RNA sequencing of acute myeloid leukemia (AML) applied to transcript quantification and fusion gene detection

Dissertation
at the Faculty of Biology
of the Ludwig Maximilian University
Munich

by

Paul Kerbs

Munich, 2021

First reviewer:	Prof. Dr. Dirk Metzler
Second reviewer:	PD Dr. Philipp Greif
Date of submission:	30.06.2021
Date of oral examination:	30.09.2021

Statutory declaration and statement

I herewith declare that I have composed the present thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance.

Munich, 20.10.2021

Place, Date

Paul Kerbs

Table of contents

Glossary

A

ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase
AML	Acute myeloid leukemia
APL	Acute promyelocytic leukemia
ATO	Arsenic trioxide
ATRA	All-trans retinoic acid
AURKA	Aurora kinase A

B

BAALC	Brain and acute leukemia cytoplasmic
BCR	Breakpoint cluster region gene

C

CBFB	Core-binding factor subunit beta
CCNA2	Cyclin A2
CD34	CD34 molecule (cluster of differentiation)
CENPF	Centromere protein F
CML	Chronic myeloid leukemia
CN	Cytogenetically normal
CR	Complete remission

D

DEK	DEK proto-oncogene
DNA-seq	DNA sequencing
DNMT3A	DNA methyltransferase 3 alpha

E

ELN	European LeukemiaNet
ERG	ETS (erythroblast transformation-specific) transcription factor

F

FAB	French-American-British
FES	Feline sarcoma proto-oncogene, tyrosine kinase
FISH	Fluorescence *in situ* hybridization
FLT3	FMS-related receptor tyrosine kinase 3
FOS	Fos proto-oncogene, AP-1 transcription factor subunit
FTS	Fusion Transcript Score

G

GATA2	GATA binding protein 2
GTF2I	General transcription factor IIi

I

IDH	Isocitrate dehydrogenase
ISWI	Imitation switch

K

KIT	KIT proto-oncogene, receptor tyrosine kinase
KMT2A	Lysine methyltransferase 2A

L

LINC00173	Long intergenic non-protein coding RNA 173
lncRNA	Long non-coding RNA

M

MECOM	MDS1 and EVI1 complex locus
MIR99AHG	mir-99a-let-7c cluster host gene
MN1	Meningioma (disrupted in balanced translocation) 1 proto-oncogene, transcriptional regulator
MYB	MYB proto-oncogene, transcription factor
MYC	MYC proto-oncogene, bHLH transcription factor
MYH11	Myosin heavy chain 11

N

ncRNA	Non-coding RNA
NGS	Next-generation sequencing
NPM1	Nucleophosmin 1
NRIP1	Nuclear receptor-interacting protein 1
NUP214	Nucleoporin 214

P

PCR	Polymerase chain reaction
PIM	Pim-1 proto-oncogene, serine/threonine kinase
PLK1	polo like kinase 1
PLZF	Zinc finger and BTB domain containing 16
PML	PML nuclear body scaffold
PS	Promiscuity Score

Q

qPCR	Quantitative PCR

R

RARA	Retinoic acid receptor alpha
RNA-seq	RNA sequencing
RS	Robustness Score
RUNX1	RUNX family transcription factor 1
RUNX1T1	RUNX1 partner transcriptional co-repressor 1

S

SCT	Stem cell transplantation
SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5
SNV	Single-nucleotide variant
SV	Structural variant

T

TPM	Transcripts per million

W

WHO	World Health Organization
WT1	Wilms tumor 1, transcription factor

List of publications

- **Kerbs, P.**, Vosberg, S., Krebs, S., Graf, A., Blum, H., Swoboda, A., Batcha, A. M. N., Mansmann, U., Metzler, D., Heckman, C. A., Herold, T., & Greif, P. A. (2021). Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring NRIP1-MIR99AHG rearrangements. *Haematologica*, https://doi.org/10.3324/haematol.2021.278436 [Early view].

- Redondo Monte, E., **Kerbs, P.**, & Greif, P. A. (2020). ZBTB7A links tumor metabolism to myeloid differentiation. *Experimental Hematology*, 87, 20–24.e1. https://doi.org/10.1016/j.exphem.2020.05.010

- Zikmund, T., Paszekova, H., Kokavec, J., **Kerbs, P.**, Thakur, S., Turkova, T., Tauchmanova, P., Greif, P. A., & Stopka, T. (2020). Loss of ISWI ATPase SMARCA5 (SNF2H) in Acute Myeloid Leukemia Cells Inhibits Proliferation and Chromatid Cohesion. *International Journal of Molecular Sciences*, 21(6), 2073. https://doi.org/10.3390/ijms21062073

- Redondo Monte, E., Wilding, A., Leubolt, G., **Kerbs, P**., Bagnoli, J. W., Hartmann, L., Hiddemann, W., Chen-Wichmann, L., Krebs, S., Blum, H., Cusan, M., Vick, B., Jeremias, I., Enard, W., Theurich, S., Wichmann, C., & Greif, P. A. (2020). ZBTB7A prevents RUNX1-RUNX1T1-dependent clonal expansion of human hematopoietic stem and progenitor cells. *Oncogene*, 39(15), 3195–3205. https://doi.org/10.1038/s41388-020-1209-4

- Batcha, A. M. N., Bamopoulos, S. A., **Kerbs, P**., Kumar, A., Jurinovic, V., Rothenberg-Thurley, M., Ksienzyk, B., Philippou-Massier, J., Krebs, S., Blum, H., Schneider, S., Konstandin, N., Bohlander, S. K., Heckman, C., Kontro, M., Hiddemann, W., Spiekermann, K., Braess, J., Metzeler, K. H., . . . Herold, T. (2019). Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *Scientific Reports*, 9(1). https://doi.org/10.1038/s41598-019-48167-4

- Haubner, S., Perna, F., Köhnke, T., Schmidt, C., Berman, S., Augsberger, C., Schnorfeil, F. M., Krupka, C., Lichtenegger, F. S., Liu, X., **Kerbs, P.**, Schneider, S., Metzeler, K. H., Spiekermann, K., Hiddemann, W., Greif, P. A., Herold, T., Sadelain, M., & Subklewe, M. (2018). Coexpression profile of leukemic stem cell markers for combinatorial targeted therapy in AML. *Leukemia*, 33(1), 64–74. https://doi.org/10.1038/s41375-018-0180-3

- Greif, P. A., Hartmann, L., Vosberg, S., Stief, S. M., Mattes, R., Hellmann, I., Metzeler, K. H., Herold, T., Bamopoulos, S. A., **Kerbs, P.**, Jurinovic, V., Schumacher, D., Pastore, F., Bräundl, K., Zellmeier, E., Ksienzyk, B., Konstandin, N. P., Schneider, S., Graf, A., . . . Spiekermann, K. (2018). Evolution of Cytogenetically Normal Acute Myeloid Leukemia During Therapy and Relapse: An Exome Sequencing Study of 50 Patients. *Clinical Cancer Research*, 24(7), 1716–1726. https://doi.org/10.1158/1078-0432.ccr-17-2344

Declaration of author contributions

The study "Loss of ISWI ATPase SMARCA5 (SNF2H) in Acute Myeloid Leukemia Cells Inhibits Proliferation and Chromatid Cohesion" (Chapter 1) was conducted in collaboration with the group of Tomas Stopka from the First Medical Faculty of the Charles University in Prague. I performed gene expression analysis of AML patients and provided statistical analyses.

The study "Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring *NRIP1-MIR99AHG* rearrangements" (Chapter 2 and 3) was conducted in collaboration with Caroline Heckman from the Institute for Molecular Medicine Finland and Tobias Herold from the University Hospital of the LMU Munich. I developed the RNA-seq/Filtering pipeline for fusion gene detection. I performed the analyses with support from Anja Swoboda and Stefan Krebs. I wrote the manuscript with support from Sebastian Vosberg and Philipp Greif.

Prof. Dr. Dirk Metzler confirms the correctness of the statements about the author contributions.

Prof. Dr. Dirk Metzler

Paul Kerbs

Summary

Fusion genes result from genomic rearrangements, such as translocations or inversions. On the transcript level, fusions arise from accidental read-through and trans-splicing events. In acute myeloid leukemia (AML), fusion genes are found in around 30% of patients constituting major biomarkers for diagnosis, prognosis and treatment decisions. Furthermore, studies have shown the relevance of gene expression profiles for the distinction of AML subtypes and risk assessment of the patients. This dissertation is focused on fusion gene detection and gene expression analysis by transcriptome sequencing (RNA-seq) of large AML patient cohorts.

In the first chapter, we analyzed the expression of *SMARCA5* in AML patients and performed knockout experiments using leukemic cell lines. We found a positive correlation between the expression levels of proliferation biomarkers and *SMARCA5*. In addition, we observed shorter overall survival of patients with high *SMARCA5* expression. Knockout experiments showed decreased proliferation and growth of leukemic cells lacking *SMARCA5*. Therefore, we concluded that *SMARCA5* might have prognostic relevance and constitutes a potential target for inhibition treatment of AML patients with high *SMARCA5* expression.

In chapter 2, we analyzed the performance of fusion gene detection by RNA-seq in nearly a thousand AML patient samples. Therefore, data from clinical routine diagnostics was compared to results from fusion callers (i.e., Arriba, FusionCatcher) showing high sensitivity (90%) of fusion gene detection by RNA-seq. Moreover, RNA-seq identified AML-related fusion genes in 26 cases that were not reported by routine diagnostics. However, fusion calling from sequencing data usually yields many false positive events. Therefore, we established a detection pipeline with fine-tuned filtering strategies enabling the identification of 157 robust fusion candidates and the discovery of *NRIP1-MIR99AHG*, a novel recurrent fusion gene in AML.

Chapter 3 presents the filtering strategies and the workflow for the detection of robust fusion gene candidates by RNA-seq. The filtering metrics Promiscuity Score (PS), Fusion Transcript Score (FTS) and Robustness Score (RS), which were developed in this study, use properties such as expression and frequencies of fusion events to assign evidence levels to the detected fusion genes. This enabled substantial reduction of putative false positive fusion calls and allowed for robust identification of novel fusions as demonstrated in chapter 2. Furthermore, all required tools and modules of the workflow were bundled into a publicly available software package for simple execution in different system environments.

This thesis highlights the power of RNA-seq for gene expression analyses and fusion gene detection in the context of AML diagnostics.

Introduction

Next-generation sequencing

Due to higher precision, lower costs and shorter runtime, next-generation sequencing (NGS) is becoming increasingly popular. NGS enables sensitive and comprehensive genetic analyses providing a valuable tool for clinical diagnostics of genetic disorders. Millions of reads generated by NGS contain detailed structural information of the genome (DNA-seq) or transcriptome (RNA-seq) with a resolution of single base pairs. NGS enables the identification and quantification of transcribed genes, as well as more specific analyses such as the detection of single-nucleotide variants (SNV), structural variants (SV), fusion genes, differential/alternative splicing, allelic imbalances, etc. Furthermore, targeted NGS assays allow for further cost reduction and shorter runtime while sensitivity and precision can be increased due to on-target focused sequencing power. Several studies explored the applicability of these targeted assays in clinical settings, demonstrating the added value in diagnostics of hematological malignancies[1–5]. Nevertheless, novel genetic lesions are not captured by targeted approaches and unusual aberrations, not covered by these assays, might be missed. In general, the workflow of processing NGS data (Figure 1) can be divided into the following steps: (I) Removal of low-quality reads and trimming of low-quality bases. (II) Mapping of the reads to a reference sequence. (III) Primary analyses such as quantification of sequence coverage and identification of differences between the mapped reads and the reference. (IV) Secondary analyses.

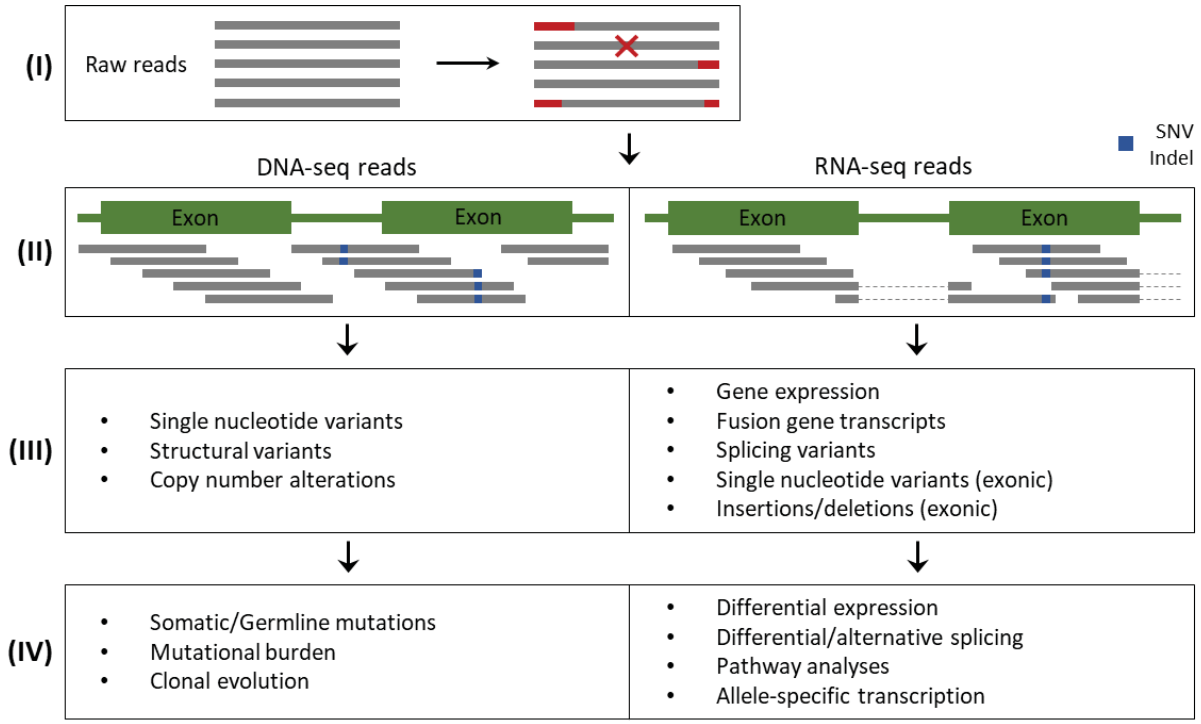


Figure 1: General pipeline for processing next-generation sequencing data divided into the steps: (I) Quality filtering and read trimming, (II) Mapping, (III) Primary Analyses, (IV) Secondary Analyses.

The vast repertoire of computational methods to process NGS data complicates the establishment of standardized analyses. In addition to the variety of sequence mapping and downstream analyses methods, many parameters within the tools can be fine-tuned to improve sensitivity and/or reduce false positives. Several studies have been conducted to comprehensively compare and evaluate different algorithms in various settings but overall, no standard approach has been established so far. It was rather concluded that performance is dependent on underlying properties of the sequencing data (e.g., organism, read length, sequencing technology) and the selection of proper settings based on study design[6–9]. Moreover, efforts have been made to develop applications merging several analyses into one pipeline providing a simplified tool for an extensive diagnostic workup of patients' genomic data[10,11].

In the last decade, NGS has tremendously advanced our knowledge about hematopoietic and other genetic diseases, enabling ever more accurate assessment of individual genetic aberrations and the development of tailored treatments. What is more, third-generation sequencing methods are on the rise e.g., Oxford Nanopore. This technology generates long reads with lengths of tens of kilobases in contrast to NGS producing reads of usually 100 to 200 bases in length. Longer reads allow for more accurate and complete assembly of the genome and genomic transcripts. By all means, the significance of NGS as an additional diagnostic method of hematological malignancies in clinical applications has already been demonstrated[12–15]. Therefore, progressive improvements and increasing application of NGS in clinical routine will inevitably become a new standard in precision oncology.

Acute myeloid leukemia

AML is a hematological disease characterized by impaired maturation and clonal expansion of myeloid progenitor cells (i.e., myeloblasts). The inability of these cells to differentiate into mature blood cells such as erythrocytes, megakaryocytes, macrophages or granulocytes, and their increased proliferation suppresses the production of normal blood cells. AML is a heterogeneous disease and subgroups were initially defined by the French-American-British (FAB) co-operative group[16] but mostly replaced by a classification of the World Health Organization (WHO) which included genetic aberrations[17,18]. In addition, the European LeukemiaNet (ELN) has published recommendations on the management and risk stratification of AML[19,20].

Cytogenetically normal (CN) patients (40-50% of AML cases) are characterized by small genetic changes e.g., single base mutations. Previous studies comprehensively explored the mutational landscape of AML and identified hundreds of recurrent disease-defining lesions[21–23] showing that the broad spectrum and variable co-occurrence of these small somatic aberrations substantially contribute to the heterogeneity of AML. Commonly mutated genes are *NPM1*, *FLT3* and *DNMT3A*. Mutated *NPM1* was defined as a distinct class by the WHO and is associated with favorable risk (without the co-occurrence of *FLT3* internal tandem duplications or other adverse risk factors). This gene primarily

resides in the nucleolus and is involved in multiple cellular processes such as ribosome biogenesis, preservation of genomic stability, p53-dependent stress response and modulation of growth-suppressive pathways[24]. Although the precise leukemogenic mechanism of mutated *NPM1* has not yet been identified, aberrations of *NPM1* were found to impair its function in maintaining genome stability[25] and induce delocalization of the protein to the cytoplasm[26]. Furthermore, mutated *NPM1* requires cooperating aberrations in leukemogenesis[24] and is mostly found in co-occurrence with mutations in *FLT3* and/or *DNMT3A*[27]. Small mutations are rarely mutually exclusive and often occur together with other larger aberrations such as translocations, inversions or deletions, which is crucial for risk stratification and treatment strategy.

AML patients are usually treated with chemotherapy to reduce the leukemic burden and achieve complete remission (CR) which is commonly defined by the abundance of ≤5% blasts in the bone marrow. The rate of patients achieving CR is around 64%, and younger patients or favorable risk group patients show higher treatment response rates (73%)[28]. Nevertheless, depending on age and risk group, relapse rates range from 30% to 80%[29] and the majority of patients decease within 5 years after diagnosis[30]. Initial treatment strategies have not substantially changed over the past decades, consisting of an induction therapy with anthracycline and cytarabine for young adults and medically fit elderly patients. After achieving CR, consolidation treatment of several cycles with high-dose cytarabine and/or stem cell transplantation (SCT) aims to prevent or to delay relapse, which eventually occurs in the majority of patients. However, alternative treatment strategies of AML subtypes have been shown to improve response and survival of patients. For example, patients with acute promyelocytic leukemia (APL), characterized by a *PML-RARA* fusion, are commonly treated with all-trans retinoic acid (ATRA) and achieve CR in 80-100% of cases[31,32]. ATRA induces degradation of the PML-RARA oncoprotein and restores transcription of genes that are involved in myeloid differentiation, thereby reconstituting normal blood production. The introduction of ATRA for the treatment of APL was a pioneering step towards personalized medicine and emphasizes the need for precise tumor diagnostics as well as targeted treatment approaches. Furthermore, inhibition of specific mutant proteins resulting from alterations in *FLT3*, *IDH* or nuclear exporters shows promising results with regards to outcome of the corresponding AML subgroups[33–36]. Besides SCT, the most established immunotherapy for AML, targeted immunotherapies are on the rise showing exciting results in other entities such as chronic lymphoblastic leukemia[37] and are currently under investigation for the use in AML[38,39]. The aim is to immunologically eradicate malignant cells in a targeted manner by e.g., chimeric antigen receptor T-cells that are designed to bind tumor-specific antigens.

Taken together, AML is a highly heterogenous disease requiring tailored treatment strategies that are based on risk stratification deduced from clinical parameters and more importantly, thorough identification of the mutational profile and other genetic abnormalities of the individual patients.

Fusion genes in AML

Fusion genes emerge from rearrangements of chromosomal regions (e.g., translocation, inversion) whereby the breakpoints are located within or in proximity of affected genes. However, fusion genes may also arise from intergenic splicing events without any disruption of the genome[40]. The first fusion genes were discovered by the identification of the recurrent translocation t(9;22)(q34;q11) in chronic myeloid leukemia (CML) resulting in a *BCR-ABL1* fusion and the recurrent translocation t(8;21)(q22;q22.1) in AML resulting in a *RUNX1-RUNX1T1* fusion[41–44]. These milestones led to further discoveries of disease-defining fusion genes, not only in hematological entities but also in solid tumors[45,46], and laid the foundation for targeted treatments. For example, the BCR-ABL1 protein can be targeted by Imatinib which inhibits ATP-binding of the fusion protein, thereby preventing its oncogenic effect and inducing apoptosis of the affected cells. In AML, fusion genes occur in a third of all cases (Figure 2) constituting important diagnostic and prognostic biomarkers.

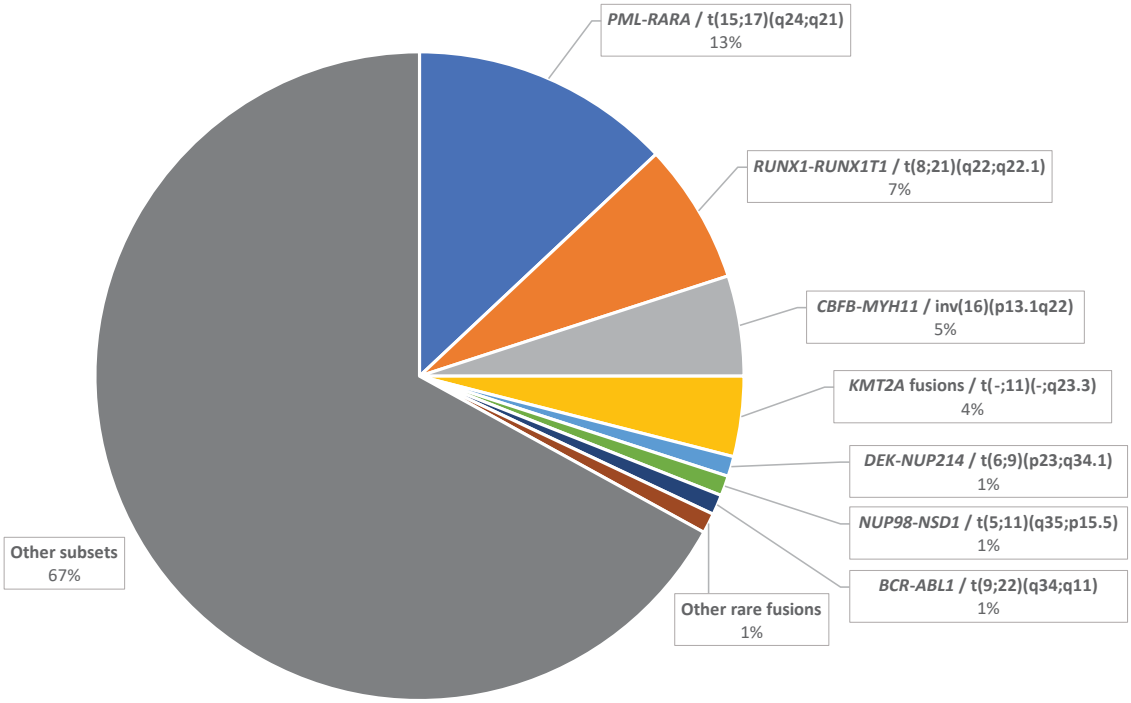


Figure 2: Distribution of biological and prognostic subgroups in a cohort study of 5876 AML patients[27]. Other subsets include subgroup-defining alterations that are not resulting in transcribed fusion genes.

The most recurring fusions are *RUNX1-RUNX1T1*, *CBFB-MYH11*, *PML-RARA*, *DEK-NUP214* and fusions involving *KMT2A*. *DEK-NUP214* and *KMT2A* fusions result from translocations t(6;9)(p23;q34.1) and t(-;11)(-;q23.3), respectively. While these fusions constitute adverse risk and affected individuals have a poor prognosis, *RUNX1-RUNX1T1* and *CBFB-MYH11*, resulting from t(8;21)(q22;q22.1) and inv(16)(p13.1q22)/t(16;16)(p13.1;q22), are fusion genes associated with favorable outcome. Moreover, t(15;17)(q24;q21) forming a *PML-RARA* fusion and found in 95% of APL cases, is regarded as the best manageable subtype of AML and treatment with ATRA in combination with N (ATO) showed

remarkable CR and cure rates[32,47]. However, resistance to ATRA and ATO has been observed. In some instances, this could be driven by pre-existing or acquired mutations in *PML* or *RARA*[48–50] but also other driver genes[51,52]. Furthermore, other *RARA* fusions in APL, such as *GTF2I-RARA* and *PLZF-RARA*, have also shown insensitivity to ATRA[53–55]. This emphasizes the critical relevance of accurate and complete capturing of genetic aberrations in clinical diagnostics for proper assessment of treatment options. So far, there are no therapies targeting other AML-related fusions.

Current standard in AML diagnostics

Clinical routine diagnostics of AML patients includes initial microscopic inspection of cells from bone marrow or peripheral blood smears and the identification of potential leukemic cells based on cytomorphology. Further, cellularity, histotopography and distribution of immature and mature hematopoietic stem cells are used to identify different hematological disorders. In example, AML is determined by a myeloblast count ≥20%[17], as defined by the WHO. However, recurrent rearrangements t(15;17), t(8;21) and inv(16)/t(16;16), resulting in *PML-RARA*, *RUNX1-RUNX1T1* and *CBFB-MYH11* fusions, respectively, are sufficient to diagnose AML, regardless of the blast count. Blast lineages and maturation state can be inferred from immunophenotyping which is the measurement of specific surface antigens. To this end, cells are labeled with fluorophore-conjugated antibodies targeting surface markers of interest and are subsequently quantified by flow cytometry. This allows for discrimination of cell populations and the identification of specific immunophenotypes, guiding more specialized aberration screenings such as FISH and PCR. Therefore, immunophenotyping by flow cytometry is a crucial initial step in diagnosis of hematological diseases and is also used for minimal residual disease monitoring.

Chromosomal G-banding (Karyotyping) is a technique based on microscopic analysis of metaphase chromosomes allowing for detection of larger SV such as deletions, duplications, translocations, inversions or aneuploidy. Of note, Karyotyping requires culture of leukemia cells which is not always successful. FISH is used to identify known or suspected rearrangements and numerical aberrations by fluorescently labeled probes. FISH can be applied to metaphase spreads from cultured cells as well as to interphase nuclei from cells directly spread on a slide with the latter allowing for higher cell counts. While FISH is a targeted approach, Karyotyping provides genome-wide analyses but is less sensitive since the resolution is approximately 5-10 Mb[56] and the analysis is typically limited to 25 metaphases. A further crucial diagnostic technique in clinical routine is PCR which is a targeted approach and offers very high sensitivity. Thereby, particular sequences of DNA or cDNA (reverse transcribed RNA) are targeted by specific primers and amplified enabling the identification of SNV, SV or the detection of specific transcripts (e.g., fusions) which can be quantified (qPCR, microarrays) and monitored. Complementary utilization of these techniques in contemporary clinical routine is essential for AML diagnostics and lays the foundation for risk stratification and treatment strategies.

Fusion gene detection by RNA-seq

Fusion gene detection by RNA-seq allows for systematic examination of the entire transcriptome, is not limited to specific targets and has the potential for discovery of novel fusion transcripts. However, fusion detection by RNA-seq is computationally challenging and is based on the detection of chimeric sequences. Therefore, fusion calling algorithms need to identify reads whose subsequences map to different locations in the transcriptome, defined as fusion spanning reads or fusion spanning pairs (Figure 3).

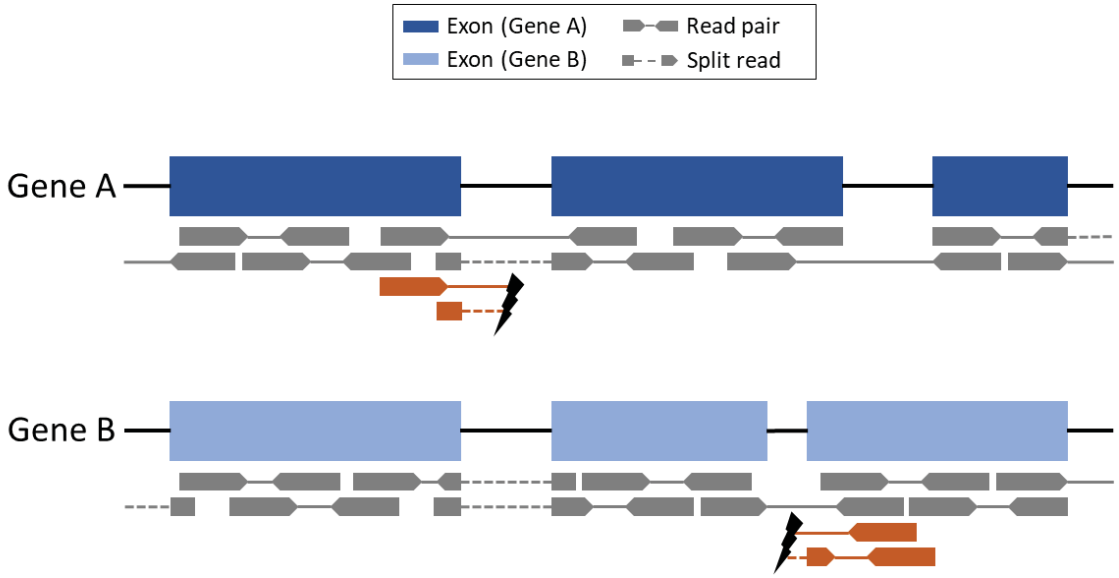


Figure 3: Illustration of paired-end RNA-seq reads mapped to two gene loci. Fusion supporting reads are highlighted in orange. A Fusion spanning pair is characterized by one read mapping to a gene locus other than its partner read. A read whose subsequences map to different gene loci is defined as a fusion spanning read.

