**Aus der Medizinischen Klinik III,**
**Klinikum der Ludwig-Maximilians-Universität München**
Vorstand: Prof. Dr. Dr. Michael von Bergwelt-Baildon

# Differential splicing and allelic imbalance as pathomechanisms of recurring mutations in Acute Myeloid Leukemia

Dissertation
zum Erwerb des Doktorgrades der Medizin
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von

Stefanos Alexandros Bamopoulos
aus Cholargos, Griechenland

Jahr
2021

Mit Genehmigung der Medizinischen Fakultät der Universität München

Berichterstatter: PD Dr. Tobias Herold

Mitberichterstatter: Prof. Dr. Michael Albert
PD Dr. Oliver J. Stötzer
PD Dr. Hanna-Mari Baldauf

Dekan: Prof. Dr. med. dent. Reinhard Hickel

Tag der mündlichen Prüfung: 18.02.2021

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Promotionsbüro
Medizinische Fakultät

MMRS MUNICH MEDICAL RESEARCH SCHOOL

## Eidesstattliche Versicherung

# Bamopoulos, Stefanos Alexandros

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Titel

Differential splicing and allelic imbalance as pathomechanisms of recurring mutations in Acute Myeloid Leukemia

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Berlin, den 03.04.2021

Ort, Datum

Stefanos Alexandros Bamopoulos

Unterschrift Doktorandin bzw. Doktorand

# CONTENTS

# ABBREVIATIONS

| | |
|---|---|
| AI | allelic imbalance |
| AML | acute myeloid leukemia |
| AMLCG | German AML Cooperative Group |
| AMLSG | German-Austrian AML Study Group |
| AS | alternative splicing |
| cDNA | complementary DNA |
| CNV, -s | copy number variation, -s |
| DNA-seq | DNA sequencing |
| dNTP, -s | deoxynucleoside triphosphate, -s |
| EHA | European Hematology Association |
| ELN | European Leukemia Net |
| INDEL, -s | insertion deletion, -s |
| mRNA | messenger RNA |
| NGS | next-generation sequencing |
| pre-mRNA | precursor messenger RNA |
| RNA-seq | RNA sequencing |
| rRNA | ribosomal RNA |
| SBS | sequencing by synthesis |
| SF | splicing factor |
| snRNP, -s | small nuclear riboprotein, -s |
| SNP, -s | single nucleotide polymorphism |
| SNV, -s | single nucleotide variation, -s |
| VAF, -s | variant allele frequency, -ies |
| VC | variant calling |
| WES | whole exome sequencing |
| WGS | whole genome sequencing |

## PUBLICATION LIST

I. **Bamopoulos, S.A.**, Batcha, A.M.N., Jurinovic, V., Clinical presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with Acute Myeloid Leukaemia. *Leukemia (article in press).* doi: 10.1038/s41375-020-0839-4

Journal Citation Report (2018): Impact Factor: 9.944, Ranking: 4/73, Category: Hematology

II. Batcha, A.M.N., **Bamopoulos, S.A.**, Kerbs, P. et al. Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *Sci Rep* 2019 Aug 13;9(1):11796. doi: 10.1038/s41598-019-48167-4

Journal Citation Report (2018): Impact Factor: 4.011, Ranking: 15/69, Category: Multidisciplinary Sciences

# INTRODUCTION

## 1. Acute myeloid leukaemia

Acute myeloid leukaemia (AML) is a malignant neoplasm with an incidence of 4.3 per 100.000 men and women every year, which corresponds to a lifetime risk of receiving an AML diagnosis of 0.5%.[1] The scientific community has invested a sizeable amount of time, effort and resources into improving the prognosis of AML. However, despite continuous improvement, the relative five-year survival rate of AML is still only at 28.7%.[1] The advent of next-generation sequencing (NGS), has revolutionized the field of cancer genomics. NGS continues to provide scientists and medical professionals with new ways to improve the diagnosis and classification of neoplasms and builds the foundation for new treatment options through the identification of cancer-specific, molecular targets.[2] In AML specifically, NGS methods have allowed scientist to detect recurring mutations with high accuracy, observe their pattern of co-expression and shed light on their often genome-wide consequences on core biological processes.[3,4] In this thesis two articles are presented, which employ two complementary NGS methods in conjunction with standard statistical tools in order to investigate the pathomechanisms of recurring mutations in AML.

Acute myeloid leukaemia refers to a group of malignant neoplasms that are derived from undifferentiated myeloid precursor cells.[5] The disease is characterized by rapid clonal expansion of myeloid blasts in the bone marrow and peripheral blood, leading to impaired haematopoiesis and bone marrow failure. Depending on the impaired cell line, patients may present with symptoms of anaemia (lethargy and fatigue, shortness of breath, pale skin), easy bruising or unusual bleeding (e.g. frequent nosebleeds, or bleeding from the gums), fever and frequent infections. Known risk factors include old age, male sex, exposure to radiation or carcinogenic chemicals (e.g. benzene), other blood disorders such as myelodysplastic syndrome, previous chemotherapeutic treatment and certain genetic disorders such as Down syndrome. Although several treatment options are available, including allogeneic stem cell transplants, the disease is often refractory or recurs after treatment.[6]

Carcinogenesis and consequently leukaemogenesis has been linked to mutations that disrupt the balance between proliferation and programmed cell death (apoptosis) leading to uncontrolled cell growth. The emergence of new technologies, especially

next-generation sequencing, has led to the discovery of genes that are recurrently mutated in AML and other haematopoietic malignancies.[7–9] This has shifted the focus from morphological classifications of AML to more accurate cytogenetic and molecular classifications that have improved the diagnosis and management of the disease. One such classification, the European Leukemia Net (ELN) 2017 classification includes a large spectrum of cytogenetic and molecular parameters and has supported physicians in the selection of treatment options for AML patients.[10] The parameters of the ELN 2017 classification were used in *Publication I* in order to create a regression model with the purpose of characterizing the prognostic relevance of splicing factor (SF) mutations, which make up a relevant portion of the recurring mutations in AML.[4] A prerequisite for interpreting the impact of SF mutations is understanding the process of splicing, a core concept of cancer genomics and molecular biology in total for that matter.[11] The next section covers this biological mechanism and its implications for the field of cancer genomics.

## 2. Splicing

### 2.1. Definition

The central dogma of molecular biology states that the information in a cell flows via transcription from DNA to RNA, which is then translated into a polypeptide sequence (a protein) that executes a specific function in the cell.[12] In the past decades scientist have taken a closer look at this process and have refined our understanding of it. The regions on the DNA that are transcribed to RNA are referred to as genes. While all genes are by definition transcribed to RNA, only messenger RNA (mRNA) is used as a template for the synthesis of a protein. However, the initial product of transcription is in the vast majority of cases a precursor messenger RNA (pre-mRNA). This primary transcript must undergo processing prior to translation. Several post-transcriptional modifications take place, such as 5'-capping, polyadenylation (which is used for poly(A) selection of mRNA in RNA-seq experiments[13]) and splicing.

