Ferrata Storti Foundation

Haematologica 2022
Volume 107(1):100-111

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] and 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, Germany; [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 and [7]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Presented in part at the 61st ASH Annual Meeting Kerbs P, et al. Blood. 2019;134(Suppl_1):4655.

Correspondence:

PHILIPP A. GREIF
pgreif@med.lmu.de

Received: January 26, 2021.

Accepted: May 3, 2021.

Pre-published: June 17, 2021.

https://doi.org/10.3324/haematol.2021.278436

ABSTRACT

Identification of fusion genes in clinical routine is mostly based on cytogenetics and targeted molecular genetics, such as metaphase karyotyping, fluorescence *in situ* hybridization and reverse-transcriptase polymerase chain reaction. 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 compared to standard diagnostic techniques, we analyzed 806 RNA-sequencing samples from patients with acute myeloid leukemia using two state-of-the-art software tools, namely Arriba and FusionCatcher. RNA-sequencing detected 90% of fusion events that were reported by routine with high evidence, while samples in which RNA-sequencing 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-sequencing. 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-sequencing 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 mechanisms of leukemogenesis in acute megakaryoblastic leukemia. Our findings demonstrate that RNA-sequencing 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] Hematopoietic malignancies are well-characterized regarding the abundance of fusion genes, e.g, chronic myeloid leukemia harbors the *BCR-ABL1* fusion in more than 95% of cases, and acute promyelocytic leukemia is characterized by *PML-RARA* in 90% of cases. In acute myeloid leukemia (AML), fusion genes are found in about 30% of patients[2] and are often regarded as major markers, defining clinically relevant subtypes.[3-6] Their identification is crucial for risk assessment and deciding treatment strategy. During the 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) 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, the 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 difficult to detect with 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 usually serve as complementary validation methods.

Over the last decade, next-generation sequencing techniques have evolved tremendously and are being increasingly used in clinical diagnostics.[7-9] Next-generation sequencing 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. Furthermore, RNA-sequencing enables 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-sequencing 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 enables the processing of even more samples in parallel with even greater 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-sequencing have been published[9,11,12] but identification of fusions using RNA-sequencing 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-sequencing and clinical routine has been reported so far. In this study, we set out to assess the potential of RNA-sequencing 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

Patients' samples

A total of 806 AML patients' samples were subjected to whole-transcriptome sequencing. The samples were collected from 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

Table 1. **Summary of the patients' characteristics.**

Cohort	Median age (range)	Sex, n (%)	ELN risk group, n (%)

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)	Favorable = 33 (47.8)
			Males = 38 (55.1)
			ntermediate-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)

ELN: European LeukemiaNet; NA: not available.

(FIMM, n=57).[18] RNA-sequencing was performed as described previously.[13-18] The patients' characteristics are summarized in Table 1 and *Online Supplementary Table S1*. Sequencing metrics are summarized in Table 2. In addition, RNA-sequencing data of healthy samples were obtained from the Gene Expression Omnibus Database (*Online Supplementary 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 evaluating performance in fusion gene detection

Comprehensive definitions and metrics are provided in the *Online 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, molecular diagnostics and RNA-sequencing (*Online Supplementary Methods, Online Supplementary Figure S1*).

Filtering strategies

Initially, built-in filters of the callers were applied and fusions were filtered by a custom-generated blacklist (*Online 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 transcripts per million, 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 *Online Supplementary Methods*.

Results

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

In 806 patients' samples, we identified 138 true fusions which provided the benchmark for the comparison of performance in fusion gene detection between routine diagnostics (Karyotyping, molecular diagnostics) and RNA-sequencing (Figure 1, *Online Supplementary Table S3*). Of 138 true fusions, Karyotyping identified 121 (87.7%) and molecular diagnostics identified 107 (77.5%) with high evidence. In addition, Karyotyping identified 11 (8%) and molecular diagnostics identified five (3.6%) true fusions with low evidence. Fusion gene detection by RNA-sequencing 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 (*Online Supplementary Figure S2*), both involved in eight of ten undetected true fusions by RNA-sequencing in those samples. Further fusions missed by RNA-sequencing 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-sequencing and clinical routine in three of four cases of true fusions missed by RNA-sequencing: (i) Karyotyping reported t(6;11)(q27;q23) resulting in *KMT2A-AFDN*, while RNA-sequencing detected *KMT2A-MLLT10* resulting from t(10;11)(p12;q23); (ii) Karyotyping reported del(2)(p21p23) resulting in *EML4-ALK*, while RNA-sequencing 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-sequencing detected *PML-RARA* but not *STAT5B-RARA*. In the fourth case, a *PML-RARA* fusion was found by FISH while Karyotyping indicated a normal karyotype in this sample.

