University of Siena

Department of Medical Biotechnologies
Doctorate in Genetics, Oncology and Clinical Medicine
XXXIV cycle
Coordinator: Prof. Francesca Ariani

Identification of Structural Variants in Acute Myeloid Leukemia with normal karyotype patients by using long-reads sequencing technology

Scientific disciplinary sector: MED/15 - Hematological Diseases

Supervisor:
Prof. Alessandro Maria Vannucchi

Doctoral Dissertation of:
Dr. Simone Romagnoli

Academic Year: 2020/2021

Preface

XXX

Abstract

Acute Myeloid Leukemia (AML) accounts for approximately 25% of all leukemias in adults in the Western world, and therefore is the most frequent form of blood *neoplasia*. Leukemic stem cells show abnormal proliferation, activation of antiapoptotic pathways and the impairment normal cell differentiation resulting in the dysregulated production of not functional blood cells, known as blast. AML is an aggressive disease, with a relative survival rate for all ages 5 years after diagnosis of $\sim 29.5\%$, the clinical manifestations of AML reflect the accumulation of malignant, poorly differentiated myeloid cells within the bone marrow, peripheral blood and in other organs. Diagnostic tests are mainly constituted by blood cells count and morphology, AML diagnosis is established by the presence of >=20% myeloid blasts in the bone marrow or peripheral blood. The prognostic assessment of AML patients is of capital importance for the management of the disease and to set up risk adapted therapies. Although clinical factors play an important role in disease development, karyotype is the most independent prognostic factor to forecast patients' survival and it is adopted to provide the framework for risk-adapted treatment approach (Deschler and Lübbert, 2006; De Kouchkovsky and Abdul-Hay, 2016). The *European Leukemia Net* (ELN) guidelines aims to standardize risk stratification in adult AML patients by incorporating cytogenetic and known molecular abnormalities in hot spot genes. Accordingly, AML patients could be stratified into distinct prognostic risk groups (favorable, intermediate or adverse) based on their cytogenetic and molecular profile. Although this classification is the gold standard for the stratification of patients, it is fulfilled for only the 75% of AML whereas it is poorly satisfying for those patients resulted with *normal karyotype* (nk) at the conventional cytogenetic analysis. *normal karyotype AML* (nkAML) patients mostly belong to the intermediate risk category but they experience an extremely heterogeneous outcome that represents an *unmet needs* in the clinical context of AML (De Kouchkovsky and Abdul-Hay, 2016; Döhner et al., 2017). In the last few years, large-scale tumour-sequencing studies have demonstrated that the majority of cancers, including hematologic *neoplasia*, are driven by *Structural Variants* (SVs) that are, for instance, genomic rearrangements larger than 50 bp. SVs include insertions, translocations,

inversions and *Copy Number Alterations* (CNAs) (deletions and duplications). The recent development of high-throughput sequencing platforms provided impressive insights into leukemia pathogenesis and contributed to consider SVs as the hallmark of the genome instability leading to the establishment of the *neoplasia*. Beside karyotype, SVs detection is currently addressed by *Next Generation Sequencing* (NGS) technologies that allow the simultaneous and accurate detection of recurrent SVs breakpoints (Schütte et al., 2019), nothwithstanding, NGS faces inaccuracy and limitations when applied to resolve wide and structurally complex SVs due to the short length (100-500 bp) of the sequencing read employed (Norris et al., 2016).
In this study, we exploited the long-reads Oxford Nanopore Sequencing technology to explore the genome of a cohort of 152 AML patient with normal cytogenetics, aiming to address the genomic analysis challenges and to identify new potential genomic biomarkers able to refine the prognostic forecasting for nkAML patients. Of 152 bone marrow samples collected at diagnosis, 85 referred to the hematology unit of the A.O.U.Careggi and 67 were prospectively collected for the AML #1310 study by the Italian Hematologic Network GIMEMA (Venditti et al., 2019).
The DNA purified from nkAML samples was used to sequence the whole genome by the nanopore long-reads approach and further analysed by the bioinformatic pipeline specifically developed for SVs calling. Two SVs caller, Sniffles (Sedlazeck et al., 2018) and cuteSV (Jiang et al., 2020), were employed for the identification of an high-confidency call-set of SVs that were further clustered and filtered before correlating them with patients' outcome data. We employed an univariate Cox proportional-hazards analysis to weight the correlation between patients' survival and each predictor variables. Further, to better estimate the cumulative impact of multiple genome and clinical variables, we developed a multivariate Cox regression model including those SVs selected by Cox univariate model (pvalue $<.05$) and other predictors such as age, white blood cells count and the known molecular abnormalities in specific hotspot genes included in the ELN guidelines (*Fms related Receptor Tyrosine Kinase 3* (FLT3)-ITD, *Nucleophosmin 1* (NPM1), *CCAAT Enhancer Binding Protein alpha* (CEBPa)). Multivariate analysis allowed to select 12 SVs, represented by genomic deletions or insertions, with high impact on patients'

leukemia free and *Overall Survival* (OS). Of those, 8 resulted with an HR >1 (also referred as *High Risk SVs* (hrSVs)), thus associated with an increased risk of death, the other with an *Hazard Ratio* (HR) <1 (also referred as *Low-risk SVs* (lrSVs)) were associated to a reduced risk of death. The following stratification of the study cohort based on the presence of hrSVs enabled the identification of a high risk group of patients (accounting for the 17% of the cohort) with an extremely poor survival (median OS time 8.27 months for the group harbouring the hrSVs compared to 62.7 month fo the other, LogRank pvalue <.0001) and a low rate of response to therapy (46% for the patients with hrSVs compared to the 80%, pvalue <.0001). Taking together, these data suggest that the employ of an emerging long-reads sequencing technology capable to detect wide SVs together with a dedicated analysis pipeline could represent a powerful tool to accurately screen the whole genome of AML patients and identify new genomic biomarkers for the prognostic assessment of nkAML patients capable to refine the actual ELN prognostic assessment in our cohort.

Contents

List of Figures

List of Tables

List of Acronyms

1.1 Acute Myeloid leukaemia

1.1.1 Definition

AML is an hematological *neoplasia* originated by the clonal proliferation of stem precursors of the myeloid lineage residing in the *Bone Marrow* (BM) and leading to the production of erythrocyte, platelets and white blood cells (Deschler and Lübbert, 2006). The clinical manifestations of AML reflect the accumulation of poorly differentiated myeloid clones, conventionally known as blast cells (Döhner et al., 2015)(Figure 1). The myeloid blasts are characterized by a maturative arrest, preventing physiological hemopoiesis and leading to BM failure. Malignant blast can spread to other organs as brain and lung, this is particularly true for patients with high peripheral blast count (*e.g*, >50000/µL) (Estey, 2018).
AML classification relies on the blast morphology and the type of physiological precursor it most closely resembles. The term "acute" refers to the rapid disease progression, often resulting in poor survival.

1.1.2 Epidemiology

Taking into the account the whole human cancers, AML represents the 1.2% of all new cancer diagnosis per year in the *United States* (U.S.) and it accounts for about one third of all leukaemia cases, thereby constituting the most common leukaemia in adults (Pelcovits and Niroula, 2020). The age-adjusted incidence of AML in U.S. is 4.3 per 100000 subjects with a mortality-related of 2.8 per 100000 every years. The median age at diagnosis is 65 years with an incidence rate in people age <65 of only 2 per 100000 people while the incidence rate in people age ≥ 65 is 20 per 100000 people. (Shallis et al., 2019) (noa).
AML in adults has a male predominance, in the *Surveillance, Epidemiology, and End Result* (SEER) database is reported that males are 1.6 times more likely to be diagnosed with AML than females (age-adjusted incidence of 5.42 and 3.47 per 100000 people per years in males and females) (Shallis et al., 2019).

1.1.3 Etiology

The onset and the development of AML has been associated with several perturbations (summarized in Table 1) of hematopoietic progenitors which boost the clonal and the malignant expansion of immature myeloblasts. Increasing age, genetic disorders, physical and chemical exposure, radiation exposure and previous chemotherapy are the most frequently AML-associated factors.
In the majority of cases, AML appears as a *de novo* malignancy through a multistep process, called leukemogenesis, leading to the leukemic transformation of hematopoietic progenitor stem cells caused by the accumulation of genetic mutation (Shlush et al., 2014; De Kouchkovsky and Abdul-Hay, 2016). A considerable number of patients with chronic myeloid disorders, as *Myelodysplasia* (MDS), *Myeloproliferative Disorder* (MPN) or *Aplastic Anemia* (AA), could evolve to a *Secondary AML* (sAML), regardless previous treatments or exposure to a proven leukemogenic chemotherapeutic agents (therapy-related AML) (Boddu et al., 2017). sAML accounts for 10-30% of all cases of AML, but this varies from study to study, given the heterogeneity and the high percentage of cases arising from undiagnosed MDS. (Soulier, 2020).

1.1.4 Pathogenesis

The pathogenesis of AML is a multistep process through which the genetic lesions accumulate in hematopoietic stem cells (Rubnitz et al., 2008) and give rise to a malignant haemopoietic clone with deregulated cell functions able to outcompete or suppress normal haemopoiesis. (Merker et al., 2012).
The first insights into AML pathogenesis was furnished by patients' karyotype (also known as conventional cytogenetics) showing chromosomal alterations by the visual inspection of chromosomal banding. Notwithstanding these aberrations affect large sequences of DNA, and the pathologic effects are still likely to be due to changes in just one or a few genes that are disrupted by the alteration. Recurrent chromosomal abnormalities was found in approximately 55% of adult patients with AML, they are classified into different groups based on type (Meyer and Levine, 2014):

- **Translocations**, these are the most common chromosomal aberrations found in AML. In this type of alteration a piece of chromosome breaks off and fuse with part of another chromosome, thus originating a "new" chimeric chromosome. The point at which the breakpoints occur can affect nearby genes;

- **Deletions**, in this type of chromosomal aberrations a piece of chromosome is lost;

- **Inversions**, in this type of chromosomal aberrations a part of chromosome is reversed end to end. This type of alterations could result in the loss of one or more genes caused by the alteration of coding sequence;

- **Duplications**, in this type of chromosomal aberrations a part of chromosome is duplicated.

Such chromosomal aberration are found in about the 50% of AML, conversely, about half of diagnosed patients result normal karyotype, since they lack structural variations at cytogenetics inspection by karyotyping, high-density comparative genomic hybridization or SNPs arrays..

1.1.4.1 Multi-step hypothesis of AML development

In 2002, Speck and Gilliland proposed a model of leukemogenesis termed "two hit model" according to which leukaemia is the consequence of the accumulation of subsequent genetic lesions ascribable to two broad classes of mutations (Figure 2). The Class I one mutation (mutations of *FLT3*, *tyrosine-protein kinase KIT* (KIT), oncogenic *Rat Sarcoma Virus* (RAS) and *Protein Tyrosine Phosphatase non-receptor Type 11* (PTPN11), and the *Breakpoint Cluster Region* (BCR)/*Tyrosine-protein kinase ABL1* (ABL1) and *Ets Variant Gene 6* (ETV6)/*Platelet Derived Growth Factor Receptor beta* (PDGFRb) gene fusions), are mutations that confer survival advantages but do not affect cellular differentiation, while the Class II mutations (*Runt-related Transcription Factor 1* (RUNX1)-*RUNX1 Partner Transcriptional Co-Repressor 1* (RUNX1T1) and *Promyelocytic Leukemia* (PML)-*Retinoic Acid Receptor* α (RAR) fusions, *Mixed Lineage Leukaemia* (MLL) rearrangements, and mutations in *CEBPa*, *Core Binding Factor* (CBF), *Homeobox Genes* (HOX) family members, *CREB Binding Protein* (CBP)/*E1A Binding Protein P300* (P300), and co-activators of *RRN3 Homolog, RNA Polymerase I Transcription Factor* (TIF)) impair cell differentiation and apoptosis. The accumulation of both Class I and Class II mutations in the genome of the hematopoietic stem cells cause the clonal transformation of myeloid progenitors, leading to the onset of neoplasia (Speck and Gilliland, 2002).

1.1.4.2 Modern Genomic Landscape in AML

The co-occurence of genetic alterations with different functional effects and the stepwise acquisition of genetic changes by malignant cells paved the way to the concept of clonal architecture and the related heterogeneity of the developing subclones. Beside diagnostic and therapeutic implications, increasing clonal diversity is associated with adverse outcome, mostly due to the probability that one of the subclones acquire resistance to therapy (Bochtler et al., 2013). The study of Li Ding and colleagues evaluated AML samples at diagnosis and relapse and found two major clonal evolution patterns during AML progression: (1) the founding clone

evolving into relapsed clone by acquisition of mutations, (2) a subclone of the founding clone gained additional mutations and expanded at relapse (Ding et al., 2012). In another study it has been demonstrated that in most patients the karyotype aberrations found at diagnosis were stable at the time of relapse except for patients with unfavorable aberrations at diagnosis. Anyhow the acquisition of mutation appears to be less important that the incomplete eradication of the founding clones. Taking together, these findings suggest that AML relapse emerges from incompletely eradicated founder clones, rather than from development of new malignant clones (Kern et al., 2002; Ding et al., 2012).
In 2013, the Cancer Genome Atlas profiled 200 *de novo* AML cases obtained by *Whole Genome Sequencing* (WGS) and *Whole Exome Sequencing* (WES), along with RNA and microRNA sequencing and DNA-methylation analysis (Cancer Genome Atlas Research Network et al., 2013). This study identified 2585 somatic mutations in coding regions of AML, of which 24 were have a putative role in AML pathogenesis: These 24 gene were divided into 7 functional categories (summarized also in table Table 1) (Roussel et al., 2020):

- **signaling genes**: *FLT3*, *Kirsten Rat Sarcoma* (KRAS), *Neuroblastoma RAS viral oncogene homolog* (NRAS) and *KIT* mutations;
- **epigenetic homeostasis genes**: *Putative Polycomb group protein ASXL1* (ASXL1) and *Enhancer of Zeste Homolog 2* (EZH2), *MLL fusions*, *DNA Methyltransferase 3 Alpha* (DNMT3A), *Tet methylcytosine dioxygenase 2* (TET2), *Isocitrate dehydrogenase 1* (IDH1), and *Isocitrate dehydrogenase 1* (IDH2) mutations;
- *NPM1* mutations;
- **spliceosome-complex genes**: *Splicing factor, arginine/serine-rich 2* (SRSF2), *Splicing factor 3B subunit 1* (SF3B1), *U2 Small Nuclear RNA Auxiliary Factor 1* (U2AF1), and *U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit-related protein 2* (ZRSR2) mutations;
- **cohesin-complex genes**: *RAD21 Cohesin Complex Component* (RAD21), *Stromal Antigen 1* (STAG1), *Stromal Anti-*

gen 2 (STAG2), *Structural maintenance of chromosomes protein 1A* (SMC1A), *Structural Maintenance Of Chromosomes 3* (SMC3) mutations;

- **myeloid transcription factors**: *RUNX1*, *CEBPa*, and *GATA Binding Protein 2* (GATA2) mutations, *RUNX1-RUNX1T1*, *PML-RAR*, *Myosin Heavy Chain 11* (MYH11)-*CBFβ* fusions;

- **tumor suppressive genes**, *WT1 Transcription Factor* (WT1), *Tumor Protein P53* (TP53) mutations with *Phosphatase and tensin homolog* (PTEN) and *Dynamin 2* (DNM2) deregulations.

Two or more of these driver mutations have been identified in 86% of the patients. By the way, the "two hit model" proposed by Speck and Gilliland was no longer sufficient to classify all known "AML-alleles" (Moss, 2016; Meyer and Levine, 2014).

1.1.4.3 Clonal Hierarchy

The patterns of mutations seem to follow specific and temporally ordered trajectories. Mutations in genes encoding epigenetic modifiers, such as *DNMT3A*, *ASXL1*, *TET2*, *IDH1*, and *IDH2*, are commonly acquired early and they are often present in the founding clone. The same genes are frequently found mutated in elderly subjects together with the condition of a clonal hematopoiesis and both these factors are known to increase the risk of hematologic cancers. Such mutations may persist after therapy-driven remission and further lead to a clonal expansion, eventually resulting to the relapse of disease. In contrast, mutations involving *NPM1* or signaling molecules (*e.g.*, *FLT3*, *RAS*) are typical secondary events that occur later during leukemogenesis. Genetic data are now being used to inform disease classification, risk stratification, and clinical care of patients. Two new provisional entities, AML with mutated *RUNX1* and AML with *BCR-ABL1*, have been included in the current update of the WHO classification of myeloid neoplasms and AML with mutations in three genes *RUNX1*, *ASXL1*, and *TP53* have been added in the risk stratification of the 2017 ELN recommendations for *AML*. Integrated evaluation of baseline genetics and assessment of minimal residual

disease are expected to further improve risk stratification and selection of postremission therapy.
In conclusion, the identification of disease alleles will guide the development and use of novel molecularly targeted therapies (Bullinger et al., 2017; Döhner et al., 2017).

1.1.5 Diagnosis

The variety of AML manifestations in related to the leukemic infiltration of the BM and extramedullary sites. The replacement of normal BM hematopoietic cells with malignant blasts results in neutropenia, anemia (normochromic and normocytic), thrombocytopenia and often in altered number of white blood cells.
The diagnosis and classification of AML are based on different tests: morphologic, flow cytometric immunophenotyping, cytogenetics and *Fluorescent in situ Hybridization* (FISH), and molecular testing (Rubnitz et al., 2008).
The initial diagnostic framework relies on morphological tests aiming to evaluate both BM aspirate and BM biopsies (Schiffer and Stone, 2003). BM aspirate is preferentially performed, while, in cases of dry tap or diagnostic uncertainty (*e.g.*, distinguishing whether peripheral pancytopenia is related to AML or MDS), bioptic specimens should be evaluated. The biopsy is always indicated in those patients with AML suspect and no circulating blasts in the peripheral blood. The diagnosis of AML requires a blast count (myeloblasts, monoblasts, and megakaryoblast) of at least 20% except for AML with t(15;17), t(8;21), inv(16), or t(16;16) (Döhner et al., 2017). Immunophenotyping identifies a set of cellular marker associated with AML. Cellular markers allow a proper definition of hematological malignancies' lineage and differentiation, whereas cytogenetic or molecular abnormalities further characterize subsets of AML (Vardiman et al., 2009).

1.1.6 Classification

The different classification systems for AML are based on etiology, morphology, immune-phenotype and molecular abnormalities. In 1976 Bennet and colleagues proposed a uniform system of classification and nomenclature for AML, known as the FAB classification, based on blast morphology and the expression of surface antigens

(Bennett et al., 1976). The 8 AML subtypes (FAB M0 to M7, Figure 3) distinguish by the grade of maturation of leukemic cells. The more recent WHO classification and the following updates (the latest in 2016) distinguishes AML in several different categories (Arber et al., 2016):

- AML with recurrent genetic abnormalities and with gene mutations:
  - AML with with a translocation between chromosomes 8 and 21 (*RUNX1-RUNX1T1* t(8;21)(q22;q22));
  - AML with a translocation or inversion in chromosome 16 (*CBF-MYH11* inv(16)(p13.1q22) or inv(16));
  - AML with a translocation between chromosomes 9 and 11 (*Lysine Methyltransferase 2A* (KMT2A)-*Mixed-Lineage Leukemia Translocated To Chromosome 3* (MLLT3) t(9;11)(p21;q23));
  - AML with a translocation between chromosomes 6 and 9 (*DEK Proto-Oncogene* (DEK)-*Nucleoporin 214* (NUP214) t(6;9)(p23;q34));
  - AML with a translocation or inversion in chromosome 3 (*Ribophorin I* (RPN1)-*MDS1 And EVI1 Complex Locus* (MECOM) t(3;3)(q21;q26) or inv(3));
  - AML with a translocation between chromosomes 1 and 22 (*RNA Binding Motif Protein 15* (RBM15)-*Myocardin Related Transcription Factor A* (MKL1) t(1;22)(p13;q13));
  - AML with a translocation between chromosomes 9 and 22 (*BCR-ABL1* t(9;22)(q34;q11));
  - AML with a translocation between chromosomes 15 and 22 (*PML-RAR* t(15,17)(q22;q12));
  - AML with *MLL* 11q23 abnormalities;
  - AML with *NPM1* mutations;
  - AML with bialellic mutation of the *CEBPa* gene;
  - AML with *RUNX1* mutations.
- AML with myelodysplasia-related changes;

- Therapy-related myeloid neoplasms;
- Myeloid sarcoma (also known as granulocytic sarcoma or chloroma);
- Myeloid proliferations related to Down syndrome;
- AML Not Otherwise Specified (This includes cases of AML that don't fall into one of the above groups, and is similar to the FAB classification):
  - AML with minimal differentiation (FAB M0);
  - AML without maturation (FAB M1);
  - AML with maturation (FAB M2);
  - Acute myelomonocytic leukaemia (FAB M4);
  - Acute monoblastic/monocytic leukaemia (FAB M5);
  - Pure erythroid leukaemia (FAB M6);
  - Acute megakaryoblastic leukaemia (FAB M7);
  - Acute basophilic leukaemia;
  - Acute panmyelosis with fibrosis

1.1.7 Prognosis

The prognostic assessment of AML is extremely heterogeneous and depends either on patient-specific features such as patient age, medical comorbidities and performance status, and also on underlying disease-specific features including both cytogenetic and molecular aberrations. The prognostic stratification of AML patients according to their risk of treatment resistance or *Therapy-related Mortality* (TRM) help to guide physicians in deciding between standard or increased treatment intensity, consolidation chemotherapy or allogenic HCT, or in choosing between established or investigational therapies. Among clinical factors, increased age and poor performance status are both associated with lower rates of *Complete Remission* (CR) and decreased survival (DiNardo and Cortes, 2016; Liersch et al., 2014; De Kouchkovsky and Abdul-Hay, 2016).
The ELN prognostic guidelines released in 2010, aimed to provide a prognostic classification based on cytogenetic and known

molecular abnormalities. The 4 prognostic categories with different survival considered by ELN were: favourable, intermediate I, intermediate II and adverse (Döhner et al., 2010). In 2017, the latest refinement of ELN recommendations updated the genetic risk stratification system by incorporating additional cytogenetics and molecular prognostic factors (). Following the ELN-2017 risk stratification, the number of categories were reduced to three because of the distinction between the Intermediate-I category (including only patients with normal cytogenetics) and the intermediate-II category (including patients with intermediate-risk abnormal karyotypes) was eliminated (Figure 4). The adverse risk group includes also patients with chromosomal rearrangements, such as t(6;9)(p23;q34.1), t(9;22)(q34.1;q11.2), inv(3)(q21.3q26.2), mosomy or a complex karyotype (Three or more unrelated chromosome abnormalities in the absence of 1 of the *World Health Organization* (WHO)-designated recurring translocations or inversions, that is, t(8;21), inv(16) or t(16;16), t(9;11), t(v;11)(v;q23.3), t(6;9), inv(3) or t(3;3); *BCR-ABL1*). The karyotype rearrangments t(8;21)(q22;q22.1), inv(16)(p13.1q22) are considered favourable events and associate with a better survival. Beside karyotype alterations, the new guidelines include specific leukemia gene mutations helping the prognostic assessment. In particular, mutations in *ASXL1*, *RUNX1*, *TP53* are considered adverse events. Conversely, bi-allelic mutations in *CEBPa* associate with favourable outcomes. The most common mutational events in AML occur in *NPM1* and *FLT3* genes, the latter is frequently hit by point mutations in the kinase 2 domain (the most frequent *FLT3* D835Y) or by *Internal Tandem Duplication* (ITD) in the region coding for the juxtamembrane and the kinase 1 domain. For instance, the *NPM1* mutational status and the *FLT3*-ITD presence should be considered together, since *NPM1* mutation is considered as a favourable event in case of FLT-wildtype or *FLT3*-ITD with a low allelic burden (<50%). A *FLT3*-ITD burden >50% has a negative impact, leading to a poor survival (adverse category), partially mitigated by the presence of mutation in *NPM1* (intermediate category).
The introduction of molecular alterations allowed to refine the prognostic classification based on karyotype and this is particularly true for those patients with normal karyotype mostly belonging to the intermediate risk group. By the way, the growing body of data on the prognostic relevance of gene mutations, the use of gene

mutations for risk stratification is no longer restricted to normal karyotype patients but employed for the survival forecasting of whole AML diagnosis.

1.1.7.1 Normal Karyotype AML patients

AML is an extremely heterogeneous disease driven by a complex mutational landscape constituted by alterations involving whole chromosomes or arms and hotspot genes. Among all the genetic lesions, the karyotype abnormalities are the most important independent prognostic factor. However, about the 40-45% of newly diagnosed patients show normal karyotype and the majority of nkAML patients are considered as intermediate risk (Nimer, 2008). A variable amount of intermediate risk patients, ranging from 35% to 45%, experience a 5-year overall survival, but clinical outcome may vary greatly. The lack of cytogenetic abnormalities and the heterogeneous outcome constitute a clinical matter for nkAML, since the choice of treatment strategy for consolidation therapy (chemotherapy versus autologous transplantation versus allogenous transplantation) is still debated (Bienz et al., 2005). Despite the refinement of prognostic stratification made possible by the analysis of hotspot mutations in *NPM1*, *FLT3*, *CEBPa*, *ASXL1*, *TP53* and *RUNX1*, the prognostic assessment for the intermediate risk patients results unsatisfactory.
Recent studies with nkAML patients provided intriguing insights on the potential of sequencing approaches applied to the discovery of new disease-related biomarkers. In example, Valk and colleagues carried out the gene-expression profiling of 285 nkAML patients by using Affymetrix U133A GeneChips containing approximately 13,000 unique genes (Valk et al., 2004). nkAML patients were divided into several clusters based on molecular signatures.
In the study of Bullinger et al. the complementary-DNA microarrays were used to profile 116 diagnostic AML. the unsupervised clustering analysis was applied to nkAML patients identified good-outcome and poor-outcome classes associated with significant differences in OS (Bullinger et al., 2004). This unsupervised algorithm developed a clinical outcome predictor that was validated in an independent data set. Later, the prognostic association of Bullinger's signature in normal nkAML patients was validated using a different platform Affymetrix U133 (Radmacher, 2006).

In this study *Compound Covariate Prediction* (CCP), statistical techniques that may be used to assign individual patients to poor or good outcome groups, was used and to assign patients to poor or good outcome groups based on molecular signatures found by Bullinger. Moreover, they developed a classifier that predicts outcome in terms of *Disease-free Survival* (DFS) and OS (Bullinger et al., 2004; Radmacher, 2006). Bullinger et al also profiled gene expression of 138 samples of adult AML patients with normal karyotype using DNA microarray technology (Bullinger et al., 2006). 116 genes comprising expression pattern correlated with *NPM1*-mutated and *FLT3*-ITD-negative nkAML patients were found. Furthermore, they identified the *HOX* gene cluster of potential pathogenic relevance in nkAML with *NPM1*-mutated/*FLT3*-ITD-negative pattern, the expression of *HOX* genes clearly separated the *NPM1*-wild type from the *NPM1*-mutated cases. On the other hand the *NPM1*-unmutated cases displayed higher *Brain And Acute Leukemia, Cytoplasmic* (BAALC) and *MN1 Proto-Oncogene, Transcriptional Regulator* (MN1) expression and the newly defined signature also defined a *NPM1*-mutated group that did not contain many *FLT3*-ITD-positive samples. These data support a distinct molecular mechanism associated with the favorable outcome of *NPM1*-mutated/*FLT3*-ITD-negative AML cases, thus improving the risk stratification and also the clinical mangement of nkAML patients (Bullinger et al., 2006).
In a more recent study the authors profiled 221 nkAML patients by NGS; *acNPM1*, *DNMT3A*, and *acFLT3*-ITD were the most frequently mutated genes and while *DNMT3A*, *FLT3*, *IDH1*, *PTPN11*, and *RAD21* mutations were more common in the *NPM1*-mutated, $IDH1^{(R132)}$ mutation was strictly associated with *NPM1*-mutated patients and mutually exclusive with *RUNX1* and *ASXL1* (Salmoiraghi et al., 2020). In conclusion, the authors identified mutations which are associated with different outcomes and which help to select the most appropriate consolidation strategies.
It has been demostrated by Ibáñez and colleagues that the presence of *Cryptic Cytogenetic Abnormalities* (CCAs) (CNAs and *Loss of Heterozygozity* (LOH)) detectable by high-resolution SNP-array and not by conventional cytogenetics had a negative impact on the outcome of the nkAML patients (Ibáñez et al., 2020).
While many challenges remain to be overcome, a combination of gene expression profiling with other microarray-based applications,

high-throughput mutational analyses and proteomic approaches could grant significant insights into knowledge of the AML pathogenesis in normal karyotype patients. The further characterization of the known molecular signatures could allow the setting up of tailored therapies for a risk-adapted treatment of the heterogeneous subset of nkAML patients.