### 2.2. The splicing process

The term splicing refers to the removal (cleavage) of non-coding regions of the pre-mRNA and the ligation of the remaining fragments. Regions on the pre-mRNA that

contain information for protein coding are termed exons, while non-coding regions are referred to as introns. The molecular machine responsible for splicing is the spliceosome, which is comprised of five small nuclear RNAs and a number of associated protein factors that combine and form small nuclear riboproteins (snRNPs). In eukaryotic organisms, the major spliceosome that is responsible for splicing most genes is comprised of five such snRNP units (U1, U2, U4, U5 and U6) that assemble anew on each pre-mRNA.[14,15]

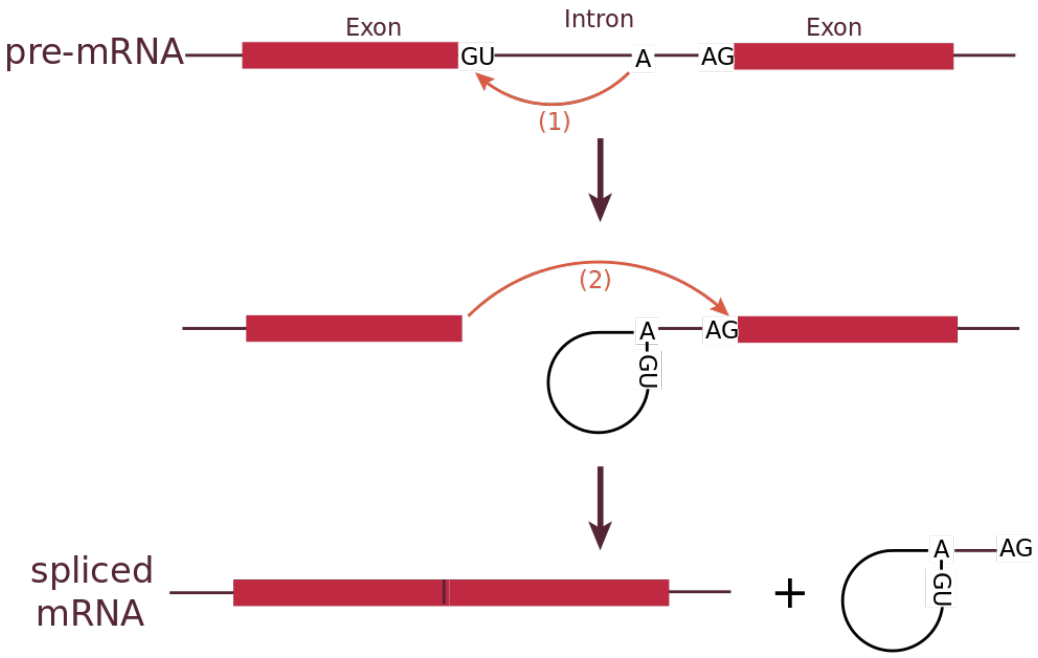


**Figure 1: mRNA splicing.** Schematic illustration depicting the core steps of the splicing process. Illustration taken from Wikimedia Commons where it is available under a Creative Commons license (https://en.wikipedia.org/wiki/RNA_splicing#/media/File:RNA_splicing_reaction.svg).

Splicing occurs at specific conserved sequences that are found at the 5' and 3' ends of the spliced intronic sequence and are known as splice sites. Another critical sequence for splicing is the branch point, which is a sequence usually located a few bases (~20-50) upstream of the 3' splice site that is loosely conserved, but always contains an adenine nucleotide.[16] The process is initiated by the binding of the U1 snRNP unit to the 5' splice site, committing the pre-mRNA to splicing. This causes the 5' splice site to pair with the adenine at the branch point through a chemical process called transesterification. This requires the pre-mRNA to "bend" and form a circular

structure called a “lariat”. The free 3’-end of the upstream exon is covalently bound to the 5’ splice site and the lariat structure is cleaved away, concluding the splicing process (Figure 1).[14]

### 2.3. Alternative splicing

In the past decades, scientist have observed that a gene can produce multiple products although it is a singular template. This capability of genes is facilitated by a process referred to as alternative splicing (AS), which is a major contributor of proteomic diversity between individuals and different tissue types in an individual.[17] While up to the beginning of the 21st century this process was believed to be an exception, current studies suggest that it is in fact a universal phenomenon that occurs in up to 95% of human genes.[18,19] Alternatively spliced genes can produce several transcripts isoforms, which are translated
into proteins that can serve different or even opposing purposes.[17,20] AS has also been linked to other cellular mechanisms, including gene regulation[21,22] and intracellular transport of mRNAs and proteins.[23,24] AS can produce transcript isoforms, where an exonic sequence is excluded or an intronic sequence is included. Four main types of AS have been described: a “cassette” exon, mutually exclusive exons, alternative 5’ or 3’ splice sites and intron retention and are shown in Figure 2. A cassette exon describes an exon that is included in one isoform, but excluded completely from another. Mutually exclusive exons are not spliced in together in one isoform, i.e. when one exon is spliced in, the other exon is spliced out. Alternative splice sites refer to the fact that splicing can occur upstream or downstream of the 5’ or 3’ splice sites at a sequence, which also contains a splice junction motif. Finally, intron retention refers to the inclusion of an intronic sequence in the final mature mRNA product.

### 2.4. Identification of disease-relevant aberrant splicing

Splicing and AS, are highly regulated processes.[25] However, mutations can induce aberrant splicing in cancer cells.[26] The resulting aberrant transcript isoforms are translated into proteins that are missing domains or are truncated. As a consequence, proteins may be degraded prematurely (e.g. through nonsense-mediated decay[27]), have severe functional impairments, or may even present with a completely new function.

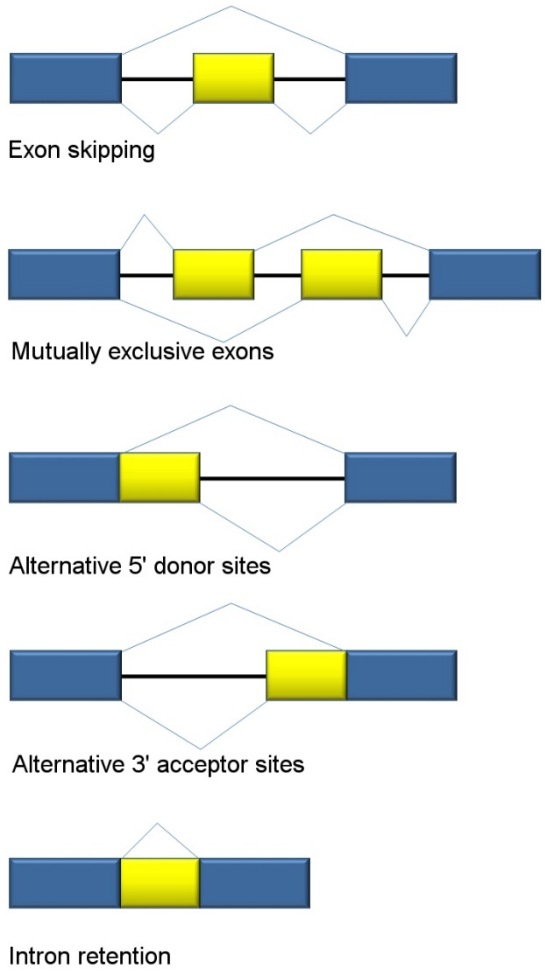


**Figure 2: Modes of alternative splicing.** The most common types of alternative splicing are shown. Exons are colored blue and introns are depicted in yellow. The blue lines connecting exons represent splice junctions. Illustration taken from Wikimedia Commons where it is available under a Creative Commons license. (https://commons.wikimedia.org/wiki/File:Alt_splicing_bestiary2.jpg)

Increasing evidence has come to light that aberrant splicing is integral to cancer development and thus a growing body of research aims at uncovering aberrant splicing patterns in tumor cells. The recent discovery of mutations in genes, which are directly involved in the transcription and splicing pathway (e.g. SF mutations), has provided an additional incentive.[28] The emergence of RNA-seq has greatly contributed to the detection of aberrant splicing, as it comes with genome-wide capabilities of splice junction detection. Through transcriptome assembly tools it is possible to reconstruct transcript isoforms of a sample or a sample group and compare them to an existing reference transcriptome, thereby detecting novel transcript isoforms.[29] The usage of tools with powerful novel splice junction detection algorithms has also supported the identification of aberrant splicing. The following sections cover the fundamentals of NGS and its applications in molecular biology using the Illumina sequencing method as an example.

## 3. Next-generation sequencing

### 3.1. Overview

Sequencing in molecular biology refers to the identification of the primary order of biological structures. Depending on the genomic region studied, DNA-sequencing (DNA-seq) can be classified into whole genome sequencing (WGS; sequencing of the

complete DNA of an organism), whole exome sequencing (WES; only gene-coding regions are sequenced) or targeted sequencing (only selected genomic regions are sequenced). RNA-sequencing (RNA-seq) provides the ability to sequence the total transcriptome of a sample (whole transcriptome sequencing; corresponds to WES). However, RNA-seq can also be used to study specific RNA populations, such as mRNA, micro RNA, transfer RNA and ribosomal RNA (rRNA) or more recently the total RNA population of a single cell (single-cell RNA sequencing[30]). Other applications include ribosomal profiling and targeted RNA-seq.[31]

The sequencing data analyzed in this thesis included both DNA and RNA data and was created on an Illumina sequencing platform, the most broadly used, commercially available NGS platform.[32] A typical Illumina NGS workflow consist of four steps: library preparation, cluster ("bridge") amplification, sequencing by synthesis, and data analysis. The studies presented in this thesis cover different ways of interpreting NGS data and thus fall under the data analysis step. However, accurate data analysis requires understanding of all steps of an NGS workflow, which is why they will briefly be covered in the following sections.

