

Aus der Medizinischen Klinik und Poliklinik III

der Ludwig-Maximilians-Universität München

Direktor: Prof. Dr. med. Dr. rer. nat. Michael von Bergwelt

**Pilot Analysis of
Single Nucleotide Polymorphisms
in Patients with
Acute Myeloid Leukaemia
identified by Targeted DNA Sequencing**

Dissertation zum Erwerb des Doktorgrades der Medizin

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

vorgelegt von

Nele Buckup

aus Hagen

2022

**Mit Genehmigung der Medizinischen Fakultät
der Universität München**

Berichterstatter: Prof. Dr. med. Tobias Herold

Mitberichterstatter: Prof. Dr. med. Christian Reis

Mitbetreuung durch den
promovierten Mitarbeiter: Dr. Aarif Mohamed Nazeer Batcha

Dekan: Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 24.11.2022

When you can't control the wind, adjust your sails.

Sailor's saying.

TABLE OF CONTENTS

SHORTCUT REGISTER

ABSTRACT

INTRODUCTION	1
Acute Myeloid Leukaemia	1
Epidemiology and Pathogenesis	1
Clinical Appearance	1
Aetiology and Risk Factors	2
Diagnosis	3
WHO Classification	3
Prognosis and Development History of Risk Prognosing Models	4
Therapy	6
Single Nucleotide Polymorphisms	7
Definition, Types and their Role in Genes	7
SNPs as potential Biomarkers in AML and Databases announcing them	9
THE AIM OF OUR STUDY	11
PATIENTS AND METHODS	12
Study Design	12
Patient Cohort, Clinical Characterisation and Data Preparation	12
Patient Selection	12
AMLCG-Cohort	13
AMLSG-Cohort	14
BEAT-AML Cohort	14
Clinical and Laboratory Patient Data	15
Data Extraction and Sequencing	16
Mapping to a Reference Genome	17
Processing	17
Workflow of the explorative Association Analysis	18
Pileup and Variant Calling	19
SNP Annotation	20
Filtering	21
Quality Control and Explorative Association Analysis	22
Explorative Results' Validation	26
Validation Analyses of previously published clinically relevant SNPs	26

RESULTS	33
Data Comparison	33
Explorative Results	33
SNPs associated with Clinical Characteristics	35
SNPs associated with Gene Mutations	38
Validation Analyses of previously published clinically relevant SNPs	40
DISCUSSION	44
Patients	44
Methods	44
Results	46
Explorative Analysis	46
Validation Analyses of previously published clinically relevant SNPs	49
Clinical Relevance	51
PERSPECTIVE	53
SUMMARY	54
ZUSAMMENFASSUNG	55
LIST OF TABLES	57
LIST OF FIGURES	57
GLOSSARY	58
REFERENCES	60
ACKNOWLEDGEMENT	
PUBLIKATIONSLISTE	
AFFIDAVIT	

SHORTCUT REGISTER

ALL	Acute Lymphoblastic Leukaemia
AML	Acute Myeloid Leukaemia
AMLCG	German AML Cooperative Group
AMLSG	German-Austrian AML Study Group
APL	Acute Promyelocytic Leukaemia
ATRA	All-Trans Retinoic Acid
BAM	Binary Alignment Map
BCL	B-Cell Lymphoma
BH	Benjamini-Hochberg
BM	Bone Marrow
BWA	Burrows-Wheeler Alignment
CR	Complete Remission
cSNP	coding Single Nucleotide Polymorphism
DFS	Disease-Free Survival
DNA	Deoxyribonucleic Acid
ECOG	Eastern Cooperative Oncology Group
ED	Early Death
EFS	Event-Free Survival
ELN	European Leukemia Net
FISH	Fluorescence In Situ Hybridisation
FLT	Fms-Related Tyrosine Kinase
GIF	Genomic Inflation Factor
GWAS	Genome-Wide Association Studies
HLA	Human Leukocyte Antigen
ICE	Idarubicin, Cytarabine, Etoposide
ID	Identification Number
InDel	Insertion and Deletion
iSNP	intergenic Single Nucleotide Polymorphism
ITD	Internal Tandem Duplication
LCL	Lymphoblastoid Cell Lines
MA	Minor Allele
MAF	Minor Allele Frequency

MDS	Myelodysplastic Syndrome
MIR	Micro Ribonucleic Acid
mRNA	messenger Ribonucleic Acid
MRC	Medical Research Council
N	Number
n.a.	not available
NGS	Next Generation Sequencing
NK	Normal Karyotype
OR	Odds Ratio
OS	Overall Survival
pAML	primary AML, de novo AML
Pat	Patients
PFS	Progression-Free Survival
pSNP	perigenic Single Nucleotide Polymorphism
PCR	Polymerase Chain Reaction
PTD	Partial Tandem Duplication
RD	Resistant Disease
RFS	Relapse-Free Survival
RNA	Ribonucleic Acid
rs	reference SNP cluster ID
SAM	Sequence Alignment Map
sAML	secondary Acute Myeloid Leukaemia
SCT	Stem Cell Transplantation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
tAML	therapy-related Acute Myeloid Leukaemia
tMDS	therapy-related Myelodysplastic Syndrome
VA	Variant Allele
VCF	Variant Calling File
WHO	World Health Organization

ABSTRACT

Studies of prognostic or predictive markers in cancer diseases often focus on somatic mutations and cytogenetic abnormalities. Concomitant Single Nucleotide Polymorphisms (SNPs) are often understood as additional influencing factors for the development and outcome of multiple diseases. Despite great efforts in recent studies, the role of SNPs in Acute Myeloid Leukaemia (AML) remains unclear. This study aimed to provide a feasibility and pilot analysis of SNPs in recurrent mutant genes among AML patients.

Sequenced deoxyribonucleic acid (DNA) data from 2,678 Northern European AML patients homogenously treated in trials of the German AML Cooperative Group (AMLCG) and the German-Austrian AML Study Group (AMLSG) were analysed. A total of 114,004 variants have been included in the targeted sequencing panels and was available for the association analysis.

First, we focussed on the interplay of SNPs and clinical prediction parameters as well as SNPs and recurrent gene mutations. Novel predictive SNPs were identified but could not be successfully validated in an independent cohort.

Second, we intended to validate SNPs from formerly conducted smaller-scale studies that were supposed to associate with outcome of AML patients. These prognostic factors could not be validated in our large data set.

In summary, our analysis showed the feasibility to use targeted sequencing panels for SNP analysis in AML. The preliminary results showed SNP associations with prognostic power and invalidates SNP associations published in previous studies.

INTRODUCTION

The introduction will explain the fundamentals of Acute Myeloid Leukaemia (AML) and provides an overview on Single Nucleotide Polymorphisms (SNPs) in the human genome.

Acute Myeloid Leukaemia

Epidemiology and Pathogenesis

With an approximate incidence of 3.7 per 100,000 people diagnosed per year, AML comprises the largest group of acute Leukaemia in European adults.¹ There is a peak of incidence at a very young age, followed by a constant rise while ageing.² The median age of diagnosis is documented to be around 69 years.^{3,4}

Healthy, functional blood cells arise by division and quiescence of pluripotent haematopoietic stem cells in the bone marrow. The underlying pathophysiology of AML is an uninhibited clonal multiplication of these haematopoietic stem cells as well as haematopoietic precursor cells. At the same time the healthy blood cell formation becomes gradually repressed. This process occurs due to chromosomal translocations or other genetic and epigenetic abnormalities that alter the activity of genes responsible for regulating the complex procedure of healthy blood cell formation.⁵⁻⁸

Clinical Appearance

The patient's clinical picture is mainly stamped by the non-existence of mature blood cells. Beside the suppressed healthy white blood cell formation, also the development of erythrocytes and thrombocytes decreases. With the decreasing number of healthy cells in the periphery patients become granulocytopenic, anaemic and thrombocytopenic. The hereby arising symptoms tend to be unspecific. Granulocytopenia leads to an increased number of infections, particularly lung and throat infections as well as systemic mycoses and sepsis. Fatigue, pallor and decreased individual body performance is explained by anaemia. Thrombocytopenia leads to bleedings and petechiae, ecchymoses, menorrhagia or epistaxis.⁹ Simultaneously the rising number of abnormal myeloblasts leads to their accumulation

in the bone marrow, blood and sometimes even in spleen and liver. In 60% of the cases, leukocytosis is found but mainly branded by non-functioning leukocytes. Leukocytosis >100.000 leucocytes/ μ l increases the risk of leukostasis and can lead to hypoxia, neurological symptoms and retinal haemorrhages.¹⁰ However, some AML patients are aleukaemic. Their blood contains a regular or even reduced number of leukocytes. In clinical examinations, some patients suffer from spleno- and hepatomegaly as well as infiltrations of leukaemic cells in gingiva and skin.^{9,10}

Aetiology and Risk Factors

Depending on the origin and prehistory, different types of AML can be diagnosed. Most common is *de novo* AML (*primary* AML, pAML) occurring without any history of previous cancer or haematological disease. If AML develops from a myelodysplastic syndrome (MDS) or other haematologic disorders, it is called *secondary* AML (sAML).¹¹ AML that develops after previous cancer treatment is referred to as *therapy-related* AML (tAML).¹²

Despite many possible risk factors, most pAML patients develop the disease without any identifiable risk factor. A distinction can be established between external and internal risk factors. Operating from extern, exposure to various substances, like ionising radiation, benzene, soot and coal dust as well as the inhalation of tobacco were reviewed.¹³⁻¹⁶ Further, internal factors as genetics play a large role in AML development. Patients with Trisomy 21, Li Fraumeni Syndrome, Fanconi Anaemia or Dyskeratosis Congenita show increased incidences of AML.^{17,18} Suffering from these congenital syndromes often coincidences with germline or somatic mutations that are involved in haematopoiesis and leukaemogenesis. Among others, predisposing mutations are found in the following genes: *GATA1*, *GATA2*, *RUNX1*, *ANKRD26*, *TP53*, *BRCA1*, *BRCA2*, *CEBPA* and *DDX41*. These genes have in common that their function is involved in healthy haematopoiesis or in the control of apoptosis of non-functioning cells.¹⁸⁻²²

Due to progression and possible transformation into sAML, diseases of the haematopoietic system are also internal predisposing factors. MDS, myeloproliferative neoplasia or paroxysmal haemoglobinuria were described in this context.²³

Exposure to chemotherapeutic cytostatic drugs (as alkylating agents and topoisomerase-II-inhibitors) due to a prior tumour suffering increase the risk of tAML.

The approximate latency period between exposure to these drugs and tAML lies between nine months and five years. Prior to tAML, patients often develop a tMDS.^{24,25}

Diagnosis

The diagnosis of AML is established based on the criteria of the World Health Organization (WHO).⁷ To be diagnosed with AML, the myeloid blast percentage in a patient's blood or bone marrow must be ≥ 20 . These blasts can be myeloblasts, monoblasts, or megakaryoblasts, serving as the precursors of granulocytic and agranulocytic leukocytes as well as macrophages and thrombocytes in functional bone marrow. Nonetheless, the diagnosis is also set with the blast percentage being < 20 in the case of some typical chromosomal aberrations (*t(8;21)*, *inv(16)*, *t(15;17)* or *t(16;16)*) being diagnosed.^{7,26}

A number of tests are routinely performed, outlined below.

- *Blood and Bone Marrow Smear and Blood Count*: Under the microscope, peripheral blood and bone marrow smears are analysed for appearance and morphology of their blood cells.²⁷
- *Genetic Analyses and Immunophenotyping*: Chromosomal tests and cytogenetic classification into subgroups are consulted as genetic markers and screened in AML patients to determine the therapy intensity and the individual prognosis (compare *table 1*). Immunophenotyping provides information on cell surface antigens which are used in AML classification. Samples are treated with antibodies and then analysed under the microscope (immunohistochemistry) or with a flow cytometer instrument.^{27,28}
- *Fluorescence in situ Hybridisation (FISH)* and *Polymerase Chain Reaction (PCR)*, as well as *Next Generation Sequencing (NGS)* can be performed in order to screen for further genetic alterations.^{27,29,30}

WHO Classification

Owing to the improved understandings of the molecular and pathogenetic backgrounds of AML, the WHO updated the *Classification of Tumours of Haematopoietic and Lymphoid Tissues*⁷ in 2016. This division is the currently clinical used classification system and allows to classify most of the AML patients based on cytogenetical and

moleculargenetical criteria. The condensed extract of the 2016 WHO classification concerning AML mainly categorises the following six groups:

- AML with recurrent genetic abnormalities
- AML with myelodysplasia-related changes
- Therapy-related myeloid neoplasms
- AML, not otherwise specified
- Myeloid sarcoma
- Myeloid proliferations related to Trisomy 21.

The more detailed classification can be consulted in the original paper of Daniel A. Arber, *Blood* (2016).⁷

Prognosis and Development History of Risk Prognosing Models

Even though the prognosis of AML patients has steadily improved in recent years, survival rates remain poor. 35 to 40% of AML patients <60 years can be cured whereas this holds true for only 5 to 15% of the patients >60 years with intensive therapy. Patients who cannot obtain intensive therapy due to advanced age or comorbidities have a median *Overall Survival* (OS)^a of five to ten months.³¹ Considering the fact that the average age of AML onset is about 69 years, the survival prognosis for most patients is detrimental.

Prognosis estimation before therapy start is essential in order to individually assign the best type of therapy, either curative or palliative. The most appropriate type of postremission treatment such as chemotherapy or Stem Cell Transplantation (SCT) also depends on the prognostic and predictive grading. *Prognostic* markers are indicators that estimate the outcome of a disease like the OS. Factors that outlook the chance of response or toxicity of an administered therapy, for instance, the chance of receiving *Complete Remission*^b (CR), are called *predictive*.^{27,32,33} These factors are

^a Overall Survival: According to Metzeler et al., *Blood* (2016) “measured from the date of study entry until the date of death”⁵⁷

^b Complete Remission: According to Metzeler et al., *Blood* (2016) “bone marrow aspirate with cellularity greater than 20% and maturation of all cell lines, less than 5% blasts and no Auer rods; and in the peripheral blood, an absolute neutrophil count of $\geq 1,500/\mu\text{L}$, platelet count of $\geq 100,000/\mu\text{L}$, and no leukemic blasts; and no evidence of extramedullary Leukaemia, all of which have to persist for at least 1 month.”⁵⁷

either patient- or disease-related. Among patient-related factors, age in particular has a strong independent prognostic influence with increasing age resulting in worse patient's outcome.³³ Other common patient-related factors are age-related specific genetic variations with an increased risk of therapy resistance. Further, coexisting diseases and poor performance status strongly lower the chances of survival. Prior MDS and prior cytotoxic treatment also influence patient's outcome.^{27,34} Genetics as a disease-related prognostic factor has become increasingly important in recent decades. Some authors even hold genetical variations accountable for two-thirds of the observed variations in *OS* and *Relapse-Free Survival*^c (RFS) of AML patients.²⁷

Prognostic and predictive classification systems have grown for more than 20 years and are still developing. These systems mainly focus on the outlook power of typical genetical combinations in AML patients. The history of involving genetics into the routine of AML diagnostics started in 1990, when expert groups released the first diagnosis and treatment recommendations based on cytogenetic findings. According to Cheson et al., *Clinical Journal of Oncology* (1990), they divided AML into the four subgroups of *Undifferentiated Acute Leukaemia*, *Mixed Lineage Leukaemia*, *Hypocellular AML* and *AML with lacking definitive myeloid Differentiation by Morphology or conventional Cytochemistry but with ultrastructural or immunophenotypic Evidence for AML*.³⁵ From this time onwards, many new criteria concerning the classification, prognostic and predictive grouping have followed. Chromosomal aberrations and abnormal karyotypes were observed, and finally, genes involved in the progress of AML were identified.³⁶ In 2003, the criteria from 1990 were revised, and new research findings were added.³⁷ In 2010, the ELN released a molecular genetic based risk stratification model. Primal genome-wide studies were conducted, and first AML-related SNPs were identified. The *ELN2010* risk stratification sorted AML patients into 4 groups according to their karyotype and molecular genetics.²⁶ The *ELN2010* stratification was revised in 2017 following new findings in the genomic landscape of AML.²⁷ The *ELN2017* risk stratification, nowadays used in clinics, is mainly based on cytogenetic and genetic markers. It provides recommendations and scores regarding the individual therapy of AML patients based

^c Relapse-Free Survival: According to Metzeler et al., *Blood* (2016) "measured from the date of CR until the date of relapse or death."⁵⁷

on his or her risk profile. *Table 1* shows the characterisation criteria of the *ELN2017* risk classification.²⁷

Favourable	t(8;21)(q22;q22.1); RUNX1-RUNX1T1
	inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBFB-MYH11
	Mutated NPM1 without FLT3-ITD or with FLT3-ITD ^{low}
	Biallelic mutated CEBPA
Intermediate	Mutated NPM1 and FLT3-ITD ^{high}
	Wild-type NPM1 without FLT3-ITD or with FLT3-ITD ^{low} (without adverse-risk genetic lesions)
	t(9;11)(p21.3;q23.3); MLLT3-KMT2A
	Cytogenetic abnormalities not classified as favourable or adverse
Adverse	t(6;9)(p23;q34.1); DEK-NUP214
	t(v;11q23.3); KMT2A rearranged
	t(9;22)(q34.1;q11.2); BCR-ABL1
	inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2,MECOM(EV11)
	-5 or del(5q); -7; -17/abn(17p)
	Complex karyotype, monosomal karyotype
	Wild-type NPM1 and FLT3-ITD ^{high}
	Mutated RUNX1
	Mutated ASXL1
	Mutated TP53

Table 1. ELN2017 Cytogenetic Risk Categories, adopted from Döhner et al., Blood (2017).²⁷

Therapy

While new facts about the pathogenesis and molecular fundamentals of AML have been identified, the major therapy design has changed only marginally since about 40 years.³⁸ The latter remains to take into account the patient's age, general health status and co-diseases alongside of the *ELN2017* recommendations deciding whether a patient can receive intensive therapy. This therapy design is complex with

chemotherapy constituting its backbone.²⁷ The essential parts of the therapy are summarised below.