RNA-sequencing 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, respectively) (Table 2). We applied filtering strategies as shown in Figure 2A. PS filter cutoffs for individual cohorts were set at 8.25 for AMLCG, 3.5 for DKTK, 16.5 for Beat AML and 3.5 for FIMM (*Online Supplementary Figure S3A, Online Supplementary Methods*). In addition to our previously described cutoffs for $\mathrm{FTS}_{5'}$ and $\mathrm{FTS}_{3'}$ (*Online Supplementary Methods*), we set a minimum FTS for unknown fusion events based on the median FTS of all detected unknown fusions (FTS ≥0.1) (*Online Supplementary Figure S3B*). Finally, we defined highly reliable fusion gene candidates based on

Table 2. **Statistics for RNA-sequencing, 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 millions (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

Avg: average.

the overlap of filtered fusion calls from Arriba and FusionCatcher. The built-in filter of Arriba excluded, on average, 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 also found 187 fusion events as robust candidates. Thirty of these 187 events have been described before, while 157 were putative novel fusion events (*Online Supplementary Table S4)*. Clinical routine showed only low evidence for four of the 30 known events (Figure 1B, *Online Supplementary Table S5*), while 26 candidates were not reported by routine diagnostics in our cohorts. In two of the

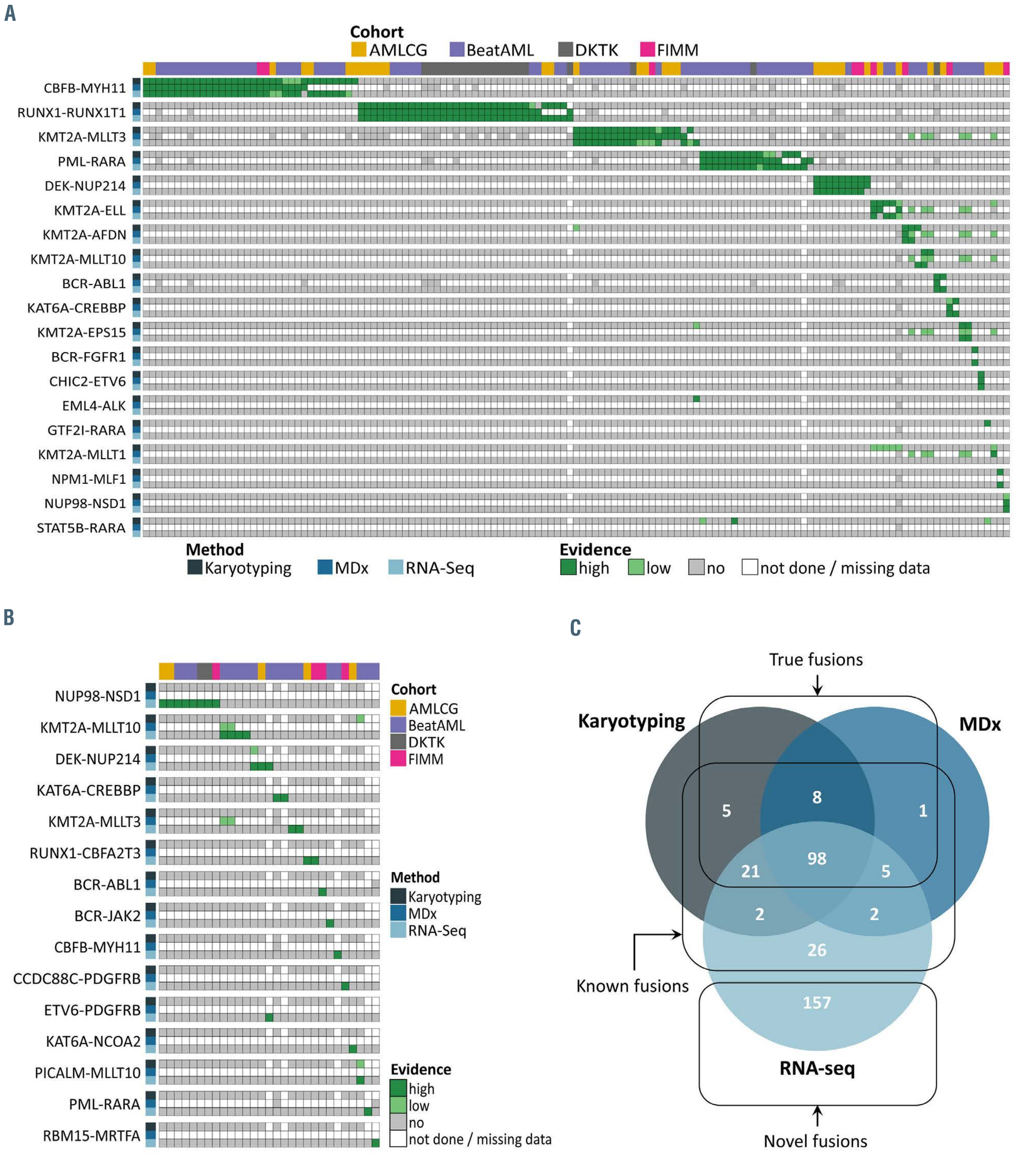

Figure 1. Evidence for fusions by routine clinical diagnostics and RNA-sequencing. (A) True fusions detected by Karyotyping, molecular diagnostics (MDx) and RNA-sequencing (RNA-seq) in the AMLCG, DKTK, Beat AML and FIMM cohorts. Dark green boxes indicate high evidence, light green boxes indicate low evidence. Gray 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 that were missed or detected with low evidence only by Karyotyping/MDx. (C) Venn diagram summarizing fusions detected with the different methods.