### 3.2. Library preparation

Prior to the actual sequencing process a sequencing library (a collection of DNA or complementary DNA (cDNA) fragments) needs to be created, which will serve as the primary sequencing input. This creation process is called library preparation and requires isolated and purified DNA or RNA.[33] Core steps in the isolation and purification of DNA include cell lysis, the removal of membrane lipids, the denaturation and removal of proteins (through protease treatment), and other cellular components, as well as the removal of RNA through ribonuclease treatment. Common DNA extraction and purification methods are organic extraction, magnetic separation, silica-based and anion exchange technology.[34,35] The use of a specific extraction kit varies depending on the sample source type. While the process of RNA isolation and purification is similar to DNA purification, the ubiquitous presence of RNases in the environment (including human skin) provide an additional challenge in the collection of intact RNA molecules. The most common RNA extraction methods are the acid guanidinium thiocyanate-phenol-chloroform extraction and silica-based methods.[36] In most RNA-seq studies only the mRNA is of interest, however rRNA accounts for most (~80%) of the RNA in eukaryotic organisms. Therefore, an additional step to enrich for mRNA is

required. Two common methods that address this issue are poly(A) selection and rRNA depletion, each having their own advantages and drawbacks.[13,37–41] In the whole-transcriptome RNA-seq protocol used to generate data for the studies presented in this thesis poly(A) selection was used to enrich for mRNA. The RNA molecules of interest are then converted into cDNA molecules through reverse transcription.

The obtained DNA (or cDNA) molecules are fragmented into smaller segments (usually a few hundred base pairs each), through mechanical methods (e.g. sonication, nebulization) or enzymatic methods (e.g. enzymatic digestion[42,43]), which is followed by the 3' and 5' ligation of synthetic oligonucleotides (known as adapters) to the fragments. The final library preparation step includes polymerase chain reaction amplification and gel purification of the adapter-ligated fragments. The inclusion of unique, library-specific index sequences in the adapters, called multiplexing, permits the pooling and concurrent sequencing of up to 96 samples in a single sequencing run and has contributed significantly to the increased output potential of NGS methods.

### 3.3. Cluster generation

The cluster generation step begins by loading the adapter-ligated fragments onto a proprietary flow cell. The flow cell is, in essence, a glass slide with up to eight physically separate lanes, each of which can host an independent analysis. The flow cell is coated with oligonucleotide sequences that are complementary to one of the adapter sequences of the fragments, which allows for the binding of the fragments to the flow cell. A polymerase then creates a complement to the bound fragment. The double-stranded DNA is then denatured and the original fragment (forward strand) is washed away, leaving only the reverse complement (reverse strand), which is covalently bound to the flow cell.

Clusters are created through a procedure known as bridge amplification. During this procedure, the free end of the bound reverse strands bends over and binds ("bridges") to the oligonucleotide bound on the flow cell, which is complementary to its free end. The strands are then amplified through repeated extension and denaturation cycles. As the fragments are continuously bound on one end to the flow cell, the clonal amplification takes place locally, resulting in millions of unique clonal clusters across the flow cell. These clusters contain both the forward and the reverse strand of each original molecule. The reverse strand is cleaved and washed away prior to the actual sequencing process.

In the case of paired-end sequencing, where sequences are created in pairs, the reverse strand is synthesized again after the sequencing of the forward strand has been concluded and the forward strand is cleaved and washed away. The reverse strand is then sequenced as well, resulting in a sequencing pair (a “paired-end” read). Paired-end sequencing is currently used for most scientific questions, including the ones addressed in the studies presented in this thesis, as it is more time efficient and more accurate in the detection of artifacts, genomic rearrangements and insertion-deletion variants, as opposed to single-end sequencing.[44] Nevertheless, single-end still finds its applications (e.g. in small RNA sequencing).

### 3.4. Illumina sequencing by synthesis (SBS)

The standard Illumina sequencing by synthesis (SBS) method uses fluorescently-labeled deoxynucleoside triphosphates (dNTPs), which contain an element that can reversibly terminate polymerization. These dNTPs each contain one of the four nitrogenous bases necessary to synthesize a DNA molecule (adenine, thymine, cytosine or guanine). Firstly, primers and polymerization enzymes are added to the flow cell. During each sequencing cycle, all four fluorescently-labeled dNTPs are added to the flow cell containing the clonally amplified forward strands. After the incorporation of a single dNTP, which happens in a massively-parallel fashion for all the clones across the flow cell, the polymerization is terminated and the remaining dNTPs are washed away. The concurrent presence of all four dNTPs leads to natural competition, thus reducing incorporation bias.[45] After laser excitation, the fluorophores on the dNTPs incorporated in each cluster emit light at a characteristic wavelength, which is captured by a camera and used to identify the dNTP type, i.e. the nitrogenous base contained within the dNTP molecule. This method is not sensitive enough to capture the light emitted from a single molecule, however the clonal amplification step prior to sequencing allows for accurate identification of the incorporated dNTP by signal enhancement. The certainty with which a base is identified depends on the signal intensity captured and is quantified through a Phred quality score (Q score).[46] After the base has been identified the reversible polymerization terminator is enzymatically cleaved along with the fluorophore, ending the sequencing cycle. The repetition of multiple sequencing cycles results in a sequence of bases (a “read”), which in Illumina sequencing usually has a length between 50 and 300 base pairs (bp). If paired-end sequencing is performed, the reverse strand is also sequenced and the end result are

two sequences of equal length with an unsequenced region of a few hundred bp in between. Millions of these pairs are created per sample and used for the subsequent data analysis.

### 3.5. Data analysis

The data analysis for DNA and RNA sequencing is computationally intensive and thus requires time- and memory-efficient tools. As the cost for sequencing decreases at a fast pace and therefore sequencing experiments tend to have larger sample sizes, the need for specialized tools that scale well with increasing input size and have multi-threading capabilities intensifies. Each analysis pipeline must be customized to answer the biological question at hand, however some core steps exist that are mandatory in most analyses.

Prior to any analysis, samples need to be demultiplexed, if multiplexing has occurred during library preparation, as each sample usually represents a biological or technical replicate. This process should be followed by a quality control step, where each sample is checked for sufficient coverage over the genomic regions covered in the experiment, as well as the average read quality elicited from the Phred score provided by the sequencer.[46] Depending on the quality of the reads some reads or rarely whole samples may need to be excluded from downstream analysis, however in most cases trimming of the reads (removal of low quality 3' or 5' bases) is sufficient. Tools that possess trimming capabilities can frequently also remove (incomplete) adapter sequences that were not removed sufficiently by the sequencer.

The quality-controlled reads are then (in most experiments) "aligned" to the reference genome, meaning that a computational tool identifies the genomic location from which these reads originated and maps them to specific genomic coordinates. This task differs between DNA- and RNA-seq, as for RNA-seq reads splicing has occurred, therefore aligners need to be "splice-aware". This procedure is followed by another quality control step to assess the quality of the alignment.

Sequencing workflows usually diverge after alignment and are tailor-made to fit the scientific research question examined. Some common applications of a sequencing data analysis are variant calling, gene and isoform expression analyses, gene fusion detection and the detection of novel isoforms. For some analyses DNA-seq data provide more accurate results (e.g. variant calling, an issue addressed in *Publication II*), while others are only possible with RNA-seq data (e.g. novel isoform detection). A

variant calling pipeline was the cornerstone of *Publication II*. The core concepts of variant calling, as well as its capabilities and limitations are covered in the following sections.

# 4. Genomic variants

## 4.1. Classification

Genomic variants are present ubiquitously across the genome and contribute to the genomic diversity of a species. The term variant is neutral in respect to the effect of the genomic variation, while the term polymorphism usually describes a benign variation. However, there is no clear distinction and the term single nucleotide polymorphism (SNP) is often used synonymously with the term single nucleotide variation (SNV). In contrast, the term mutation is usually reserved for a variant with a harmful effect often in association with a specific disease.