Induction therapy starts as soon as possible after AML diagnosis and aims to achieve *CR*. The standard induction therapy is the *7+3 regimen*: Seven days of continuous Cytarabine infusion together with three days of an Anthracycline. Patients who do not reach *CR* after a second induction cycle are classified as primary refractory. Young patients up to 60 years achieve *CR* in 60-80%. Patients >60 years achieve *CR* in 40-60%.²⁷

Postremission therapy and stem cell transplantation intend to avoid quick relapses and clear from remaining leukaemia cells. Differences in the patient's age, comorbidities, cytogenetic profile of the myeloid blasts and the availability of a matching SCT donor determine the therapy options. Usually, patients receive either high doses of Cytarabine or multiagent chemotherapy in up to 4 cycles.²⁷ In the case of an adverse risk group according to *ELN2017*, SCT is advised. Intermediate risk patients with a suitable donor are also considered to receive SCT. SCT is mostly executed during the first *CR* and is usually allogenic. Allogenic SCTs are given by a human leukocyte antigen- (HLA) compatible donor, a cord-blood- or haplo-identical donor. In rare cases, autologous SCT from the patient's own BM is administered.^{27,39}

During the maintenance therapy low doses of chemotherapy, hypomethylating agents, immune therapy, B-cell Lymphoma 2 (BCL-2) or Fms-related tyrosine kinase 3 (FLT3) inhibitors are applied for few years.⁴⁰

Single Nucleotide Polymorphisms

Definition, Types and their Role in Genes

While 99.9% of DNA is shared between all humans, the remaining 0.1% is mainly covered by SNPs.³³ SNPs are single base pair positions in the DNA at which place differing nucleotides can be found when comparing different individuals. These nucleotide alterations are called *alleles* and count as the most differing variation in our genome. SNPs are mainly detected in *Genome-Wide-Association Studies* (GWAS) and the number of newly discovered SNPs continues growing. The frequency of the least frequent allele must be equal or higher one percent to be called a SNP.⁴¹⁻⁴³ Variations that occur with unknown frequency or in less than one percent of the

population are referred to as *Single Nucleotide Variations* (SNVs).^{44,45} Most often, SNPs are referred to as biallelic markers. However, sometimes they can be tri- or even tetraallelic, describing the number of different alleles that are found in various individuals at a fixed position.⁴¹ On average, there is one SNP found in about 300 to 2,000 nucleotide-bases. SNPs are inheritedly stable and mutate quite little. The latter makes them an excellent foundation for research projects.^{46,47} The allele carried by most people at that genomic position is called *Major Allele* or *Reference Allele*. *Minor Alleles* or *Variant Alleles* are those which appear in the minority of people. Individuals can carry SNPs heterozygously or homozygously. Heterozygous carriers carry two different alleles at a particular locus, and homozygous carriers carry twice the same allele. Figure 1 illustrates the options in which SNPs can be carried. SNPs are further subdivided based on their genomic location. Located in coding regions of a gene they are called *coding* SNPs (cSNP), in between genes, they are referred to as *intergenic* SNPs (iSNPs) and in noncoding regions as *perigenic* SNPs (pSNPs).^{43,48} If a cSNP causes an amino acid change, it is referred to as *nonsynonymous* or *missense* SNP. cSNPs which do not result in an amino acid exchange are labelled *synonymous* or *silent*. A nonsynonymous SNP can change the structure of a protein and, therefore, it can have direct positive or negative influence onto enzyme activity. This, again, can influence factors like cancer susceptibility or drug effectiveness.⁴³ Nonetheless, SNPs in noncoding regions like introns can also have consequences on genes, for instance, on gene splicing or transcription factor binding.⁴⁹

Genetic factors determine the sensitivity and resistance to diseases, such as disease progress and response to therapies. As SNPs are a common type of genetic variation, they are an essential base for understanding diseases and differences in disease progress and therapy response.⁴⁷

Many research fields make use of the frequent appearance and stable inheritance of SNPs such as pharmacogenetics, evolution and population studies, forensics and, finally, SNPs in cancer research.⁴¹

<u>Individual 1, reference</u>		<u>Individual 3</u>	
CHR.1, copy1	...C G C T A A T...	CHR.1, copy1	...C G A T A A T ...
CHR.1, copy2	...C G C T A A T ...	CHR.1, copy2	...C G A T A A T ...
<u>Individual 2</u>		<u>Individual 4</u>	
CHR.1, copy1	...C G A T A A T ...	CHR.1, copy1	...C G G T A A T ...
CHR.1, copy2	...C G C T A A T ...	CHR.1, copy2	...C G C T A A T ...

Figure 1. SNPs. This DNA sequence equals in all four individuals except at one nucleotide position. Individual 1 represents an example reference sequence found in the majority of people at an unspecific location of chromosome 1. Individual 1 carries the reference allele 'C' at that certain position, printed in orange. Individual 2 also carries the reference allele 'C' at one of its chromosomes 1. On the other copy of chromosome 1, individual 2 carries allele 'A' which is not the reference nucleotide at this position and is hereby a minor allele. In this combination, individual 2 is a heterozygous carrier of the minor allele 'A'. Individual 3 is a homozygous carrier of the minor allele 'A' by carrying it at both chromosome 1 copies. Individual 4 carries a 'G' at the specific position what is also not the reference allele and makes individual 4 also being a heterozygous carrier of a minor allele. Since three different nucleotides can be found at the same position, this SNP is called a triallelic SNP.

SNPs as potential Biomarkers in AML and Databases announcing them

SNPs were identified as prognostic and predictive biomarkers in a wide variety of cancer diseases.³³ In case of AML, a multitude of studies have been conducted during the last years concerning the prognostic role of SNPs. These association studies aim to find genetic variations associated with specific traits. They form the base for identifying relevant SNPs. Open access databases summarising published SNPs and their associations are, for example, *dbSNP*⁵⁰, *SNPedia*⁵¹ and the *GWAS catalog*⁵². Together, these databases (as at may 2021) describe 7,759 SNPs to be associated with AML according to prior studies. The databases provide important information on the SNPs, including the location, the minor and major alleles, the frequency of the minor allele, the SNP type, e.g. synonymous variant, and if clinical influence of the SNP has been reported (e.g. in the *ClinVar*⁵³ database).

Many of these SNP associations arose from GWAS analysis. GWAS trials mainly serve to identify alleles that commonly emerge together with a certain disease. This information is important in order to identify biological variations leading to the disease.

We were particularly interested in the studies that suggest different SNPs to have prognostic or predictive value in AML patients. Some of these SNPs, for instance, seem to be able to prognose the *RFS* period of patients, some SNPs outlook the chance of *CR* achievement or how long *OS* will be. The results of these studies sound promising regarding the further subdivision of the *ELN2017* prognostic groups. Hereby, the treatment is supposed to get further individualised. A range of these studies is described in the chapter *Confirmation of previously identified SNP associations*.

THE AIM OF OUR STUDY

With *ELN2017*, a powerful prognosis prediction model is available that classifies AML patients according to their genetic subtypes and helps to decide about the administered type of therapy. Still, the therapy success remains variable.⁵⁴ Hence, this lets assume that more factors beyond age, health status, and the mutations mentioned in *ELN2017* influence the outcome of AML patients.

In the first part of our study we aimed to examine the sequencing data of two large and homogeneously treated AML cohorts to detect biomarkers that are capable of prognosing the outcome of AML patients. Here, we set the focus on SNPs serving as such biomarkers. We intended to research SNPs with outcome outlook capabilities by associating them, amongst others, to the following characteristics: *OS*, *RFS* and the post-induction-situation in terms of *CR*, *Refractory or Resistant Disease^d (RD)* or death during or soon after induction therapy (*Early Death^e, ED*). Furthermore, we strived to analyse whether there are SNPs which occur mostly together with specific AML-associated gene mutations. We performed all association analyses within the framework of the currently used risk classification system *ELN2017*. This enabled to find associations that have a prognostic relevance independent of any genetic mutations mentioned in the *ELN2017* prognostic grouping.

As demonstrated with the total number of more than 7,750 AML-associated SNPs, the research libraries are already rich in studies analysing the association between SNPs and AML. Nonetheless, the majority of these published associations arose from small analysed groups and has not yet been validated in a second cohort. Hence, serving as the second aim of our study, we intended to validate those 241 out of the more than 7,750 associations which were analysed to prognose or predict the outcome of AML patients in previous studies. We aimed to reproduce these association analyses identical to the respective prior executed study.

^d Refractory/Resistant Disease: According to Döhner et al., Blood (2017) “no CR after 2 courses of intensive induction treatment; excluding patients with death in aplasia or death due to indeterminate cause.”²⁷

^e Early Death: Death within 60 days after the primary diagnosis of AML.

PATIENTS AND METHODS

The following section is going to introduce the analysed patient samples. Afterwards the step by step workflow until the finding of SNP associations is depicted.

Due to the complexity and error susceptibility of SNP association analyses, I performed the statistical part with the help and supervision of Dr. Aarif Nazeer Batcha from the *Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie* (IBE), who is a bioinformatics specialist and has considerable expertise with SNP association projects.

Study Design

This study is a retrospective analysis of AML patients. It includes five trials collected and sequenced by the German AML Cooperative Group (AMLCG) and German-Austrian AML Study Group (AMLSG). Clinical data, as well as sequencing (genotypic) data from the below described cohorts were available for statistical evaluation.

Patient Cohort, Clinical Characterisation and Data Preparation

The genetic material of 2,678 intensively treated AML patients enrolled in the trials of *AMLCG* and *AMLSG* was statistically analysed. All patients had been diagnosed with AML based on the WHO criteria.⁵⁵ The patients were treated with comparable intensive therapy schemes. Moreover, all study protocols were in accordance with the *Declaration of Helsinki*⁵⁶. Vast majority of the patients whose national backgrounds were recorded originated from northern Europe. All study participants provided their written informed consent.^{57,58}

Patient Selection

The exclusion criteria rarely differed between the studies. The most common reasons for exclusion were: children (under 16 years of age), serious previous or concurrent illnesses, pregnancy and lack of consent. The samples involved in the respective trials were then included in our cohort. To increase the comparability of the data, we applied

additional selection criteria like the exclusion of pretreated patients and of patients suffering from the AML subtype *Acute Promyelocytic Leukaemia* (APL) during the statistical work process. A more detailed description of these criteria can be found in the chapter *Workflow of the explorative Association Analysis*.

AMLCG-Cohort

The *AMLCG* cohort involved 1,138 patients from the *AMLCG-1999* and *AMLCG-2008* trials. The age range was from 18 to 86 years.⁵⁹

AMLCG-1999 (clinicaltrials.gov identifier: NCT00266136):

864/1,138 patients were enrolled in the *AMLCG-1999 study*. These patients were recruited between 1999 and 2005. All patients were treated with intensive Cytarabine-based induction chemotherapy. For consolidation therapy, the patients were assigned to further Cytarabine-based therapy, allogenic, or autologous SCT. The assignment was based on the patient's age, the availability of a matching donor, and randomisation. Also, some patients received maintenance therapy according to the study protocol. For more details see Büchner et al., *Clinical Journal of Oncology* (2006)⁶⁰ and *Leukemia* (2016)⁶¹.

AMLCG-2008 (clinicaltrials.gov identifier: NCT01382147):

274/1,138 patients were recruited from *AMLCG-2008* between 2009 and 2012 and included in this study. All patients received intensive Cytarabine-based chemotherapy as induction treatment. Consolidation therapy differed based on the genetic risk profile. Allogenic SCT was offered to all patients except those with favourable genetics and good response to induction chemotherapy. Patients who did not receive SCT obtained further Cytarabine-based treatment as consolidation and maintenance therapy. For more details see Braess et al., *Blood* (2013)⁶².

Additional details of the *AMLCG-1999* and *AMLCG-2008* study protocols can be found in the supplemental appendix of Metzeler et al., *Blood* (2016)⁵⁷.

AMLSG-Cohort

The *AMLSG* cohort encompassed 1,540 AML patients. These were enrolled in three clinical trials. The patients were aged between 18 and 84 years at the time of diagnosis.

AML-HD98B (clinicaltrials.gov identifier: not available):

173/1,540 participants were included from this study. They were aged older than 61 years and enrolled between 1997 and 2003. The patients were randomised to receive Idarubicin, Cytarabine, and Etoposide (ICE) with or without All-Trans Retinoic Acid (ATRA) for induction therapy. The subsequent therapy was determined by the individual response.

AML-HD98A (clinicaltrials.gov identifier: NCT00146120):

This study involved 627/1,540 patients, recruited between 1998 and 2004. They received ICE for induction therapy. Depending on the risk stratification, patients were either treated with allogeneic stem cell transplant (high risk and intermediate risk with matching donors) or intense consolidation chemotherapy for postremission therapy (low risk and intermediate risk without matching donor).

AMLSG-07-04 (clinicaltrials.gov identifier: NCT00151242):

740/1,540 patients were recruited between 2004 and 2011. A similar therapy design as in *AML-HD98A* was offered except that, here, the patients received induction ICE with or without ATRA based on randomisation.

The appendix of Papaemmanuil et al., *NEJM* (2016)⁵⁸ gives detailed insights into all three trials and treatment schemes.

BEAT-AML Cohort

(clinicaltrials.gov identifier: NCT03013998)

The *BEAT-AML* cohort served as validation cohort. This cohort had incorporated 175 adult AML patients (APL excluded) who received intensive standard treatment.⁶³

Clinical and Laboratory Patient Data

Among others, following data were acquired from the patients in the above mentioned studies:

Information	Description	AMLCG	AMLSG
Study samples		AMLCG 1999: N=864 AMLCG 2008: N=274	AMLHD98A: N=627 AMLHD98B: N=173 AMLSG0704: N=740
AML type		pAML*: N=954 (84%) sAML*: N=126 (11%) tAML*: N=58 (5%)	pAML: N=1,376 (91%) sAML: N=61 (4%) tAML: N=68 (5%)
Age (years)	By the Point of Diagnosis	58 (18-86)	50 (18-84)
Sex		Female: N=554 (49%) Male: N=584 (51%)	Female: N=719 (47%) Male: N=821 (53%)
Hemoglobine (g/dl)	By the Point of Diagnosis	9 (3.5-16.0)	9.1 (2.5-17.6)
White Blood Cells (G/l)	By the Point of Diagnosis	20.6 (0.1-798.2)	14.2 (0.2-532.7)
Platelets (G/l)	By the Point of Diagnosis	54 (0-1,760)	53 (2-916)
LDH (U/l)	By the Point of Diagnosis	440 (76-19,624)	435 (83-7,627)
Bone Marrow Blasts (%)**	By the Point of Diagnosis	80 (6-100)	75 (0-100)
ECOG Performance Status	By the Point of Diagnosis 0=fully active without restriction 4=completely disabled, no selfcare	0: N=203 (27%) 1: N=354 (47%) 2: N=147 (20%) 3: N=38 (5%) 4: N=6 (<1%)	0: N=350 (26%) 1: N=846 (62%) 2: N=156 (11%) 3: N=16 (1%) 4: N=2 (<1%)
ELN 2017 Classification		Favourable: N=428 (38%) Intermediate: N=290 (26%) Adverse: N=400 (36%) t_15_17 (APL): N=0	Favourable: N=471 (38%) Intermediate: N=292 (24%) Adverse: N=429 (35%) t_15_17 (APL): N=38 (<1%)
MRC	Prognostic Groups according to the Medical Research Council Criteria	Favourable: N=81 (7%) Intermediate: N=877 (78%) Adverse: N=161 (14%)	Favourable: N=208 (15%) Intermediate: N=960 (68%) Adverse: N=253 (18%)
OS (days)	Days of Survival after Diagnosis until Death of any Cause	513.5 (1-5,023)	773.5 (1-5,384)
Post Induction Therapy	Status after Induction Therapy	Early Death: N=234 (21%) Complete Remission: N=756 (66%)	Early Death: N=123 (8%) Complete Remission: N=1078 (71%)

		Resistant Disease: N=148 (13%)	Resistant Disease: N=326 (21%)
RFS (days)	Days after the point of CR achievement until relapse	509 (9-4,995)	554 (3-5,357)
Relapse	After achievement of Complete Remission	No Relapse: N=226 (34%) Relapse: N=440 (66%)	No Relapse: N=489 (38%) Relapse: N=784 (62%)
Normal Karyotype (NK AML)	Cytogenetical normal karyotype	Normal: N=732 (65%) Abnormal: N=386 (35%)	Normal: N=695 (48%) Abnormal: N=754 (52%)
Complex Karyotype	≥3 Chromosome Abnormalities ⁶⁴	Complex: N=101 (9%) Non-complex: N=1,019 (91%)	Complex: N=159 (11%) Non-complex: N=1,267 (89%)
Mutations	Screening for cytogenetic abnormalities and mutations	No/Yes for each Mutation	No/Yes for each Mutation

Table 2. Clinical and Laboratory Patient Data. This table provides information on all 2,678 samples. If the sum of patients is lower in a row, the remaining data are not available. For all numerical values, the median value is given as well as the ranges in brackets.

*pAML=primary/de novo AML, sAML=secondary AML, tAML=therapy-related AML.

** The diagnosis of AML was centrally reviewed according to WHO 2016 classification. In case of a blast count below 20%, recurrent genetic abnormalities (t(8;21), inv(16)) irrespective of blast percent lead to the diagnosis of AML.

Data Extraction and Sequencing

Suitable bone marrow or peripheral blood was collected from the study participants before the start of induction treatment. Both cohorts used targeted DNA sequencing techniques and focussed on recurrently mutated genes in AML. Although the sequenced regions of interest differed between the cohorts, some genes were covered in both, *AMLCG*⁵⁷ and *AMLSG*⁵⁸ as illustrated in *Figure 2*.

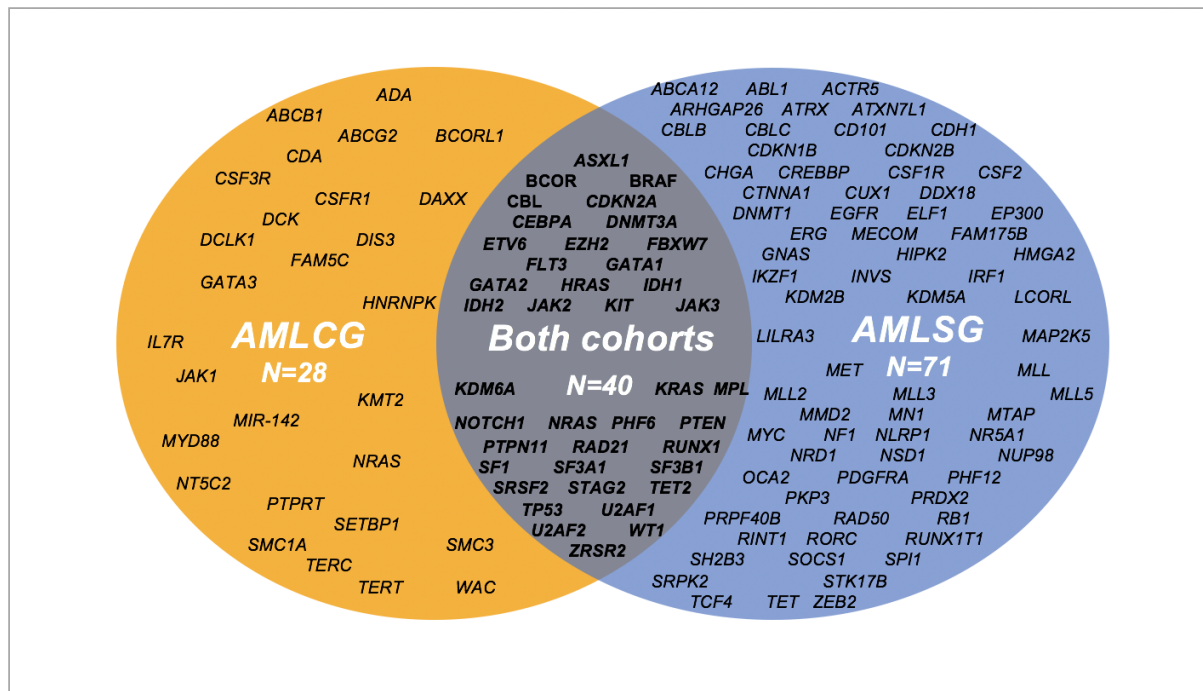


Figure 2. Sequenced Genes. The illustration shows the specific genes that were sequenced in *AMLCG* and *AMLSG* samples. From those sequenced in *AMLCG*, the entire coding region was sequenced for some genes, whereas for other genes, only specific exons were sequenced.⁵⁷ In *AMLSG* samples, all coding exons were sequenced in the demonstrated genes.⁵⁸ In total, 40 genes were sequenced by both cohorts.

