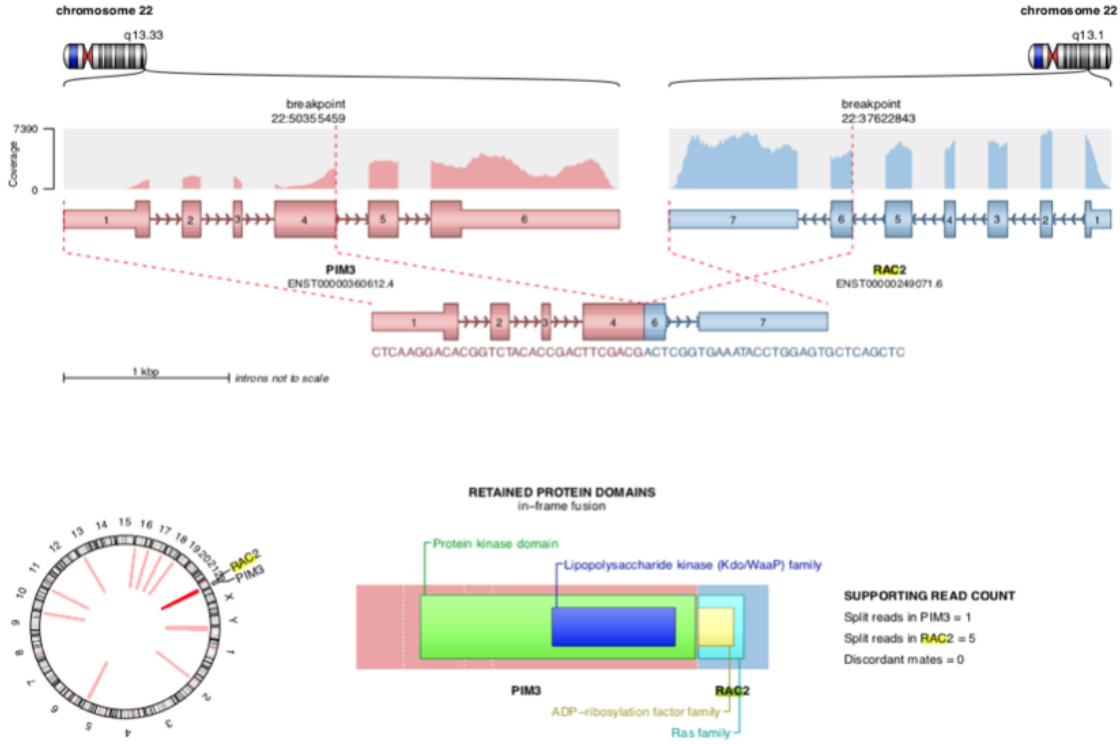


Figure 21: The figures represents the Arriba's predictions of the exons retained in the fusion, as well as protein domains that are retained in the fusion.

3.3.6 Clinical implementation of RNA sequencing for AML gene drivers

A significant element in translational cancer research is biomarker discovery using gene expression profiling [184]. The transcriptomic analysis is an essential and complementary tool for the whole exome and genome sequencing and can potentially uncover new biomarkers, which will significantly affect early diagnosis and treatment strategies and prevent complications. Besides improving the diagnostic yield, RNA-seq can enhance knowledge of molecular pathomechanism and basic genetic mechanisms [185]. Furthermore, In a rare disease diagnostic setting, the methodological design differs from the well-established differential expression analysis workflows. In rare diseases, every case has its genetic cause of disease, even though the consequent phenotype can be similar. Therefore identifying exceptional down- or upregulated genes may lead to new candidates for therapeutic intervention and reply biomarkers for partner precision medicine approach. Furthermore, several tumor suppressors have an outlier sensitivity pattern, supporting and generalizing the notion that tumor suppressors can play context-dependent oncogenic roles. Therefore the performance of aberrant gene expression is still required for adopting RNA-seq in systematic diagnostics [185]I prioritized genes demonstrating transcript with significant allelic imbalance and aberrant expression.

For each sample, in every expressed mutated position, I estimate the counts of the variant and reference reads across the DNA and RNA matching datasets. Variant probability is a biologically interpretable quantitative cognition of the VAF distribution [186]. For example, in VAF (DNA) from a diploid genome, I assume variant probability equal to 0.5 (meaning that both alleles are equally probable). Differences between VAF (DNA) and VAF (RNA) values indicate if the mutated allele is generally expressed and where one of the alleles is preferentially transcribed over the other. Next, I estimated the outliers for proof of biological processes. Finally, I developed a metric for estimating RNA versus DNA mutation allele imbalance:

Imbalance = (DNA mutation allele frequency) minus (RNA mutation allele frequency)

The RNA and DNA mutation allele fractions are comparable for the most found mutations, and the imbalance is close to zero. However, there are four different genes where the alternative or the reference allele is preferably transcribed (Fig. 22) . For example, I have 12 TET2 mutations, all in different positions. However, only in the patient ID (PID) 1XWDA1, the TET2 mutation, with a protein change D1376N, the alternative allele is preferentially expressed than the reference ones. The other alternative allele overexpression is G to C mutations causing E79E protein change in the PPM1L gene, 7YYM5G patient id. This finding is exciting since, in the same gene position, but an E79K protein change is found in Skin Cutaneous Melanoma TCGA PanCancer data. The protein encoded by this gene is a phosphatase that play a role in apoptosis and cytotoxic stresses [187]. The function of this mutation is hard to predict since it is out of the Protein phosphatase 2C protein domain. However Q72L protein change of the RRAS gene and the G39R protein change of the IWS1 gene, the alternative alleles are less expressed than the other normal allele, although the mechanism for this is unclear. Disease-causing RRAS mutations activate and act by preserving the GTPase in its GTP-bound active state [188]. Therefore even if the reference allele is preferentially expressed over the alternative, mutations in this gene often contribute to leukemogenesis, and there still could be an altering effect caused by this mutation. A leukemia study shows that allele-specific expression could be determined by the DNA methylation status of the corresponding gene [189]. However, I did not find evidence of any DNA methylation epigenetic silencing allele-specific expression, preferentially promoting mutated or wild-type alleles. Instead, the beta value in each cpg from the gene bodies or promoters of those genes has similar values.

Dominant expression of the alternative allele

GENE	PID	VAF	allele fraction	diff
TET2	1XWDA1	0.5142857	1.0000000	0.4857143
PPM1L	7YYM5G	0.4568966	0.8095238	0.3526273

PPM1L Regulates Hematopoiesis

Dominant expression of the reference allele

GENE	PID	VAF	allele fraction	diff
RRAS2	1S5RRR	0.5689655	0.3137255	-0.2552400
IWS1	BCRRB9	0.4634146	0.2195122	-0.2439024

RRAS2 gene: hotspot package

Figure 22: List of genes in my cohort that tend to be expressed dominantly from alternative and reference allele

3.3.7 expression outliers

Overexpression of wild-type genes, even without oncogenic mutations, could serve as an oncogene and significantly increase cell survival [190][191]. For example, it has already been found that in datasets from patients with acute myeloid leukemia (AML) with normal karyotype, MN1 and NCALD overexpression are prognostic markers and foreshadow poor prognosis [192][193]. I used different approaches to determine expression outliers. Similar to [194], I computed Z scores on the TPN-normalized read counts by subtracting the mean count and dividing by the standard deviation. Expression outliers were identified as reading counts with an absolute Z score greater than 3. However, this approach did not find any plausible pathogenic candidates. The reason for this approach is that one possessed for variation in sequencing depth by using transcripts per million (TPN) values in the z-score calculation while disregarding other possible confounders. I also used DESeq2 [195], a method designed for differential expression analysis, to test each patient against the remains of the AML cohort using a negative binomial (NB) distribution. While both methods regulated the covariations in the RNA-seq read count data, they neither estimated aberrantly expressed genes. As with most cancer cohorts, replicates from ours are not available. Therefore, a standard case versus control comparison cannot be applied here. However, it is notable that DESeq2 [195] and edgeR [196] already have approaches to down-weight outliers.

Nevertheless, these procedures seek to improve the robustness of the model fit by extracting outliers instead of considering their significance. Therefore, I used OUTRIDER (Outlier in RNA-Seq Finder), a deep learning method for outlier detection in RNA-seq samples, which controls for covariations among the gene read counts [103]. The package applied a standard and denoising autoencoder schema to control for known and unknown covariation in the read count data [197]. Furthermore, the outrider reduces the number of outliers per sample, which I also observed in the data. The result shows 4 upregulated genes in 3 samples and 13 downregulated genes in 8 samples (Fig 23 A). To compare the

findings with the expression profiling of other cancer entities, especially leukemia cohorts, I used the GEPIA, a web server, to compare the expression outlier findings with expression TCGA data [198]. Intriguingly in 2 AML AML samples, there is overexpression of LIN28B. LIN28B is an RNA-binding protein whose expression is crucial during human embryogenesis and is down-regulated in most tissues after birth. LIN28B is an RNA-binding protein whose expression is crucial during human embryogenesis and is down-regulated in most tissues after birth [199]. Consequently, LIN28B is implicated in malignant transformation partly because it promotes the degradation of the let-7 family of miRs. Lin28b expression is associated with undifferentiated cell states, and its overexpression has been reported in around 15% of human malignancies enclosing a broad spectrum of tumor types [200]. In particular, LIN28B is oft over-expressed in AML, more frequently in pediatric acute myeloid leukemia [201]. Furthermore, a study already has demonstrated that targeting LIN28B in AML cells results in cell cycle arrest and inhibition of cell proliferation [202]. Another overexpressed gene in the AML cohort is another RNA-binding protein, KHNRNPK. It has already been identified that hnRNPK overexpression is a recurrent abnormality in AML that is associated with poor patient outcomes [203]. There are 13 downregulated expressed outliers. Two of them, HSD17B8 and VOPP1, are also found to be downregulated in AML cohorts, according to the GEPIA server. On the other hand, other genes like RAB34 and MGMT are shown to be upregulated in AML and cancer in general but downregulated in the cohort in corresponding samples: FG8SS4 and EF6TXR. Therefore, I analyzed AML expression data from TCGA with related survival data. I aimed to look for a sample with the same outlier genes (up- and downregulated) found in the dataset. The outers expression survival were statistically evaluated using the Kaplan–Meier method, the log-rank test, and multivariate evaluation by Cox regression analysis. Survival curves showed no significant difference in prognosis related to the expression of mine up- and downregulated genes.

A.

Upregulated genes		
	sampleID	geneID
1	E22FSF	UBA52
2	E22FSF	TCEB2
3	E22FSF	POLR1D
4	7YYM5G	RP11-506O24.1
5	QXYLN9	RP11-71N10.1

Downregulated genes		
	sampleID	geneID
1	FG8SS4	PCCA
2	FG8SS4	RAB34
3	2D8P91	HSD17B8
4	1S5RRR	ARHGEF11
5	1S5RRR	CTIF
6	FLR9AK	MDFIC
7	FLR9AK	ZNF597
8	FLR9AK	RP3-449O17.1
9	QXYLN9	ZPR1

B.

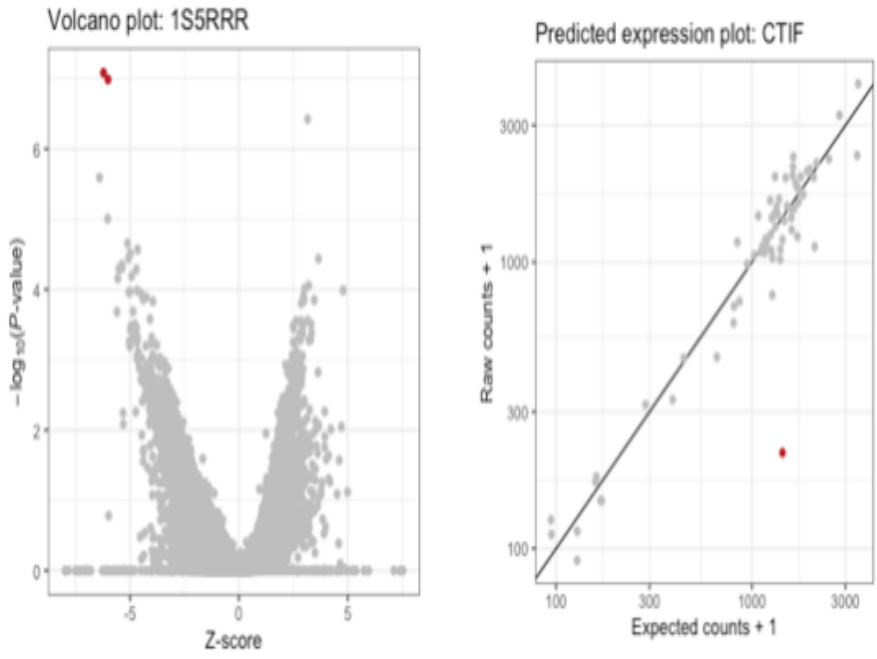


Figure 23: Expression outliers. A. Table displays the over- and downregulated genes. B. Left: volcano plot displaying the distribution of P-values on a sample level. The chosen example is patient id 1S5RRR, with two downregulated outliers marked in red. Right: Dot plot displaying the observed versus expected gene expression. Displayed is the result for the gene CTIF in patient id 1S5RRR.