1.1.8 Treatment

The growing knowledge of the molecular landscape of AML fosters the development of different treatment strategies specifically targeting the recurrent disease biomarkers. The initial therapeutic framework has not changed substantially in the past 30 years (Döhner et al., 2017), nonetheless, a variety of scoring systems basing on patient and disease -specific factors help the clinician to assign patients to intensive or alternative treatment (Ossenkoppele and Löwenberg, 2015; Walter et al., 2011; Klepin, 2014; Klepin et al., 2013).
Age is one of the most important parameters to evaluate relate to treatment choice, in particular when we want to consider whether a patient could be consider a suitable candidate for the intensive induction chemotherapy. The age is also an important parameter for assessing the risk of TRM after intensive therapy and it is usually most relevant in older patients (commonly age 65) (Krug et al., 2010). However, age alone should not be decisive determinant for the choice of the most appropriate treatment.
The first-line treatment (induction therapy) for AML young adults (<65) and for older patients (age >65) is the intensitive anthracycline and cytarabine regimen, "7 + 3", especially for those harbouring NPM1 mutations and CBF leukaemia. Several studies demonstrated that older patients may benefit more from "intensive" than "non-intensive" induction therapy . The goal of the induction therapy is the achievement of morphologic CR. CR is achieved in 60%-80% of younger adults and in 40%-60% of older adults after 3 days classic induction therapy.
Several clinical trials with novel agents targeting specific mutations are under evaluation:

- **FLT3 inhibitors**. FLT3 is a receptor tyrosine kinase and it is mutated in at least 30% of AML. The most frequent

mutation in FLT3 is the ITD, which is a gain-of-function reported in about 25-35% of newly diagnosed AML (Antar et al., 2020). The first generation drugs belonging to this group includes midostaurin, lestaurtinib, tandutinib sunitinib and sorafenib. It has been demonstrated that these drugs, used as single agent, have limited effects showing only transient reduction of blood and BM blasts but an increased toxicity (Sudhindra and Smith, 2014). In the RAFITY trial the application of intensive induction and consolidation chemotherapy plus midostaurin or placebo followed by a 1-year midostaurin/placebo improves the OS;

- **Gemtuzumab ozogamicin**. Gemtuzumab ozogamicin is a monoclonal antibody conjugated with calicheamicin that targets CD33, an antigen found on the blast cells' membrane. CD33 is an antigen expressed in about 85–90% of AML cases (De Propris et al., 2011).
  Two different studies using two different single Gemtuzumab ozogamicin dose on days 1, 4, and 7 of induction chemotherapy in older patients belonging to favorable or intermediate risk group shown survival benefit (Burnett et al., 2012; Castaigne et al., 2012);

- **CPX-351**. CPX-351 is a dual-drug liposomal encapsulation of cytarabine and daunorubicin at 5:1 molar ratio (Tardi et al., 2009). In the randomized phase II study of Jeffrey E. Lancet and colleagues the CPX-351 improved OS and *Event-free Survival* (EFS) in the group of patients (age 60 to 75 years) with sAML against the gold standard "7+3" regimen (Lancet et al., 2014). The following phase III evaluates the response to CPX-351 against "7+3." regimen in older patients (age 60 to 75 years) with newly diagnosed in therapy-related AML or AML with myelodysplasia-related changes. CPX-351 produced a higher response rate and longer OS. Taking together these data suggest how CPX-351 could improve therapy for older patients with high-risk AML (Lancet et al., 2018);

- **Isocitrate Dehydrogenase Inhibitors**. *IDH* mutations, either *IDH1* (*IDH1*$^{\mathrm{R132}}$) or *IDH2* (*IDH2*$^{\mathrm{R140}}$, *IDH2*$^{\mathrm{R172}}$), occurs in at least 20% of AML (Liu and Gong, 2019). Ivosi-

denib, the *IDH1*^{R132} inhibitor, was approved in refractory or relapse AML as monoterapy. Ivosidenib could induce the cell differentiation with no significant cytotoxic events in mutant *IDH1* AML cells . A phase III study evaluating the efficacy and safety of Ivosidenib combined with azacytidine in newly diagnosed AML who are not suitable for intense chemotherapy is ongoing (Fernandez et al., 2019). Enasidenib, the *IDH2* inhibitors, shows a more powerful inhibitory effect on R172 rather than R140. The results from a phase I/II study indicate that Enasidenib as monotherapy in adult refractory or relapse AML is efficacious and safe (Stein et al., 2020). There are several ongoing clinical trial that assess the combination between Enasideninb and azacytidine. Azacytidine and IDH inhibitors can directly or indirectly reduce DNA methylation levels and have synergistic effects on inducing cell differentiation (DiNardo et al., 2021);

- **Purine analogs**. The purine analogues are antimetabolites that mimic the structure of metabolic purines. It has previously demonstrated that the addition of cladribine to "7+3" in adults up to age 60 years produced a higher CR rate and better OS than 7+3, particularly in patients age 50 to 60 years and in those with adverse-risk cytogenetics (Holowiecki et al., 2012). In the study of of Burnett and colleagues they compare clofarabine plus daunorubicin vs daunorubicin/ara-C in older patients with AML or high-risk MDS without any statistical differences in OS, relapse and CR (on behalf of the UK NCRI AML Study Group et al., 2017).

The induction therapy is usually followed by a consolidation therapy with the aim to prevent disease relapse and to control *Minimal Residual Disease* (MRD) in the BM. The induction therapy is substantially the same across the prognostic group of AML, conversely, the consolidation therapy mostly depends on the risk assessment of the patient. The consolidation therapy could include intensive chemotherapy followed by autologous or allogenic HCT when available. Despite the initial choice of consolidation therapy, it is important to determine the availability of a marrow or stem cell donor as soon as possible following the initial diagnosis

of AML. The availability of a matched donor allows the timely transplantation for those patients who does not meet a clinical remission and defines the therapeutic option once a remission is achieved.
In patients with favorable risk, consolidation therapy is based on the use of cytarabine. In those belonging to the intermediate risk group the choice between conventional chemotherapy and allogenenic HCT is based on: the individual risk of relapse, donor availability, performance status, comorbidities and also patient preferences. For patients in the adverse risk group, the allogeneic HCT is the main choice (Schlenk, 2014). For patients not suitable for the intensive chemoterapy the treatment strategy is limited to the best supportive care, low-intensity treatment, or clinical trials with investigational drugs. Low-intensity options are either low-dose cytarabine or therapy with hypomethylating agents.

1.2 Nanopore Sequencing

1.2.1 First Generation Sequencing

In 1977, Sanger and colleagues announced a new method to determine the nucleotide sequence of DNA strands, which is nowadays known as Sanger sequencing. (Sanger et al., 1977).
The Sanger method takes inspiration from a previous work demonstrating the inhibitory activity of *Dideoxythymidine Triphosphates* (ddTTPs) on DNA polymerase I. Indeed, ddTTPs lack the $3'$ hydroxyl group needed to form the phosphodiester bond between a nucleotide and the following one during DNA strand elongation and hence cause the chain termination when incorporated into the nascent fragment by the DNA polymerase (Atkinson et al., 1969).
When an oligonucleotide primer and single-stranded target DNA are incubated in the presence of a mixture of *Deoxythymidine Triphosphates* (dTTPs) and ^{32}P-radiolabeled, corresponding to the four DNA bases, the newly produced fragments will result in strands having all the same $5'$ and terminating with a ddTTPs residue at the $3'$. The mixture of fragments is then fractioned by electrophoresis on acrylamide gel to distinguish pattern of bands revealing the distribution of dTTPs in the newly synthetized DNA. By using ddTTPs terminators for each of 4 nucleotide types in

separate incubations followed by acrylamide gel (*i.e.* one lane for each type of dNTP), a pattern of bands is obtained, from which the entire sequence of the newly synthetized DNA can be deducted (Metzker, 2005).
The Sanger sequencing has been implemented over the years, by including: (1) the development of fluorescent terminator dyes to eliminate the risk caused by the radioisotopes used for labelling; (2) the introduction of thermal-cycle sequencing to reduce the quantity of required input DNA and thermostable polymerases to efficiently and accurately incorporate the terminator dyes into the growing DNA strands; (3) the replacement of acrylamide gel electophoresis with multichannel capillary electrophoresis powered by automated, refillable and reusable capillaries, and the introduction of electrokinetic sample loading.
The automated Sanger sequencing platforms (Applied Biosystems) were successfully exploited for the sequencing of the first human genome , completed in 13 years by the Human Genome Project consortium with an estimated cost of 2.7 billion of $ (Lander et al., 2001). Although relatively slow and not as cost-effective for high numbers of targets when compared to current NGS standards, the Sanger method remains the most appropriate sequencing strategy for low throughput applications (*e.g.* verify plasmid constructs). Moreover, Sanger sequencing is currently used to complement NGS for those regions notoriously difficult-to-sequence (*e.g.* GC-rich and low-complexity regions), and to confirm NGS results (Behdad et al., 2015; Mu et al., 2016).
The first Sanger sequencing projects determined the sequence of small DNA regions, comprising at most a single gene, or very small genome ($\sim$ 5000% bases at most). The following introduction of the Staden Package, developed by Roger Staden in concert with Sanger's laboratory, led to the possibility to sequence long DNA molecules. The Staden Package, compiled for use on early Unix operating systems, allowed to randomly sheared a large original DNA source in order to be sequenced randomly, whose original sequence reconstruction was further accomplished by computational overlapping of multiple sequenced strands (Staden, 1984, 1979, 1996).
As sequencing projects became focused on longer DNA sequences and on larger genomes, the phred/phrap/consed suite from Phil Green's laboratory rapidly supplanted the Staden Package. The

phred provided statistics on basecalling accuracy of Sanger reads, phrap was a read assembly program, and consed was an assembly viewer and editing program (Ewing and Green, 1998; Gordon et al., 1998).

1.2.2 Second Generation Sequencing

Large-scale massively parallel sequencing, also known as *Second Generation Sequencing* (SGS) or NGS, was commercialized since 10 years. The Sanger sequencing method could be applied to short DNA strands (ranging from 100 to 1000 bp), thus long DNA sequences must be fragmented into smaller fragments and sequenced separately. Once the sequencing is done these short sequences are assembled to give the overall sequence. The fragmentation of long DNA molecules could be performed by genome walking, a DNA-cloning methodology used to isolate unknown genomic regions adjacent to known sequences, or by the shotgun sequencing strategy, which was introduced during the Human Genome Project, laid the foundation for massively parallel sequencing (Li et al., 2019). In shotgun sequencing, the DNA sample is randomly broken up into many small pieces, sort of shotgun fashion, further sequenced individually. The resulting sequencing reads generated from the different pieces are then analyzed by means of dedicated software aiming to look for stretches of sequence from different reads that are identical with one another. When identical regions are identified, they are all overlapped, allowing the two sequence reads to be stitched together. This process is repeated yielding the complete sequence of the origin DNA (Fleischmann et al., 1995).
The release of the first truly high-throughput sequencing platform by Lynx Therapeutics (later purchased by Illumina) marked the beginning of the SGS technologies. The SGS platforms could produce large amounts of DNA reads (typically millions to billions), with a length between 25 and 400 bp, conversely, Sanger sequencing reads range from 300 to 750 bp (Barba et al., 2013, 2014).
SGS methods can be broadly grouped into *Sequencing By Hybridization* (SBH) and SBS approaches.
On one hand, SBH approach exploits an arrayed of oligonucleotides that allow the detection of complementary sequences in the target

template (Drmanac et al., 2002). On the other, in SBS approach, a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into the elongating strand.
In both SBH and SBS approaches, the template is first clonally amplified by *Polymerase Chain Reaction* (PCR). In SBS technologies the individual DNA molecules are distributed in millions separate wells or tethered to specific locations on a solid substrate. The first step of template generation is fragmentation (usually by sonication or enzymatically) followed by ligation to common adaptors for clonal amplification and sequencing. The DNA molecules were amplified by PCR or by isothermal modified "rolling circle" amplification methods prior to be subjected to DNA synthesis reactions. In this reactions labeled nucleotides, or chemical reactions based on the incorporation of a particular nucleotide, are imaged or otherwise detected on a solid surface (Slatko et al., 2018). Overall the shortest reads produced by SGS platforms have an intrinsic higher error rate compared to Sanger sequencing, so many thousands of identical copies of a DNA fragment resulted from the massively parallel sequencing and the following identification of a consensus sequence allow the reconstruction of an accurate sequence (Goodwin et al., 2016). The available strategies for clonal amplification of a template are summarized in Figure 5.
In bead-based preparations (Figure 5, panel A), the template is hybridized to bead-bound primers. The template is amplified by *Emulsion PCR* (emPCR) in order to obtain thousands of clonal DNA fragments immobilized on a single bead. Beads can in turn be distributed onto a glass surface (Jae et al., 2007) or arrayed on a PicoTiterPlate (Leamon et al., 2003).
In solid-state strategies (Figure 6, panel B and C), amplification is directly achieved on a slide. Forward and reverse primers are covalently bound to the slide surface, either randomly or on a patterned slide (*i.e.* a flow cell), and provide complementary ends to which template can bind.
The only approach that achieves template enrichment in solution is currently the Complete Genomics technology used by the Beijing Genomics Institute (Figure 6, panel D). In this kind of approach the DNA forms a circular template, also known as rolling circle amplification, which generates up to 20 billion discrete DNA nanoballs that are in turn distributed onto a patterned slide sur-

face containing features that allow a single nanoball to associate with each location (Drmanac et al., 2010).

1.2.3 Sequencing By Hybridization

In the SBH approaches a collection of overlapping oligonucleotide sequences is assembled together to determine the specimen's DNA sequence (Church, 2006).
More in depth, an oligomer constituted by a known sequence hybridizes specifically with a target DNA and reveals the existence of one or more complementary *n-mer* sequences within the chain. A particular *n-mer* sequence occurs roughly once in every 4n/2 base pairs of double-stranded DNA. As easily demonstrated, wide DNA molecules require longer probes in order to avoid situations in which every probe has one or more complementary sequences in the target, leading to the complete loss of specific information. Similarly, short probes could lead to loss of specific information because their intrinsic ability to frequently bind most of the targets. On the other hand, very long probes occur so rarely, hence it is difficult to design probes if the target sequence is not known *a priori* (Drmanac et al., 2010). In general, SBH experiments may use probe sets of length ranging from 5 to 25 bases. A probe can have either one (*i.e.* one-base-encoded probes) or two (*i.e.* two-base-encoded probes) known bases followed by a series of degenerate bases that drive complementarity between probe and template. After hybridization, the template is imaged and the known base or bases in the probe are identified. A new cycle begins after complete removal of the anchor-probe complex or through cleavaging fluorophore and to regenerate the ligation site. Figure 6 illustrates these details.
SOLiD SBH sequencing platforms are based on two-base-encoded probes (Figure 6, panel A) where each fluorometric signal represents a dinucleotide. The 16 possible dinucleotide combinations cannot be identified with spectrally-resolvable fluorophores, for this reason, four signals, each representing a subset of four dinucleotide combinations, are used as decoding markers, further deconvoluted during data analysis (Valouev et al., 2008).
Complete Genomics technology performs DNA sequencing by using cPAL or cPAS, the latter is is a modification of cPAL in which hybridization probes are from a pool of one-base-encoded probes

(Figure 6, panel B) (Fehlmann et al., 2016).
SOLiD platform could be used for whole genome resequencing, targeted resequencing (sequencing a sample for which a reference genome already exists), transcriptomics (including gene expression profiling, small RNA analysis, and whole transcriptome analysis), and epigenomics (like ChIP-Seq and methylation) (Liu et al., 2012).

1.2.4 Sequencing By Synthesis

in SBS approaches, a DNA polymerase is used to add nucleotides into the elongating strand. Each added nucleotide is recognized as a signal by the release of a fluorophore or a change in ionic concentration.
The sequencing library preparation has the goal to conjugate universal platform-specific adaptors to each strand end of previously fragmented DNA/RNA, allowing the binding of the template to the solid surface. The DNA template covalently bound to the solid is further amplified by bridge PCR and primed by sequences complementary to the adaptor region that enable the polymerase binding. In each cycle, a mixture of four individually labelled and -blocked *Deoxyribonucleotide Triphosphates* (dNTPs) are added. Once the labelled base is added to the forming chain, an imager detetcts the type of dNTPs bound. After the imaging step the $3'$-blocked dNTPs and the unbound ones are removed by washing and a new cycle begins (Mardis, 2008).
In Illumina platforms SBS take place on a solid support, named flowcell, where DNA molecules are tethered to specific locations. The Illumina sequencing approach is summarized in Figure 7. In most of Illumina devices, all four nucleotides are added simultaneously, each bound to a single fluorophore, along with DNA polymerase to give the start to the solid-phase bridge amplification. The internal reflection fluorescence microscopy reveals the dNTP incorporated using four different laser channels. Illumina HiSeq series rely on Tour imaging channels, he NextSeq and Miniseq implemented synthesis detection through two-fluorophore system (`https://www.illumina.com/`).
The NGS platforms that rely on SBS differ from each other by the different method to detect dNTPs incorporation There are substantially two main methods classified as direct and indirect

methods (McCombie et al., 2019). The indirect one differs from the direct, mentioned above, because it is based on the indirect detection of signal produced by each nucleotide incorporation. This indirect method does not require the 3′-blocked, as the absence of the following nucleotide prevents the elongation. However, the sequencing instruments based on this indirect detection show lower accuracy when applied to low-complexity region, due to presence of identical adjacent nucleotides composing the sequence. Figure 8 summarized the two groups by which the indirect method is divided. The two group differ by type of the metabolite released during the incorporation of the dNTPs in: (1) detection of pyrophosphate released during nucleotide incorporation by a highly sensitive *Charge-Coupled Device* (CCD) camera (*e.g.* Roche 454), (2) detection of the pH shift, by a miniaturized pH meter, occurring at every dNTP incorporation (*e.g.*, Ion Torrent). The Roche 454 Pyrosequencer is the first NGS platform commercially released in 2004 (Figure 8, panel A). The fragments used for the library are incorporated with agarose beads thanks to the 454-specific adapter sequence, in order to obtain one fragments each bead. This complex is sequestrated into an oil micelle that contains all the PCR reagents necessary to the following emPCR reaction. The template-bound beads, together with an cocktail of enzymes are arrayed, one bead per well, into a PicoTiterPlate were take place the enzymatic reaction generating a bioluminescence signal. The CCD records the light emission from each bead. The 454 basecalling software calibrated itself by recording the light emitted by the first four nucleotides (TCGA) added adjacent to the primer. On one hand, this calibration is not fully satisfactory when applied to interpret long stretches (>6) of the same nucleotide, in this case the resulting sequence could contain insertion or deletion errors; on the other hand, since the pyrophosphate release is nucleotide specific, substitution errors are extremely rare (Goodwin et al., 2016; Mardis, 2008).

The Ion Torrent (Figure 8, panel B) sequencing technology was the first platform commercially released without optical sensing (Rothberg et al., 2011). The Ion Torrent sequencers, distributed by Thermo Fisher Scientific (Figure 8, panel B), rely on the detection of H^+ ions released at every incorporation of a dNTP, by metal-oxide-semiconductor and an ion-sensitive field-effect transistor. Like the 454 Pyrosequencer, the sensor resulted inaccurate

in detecting the pH change of homopolymer.

Given the high amount of data generated by NGS platforms, multiple and different analytical steps are needed. The raw data produced by NGS platforms are more suitable for resequencing approaches rather than for *de novo* assembly (sequencing an entirely new species, there is no pre-existing, similar sequence to help us out in determining where genes are located) due to their typical reads length. Resequencing of candidate genes or other genomic regions of interest in paired samples, patients and controls, is of key importance to identify pathologic mutations. Resequencing techniques could test for known mutations (genotyping) or scan for any mutations in a given target region (variation analysis). Once data generated are checked for quality, the alignment of the reads to the reference genome is crucial for the sample genotyping. To this purpose several read alignment algorithms were developed to map sequencing reads to an existing genome reference. Aligned data are then inspected by variant-callers to detect *Single-Nucleotide Polimorphisms* (SNPs), known as substitutions or point mutations, and other frameshift, in which one or more nucleotides were either added or deleted. The typical structure of *insertion or deletion* (indel) could challenge their alignment to the reference resulting often under-detected; the introduction of paired-end reads constitutes one of the most important technical improvements that facilitate the detection of such structural abnormalities. The paired-end reads sequencing is adopted by NGS platforms (*e.g.* Illumina) in order to boost the quality of data; the analysis of both ends of the same fragments, made possible by a asecond set of reads with the opposite orientation respect to the first set generated `https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html`. The paired-end approach facilitate also the identification of the other genomic rearrangements, such as duplications or amplifications, large deletions or more complex rearrangements such as translocations and inversions (Mardis, 2008; McCombie et al., 2019).

1.3 Third Generation Sequencing

The technical advantages provided by massive parallel sequencing made such technology spread to worldwide laboratories and NGS became a standard for many applications in basic and clinical biology. Despite the undeniable advantages granted by NGS, the short-reads approaches pose several limitations, mostly due to the typical reads length. As an example, the resolution of sequences rich for repeated elements, as those of *CEBPa* or *FLT3*, results ambiguously erratic when addressed by short-reads sizing not more than 500 bp. As a matter of fact, a repeated region could be erroneously collapsed on top of one another, causing complex, misassembled rearrangements (Treangen and Salzberg, 2012b; Mantere et al., 2019). Overall, the repeat content in the human genome is ∼50% and , in contrast, the percentage of short reads mapping to unique region on the human genome is typically reported to be ∼80%, although this number varies depending on the read length, the approach used (*e.g.* single-end reads or paired-end reads) and and the performance of the aligner used (Treangen and Salzberg, 2012a). This discrepancy mainly depends on that most repeats are inexact, which implies that they will have a unique best match even if the same sequence occur with slight variations in other locations, as shown in panel A of Figure 9. Assigning reads to the location of their best alignment, is the simplest way to resolve repeats, although it is not always correct. For example, assume that the same read map to two locations, A and B, where the read aligns with one mismatch at A and with one deletion at B. If the aligner considers mismatches more likely to happen than deletions, then it will put the read in location A. However, if the source DNA has a true deletion, then the read would perfectly match position B. This true-to-life problem, that is inherent in the process of aligning reads to a reference genome, is also illustrated in panel B of Figure 9. This situation happens not only in the context of repetited element but whenever a read maps to multiple locations.
Short reads sequencing could also limit the phasing of haplotypes (the process of estimation of haplotypes, maternal or paternal, from genotype data), this is mostly due to the physiological different location of certain alleles in maternal and paternal genomes, preventing the coverage of the whole region of interest by reads of

limited extent (Tewhey et al., 2011).
Taking into the account all the NGS advantages and limitations, short reads approaches fit at best for SNPs and few bp indel identification, but in turn, they results inadequate to resolve wide and complex SVs (Weischenfeldt et al., 2013a).
In addition to the above mentioned limitations due to the short reads, the fact that NGS methods rely on PCR causes biases with regions of extreme GC%.
In this scenario the so-called TGS technologies emerged rapidly providing reads length in excess of several kilobases and may allow to overcome the above mentioned limitations (van Dijk et al., 2018). The first TGS platforms commercially available were produced in chronologically order by PacBio and ONT.

1.3.1 PacBio Single-Molecule Real-Time Sequencing

The PacBio released the PacBio RS sequencer that exploits the *Single Molecule real Time* (SMRT) sequencing technology which enables the real-time detection of nucleotide incorporation events during the elongation of the replicated strand (Eid et al., 2009). The template to be sequenced, also called SMRTbell, is a single-stranded circular DNA generated by the ligation of hairpin adaptors to both ends of the *Double-stranded DNA* (dsDNA) molecules. The real time detection of labeled nucleotide incorporation events occurs in a cavity tens of nanometers in diameter deposited on a glass substrate , also called *Zero Mode Waveguide* (ZMW). The key reagents of SMRT, labelled dNTPs, an anchored DNA polymerase and SMRTbell, are mixed together in the ZMW where the emission of light with distinct spectrum caused by nucleotides incorporation is measured ate every cycle. More in depth, in each ZMW a DNA polymerase, anchored to the bottom glass surface, bind the hairpin adaptors of the SMRTbell and start to add nucleotides to the strand. Once a labelled base used by polymerase, a laser light exciting the dNTPs results in light signal emission that further computed produce the sequence of the incorporated bases (Rhoads and Au, 2015; Eid et al., 2009). This process take place in every ZMW of the SMRT cell and the pulses in each ZMW correspond to a sequence of bases (called *Continuous Long Read* (CLR)). Figure 10 panel Aa gives an overview of the SMRT

sequencing approaches commercially available.
PacBio sequencer is also capable to detect modified nucleotides (*e.g.* 6-methyladenosine) through the registration of the time-shift between nucleotides incorporation which results delayed in case of modified nucleotides (Figure 11, panel A,B and C).
The average length of the reads produced by PacBio RS sequencer was ∼1500 bases with a mean error rate of ∼13%. The latest PacBio sequencer, the PacBio RS II together with the introduction of *Circular Consensus Sequence* (CCS) technology improved these parameters raising the average read length to 13.5 kilobases and the accuracy to ∼ 99.91% (Wenger et al., 2019). The key advantage of CCS technology is represented by the increased lifetime of the DNA polymerase. The SMRTbell forms a closed circle in which each polymerase replicates the strands multiple times depending on its lifetime. In this scenario, multiple reads could be obtained by splitting the CLR after adapter sequences is recognized and cleaved. The resulted consensus sequence of multiple subreads in a single ZMW is further assembled to form the CCS (Rhoads and Au, 2015; Eid et al., 2009).

1.3.2 Nanopore Sequencing

ONT developed the first sequencing technology that uses nanopore as biosensor to sequence long DNA molecules. The first commercially released ONT sequencing device was the MinION, a pocket-sized cost-affordable instrument producing high-throughput sequencing data in real time (Ip et al., 2015)(`https://nanoporetech.com/products/minion`).
Each nanopore hosted in the flowcell is connected to an electrode and a sensor chip (called *Application-specific Integrated Circuit* (ASIC)) that measures the electric current flowing through the nanopore channel. As the DNA molecule translocate through the pore, the different combination of the nucleotides creates a characteristic disruption in the ionic current, also referred as "squiggle". The observed shift in the ionic current is not influenced by a single nucleotide, but rather than k-mers. A k-mer is subsequence of length k part of a nucleic acid strand; ONT exploits the signal derived from 5-mers (Lu et al., 2016). When a DNA molecule comes in proximity of the pore, an helicase enzyme unwind the paired double strands and foster the translocation

of a single strand through the pore (Figure 10 panel Ab). The "squiggle" resulted from the passage of the 5-mers composing the strand is decoded in real time by the basecalling algorithms to output the DNA (RNA or cDNA) sequence. The changes in the ionic current are influenced also by epigenetically-modified bases, as shown in panels D, E and F of Figure 11.
Basecalling is a crucial step for nanopore sequencing workflow as it allows the conversion of raw current signal into a string of nucleotides.
The first Nanopore basecalling algorithm was provided on the metrichor cloud and was based on *Hidden Markov Models* (HMM). Metrichor for R7.3 version of flowcell recognized the electric signal from 6-mers. Since ONT growing rapidly and basecalling algorithms developed dynamically, actually most of them are deprecated (*e.g.* metrichor and albacore). The study of Wick and colleagues compared 4 basecalling algorithms developed by ONT: Albacore, Guppy, Scrappie and Flappie and third-party basecaller, Chiron (Teng et al., 2018). They concluded that Guppy performs best, in terms of both read and consensus accuracy. The Table3 shows the features of ten basecallers specifically developed specifically for nanopore sequencing (Wick et al., 2019; Makałowski and Shabardina, 2020).
The read length of nanopore sequencing has no apparent technical limit but it is highly affected by the quality and the fragmentation of the input sample, therefore the nucleic acids extraction is a key step to control in order to maximise the throughput of sequencing.
The main drawback of nanopore sequencing is the relative high error rate (range from 5% to 20%) compared to other sequencing technologies. To increase the accuracy, ONT developed a method to sequence both strands of a double-stranded DNA molecule. In this method, called $1D^2$, an adaptor with a specialized sequence promotes the entry of the second strand into the pore after the first one was flowed. The $1D^2$ protocol could increase the accuracy up to 97% and lower the error rate to 6.7%. Given that both strands of each molecules are sequenced, the consumption of pores is doubled and the boost in terms of accuracy comes at the cost of a lower sequencing throughput (Silvestre-Ryan and Holmes, 2021).

Aims And Objectives 2

In current practice, karyotype drives the prognostisc stratification and the choice of risk-adapted therapies for AML patients. The 25-30% of total AML diagnosis result with normal karyotype by standard cytogenetics test. The prognostic assessment of nkAML patients relies on the analysis of the molecular determinants as NPM1, FLT3, CEBPA, ASXL1, TP53 and RUNX1. Nonetheless, the assignment of the risk probability following ELN guidelines is poor satisfactory for nkAML patients and this is particularly true for those that result negative for the mentioned hotspot mutations. As a matter of fact, nkAML shows a high mutational diversity that reflects the heterogeneous outcomes observed in patients. To address such clinical need, in the present study I extensively screened the whole genome of nkAML patients with the aim to identify new genomic aberrations of impact on prognosis and response to therapy, eventually able to refine the actual risk classification largely based on a limited number of gene mutations.