Mapping to a Reference Genome

The sequencing data were mapped to the human reference genome. The human reference genome serves as a comparison base for researchers who analyse DNA variations like SNPs. This study used the reference genome from *Genome Reference Consortium human build 37 (GRCh37)*.⁶⁵ Both datasets were mapped with the *Burrows-Wheeler Alignment (BWA)* tool.⁶⁶ The mapped output file contained every sequenced position from every sample, as well as unmapped reads. These data packed in *Sequence Alignment Map (SAM)* and in *Binary Alignment Map (BAM)* formats were subsequently processed.^{57,58}

Processing

Data processing constituted an interim step that served to improve the accuracy of mapping and removed mis-mapped reads. Processing was conducted differently in both cohorts. In *AMLCG*, Insertions and Deletions (InDels) realignment was conducted.⁵⁷ Realignment increased the precision of the mapping results from these

regions. 12 *AML*CG samples had to be removed due to the unavailability of genotype data. During the processing of the *AML*SG dataset, unmapped reads, PCR duplicates as well as regions that were mapped outside of the target regions, were excluded.⁵⁸

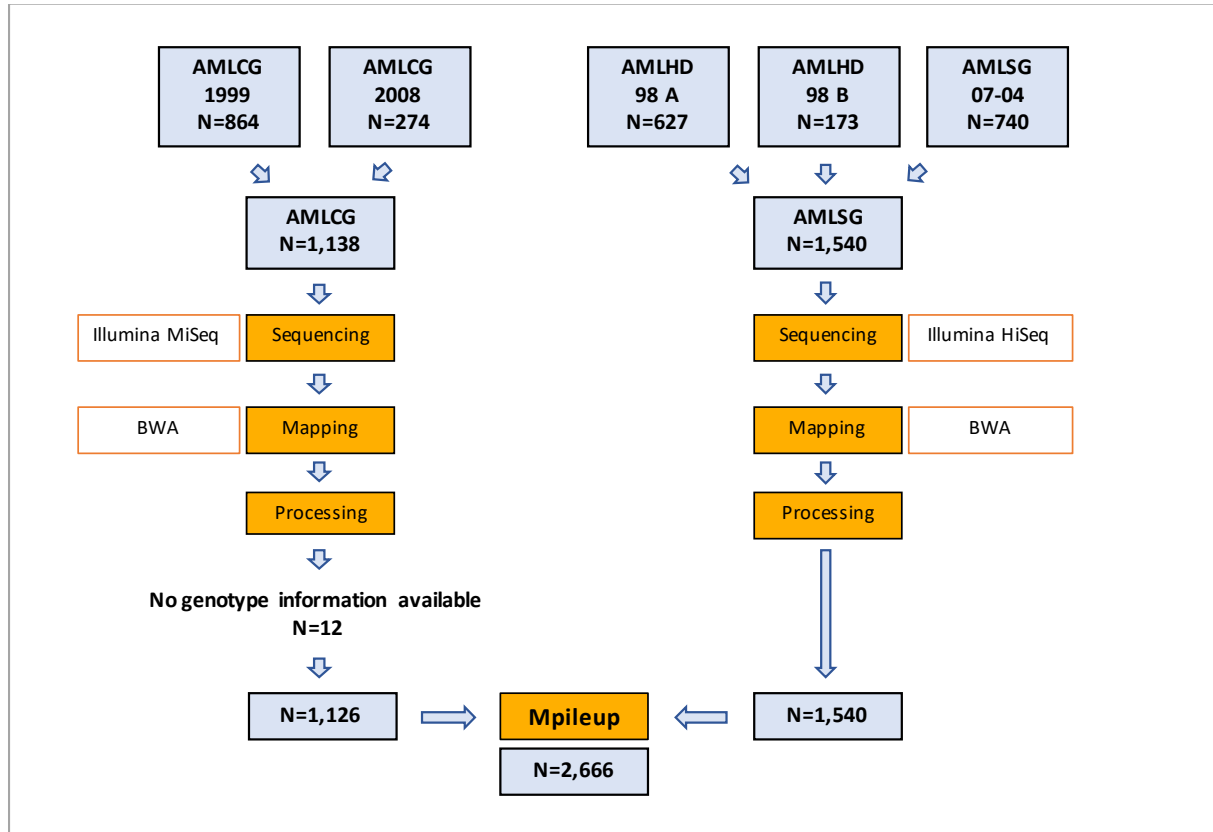


Figure 3. Cohort Composition and Data Preparation. The diagram shows the build-up of the cohort and the data preparation executed before we received the data for the statistical analysis. Detailed information on the prior executed workflow can be received from the papers of Papaemmanuil et al., *NEJM* (2016)⁵⁸ and Metzeler et al., *Blood* (2016)⁵⁷.

Workflow of the explorative Association Analysis

All steps intended to extract the necessary information from the attained data sets, to identify meaningful SNPs and finally link these SNPs to specific outcome parameters described in detail in the following section.

Pileup and Variant Calling

We combined the *AMLCG* and *AMLSG* datasets and executed all following steps as one cohort.

We performed *Multiway pileup (Mpileup)* with *Samtools*⁶⁷. *Mpileup* aimed to visualise the mapped data. The bases of the mapped reads compared to the reference genome got visible. Single positions that differed between the reference genome and the mapped reads were recorded. *Mpileup* incorporated all positions in all samples and summarised the results in *pileup format*. The following information were available for each genomic position:

- Sequence ID
- Sequence position
- Reference base
- Read count (number of reads which covered the position)
- Read result (base which was found in the mapped read)
- Quality of the read

With this information on every mapped position, we performed variant calling with *Varscan2*⁶⁸. Variant calling calculated if specific base changes could be considered as SNVs. As long as the frequencies of these base variations were not evaluated, they cannot be termed as *SNPs*. Also, in variant calling the Minor Allele Frequencies (MAFs) were calculated. If patient one had 150 SNVs that differed from the reference genome, these positions would be recorded. If patient two shared 55 SNV positions with patient one while additionally showing 80 further SNV positions, then, these latter positions were added to the variant file. *Varscan2* added the information on all SNVs from the whole cohort and wrote it onto a *Variant Calling File (VCF)*. This file contained each SNV (together with its pileup information) that was scanned in at least one sample and finally comprised 114,004 different SNVs.

We applied the following parameters to avoid incorrect variant calls.

*Mpileup criteria (adapted to Samtools*⁶⁷*):*

- Reference Genome: *GRCh37*
- Maximal per file read depth: 1,000,000
- Minimum mapping quality for an alignment to be considered: 10

- Minimum base quality for a base to be considered: 20

*Varscan2*⁶⁸ criteria:

- Minimum coverage (minimum depth at a position to make a call): 30
- Minimum supporting reads (minimum number of variant-supporting reads at a position required to make a call): 6
- Minimum base quality at the variant position required to use a read for calling: 20
- Minimum variant allele (minor allele) frequency for calling a variant: 0.01
- Default p-value for calling variants: 0.1
- We ignored variants with >90% support in one strand
- We called only *variants* (positions without changes were skipped)

The overall aim of applying these threshold values was to filter out low-quality reads before the evaluation of variants started.

We conducted the steps of *Mpileup* and variant calling twice in series. During the second approach, each of the 114,004 called SNVs was compared with each corresponding sample from the cohort. This implementation provided the exact information on the individual genotype of each sample and how often each allele appeared in the sample set. To determine each individual's genotype for a certain SNV, we utilised the following frequency thresholds: If a minor allele was found in more than 90 percent of the reads of one sample, this sample was referred to as homozygous minor allele carrier for this specific SNV. If the occurrence of the minor allele was situated between 40 and 60 percent, the sample was a heterozygous carrier, and if the minor allele percentage was less than ten percent within one sample, it was referred to as homozygous reference allele carrier. Samples with genotypes that could not be assigned to one of these groups due to its percentages exceeding the mentioned ranges were excluded for the respective assessed SNV. SNVs had to be found in both strands of a read, the forward and reverse strand. If they were found in only one strand, *Varscan2* filtered them out. This step vastly reduced the number of considered SNVs. The SNVs were presented on a VCF.

SNP Annotation

Annotation of the called SNVs was an indispensable step in the process of linking the SNVs with clinical characteristics. In order to estimate which SNV would be more likely

to influence clinical factors in AML patients, we needed to perform effect prediction. This step was carried out with the *SNPeff*⁶⁹ tool. Every SNV got annotated depending on its location (cSNP/pSNP/iSNP) and type (synonymous/nonsynonymous). The combination of these factors assumed the probable influence of a variation onto a gene. A SNV which does not change the amino acid determined through its codon is less likely to have any clinical influence than a SNV that e.g., leads to the stop gain of an important gene. By changing important functions of a gene, these SNVs could potentially result in longer OS or higher resistance to chemotherapy, for instance^{41,49}.

Incorrectly or incompletely executed annotation could lead to overestimation of variants which do not have significant influence, or to dilution of significant correlated SNVs inside many wrongly assumed influencing SNPs. We displayed the results in VCF format.

Applied annotation criteria:

- Upstream and downstream length: 5,000 bases
- Size for splice sites (donor and acceptor): 2 bases.

Filtering

*GEMINI*⁷⁰ is an online tool that enabled to break down the vast number of called SNVs into those variants which contained all necessary information for the subsequent steps. The previously annotated variants were extracted and served as input. Only SNVs with a correct ID were considered. To filter for the informative SNVs, we instructed *GEMINI* to register only those SNVs whose position was covered in at least 30% of the samples. For all variations that fitted to these designations we commanded *GEMINI* to query, select, and table the following parameters: The *chromosome number*, the *SNV ID*, '0', the *position of the SNV*, the *reference* and *alternative allele* as well as the *MAF*.

Also, within *GEMINI* we transformed the genotype information of our samples. The percentages that determined the genotypes mentioned in variant calling, were converted to characters in order to facilitate grouping them in the following steps of the association analysis. We grouped all homozygous minor allele carriers as *BB*, heterozygous carriers as *AB*, and homozygous reference allele carriers were set to *AA*.

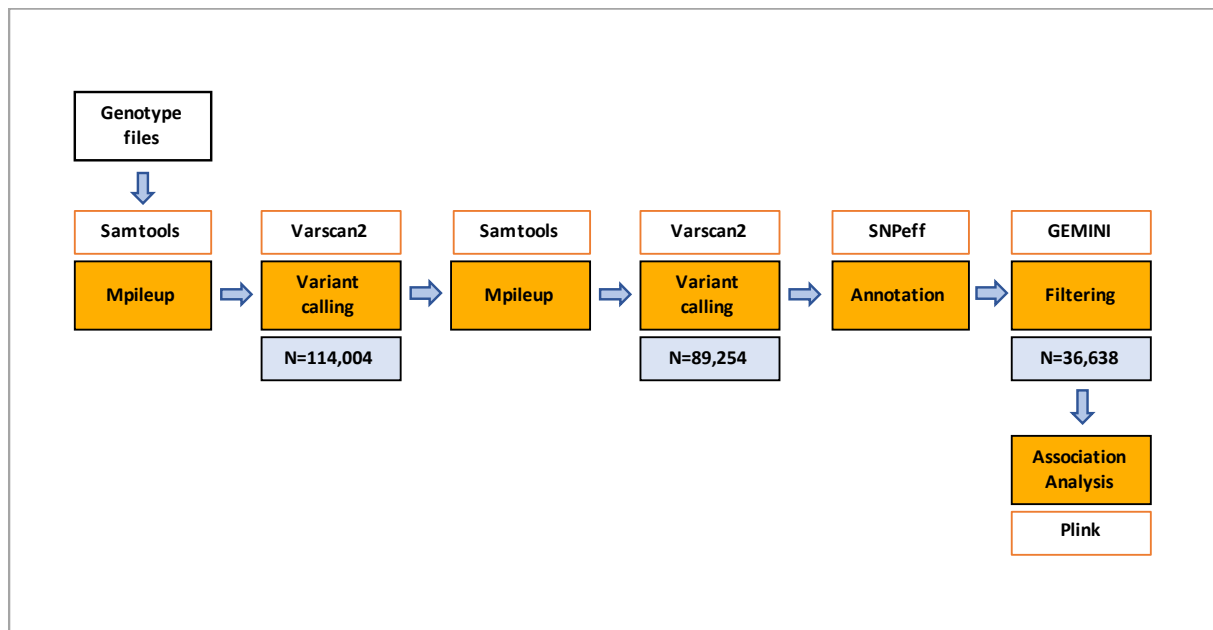


Figure 4. Workflow until Plink. The diagram shows the workflow executed in order to call the variants and prepare the data for the association analysis. The given numbers show the registered SNVs.

Quality Control and Explorative Association Analysis

Plink^{71,72} is a tool for genome association analysis that is capable of analysing genotype and clinical data. We used *Plink* version 2.00 Alpha. The *GEMINI* output file served as input. The desired actions were specified by defining commands. We used *Plink* to perform both, the quality control and the final association analysis. The survival analyses were carried out in *R*.

Quality Control

To minimise biases, false-positive associations and masking of positive association results, we performed SNV and sample input filtering before starting the association analysis. Duplicated SNVs were excluded. We removed patients with the AML subtype *APL* because of their different treatment and better chances of permanent recovery.⁷³ Furthermore, patients that had been pre-treated before study entry were excluded.

Some of the SNVs had a low frequency and appeared only in the out-filtered samples. Hereby, no information was extractable from them anymore. We excluded these uninformative SNVs.

After the step of filtering, we evaluated the *Hardy Weinberg Equilibrium* (HWE) and the *Genomic Inflation Factor* (GIF) for each SNV. Both methods were applied to

visualise population stratification. Stratification would bias the following association analysis and could produce false positive results. None of the SNPs deviated significantly from the HWE. SNPs that showed GIFs $\lambda > 1.04$ got adjusted by dividing the expected test statistics (calculated in null distribution population) by the GIF λ of the SNV. Hereby, the number of false positive results during the later association analysis decreased.^{74,75}

Ensuing, we prepared two *obligatory missing lists* containing all SNVs that were only covered in one, the *AMLCG* or *AMLSG* cohort. All SNVs that were only covered in one cohort were also only investigated in this group during the following association analysis. We further produced a *SNV attribute file* according to the MAFs. For this, we divided the SNVs into three groups. If the minor allele of one variant was found in more than five percent of the covered samples, we have defined it as *common* SNP. If present in one to five percent of the covered samples we have called it a *less common* SNP, and in case of less than one percent prevalence we set it to be a *rare* SNV. We wrote separate files that included the SNVs from each attribute group. The subsequent analyses focussed on *common* SNPs. We also analysed the *less common* and *rare* files but, due to their less frequent occurrence, the power of the results is lower compared to the *common* SNPs.

We divided the clinical patient variables/parameters into two files. One file included the binary variables like *CR* or *ABCB1* mutation, the other file enclosed the quantitative clinical variables like *OS* and *RFS*. The subsequently executed analyses were done on these files separately.

The genetic mutations categorising the *ELN2017* groups *adverse*, *intermediate*, and *favourable* were attached. This was necessary for the multivariate analysis carried out later.

Association analysis

As primal part of the exploratory data analysis, we performed univariate association analysis. Different models were available for the implementation of association analyses. These models differ in terms of the genotypes to be compared. During the explorative analysis, we applied an *allelic* model. This model describes if there is any association found between the presence or absence of the minor allele in a certain SNP. It does not give the information whether the presence of the SNP in a

homozygous manner has a stronger association with the clinical parameter than the presence of the SNP in the heterozygous form.

Model	Genotype/allele	Compared to (vs.)	Genotype/allele	Compared to (vs.)	Genotype/allele
Allelic	B	vs.	A		
Dominant	BB,AB	vs.	AA		
Recessive	BB	vs.	AB,AA		
Genotypic	BB	vs.	AB	vs.	AA

Table 3. Allelic Association Model. The 'B' demonstrates the minor allele, while the 'A' shows the major allele.

With *Likelihood-Ratio tests* we executed the association analyses of SNPs to continuous variables. For the association analyses to binary variables, we applied *Chi-Square tests*. Every SNV was set into association with every continuous and binary variable. For univariate analysis, we set the p-value significance level to 0.1.

In the next step, all SNVs significantly associated according to the univariate analysis were subjected to a multivariate association analysis. Here, the influence of age and the risk groups from *ELN2017*²⁷ were considered. Assuming that after univariate analysis there was a significant association between one SNP and OS, and that this association would no longer be significant after age and *ELN* adjustment in multivariate analysis. This would indicate that the association was not induced by the SNP but, for example, by a higher age of the patients carrying the SNP. We adjusted for both, age and *ELN2017*²⁷ and removed their influence on the clinical characteristics of AML patients.

In multivariate analysis binary variables were associated with the help of *logistic regression* and continuous variables with *linear regression models*. For multivariate analysis, we set the p-value significance level to 0.05. The p-values were adjusted with the *Benjamini-Hochberg* (BH) method, as the BH technique has proven to be a particularly good method of suppressing the detection of false significant p-values.⁷⁶

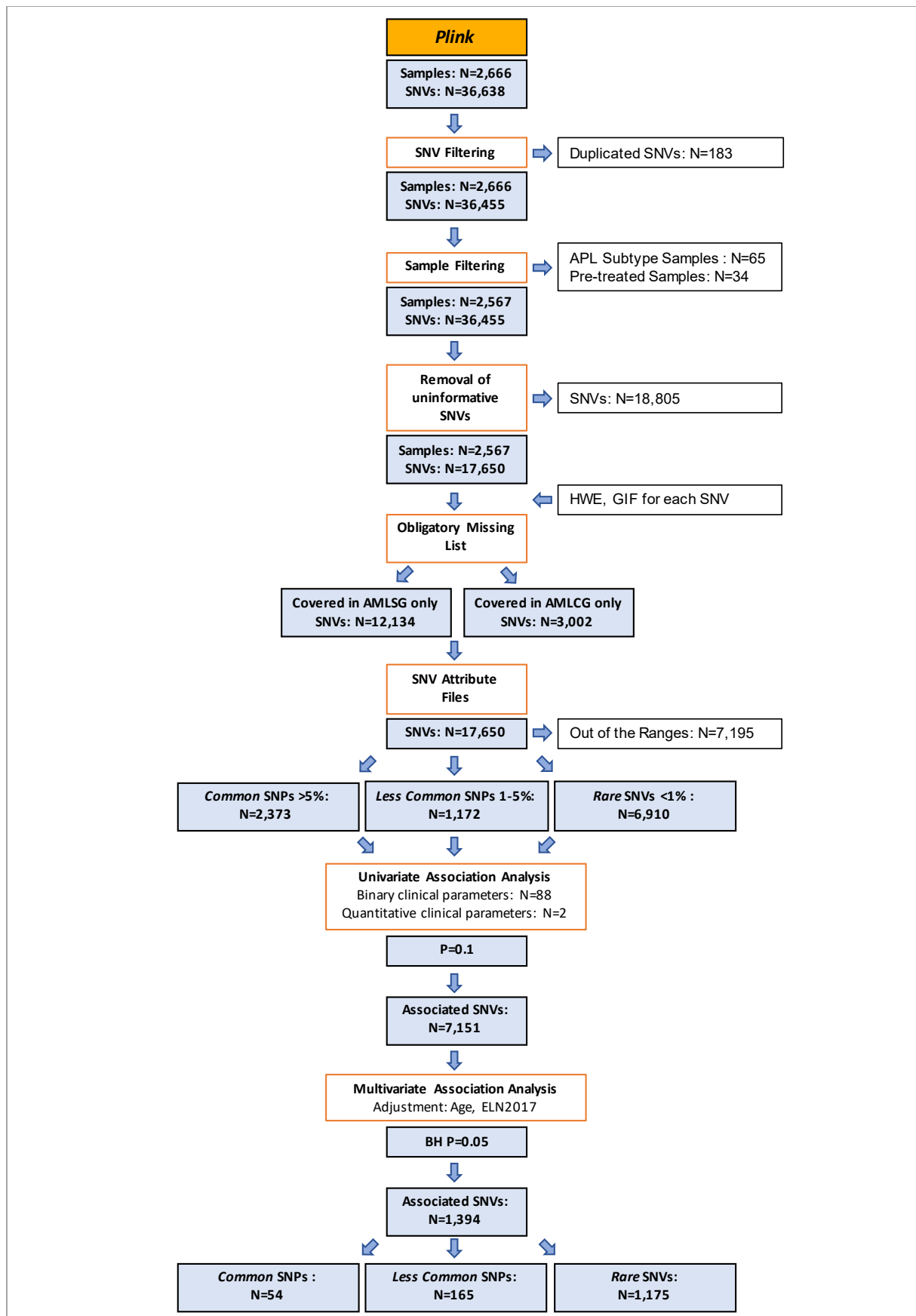


Figure 5. Quality Control and Association Analysis in Plink. The diagram shows the executed workflow in Plink with the variation numbers that were excluded during the different steps.

