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;31(3):448–460.
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;11650.
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*. http://broadinstitute.github.io/picard/ (2019).
7. Jang YE, Jang I, Kim S, et al. ChimerDB 4.0: an updated and expanded database of fusion genes. Nucleic Acids Res 2019;48(D1):D817–D824.
8. Mitelman F, Johansson B, Mertens F (Eds.). Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. https://mitelmandatabase.isb-cgc.org (2021).
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

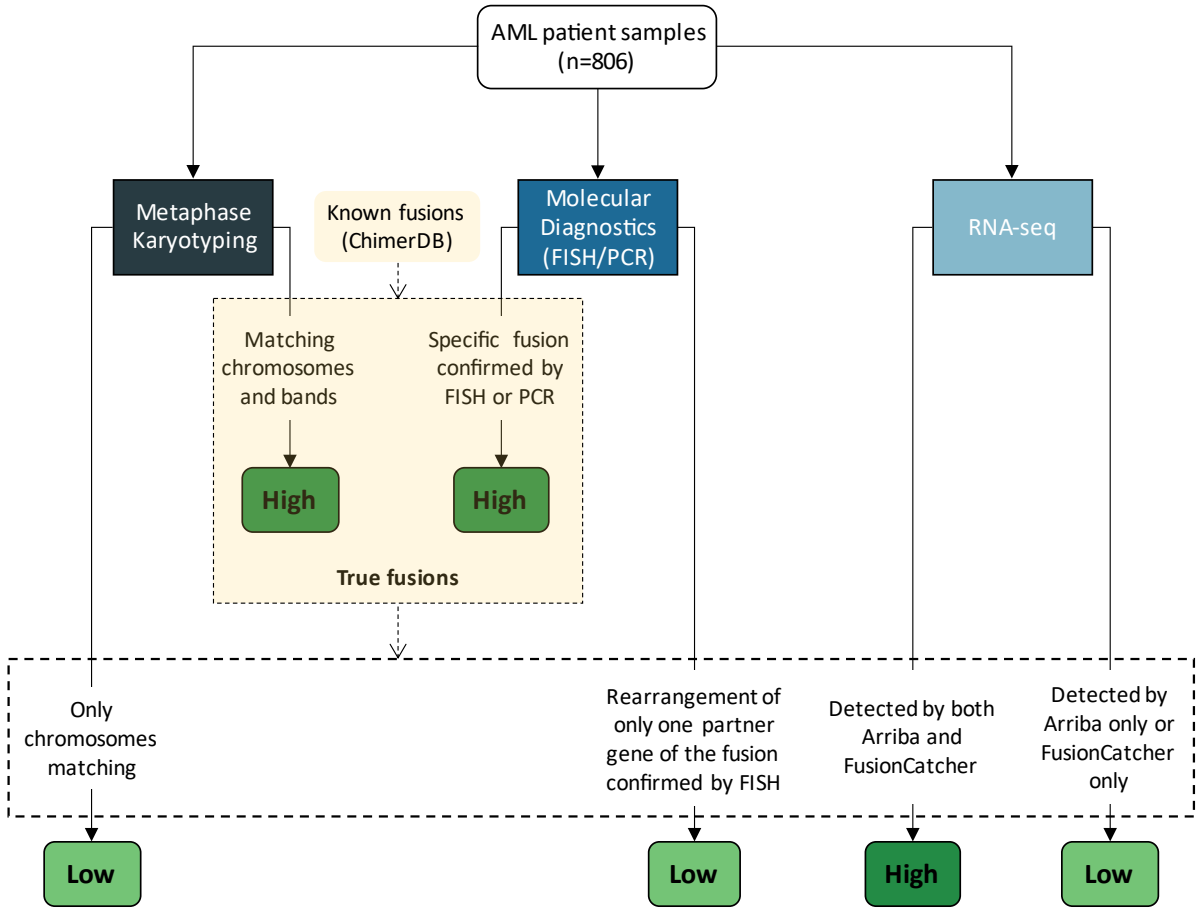

Figure S1: Illustration of the definitions for known/true fusions and high/low evidence for detected fusion events by metaphase karyotyping, molecular diagnostics and RNA-seq.