Several technical factors might influence the accuracy of this procedure such as sequencing errors and artifacts derived from erroneous sequence amplification, potentially resulting in false mappings. Furthermore, biological factors such as polymorphic genes, homologous regions and highly expressed genes might contribute to false positive fusion calls[7]. Over the last decade, many tools have been developed for the identification of fusion genes in RNA-seq data. Comparative evaluation of these tools on real and synthetic sequencing data demonstrated overall good performance with a sensitivity of around 90%[57–59]. However, only a low proportion of fusion events were called consistently by the different tools and therefore, the authors recommended to utilize several callers for robust fusion detection analyses. The low overlap of identified fusion events between the tools suggests a high rate of false positive calls which requires proper filtering strategies. Current filtering approaches offered by the tools are based on read coverage, built-in scoring dependent on individual parameters, proportion of spanning reads and spanning pairs, annotation of partner genes involved in the fusion or blacklists generated from databases of fusions found in healthy samples.

Gene expression as a biomarker in AML

On a molecular level, the state of a cell is largely characterized by its gene expression. Furthermore, it is commonly known that changes in expression of certain genes trigger alterations in the state or behavior of a cell. Transcription factors *MYC*, *MYB*, *FOS* and tyrosine kinases *ABL1*, *FES*, *KIT*, *PIM* play a crucial role in hematopoiesis and were the first genes whose expression was studied in AML-derived cells[60–63]. Overexpression of *RUNX1*, another important transcription factor involved in hematopoiesis, enhanced cell proliferation while suppressing granulocytic differentiation[64]. Increased expression rates of growth factor *FLT3* were observed in leukemic blasts[65] and elevated expression of the *CD34* gene in leukemic patient samples was associated with lower CR rates and adverse outcome[66]. Moreover, increased expression of *WT1*[67], *MN1*[68], *BAALC*[69], *ERG*[70], and *MECOM*[71] was shown to be significantly correlated with poor prognosis. Besides prognostic relevance, studies have shown that AML subtypes can be reliably distinguished based on gene expression[72,73].

In addition to expression analysis by qPCR, microarrays paved the way for simultaneous expression analyses of thousands of genes evolving to a popular tool in hematological research throughout the first decade of this century. With the advent of NGS, a new milestone in high-throughput screening for gene expression was set. Since then, ever more new markers have been identified and gene expression profiles have been proven as independent classifiers and prognostic indicators[74]. Furthermore, several studies proposed scoring models based on the expression of gene sets as predictors for therapy resistance and survival, providing significant impact for risk assessment[75–77].

A big leap forward in understanding the regulatory mechanisms of a cell was made by the discovery of non-coding RNAs (ncRNA). Coding genes comprise only a small fraction of the RNA pool while non-coding elements constitute over 90% of the processed RNA, playing a central role in the regulation of cellular processes[78,79]. Numerous studies have explored the significance of ncRNAs in different cancers but amongst hematopoietic diseases, AML is the most studied entity regarding long ncRNAs (lncRNA)[80]. LncRNA expression profiles have been associated with clinical characteristics, recurrent mutations and survival[81]. Furthermore Schwarzer et al.[82] conducted a comprehensive study in order to establish a ncRNA expression atlas demonstrating specific signatures for different hematopoietic cell populations. For example, the authors identified *LINC00173* as a regulator of granulocytic proliferation and differentiation.

Gene expression can be influenced by various factors such as altered methylation of regulatory regions, abundance/absence of certain transcription factors, or even chromosomal rearrangements. For example, the rearrangement inv(3)(q21q26)/t(3;3)(q21;26) causes the reallocation of a *GATA2* enhancer resulting in overexpression of *MECOM* and *GATA2* haploinsufficiency[83,84]. In case that a chromosomal rearrangement results in a fusion gene (Figure 4), it is expected that the expression of the 3' partner gene gets under the control of the 5' partner's promoter.

Thus, extensive gene expression analysis of coding and non-coding genes provides critical information with diagnostic relevance and can be easily derived from RNA-seq data. Further expression studies on larger cohorts are needed to gain deeper insight into the regulatory mechanism of gene expression and the correlation to leukemogenesis, which could lead to the discovery of new drug targets and more effective individualized treatment.

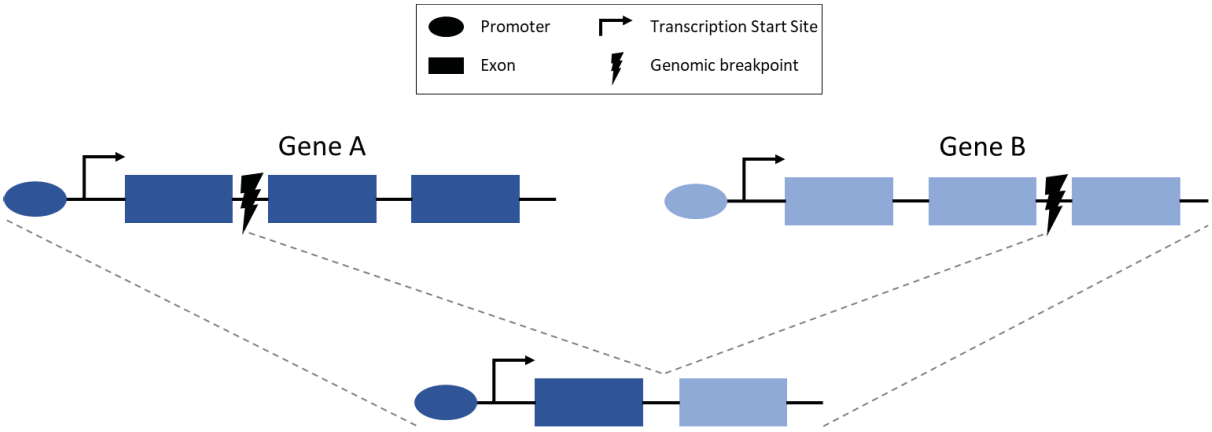


Figure 4: Illustration of a fusion gene resulting from a chromosomal rearrangement. Expression of the rear part of Gene B (3' partner) is supposed to be controlled by the promoter of Gene A (5' partner).

Objectives

This thesis is focused on fusion gene detection and gene expression analysis by RNA-seq in the context of the application in clinical diagnostics of AML. The specific aims are:

1. Analysis of RNA-seq data from AML patients to measure *SMARCA5* expression. Test for correlation between gene expression of proliferation biomarker genes and *SMARCA5* expression. Test for correlation between overall survival of patients and *SMARCA5* expression levels.
2. Evaluation of the performance of RNA-seq in fusion gene detection in comparison to methods from clinical routine. Analysis of RNA-seq data from cohorts of nearly a thousand AML patients in order to identify putative novel fusion genes.
3. Development of a filtering concept integrated into a detection workflow to enable robust identification of fusion genes.

Chapter 1

Loss of ISWI ATPase SMARCA5 (SNF2H) in Acute Myeloid Leukemia Cells Inhibits Proliferation and Chromatid Cohesion

Tomas Zikmund, Helena Paszekova, Juraj Kokavec, **Paul Kerbs**, Shefali Thakur, Tereza Turkova, Petra Tauchmanova, Philipp A. Greif, Tomas Stopka

(*Int. J. Mol. Sci.* 2020, *21*(6), 2073; https://doi.org/10.3390/ijms21062073)

In this chapter, we analyzed the relevance of *SMARCA5* expression for proliferation of leukemic cells. The ATPase *SMARCA5* is a member of the imitation switch (ISWI) gene family that are involved in chromatin remodeling and play an essential role in DNA repair, transcription, and replication. RNA-seq data from AML patients showed significantly elevated *SMARCA5* expression in diagnostic samples (high amount of immature blast cells) compared to matched remission samples (less than 5% blast cells) confirming previous reports of higher *SMARCA5* expression in CD34+ AML cells. Moreover, we observed shorter overall survival of patients with higher *SMARCA5* expression, but this finding was statistically not significant. However, we saw a positive correlation of *SMARCA5* levels and expression of proliferation biomarkers (*AURKA*, *PLK1*, *CCNA2*, *CENPF*). To examine the effects of *SMARCA5* depletion, *SMARCA5* knockout clones were generated from leukemia cell lines (K562, OCI-M2, NB4, SKM1, MOLM-13) by CRISPR/Cas9 genome editing. Clones lacking *SMARCA5* demonstrated impaired proliferation in K562 cells, while knockout clones from other cell lines even seemed not to tolerate *SMARCA5* depletion at all. Transplantation of *SMARCA5* expressing and lacking murine fetal liver cells into lethally irradiated mice showed reconstitution of hematopoiesis only in mice that were transplanted with *SMARCA5* expressing cells, suggesting an essential role in normal blood production and implicating that *SMARCA5* might play a role in early leukemia-initiating compartments. Additionally, *SMARCA5* lacking cells were frequently found to have nucleic abnormalities such as polyploidy, nucleic budding, karyorrhexis, and multinuclearity. Although decreased cell growth and proliferation defects were also observed in healthy non-hematopoietic cells mediated by *SMARCA5* deletion, *SMARCA5* might be a target for inhibition treatments, which requires further extensive studies.

Chapter 2

Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring *NRIP1-MIR99AHG* rearrangements

Paul Kerbs, Sebastian Vosberg, Stefan Krebs, Alexander Graf, Helmut Blum, Anja Swoboda, Aarif M. N. Batcha, Ulrich Mansmann, Dirk Metzler, Caroline A. Heckman, Tobias Herold, and Philipp A. Greif

(*Haematologica* Early view Jun 17, 2021; https://doi.org/10.3324/haematol.2021.278436)

In this chapter, we collected RNA-seq data of nearly a thousand clinically well-characterized AML patients from four different cohorts. We applied two fusion calling methods, namely Arriba and FusionCatcher, to identify fusion events and compared detection performance between calls from RNA-seq and data from clinical routine diagnostics. Around 90% of fusion genes reported by routine were also detected by the RNA-seq methods, while we observed lower sequence read depth in samples in which RNA-seq did not detect any fusion despite evidence from routine data. On the other hand, we identified 26 known recurrent fusion genes that were not reported by routine diagnostics. In general, algorithms for the detection of fusions by RNA-seq tend to report many false positives. Therefore, we developed a fusion detection workflow together with several filtering strategies including blacklists generated from healthy samples and several metrics assessing evidence levels for individual fusion calls. Evidence level cutoffs were derived from known fusion events enabling a substantial reduction of putative false positive calls. On average, we detected 51 fusion events per patient. Although roughly 70% of these events were excluded by the built-in filters of the callers, the number of remaining events indicated a high proportion of false positives. Based on our filtering strategies, we excluded approximately 95% of fusion calls that were most likely artifacts. In addition, we observed elevated expression of genes in specific cases where they form the 3' end of a fusion gene, which can provide further evidence for a fusion event. Finally, we discovered a novel recurrent inversion on chromosome 21 resulting in a *NRIP1-MIR99AHG* fusion transcript which was validated by PCR and Nanopore sequencing. Both genes involved in the fusion have already been associated to leukemogenesis. Furthermore, we identified 157 putatively novel fusion transcripts with high evidence according to our detection workflow.

Chapter 3

A workflow for the detection of robust fusion gene candidates by RNA-seq

This chapter describes the workflow for the detection of fusion genes by RNA-seq and addresses filtering approaches to reduce fusion calls that are most likely artifacts. In our study of RNA-seq data of 806 AML patient samples (Chapter 2), we observed a high number of reported fusion genes per patient, suggesting a high false positive rate of the RNA-seq based tools. The tools provide built-in filters such as a minimum number of fusion-supporting reads, annotation-based filtering, internal evidence scores and blacklists compiled from public databases of fusions that were found in healthy samples. However, the number of fusion genes per patient remained high, even after exclusion of fusion genes by the built-in filters (Appendix 2, Figure 2).

Therefore, we established a filtering pipeline (Figure 5) based on several in-house developed metrics in addition to the built-in filters of the fusion callers. First, we included a custom generated blacklist of fusion genes that were called by RNA-seq data of 39 healthy samples. Moreover, we observed certain genes that are reported in a multitude of different fusion events, which we measured by our Promiscuity Score (PS). The assumption is that the higher the PS, the higher the probability for a fusion event to be a false positive. In order to test this hypothesis, we compared PS values between known and unknown fusion events. Indeed, we observed a maximum PS of 16.5 among known fusions while unknown fusions showed significant higher PS values (Appendix 2, Figure S3A). This allowed for the definition of distinct cutoffs to filter for fusions resembling observed PS values of known fusion genes. Furthermore, we found fusion events consisting of high expressed genes while the respective fusion showed only low expression, which might indicate artifacts. In order to assess this discrepancy, we captured the relative expression of a fusion by our Fusion Transcript Score (FTS). Known fusions were characterized by a median FTS value of 0.35 while unknown fusions showed a median FTS value of 0.1 (Appendix 2, Figure S3B). This finding allowed for the selection of fusion events resembling higher FTS values of known fusions and the exclusion of likely false positive calls characterized by low FTS values. Occasionally, we observed that fusion events slip through the FTS filter. Capture of these events was addressed by our Robustness Score (RS). Together, these filtering metrics enabled a substantial reduction of putative false positive events (Appendix 2, Figure 2) using characteristics of fusion calls and estimated gene expression which can also be obtained from RNA-seq data. A detailed description of the PS, FTS and RS filter metrics are provided in the following sections.

Our workflow included two detection streams via Arriba[85] and FusionCatcher[86]. In the final step, only overlapping fusion calls from both streams were regarded as robust fusion candidates. Based on the

filtering metrics developed in this study, evidence levels were assigned to all identified fusion events, which allowed for adjustment of the filter stringency.

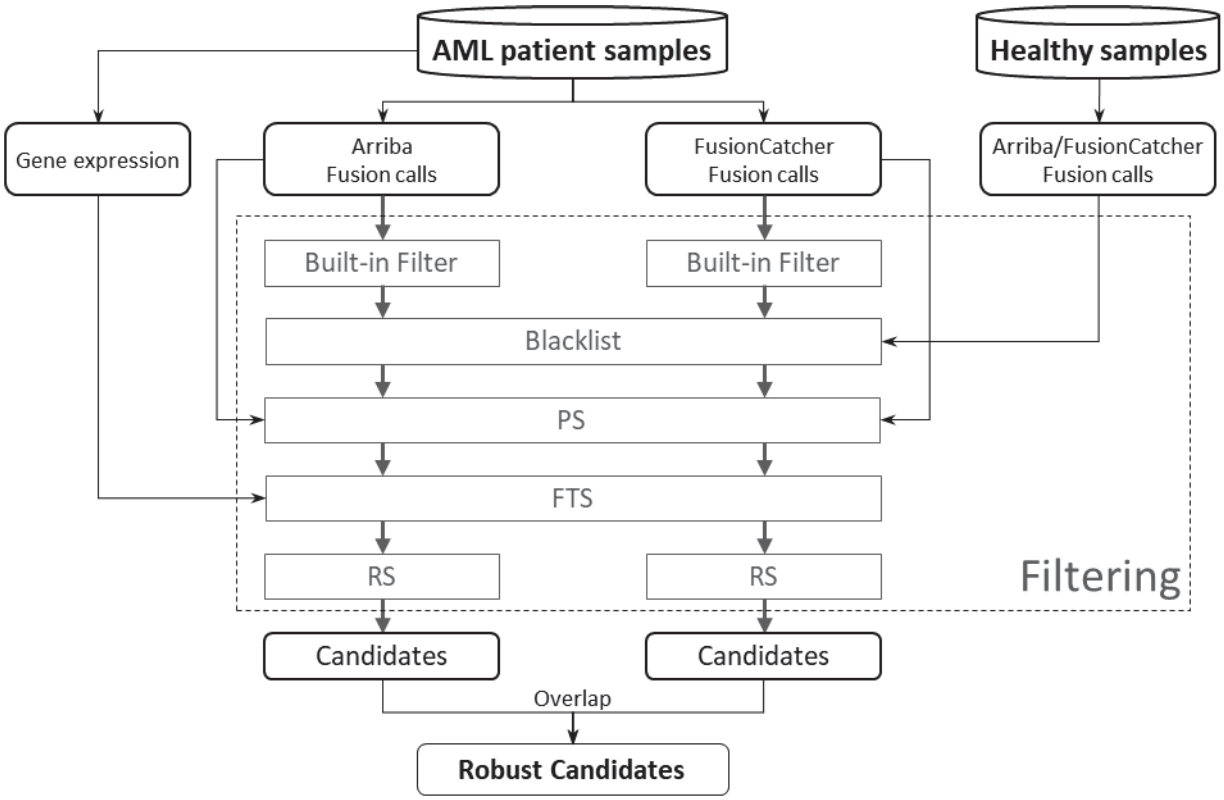


Figure 5: Fusion gene detection and filtering workflow. Firstly, fusion calls from Arriba and FusionCatcher are filtered by built-in filters of the callers. Afterwards, detected fusion events are filtered by a custom blacklist, the Promiscuity Score (PS), the Fusion Transcript Score (FTS) and the Robustness Score (RS). Finally, only consistently called fusion events by Arriba and FusionCatcher are regarded as robust fusion gene candidates.

Promiscuity Score

As mentioned in the introduction, certain genes are prone to be detected as part of false fusion events. Characteristically, these genes are found to form fusions with many different partner genes. Therefore, the PS of a fusion event (PS_{fusion}) measures the average number of different partner genes, that were detected in a set of samples, for the two genes involved in that fusion event. In more detail, the average number of varying partners, that were detected by Arriba and FusionCatcher for the individual genes at the 5' and 3' end of the specific fusion, was defined as P_x. Thus, the PS of a fusion event can be formalized as follows:

$$PS_{fusion} = mean(P_{5'}, P_{3'})$$

$$with\ P_x = mean\left(Ptr_{Arriba,x}, Ptr_{FusionCatcher,x}\right)\ for\ x\ in\ \{5', 3'\}$$

$$and\ Ptr_{M,x} = amount\ of\ different\ fusion\ partners$$

$$for\ M\ in\ \{Arriba, FusionCatcher\}$$

Fusion Transcript Score

Sequence reads of highly expressed genes are likely to accumulate sequencing errors or produce fusion artifacts during the amplification steps. This could result in false mapping and false positive fusion calls. It is fair to assume that expression of a fusion correlates to the expression of its partner genes. Therefore, we developed the FTS which provides a metric to measure, in transcripts per million (TPM), the expression of a fusion relative to the expression of its partner genes:

$$FTS_{fusion} = mean(FTS_{5'}, FTS_{3'})$$

$$with\ FTS_x = \frac{TPM_{fusion}}{TPM_{fusion} + TPM_x}\ for\ x\ in\ \{5', 3'\}$$

Calculation of TPM expression requires read counts and the length of the respective gene transcript. While these values are available for single transcripts from mapping and gene annotations, this is not the case for fusion genes. Due to limited length of the read fragments and the fact that only reads covering the fusion breakpoint can be accounted for the expression of the fusion gene transcript, exact length and therefore, TPM expression of the fusion transcript cannot be determined. Therefore, TPM values for a fusion transcript were approximated by using estimated median insert size of the mapped read fragments.

Robustness Score

We observed recurrently detected fusion genes being excluded by the FTS filter in most affected samples but occasionally passing the filter in a few samples. These fusion genes were characterized by a low FTS close to the defined cutoff and an unusual high recurrence among the patients. Most likely, these fusion genes represent false positive events. Therefore, we developed the RS which is defined as the ratio between the number of samples in which a fusion gene passed the FTS filter and the total number of samples in which this fusion gene was called. Only fusion genes passing the FTS filter in at least half of the reported samples (RS ≥ 0.5) were considered.

Detection and filtering workflow as a single software package

Our workflow for fusion gene detection consists of quality filtering, trimming, mapping and insert size estimation of the RNA-seq reads, fusion calling, estimation of gene expression and calculation of the PS, FTS and RS. These steps require installation of several tools and various packages in the computational environment. Therefore, we bundled all software dependencies in a Singularity[87] container and programmatically combined the aforementioned steps into one single analysis. The portability of the Singularity software allows for simple execution of our fusion detection workflow on different computer systems without the necessity for the installation of further software. The package of the fusion detection workflow and a documentation is publicly available for download:
https://sourceforge.net/projects/fusion-detection-pipeline/

Discussion

In the first chapter of this thesis, we analyzed RNA-seq data from large cohorts of AML patients focusing on *SMARCA5*, a gene involved in ATP-dependent chromatin remodeling, and its role in leukemic cell proliferation and differentiation. We observed higher expression of *SMARCA5* in patient samples at the time of diagnosis compared to matched samples at the time of CR, which is consistent with a previous study showing upregulation of *SMARCA5* in CD34+ AML cells[88]. Moreover, data suggested a trend to shorter overall survival of patients with high *SMARCA5* expression. We showed that deletion of *SMARCA5* reduced cell growth and proliferation of leukemic cells, but also proliferation of healthy cells was affected by the absence of *SMARCA5*. Nevertheless, our findings suggest that *SMARCA5* could be a potential target for treatment, which needs to be investigated in further studies. Although statistically not significant, higher expression of *SMARCA5* was associated with shorter survival, which is consistent with previous studies in solid tumors reporting correlation of high *SMARCA5* expression with disease aggressiveness and resistance to chemotherapy[89,90]. Our findings suggest that *SMARCA5* expression rates should be taken into consideration for the prognosis and treatment of AML patients. Furthermore, aberrant expression of certain genes can provide evidence for fusion events, as will be discussed in the following section.

In chapter 2, we studied RNA-seq data derived from bone marrow or peripheral blood samples of nearly a thousand well-characterized AML patients. To evaluate the performance of RNA-seq regarding the detection of fusion genes, we defined a benchmark of true fusions that were reported by clinical routine diagnostics. RNA-seq based methods showed high sensitivity by detecting 90% of the true fusion set and the identification of a relevant number (n=26) of recurrent AML-related fusion genes that were not reported by routine diagnostics. This demonstrates the strong potential of RNA-seq for complementary application in clinical diagnostics of AML. In most cases in which true fusions were missed by RNA-seq, affected samples showed overall lower read coverage, especially at the loci of genes involved in known recurrent fusions. Although sequencing depth of these samples (~30 mio. reads) is sufficient for overall analysis of gene expression, transcript discovery (e.g., fusion gene transcripts) requires higher sequencing depth, according to the data standards of the ENCODE consortium[91]. Thus, fusion gene detection might have been impaired by lower sequencing depth of these samples. Furthermore, 71% of samples in which no true fusion could be detected by RNA-seq were from the same cohort, indicating cohort-specific sequencing issues.

In the effort to identify novel fusion genes, we developed a detection workflow including several filtering steps, which allowed for substantial reduction of reported fusion genes that are most likely artifacts. In addition to our filtering strategies, we showed that gene expression alone could already

provide evidence for certain fusion events. As described in the introduction, expression of the partner gene at the 3' end of a fusion is supposedly controlled by the promoter of the partner gene at the 5' end. Therefore, it is expected that expression of the 3' partner gene should adjust to expression levels of the 5' partner. Especially in cases in which the 3' partner is usually not expressed or expressed at low levels, the 3' partner should show increased expression. Indeed, we observed this effect in cases of known recurrent fusion genes as well as in cases of novel fusion candidates.

Based on our detection workflow, we identified 157 novel fusion gene candidates. The most interesting fusion gene among those candidates was *NRIP1-MIR99AHG*, resulting from inv(21)(q11.2;q21.1) and recurrently found in nine patients. Based on available cDNA and gDNA of some patient samples, we validated the *NRIP1-MIR99AHG* rearrangement by PCR and Nanopore sequencing. Long reads from Nanopore sequencing revealed several distinct *NRIP1-MIR99AHG* fusion transcripts. None of these transcripts included an annotated open reading frame suggesting no resulting protein products. One of the fusion breakpoints is located upstream of *MIR125B2* which belongs to the *miR-99a*/*let-7c*/*miR-125b-2* tricistronic gene cluster residing in an intronic region of *MIR99AHG*. This miRNA cluster was shown to play a role in homeostasis of hematopoietic stem and progenitor cells[92], and therefore, disruption of this cluster caused by the *NRIP1-MIR99AHG* rearrangement might contribute to leukemogenesis. Furthermore, overexpression of *MIR99AHG* was shown to increase proliferation in acute megakaryoblastic leukemia cell lines[93]. In hematopoietic cells, *MIR99AHG* is usually not expressed or expressed at low levels only. We demonstrated that the *NRIP1-MIR99AHG* fusion drives transcription of the 3' end of *MIR99AHG*, which might also be a factor in leukemogenesis. However, disruption of *NRIP1* might also play a role since it was found to be involved in other fusions[94,95] and has been linked to hematological malignancies in previous studies[96,97]. Further studies are needed to untangle the mechanism and determine the clinical implications of this novel recurrent rearrangement in AML and other hematological entities.

The presented workflow for the detection of fusion genes by RNA-seq (Chapter 3) includes several computational tools and filtering metrics which were developed in this study. Implementation of these tools requires preceding installation steps and the resolution of software dependencies. This can be troublesome and might lead to conflicts within certain system environments. Therefore, a software package was developed including all steps of the workflow which were programmatically bundled into a streamlined and easy-to-use pipeline for reproducibility and robust fusion gene analyses.

The main feature of this workflow was the filtering procedure for reducing the number of false positive fusion calls and included: (1) Built-in filters of the callers, (2) Custom fusion blacklist, (3) PS filter, (4) FTS filter, (5) RS filter, (6) Consistently called fusion genes between Arriba and FusionCatcher. The fusion callers already provide simple filters such as blacklists, fusion supporting reads or annotation-

based filtering, etc. However, even after utilization of these filters many fusion events per patient were reported indicating a high proportion of false positive calls. The built-in blacklists of the callers are compiled from public databases and might therefore not be complete. We showed that an additional blacklist, generated by fusion gene detection in publicly available healthy samples, can further reduce irrelevant fusion events. In order to filter for robust fusion candidates, we excluded further events based on evidence levels derived from three in-house developed metrics i.e., PS, FTS and RS. These metrics are based on gene expression measurements and frequencies of called fusion genes, data that are concurrently retrieved in the fusion detection process. We observed distinctive differences between known and unknown fusion events that were evaluated by these metrics, which allowed for differentiation of likely real fusions and fusions with a high probability of being an artifact.

The PS of a fusion gene measures how frequent the respective partner genes were found to be involved in other fusion events. Events with high PS are likely artifacts and cutoffs were defined based on PS of known fusion genes. It must be noted that PS values strongly depend on sample size. The more samples are used to estimate PS, the better the estimation gets, enabling more accurate differentiation between true and false fusion events. Moreover, different sequencing procedures (e.g., library preparation kit, sequencing platform, sequencing depth) or different detection algorithms have impact on fusion calling. Therefore, application of the PS filter will perform best on uniformly called fusion events in uniformly sequenced cohorts comprising a proper number of samples.

Furthermore, our FTS estimates the expression of a fusion event (based on breakpoint spanning reads) in relation to the expression of the respective partner genes (excluding breakpoint spanning reads). The underlying assumption is that a relatively low number of fusion-supporting reads compared to the number of reads supporting the individual partner genes (low FTS) is an indicator for fusion artifacts. However, considering that RNA-seq data is usually generated from a mixed cell population, real fusion genes might also be characterized by a low FTS since they might reside in small subclones only. Nevertheless, comparative analysis of FTS values between known and unknown fusion events provided an indicative cutoff for maximizing specificity while maintaining sensitivity.

Finally, we included our RS as another quality feature for fusion calls. We noticed that in some cases certain fusion genes get past the FTS filter while the same fusion gene, detected in many other samples, gets filtered out. This might be explained by one of the two following reasons: (I) The fusion gene is present only in a very small subclone, so that expression of this fusion is usually too low to pass the FTS filter, but in some patients this fusion gene happens to show enough expression (e.g., present in larger clones) for reaching the FTS cutoff. (II) Under certain conditions, fusion-supporting read artifacts mimic sufficient expression to pass the FTS filter. In either case, the identified fusion event does not provide sufficient relevance to be considered as a robust candidate.

Taken together, our filtering workflow provided a well-founded procedure for the exclusion of putative false positive fusion events that were detected by current RNA-seq methods. This was supported by our study of nearly a thousand AML patients (Chapter 2) presenting the discovery of a novel recurrent fusion gene and further robust fusion candidates.