Explorative Results' Validation

After multivariate analysis significantly associated SNPs were subjected to validation in an unrelated cohort - the *BEAT-AML* cohort. The validation analysis was done with comparable filtering criteria and workflow as in the initial association transaction. In validation statistics, we evaluated unadjusted p-values.

Validation Analyses of previously published clinically relevant SNPs

The validation of associations from literature constituted the second part of this project. Before we started the validation analyses, varying reasons decreased the number of validatable SNP associations. Most SNP association studies were GWA studies, investigating which SNPs were found more frequently in AML compared to healthy control groups. This was not the focus of our project. Of those associations with prognostic relevance many SNPs were located in regions not covered by our sequencing panels. Among the identified prognostically relevant plus covered SNPs, some SNPs did not pass the executed filtering steps and were therefore eliminated. Furthermore, in a couple of SNPs, the number of available data was too small for obtaining significant results. *Figure 6* shows the reasons and sample numbers of the filtering process.

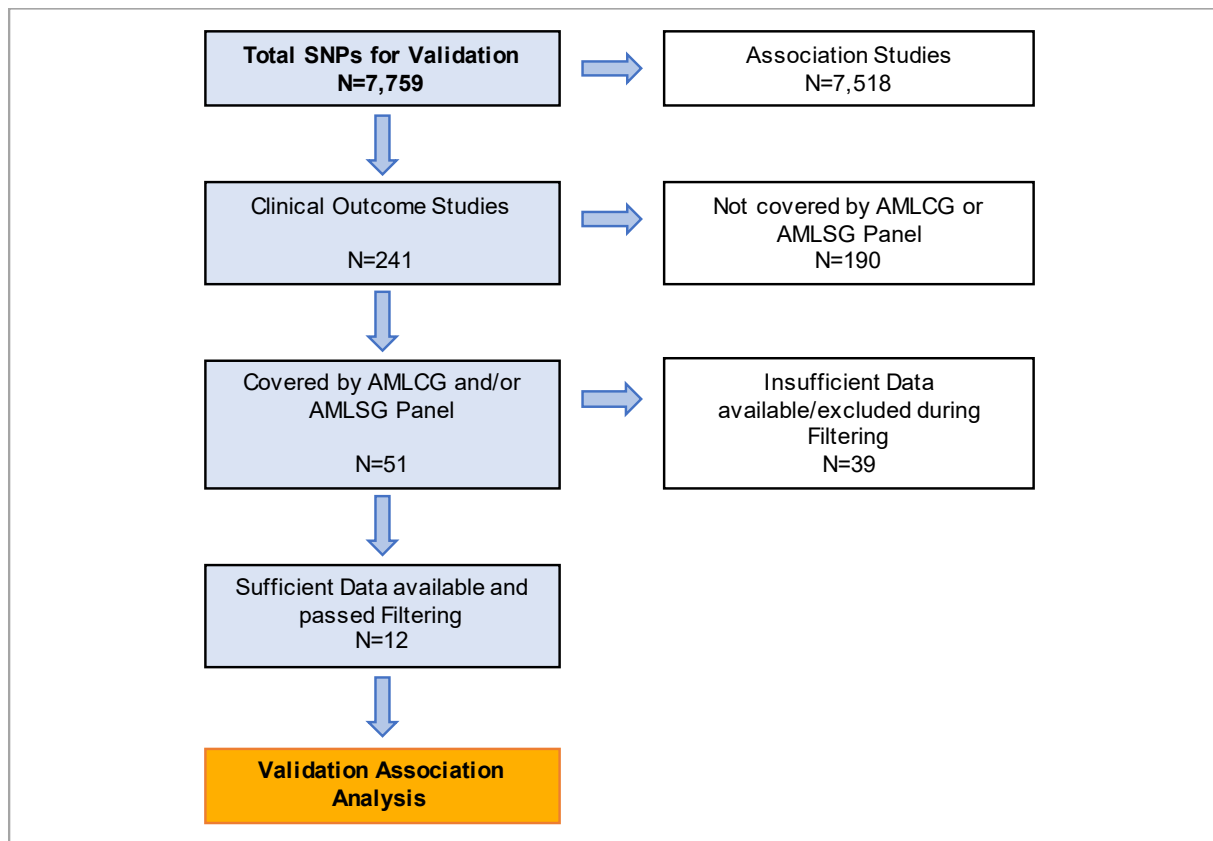


Figure 6. SNPs from previous Studies. The diagram shows the reasons due to which SNPs from previous studies were excluded from the validation analysis.

Twelve associations of prognostic relevance reported in previous studies were sufficiently covered in the *AMLCG* and/or *AMLSG* cohort to allow further analysis. Some of these SNP associations were previously described by one scientific group only, while others were found by different groups. All studies had in common that they analysed much smaller patient cohorts. Except the study team around Kutny, Leukaemia (2015)⁷⁷, no study group validated their results with help of another uncoupled cohort. Some studies adjusted for certain risk categories or single risk mutations, but no study adjusted for the currently used *ELN2017* prognostic scoring system. *Table 4* illustrates the previous studies and the SNP associations discovered there.

SNP ID (gene)	Publication	Alleles (study)	Samples	Stratification	Association result
rs532545 (CDA)	Falk et al., Am J Hematol. (2013) ⁷⁸	C>T	N=205	NK; pAML; <i>FLT3-ITD</i> ^f	TT: ↓OS, ↓Progression Free Survival (PFS)
				NK; pAML; <i>FLT3-ITD</i> ; <i>NPM1</i> mutation (mut.)	TT: Most pronounced ↓OS, ↓PFS
rs602950 (CDA)	Megías-Vericat et al., Leuk Lymphoma (2017) ⁷⁹	T>C	N=225	pAML	Minor Allele (MA): ↑CR, ↑OS, ↑Event Free Survival (EFS) at 5 years
rs2072671 (CDA)	Falk et al., Am J Hematol. (2013) ⁷⁸	A>C	N=205	NK; pAML; <i>FLT3-ITD</i>	CC: ↓OS
				NK; pAML; <i>FLT3-ITD</i> ; <i>NPM1</i> mut.	CC: Most pronounced ↓OS
	Kim et al., J Hum Genet. (2015) ⁸⁰	A>C	N=50	NK	MA: ↓OS ↓CR duration
	Megías-Vericat et al., Leuk Lymphoma (2017) ⁷⁹	A>C	N=225	pAML	AC: ↓OS, ↓EFS, ↓RFS
rs12036333 (CDA)	Gamazon et al., Blood (2013) ⁸¹	G>A	N=232 (in vivo, paediatric); N=523 (in vitro)		AA: ↓OS, ↑Treatment Related Mortality (TRM)
	Megías-Vericat et al., Pharmacogenet. Genomics (2017) ⁸²	G>A	N=225		MA: ↓OS, ↓Disease Free Survival (DFS), ↓RFS at 5 years

rs11554137 (IDH1)	Wagner et al., J Clin Oncol. (2010) ⁸³	C>T	N=275	NK; pAML/sAML ^g	MA: ↓OS, ↓RFS
	Ho et al., Blood (2011) ⁸⁴	C>T	N=274 (adult); N=253 (paediatric)	NK; pAML/sAML; <i>FLT3-ITD</i> a/o <i>NPM1</i> wildtype ^h pAML; CR	MA: Most pronounced ↓OS, ↓RFS MA: ↓OS, ↓RFS
rs2454206 (TET2)	Kutny et al., Leukaemia (2015) ⁷⁷	A>G	N=169 (paediatric)	pAML	MA: ↑OS
	Wang et al., Genes Chromosomes Cancer (2018) ⁸⁵	A>G	N=254 (paediatric)	MRC ⁸⁶ intermediate risk ⁱ	MA: ↑OS, ↑EFS
rs2897047 (intergenic, near IRX2)	Gamazon et al., Blood (2013) ⁸¹	C>T	N=232 (in vivo, paediatric); N=523 (in vitro)		TT: ↓Cytarabine response, ↑Minimal residual disease (MRD), ↓RFS
	Megías-Vericat et al., Pharmacogenet. Genomics (2017) ⁸²	C>T	N=225		CT: ↑OS
rs7729269 (MCC)	Gamazon et al., Blood (2013) ⁸¹	T>C	N=523 (in vitro)		Cytarabine sensitivity in lymphoblastoid Cell Lines (LCLs)
	Megías-Vericat et al., Pharmacogenet Genomics (2017) ⁸²	T>C	N=225		MA: Cytarabine- associated toxicities
rs1045642 (ABCB1)	Megías-Vericat et al., Pharmacogenomics J. (2015) ⁸⁷	C>T	N=1,221	pAML/sAML	MA: ↑OS
	Rafiee et al., Blood Cancer J. (2019) ⁸⁸	C>T	N=942	pAML	MA: ↑EFS, ↓Relapse rates

rs1128503 (ABCB1)	Megías-Vericat et al., Pharmacogenomics J. (2015) ⁸⁷	C>T	N=925	pAML/sAML	MA: ↑OS
rs2229109 (ABCB1)	Gregers et al., Pharmacogenomics J. (2015) ⁸⁹ Dessilly et al., Pharmacogenomics (2016) ⁹⁰	G>A G>A	N=522 (ALL) n.a. (CML LCLs ^f)		GA: ↑Relapse risk, ↓EFS MA: ↑Drug metabolism, ↓Intracellular accumulation
rs10883841 (NT5C2)	Falk et al, Am J Hematol. (2013) ⁷⁸	A>G	N=205	NK; pAML; <i>FLT3-ITD</i> neg.	MA: ↓OS
<p>^f <i>FLT3-ITD</i>: Internal Tandem Duplication of the <i>FLT3</i> gene. ^g pAML/sAML: samples were stratified for having either <i>de novo</i> AML, or <i>secondary</i> AML. ^h a/o: Wagner et al. stratified for samples that had a <i>FLT3-ITD</i> mutation and wildtype (unmutated) <i>NPM1</i> or for patients that showed one of those phenomena, the <i>FLT3-ITD</i> or the unmutated <i>NPM1</i> form. ⁱ MRC: a risk score applied by the Medical Research Council, UK.⁸⁶ ^j Chronic Myeloid Leukaemia Lymphoblastoid Cell Lines.</p>					

Table 4. Detailed Literature Studies Information. The table shows the SNP associations from previous publications that underwent validation analysis in our cohort. Description: The *SNP ID* shows the SNP for which an association was found. The corresponding gene name is shown in brackets. *Publication* states authors and journals. *Alleles (study)* show the major and minor alleles resumed from the studies. *Samples* shows the number of samples that were analysed in the respective study. *Stratification* describes the cohort subgroup the result was valid for. This includes primal cohort restrictions as well as further stratification of the analysed group. The column *Association result* shows the results of the analysed cohort. In case *MA* is indicated in this column, the minor allele is associated with the shown clinics independent of its present genotype (according to an allelic association model). Specific genotypes like, for example, *AA* or *CT* show that only this genotype was significantly associated in the respective study.

To validate the reported associations, we had to mimic the previous studies. In majority, the results of the previously conducted studies arose from subgroup analyses. Hence, many of the SNPs associated in a certain analysed subgroup were not associated in an unstratified cohort. We succeeded in stratifying our cohort according to all subgroups on which the studies were executed. These subgroups were normal karyotype AML, primary (*de novo*) AML, secondary AML, *FLT3-ITD* status, *NPM1* status and MRC risk groups.

The stratification criteria mentioned in the study description (compare *table 4*) were applied to all samples of the *AML CG* and *AML SG* cohorts. Accordingly, during validation analyses all samples in our cohort were filtered by equal criteria.

Also, we aimed to imitate the association models used by the prior study conductors. We concluded the used models from the resulted associations since the information on those models were not available in most publications. If a study identified that the SNP's minor allele has generally been associated with a clinical parameter, it was likely that an allelic model was performed. Studies that found specific genotypes of a SNP to be associated, were imitated by applying a genotypic model.

Model	Genotype/allele	Compared to (vs.)	Genotype/allele	Compared to (vs.)	Genotype/allele
Allelic	B	vs.	A		
Dominant	BB, AB	vs.	AA		
Recessive	BB	vs.	AB, AA		
Genotypic	BB	vs.	AB	vs.	AA

Table 5. Genotypic Association Model. The 'B' demonstrates the minor allele, while the 'A' shows the major allele.

In contrast to the successfully reproduced stratifications of the previous studies, we were only partially able to reproduce the associated clinical parameters. Following clinical parameters associated in the prior studies could be analysed in the *AMLCG* and *AMLSG* datasets: *OS*, *CR*, *RFS*, *ED* and *RD*. Other clinical characteristics were not available in *AMLCG* and *AMLSG* and could not be reproduced. In these cases, we associated the certain previously associated SNP with the closest screened variable in our cohort. Parameters that showed no close variable in our cohort were associated with the major outcome variables (*OS*, *RFS*, *CR*). Since in vitro variables were generally not evaluated in our cohort, previous studies focussing on the latter were also reproduced by associating their SNPs with the major outcome variables for validation analysis.

A further challenge was that some of the studies found SNP alleles other than the SNP alleles we mentioned. In the studies of Megías-Vericat et al., Leuk Lymphoma (2017)⁷⁹ (SNP rs602950), Wagner et al., J Clin Oncol (2010)⁸³ and Ho et al., Blood (2011)⁸⁴ (both rs11554137), Gregers et al., Pharmacogenomics (2015)⁸⁹ and Desilly et al., Pharmacogenomics (2016)⁹⁰ (both rs2229109) the shown alleles (compare table 4) were the paired alleles from the opposite strand. This could be explained by sequencing the reverse strands in the literature studies. Alleles from the opposite

strand should not influence the comparability of the literature results and the validation study results. Anyway, some other allele differences could not be explained by sequencing the opposite strand. These were the alleles mentioned in the studies of Gamazon et al., *Blood* (2013)⁸¹ and Megías-Vericat et al., *Pharmacogenet Genomics* (2017)⁸² (both rs2897047) as well as Megías-Vericat et al., *Pharmacogenomics J.* (2015)⁸⁷ and Rafiee et al., *Blood Cancer J.* (2019)⁸⁸ (both rs1045642) and Megías-Vericat et al., *Pharmacogenomics J.* (2015)⁸⁷ (rs1128503). For validating those associations, we used the variations analysed in our cohort. It is important to mention that all alleles from our analysis accorded with alleles found in the current *dbSNP* version (as of May 2021).

In addition, some of the previous studies were conducted in children and patients with other types of leukaemia but AML.

Beside the factors already mentioned, the validation analyses were carried out in the same way as the respective explorative analyses. The certain SNP(s) found in prior studies underwent association analysis with the previously noticed outcome factor(s).

Univariate analysis without adjustment was performed. In case, a SNP showed $p < 0.1$, multivariate analysis followed. Here, many different adjustment factors were applied in previous studies. Some examples were white blood cell count, year of diagnosis, or performance status. We did not mimic these adjustment parameters and adjusted for the clinically relevant parameters *ELN2017* and age instead.

RESULTS

In this section I will describe the results of our pilot project. I will compare our data set with previously published data sets and hereby verify the transferability of our dataset to clinical AML patients. This will be followed by findings from both, the explorative analyses and the validation analyses from previous studies.

Data Comparison

Comparing the *AML*CG and *AML*SG data sets revealed a number of differences. The median *AML*CG patient was eight years older than the median *AML*SG patient. Since age is considered to be one of the major influencing factors in terms of outcome, differences in OS and the post-induction situation could be detected between both data sets. *AML*CG patients had slightly lower rates of CR achievements, more EDs, and a shorter OS compared to *AML*SG patients. Nevertheless, the cases of RD were proportionally higher in *AML*SG.

Compared to other published data on AML patients, both, the *AML*CG and *AML*SG cohort, incorporated rather young patients with the average of patients being in their fifties while the median age of AML onset is often published to be 69 years (compare chapter *Epidemiology and Pathogenesis*). Further differences compared to other publications were comparably high rates of pAML and low rates of sAML samples in both, the *AML*CG and the *AML*SG cohort. Nevertheless, the literature on the occurrence of pAML and sAML has strongly varied.^{91,92} With slightly more male patients in both cohorts, the sex distribution corresponded with other studies.^{93,94} Broadly, most other parameters, such as blood parameters and karyotype, were consistent with other published data on AML patients.

Explorative Results

After conduction of the data preparation steps in terms of variant calling, pileup, annotation and *GEMINI* filtering 36,638 SNVs were available. Comparable with the gene coverage distribution (compare *figure 2*) only a smaller number of the SNVs was covered by both, the *AML*CG and the *AML*SG cohort. *Figure 7* shows the distribution of these SNVs which then passed to *plink*.

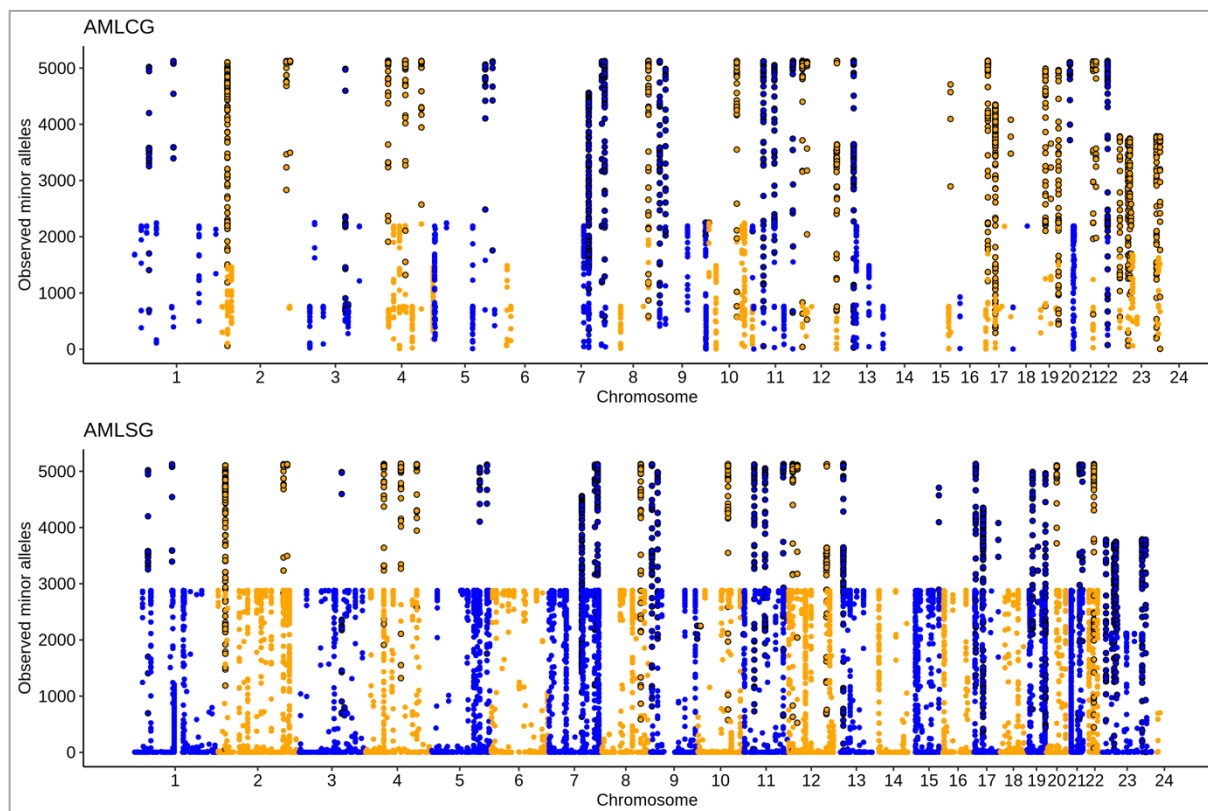


Figure 7. SNV Distribution. The upper figure shows the SNVs called in the AMLCG cohort; the lower figure shows the SNVs called in the AMLSG cohort. The X-axis indicates the relative positions of the SNVs on the chromosomes whereas 23=X and 24=Y chromosome. The Y-axis shows the overall number of minor alleles screened at the certain position. The colour coding (blue and orange) is used for the clarity between the different chromosomes. Uncircled dots indicate SNVs present in the shown cohort only. Black circled dots indicate SNVs that were covered in both, the AMLCG and AMLSG cohorts.