Despite the undeniable advantages introduced by the massive parallel sequencing through NGS platforms (i.e. Illumina or Ion Torrent), such technologies result inadequate to resolve complex and wide genomic structures such as long repetitive elements or structural variations (inversions, translocations, insertions and CNAs longer than 50bp) mostly due to their ability to only gener-

ate short reads comprised between 100-300 bp. For instance, SVs can extend to well over megabases of sequence, accounting for more varying base pairs than any other class of sequence variants (Ho et al., 2020). SVs are involved, and eventually driver of, several human *neoplasia* and pathologic conditions, such as cognitive disorders (Rovelet-Lecrux et al., 2006), obesity ((Walter et al., 2011) and cancer (PCAWG Structural Variation Working Group et al., 2020) among others (Weischenfeldt et al., 2013b). Despite SVs are of major relevance to understand the etiology and the manifestations of disorders, they have been largely understudied, mostly because their identification is hindered by technical challenges of short-read based technologies, especially for repetited DNA elements, also known as low-complexity regions, which are known to be SVs hotspots (1000 Genomes Project et al., 2011). Indeed, it has been shown that, from a computational perspective, short-read alignment and assembly of genomic repeated elements result ambiguously erratic and, in turn, could introduce errors in the genetic variant calls (Treangen and Salzberg, 2012b). In this scenario, the rise of TGS technologies aims to overcome such limitations by the introduction of a new sequencing approach based on long reads. The employ of TGS technologies could facilitate the study of genome "twilight regions" unsatisfactory covered by NGS or, on the other hand, too small in size to be addressed by cytogenetics.

Among TGS platforms, ONT developed the first devices that use nanopores as biosensor capable to sequence long molecules of DNA (single or double stranded) sizing up to hundreds Kbs. ONT is based on 2048 individual protein nanopores embedded in a matrix and constituting the flowcell. The nanopores allow the translocation of DNA strands through their inner channel by a drift-diffusion process that result in the perturbation of the ionic current constantly applied to the matrix. The measure of the current changes generated by the passage of the oligonucleotides is fulfilled by an ASIC with a constant sampling frequency. Raw current data are collected and analyzed by MinKNOW, the administration software, that manages every step of the sequencing run since the loading of the library into the flowcell to the output of raw data. Of interest, ONT devices can directly sequence purified nucleic acids, not pre-amplified, thereby allowing the analysis of "un-touched" samples and minimizing the introduction

of biases possibly arising from PCR reactions. The long reads sequencing has the potential to study complex alterations involving wide genomic regions improving the detection potential SVs. In this context, ONT suggests a great potential ability in the SVs identification with higher precision than current NGS SVs detection tools. Given the intriguing possibilities offered by the nanopore-based sequencing, I reasoned to use long-read ONT technology to comprehensively screen the genome of a/two cohort(s) of nkAML patients with the aim to identify possible SVs went undetected with conventional methods. The further analysis of genomic data was arranged to identify those newly discovered SVs with prognostic impact by correlating clinical, molecular and sequencing information of nkAML cohort(s). Moreover, the aberrant molecules that could be generated by such SVs may represent new targets for future therapies. The integration of new molecular markers, detected by an innovative approach, with known mutations has the potential to improve the current prognostic stratification for nkAML patients, representing a step toward a tailored medicine aiming to improve the clinical management of nkAML patients.

The pipeline developed in this study is further described in detail in Methods. The Results shows the findings of the study, illustrating the sequencing data, the bioinformatics and the correlation analysis. The Discussion provides a critical revision of results and their application to the AML clinical context.

3.1 Samples Collection

To the purpose of this study, a cohorts of patients diagnosed with AML resulted normal karyotype according to conventional cytogenetics were enrolled from Florence hematology unit and from the AML 1310 GIMEMA study. Clinical, laboratory, immunophenotypic and molecular data are examined at diagnosis, and information relating to the treatment and outcome of patients for the duration of follow-up.
The whole cohort is composed by 152 samples diagnosed with AML from 2010 to 2017, the median age at diagnosis is 51 years (ranging from 19 to 74), the mean white blood cells count was 22700 (ranging from 1060 to 435000) and the median OS time is 44 months (lower 95% 37.4, upper 95% 48.8); the median DFS time was 40 months (lower 95% 31.86, upper 95% 46.13). Overall the 37.6% of patients undergo to allogenic HCT. All the samples' clinic characteristics were summarized in table Table 4.
Patients provided informed consent for the use of archival material for the use of archival material for SVs analysis and the study was performed under the Florence University Institutional Review Board-approved protocol.

3.2 Sample Preparation

For this study we used purified DNA from mononuclear cells recovered from BM from patients found to be affected by AML resulted normal karyotype as input.
Mononuclear cells were purified from BM blood by density gradient stratification using Lymphoprep®, also known as Ficoll. BM whole blood was collected into EDTA-containing polypropylene tubes and procecessed within 4 hours. Blood was diluited 1:1 with Ca^{2+}/Mg^{2+} free phosphate-buffered saline, carefully layered over Ficoll (in a blood-to-Ficoll ratio 2:1) in a 50-ml tube and centrifuged at 800 x g at room temperature for 20 minutes. After the centrifugation we obtain different phases: the first phase consists of plasma and thrombocytes; the second consists of a thiny opalescent ring, contains *Mononuclear Cells* (MNCs) below which is the third phase containing mainly the stratification medium. The last phase is represented by erythrocytes on whose contact surface with the third phase, there is a whitish ring consisting of granulocytes. The Ficoll layer was carefully removed and the opalescent ring transferred to a fresh tube. Lysis of red blood cells was performed by the addition of a 10X volume of 1X BD PharmLyse solution (Becton Dickinson DB®), centrifugation of the tube after 15 minutes incubation at room temperature, and removal of the supernatant. This step was repeated twice. After two washes in Ca^{2+}/Mg^{2+} free phosphate-buffered saline the dry mononucloear pellet was stored at -20℃.

3.3 DNA extraction from pellet

The purification of *Genomic DNA* (gDNA) from MNCs' pellet was performed using Wizard HMW DNA Extraction Kit (Promega, Madison, WI, US) optimized to extract high molecular weight DNA. The various manual steps required for the purification of nucleic acids are aimed at the enrichment and preservation of large oligonucleotides, particularly appropriate for long read sequencing purposes.

3.3.1 Quantification and Evaluation of the gDNA

The concentration, quality and the analysis of the DNA fragmentation obtained from the extraction process were assessed by Qubit® 2.0 (Life Tech., MA, US), Nanodrop® One (Thermo Fisher Scientific, MA, US) and Agilent® Bioanalyzer respectively. (Agilent, CA, US). The NanoDrop® One instrument is a spectrophotometer allowed the evaluation of the purity of the DNA through the 260nm (DNA absorbance) / 280nm (absorbance ratios constitutes the protein absorption peak) and 260nm / 230nm (it reflects contamination due to substances such as carbohydrates, phenols or aromatic compounds). The Qubit® 2.0 fluorometer was used for the quantification of purified dsDNA. The extracted DNA was also evaluated in order to establish its by Agilent® Bioanalizer. This technology relies on capillary electrophoresis; each chip used for this instruments contains a series of tightly interconnected microchannels: the nucleic acid fragments are well separated according to their molecular weights as in a standard electrophoresis in agarose gel. Specifically, the DNA chip 12000 was used for the sizing and quantification of dsDNA fragments from 100 to 12000 bp.

3.4 ONT Library Preparation

After the extraction and the quality control, only the samples that meet the quantity and quality requirements are used for the library prepration step. According to the ONT 1D Amplicon by Ligation Protocol (SQK-LSK109), 1,5µg of gDNA was used as input. Long fragments-enriched samples was subjected to the FFPE DNA repair step, in order to repair possible *nick* (discontinuity in a double stranded DNA molecule) along the DNA molecule; the gDNA was further end-repaired and dA-tailed (NEB®, MA, US), this step is crucial to prepare the DNA ends for adapter attachment. The next step is the ligation of the ONT adapter that are recognized by the Motor protein on the surface of each nanopore. Each of the described steps will be followed by Agentcourt AMPure XP beads (Beckman Coulter, CA, US) purification. This protocol is optimized for selectively binding amplicons greater than 100bp. Two washes in 70% ethanol are provided to remove the primers, nucleotides, salts and enzymes required in the sample

preparation step. The purified DNA is mixed with the sequencing buffer and the loading beads constituting the sequencing mix. Once the flowcell is primed with specific buffer the sequencing library is loaded in the flowcells (9.4.1 chemistry) for a standard 48 hours sequencing experiments. The library preparation step is summarized in Figure 12.
For this study we used the ONT GridION X5 that allows up to five MinION sequencing individual experiments to be run concurrently.

3.5 Data Analysis

The Nanopore sequencing platforms GridION outputs fast5 files organized in a hierarchical structure with groups, datasets and attributes, a sequencing_summary file and fastq files generated locally starting from the fast5 files by the GPU-enhanced basecaller Guppy. The sequencing summary file is used for QC assessment by the conda based environment Summary Statistics and QC tutorial (`https://github.com/nanoporetech/ont_tutorial_basicqc`). The QC analysis produces a report in which the throughput of the run, the number of reads passes the QC threshold, the channel activity and the reads length distribution are the most important parameters. Regarding the average quality of fastq reads we select for downstream analysis those reads with an average quality of 7. For the reads length distribution it outputs two histograms, one shows the number of sequenced bases against sequence length in which the N50 and the mean are annotated(the N50 is the sequence length where the 50% of the sequenced bases are contained) and the other the distribution of read lengths across the quality. This last two parameters are important because they represent the library preparation kit depending parameters.
The high-quality reads were aligned to the human reference genome GRCh37 (`ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz`) by minimap2 (Li, 2018), the aligner most commonly used with ONT data. The *Sequence Alignment Map* (SAM) files were compressed, sorted and indexed by samtools (v. 1.9) (Danecek et al., 2021; Li et al., 2009) to obtain the *Binary Alignment Map* (BAM) files. A more powerful programmatic access

to a SAM/BAM files is provided by Pysam (Li et al., 2009), a lighweigth wrapper of the htslib C-API, that enables python parsing of the alignment files. Mosdepth (v. 0.3.1) (Pedersen and Quinlan, 2018) is the tools used to investigate the depth of coverage across the genome. Mosdepth is able to output coverage per-base, per-region and per-chromosome. In this pipeline we calculated the using windows size of 500 bp. Samtools provides the depth module for performing these checks, but mosdepth is simpler to use and also being faster, making it a natural choice for running coverage analysis. The coverage analysis enables to determine how well a sequencing run performed in terms of on-target throughput, to check if we have enough depth of coverage to support variant calling, or to find out if there are any areas of the genome that have been under or over-represented in our sequencing run. Moreover, the coverage analysis in addition to the determination of the average coverage for each sample allowing the determination of genomic regions (of 500 bp) that have a depth departing from the overall distribution (referred as outliers). For each 500 bp of the whole genome of each sample we calculated the z-score in order to obtain a robust and non constant thresholds for each region. Finally, based on the work of Desvillechabrol we flag as outlier those genomic regions that have a z-score ≥ 1.5 (Desvillechabrol et al., 2018). These outlier regions were filtered out and not subjected to SVs calling.

SVs are large genomic alterations classified as deletions, duplications, insertions, inversions, and translocations describing different combinations of DNA gains, losses, or rearrangements (shown in Figure 13). CNAs are a particular subtype of SVs represented by deletions and duplications.

For SVs, such as insertion, deletion,inversion, duplication and translocation we employed Sniffles (v. 1.0.12) (Sedlazeck et al., 2018) and cuteSV (v. 1.0.10) (Jiang et al., 2020) while for large CNAs profiling NanoGLADIATOR (v. 1.0) (Magi et al., 2019). Since now with the term SVs I will refer to the alterations identified by sniffles and cuteSV whereas with the term CNAs those identified by NanoGLADIATOR.

For the CNAs profiling, we run CNAs with Nano-GLADIATOR with 100 kb as windows size parameters so 100 kb is the smallest CNAs identified. We used this threshold because the mean coverage was 3.0X and 2.7X for the exploratory and validation

respectively. Hence, we take 100 Kb as windows size since it could be a window large enough to allow the read count normalization and identification operated by NanoGLADIATOR. We annotated the resulted CNAs using AnnotSV and we filter out those alteration that overlap (50% in length) and have a frequency in the population $\geq 1\%$ with CNAs present in *Database of Genomic Variants* (DGV) (MacDonald et al., 2014).
For SVs identification we run Sniffles and cuteSV with the same parameters, in particular we called SVs only in regions that are at least 5X in coverage and 50 bp in length. As for CNAs profiling we decide to use this threshold ($\sim$ double the average coverage for both cohort) in order to filter out those SVs that could repersent a false positive calling. For each sample we obtain two *Variant Call Format* (VCF) files, one per caller, and we select only the "PASS" alterations that are merged together with SURVIVOR (Jeffares et al., 2017). In the merged VCF file we filter out those alterations that are supported by only one caller to make a high-confidency callset. The high-confidency callsets are then annotated by AnnotSV (Geoffroy et al., 2018) and filtered. as for the CNAs we remove those duplications and deletions that overlap (50% in length) to an existing alterations in DGV database that harbour a population frequency $\geq 1\%$.
All of the tools mentioned above used sorted BAM files as input, while sniffles and cuteSV extract the putative SVs from the input that are then clustered and analyzed to call and genotype SVs, NanoGLADIATOR operates a fragmentation of the genome in discrete windows, in which the number of reads should be calculated and reflects the copy number in the specific genomic location. We decided to profile SVs and CNAs with 2 different approaches for mainly two reasons. On one hand, NanoGLADIATOR is a tool specifically developed for only CNAs; on the other the smallest CNAs identified by NanoGLADIATOR was 100 Kb, thus for an accurate SVs profiling we employed sniffles and cuteSV that are able to identify SVs with a length < 100 Kb.
For both SVs and CNAs callset we filtered out those alterations that matched with a pool of references (Human Reference DNA, Agilent) that we had already sequenced by GridION.
The SVs and CNAs analysis pipeline is shown in Figure 14.

3.6 Statistical Analysis

All the statistical analysis were performed usign R (v 4.0) (R Core Team, 2020). Comparison between two group was determined by t test (when the data distribution would follow a normal distribution and we want to determine if the means of two sets of data are significantly different from each other), Wilcoxon rank sum test (is similar to t-test but it is a non-parametric test so it does not assume that the data is normally distributed) or Chi square test (chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table). The survival analysis was carried out by the R package survival (v. 3.2.13) (Therneau, 2021) using the Kaplan-Meier plots to visualize survival curves, the Log-rank test to compare the survival curves of two or more groups and the Cox proportional hazards regression to describe the effect of variables on survival. The survival analysis focuses on the expected duration of time until occurrence of an event of interest (in this case death). However, the event may not be observed for some individuals within the study time period, producing the so-called censored observations. Censoring may arise in the following ways: a patient has not (yet) experienced the event of interest, such as relapse or death, within the study time period, a patient is lost to follow-up during the study period a patient experiences a different event that makes further follow-up impossible. This type of censoring, named right censoring, is handled in survival analysis.
We performed univariate Cox proportional-hazards analysis by the function coxph of the survival R package (Therneau, 2021) to weight the correlation between the survival time of patients and each predictor variables. This univariate approach describes the survival according to the single factor under investigation, not taking into the account the impact of the other parameter. It allows us to examine how specified factors influence the rate of a particular event happening (*e.g.* death) at a particular point in time. This rate is commonly referred as the hazard rate. Predictor variables (or factors) are usually termed covariates or predictor in the survival-analysis literature.
To better estimate the cumulative impact of multiple genome and clinical variables, we further performed a multivariate Cox

regression analysis selecting those alteration with a LogRank pvalue<0.05 in Cox univariate model, and other predictors such as age and the known molecular abnormalities in specific hotspot genes as recommended by ELN. The output of the Cox model gives:

- **Wald statistic value**. It corresponds to the ratio of each regression coefficient to its standard error (z = coef/se(coef)). The wald statistic evaluates, whether the β coefficient of a given variable is statistically significantly different from 0;
- **the regression coefficients** (β coefficient). Important in the regression coefficient is the sign. A positive sign means that the hazard (*e.g.* risk of death, risk of relapse) is higher, and thus the prognosis worse, for subjects with higher values of that variable;
- **Hazard ratio**. The exponentiated β coefficients, also known as HR, give the effect size of covariates. In another word, a HR >1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival. Conversely, a covariate with an HR <1 is positively associated with the length of survival;
- **Confidence intervals of the hazard ratios**. it gives upper and lower 95% confidence intervals for the hazard ratio.
- **Global statistical significance of the model**. Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score log-rank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

the Forest Plot for Cox proportional hazards models were obtained used the function ggforest of the R package survimer (v. 0.4.9) (Kassambara et al., 2017) for the multivariate models and forest-model (v. 0.6.2) (`https://github.com/NikNakk/forestmodel/`) for the univariate ones.

3.7 Diagnostic of Cox multivariate model

The Cox proportional hazards model makes several assumptions. Thus, it is important to assess whether a fitted Cox regression multivariate model adequately describes the data. We checked the assumptions of the Cox proportional model by testing the proportional hazards assumption, examining the outliers and detecting non linearity in relationship between the log hazard and the statistically significant covariates. In order to check these model assumptions, Residuals method are used and the common residuals for the Cox model include the Schoenfeld residuals (to check the *Proportional Hazard* (PH) assumption), the Martingale residual (to assess the non linearity) and the Deviance residual (symmetric transformation of the Martinguale residuals, to examine the outlier observations).
In principle, the Schoenfeld residuals are independent of time thus a non-random pattern against time is evidence of violation of the PH assumption. The function cox.zph() of the *survival* package test the PH assumption and the function ggcoxzph() of the R package *survimer* offered a graphical visualization for each covariate, of the scaled Schoenfeld residuals against the transformed time. The function cox.zph() provides a convenient solution to test the proportional hazards assumption for each covariate included in a Cox regression model fit. For each covariate, the function cox.zph() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole. The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refused by a significant relationship. In the ggcoxzph() resulted plot the systematic departures from a horizontal line are indicative of non-proportional hazards, since PH assumes that estimates $\beta 1$,$\beta 2$,$\beta 3$ do not vary much over time.
To test influential observations or outliers, we can visualize either: the deviance residuals or the dfbeta values. The ggcoxdiagnostics(type = "dfbeta") function of the package *survminer* allows to check influential observations. It is also possible to check outliers by visualizing the deviance residuals. The deviance residual is a normalized transform of the martingale residual. These residuals should be roughly symmetrically distributed about zero with a

standard deviation of 1. Positive values correspond to individuals that "died too soon" compared to expected survival times. Negative values correspond to individual that "lived too long". Very large or small values are outliers, which are poorly predicted by the model.
Plotting the Martingale residuals against continuous covariates is a common approach used to detect nonlinearity or, in other words, to assess the functional form of a covariate. However, the non linearity is not an issue for categorical variables, anyhow we plotted the Martingale residual for all the covariate statistically significant in the Cox multivariate model.

To the purpose of the study, we performed genome-wide sequencing on a cohort of 152 nkAML patients by long read ONT. Raw data analysis was carried out via the *ad hoc* pipeline we developed for Nanopore data to identify genomic variants of possible impact on patients outcome and estimate their correlation with known molecular abnormalities.

4.1 ELN molecular assessment

We first classified nkAML patients risk according to ELN guidelines basing on the available information on the mutational status of *FLT3*, *NPM1*, *CEBPa*, *ASXL1*, *TP53* and *RUNX1*. The frequency of such mutations in the cohort is reported in Table 5. *NPM1* mutations were the most frequent and it was found in the 54.6% of patients. The 23.68% of the patients presented an ITD in *FLT3*, of those, the 67% with allelic burden >50% (ITD high) and the 33% with ITD low. Bi-allelic mutation in *CEBPa* affected the 7.9% of the cohort. Concerning *ASXL1*, *TP53* and *RUNX1* mutational status, information were available for only 85/152 patients belonging to the Florence cohort. The prognostic stratification of the 152 in the cohort were accomplished following ELN guidelines taking into the account the mutational status of

NPM1, bi-allelic mutations of *CEBPa* and the allelic ratio of ITD mutation in *FLT3* (Figure 15). As shown by the Kaplan- Meier plot for OS (LogRank pvalue .00031), the group of low-risk patients (red line) did not reach the median time for overall survival, conversely those belonging to intermediate (blue line) and adverse (green line) risk showed comparable median OS time, respectively of 15.9 and 16.9 months. The prognostic stratification of the intermediate and adverse populations resulted unsatisfactory in our cohort of nkAML patients. The blue line and the green line intersected making a clear separation between these two populations extremely difficult, this observation largely reflects the inadequate risk assessment for nkAML patients.

4.2 Structural Variants identification in nkAML cohorts

Long reads genome-wide sequencing analysis (decribed in detail in the *Methods* section) was carried out through a multistep analytical pipeline starting from GridION X5 sequencer outputs (raw data) to statistics correlations. Briefly, we selected the *fastq* reads based on the quality score, in ASCII code, assigned to each base; in particular we took only the *fastq* reads with an average quality score 7 for downstream analysis. The high quality reads were aligned to the reference genome (GRCh37) and further subjected to SVs calling by Sniffles and cuteSV. Those SVs reported by both callers (high-confidency callset) were selected for the correlation with patients' outcome by Cox regression model. As explained in detail in the section *Methods*, the high-confidency SVs callset was obtained by merging the SVs callset obtained by Sniffles and cuteSV taking only those alterations supported by both SVs caller. The Venn Diagram in Figure 16 shows the number of SVs identified by each caller and the consensus between them. The SVs identified by cuteSV were 43625 whereas 37584 by Sniffles, the consensus SVs callset was composed by 25502 SVs (high confidency callset) as visualized by the circos plot (Figure 17). The SVs called by only one of two callers, respectively 18125 and 12082, were filtered out. The high confidency SVs callset comprised 13104 insertions (purple dots), 12198 deletions (red dots), 118 duplications (green dots) and 82 inversions (yellow dots) (Figure 17).

Considering high-confidency SVs, the median number of mutation per sample was 396, the histogram in Figure 18 reports the SVs type (insertions, deletions, duplication and inversion) for each sample. The following analytical step was settled to filter out those alterations already reported in healthy population and those with significant overlap (at least 50% in length) with SVs registered in DGV database. Moreover, after z-score calculation for each genomic interval (500 bp) we were able to remove those regions that have a coverage departing from the overall distribution. The filtering strategy allowed to retrieve 14869 SVs not previously reported, whose 7291 were insertions, 7416 deletions, 59 inversions and 102 duplications Figure 19. Once filters were applied, an average number of 112 SVs per patient was calculated (Figure 20). In order extend the landscape of the genomic lesions, we further analyzed raw data to generate a specific CNAs profiling of nkAML patients. To this aim, CNAs spanning at least 100 Kbp were reported after NanoGLADIATOR analysis. The whole identified CNAs were filtered removing those overlapping with donor reference and DGV datasets. We obtained a total of 186 CNAs including 101 deletions and 85 duplications (Figure 21). Regardless CNAs type, the median number of CNAs per sample was 2, as we can see in the histogram of Figure 22, in which the number of CNAs, labeled as deletions and duplication, were shown for each sample.

4.3 Cox Regression Analysis

In order to investigate the potential correlation of selected SVs (also referred as covariates) with patients' survival, we perform univariate Cox regression analysis. The univariate Cox regression outputted the statistical significance, the regression coefficient, the HR, the Confidence intervals of the hazard ratios and the Global statistical significance of the model for each covariate. Based on this parameters, we selected only those univariate models with a LogRank pvalue$<$ .01 taking only the SVs with a highly statistically significant impact on OS. As resulted from univariate analysis, 51 covariates associated with an increased risk of death with an HR $>$1, those are reported in the forest plot in Figure 23. The further multivariate Cox regression prompted us to describe

how the covariates jointly impact on survival. For multivariate regressions, the input variables were the 51 covariates previously selected together with the mutational status of *FLT3* (with the information of ITD allele burden), *CEBPa* and *NPM1*. We repeated the construction of multivariate model until we obtained a model with all statistically significant covariates (selecting at every step the covariates with LogRank pvalue <.05).
The model we developed was composed by 12 SVs, 8 with an HR >1, thus associated with an increased risk of death, and 4 with an HR <1, associated with reduced hazard (summarized in the Forest plot in Figure 24). Overall, the model resulted highly significant (Likelihood ratio test=66.88 on 13 df p=3e-09,Wald test=73.87 on 13 df p=2e-10 and LogRank test=144.6 on 13 df p=<2e-16). The 12 model-selected SVs involved insertions and deletions affecting several chromosomes and summarized in Table 6. The plot in Figure 25 shows the baseline predicted survival curve at any given point in time. Next, we displayed the correlation of each covariate with patients survival. Given that, we estimated the impact of each single SVs with an HR >1 (referred as hrSVs) at any time point on the estimated survival probability. As shown In the Figure 26 the probability of fatal events (death) is significantly increased in patients harboring at least 1 hrSVs at time compared to hrSVs-negative group.
The plot in Figure 27 shows the Schoenfeld residuals against the transformed time for each covariate. In the figure the solid line is a smoothing spline fit to the plot, with the dashed lines representing a $\pm 2\sigma$ band around the fit. From the graphical inspection, there is no pattern with time for all the covariates except a slightly dependence for the insertion in chr15:72061719, nonetheless the global Schoenfeld test is not significant (p=.06) thus we can accept the PH assumption. By comparing the magnitudes of the largest dfbeta values to the regression coefficients, we can speculated that none of the observations was terribly influential individually, even though some of the dfbeta values for the deletions at chr11:2026821-2026936 and the deletions at chr14:80106289-80115050 were large compared with the others (Figure 28). The deviance residual was a normalized transform of the martingale residual and were visualized in the Figure 29, where we can see that the pattern looks fairly symmetric around 0.
The plot in Figure 30 shows the Martingale residuals of null Cox

proportional hazards model.

4.4 Evaluation of the Cox multivariate model

Considering the cluster of patients with the 8 hrSVs, we further investigated the clinical and molecular characteristics of those patients (from now referred as high-risk patients). The Kaplan-Meier in Figure 31 shows the survival analysis by stratifying patients basing on the hrSVs presence, comparing those patients with at least one of the above mentioned SVs and the negative-ones (from now referred as low-risk patients). The two distinct populations showed a statistically significant different median OS time of 8.27 and 62.67 months (LogRank pvalue <.0001). To deepen the analysis, we categorized the whole cohort basing on the hrSVs number per patient. The Figure 32 shows the output of such stratification identifying three population with statistically significant (LogRank pvalue <.0001) median OS time of 52, 9 and 5 months, respectively for the low-risk group, high-risk group harbouring 1 hrSVs and high-risk group with more than 1 hrSVs (shown in red, green and blue line respectively). However, the high risk groups with 1 hrSVs or more than 1 hrSVs were composed by 16 and 10 patients respectively, thus, given the poor numerosity of such selected casistic, the LogRank test was not statistically robust. As we can see in the Figure 32, the green line and the blue line intersected, probably due to the small number of patients in each groups.
High risk patients had a median age at diagnosis of 54 years compared to 49 years for low-risk (not statistically significant, Wilcoxon test p-value = 0.07937); mean white blood cell count was 48231 for high-risk patients and 43977 for low-risk (not statistically significant, T-test p-value = 0.7012). CR at first induction therapy was achieved by 13 high-risk patients (the 46% of the high-risk cohort) and 102 patients low-risk patients (the 80% of the low-risk cohort) (Chi square test pvalue <.00000000000000022). Considering the 49 patients that benefitted of allogenic HCT, 6 were high-risk patients, corresponding to the 23%, and 46 low-risk patients, corresponding to the 36%. The clinical information of the high-risk group were summarized in Table 7 while Table 8 reports low-risk group clinical data.

Only 3 high-risk patients were previously classified with adverse prognosis following ELN recommendations, whereas 13 and 10 resulted with intermediate and favourable risk, respectively.