When variants are mentioned in the context of sequencing experiments they refer to genomic variations that exist in the sample or samples studied compared to the reference genome (sequencing variant). They can be classified according to their type into single nucleotide variations, insertions, deletions and substitutions, as well as structural variants and chromosomal aberrations. SNVs are sequence variations where the sample examined has a single different nucleotide compared to the reference genome. Insertions describe the addition of a nucleotide sequence to the genome, while deletions describe the removal of a nucleotide sequence. Substitutions occur when a sequence of nucleotides is replaced by another nucleotide sequence and is usually reserved for cases where both the replacing and the replaced sequences have the same length. In contrast INDELs, short for insertion and deletion polymorphisms, describe the insertion or deletion of a sequence, resulting in an absolute difference in the number of nucleotides. The term is also used for cases where both an insertion and a deletion occur at the same genomic coordinates or it is not possible to clearly distinguish between them. Structural variants affect a large genomic region (usually > 1000bp) and can present themselves as deletions, duplications, insertions, inversions, translocations and copy number variations (CNVs). CNVs are repetitions of a specific genomic segment a set number of times that can differ between individuals. Similarly,

chromosomal abnormalities affect a large genomic sequence (the term is loosely defined) with often severe functional impairments (e.g. Trisomy 21).

INDELs can be further classified into frameshift and non-frameshift INDELs. A frameshift is defined as the occurrence of an INDEL in a coding region of the genome, which causes a nucleotide difference that is not a multiple of three. As a triplet of nucleotides (a codon) is translated into an amino acid, a shift in the reading frame of a coding region leads to a completely altered sequence of amino acids in the resulting protein. As the stop codon (the last codon signifying the end of the translation process) is also altered, the protein can be truncated or abnormally long and is very like to have an impaired or completely lacking function. The earlier in the reading frame a frameshift occurs, the more detrimental the effect on the protein function will be. Non-frameshift or in-frame INDELs cause an insertion or deletion (or both) that is divisible by three.

Lastly, mutations can be classified according to their effect on the amino acid sequence of a protein (Figure 3). Mutations that do not alter the amino acid sequence are called “silent” mutations and are usually benign. Conversely, mutations that lead to a different amino acid sequence are termed “missense” mutations. Frameshift mutations or substitutions that introduce a stop codon usually lead to a completely altered or truncated amino acid sequence and are called “nonsense” mutations, as the protein can no longer fulfill its function.

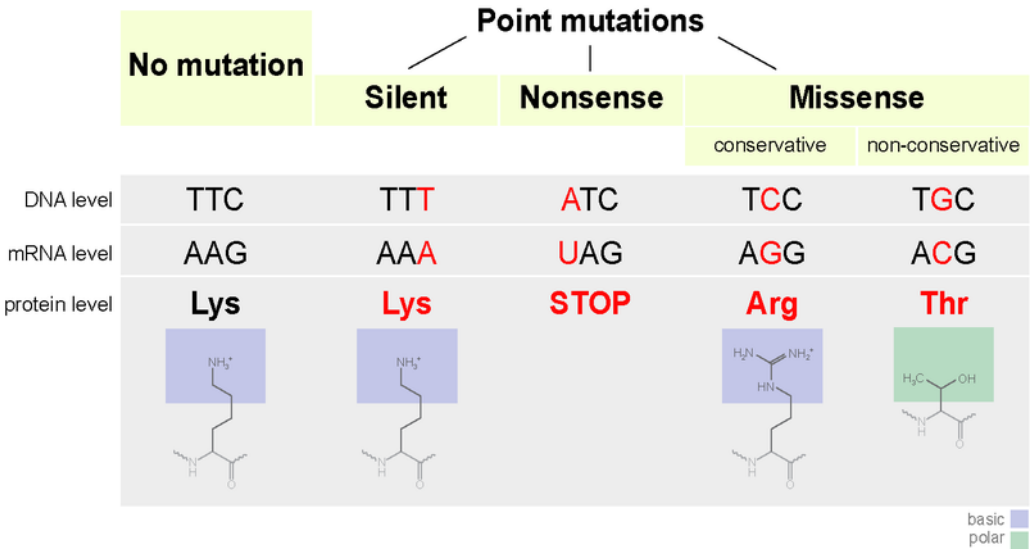


**Figure 3. Effect of point mutations on protein structure.** Mutations can be characterized according to their effect on the amino acid sequence of the resulting protein into silent, nonsense or missense mutations. Missense mutations can be further classified based on the similarity (in terms of chemical properties) between the resulting amino acid and the amino acid that would have resulted from the

unmutated RNA sequence. Illustration taken from Wikimedia Commons where it is available under a Creative Commons license (https://en.wikipedia.org/wiki/File:Different_Types_of_Mutations.png).

### 4.2. Variant calling

Variant calling (VC) describes the *in silico* process by which sequencing variants are identified from sequencing data. After a sample has been sequenced and aligned to its corresponding reference genome, the alignment file undergoes quality control and pipeline specific filtering steps. The alignment file then serves as input for VC tools. The underlying algorithm of VC tools checks for sequence variations between aligned reads and the reference genome. If a variant is present in a sufficient number of reads (criteria are provided by the user), the detected variant is "called" and its genomic coordinates are provided along with the type of variation (e.g. SNV or INDEL). One additional critical information that is provided by VC tools, is the variant allele frequency (VAF). The VAF is a measure of frequency, which represents the relative frequency of a variant in a specific gene locus. The gene locus containing a variant is called an allele (therefore variant *allele* frequency).

In a sequencing experiment, if a variant is detected in 20 out of 100 reads, it has a VAF of 20%. As the VAF depends on the number of reads covering a genomic region, the more coverage a region has (the higher the sequencing depth) the more accurate the VAF will be. Additionally with increasing coverage, variants with low VAF can be detected. The VAF is often used to determine whether a variant is valid or the result of a sequencing error. Due to the direct dependence of the VAF on the coverage of the genomic locus, targeted sequencing experiments, where a few loci are sequenced with a high sequencing depth, are superior to whole genome/transcriptome sequencing, where the coverage over individual genomic regions tends to be low. Furthermore, as mRNA undergoes post-translational modifications and RNA-seq is more prone to sequencing errors, DNA-seq is the preferred sequencing method for VC.[47]

When both the variants and the RNA expression profile of an individual is of interest, as is often the case in cancer samples, a joint approach using both DNA- and RNA-seq is used. A limited amount of studies have compared VC in samples where both DNA and RNA sequencing data is available and observed that the VAFs of variants can differ significantly between DNA- and RNA-seq.[47] When such a difference is observed it is coined allelic imbalance (AI).[48] For example, in a sample where a variant has a VAF of 10% in DNA-Seq and a VAF of 50% in RNA-seq, AI is present. One of

the underlying biological mechanisms that leads to an AI is termed allele-specific transcription. Allele-specific transcription describes the fact that one allele is transcribed disproportionally with respect to its VAF, a phenomenon especially interesting when it occurs in mutated cancer-related genes.[48] The next section covers allelic imbalance along with alternative splicing as pathomechanisms of recurring mutations in AML patients.

## 5. Differential splicing and allelic imbalance as pathomechanisms of recurring mutations in AML

In the studies presented in this thesis the authors attempted to illuminate the pathomechanism of recurring mutations in AML. Both studies included an *in silico* analysis using whole-transcriptome RNA-sequencing and shared several common data analysis steps. All patients who received RNA-sequencing were participants in two trials of the German AML Cooperative Group (AMLCG). The trials were randomized phase III trials and patients were treated with intensive induction chemotherapy. Matched targeted DNA-seq and cytogenetic data was available for all RNA-seq patients. Additional cohorts were used for validation and are highlighted below. The first study focused on SF genes harboring recurring mutations in AML, while the second study analyzed 11 recurrently mutated genes with respect to the presence of AI.