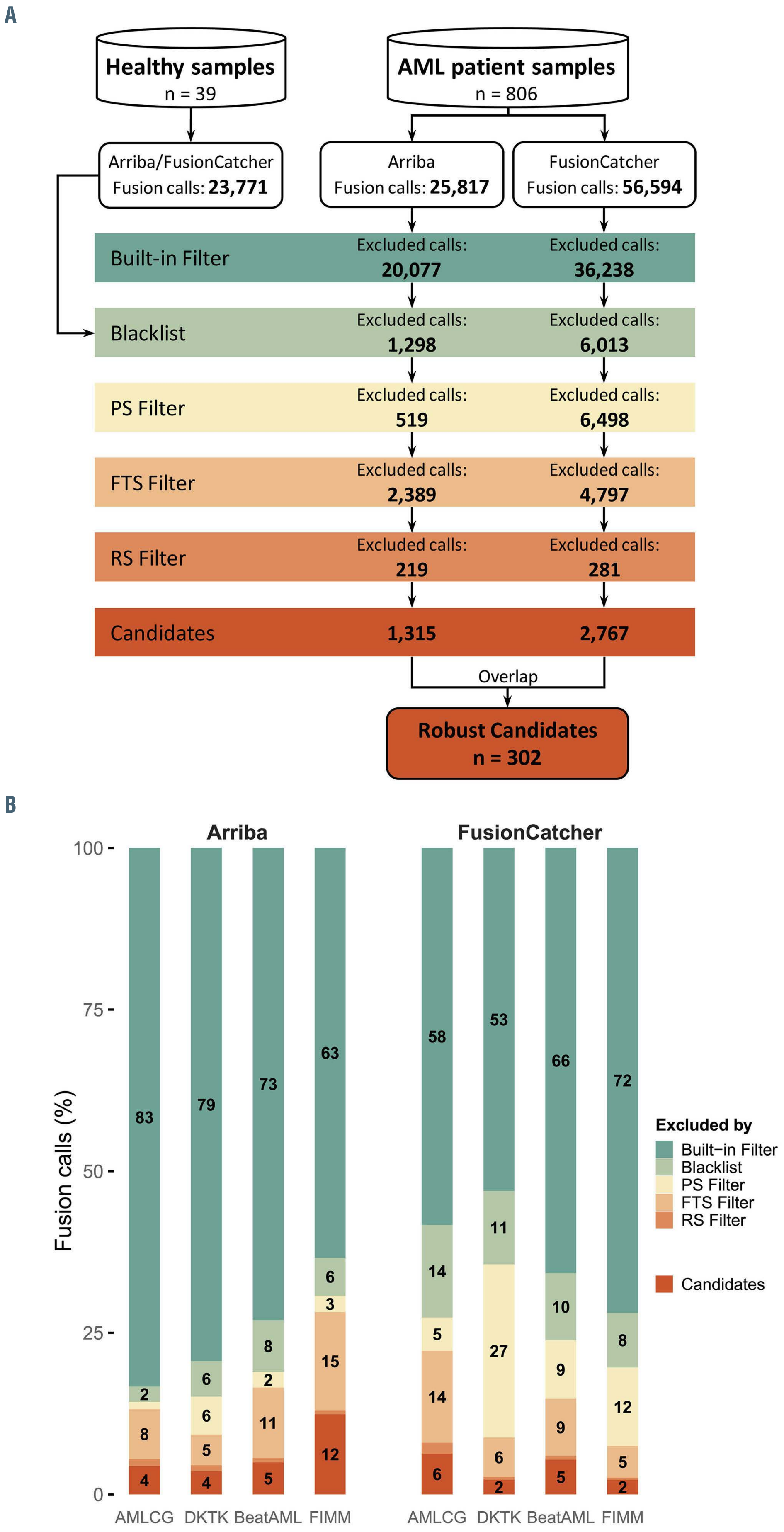

Figure 2. Detection workflow and filtering of fusion events. (A) Detection workflow and number of filtered fusion events by filtering strategies. (B) Ratios of fusion events excluded by Arriba and FusionCatcher in each filter step and cohort. AML: acute myeloid leukemia; FTS: Fusion Transcript Score; PS: Promiscuity Score; RS: Robustness Score.

four events described by clinical routine, a rearrangement of *KMT2A* was detected 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 European LeukemiaNet risk group in six of the 30 cases (*Online Supplementary Table S5*). Chromosomal locations of detected true, known and puta-