4.5 Refinement of ELN prognostic stratification

Considering that ELN prognostic assessment for the 26 high-risk patients classified 3 patients as favourable, 13 patients as intermediate and 10 patients as adverse (Table 9). We incorporate hrSVs information along with ELN molecular alterations (*FLT3*, *NPM1* and *CEBPa*) in order to re-fine their risk-assessment. As described in the section *ELN molecular assessment*, the prognostic assessment based on ELN recommendation resulted in an unsatisfactory distinction between patients belonging to the intermediate and adverse categories. The survival analysis in Figure 33 shows the stratification by merging together the intermediate and the adverse populations, resulting in two populations: one with a good prognosis and the other experiencing a poor survival (median OS = 16 months, LogRank pvalue = <.0001).
The addition of hrSVs to the ELN prognostic model allowed us to identify three different clearly distinct populations (Figure 34, right panel, LogRank pvalue <.0001). The high-risk patients (blue line) showed the poorest prognosis (median OS = 8.27 months), conversely, patients belonging to ELN intermediate and adverse group (green line) had a median OS time of 22.57 months and ELN favourable were still those with the best prognosis (median OS not reached). We also stratified our cohort with the 3 categories identified by ELN identifying 4 populations with a statistically significant OS (LogRank pvalue<.0001); the favourable group did not reach the median OS while for the intermediate, adverse, Cox multivariate model's SVs were 22.57, 16.90 and 8.27 months respectively (Figure 34, left panel)

Discussion 5

The advent of the massive parallel sequencing by technologies boost our capability to investigate genomic region at base resolution level allowing the discovery of novel molecular abnormalities and fostering considerable progress into the understanding of disease pathogenesis as well as in the development of diagnostic assays and novel therapies. The undeniable advantages granted by NGS are, on the other hand, limited by the relatively small length of the reads produced that impairs genome-wide studies and shows inadequate to resolve genomic SVs spanning several kilobases or enriched in repetitive elements. Recent studies pointed out the prominent role of SVs in several human *neoplasia* development but, in turn, such type of alteration is largely understudied and constitute a technical challenge for the conventional sequencing technologies. SVs are defined as DNA rearrangements 50 bp and include (deletions and duplications) as well as insertions, inversions, translocations and more complex combinations of these described events. It has been demonstrated that cancer onset and progression could be triggered by the accumulation of structural abnormalities in the genome as the result of an increased genome instability. Somatically acquired SVs could lead to cancer onset by deactivating tumor suppressor genes and upregulating oncogenes. The detection and classification of these variants could improve

our understanding of pathologic mechanisms and ameliorate diagnosis, prognosis and therapy strategies for cancer patients (Hayes, 2019). In the context of hematologic malignancies, AML is the most frequent leukemia in adults worldwide. Karyotype is the most important independent factor to forecast patients' outcome, notwithstanding, nearly half of AML patients lack a known genetic determinants allowing a satisfactory prognostic assessment. These cases, mostly belonging to the ELN intermediate risk category, are defined as normal karyotype . As sketched before, in 2010 ELN reccommendations for prognosis included additional molecular biomarkers (*NPM1*, *FLT3* and *CEBPa*) together with known chromosomal aberrations aiming to provide a prognostic classification based on cytogenetic and specific genes mutations. The further refinement of ELN guidelines in 2017 widens the list of prognostically relevant genes by including *ASXL1*, *RUNX1* and *TP53*. In spite of that, the extremely heterogeneous survival and response to treatment observed in nkAML patients underline the need to better subgroup such normal karyotype population and let hypothesize that additional, not yet detected, molecular determinants play a driver role into disease development. Long-read sequencing has the potential to overcome short-reads limitations and to improve SVs resolution both in comparative as well as in clinical studies. It has been shown that, from a computational perspective, repeats create ambiguities in short-read alignment and assembly which, in turn, introduces errors in calling genetic variants. The nanopore sequencing technology relies on extremely long-reads (up to 20 kb) enabling the study of those challenging twilight region of human genome . Taking into the account all these observations, we reasoned to investigate the pathologic genomes of nkAML patients by exploiting a novel sequencing technology based on long reads with the aim to detect those hidden SVs too small in size to be addressed by conventional cytogenetics and too big and/or structurally complex to be resolved by NGS. The data generated in this study demonstrate the high potential of the depicted sequencing and analytical approaches, granting an accurate (∼97%) and comprehensive characterization of genomic SVs of prognostic impact in the context of nkAML. The analysis of 152 pathologic genomes led to the identification of a cluster of SVs ranging in size between 80 to 8000 bp significantly correlated to the survival of the selected cohort. Despite the

study limitations, constituted by the lack of *ASXL1*, *RUNX1* and *TP53* data availability for 67 patients, the reported findings on the cohort of 152 patients has the potential to refine the ELN prognostic stratification for 26 patients (high-risk), corresponding to the 17% of the total included in the study. Those 26 patients shared a variable number of hrSVs identified by multistep statistical analysis. In particular, the multivariate Cox model we developed was composed by 12 SVs that jointly impact on survival. Of those 12 SVs, 8 were defined as hrSVs because their negative impact on patients' survival, conversely, the remaining 4 covariates (including *CEBPa*) had a slightly positive impact on survival patients and they were found together with hrSVs in patients of the cohort, additionally *CEBPa* and hrSVs were mutually exclusive in our cohort. The 26 patients re-categorized as high-risk nkAML were previously assessed with adverse risk in 3 cases, intermediate risk in 13 cases and favourable risk in 10 cases following ELN parameters only. The median OS time of high-risk patients (8.27 months) significantly differ from the median survival of ELN categories, in particular for those patients assigned to the favourable and intermediate risk groups (n=23, 15.1%). Moreover, the rate of CR achieved in the high risk group was the 46% (n=13/26) compared to the 80% (102/152) of the low-risk group; this observation let us speculate that the identified hrSVs also play a role on the response to therapy and may be further investigated in order to unravel the molecular mechanisms underlying therapy refractoriness. Beside the new genomic findings in the nkAML cohort, the present study also delineates a robust and reproducible analytical workflow specifically settled for nanopore long-read sequencing data that could be easily implemented and applied to the analysis of various datasets. The developed pipeline comprises all the bioinformatics steps from raw sequencing data to SVs identification. More in depth, the pipeline is composed by (a) the basecalling, in which the raw current signal is converted to a string of nucleotides of a read and the quality control, (b) the reads quality filtering, to remove low quality reads (quality score >7), (c) the genome anchorage to a reference, (c) the variant calling, (d) the building up of a high-confidency callset of SVs, (e) the SVs filtering for false positive calls, variants of populations and poor/high-coverage regions departing from the overall distribution. In conclusion, this study provides a proof of applicability of long

reads sequencing to the identification of novel biomarkers with pathologic relevance in the context of AML and further applicable to other hematologic or solid *neoplasia*. The presented data report for the first time 8 SVs with adverse impact on nkAML survival and response to therapy, eventually enabling the refinement of the prognostic forecasting and related risk adapted therapies for a considerable amount of patients, moving a step towards a tailored medicine in such uncertain *scenario* of normal karyotype Acute Myeloid Leukemia.

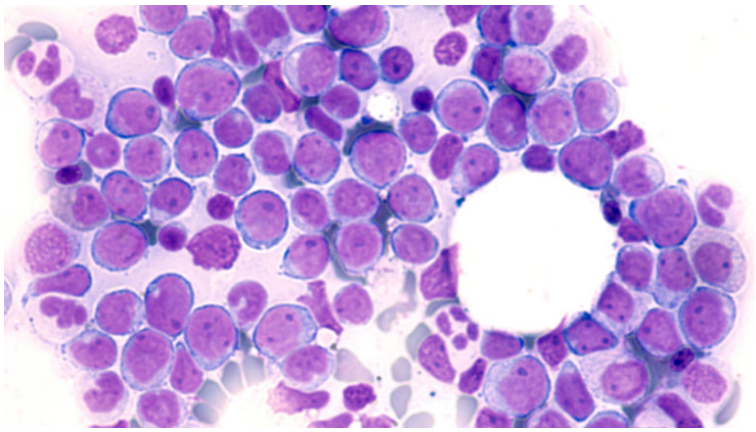

FIGURE 1: `Acute Myeloid Leukemia`. Bone Marrow aspirate cytology (FAB M1). Figure is taken from `https://www.oncotarget.org/2021/05/12/tomivosertib-versus-acute-myeloid-leukemia/`

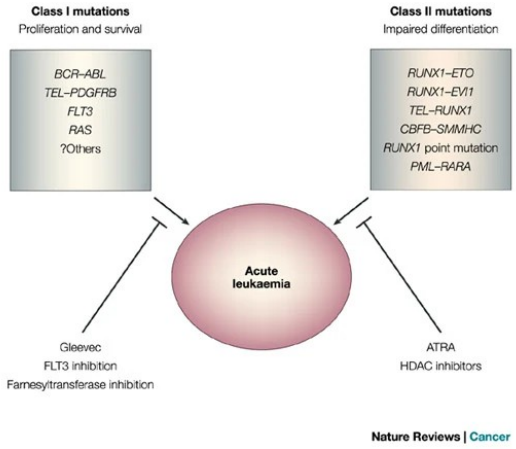

Figure 2: `The two-hit model for leukemogenesis.` The class I mutations result in enhanced proliferative and survival advantage for haematopoietic progenitors whereas class II mutations are associated with impaired haematopoietic differentiation. Figure is taken from *Core-binding factors in haematopoiesis and leukaemia* (Speck and Gilliland, 2002)

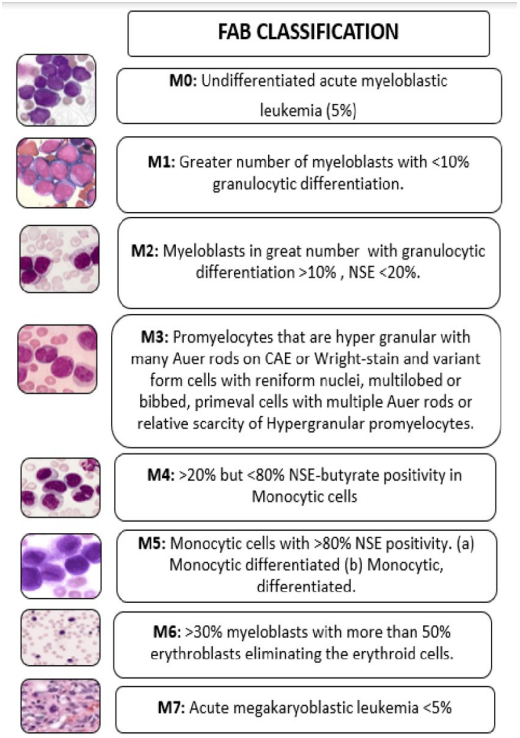

FIGURE 3: `FAB classification of AML`. Figure is taken from *Insights into Acute Myeloid Leukemia: Critical Analysis on its Wide Aspects* (Khan et al., 2020)

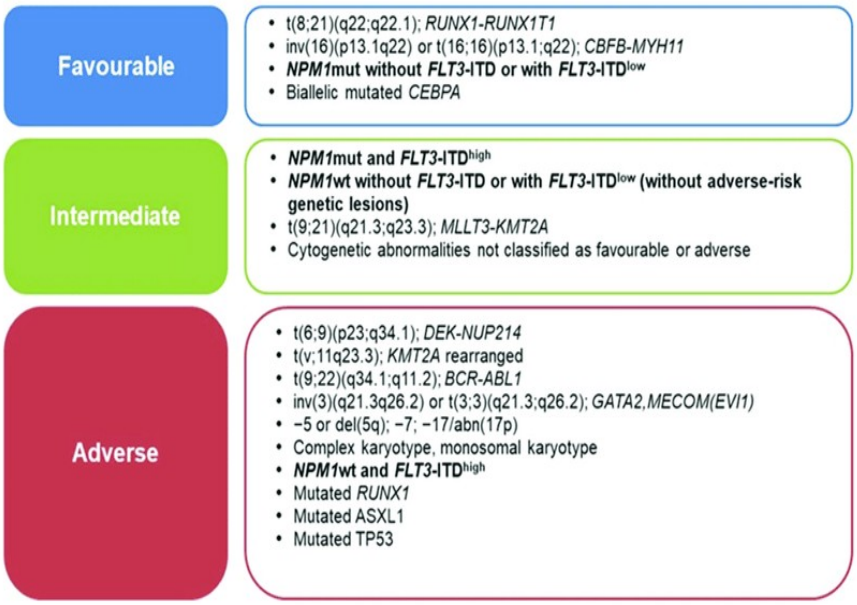

Figure 4: `ELN 2017 prognostic assessment of AML`. Figure is taken from *MRD in AML: The Role of New Techniques* (Voso et al., 2019)

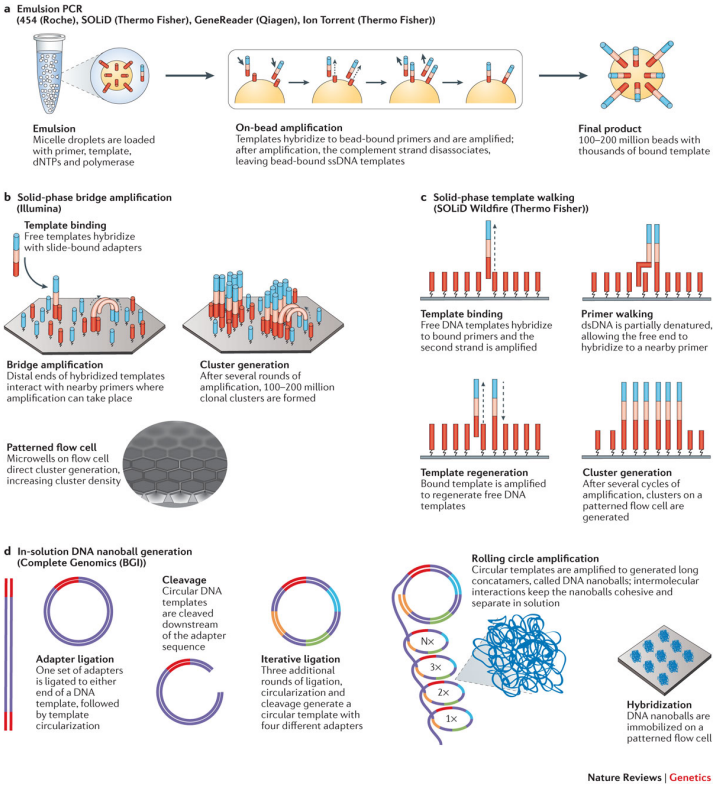

FIGURE 5: `Template amplification strategies`. Different strategies used to generate clonal DNA template populations: bead-based generation (**a**), solid-state generation (**b**,**c**), DNA nanoball generation (**d**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

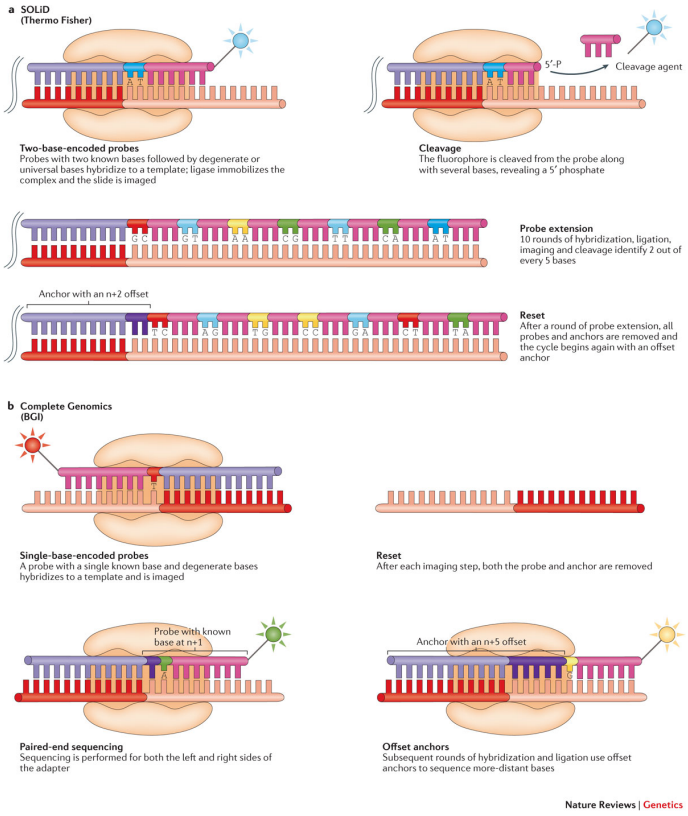

FIGURE 6: `SBL methods`. Summary of the SBL approaches by SOLiD (**a**) and Complete Genomics (**b**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

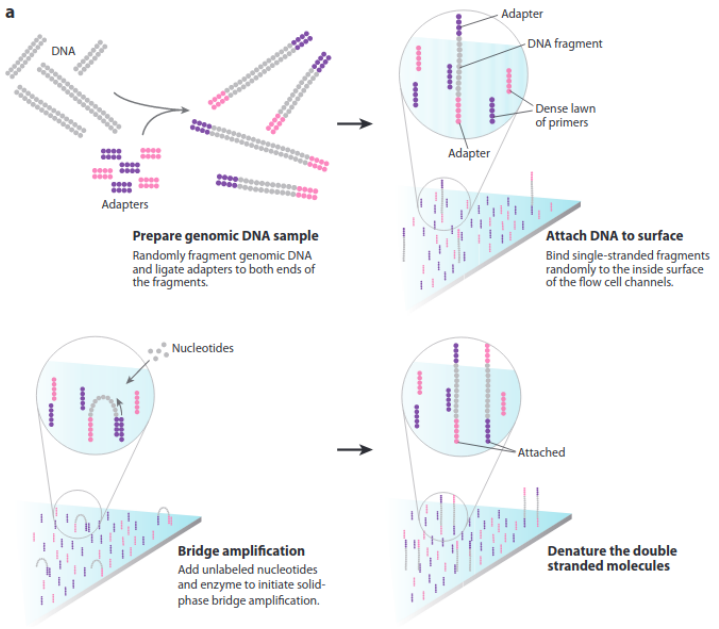

FIGURE 7: `Illumina SBS approach`. SBS approach by Illumina platforms (**a**) fragmentation of genomic DNA and ligation to adapter (upper left panel); attaching of the DNA to the solid surface (upper right panel); bridge PCR amplification (bottom left panel); fluorescence label and 3′-blocked cleavage. Figure is taken from *Next-Generation DNASequencing Methods* (Mardis, 2008)

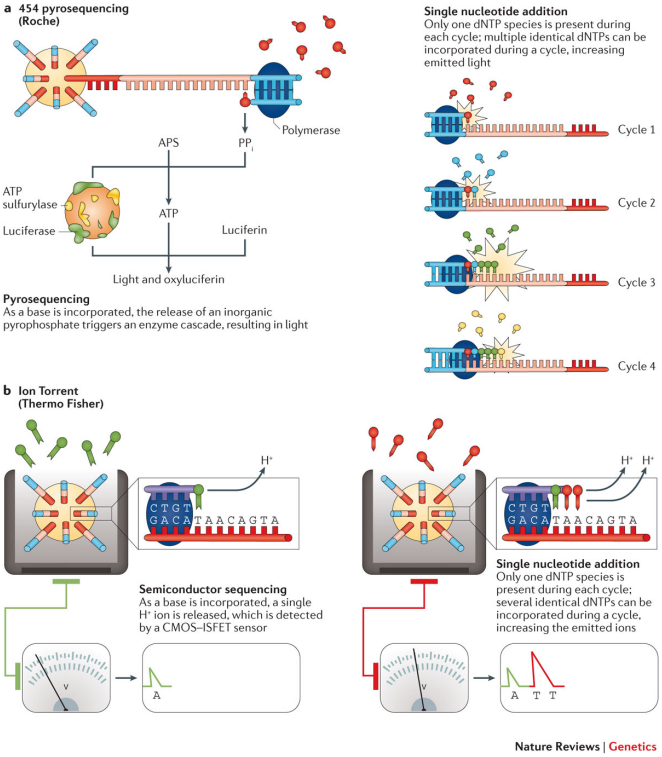

FIGURE 8: `Roche 454 pyrosequencing and Ion Torrent`. Summary of the NGS technology by Roche (**a**) and Thermo Fisher Scientific (**b**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

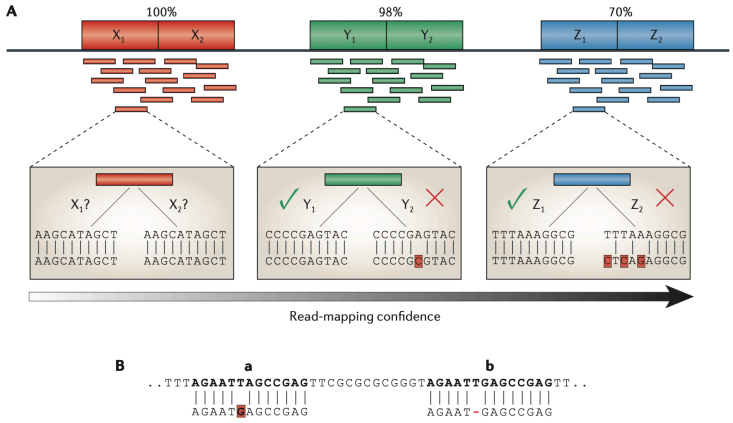

FIGURE 9: `Ambiguities in read mapping`. As the difference between two copies of a repeat increases, the confidence in any read placement within the repeat increases as well (**A**). When a read maps equally well to two different locations, this is assigned to either the first or the second depending on the score given by the aligner to mismatches and gaps (**B**). Figure is taken from *Repetitive DNA and next-generation sequencing: computational challenges and solutions* (Treangen and Salzberg, 2012a).

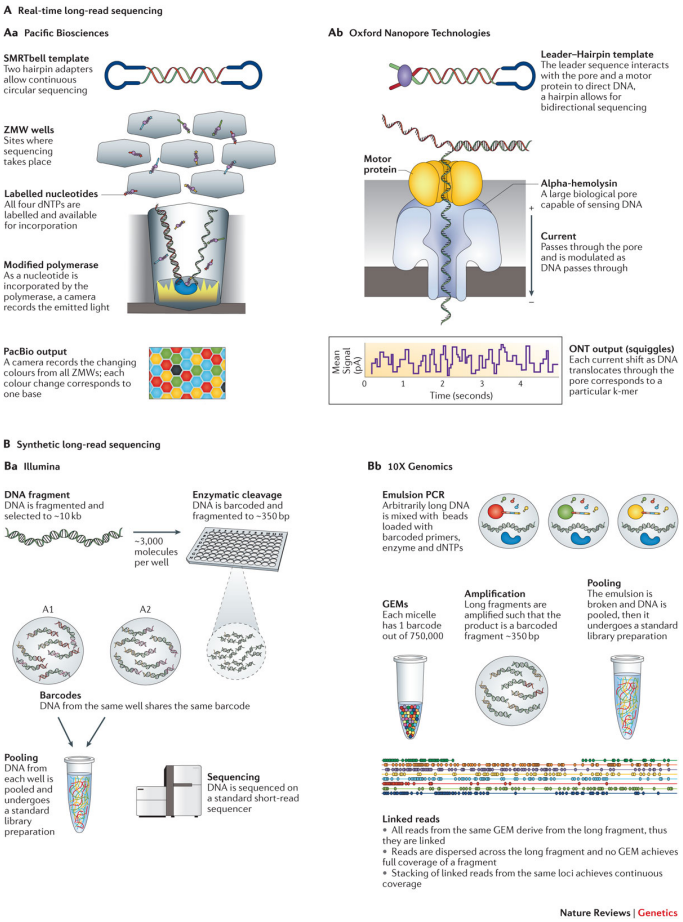

FIGURE 10: `Long-read sequencing approaches`. Different strategies used to generate long reads: SMRT sequencing by PacBio (**Aa**), nanopore sequencing by ONT (**Ab**), Synthetic long-read sequencing by Illumina (**Ba**) and 10X Genomics (**Bb**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

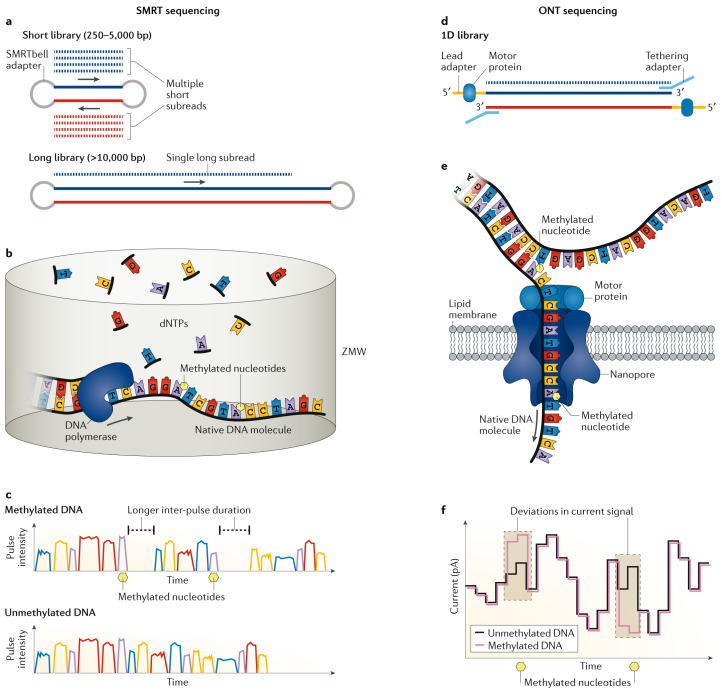

FIGURE 11: `Detection of base modifications by TGS technologies PacBio and ONT`. Different strategies used to identify nucleotides epigenetically mofified using SMS: SMRT sequencing by PacBio (**a,b,c**) and nanopore sequencing by ONT (**d,e,f**). Figure is taken from *Deciphering bacterial epigenomes using modern sequencing technologies* (Beaulaurier et al., 2019).

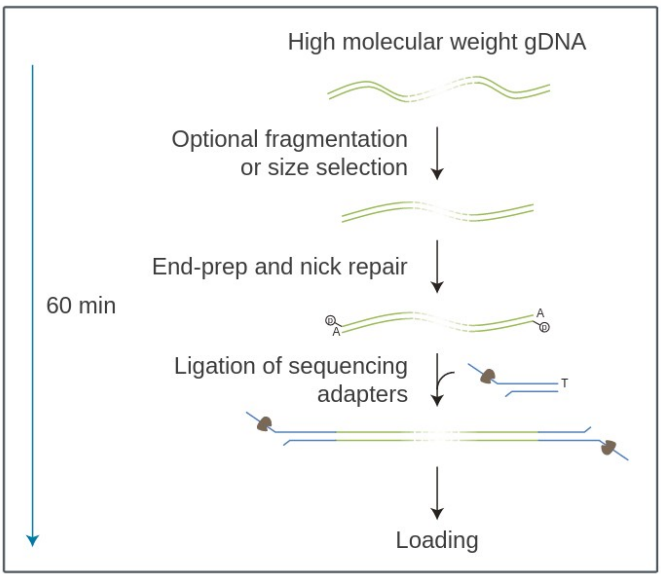

Figure 12: `Overview of the protocol SQK-LSK109`. The high molecular weight DNA is end-prep and nick repaired prior to adapter ligation. After the ligation of the adapter the library will be loaded into the flowcell. Figure is taken from `https://store.nanoporetech.com/eu/ligation-sequencing-kit.html`

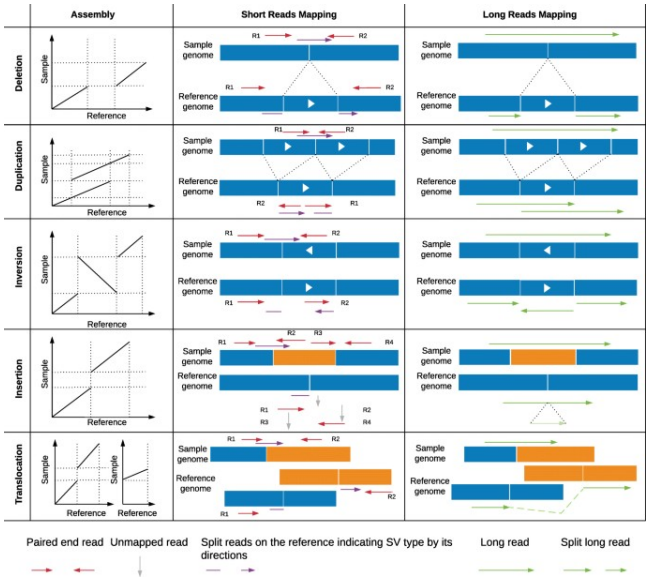

FIGURE 13: `Comparison of SVs detection approaches by short reads and long reads`. On one hand, in the short reads approaches (central panel) the identification of the type and the size of SVs are performed by paired-end (red) and split reads (purple). Moreover, the coverage can be used to improve the detection of deletions and duplications; on the other (long-read-based mapping approaches, right panel) the alignment patterns of long reads (green) are used to detect the different types of SVs. Figure is taken from *Structural variant calling: the long and the short of it* (Mahmoud et al., 2019)

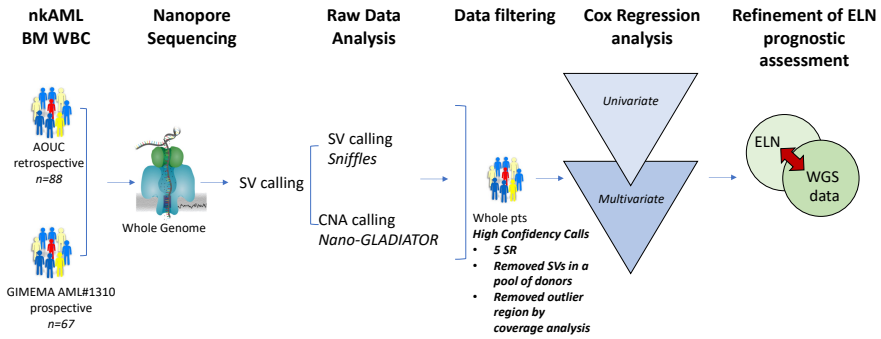

FIGURE 14: `Pipeline applied to analyze Nanopore data`. 85 samples were collected from Florence hematology unit whereas 67 from the GIMEMA AML #1310. The SVs calling was performed by Sniffles and cuteSV taking the consensus SVs callset. After data filtering we performed univariate and multivariate Cox regrssion models in order to refine the ELN prognostic assessment.