3.3.8 methylation profiling

Unlike genetic mutations, the reversible character of epigenetic changes is appealing in cancer medicine. DNA methylation is an epigenetic conversion that controls gene expression and plays a crucial role in many processes in the hematopoietic system. DNA methyltransferases (DNMTs) and Ten-eleven-translocation (TET) dioxygenases are reliable enzymes for regulating DNA methylation. Mutations of DNMTs or TETs alter normal hematopoiesis and result in preleukemic and leukemic states.

A total of 142 DNA samples from 34 patients were analyzed on Illumina EPIC DNA methylation arrays.

As explained in Materials and Methods, the array has two types of probes: Type I and II. Type I investigations estimate methylated and unmethylated CpGs in the same color channel, demanding presentation by two beads at the exact location. In contrast, using a single bead, Type II probes measure methylation status in separate color channels. The distribution of beta values for both probe types is bimodal and, when plotted as a function of density (using a Gaussian smoothing function and bandwidth = 0.05), displays two clear peaks resembling low methylated and high methylated probes (see supplementary). Both biological and technical factors probably drive the divergence in distributions. A subset quantile normalization approach was used to normalize sample technical differences. I further remove any probes that could interfere with methylation level variation due to genetic differences between samples, namely the gender and mutation difference that can disrupt the methylation outcome.

A technically sound dataset would allow verification of known facts. Dimension reduction is a decisive way of visualizing relationships between variables and global DNA methylation data trends. However, when I applied PCA on the 10000, most variable CpGs in the data did not show in a visible grouping of samples (see supplementary). I could not observe evident clustering of samples based on a patient on specific mutation or driver status.

Somatic mutations in the epigenetic modifiers DNMT3A, IDH1/2, and TET2 are triggering mutations in AML, particularly those with normal karyotypes (welch2012origin). Furthermore, their modification and loss of function could recapitulate the human epigenetic profiles. However, in the data, I didn't find a correlation between mutation and the global methylation change of the corresponding samples (see suplementary). Furthermore, I were unable to detect TET2 mutation subtype-specific DNA methylation. In Fig. 24 A. I can observe different methylation percentages in the samples, most probably coming from a noise or batch effect. F2Q2AA shows a clear shift in his methylation signature and is one ample pretreated with Daunorubicine. The changes come from the unknown or possible accumulation of batch effect. Therefore I performed the PCA of the potential biological and technical batch effect that could explain the following results (see S5). I used the undocumented normalization method established by Prof. Plass's lab (DKFZ) for batch correction from an unknown source. I used this flow to generate newly normalized beta values and, afterward, performed the exploratory downstream analysis (Fig. 24B.). Epigenetic changes have been shown to contribute to the pathophysiology of AML [37]. I thus investigated DNA methylation patterns along the NKAML disease. I first conducted a differential methylation analysis between the nkAMLs and 12 BM-MSC samples derived from healthy donor controls from a publicly accessible dataset with an accession number GSE79695. As explained in Material and Methods, the threshold for DMRs was established as a total value of $p<0.05$; $\delta\text{-}\beta > 0.2$ was described as hypermethylated sites, and $\delta\text{-}\beta < -0.2$ was evaluated to display hypomethylated sites. As expected, the AML methylome was shifted to hypomethylation in many positions. From the 413006 sites that were compatible between the two types of array, 260825 of them are hypomethylated. AML, in general, and the nkAML cohort shows a global hypomethylated genome, which may lead to genomic instability and aberrant gene expression. Therefore, I further make a combinatory analysis between methylation, expression, and fusion genes. Finally, I investigated if the methylation in or around a specific gene could explain the expression outlier or a

reason for read-through fusions. However, I did not find a tangible link between DNA methylation and gene expression. Furthermore, the altered expressed genes and fused transcripts are without significant alterations in the DNA methylation pattern, suggesting that DNA methylation may not be crucial for these findings.

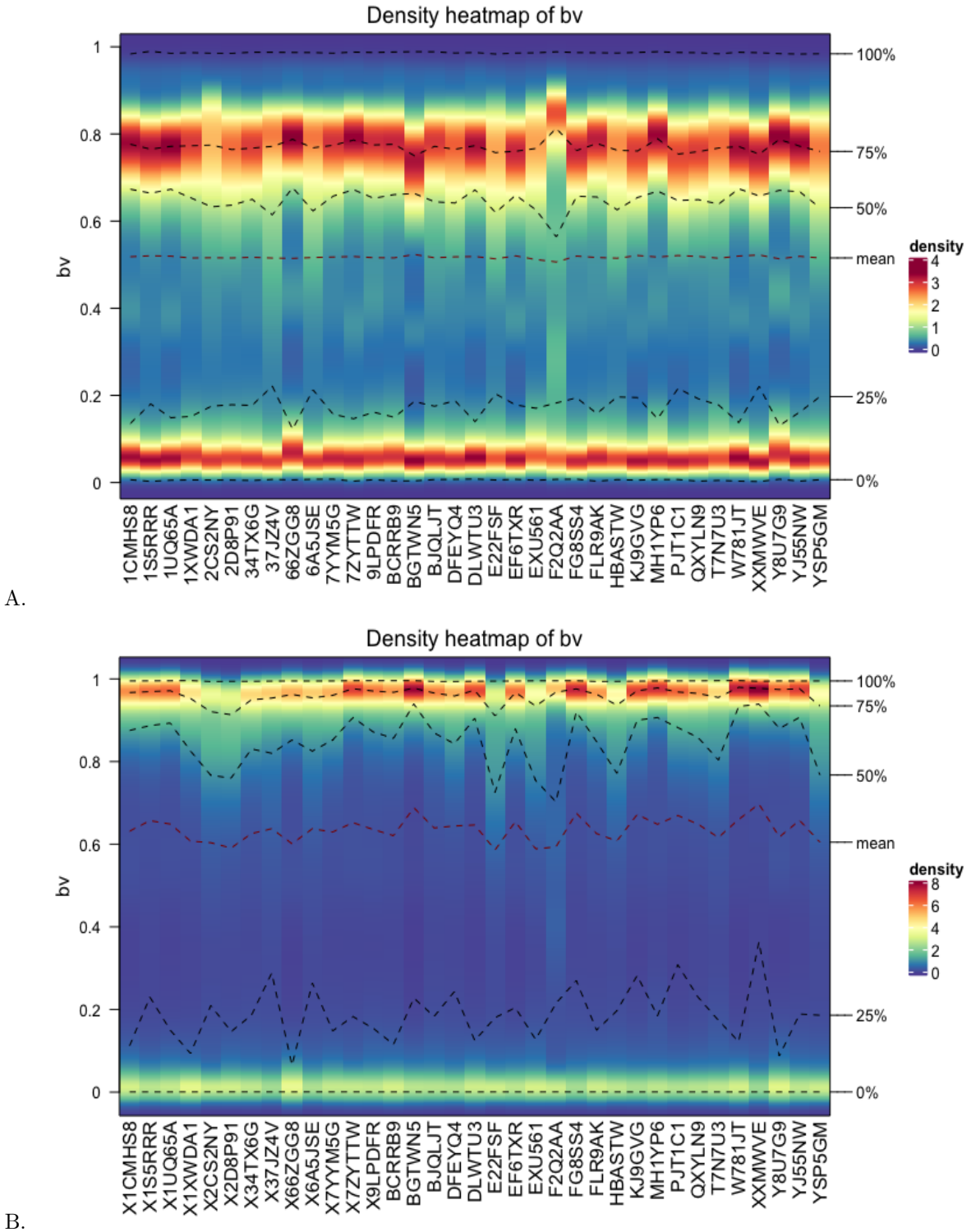


Figure 24: Density heatmaps show the distribution of the beta values (bv) for each sample before (A.) and after normalization (B.).

4 Discussion

4.1 Applying the no-control pipeline

4.1.1 The no-control pipeline successfully removes most germline variants and can find recurrent mutations in a cohort

Coding somatic mutations play the leading part in tumor initiation and progression. With the advent of whole genome sequencing, the systematic screening of such modifications is possible. Although mutation detection within next-generation sequencing information have been the basis of many analyses, cracks and weaknesses remain. The detection of population variants or somatic mutations is usually performed in matched tumor-normal samples. Especially in cancer research, it is essential to determine genetic variants in the tumor compared to the non-malignant tissue of the same individual (i.e., matched control), part of which are considered to be responsible for the carcinogenesis process. Therefore, somatic variant calling pipelines determine the cancer-specific somatic variants. However, in clinical settings not every sample has matched control due to sample availability, quality issues, limited budget, and others.

Notably, leukemia samples often come without a matched control sample. This study demonstrates the detection of frequent mutation from WGS using the no-control bioinformatics tool. I applied and adapted the pipeline and evaluated the results. Although common germline variants can be recognized with germline databases like dbSNP [93] and ExAC [92], identifying rare variants demands better methods. There are only a few described tools for performing this job: proposed by Hiltemann et al. [204], LumosVar [205], and MuTect [206]. I have applied and adapted the no-control mutation calling pipeline to detect cancer driver elements and gain insight into the underlying mechanisms of tumorigenesis.

Besides the above-mentioned germline databases, I removed mutations found in our in-house control dataset (among 280 controls). However, any clinical-related variants were rescued if they had corresponding OMIM records[207]. Thanks to the workflow system Roddy, the pipeline is employing the reliable DKFZ cluster system to complete the analysis in less than a week. In General, the no-control somatic variant calling pipelines are a huge help when the matched control is not available or unusable due to bad quality. First, after subtracting millions of common variants from public variant databases, most parts of germline variants were successfully filtered, resulting in a relatively small number of individual, rare, and somatic-like variants. After subtracting the common variants, the SNVs found in coding regions (functional SNVs) was around 200, from the expectancy with a matched control output would be about 50. The pipeline was previously used for other published studies [208] [209] and showed similar variation numbers. It was validated that the no-control pipelines can determine most somatic variants uncovered by somatic variant calling with a matched control. Since the remaining variants can retain some germline variants, the no-control method is beneficial for finding frequent mutations in a cohort. It was used for two projects requiring the modification calling without a matching control: t(8;16) AML and no-driver nkAML.

4.1.2 Evaluation of mutation of t(8;16) AML

t(8;16) AML is a rare and distinguishable clinicopathological entity. Only a few previous reports represented the mutational characteristics of patients with this type of AML [210][211][212][213]. Studies suggest that the t(8;16) translocation could be the only reason to induce hematopoietic cell transformation to AML without obtaining other genetic alterations[213][44]. Here I evaluated the frequently mutated genes and compared them with the most frequent mutated genes in AML in general and AML carrying t(8;16) translocation.

I could further confirm some mutations that are common for AML and found explicitly in the t(8;16) AML The only co-occurrent mutation in this type of AML is RUNX1 [213]. I found this mutation in 2 of our patients. However, in 10 patient cohorts, I found other candidate genes which might collaborate

with t(8;16) to result in leukemia. Even further, some of the mutations can also review the therepy prediction. Although TP53 is the most commonly mutated gene in cancer, its occurrence is observed in only 5-10% of de novo AML [214]. Even though it is not typical for this type of AML, I found TP53 mutations in 3 patients. Mutations in FLT3 are the most common genetic alteration in AML, identified in approximately one-third of newly diagnosed patients, and are found in (3.5%) of the patients with t(8;21) AML [215]. FLT3 mutation screening for internal tandem duplications (ITD) and point mutations within the tyrosine kinase domain (TKD) was found in 3 patients of our cohort. As a potential target for therapy with tyrosine kinase inhibitors, studies have already described using of midostaurin or gilteritinib [216]. Mutations in SETD2 in one of our patients involves the progression and development of chemotherapy resistance in acute myeloid leukemia [213].

In addition, neither mutations in EYS, KRTAP9-1, PSIP1, nor SPTBN5 were depicted earlier in AML. However, recent publications on MLH1 mutations indicate that they may pla y an essential role in AML [217]. From this list, KRTAP9-1 mutations was found in 5 patients. Although KRTAP9-1 has no widespread registered role in cancer, a recent study reported that they could play a role in malignant progression [218]. SPTBN5, EYS was found in 4 of 10 patients, indicating the significance of these mutations in our cohort.

4.2 nkAML

The reporting of potential oncological relevance findings from mutations, gene expression, fusion genes, and methylation changes rapidly expands into the clinical area [219] [220][221][222][223]. In this work, I aimed to present cancer drivers from different types of data and find new drivers mechanisms for nkAML. Furthermore, the approach attempts sequential filtering of various layers of multi-omic information to assist in the driver decision process. Therefore, we need a robust method for personalized driver prioritization. Despite the importance of driver calling in research and clinical care, no previous precision methodology has been built in an automated fashion. Here I provide a multi-omic tool to address several facets of this field. I developed a pipeline to look for driver events on different omic layers, integrate the output, and compare the publicly available datasets. His complexity includes prioritization of DNA events, fusion genes, as well outlier expression events, and DNA methylation modifications.

4.2.1 prioritization of mutations

First, our mutation prioritization is methodical, involving expected and unexpected challenges while estimating the expected impact of each position of the mutation. Despite years of advancement, mutation detection in cancer requires many manual examinations as a final step. In addition, expert assessment is incredibly challenging in cases where tumors are sequenced without matched control DNA. As discussed previously, the no-control pipeline removes many germline variants and could be used to find frequently mutated genes. However, estimating the driver mutation among 200 variations output in each patient is incomprehensible.