Conclusion

RNA-seq is a powerful tool for fusion gene detection and concurrent measurement of gene expression. Detection performance was shown to be influenced by sequencing properties such as sequencing depth or protocol. Furthermore, a high number of fusion calls reported by current detection algorithms are artifacts and careful filtering is required for robust fusion gene identification. Therefore, we developed a fusion detection workflow with integrated filtering strategies and identified many clinically relevant fusion genes that were not reported by routine diagnostics. We showed that RNA-seq constitutes a valuable complementary tool in clinical diagnostics for reliable transcriptome-wide identification of fusion genes and comprehensive gene expression analysis. Moreover, RNA-seq has the potential to discover novel fusion events painting a more complete picture of the genetic landscape in malignancy, which we demonstrated by the detection of *NRIP1-MIR99AHG*, a novel recurrent fusion gene in AML resulting from an inversion of chromosome 21.

References

1. Alonso CM, Llop M, Sargas C, et al. Clinical Utility of a Next-Generation Sequencing Panel for Acute Myeloid Leukemia Diagnostics. J Mol Diagnostics 2019;21(2):228–240.
2. Dillon LW, Hayati S, Roloff GW, et al. Targeted RNA-sequencing for the quantification of measurable residual disease in acute myeloid leukemia. Haematologica 2019;104(2):297–304.
3. Onecha E, Linares M, Rapado I, et al. A novel deep targeted sequencing method for minimal residual disease monitoring in acute myeloid leukemia. Haematologica 2019;104(2):288–296.
4. Ishida H, Iguchi A, Aoe M, et al. Panel-based next-generation sequencing facilitates the characterization of childhood acute myeloid leukemia in clinical settings. Biomed Reports 2020;13(5):1–10.
5. Engvall M, Cahill N, Jonsson BI, Höglund M, Hallböök H, Cavelier L. Detection of leukemia gene fusions by targeted RNA-sequencing in routine diagnostics. BMC Med Genomics 2020;13(1):106.
6. Su Z, Łabaj PP, Li S, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol 2014;32(9):903–914.
7. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17(1):13.
8. Zhang C, Zhang B, Lin LL, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics 2017;18(1):583.
9. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun 2019;10(1):3240.
10. Sahraeian SME, Mohiyuddin M, Sebra R, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun 2017;8(1):1–15.
11. Arindrarto W, Borràs DM, de Groen RAL, et al. Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. Leukemia 2020;1–15.

12. Duncavage EJ, Tandon B. The utility of next-generation sequencing in diagnosis and monitoring of acute myeloid leukemia and myelodysplastic syndromes. Int J Lab Hematol 2015;37(S1):115–121.

13. Shumilov E, Flach J, Kohlmann A, et al. Current status and trends in the diagnostics of AML and MDS. *Blood Reviews* 2018;32(6):508–519.

14. Carbonell D, Suárez-González J, Chicano M, et al. Next-generation sequencing improves diagnosis, prognosis and clinical management of myeloid neoplasms. Cancers (Basel) 2019;11(9):1364.

15. Haferlach T. Advancing leukemia diagnostics: Role of next generation sequencing (ngs) in acute myeloid leukemia. Hematol Rep 2020;12(S1):1–12.

16. Bennett JM, Catovsky D, Daniel M -T, et al. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. Br J Haematol 1976;33(4):451–458.

17. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: Rationale and important changes. *Blood* 2009;114(5):937–951.

18. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 2016;127(20):2391–2405.

19. Döhner H, Estey EH, 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.

20. 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.

21. Network TCGAR. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med 2013;368(22):2059–2074.

22. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med 2016;374(23):2209–2221.

23. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018;562(7728):526–531.

24. Heath EM, Chan SM, Minden MD, Murphy T, Shlush LI, Schimmer AD. Biological and

clinical consequences of NPM1 mutations in AML. *Leukemia* 2017;31(4):798–807.

25. Grisendi S, Bernardi R, Rossi M, et al. Role of nucleophosmin in embryonic development and tumorigenesis. Nature 2005;437(7055):147–153.

26. Bolli N, Nicoletti I, De Marco MF, et al. Born to be exported: COOH-terminal nuclear export signals of different strength ensure cytoplasmic accumulation of nucleophosmin leukemic mutants. Cancer Res 2007;67(13):6230–6237.

27. Grimwade D, Hills RK, Moorman A V., et al. Refinement of cytogenetic classification in acute myeloid leukemia: Determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. Blood 2010;116(3):354–365.

28. Löwenberg B, Ossenkoppele GJ, van Putten W, et al. High-Dose Daunorubicin in Older Patients with Acute Myeloid Leukemia. N Engl J Med 2009;361(13):1235–1248.

29. Röllig C, Bornhäuser M, Thiede C, et al. Long-term prognosis of acute myeloid leukemia according to the new genetic risk classification of the European leukemianet recommendations: Evaluation of the proposed reporting system. J Clin Oncol 2011;29(20):2758–2765.

30. Kantarjian H. Acute myeloid leukemia-Major progress over four decades and glimpses into the future. Am J Hematol 2016;91(1):131–145.

31. Abaza Y, Kantarjian H, Garcia-Manero G, et al. Long-term outcome of acute promyelocytic leukemia treated with all-trans-retinoic acid, arsenic trioxide, and gemtuzumab. Blood 2017;129(10):1275–1283.

32. Coombs CC, Tavakkoli M, Tallman MS. Acute promyelocytic leukemia: where did we start, where are we now, and the future. Blood Cancer J 2015;5(4):e304–e304.

33. Stone RM, Mandrekar SJ, Sanford BL, et al. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation . N Engl J Med 2017;377(5):454–464.

34. Perl AE, Martinelli G, Cortes JE, et al. Gilteritinib or Chemotherapy for Relapsed or Refractory FLT3 -Mutated AML . N Engl J Med 2019;381(18):1728–1740.

35. Chifotides HT, Masarova L, Alfayez M, et al. Outcome of patients with IDH1/2-mutated post–myeloproliferative neoplasm AML in the era of IDH inhibitors. *Blood Advances* 2020;4(21):5336–5342.

36. Talati C, Sweet KL. Nuclear transport inhibition in acute myeloid leukemia: recent

advances and future perspectives. Int J Hematol Oncol 2018;7(3):IJH04.

37. Maude SL, Frey N, Shaw PA, et al. Chimeric Antigen Receptor T Cells for Sustained Remissions in Leukemia. N Engl J Med 2014;371(16):1507–1517.

38. Liu F, Cao Y, Pinz K, et al. First-in-Human CLL1-CD33 Compound CAR T Cell Therapy Induces Complete Remission in Patients with Refractory Acute Myeloid Leukemia: Update on Phase 1 Clinical Trial. Blood 2018;132(Supplement 1):901–901.

39. US National Library of Medicine. *ClinicalTrials.gov [Internet]*. https://clinicaltrials.gov/ct2/results?cond=Acute+Myeloid+Leukemia&term=CAR-T&cntry=&state=&city=&dist= (accessed January 12, 2021).

40. Wu H, Li X, Li H. Gene fusions and chimeric RNAs, and their implications in cancer. *Genes and Diseases* 2019;6(4):385–390.

41. Rowley JD. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. Ann Genet 1973;16(2):109–112.

42. Rowley JD. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243(5405):290–293.

43. Erickson P, Gao J, Chang KS, et al. Identification of breakpoints in t(8;21) acute myelogenous leukemia and isolation of a fusion transcript, AML1/ETO, with similarity to Drosophila segmentation gene, runt. Blood 1992;80(7):1825–1831.

44. Shtivelman E, Lifshitz B, Gale RP, Canaani E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature 1985;315(6020):550–554.

45. Bohlander SK. Fusion genes in leukemia: An emerging network. Cytogenet Cell Genet 2000;91(1–4):52–56.

46. Teixeira MR. Recurrent fusion oncogenes in carcinomas. *Critical Reviews in Oncogenesis* 2006;12(3–4):257–271.

47. Jimenez JJ, Chale RS, Abad AC, Schally A V, Frost P. Acute promyelocytic leukemia (APL): a review of the literature. Impact Journals.; 992–1003 p.

48. Gurrieri C, Nafa K, Merghoub T, et al. Mutations of the PML tumor suppressor gene in acute promyelocytic leukemia. Blood 2004;103(6):2358–2362.

49. Schachter-Tokarz E, Kelaidi C, Cassinat B, et al. PML-RARα ligand-binding domain deletion mutations associated with reduced disease control and outcome after first

relapse of APL. *Leukemia* 2010;24(2):473–476.

50. Goto E, Tomita A, Hayakawa F, Atsumi A, Kiyoi H, Naoe T. Missense mutations in PML-RARA are critical for the lack of responsiveness to arsenic trioxide treatment. Blood 2011;118(6):1600–1609.

51. Madan V, Shyamsunder P, Han L, et al. Comprehensive mutational analysis of primary and relapse acute promyelocytic leukemia. Leukemia 2016;30(8):1672–1681.

52. Iaccarino L, Ottone T, Alfonso V, et al. Mutational landscape of patients with acute promyelocytic leukemia at diagnosis and relapse. Am J Hematol 2019;94(10):1091–1097.

53. Li J, Zhong H-Y, Zhang Y, et al. *GTF2I-RARA* is a novel fusion transcript in a t(7;17) variant of acute promyelocytic leukaemia with clinical resistance to retinoic acid. Br J Haematol 2015;168(6):904–908.

54. Yan W, Li J, Zhang Y, et al. RNF8 is responsible for ATRA resistance in variant acute promyelocytic leukemia with GTF2I/RARA fusion, and inhibition of the ubiquitin-proteasome pathway contributes to the reversion of ATRA resistance. Cancer Cell Int 2019;19(1):84.

55. Sobas M, Talarn-Forcadell MC, Martínez-Cuadrón D, et al. PLZF-RARα, NPM1-RARα, and Other Acute Promyelocytic Leukemia Variants: The PETHEMA Registry Experience and Systematic Literature Review. Cancers (Basel) 2020;12(5):1313.

56. Gelehrter TD, Collins FS, Ginsburg D. Principles of Medical Genetics. 2nd ed. Baltimore: Lippincott Williams & Wilkins; 1998. 153–194 p.

57. Liu S, Tsai W-H, Ding Y, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Res 2016;44(5):e47.

58. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep 2016;6(1):21597.

59. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol 2019;20(1):213.

60. Ferrari S, Torelli U, Selleri L, et al. Study of the levels of expression of two oncogenes, c-myc and c-myb, in acute and chronic leukemias of both lymphoid and myeloid lineage.

Leuk Res 1985;9(7):833–842.

61. Mavilio F, Sposi NM, Petrini M, et al. Expression of cellular oncogenes in primary cells from human acute leukemias. Proc Natl Acad Sci U S A 1986;83(12):4394–4398.

62. Wang C, Curtis JE, Geissler EN, McCulloch EA, Minden MD. The expression of the proto-oncogene C-kit in the blast cells of acute myeloblastic leukemia. Leukemia 1989;3(10):699–702.

63. Amson R, Sigaux F, Przedborski S, Flandrin G, Givol D, Telerman A. The human protooncogene product p33pim is expressed during fetal hematopoiesis and in diverse leukemias. Proc Natl Acad Sci 1989;86(22):8857 LP – 8861.

64. Tanaka T, Tanaka K, Ogawa S, et al. An acute myeloid leukemia gene, AML1, regulates hemopoietic myeloid cell differentiation and transcriptional activation antagonistically by two alternative spliced forms. EMBO J 1995;14(2):341–350.

65. Carow CE, Levenstein M, Kaufmann SH, et al. Expression of the hematopoietic growth factor receptor FLT3 (STK-1/Flk2) in human leukemias. Blood 1996;87(3):1089–1096.

66. Raspadori D, Lauria F, Ventura MA, et al. Incidence and prognostic relevance of CD34 expression in acute myeloblastic leukemia: Analysis of 141 cases. Leuk Res 1997;21(7):603–607.

67. Inoue K, Sugiyama H, Ogawa H, et al. WT1 as a new prognostic factor and a new marker for the detection of minimal residual disease in acute leukemia. Blood 1994;84(9):3071–3079.

68. Heuser M, Beutel G, Krauter J, et al. High meningioma 1 (MN1) expression as a predictor for poor outcome in acute myeloid leukemia with normal cytogenetics. Blood 2006;108(12):3898–3905.

69. Baldus CD, Tanner SM, Ruppert AS, et al. BAALC expression predicts clinical outcome of de novo acute myeloid leukemia patients with normal cytogenetics: A Cancer and Leukemia Group B study. Blood 2003;102(5):1613–1618.

70. Marcucci G, Baldus CD, Ruppert AS, et al. Overexpression of the ETS-related gene, ERG, predicts a worse outcome in acute myeloid leukemia with normal karyotype: A Cancer and Leukemia Group B study. J Clin Oncol 2005;23(36):9234–9242.

71. Van Waalwijk van Doorn-Khosrovani SB, Erpelinck C, Van Putten WLJ, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: A study of 319 de novo AML

patients. Blood 2003;101(3):837–845.

72. Haferlach T, Kohlmann A, Wieczorek L, et al. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group. J Clin Oncol 2010;28(15):2529–2537.

73. Shivarov V, Bullinger L. Expression profiling of leukemia patients: Key lessons and future directions. Exp Hematol 2014;42(8):651–660.

74. Heuser M, Wingen LU, Steinemann D, et al. Gene-expression profiles and their association with drug resistance in adult acute myeloid leukemia. Haematologica 2005;90(11):1484–1492.

75. Metzeler KH, Hummel M, Bloomfield CD, et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. Blood 2008;112(10):4193–4201.

76. Ng SWK, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature 2016;540(7633):433–437.

77. Herold T, Jurinovic V, Batcha AMN, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. Haematologica 2018;103(3):456–465.

78. Matera AG, Terns RM, Terns MP. Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews Molecular Cell Biology* 2007;8(3):209–220.

79. Morris K V., Mattick JS. The rise of regulatory RNA. *Nature Reviews Genetics* 2014;15(6):423–437.

80. Zimta A-A, Tomuleasa C, Sahnoune I, Calin GA, Berindan-Neagoe I. Long Non-coding RNAs in Myeloid Malignancies. Front Oncol 2019;91048.

81. Garzon R, Volinia S, Papaioannou D, et al. Expression and prognostic impact of lncRNAs in acute myeloid leukemia. Proc Natl Acad Sci U S A 2014;111(52):18679–18684.

82. Schwarzer A, Emmrich S, Schmidt F, et al. The non-coding RNA landscape of human hematopoiesis and leukemia. Nat Commun 2017;8(1):1–17.

83. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. Cell 2014;157(2):369–381.

84. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. Cancer Cell 2014;25(4):415–427.

85. Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res 2021;31(3):448–460.

86. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 2014;11650.

87. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One 2017;12(5):e0177459.

88. Stopka T, Zakova D, Fuchs O, et al. Chromatin remodeling gene SMARCA5 is dysregulated in primitive hematopoietic cells of acute leukemia. Leukemia 2000;14(7):1247–1252.

89. Jin Q, Mao X, Li B, Guan S, Yao F, Jin F. Overexpression of SMARCA5 correlates with cell proliferation and migration in breast cancer. Tumor Biol 2015;36(3):1895–1902.

90. Zhao XC, An P, Wu XY, et al. Overexpression of hSNF2H in glioma promotes cell proliferation, invasion, and chemoresistance through its interaction with Rsf-1. Tumor Biol 2016;37(6):7203–7212.

91. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Res 2018;46(D1):D794–D801.

92. Emmrich S, Rasche M, Schöning J, et al. miR-99a/100~125b tricistrons regulate hematopoietic stem and progenitor cell homeostasis by shifting the balance between TGFβ and Wnt signaling. Genes Dev 2014;28(8):858–874.

93. Emmrich S, Streltsov A, Schmidt F, Thangapandi VR, Reinhardt D, Klusmann JH. LincRNAs MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. Mol Cancer 2014;13(1):171.

94. Zhang R, Kim YM, Yang X, Li Y, Li S, Lee JY. A possible 5'-NRIP1/UHRF1-3' fusion gene detected by array CGH analysis in a Ph+ ALL patient. Cancer Genet 2011;204(12):687–691.

95. Stengel A, Shahswar R, Haferlach T, et al. Whole transcriptome sequencing detects a large number of novel fusion transcripts in patients with AML and MDS. Blood Adv 2020;4(21):5393–5401.

96. Herold T, Jurinovic V, Metzeler KH, et al. An eight-gene expression signature for the

prediction of survival and time to treatment in chronic lymphocytic leukemia. Leukemia 2011;25(10):1639–1645.

97. Lapierre M, Castet-Nicolas A, Gitenay D, et al. Expression and role of RIP140/NRIP1 in chronic lymphocytic leukemia. J Hematol Oncol 2015;8(1):20.

Acknowledgements

I am very grateful for the opportunity to work on highly interesting projects in the field of leukemia research. It was a wonderful experience to be part of this immensely motivated scientific community who allowed me to learn, exchange knowledge and who supported me in every way. I want to thank my doctoral supervisor Prof. Dirk Metzler and my group leader PD Dr. Philipp Greif for their important guidance throughout my studies. Also, I would like to thank Dr. Ines Hellmann for her support and commitment as member of my thesis advisory committee. Furthermore, I would like to thank all my lab colleagues who were always ready to provide a helping hand. I highly appreciate your kindness, patience and encouragement that I received and would like to especially thank Anja Swoboda, William Keay and Bianka Ksienzyk who helped me with my wet-lab experiments and many other things. Also, I would like to thank Vanessa Arfelli, Enric Redondo Monte and Alessandra Caroleo for the solicitude, solidarity and the tremendous support that you have given me. I hugely value your friendship and the amazing time that I had with you in the last years. Most importantly, I would like to thank my family who was always there for me, especially my mother Mila Kerbs. You made it possible that I can pursue my dreams and interests. I am deeply thankful to have you.

Appendix

International Journal of
Molecular Sciences

MDPI

Article

Loss of ISWI ATPase SMARCA5 (SNF2H) in Acute Myeloid Leukemia Cells Inhibits Proliferation and Chromatid Cohesion

Tomas Zikmund [1,†], Helena Paszekova [1,†], Juraj Kokavec [1], Paul Kerbs [2,3,4], Shefali Thakur [1], Tereza Turkova [1], Petra Tauchmanova [1], Philipp A. Greif [2,3,4] and Tomas Stopka [1,*]

1 Biocev, 1st Medical Faculty, Charles University, 25250 Vestec, Czech Republic; tomzikmund@gmail.com (T.Z.); paszekova.helena@gmail.com (H.P.); juraj.kokavec@gmail.com (J.K.); shefalithakur.st@gmail.com (S.T.); tereza.turkova@volny.cz (T.T.); petra-tauchmanova@seznam.cz (P.T.)
2 Department of Medicine III, University Hospital, LMU Munich, D-80539 Munich, Germany; paul.kerbs@med.uni-muenchen.de (P.K.); pgreif@med.uni-muenchen.de (P.A.G.)
3 German Cancer Consortium (DKTK), partner site Munich, D-80336 Munich, Germany
4 German Cancer Research Center (DKFZ), D-69120 Heidelberg, Germany
* Correspondence: tstopka@lf1.cuni.cz; Tel.: +420-32587-3001
† These authors contributed equally.

Received: 26 February 2020; Accepted: 16 March 2020; Published: 18 March 2020

Abstract: ISWI chromatin remodeling ATPase SMARCA5 (SNF2H) is a well-known factor for its role in regulation of DNA access via nucleosome sliding and assembly. SMARCA5 transcriptionally inhibits the myeloid master regulator PU.1. Upregulation of SMARCA5 was previously observed in CD34+ hematopoietic progenitors of acute myeloid leukemia (AML) patients. Since high levels of SMARCA5 are necessary for intensive cell proliferation and cell cycle progression of developing hematopoietic stem and progenitor cells in mice, we reasoned that removal of SMARCA5 enzymatic activity could affect the cycling or undifferentiated state of leukemic progenitor-like clones. Indeed, we observed that CRISPR/cas9-mediated *SMARCA5* knockout in AML cell lines (S5KO) inhibited the cell cycle progression. We also observed that the *SMARCA5* deletion induced karyorrhexis and nuclear budding as well as increased the ploidy, indicating its role in mitotic division of AML cells. The cytogenetic analysis of S5KO cells revealed the premature chromatid separation. We conclude that deleting SMARCA5 in AML blocks leukemic proliferation and chromatid cohesion.

Keywords: SMARCA5; SNF2H; AML; leukemia; CRISPR; therapeutic target

1. Introduction

Acute myeloid leukemia (AML) is a malignant hematopoietic disease derived from myeloid-primed stem cells resulting in accumulation of myeloid blasts. AML patients have a poor prognosis and the only known efficient therapy is bone marrow transplantation combined with chemotherapy. Next-generation sequencing revealed that despite similar cytology and cellular features, the mutational profile of AML clones can be very heterogenic. Leukemogenesis involves multiple types of genomic alterations from single nucleotide variants to large chromosomal abnormalities (involving deletions, translocations, or chromosomal gains and losses). Targets of mutagenesis are often genes encoding regulators of gene transcription (e.g., *RUNX1*, *CEBPA*, *GATA2*), DNA methylation (e.g., *DNMT3A*, *IDH1*, *IDH2*), and genome organization (e.g., *CTCF*, *RAD21*, *SMC3*).

Immature cells during tissue development require ATP-dependent chromatin remodeling activities to ensure accession of regulatory proteins to DNA in order to control replication, transcription, or DNA repair. Activities that facilitate nucleosome spacing and assembly during tissue development are

provided mainly by evolutionary conserved Swi2/Snf2 family helicases. Smarca5 (also known as Snf2h) belongs to important enzymes of the Swi2/Snf2 family with remodeling activity that is required for successful hematopoietic development in mammals [1–3]. In mouse, Smarca5 represents the catalytic subunit of ISWI remodeling complexes that is indispensable for developing embryo and later for fetal hematopoiesis [1,2]. Interestingly, Smarca5 loss was accompanied by upregulation of p53 and of its transcriptional targets that are usually linked to the induction of apoptosis in response to DNA damage (e.g., p21/Cdkn1a, Noxa/Pmaip1, and Bax) [1]. Our work and the work of others suggested that Smarca5 not only facilitates proliferation-associated events but also helps to activate transcriptional programs of particular developmental stages to set proper expression identity of immature cells [4,5]. Additional evidence implicated that Smarca5 regulates global gene expression programs and function of many human gene regulatory elements by cooperating with CTCF [6–8].

Smarca5 represents an integral part of heterodimeric ISWI complexes that contain usually a bromodomain-containing protein (BAZ1A, BAZ1B, BAZ2A, BAZ2B). ISWI complexes were originally identified in *Drosophila* but later they were discovered also in humans, namely, NURF (ATPase motor of the nucleosome remodeling factor), ACF (ATP-utilizing chromatin assembly and remodeling factor), and CHRAC (chromatin assembly complex). Later, additional human complexes were found, such as RSF, NoRC, WICH, CERF, and finally, SNF2H-cohesin [9]. Most ISWI complexes are involved in regulating cell cycle progression albeit via different mechanisms. While many ISWI complexes regulate transcription by nucleosome sliding mechanism utilizing either RNA-Polymerase 1 (RNAP1) (NoRC, B-WICH) or RNAP2 (ACF, NURF, CERF, WINAC), other complexes are linked to replication/repair (CHRAC, WICH) or chromatid cohesion (SNF2H-cohesin) [10]. It appears that SMARCA5 plays an indispensable part in the ISWI complexes (albeit it can remodel chromatin alone in acellular systems); however, in certain situations, it may be replaced within ISWI complexes by its close homologue SMARCA1 (SNF2L) as shown in rather differentiated cells of the cerebellum [4].

Currently, over 20% of all malignancies carry mutations in one of the subunits of chromatin remodeling complexes of the SWI/SNF family (see [11,12]). These mutations often decrease protein stability and cause loss of the particular subunit, which leads to the assembly of incomplete remodeling complexes with different functions in vivo and altered capability to precisely regulate gene expression [13]. In the case of the ISWI subfamily, the mutations of various ISWI subunits identified in oncologic diseases have still yet unknown impact on tumorigenesis. In solid tumors the overexpression of SMARCA5 [14–18] has been associated with disease aggressiveness, chemoresistance and proliferation activity [7]. SMARCA5 expression was found dysregulated in many human malignant tumors, such as aggressive gastric cancer, breast cancer, or prostate cancer. In addition, the *SMARCA5* gene is a target of cancer-associating miRNA regulation [14–18]. SMARCA5 overexpression has been also observed in AML CD34+ progenitors [7,19]. SMARCA5, through the interaction with CTCF in leukemic cells, actively inhibits expression of the *SPI1/PU.1* gene [7] that represents key hematopoietic transcription factor and dose-dependent leukemia suppressor [20]. Additional work utilizing the CRISPR/Cas9 genome editing in vitro revealed that among hematopoietic cancer cell lines, those derived from AML patients were the most SMARCA5 dependent [21]. We herein studied the consequences of *SMARCA5* deletion in AML cells and showed that SMARCA5 targeting affected proliferation and resulted in chromosomal aberrations and polyploidy pointing to the role of SMARCA5 in mitotic division. We believe that delineating the effects of *SMARCA5* targeting might pave the way for new approaches in the therapy of AML.

2. Results

2.1. SMARCA5 Overexpression Marks the Hyperproliferation and Cytogenetically Abnormal AML Patients

Based on previous evidence documenting SMARCA5 overexpression in small AML patient subset [19], we examined RNAseq data of bone marrow samples from AML patients with recorded overall survival (OS). We confirmed our previous observation [19] that *SMARCA5* levels are significantly

Int. J. Mol. Sci. **2020**, *21*, 2073 3 of 13

elevated at the time of diagnosis and decreased after the patients achieved complete hematologic remission (Figure 1A). We next associated SMARCA5 expression and clinical parameters and (due to genetic AML heterogeneity) followed separately cytogenetically normal (CN) and abnormal (CX) AML patients. Hence, we could observe a trend for decreased OS in the AML patient population with higher SMARCA5 expression and carrying cytogenetic abnormalities (Figure 1B). We also observed that higher *SMARCA5* levels correlated with mRNA expression of proliferation biomarkers such as AURKA, PLK1, CCNA2, CENPF (Figure 1C).

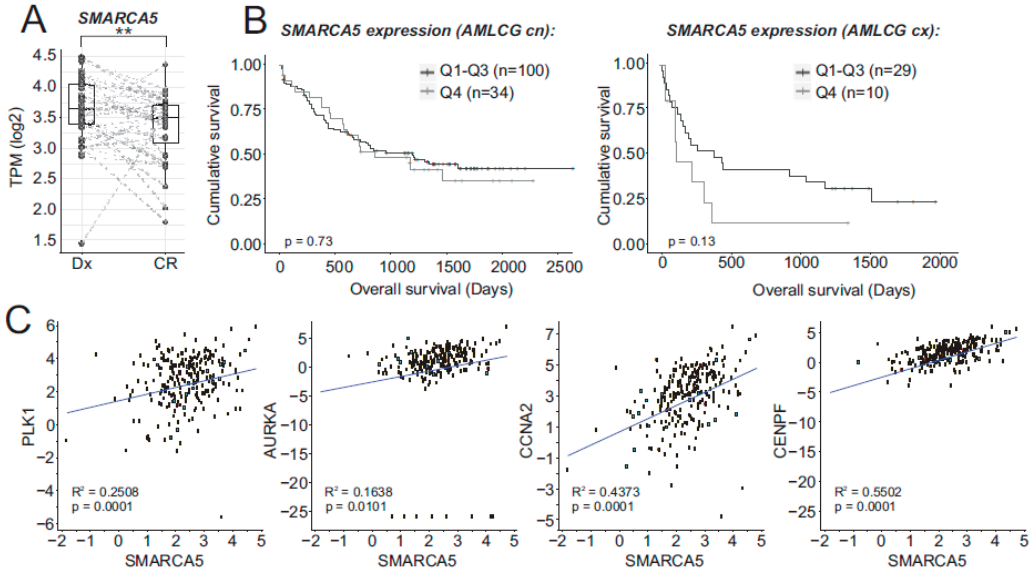


Figure 1. (**A**) SMARCA5 expression of matched AML samples at the time of diagnosis (Dx) and complete remission (CR). Dots represent individual samples; dashed lines connect matched patient samples. Boxes: distribution of the Dx and CR groups; intermediate line = median. Significance was estimated using a paired Wilcoxon test. (**B**) Survival analysis of AML patients divided into quartiles (from low Q1 to high Q4; Q1: 0–25% + Q2: 25–50% + Q3: 50–75% vs. Q4: 75–100%) based on SMARCA5 mRNA levels (cn: cytogenetically normal, cx: cytogenetic abnormalities). (**C**) Correlation of mRNA levels of PLK1, AURKA, CCNA2, CENPF, and SMARCA5 (R^2 and p-value indicated).