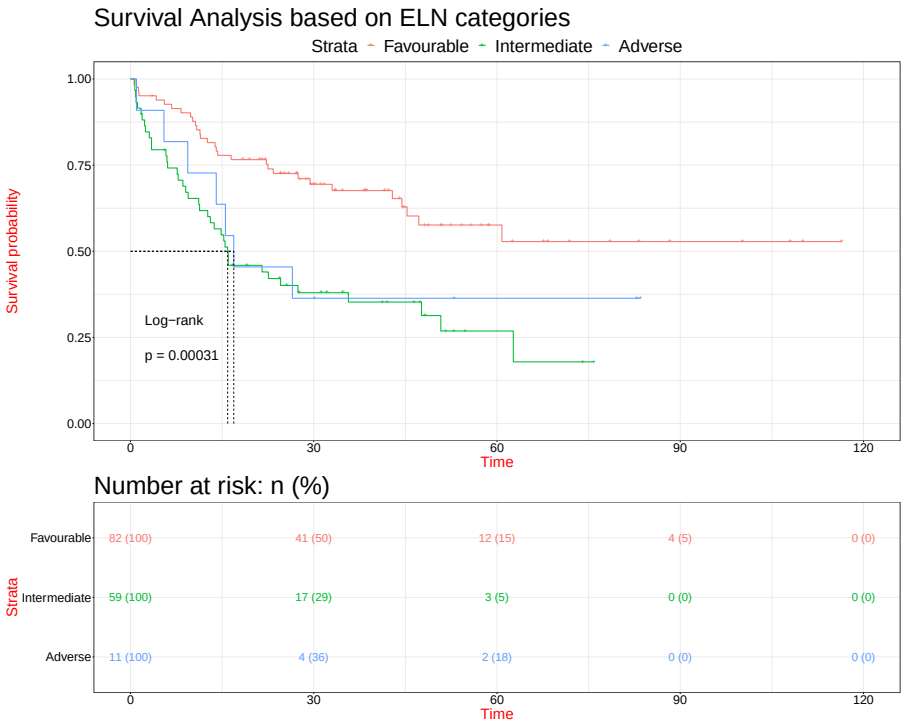

FIGURE 15: `ELN prognostic assessment of whole cohort`. The survival analysis based on ELN gene panel identified three population, named favourable, intermediate and adverse, with a statistically different median OS.

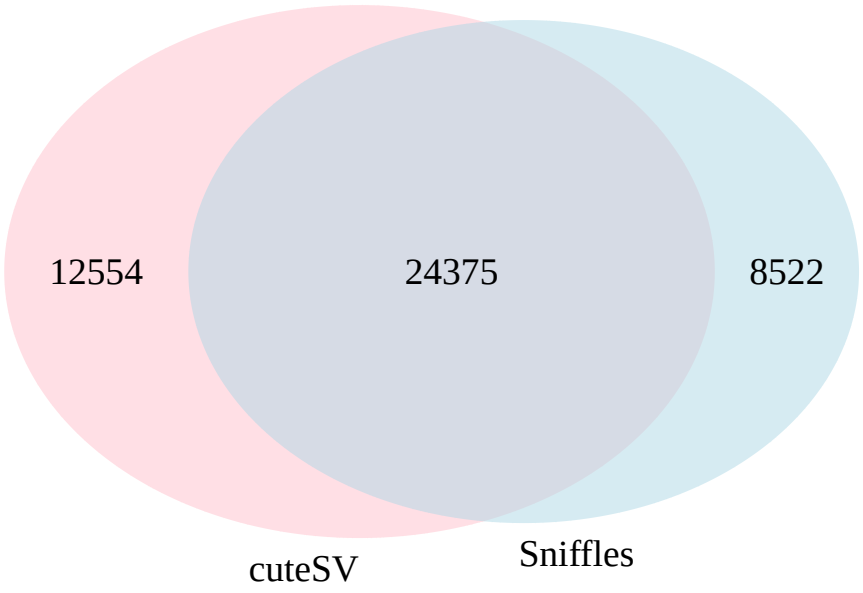

Figure 16: `Venn Diagram of the SVs`. The Venn Diagram shows the number SVs called by Sniffles and cuteSV separately and the number SVs shared by both.

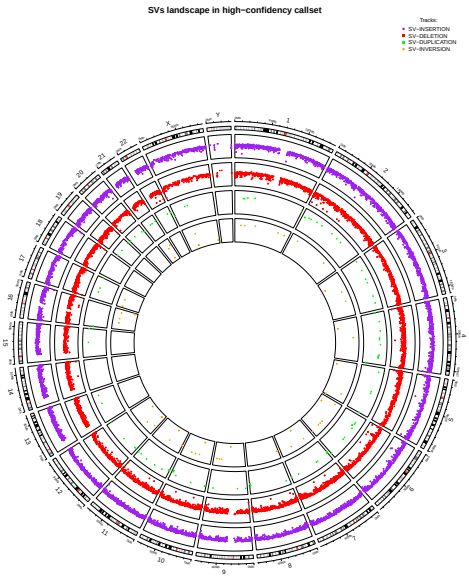

FIGURE 17: `Circos Plot of the SVs in the high-confidency callset`. The total number of SVs was 25502 divided in 13104 insertions (purple dots), 12198 deletions (red dots), 118 duplications (green dots) and 82 inversions (yellow dots)

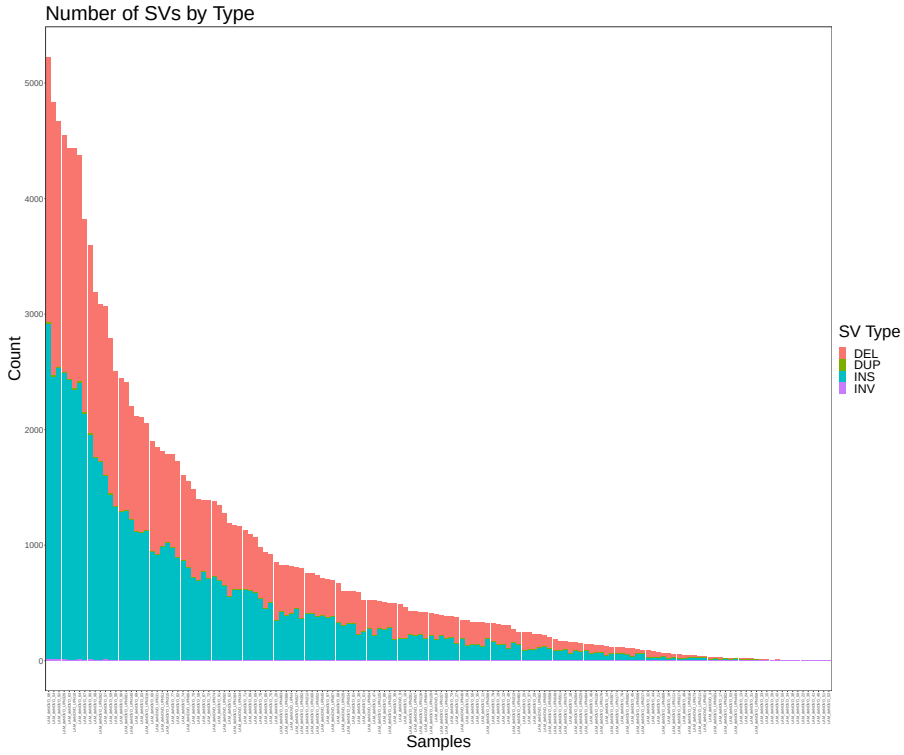

Figure 18: `Histogram of the SVs in high-confidency callset for each sample`. The plot shows the number of SVs in each sample splitted by insertions, deletions, inversions and duplications. As we can see the majority of SVs is represented by insertion and deletion. The red line represent the median number of SVs per sample

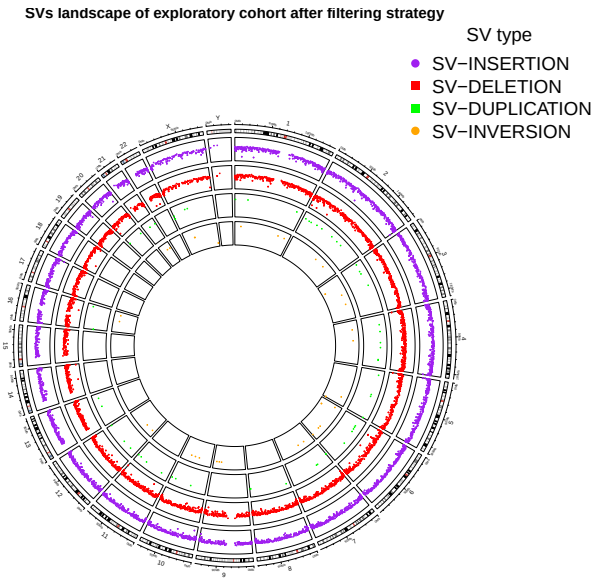

FIGURE 19: `Circos Plot of the SVs after the filtering`. The total number of SVs was 14869 divided in 7291 insertions (purple dots), 7416 deletions (red dots), 102 duplications (green dots) and 59 inversions (yellow dots)

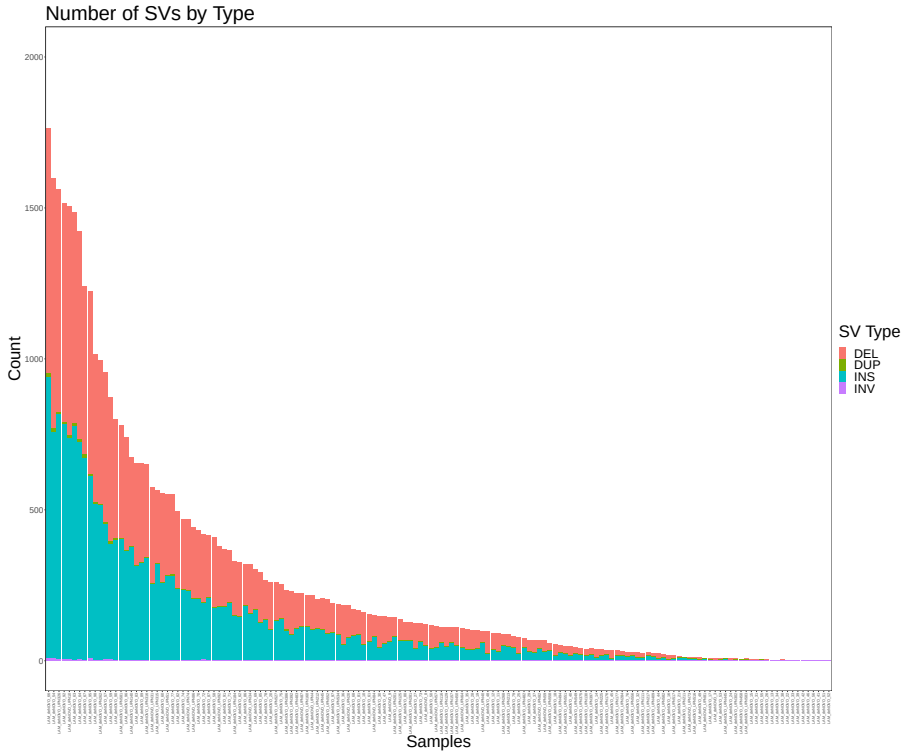

Figure 20: `Histogram of the SVs after filtering in each sample`. The plot shows the number of SVs in each sample splitted by insertions, deletions, inversions and duplications. As we can see the majority of SVs is represented by insertion and deletion. The red line represent the median number of SVs per sample

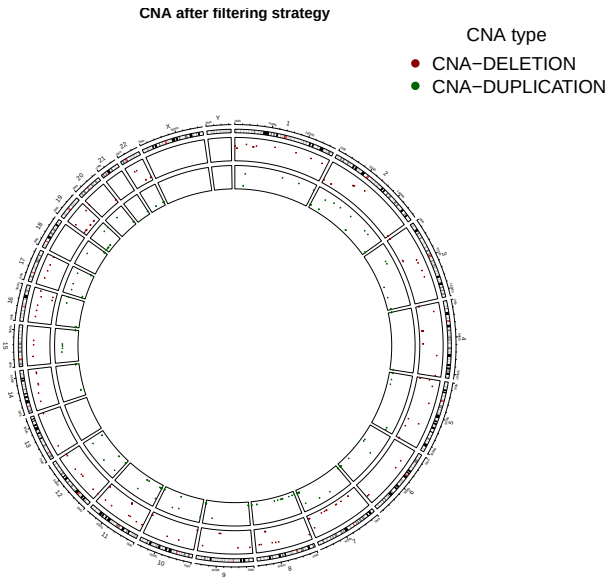

FIGURE 21: `Circos Plot of the CNAs after the filtering`. The total number of CNAs was 186 divided in 101 deletions (red dots) and 85 duplications (green dots).

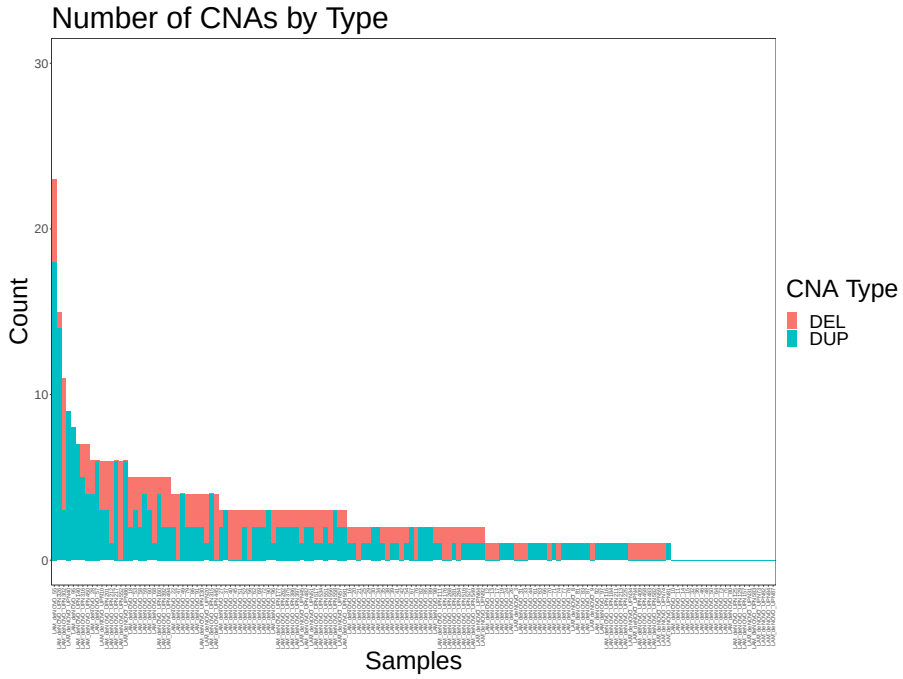

Figure 22: `Histogram of the CNAs after filtering in each sample`. The plot shows the number of SVs in each sample splitted by insertions and deletions. The red line represent the median number of SVs per sample

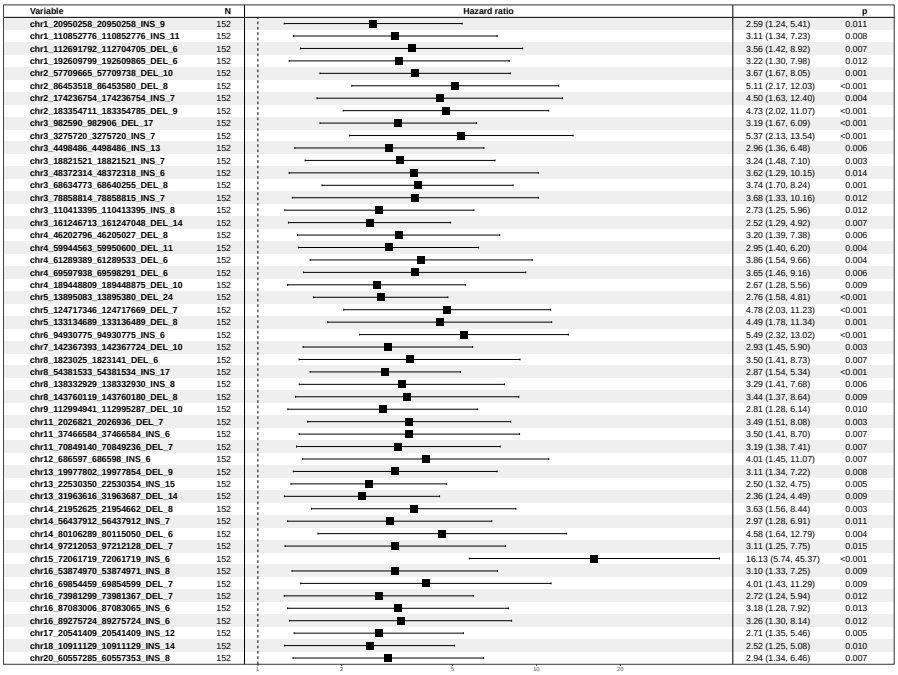

FIGURE 23: `Forest Plot of the covariates with p<.01 in univariate analysis`. The Forest Plot shows the HR and the relative LogRank pvalue

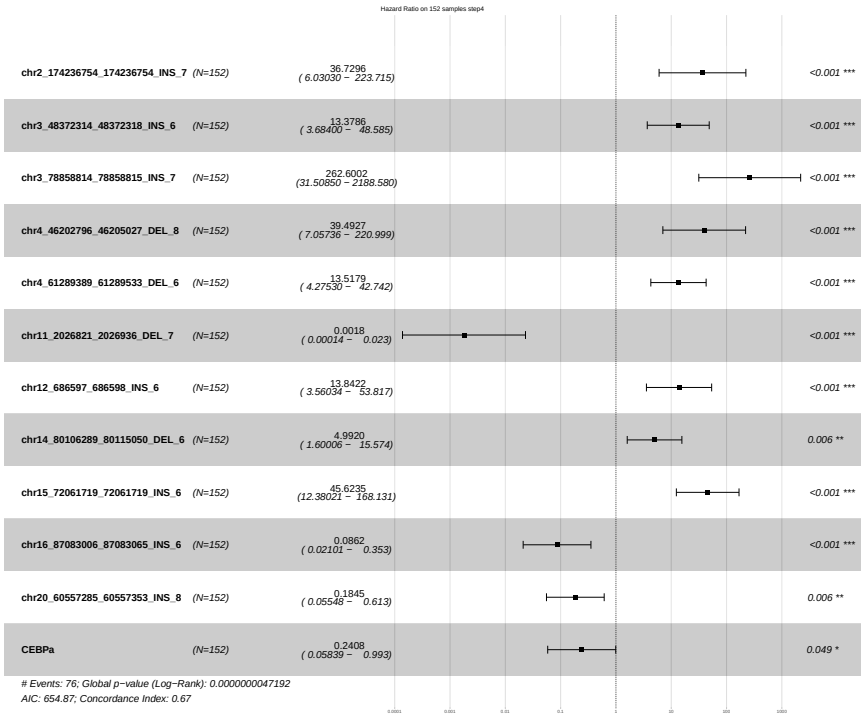

FIGURE 24: Forest Plot of the covariates with p<.05 in the final step of the Cox multivariate regression analysis

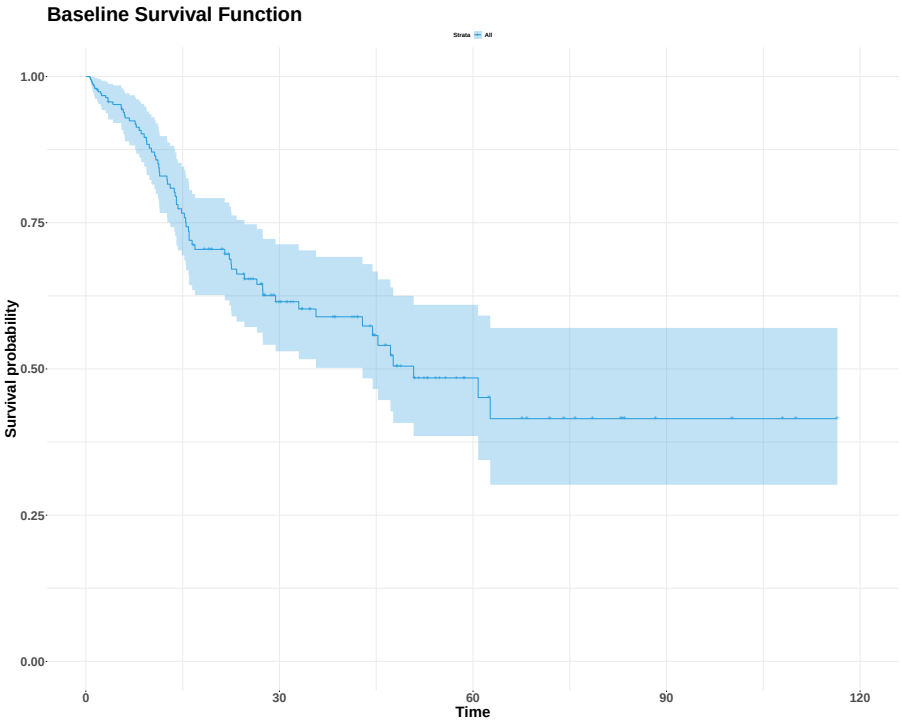

FIGURE 25: `Baseline of survival curves`. The plot visualize the predicted survival proportion at any given time point.

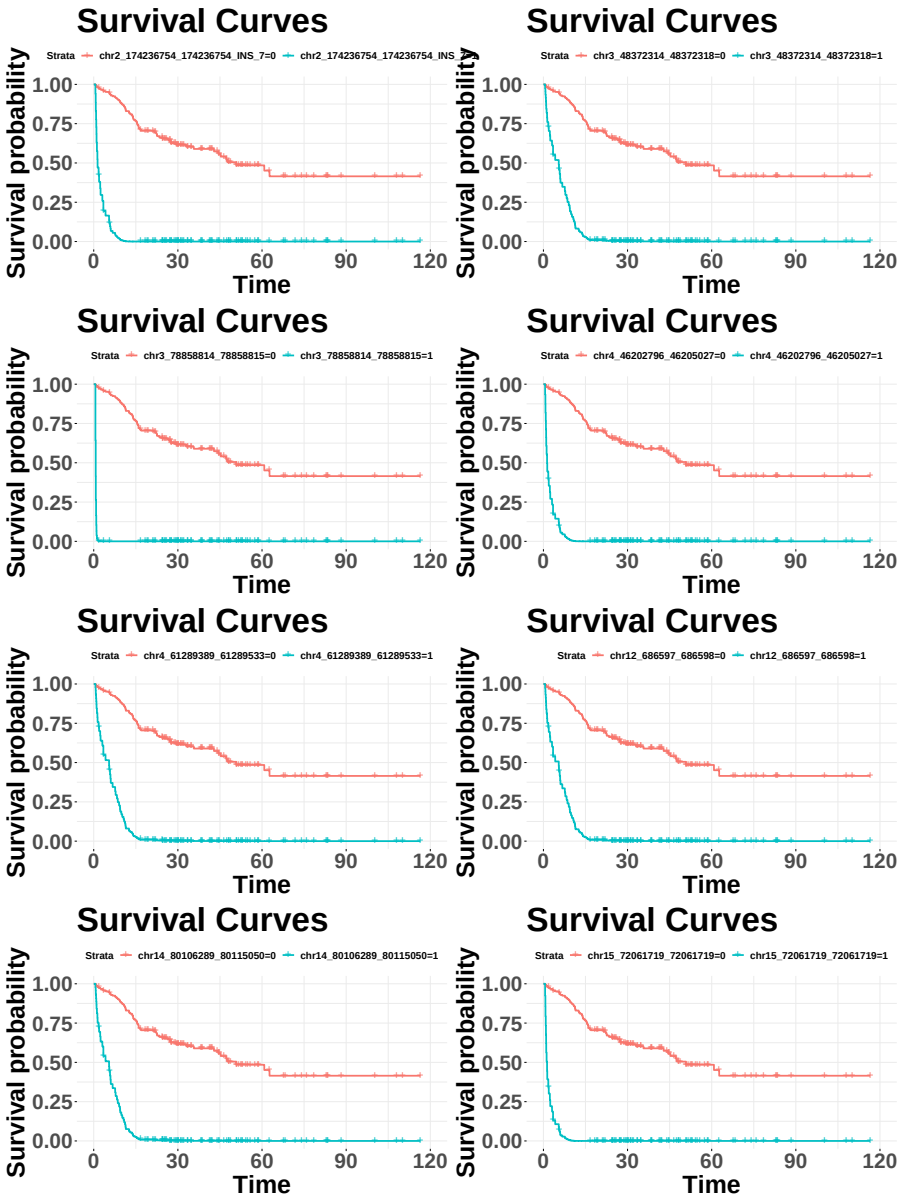

FIGURE 26: `Survival Curves`. The plot shows the probability that the event occurs in patients harbouring each covariates once at time

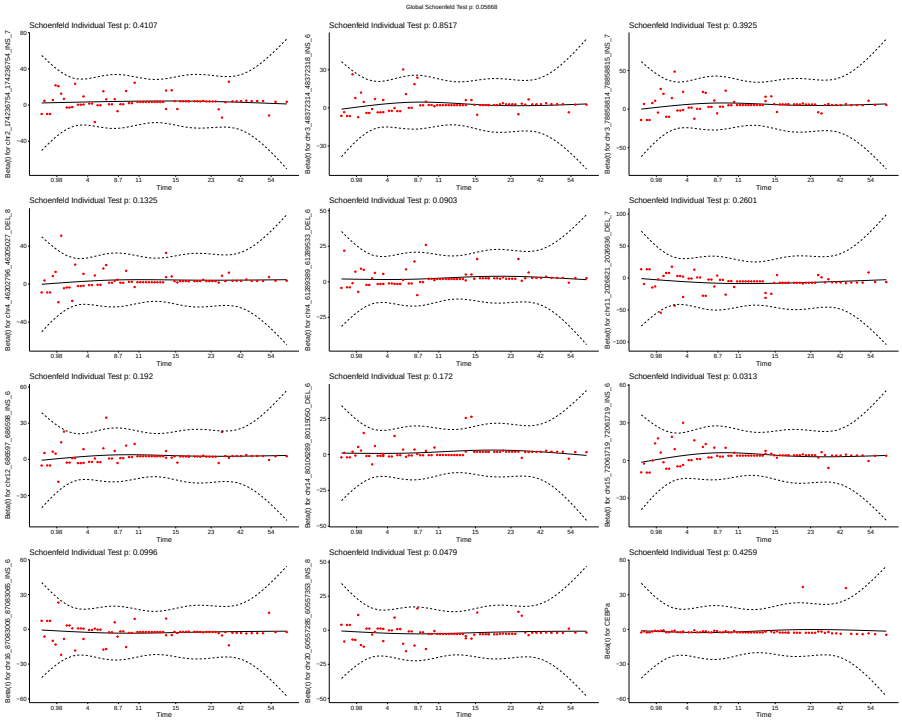

FIGURE 27: `Schoenfeld residuals`. The plot shows the the Schoenfeld residuals against the transformed time for each covariates, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a +/- 2-standard-error band around the fit.

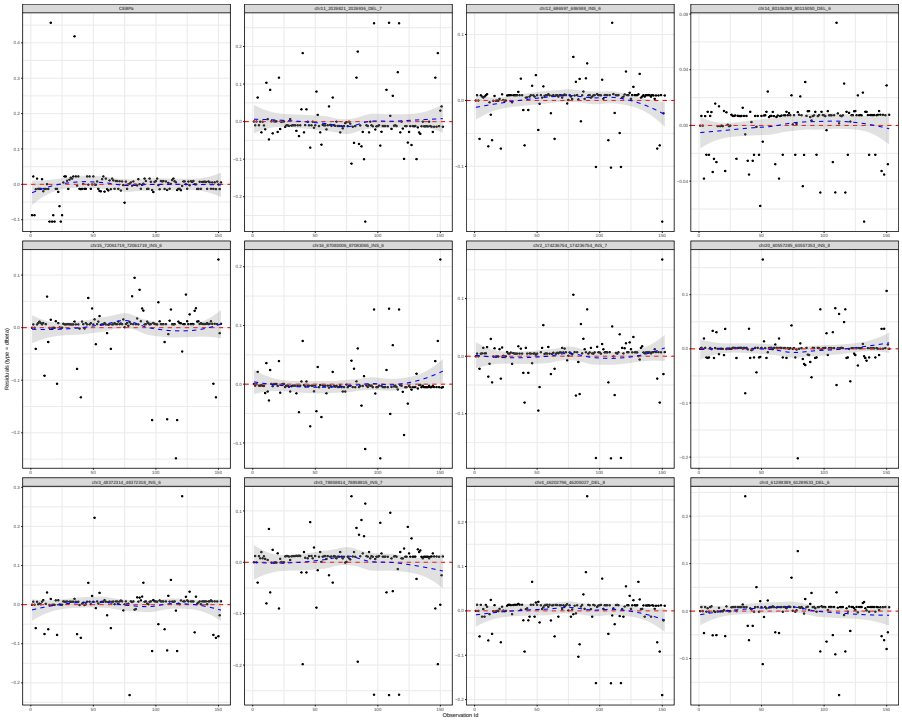

Figure 28: `Index plots of dfbeta for the Cox regression of time to death for each covariate.` The plot shows the estimated changes in the regression coefficients

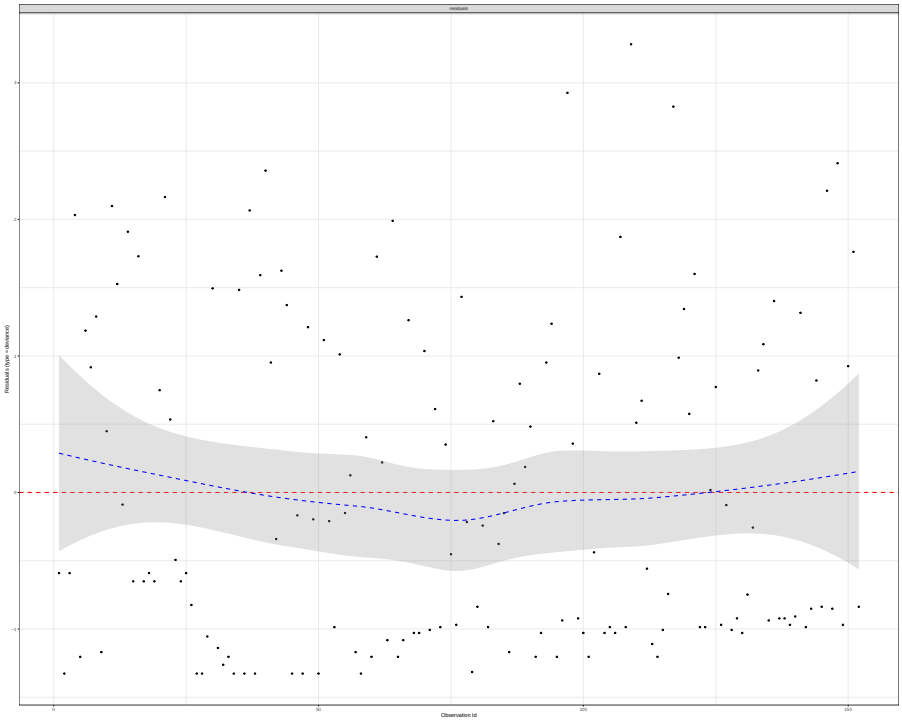

FIGURE 29: `Deviance residual (symmetric transformation of the Martinguale residuals)`. This plot shows the index number observation against the residual deviance.