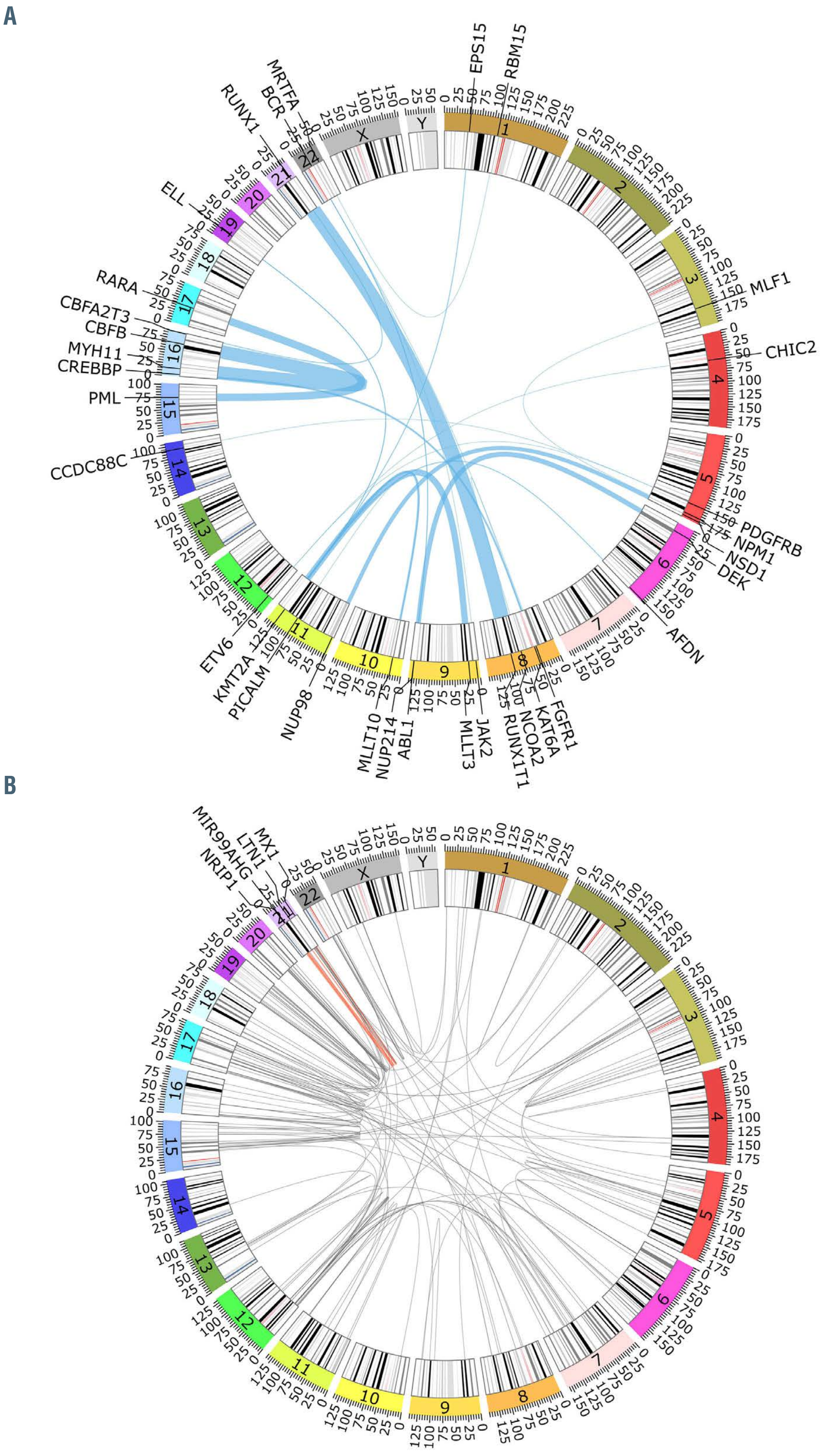

Figure 3. Genomic origin of fusion events detected by RNA-sequencing. Circos plots of (A) known and (B) unknown fusion gene candidates found in the AMLCG, DKTK, Beat AML and FIMM cohorts, illustrating chromosomal origin of the fusion events. Lines connect the positions of fusion partners. Thickness of 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 gray lines show non-recurrent novel fusion events.