2.2. *SMARCA5 Deletion Inhibits AML Cell Proliferation*

To test requirement of SMARCA5 for AML cell growth, we produced a null allele using CRIPSR/Cas9 genome editing technology (Figure 2A). Targeted was exon5, which codes a portion of evolutionarily conserved ATPase domain and that was previously shown to be a targetable region using the Cre-loxP1 system. Deletion of exon5 results in a frame shift mutation disabling expression of Smarca5 protein in mouse [1]. For the experiments, human K562 cells (AML M6 subtype) were initially utilized as they were previously used for antisense oligonucleotide-mediated transient knockdown of SMARCA5 [2]. K562 cells were transfected by a pair of pX330-mVenus vectors containing sgRNAs complementary to a sequence in the SMARCA5 introns 4 & 5 and the the effect of CRISPR/Cas9-mediated deletion of exon5 was tested by PCR. Analysis of fragments amplified from genomic DNA of FACS-sorted mVenus-positive clonal populations identified 5 clones (#H10, D7, H4, E7, H7) with a single shortened PCR product (~632bp compared to 1175bp in controls) that were homozygously mutated (Figure 2B). Sanger sequencing of PCR products confirmed that clones H10, D7, E7, and H5 contained the same deletion (543bp) and clone H4 an even larger deletion (582bp) within SMARCA5 exon5 (Figure 2C). In addition, quantitative PCR and Western blot analyses of the cellular extracts confirmed that the Cas9-mediated deletion of the SMARCA5 gene resulted in loss of SMARCA5 expression (Figure 2D,E). The resulting subclones had no expression of vector-coded & episomally expressed Cas9 nuclease. In addition, eight predicted off-target candidates (SRGAP2, RNF17, PRG4,

Int. J. Mol. Sci. **2020**, *21*, 2073

GYPA, POLQ, CYB5R4, BCKDHB, NAV2) had no alteration of their sequences. Thus, we managed to effectively delete SMARCA5 in the K562 subclones to create a cellular model for studying how SMARCA5 loss affected AML cell growth.

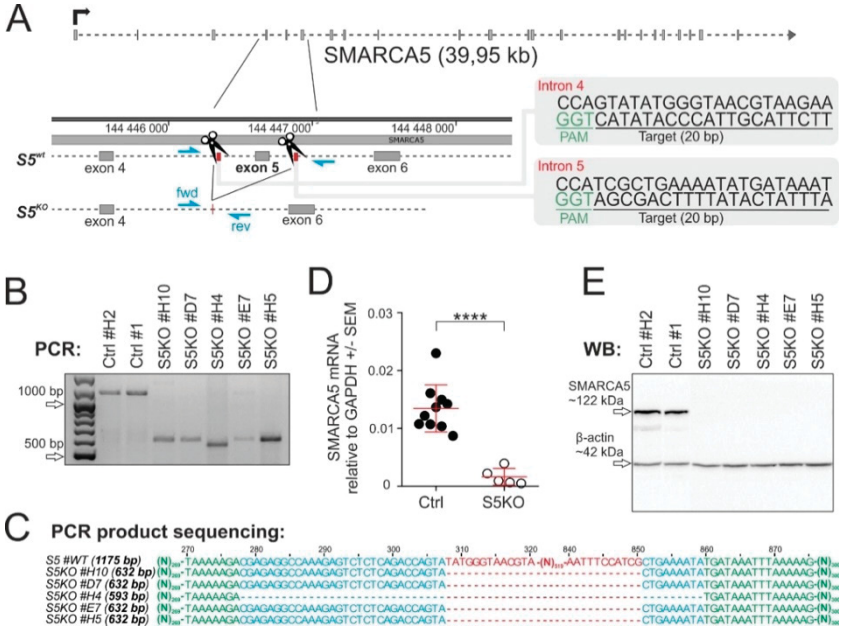


Figure 2. Inactivation of *SMARCA5* gene expression (S5KO) in AML cells. (**A**) Scheme of generating the S5KO using CRISPR/Cas9 technology. The Cas9 nuclease was targeted into two intronic sites (scissors) surrounding exon5 of the *SMARCA5* gene. The sequences of guide RNAs are depicted in gray boxes on the right. Indicated are exons 4-6 (small rectangles) and genotyping primers (blue arrowheads). (**B**) PCR verification of the exon5 deletion in the indicated S5KO clones. (**C**) Analysis of *SMARCA5* gene region following the Cas9 nuclease deletion. PCR products (same as in B) were Sanger-sequenced and aligned with the wt sequence using the Kalign web tool. After sequencing, the precise length of the resultant PCR amplified region was determined (on the left in brackets). (**D**) Quantitative PCR analysis of *SMARCA5* mRNA expression in the S5KO clones (n = 5) compared to controls (n = 10). Data normalized to the GAPDH mRNA. Student's t-test, $p < 0.00001$ ****. (**E**) Immunoblotting of SMARCA5 expression in CRISPR/Cas9-treated K562 or controls. β-actin controlled the load.

2.3. Smarca5 Deletion Inhibits Proliferation of Myeloblasts and Affects Function of Normal Stem Cells

To characterize the effect of *SMARCA5* deletion in the AML-S5KO subclones, we monitored their growth in culture by the WST-1 assay correlating the number of metabolically active cells in the 72-hr culture within a 96-well plate. We quantitated the data with a scanning multiwell spectrophotometer (ELISA reader) (Figure 3A, upper panel) and also in parallel counted the viable cells with an automated cell counter (Figure 3A, lower panel). We observed that starting day 1, the S5KO subclones produced less formazan product/s compared to AML 'control' cells, indicating that loss of SMARCA5 impaired proliferation of leukemic cells. We also attempted to create S5KO clones from additional AML cell lines. We repeatedly used OCI-M2, NB4, SKM1, MOLM-13, however, despite the fact that these AML cell lines grew normally in tissue culture conditions, the recombined cells by pX330-mVenus vectors followed by the single cell sorting could not produce clones with exon5 deletion. We therefore used the method of serial dilution of transfected cells. This approach, in contrast to the previous approach, produced populations of OCI-M2 and SKM1 cell lines with detectable Cas9-edited *SMARCA5* loci. However, the signals of mutated alleles markedly decreased during long-term cultivation, suggesting that the S5KO cells were overgrown by cells containing at least one intact SMARCA5 allele. Thus, the deletion of the *SMARCA5* gene completely impaired leukemic cell proliferation in most of the AML cell lines,

while in K562 cells it was tolerated albeit under markedly lower proliferation activity, which allowed us to study it in more detail.

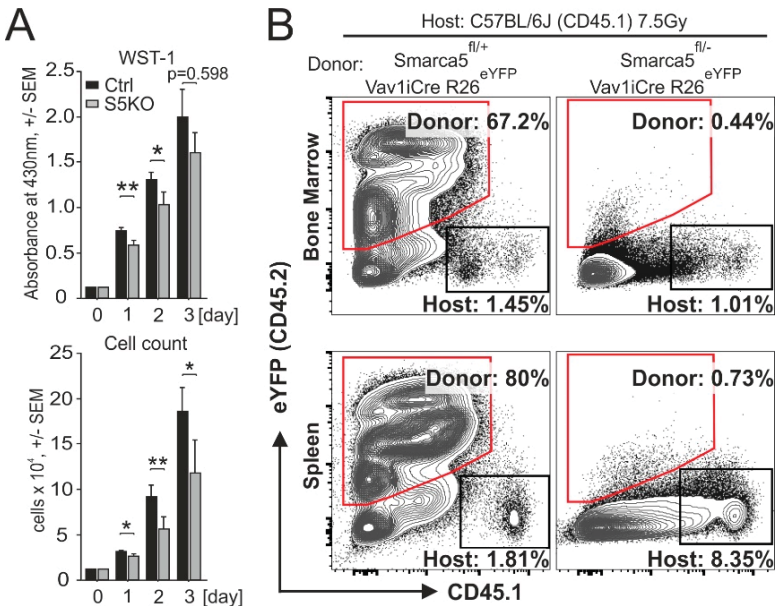


Figure 3. Proliferation of AML and progenitor cells upon *Smarca5* gene deletion. (**A**) Proliferation of S5KO clone #D7 and control cells analyzed by WST-1 assay. Mean ± SEM of formazan absorbance (top) and cell count (bottom) (pentaplicates). Student's *t*-test, $p < 0.05$ *, $p < 0.01$ **. (**B**) Flow cytometry analysis of donor (CD45.2) and host (CD45.1) derived hematopoietic cells at 14 days following the transplantation of donor fetal liver cells into lethally (7.5 Gy) irradiated host animals. Donor (red trapezoid) and host-derived (black rectangles) bone marrow cells (upper dot plots) and splenocytes (lower dot plots) were distinguished by the expression of yellow fluorescent protein (eYFP) or surface variant of CD45. Control mice: Smarca5$^{fl/+}$ Vav1iCre R26^{eYFP}; Smarca5 mutant mice: Smarca5$^{fl/fl}$ Vav1iCre R26^{eYFP}. Data are representative of repeated experiments.

AML cell population resembles early hematopoietic progenitors. Thus, as controls to AML cells, we studied early murine blood progenitors. Previously it was shown that Smarca5 loss in mouse partially inhibits differentiation of early Lin^{-}Sca-1^{+}c-Kit^{+} hematopoietic progenitors [1]. To test whether Smarca5 deletion affects reconstitution of early blood progenitors after transplanting them into normal murine recipients, we utilized the hematopoietic reconstitution assay. We transferred E13.5 mouse fetal liver cells (C57Bl/6J Ly5.2 background) isolated either from control Smarca5$^{flox/+}$ Rosa26$^{eYFP/+}$ Vav1-iCRE or *Smarca5*-deficient (Smarca5$^{flox/-}$ Rosa26$^{eYFP/+}$ Vav1-iCRE) embryos into lethally irradiated adult C57Bl/6J Ly5.1 recipients. Flow cytometric analyses of bone marrow and spleen at several weeks after transplantation revealed that repopulation was detected only in animals transplanted with cells in which the *Smarca5* gene was preserved. Thus, homeostatic expression of Smarca5 is very important for hematopoietic reconstitution (Figure 3B), implicating a possibility that the Smarca5 role in AML cells might also involve a very early leukemia-initiating compartment.

2.4. Inactivation of Smarca5 Causes Nuclear Abnormalities and Polyploidy

To gain insight into the subcellular structures of the AML S5KO cells, we utilized hematology staining using a standardized May–Grunwald and Giemsa–Romanowski stain procedure. As indicated within Figure 4A, the control AML cells were represented by a uniform layer of myeloblasts with large round nuclei, fine chromatin structure, and prominent nucleoli. Significantly more frequent nuclear abnormalities were observed in the S5KO cells compared to controls. These included nuclear budding, internuclear bridging, karyorrhexis, and multinuclearity seen in 10% to 65% of all analyzed cells (Figure 4B). To study effect/s of S5 depletion in nonhematopoietic cells, we derived mouse embryonic

Int. J. Mol. Sci. **2020**, *21*, 2073 6 of 13

fibroblast (MEF) with Tamoxifen-regulated Cre-recombinase activity (Cre-Esr1) from *Smarca5*$^{fl/fl}$ *Trp53*$^{-/-}$ animals. *Trp53*-mutated MEFs were chosen because of their lower propensity to enter proliferation senescence and because most AML cell lines including K562 have *TP53* gene inactivation [22]. After 6 h incubation with 100 nM 4-hydroxy-tamoxifen (4OHT) and additional 90 h of culture, the MEF cells were depleted from Smarca5 protein (Figure 4C). Decrease of Smarca5 protein level negatively influenced the cell growth and the proliferation defect had already occurred within 40 h from the start of the 4OHT treatment while 4OHT untreated and control Cre-Esr1 lacking cells proliferated normally (Figure 4D). This proliferative defect resembled one observed in AML S5KO clones. The flow cytometry analysis revealed that aberrant proliferation was accompanied by lower proportion of S-progressing and mitotic (pH3S10^{+}) cells. In addition, we noted a higher number of cells with polyploid nuclei (Figure 4E) that was concomitant to a decreased proportion of diploid cells upon S5 deficiency in MEFs. Taken together, inactivation of SMARCA5 triggers a cell proliferation blockade and results in nuclear abnormalities of exceedingly cycling leukemic as well as normal hematopoietic cells.

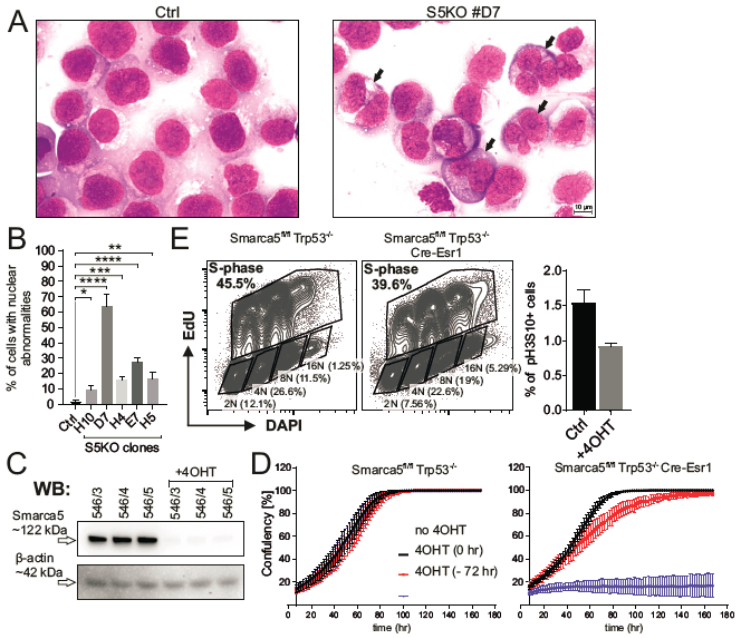
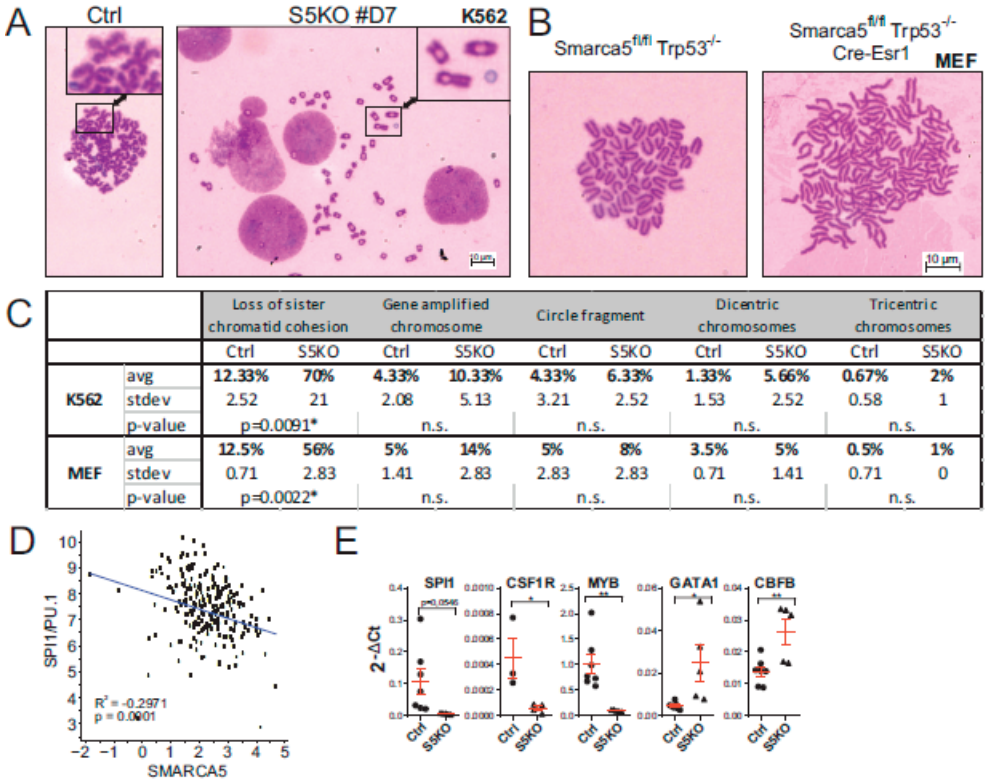


Figure 4. Nuclear abnormalities in S5KO cells. (**A**) Cytology of control (left) and S5KO clone #D7 (right), nuclear abnormalities indicated and shown (**B**) as mean % ± Stdev of control, 400 cells/subclone analyzed. Student's *t*-test, $p < 0.05$ *, $p < 0.01$ **, $p < 0.0001$ ****. (**C**) Immunoblotting of Smarca5: MEF cell lines (Smarca5$^{fl/fl}$ Cre-Esr1: untreated, 4OHT-treated (100 nM, 6 h exposure, 4 days of culture). β-actin = loading control. (**D**) IncuCyte cell proliferation analysis; control Smarca5$^{fl/fl}$ (upper panel) vs. Smarca5$^{fl/fl}$ Cre-Esr1 (lower panel) MEFs in absence/presence of 4OHT (100 nM, 6 h exposure), or alternatively, 4OHT was added 72 h prior to IncuCyte monitoring (4OHT—72 h). *Y*-axis: mean confluency (%) and ± Stdev of at least 16 different regions of the cultivation plate, *X*-axis: time (h). (**E**) Flow cytometry analysis of control and Smarca5$^{fl/fl}$ Cre-Esr1 MEF population cell cycle progression using EdU/DAPI double staining (upper dot plots). Black rectangles depict all S-phase and non-S-phase cells with different ploidy (2N-16N). Histograms show percentage of phospho-histone H3 (Ser10) positive mitotic events in experimental cell lines. (**D**) and (**E**) represent biological triplicates.

2.5. Cytogenetic Abnormalities and Gene Expression Dysregulation in the S5KO AML Cells

As pointed out in the Introduction section, SMARCA5 protein was previously shown to load cohesin complex onto human chromosomes [23]. As the canonical role of cohesin is the sister chromatid cohesion, we next analyzed the structures of mitotic chromosomes in the AML S5KO cells on metaphase spreads. The analysis of the S5KO subclone D7 consistently showed (Figure 5A) that among other chromosomal abnormalities, the cohesion defects were by far the most frequent involving premature

chromatid separation and loss of cohesion. Compared to the controls that contained only 12%, the S5KO mitotic cells displayed defects in chromatin cohesion in 70% of all cases. Similarly, the defects of chromatid cohesion were seen also in MEF cell-derived mitotic chromosome spreads (Figure 5B,C). These data suggest that SMARCA5 inhibition affects cohesin function in general.



C

		Loss of sister chromatid cohesion		Gene amplified chromosome		Circle fragment		Dicentric chromosomes		Tricentric chromosomes	
		Ctrl	S5KO	Ctrl	S5KO	Ctrl	S5KO	Ctrl	S5KO	Ctrl	S5KO
K562	avg	12.33%	70%	4.33%	10.33%	4.33%	6.33%	1.33%	5.66%	0.67%	2%
	stdev	2.52	21	2.08	5.13	3.21	2.52	1.53	2.52	0.58	1
	p-value	p=0.0091*		n.s.		n.s.		n.s.		n.s.	
MEF	avg	12.5%	56%	5%	14%	5%	8%	3.5%	5%	0.5%	1%
	stdev	0.71	2.83	1.41	2.83	2.83	2.83	0.71	1.41	0.71	0
	p-value	p=0.0022*		n.s.		n.s.		n.s.		n.s.	

Figure 5. SMARCA5 loss causes karyotypic changes in K562 cells. (**A**,**B**) Mitotic chromosome analysis of S5KO cells vs control K562 cells (clone #D7, (**A**)) or MEF cells (**B**). 1000X magnification. (**C**) Table summarizes all chromosomal aberrations; data from technical triplicates, for each replicate a total of 100 mitotic nuclei were counted. Mean percentage of chromosomal abnormalities with Stdev, Student's *t*-test, $p < 0.05$ *. (**D**) Computational analysis of correlations between expression of SPI1/PU.1 and SMARCA5 in samples of adult AML patient samples; for details, see Materials and Methods section. (**E**) Quantitative PCR analysis of SPI1, CSF1R, MYB, GATA1, and CBFB mRNAs expression in the S5KO clones ($n = 5$) compared to controls ($n = 7$). Data were normalized to the GAPDH mRNA. Student's *t*-test, $p < 0.05$ *, $p < 0.001$ **.

In order to better understand the cooperative nature of SMARCA5 and its interacting partners in AML, we correlated their expression using RNAseq data in AML patients. Hence, significant association exists between the expression pattern of SMARCA5 and BAZ proteins (BAZ1A, BAZ1B, BAZ2A, BAZ2B) as well as the members of the CTCF/cohesin complex across human AML samples. This implicates, albeit indirectly, a role of SMARCA5 in CTCF/cohesin function in AML that also coincides with karyotype abnormalities imposed by a SMARCA5 loss.

We recently showed that SMARCA5 (together with the CTCF/cohesin complex) represses PU.1-mediated myeloid differentiation [7] and similarly, we noted that SMARCA5 regulates GATA1-mediated erythropoiesis [1]. We therefore next decided to analyze the levels of SPI1/PU.1 and GATA-1 transcripts with respect to SMARCA5. As expected, transcriptomic data from AML Cooperative Group München (Figure 5D) showed an inverse correlation between SPI1/PU.1 and SMARCA5 expression in AML patient samples. To further assess the role of SMARCA5 in regulation of the hematopoietic transcription program, we determined the expression of a set of selected mRNAs upon the genetic ablation of the *SMARCA5* gene in K562 cells. Compared with previously published

data documenting an inverse relationship between SMARCA5 and hematopoietic transcription factors PU.1 or GATA-1, we observed that upon SMARCA5 deletion in K562 cells the level of SPI1/PU.1 and some of its targets (CSF1R) became downregulated while other transcription factors (GATA1, CBFB) were upregulated. The dysregulation of mRNA pattern of SMARCA5 targets upon SMARCA5 deletion can be attributed to the heterogeneity of the AML cell lines and also possibly to multiple genetic/cytogenetic abnormalities imposed by the *SMARCA5* loss.

3. Discussion

We herein studied how ISWI ATPase SMARCA5/SNF2H controls in AML the proliferation and gene expression of myeloblasts as SMARCA5 appeared to be an interesting target for anti-AML therapy. Our previous work demonstrated a pattern of SMARCA5 upregulation at AML diagnosis followed by its normalization upon achieving the hematologic remission. Importantly, additional work has not identified recurrent mutations of SMARCA5 in AML or any malignant disease (so far analyzed by next-generation sequencing-based techniques). For example, for the *SMARCA5* gene, only 186 variants with an amino acid residue substitution exist in nearly ~20 thousand oncologic patients (<1%). There also exist infrequently the variants in ISWI-interacting BAZ proteins detected in cancer, however, the significance of these variants remains also unknown. Importantly, among the AML-associated variants, only the SMARCA5-interacting proteins, CTCF and members of the cohesin complex, were shown consistently mutated in AML [24]. Based on this, we expected SMARCA5 indispensability for AML proliferation and its levels possibly reflecting the proliferative nature of AML cells. Indeed, the RNAseq analysis of a large set of AML patients confirmed that AML cells overexpressed SMARCA5 and its levels correlated with many ISWI-complex members including also cohesin complex, and finally, that the proliferative nature of AML cells marked by upregulation of SMARCA5 was supported by a trend in shorter OS albeit only in those AML patients that were marked by cytogenetic aberrations (see Figure 1).

Upon targeting of the *SMARCA5* gene in AML cell lines with a CRISPR/Cas9-mediated deletion strategy, we could observe that AML cells lacking *SMARCA5* markedly slowed the proliferation rate and became dysplastic with multiple karyotypic abnormalities. Inhibiting SMARCA5 to achieve suppression of AML growth may be thus a very efficient strategy as AML cells that are likely addicted to SMARCA5 in order to overcome various chromatin obstacles such as complex karyotype or also polyploidy often seen during progression of AML. Other data further implicated that SMARCA5 is very important also at the stem cell level to regulate their innate function: to repopulate the progeny. Indeed (as shown by Figure 3), repopulation activities were greatly reduced in normal hematopoietic stem cells in which the *Smarca5* gene was genetically deleted. Our observation, however, does not rule out the possibility of SMARCA5 being an AML target as i) the AML cells are highly proliferating compared to their normal counterparts, and ii) SMARCA5 being expressed in stem cells implicates that antiSMARCA5 therapy would preferentially target the leukemia stem and progenitor cells.

While SMARCA5 expression represents a potential target for AML therapy, it may also serve as a factor of therapeutic resistance in AML. It is likely that additional factors will be involved in modulating therapy efficacy using SMARCA5 inhibitors in the future. As the Smarca5 loss was sensed in a mouse model by a) increased p53 levels and b) associated with DNA damage response (DDR), and c) activation of the p53 targets [1], very likely the tumor cells with DDR sensing defect would have a higher propensity to tolerate SMARCA5 level downregulation. This notion is supported by our other study demonstrating that proliferation defect imposed by Smarca5 deficiency can be partly restored with concomitant *Trp53* deletion in murine thymocytes [3].

Our herein presented data indicate that AML growth is dependent on the expression of chromatin remodeling protein SMARCA5 that is a known partner of AML-associated targets: cohesin complex and CTCF [23]. Data presented in Figures 4 and 5 implicate that proliferation inhibition upon *SMARCA5* targeting is at least in part caused by karyotype abnormalities, especially cohesion defects, and possibly also by a putative replication defect due to defective chromatin compaction as well as

dysregulation of gene expression pattern of the key hematopoietic lineage restricted transcription factors. Interestingly, the nuclear changes after S5 deletion such as polyploidy were also described in other cell lines of hematologic origin [1,3] but not as a result of *Smarca5* deletion of developing brain or eye lens [4,5]. Similar evidence was noted upon experimental manipulation with cohesin complex members; for example, the nonsense mutations in STAG2 (generated in the THP1 AML cell line) led to defects in sister chromatid cohesion and induced anaphase defects, which resulted in proliferation blockade [25]. Important connections between replication and cohesion have been established in the HeLa tumor cells, in which the interfering with replication affected chromatid cohesion and caused a defect in mitotic progression [26]. Others suggested that cohesion defects depend on a functional mitotic spindle checkpoint in regulating mitotic progression [27]. It seems that the strategy of inhibiting SMARCA5 in AML to block leukemogenesis becomes even more vital as shown recently using inhibitors of SMARCA5 (ED2-AD101) that target the HELICc-DExx domain to release the terminal AML cells into differentiation while sparing normal hematopoiesis in preclinical animal models [28]. Our work also suggests that upon inhibiting SMARCA5-mediated proliferation of AML cells, we also can face the problem of inhibiting proliferation of normal cells. Further work in this respect on experimental animals is under way. An additional strategy to inhibit AML cell growth specifically could be to target the *SMARCA5* exon5 in AML cells by CRISPR/Cas9 as evidenced by the herein presented data. Data from global CRISPR/Cas9 screen identified that SMARCA5 targeting was very efficient and caused cell growth inhibition in several additional AML cell lines (OCI-AML2, OCI-AML3) and also in lymphoma and carcinoma cell lines [21]. Together, our as well as others' data demonstrate that SMARCA5 is a valuable epigenetic target suitable for inhibitor discovery projects and subsequent validation in MDS/AML and potentially also in other types of cancer.

4. Materials and Methods

4.1. CRISPR Vector Design

pX330-Venus (kindly provided by Dr. Bjoern Schuster) produces CRISPR/Cas9 enzyme that cleaves at a specific location based on sequence guide sgRNA defined target sequences in SMARCA5 intron4 (5'-TTCTTACGTTACCCATATACTGG-3') and SMARCA5 intron5 (5'-ATTTATCATATTTTCAGCGATGG-3'). CRISPR/Cas9 enzyme is also fused with fluorescent protein mVenus, that enables selection of successfully transfected clones by FACS sorting. The DNA sequences for the sgRNA SMARCA5 intron4 and sgRNA SMARCA5 intron5 were synthesized by Sigma-Aldrich as four oligonucleotides with modifications at position 1 (to encode a Guanine due to the transcription initiation requirement of the human U6 promoter). These two pairs of complementary oligos were mixed together, boiled at 95 °C for 10 min, and allowed to cool down to RT to hybridize. Double-stranded oligos also designed with complementary BbsI overhangs on 3' and 5' ends were ligated into BbsI linearized pX330-Venus vector using T4 Ligase enzyme (Thermo Fisher Scientific, Waltham, MA, USA). Ligation mixtures were transformed into Subcloning Efficiency DH5α Competent Cells (Invitrogen, Carlsbad, CA, USA) following the manufacturer's protocol. pX330-Venus sgRNA hSMARCA5 intron4 and pX330-Venus sgRNA hSMARCA5 intron5 were isolated and purified by GenElute HP Plasmid Midiprep kit (Sigma-Aldrich, St. Louis, MO, USA) and correct oligo insertion verified by Sanger sequencing.