*Publication I*:

In this study we examined mutations in the three most commonly affected SF genes in AML, namely the Serine/arginine-Rich Splicing Factor 2 (*SRSF2*), the U2 small nuclear RNA Auxiliary Factor 1 (*U2AF1*) and the Splicing Factor 3B Subunit 1 (*SF3B1*). Mutations in these genes are, in the majority of cases, heterozygous point mutations and rarely co-occur within the same patient.[49,50] They have been shown to occur early on in cancer evolution and are common events in MDS and AML that frequently coincide with other recurring mutations.[8,51] Several studies have characterized them thoroughly in MDS and have highlighted their importance as independent prognostic markers.[3,52,53] Furthermore, two large RNA-seq studies have been performed on MDS patients delineating the splicing changes that they induce, leading to the dysregulation of several disease-relevant genes.[52,54] Corresponding previous literature in AML patients is limited to small sample sizes and specific AML subgroups.[55]

In the first part of the study the clinical characteristics of SF mutated patients are outlined in two large cohorts (AMLCG cohort, n = 1119 and AMLSG cohort, n = 1540, Figure 4) treated in randomized clinical trials. The large number of SF mutated patients (n = 216; AMLCG cohort) enabled the analysis of the four most frequent individual point mutations: *SRSF2*(P95H), *SRSF2*(P95L), *U2AF1*(S34F) and *SF3B1*(K700E). SF mutations were correlated with demographic and molecular parameters using standard statistical tests. Briefly, our findings show that SF mutations are frequent in AML and even more frequent in elderly and secondary AML patients and are co-expressed with a number of recurrent mutations. A survival analysis using Kaplan-Meier curves and log-rank testing showed that SF mutated patients presented with inferior relapse-free survival and overall survival (Figure 4), which could be validated with simple Cox regression models. However, multiple Cox regression models incorporating parameters of the ELN 2017 classification did not show that SF mutations are individual prognostic markers.

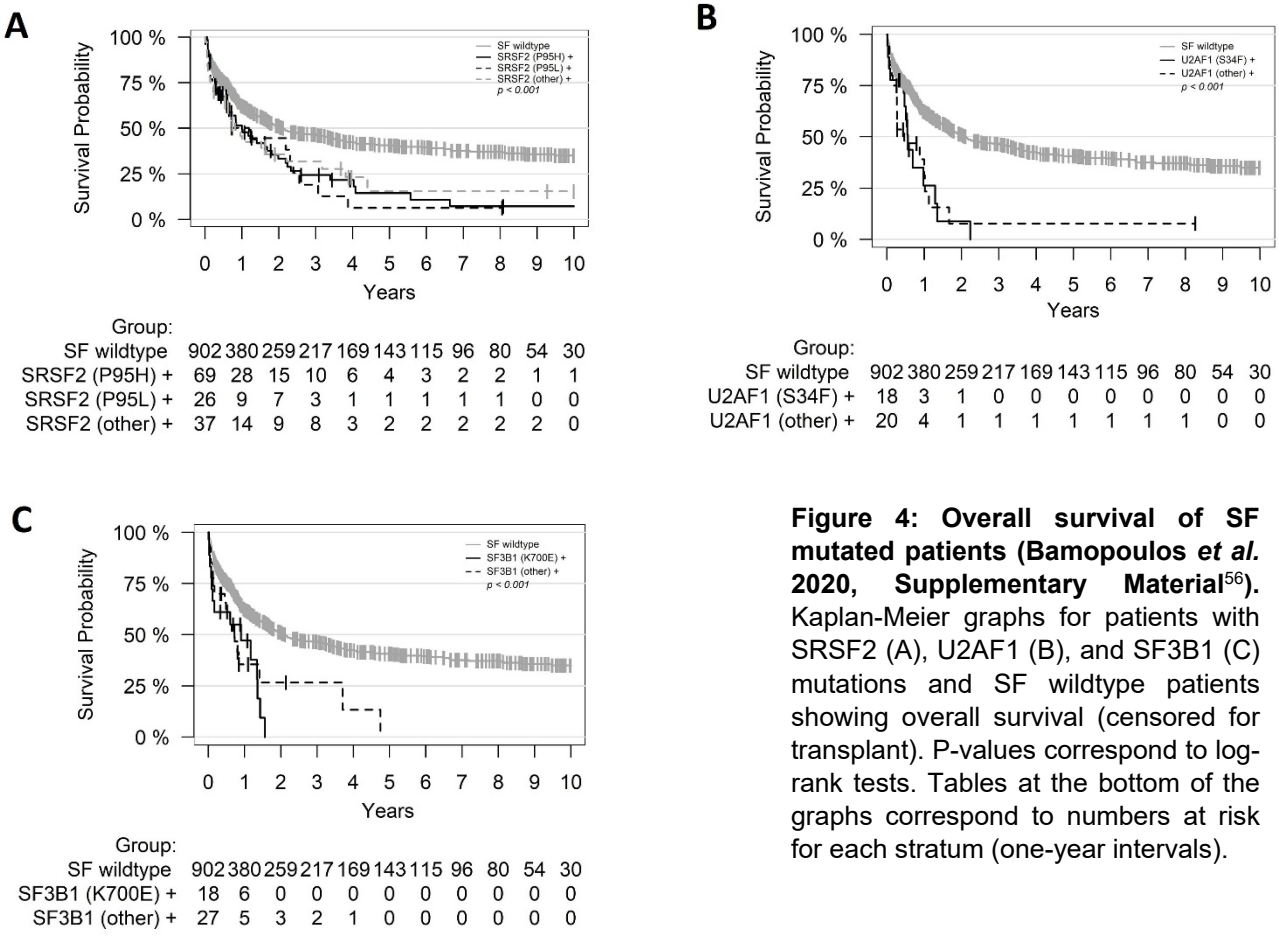


**Figure 4: Overall survival of SF mutated patients (Bamopoulos *et al.* 2020, Supplementary Material[56]).** Kaplan-Meier graphs for patients with SRSF2 (A), U2AF1 (B), and SF3B1 (C) mutations and SF wildtype patients showing overall survival (censored for transplant). P-values correspond to log-rank tests. Tables at the bottom of the graphs correspond to numbers at risk for each stratum (one-year intervals).

For the functional analysis of SF mutations an RNA-seq analysis was performed on a subgroup of the AMLCG cohort (n = 246). In addition, a subset of the Beat AML cohort (n = 177)[9] was used for validation. In a first step, a differential isoform expression analysis showed little overlap in the genes differentially expressed in each SF mutant subgroup. This was confirmed through unsupervised clustering, which showed an expression profile characteristic for each SF mutation, with several dysregulated genes being cancer-relevant genes. We next developed a novel differential splice junction usage pipeline to identify splice junctions that are over- or underused in SF mutants. Importantly, this pipeline allows for the accurate detection of both known and novel splice junctions, which we were able to show by validating our findings in the Beat AML cohort. A gene ontology analysis combing the differential isoform expression and differential splice junction usage analysis results supported a strong dysregulation of the splicing pathway in SF mutants. Lastly, we identified two splice junctions in the genes EVL and NBEAL2 the usage of which correlated with worse prognosis in both datasets analyzed, thereby exemplifying the clinical relevance of our approach (Figure 5).

In summary, this study provides a comprehensive overview of *SRSF2*, *U2AF1* and *SF3B1* mutations in AML patients. The first half covers their frequency and association with patient characteristics and other molecular markers in AML, while also showing that they do not have independent prognostic value for patient survival in AML. The latter half consist of a functional RNA-seq analysis highlighting differentially expressed isoforms in AML-related genes, while also shedding light on the genome-wide effect of SF mutations on splicing and identifying splice junctions with clinical relevance in AML.