A driver gene is expected to be mutated in a high percentage of samples (more than 20%, for example) to identify such a gene as recurrently mutated correctly. However, in heterogeneous diseases like AML, that is a rare case. Our results imply that the sequencing data of an individual could help recover the signatures of cancer-related genes. However, for a limited sample size, detecting all the driver genes or distinguishing oncogenes by population mutation frequency is hard. Even more, since the output of a no-control pipeline still contains has many rare and germline variants, it is often the case that recurrent findings are artifacts, germline variants, and, most often, passenger events. A method for evaluating hotspot significance must adequately account for site-specific mutability variation. In many research projects, gene set enrichment methods have been developed and widely used [224] [225]. These tests may overcome some difficulties, but they have significant limitations. Many annotated gene groups

are extensive, holding dozens of genes. Furthermore, pathways are interconnected in more considerable signaling and regulatory networks and never act in isolation. Therefore, gene-set methods need to pay more attention to the topology of interactions instead of regarding all genes within a pathway equally, and also, they reduce the capacity to catch driver mutations in not-so-well-studied pathways. Several other methods have been developed to uncover recurrent mutations in a bigger cohort of cancer patients, including MutSigCV [226], MuSiC [227], and others [228], [229]. The methods above concentrate on known driver gene detection but do not aim to propose personalized standards of diagnosis or treatment since a particular patient could have other compositions of driver genes. Even though a database for hotspots has been already established [111], it doesn't assign them to drivers or passengers. Besides, the method applied a computational or manual filtering criterion that may remove hotspots without mechanistic evidence. In our pipeline, I focused on patient-specific driver prioritization.

In the no-control setting, the analysis workflow should deliver comprehensive and interpretable outcomes. Therefore, I created a workflow tailored for heterogenous diseases like AML with no available healthy sample. Most methods for driver gene prioritization are developed for investigating tumor cohorts. However, these techniques usually could not recognize low-frequency driver genes and certainly patient-specific driver genes. Furthermore, patient-specific driver genes could be rare or not match the specific AML driver genes. This study demonstrates that our pipeline can fast and successfully analyze patient-specific data to help prioritize cancer drivers. This approach consecutively filters and presents layers of findings relevant to cancer. Our method uses AML- and cancer-associated variants datasets to report clinically relevant results successively.

Non-synonymous mutations are considered largely deleterious due to their property of changing amino acids. The same goes for nonsense mutations that generate truncated proteins. Although even synonymous mutations undergo natural selection, they are rarely the cancer drivers. I used the TCGA cancer datasets covering 33 cancer types, 10 182 patients, and 3 175 929 mutations and the COSMIC database [95]. I pay attention to which are FLAGS [97] genes and find as "benign" or "likely benign" in ClinVar [230]. It is established that individual tumors may harbor clinically relevant alterations, which are not observed frequently in tumors of the exact cancer type [231]. Our approach prioritizes alterations, and only cancer-relevant positions are retained. Cancer-related positions dramatically reduce the number of reported variants while containing the most driver relevant variants and other variants of potential significance.

- Driver genes prioritization based on their allele frequency

Extensive genetic variation exists within individual tumors due to mutations that cause malignant states. During this process, driver genes supply a particular growth advantage to the cell. Therefore, the allele frequency of somatic mutations, which estimates the ratios of cells in which the mutations reside, could mirror both selective growth advantage and the chronology of the modifications. Therefore, I tried to reveal the signatures of cancer-related genes using allele frequency of somatic mutations in individuals. Considering that driver genes' allele frequency is high and not subclonal, I can use the allele frequency of somatic mutations in each patient to find potential driver genes and indicate oncogenes. Therefore I used allele frequency of somatic mutations in each patient to predict mutation progression and find and distinguish driver genes. The influence of allele frequency estimation in practice is the bias of sample purity and homogeneity, library construction error, and copy number variation. The copy number variation in our case is the easiest to solve since I verified with our no-control pipeline for copy number variation that all the samples are real normal karyotypes and do not harbor any copy number alterations. The WGS samples were all good quality, and I didn't need to exclude or adapt any of the samples. I concentrate on the mutation with an allele frequency greater than 0.3.

- Analysis of DNA and RNA allele frequency distributions

Accompanying analysis of RNA and DNA VAF is evolving and possible with the growing accessibility

of paired RNA and DNA sequencing from the same individual. Transcribing tumor mutations from DNA into RNA is essential for the mutation's functionality. However, it is unclear if mutations are generally transcribed and, if so, at what proportion from the wild-type allele. Here, I dissected the correlation between DNA mutation allele frequency and RNA mutation allele frequency. 281 of the 578 mutations are expressed (around 50 %). The RNA and DNA mutation allele fractions are equivalent for most mutations, and the inequality is almost zero. That coincides with other studies about the disbalance between the allele frequency of RNA and DNA mutations [232]. In expressed heterozygote locations in DNA, alleles not regulated by traits are predicted to have expression rates with a likelihood 0.5, which will resemble the DNA allele distribution. Differences between DNA and RNA allele frequency values are seen in exceptional points of transcriptional regulation where one of the alleles is preferentially transcribed over the other. DNA and RNA are expected to have similar allele frequencies without allele-preferential transcription. In the complex karyotype of AML, the discrepancy in allele frequencies can be a preliminary indicator for breakpoints of copy number alterations if they fall within the regions covered by sequencing. However, since I proved that there is no copy number variation in none of our patients, these conclusions would be excluded.

Furthermore, allele-specific instruments have been assumed to play a role in limiting the penetrance of deleterious variants. A study examined more than a thousand likely driver mutations across 69 oncogenes in 13 448 tumors. It concluded that 55% of all oncogenic driver mutations were heterozygous, and 45% exhibited allelic imbalance. Furthermore, among all tumors, 41% showed mutant allele imbalance of one or more driver mutations [233]. All these extensive investigations demonstrate that a substantial fraction of cancer-related driver alterations are associated with allele-specific expression events and that the allelic imbalance state of cancer-associated genes may deliver further prognostic knowledge. Our analysis indicated preferential transcription of alleles in some of the genes that could add additional information to the driver events for some patients.

- Analysis of protein domains

Most approaches to identifying cancer-driver genes focus on entire genes or specific widespread mutated positions. Those approaches may be correct to describe most occurrent driver genes, but fail to represent their functionality, could prioritize passenger events, and also fail to pick up more rare events that could have active driver activity. Even more, mutations may have different consequences, including their relevancy to cancer, depending on which part of the gene they impact. I exploited the internal somatic missense mutations within the protein's functional domains to find those that indicate a bias in their mutation rate compared with other regions of the same protein, proving positive selection and meaning that these proteins may be cancer drivers.

Proteins typically possess one or more functional regions in their sequence, commonly termed 'domains'. PDs are evolutionarily conserved in representations of amino acid sequence, 3D structure, and function [234]. The usefulness of this tool has been tested by analyzing a dataset with a matching control. My prioritization technique can pinpoint expressed cancer driver mutation within the functional protein domain. Even further, the control pipeline validated the output in the samples with a matching control, which shows us the fundings are somatic. Furthermore, allowing the filter of the point mutation characterized in leukemia and other cancer types enables us to find new driver candidates in our AML cohort. It was found that more than 60% were truly propagated onto protein domains.

The mutation variant of RRAS2 is expressed relatively less than the reference allele. However, the nonsynonymous mutation falls into the protein domain, which allows us to presume that this mutation could act as a driver in case this mutation induces the locking of the protein at a GTP-bound state and causes signaling activation. PIK3CG is mutated in the Ras-binding domain. Mutations in these genes were already described in relation to leukemia [235]. FN3K Protein Kinases, catalytic domain, is described in oncogenic function and cancer development. Both contribute to the MAPK signaling [235],

[236]. In addition, from the same pathway in another patient, MAPK12 is mutated in the catalytic domain of the Serine/Threonine Kinase, which could alter its activity. The mitogen-activated protein kinase (MAPK) pathway is an essential integration pinpoint along the signal transduction cascade that connects various extracellular triggers to proliferation and survival. Unsuitable MAPK activation may also contribute to the leukemic transformation of myeloid cells [237], [238]. AML cell line studies prove that small-molecule inhibitors of the MEK/MAPK pathway significantly impact the growth of primary AML samples with constitutive MAPK activation but have negligible impact on cells with low steady-state MAPK activity [239]. Thus, further studies are needed to clarify the role of and interplay of these mutations to identify potential therapeutic targets for these patients.

The enormous number of variants that NGS can detect requires the development of computational approaches to prioritize mutations to determine the ones with a high chance of being oncogenic. In conclusion, our pipeline that combines several gene-based and public databases approaches and later adds expression and protein domains information increases the power of the analyses and enables accurate identifying hotspots.

4.2.2 fusion genes

Recurrent fusion genes in hematological malignancies are a significant part that contributes to tumorigenesis. Several studies have represented a heterogeneous landscape of fusion genes in AML [240] [241],[242] where only a few genes were recurrently rearranged or adjusted. Here I used RNA sequencing to analyze to characterize the presence of rare or never before reported transcribed fusions. As a result, I determined novel and rare fusion events with an expected pathogenic function in AML.

The description of multiple fusion transcripts among AML contributes to the molecular portrayal of that heterogeneous disease. However, the estimation of these fusions and prediction of their oncogenic prognosis have yet to be extensively investigated, and the detection of these rearrangements needs to be integrated into routine practice. In this cohort, I investigated fusion genes from paired-end RNA sequencing, and afterward, I evaluated the output extract of the high-confidence fusion genes and pinpointed them to possible cancer drivers. The advantages of RNA-seq in catching fusion events lean on the capability to recognize fusions whose associate genes are unexplored and capture those rearrangements that stay cryptic at cytogenetic analysis or even whole genome sequence (read-through and trans-spicing fusions). Several bioinformatics tools have been established to detect fusion genes using RNA sequencing data in recent years.

I applied the Ariba tool [105]to detect fusion transcripts in our cohort reliably. It reports any variant whose breakpoints involve two gene coding regions as a probable fusion in the results of whole genome analysis. Sometimes this leads to hundreds of fusions being registered in a single patient, and all are left for the user to interpret.

The total number of fusions seen per patient was relatively high. The general output produces 2049 fusion candidates. However, even after the method's filtering step, the tools still generate many predicted fusion transcripts, most of which may be false positives or low biological interest. Moreover, I found no significant correlation between the number of fusions and genomic breakpoints, following the view that those fundings with low supporting reads are not actual events. Knowledge of the fusion junction sequence is essential for forecasting frame shifts or substitution mutations that could accompany fusion events. Our data found no indels or SNV variants within the fusion breakpoints or the gene. I also checked for any DNA methylation abnormalities within the fusion genes. None of the fused genes show any significant difference in their methylation level. The results suggested that the fusion genes are not driven by epigenetic mechanisms and regulation. Since other studies indicated that successful confirmation of candidates tends to a high oncoscore rather than high read numbers [243], I first allowed the fusions with five or more supporting fusion reads. Therefore I primarily concentrated on fusion

candidates from the high-confidence group filtered by fusion-supporting reads representing 149 fusion transcripts (7.3%). Therefore, I prioritize fusion events by estimating their confidence score, type of fusion (in and out of frame), and retained protein retained domains within the fusion. In addition, sometimes fusion genes give rise to a different transcript isoform. Therefore, when evaluating the fusion, I pay attention to the number of supporting reads and different isoforms of the same fusion gene.

The leading purpose of fusion gene identification is to specify the transcript outcomes of fusions (i.e., isoforms of fusion genes). Accurately characterizing fusion transcript sequences is the basis of all succeeding downstream analyses. I found the pair of involved fusion genes, determined the fusion sites between the paired genes, and identified the expressed isoforms. From them are 22 inframe, and 16 have retained protein domains, mainly concentrating on those fusions. I discovered multiple fusion genes, likely generated by the alternative splicing during transcription or read-through. The fusion variety and the original genes' regular regions can drive the fusion isoform expression's diversity. However, the need for an evident knowledge of the variety of isoform expression of fusion genes intercepts investigating the functions of fusion products. I observed different fusion isoforms from the high-confidence fusions for the same fusion funding. Studies revealed that the Ssu72 phosphatase could sustain cohesion between sister chromatid arms, and inactivation of the Ssu72 phosphatase can generate the sudden separation of the sister chromatid arm (kim2013functional). Furthermore, mutations in the cohesin complex have been recognized in up to 20% of cases of AML [244]. However, fusion involving SSU72 is not described in relation to AML. Kinases activated by gene fusions define an essential category of oncogenes in hematopoietic malignancies [245]. From our output, I uncovered fusions involving different genes that have been shown to contribute to MAPK signaling. This conserved signaling cascade drives a sequence of kinases to transduce signals from the cell membrane to the nucleus, thereby mediating cell growth and survival [246]. Some of the in-frame fusions of this category have contained the protein kinase domain. For instance, the fusion gene PIM-RAC2 may be of interest due to the role of both fusion partners. PIM is serine/threonine kinase, often associated with the pathogenesis and treatment of hematological malignancies and solid cancers [246]. Rac2, from another hand, is a hematopoietic-specific Rho family GTPase involved in cellular events, including regulating cell growth and activating protein kinases [247]. Our analysis has concentrated mainly on fusion transcripts likely to be translated into fusion proteins that could display altered roles. The fusion that was discarded from the final step in the analysis is one fusion that is worth mentioning and discussing. RXRA is part of the retinoic acid receptors and dynamically regulates during myeloid maturation in normal hematopoiesis [248]. I found RXRA-BRD3 out-of-frame fusion in 9 patients regardless of supporting read numbers. From them, only one patient, I consider this fusion to be of high-confidence. The consequences of this out-of-frame fusion are more difficult to speculate. One possible explanation is that frameshift transcripts may be rapidly degraded and, therefore, difficult to detect. Furthermore, out-of-frame fusion may partially mute one allele of two different genes, donate to haploinsufficiency, or be associated with the loss of heterozygosity of one or both unrearranged alleles. Thus, further studies are required to elucidate the role of fusion genes in normal karyotype non-driver acute leukemia to identify potential therapeutic targets for these patients.