4.2. Cell Lines

K562 cells (ATCC, Manassas, Virginia, USA) were cultured in 90% Iscove's Modified Dulbecco's Medium supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin. NB4, SKM-1, and MOLM-13 were cultured in 90% RPMI-1640 medium (Sigma-Aldrich), OCI-M2 in 80% Iscove's Modified Dulbecco's medium (Biosera, Kansas City, MO, USA) at 37 °C and 5% CO2. The media were supplemented with 10-20% fetal bovine serum (Biosera) and 1% penicillin/streptomycin (Biosera). Cell lines were purchased from DSMZ. Both pX330-Venus sgRNA SMARCA5 intron4 (1 μg) and

px330-Venus sgRNA SMARCA5 intron5 (1 μg) were transfected into 2.5×10^6 K562 cells using Amaxa Cell Line Nucleofector kit (Lonza, Basel, Switzerland) and 1×10^6 K562 cells using Neon Transfection System 10 μL Kit (Invitrogen). Cells were cultivated for 48 h, Venus-positive cells sorted on BD FACS Aria Fusion and divided to form single cell clones on 96-well plates. DNA from growing clones was used as a template for PCR with the following primers: forward 5′-GAGATGGAGGGCTACACTGTG-3′and reverse 5′-GACATTCCCAAAGTCATCTAGCAG-3′. The resulting amplification produced 1175 bp fragment from wild-type and approximately 632 bp long fragment from CRISPR/Cas9 edited allele of the SMARCA5 gene. Cell smears ($0.5–1 \times 10^5$ cells) were fixed with methanol and stained with May–Grünwald solution (mixed 1:1 with distilled water, Penta, Limassol, Cyprus) for 5 min and Giemsa–Romanowski solution (mixed 1:13 with distilled water, Penta) for 12 min. Cell Proliferation Reagent WST-1 (Roche, Basel, Switzerland) was used following manufacturer's protocol starting from day 0 with seeding 0.5 × 104 cell/100 μL/well in triplicates and continued by daily measurement of absorbance at 430 nm on microplate reader Infinite 200 PRO (Tecan, Männedorf, Switzerland). Cells were simultaneously counted by Luna Automated Cell Counter (Logos Biosystems, Dongan, South Korea).

4.3. AML Patients and Statistics

RNA-Seq data sets from AML patient samples were previously described including the informed consent and ethical issues [29–31]. Reads were mapped with STAR aligner version 2.7.2d using GRCh37 reference and annotation version 32 from GENCODE (www.gencodegenes.org). Reads were counted using FeatureCounts version 1.6.5, normalized to transcripts per million (TPM) and log2 transformed. Log-rank test was performed in survival analysis, Wilcoxon test was used to assess differences in gene expression.

4.4. Real-Time qPCR

Total RNA from wild-type (n = 10) and knockout (n = 5) K562 clones was isolated by TRIzol Reagent (Invitrogen) and reverse-transcribed by High Capacity cDNA Reverse Transcription kit (Thermo Fisher Scientific). Quantitative PCR was run in triplicates on LightCycler 480 (Roche) using LightCycler 480 SYBR Green I Master (Roche) and specific primers for human SMARCA5 (forward primer 5′-AACTTACTATCCGTTGGCGATT-3′, reverse primer 5′-GGTTGCTTTGGAGCTTTCTG-3′) and GADPH (forward 5′-AGCCACATCGCTCAGACAC-3′, reverse primer 5′-GCCCAATACGACCAAATCC-3′) gene. Ct values served for fold-change calculation using 2-ΔΔCt equation. Student's t-test was used for statistical analysis.

4.5. Western Blot

Wild-type and S5KO K562 clones (1 × 10E7) were lysed in RIPA Buffer (Sigma-Aldrich) supplemented with protease and phosphatase inhibitors (Roche). Denatured cell lysates were run on 1 mm thick 10% SDS-PAGE gel (40 μg/lane) in Mini-Protean Electrophoresis system (Bio-Rad, Hercules, CA, USA) and semi-dry-blotted onto PVDF membrane (Bio-Rad) using Trans-Blot Turbo transfer system (Bio-Rad). PVDF membrane was blocked for 1 h in 5% nonfat milk in 1x TBS/0.1% Tween-20 and incubated with primary antibodies: Snf2h/ISWI (Bethyl Laboratories Inc., #A301-017A-1, Montgomery, TX, USA) and β-actin (Santa Cruz Biotechnology, #sc-1616-R, Dallas, Texas, USA) overnight at 4 °C. Horseradish peroxidase-conjugated secondary antibodies (anti-rabbit, anti-goat) visualized bands using Pierce ECL Western Blotting substrate (Thermo Fisher Scientific).

4.6. Cytogenetics

Standard cytogenetic methods published previously [10,11] were used for preparation of slides, with few modifications. Briefly, the K562 cells were synchronized with colcemid (10 μl/mL) at 37 °C and hypotonized in 0.075 M KCl for 20 min. The cells were then fixed in three changes of cold Carnoy's fixative (ethanol: glacial acetic acid, 3:1) and dropped onto a slide inclined at an angle of 45 degrees

from a height. The chromosomal preparations were air-dried overnight and stained using 5% Giemsa blue solution (Sigma-Aldrich) prepared in standard Sorenson buffer. Preparations were inspected under a light microscope BX43 (Olympus, Sony, Shinjuku, Japan) with microscope camera Infinity 2-2 (Lumenera, Ottawa, ON, Canada). Selected plates were photographed under a 100x immersion oil objective using software QuickPHOTO CAMERA 3.1 (Olympus).

4.7. Hematopoietic Reconstitution

For hematopoietic reconstitution experiments, 2.5×10^6 fetal liver cells isolated from E13.5 control (Smarca5$^{fl/+}$ Rosa26$^{eYFP/+}$ Vav1-iCRE) and Smarca5-deficient (Smarca5$^{fl/-}$ Rosa26$^{eYFP/+}$ Vav1-iCRE) with C57Bl/6J Ly5.2 background were transplanted into lethally irradiated (7.5 Gy) adult (8 weeks) C57Bl/6J Ly5.1 recipients. After 12 days, the recipients were euthanized, and their bone marrow and spleen were tested for the presence of donor-derived eYFP+ hematopoietic cells using flow cytometry. The antibody panel included CD45.1, CD45.2, c-Kit, Sca1, and lineage cocktail (CD3, B220, Mac-1, Gr-1, Ter119).

4.8. Analysis of S5KO MEF Cells

S5KO MEF cells ($n = 3$) were isolated from E14.5 embryos, in which the *Smarca5* gene contained the LoxP1 sites upstream and downstream of exon5 and also expressed Cre Recombinase-Estrogen receptor fusion protein that translocated into the nucleus upon addition of 4OHT into the cultures for 6 h. Deletion of *Smarca5*-exon5 represents a null allele [2]. Production of stable MEF cells was enabled by concurrent deletion of *Tp53* gene [32]. Gene targeting of the Smarca5$^{flox/flox}$ Cre-Esr1 cells upon 4OHT addition was confirmed by previously published detection methods [2]. Analysis of cell proliferation of MEFs was determined by IncuCyte (Sartorius, Göttingen, Germany) that enables analysis in 96 wells under real-time continuous visualization and monitoring.

Author Contributions: CRISPR design and mouse transgenics and writing: T.Z., clone preparation and functional analysis: H.P., MEF cells: T.T., cytogenetics: S.T., IncuCyte: P.T., AML patient data and statistics: P.K. and P.A.G., hematopoietic reconstitution: J.K., supervision and writing: T.S. All authors have read and agree to the published version of the manuscript.

Funding: Specific grants: 18-01687S, 19-03586S, NV19-08-00144, Student grants: GAUK 228316 & SVV 260374/2017, Institutional: UNCE/MED/016, Progres Q26, LM2015040, NPU II LQ1604 (MEYS), OP RDI CZ.1.05/2.1.00/19.0395, CZ.1.05/1.1.00/02.0109 (ERDF, MEYS).

Acknowledgments: We are thankful to Kristina Leblova for technical support.

Conflicts of Interest: The authors declare no conflict of interest

Abbreviations

SMARCA5	SWI/SNF-Related, Matrix-Associated, Actin-Dependent Regulator of Chromatin, Subfamily A, Member 5
SNF2H	Sucrose Nonfermenting Protein 2 Homolog
AML	Acute Myeloid Leukemia
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats

References

1. Kokavec, J.; Zikmund, T.; Savvulidi, F.; Kulvait, V.; Edelmann, W.; Skoultchi, A.I.; Stopka, T. The ISWI ATPase Smarca5 (Snf2h) Is Required for Proliferation and Differentiation of Hematopoietic Stem and Progenitor Cells. *Stem Cells* **2017**, *35*, 1614–1623. [CrossRef]
2. Stopka, T.; Skoultchi, A.I. The ISWI ATPase Snf2h is required for early mouse development. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 14097–14102. [CrossRef]
3. Zikmund, T.; Kokavec, J.; Turkova, T.; Savvulidi, F.; Paszekova, H.; Vodenkova, S.; Sedlacek, R.; Skoultchi, A.I.; Stopka, T. ISWI ATPase Smarca5 Regulates Differentiation of Thymocytes Undergoing beta-Selection. *J. Immunol.* **2019**, *202*, 3434–3446. [CrossRef]

4. Alvarez-Saavedra, M.; De Repentigny, Y.; Lagali, P.S.; Raghu Ram, E.V.; Yan, K.; Hashem, E.; Ivanochko, D.; Huh, M.S.; Yang, D.; Mears, A.J.; et al. Snf2h-mediated chromatin organization and histone H1 dynamics govern cerebellar morphogenesis and neural maturation. *Nat. Commun.* **2014**, *5*, 4181. [CrossRef]
5. He, S.; Limi, S.; McGreal, R.S.; Xie, Q.; Brennan, L.A.; Kantorow, W.L.; Kokavec, J.; Majumdar, R.; Hou, H., Jr.; Edelmann, W.; et al. Chromatin remodeling enzyme Snf2h regulates embryonic lens differentiation and denucleation. *Development* **2016**, *143*, 1937–1947. [CrossRef]
6. Barisic, D.; Stadler, M.B.; Iurlaro, M.; Schubeler, D. Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. *Nature* **2019**, *569*, 136–140. [CrossRef]
7. Dluhosova, M.; Curik, N.; Vargova, J.; Jonasova, A.; Zikmund, T.; Stopka, T. Epigenetic control of SPI1 gene by CTCF and ISWI ATPase SMARCA5. *PLoS ONE* **2014**, *9*, e87448. [CrossRef]
8. Morris, S.A.; Baek, S.; Sung, M.H.; John, S.; Wiench, M.; Johnson, T.A.; Schiltz, R.L.; Hager, G.L. Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions. *Nat. Struct. Mol. Biol.* **2014**, *21*, 73–81. [CrossRef]
9. Goodwin, L.R.; Picketts, D.J. The role of ISWI chromatin remodeling complexes in brain development and neurodevelopmental disorders. *Mol. Cell Neurosci.* **2018**, *87*, 55–64. [CrossRef]
10. Erdel, F.; Rippe, K. Chromatin remodelling in mammalian cells by ISWI-type complexes—Where, when and why? *FEBS J.* **2011**, *278*, 3608–3618. [CrossRef]
11. Kadoch, C.; Crabtree, G.R. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Sci. Adv.* **2015**, *1*, e1500447. [CrossRef]
12. Garraway, L.A.; Lander, E.S. Lessons from the cancer genome. *Cell* **2013**, *153*, 17–37. [CrossRef]
13. Dutta, A.; Sardiu, M.; Gogol, M.; Gilmore, J.; Zhang, D.; Florens, L.; Abmayr, S.M.; Washburn, M.P.; Workman, J.L. Composition and Function of Mutant Swi/Snf Complexes. *Cell Rep.* **2017**, *18*, 2124–2134. [CrossRef]
14. Gigek, C.O.; Lisboa, L.C.; Leal, M.F.; Silva, P.N.; Lima, E.M.; Khayat, A.S.; Assumpcao, P.P.; Burbano, R.R.; Smith Mde, A. SMARCA5 methylation and expression in gastric cancer. *Cancer Investig.* **2011**, *29*, 162–166. [CrossRef]
15. Reis, S.T.; Timoszczuk, L.S.; Pontes-Junior, J.; Viana, N.; Silva, I.A.; Dip, N.; Srougi, M.; Leite, K.R. The role of micro RNAs let7c, 100 and 218 expression and their target RAS, C-MYC, BUB1, RB, SMARCA5, LAMB3 and Ki-67 in prostate cancer. *Clinics* **2013**, *68*, 652–657. [CrossRef]
16. Sheu, J.J.; Choi, J.H.; Yildiz, I.; Tsai, F.J.; Shaul, Y.; Wang, T.L.; Shih Ie, M. The roles of human sucrose nonfermenting protein 2 homologue in the tumor-promoting functions of Rsf-1. *Cancer Res.* **2008**, *68*, 4050–4057. [CrossRef]
17. Jin, Q.; Mao, X.; Li, B.; Guan, S.; Yao, F.; Jin, F. Overexpression of SMARCA5 correlates with cell proliferation and migration in breast cancer. *Tumour. Biol.* **2015**, *36*, 1895–1902. [CrossRef]
18. Zhao, X.C.; An, P.; Wu, X.Y.; Zhang, L.M.; Long, B.; Tian, Y.; Chi, X.Y.; Tong, D.Y. Overexpression of hSNF2H in glioma promotes cell proliferation, invasion, and chemoresistance through its interaction with Rsf-1. *Tumour. Biol.* **2016**, *37*, 7203–7212. [CrossRef]
19. Stopka, T.; Zakova, D.; Fuchs, O.; Kubrova, O.; Blafkova, J.; Jelinek, J.; Necas, E.; Zivny, J. Chromatin remodeling gene SMARCA5 is dysregulated in primitive hematopoietic cells of acute leukemia. *Leukemia* **2000**, *14*, 1247–1252. [CrossRef]
20. Rosenbauer, F.; Wagner, K.; Kutok, J.L.; Iwasaki, H.; Le Beau, M.M.; Okuno, Y.; Akashi, K.; Fiering, S.; Tenen, D.G. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, PU.1. *Nat. Genet.* **2004**, *36*, 624–630. [CrossRef]
21. Behan, F.M.; Iorio, F.; Picco, G.; Goncalves, E.; Beaver, C.M.; Migliardi, G.; Santos, R.; Rao, Y.; Sassi, F.; Pinnelli, M.; et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **2019**, *568*, 511. [CrossRef]
22. Law, J.C.; Ritke, M.K.; Yalowich, J.C.; Leder, G.H.; Ferrell, R.E. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. *Leuk. Res.* **1993**, *17*, 1045–1050. [CrossRef]
23. Hakimi, M.A.; Bochar, D.A.; Schmiesing, J.A.; Dong, Y.; Barak, O.G.; Speicher, D.W.; Yokomori, K.; Shiekhattar, R. A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature* **2002**, *418*, 994–998. [CrossRef]

24. Welch, J.S.; Ley, T.J.; Link, D.C.; Miller, C.A.; Larson, D.E.; Koboldt, D.C.; Wartman, L.D.; Lamprecht, T.L.; Liu, F.; Xia, J.; et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **2012**, *150*, 264–278. [CrossRef]
25. Kim, J.-S.; He, X.; Orr, B.; Wutz, G.; Hill, V.; Peters, J.-M.; Compton, D.A.; Waldman, T. Intact Cohesion, Anaphase, and Chromosome Segregation in Human Cells Harboring Tumor-Derived Mutations in STAG2. *PLOS Genet.* **2016**, *12*, e1005865. [CrossRef]
26. Leman, A.R.; Noguchi, C.; Lee, C.Y.; Noguchi, E. Human Timeless and Tipin stabilize replication forks and facilitate sister-chromatid cohesion. *J. Cell Sci.* **2010**, *123*, 660–670. [CrossRef]
27. De Lange, J.; Faramarz, A.; Oostra, A.B.; de Menezes, R.X.; van der Meulen, I.H.; Rooimans, M.A.; Rockx, D.A.; Brakenhoff, R.H.; van Beusechem, V.W.; King, R.W.; et al. Defective sister chromatid cohesion is synthetically lethal with impaired APC/C function. *Nat. Commun.* **2015**, *6*, 8399. [CrossRef]
28. Kishtagari, A.N.; Jarman, C.; Tiwari, A.D.; Phillips, J.G.; Schuerger, C.; Jha, B.K.; Saunthararajah, Y. A First-in-Class Inhibitor of ISWI-Mediated (ATP-Dependent) Transcription Repression Releases Terminal-Differentiation in AML Cells While Sparing Normal Hematopoiesis. *Blood* **2018**, *132*, 216. [CrossRef]
29. Herold, T.; Jurinovic, V.; Batcha, A.M.N.; Bamopoulos, S.A.; Rothenberg-Thurley, M.; Ksienzyk, B.; Hartmann, L.; Greif, P.A.; Phillippou-Massier, J.; Krebs, S.; et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica* **2018**, *103*, 456–465. [CrossRef]
30. Stief, S.M.; Hanneforth, A.L.; Weser, S.; Mattes, R.; Carlet, M.; Liu, W.H.; Bartoschek, M.D.; Dominguez Moreno, H.; Oettle, M.; Kempf, J.; et al. Loss of KDM6A confers drug resistance in acute myeloid leukemia. *Leukemia* **2020**, *34*, 50–62. [CrossRef]
31. Batcha, A.M.N.; Bamopoulos, S.A.; Kerbs, P.; Kumar, A.; Jurinovic, V.; Rothenberg-Thurley, M.; Ksienzyk, B.; Philippou-Massier, J.; Krebs, S.; Blum, H.; et al. Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *Sci. Rep.* **2019**, *9*, 11796. [CrossRef] [PubMed]
32. Jacks, T.; Remington, L.; Williams, B.O.; Schmitt, E.M.; Halachmi, S.; Bronson, R.T.; Weinberg, R.A. Tumor spectrum analysis in p53-mutant mice. *Curr. Biol.* **1994**, *4*, 1–7. [CrossRef]

haematologica
Journal of The Ferrata Storti Foundation

Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring NRIP1-MIR99AHG rearrangements

by Paul Kerbs, Sebastian Vosberg, Stefan Krebs, Alexander Graf, Helmut Blum, Anja Swoboda, Aarif M.N. Batcha, Ulrich Mansmann, Dirk Metzler, Caroline A. Heckman, Tobias Herold, and Philipp A. Greif

Haematologica 2021 [Epub ahead of print]

Citation: Paul Kerbs, Sebastian Vosberg, Stefan Krebs, Alexander Graf, Helmut Blum, Anja Swoboda, Aarif M.N. Batcha, Ulrich Mansmann, Dirk Metzler, Caroline A. Heckman, Tobias Herold, and Philipp A. Greif. Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring NRIP1-MIR99AHG rearrangements.
Haematologica. 2021; 106:xxx
doi:10.3324/haematol.2021.278436

Running head: RNA-SEQ FUSION GENES IN AML

Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring *NRIP1-MIR99AHG* rearrangements*

Paul Kerbs[1,2,3], Sebastian Vosberg[1,2,3], Stefan Krebs[4], Alexander Graf[4], Helmut Blum[4], Anja Swoboda[1], Aarif M. N. Batcha[5], Ulrich Mansmann[5], Dirk Metzler[6], Caroline A. Heckman[7], Tobias Herold[1,2,3] & Philipp A. Greif[1,2,3]

[1]Department of Medicine III, University Hospital, LMU Munich, Munich, Germany
[2]German Cancer Consortium (DKTK), partner site Munich; and
[3]German Cancer Research Center (DKFZ), Heidelberg, Germany
[4]Laboratory for Functional Genome Analysis (LAFUGA), Gene Center, LMU Munich, Munich, Germany
[5]Department of Medical Data Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany
[6]Division of Evolutionary Biology, Faculty of Biology, LMU Munich, Planegg-Martinsried, Germany
[7]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Correspondence: pgreif@med.lmu.de

Acknowledgements

This study was supported by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Centre (SFB) 1243 "Cancer Evolution" (Projects A08 and A16). PAG acknowledges support by the Munich Clinician Scientist Program (MCSP). SV was supported by the Deutsche José Carreras Leukämie-Stiftung. We thank all participants and recruiting centers of the AMLCG and Beat AML trial. We thank Bianka Ksienzyk and William Keay for technical support.

Author contributions

PK, SV, SK and AS performed research and analyzed data. AMNB and AG provided computational support. CH, PAG and TH provided RNA-Seq data and clinical annotations. SV, HB, UM, DM and PAG supervised research. PK, SV and PAG wrote the manuscript. All authors approved the final manuscript.

Additional Information

The authors declare no conflicts of interest.

Word count: 3982, Reference count: 40

Figure count: 5, Table count: 2

* presented at the 61st ASH Annual Meeting[40]

ABSTRACT

Identification of fusion genes in clinical routine is mostly based on cytogenetics and targeted molecular genetics, such as metaphase karyotyping, FISH and RT-PCR. However, sequencing technologies are becoming more important in clinical routine as processing-time and costs per sample decrease. To evaluate the performance of fusion gene detection by RNA sequencing (RNA-seq) compared to standard diagnostic techniques, we analyzed 806 RNA-seq samples from acute myeloid leukemia (AML) patients using two state-of-the-art software tools, namely Arriba and FusionCatcher. RNA-seq detected 90% of fusion events that were reported by routine with high evidence, while samples in which RNA-seq failed to detect fusion genes had overall lower and inhomogeneous sequence coverage. Based on properties of known and unknown fusion events, we developed a workflow with integrated filtering strategies for the identification of robust fusion gene candidates by RNA-seq. Thereby, we detected known recurrent fusion events in 26 cases that were not reported by routine and found discrepancies in evidence for known fusion events between routine and RNA-seq in three cases. Moreover, we identified 157 fusion genes as novel robust candidates and comparison to entries from ChimerDB or Mitelman Database showed novel recurrence of fusion genes in 14 cases. Finally, we detected the novel recurrent fusion gene *NRIP1-MIR99AHG* resulting from inv(21)(q11.2;q21.1) in nine patients (1.1%) and *LTN1-MX1* resulting from inv(21)(q21.3;q22.3) in two patients (0.25%). We demonstrated that *NRIP1-MIR99AHG* results in overexpression of the 3' region of *MIR99AHG* and the disruption of the tricistronic miRNA cluster *miR-99a/let-7c/miR-125b-2*. Interestingly, upregulation of *MIR99AHG* and deregulation of the miRNA cluster, residing in the *MIR99AHG* locus, are known mechanism of leukemogenesis in acute megakaryoblastic leukemia. Our findings demonstrate that RNA-seq has a strong potential to improve the systematic detection of fusion genes in clinical applications and provides a valuable tool for fusion discovery.

INTRODUCTION

Fusion genes result from chromosomal aberrations, such as translocations, duplications, inversions or small interstitial deletions. On the transcript level, fusion genes may not only reflect underlying genomic rearrangements but may also arise due to aberrant transcription or trans-splicing events. Many fusion genes have been described as drivers across multiple human cancer entities[1]. Particularly, hematopoietic malignancies are well characterized regarding the abundance of fusion genes, including chronic myeloid leukemia, harboring the *BCR-ABL1* fusion in more than 95% of cases, and acute promyelocytic leukemia which is characterized by *PML-RARA* (90%). In acute myeloid leukemia (AML), fusion genes are found in about 30% of the patients[2] and are often regarded as major markers, defining clinically relevant subtypes[3–6]. Their identification is crucial for risk assessment and treatment strategy. During initial diagnosis of AML, fusion genes are detected using conventional metaphase karyotyping (hereafter referred to as Karyotyping) and/or targeted molecular assays (hereafter referred to as molecular diagnostics, MDx) such as fluorescence in situ hybridization (FISH) or reverse-transcriptase polymerase chain reaction (hereafter referred to as PCR). On a chromosomal level, Karyotyping detects abnormalities by light microscopy of metaphase spreads, whereas FISH labels chromosomal alterations using specifically designed probes that bind to particular genomic regions of interest. On a molecular level, PCR may confirm the presence of a specific genomic or transcriptomic sequence by targeted amplification. However, these methods are laborious and time consuming, depend on the experience of the analyst and might be subject to erroneous assessments. Furthermore, resolution of Karyotyping is limited to the microscopic level of chromosomal arms/bands and PCR/FISH can only be used to analyze predefined targets. Small inversions, duplications or short interstitial deletions as well as cryptic fusions are hardly detectable by these procedures. Although FISH and PCR are suitable for the targeted detection of submicroscopic lesions, they are not routinely applied to the systematic identification of previously uncharacterized aberrations and rather serve as complementary validation methods.

Over the last decade, next-generation sequencing (NGS) techniques have evolved tremendously and are increasingly used in clinical diagnostics[7–9]. NGS methods enable scalable genomic analyses, ranging from single genes and gene sets of interest up to genome-wide analyses, covering the entire genome at single base pair resolution. Further, RNA sequencing (RNA-seq) allows for transcriptome-wide studies, covering all transcribed genes present in a cell. Recently, a study proposed a single bioinformatic pipeline for AML diagnostics which integrates detection of fusion genes, small variants, tandem duplications and gene expression from RNA-seq data[10]. Thus, DNA and RNA sequencing allow for the examination of a wide range of genetic lesions, including the discovery of novel aberrations. Sequencing technologies are improving quickly and innovation in this field continuously reduces time and costs for genomic analyses, which allows for the processing of even more samples in parallel

with even higher precision. Simultaneously, developments in computational biology can exploit these advancements for accurate detection of genetic aberrations.

To date, more than 20 algorithms for fusion gene detection by RNA-seq have been published[9,11,12] but identification of fusions using RNA-seq remains challenging and a high rate of false positives is common. Therefore, careful evaluation of fusion calls and appropriate filtering strategies are needed to enable reliable application of this technology in diagnostics. In AML, no comprehensive comparison of fusion gene detection by RNA-seq and clinical routine has been reported so far. In this study, we set out to assess the potential of RNA-seq for the detection of clinically relevant fusion genes in comparison to standard diagnostic methods. Additionally, we developed several filters for robust fusion gene identification and the discovery of novel rearrangements in AML patients.

METHODS

Patient samples

A total of 806 AML patient samples underwent whole-transcriptome sequencing, collected within four different cohorts: (I) the German AML cooperative group (AMLCG 2008 and AMLCG 1999, n=257)[13,14]; (II) the German Cancer Consortium (DKTK, n=69)[15,16]; (III) the Beat AML program (n=423)[17]; and (IV) the Institute for Molecular Medicine Finland (FIMM, n=57)[18]. RNA-seq was performed as described previously[13–18]. Patient characteristics are summarized in Table 1 and Table S1. Sequencing metrics are summarized in Table 2. In addition, RNA-seq data of healthy samples were obtained from Gene Expression Omnibus Database (Table S2; n=36) and the FIMM cohort (n=3). All study protocols were approved by the institutional review boards of the participating centers. All patients provided written informed consent for scientific use of surplus samples in accordance with the Declaration of Helsinki.

Definitions and metrics for the evaluation of the performance in fusion gene detection

Comprehensive definitions and metrics are provided in the Supplementary Methods. In brief, recurrent and reliably detected fusion genes that were reported by public databases were defined as *known fusions*. Furthermore, fusion genes that were found with high evidence by at least one method in routine diagnostics were defined as a benchmark (*true fusions*). High and low evidence were defined separately for Karyotyping, MDx and RNA-seq (Figure S1, Supplementary Methods).

Filtering strategies

Initially, built-in filters of the callers were applied and fusions were filtered by a custom-generated blacklist (Supplementary Methods). The Promiscuity Score (PS), developed in this study, excluded fusion events whose respective partner genes were frequently called in other distinct fusion events, since these are likely artifacts. Furthermore, low read coverage of a fusion event relative to the read coverage of its partner genes indicates an artifact. Our custom Fusion Transcript Score (FTS) measures, in TPM, the expression of a fusion relative to the expression of its partner genes. Fusion events with a low FTS were excluded. The Robustness Score (RS) of a fusion gene is defined as the ratio between the number of samples in which this fusion gene passed all applied filters and the total number of samples in which this fusion gene was called. Only fusion genes passing all filters in at least half of the reported samples were considered. A comprehensive description of the filtering is enclosed in the Supplementary Methods.

RESULTS

Close correlation of fusion detection by routine diagnostics and RNA-seq

In 806 patient samples, we identified 138 true fusions which provided the benchmark for the comparison of performance in fusion gene detection between routine diagnostics (Karyotyping, MDx) and RNA-seq (Figure 1, Table S3). Of 138 true fusions, Karyotyping identified 121 (87.7%) and MDx identified 107 (77.5%) with high evidence. In addition, Karyotyping identified 11 (8%) and MDx identified 5 (3.6%) true fusions with low evidence. Fusion gene detection by RNA-seq resulted in 124 (89.9%) positive findings (high evidence: 115, low evidence: 9), thereby missing 14 true fusions (AMLCG: n=10; Beat AML: n=4).

Notably, samples from the AMLCG cohort showed less overall coverage and mappability of sequencing reads as compared to other samples (Table 2). In particular, *CBFB* and *KMT2A* showed poor coverage (Figure S2), both involved in 8/10 undetected true fusions by RNA-seq in those samples. Further fusions missed by RNA-seq were *DEK-NUP214* and *GTF2I-RARA*. Overall, in samples from the AMLCG cohort, substantially fewer fusion events were detected by FusionCatcher while Arriba detected twice as many compared to samples from other cohorts (Table 2).

