NAR Genomics and Bioinformatics, 2024, **6**, lqae117
https://doi.org/10.1093/nargab/lqae117
Advance access publication date: 24 September 2024
Standard Article

Effective requesting method to detect fusion transcripts in chronic myelomonocytic leukemia RNA-seq

Florence Rufflé[1], Jérôme Reboul[1], Anthony Boureux[1], Benoit Guibert[1], Chloé Bessière[1,2], Raissa Silva[1], Eric Jourdan[3], Jean-Baptiste Gaillard[4], Anne Boland[5], Jean-François Deleuze[5], Catherine Sénamaud-Beaufort[6], Dorothée Selimoglu-Buet[7], Eric Solary[7], Nicolas Gilbert[1,*,†] and Thérèse Commes[1,*,†]

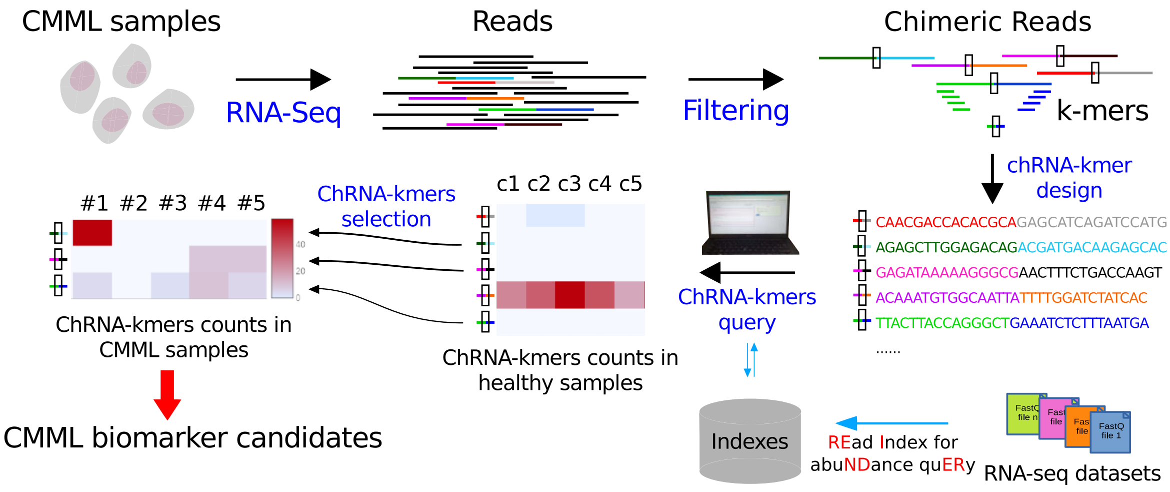
[1]IRMB, University of Montpellier, INSERM, 80 rue Augustin Fliche, 34295 Montpellier, France
[2]CRCT, Inserm, CNRS, University Toulouse III-Paul Sabatier, 31100 Toulouse, France
[3]Department of Hematology, Nîmes University Hospital, 30900 Nîmes, France
[4]Department of Molecular Genetics and Cytogenomics, Montpellier university hospital, 34295 Montpellier, France
[5]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057 Evry, France
[6]GenomiqueENS, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[7]Department of Hematology, Gustave Roussy Cancer Center, Université Paris-Saclay, 94805 Villejuif, France

[*]To whom correspondence should be addressed. Tel: +33 4 67 33 94 88; Fax: +33 4 67 330 113; Email: nicolas.gilbert@inserm.fr
Correspondence may also be addressed to Thérèse Commes. Email: Therese.Commes@inserm.fr
[†]The last two authors should be considered as Joint Last Authors.