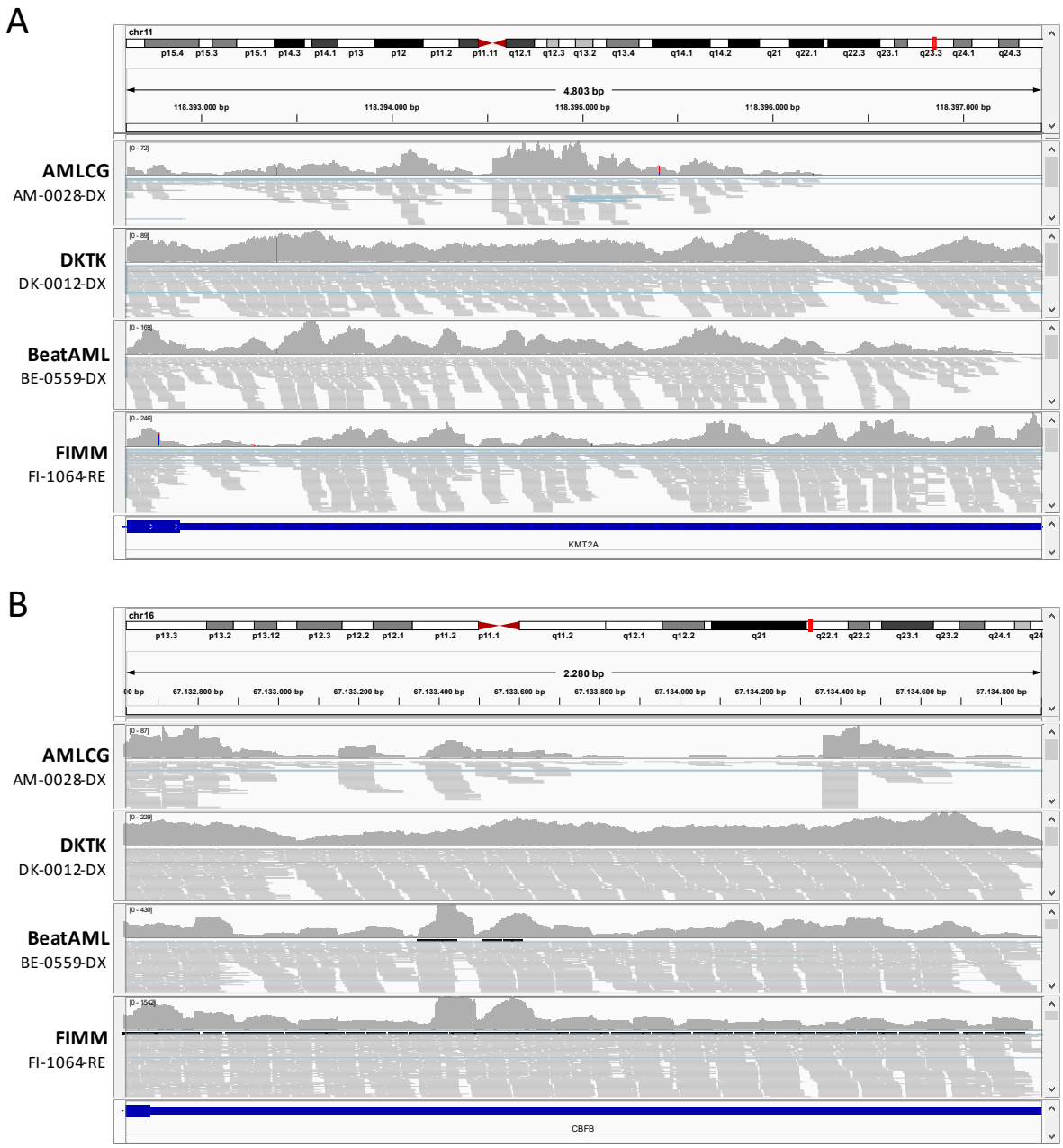

Figure S2: Mapped RNA-seq reads of samples from the AMLCG, DKTK, Beat AML and FIMM cohort, respectively, displayed by the IGV browser. Reads mapped to the locus of the gene A) *KMT2A* and B) *CBFB*.