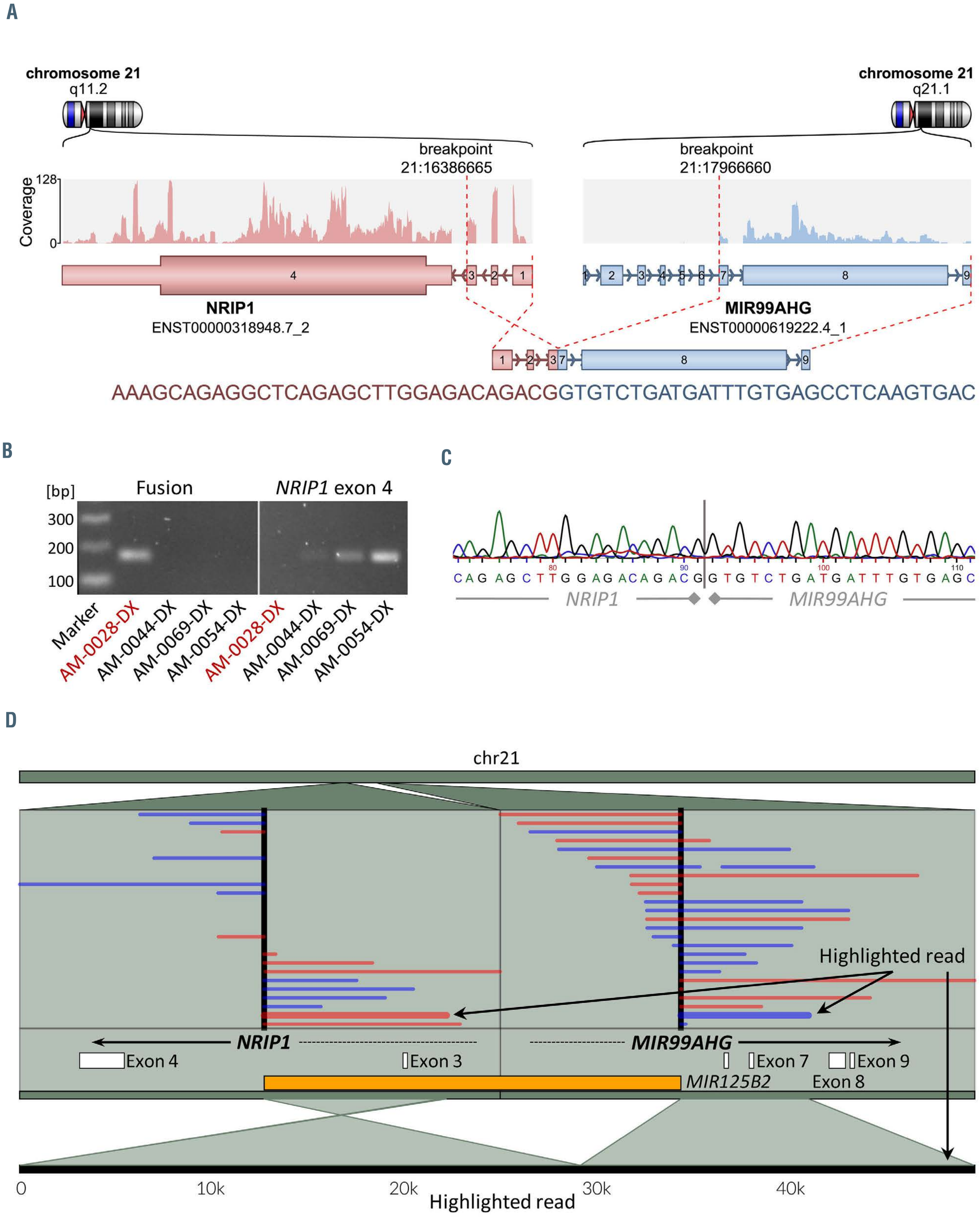

Figure 4. Detection and validation of the novel *NRIP1-MIR99AHG* fusion gene. Evidence for the *NRIP1-MIR99AHG* fusion gene in sample AM-0028-DX determined by various methods. (A) Schematic representation of the fusion transcript as predicted by RNA-sequencing. (B) Gel-electrophoresis of reverse transcriptase polymerase chain reaction analysis of fusion breakpoint and *NRIP1* exon 4. Three samples from cytogenetically normal patients with acute myeloid leukemia were used as negative controls. (C) A trace from Sanger sequencing of the 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. The 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.

tive 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-sequencing in one sample (AM-0292-DX: *DEK-NUP214*) (*Online Supplementary Figure S4*). 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* (*Online Supplementary Figure S5*) resulting from inv(21)(q21.3;q22.3) in two patients' sam-