4.2.3 outlier expression

In our analysis, I examined the gene expression signatures of our cohort, trying to anticipate the abbreviation from a different angle. For detecting expression outliers for rare-disease diagnostic, I applied OUTRIDER [103], a deep learning tool for identifying expression outliers, regulating for hidden confounders, and providing estimations of statistical significance. OUTRIDER incorporates an autoencoder that automatically controls technical and biological variations among genes and a statistical test based on the NB distribution. OUTRIDER outperforms the previous techniques in recalling simulated and pathogenic outliers from rare-disease cohorts. OUTRIDER has two advantages over the primary

overall used methods. First, it computes p values that can be adjusted to control the FDR. Second, z-score-based approaches lack p values, so the set of cutoffs is incidental [103]. To manage our output, I downloaded and analyzed healthy individuals' RNA seq data and implemented it into the analysis. I also compared the results to TCGA datasets and other AML expression studies to validate our funding.

Our study identified genes already demonstrated to be aberrantly expressed in AML cohorts as well as new candidates. An unexpected finding of this analysis was that the gene expression of Two of them, RAB34 and MGMT, were downregulated in our cohort, while they have been shown to be upregulated in AML and cancer in general. Many factors are involved in initiating and establishing of regulation of gene expression. For example, DNA methylation and histone modifications act with regulatory proteins to remodel chromatin structure, suppress gene expression, form euchromatic structures, and enable gene expression. However, I didn't find any epigenetic explanation in our DNA methylation data. The possible reason is that outlier expression may not be an epigenetically driven process.

Our study investigated three data layers to check for possible drivers that can cause the specific filtered driver mutations in AML. First, I remove most germline and rare variants from the no-control tool, allowing us to handle just a hundred possible variations for further prioritization. Since I had different types of data, whole genome sequencing, RNA-seq, and DNA methylation array, I decided to use and integrate this information to achieve interpretable and integrated results. My pipeline allows us to combine all three data layers and simultaneously search and compare the output results with publicly available data from cancer in general, leukemias, and healthy individual. For instance, Patient E22FSF has only one mutation, a stopgain in MAN2A1, that may or may not be the only cause of the disease. However, the same patient shows aberrant expression of genes regulating translation and ubiquitination of proteins. Therefore, the output provides a broader picture of each patient, and each result is further compared with publicly available TCGA dataset.

4.3 The antileukemic activity of decitabine upon PML/RARA-negative AML blasts is supported by all-trans retinoic acid

Many studies, including ours from both previous projects, described recurrent mutations in epigenetic master regulators (such as TET2) of AML cases, which donated to pathogenesis and held prognostic applicability to AML. Until now, several studies have demonstrated the probable efficacy of targeting the epigenetic machinery with hypomethylating agent DAC in older AML patients [249], [250]. Decitabine is a DNA-hypomethylating agent that causes apoptosis of leukemic cells. However, in elderly AML patients, DAC demonstrated no significant survival difference [251]. Therefore, our project aimed to investigate the combinatorical methylation mechanism of decitabine (DAC) and all-trans retinoic acid (ATRA) in AML-cultured cells. Separate cultures were treated with ATRA, DAC, and the combination of DAC and ATRA. In addition, limited studies described the joint forces on the methylation level of these two drugs [89]. I hypothesized that the ATRA might support the DNMT-inhibitory activity of DAC since other studies have represented partial demethylation of ATRA alone [252], [253]. I observed that the DNA methylation profiles in U937 cells exhibited global hypomethylation after treatment with DAC and DAC in combination with ATRA. However, our analysis does not detect global demethylation upon single-agent ATRA of either timepoint nor did ATRA enhance the demethylation mediated by DAC on a global scale. However, since I used methylation arrays and not whole bisulfite sequencing, I cannot exclude that the mix of DAC + ATRA more effectively demethylates some of the positions not covered in the array compared to DAC alone. In addition, I demonstrated how ATRA cooperates with DAC on epigenetic levels triggering antileukemic activity, including dysregulation of transposable elements. The results showed potential for hypomethylation and differentiation combinations in the clinical setting.

4.4 Conclusions

In this thesis, I worked from various angles on AML. First, I showed different mutations that can accompany the translocation (8;16) and found known AML drivers like FLT3 and SETD2, as well as never depicted earlier in AML genes such as EYS, PSIP1, and SPTBN5. Then, in our main project, I developed a pipeline that takes advantage of 4 layers of data (mutations, fusion genes, gene expression, and DNA methylation) and prioritizes possible driver events by utilizing the publicly available dataset. In different data layers, I found altered genomic and transcriptomic signatures of different GTPases, which are known to be involved in many stages of tumorigenesis. I found new stopgain mutations in cancer genes (NIPBL and NF1). In addition, I identified outlier gene expression and several novel and recurrent fusions involving kinases that potentially play a role in leukemogenesis. I detected previously unreported fusions involving known cancer-related genes, PIM3- RAC2, PROK2- EIF4E3. These analyses can nominate novel disease genes and disclose mechanisms of gene deregulation. The candidates described above can be considered for further functional and mechanistic studies in the applicants' laboratories.

5 REFERENCES

1. Song, Y. *et al.* Comparison of multi-omics results between patients with acute myeloid leukemia with long-term survival and healthy controls. *Annals of Translational Medicine* **10** (2022).
2. Kantarjian, H. *et al.* Acute myeloid leukemia: Current progress and future directions. *Blood cancer journal* **11,** 1–25 (2021).
3. Behrmann, L., Wellbrock, J. & Fiedler, W. Acute myeloid leukemia and the bone marrow nicheâtake a closer look. *Frontiers in Oncology,* 444 (2018).
4. Heaney, M. L. & Soriano, G. Acute myeloid leukemia following a myeloproliferative neoplasm: clinical characteristics, genetic features and effects of therapy. *Current hematologic malignancy reports* **8,** 116–122 (2013).
5. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood, The Journal of the American Society of Hematology* **129,** 424–447 (2017).
6. Ramdas, B. *et al.* Driver mutations in leukemia promote disease pathogenesis through a combination of cell-autonomous and niche modulation. *Stem cell reports* **15,** 95–109 (2020).
7. Kishtagari, A., Levine, R. L. & Viny, A. D. Driver mutations in acute myeloid leukemia. *Current opinion in hematology* **27,** 49–57 (2020).
8. DiNardo, C. D. & Cortes, J. E. Mutations in AML: prognostic and therapeutic implications. *Hematology 2014, the American Society of Hematology Education Program Book* **2016,** 348–355 (2016).
9. Naoe, T. & Kiyoi, H. Gene mutations of acute myeloid leukemia in the genome era. *International journal of hematology* **97,** 165–174 (2013).
10. Daver, N., Schlenk, R. F., Russell, N. H. & Levis, M. J. Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia* **33,** 299–312 (2019).
11. Pomeroy, E. J. & Eckfeldt, C. E. Targeting Ras signaling in AML: RALB is a small GTPase with big potential. *Small GTPases* **11,** 39–44 (2020).
12. Sawada, J. *et al.* Small GTPase R-Ras regulates integrity and functionality of tumor blood vessels. *Cancer cell* **22,** 235–249 (2012).
13. Prior, I. A., Lewis, P. D. & Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer research* **72,** 2457–2467 (2012).
14. Muñoz-Maldonado, C., Zimmer, Y. & Medová, M. *A Comparative Analysis of Individual RAS Mutations in Cancer Biology. Front Oncol. 2019; 9 (1088)* 2019.
15. Alfayez, M. *et al.* The Clinical impact of PTPN11 mutations in adults with acute myeloid leukemia. *Leukemia* **35,** 691–700 (2021).
16. Swoboda, D. *et al.* MDS-144: PTPN11 Mutations are Associated with Poor Outcomes Across Myeloid Malignancies. *Clinical Lymphoma Myeloma and Leukemia* **20,** S316–S317 (2020).
17. Di Costanzo, A., Del Gaudio, N., Conte, L. & Altucci, L. The ubiquitin proteasome system in hematological malignancies: New insight into its functional role and therapeutic options. *Cancers* **12,** 1898 (2020).
18. Husnjak, K. & Dikic, I. Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions. *Annual review of biochemistry* **81,** 291–322 (2012).

19. Ciechanover, A. & Schwartz, A. L. The ubiquitin-mediated proteolytic pathway: mechanisms of recognition of the proteolytic substrate and involvement in the degradation of native cellular proteins. *The FASEB journal* **8,** 182–191 (1994).

20. Mackmull, M.-T. *et al.* Landscape of nuclear transport receptor cargo specificity. *Molecular systems biology* **13,** 962 (2017).

21. Falini, B., Brunetti, L., Sportoletti, P. & Martelli, M. P. NPM1-mutated acute myeloid leukemia: from bench to bedside. *Blood* **136,** 1707–1721 (2020).

22. Pabst, T. & Müller, B. U. Transcriptional dysregulation during myeloid transformation in AML. *Oncogene* **26,** 6829–6837 (2007).

23. Tenen, D. G., Hromas, R., Licht, J. D. & Zhang, D.-E. Transcription factors, normal myeloid development, and leukemia. *Blood, The Journal of the American Society of Hematology* **90,** 489–519 (1997).

24. Zhang, P. *et al.* Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBPα. *Immunity* **21,** 853–863 (2004).

25. Jalili, M. *et al.* Prognostic value of RUNX1 mutations in AML: a meta-analysis. *Asian Pacific Journal of Cancer Prevention: APJCP* **19,** 325 (2018).

26. Mondesir, J., Willekens, C., Touat, M. & de Botton, S. IDH1 and IDH2 mutations as novel therapeutic targets: current perspectives. *Journal of blood medicine* **7,** 171 (2016).

27. Nassereddine, S., Lap, C. J., Haroun, F. & Tabbara, I. The role of mutant IDH1 and IDH2 inhibitors in the treatment of acute myeloid leukemia. *Annals of hematology* **96,** 1983–1991 (2017).

28. Zhang, X. *et al.* DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nature genetics* **48,** 1014–1023 (2016).

29. Han, M., Jia, L., Lv, W., Wang, L. & Cui, W. Epigenetic enzyme mutations: role in tumorigenesis and molecular inhibitors. *Frontiers in Oncology* **9,** 194 (2019).

30. Monaghan, L. *et al.* The emerging role of H3K9me3 as a potential therapeutic target in acute myeloid leukemia. *Frontiers in oncology,* 705 (2019).

31. Gelsi-Boyer, V. *et al.* Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. *Journal of hematology & oncology* **5,** 1–6 (2012).

32. De Rooij, J. D. *et al.* BCOR and BCORL1 mutations in pediatric acute myeloid leukemia. *Haematologica* **100,** e194 (2015).

33. Tsai, C.-H. *et al.* Prognostic impacts and dynamic changes of cohesin complex gene mutations in de novo acute myeloid leukemia. *Blood cancer journal* **7,** 1–7 (2017).

34. Molica, M., Mazzone, C., Niscola, P. & De Fabritiis, P. TP53 mutations in acute myeloid leukemia: still a daunting challenge? *Frontiers in Oncology* **10,** 3368 (2021).

35. Massah, S., Beischlag, T. V. & Prefontaine, G. G. Epigenetic events regulating monoallelic gene expression. *Critical reviews in biochemistry and molecular biology* **50,** 337–358 (2015).

36. Huang, Y.-W., Huang, T. H.-M. & Wang, L.-S. Profiling DNA methylomes from microarray to genome-scale sequencing. *Technology in cancer research & treatment* **9,** 139–147 (2010).

37. Goldman, S. L. *et al.* Epigenetic modifications in acute myeloid leukemia: prognosis, treatment, and heterogeneity. *Frontiers in genetics* **10,** 133 (2019).

38. Dedeurwaerder, S. *et al.* A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics* **15,** 929–941 (2014).

39. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome biology* **17,** 1–17 (2016).

40. Li, E. & Zhang, Y. DNA methylation in mammals. *Cold Spring Harbor perspectives in biology* **6,** a019133 (2014).

41. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13,** 484–492 (2012).

42. Zhang, W. & Xu, J. DNA methyltransferases and their roles in tumorigenesis. *Biomarker research* **5,** 1–8 (2017).

43. Melki, J. R., Vincent, P. C., Brown, R. D. & Clark, S. J. Hypermethylation of E-cadherin in leukemia. *Blood, The Journal of the American Society of Hematology* **95,** 3208–3213 (2000).

44. Network, C. G. A. R. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* **368,** 2059–2074 (2013).

45. Kulis, M. & Esteller, M. DNA methylation and cancer. *Advances in genetics* **70,** 27–56 (2010).

46. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & development* **25,** 1010–1022 (2011).

47. Villa, R. *et al.* Epigenetic gene silencing in acute promyelocytic leukemia. *Biochemical pharmacology* **68,** 1247–1254 (2004).

48. Jiang, H. *et al.* DNA methylation markers in the diagnosis and prognosis of common leukemias. *Signal transduction and targeted therapy* **5,** 1–10 (2020).