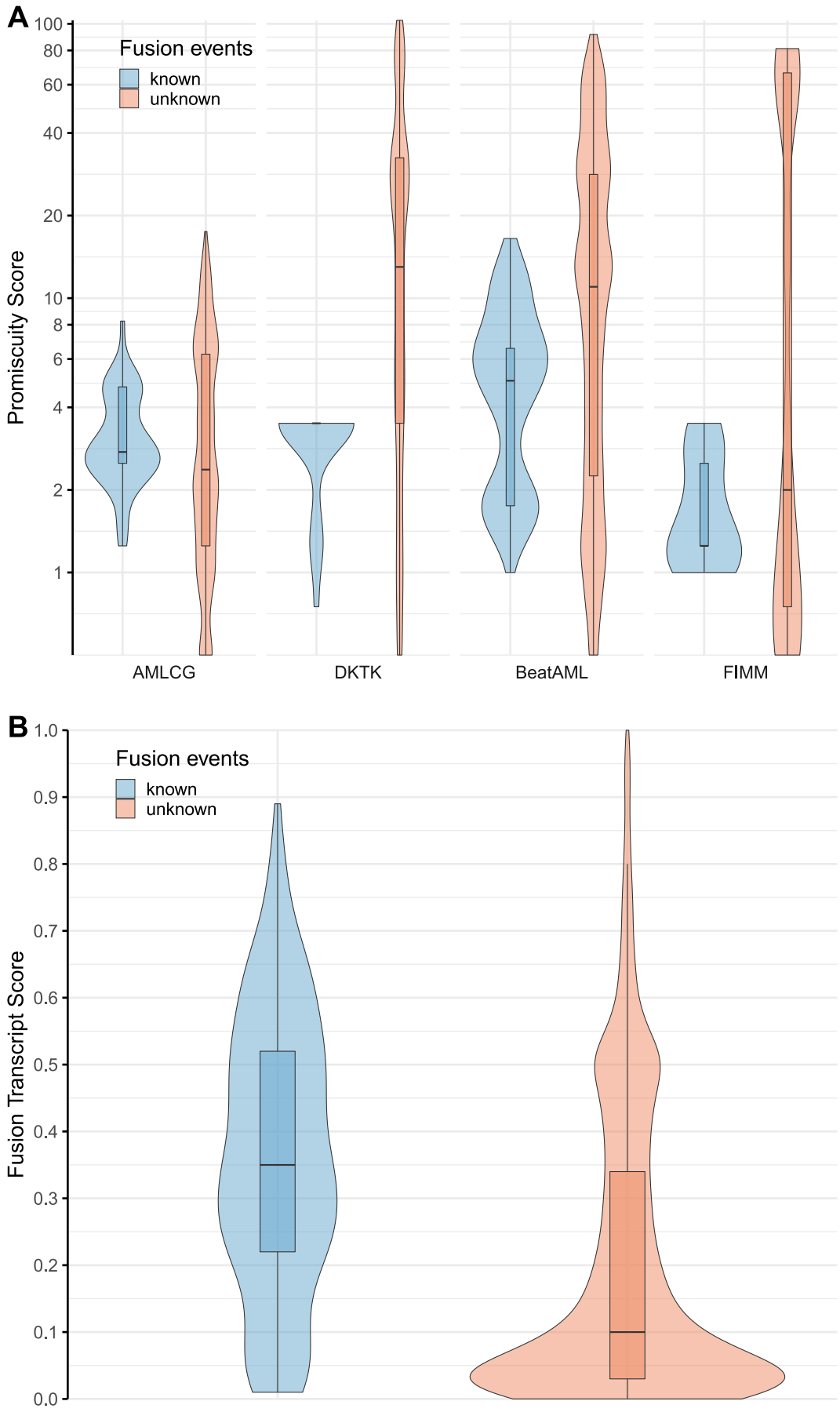

Figure S3: Distributions of A) Promiscuity Score by cohort and B) Fusion Transcript Score.