In the Beat AML cohort, we observed discrepancies in reported fusions between RNA-seq and clinical routine in 3/4 cases of true fusions missed by RNA-seq: (I) Karyotyping reported t(6;11)(q27;q23) resulting in *KMT2A-AFDN*, while RNA-seq detected *KMT2A-MLLT10* resulting from t(10;11)(p12;q23). (II) Karyotyping reported del(2)(p21p23) resulting in *EML4-ALK*, while RNA-seq detected *KMT2A-MLLT3* resulting from t(9;11)(p21;q23). (III) Karyotyping reported der(17)t(15;17)(q24;q21) and inv(17)(q21q21) resulting in *PML-RARA* and *STAT5B-RARA*, respectively, while RNA-seq detected *PML-RARA* but not *STAT5B-RARA*. In the fourth case, a *PML-RARA* fusion was reported by FISH while Karyotyping reported a normal karyotype in this sample.

RNA-seq identifies known fusions missed by routine and yields additional candidates

Before filtering, a total of 25,817 and 56,594 fusion events were detected in 806 samples by Arriba and FusionCatcher, respectively (mean 32 and 70 per sample; Table 2). We applied filtering strategies as shown in Figure 2A. PS filter cutoffs for individual cohorts were set to: AMLCG=8.25, DKTK=3.5, Beat AML=16.5 and FIMM=3.5 (Figure S3A, Supplementary Methods). In addition to our previously described cutoffs for $FTS_{5'}$ and $FTS_{3'}$ (Supplementary Methods), we set a minimum FTS for unknown fusion events based on the median FTS of all detected unknown fusions (FTS ≥ 0.1, Figure S3B). Finally, we defined highly reliable fusion gene candidates based on the overlap of filtered fusion calls from Arriba and FusionCatcher. The built-in filter of Arriba, on average, excluded more putative false fusion events (74.8%) as compared to the built-in filter of FusionCatcher (62.3%). By applying our additional filtering strategies, we further reduced the amount of putative false fusion events substantially, resulting in an average of around 94% excluded fusion events from Arriba calls and

around 96% from FusionCatcher calls (Figure 2B). Besides detected true fusions (n=115), we found 187 fusion events as robust candidates. A total of 30/187 events have been described before, while 157 were putative novel fusion events (Table S4). Clinical routine showed only low evidence in 4/30 known events (Figure 1B, Table S5), while 26 candidates were not reported by routine diagnostics in our cohorts. In two of the four events described by clinical routine, a rearrangement of *KMT2A* was reported using FISH without any evidence from analyses by Karyotyping and in the other two events, Karyotyping reported rearranged chromosomes matching the chromosomal location of the fusion partner genes but different chromosomal bands. The 30 known fusion events having no or only low evidence by routine diagnostics include recurrent fusions *NUP98-NSD1* (n=8), *KMT2A-MLLT10* (n=4), *DEK-NUP214* (n=3), *KAT6A-CREBBP* (n=2), *KMT2A-MLLT3* (n=2) and *RUNX1-CBFA2T3* (n=2). Based on the newly identified fusion genes, patients would be assigned to a different ELN risk group in 6/30 cases (Table S5). Chromosomal locations of detected true, known and putative novel fusion events are presented as Circos plots[19] in Figure 3. Based on sample availability, we validated known fusion genes by PCR analysis that were exclusively found by RNA-seq in one sample (Figure S4; AM-0292-DX: *DEK-NUP214*). Moreover, 14 out of the 157 putative novel fusion events had an entry in ChimerDB or Mitelman Database but were not classified as known based on the criteria in the present study.

***NRIP1-MIR99AHG* is a novel recurrent fusion gene resulting from inv(21)(q11.2;q21.1)**

Beyond the detection of known rearrangements, we sought to identify novel recurrent fusion genes. Among our 157 putative novel fusion genes, we found *NRIP1-MIR99AHG* (Figure 4A) resulting from inv(21)(q11.2;q21.1) in six and *LTN1-MX1* (Figure S5) resulting from inv(21)(q21.3;q22.3) in two patient samples. Notably, *LTN1-MX1* was only found in co-occurrence with *NRIP1-MIR99AHG*. Further recurrence of *NRIP1-MIR99AHG* was reported by FusionCatcher alone in two patient samples (AM-0013-DX, FI-1216-RE).

Based on cDNA availability, we validated the junction of the *NRIP1-MIR99AHG* fusion transcript by PCR in sample AM-0028-DX. Three cytogenetically normal samples (AM-0044-DX, AM-0054-DX, AM-0069-DX) were used as negative controls (Figure 4B). Sanger sequencing of the PCR product confirmed a junction spanning sequence which matched the prediction of the RNA-seq fusion callers (Figure 4C). Nanopore sequencing of available gDNA from *NRIP1-MIR99AHG* positive samples AM-0028-DX (Figure 4D) and AM-0013-DX (Figure S6) identified the breakpoints (Table S6) and confirmed an inversion on the genomic level. Aiming to determine the complete fusion transcript, we generated a customized reference sequence of the inversion based on the identified breakpoints. Reads from Nanopore cDNA sequencing (median length: 883 bp) of the two *NRIP1-MIR99AHG* positive samples were mapped to this reference. Only unique mappings were considered to obtain reads spanning the junction of the fusion. We observed high coverage of the custom reference by

junction-spanning reads in the two fusion positive patients (Figure S7), while there was no coverage in negative controls. *NRIP1* includes a consensus coding sequence with an open reading frame (ORF) starting in exon 4, whereas *MIR99AHG* is non-coding. The identified breakpoint in the *NRIP1* locus in AM-0028-DX was located between exon 3 and 4, while the breakpoint in AM-0013-DX was located between exon 1 and 2, consistent with reports from RNA-seq fusion callers. In both cases, no annotated ORF was included in the putative fusion transcripts. A validation in samples from the Beat AML cohort was not possible due to lack of access to the patients' material. Literature research yielded the report[20] of a chronic myelomonocytic leukemia (CMML) patient with trisomy 21. The authors identified an inversion of chromosome 21 with breakpoints in the *NRIP1* locus and in a region upstream of *MIR125B2* (overlapping with an intronic region of *MIR99AHG*). We analyzed RNA-seq data from this patient (FI-0564-RE) with our fusion detection workflow and found high evidence for a *NRIP1-MIR99AHG* fusion. In total, *NRIP1-MIR99AHG* was found in nine (1.1%) out of 806 AML patients (AMLCG, n=2; Beat AML, n=5; FIMM, n=1) and one CMML patient.

Increased expression of the 3' partner gene in *NRIP1-MIR99AHG* and other fusions

In addition to the detection of fusion transcripts, we examined the expression rate of the single partner genes of a fusion and compared it between samples with and without this specific fusion. Sequence coverage of a gene as obtained from mapping but not read coverage of the fusion junction was considered as expression of this gene. Samples harboring a fusion, whose 3' partner gene is usually not expressed or at low levels only, showed an increased expression of the 3' partner gene up to the levels of the 5' partner gene, which is expressed at reasonable levels regardless of the fusion (Figure 5A-B). We did not observe an increase in the expression of the 3' partner in fusion events with similar expression rates between the 5' and 3' partner genes (Figure 5C-D). Accordingly, *MIR99AHG*, which is usually not expressed or at low levels only, showed increased expression levels in *NRIP1-MIR99AHG* positive samples (Figure 5E). On the other hand, *MX1*, which is inherently fairly expressed, only showed a slight elevation of expression levels in *LTN1-MX1* positive samples (Figure 5F).

Clinical and genetic characteristics of patients with *NRIP1-MIR99AHG* fusion

All patients found to harbor *NRIP1-MIR99AHG* had a poor survival with a median of 296 days (range: 36-1650 days). Interestingly, most of the patients were male (6/9) and had a median age of 59 (Table S7). Karyotyping showed a complex karyotype in four patients and five patients were refractory to intensive induction therapy. Furthermore, three patients showed a gain, and one patient showed a loss of chromosome 21. Unfortunately, we have no information whether these patients had a constitutional or somatic monosomy/trisomy 21. Cytomorphology was available for 3/9 patients without any evidence for megakaryoblastic leukemia (FAB M7). Mutational status was

available for 6/9 patients, but no apparent pattern was observed. However, recurrently mutated genes among those patients were *NRAS* (n=2) and *ASXL1* (n=2) (Table S7).

DISCUSSION

This study aimed to test the potential of fusion gene detection by RNA-seq in several cohorts of AML patient samples and to assess its diagnostic applicability by comparison to current standard techniques used in clinical routine. Based on our benchmark, the vast majority of true fusions reported by routine diagnostics was also detected by RNA-seq, underscoring the high sensitivity of this method. Notably, most of the samples in which a true fusion could not be detected by RNA-seq had a low read depth (median = 24 mio. mapped reads), while a minimum of 30 million mapped reads is recommended by the ENCODE consortium[21] for general expression analyses and even deeper sequencing for transcript discovery (e.g., fusion transcripts). Therefore, fusion gene detection was most likely hampered by the low read depth of these samples.

Limitations of fusion gene detection by RNA-seq are governed by library preparation steps, read depth, expression rates of the affected genes and the applied bioinformatic algorithms. On the other hand, Karyotyping is limited to a resolution of 5-10 Mb[22], which hampers the identification of small or cryptic rearrangements as well as rearrangements in specific locations (e.g. centromeric, telomeric)[23]. Furthermore, break-apart FISH probes identify genomic rearrangements in targeted regions through the visual separation of fluorescent labels. Although, this can indicate the rearrangement of a targeted locus, the detection of a specific aberration is still limited by the resolution of microscopic inspection, and the identification of the involved partner locus requires additional assays. In contrast to break-apart FISH, dual fusion probes target two partner loci and thereby can detect specific rearrangements but are restricted to the candidate loci of interest. In analogy, targeted PCR amplification of fusion transcripts requires prior knowledge of the affected genes and the corresponding break-point regions. In contrast, diagnostic application of RNA-seq has the potential to overcome these limitations through systematic detection of fusion genes on a transcriptome-wide level, as demonstrated in these three examples: (I) *NUP98-NSD1* is a biomarker for poor prognosis and *NUP98* fusions in general were found to define a clinically relevant distinct subgroup in AML[24–26] but reliable detection of the underlying cryptic translocation t(5;11)(q35.2;p15.4) by Karyotyping is not possible[27]. Of note, we identified *NUP98-NSD1* in eight samples using RNA-seq, as well as further known fusion genes in 22 samples that showed no or only low evidence for these fusions by either Karyotyping or MDx. (II) We observed discrepancies between results from routine and RNA-seq, i.e., one sample showing a translocation t(6;11)(q27;q23), according to Karyotyping. This translocation results in a *KMT2A-AFDN* fusion but RNA-seq reported a *KMT2A-MLLT10* fusion with high evidence, corresponding to translocation t(10;11)(p12;q23). Furthermore, *KMT2A* rearrangements were reported by break-apart FISH without any evidence for a rearrangement by Karyotyping in two cases. Fusion detection by RNA-seq identified a *KMT2A-MLLT10* fusion in these samples. Since various *KMT2A* fusions may reflect

different risk assessment based on the European LeukemiaNet classification[6], the correct description of the fusion may have therapeutic consequences. (III) In another sample, Karyotyping and FISH reported a t(15;17)(q24;q21) translocation, typically resulting in a *PML-RARA* fusion transcript (no information on PCR status was available), while RNA-seq identified a *PML-CASC3* fusion, with *CASC3* being located ~170 kb upstream of *RARA*. Unfortunately, no information on response to ATRA treatment of this patient was available.

In addition to standard diagnostic methods that are used in clinical routine, targeted RNA-seq panels are becoming increasingly popular for high-throughput detection of annotated fusion genes and were shown to be more sensitive than classical approaches[28].

Admittedly, RNA-seq based fusion callers report many false positive events due to technical and biological properties like sequencing errors, false mapping, homologous genomic regions, polymorphic genes, or exceptionally high gene expression[29]. Some genes are therefore prone to be reported in fusion gene artifacts, requiring reasonable filtering to maintain sensitivity while increasing specificity of the fusion detection analysis. Current callers integrate blacklists of fusion events into their built-in filters, which are compiled from public databases. However, technical differences between sequencing protocols and fusion calling algorithms may result in specific fusion artifacts that are not covered by those blacklists. Therefore, the generation of an additional custom blacklist further improves the specificity in RNA-seq based fusion analyses. Furthermore, we found genes which form fusions with many distinct partners indicating that these events are likely artifacts. The PS, developed in the present study, evaluates fusion events using this characteristic and filters events based on scores obtained from known fusions. However, the PS depends on the sequencing properties and the number of samples from which the score was derived. Thus, we defined cutoffs for the individual cohorts separately. Furthermore, the amount of fusion supporting reads correlates with the number of reads supporting the expression of the individual partner genes. The FTS, also developed in this study, measures the abundance of fusion transcripts relative to their respective partner gene transcripts. Most known fusions had an FTS around 0.3, but fusions present in subclones only, or fusions found in samples with lower tumor load will yield lower scores. As a tradeoff between specificity and sensitivity, we defined the median of all FTS detected in unknown fusion events as a cutoff. Besides, we observed unknown fusion events with high recurrence that passed all preceding filter steps in some samples, while these fusion events were filtered out in most other samples. This may indicate transcript artifacts of error prone genes. Therefore, the RS filter excludes fusion events that failed at least one preceding filter in most of the identified cases. The integration of our PS, FTS, custom blacklist and RS Filter into our detection strategy allowed for substantial reduction of fusion calls that are most likely false or irrelevant. Selection of fusion events consistently found between Arriba and FusionCatcher further increased the evidence of fusion

candidates. As an additional source of evidence for fusion events, we utilized individual gene expression values of the partner genes. The expression of a fusion transcript is mostly driven by the promoter of the 5' partner gene and the expression of the 3' partner should therefore adjust to the levels of the 5' partner. Although this simplified assumption neglects the influence of 3' enhancers and other regulatory elements, we observed substantially elevated expression of the 3' partner if it is usually not expressed or expressed at low levels only. Consistently, 3' partner genes with inherently similar expression to the 5' partner, showed no or only marginal adjustments in expression levels. However, genomic rearrangements do not necessarily result in a fusion transcript but may have other effects, e.g., the reallocation of the 3' enhancer of *GATA2* in inv(3)(q21.3q26.2)/t(3;3)(q21.3;q26.2) positive leukemia, causing overexpression of *MECOM* and *GATA2* haploinsufficiency[30,31]. Although, there is usually no fusion transcript in these patients, we found evidence for the transposition of *MECOM* by chimeric reads found in several affected samples (data not shown).

Among our fusion candidates, we identified the novel recurrent fusion gene *NRIP1-MIR99AHG*, which results from inversion inv(21)(q11.2;q21.1). Interestingly, both Nanopore sequencing and RNA-seq revealed different breakpoint positions in *NRIP1-MIR99AHG* positive samples. Moreover, none of the identified fusion transcripts included an annotated consensus coding sequence, and therefore translation to a protein product is rather unlikely. *NRIP1* was described as a transcriptional repressor[32], playing a role in hematological malignancies[33,34], and was found to be involved in other fusions[35]. A disruption of this gene by the *NRIP1-MIR99AHG* rearrangement might therefore contribute to leukemogenesis. On the other hand, overexpression of *MIR99AHG* and accompanying enhanced proliferation was previously demonstrated in acute megakaryoblastic leukemia cell lines (with *MIR99AHG* referred to as *MONC*)[36]. Furthermore, *MIR99AHG* is the host gene of *miR-99a/let-7c/miR-125b-2*, a miRNA cluster, also shown to influence homeostasis of hematopoietic stem and progenitor cells[37]. Interestingly, the identified fusion breakpoint in the *MIR99AHG* locus was located between *let-7c* and *miR-125b-2*, thereby disrupting the tricistronic gene cluster. This aberration as well as fusion-induced transcription of the 3’ region of *MIR99AHG* may constitute a mechanism of leukemogenesis. In the present study, *NRIP1-MIR99AHG* was found in eight AML patients, as well as in one CMML patient, all of which showed poor survival and were mostly refractory to intensive induction treatment. However, this might also be related to the complex karyotype in several patients. Of note, a recent whole transcriptome study of 572 AML and 630 MDS patients did not report any *NRIP1-MIR99AHG* fusion[38]. An extended analysis by the same authors[39] of overlapping cohorts, presented at the recent Annual Meeting of the American Society of Hematology, reported recurring *NRIP1-MIR99AHG* in AML and MDS but not in lymphoid malignancies (with *MIR99AHG* referred to as *LINC00478*). Further studies are needed to gain more insight into the pathogenic,

diagnostic and prognostic significance of the *NRIP1-MIR99AHG* fusion in AML and other hematological malignancies.

In conclusion, RNA-seq allows for accurate and more exhaustive identification of fusion transcripts as compared to classical cytogenetics or molecular diagnostics alone. We demonstrated that crucial AML-related fusions can be reliably identified by RNA-seq, but low sequence coverage was limiting sensitivity in a subset of samples. These findings underscore the need for stringent quality metrics in diagnostic RNA-seq applications. Nevertheless, we found several AML-related fusions that are hardly detectable by clinical routine. Furthermore, our workflow allowed for the identification of novel recurrent fusion transcripts such as *NRIP1-MIR99AHG* which results from the chromosomal rearrangement inv(21)(q11.2;q21.1). This study presents RNA-seq as a valuable complementary method to current standard techniques for the detection of fusion genes and we recommend the integration of RNA-seq applications into clinical routine for more comprehensive and precise diagnostics of hematological malignancies.

REFERENCES

1. Gao Q, Liang W-W, Foltz SM, et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. 2018;23(1):227-238.
2. Grimwade D, Hills RK, Moorman AV, et al. Refinement of cytogenetic classification in acute myeloid leukemia: Determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. Blood. 2010;116(3):354-365.
3. Döhner H, Estey EH, 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.
4. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 2016;127(20):2391-405.
5. Wang Y, Wu N, Liu D, Jin Y. Recurrent Fusion Genes in Leukemia: An Attractive Target for Diagnosis and Treatment. Curr Genomics. 2017;18(5):378-384.
6. 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.
7. Mack E, Langer D, Marquardt A, et al. Comprehensive Genetic Diagnostics of Acute Myeloid Leukemia By Next Generation Sequencing. Blood. 2016;128(22):1665.
8. Bacher U, Shumilov E, Flach J, et al. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. Blood Cancer J. 2019 811 2018;8(11):113.
9. Liu S, Tsai W-H, Ding Y, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Res. 2016;44(5):e47.
10. Arindrarto W, Borràs DM, de Groen RAL, et al. Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. Leukemia. 2020;35(1):47-61.
11. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep. 2016;6:21597.
12. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20(1):213.
13. 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.
14. 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. J Clin Oncol. 2006;24(16):2480-2489.

15. Hartmann L, Dutta S, Opatz S, et al. ZBTB7A mutations in acute myeloid leukaemia with t(8;21) translocation. Nat Commun. 2016;7(1):1-7.
16. Greif PA, Hartmann L, Vosberg S, et al. Evolution of cytogenetically normal acute myeloid leukemia during therapy and relapse: An exome sequencing study of 50 patients. Clin Cancer Res. 2018;24(7):1716-1726.
17. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018;562(7728):526-531.
18. Pemovska T, Kontro M, Yadav B, et al. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. Cancer Discov. 2013;3(12):1416-1429.
19. Krzywinski M, Schein J, Birol I, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-1645.
20. Majumder MM, Kontro M, Edgren H, et al. Genomic and transcriptomic data integration in chronic myelomonocytic leukemia reveals a novel fusion gene involving onco-miR-125b-2. Cancer Res. 2012;72(8 Supplement):3175.
21. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Res. 2018;46(D1):D794-D801.
22. Gelehrter TD, Collins FS, Ginsburg D. Principles of Medical Genetics. Williams & Wilkins. 153-194 p.
23. De Braekeleer E, Meyer C, Douet-Guilbert N, et al. Complex and cryptic chromosomal rearrangements involving the MLL gene in acute leukemia: A study of 7 patients and review of the literature. Blood Cells Mol Dis. 2010;44(4):268-274.
24. Kivioja JL, Lopez Martí JM, Kumar A, et al. Chimeric *NUP98–NSD1* transcripts from the cryptic t(5;11)(q35.2;p15.4) in adult de novo acute myeloid leukemia. Leuk Lymphoma. 2018;59(3):725-732.
25. Hollink IHIM, van den Heuvel-Eibrink MM, Arentsen-Peters STCJM, et al. NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. Blood. 2011;118(13):3645-3656.
26. Bisio V, Zampini M, Tregnago C, et al. NUP98-fusion transcripts characterize different biological entities within acute myeloid leukemia: a report from the AIEOP-AML group. Leukemia. 2017;31(4):974-977.

27. Kearney L. t(5;11)(q35;p15.5) NUP98/NSD1. *Atlas Genet Cytogenet Oncol Haematol.* 2002;6(3):209-211. http://atlasgeneticsoncology.org/Anomalies/t0511q35p15ID1209.html (2002, accessed April 28, 2020).
28. Heyer EE, Deveson IW, Wooi D, et al. Diagnosis of fusion genes using targeted RNA sequencing. Nat Commun. 2020;11(1):1810.
29. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13.
30. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. Cell. 2014;157(2):369-381.
31. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. Cancer Cell. 2014;25(4):415-427.
32. Castet A, Boulahtouf A, Versini G, et al. Multiple domains of the Receptor-Interacting Protein 140 contribute to transcription inhibition. Nucleic Acids Res. 2004;32(6):1957-1966.
33. Lapierre M, Castet-Nicolas A, Gitenay D, et al. Expression and role of RIP140/NRIP1 in chronic lymphocytic leukemia. J Hematol Oncol. 2015;8:20.
34. Herold T, Jurinovic V, Metzeler KH, et al. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. Leukemia. 2011;25(10):1639-1645.
35. Zhang R, Kim YM, Yang X, Li Y, Li S, Lee JY. A possible 5'-NRIP1/UHRF1-3' fusion gene detected by array CGH analysis in a Ph+ ALL patient. Cancer Genet. 2011;204(12):687-691.
36. Emmrich S, Streltsov A, Schmidt F, Thangapandi VR, Reinhardt D, Klusmann JH. LincRNAs MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. Mol Cancer. 2014;13(1):171.
37. Emmrich S, Rasche M, Schöning J, et al. miR-99a/100~125b tricistrons regulate hematopoietic stem and progenitor cell homeostasis by shifting the balance between TGFβ and Wnt signaling. Genes Dev. 2014;28(8):858-874.
38. Stengel A, Shahswar R, Haferlach T, et al. Whole transcriptome sequencing detects a large number of novel fusion transcripts in patients with AML and MDS. Blood Adv. 2020;4(21):5393-5401.
39. Haferlach C, Walter W, Meggendorfer M, et al. The Diverse Landscape of Fusion Transcripts in 25 Different Hematological Entities. Blood. 2020;136(Supplement 1):16-17.
40. Kerbs P, Nazeer Batcha AM, Vosberg S, Metzler D, Herold T, Greif PA. Gene Fusion Detection By RNA-Seq in Acute Myeloid Leukemia (AML). Blood. 2019;134(Supplement_1):4655.

TABLES

Table 1: Summary of patient characteristics.

Cohort	Median Age (Range)	Sex (%)	ELN risk group (%)
AMLCG (n=257)	58 (18-79)	Females = 131 (51.0) Males = 126 (49.0)	Favorable = 75 (29.2) Intermediate = 61 (23.7) Adverse = 107 (41.6) NA = 14 (5.5)
DKTK (n=69)	61 (21-85)	Females = 31 (44.9) Males = 38 (55.1)	Favorable = 33 (47.8) Intermediate-I = 25 (36.2) Intermediate-II = 7 (10.1) Adverse = 3 (4.4) NA = 1 (1.5)
Beat AML (n=423)	61 (2-87)	Females = 186 (44.0) Males = 237 (56.0)	Favorable = 112 (26.5) Intermediate = 141 (33.3) Adverse = 148 (35.0) Favorable or Intermediate = 13 (3.1) Intermediate or Adverse = 7 (1.6) NA = 2 (0.5)
FIMM (n=57)	58.5 (21-77)	Females = 29 (50.0) Males = 29 (50.0)	Favorable = 9 (15.8) Intermediate = 19 (33.3) Adverse = 18 (31.6) NA = 11 (19.3)

Table 2: Statistics for RNA-seq, mapping and fusion calling.

	AMLCG	DKTK	Beat AML	FIMM
RNA selection	poly(A)	poly(A)	poly(A)	Total RNA (rRNA depleted)
Avg. library size in mio. (range)	30.6 (19.1-97.8)	33.7 (23.4-43.3)	55.1 (24.7-126.8)	57.4 (23.9-119.9)
Avg. % uniquely mapped reads (range)	80 (44.2-94.1)	90.7 (82-93.7)	91.4 (80.9-94.3)	86.3 (70.4-93.3)
Avg. % reads mapped to exons (range)	72.4 (40.5-87.6)	81.5 (75.6-85.7)	76.8 (60.1-86.8)	51 (20.2-67.9)
Avg. insert size (range)	248.1 (97-455)	257.1 (217-289)	186.7 (131-246)	219.5 (141-289)
Avg. fusion events called by Arriba	48.3	23.2	24.1	27.8
Avg. fusion events called by FusionCatcher	12.9	113.1	97.8	71.4

FIGURE LEGENDS

Figure 1: Evidence for fusions by clinical routine diagnostics and RNA-seq

A) True fusions detected by Karyotyping, MDx and RNA-seq in the AMLCG, DKTK, Beat AML and FIMM cohorts. Dark green boxes indicate high evidence, light green boxes indicate low evidence. Grey boxes represent no evidence although the respective method was performed. White boxes indicate that the respective method was not performed, or information was missing. B) Known fusions detected with high evidence by RNA-seq which were missed or detected with low evidence only by Karyotyping/MDx. C) Venn diagram summarizing detected fusions according to the different methods.

Figure 2: Detection workflow and filtering of fusion events

A) Detection workflow and number of filtered fusion events by filtering strategies. B) Ratios of fusion events excluded by Arriba and FusionCatcher in each filter step and cohort.

Figure 3: Genomic origin of fusion events detected by RNA-Seq

Circos plots of A) known and B) unknown fusion gene candidates found in the AMLCG, DKTK, Beat AML and FIMM cohorts, illustrating chromosomal origin of the fusion events. Lines connect the positions of fusion partners. Thickness of lines indicates recurrence. Recurrent fusions are labeled with gene symbols of the partner genes. Blue lines indicate known fusion events, red lines indicate recurrent novel and grey lines show non-recurrent novel fusion events.

Figure 4: Detection and validation of novel *NRIP1-MIR99AHG* fusion gene

Evidence for *NRIP1-MIR99AHG* fusion gene in sample AM-0028-DX by various methods. A) Schematic representation of the fusion transcript as predicted by RNA-seq. B) Gel-electrophoresis of RT-PCR analysis of fusion breakpoint and *NRIP1* exon 4. Three cytogenetically normal AML patient samples were used as negative controls. C) Trace from Sanger sequencing of fusion breakpoint. D) Mapping of long reads from Nanopore sequencing of genomic DNA. Each line represents one read, which can be divided at the breakpoints of the fusion. Single parts of the read can be mapped to the positive strand (blue) at one locus with the other part mapped to the negative strand (red) at the other locus. The consensus inversed region is indicated by orange. Mapping structure of a highlighted read at the bottom shows that one part of the read was inversely mapped to the *NRIP1* locus, while the other part was mapped to the *MIR99AHG* locus.

Figure 5: Gene expression of genes involved in fusions

Gene expression of the 5' and 3' partner genes of the respective fusion. Red dots indicate samples positive for the respective fusion, grey dots represent samples negative for the respective fusion.