49. Jiang, D. *et al.* The diagnostic value of DNA methylation in leukemia: a systematic review and meta-analysis. *PloS one* **9,** e96822 (2014).

50. Kroeze, L. I. *et al.* Characterization of acute myeloid leukemia based on levels of global hydroxymethylation. *Blood, The Journal of the American Society of Hematology* **124,** 1110–1118 (2014).

51. Yang, X., Wong, M. P. M. & Ng, R. K. Aberrant DNA methylation in acute myeloid leukemia and its clinical implications. *International journal of molecular sciences* **20,** 4576 (2019).

52. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic acids research* **44,** 4487–4503 (2016).

53. Di Rorà, G. L., Ficarra, E., *et al.* Novel and Rare Fusion Transcripts Involving Transcription Factors and Tumor Suppressor Genes in Acute Myeloid Leukemia (2019).

54. Matsukawa, T. & Aplan, P. D. Clinical and molecular consequences of fusion genes in myeloid malignancies. *Stem Cells* **38,** 1366–1374 (2020).

55. Chen, X. *et al.* Retrospective analysis of 36 fusion genes in 2479 Chinese patients of de novo acute lymphoblastic leukemia. *Leukemia Research* **72,** 99–104 (2018).

56. Li, H., Wang, J., Ma, X. & Sklar, J. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle* **8,** 218–222 (2009).

57. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* **12,** 1–15 (2011).

58. Kumar, S., Razzaq, S. K., Vo, A. D., Gautam, M. & Li, H. Identifying fusion transcripts using next generation sequencing. *Wiley Interdisciplinary Reviews: RNA* **7,** 811–823 (2016).

59. Kim, P., Ballester, L. Y. & Zhao, Z. Domain retention in transcription factor fusion genes and its biological and clinical implications: a pan-cancer study. *Oncotarget* **8,** 110103 (2017).

60. Miyoshi, H. *et al.* The t (8; 21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript. *The EMBO journal* **12,** 2715–2721 (1993).

61. Langabeer, S. *et al.* Incidence of AML1/ETO fusion transcripts in patients entered into the MRC AML trials. *British journal of haematology* **99,** 925–928 (1997).

62. Okuda, T., Van Deursen, J., Hiebert, S. W., Grosveld, G. & Downing, J. R. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84,** 321–330 (1996).

63. Wright, R. L. & Vaughan, A. T. A systematic description of MLL fusion gene formation. *Critical reviews in oncology/hematology* **91,** 283–291 (2014).

64. Grimwade, D. *et al.* The predictive value of hierarchical cytogenetic classification in older adults with acute myeloid leukemia (AML): analysis of 1065 patients entered into the United Kingdom Medical Research Council AML11 trial. *Blood, The Journal of the American Society of Hematology* **98,** 1312–1320 (2001).

65. Krivtsov, A. V. & Armstrong, S. A. MLL translocations, histone modifications and leukaemia stem-cell development. *Nature reviews cancer* **7,** 823–833 (2007).

66. Perez-Campo, F. M., Costa, G., Lie-a-Ling, M., Kouskoff, V. & Lacaud, G. The MYST erious MOZ, a histone acetyltransferase with a key role in haematopoiesis. *Immunology* **139,** 161–165 (2013).

67. Holmstrom, S. R., Wijayatunge, R., McCrum, K., Mgbemena, V. E. & Ross, T. S. Functional Interaction of BRCA1 and CREBBP in Murine Hematopoiesis. *Iscience* **19,** 809–820 (2019).

68. Forgione, M. O., McClure, B. J., Eadie, L. N., Yeung, D. T. & White, D. L. KMT2A rearranged acute lymphoblastic leukaemia: Unravelling the genomic complexity and heterogeneity of this high-risk disease. *Cancer Letters* **469,** 410–418 (2020).

69. Saad, A. A., Beshlawi, I., Zachariah, M. & Wali, Y. KMT2A-MLLT3 AML masquerading as JMML may disguise fatal leukemia. *Oman Medical Journal* **34,** 553 (2019).

70. Miller, J. *et al.* NUP98-KDM5A Fusion Induces Hematopoietic Cell Proliferation and Alters Myelo-Erythropoietic Differentiation. *Blood* **134,** 3775 (2019).

71. Chen, X. *et al.* Fusion gene map of acute leukemia revealed by transcriptome sequencing of a consecutive cohort of 1000 cases in a single center. *Blood Cancer Journal* **11,** 1–10 (2021).

72. Lin, M. & Chen, B. Advances in the drug therapies of acute myeloid leukemia (except acute wpromyelocytic leukemia). *Drug Design, Development and Therapy* **12,** 1009 (2018).

73. Yu, J. *et al.* Advances in targeted therapy for acute myeloid leukemia. *Biomarker research* **8,** 1–11 (2020).

74. Kayser, S. & Levis, M. J. Advances in targeted therapy for acute myeloid leukaemia. *British journal of haematology* **180,** 484–500 (2018).

75. Kayser, S. & Levis, M. J. FLT3 tyrosine kinase inhibitors in acute myeloid leukemia: clinical implications and limitations. *Leukemia & lymphoma* **55,** 243–255 (2014).

76. Zhao, P. *et al.* Recent advances of antibody drug conjugates for clinical applications. *Acta Pharmaceutica Sinica B* **10,** 1589–1600 (2020).

77. Stokke, J. L. & Bhojwani, D. Antibody–drug conjugates for the treatment of acute pediatric leukemia. *Journal of Clinical Medicine* **10,** 3556 (2021).

78. Li, S. *et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nature medicine* **22,** 792–799 (2016).

79. Christopher, M. J. *et al.* Immune escape of relapsed AML cells after allogeneic transplantation. *New England Journal of Medicine* **379,** 2330–2341 (2018).

80. Lue, J. K. *et al.* Precision targeting with EZH2 and HDAC inhibitors in epigenetically dysregulated lymphomas. *Clinical Cancer Research* **25,** 5271–5283 (2019).

81. Presant, C. A., Coulter, D., Valeriote, F. & Teresa J, V. Contrasting cytotoxicity kinetics of 5-azacytidine and dihydro-5-azacytidine hydrochloride in L1210 leukemia in mice. *JNCI: Journal of the National Cancer Institute* **66,** 1151–1154 (1981).

82. Stresemann, C. & Lyko, F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. *International journal of cancer* **123,** 8–13 (2008).

83. Agrawal, K., Das, V., Vyas, P. & Hajdúch, M. Nucleosidic DNA demethylating epigenetic drugs–a comprehensive review from discovery to clinic. *Pharmacology & therapeutics* **188,** 45–79 (2018).

84. Cany, J. *et al.* Decitabine enhances targeting of AML cells by CD34+ progenitor-derived NK cells in NOD/SCID/IL2Rgnull mice. *Blood, The Journal of the American Society of Hematology* **131,** 202–214 (2018).

85. Li, L. *et al.* Low-dose hypomethylating agent decitabine in combination with aclacinomycin and cytarabine achieves a better outcome than standard FLAG chemotherapy in refractory/relapsed acute myeloid leukemia patients with poor-risk cytogenetics and mutations. *OncoTargets and therapy* **11,** 6863 (2018).

86. Lübbert, M. *et al.* Non-intensive treatment with low-dose 5-aza-2â^2−*deoxycytidine*(*DAC*)*priortoallogeneicbloodSC* *Bone marrow transplantation* **44,** 585–588 (2009).

87. Schenk, T., Stengel, S. & Zelent, A. Unlocking the potential of retinoic acid in anticancer therapy. *British journal of cancer* **111,** 2039–2045 (2014).

88. Germain, P., Iyer, J., Zechel, C. & Gronemeyer, H. Co-regulator recruitment and the mechanism of retinoic acid receptor synergy. *Nature* **415,** 187–192 (2002).

89. Cao, Y. *et al.* Decitabine and all-trans retinoic acid synergistically exhibit cytotoxicity against elderly AML patients via miR-34a/MYCN axis. *Biomedicine & Pharmacotherapy* **125,** 109878 (2020).

90. Kayser, S. *et al.* Characteristics and outcome of patients with acute myeloid leukaemia and t (8; 16)(p11; p13): results from an International Collaborative Study. *British journal of haematology* **192,** 832–842 (2021).

91. Reisinger, E. *et al.* OTP: An automatized system for managing and processing NGS data. *Journal of biotechnology* **261,** 53–62 (2017).

92. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

93. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29,** 308–311 (2001).

94. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38,** e164–e164 (2010).

95. Tate, J. G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research* **47,** D941–D947 (2019).

96. Chang, M. T. *et al.* Accelerating Discovery of Functional Mutant Alleles in CancerAccelerating Discovery of Mutant Alleles in Cancer. *Cancer discovery* **8,** 174–183 (2018).

97. Shyr, C. *et al.* FLAGS, frequently mutated genes in public exomes. *BMC medical genomics* **7,** 1–14 (2014).

98. Stanek, D. *et al.* Prot2HG: a database of protein domains mapped to the human genome. *Database* **2020** (2020).

99. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).

100. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nature methods* **11,** 1138–1140 (2014).

101. Von der Heide, E. *et al.* Molecular alterations in bone marrow mesenchymal stromal cells derived from acute myeloid leukemia patients. *Leukemia* **31,** 1069–1078 (2017).

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

103. Brechtmann, F. *et al.* OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *The American Journal of Human Genetics* **103,** 907–917 (2018).

104. Hotelling, H. in *Breakthroughs in statistics* 162–190 (Springer, 1992).

105. Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome research* **31,** 448–460 (2021).

106. Collins, A. & Project Team, T. C. G. A. The cancer genome atlas (TCGA) pilot project. *Cancer Research* **67,** LB–247 (2007).

107. Rozman, M. *et al.* Type I MOZ/CBP (MYST3/CREBBP) is the most common chimeric transcript in acute myeloid leukemia with t (8; 16)(p11; p13) translocation. *Genes, Chromosomes and Cancer* **40,** 140–145 (2004).

108. Xie, W. *et al.* Acute myeloid leukemia with t (8; 16)(p11. 2; p13. 3)/KAT6A-CREBBP in adults. *Annals of hematology* **98,** 1149–1157 (2019).

109. Velloso, E. R. P. *et al.* Translocation t (8; 16)(p11; p13) in acute non-lymphocytic leukemia: report on two new cases and review of the literature. *Leukemia & lymphoma* **21,** 137–142 (1996).

110. Jiang, J. *et al.* Identifying and characterizing a novel activating mutation of the FLT3 tyrosine kinase in AML. *Blood* **104,** 1855–1858 (2004).

111. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology* **34,** 155–163 (2016).

112. Paschka, P. *et al.* ASXL1 mutations in younger adult patients with acute myeloid leukemia: a study by the German-Austrian Acute Myeloid Leukemia Study Group. *Haematologica* **100,** 324 (2015).

113. Chaudhary, S. *et al.* Mitochondrial biogenesis gene POLG correlates with outcome in pediatric acute myeloid leukemia. *Leukemia & Lymphoma* **63,** 1005–1008 (2022).

114. Nair, R., Salinas-Illarena, A. & Baldauf, H.-M. New strategies to treat AML: novel insights into AML survival pathways and combination therapies. *Leukemia* **35,** 299–311 (2021).

115. Löwenberg, B. *et al.* High-dose daunorubicin in older patients with acute myeloid leukemia. *New England Journal of Medicine* **361,** 1235–1248 (2009).

116. Su, M. *et al.* All-trans retinoic acid activity in acute myeloid leukemia: role of cytochrome P450 enzyme expression by the microenvironment. *PloS one* **10,** e0127790 (2015).

117. Schlenk, R. *et al.* Phase III study of all-trans retinoic acid in previously untreated patients 61 years or older with acute myeloid leukemia. *Leukemia* **18,** 1798–1803 (2004).

118. Estey, E. H. *et al.* Randomized phase II study of fludarabine+ cytosine arabinoside+ idarubicin±all-trans retinoic acid±granulocyte colony-stimulating factor in poor prognosis newly diagnosed acute myeloid leukemia and myelodysplastic syndrome. *Blood, The Journal of the American Society of Hematology* **93,** 2478–2484 (1999).

119. Milligan, D. W. *et al.* Fludarabine and cytosine are less effective than standard ADE chemotherapy in high-risk acute myeloid leukemia, and addition of G-CSF and ATRA are not beneficial: results of the MRC AML-HR randomized trial. *Blood* **107,** 4614–4622 (2006).

120. Luebbert, M. *et al.* Activity of Decitabine (DAC) Combined with All-Trans Retinoic Acid (ATRA) in Oligoblastic AML: Subgroup Analysis of a Randomized 2x2 Phase II Trial. *Blood* **136,** 9–10 (2020).

121. Blagitko-Dorfs, N. *et al.* Epigenetic priming of AML blasts for all-trans retinoic acid-induced differentiation by the HDAC class-I selective inhibitor entinostat. *PLoS One* **8,** e75258 (2013).

122. Liau, M., Kim, J. & Fruehauf, J. Potentiation of ATRA activity i n HL-60 cells by targeting methylation enzymes. *Pharmacol Pharmace u Pharmacovigi* **3,** 1–9 (2019).

123. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome biology* **19,** 1–12 (2018).

124. Lynch-Sutherland, C. F., Chatterjee, A., Stockwell, P. A., Eccles, M. R. & Macaulay, E. C. Reawakening the developmental origins of cancer through transposable elements. *Frontiers in Oncology* **10,** 468 (2020).

125. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98,** 288–295 (2011).

126. Bujko, M. *et al.* Repetitive genomic elements and overall DNA methylation changes in acute myeloid and childhood B-cell lymphoblastic leukemia patients. *International journal of hematology* **100,** 79–87 (2014).

127. Makambi, K. Weighted inverse chi-square method for correlated significance tests. *Journal of Applied Statistics* **30,** 225–234 (2003).

128. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).

129. Grimwade, D. *et al.* Impact of karyotype on treatment outcome in acute myeloid leukemia. *Annals of hematology* **83,** S45–8 (2004).

130. Ahn, J.-S. *et al.* Assessment of a new genomic classification system in acute myeloid leukemia with a normal karyotype. *Oncotarget* **9,** 4961 (2018).

131. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).

132. Robert, F. & Pelletier, J. Exploring the impact of single-nucleotide polymorphisms on translation. *Frontiers in genetics* **9,** 507 (2018).

133. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* **12,** 745–755 (2011).

134. Davnall, F. *et al.* Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights into imaging* **3,** 573–589 (2012).

135. Thomas, M. *et al.* Integration of technical, bioinformatic, and variant assessment approaches in the validation of a targeted next-generation sequencing panel for myeloid malignancies. *Archives of Pathology and Laboratory Medicine* **141,** 759–775 (2017).

136. Li, X. *et al.* OncoBase: a platform for decoding regulatory somatic mutations in human cancers. *Nucleic Acids Research* **47,** D1044–D1055 (2019).

137. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences* **112,** E5189–E5198 (2015).

138. Azad, A., Poloni, G., Sontayananon, N., Jiang, H. & Gehmlich, K. The giant titin: how to evaluate its role in cardiomyopathies. *Journal of muscle research and cell motility* **40,** 159–167 (2019).

139. Chakravarty, D. *et al.* OncoKB: a precision oncology knowledge base. *JCO precision oncology* **1,** 1–16 (2017).

140. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics* **14,** 1–13 (2013).

141. Freese, E. The difference between spontaneous and base-analogue induced mutations of phage T4. *Proceedings of the National Academy of Sciences* **45,** 622–633 (1959).

142. Pfeifer, G. Mutagenesis at methylated CpG sequences. *DNA methylation: basic mechanisms,* 259–281 (2006).

143. Rubin, A. F. & Green, P. Mutation patterns in cancer genomes. *Proceedings of the National Academy of Sciences* **106,** 21766–21770 (2009).

144. Walser, J.-C. & Furano, A. V. The mutational spectrum of non-CpG DNA varies with CpG content. *Genome research* **20,** 875–882 (2010).

145. Liu, J.-J. *et al.* The structure-based cancer-related single amino acid variation prediction. *Scientific reports* **11,** 1–17 (2021).

146. Niroula, A. & Vihinen, M. Harmful somatic amino acid substitutions affect key pathways in cancers. *BMC medical genomics* **8,** 1–12 (2015).

147. Nelakurti, D. D., Rossetti, T., Husbands, A. Y. & Petreaca, R. C. Arginine Depletion in Human Cancers. *Cancers* **13,** 6274 (2021).

148. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *BioRxiv,* 322859 (2019).

149. Vitkup, D., Sander, C. & Church, G. M. The amino-acid mutational spectrum of human genetic disease. *Genome biology* **4,** 1–10 (2003).

150. Rainer, J., Gatto, L. & Weichenberger, C. X. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* **35,** 3151–3153 (2019).

151. Hwang, J. W., Cho, Y., Bae, G.-U., Kim, S.-N. & Kim, Y. K. Protein arginine methyltransferases: Promising targets for cancer therapy. *Experimental & molecular medicine* **53,** 788–808 (2021).

152. Tsuber, V., Kadamov, Y., Brautigam, L., Warpman Berglund, U. & Helleday, T. Mutations in cancer cause gain of cysteine, histidine, and tryptophan at the expense of a net loss of arginine on the proteome level. *Biomolecules* **7,** 49 (2017).

153. Peterson, T. A., Park, D. & Kann, M. G. A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC genomics* **14,** 1–16 (2013).

154. Gilad, O. *et al.* Syndromes predisposing to leukemia are a major cause of inherited cytopenias in children. *Haematologica* (2020).

155. Chiang, P.-W. *et al.* The Hermansky-Pudlak syndrome 1 (HPS1) and HPS4 proteins are components of two complexes, BLOC-3 and BLOC-4, involved in the biogenesis of lysosome-related organelles. *Journal of Biological Chemistry* **278,** 20332–20337 (2003).

156. Foltz, S. M. *et al.* Evolution and structure of clinically relevant gene fusions in multiple myeloma. *Nature communications* **11,** 1–12 (2020).

157. Long, M. A new function evolved from gene fusion. *Genome research* **10,** 1655–1657 (2000).

158. Latysheva, N. S. *et al.* Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Molecular cell* **63,** 579–592 (2016).

159. Melnick, A. & Licht, J. D. Deconstructing a Disease: RAR, Its Fusion Partners, and Their Roles in the Pathogenesis of Acute Promyelocytic Leukemia. *Blood, The Journal of the American Society of Hematology* **93,** 3167–3215 (1999).

160. Barresi, V. *et al.* Fusion transcripts of adjacent genes: New insights into the world of human complex transcripts in cancer. *International journal of molecular sciences* **20,** 5252 (2019).

161. Wang, Y., Wu, N., Liu, D. & Jin, Y. Recurrent fusion genes in leukemia: an attractive target for diagnosis and treatment. *Current Genomics* **18,** 378–384 (2017).

162. Thol, F. Fusion genes in acute myeloid leukemia: do acute myeloid leukemia diagnostics need to fuse with RNA-sequencing? *haematologica* **107,** 44 (2022).

163. Cicconi, L. *et al.* Characteristics and outcome of acute myeloid leukemia with uncommon retinoic acid receptor-alpha (RARA) fusion variants. *Blood cancer journal* **11,** 1–4 (2021).

164. Panagopoulos, I. *et al.* Myeloid leukemia with t (7; 21)(p22; q22) and 5q deletion. *Oncology reports* **30,** 1549–1552 (2013).

165. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34,** 4845–4854 (2015).

166. Zuo, Z.-H. *et al.* Oncogenic Activity of Solute Carrier Family 45 Member 2 and Alpha-Methylacyl-Coenzyme A Racemase Gene Fusion Is Mediated by Mitogen-Activated Protein Kinase. *Hepatology communications* **6,** 209–222 (2022).

167. Rochette, L. *et al.* Mitochondrial SLC25 carriers: Novel targets for cancer therapy. *Molecules* **25,** 2417 (2020).

168. Kim, P. *et al.* FusionGDB 2.0: fusion gene annotation updates aided by deep learning. *Nucleic acids research* **50,** D1221–D1230 (2022).

169. Cao, Y. Tumorigenesis as a process of gradual loss of original cell identity and gain of properties of neural precursor/progenitor cells. *Cell & Bioscience* **7,** 1–14 (2017).

170. Bose, P. *et al.* ING1 induces apoptosis through direct effects at the mitochondria. *Cell death & disease* **4,** e788–e788 (2013).

171. Thiel, J. T., Daigeler, A., Kolbenschlag, J., Rachunek, K. & Hoffmann, S. The Role of CDK Pathway Dysregulation and Its Therapeutic Potential in Soft Tissue Sarcoma. *Cancers* **14,** 3380 (2022).

172. Barry, S. *et al.* Tumor microenvironment defines the invasive phenotype of AIP-mutation-positive pituitary tumors. *Oncogene* **38,** 5381–5395 (2019).

173. Wu, M.-H. *et al.* Silencing PROK2 Inhibits Invasion of Human Cervical Cancer Cells by Targeting MMP15 Expression. *International Journal of Molecular Sciences* **21,** 6391 (2020).

174. Osborne, M. J. *et al.* eIF4E3 acts as a tumor suppressor by utilizing an atypical mode of methyl-7-guanosine cap recognition. *Proceedings of the National Academy of Sciences* **110,** 3877–3882 (2013).

175. Yun, J. W. *et al.* Dysregulation of cancer genes by recurrent intergenic fusions. *Genome biology* **21,** 1–20 (2020).

176. Guérillon, C., Larrieu, D. & Pedeux, R. ING1 and ING2: multifaceted tumor suppressor genes. *Cellular and molecular life sciences* **70,** 3753–3772 (2013).

177. Latysheva, N. S. & Babu, M. M. Molecular signatures of fusion proteins in cancer. *ACS pharmacology & translational science* **2,** 122–133 (2019).

178. Pichiorri, F. *et al.* In vivo NCL targeting affects breast cancer aggressiveness through miRNA regulation. *Journal of Experimental Medicine* **210,** 951–968 (2013).

179. Joshi, S. *et al.* Rac2 controls tumor growth, metastasis and M1-M2 macrophage differentiation in vivo. *PloS one* **9,** e95893 (2014).

180. Wang, C. *et al.* Phosphorylation of ULK1 affects autophagosome fusion and links chaperone-mediated autophagy to macroautophagy. *Nature communications* **9,** 1–15 (2018).

181. Ufartes, R. *et al.* De novo mutations in FBRSL1 cause a novel recognizable malformation and intellectual disability syndrome. *Human genetics* **139,** 1363–1379 (2020).

182. He, X. *et al.* Functional interactions between the transcription and mRNA 3â^2*endprocessingmachineriesmediatedb*
Genes & development **17,** 1030–1042 (2003).

183. Zhou, Y., Shen, J. K., Hornicek, F. J., Kan, Q. & Duan, Z. The emerging roles and therapeutic potential of cyclin-dependent kinase 11 (CDK11) in human cancer. *Oncotarget* **7,** 40846 (2016).

184. Karrila, S., Lee, J. H. E. & Tucker-Kellogg, G. A comparison of methods for data-driven cancer outlier discovery, and an application scheme to semisupervised predictive biomarker discovery. *Cancer informatics* **10,** CIN–S6868 (2011).

185. Yépez, V. A. *et al.* Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome medicine* **14,** 1–26 (2022).

186. Słowiński, P. *et al.* GeTallele: A Method for Analysis of DNA and RNA Allele Frequency Distributions. *Frontiers in bioengineering and biotechnology* **8,** 1021 (2020).

187. Lu, G. *et al.* PPM1l encodes an inositol requiring-protein 1 (IRE1) specific phosphatase that regulates the functional outcome of the ER stress response. *Molecular metabolism* **2,** 405–416 (2013).

188. Flex, E. *et al.* Activating mutations in RRAS underlie a phenotype within the RASopathy spectrum and contribute to leukaemogenesis. *Human molecular genetics* **23,** 4315–4327 (2014).

189. Milani, L. *et al.* Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome research* **19,** 1–11 (2009).

190. Liang, H. *et al.* Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome research* **22,** 2120–2129 (2012).

191. Hortal, A. M. *et al.* Overexpression of wild type RRAS2, without oncogenic mutations, drives chronic lymphocytic leukemia. *Molecular cancer* **21,** 1–24 (2022).

192. Heuser, M. *et al.* High meningioma 1 (MN1) expression as a predictor for poor outcome in acute myeloid leukemia with normal cytogenetics. *Blood* **108,** 3898–3905 (2006).

193. Song, Y. *et al.* High NCALD expression predicts poor prognosis of cytogenetic normal acute myeloid leukemia. *Journal of translational medicine* **17,** 1–12 (2019).

194. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature communications* **8,** 1–11 (2017).

195. Love, M., Anders, S. & Huber, W. Differential analysis of RNA-Seq data at the gene level using the DESeq2 package. *Heidelberg: European Molecular Biology Laboratory (EMBL)* (2013).

196. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic acids research* **42,** e91–e91 (2014).

197. Bourlard, H. & Kabil, S. H. Autoencoders reloaded. *Biological Cybernetics* **116,** 389–406 (2022).

198. Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic acids research* **45,** W98–W102 (2017).

199. Zhu, H. *et al.* The Lin28/let-7 axis regulates glucose metabolism. *Cell* **147,** 81–94 (2011).

200. Viswanathan, S. R. *et al.* Lin28 promotes transformation and is associated with advanced human malignancies. *Nature genetics* **41,** 843–848 (2009).

201. Helsmoortel, H. H. *et al.* LIN28B is over-expressed in specific subtypes of pediatric leukemia and regulates lncRNA H19. *Haematologica* **101,** e240 (2016).

202. Zhou, J. *et al.* Inhibition of LIN28B impairs leukemia cell growth and metabolism in acute myeloid leukemia. *Journal of hematology & oncology* **10,** 1–13 (2017).

203. Aitken, M. J. *et al.* Heterogeneous nuclear ribonucleoprotein K is overexpressed in acute myeloid leukemia and causes myeloproliferative disease in mice via altered Runx1 splicing. *bioRxiv* (2021).

204. Hiltemann, S., Jenster, G., Trapman, J., Van Der Spek, P. & Stubbs, A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome research* **25,** 1382–1390 (2015).

205. Halperin, R. F. *et al.* A method to reduce ancestry related germline false positives in tumor only somatic variant calling. *BMC medical genomics* **10,** 1–17 (2017).

206. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31,** 213–219 (2013).

207. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM. org: Online Mendelian Inheritance in Man, an online catalog of human genes and genetic disorders. *Nucleic acids research* **43,** D789–D798 (2015).