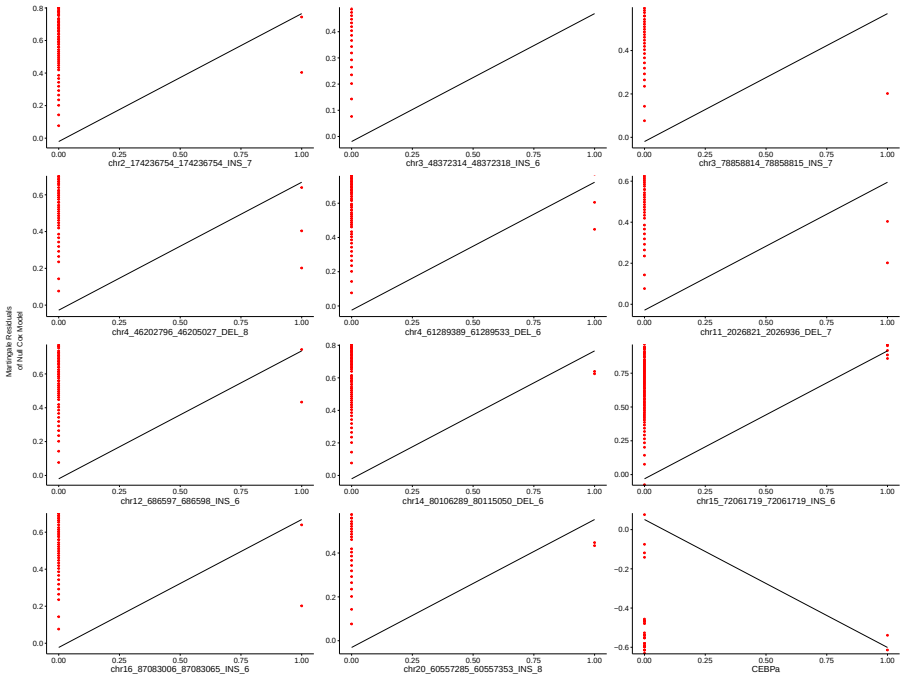

FIGURE 30: `Martingale residuals against the covariates`. This approach is used to detect the non linearity in order to assess the functional form of a covariate

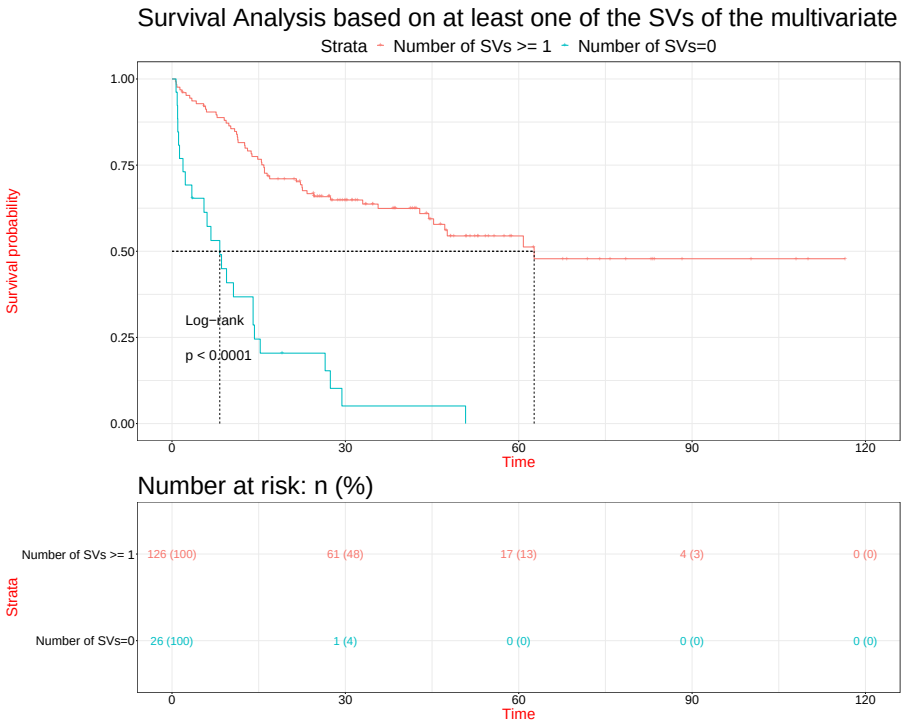

FIGURE 31: `Survival Analysis`. This survival analysis stratified our cohort based on the presence of at least 1 of the Cox multivariate model's SVs

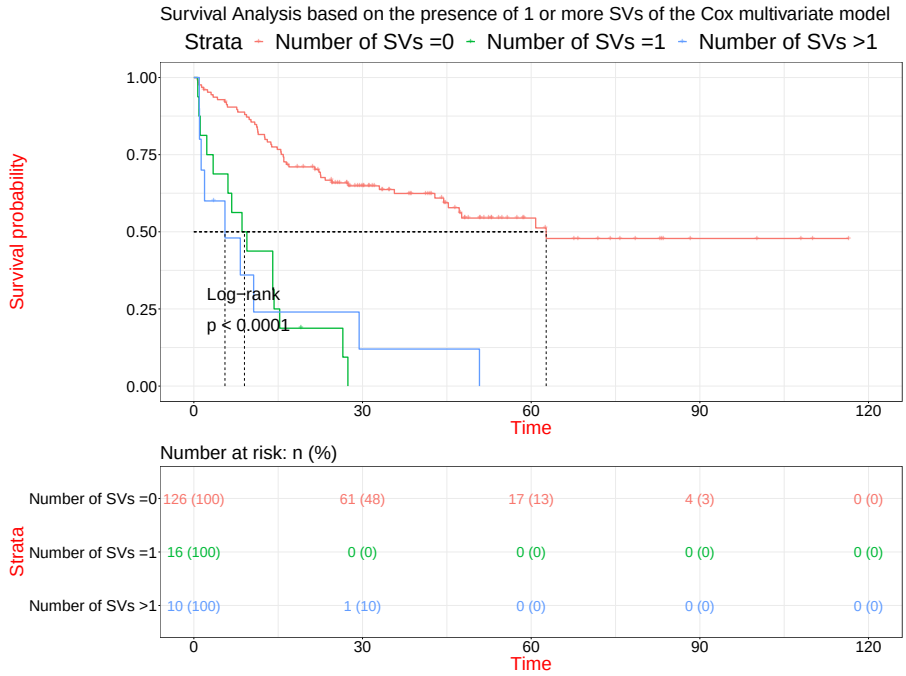

FIGURE 32: `Survival Analysis`. This survival analysis stratified our cohort based on the presence of 1 or more Cox multivariate model's SVs

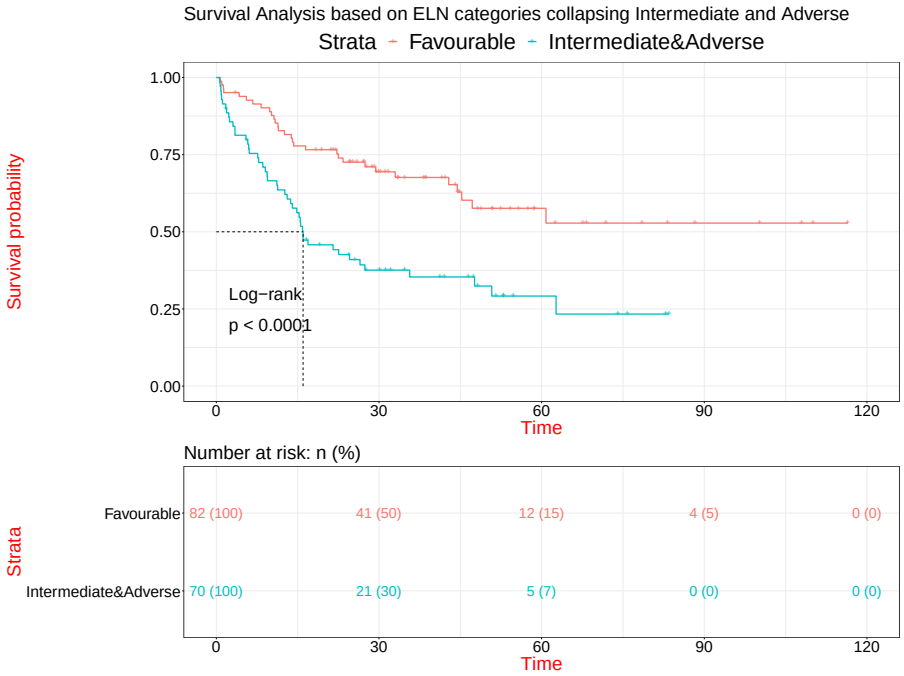

FIGURE 33: `ELN prognostic assessment of whole cohort`. The survival analysis based on ELN gene panel by merging the Intermediate and the Adverse population.

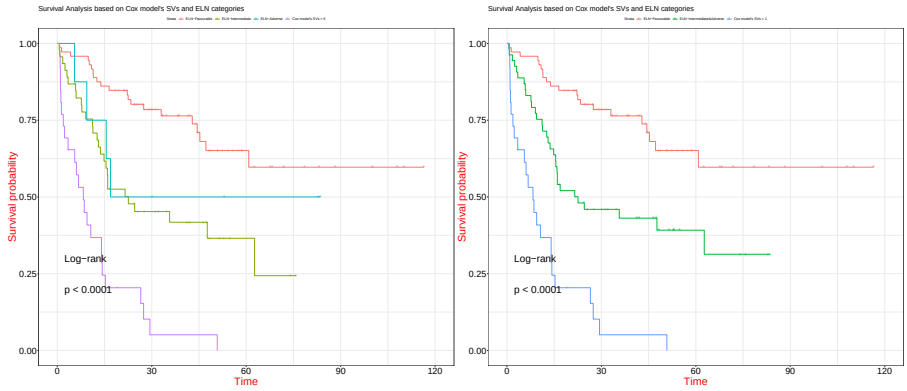

Figure 34: `ELN refinement`. This survival analysis stratified our cohort based on the presence of Cox multivariate model's SVs and the 3 categories identified by ELN (favourable intermediate and adverse), left panel and by merging the intermediate and adverse categories, right panel

Selected Risk Factors Associated With Acute Myeloid Leukemia

Genetic disorders	Down syndrome Klinefelter syndrome Patau syndrome Ataxia telangiectasia Shwachman syndrome Kostman syndrome Neurofibromatosis Fanconi anemia Li-Fraumeni syndrome
Physical and chemical exposures	Benzene Drugs such as pipobroman Pesticides Cigarette smoking Embalming fluids Herbicides
Radiation exposure	Nontherapeutic, therapeutic radiation
Chemotherapy	Alkylating agents Topoisomerase-II inhibitors Anthracyclines Taxanes

TABLE 1: `AML-associated risk factors`. Table is adapted from (Deschler and Lübbert, 2006)

Functional categories		Genes involved	Frequency
Signaling genes		FLT3, KRAS, NRAS, and KIT mutations	59%
Epigenetic homeostasis genes	Chromatin-modifying genes	ASXL1 and EZH2 mutations, MLL fusions	30%
	Methylation-related genes	DNMT3A, TET2, IDH1, and IDH2 mutations	44%
Nucleophosmin gene		NPM1 mutations	27%
Spliceosome-complex genes		SRSF2, SF3B1, U2AF1, and ZRSR2 mutations	14%
Cohesin-complex genes		RAD21, STAG1, STAG2, SMC1A, SMC3 mutations	13%
Myeloid transcription factors		RUNX1, CEBPA, GATA2 mutations	22%
		RUNX1-RUNX1T1, PML-RARA, MYH11-CBFB fusions	18%
Tumor suppressive genes		WT1, TP53 and PHF6 mutations (PTEN and DMM2 deregulations)	16%

TABLE 2: `The table summarizes the functional categories of gene mutations in AML.` Table is taken from *Acute Myeloid Leukemia: From Biology to Clinical Practices Through Development and Pre-Clinical Therapeutics* (Roussel et al., 2020)

Tool	Read qscore[a]	Consensus qscore[a#]	Availability
Albacore	9.2	21.9	Only to ONT customers
BasecRAWller	N/A	N/A	https://basecrawller.lbl.gov/ (seems to be down)
Chiron	7.7	21.4	https://github.com/haotianteng/Chiron
DeepNano	N/A	N/A	https://bitbucket.org/vboza/deepnano/src/master/
FastQC	A quality control tool for high throughput sequence data.		https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Flappie	9.6	22.0	https://github.com/nanoporetech/flappie
Guppy	9.7	23.0	Only to ONT customers
Metrichor	N/A	N/A	Only to ONT customers
Nanocall	N/A	N/A	https://github.com/mateidavid/nanocall
Scrappie	9.3	22.4	https://github.com/nanoporetech/scrappie

TABLE 3: `Base callers developed for nanopore sequencing`. The table summarized the reads qscore and the consensus qscore associated with 10 basecallers specifically developed for nanopore sequencing. This table is taken from from *Bioinformatics of nanopore sequencing* (Makałowski and Shabardina, 2020)

	Outcome Parameters	nkAML cohort
1	Patients,(n)	152
2	OS,median (min - max)	25.38(0.6 -114.46)
3	BMT,n(%)	52.00(34.2%)
4	WBC,median (min - max)	21300(1060 – 435000)
5	CR,n(%)	114(75%)
6	Not_responder,n(%)	38(25%)
7	Relapse,n(%)	48(31.6%)
8	DFS,median (min - max)	17.10(0.3 -115)

TABLE 4: `Table summarized the clinical features of nkAML cohort.`

	ELN molecular parameters	nkAML cohort
1	FLT3-ITD,n(%)	36(23.7%)
2	ITD ratio HIGH,n(%)	24 (15.8%)
3	NPM1,n(%)	83(54.6%)
4	CEBPa,n(%)	12(9.6%)

TABLE 5: `Table summarized the molecular features of nkAML cohort.`

	Covariates	type	beta coef	HR	SE	Wald statistic	Wald pvalue
1	chr2:174236754-174236754	INS	3.60	36.73	0.92	3.91	9.264443e-05
2	chr3:48372314-48372318	INS	2.59	13.38	0.66	3.94	8.089724e-05
3	chr3:78858814-78858815	INS	5.57	262.60	1.08	5.15	2.615987e-07
4	chr4:46202796-46205027	DEL	3.68	39.49	0.88	4.18	2.864080e-05
5	chr4:61289389-61289533	DEL	2.60	13.52	0.59	4.43	9.267471e-06
6	chr11:2026821-2026936	DEL	-6.33	0.00	1.31	-4.84	1.282537e-06
7	chr12:686597-686598	INS	2.63	13.84	0.69	3.79	1.488980e-04
8	chr14:80106289-80115050	DEL	1.61	4.99	0.58	2.77	5.611651e-03
9	chr15:72061719-72061719	INS	3.82	45.62	0.67	5.74	9.421669e-09
10	chr16:87083006-87083065	INS	-2.45	0.09	0.72	-3.41	6.603624e-04
11	chr20:60557285-60557353	INS	-1.69	0.18	0.61	-2.76	5.818547e-03
12	CEBPa	biallelic	-1.42	0.24	0.72	-1.97	4.884707e-02

TABLE 6: `Table summarized the SVs in the Cox multivariate model`

SAMPLE	OS time	OS status	HCT	HCT status	HCT Time	CR	Data CR	Data RIC	Last FU	RIC	DFS time	WBC	AGE
19	6.73	1.00	0.00	1	6.73	1			17/08/2012			67000	61.00
45	0.70	1.00	0.00	1	0.70	1			01/08/2011			147000	70.00
54	10.63	1.00	0.00	1	10.63	0	02/03/2012	17/07/2012	23/09/2012	1.00	4.567	58500	67.00
57	14.03	1.00	1.00	0	4.03	0	06/11/2012	12/03/2013	30/10/2013	1.00	4.2	52800	53.00
59	8.57	1.00	0.00	1	8.56	0	08/08/2012	06/02/2013	06/03/2013	1.00	6.067	21100	36.00
60	0.97	1.00	0.00	1	0.96	1			02/02/2010			125000	63.00
63	50.80	1.00	1.00	0	33.53	0	28/03/2014	16/08/2015	10/04/2018	1.00	16.867	1360	61.00
83	15.27	1.00	1.00	0	6.10	0	22/02/2017	14/12/2017	10/03/2018	1.00	9.833	2260	65.00
90	19.07	0.00	1.00	0	5.60	1			09/09/2019			6910	34.00
92	29.40	1.00	1.00	0	17.50	0	19/09/2016	20/09/2017	16/01/2019	1.00	12.2	1150	52.00
96	1.90	1.00	0.00	NA	1.90	1			19/04/2017			24860	74.00
97	2.30	1.00	0.00	NA	2.30	1			13/02/2017			13600	68.00
98	27.40	1.00	0.00	NA	27.40	1			06/08/2018			12880	62.00
10	5.53	1.00	0.00	1	5.53	1			06/08/2012			62900	57.00
31	14.03	1.00	0.00	1	14.03	0	NA		13/06/2013	0.00	NA	15020	57.00
34	1.13	1.00	0.00	1	1.13	1			10/06/2012			62000	57.00
86	8.27	1.00	0.00	1	8.27	0	27/09/2012	10/04/2013	30/04/2013	1.00	6.5	33800	36.00
140	1.30	1.00	0.00	1	1.30	NA	NA		06/01/2013	0.00	NA	2600	48.00
160	9.43	1.00	1.00	0	7.17	0	06/03/2013		13/10/2013	0.00	7.367	10390	38.00
201	3.43	1.00	0.00	1	3.43	1			01/06/2013			NA	56.00
225	0.90	1.00	0.00	1	0.90	1			21/04/2013			111000	49.00
281	6.07	1.00	0.00	1	6.07	0	14/02/2014	29/05/2014	08/07/2014	1.00	3.467	101000	48.00
315	26.50	1.00	0.00	1	26.50	1			27/05/2016			167000	38.00
386	3.53	0.00	0.00	0	3.53	0	03/09/2014		14/11/2014	0.00	2.4	46450	52.00
392	14.27	1.00	0.00	1	14.27	0	27/08/2014	09/03/2015	06/10/2015	1.00	6.467	57000	44.00
581	1.00	1.00	0.00	1	1.00	1			14/06/2015			2200	53.00

TABLE 7: `Table summarized the Clinical characteristics of high-risk patients.` The column "OS time" described the overall survival censored at latest follow-up or death, "OS status" described the status of the patients at the last follow-up (0:alive, 1:death), "HCT" described the allogenic HCT (0:no, 1:yes), "HCT status" described the status censored at HCT and "HCT time" described the time from diagnosis to allogenic HCT or last follow-up or death, "CR" described the complete remission at the induction therapy (0:no, 1:yes), "Data CR" described the complete remission date, "Data RIC" described the relapse date, "Last fu" described the last data point of follow-up available, "RIC" described the presence of relapse (0:no, 1:yes), "DFS Time" described the disease-free survival, "WBC" described white-blood cell count, "AGE" the age at diagnosis.

SAMPLE	OS time	OS status	HCT	HCT status	HCT Time	CR	Data CR	Data RIC	Last FU	RIC	DFS time	WBC	AGE
3	0.10	0.00	0	0	50.86	0.00	16343.00	17836.00	17836.00	0.00	49.80	9600.00	54.00
8	0.30	0.00	1	0	12.1	0.00	13522.00	13759.00	15711.00	1.00	7.90	166000.00	37.00
9	0.30	0.00	0	0	55.76	0.00	16210.00	17836.00	17836.00	0.00	54.20	3750.00	44.00
10	0.30	1.00	0	1	4.2	1.00			15971.00			160000.00	68.00
11	0.40	0.00	1	0	4.2	0.00	14936.00	16526.00	16526.00	0.00	53.00	54700.00	45.00
12	0.40	1.00	0	1	13.83	0.00	15467.00	15793.00	15809.00	1.00	10.90	48000.00	58.00
13	0.40	1.00	0	1	16.46	0.00	13675.00	13970.00	14123.00	1.00	9.80	84300.00	47.00
14	0.50	1.00	0	1	12.56	0.00	14767.00	15006.00	15083.00	1.00	8.00	47300.00	49.00
15	0.50	0.00	0	0	48.8	0.00	15111.00	16526.00	16526.00	0.00	47.20	116000.00	40.00
16	0.50	1.00	0	1	45.26	0.00	15280.00	16526.00	16587.00	1.00	41.50	21200.00	66.00
17	0.60	1.00	0	1	3.43	1.00			15400.00			1950.00	49.00
18	0.60	1.00	1	0	8.5	0.00	15764.00	15957.00	15957.00	0.00	6.40	66000.00	51.00
20	0.70	1.00	0	1	5.8	0.00	15853.00	15963.00	15972.00	1.00	3.70	47300.00	69.00
21	0.70	0.00	1	0	6.133333333333333	0.00	14376.00	17836.00	17836.00	0.00	115.30	50400.00	40.00
22	0.70	1.00	1	0	20.96666666666667	0.00	13765.00	14162.00	14400.00	1.00	13.20	37800.00	32.00
24	0.80	0.00	1	0	5.866666666666667	0.00	14882.00	17836.00	17836.00	0.00	98.50	14900.00	40.00
25	0.80	0.00	0	0	54.23	0.00	16259.00	17836.00	17836.00	0.00	52.60	6000.00	61.00
26	0.90	0.00	0	0	88.3	0.00	15244.00	17836.00	17836.00	0.00	86.40	2590.00	56.00
27	0.90	1.00	0	1	24.56	0.00	15429.00	15651.00	16038.00	1.00	7.40	81700.00	70.00
28	0.90	1.00	1	1	3.066666666666667	1.00			15934.00			142000.00	66.00
29	1.00	1.00	0	1	42.86	0.00	15664.00	16379.00	16902.00	1.00	23.80	7000.00	62.00
30	1.00	0.00	1	0	13.56666666666667	0.00	16869.00	17109.00	17840.00	1.00	8.00	7500.00	30.00
31	1.00	0.00	0	0	83.23	0.00	15394.00	17836.00	17836.00	0.00	81.40	44000.00	40.00
32	1.10	0.00	0	0	57.46	0.00	14853.00	16526.00	16526.00	0.00	55.80	4420.00	29.00
33	1.10	0.00	0	0	16.7	1.00			15965.00			105000.00	65.00
34	1.10	0.00	1	0	5.833333333333333	0.00	14279.00	16526.00	16526.00	0.00	74.90	191000.00	40.00
35	1.20	0.00	1	0	8.9	0.00	14320.00	17518.00	17518.00	0.00	106.60	105000.00	49.00
36	1.20	0.00	1	0	13.73333333333333	0.00	16559.00	16771.00	17836.00	1.00	7.10	1100.00	36.00
37	1.20	1.00	0	1	10.16	0.00	16394.00	16509.00	16661.00	1.00	3.80	7320.00	54.00
38	1.30	0.00	0	0	47.43	0.00	16466.00	17836.00	17836.00	0.00	45.70	85000.00	49.00
39	1.30	0.00	0	0	62.53	0.00	14277.00	16099.00	16099.00	0.00	60.70	29000.00	39.00
40	1.30	0.00	0	0	58.73	0.00	14510.00	16222.00	16222.00	0.00	57.10	36900.00	46.00
41	1.40	0.00	0	0	68.33	0.00	14529.00	16526.00	16526.00	0.00	66.60	22700.00	35.00
42	1.40	1.00	0	1	44.43	0.00	13434.00	13665.00	14713.00	1.00	7.70	73100.00	65.00
44	1.50	0.00	0	0	78.53	0.00	15196.00	17512.00	17512.00	0.00	77.20	1210.00	61.00
46	1.50	0.00	1	0	9.4	0.00	14540.00	17800.00	17800.00	0.00	108.70	14800.00	58.00
47	1.60	1.00	1	0	2.9	1.00			14909.00			81500.00	39.00
48	1.60	1.00	0	1	1.73	1.00			16178.00			13000.00	44.00
49	1.60	1.00	0	1	16	1.00			16357.00			50900.00	47.00
50	1.70	0.00	0	0	5.6	1.00			14956.00			68800.00	47.00
51	1.70	1.00	1	0	3.56666666666667	0.00	14616.00	14805.00	14836.00	1.00	6.30	1060.00	47.00
52	1.70	1.00	1	0	6.7	1.00			15978.00			90900.00	36.00
53	1.80	0.00	0	0	71.9	0.00	15572.00	17659.00	17659.00	0.00	69.60	55000.00	60.00
55	1.80	0.00	1	0	3.433333333333333	0.00	15497.00	17931.00	17931.00	0.00	81.10	13700.00	45.00
56	1.90	1.00	1	0	8.633333333333333	0.00	14973.00	15161.00	15346.00	1.00	6.30	25900.00	23.00
58	1.90	0.00	1	0	5.833333333333333	0.00	15254.00	17700.00	17700.00	0.00	81.50	3560.00	39.00
61	2.00	0.00	0	0	30.3	0.00	15231.00	16141.00	16099.00	1.00	30.30	7600.00	68.00
62	2.10	1.00	1	0	7.433333333333333	0.00	15246.00	15670.00	15670.00	0.00	14.10	65500.00	54.00
64	2.10	1.00	1	0	27.93333333333333	0.00	16167.00	16792.00	18006.00	1.00	20.80	1600.00	61.00
66	2.20	0.00	0	0	48.2	0.00	16680.00	18064.00	18064.00	0.00	46.10	34000.00	40.00
67	2.20	0.00	1	0	3.266666666666667	0.00	15972.00	17966.00	17966.00	0.00	66.50	435000.00	59.00
68	2.30	1.00	0	1	47.2	0.00	14671.00	15656.00	16044.00	1.00	32.80	25400.00	63.00
69	2.30	0.00	1	0	4.4	0.00	16580.00	18120.00	18120.00	0.00	51.30	17800.00	36.00
70	2.30	1.00	0	1	7.8	1.00			16632.00			63200.00	42.00
71	2.40	1.00	1	0	45.3666666666667	0.00	15854.00	16626.00	17626.00	1.00	25.70	35900.00	55.00
72	2.40	0.00	0	0	44.56	0.00	16063.00	17357.00	17357.00	0.00	43.10	3480.00	35.00
73	2.40	1.00	1	0	4.666666666666667	0.00	16134.00	16270.00	16270.00	0.00	4.50	19300.00	55.00
74	2.50	0.00	1	0	48.4666666666667	0.00	16075.00	16965.00	18143.00	1.00	29.70	18800.00	53.00
75	2.50	0.00	1	0	22.13333333333333	0.00	16818.00	17210.00	18136.00	1.00	13.10	44600.00	56.00
76	2.50	1.00	0	1	12.63	0.00	16929.00	17140.00	17256.00	1.00	7.00	5130.00	60.00
77	2.60	0.00	1	0	6.866666666666667	0.00	16951.00	18022.00	18022.00	0.00	35.70	15100.00	63.00
78	2.60	0.00	0	0	41.63	0.00	16937.00	18130.00	18130.00	0.00	39.80	104000.00	62.00
79	2.60	1.00	1	0	3.666666666666667	0.00	17036.00	17449.00	17462.00	1.00	13.80	34100.00	57.00
80	2.70	0.00	0	0	34.73	0.00	17057.00	18050.00	18050.00	0.00	33.10	58300.00	57.00
81	2.70	1.00	1	0	4.833333333333333	1.00			18140.00			177000.00	43.00
82	2.70	0.00	1	0	6.266666666666667	0.00	17181.00	18092.00	18092.00	0.00	30.40	14700.00	38.00
85	2.80	0.00	0	0	24.63	0.00	17283.00	17980.00	17980.00	0.00	23.20	94000.00	43.00
86	2.90	0.00	1	0	6.4	0.00	17320.00	18087.00	18087.00	0.00	25.60	2370.00	48.00
87	2.90	1.00	1	0	7.466666666666667	0.00	17429.00	17548.00	17714.00	1.00	4.00	183000.00	63.00
89	3.00	0.00	0	0	1.8	0.00	17203.00	17212.00	17212.00	0.00	0.30	10300.00	56.00
91	3.00	0.00	0	0	18.4	0.00	17619.00	18141.00	18141.00	0.00	17.40	21300.00	32.00
95	3.20	0.00	1	NA	15.96666666666667	0.00	17283.00	18074.00	18074.00		26.40	3730.00	61.00
4	48.20	0.00	1	0	5.2	1.00			16811.00	0.00		2700.00	36.60
20	22.60	1.00	1	0	5.4	0.00	15470.00	15712.00	16112.00	1.00	8.10	1780.00	43.10
40	1.30	0.00	0	0	50.96666666666667	0.00	15524.00	15755.00	17006.00	1.00	7.70	58000.00	58.60
45	42.00	0.00	0	0	42.03333333333333	0.00	15516.00		16743.00	0.00	40.90	6220.00	48.00
51	16.00	1.00	0	1	16	0.00	15518.00	15736.00	15968.00	1.00	7.30	38450.00	34.60
57	13.10	1.00	1	0	2.86666666666667	1.00		15782.00	15887.00	1.00		62000.00	49.20
59	53.00	0.00	0	0	53	0.00	15578.00		17086.00	0.00	50.30	84000.00	41.40
64	27.30	0.00	0	0	27.2666666666667	0.00	15551.00		16323.00	0.00	25.70	3600.00	60.90
73	0.70	1.00	0	1	0.733333333333333	1.00			15550.00	0.00		109000.00	56.90
74	47.60	1.00	0	1	47.63333333333333	0.00	15565.00		16961.00	0.00	46.50	18600.00	44.60
81	25.60	0.00	1	0	6.9	0.00	15604.00		16315.00	0.00	23.70	2350.00	49.70
82	41.30	0.00	1	0	4.466666666666667	0.00	15602.00		16799.00	0.00	39.90	16850.00	43.90
84	51.60	0.00	0	0	51.6	0.00	15607.00		17117.00	0.00	50.30	4000.00	32.10
87	21.50	1.00	0	1	21.53333333333333	0.00	15621.00	15707.00	16227.00	1.00	2.90	22500.00	56.10
92	42.20	0.00	0	0	42.23333333333333	0.00	15614.00	16078.00	16850.00	1.00	15.50	2200.00	27.40
111	31.70	0.00	0	0	31.73333333333333	0.00	15663.00	15985.00	16581.00	1.00	10.70	27320.00	58.10
116	38.70	0.00	0	0	38.7	0.00	15679.00		16799.00	0.00	37.30	28500.00	59.30
125	6.00	1.00	0	1	5.966666666666667	0.00	15691.00		15828.00	0.00	4.60	3030.00	51.00
129	31.10	0.00	0	0	31.1	0.00	15701.00		16597.00	0.00	29.90	2450.00	58.10
130	0.60	1.00	0	1	0.6	1.00			15681.00	0.00		2160.00	54.90
157	27.40	1.00	0	1	27.43333333333333	0.00	15757.00	16027.00	16539.00	1.00	9.00	16500.00	45.70
176	46.50	0.00	1	0	4.866666666666667	0.00	15769.00		17122.00	0.00	45.10	1640.00	47.80
177	52.40	0.00	0	0	52.43333333333333	0.00	15761.00		17301.00	0.00	51.30	17210.00	47.00
184	34.80	0.00	1	0	4.4	0.00	15776.00		16780.00	0.00	33.50	7070.00	34.80
211	0.80	1.00	0	1	0.8333333333333333	1.00			15787.00	0.00		113000.00	19.60
212	15.90	1.00	0	1	15.9	1.00		15901.00	16242.00	1.00		17490.00	56.40
213	11.40	1.00	0	1	11.3666666666667	0.00	15812.00	15978.00	16116.00	1.00	5.50	20810.00	57.70
217	35.70	1.00	1	0	9.033333333333333	0.00	15806.00	16694.00	16848.00	1.00	29.60	43500.00	58.20
262	31.20	0.00	1	0		1.00		16267.00	16771.00	1.00		94730.00	46.70
268	30.10	0.00	0	0	30.13333333333333	1.00			16751.00	0.00		3420.00	43.30
284	23.40	1.00	1	0	4.9	0.00	16121.00		16786.00	0.00	22.20	151000.00	53.10
302	27.80	0.00	0	0	27.7666666666667	0.00	16167.00		16966.00	0.00	26.60	6430.00	26.30
316	33.50	0.00	0	0	33.4666666666667	0.00	16202.00		17158.00	0.00	31.90	11600.00	55.10
326	25.20	0.00	0	0	25.2	0.00	16209.00		16931.00	0.00	24.10	13640.00	56.90
375	38.50	0.00	0	0	38.53333333333333	0.00	16295.00		17409.00	0.00	37.10	66700.00	53.20
397	16.90	1.00	0	1	16.9	1.00			16800.00	0.00		136000.00	32.70
409	14.80	1.00	0	1	14.83333333333333	0.00	16345.00	16542.00	16760.00	1.00	6.60	7500.00	19.50
445	27.20	0.00	0	0	27.1666666666667	0.00	16463.00	16519.00	17194.00	1.00	1.90	8000.00	46.10
450	11.20	1.00	0	1	11.16666666666667	1.00			16722.00	0.00		84000.00	45.80
458	25.90	0.00	0	0	25.93333333333333	0.00	16434.00		17176.00	0.00	24.70	1810.00	56.50
482	24.60	0.00	1	0	6.066666666666667	0.00	16477.00		17181.00	0.00	23.50	110090.00	46.90
483	29.10	0.00	0	0	29.13333333333333	0.00	16482.00		17316.00	0.00	27.80	2080.00	44.60
494	24.50	0.00	0	0	24.4666666666667	0.00	16497.00		17195.00	0.00	23.30	88900.00	44.20
516	11.40	1.00	1	0	5.2	0.00	16524.00		16828.00	0.00	10.10	2100.00	51.10
517	29.50	0.00	0	0	29.5	0.00	16520.00		17372.00	0.00	28.40	3850.00	53.10
527	22.10	0.00	0	0	22.1	0.00	16532.00	16911.00	17155.00	1.00	12.60	30590.00	53.80
534	22.50	1.00	0	1	22.5	0.00	16533.00	17148.00	17175.00	1.00	20.50	6900.00	54.20
549	19.50	0.00	0	0	19.4666666666667	0.00	16552.00	16988.00	17109.00	1.00	14.50	3200.00	58.50
551	2.40	1.00	0	1	2.433333333333333	1.00			16599.00	0.00		25000.00	57.90
552	21.60	0.00	1	0	3.733333333333333	0.00	16567.00		17175.00	0.00	20.30	3040.00	49.00
556	1.40	1.00	0	1	1.366666666666667	1.00			16579.00	0.00		178000.00	57.20
558	28.70	0.00	0	0	28.7	0.00	16584.00	17276.00	17407.00	1.00	23.10	5390.00	57.90
582	7.60	1.00	0	1	7.633333333333333	1.00		16660.00	16790.00	1.00		38100.00	54.90
583	21.10	0.00	0	0	21.13333333333333	0.00	16602.00		17207.00	0.00	20.20	3620.00	42.80