Abstract

RNA sequencing technology combining short read and long read analysis can be used to detect chimeric RNAs in malignant cells. Here, we propose an integrated approach that uses k-mers to analyze indexed datasets. This approach is used to identify chimeric RNA in chronic myelomonocytic leukemia (CMML) cells, a myeloid malignancy that associates features of myelodysplastic and myeloproliferative neoplasms. In virtually every CMML patient, new generation sequencing identifies one or several somatic driver mutations, typically affecting epigenetic, splicing and signaling genes. In contrast, cytogenetic aberrations are currently detected in only one third of the cases. Nevertheless, chromosomal abnormalities contribute to patient stratification, some of them being associated with higher risk of poor outcome, e.g. through transformation into acute myeloid leukemia (AML). Our approach selects four chimeric RNAs that have been detected and validated in CMML cells. We further focus on *NRIP1-MIR99AHG*, as this fusion has also recently been detected in AML cells. We show that this fusion encodes three isoforms, including a novel one. Further studies will decipher the biological significance of such a fusion and its potential to improve disease stratification. Taken together, this report demonstrates the ability of a large-scale approach to detect chimeric RNAs in cancer cells.

Graphical abstract


CMML samples
RNA-Seq
Reads
Filtering
Chimeric Reads
k-mers
chRNA-kmer design
CAACGACCACACGCAGAGCATCAGATCCATG
AGAGCTTGGAGACAGACGATGACAAGAGCAC
GAGATAAAAAGGGCGAACTTTCTGACCAAGT
ACAAATGTGGCAATTATTTTGGATCTATCAC
TTACTTACCAGGGCTGAAATCTCTTTAATGA
......
FastQ file n
FastQ file
FastQ file
FastQ file 1
RNA-seq datasets
REad Index for abuNDance quERy
Indexes
ChRNA-kmers query
c1 c2 c3 c4 c5
ChRNA-kmers counts in healthy samples
ChRNA-kmers selection
#1 #2 #3 #4 #5
40
20
0
ChRNA-kmers counts in CMML samples
CMML biomarker candidates

Received: January 29, 2024. **Revised:** August 4, 2024. **Editorial Decision:** August 6, 2024. **Accepted:** August 28, 2024

Introduction

New sequencing technologies offer opportunities to refine the detection of fusion transcripts that cannot be detected with classical cytogenetic approaches (1–3). Difficulties in identifying new chimeric RNAs (chRNAs) are mostly due to false positive predictions as a consequence of technical and algorithmic artifacts. Such chRNAs are transcripts generated by either two distinct chromosomes or a single chromosome through complex biogenesis processes induced by either intra or inter chromosomal rearrangement, transcript read-through, exonization, trans-splicing or other mechanisms (4–6). Therefore, given this complexity, it is always challenging to characterize chimeric transcripts and the identification process requires multiple steps to extract 'raw' predictions and then filter the one that would be of specific interest.

The detection of chRNAs can be of great interest in chronic myelomonocytic leukemia (CMML), a myeloid malignancy that occurs most frequently in the elderly, with a mean age at diagnosis of about 72 years, and more frequently in males. The hallmark of the disease is an increased number of peripheral blood monocytes, resulting in both absolute (monocyte count $\geq 0.5\ 10^9$/L) and relative (monocyte count $\geq 10\%$ of white blood cell count) monocytosis. The recently revised World Health Organization (WHO) classification (7), distinguishes between dysplastic and proliferative CMML subtypes, with a white blood cell count cut-off value at 13 10^9/L. This classification also separates CMML1 from CMML2, based on bone marrow and peripheral blood blast cell counts. Monocytes that accumulate in the peripheral blood are mostly classical monocytes that express CD14 but not CD16 (8–10). Bone marrow cytological examination typically identifies dysplastic features that affects one or more lineages. The diagnostic workup of CMML must exclude other myeloid malignancies, including chronic myeloid leukemia (by demonstrating the absence of the *BCR::ABL1* fusion gene in leukemic cells), other classical myeloproliferative neoplasms, myeloid or lymphoid neoplasms with a fusion gene involving a tyrosine kinase gene (such as *FIP1L1::PDGFRA*), and an acute myeloid leukemia (AML) (7). Targeted gene sequencing identifies one or more somatically acquired gene mutations that typically affect genes involved in epigenetic regulation (such as TET2 and ASXL1), pre-mRNA splicing (such as SRSF2, SF3B1 and UAF1) and intracellular signaling (mostly in the RAS pathway, sometimes in JAK2) (11). The ASXL1 mutation, present in about 40% of patients, is typically associated with a poor outcome and an increased risk of AML transformation (12–14).