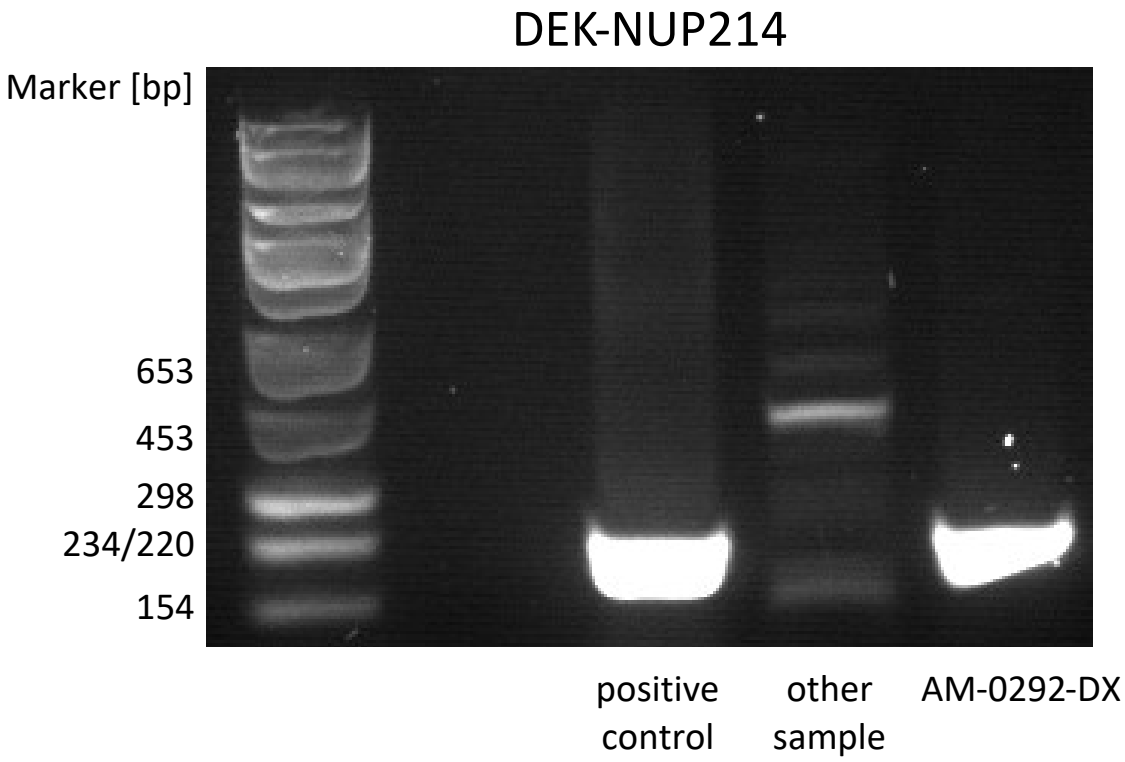

Figure S4: Electrophoresis of RT-PCR amplicons of *DEK-NUP214* fusion in sample AM-0292-DX.