TABLE 8: `Table summarized the Clinical characteristics of low-risk patients`. The column "OS time" described the overall survival censored at latest follow-up or death, "OS status" described the status of the patients at the last follow-up (0:alive, 1:death), "HCT" described the allogenic HCT (0:no, 1:yes), "HCT status" described the status censored at HCT and "HCT time" described the time from diagnosis to allogenic HCT or last follow-up or death, "CR" described the complete remission at the induction therapy (0:no, 1:yes), "Data CR" described the complete remission date, "Data RIC" described the relapse date, "Last fu" described the last data point of follow-up available, "RIC" described the presence of relapse (0:no, 1:yes), "DFS Time" described the disease-free survival, "WBC" described white-blood cell count, "AGE" the age at diagnosis.

TABLE 9: Table summarized the Molecular characteristics of high-risk patients. The columns described the presence of the 8 hrSVs (0:not present, 1:present), the number of the hrSVs for each patients ("HR_sum"), the number of the lrSVs for each patients ("LR_sum"), the presence of the CEBPa biallelic mutation (0:not present, 1:present), "NPM1" (0:not present, 1:present) and "FLT3.ITD.ratio" described the presence of the ITD in FLT3 with ratio (0:not present, 1:low allelic ratio, 2:high allelic ratio) and the column ELN described the prognostic stratification based on ELN recommendations (0:favourable,1:intermediate, 2:adverse)

SAMPLE	chr2_INS	chr3_INS	chr3_INS.1	chr4_DEL	chr4_DEL.1	chr11_DEL	chr12_INS	chr14_DEL	chr15_INS	chr16_INS	chr20_INS	CEBPA	NPM1	FLT3.ITD.ratio	ELN	HR.SUM	LR.SUM
19	0	0	0	0	0	0	0	0	0	0	0	1	2	0	1	1	0
45	0	0	0	1	1	1	0	0	0	1	0	0	2	0	1	2	0
54	1	0	0	0	0	1	1	0	0	0	1	1	0	2	2	2	2
57	0	0	0	0	0	0	1	0	0	0	0	0	2	2	1	1	0
59	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	1
60	1	0	0	0	1	0	0	1	0	0	0	0	2	2	1	1	0
63	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	3	1
83	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
92	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1	1	1
96	0	1	0	0	0	0	0	0	0	0	0	1	0	0	3	3	0
97	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1	1	0
98	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	1
10	0	0	0	1	0	1	0	1	0	0	0	1	0	1	1	1	1
31	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	1	0
34	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	4	0
86	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	1	1
140	0	0	0	0	1	0	1	0	0	0	0	1	0	0	2	2	0
160	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	4	0
201	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0
225	0	1	0	0	0	1	0	0	0	0	1	0	0	1	1	1	0
281	0	1	0	0	0	0	0	0	0	0	0	1	2	1	1	1	0
515	0	1	0	0	0	0	0	0	0	0	0	0	2	2	1	1	0
3486	0	1	0	0	0	0	0	0	0	0	0	1	0	0	2	2	1
3492	0	0	0	0	0	0	1	0	0	0	0	1	0	0	3	3	0
581	1	0	0	0	0	0	0	1	0	1	1	1	0	0	2	2	2

Bibliography

SEER Cancer Statistics Review, 1975-2018. URL https://seer.cancer.gov/csr/1975_2018/index.html.

1000 Genomes Project, Ryan E. Mills, Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R. Keira Cheetham, Asif Chinwalla, Donald F. Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M. Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M. Kidd, Miriam K. Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y. K. Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E. Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P. Stromberg, Adrian M. Stütz, Alexander Eckehart Urban, Jerilyn A. Walker, Jiantao Wu, Yujun Zhang, Zhengdong D. Zhang, Mark A. Batzer, Li Ding, Gabor T. Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E. Eichler, Mark B. Gerstein, Matthew E. Hurles, Charles Lee, Steven A. McCarroll, and Jan O. Korbel. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09708. URL http://www.nature.com/articles/nature09708.

Ahmad I. Antar, Zaher K. Otrock, Elias Jabbour, Mohamad Mohty, and Ali Bazarbachi. FLT3 inhibitors in acute myeloid leukemia: ten frequently asked questions. *Leukemia*, 34(3):682–696, March 2020. ISSN 0887-6924, 1476-5551. doi: 10.1038/s41375-019-0694-3. URL http://www.nature.com/articles/s41375-019-0694-3.

Daniel A. Arber, Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J. Borowitz, Michelle M. Le Beau, Clara D. Bloomfield, Mario Cazzola, and James W. Vardiman. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405, May 2016. ISSN 0006-4971. doi: 10.1182/blood-2016-03-643544. URL `https://doi.org/10.1182/blood-2016-03-643544`.

Maurice R. Atkinson, Murray P. Deutscher, Arthur Kornberg, Alan F. Russell, and J. G. Moffatt. Enzymatic Synthesis of Deoxyribonucleic Acid. XXXIV. Termination of Chain Growth by a 2′,3′-Dideoxyribonucleotide. *Biochemistry*, 1969. ISSN 15204995. doi: 10.1021/bi00840a037.

Marina Barba, Henryk Czosnek, and Ahmed Hadidi. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 2013. ISSN 19994915. doi: 10.3390/v6010106.

Marina Barba, Henryk Czosnek, and Ahmed Hadidi. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6(1):106–136, January 2014. ISSN 1999-4915. doi: 10.3390/v6010106.

John Beaulaurier, Eric E. Schadt, and Gang Fang. Deciphering bacterial epigenomes using modern sequencing technologies, 2019. ISSN 14710064.

Amir Behdad, Helmut C. Weigelin, Kojo S.J. Elenitoba-Johnson, and Bryan L. Betz. A clinical grade sequencing-based assay for CEBPA mutation testing: Report of a large series of myeloid neoplasms. *Journal of Molecular Diagnostics*, 2015. ISSN 19437811. doi: 10.1016/j.jmoldx.2014.09.007.

J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *British Journal of Haematology*, 33 (4):451–458, August 1976. ISSN 0007-1048. doi: 10.1111/j.1365-2141.1976.tb03563.x.

Marianne Bienz, Madleina Ludwig, Beatrice U. Mueller, Elisabeth Oppliger Leibundgut, Daniel Ratschiller, Max Solenthaler, Martin F. Fey, and Thomas Pabst. Risk Assessment in Patients with Acute Myeloid Leukemia and a Normal Karyotype. *Clinical Cancer Research*, 11(4):1416–1424, February 2005. ISSN 1078-0432, 1557-3265. doi: 10.1158/1078-0432.CCR-04-1552. URL `http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-04-1552`.

Tilmann Bochtler, Friedrich Stölzel, Christoph E. Heilig, Christina Kunz, Brigitte Mohr, Anna Jauch, Johannes W.G. Janssen, Michael Kramer, Axel Benner, Martin Bornhäuser, Anthony D. Ho, Gerhard Ehninger, Markus Schaich, and Alwin Krämer. Clonal Heterogeneity As Detected by Metaphase Karyotyping Is an Indicator of Poor Prognosis in Acute Myeloid Leukemia. *Journal of Clinical Oncology*, 31(31):3898–3905, November 2013. ISSN 0732-183X, 1527-7755. doi: 10.1200/JCO.2013.50.7921. URL `http://ascopubs.org/doi/10.1200/JCO.2013.50.7921`.

Prajwal Boddu, Hagop M. Kantarjian, Guillermo Garcia-Manero, Farhad Ravandi, Srdan Verstovsek, Elias Jabbour, Gautam Borthakur, Marina Konopleva, Kapil N. Bhalla, Naval Daver, Courtney D. DiNardo, Christopher B. Benton, Koichi Takahashi, Zeev Estrov, Sherry R. Pierce, Michael Andreeff, Jorge E. Cortes, and Tapan M. Kadia. Treated secondary acute myeloid leukemia: a distinct high-risk subset of AML with adverse prognosis. *Blood Advances*, 1(17):1312–1323, July 2017. ISSN 2473-9529. doi: 10.1182/bloodadvances.2017008227.

Lars Bullinger, Konstanze Döhner, Eric Bair, Stefan Fröhling, Richard F. Schlenk, Robert Tibshirani, Hartmut Döhner, and Jonathan R. Pollack. Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. *New England Journal of Medicine*, 350 (16):1605–1616, April 2004. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa031046. URL `http://www.nejm.org/doi/abs/10.1056/NEJMoa031046`.

Lars Bullinger, Konstanze Dohner, Raphael Kranz, Frank G. Rucker, Stefan Frohling, Richard F. Schlenk, Jonathan R.

Pollack, and Hartmut Dohner. Characterization of NPM1-Mutated/FLT3 ITD-Negative Acute Myeloid Leukemia with Normal Karyotype by Gene Expression Profiling. *Blood*, 108(11):155–155, November 2006. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood.V108.11.155.155. URL `https://ashpublications.org/blood/article/108/11/155/126612/Characterization-of-NPM1MutatedFLT3-ITDNegative`.

Lars Bullinger, Konstanze Döhner, and Hartmut Döhner. Genomics of Acute Myeloid Leukemia Diagnosis and Pathways. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 35(9):934–946, March 2017. ISSN 1527-7755. doi: 10.1200/JCO.2016.71.2208.

Alan K. Burnett, Nigel H. Russell, Robert K. Hills, Jonathan Kell, Sylvie Freeman, Lars Kjeldsen, Ann E. Hunter, John Yin, Charles F. Craddock, Inge Hoegh Dufva, Keith Wheatley, and Donald Milligan. Addition of Gemtuzumab Ozogamicin to Induction Chemotherapy Improves Survival in Older Patients With Acute Myeloid Leukemia. *Journal of Clinical Oncology*, 30(32):3924–3931, November 2012. ISSN 0732-183X, 1527-7755. doi: 10.1200/JCO.2012.42.2964. URL `http://ascopubs.org/doi/10.1200/JCO.2012.42.2964`.

Cancer Genome Atlas Research Network, Timothy J. Ley, Christopher Miller, Li Ding, Benjamin J. Raphael, Andrew J. Mungall, A. Gordon Robertson, Katherine Hoadley, Timothy J. Triche, Peter W. Laird, Jack D. Baty, Lucinda L. Fulton, Robert Fulton, Sharon E. Heath, Joelle Kalicki-Veizer, Cyriac Kandoth, Jeffery M. Klco, Daniel C. Koboldt, Krishna-Latha Kanchi, Shashikant Kulkarni, Tamara L. Lamprecht, David E. Larson, Ling Lin, Charles Lu, Michael D. McLellan, Joshua F. McMichael, Jacqueline Payton, Heather Schmidt, David H. Spencer, Michael H. Tomasson, John W. Wallis, Lukas D. Wartman, Mark A. Watson, John Welch, Michael C. Wendl, Adrian Ally, Miruna Balasundaram, Inanc Birol, Yaron Butterfield, Readman Chiu, Andy Chu, Eric Chuah, Hye-Jung Chun, Richard Corbett, Noreen Dhalla, Ranabir Guin, An He, Carrie Hirst, Martin Hirst, Robert A. Holt, Steven Jones, Aly Karsan, Darlene Lee, Haiyan I. Li, Marco A. Marra, Michael Mayo, Richard A. Moore, Karen Mungall, Jeremy Parker, Erin

Pleasance, Patrick Plettner, Jacquie Schein, Dominik Stoll, Lucas Swanson, Angela Tam, Nina Thiessen, Richard Varhol, Natasja Wye, Yongjun Zhao, Stacey Gabriel, Gad Getz, Carrie Sougnez, Lihua Zou, Mark D. M. Leiserson, Fabio Vandin, Hsin-Ta Wu, Frederick Applebaum, Stephen B. Baylin, Rehan Akbani, Bradley M. Broom, Ken Chen, Thomas C. Motter, Khanh Nguyen, John N. Weinstein, Nianziang Zhang, Martin L. Ferguson, Christopher Adams, Aaron Black, Jay Bowen, Julie Gastier-Foster, Thomas Grossman, Tara Lichtenberg, Lisa Wise, Tanja Davidsen, John A. Demchok, Kenna R. Mills Shaw, Margi Sheth, Heidi J. Sofia, Liming Yang, James R. Downing, and Greg Eley. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England Journal of Medicine*, 368(22):2059–2074, May 2013. ISSN 1533-4406. doi: 10.1056/NEJMoa1301689.

Sylvie Castaigne, Cécile Pautas, Christine Terré, Emmanuel Raffoux, Dominique Bordessoule, Jean-Noel Bastie, Ollivier Legrand, Xavier Thomas, Pascal Turlure, Oumedaly Reman, Thierry de Revel, Lauris Gastaud, Noémie de Gunzburg, Nathalie Contentin, Estelle Henry, Jean-Pierre Marolleau, Ahmad Aljijakli, Philippe Rousselot, Pierre Fenaux, Claude Preudhomme, Sylvie Chevret, and Hervé Dombret. Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): a randomised, open-label, phase 3 study. *The Lancet*, 379 (9825):1508–1516, April 2012. ISSN 01406736. doi: 10.1016/S0140-6736(12)60485-1. URL `https://linkinghub.elsevier.com/retrieve/pii/S0140673612604851`.

George M. Church. Genomes for All. *Scientific American*, 294(1):46–54, January 2006. ISSN 0036-8733. doi: 10.1038/scientificamerican0106-46. URL `https://www.scientificamerican.com/article/genomes-for-all`.

Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 02 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008. URL `https://doi.org/10.1093/gigascience/giab008`. giab008.

I. De Kouchkovsky and M. Abdul-Hay. 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood Cancer Journal*, 6(7):e441, July 2016. ISSN 2044-5385. doi: 10.1038/bcj.2016.50.

Maria Stefania De Propris, Sara Raponi, Daniela Diverio, Maria Laura Milani, Giovanna Meloni, Brunangelo Falini, Robin Foà, and Anna Guarini. High CD33 expression levels in acute myeloid leukemia cells carrying the nucleophosmin (NPM1) mutation. *Haematologica*, 96(10):1548–1551, October 2011. ISSN 1592-8721. doi: 10.3324/haematol.2011.043786.

Barbara Deschler and Michael Lübbert. Acute myeloid leukemia: epidemiology and etiology. *Cancer*, 107(9):2099–2107, November 2006. ISSN 0008-543X. doi: 10.1002/cncr.22233.

Dimitri Desvillechabrol, Christiane Bouchier, Sean Kennedy, and Thomas Cokelaer. Sequana coverage: detection and characterization of genomic variations using running median and mixture models. *GigaScience*, 7(12), December 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy110.

Courtney D. DiNardo and Jorge E. Cortes. Mutations in AML: prognostic and therapeutic implications. *Hematology. American Society of Hematology. Education Program*, 2016(1):348–355, December 2016. ISSN 1520-4383. doi: 10.1182/asheducation-2016.1.348.

Courtney D. DiNardo, Anthony S. Stein, Eytan M. Stein, Amir T. Fathi, Olga Frankfurt, Andre C. Schuh, Hartmut Döhner, Giovanni Martinelli, Prapti A. Patel, Emmanuel Raffoux, Peter Tan, Amer M. Zeidan, Stéphane de Botton, Hagop M. Kantarjian, Richard M. Stone, Mark G. Frattini, Frederik Lersch, Jing Gong, Diego A. Gianolio, Vickie Zhang, Aleksandra Franovic, Bin Fan, Meredith Goldwasser, Scott Daigle, Sung Choe, Bin Wu, Thomas Winkler, and Paresh Vyas. Mutant Isocitrate Dehydrogenase 1 Inhibitor Ivosidenib in Combination With Azacitidine for Newly Diagnosed Acute Myeloid Leukemia. *Journal of Clinical Oncology*, 39(1):57–65, January 2021. ISSN 0732-183X, 1527-7755. doi: 10.1200/JCO.20.01632. URL `https://ascopubs.org/doi/10.1200/JCO.20.01632`.

Li Ding, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, Julie K. Ritchey, Margaret A. Young, Tamara Lamprecht, Michael D. McLellan, Joshua F. McMichael, John W. Wallis, Charles Lu, Dong Shen, Christopher C. Harris, David J. Dooling, Robert S. Fulton, Lucinda L. Fulton, Ken Chen, Heather Schmidt, Joelle Kalicki-Veizer, Vincent J. Magrini, Lisa Cook, Sean D. McGrath, Tammi L. Vickery, Michael C. Wendl, Sharon Heath, Mark A. Watson, Daniel C. Link, Michael H. Tomasson, William D. Shannon, Jacqueline E. Payton, Shashikant Kulkarni, Peter Westervelt, Matthew J. Walter, Timothy A. Graubert, Elaine R. Mardis, Richard K. Wilson, and John F. DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, January 2012. ISSN 1476-4687. doi: 10.1038/nature10738.

Radoje Drmanac, Snezana Drmanac, Gloria Chui, Robert Diaz, Aaron Hou, Hui Jin, Paul Jin, Sunhee Kwon, Scott Lacy, Bill Moeur, Jay Shafto, Don Swanson, Tatjana Ukrainczyk, Chongjun Xu, and Deane Little. Sequencing by Hybridization (SBH): Advantages, Achievements, and Opportunities. In T. Scheper, W. Babel, H. W. Blanch, I. Endo, S. O. Enfors, A. Fiechter, M. Hoare, B. Mattiasson, H. Sahm, K. Schügerl, G. Stephanopoulos, U. von Stockar, G. T. Tsao Director, J. Villadsen, C. Wandrey, Jörg Hoheisel, A. Brazma, K. Büssow, C. R. Cantor, F. C. Christians, G. Chui, R. Diaz, R. Drmanac, S. Drmanac, H. Eickhoff, K. Fellenberg, S. Hannenhalli, J. Hoheisel, A. Hou, E. Hubbell, H. Jin, P. Jin, C. Jurinke, Z. Konthur, H. Köster, S. Kwon, S. Lacy, H. Lehrach, R. Lipshutz, D. Little, A. Lueking, G. H. McGall, B. Moeur, E. Nordhoff, L. Nyarsik, P. A. Pevzner, A. Robinson, U. Sarkans, J. Shafto, M. Sohail, E. M. Southern, D. Swanson, T. Ukrainczyk, D. van den Boom, J. Vilo, M. Vingron, G. Walter, and C. Xu, editors, *Chip Technology*, volume 77, pages 75–101. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. ISBN 9783540432159 9783540457138. doi: 10.1007/3-540-45713-5_5. URL `http://link.springer.com/10.1007/3-540-45713-5_5`.

Radoje Drmanac, Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B. Nilsen, George Yeung,

Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P. Pant, Jonathan Baccash, Adam P. Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C. Ebert, Coleen R. Hacker, Robert Hartlage, Brian Huser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E. McBride, Matt Morenzoni, Robert E. Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A. Peters, Joe Peterson, Charit L. Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M. Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W. Shannon, Conrad G. Sheppy, Michel Sun, Joseph V. Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R. Oliphant, William C. Banyai, Bruce Martin, Dennis G. Ballinger, George M. Church, and Clifford A. Reid. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 2010. ISSN 00368075. doi: 10.1126/science.1181498.

Hartmut Döhner, Elihu H. Estey, Sergio Amadori, Frederick R. Appelbaum, Thomas Büchner, Alan K. Burnett, Hervé Dombret, Pierre Fenaux, David Grimwade, Richard A. Larson, Francesco Lo-Coco, Tomoki Naoe, Dietger Niederwieser, Gert J. Ossenkoppele, Miguel A. Sanz, Jorge Sierra, Martin S. Tallman, Bob Löwenberg, and Clara D. Bloomfield. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*, 115(3):453–474, January 2010. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2009-07-235358. URL `https://ashpublications.org/blood/article/115/3/453/27145/Diagnosis-and-management-of-acute-myeloid-leukemia`.

Hartmut Döhner, Daniel J. Weisdorf, and Clara D. Bloomfield. Acute Myeloid Leukemia. *New England Journal of Medicine*, 373(12):1136–1152, September 2015. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMra1406184. URL `http://www.nejm.org/doi/10.1056/NEJMra1406184`.

Hartmut Döhner, Elihu Estey, David Grimwade, Sergio Amadori, Frederick R. Appelbaum, Thomas Büchner, Hervé Dombret,

Benjamin L. Ebert, Pierre Fenaux, Richard A. Larson, Ross L. Levine, Francesco Lo-Coco, Tomoki Naoe, Dietger Niederwieser, Gert J. Ossenkoppele, Miguel Sanz, Jorge Sierra, Martin S. Tallman, Hwei-Fang Tien, Andrew H. Wei, Bob Löwenberg, and Clara D. Bloomfield. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129(4):424–447, January 2017. ISSN 1528-0020. doi: 10.1182/blood-2016-08-733196.

J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, January 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1162986. URL `https://www.sciencemag.org/lookup/doi/10.1126/science.1162986`.

Elihu H. Estey. Acute myeloid leukemia: 2019 update on risk-stratification and management. *American Journal of Hematology*, 93(10):1267–1291, October 2018. ISSN 03618609. doi: 10.1002/ajh.25214. URL `https://onlinelibrary.wiley.com/doi/10.1002/ajh.25214`.

Brent Ewing and Phil Green. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. *Genome Research*, 8(3):186–194, March 1998. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.8.3.186. URL `http://genome.cshlp.org/lookup/doi/10.1101/gr.8.3.186`.

Tobias Fehlmann, Stefanie Reinheimer, Chunyu Geng, Xiaoshan Su, Snezana Drmanac, Andrei Alexeev, Chunyan Zhang, Christina Backes, Nicole Ludwig, Martin Hart, Dan An, Zhenzhen Zhu, Chongjun Xu, Ao Chen, Ming Ni, Jian Liu, Yuxiang Li, Matthew Poulter, Yongping Li, Cord Stähler, Radoje

Drmanac, Xun Xu, Eckart Meese, and Andreas Keller. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clinical Epigenetics*, 2016. ISSN 18687083. doi: 10.1186/s13148-016-0287-1.