CMML can be a severe disease, with median survival ranging from 2 to 3 years depending on the series, with a large heterogeneity between patients. Multiple scores have been generated to stratify patients at diagnosis. Conventional cytogenetics and Fluorescence *in situ* Hybridization (FISH) analyses identify chromosomal abnormalities in approximately one third of the patients, and the nature of these aberrations also correlates with disease outcome and has therefore been incorporated into prognostic scores (14). However, FISH detection of a gene fusion requires prior knowledge of chromosomal positions affected by the rearrangement. In addition, this approach fails to identify short insertions, deletions and small tandem repeat variations. These later events can be detected by polymerization chain reaction (PCR) only when involved genes are well-identified targets for rearrangements in hematological malignancies. These limitations can be circumvented with NGS methods, especially with RNA-sequencing (RNA-seq). Indeed, a recent study compared the potential of RNA-seq for detecting clinically relevant fusion genes with standard routine diagnostic methods (karyotyping and molecular diagnostics) and demonstrated that RNA-seq identifies known fusions missed by routine methods (15). They also demonstrated that RNA-seq can yield additional candidates.

Therefore, to continue in this direction and in an attempt to mitigate the above technical limitations, we have developed a multi-step pipeline to detect new fusion transcripts in RNA-seq data and applied this approach to CMML samples. We detected some chRNAs that have been validated as disease biomarkers. Such a chimera detection method, coupled with an efficient and versatile large-scale validation process, might be applicable to any cancer sample.

Materials and methods

CMML samples

Two sets of 10 samples from patients with CMML were used in this study. The first set (CMML1 to CMML10), provided by ES (Gustave Roussy Institute) consisted of 10 samples from bone marrow or peripheral blood separated on Fycoll-Hypaque. CD14+ monocytes were then sorted from PBMCs using magnetic beads and the AutoMacs system as previously described by Merlevede *et al.* (11). The second set (CMML11 to CMML20) includes 10 blood samples provided by EJ (Biological Resource Center CHU-Nimes, France). These 10 samples correspond to 8 different patients, two of whom were sampled on different dates. Peripheral blood mononuclear cells were separated by Ficoll-Hypaque density gradient and stored in RNA Later. Patients gave written informed consent to be included in the study, which was approved by the Ile de France Ethics Committee and the Nîmes CHU Ethics Board. Sample collection was approved by the Ile de France Ethics Committee (DC-2014-2091) and data collection and storage was approved by the CNIL (N°DR-2016-256).

RNA-seq

RNA-seq experiments were performed on the twenty samples described above. The procedure for RNA extraction and quality assessment was performed as previously described in Rufflé *et al.* (16). 5 μg of total RNA was sent to France genomique for sequencing. Total stranded RNA-seq was performed at the Centre National de Recherche en Génomique Humaine (CNRGH, CEA). After full RNA quality control on each sample (quantification in duplicate on a NanoDrop™ 8000 spectrophotometer and RNA6000 Nano LabChip analysis on an Agilent Bioanalyzer), libraries were prepared using Illumina's 'TruSeq Stranded Total RNA Gold' Kit. An input of 1 μg total RNA was used for all samples, and libraries were prepared on an automated platform according to the manufacturer's instructions. Library quality was verified using LabGx (Perkin Elmer) analysis for profile analysis and quantification, and sample libraries were pooled prior to sequencing to achieve the expected sequencing depth (typically 4 samples per lane). Sequencing was performed on an Illumina HiSeq2000 as 100 bp paired-end reads, using Illumina sequencing reagents. Sequence quality parameters were assessed throughout the sequencing run, and standard bioinformatic analysis of the sequencing data was based on the Illumina pipeline to generate a FASTQ file for each sample. FASTQ files generated after RNA-seq sequencing were processed using in-house

CNRGH tools to assess the quality of raw and genomic-aligned nucleotides.