The total number of shown SNVs is N=36,638.

After the univariate analysis of all SNVs, a total of 7,151 associations with clinical variables as well as with curated mutations resulted. Out of these, 1,394 SNVs remained associated after the multivariate analysis.

Since we focussed on frequent, clinically significant SNPs, the number of interesting associations decreased considerably after exclusion of the rare SNVs. Concerning gene mutations, only those associations with recurrent mutations were viewed for the result evaluation. By applying these filters, out of the 1,394 associations from multivariate analysis, solely 23 associations remained for further consideration.

Figure 8 gives an insight on the composition of the multivariate association results.

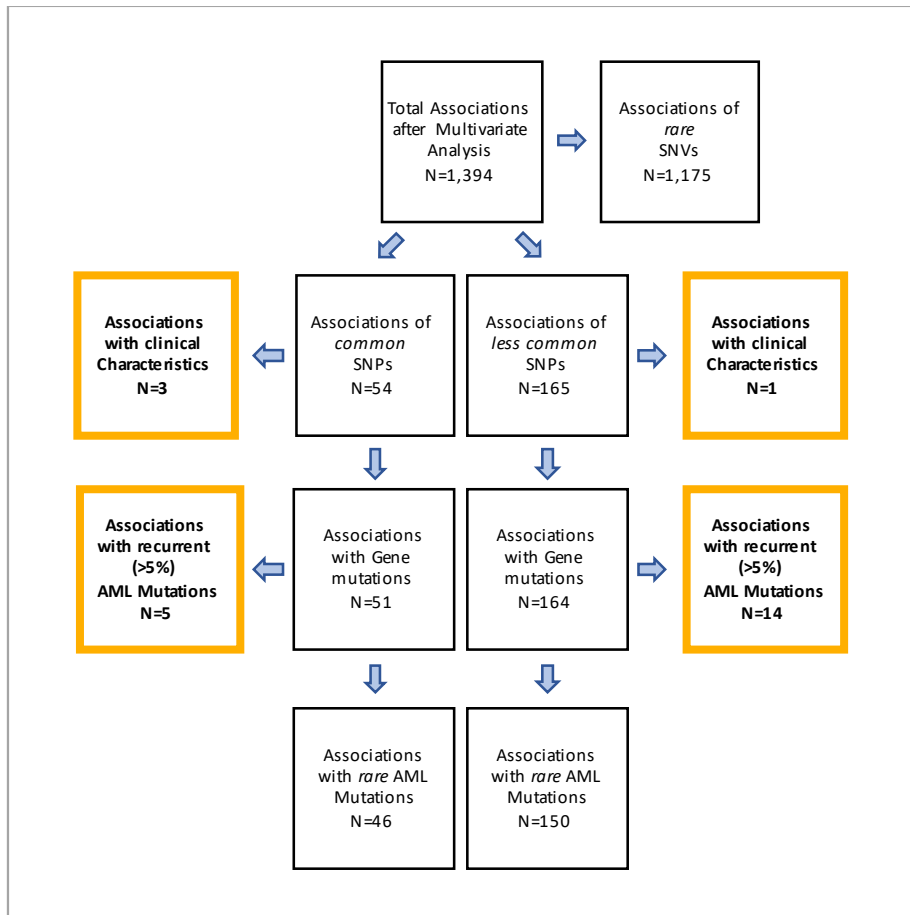


Figure 8. Explorative Multivariate Association Results Overview. The figure shows the composition of the multivariate association results from *common* and *less common* SNPs.

In the following section I will describe in detail the associations marked in orange. First, I will describe the associations with clinical characteristics and then the associations with recurrent gene mutations.

SNPs associated with Clinical Characteristics

The following outcome variables were analysed: *Overall Survival (OS)*, *Refractory-Free Survival (RFS)*, *Complete Remission (CR)*, *Early Death (ED)* and *Resistant Disease (RD)*. When focussing on these characteristics, we found three variants within the *common* and *less common* SNPs that were significantly ($P < 0.05$) associated after multivariate analysis and correction for multiple testing.

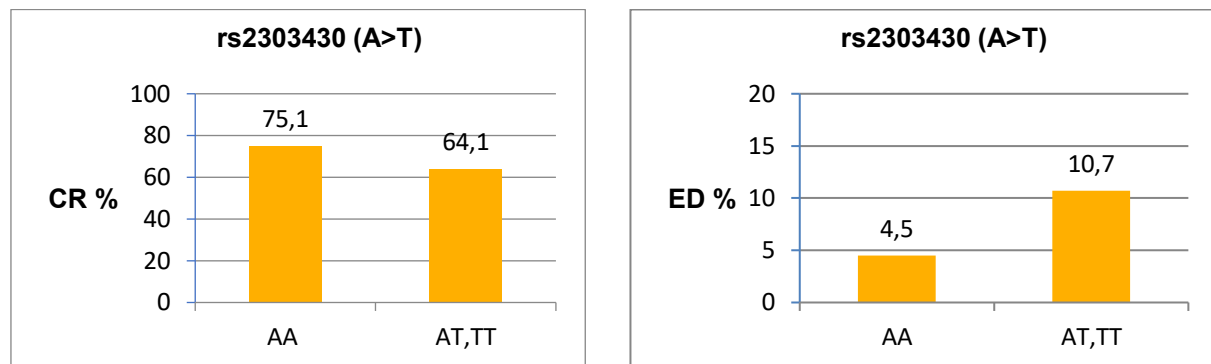
rs2303430 (A>T, A=major allele, T=minor allele)

The most prominent SNP associations concerned rs2303430. This variant is located in the *PDGFRA* gene on chromosome 4.⁹⁵ The region was covered in the *AML*SG

population (N=1,366). With a MAF of 0.298, almost one-third of the sequenced reads carried the minor allele *T* at this specific position. Hereby, the SNP was categorised as *common*. *dbSNP* confirms the high MAF of rs2303430.⁹⁵

Within the AML patients from our cohort, there was a significant association found between samples carrying the minor allele and the patient's outcome after induction treatment. A negative association could be seen between people carrying the minor allele and the rate of achieved *CRs* (P=0.003; OR=0.626). Simultaneously, there was a positive association between minor allele carriers and the rate of *EDs* (P<0.001; OR: 2.004). *Figure 9a* and *9b* visualise these associations.

To the best of our knowledge, this SNP has not been mentioned in literature before neither in association with AML nor with any other variable or disease.



9a.

9b.

Figure 9a. rs2303430~CR. Figure 9b. rs2303430~ED. The number of patients sequenced at the rs2303430 position who carried the major allele karyotype (AA) was N=694 in our cohort. N=672 patients carried either one or two minor alleles (AT: N=530, TT: N=142).

The *BEAT-AML* cohort covered rs2303430. We were not able to validate the associations to *CR* in the *BEAT-AML* cohort (P=0.56; OR=1.19). Information on *ED* was not available in the *BEAT-AML* cohort.

rs28489067 (C>T)

The SNP rs28489067 is located in the *PDGFRA* gene on chromosome 4.⁹⁶ This region was covered in N=1,148 *AML*SG samples. The MAF of 0.1 assigned this SNP to the group of *common* SNPs. With a MAF of 0.18 the *dbSNP* database indicates the minor allele to be more frequent as in our dataset.⁹⁶ The association's OR was 1.88, stating that the minor allele of the SNP was associated with higher rates of *CR* (P<0.001).

Figure 10 visualises the higher rate of CR in patients who have carried at least one minor allele compared to patients who are homozygous for the major allele.

The SNP has not been reviewed in association with AML before. Solely, one article concerning the interpretation of variations in general, mentioned this SNP before.⁹⁷

The *BEAT-AML* cohort did not cover the SNP's position for validation.

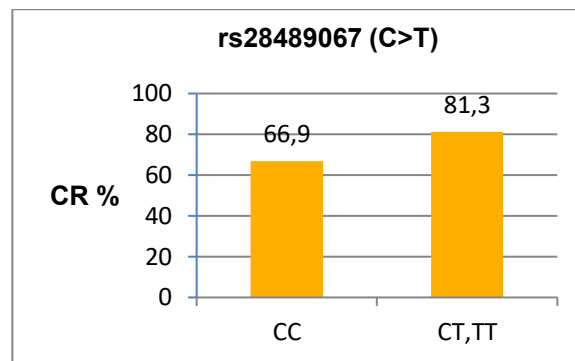


Figure 10. *rs28489067*~CR. The total number of *rs28489067* sequenced patients who carried the major CC-genotype was N=934 while the minor allele (CT, TT) was carried in N=214 patients.

rs145370659 (A>C)

This SNP belonged to the group of *less common* SNPs in our cohort (MAF: 0.015). The MAF was quoted similar by the *dbSNP* database. It is located on chromosome 1 in the intergenic region of the protein-coding gene *NDRC* and the downstream variant of the MicroRNA *MIR761*.⁹⁸ In our cohort, this region was covered in N=1,425 samples of the *AML*SG cohort. We detected an association between the SNP and *RD* ($P < 0.001$; Odds Ratio (OR)=4.29). This result revealed that holders of the minor allele had higher rates of *RD* in comparison to patients who were homozygous for the major allele as shown in *figure 11*. This SNP has not been mentioned in the literature before.

The *BEAT-AML* cohort did not cover the position of *rs145370659*, therefore, we were not able to validate the association in an unrelated cohort.

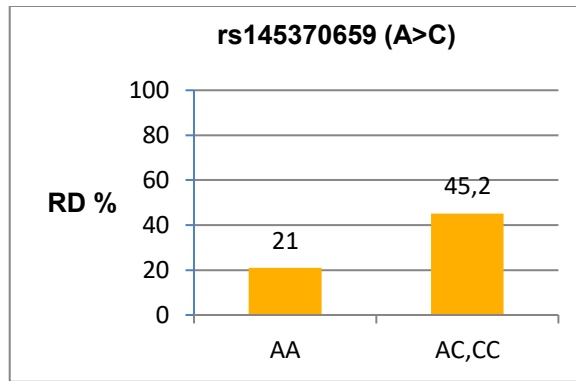


Figure 11. *rs145370659~RD*. The total number of screened samples with the major AA-genotype was N=1,383. N=42 Samples carried the minor allele (AC, CC).

SNPs associated with Gene Mutations

Focussing on *common* and *less common* SNPs, 215 out of the 219 total associations from multivariate analysis were found between SNPs and gene mutations (compare *figure 8*). These associations occurred between SNPs and curated mutations of 38 different genes. Most of these mutations were rare in the patient samples (found in less than five percent). Focussing on mutations which affected more than five percent of the patients in our cohort resulted in 19 significant associations between SNPs and gene mutations. All these associations were detected with SNPs only covered in the AMLSG cohort and shown in *table 6*.

SNP	Gene	Samples N=	Associated gene	OR	Adjusted P-Value	MAF group
rs1052639 (G>A)	DDX18	1,435	IDH2	1.865	0.002	common
rs114734174 (C>T)	perigenic	1,412	IDH2	1.954	0.001	common
rs7628252 (G>C)	perigenic	1,429	IDH2	2.152	<0.001	common
rs490052 (A>G)	RBFOX1	1,201	TP53	0.553	0.002	common
rs16845518 (C>T)	perigenic	1,362	KRAS	2.427	<0.001	common
rs726070 (C>T)	ABCA12; SNHG31	1,435	MLL	3.614	0.002	less common
rs17501532 (C>A)	ABCA12	1,424	MLL	4.541	0.002	less common
rs11890512 (T>G)	ABCA12	1,433	MLL	2.59	0.035	less common
rs10179876 (T>C)	ABCA12	1,424	MLL	3.852	0.002	less common
rs10167501 (G>A)	ABCA12	1,424	MLL	3.248	0.009	less common
rs11890512 (T>G)	ABCA12	1,433	MLL-PTD	2.522	0.042	less common
rs10179876 (T>C)	ABCA12	1,424	MLL-PTD	3.574	0.008	less common
rs10167501 (G>A)	ABCA12	1,424	MLL-PTD	3.071	0.016	less common
rs7789818 (T>C)	KMT2C	1,343	NRAS	2.724	0.002	less common
rs140378196 (A>G)	KDM5A	1,386	NRAS	3.042	<0.001	less common
rs77123954 (T>C)	perigenic	1,380	NRAS	3.751	<0.001	less common
rs75395837 (G>C)	ZBTB44	1,381	ASXL1	11.45	<0.001	less common
rs35918540 (C>T)	TCF4	1,417	IDH2	3.114	<0.001	less common
rs370821688 (G>T)	perigenic	1,127	KRAS	4.309	0.002	less common

Table 6. SNP Associations with Gene Mutations. The table shows the SNP associations with recurrent gene mutations after multivariate analysis. The given *Gene* is the name of the gene in which the SNP is located. *Samples* describes the number of samples in which the region was covered. The column *Associated gene* shows the name of the gene mutation that was associated with this certain SNP. *OR* and *Adjusted P-value* show the statistical results from multivariate association analysis in *Plink*. The *MAF group* gives information on the frequency of the minor allele and relates to the ranges for *common* and *less common* defined within *Plink*.

Besides rs726070 none of the SNPs have been mentioned in literature before. rs726070 was described as associated with congenital ichthyosis and mentioned in a publication about the interpretation of sequence variants.^{97,99,100} Up to now, it has not been associated with AML before.

Since SNP to gene mutation associations were not the main target of our study, we did not try to validate them in the *BEAT-AML* cohort.

Validation Analyses of previously published clinically relevant SNPs

We were able to identify 12 previous studies which could be re-analysed in our data sets (compare figure 4). Below, table 7 shows the validation studies and the results from our data sets. Some association parameters were replaced by other parameters for the validation analyses. Also, some alleles mentioned in literature had to be validated with other alleles as identified in our cohort (compare chapter *Confirmation of previously identified SNP associations*).

SNP	Study	Genotype	Stratification	Model	Samples N=	Association	P	
rs532545 (C>T)	Falk et al., Am J Hematol. (2013) ⁷⁸	TT	NK; pAML; <i>FLT3-ITD</i>	genotypic	212	OS	0.99	
						RFS	0.77	
			NK; pAML; <i>FLT3-ITD</i> ; <i>NPM1</i>	genotypic	157	OS	0.82	
						RFS	0.69	
rs602950 (A>G)	Megías-Vericat et al., Leuk Lymphoma (2017) ⁷⁹	MA	pAML	allelic	892	OS	0.32	
						RFS	0.92	
						CR	0.33	
rs2072671 (A>C)	Falk et al., Am J Hematol. (2013) ⁷⁸	CC	NK; pAML; <i>FLT3-ITD</i>	genotypic	216	OS	0.65	
						NK; pAML; <i>FLT3-ITD</i> ; <i>NPM1</i>	genotypic	160
	Kim et al., J Hum Genet. (2015) ⁸⁰	CC	NK	genotypic	696			
		Megías-Vericat et al., Leuk Lymphoma (2017) ⁷⁹	MA	NK	allelic	696	OS	0.49
	AC		pAML	genotypic	900	OS	0.88	
		RFS				0.12		
rs12036333 (G>A)	Gamazon et al., Blood (2013) ⁸¹	AA		genotypic	1,022	OS	0.21	
						RFS	0.46	
	Megías-Vericat et al., Pharmacogenet. Genomics (2017) ⁸²	MA			allelic	1,022	OS	0.46
							RFS	0.91

rs11554137 (G>A)	Wagner et al., J Clin Oncol. (2010) ⁸³	MA	NK; pAML/sAML	allelic	559	OS RFS	0.65 0.33
	Ho et al., Blood (2011) ⁸⁴	MA	NK; pAML/sAML; <i>FLT3-ITD a/o</i> <i>NPM1</i> wildtype	allelic	445	OS	0.89
						RFS	0.64
							OS RFS
rs2454206 (A>G)	Kutny et al., Leukemia (2015) ⁷⁷	MA	pAML	allelic	2,207	OS	0.41
	Wang et al., Genes Chromosomes Cancer (2018) ⁸⁵	MA	MRC intermediate risk	allelic	1,786	OS	0.11
						RFS	0.06
rs2897047 (A>G)	Gamazon et al., Blood (2013) ⁸¹	GG		genotypic	620	RFS	0.25
	Megías-Vericat et al., Pharmacogenet Genomics (2017) ⁸²	AG		genotypic	1,070	OS	1
rs7729269 (T>C)	Megías-Vericat et al., Pharmacogenet Genomics (2017) ⁸² ; Gamazon et al., Blood (2013) ⁸¹	MA		allelic	1,076	OS	0.37
						RFS	0.89
						CR	0.06
rs1045642 (A>G)	Megías-Vericat et al., Pharmacogenomics J. (2015) ⁸⁷	MA	pAML/sAML	allelic	1,013	OS	0.85
	Rafiee et al., Blood Cancer J. (2019) ⁸⁸	MA	pAML	allelic	561	RFS	0.83
rs1128503 (A>G)	Megías-Vericat et al., Pharmacogenomics J. (2015) ⁸⁷	MA	pAML/sAML	allelic	1,007	OS	0.47
rs2229109 (C>T)	Gregers et al., Pharmacogenomics J. (2015) ⁸⁹	CT		genotypic	636	RFS	0.99
	Dessilly et al., Pharmacogenomics (2016) ⁹⁰	MA		allelic	1,089	OS	0.45
						RFS	0.7
						CR	0.81
rs10883841 (T>C)	Falk et al., Am J Hematol. (2013) ⁷⁸	MA	NK; pAML; <i>FLT3-ITD</i> pos.	allelic	376	OS	0.72

Table 7. Previous Studies' Validation Results. The table shows the results from our validation study. The analyses printed in **bold** are those we fully succeeded to imitate. *Study* states which previous study was aimed to get validated. *Genotype* shows which genotype we used for the validation association analysis. In case a genotype differed from the literature study (compare *table 4*)

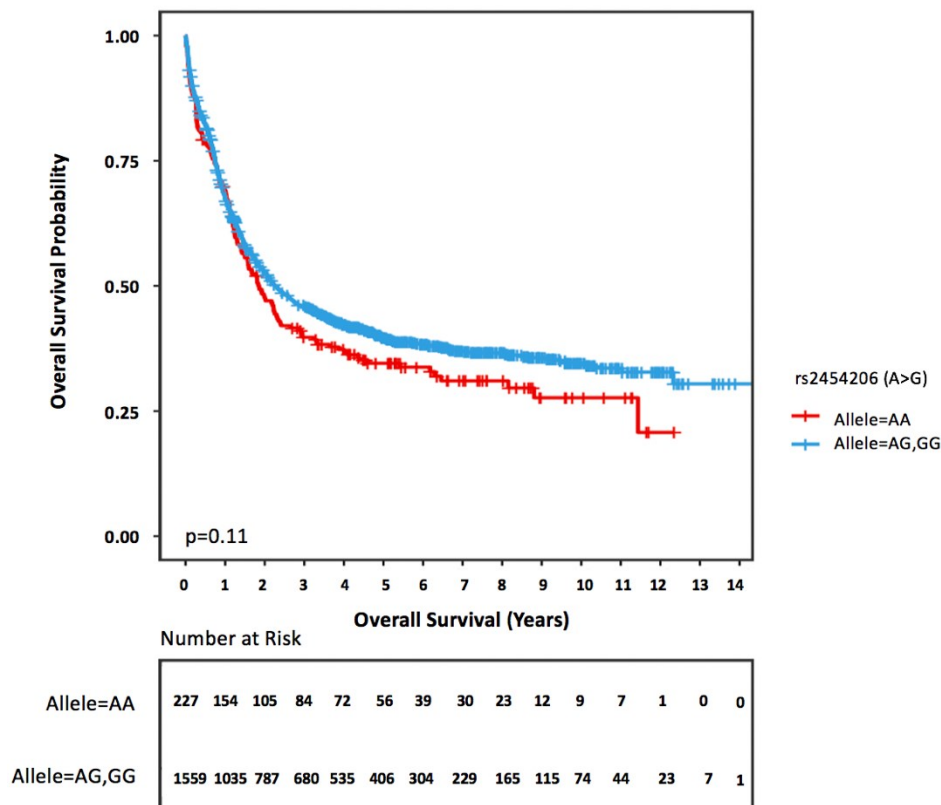
different alleles were sequenced in the literature cohort and our cohort. If *MA* is shown, we associated all minor allele carrying (heterozygous and homozygous) genotypes with the SNP. *Stratification* explains how we stratified our cohort in the validation analysis. *Model* shows which association model was utilised for validation. The *samples* column illustrates number of samples in our cohort who fitted to the designated stratification criteria. *Re-analysed association* shows the variable to which we associated the SNP. In case, the association factor from the prior study was not recorded in our data set, this factor varies comparing to *table 4*. *P* shows the statistical results of the validation analysis with *P* being unadjusted.