Figure 1

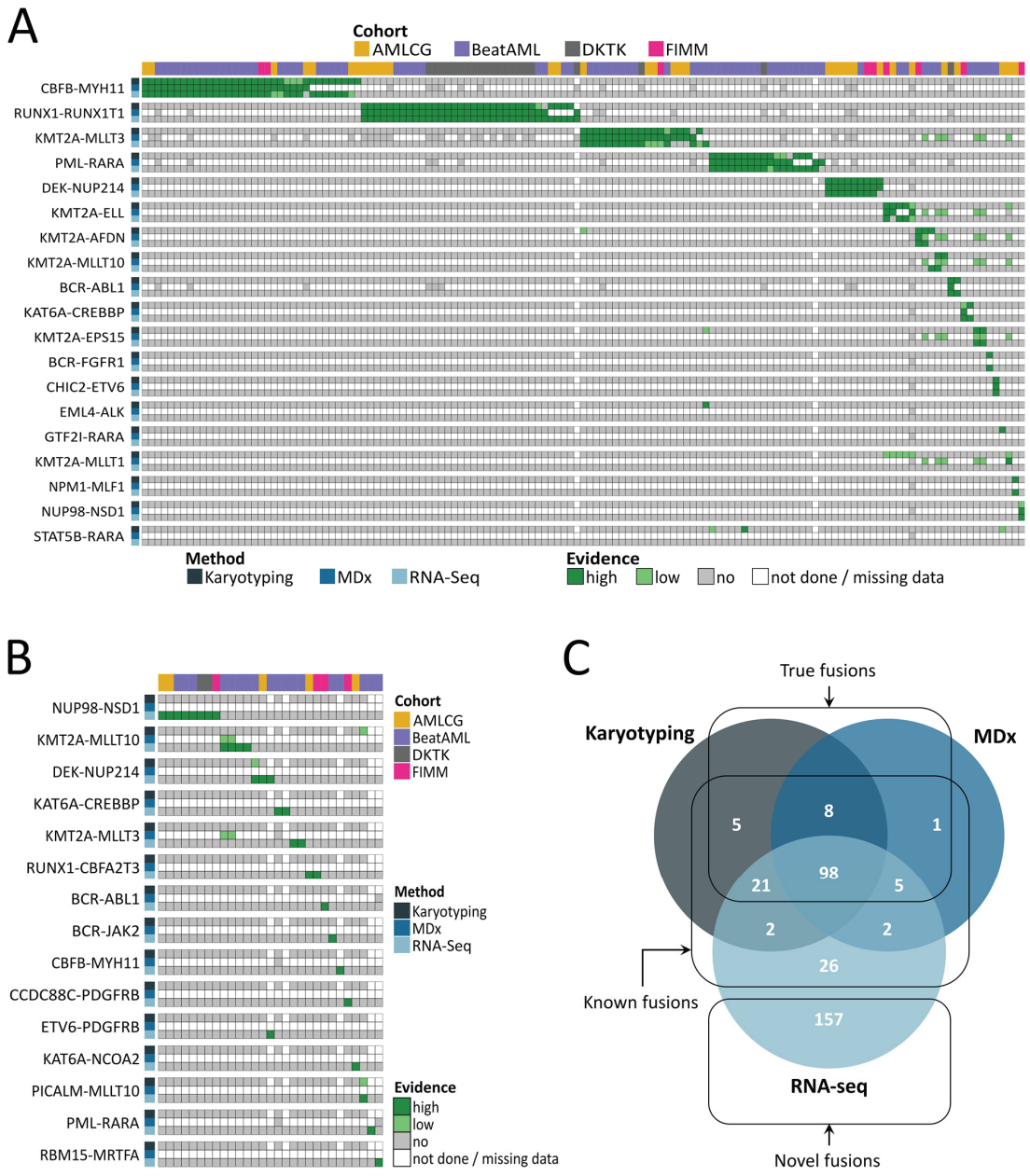

A
Cohort
AMLCG
BeatAML
DKTK
FIMM
CBFB-MYH11
RUNX1-RUNX1T1
KMT2A-MLLT3
PML-RARA
DEK-NUP214
KMT2A-ELL
KMT2A-AFDN
KMT2A-MLLT10
BCR-ABL1
KAT6A-CREBBP
KMT2A-EPS15
BCR-FGFR1
CHIC2-ETV6
EML4-ALK
GTF2I-RARA
KMT2A-MLLT1
NPM1-MLF1
NUP98-NSD1
STAT5B-RARA
Method
Karyotyping
MDx
RNA-Seq
Evidence
high
low
no
not done / missing data
B
NUP98-NSD1
KMT2A-MLLT10
DEK-NUP214
KAT6A-CREBBP
KMT2A-MLLT3
RUNX1-CBFA2T3
BCR-ABL1
BCR-JAK2
CBFB-MYH11
CCDC88C-PDGFRB
ETV6-PDGFRB
KAT6A-NCOA2
PICALM-MLLT10
PML-RARA
RBM15-MRTFA
Cohort
AMLCG
BeatAML
DKTK
FIMM
Method
Karyotyping
MDx
RNA-Seq
Evidence
high
low
no
not done / missing data
C
True fusions
Karyotyping
MDx
5
8
1
21
98
5
2
2
26
157
Known fusions
RNA-seq
Novel fusions

Figure 2

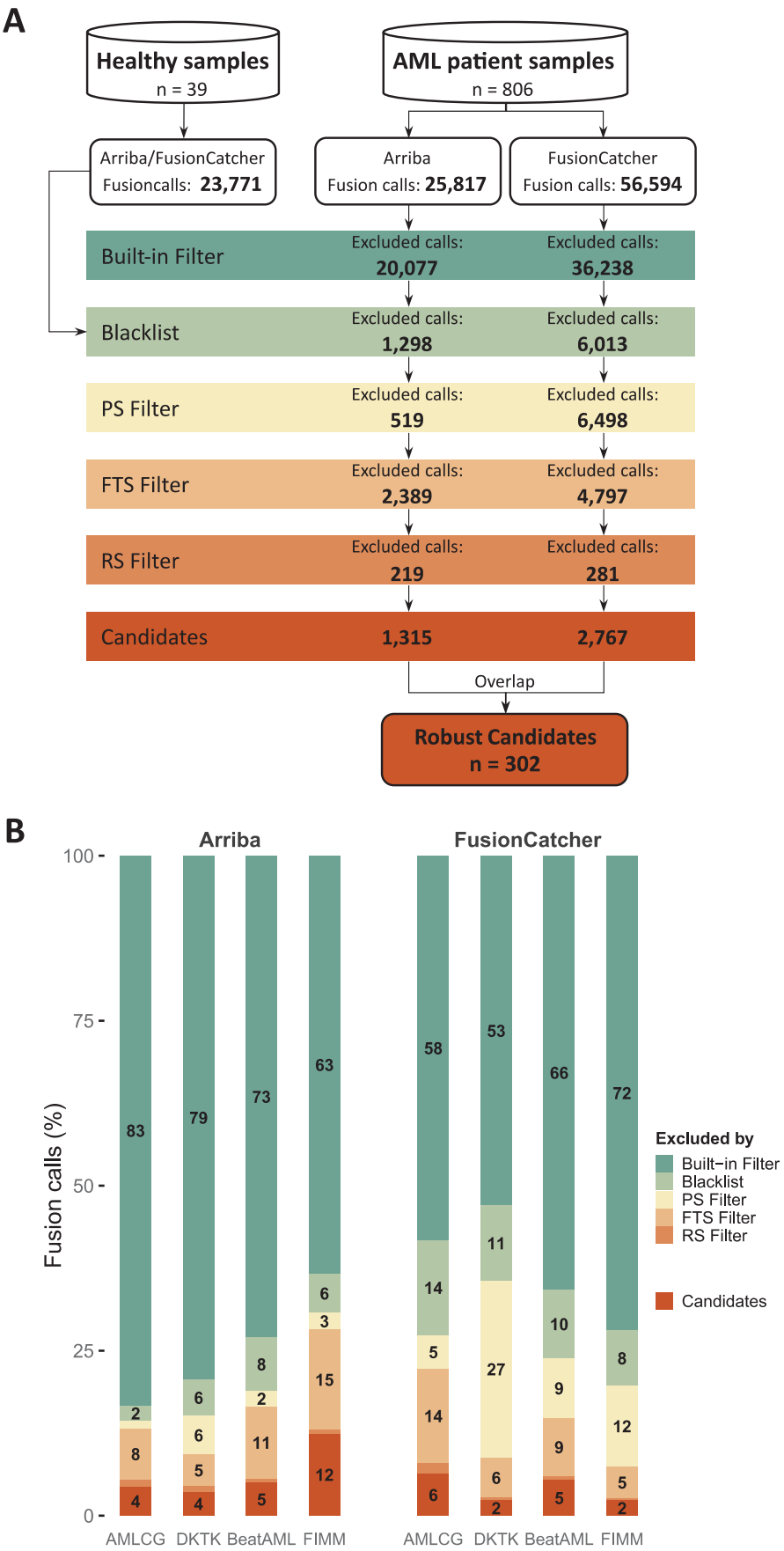

A
Healthy samples
n = 39
AML patient samples
n = 806
Arriba/FusionCatcher
Fusioncalls: 23,771
Arriba
Fusion calls: 25,817
FusionCatcher
Fusion calls: 56,594
Built-in Filter
Excluded calls: 20,077
Excluded calls: 36,238
Blacklist
Excluded calls: 1,298
Excluded calls: 6,013
PS Filter
Excluded calls: 519
Excluded calls: 6,498
FTS Filter
Excluded calls: 2,389
Excluded calls: 4,797
RS Filter
Excluded calls: 219
Excluded calls: 281
Candidates
1,315
2,767
Overlap
Robust Candidates
n = 302
B
Arriba
FusionCatcher
Fusion calls (%)
100
75
50
25
0
AMLCG
DKTK
BeatAML
FIMM
Excluded by
Built−in Filter
Blacklist
PS Filter
FTS Filter
RS Filter
Candidates

Figure 3

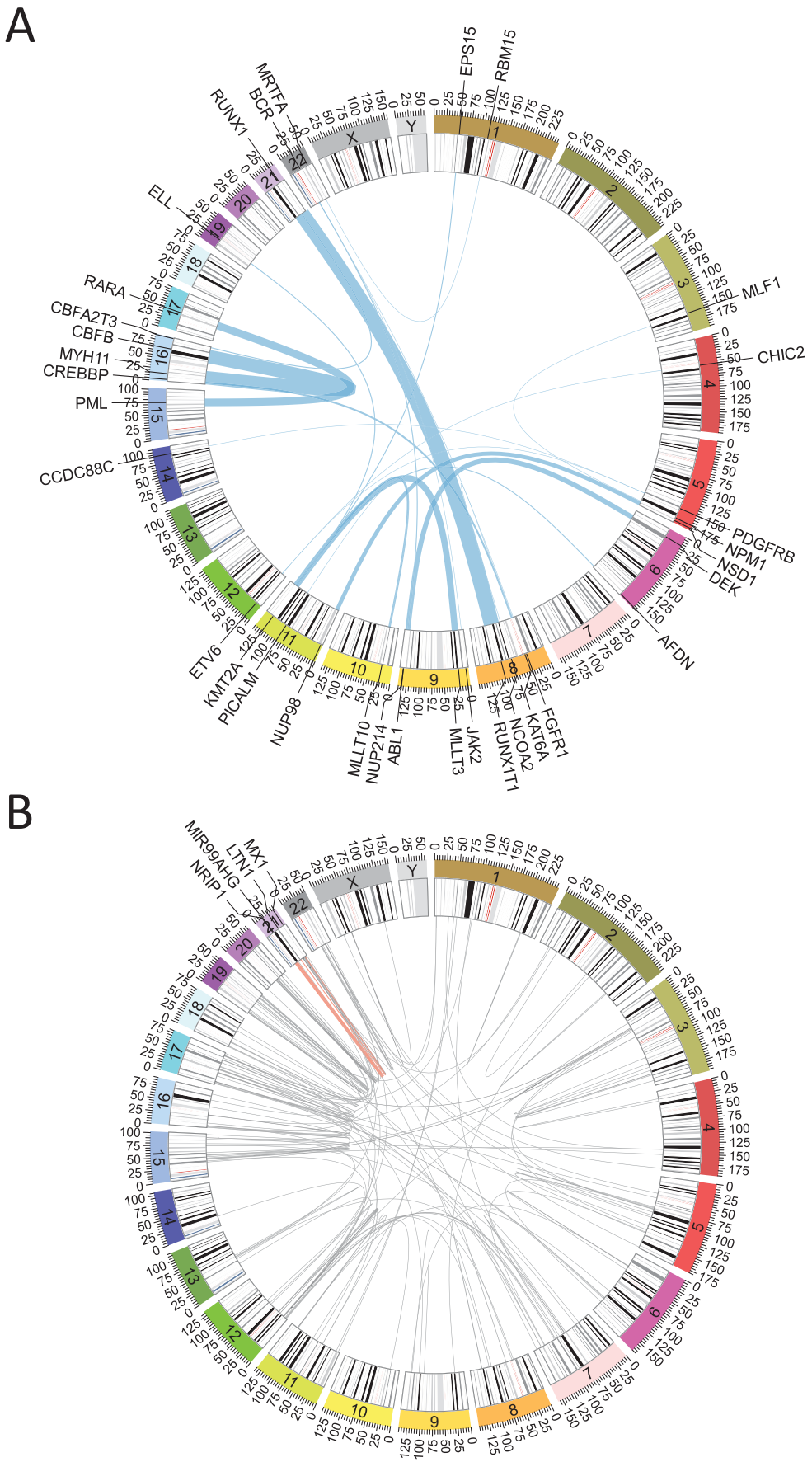

A
EPS15
RBM15
MLF1
CHIC2
PDGFRB
NPM1
NSD1
DEK
AFDN
FGFR1
KAT6A
NCOA2
RUNX1T1
JAK2
MLLT3
MLLT10
NUP214
ABL1
NUP98
KMT2A
PICALM
ETV6
CCDC88C
PML
CREBBP
MYH11
CBFB
CBFA2T3
RARA
ELL
RUNX1
MRTFA
BCR
B
MIR99AHG
LTN1
MX1
NRIP1

Figure 4

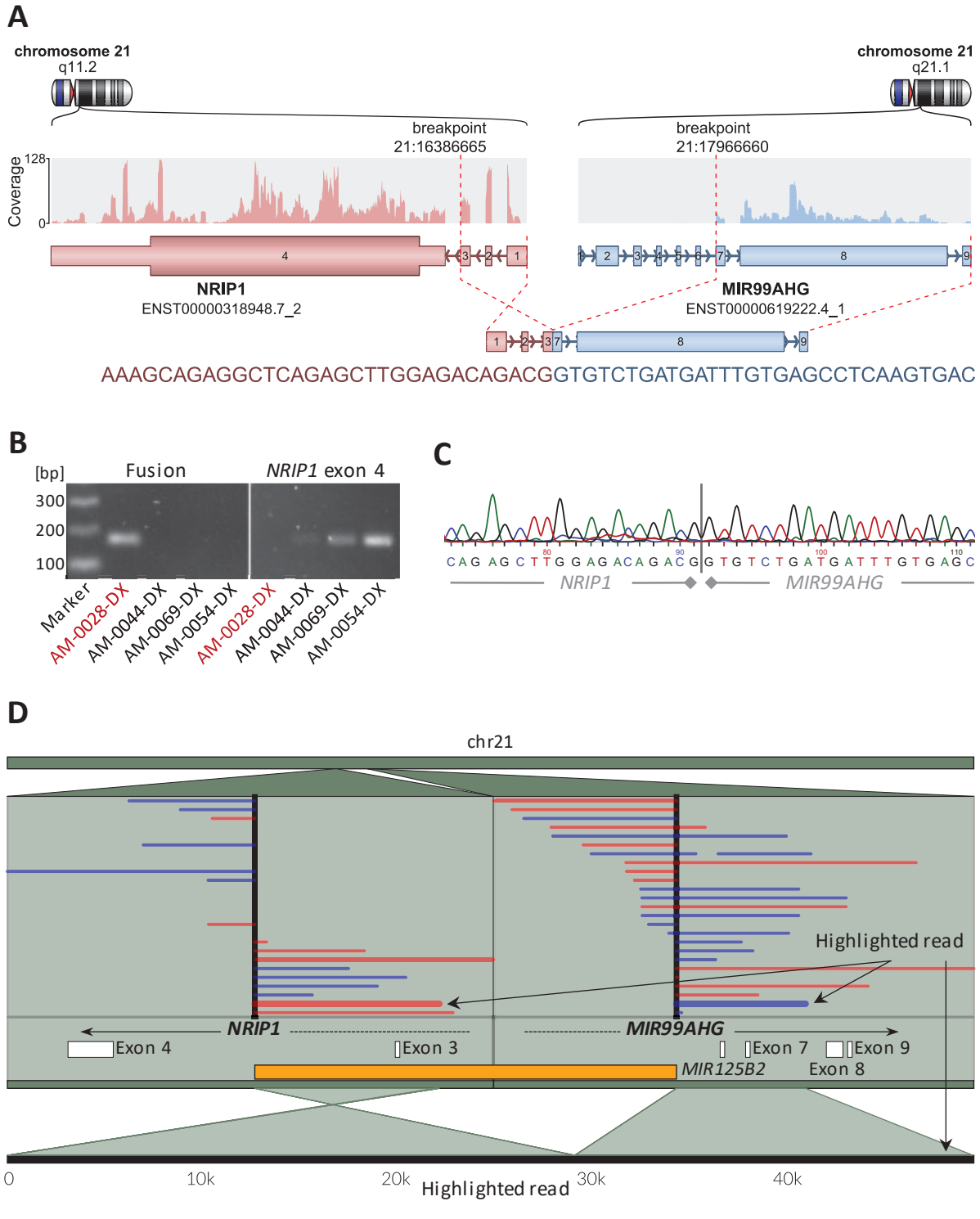
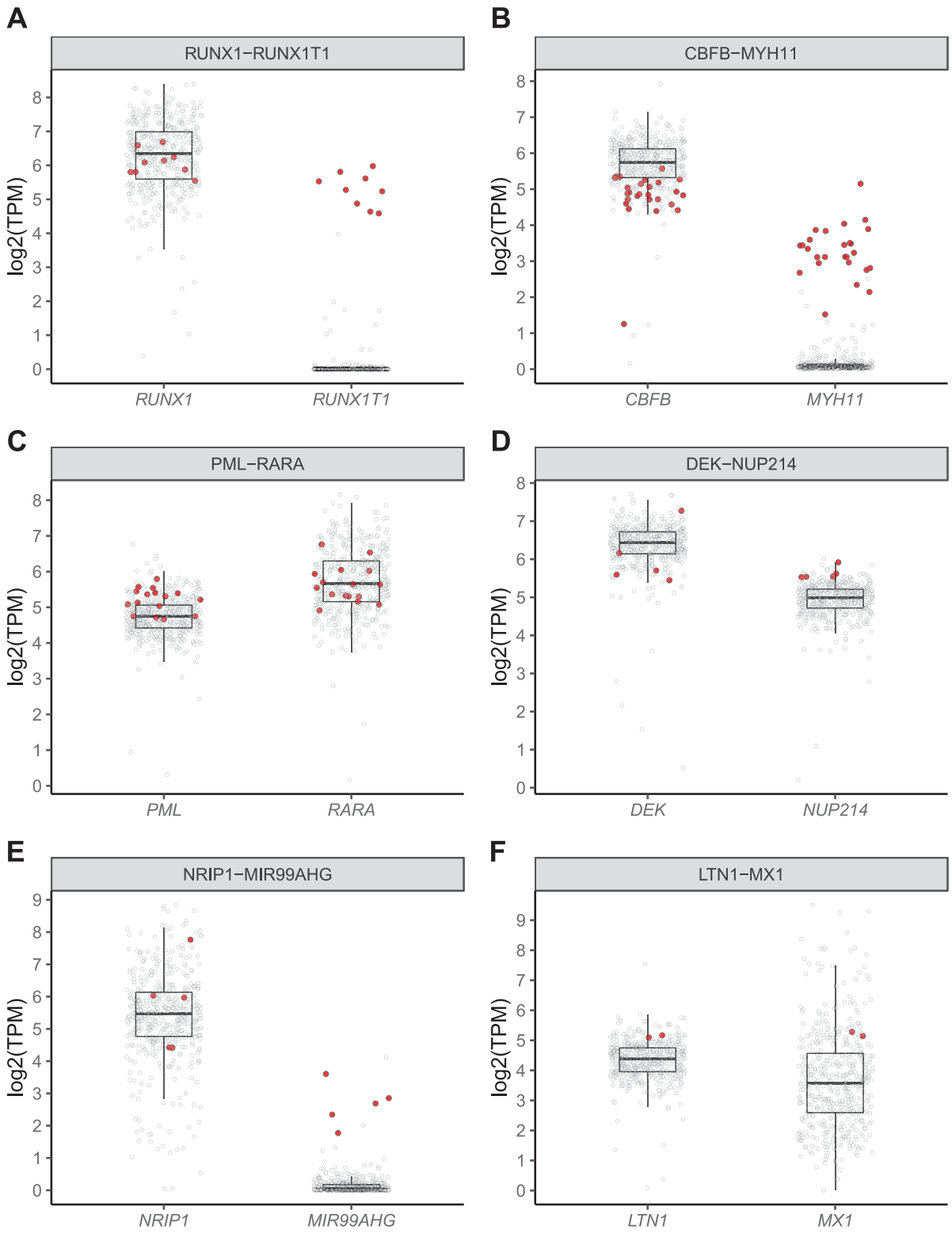

A
chromosome 21
q11.2
chromosome 21
q21.1
breakpoint
21:16386665
breakpoint
21:17966660
Coverage
128
0
NRIP1
ENST00000318948.7_2
MIR99AHG
ENST00000619222.4_1
AAAGCAGAGGCTCAGAGCTTGGAGACAGACGGTGTCTGATGATTTGTGAGCCTCAAGTGAC
B
[bp]
Fusion
NRIP1 exon 4
300
200
100
Marker
AM-0028-DX
AM-0044-DX
AM-0069-DX
AM-0054-DX
AM-0028-DX
AM-0044-DX
AM-0069-DX
AM-0054-DX
C
CAG AG C TTG G AG ACA G AC G G TG TC T G AT G ATTT G TG AG C
NRIP1
MIR99AHG
D
chr21
Highlighted read
NRIP1
Exon 4
Exon 3
MIR99AHG
Exon 7
Exon 9
MIR125B2
Exon 8
0
10k
20k
30k
40k
Highlighted read

Figure 5


A
RUNX1−RUNX1T1
log2(TPM)
RUNX1
RUNX1T1
B
CBFB−MYH11
log2(TPM)
CBFB
MYH11
C
PML−RARA
log2(TPM)
PML
RARA
D
DEK−NUP214
log2(TPM)
DEK
NUP214
E
NRIP1−MIR99AHG
log2(TPM)
NRIP1
MIR99AHG
F
LTN1−MX1
log2(TPM)
LTN1
MX1

SUPPLEMENTARY MATERIALS

Fusion gene detection by RNA sequencing complements diagnostics of acute myeloid leukemia and identifies recurring *NRIP1-MIR99AHG* rearrangements

Paul Kerbs, Sebastian Vosberg, Stefan Krebs, Alexander Graf, Helmut Blum,
Anja Swoboda, Aarif M. N. Batcha, Ulrich Mansmann, Dirk Metzler, Caroline A. Heckman,
Tobias Herold & Philipp A. Greif

Table of Contents

* supplementary tables are provided as a separate Excel file

Supplementary Methods

RNA-seq analysis and fusion calling

Fusion gene detection was performed using Arriba[1] and FusionCatcher[2]. FusionCatcher was applied to untrimmed and unmapped reads, as recommended by the authors. Ensembl release 98 was used as reference/annotation in FusionCatcher analyses (required resources were generated by using the 'fusioncatcher-build' module). Arriba was applied to trimmed and mapped sequence reads, as recommended by the authors. Trimming of adapter and low-quality sequences was done using Trimmomatic[3]. Reads were mapped to the human genome GRCh37 (GENCODE release 32) using STAR[4]. Gene expression analysis was done using FeatureCounts[5]. Read counts were normalized to transcripts per million (TPM). Insert size per sample was estimated by Picard toolkit[6]. Detailed parameters are available in Table S8.

Definition of known/true fusions and high/low evidence

Highly reliable fusion genes (recurrently reported, validated by PCR, part of ChimerSeq-Plus) from ChimerDB[7] were defined as known fusions. Corresponding karyotypes were obtained from the Mitelman Database[8]. Known fusions, identified from all samples in the present study, which were supported with high evidence by at least one method used in routine diagnostics (i.e., Karyotyping and/or MDx), were defined as benchmark (true fusions). High and low evidence for a fusion gene were defined separately for Karyotyping, MDx and RNA-seq, based on the following criteria: High evidence by Karyotyping was defined as chromosomes as well as chromosomal bands matching the localization of the two partner genes in the respective fusion; low evidence by Karyotyping was defined as a match of chromosomes only, while chromosomal bands did not match or information on bands was missing. High evidence by MDx was defined as confirmation of a specific fusion gene by FISH or PCR; low evidence by MDx was defined as the confirmation of a rearrangement by FISH of only one fusion partner (e.g., using a break-apart probe). High evidence by RNA-seq was defined as fusion genes found by both RNA-seq based algorithms; low evidence by RNA-seq was defined as fusion genes found by either Arriba or FusionCatcher alone (Figure S1).

Built-in filters of fusion callers and custom blacklist of fusion genes

All reported fusion events were filtered by the number of supporting reads (minimum 3). Based on FusionCatcher reports, fusion events with an annotation (Table S9) that implies irrelevant, non-somatic or false-positive events, as well as fusions whose partner genes showed sequence homology by common mapping reads were excluded. Based on Arriba reports, we excluded fusion events scored with a "low" confidence. Further, we defined a blacklist of fusion genes detected in 39 healthy samples (Table S10).

Promiscuity Score

Due to biological or technical reasons, certain genes are prone to be falsely detected as part of fusion events with many different partners. Therefore, we defined a custom Promiscuity Score (PS) which measures, for each fusion event detected, the average amount of varying fusion partners of the two partner genes involved in that fusion. First, P_{gene} was defined as the average number of varying fusion partners of a specific gene that were identified by Arriba and FusionCatcher within the cohorts. Second, PS_{fusion} was defined as the average of P_{gene} values of the two genes forming the 5' and 3' end of the specific fusion:

$$PS_{fusion} = mean(P_{5'}, P_{3'})$$

$$with\ P_x = mean\left(Ptr_{Arriba,x}, Ptr_{FusionCatcher,x}\right)\ for\ x\ in\ \{5', 3'\}$$

$$and\ Ptr_{M,x} = amount\ of\ different\ fusion\ partners$$

$$for\ M\ in\ \{Arriba, FusionCatcher\}$$

Since the PS is dependent on sequencing characteristics and the number of samples from which it was derived, cutoffs were set based on the highest PS detected for known fusions in each cohort individually.

Fusion Transcript Score

It is fair to assume that expression of a fusion gene is closely correlated to the expression of its partner genes. Therefore, we defined a custom Fusion Transcript Score (FTS) which measures, in TPM, the expression of a fusion relative to the expression of its partner genes:

$$FTS_{fusion} = mean(FTS_{5'}, FTS_{3'})$$

$$with\ FTS_x = \frac{TPM_{fusion}}{TPM_{fusion} + TPM_x}\ for\ x\ in\ \{5', 3'\}$$

Calculation of expression in TPM requires the length of the respective transcript. Due to limited length of the sequenced fragments and the fact that only reads covering the fusion breakpoint can be accounted for the expression of the fusion gene transcript, exact length and expression of the fusion transcript cannot be determined. Therefore, TPM values for a fusion transcript were approximated by using estimated median insert size from mapping. Fusion genes with $TPM_{5'} = 0$ or $TPM_{3'} = 0$ are regarded as artifacts since it is highly unlikely that the partner genes of the fusion show no read coverage. A minimum cutoff of 0.025 was set for $FTS_{5'}$ and $FTS_{3'}$, which corresponds to one out of two alleles being affected in a tumor population, making up more than 5% of a bulk sample, which is representing the normal levels of myeloid blasts in healthy hematopoiesis.

Robustness Score

Moreover, particular fusion genes eventually pass all filters in some samples but are filtered out in many other samples that were reported to harbor these fusion genes, indicating false positives. The Robustness Score (RS) of a fusion gene is defined as the ratio between the number of samples in which this fusion gene passed all applied filters and the total number of samples in which this fusion gene was called. Only fusion genes passing all filters in at least half of the reported samples (RS ≥ 0.5) were considered.

PCR and Sanger sequencing

Primers for PCR validation of the *NRIP1-MIR99AHG* fusion gene were designed using Primer-Blast[9] and a customized reference of the fusion transcript predicted by RNA-seq. We generated two primer pairs, one spanning the breakpoint of the fusion, and another one capturing exon 4 of *NRIP1* as control (Table S6). Available cDNA from patient samples was amplified using the KOD Xtreme Hot Start DNA Polymerase (Sigma-Aldrich, St. Louis, MO, USA) in 35 Stepdown cycles. Denaturation temperature was 95°C, annealing temperature was decreased stepwise during the first 12 cycles from 74°C to 62°C and elongation temperature was set to 68°C. PCR products were electrophoresed on a 1.8% agarose gel. Purification of the PCR products was done with QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and sent to Eurofins Genomics (http://www.eurofinsgenomics.eu) for Sanger sequencing.

Nanopore sequencing

Starting from 50ng of total RNA, 1st strand cDNA was synthesized with Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA, USA), an oligoT anchor primer and a strand-switching primer from PCR cDNA Barcoding kit (Oxford Nanopore Technologies, Oxford, UK). Full-length cDNA was enriched and amplified by PCR with barcoded, coupling-activated primers (Oxford Nanopore Technologies, Oxford, UK) and SeqAmp DNA polymerase (Takara, Kusatsu, Japan) for 20 cycles. After exonuclease I digestion of unincorporated primers and purification using Ampure XP magnetic beads (Beckman Coulter, Brea, USA), an equimolar amount of barcoded cDNA library was linked to coupling-activated sequencing adapter (PCR cDNA Barcoding kit, Oxford Nanopore Technologies, Oxford, UK) and sequenced for 24h on a R9.4.1 flowcell on a PromethION24 instrument (Oxford Nanopore Technologies, Oxford, UK). Sequencing reads were mapped with Minimap2[10] version 2.17 using default parameters. Genomic breakpoints were identified using the inversion caller npInv[11] version 1.24 with default parameters. The genomic rearrangement was visualized using Ribbon[12].