Whole genome sequencing

Whole genome sequencing was performed on the CMML10 sample by the Centre National de Recherche en Génomique Humaine (CNRGH, CEA). After a complete quality control (quantification in duplicate using Quant-IT kits, quality control by migration on agarose gel), genomic DNA (1µg) was used to prepare a library for whole genome sequencing, using the Illumina TruSeq DNA PCR-Free Library Preparation Kit, according to the manufacturer's instructions. After normalization and quality control, qualified libraries were sequenced as 100 bp paired-end reads on an Illumina HiSeq2000 platform (Illumina Inc., CA, USA). A minimum of 3 lanes of the HiSeq2000 flow cell were generated for each sample to achieve an average sequencing depth of 30× for each sample. Sequence quality parameters were assessed throughout the sequencing run, and standard bioinformatic analysis of sequencing data was based on the Illumina pipeline to generate a FASTQ file for each sample. FASTQ files generated after whole genome sequencing were processed using in-house CNRGH tools in order to assess the quality of raw and genomic aligned nucleotides.

RNA-seq from public datasets

The Beat-AML dataset (17) consisting of acute myeloid leukemia bone marrow or peripheral blood mononuclear cells ($n = 474$), healthy CD34 cells ($n = 14$) and healthy bone marrow mononuclear cells BMMNCs ($n = 19$) was used in this study. Five additional datasets retrieved from the following projects were downloaded and added to the healthy group of the Beat-AML project to build a collection of 132 healthy hematopoietic cells : (i) 41 samples from LEUCEGENE (GSE48846 and GSE51984), (ii) 22 samples from GSE117970 (18), (iii) 28 samples from GSE135902 (19) and 8 pooled samples from different lineages (CRCT Toulouse, see sample details in Supplementary Tables S1–S3). RNA-seq data from the GTEx project ($n = 1023$) were used as healthy tissue controls (20).

RNA-seq analysis

Quality

The Quality of the Hiseq2000 sorting reads for the twenty CMML samples was assessed using fastQC and multiQC and is reported in Supplementary Table S4. An additional quality step was performed with KmerExploR (https://github.com/Transipedia/kmerexplor), an application of the Kmerator Suite (21), to assess the level of contamination with other species in the sequencing fastq results (Supplementary Figure S1).

Mapping

Reads were aligned to the reference genome (human GRCh38) using the CRAC software (VN:2.5.0) with the following command line {crac –no-ambiguity –stranded –detailed-sam -i /data/indexes/crac/GRCh38 -k 22 –bam –nb-tags-infos-stored 10000 –nb-threads 15 -r fastq/*.fastq.gz} (22). The CRAC analysis software is based on k-mer decomposition of reads prior to alignment. Reads are then affected depending on their location on the human reference genome. The statistics from the CRAC summary show the number of reads and their repartition for each sample. The majority is represented by single location reads (around 85%), as is common for this type of dataset and consistent with our subsequent questions (Supplementary Figure S2 and Supplementary Table S5).

ChRNAs annotation, classification

Annotations, classifications and metrics used in the filtering process were provided by CracTools-chimCT (https://github.com/Bio2M/cractools-chimct). The CRAC software generates BAM files with read information to detect mutations, splice or chimeric junctions (22). The BAM files are then submitted to the CracTools core (V 1.251), which allows the separation of biological events and the collection of annotation information through a multi-step snakemake process. Within CracTools, a ChimCT module supported by a GFF file (Ensembl genome browser) is used to classify and annotate the detected chimeric RNAs. Another alignment tool, GSNAP, is used by ChimCT to identify splicing events that should be distinguished from chRNAs. For each sample, ChimCT (V0.14) returns a tsv file with chimeric description information. The file lists all predicted chRNAs organized into four classes, and each chRNA is associated with a ChimValue, which corresponds to a confidence value that takes into account mapping information from CRAC and other analyzers such as annotation, GSNAP alignment tool, fusion distance, and others (https://github.com/Bio2M/cractools-chimct/blob/master/bin/chimCT). The ChimValue (columns i in Supplementary Tables S7–S10), depends on methodological parameters including the mapping score quality and the number of reads from the predicted chRNA, also considering the support of Spanning Reads (SR), i.e. the read containing the chimeric junction, the warning flags given by the annotation (pseudogenes, anchored reads, superfamily genes…) and the spanning paired-end information (SPE). The classification of chRNAs and features provided by ChimCT have been previously described by Ruffle *et al.* and Bouge *et al.* (16,23). The tsv files are aggregated into one large file to merge all the chRNAs found in the 20 samples. This file is then divided into the four classes of chRNAs (Supplementary Figure S3).