In summary, we could not validate any of the findings from previous studies.

The validation from two studies showed interesting, though not significant, associations.

First, rs2454206 (A>G) was associated with *RFS* and *OS* in MRC intermediate risk patients. The analysis aimed to validate the results prior published by Wang et al., *Genes Chromosomes Cancer* (2018)⁸⁵. While Wang found higher *RFS* and better *OS* in paediatric patients carrying the SNP's minor allele, we were able to reproduce these results in an adult cohort. The univariate allelic association analysis showed significant results for *RFS* and borderline significant results for *OS* in our cohort. However, the associations were not significant in multivariate analysis. *Figure 12a* and *12b* visualise the association.

Figure 12a.



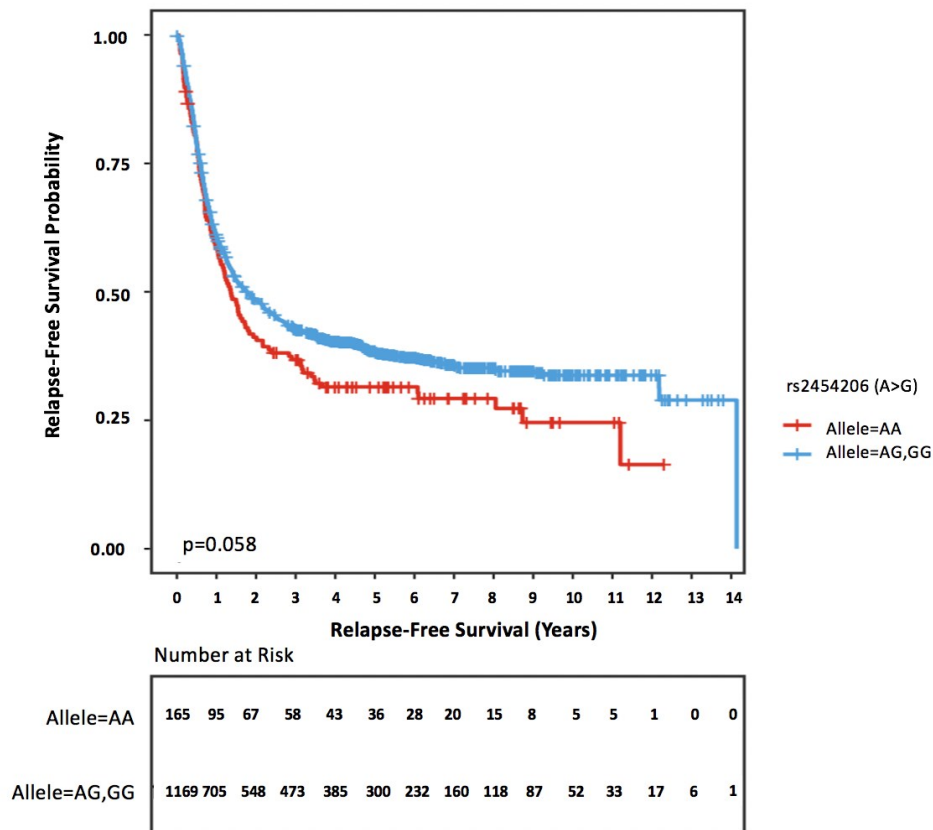


Figure 12b.

Figure 12a. rs2454206~OS 12b. rs2454206~RFS. The plots show the results from the univariate association analysis of MRC intermediate risk group patients.

Second, the association analysis of rs7729269 (T>C) showed interesting results. We performed it to reproduce the studies of Gamazon et al., Blood (2013)⁸¹ and Megías-Vericat et al., Pharmacogenetic Genomics (2017)⁸². In the previous studies, associations with Cytarabine response and toxicity were reported, neither of which were evaluated in our cohort. In our cohort the SNP associated with CR. Lower rates of CR in patients carrying the minor allele (P=0.06; OR=0.8) were observed. Again, these findings were not left as significant variables in multivariate analysis.

In summary, our work was able to reproduce some previous findings, however, after considering current standard prognostic variables like ELN2017 none proved clinical relevance.

DISCUSSION

Patients

A detailed characterisation of the patient samples was essential to correctly process the data and to minimise the risk of biases. All samples were broadly clinically characterised, and the disease process was continuously documented by the *AML CG* and *AML SG* study conductors. The data comparability was high with all patients being intensively treated within clinical trials. Since the therapy backbone did not change for 30-40 years^{63,101} the results obtained from rather old samples (beginning from 1999) could be transferred to AML patients in these days. However, considerable improvements in supportive care and the introduction of targeted treatment approaches in the last years might influence comparability to more recently conducted trials.

Median age of our cohort was significantly lower than the age of average, unselected AML patients.^{3,4} This age difference is presumably reasoned by the inclusion of only intensively treated patients in the *AML CG* and *AML SG* datasets. The intensive treatment is not offered to most of older or comorbid patients.¹⁰² As many patients do not accord with the age range in our study, the results might not be completely transferable to the large group of older AML patients. Also, the age difference between *AML CG* and *AML SG* patients might have influenced the comparability of the data. Since age is an important outcome-determining factor³⁴, age was uncoupled from other variables during the adjustment in multivariate analysis.

Since the transferability of a prognostic or predictive marker is always coupled with a specific therapy, the results of this study can only be applied on patients receiving the same intensive treatment protocols regardless of a patient's age.

The other clinical parameters of our cohort were largely comparable to those published in previous studies that included both younger and older AML patients, as indicated in the results section.

Methods

All tools and statistical methods have complied to commonly executed algorithms in SNP association analyses. The team chose the tools on base of good experience

considering speed, accuracy, and handling of large datasets. Also, the chosen programs interlocked well with each other. However, concerning the way of sequencing, the chosen reference genome, and the cut-off percentage for excluding variants with low coverage, different applications might have been an equally good option. Following, I will discuss some of these alternate options. I will also briefly comment and reflect the way we performed the association re-analyses of previous studies and what consequences resulted from this way of performing validation analyses.

First, the applied method of targeted amplicon sequencing required the knowledge of informative AML genome segments before the start of sequencing.¹⁰³ Since recurrently mutated genes in AML have been identified, sequencing sites could be defined. Despite, potentially significant associations outside the targeted locations have been missed. Also, the sequenced regions differed between *AMLCG* and *AML_{SG}* samples. With this, the majority of the called SNPs also did not overlap between the cohorts. Due to the low overlap, an alternate option would have been to perform the workflow separately in both cohorts. This would have improved homogeneity of the data and thus reduced the error susceptibility of the subsequent association analyses. Meanwhile, the number of samples analysed for several SNPs would have been reduced, therefore reducing the significance and power of the results. In this context we kept the dataset as one cohort and decreased the error susceptibility by applying various filtering steps (compare chapter *Workflow of the explorative Association Analysis*).

Second, even though many SNPs were only covered in half of our cohort, the SNPs were nevertheless analysed in a greater number of patients than in most previous AML SNP studies. Instead, a couple of previous studies sequenced larger genomic regions, providing a higher number of analysable SNPs.

Third, the choice of the reference genome fell on the no longer current version *GRCh37*. This version was used to keep this study comparable to former studies executed on the same data set and utilised *GRCh37*. Also, a genomic position could always be viewed from the sight of *GRCh37* and the actual *GRCh38* in the *dbSNP* database. Nonetheless, we had to be careful when comparing genomic locations between our and other studies, as sequences may have differed depending on the reference genome chosen.

Fourth, during the data preparation we removed SNVs whose genomic locations were covered in <30% of the samples. In many association studies, filters were applied that also exclude variants with higher missingness cut-offs as <50%. However, most of the other association studies were conducted on GWAS data, consequentially the coverage for the single samples was much higher compared to the amplicon sequencing in our cohort. Especially, variations that were covered in one of the cohorts, for example the *AML*CG cohort and in only few samples of the second, *AML*SG cohort, would have been excluded by applying a higher cut-off for missingness.

During the validation analysis of previous study findings, we aimed to exactly re-analyse the previous studies. Since we did not screen for all clinical parameters mentioned in the prior studies, we had to focus on the validation of few associations. With regard to the parameters that were not assessed in our cohort, replacing them by related variables reduced the power of our re-analysis results as a validation analysis on the one hand, on the other hand we still proved the clinical relevance of those variants.

Results

Explorative Analysis

In this section, I will elaborate on possible causal explanations for the established associations between rs2303430, rs28489067, rs145370659, and prognostic factors. Furthermore, I will have a closer look at the associations between SNPs and gene mutations.

With respect to rs2303430 (A>T), we found two associations between the SNP's minor allele and lower *CR* rates as well as higher *ED* rates. While the reason of these associations has stayed speculative, it is worthy to discuss potential causes.

Although rs2303430 is an intron variant and thus does not per se change the function of the *Platelet Derived Growth Factor Receptor Alpha* (PDGFRA) gene protein, intron variants do influence gene activity in other ways. For example, previous

research has described SNPs which, though located far from any splice site in the intron, influenced the transcriptional factor binding or the splicing efficiency of genes.⁴⁹

The *PDGFRA* gene on whose intronic region SNP rs2303430 is located is particularly responsible for the stimulation of cell growth and proliferation processes.^{95,104} Specifically, the gene induces the formation of a protein belonging to the family of receptor tyrosine kinases (RTKs). Among others, these *RTKs* are responsible for reactions that (de-)activate enzymes¹⁰⁴⁻¹⁰⁷. Enzymes as regulators of cell growth could influence the remission rates and mortality in leukaemic patients. Also, the potentially altered role of enzymes in drug metabolism may contribute to decreased *CR* and increased *ED* rates.

A further causal link could be established between rs2303430 and a particular leukaemia entity called *PDGFRA-associated chronic eosinophilic leukaemia*. The patients who suffer from this type of leukaemia carry a mutated *PDGFRA* gene (often in fusion with the *FIP1L1* gene) that encodes a fusion protein which does not need cytokine bindings for being activated. Via signalling molecules, the mutated *PDGFRA* gene leads to the continuous myeloproliferation and a more probable survival of the affected cells. Cases in AML patients carrying the *PDGFRA-FIP1L1* gene have been described before.¹⁰⁸⁻¹¹⁰

Translated to our cohort, this might mean that changes in the *PDGFRA* gene, potentially co-initiated by the SNP rs2303430, might increase the growth rate of leukaemic cells and hereby influence the achievement of *CR* and *ED*.

Also associated with *CR* was the SNP rs28489067 (C>T). Patients with the minor allele showed an increased chance of achieving *CR*. Like rs2303430, this SNP is located in an intronic region of the *PDGFRA* gene.⁹⁶ Therefore, the same possible association explanations can be constructed as for rs2303430.

With two SNPs located in the *PDGFRA* gene being associated with *CR*, the theory of changes in this gene as a potential cause of the association with the postinduction situation becomes more likely. Surprisingly, previous studies did not review the latter two SNPs to be associated with the outcome of AML patients. A potential explanation for this may relate to the much lower patient sample numbers in studies analysing these associations up to now and to the intronic location of the respective SNPs.

In contrast to the two *common* SNPs discussed above, rs145370659 (A>C) was *less common* in our cohort. Our analysis indicated that patients carrying this SNP's minor allele had higher rates of *RD* compared to those carrying the major allele homozygously. rs145370659 is located on chromosome 1 in the intronic region of the Nardilysin Convertase (*NDRC*) gene.⁹⁸ To our knowledge, no studies have established a link between either *NDRC* and leukaemia or *NDRC* and drug metabolism. Further, rs145370659 falls to the genomic location of the microRNA *MIR671*.⁹⁸ Micro-RNAs participate in the post-transcriptional control of gene expression. They influence the stability and translation of messenger RNAs (mRNA) which promote gene transcription. *MIR671* has been described in different manuscripts to be associated with tumour development and proliferation of other forms of cancer. For example, studies have revealed that *MIR671* plays a regulating role in breast cancer, osteosarcomas and glioblastomas.¹¹¹⁻¹¹³ Therefore, an influence of rs145370659 on the proliferation of AML cannot be ruled out either. MicroRNAs, in general, are proven to influence drug metabolism and efficacy.¹¹⁴ This was not yet proven for *MIR671* in particular. Nevertheless, by being located on *MIR671*, rs145370659 might influence the chemotherapeutic drug metabolism in AML patients, leading to higher rates of *RD*.

However, our findings remain speculative as long as they are not validated in an independent data set.

Considering the results of the explorative analysis part, one can see that most associations existed between SNPs and gene mutations. In the next part, I will discuss the potential role of these associations.

Many SNPs were associated with gene mutations located on chromosomes other than the SNP. For clarity, we focussed on *common* and *less common* SNPs in recurrently mutated genes. With *ASXL1* and *TP53*, we found two AML mutations, incorporated in the *ELN2017* scoring (compare *table 1*), that were significantly associated with analysed SNPs. Some other SNP associations were found with genes that are in the present field of AML research, for instance, *IDH2* and *RAS*. We found the minor alleles of three SNPs to be associated with a higher occurrence of *IDH2* mutations. *IDH* inhibitors are one of the targets of new molecular therapies and first studies showed promising results for their success in *IDH* mutated AML patients.¹¹⁵ Also subject to current AML research have been *RAS* (*NRAS*, *KRAS*, *HRAS*)

mutations.¹¹⁶⁻¹¹⁸ We identified that the minor alleles of various SNPs are associated with an increased incidence of *NRAS* and *KRAS* mutations in our cohort. We further found several SNPs associated with *MLL* mutations. The latter are said to lead to decreased *RFS* periods and higher refractory rates.¹¹⁹

Even if all the associated genes potentially influence AML in different manners, SNP with gene mutation associations are not of high relevance in diagnostics or prognosis estimation of AML. The following applies to all of the above-described SNP with gene associations: Either they are associated with gene mutations that are already being screened during AML diagnostics, or the prognostic significance of the target gene in AML is unclear.

Nonetheless, SNP to gene associations should be viewed from another perspective: Genes located on the same chromosome are usually inherited together as long as there is no crossing over. On the other hand, genes located on different chromosomes are inherited unattached and uninfluenced by genes on other chromosomes.¹²⁰ Hence, the question arises as to how a SNP can be associated with a mutation that is found elsewhere on the genome. For instance, the incidence of *ASXL1* mutations on chromosome 20 is more than eleven times higher in patient samples carrying the minor allele of rs75395837 on chromosome eleven in the *ZBTB44* gene ($P < 0.0001$). Since the association resulted from the large number of 1,381 screened samples and the frequency of *ASXL1* mutations in AML is high, it is not likely to be coincidence. Concluding, the inheritance of mutations seems more complex than generally assumed. As an already known example the simultaneous inheritance of mutations is evident in Trisomy 21 patients. These patients have an above-average incidence of *Philadelphia-Chromosome-like Acute Lymphoblastic Leukaemia* without a genomic link between Trisomy 21 and the underlying translocation.¹²¹ In this context, the role of SNPs as potential influencing factors in mutation inheritance and predisposition should be considered more closely.

Validation Analyses of previously published clinically relevant SNPs

In the following part, I will discuss the relevance of the results from the validation analyses. I will mention reasons that may have contributed to the invalidation as well as possible sources of biases during the validation analyses.

I performed an intensive literature research by using the *dbSNP*, *SNPedia*, and *GWAS catalogue* databases to identify those SNPs associated with the outcome of AML in previously conducted studies. Due to the lack of a central platform that summarises all SNP studies, my literature search might have missed some relevant studies.

Out of the final 12 SNPs that were sufficiently covered in our cohort for validation, merely rs2454206⁸⁵ and rs7729269^{81,82} manifested a comparable association in the univariate validation analysis. However, the validation analyses of both associations could not be performed by exactly mimicking the respective previous studies. Furthermore, both associations were not significant after multivariate analysis. For most of the other reported SNP associations, we found a tendency of association, but the p-values did not approach the significance threshold of 0.1 in univariate analysis.

The non-validation of the prior study results could be explained by various reasons.

First, some literature results were based on cohorts that differed from our cohort either in terms of age or in terms of the type of leukaemia. Also, we reproduced studies that investigated parameters not assessed in our dataset and we associated the respective SNPs with clinically relevant parameters from our cohort. We chose to validate these associations with our differing data in order not to miss associations that overlapped between the different leukaemia cohorts.

Second, the prior results were, with the exception of two studies of Megías-Vericat et al., *Pharmacogenomics J.* (2015)⁸⁷ and Rafiee et al., *Blood Cancer J.* (2019)⁸⁸ based on small cohorts (compare *table 4*). These small cohorts were additionally stratified, further reducing the sample size analysed. Study results from small cohorts have a low statistical power.¹²² However, the size of our cohort should be relativised, too, as many regions and hereby many SNPs were only covered by about half of our cohort. Additionally, by stratification, our cohort size decreased analogously to the previous study cohorts.

Third, some of the published association papers did not inform on the applied association models. For performing the validation analysis, we had to deduce the chosen model from the results of the previous association analysis. These deductions might have led to other than the before applied association model.

All before described points could have served as source of biases during the validation process. Therefore, like previous studies, our results are not definitive and need to be interpreted with caution.

However, reinforced by the invalidation of several SNP associations, it can be concluded that results from small cohorts should only be considered as association tendencies as long as they have not been independently validated in a large cohort.

Also, results from stratified cohorts should be questioned since they cannot be transferred to those patients who differ from the specific investigated parameters. Referring to those studies that have repeatedly stratified their cohorts, one might conclude that the results are applicable to only a relatively small minority of AML patients. In addition, the inclusion of *ELN2017*, a meaningful stratification model, challenges association results from studies which did not adjust for pre-known risk models during multivariate analysis.

Our project encourages to critically reconsider the value of association studies on small cohorts. It further emphasises the importance of validating results with unrelated cohorts – especially in the context of currently established markers.