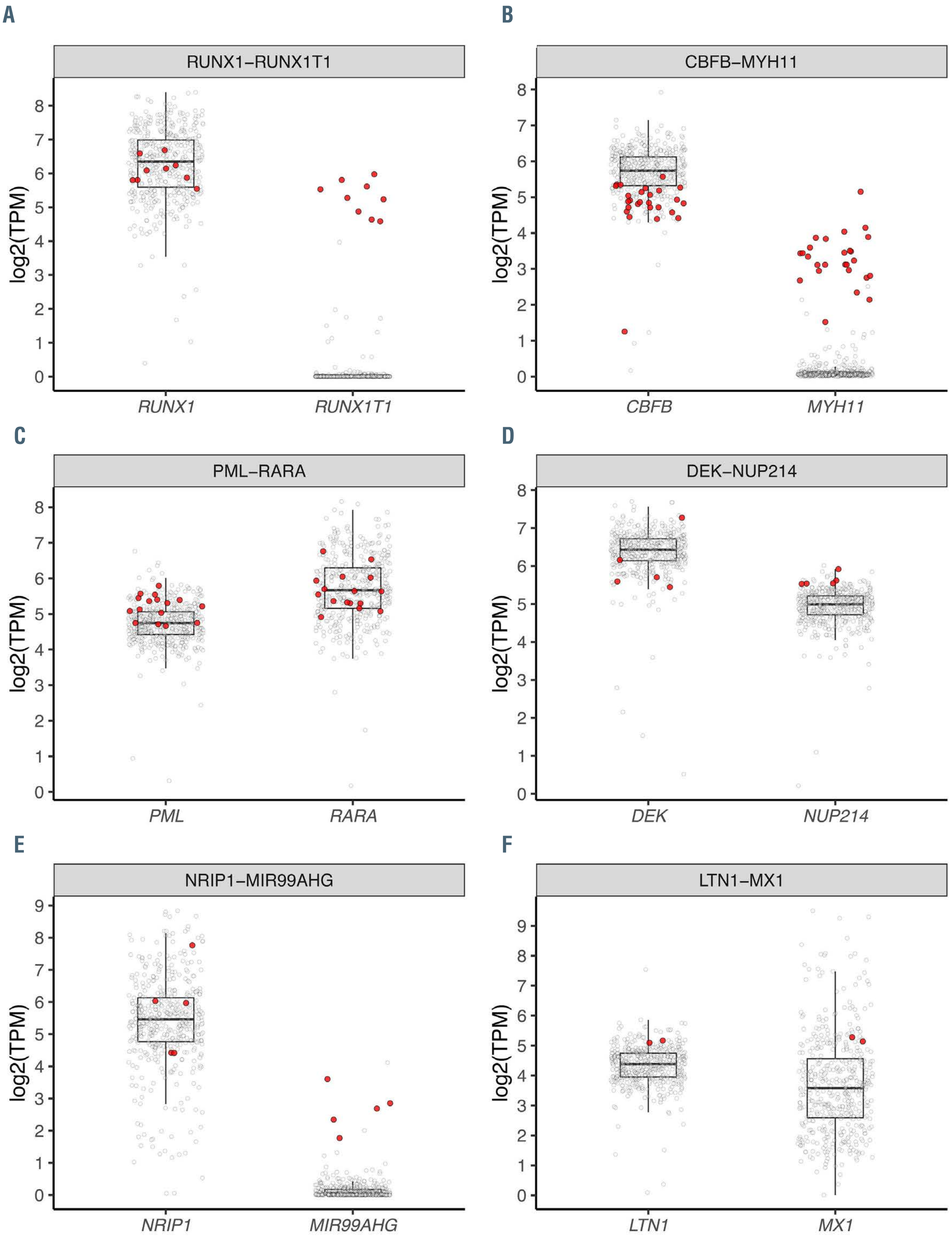

Figure 5. Gene expression of genes involved in fusions. Gene expression of the 5' and 3' partner genes of the respective fusion. Red dots indicate samples positive for the respective fusion, gray dots represent samples negative for the respective fusion. TPM: transcripts per million.

ples. Notably, *LTN1-MX1* was only found in co-occurrence with *NRIP1-MIR99AHG*. Further recurrence of *NRIP1-MIR99AHG* was reported by FusionCatcher alone in two patients' 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-sequencing fusion callers (Figure 4C). Nanopore sequencing of available gDNA from *NRIP1-MIR99AHG*-positive samples AM-0028-DX (Figure 4D) and AM-0013-DX (*Online Supplementary Figure S6*) identified the breakpoints (*Online Supplementary Table S6*) and confirmed an inversion on the genomic level. With the aim of determining 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 (*Online Supplementary Figure S7*), while there was no coverage in negative controls. *NRIP1* includes a consensus coding sequence with an open reading frame starting in exon 4, whereas *MIR99AHG* is non-coding. The identified breakpoint in the *NRIP1* locus in AM-0028-DX was located between exons 3 and 4, while the breakpoint in AM-0013-DX was located between exons 1 and 2, consistent with reports from RNA-sequencing fusion callers. In both cases, no annotated open reading frame was included in the putative fusion transcripts. A validation in samples from the Beat AML cohort was not possible because of 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-sequencing 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%) 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 expressed at low levels only, showed 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 expressed 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 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 years (*Online Supplementary 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 about whether these patients had a constitutional or somatic monosomy/trisomy 21. Cytomorphology was available for three of the nine patients without there being any evidence of megakaryoblastic leukemia (French-American-British classification, M7). Mutational status was available for six of the nine patients, but no apparent pattern was observed. However, recurrently mutated genes among those patients were *NRAS* (n=2) and *ASXL1* (n=2) (*Online Supplementary Table S7*).