Filtering process

The best (highest) ChimValue is 100, which meets all the selection criteria described above. From this value, penalties are applied if some criteria are not fulfilled (see documentation provided at https://github.com/Bio2M/cractools-chimct/blob/master/bin/chimCT for details). The ChimCT value, as well as the information provided by CracTools-chimCT, was used to perform the first filter to select chRNAs (Standard Filters Figure 1). The parameters were chosen to minimize false positive selection as previously described in Bouge et al. (23). Each chRNA class has its specific parameters listed below and chRNAs were retained if:

- Class 1 : ChimValue > 60; SR ≥ 3; without annotations of pseudogene or strange paired end (PE) support *.

* means that there are fewer reads on both sides of the junction than reads covering the junction.

- Class 2: ChimValue $= 100$; SR ≥ 4 and (spanning paired end SPE ≥ 4); no NA—NA and HLA in chRNAs gene names. (NA; Not Annotated)

- Class 3: ChimValue ≥ 70; no NA—NA and HLA in chRNAs gene names. These class 3 chRNAs are then divided

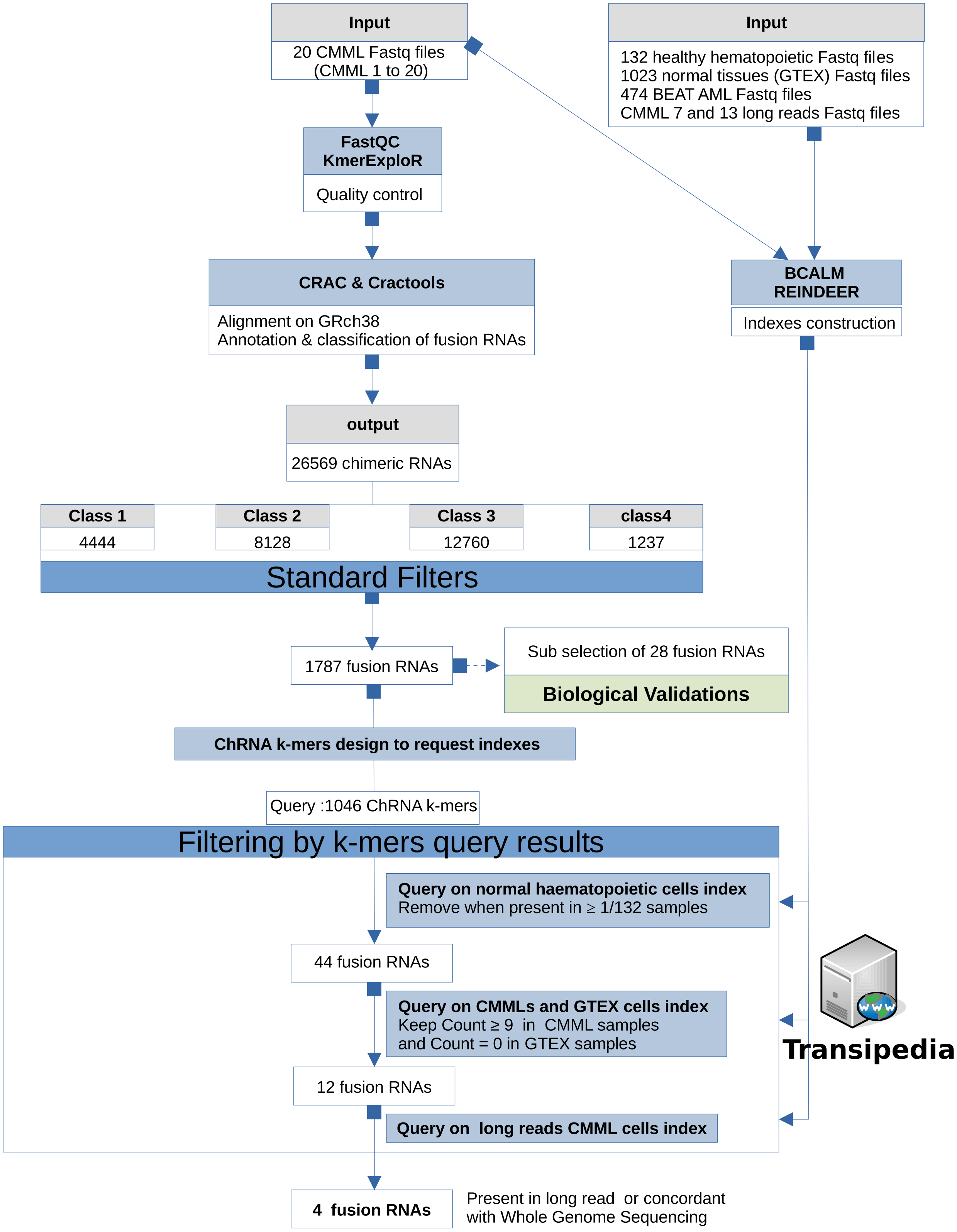

Input
20 CMML Fastq files
(CMML 1 to 20)
Input
132 healthy hematopoietic Fastq files
1023 normal tissues (GTEX) Fastq files
474 BEAT AML Fastq files
CMML 7 and 13 long reads Fastq files
FastQC
KmerExploR
Quality control
CRAC & Cractools
Alignment on GRch38
Annotation & classification of fusion RNAs
BCALM
REINDEER
Indexes construction
output
26569 chimeric RNAs
Class 1
4444
Class 2
8128
Class 3
12760