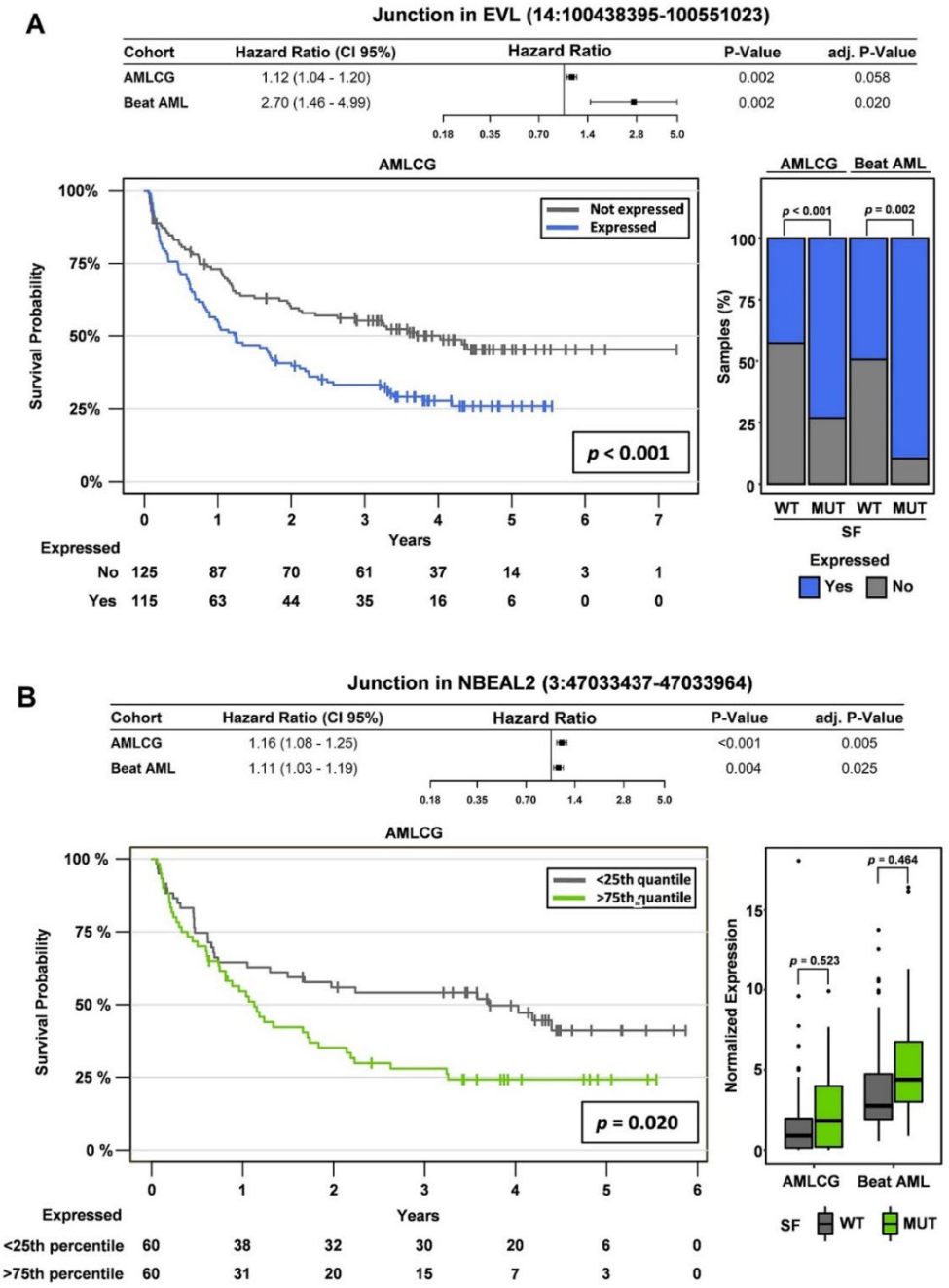


**Figure 5. Differential splice junction usage correlates with patient survival (Bamopoulos *et al.* 2020[56]).** Cox regression models incorporating normalized measurements of splice junction usage for one junction in the gene *EVL* and one in *NBEAL2* showed that the usage of these splice junctions can predict patient overall survival (also visualized with Kaplan-Meier graphs). The junction in *EVL* showed clear overusage in SF mutated patients.

Author Contribution:

The author of this thesis contributed to the following elements of *Publication I*:

- Design of the clinical analysis, preparation and parsing of all available clinical data of the AMLCG and AMLSG cohorts
- Statistical correlation of SF mutations with other molecular and cytogenetic markers in AML
- Survival analysis of SF mutations
- Conception and design of the RNA-Seq pipeline, automatization, optimization and execution of the RNA-Seq workflow, establishment and performance testing of all necessary tools
- Assistance in the storage and management of all genomic data used in the study, quality assessment of the data
- Design and execution of the differential isoform expression analysis, hierarchical clustering and gene ontology analysis
- Design of a novel differential splice junction usage pipeline and survival analysis of significantly differentially spliced junctions in SF mutants
- Writing and submission of the publication manuscript and supplementary information along with the creation of all graphic elements and tables

In addition, the contents of this publication were presented as poster at the 24$^{th}$ European Hematology Association (EHA) congress 2019 for which a travel grant was awarded to the author by the EHA.

*Publication II*:

Mutations and their contribution to cancer are a central focus of current cancer research. However, few studies have focused on whether recurrent mutations are transcribed from DNA to RNA and whether the observed transcripts are proportional to the DNA VAF of those mutations.[48,57,58] One reason for this omission is the difficulty of identifying true genomic variants in RNA-seq data, mostly due to the large number of false positive variants, which is why DNA-seq is still considered the gold standard for variant discovery.[59] In this study, we established a pipeline that enables the comparison of DNA-seq and RNA-seq variant calling results to determine allelic imbalance using mutations in recurrently mutated AML genes.

For our analysis we considered 36 genes recurrently mutated in our dataset with a VAF of >1%. Out of those, 11 met our filtering criteria and were considered for further analysis. Variants were categorized to DNA-exclusive, RNA-exclusive or transcribed variants (when detected in both DNA- and RNA-seq). As expected due to sequencing errors, RNA-exclusive variants were relatively high (47.9% of all variants), which necessitated strict filtering criteria to remove false positives. For this purpose we calculated the proportion of heterozygous to homozygous mutations, which stabilized at a cut-off of around 10x. Additional filtering criteria further reduced false positives. After application of the filtering criteria, most variants detected in DNA-seq could also be detected in RNA-seq (95.4%). Of these around 5.3% were overrepresented in RNA-seq, while 9.9% could not be detected in RNA-seq.

The large differences in coverage between targeted DNA-seq and whole-transcriptome RNA-seq, which are common in these approaches require a workflow that addresses this issue. We solved this by defining a “weighted” AI based on the differences between expected and observed mutant allele reads. Using weighted AI in a regression model we showed imbalance towards wild-type transcript abundance in *CEBPA*, *PTPN11* and *WT1*. In contrast, in *GATA2*, *IDH2*, *NPM1, RUNX1*, *SRSF2* and *TET2* an AI towards the mutant allele was observed (Figure 6). An analysis of SNPs in these genes did not identify any AI, suggesting that AI is specific to recurrent mutations in these genes. Using pooled sample data from three additional cohorts (DKTK, TCGA and HELSINKI cohorts) we were able to validate an AI towards the mutant-allele for *GATA2*. Differential isoform expression analysis was performed to check whether preferential expression of the mutant alleles correlated with the differential expression of specific isoforms, which however was not the case.