Clinical Relevance

Ensuing, I will outline the meaning of our results for the clinical management of AML patients.

Since *ELN* risk stratification models were established, the individualised therapy opportunities of AML patients vastly increased. *ELN2017* is an efficient scoring system, hence, additional prognostic biomarkers must show clear additive benefit. Such biomarkers could be of various origins. Our pilot study demonstrated that SNPs incorporated in current targeted sequencing panels cannot yet improve the further therapy individualisation, even though they are a promising area of research.

I would like to point out that, with regard to the 12 reproduced SNP association studies, we were the only group to carry out the association analysis in the context of *ELN2017*. The same applies to most of the prior studies, which we could not validate due to the lack of coverage, but most of which were also not adjusted for *ELN2017*. Hereby, they are not applicable in times of *ELN2017*. Furthermore, compared to previous AML SNP studies, our explorative results are especially clinically interesting because, in contrast to most of the other studies, they were based on the analysis of an unstratified AML cohort. This makes our results being transferable to all younger, intensively treated adult AML patients.

To be applicable, a biomarker must meet certain requirements. Low costs, quick results as well as high sensitivity and specificity are of main interest. All this vastly improved since Next Generation Sequencing techniques replaced Sanger sequencing techniques as standard.¹²³ Hereby, SNP screening of patients with AML in the clinical diagnostic routine is realistic and could, in principle, be easily incorporated in targeted sequencing panels.

Especially reliable biomarkers predicting the likelihood of *CR* achievement before the start of induction therapy would be of high value. rs2303430 could be one SNP that delivers this additional information. If the association could be validated in independent AML cohorts, studies on minor allele carriers can be conducted. Hereby, different treatment regimens could be compared in terms of outcome. In case these studies show positive results, screening for rs2303430 alleles could potentially improve the therapy success for AML patients as rs2303430 is a frequent SNP in AML cohorts. However, as mentioned before, since these findings were not yet validated, they should be considered preliminary and only hypothesis generating.

Besides the honest discussion about a patient's prognostic chances, the primary use of alternative second-line therapies in patients with an adverse constellation for *CR* achievement, could avoid burdensome therapies. Thus, in the best case, the survival rate could be improved through more individualised therapy, the patient's quality of life would not be unnecessarily impaired by measures that do not promise success and the health system would not be burdened.

Summarised, the results of our pilot study, particularly concerning rs2303430, encompassed promising associations relating to the outcome of AML patients. Yet, there is a need to conduct validation studies before they can evaluate for clinical use.

PERSPECTIVE

The here reported results can only be considered preliminary. The *BEAT* cohort, which had been used as a validation cohort up to this point, included genomic regions other than *AMLCG* and *AMLSG*. Furthermore, the *BEAT* cohort proved to be too small for being reliably used to validate the results of this study. Currently, under the direction of Dr. Aarif Nazeer Batcha, the results are being validated using the data from a further AML cohort. However, by May 2022, within our results no additional SNP associations were found that could have been validated with the additional dataset.

SUMMARY

The primary aim of this retrospective pilot study was to identify SNPs incorporated in targeted DNA sequencing panels of AML patients that can predict disease progression and/or therapy efficacy. Additionally, we intended to validate several SNP associations already described in previous publications.

This comprehensive analysis constituted one of the largest SNP projects in patient samples with AML but is limited to a small part of the genome. Data of 2,678 Northern European patients enrolled in phase III trials of the *AMLCG* and *AMLSG* study groups were included in the study. These patients sequencing data were first compared with a reference genome to identify variants. The variants were annotated, filtered and then quality-controlled. Ensuing, univariate association analysis was conducted, followed by multivariate analysis of the significant association results. For associated SNPs, sequencing data of the *BEAT* cohort served as validation.

With rs2303430 and rs28489067 (both located in the *PDGFRA* - gene) as well as rs145370659 (*NDRC* - gene), we found three previously unknown SNP associations with prognosis-predicting parameters among the analysed AML cohort. rs2303430 was associated with lower rates of *Complete Remission* and higher rates of *Early Death*. rs145370659 was more frequently found in patients who achieved *Complete Remission* while rs28489067 was associated with higher rates of *Resistant Disease*. Yet, these given associations could not be validated.

In the second part of our project, we reproduced various previously published SNP studies in our larger cohort. None of these previous association results could be validated. Since most of the published associations resulted from small and stratified cohorts, the invalidation of these SNPs in a large cohort underlines the relevance of large and homogeneously treated patient cohorts. It further highlights the need for independent validation of association analyses to achieve reliable and reproducible results.

If the SNP associations described as relevant in this project can be validated in further cohorts, they might serve as biomarkers in the future. The SNPs could be used in addition to the *ELN2017* risk scoring in clinical practice. Used at the time of diagnosis, they could gain importance in determining the therapeutic procedure and support the individualisation of AML patients' treatment.

ZUSAMMENFASSUNG

Das primäre Ziel dieser retrospektiven Pilotstudie war, anhand von DNA-Sequenzierungsdaten SNPs zu identifizieren, welche den Krankheitsverlauf sowie den Therapieerfolg von Patient*innen mit AML prognostizieren können. Ein weiteres Studienziel war die Validierung mehrerer SNP-Assoziationen, welche bereits in früheren Publikationen durch andere Autor*innen beschrieben wurden.

Diese umfassende Analyse stellt eines der größten an AML Patient*innen durchgeführten SNP-Assoziations-Projekte dar. Dennoch konnte in unserer Studie nur ein kleiner Teil des Genoms analysiert werden. Daten von 2678 nordeuropäischen Patient*innen, behandelt in Phase-III-Studien der AMLCG und AMLSG Studiengruppen, wurden eingeschlossen. Zunächst wurden die Sequenzierungsdaten dieser Patient*innen mit einem Referenzgenom verglichen, um Varianten zu identifizieren. Die Varianten wurden annotiert, gefiltert und ihre Qualität überprüft. Anschließend wurden univariate Assoziationsanalysen durchgeführt, gefolgt von multivariaten Analysen der signifikanten Assoziationsergebnisse. Für die assoziierten SNPs dienten Sequenzierungsdaten der *BEAT*-Kohorte als Validierung.

Mit den SNPs rs2303430 und rs28489067 (beide im *PDGFRA* - Gen lokalisiert), sowie rs145370659 (*NDRC* - Gen), fanden wir drei bisher unbekannte SNP-Assoziationen mit prognostisch relevanten Parametern in der untersuchten Kohorte. SNP rs2303430 war mit niedrigeren Raten von *kompletten Remissionen* und mit höheren Raten von *frühen Todesfällen* assoziiert. rs145370659 wurde häufiger bei Patient*innen gefunden, die eine *komplette Remission* erreichten, während rs28489067 mit höheren Raten einer *therapieresistenten Erkrankung* assoziiert war. Diese gefundenen Assoziationen konnten jedoch bisher nicht validiert werden.

Im zweiten Teil unseres Projekts reproduzierten wir verschiedene zuvor durchgeführte SNP-Studien in unserer größeren Kohorte. Wir konnten keine der beschriebenen Assoziationen validieren. Da die meisten der zuvor publizierten Assoziationen aus kleinen und stratifizierten Kohorten stammten, unterstrich deren nicht erfolgreiche Validierung erneut die Notwendigkeit, Forschungsprojekte an großen und homogen behandelten Patient*innengruppen durchzuführen. Es bestätigt zudem das Erfordernis einer unabhängigen Validierung von Assoziationsergebnissen, um belastbare Daten zu erzielen.

Vorausgesetzt, die von uns als relevant beschriebenen SNP-Assoziationen können in weiteren unabhängigen Kohorten validiert werden, könnten sie in Zukunft als Biomarker dienen. Die SNPs könnten zusätzlich zu der *ELN2017* Risikostratifizierung im klinischen Alltag angewendet werden. Bereits bei Diagnosestellung eingesetzt, könnten sie Bedeutung bei der Festlegung des therapeutischen Prozedere gewinnen und die Individualisierung der Behandlung von AML-Patient*innen fördern.

LIST OF TABLES

		Page:
<i>Table 1</i>	<i>ELN2017</i> Cytogenetic Risk Categories.	6
<i>Table 2</i>	Cinical and Laboratory Patient Data.	15f.
<i>Table 3</i>	Allelic Association Model.	24
<i>Table 4</i>	Detailed Literature Studies Information.	28ff.
<i>Table 5</i>	Genotypic Association Model.	31
<i>Table 6</i>	SNP Associations with Gene Mutations.	39
<i>Table 7</i>	Previous Studies' Validation Results.	40f.

LIST OF FIGURES

<i>Figure 1</i>	SNPs.	9
<i>Figure 2</i>	Sequenced Genes.	17
<i>Figure 3</i>	Cohort Composition and Data Preparation.	18
<i>Figure 4</i>	Workflow until Plink.	22
<i>Figure 5</i>	Quality Control and Association Analysis in Plink.	25
<i>Figure 6</i>	SNPs from previous Studies.	27
<i>Figure 7</i>	SNV Distribution.	34
<i>Figure 8</i>	Explorative Multivariate Association Results Overview.	35
<i>Figure 9a</i>	rs2303430~CR.	36
<i>Figure 9b</i>	rs2303430~ED.	36
<i>Figure 10</i>	rs28489067~CR.	37
<i>Figure 11</i>	rs145370659~RD.	37
<i>Figure 12a</i>	rs2454206~OS.	42
<i>Figure 12b</i>	rs2454206~RFS.	43

GLOSSARY

This is an explanation of terms, in what meaning they were used in this work.

Allele	The different variants (bases) that exist for one SNV/SNP.
Biomarker	Variable that can be measured to forecast a particular outcome of the disease.
Called Variants/ SNVs/ SNPs	Genomic locations that were identified as different compared to the reference genome during variant calling.
Clinical variables/ parameters/ characteristics/	Disease factors that can be measured from the patient's record. Comparable to <i>Outcome</i> .
ELN2017	Risk scoring system from the European Leukemia Net that classifies AML by genetical characteristics.
Exome	The totality of the exons of an organism.
Exon	Section of a gene containing the information necessary to produce proteins.
Genome	All genetic information present in a cell.
Genomic location	Specifies the physical location of a SNV/SNP on the genome
Genotype	The two alleles present for a particular SNV/SNP in one individual.
Heterozygous	Carrying two different alleles at a particular SNV/SNP location.
Homozygous	Carrying two identical alleles at a particular SNV/SNP location.
Intron	The non-coding sections of DNA within a gene.

Mapping	The genetical comparison between the reference genome and each individual of the cohort.
Minor Allele Frequency	The frequency with which the second most common allele appears in a given population.
Multivariate analysis	Association analysis of data that contain different variables which might influence each other. In this study, the term describes association analysis with adjustment for age and <i>ELN2017</i> . This was only done for associations that were significant ($p < 0.1$) in univariate analysis.
Outcome	Measurable parameters that describe the health status and the therapy success of a patient. In our analysis mainly: <i>Overall Survival, Refractory-Free Survival, Complete Remission, Early Death and Resistant Disease</i> .
rs...	Prefix for SNVs/SNPs (see SNV/SNP). <i>rs=reference SNP cluster ID</i> .
SNP	SNV that occurs in >1% of the organisms in a given population.
SNV	Variation of a single base pair in a complementary DNA double strand.
Stratification	Association analysis on parts of a cohort only. Often performed in studies found in publications. It was mainly stratified for: primary AML, Normal Karyotype AML, or patients carrying certain gene mutations.
Univariate analysis	Association analysis between two parameters. In this study it was done for every called SNV and every clinical outcome parameter as well as every gene mutation evaluated in the cohort.

REFERENCES

1. Ries L, Harkins D et al. SEER Cancer Statistics Review 1975-2003. *National Cancer Institute*. Bethesda. Available from: https://seer.cancer.gov/csr/1975_2003/. 2006.
2. Sitzmann FC, Bauer C-P. Pädiatrie. Hippokrates-Verlag. Stuttgart. 1995.
3. Roushangar R, Mias GI. Multi-study reanalysis of 2,213 acute myeloid leukemia patients reveals age- and sex-dependent gene expression signatures. *Sci Rep*. 2019;9(1):12413.
4. Juliusson G, Antunovic P, et al. Age and acute myeloid leukemia: real world data on decision to treat and outcomes from the Swedish Acute Leukemia Registry. *Blood*. 2009;(113):4179–4187.
5. Smith MT, Skibola CF, Allan JM, Morgan GJ. Causal models of leukaemia and lymphoma. *IARC Sci Publ*. 2004;(157):373-392.
6. Ghiaur G, Wroblewski M, Loges S. Acute Myelogenous Leukemia and its Microenvironment: A Molecular Conversation. *Semin Hematol*. 2015;52(3):200-206.
7. Arber D, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-405.
8. Irons RD, Stillman WS. The process of leukemogenesis. *Environ Health Perspect*. 1996;104(Suppl 6):1239-1246.
9. Ellegast JM, Gerber B, Manz M. Diagnose und Therapie der akuten myeloischen Leukämie. *Swiss Medical Forum*. 2013;13(06):112-119.
10. Röllig C, Thiede C, Ehninger G. Acute myeloid leukemia. *Springer Medizin*. München. 2017.
11. Lowenberg B, Downing JR, Burnett A. Acute Myeloid Leukemia. *Massachusetts Medical Society*. 1999;341(14):1051-62.
12. Vyas C, Jain S, Kapoor G. Therapy Related AML/MDS Following Treatment for Childhood Cancer: Experience from a Tertiary Care Centre in North India. *Springer India*. 2018;34(1):78-82.
13. Beebe G, Kato H, Land C. Studies of the mortality of A-bomb survivors: 6. mortality and radiation dose, 1950-1974. *Radiation Research*. 1978;(75):138-201.
14. Austin H, Delzell E, Cole P. Benzene and leukemia. A review of the literature and a risk assessment. *Am J Epidemiol*. 1988;127(3):419-439.
15. Brownson R, Chang J, Davis J. Cigarette smoking and risk of adult leukemia. *Am J Epidemiol*. 1991;134(9):938-941.
16. Poynter JN, Richardson M, Roesler M, et al. Chemical exposures and risk of acute myeloid leukemia and myelodysplastic syndromes in a population-based study. *Int J Cancer*. 2017;140(1):23-33.
17. Roizen NJ, Patterson D. *Down's syndrome*. Elsevier Ltd; 2003;361(9365):1281-1289.
18. Porter CC. Germ line mutations associated with leukemias. *Hematology Am Soc Hematol Educ Program*. 2016;2016(1):302-308.
19. Banno K, Omori S, Hirata K, et al. Systematic Cellular Disease Models Reveal Synergistic Interaction of Trisomy 21 and GATA1 Mutations in Hematopoietic Abnormalities. *Cell Rep*. 2016;15(6):1228-1241.
20. Gao J, Gentzler RD, Timms AE. Heritable GATA2 mutations associated with familial AML-MDS: a case report and review of literature. *BioMed Central*. 2014; 7(36).

21. Theis F, Corbacioglu A, Gaidzik VI, et al. Clinical impact of GATA2 mutations in acute myeloid leukemia patients harboring CEBPA mutations: a study of the AML study group. *Leukemia*. 2016;30(11):2248-2250.
22. Roumier C, Fenaux P, Lafage M, Imbert M, Eclache V, Preudhomme C. New mechanisms of AML1 gene alteration in hematological malignancies. *Leukemia*. 2003;17(1):9-16.
23. Leone G, et al. The incidence of secondary leukemias. *Haematologica*. 1999;84(10):937-945.
24. Andersen M, et al. Chromosomal abnormalities in secondary MDS and AML. Relationship to drugs and radiation with specific emphasis on the balanced rearrangements. *Haematologica* 1998;83(6):483-488.
25. Pedersen-Bjergaard J, et al. Therapy-related myelodysplasia and acute myeloid leukemia. Cytogenetic characteristics of 115 consecutive cases and risk in seven cohorts of patients treated intensively for malignant diseases in the Copenhagen series. *Leukemia*. 1993;7(12):1975-1986.
26. Döhner H, Estey E, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*. 2010;115(3):453-474.
27. 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.
28. Heilmeier B, Buske C, Spiekermann K, et al. Diagnostics, classification and prognostic criteria of acute myeloid leukemia. *Med Klin (Munich)*. 2007;102(4):296-308.
29. Leisch M, Jansko B, Zaborsky N. Next Generation Sequencing in AML-On the Way to Becoming a New Standard for Treatment Initiation and/or Modulation?. *Cancers*. 2019;11(2):252.
30. Yohe S, Ustun C, Godley LA. Molecular Genetic Markers in Acute Myeloid Leukemia. *J. Clin Med*. 2015;4(3):460-478.
31. Döhner H, Weisdorf DJ, Bloomfield CD. Acute Myeloid Leukemia. *N Engl J Med*. 2015;373(12):1136-1152.
32. Mehta S, Shelling A, Muthukaruppan A, et al. Predictive and prognostic molecular markers for cancer medicine. *Ther Adv Med Oncol*. 2010;2(2):125-148.
33. Hirshfield K. SNPs as prognostic indicators in cancer outcomes. *Pharmacogenetics*. 2008;14(15):4
34. Fröhling S, Schlenk RF, Kayser S, et al. Cytogenetics and age are major determinants of outcome in intensively treated acute myeloid leukemia patients older than 60 years: results from AMLSG trial AML HD98-B. *Blood*. 2006;108(10):3280-3288.
35. Cheson BD, Cassileth PA, Head DR, et al. Report of the National Cancer Institute-sponsored workshop on definitions of diagnosis and response in acute myeloid leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 1990;8(5):813-819.
36. Schoch C, Haferlach T. Cytogenetics in acute myeloid leukemia. *Curr Oncol Rep*. 2002;4(5):390-397.
37. Cheson BD, Bennett JM, Kopecky KJ, et al. Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute

- Myeloid Leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2003;21(24):4642-4649.
38. Winer ES, Stone RM. Novel therapy in Acute myeloid leukemia (AML): moving toward targeted approaches. *Ther Adv Hematol*. 2019;10:1-18
 39. Bornhauser M, Ehninger G. Diagnostics and therapy of acute myeloid leukemia. *Dtsch Med Wochenschr*. 2009;134(39):1935-1941.
 40. Barouch S. Maintenance Therapy in Acute Myeloid Leukemia: An Overview. *Hematology Advisor*. Available from: <https://www.hematologyadvisor.com/home/topics/leukemia/treatment-strategies-for-maintenance-therapy-in-acute-myeloid-leukemia/>. 2019.
 41. Brookes AJ. The essence of SNPs. *Gene*. 1999;234(2):177-186.
 42. Witte JS. Genome-wide association studies and beyond. *Annu Rev Public Health*. 2010;31:9-20.
 43. Deng N, Zhou H, Fan H, Yuan Y. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget*. 2017;8(66):110635-110649.
 44. Spencer D, Zhang B, Pfeifer J. Single Nucleotide Variant Detection Using Next Generation Sequencing. *Clinical Genomics*. 2015;109-127.
 45. Dictionary of Genetics Terms. *National Cancer Institute*. Available from: <https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/snv>. 2021.
 46. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;409(6822):928-933.
 47. Jackson M, Marks L, May GHW, Wilson JB. The genetic basis of disease. *Essays Biochem*. 2018;62(5):643-723.
 48. Nebert DW. Suggestions for the nomenclature of human alleles: relevance to ecogenetics, pharmacogenetics and molecular epidemiology. *Pharmacogenetics*. 2000;10(4):279-290.
 49. Cooper DN. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. *Hum Genomics*. 2010;4(5):284-288.
 50. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311.
 51. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research*. 2012;40:1308-1312.
 52. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(1):1005-1012.
 53. Landrum MJ, Lee JM, Benson M. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46:1062-1067.
 54. Ferguson P, Hills RK, Grech A, et al. An operational definition of primary refractory acute myeloid leukemia allowing early identification of patients who may benefit from allogeneic stem cell transplantation. *Haematologica*. 2016;101(11):1351-1358.
 55. Sabattini E, Bacci F, Sagrmoso C, Pileri SA. WHO classification of tumours of haematopoietic and lymphoid tissues: an overview. *Pathologica*. 2010;102(3):83-87.
 56. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama*. 2013;310(20):2191-2194.
 57. 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):686-698.
58. Papaemmanuil E, Gerstung M, Bullinger L. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med*. 2016;374(23):2209-2221.
 59. Herold T, Rothenberg-Thurley M, Grunwald VV, et al. Validation and Refinement of the Revised 2017 European LeukemiaNet Genetic Risk Stratification of Acute Myeloid Leukemia. *Leukemia*. 2020;34(12):3161-3172.
 60. Büchner T, Berdel WE, Schoch C, et al. Double induction containing either two courses or one course of high-dose cytarabine plus mitoxantrone and postremission therapy by either autologous stem-cell transplantation or by prolonged maintenance for acute myeloid leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2006;24(16):2480-2489.
 61. Büchner T, Krug UO, Peter Gale R, et al. Age, not therapy intensity, determines outcomes of adults with acute myeloid leukemia. *Leukemia*. 2016;30(8):1781-1784.
 62. Braess J, Amler S, Kreuzer KA, et al. Sequential high-dose cytarabine and mitoxantrone (S-HAM) versus standard double induction in acute myeloid leukemia-a phase 3 study. *Leukemia*. 2018;32(12):2558-2571.
 63. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*. 2018;562(7728):526-531.
 64. Mrózek K. Cytogenetic, molecular genetic, and clinical characteristics of acute myeloid leukemia with a complex karyotype. *Semin Oncol*. 2008;35(4):365-377.
 65. Church DM, Schneider VA, Graves T, et al. Modernizing reference genome assemblies. *PLoS Biol*. 2011;9(7):1-5.
 66. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
 67. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
 68. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.
 69. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92.
 70. Paila U, Chapman BA, Kirchner R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput Biol*. 2013;9(7):n.a.
 71. Purcell S, Neale B, Todd-Brown K. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *AM J Hum Genet*. 2017;81(3):559-575.
 72. Purcell S, Chang CC. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga Science*. 2015;4(1).
 73. Noguera NI, Catalano G, Banella C, et al. Acute Promyelocytic Leukemia: Update on the Mechanisms of Leukemogenesis, Resistance and on Innovative Treatment Strategies. *Cancers (Basel)*. 2019;11(10):1591.
 74. Smith G, Newton-Cheh C. Genome-wide association study in humans. *Methods Mol Biol*. 2009;573:231-258.
 75. Van den Berg S, Vandenplas J, van Eeuwijk FA, Lopes MS, Veerkamp RF. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. *J Anim Breed Genet*. 2019;136(6):418-429.

76. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? *Korean J Anesthesiol.* 2018;71(5):353-360.
77. Kutny MA, Alonzo TA, Gamazon ER, et al. Ethnic variation of TET2 SNP rs2454206 and association with clinical outcome in childhood AML: a report from the Children's Oncology Group. *Leukemia.* 2015;29(12):2424-2426.
78. Falk IJ, Fyrberg A, Paul E. Decreased survival in normal karyotype AML with single-nucleotide polymorphisms in genes encoding the AraC metabolizing enzymes cytidine deaminase and 5'-nucleotidase. 2013;88(12):1001-1006.
79. Megías-Vericat JE, Montesinos P, Herrero MJ. Influence of cytarabine metabolic pathway polymorphisms in acute myeloid leukemia induction treatment. *Leuk Lymphoma.* 2017;58(12):2880-2894.
80. Kim LH, Cheong HS, Koh Y. Cytidine deaminase polymorphisms and worse treatment response in normal karyotype AML. *J Hum Genet.* 2015;60(12):749-754.
81. Gamazon ER, Lamba JK, Pounds S, et al. Comprehensive genetic analysis of cytarabine sensitivity in a cell-based model identifies polymorphisms associated with outcome in AML patients. *Blood.* 2013;121(21):4366-4376.
82. Megías-Vericat JE, Montesinos P, Herrero MJ. Impact of novel polymorphisms related to cytotoxicity of cytarabine in the induction treatment of acute myeloid leukemia. 2017;27(7):270-274.
83. Wagner K, Damm F, Gohring G, et al. Impact of IDH1 R132 mutations and an IDH1 single nucleotide polymorphism in cytogenetically normal acute myeloid leukemia: SNP rs11554137 is an adverse prognostic factor. *J Clin Oncol.* 2010;28(14):2356-2364.
84. Ho PA, Kopecky KJ, Alonzo TA, et al. Prognostic implications of the IDH1 synonymous SNP rs11554137 in pediatric and adult AML: a report from the Children's Oncology Group and SWOG. *Blood.* 2011;118(17):4561-4566.
85. Wang X, Chen X, Yang Z, et al. Correlation of TET2 SNP rs2454206 with improved survival in children with acute myeloid leukemia featuring intermediate-risk cytogenetics. *Genes Chromosomes Cancer.* 2018;57(8):379-386.
86. Kadia T, Kantarjian H, Ravandi F, et al. Prognostic Significance of the Medical Research Council (MRC) Cytogenetic Classification Compared with the European LeukaemiaNet (ELN) Risk Classification System in Acute Myeloid Leukaemia (AML). *British Journal of Haematology.* 2015;170(4):590-593.
87. Megías-Vericat JE, Rojas L, Herrero MJ. Influence of ABCB1 polymorphisms upon the effectiveness of standard treatment for acute myeloid leukemia: A systematic review and meta-analysis of observational studies. *Pharmacogenomics J.* 2015;15(2):109-118.
88. Rafiee R, Chauhan L, Alonzo TA, et al. ABCB1 SNP predicts outcome in patients with acute myeloid leukemia treated with Gemtuzumab ozogamicin: a report from Children's Oncology Group AAML0531 Trial. *Blood Cancer J.* 2019;9(6):51.
89. Gregers J, Gréen H, Christensen IJ. Polymorphisms in the ABCB1 gene and effect on outcome and toxicity in childhood acute lymphoblastic leukemia. *Pharmacogenomics J.* 2015;15(4):372-379.
90. Dessilly G, Elens L, Panin N, Karmani L, Demoulin JB, Haufroid V. ABCB1 1199G>A polymorphism (rs2229109) affects the transport of imatinib, nilotinib and dasatinib. *Pharmacogenomics.* 2016;17(8):883-890.
91. Bauduer F, Ducout L, Dastugue N, Capdupuy C, Renoux M. De novo and secondary acute myeloid leukemia in patients over the age of 65: a review of fifty-six successive and unselected cases from a general hospital. *Leuk Lymphoma.* 1999;35(3-4):289-

296.

92. Hulegardh E, Nilsson C, Lazarevic V, et al. Characterization and prognostic features of secondary acute myeloid leukemia in a population-based setting: a report from the Swedish Acute Leukemia Registry. *Am J Hematol*. 2015;90(3):208-214.
93. Acharya UH, Halpern AB, Wu QV, et al. Impact of region of diagnosis, ethnicity, age, and gender on survival in acute myeloid leukemia (AML). *J Drug Assess*. 2018;7(1):51-53.
94. Shysh AC, Nguyen LT, Guo M, Vaska M, Naugler C, Rashid-Kolvear F. The incidence of acute myeloid leukemia in Calgary, Alberta, Canada: a retrospective cohort study. *BMC Public Health*. 2017;18(1):94.
95. Database of Single Nucleotide Polymorphisms (dbSNP) rs2303430. Build: 153. *National Library of Medicine*. Available from: <https://www.ncbi.nlm.nih.gov/snp/rs2303430>. 2020.
96. Database of Single Nucleotide Polymorphisms (dbSNP) rs28489067. Build: 153. *National Library of Medicine*. Available from: <https://www.ncbi.nlm.nih.gov/snp/rs28489067>. 2020.
97. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.
98. Database of Single Nucleotide Polymorphisms (dbSNP) rs145370659. Build: 153. *National Library of Medicine*. Available from: <https://www.ncbi.nlm.nih.gov/snp/rs145370659>. 2020.
99. Akiyama M. ABCA12 mutations and autosomal recessive congenital ichthyosis: a review of genotype/phenotype correlations and of pathogenetic concepts. *Hum Mutat*. 2010;31(10):1090-1096.
100. Kelsell DP, Norgett EE, Unsworth H, et al. Mutations in ABCA12 underlie the severe congenital skin disease harlequin ichthyosis. *Am J Hum Genet*. 2005;76(5):794-803.
101. Mahlknecht U, Dransfeld CL, Bulut N, et al. SNP analyses in cytarabine metabolizing enzymes in AML patients and their impact on treatment response and patient survival: identification of CDA SNP C-451T as an independent prognostic parameter for survival. *Leukemia*. 2009;23(10):1929-1932.
102. Isidori A, Venditti A, Maurillo L, et al. Alternative novel therapies for the treatment of elderly acute myeloid leukemia patients. *Expert Rev Hematol*. 2013;6(6):767-784.
103. Metzker ML. Sequencing technologies — the next generation. *Nat Ref Genet*. 2009;11(1):31-46.
104. PDGFRA gene. *Medline Plus, National Library of Medicine*. Bethesda. Available from: <https://ghr.nlm.nih.gov/gene/PDGFRA>. 2014
105. PDGFRA platelet derived growth factor receptor alpha [Homo sapiens (human)]. *National Library of Medicine*. Bethesda. Available from: <https://www.ncbi.nlm.nih.gov/gene/5156>. 2020.
106. PDGFRA Gene. *Gene Cards, the Human Gene Database*. Available from: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=PDGFRA>. 2020.
107. Berenstein R. Class III Receptor Tyrosine Kinases in Acute Leukemia - Biological Functions and Modern Laboratory Analysis. *Biomark Insights*. 2015;10(Suppl 3):1-14.
108. Metzgeroth G, Walz C, Score J, et al. Recurrent finding of the FIP1L1-PDGFRA fusion gene in eosinophilia-associated acute myeloid leukemia and lymphoblastic T-cell lymphoma. *Leukemia*. 2007;21(6):1183-1188.

109. Buitenhuis M, Verhagen LP, Cools J, Coffey PJ. Molecular mechanisms underlying FIP1L1-PDGFR α -mediated myeloproliferation. *Cancer Res.* 2007;67(8):3759-3766.
110. PDGFR α -associated chronic eosinophilic leukemia. *Medline Plus, National Library of Medicine*. Bethesda. Available from: <https://medlineplus.gov/genetics/condition/pdgfra-associated-chronic-eosinophilic-leukemia/>. 2020.
111. Xin C, Lu S, Li Y, et al. miR-671-5p Inhibits Tumor Proliferation by Blocking Cell Cycle in Osteosarcoma. *DNA Cell Biol.* 2019;38(9):996-1004.
112. Barbagallo D, Condorelli A, Ragusa M, et al. Dysregulated miR-671-5p / CDR1-AS / CDR1 / VSNL1 axis is involved in glioblastoma multiforme. *Oncotarget.* 2016;7(4):4746-4759.
113. Tan X, Fu Y, Chen L, et al. miR-671-5p inhibits epithelial-to-mesenchymal transition by downregulating FOXM1 expression in breast cancer. *Oncotarget.* 2016;7(1):293-307.
114. He Y, Chevillet JR, Liu G, Kim TK, Wang K. The effects of microRNA on the absorption, distribution, metabolism and excretion of drugs. *Br J Pharmacol.* 2015;172(11):2733-2747.
115. Yen K, Travins J, Wang F, et al. AG-221, a First-in-Class Therapy Targeting Acute Myeloid Leukemia Harboring Oncogenic IDH2 Mutations. *Cancer Discov.* 2017;7(5):478-493.
116. Dunna NR, Vuree S, Anuradha C, et al. NRAS mutations in de novo acute leukemia: prevalence and clinical significance. *Indian J Biochem Biophys.* 2014;51(3):207-210.
117. Bacher U, Haferlach T, Schoch C, Kern W, Schnittger S. Implications of NRAS mutations in AML: a study of 2502 patients. *Blood.* 2006;107(10):3847-3853.
118. Bowen DT, Frew ME, Hills R, et al. RAS mutation in acute myeloid leukemia is associated with distinct cytogenetic subgroups but does not influence outcome in patients younger than 60 years. *Blood.* 2005;106(6):2113-2119.
119. Kotani S, Yoda A, Kon A, et al. Molecular pathogenesis of disease progression in MLL-rearranged AML. *Leukemia.* 2019;33(3):612-624.
120. Miko I. Mitosis, meiosis, and inheritance. *Nature Education.* 2008;1(1):206.
121. Pui CH, Roberts KG, Yang JJ, Mullighan CG. Philadelphia Chromosome-like Acute Lymphoblastic Leukemia. *Clin Lymphoma Myeloma Leuk.* 2017;17(8):464-470.
122. Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafò MR. Low statistical power in biomedical science: a review of three human research domains. *R Soc Open Sci.* 2017;4(2):160254.
123. Marino P, Touzani R, Perrier L, et al. Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study. *Eur J Hum Genet.* 2018;26(3):314-323.

ACKNOWLEDGEMENT

At the beginning, when I embarked on composing this thesis, I felt like being on a fragile sailing boat with which I had to cross the ocean, to reach a far-away place whilst not even knowing how the practice of sailing works and not having a crew that could help me navigate deep seas and overcome stormy weathers.

Then, I realised that, actually, I had this much-needed crew. Now, after having spent the past three years together with this very crew, I would like to deeply thank all of those crew members without whom I could not have reached this far-away doctoral place:

You, first officer, Prof. Dr Tobias Herold, you were the best doctorate supervisor anyone could ask for. Thank you for all the meetings, the tips with regards to the technical approach to this project, the time you took for proofreading, and, above all, for your motivational impetus! I am grateful to you for explaining me the navigation I needed to reach my destination.

You, steersman Dr Aarif Nazeer Batcha for the two years in which you supported me with your skills in statistics and information technology. Thank you for always keeping the door of your steering house open to me. During the whole journey, you were a loyal companion, and sometimes took the rudder when a storm appeared on the horizon.

PD Dr. Klaus Metzeler for the technical support of the project. You had a special eye for the fine-tuning and made yourself noticeable just before I almost mixed up the wind directions.

You, Markus, were always by my side. You were a strong buoy, buffering everything. Thank you for never allowing me to enter a harbour that was before my actual destination for the day. You were also the ship's cook, making sure that I was strong enough not to let the ship run aground.

My siblings Eike and Rieke, who joined me on the boat towards the end of the passage. Without you, this work could not have been written fluently. Thank you for your linguistic support and the continuous emotional and motivational assistance.

A special thanks goes also to my parents Jutta and Jan! You built the ship and placed it with me at the start. You supported me all the route and repeatedly mended the sails. Thank you for being there for me day and night! The biggest thanks goes to my mum for supporting this project up to the very last minute.

Without all of you, I would never have been able to finish this long and exhausting yet fruitful and fulfilling sailing trip.

PUBLIKATIONSLISTE

1. Vortrag beim *Herrsching Meeting* im Juli 2019 mit Präsentation der Kohorte, des Projektziels sowie erster, nicht im Fließtext dargestellter Zwischenergebnisse:

„Analysis of “Single Nucleotide Polymorphisms” (SNPs) in Patients with Acute Myeloid Leukemia“

2. Poster zum Herrsching Meeting 2019:

„Analysis of “Single Nucleotide Polymorphisms” (SNPs) in Patients with Acute Myeloid Leukemia“

A. M. N. Batcha^{1,2 *}, N. Buckup^{3*}, S. A. Bamopoulos³, V. Jurinovic^{1,3}, M. Rothenberg-Thurley³, H. Janke³, B. Ksienzyk³, S. Schneider^{3,4}, N. Konstandin³, M. C. Sauerland⁵, D. Görlich⁵, W. E. Berdel⁶, B. J. Woermann⁷, J. Braess⁸, U. Mansmann^{1,2,9,10}, W. Hiddemann^{3,9,10}, K. Spiekermann^{3,9,10}, K. H Metzeler^{3,9,10,§} and T. Herold^{3,9-11,§}

¹ Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich

² Medical Data Integration Center (MeDIC), University Hospital, LMU Munich, Germany

³ Laboratory for Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Germany

⁴ Institute of Human Genetics, University Hospital, LMU Munich, Munich, Germany

⁵ Institute of Biostatistics and Clinical Research, University of Münster

⁶ Department of Medicine, Hematology and Oncology, University of Münster

⁷ German Society of Hematology and Oncology, Berlin

⁸ Department of Oncology and Hematology, Hospital Barmherzige Brüder, Regensburg

⁹ German Cancer Consortium (DKTK), Heidelberg

¹⁰ German Cancer Research Center (DKFZ), Heidelberg

¹¹ Research Unit Apoptosis in Hematopoietic Stem Cells, Helmholtz Zentrum München, German Center for Environmental Health (HMGU)

* ,§ contributed equally

3. Publikation, aktuell in Revision:

„Single nucleotide polymorphisms and outcomes in intensively treated acute myeloid leukemia – a validation study“

Aarif M. N. Batcha^{1,2}, Nele Buckup³, Stefanos A. Bamopoulos^{3,4}, Vindi Jurinovic^{1,3}, Maja Rothenberg-Thurley³, Hanna Janke³, Bianka Ksienzyk³, Annika Dufour³, Stefanie Schneider^{3,5}, Mika Kontro⁶, Joseph Saad⁷, Caroline A. Heckmann⁷, Cristina M. Sauerland⁸, Dennis Görlich⁸, Wolfgang E. Berdel⁹, Bernhard J. Woermann¹⁰, Utz Krug¹¹, Jan Braess¹², Ulrich Mansmann^{1,2,13,14}, Wolfgang Hiddemann^{3,13,14}, Karsten Spiekermann^{3,13,14}, Klaus H Metzeler^{3,13,14,15*} and Tobias Herold^{3,13,14,*}

¹ Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany

² DIFUTURE, Data integration for Future Medicine (DiFuture, www.difuture.de), LMU Munich, Munich, Germany

³ Laboratory for Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Germany

⁴ Department of Hematology, Oncology and Tumor Immunology (Campus Benjamin Franklin), Charité University Medicine Berlin, Berlin, Germany

⁵ Institute of Human Genetics, University Hospital, LMU Munich, Munich, Germany

⁶ Department of Haematology, Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland

⁷ Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

⁸ Institute of Biostatistics and Clinical Research, University of Münster

⁹ Department of Medicine, Hematology and Oncology, University of Münster, Germany

¹⁰ German Society of Hematology and Oncology, Berlin, Germany

¹¹ Department of Medicine III, Hospital Leverkusen, Leverkusen, Germany

¹² Department of Oncology and Hematology, Hospital Barmherzige Brüder, Regensburg, Germany




¹³ German Cancer Consortium (DKTK), Heidelberg, Germany

¹⁴ German Cancer Research Center (DKFZ), Heidelberg, Germany

¹⁵ Department of Hematology and Cellular Therapy, University Hospital Leipzig, Leipzig, Germany

* contributed equally

AFFIDAVIT

	LUDWIG- MAXIMILIANS- UNIVERSITÄT MÜNCHEN	Promotionsbüro Medizinische Fakultät		
Eidesstattliche Versicherung				

Buckup, Nele

Name, Vorname

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel:

Pilot Analysis of Single Nucleotide Polymorphisms in Patients with Acute Myeloid Leukaemia identified by Targeted DNA Sequencing

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.

Kiel, 08.12.2022

Ort, Datum

Nele Buckup

Unterschrift Doktorandin