Discussion

The aim of this study was to test the potential of fusion gene detection by RNA-sequencing in several cohorts of AML patients' 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-sequencing, underscoring the high sensitivity of this method. Notably, most of the samples in which a true fusion could not be detected by RNA-sequencing had a low read depth (median = 24 million 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-sequencing 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. Diagnostic application of RNA-sequencing 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-sequencing, as well as further known fusion genes in 22 samples that showed no or only low evidence for these fusions by either Karyotyping or molecular diagnostics. (ii) We observed discrepancies between results from routine and RNA-sequencing, 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-sequencing 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-sequencing identified a *KMT2A-MLLT10* fusion in these samples. Since various *KMT2A* fusions may reflect different risk assessments based on the European LeukemiaNet classification,[6] the correct description of the fusion may have therapeutic consequences. (iii) In another sample, Karyotyping and FISH identified a t(15;17)(q24;q21) translocation, typically resulting in a *PML-RARA* fusion transcript (no information on PCR status was available), while RNA-sequencing identified a *PML-CASC3* fusion, with *CASC3* being located ~170 kb upstream of *RARA*. Unfortunately, no information was available on this patient's response to all-*trans* retinoic acid treatment.

In addition to standard diagnostic methods that are used in clinical routine, targeted RNA-sequencing 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-sequencing-based fusion callers report many false positive events due to technical and biological properties, such as 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 customized blacklist further improves the specificity in RNA-sequencing-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. The RS filter therefore excludes fusion events that failed at least one preceding filter in most of the identified cases. The integration of our PS, FTS, customized blacklist and RS Filter into our detection strategy substantially reduced the fusion calls that were 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 as 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-sequencing revealed different breakpoint positions in *NRIP1-MIR99AHG*-positive samples. 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 hematologic malignancies,[33,34] and was found to be involved in other fusions.[35] A disruption of the corresponding gene by the *NRIP1-MIR99AHG* rearrangement might therefore contribute to leukemogenesis. On the other hand, overexpression of *MIR99AHG* and accompanying enhanced proliferation were 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 microRNA 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 whom had 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 patients with AML and 630 with myelodysplastic syndromes did not find any *NRIP1-MIR99AHG* fusions.[38] An extended analysis by the same authors[39] of overlapping cohorts, presented at the recent Annual Meeting of the American Society of Hematology, identified recurring *NRIP1-MIR99AHG* in AML and myelodysplastic syndromes 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 hematologic malignancies.

In conclusion, RNA-sequencing 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-sequencing, but low sequence coverage limited sensitivity in a subset of samples. These findings underscore the need for stringent quality metrics in diagnostic RNA-sequencing applications. Nevertheless, we found several AML-related fusions that are difficult to detect 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-sequencing as a valuable complementary method to current standard techniques for the detection of fusion genes and we recommend the integration of RNA-sequencing applications into clinical routine for more comprehensive and precise diagnostics of hematologic malignancies.

Disclosures

No conflicts of interest to disclose.

Contributions

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

Acknowledgments

We thank all participants and recruiting centers of the AMLCG and Beat AML trials. We also thank Bianka Ksienzyk and William Keay for technical support.

Funding

This study was supported by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center (SFB) 1243 "Cancer Evolution" (projects A08 and A16). PAG acknowledges support from the Munich Clinician Scientist Program (MCSP). SV was supported by the Deutsche José Carreras Leukämie-Stiftung.

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. 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 Suppl):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. pp 153-194.
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.ht ml (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(Suppl 1):16-17.