208. Wagener, R. *et al.* IG-MYC+ neoplasms with precursor B-cell phenotype are molecularly distinct from Burkitt lymphomas. *Blood, The Journal of the American Society of Hematology* **132,** 2280–2285 (2018).

209. Lipka, D. B. *et al.* RAS-pathway mutation patterns define epigenetic subclasses in juvenile myelomonocytic leukemia. *Nature communications* **8,** 1–14 (2017).

210. Haferlach, T. *et al.* AML with translocation t (8; 16)(p11; p13) demonstrates unique cytomorphological, cytogenetic, molecular and prognostic features. *Leukemia* **23,** 934–943 (2009).

211. Coenen, E. A. *et al.* Pediatric acute myeloid leukemia with t (8; 16)(p11; p13), a distinct clinical and biological entity: a collaborative study by the International-Berlin-Frankfurt-Münster AML-study group. *Blood, The Journal of the American Society of Hematology* **122,** 2704–2713 (2013).

212. Quesnel, B. *et al.* Therapy-related acute myeloid leukemia with t (8; 21), inv (16), and t (8; 16): a report on 25 cases and review of the literature. *Journal of clinical oncology* **11,** 2370–2379 (1993).

213. Diab, A. *et al.* Acute myeloid leukemia with translocation t (8; 16) presents with features which mimic acute promyelocytic leukemia and is associated with poor prognosis. *Leukemia research* **37,** 32–36 (2013).

214. George, B., Kantarjian, H., Baran, N., Krocker, J. D. & Rios, A. TP53 in acute myeloid leukemia: molecular aspects and patterns of mutation. *International journal of molecular sciences* **22,** 10782 (2021).

215. Kennedy, V. E. & Smith, C. C. FLT3 mutations in acute myeloid leukemia: key concepts and emerging controversies. *Frontiers in Oncology* **10,** 612880 (2020).

216. Perl, A. E. & Pratz, K. W. Incorporation of FLT3 Inhibitors Into the Treatment Regimens for FLT3 Mutated Acute Myeloid Leukemia: The Case for Total Therapy. *The Cancer Journal* **28,** 14–20 (2022).

217. Yu, G. *et al.* Gene mutation profile and risk stratification in AML1-ETO-positive acute myeloid leukemia based on next-generation sequencing. *Oncology reports* **42,** 2333–2344 (2019).

218. Berens, E. B. *et al.* Keratin-associated protein 5-5 controls cytoskeletal function and cancer cell vascular invasion. *Oncogene* **36,** 593–605 (2017).

219. Lapointe, J. *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences* **101,** 811–816 (2004).

220. Gore, M. & Larkin, J. Precision oncology: where next? *The Lancet Oncology* **16,** 1593–1595 (2015).

221. Yang, H.-T., Shah, R. H., Tegay, D. & Onel, K. Precision oncology: Lessons learned and challenges for the future. *Cancer Management and Research* **11,** 7525 (2019).

222. Tsang, E. S. *et al.* Uncovering Clinically Relevant Gene Fusions with Integrated Genomic and Transcriptomic Profiling of Metastatic CancersLandscape of Genomic Fusions. *Clinical Cancer Research* **27,** 522–531 (2021).

223. Lietz, C. E. *et al.* Genome-wide DNA methylation patterns reveal clinically relevant predictive and prognostic subtypes in human osteosarcoma. *Communications biology* **5,** 213 (2022).

224. Boca, S. M., Kinzler, K. W., Velculescu, V. E., Vogelstein, B. & Parmigiani, G. Patient-oriented gene set analysis for cancer mutation data. *Genome biology* **11,** 1–10 (2010).

225. Wendl, M. C. *et al.* PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* **27,** 1595–1602 (2011).

226. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

227. Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome research* **22,** 1589–1598 (2012).

228. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173,** 2187–2198 (2006).

229. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).

230. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **46,** D1062–D1067 (2018).

231. Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature medicine* **20,** 682–688 (2014).

232. Castle, J. C. *et al.* Mutated tumor alleles are expressed according to their DNA frequency. *Scientific reports* **4,** 4743 (2014).

233. Bielski, C. M. *et al.* Widespread selection for oncogenic mutant allele imbalance in cancer. *Cancer cell* **34,** 852–862 (2018).

234. Chothia, C. & Gough, J. Genomic and structural aspects of protein evolution. *Biochemical Journal* **419,** 15–28 (2009).

235. Pillinger, G. *et al.* Targeting PI3Kδ and PI3Kγ signalling disrupts human AML survival and bone marrow stromal cell mediated protection. *Oncotarget* **7,** 39784 (2016).

236. Beeraka, N. M. *et al.* The taming of nuclear factor erythroid-2-related factor-2 (Nrf2) deglycation by fructosamine-3-kinase (FN3K)-inhibitors-a novel strategy to combat cancers. *Cancers* **13,** 281 (2021).

237. Kim, S.-C. *et al.* Constitutive activation of extracellular signal-regulated kinase in human acute leukemias: combined role of activation of MEK, hyperexpression of extracellular signal-regulated kinase, and downregulation of a phosphatase, PAC1. *Blood, The Journal of the American Society of Hematology* **93,** 3893–3899 (1999).

238. Towatari, M. *et al.* Constitutive activation of mitogen-activated protein kinase pathway in acute leukemia cells. *Leukemia* **11,** 479–484 (1997).

239. Milella, M. *et al.* Therapeutic targeting of the MEK/MAPK signal transduction module in acute myeloid leukemia. *The Journal of clinical investigation* **108,** 851–859 (2001).

240. Stengel, A. *et al.* Detection of recurrent and of novel fusion transcripts in myeloid malignancies by targeted RNA sequencing. *Leukemia* **32,** 1229–1238 (2018).

241. Iacobucci, I. *et al.* Genomic subtyping and therapeutic targeting of acute erythroleukemia. *Nature genetics* **51,** 694–704 (2019).

242. Brunner, A. M. No Mutation left behind: the impact of reporting recurrent genetic abnormalities on outcomes of patients with Acute Myeloid Leukemia. *Acta Haematologica* **139,** 128–130 (2018).

243. Dupain, C. *et al.* Discovery of new fusion transcripts in a cohort of pediatric solid cancers at relapse and relevance for personalized medicine. *Molecular Therapy* **27,** 200–218 (2019).

244. Heimbruch, K. E., Meyer, A. E., Agrawal, P., Viny, A. D. & Rao, S. A cohesive look at leukemogenesis: The cohesin complex and other driving mutations in AML. *Neoplasia* **23,** 337–347 (2021).

245. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nature communications* **5,** 4846 (2014).

246. Brault, L. *et al.* PIM serine/threonine kinases in the pathogenesis and therapy of hematologic malignancies and solid cancers. *haematologica* **95,** 1004 (2010).

247. Kim, C. & Dinauer, M. C. Rac2 is an essential regulator of neutrophil nicotinamide adenine dinucleotide phosphate oxidase activation in response to specific signaling pathways. *The Journal of Immunology* **166,** 1223–1232 (2001).

248. De Braekeleer, E., Douet-Guilbert, N. & De Braekeleer, M. RARA fusion genes in acute promyelocytic leukemia: a review. *Expert review of hematology* **7,** 347–357 (2014).

249. Cashen, A. F., Schiller, G. J., O'Donnell, M. R. & DiPersio, J. F. Multicenter, phase II study of decitabine for the first-line treatment of older patients with acute myeloid leukemia. *Journal of Clinical Oncology* **28,** 556–561 (2010).

250. Garcia-Manero, G. *et al.* Phase 1/2 study of the combination of 5-aza-2â^2−*deoxycytidinewithvalproicacidinpatient*. *Blood* **108,** 3271–3279 (2006).

251. Malik, P. & Cashen, A. F. Decitabine in the treatment of acute myeloid leukemia in elderly patients. *Cancer management and research,* 53–61 (2014).

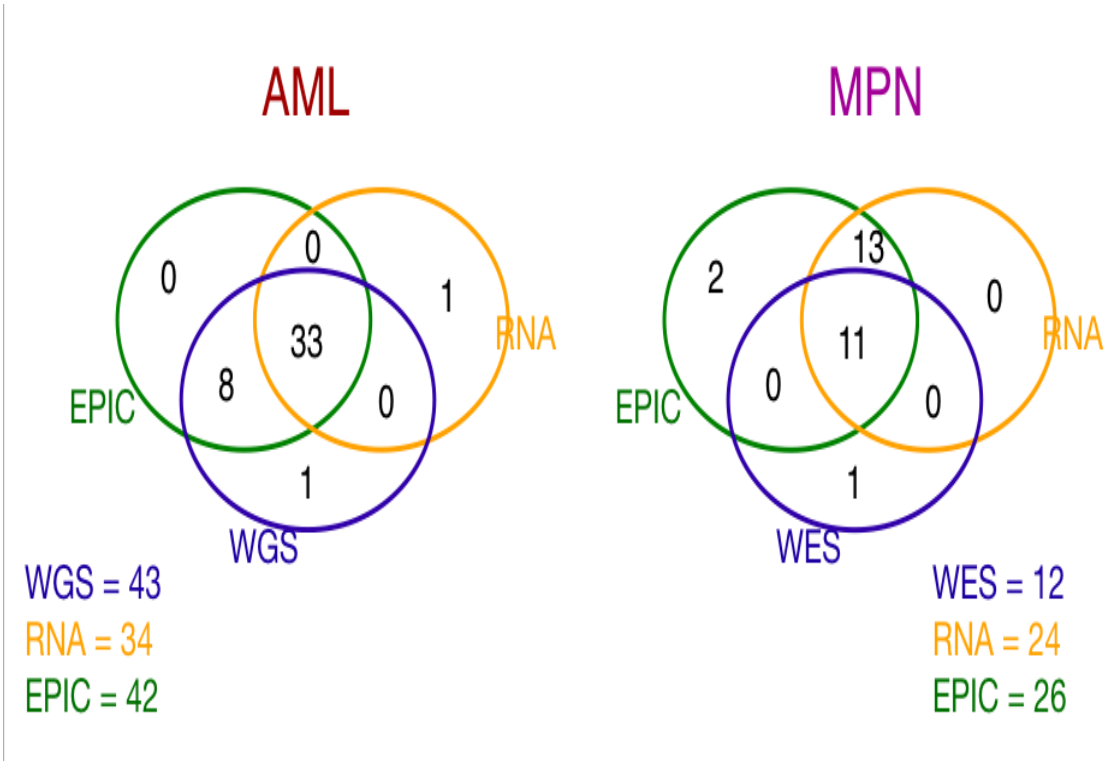
252. Fazi, F. *et al.* Retinoic acid targets DNA-methyltransferases and histone deacetylases during APL blast differentiation in vitro and in vivo. *Oncogene* **24,** 1820–1830 (2005).

253. Das, S. *et al.* MicroRNA Mediates DNA Demethylation Events Triggered by Retinoic Acid during Neuroblastoma Cell DifferentiationEpigenomic Alterations in Neuroblastoma Cell Differentiation. *Cancer research* **70,** 7874–7881 (2010).

6 APPENDICES

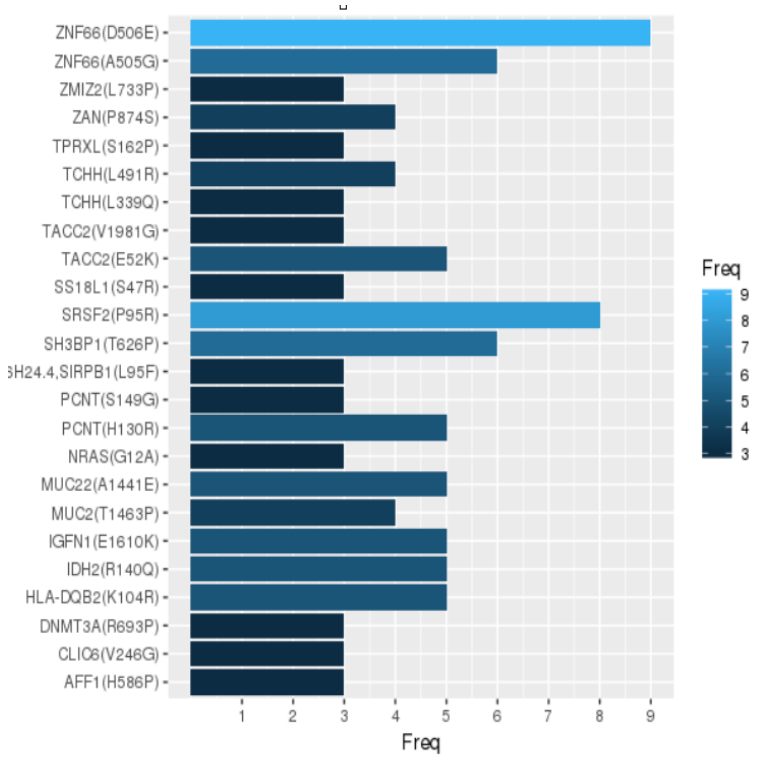
6.1 Supplementary Figures

All the supplementary figures are related to the main project, prospective driver events in mutation-negative normal karyotype AML.