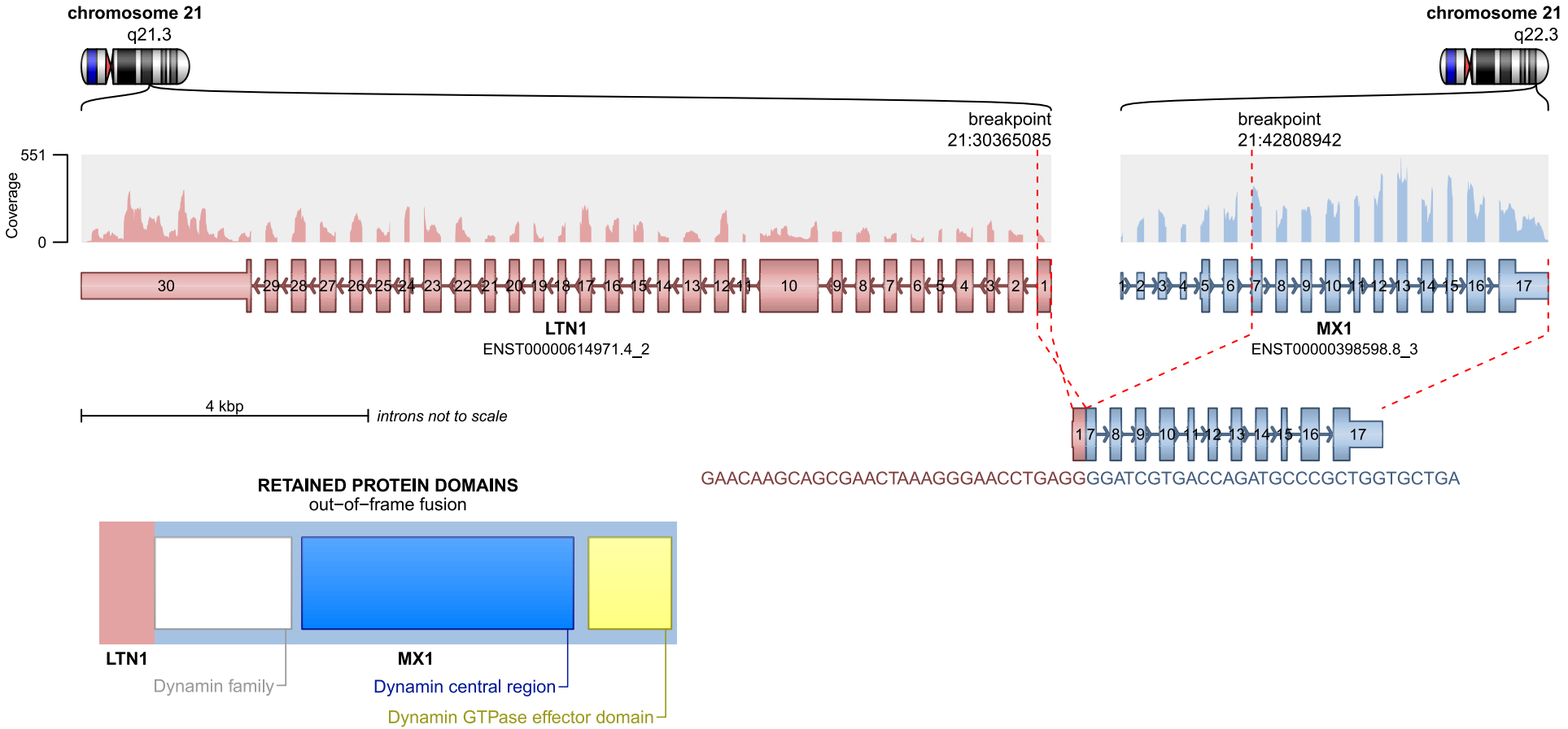

Figure S5: Schematic representation of the putative gene fusion transcript *LTN1-MX1* as predicted by RNA-seq in sample BE-1233-RD.