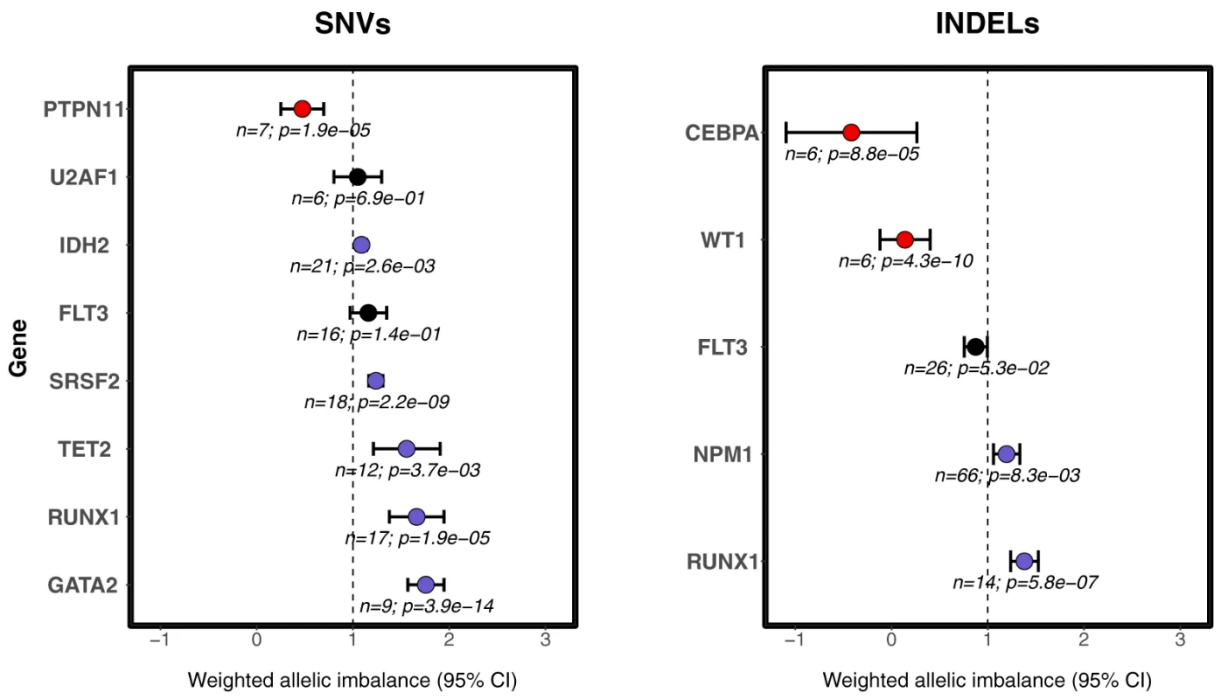


**Figure 6. Weighted allelic imbalance of recurring AML mutations (adapted from Batcha *et al.* 2019[60]).** Weighted allelic imbalance is shown separately for SNVs and INDELs. A value below one signifies an allelic imbalance towards the wildtype allele. Conversely, a value above one denotes an allelic imbalance towards the mutant allele.

In summary, we established a method to compare the VAFs of variants in DNA and RNA and defined several filtering criteria to remove false positive variants while retaining as many likely true variants as possible. Using this method we were able to show AI for 9/11 recurrently mutated genes examined in this study, which correlated with the presence of recurrent mutations, but not with common SNPs present in these recurrently mutated genes. Our analysis suggests that AI is a common phenomenon in AML and further studies are required to reveal its potential implication in leukemogenesis.

Author Contribution:

The author of this thesis contributed to the following elements of *Publication II*:

- Assistance in the study design
- Assistance in the storage and management of all genomic data used in the study, quality assessment of the data
- Design and execution of several steps in the RNA-seq analysis, including quality-trimming, adaptor-clipping, alignment and quality control
- Design of the differential gene and isoform expression pipeline
- Proof-reading and correction on the publication manuscript
- Assistance in the creation of several graphic elements

# ENGLISH SUMMARY

Acute myeloid leukemia is an aggressive malignancy which proves fatal if left untreated. Most patients respond to intensive chemotherapy, however refractory or relapsing disease is still a major contributor of poor patient outcome. New generation sequencing methods enabled the identification of genes harboring recurrent mutations in this disease, and they are being used to inform clinical decisions. In the studies presented in this thesis the aim was to improve our understanding of these mutations to further refine clinical decision making. The first study provided an overview of splicing factor mutations, which affect around 20% of all acute myeloid leukemia patients. It highlighted the association of splicing factor mutations with clinical and molecular parameters and further showed that splicing factor mutations are not independent prognostic markers in acute myeloid leukemia. A novel differential splice junction usage pipeline was used to quantify aberrant splicing patterns in mutated patients in two large sequencing datasets. The usage of two splice junctions was shown to identify patients with poor prognosis thereby providing an example of how our findings can be translated to clinical practice. The purpose of the second study was to examine allelic imbalance of recurrent mutations, a currently underappreciated phenomenon in acute myeloid leukemia. Using a large patient sample pool with matched DNA- and RNA-sequencing data we were able to compare variant calling pipelines between both sequencing methods to determine whether recurrent mutations are over- or underrepresented in RNA. We defined weighted allelic imbalance as a parameter for statistically comparing variant allele frequencies between DNA and RNA and identified allelic imbalance in nine out of eleven recurrently mutated genes examined in this study. Furthermore, recurrent mutations in *GATA2* were also shown to exhibit preferential transcription for the mutant allele in a pooled validation cohort of three independent datasets. In summary, our studies show how customized bioinformatics pipelines can lead to an improved pathomechanistic understanding of recurrent mutations in acute myeloid leukemia and provide a foothold for further study of these mutations in high throughput sequencing experiments.

# GERMAN SUMMARY / DEUTSCHE ZUSAMMENFASSUNG

Die Akute Myeloische Leukämie ist eine aggressive Krebserkrankung die unbehandelt tödlich verläuft. Die Mehrheit der Patienten spricht auf eine intensive Chemotherapie an, jedoch resultieren refraktäre Erkrankungsverläufe oder Rezidive immer noch in einer schlechten Gesamtprognose. Hochdurchsatz-Sequenzierungsverfahren erlaubten die Identifikation von Genen, die in dieser Erkrankung häufig mutiert sind. Diese Mutationen ermöglichen eine Risikostratifizierung der Patienten und fließen in Therapie-Entscheidungen ein. Das Ziel der in dieser Dissertation präsentierten Studien war es, die funktionelle Bedeutung einiger dieser Mutationen genauer zu charakterisieren. Die erste Studie charakterisierte Spliceosom-Mutationen, die bei etwa 20% aller Patienten mit einer Akuten Myeloischen Leukämie beobachtet werden. Die Assoziation von Spliceosom-Mutationen mit klinischen und molekularen Parametern wurde untersucht und zeigte, dass Spliceosom-Mutationen keine unabhängige prognostische Wertigkeit besitzen. Eine neue Analyse-Methode zur Splicing-Quantifizierung wurde zur Untersuchung von aberranten Splicing-Mustern in Patienten mit Mutationen in diesen Genen entwickelt. Diese wurde in der Folge auf zwei große Sequenzierdatensätze angewandt. Zwei der aberranten Splicing-Muster konnten genutzt werden, um Patienten mit einer schlechten Prognose zu identifizieren und stellen damit die klinische Bedeutung der Ergebnisse beispielhaft dar.

Das Ziel der zweiten Studie war es, ein allelisches Ungleichgewicht von häufigen Mutationen zu untersuchen. Mittels eines großen Patientenkollektivs mit gepaarten DNA- und RNA-Sequenzierungsdaten konnten eine Über- oder Unterrepräsentation von häufigen bei AML Patienten beobachteten Mutationen auf RNA-Ebene bestimmt werden. Wir definierten die “weighted allelic imbalance” als einen Parameter für den statistischen Vergleich der Allelfrequenzen von rekurrenten Mutationen in DNA und RNA und stellten ein allelisches Ungleichgewicht in neun von elf untersuchten Gen-Mutationen fest. Weiterhin konnte die bevorzugte Transkription des mutierten Allels von *GATA2* in einer Validierungskohorte, bestehend aus drei unabhängigen Datensätzen, gezeigt werden.

Zusammenfassend, zeigen diese Studien wie maßgeschneiderte bioinformatische Arbeitsabläufe zu einem verbesserten pathomechanistischen Verständnis von rekurrenten Mutationen in der Akuten Myeloischen Leukämie führen können und stellen einen Baustein für die weitere Erforschung solcher Mutationen mit Hilfe von Hochdurchsatz-Experimenten dar.

# PUBLICATION I

Title:

Clinical presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with Acute Myeloid Leukaemia.