Pau Montesinos Fernandez, Christian Recher, Vadim Doronin, Rodrigo T. Calado, Jun Ho Jang, Yasushi Miyazaki, Jianxiang Wang, Diego A Gianolio, Scott R. Daigle, Thomas Winkler, Vickie Zhang, and Peter Paschka. AGILE: A Phase 3, Multicenter, Double-Blind, Randomized, Placebo-Controlled Study of Ivosidenib in Combination with Azacitidine in Adult Patients with Previously Untreated Acute Myeloid Leukemia with an IDH1 Mutation. *Blood*, 134(Supplement_1): 2593–2593, November 2019. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2019-123045. URL `https://ashpublications.org/blood/article/134/Supplement_1/2593/423323/AGILE-A-Phase-3-Multicenter-DoubleBlind-Randomized`.

R. Fleischmann, M. Adams, O White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, July 1995. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.7542800. URL `https://www.sciencemag.org/lookup/doi/10.1126/science.7542800`.

Véronique Geoffroy, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, and Jean Muller. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, 34(20):3572–3574, October 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty304. URL `https://academic.oup.com/bioinformatics/article/34/20/3572/4970516`.

Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: Ten years of next-generation sequencing technologies, 2016. ISSN 14710064.

David Gordon, Chris Abajian, and Phil Green. *Consed:* A Graphical Tool for Sequence Finishing. *Genome Research*, 8 (3):195–202, March 1998. ISSN 1088-9051, 1549-5469. doi:

10.1101/gr.8.3.195. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.8.3.195.

Matthew Hayes. Computational Analysis of Structural Variation in Cancer Genomes. In Alexander Krasnitz, editor, *Cancer Bioinformatics*, volume 1878, pages 65–83. Springer New York, New York, NY, 2019. ISBN 9781493988662 9781493988686. doi: 10.1007/978-1-4939-8868-6_3. URL http://link.springer.com/10.1007/978-1-4939-8868-6_3.

Steve S. Ho, Alexander E. Urban, and Ryan E. Mills. Structural variation in the sequencing era. *Nature Reviews Genetics*, 21 (3):171–189, March 2020. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-019-0180-9. URL http://www.nature.com/articles/s41576-019-0180-9.

Jerzy Holowiecki, Sebastian Grosicki, Sebastian Giebel, Tadeusz Robak, Slawomira Kyrcz-Krzemien, Kazimierz Kuliczkowski, Aleksander B. Skotnicki, Andrzej Hellmann, Kazimierz Sulek, Anna Dmoszynska, Janusz Kloczko, Wieslaw W. Jedrzejczak, Barbara Zdziarska, Krzysztof Warzocha, Krystyna Zawilska, Mieczysław Komarnicki, Marek Kielbinski, Beata Piatkowska-Jakubas, Agnieszka Wierzbowska, Malgorzata Wach, and Olga Haus. Cladribine, But Not Fludarabine, Added to Daunorubicin and Cytarabine During Induction Prolongs Survival of Patients With Acute Myeloid Leukemia: A Multicenter, Randomized Phase III Study. *Journal of Clinical Oncology*, 30(20):2441–2448, July 2012. ISSN 0732-183X, 1527-7755. doi: 10.1200/JCO.2011.37.1286. URL http://ascopubs.org/doi/10.1200/JCO.2011.37.1286.

Mariam Ibáñez, Esperanza Such, Esther Onecha, Inés Gómez-Seguí, Alessandro Liquori, Jorge Sellés, David Hervás-Marín, Eva Barragán, Rosa Ayala, Marta LLop, María López-Pavía, Inmaculada Rapado, Alex Neef, Alejandra Sanjuan-Pla, Claudia Sargas, Elisa Gonzalez-Romero, Mireia Boluda-Navarro, Rafa Andreu, Leonor Senent, Pau Montesinos, Joaquín Martínez-López, Miguel Angel Sanz, Guillermo Sanz, and José Cervera. Analysis of SNP Array Abnormalities in Patients with DE NOVO Acute Myeloid Leukemia with Normal Karyotype. *Scientific Reports*, 10(1):5904, April 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61589-9.

Camilla L.C. Ip, Matthew Loose, John R. Tyson, Mariateresa de Cesare, Bonnie L. Brown, Miten Jain, Richard M. Leggett, David A. Eccles, Vadim Zalunin, John M. Urban, Paolo Piazza, Rory J. Bowden, Benedict Paten, Solomon Mwaigwisya, Elizabeth M. Batty, Jared T. Simpson, Terrance P. Snutch, Ewan Birney, David Buck, Sara Goodwin, Hans J. Jansen, Justin O'Grady, and Hugh E. Olsen. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, 2015. ISSN 1759796X. doi: 10.12688/f1000research.7201.1.

Bum Kim Jae, Gregory J. Porreca, Lei Song, Steven C. Greenway, Joshua M. Gorham, George M. Church, Christine E. Seidman, and J. G. Seidman. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, 2007. ISSN 00368075. doi: 10.1126/science.1137325.

Daniel C. Jeffares, Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J. Sedlazeck. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, 8(1):14061, April 2017. ISSN 2041-1723. doi: 10.1038/ncomms14061. URL `http://www.nature.com/articles/ncomms14061`.

Tao Jiang, Yongzhuang Liu, Yue Jiang, Junyi Li, Yan Gao, Zhe Cui, Yadong Liu, Bo Liu, and Yadong Wang. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, 21(1):189, December 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02107-y. URL `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02107-y`.

A Kassambara, M Kosinski, P Biecek, and others. survminer: Drawing survival curves using'ggplot2'. *R package version 0.3*, 1, 2017.

W Kern, T Haferlach, S Schnittger, Wd Ludwig, W Hiddemann, and C Schoch. Karyotype instability between diagnosis and relapse in 117 patients with acute myeloid leukemia: implications for resistance against therapy. *Leukemia*, 16(10):2084–2091,

October 2002. ISSN 0887-6924, 1476-5551. doi: 10.1038/sj.leu.2402654. URL http://www.nature.com/articles/2402654.

Mudassir Khan, Misbahud Din, Zerbab Naeem, Zahra Sajid, Dilawar Khan, Muhammad Amjad, Aurang Zeb, Faheem Anwar, Mehran Akhtar, and Sana Noreen. Insights into acute myeloid leukemia: Critical analysis on its wide aspects. Volume 3:1–9, 12 2020. doi: 10.34091/AJLS.3.2.1.

Heidi D. Klepin. Geriatric perspective: how to assess fitness for chemotherapy in acute myeloid leukemia. *Hematology*, 2014(1):8–13, December 2014. ISSN 1520-4391, 1520-4383. doi: 10.1182/asheducation-2014.1.8. URL https://ashpublications.org/hematology/article/2014/1/8/20519/Geriatric-perspective-how-to-assess-fitness-for.

Heidi D. Klepin, Ann M. Geiger, Janet A. Tooze, Stephen B. Kritchevsky, Jeff D. Williamson, Timothy S. Pardee, Leslie R. Ellis, and Bayard L. Powell. Geriatric assessment predicts survival for older adults receiving induction chemotherapy for acute myelogenous leukemia. *Blood*, 121(21):4287–4294, May 2013. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2012-12-471680. URL https://ashpublications.org/blood/article/121/21/4287/31319/Geriatric-assessment-predicts-survival-for-older.

Utz Krug, Christoph Röllig, Anja Koschmieder, Achim Heinecke, Maria Cristina Sauerland, Markus Schaich, Christian Thiede, Michael Kramer, Jan Braess, Karsten Spiekermann, Torsten Haferlach, Claudia Haferlach, Steffen Koschmieder, Christian Rohde, Hubert Serve, Bernhard Wörmann, Wolfgang Hiddemann, Gerhard Ehninger, Wolfgang E Berdel, Thomas Büchner, and Carsten Müller-Tidow. Complete remission and early death after intensive chemotherapy in patients aged 60 years or older with acute myeloid leukaemia: a web-based application for prediction of outcomes. *The Lancet*, 376(9757):2000–2008, December 2010. ISSN 01406736. doi: 10.1016/S0140-6736(10)62105-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673610621058.

Jeffrey E. Lancet, Jorge E. Cortes, Donna E. Hogge, Martin S. Tallman, Tibor J. Kovacsovics, Lloyd E. Damon, Rami Komrokji,

Scott R. Solomon, Jonathan E. Kolitz, Maureen Cooper, Andrew M. Yeager, Arthur C. Louie, and Eric J. Feldman. Phase 2 trial of CPX-351, a fixed 5:1 molar ratio of cytarabine/daunorubicin, vs cytarabine/daunorubicin in older adults with untreated AML. *Blood*, 123(21):3239–3246, May 2014. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2013-12-540971. URL `https://ashpublications.org/blood/article/123/21/3239/32772/Phase-2-trial-of-CPX351-a-fixed-51-molar-ratio-of`.

Jeffrey E. Lancet, Geoffrey L. Uy, Jorge E. Cortes, Laura F. Newell, Tara L. Lin, Ellen K. Ritchie, Robert K. Stuart, Stephen A. Strickland, Donna Hogge, Scott R. Solomon, Richard M. Stone, Dale L. Bixby, Jonathan E. Kolitz, Gary J. Schiller, Matthew J. Wieduwilt, Daniel H. Ryan, Antje Hoering, Kamalika Banerjee, Michael Chiarella, Arthur C. Louie, and Bruno C. Medeiros. CPX-351 (cytarabine and daunorubicin) Liposome for Injection Versus Conventional Cytarabine Plus Daunorubicin in Older Patients With Newly Diagnosed Secondary Acute Myeloid Leukemia. *Journal of Clinical Oncology*, 36(26):2684–2692, September 2018. ISSN 0732-183X, 1527-7755. doi: 10.1200/JCO.2017.77.6112. URL `https://ascopubs.org/doi/10.1200/JCO.2017.77.6112`.

Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier,

John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Mei Lee Hong, Joann Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa De La Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G.R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg

Schuler, Jörg Schultz, Guy Slater, Arian F.A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw Pyng Yang, Ru Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 2001. ISSN 00280836. doi: 10.1038/35057062.

John H. Leamon, William L. Lee, Karrie R. Tartaro, Janna R. Lanza, Gary J. Sarkis, Alex D. DeWinter, Jan Berka, and Kenton L. Lohman. A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, 2003. ISSN 01730835. doi: 10.1002/elps.200305646.

Fajun Li, Chunpeng Fu, and Qunfeng Li. A Simple Genome Walking Strategy to Isolate Unknown Genomic Regions Using Long Primer and RAPD Primer. *Iranian Journal of Biotechnology*, 17 (2):e2183, April 2019. ISSN 1728-3043. doi: 10.21859/ijb.2183.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp352. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352`.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty191. URL `https://academic.oup.com/bioinformatics/article/34/18/3094/4994778`.

Ruediger Liersch, Carsten Müller-Tidow, Wolfgang E. Berdel, and Utz Krug. Prognostic factors for acute myeloid leukaemia in adults - biological significance and clinical use. *British Journal of Haematology*, 165(1):17–38, April 2014. ISSN 00071048.

doi: 10.1111/bjh.12750. URL https://onlinelibrary.wiley.com/doi/10.1111/bjh.12750.

Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, 2012:251364, 2012. ISSN 1110-7251. doi: 10.1155/2012/251364.

Xiaoyan Liu and Yuping Gong. Isocitrate dehydrogenase inhibitors in acute myeloid leukemia. *Biomarker Research*, 7:22, 2019. ISSN 2050-7771. doi: 10.1186/s40364-019-0173-z.

Hengyun Lu, Francesca Giordano, and Zemin Ning. Oxford Nanopore MinION Sequencing and Genome Assembly, 2016. ISSN 22103244.

Jeffrey R. MacDonald, Robert Ziman, Ryan K. C. Yuen, Lars Feuk, and Stephen W. Scherer. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(Database issue):D986–992, January 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt958.

Alberto Magi, Davide Bolognini, Niccoló Bartalucci, Alessandra Mingrino, Roberto Semeraro, Luna Giovannini, Stefania Bonifacio, Daniela Parrini, Elisabetta Pelo, Francesco Mannelli, Paola Guglielmelli, and Alessandro Maria Vannucchi. Nano-GLADIATOR: real-time detection of copy number alterations from nanopore sequencing data. *Bioinformatics*, 35 (21):4213–4221, November 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz241. URL https://academic.oup.com/bioinformatics/article/35/21/4213/5428178.

Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. Structural variant calling: the long and the short of it. *Genome Biology*, 20(1):246, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1828-7. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1828-7.

Wojciech Makałowski and Victoria Shabardina. Bioinformatics of nanopore sequencing. *Journal of Human Genetics*, 65(1):

61–67, January 2020. ISSN 1434-5161, 1435-232X. doi: 10.1038/s10038-019-0659-4. URL http://www.nature.com/articles/s10038-019-0659-4.

Tuomo Mantere, Simone Kersten, and Alexander Hoischen. Long-read sequencing emerging in medical genetics, 2019. ISSN 16648021.

Elaine R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, September 2008. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev.genom.9.081307.164359. URL http://www.annualreviews.org/doi/10.1146/annurev.genom.9.081307.164359.

W. Richard McCombie, John D. McPherson, and Elaine R. Mardis. Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, 9(11):a036798, November 2019. ISSN 2157-1422. doi: 10.1101/cshperspect.a036798. URL http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a036798.

Jason D. Merker, Anton Valouev, and Jason Gotlib. Next-generation sequencing in hematologic malignancies: what will be the dividends? *Therapeutic Advances in Hematology*, 3(6):333–339, December 2012. ISSN 2040-6207. doi: 10.1177/2040620712458948.

Michael L. Metzker. Emerging technologies in DNA sequencing, 2005. ISSN 10889051.

Sara C Meyer and Ross L Levine. Translational implications of somatic genomics in acute myeloid leukaemia. *The Lancet Oncology*, 15(9):e382–e394, August 2014. ISSN 14702045. doi: 10.1016/S1470-2045(14)70008-7. URL https://linkinghub.elsevier.com/retrieve/pii/S1470204514700087.

Paul Moss. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The Hematologist*, 13(5), September 2016. ISSN 1551-8779. doi: 10.1182/hem.V13.5.6435. URL https://ashpublications.org/thehematologist/article/doi/10.1182/hem.V13.5.6435/462716/Genomic-Classification-and-Prognosis-in-Acute.

Wenbo Mu, Hsiao Mei Lu, Jefferey Chen, Shuwei Li, and Aaron M. Elliott. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *Journal of Molecular Diagnostics*, 2016. ISSN 19437811. doi: 10.1016/j.jmoldx.2016.07.006.

Stephen D. Nimer. Is it important to decipher the heterogeneity of "normal karyotype AML"? *Best Practice & Research. Clinical Haematology*, 21(1):43–52, March 2008. ISSN 1521-6926. doi: 10.1016/j.beha.2007.11.010.

Alexis L. Norris, Rachael E. Workman, Yunfan Fan, James R. Eshleman, and Winston Timp. Nanopore sequencing detects structural variants in cancer. *Cancer Biology & Therapy*, 17(3): 246–253, 2016. ISSN 1555-8576. doi: 10.1080/15384047.2016.1139236.

on behalf of the UK NCRI AML Study Group, A K Burnett, N H Russell, R K Hills, J Kell, O J Nielsen, M Dennis, P Cahalin, C Pocock, S Ali, S Burns, S Freeman, D Milligan, and R E Clark. A comparison of clofarabine with ara-C, each in combination with daunorubicin as induction treatment in older patients with acute myeloid leukaemia. *Leukemia*, 31(2):310–317, February 2017. ISSN 0887-6924, 1476-5551. doi: 10.1038/leu.2016.225. URL `http://www.nature.com/articles/leu2016225`.

Gert Ossenkoppele and Bob Löwenberg. How I treat the older patient with acute myeloid leukemia. *Blood*, 125(5):767–774, January 2015. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2014-08-551499. URL `https://ashpublications.org/blood/article/125/5/767/34072/How-I-treat-the-older-patient-with-acute-myeloid`.

PCAWG Structural Variation Working Group, PCAWG Consortium, Yilong Li, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O. Korbel, James E. Haber, Marcin Imielinski, Joachim Weischenfeldt, Rameen Beroukhim, and Peter J. Campbell. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, February 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/

s41586-019-1913-9. URL http://www.nature.com/articles/s41586-019-1913-9.

Brent S Pedersen and Aaron R Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34 (5):867–868, March 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btx699. URL https://academic.oup.com/bioinformatics/article/34/5/867/4583630.

Ari Pelcovits and Rabin Niroula. Acute Myeloid Leukemia: A Review. *Rhode Island Medical Journal (2013)*, 103(3):38–40, April 2020. ISSN 2327-2228.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

M. D. Radmacher. Independent confirmation of a prognostic gene-expression signature in adult acute myeloid leukemia with a normal karyotype: a Cancer and Leukemia Group B study. *Blood*, 108(5):1677–1683, May 2006. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2006-02-005538. URL http://www.bloodjournal.org/cgi/doi/10.1182/blood-2006-02-005538.

Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5): 278–289, October 2015. ISSN 2210-3244. doi: 10.1016/j.gpb.2015.08.002.

Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, Jeremy Hoon, Jan F. Simons, David Marran, Jason W. Myers, John F. Davidson, Annika Branting, John R. Nobile, Bernard P. Puc, David Light, Travis A. Clark, Martin Huber, Jeffrey T. Branciforte, Isaac B. Stoner, Simon E. Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline A. Fidanza, Eugeni Namsaraev, Kevin J. McKernan, Alan Williams, G. Thomas Roth, and James

Bustillo. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, July 2011. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10242. URL http://www.nature.com/articles/nature10242.

Xavier Roussel, Etienne Daguindau, Ana Berceanu, Yohan Desbrosses, Walid Warda, Mathieu Neto da Rocha, Rim Trad, Eric Deconinck, Marina Deschamps, and Christophe Ferrand. Acute Myeloid Leukemia: From Biology to Clinical Practices Through Development and Pre-Clinical Therapeutics. *Frontiers in Oncology*, 10:599933, December 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.599933. URL https://www.frontiersin.org/articles/10.3389/fonc.2020.599933/full.

Anne Rovelet-Lecrux, Didier Hannequin, Gregory Raux, Nathalie Le Meur, Annie Laquerrière, Anne Vital, Cécile Dumanchin, Sébastien Feuillette, Alexis Brice, Martine Vercelletto, Frédéric Dubas, Thierry Frebourg, and Dominique Campion. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics*, 38(1):24–26, January 2006. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng1718. URL http://www.nature.com/articles/ng1718.

Jeffrey E. Rubnitz, Brenda Gibson, and Franklin O. Smith. Acute Myeloid Leukemia. *Pediatric Clinics of North America*, 55(1):21–51, February 2008. ISSN 00313955. doi: 10.1016/j.pcl.2007.11.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0031395507001733.

Silvia Salmoiraghi, Roberta Cavagna, Pamela Zanghì, Chiara Pavoni, Anna Michelato, Ksenija Buklijas, Lara Elidi, Tamara Intermesoli, Federico Lussana, Elena Oldani, Chiara Caprioli, Paola Stefanoni, Giacomo Gianfaldoni, Ernesta Audisio, Elisabetta Terruzzi, Lorella De Paoli, Erika Borlenghi, Irene Cavattoni, Daniele Mattei, Annamaria Scattolin, Monica Tajana, Fabio Ciceri, Elisabetta Todisco, Leonardo Campiotti, Paolo Corradini, Nicola Fracchiolla, Renato Bassan, Alessandro Rambaldi, and Orietta Spinelli. High Throughput Molecular Characterization of Normal Karyotype Acute Myeloid Leukemia in the Context of the Prospective Trial 02/06 of the Northern

Italy Leukemia Group (NILG). *Cancers*, 12(8):E2242, August 2020. ISSN 2072-6694. doi: 10.3390/cancers12082242.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12): 5463–5467, December 1977. ISSN 0027-8424. doi: 10.1073/pnas.74.12.5463.

Charles A. Schiffer and Richard M. Stone. Morphologic Classification and Clinical and Laboratory Correlates. *Holland-Frei Cancer Medicine. 6th edition*, 2003. URL https://www.ncbi.nlm.nih.gov/books/NBK13452/.

R. F. Schlenk. Post-remission therapy for acute myeloid leukemia. *Haematologica*, 99(11):1663–1670, November 2014. ISSN 0390-6078, 1592-8721. doi: 10.3324/haematol.2014.114611. URL http://www.haematologica.org/cgi/doi/10.3324/haematol.2014.114611.

Judith Schütte, Julia Reusch, Cyrus Khandanpour, and Christine Eisfeld. Structural Variants as a Basis for Targeted Therapies in Hematological Malignancies. *Frontiers in Oncology*, 9:839, August 2019. ISSN 2234-943X. doi: 10.3389/fonc.2019.00839. URL https://www.frontiersin.org/article/10.3389/fonc.2019.00839/full.

Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, June 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0001-7. URL http://www.nature.com/articles/s41592-018-0001-7.

Rory M. Shallis, Rong Wang, Amy Davidoff, Xiaomei Ma, and Amer M. Zeidan. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood Reviews*, 36: 70–87, July 2019. ISSN 0268960X. doi: 10.1016/j.blre.2019.04.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0268960X18301395.

Liran I. Shlush, Sasan Zandi, Amanda Mitchell, Weihsu Claire Chen, Joseph M. Brandwein, Vikas Gupta, James A. Kennedy, Aaron D. Schimmer, Andre C. Schuh, Karen W. Yee, Jessica L. McLeod, Monica Doedens, Jessie J. F. Medeiros, Rene Marke, Hyeoung Joon Kim, Kwon Lee, John D. McPherson, Thomas J. Hudson, The HALT Pan-Leukemia Gene Panel Consortium, Andrew M. K. Brown, Fouad Yousif, Quang M. Trinh, Lincoln D. Stein, Mark D. Minden, Jean C. Y. Wang, and John E. Dick. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*, 506(7488):328–333, February 2014. ISSN 1476-4687. doi: 10.1038/nature13038. URL `https://www.nature.com/articles/nature13038`.

Jordi Silvestre-Ryan and Ian Holmes. Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. *Genome Biology*, 22(1):38, December 2021. ISSN 1474-760X. doi: 10.1186/s13059-020-02255-1. URL `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02255-1`.

Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1):e59, April 2018. ISSN 1934-3647. doi: 10.1002/cpmb.59.

Jean Soulier. Introduction to a review series on secondary leukemia. *Blood*, 136(1):1–1, July 2020. ISSN 0006-4971. doi: 10.1182/blood.2019004171. URL `https://doi.org/10.1182/blood.2019004171`.

Nancy A. Speck and D. Gary Gilliland. Core-binding factors in haematopoiesis and leukaemia. *Nature Reviews Cancer*, 2(7):502–513, July 2002. ISSN 1474-175X, 1474-1768. doi: 10.1038/nrc840. URL `http://www.nature.com/articles/nrc840`.

R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, 1979. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/6.7.2601. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/6.7.2601`.

Rodger Staden. Computer methods to aid the determination and analysis of DNA sequences. *Biochemical Society Transactions*, 12(6):1005–1008, December 1984. ISSN 0300-5127, 1470-8752. doi: 10.1042/bst0121005. URL https://portlandpress.com/biochemsoctrans/article/12/6/1005/79016/Computer-methods-to-aid-the-determination-and.

Rodger Staden. The staden sequence analysis package. *Molecular Biotechnology*, 5(3):233–241, June 1996. ISSN 1073-6085, 1559-0305. doi: 10.1007/BF02900361. URL http://link.springer.com/10.1007/BF02900361.

Eytan M Stein, Amir T Fathi, Courtney D DiNardo, Daniel A Pollyea, Gail J Roboz, Robert Collins, Mikkael A Sekeres, Richard M Stone, Eyal C Attar, Mark G Frattini, Alessandra Tosolini, Qiang Xu, Wendy L See, Kyle J MacBeth, Stéphane de Botton, Martin S Tallman, and Hagop M Kantarjian. Enasidenib in patients with mutant IDH2 myelodysplastic syndromes: a phase 1 subgroup analysis of the multicentre, AG221-C-001 trial. *The Lancet Haematology*, 7 (4):e309–e319, April 2020. ISSN 23523026. doi: 10.1016/S2352-3026(19)30284-4. URL https://linkinghub.elsevier.com/retrieve/pii/S2352302619302844.

Akshay Sudhindra and Catherine Choy Smith. FLT3 Inhibitors in AML: Are We There Yet? *Current Hematologic Malignancy Reports*, 9(2):174–185, June 2014. ISSN 1558-8211, 1558-822X. doi: 10.1007/s11899-014-0203-8. URL http://link.springer.com/10.1007/s11899-014-0203-8.

Paul Tardi, Sharon Johnstone, Natashia Harasym, Sherwin Xie, Troy Harasym, Natalia Zisman, Pierrot Harvie, David Bermudes, and Lawrence Mayer. In vivo maintenance of synergistic cytarabine:daunorubicin ratios greatly enhances therapeutic efficacy. *Leukemia Research*, 33(1):129–139, January 2009. ISSN 01452126. doi: 10.1016/j.leukres.2008.06.028. URL https://linkinghub.elsevier.com/retrieve/pii/S0145212608003056.

Haotian Teng, Minh Duc Cao, Michael B Hall, Tania Duarte, Sheng Wang, and Lachlan J M Coin. Chiron: translating nanopore raw signal directly into nucleotide se-

quence using deep learning. *GigaScience*, 7(5):giy037, May 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy037. URL https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy037/4966989.

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. The importance of phase information for human genomics, 2011. ISSN 14710056.

Terry M Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-11.

Todd J. Treangen and Steven L. Salzberg. Repetitive DNA and next-generation sequencing: Computational challenges and solutions, 2012a. ISSN 14710056.

Todd J. Treangen and Steven L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, January 2012b. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3117. URL http://www.nature.com/articles/nrg3117.

Peter J.M. Valk, Roel G.W. Verhaak, M. Antoinette Beijen, Claudia A.J. Erpelinck, Sahar Barjesteh van Waalwijk van Doorn-Khosrovani, Judith M. Boer, H. Berna Beverloo, Michael J. Moorhouse, Peter J. van der Spek, Bob Löwenberg, and Ruud Delwel. Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *New England Journal of Medicine*, 350 (16):1617–1628, April 2004. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa040465. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa040465.

Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, Joel A. Malek, Gina Costa, Kevin McKernan, Arend Sidow, Andrew Fire, and Steven M. Johnson. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 2008. ISSN 10889051. doi: 10.1101/gr.076463.108.

Erwin L. van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The Third Revolution in Sequencing Technology, 2018. ISSN 13624555.

James W. Vardiman, Jüergen Thiele, Daniel A. Arber, Richard D. Brunning, Michael J. Borowitz, Anna Porwit, Nancy Lee Harris, Michelle M. Le Beau, Eva Hellström-Lindberg, Ayalew Tefferi, and Clara D. Bloomfield. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*, 114(5):937–951, July 2009. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2009-03-209262. URL `https://ashpublications.org/blood/article/114/5/937/103719/The-2008-revision-of-the-World-Health-Organization`.

Adriano Venditti, Alfonso Piciocchi, Anna Candoni, Lorella Melillo, Valeria Calafiore, Roberto Cairoli, Paolo de Fabritiis, Gabriella Storti, Prassede Salutari, Francesco Lanza, Giovanni Martinelli, Mario Luppi, Patrizio Mazza, Maria Paola Martelli, Antonio Cuneo, Francesco Albano, Francesco Fabbiano, Agostino Tafuri, Anna Chierichini, Alessia Tieghi, Nicola Stefano Fracchiolla, Debora Capelli, Robin Foà, Caterina Alati, Edoardo La Sala, Paola Fazi, Marco Vignetti, Luca Maurillo, Francesco Buccisano, Maria Ilaria Del Principe, Maria Irno-Consalvo, Tiziana Ottone, Serena Lavorgna, Maria Teresa Voso, Francesco Lo-Coco, William Arcese, and Sergio Amadori. GIMEMA AML1310 trial of risk-adapted, MRD-directed therapy for young adults with newly diagnosed acute myeloid leukemia. *Blood*, 134 (12):935–945, September 2019. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood.2018886960. URL `https://ashpublications.org/blood/article/134/12/935/374904/GIMEMA-AML1310-trial-of-riskadapted-MRDdirected`.

Maria Teresa Voso, Tiziana Ottone, Serena Lavorgna, Adriano Venditti, Luca Maurillo, Francesco Lo-Coco, and Francesco Buccisano. MRD in AML: The Role of New Techniques. *Frontiers in Oncology*, 9:655, July 2019. ISSN 2234-943X. doi: 10.3389/fonc.2019.00655. URL `https://www.frontiersin.org/article/10.3389/fonc.2019.00655/full`.

Roland B. Walter, Megan Othus, Gautam Borthakur, Farhad Ravandi, Jorge E. Cortes, Sherry A. Pierce, Frederick R. Appelbaum, Hagop A. Kantarjian, and Elihu H. Estey. Prediction of Early Death After Induction Therapy for Newly Diagnosed Acute Myeloid Leukemia With Pretreatment Risk Scores: A Novel Paradigm for Treatment Assignment. *Journal of Clinical Oncology*, 29(33):4417–4424, November 2011. ISSN 0732-183X, 1527-7755. doi: 10.1200/JCO.2011.35.7525. URL `http://ascopubs.org/doi/10.1200/JCO.2011.35.7525`.

Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O. Korbel. Phenotypic impact of genomic structural variation: Insights from and for human disease, 2013a. ISSN 14710056.

Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O. Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138, February 2013b. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3373. URL `http://www.nature.com/articles/nrg3373`.

Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 2019. ISSN 15461696. doi: 10.1038/s41587-019-0217-9.

Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, 20(1):129, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1727-y. URL `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1727-y`.