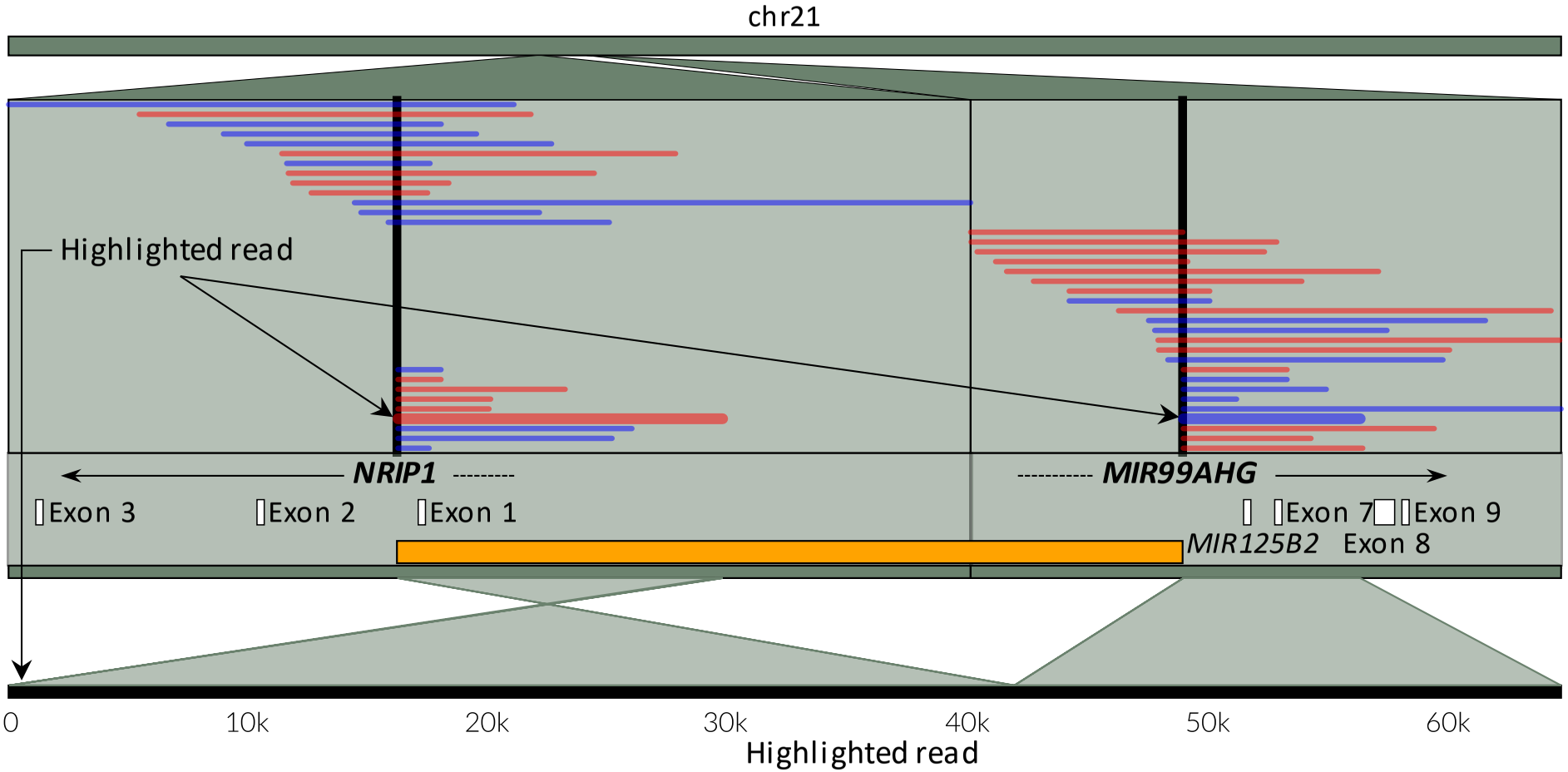

Figure S6: Mapping of long reads from Nanopore sequencing of genomic DNA of sample AM-0013-DX. Each line represents one read, which can be divided at the breakpoints of the fusion. Single parts of the read can be mapped to the positive strand (blue) at one locus with the other part mapped to the negative 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.