Authors:

**Bamopoulos SA**, Batcha AMN, Jurinovic V, Rothenberg-Thurley M, Janke H, Ksienzyk B, Philippou-Massier J, Graf A, Krebs S, Blum H, Schneider S, Konstandin N, Sauerland MC, Görlich D, Berdel WE, Woermann BJ, Bohlander SK, Canzar S, Mansmann U, Hiddemann W, Braess J, Spiekermann K, Metzeler KH and Herold T

Journal:

Leukemia (2020)

**doi: 10.1038/s41375-020-0839-4**

## PUBLICATION II

Title:

Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia

Authors:

Batcha AMN, **Bamopoulos SA**, Kerbs P, Kumar A, Jurinovic V, Rothenberg-Thurley M, Ksienzyk B, Philippou-Massier J, Krebs S, Blum H, Schneider S, Konstandin N, Bohlander SK, Heckman C, Kontro M, Hiddemann W, Spiekermann K, Braess J, Metzeler KH, Greif PA, Mansmann U and Herold T

Journal:

Scientific Reports (2019)

**doi: 10.1038/s41598-019-48167-4**

## REFERENCES


1. Cancer Statistics. https://seer.cancer.gov/statistics/ (accessed May 3, 2020).
2. Guan YF, Li GR, Wang RJ, et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. *Chinese Journal of Cancer* 2012;31(10):463–470.
3. Papaemmanuil E, Gerstung M, Malcovati L, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood 2013;122(22):3616–3627.
4. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med 2016;374(23):2209–2221.
5. Al-Anazi KA. Update on Non-M3 Acute Myeloid Leukemia — Etiology, Classification, Risk Stratification, Emergencies, Complications, Disease in Special Circumstances and Current and Future Therapeutics. In: Leukemias - Updates and New Insights. InTech; 2015. p.
6. Nath R, Reddy V, Kapur A, et al. Survival of Relapsed/Refractory Acute Myeloid Leukemia (R/R AML) Patients Receiving Stem Cell Transplantation (SCT). Biol Blood Marrow Transplant 2019;25(3):S125.
7. Konstandin NP, Pastore F, Herold T, et al. Genetic heterogeneity of cytogenetically normal AML with mutations of *CEBPA*. Blood Adv 2018;2(20):2724–2731.
8. Larsson CA, Cote G, Quintás-Cardama A. The changing mutational landscape of acute myeloid leukemia and myelodysplastic syndrome. Mol Cancer Res 2013;11(8):815–27.
9. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018;562(7728):526–531.
10. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* 2017;129(4):424–447.
11. Escobar-Hoyos L, Knorr K, Abdel-Wahab O. Aberrant RNA Splicing in Cancer. Annu Rev Cancer Biol 2019;3(1):167–185.
12. Crick F. Central dogma of molecular biology. Nature 1970;227(5258):561–563.

13. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA;8(1):.
14. Wahl MC, Will CL, Lührmann R, et al. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell 2009;136(4):701–718.
15. Ohi MD. Structural and functional analyses of the spliceosome requires a multi-disciplinary approach. *Methods* 2017;1251–2.
16. Mercer TR, Clark MB, Andersen SB, et al. Genome-wide discovery of human splicing branchpoints. Genome Res 2015;25(2):290–303.
17. Gabut M, Samavarchi-Tehrani P, Wang X, et al. An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. Cell 2011;147(1):132–146.
18. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456(7221):470–476.
19. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature 2010;463(7280):457–463.
20. Boise LH, González-García M, Postema CE, et al. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 1993;74(4):597–608.
21. Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev E V. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev 2012;26(11):1209–1223.
22. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci 2003;100(1):189–192.
23. Buckley PT, Lee MT, Sul J-Y, et al. Cytoplasmic Intron Sequence-Retaining Transcripts Can Be Dendritically Targeted via ID Element Retrotransposons. Neuron 2011;69(5):877–884.
24. Freitag J, Ast J, Bölker M. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. Nature 2012;485(7399):522–525.
25. Wang Y, Liu J, Huang BO, et al. Mechanism of alternative splicing and its regulation. Biomed reports 2015;3(2):152–158.
26. Fackenthal JD, Godley LA. Aberrant RNA splicing and its functional consequences in cancer cells. Dis Model Mech 2008;1(1):37–42.
27. Baker KE, Parker R. Nonsense-mediated mRNA decay: terminating erroneous

gene expression. Curr Opin Cell Biol 2004;16(3):293–299.
28. Bejar R. Splicing factor mutations in cancer. In: Advances in Experimental Medicine and Biology. Springer New York LLC; 2016. p215–228.
29. Voshall A, Moriyama EN. Next-Generation Transcriptome Assembly: Strategies and Performance Analaysis. In: Bioinformatics in the Era of Post Genomics and Big Data. InTech; 2018. p.
30. Eberwine J, Sul J-Y, Bartfai T, Kim J. The promise of single-cell sequencing. Nat Methods 2014;11(1):25–27.
31. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology* 2015;16(11):651–664.
32. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet 2010;11(1):31–46.
33. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: Tone down the bias. Exp Cell Res 2014;322(1):12–20.
34. Dhaliwal A. DNA Extraction and Purification. Mater Methods;3.
35. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One 2012;7(3):e33865.
36. Johnson M. Kits for RNA Extraction, Isolation, and Purification. Mater Methods;2.
37. Nam DK, Lee S, Zhou G, et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc Natl Acad Sci 2002;99(9):6152–6156.
38. Adiconis X, Borges-Rivera D, Satija R, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods 2013;10(7):623–629.
39. Sultan M, Amstislavskiy V, Risch T, et al. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics 2014;15(1):675.
40. He S, Wurtzel O, Singh K, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods 2010;7(10):807–812.
41. Zhao W, He X, Hoadley KA, Parker JS, Hayes D, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray

for expression profiling. BMC Genomics 2014;15(1):419.
42. Poptsova MS, Il'icheva IA, Nechipurenko DY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep 2014;44532.
43. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One 2011;6(11):e28240.
44. Nakazato T, Ohta T, Bono H. Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. PLoS One 2013;8(10):e77910.
45. Ross MG, Russ C, Costello M, et al. Characterizing and measuring bias in sequence data. Genome Biol 2013;14(5):R51.
46. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8(3):186–94.
47. O'Brien TD, Jia P, Xia J, et al. Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. Methods 2015;83118–127.
48. Rhee J-K, Lee S, Park W-Y, Kim Y-H, Kim T-M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. Sci Rep 2017;7(1):1653.
49. Seiler M, Peng S, Agrawal AA, et al. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell Rep 2018;23(1):282-296.e4.
50. Thol F, Kade S, Schlarmann C, et al. Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes. Blood 2012;119(15):3578–3584.
51. Metzeler KH, Herold T, Rothenberg-Thurley M, et al. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. Blood 2016;128(5):686–98.
52. Pellagatti A, Armstrong RN, Steeples V, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: Dysregulated genes/pathways and clinical associations. Blood 2018;132(12):1225–1240.
53. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med 2011;365(15):1384–95.
54. Shiozawa Y, Malcovati L, Gallì A, et al. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. Nat

Commun 2018;9(1):3649.
55. Hou H-A, Liu C-Y, Kuo Y-Y, et al. Splicing factor mutations predict poor prognosis in patients with de novo acute myeloid leukemia. Oncotarget;7(8):.
56. Bamopoulos SA, Batcha AMN, Jurinovic V, et al. Clinical presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with acute myeloid leukemia. Leukemia 2020;1–14.
57. Castle JC, Loewer M, Boegel S, et al. Mutated tumor alleles are expressed according to their DNA frequency. Sci Rep 2014;4(1):1–6.
58. Ley TJ, Miller C, Ding L, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 2013;368(22):2059–2074.
59. Sun Z, Bhagwate A, Prodduturi N, Yang P, Kocher J-PA. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Brief Bioinform 2017;18(6):973–983.
60. Batcha AMN, Bamopoulos SA, Kerbs P, et al. Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. Sci Rep 2019;9(1):1–11.

# ACKNOWLEDGEMENTS

First and foremost I would like to express my deepest gratitude to my supervisor PD Dr. med Tobias Herold for his continuous support throughout the duration of this thesis. Without his help, guidance and unwavering optimism this project would not have been realized. I would also like to thank PhD Aarif Mohammed Nazeer Batcha with whom I worked closely during my thesis both as a colleague and as a friend.

I would like to pay my special regards to all members of the Medical Clinic III and the Institute for Medical Information Processing, Biometry, and Epidemiology (IBE) that enabled the completion of this thesis and all associated scientific work, with a special gratitude to Prof. Ulrich Mansmann, Prof. Karsten Spiekermann, Prof. Wolfgang Hiddemann and Prof. Michael von Bergwelt for providing me with the facilities and means to carry out my research.

Lastly, I would like to thank my parents and brother for giving me the opportunity to write and the encouragement needed to complete this thesis.