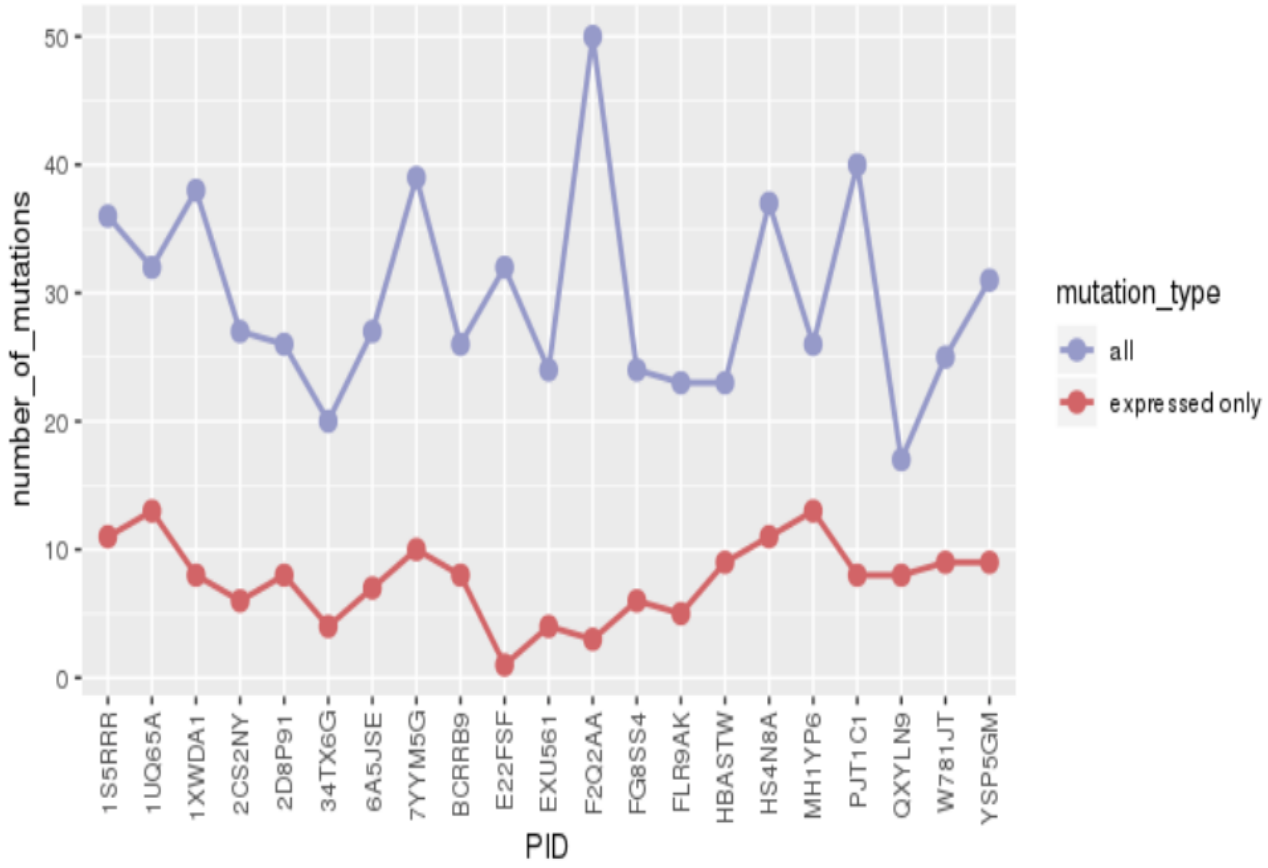


Supplementary Figure S1: The Venn diagram illustrates the extent to which the different layers of data (WGS, RNA-seq, EPIC array) overlap with each other in terms of common samples.

6.1.1 Mutation signature

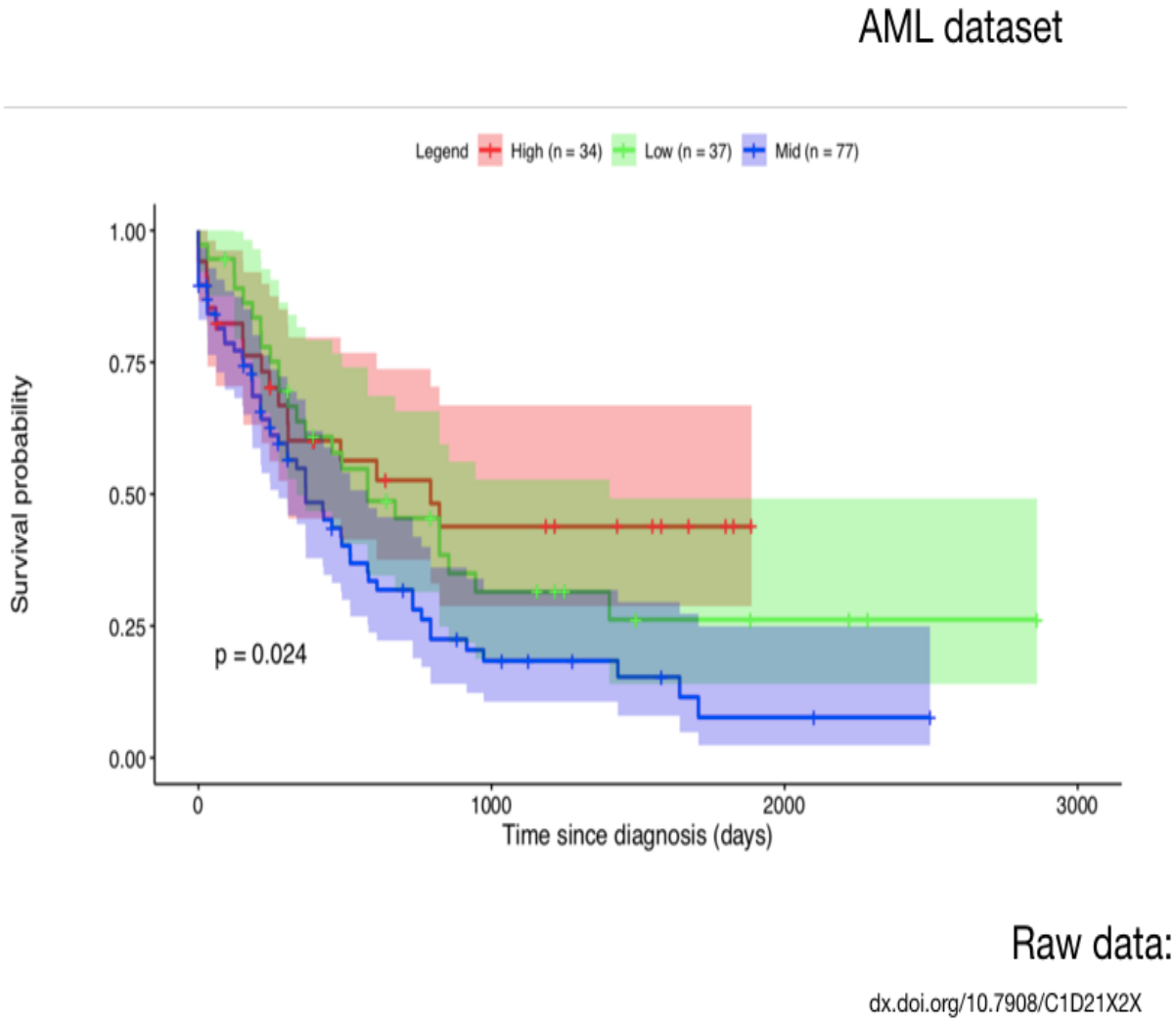


Supplementary Figure S2: SNV point mutation frequency at the position level. The information about the amino acid change is found between parenthesis.



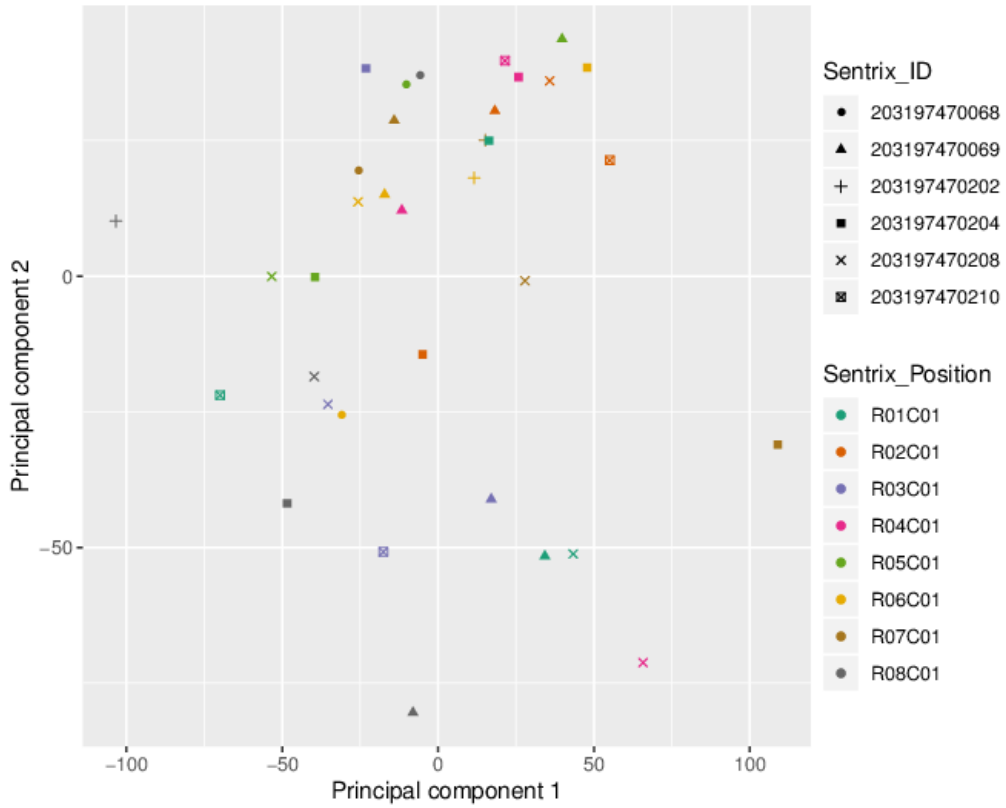
Supplementary Figure S3: The number of mutations per PID after cancer-related filtering of the prioritization pipeline. The blue line shows the number of mutations per sample regardless of whether they are presented on the RNA transcript. The red line represents the number of modifications per sample if only those positions are represented on the RNA level. Seven patients have no expressed mutations: 7ZYTTW, 37JZ4V, 9CJL9K DFEYQ4, DLWTU3, XXMWVE, Y8U7G9, and therefore are not depicted on the graph.

6.2 Outlier expression



Supplementary Figure S4: The survival model of AML using RNA seq from TCGA data for CTIF is a downregulated outlier in our dataset. In that case, lower expression of CTIF correlates with longer survival of AML patients.

6.3 Methylation data



Supplementary Figure S5: PCA of the EPIC array data. It can be observed there is no significant batch effect. There is no strong bias on both covariable: Sentrix ID and Sentrix Position.

7 AUTHOR'S PUBLICATIONS

The antileukemic activity of decitabine upon PML/RARA-negative AML blasts is supported by all-trans retinoic acid: in vitro and in vivo evidence for cooperation

Ruth Meier, Gabriele Greve, Dennis Zimmer, Helena Bresser, Bettina Berberich, **Ralitsa Langova**, Julia Stomper, Anne Rubarth, Lars Feuerbach, Daniel Lipka, Joschka Hey, Björn Grüning, Benedikt Brors, Justus Duyster, Christoph Plass, Heiko Becker, and Michael Lübbert

Blood Cancer Journal 12.8 (2022): 122

Characteristics and outcome of patients with acute myeloid leukaemia and t(8;16)(p11;p13): results from an International Collaborative Study

Sabine Kayser, Robert K Hills, **Ralitsa Langova**, Michael Kramer, Francesca Guijarro, Zuzana Sustkova, Elihu H Estey, Carole M Shaw, Zdeněk Ráčil, Jiri Mayer, Pavel Zak, Maria R Baer, An- drew M Brunner, Tomas Szotkowski, Petr Cetkovsky, David Grimwade, Roland B Walter, Alan K Burnett, Anthony D Ho, Gerhard Ehninger, Carsten Müller-Tidow, Uwe Platzbecker, Christian Thiede, Christoph Röllig, Angela Schulz, Gregor Warsow, Benedikt Brors, Jordi Esteve, Nigel H Russell, RichardFSchlenk, MarkJ Levis

British journal of haematology 192.5 (2021): 832-842.

8 ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Prof. Benedikt Brors, for the opportunity to join the ABI research group and for his great supervision and constant support. His plentiful experience has encouraged me throughout the time of my PhD research. It was a great privilege to be under his supervision. I would like to thank Prof. Stefan Fröhling for being in my TAC and, with Prof. Claudia Scholl, for the great collaboration and valuable advice. I would like to thank my other collaborators Prof. Michael Lübbert from the University of Freiburg and Dr. Sabine Kayser from University Hospital Leipzig, for their great effort and energy when working together. Special thanks go to Dr. Lars Feuerbach and Dr. Yassen Assenov.

I would like to thank Prof. Lars Steinmetz and Prof. Oliver Stegle, who sparked my interest in Personalized Medicine and Bioinformatics, for joining my TAC and for their valuable scientific advice. I would also like to thank Dr. Sevin Turcan, who agreed to be on my examination committee. Special thanks go to my office-mates Pitithat and Sebastian for their friendship and discussions of various scientific and less-scientific topics. I would also like to thank all my ABI friends and colleagues for their suggestions, encouragement, and friendly working environment. I would like to thank Birgit Vey and Corinna Sprengart for helping with administrative matters and support.

Thank you to my parents, Ani and Dimitar, who have supported me in every step of my life and career.