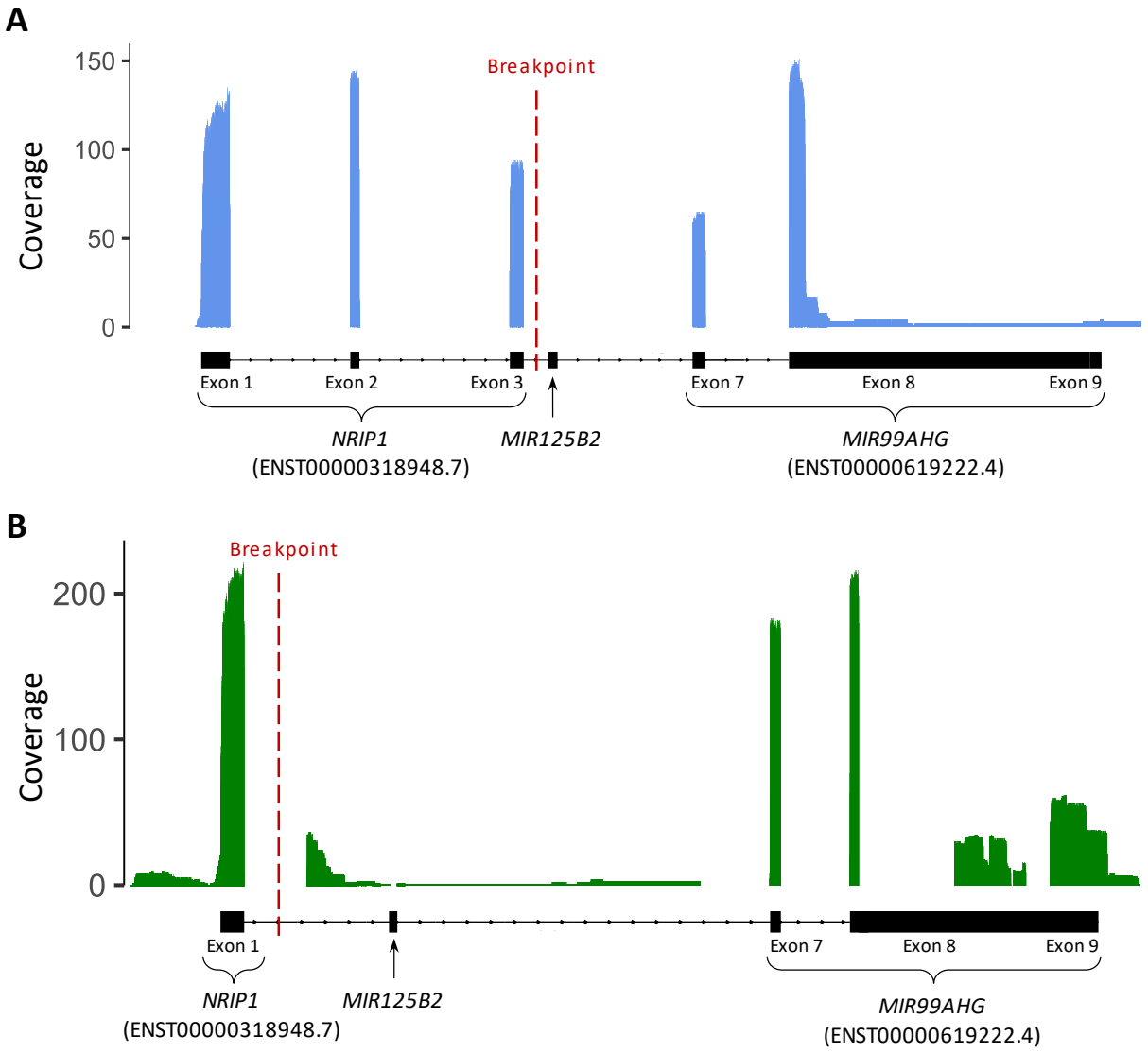

Figure S7: Read coverage of the customized reference of the *NRIP1-MIR99AHG* rearrangement by long reads from Nanopore sequencing of cDNA from samples A) AM-0028-DX B) AM-0013-DX. Control samples from two negative patients did not show any coverage and are therefore not shown.

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.