Supplementary References

1. Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res 2021;gr.257246.119.
2. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 2014;011650.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–2120.
4. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
5. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30(7):923–930.
6. Picard toolkit. *Broad Institute, GitHub repository*.
7. Jang YE, Jang I, Kim S, et al. ChimerDB 4.0: An updated and expanded database of fusion genes. Nucleic Acids Res;48(D1):.
8. Mitelman F JB and MF (Eds. . Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2020). https://mitelmandatabase.isb-cgc.org.
9. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 2012;13134.
10. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–3100.
11. Shao H, Ganesamoorthy D, Duarte T, Cao MD, Hoggart CJ, Coin LJM. npInv: Accurate detection and genotyping of inversions using long read sub-alignment. BMC Bioinformatics 2018;19(1):261.
12. Nattestad M, Aboukhalil R, Chin C-S, Schatz MC. Ribbon: intuitive visualization for complex genomic variation. Bioinformatics [Epub ahead of print].

Supplementary Figures

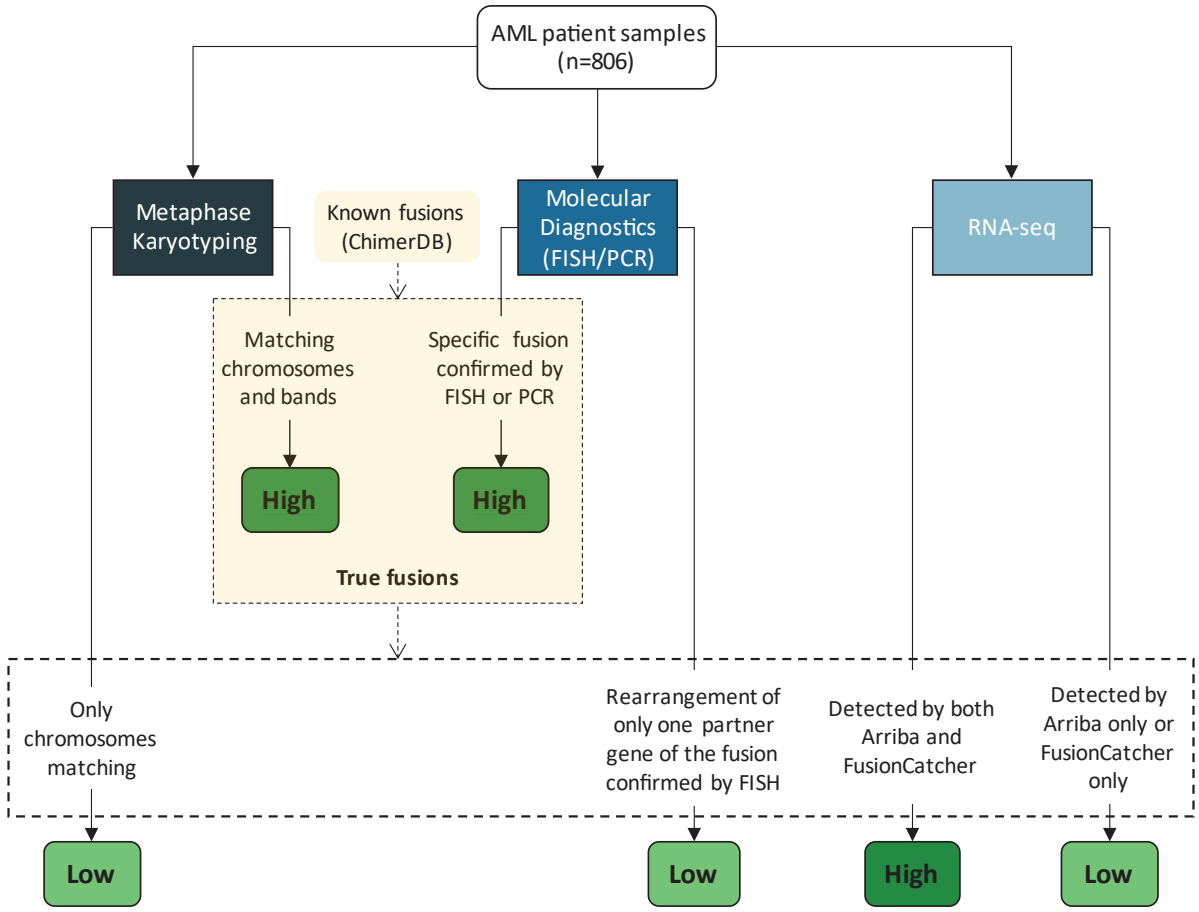


Figure S1: Illustration of the definitions for known/true fusions and high/low evidence for detected fusion events by metaphase karyotyping, molecular diagnostics and RNA-seq.

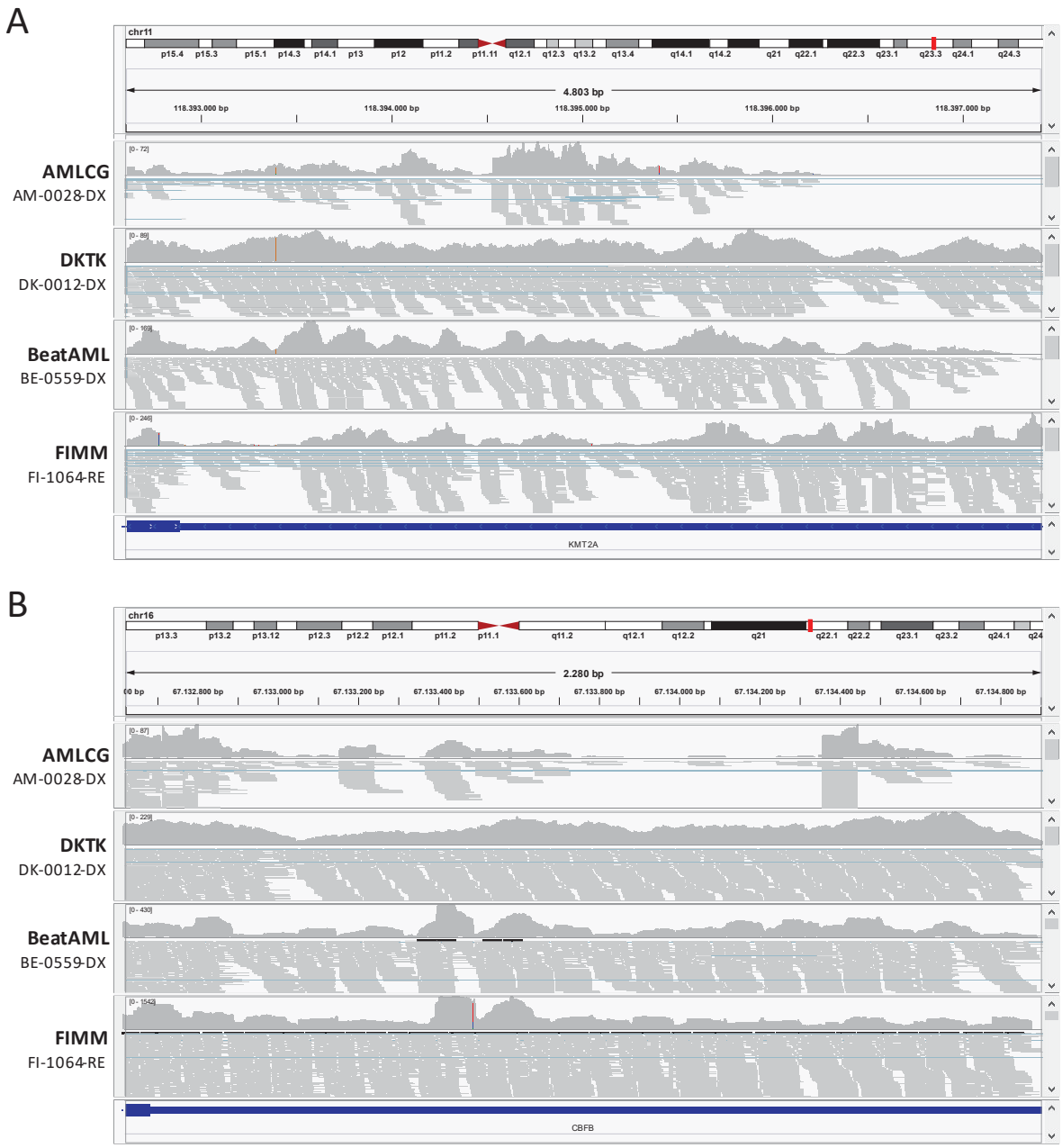


Figure S2: Mapped RNA-seq reads of samples from the AMLCG, DKTK, Beat AML and FIMM cohort, respectively, displayed by the IGV browser. Reads mapped to the locus of the gene A) *KMT2A* and B) *CBFB*.

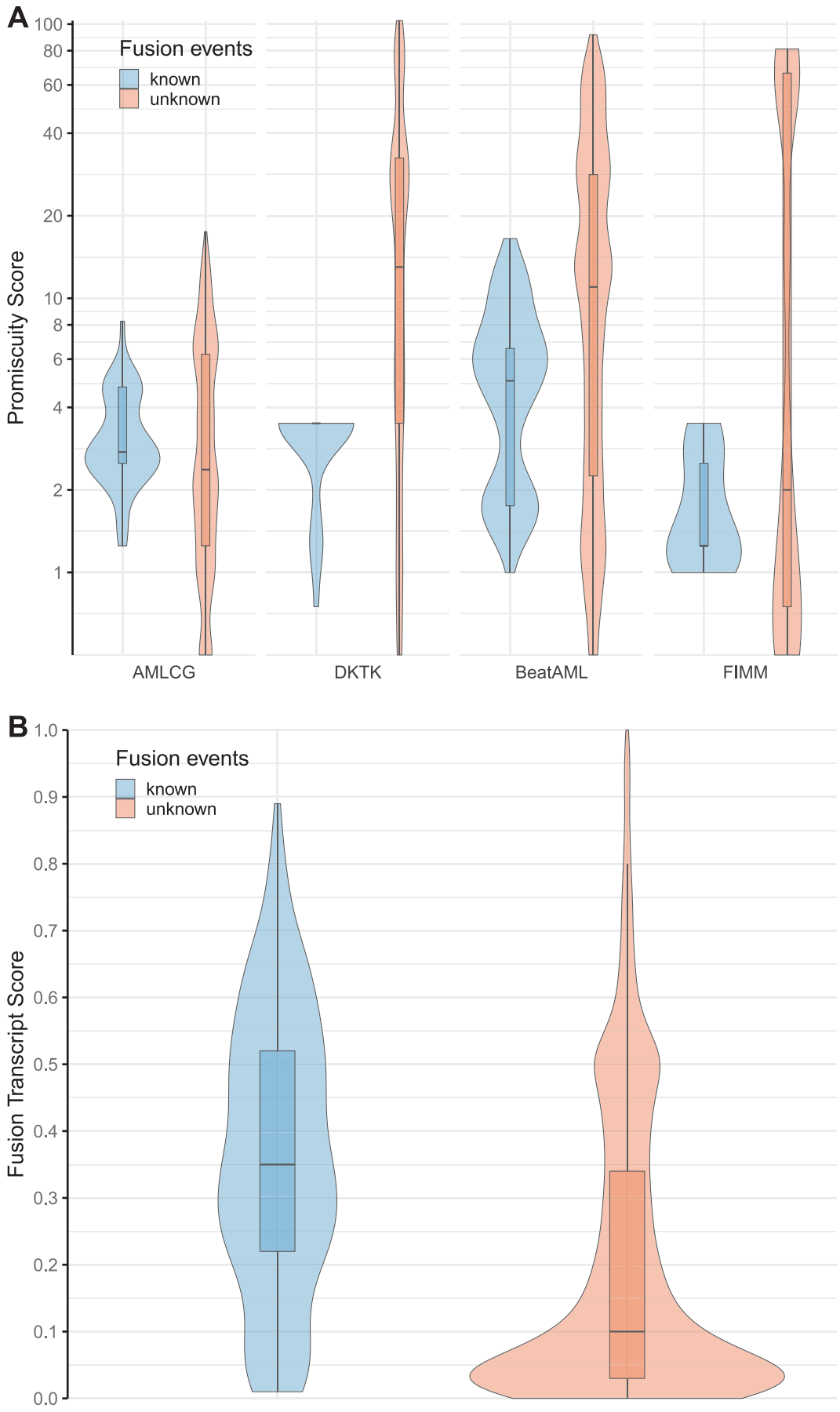


Figure S3: Distributions of A) Promiscuity Score by cohort and B) Fusion Transcript Score.

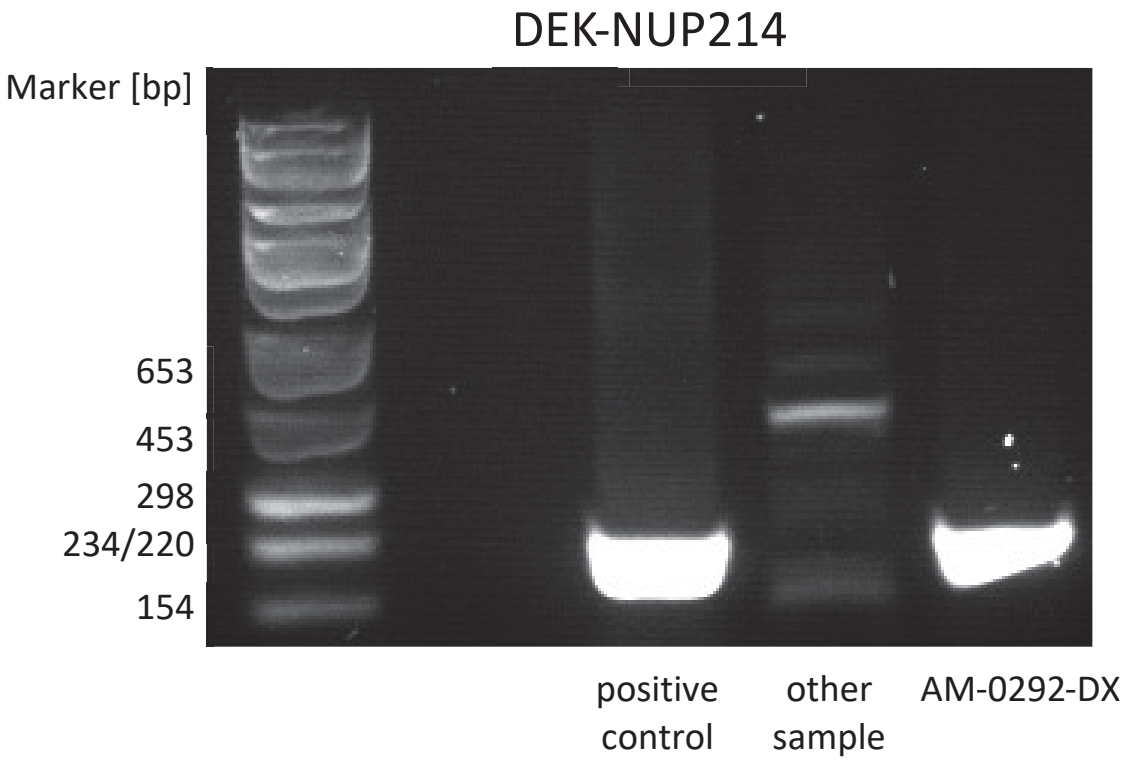


Figure S4: Electrophoresis of RT-PCR amplicons of *DEK-NUP214* fusion in sample AM-0292-DX.

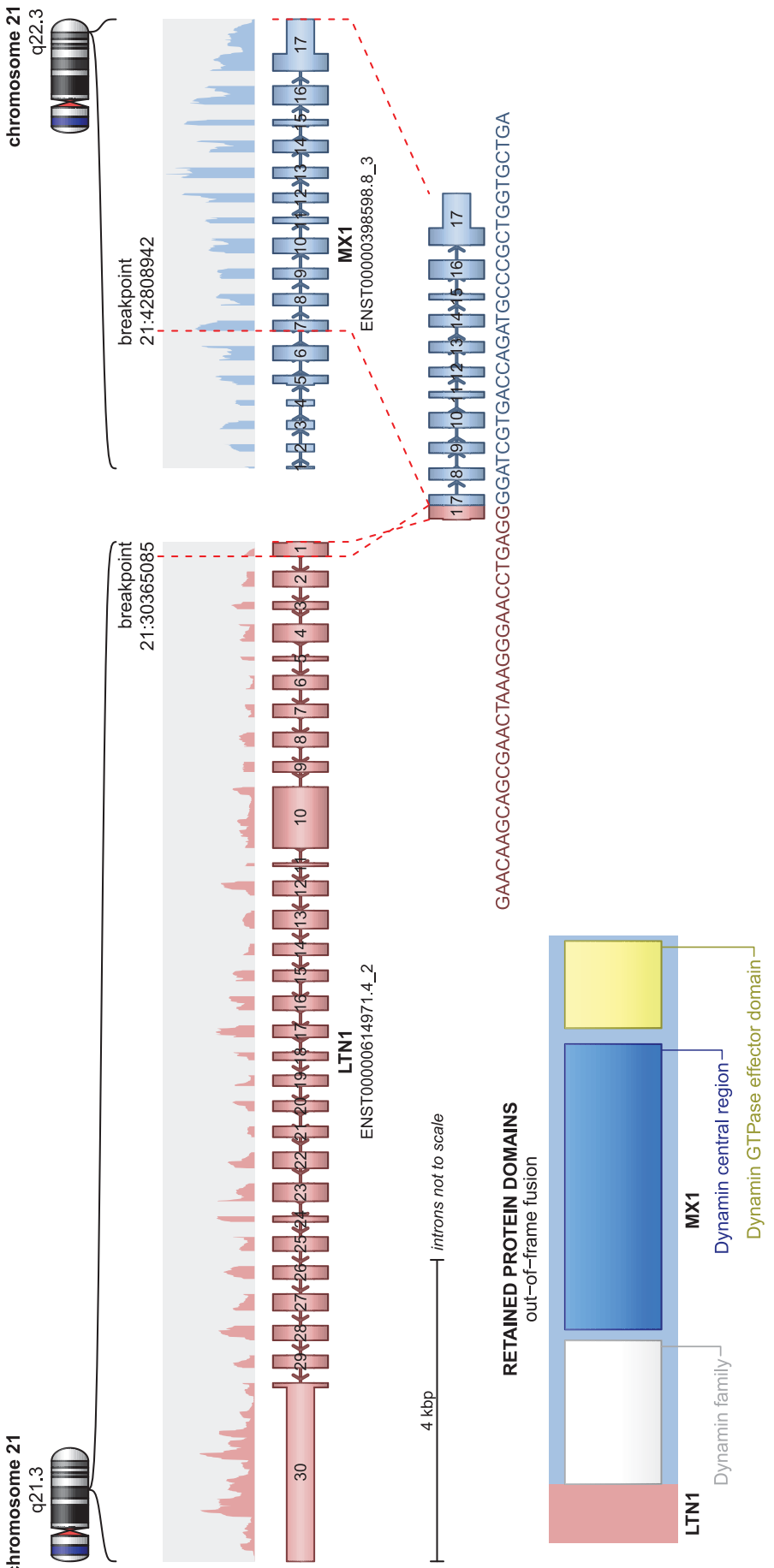


Figure S5: Schematic representation of the putative gene fusion transcript *LTN1-MX1* as predicted by RNA-seq in sample BE-1233-RD.

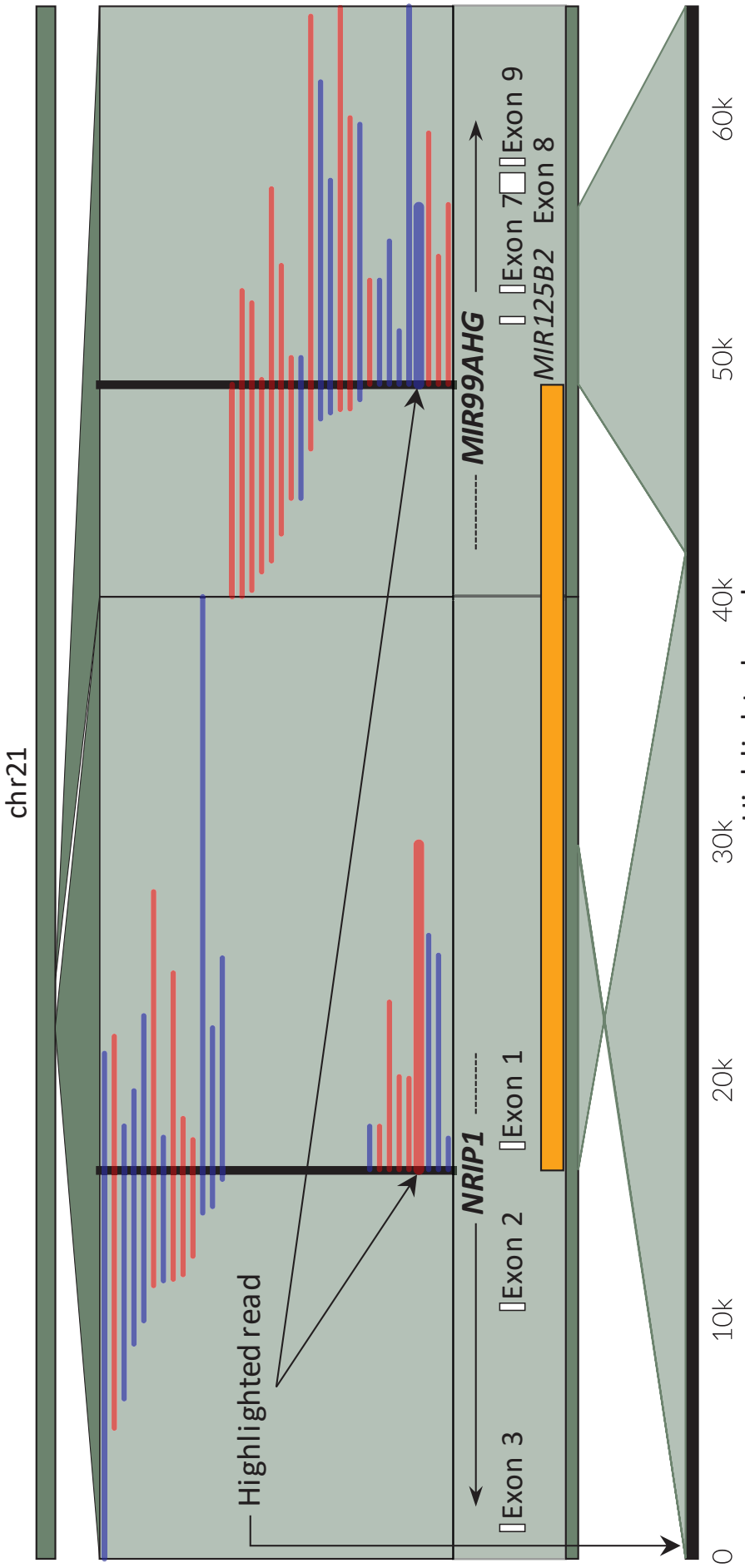


Figure S6: Mapping of long reads from Nanopore sequencing of genomic DNA of sample AM-0013-DX. Each line represents one read, which can be divided at the breakpoints of the fusion. Single parts of the read can be mapped to the positive strand (blue) at one locus with the other part mapped to the negative strand (red) at the other locus of chromosome 21. The consensus inversed region is marked in orange. Mapping structure of a highlighted read at the bottom shows that one part of the read was inversely mapped to the *NRIP1* locus, while the other part was mapped to the *MIR99AHG* locus.

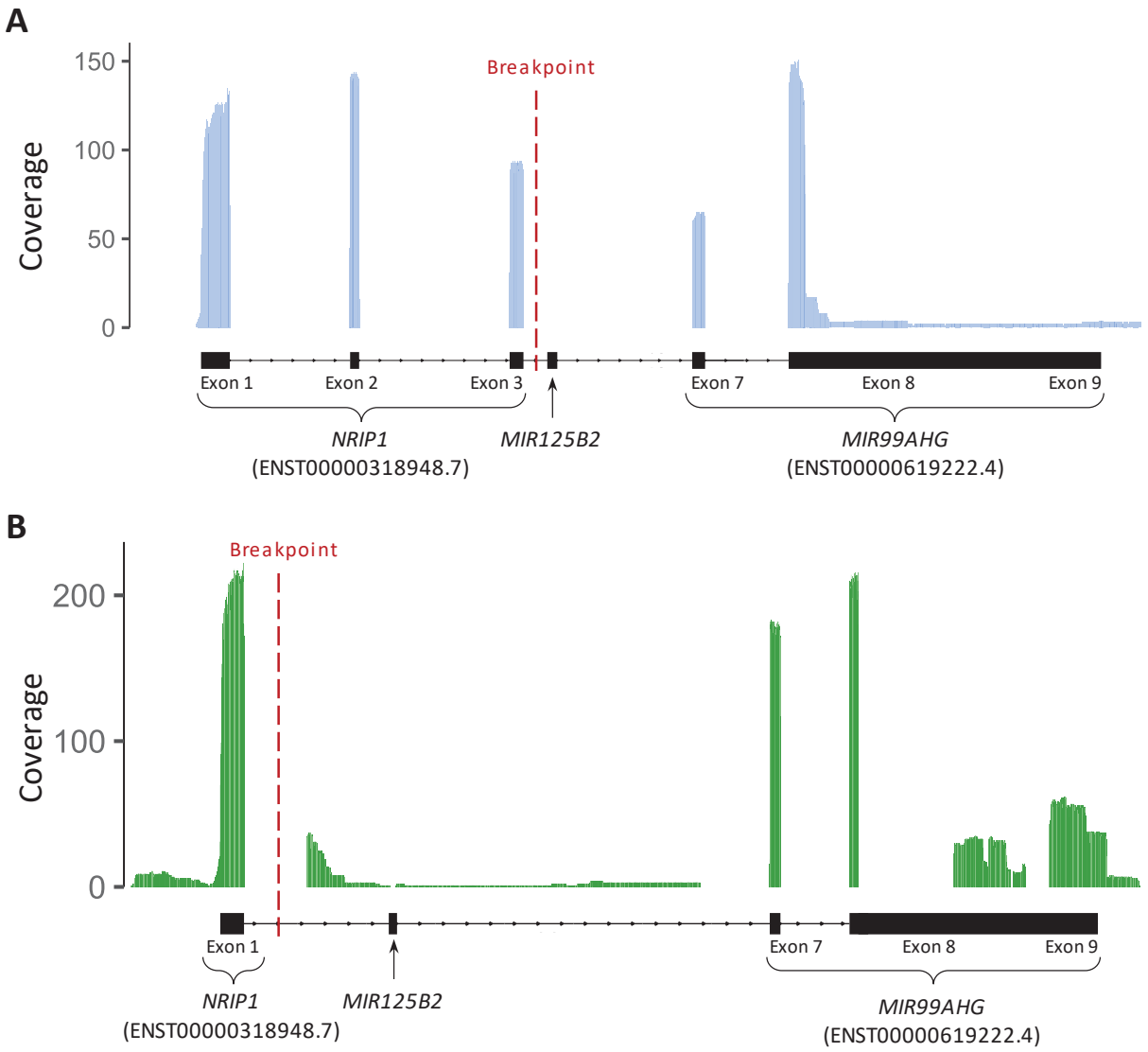


Figure S7: Read coverage of the customized reference of the *NRIP1-MIR99AHG* rearrangement by long reads from Nanopore sequencing of cDNA from samples A) AM-0028-DX B) AM-0013-DX. Control samples from two negative patients did not show any coverage and are therefore not shown.

Legends for Supplementary Tables

Table S1: Clinical data of patient samples from the AMLCG, DKTK, Beat AML and FIMM cohort.

Table S2: Summary of publicly available RNA-seq data of healthy bone marrow samples.

Table S3: List of samples harboring true fusions and evidence by Karyotyping, MDx and RNA-seq for each case. Dark green indicates high evidence, light green indicates low evidence. Grey represents no evidence although the respective method was performed.

Table S4: Novel fusion candidates that passed all filter steps and were consistently called between Arriba and FusionCatcher.

Table S5: List of samples harboring known fusions as reported by RNA-seq that had no or low evidence only by Karyotyping or MDx. Dark green indicates high evidence, light green indicates low evidence. Grey represents no evidence although the respective method was performed.

Table S6: Primer sequences capturing the junction of a *NRIP1-MIR99AHG* fusion transcript and exon 4 of *NRIP1* in sample AM-0028-DX and AM-0013-DX. Genomic positions of inversion breakpoints identified by long reads from Nanopore sequencing.

Table S7: Clinical and genetic characteristics of patients with *NRIP1-MIR99AHG* fusion.

Table S8: Detailed parameters of tools used in the fusion detection workflow in the present study.

Table S9: Annotations for fusion events as reported by FusionCatcher that indicate artifacts or fusion events that were detected in healthy samples.

Table S10: Blacklist of fusion genes generated from fusion events that were detected in RNA-seq data of healthy bone marrow samples.