class4
1237
Standard Filters
1787 fusion RNAs
Sub selection of 28 fusion RNAs
Biological Validations
ChRNA k-mers design to request indexes
Query :1046 ChRNA k-mers
Filtering by k-mers query results
Query on normal haematopoietic cells index
Remove when present in ≥ 1/132 samples
44 fusion RNAs
Query on CMMLs and GTEX cells index
Keep Count ≥ 9 in CMML samples
and Count = 0 in GTEX samples
12 fusion RNAs
Query on long reads CMML cells index
Transipedia
www
4 fusion RNAs
Present in long read or concordant
with Whole Genome Sequencing

Figure 1. Description of the different steps leading to chRNAs selection.

into 3 subgroups, group with chRNAs without specific annotations SR ≥ 3, group with fusionDistance = Overlap* SR ≥ 6, group with short fusionDistance SR ≥ 3; without comment anchored or low support.

* means that a 5′ segment of the read alignment overlaps a 3′ segment, see example below.

Example read TRIM28—TRIM28

Read_seq = TTGTTATCTCTAGAAGCTAGAAGAAA GGGATGTGTTTCTCAGCTATGTTG∗AGAAGCTAGAA GAAAGGGATGTGTTTCTCAGCTATGTTGGGGCAGAG GATT

START END QSIZE IDENTITY CHROM STRAND START END

1 50 50 100.0% chr19 + 58547136 58547185

1 51 51 100.0% chr19 + 58547147 58547197

- Class 4: ChimValue ≥ 70; SR ≥ 2 and SPE ≥ 2; no HLA in chRNAs gene names; no annotation strange PE support.

Finally, a chRNA that does not meet the selection criteria can be recovered if a similar chRNA with the same gene names in another sample as reach the filtering step. The lists of all selected chimeric RNAs can be found in Supplementary Tables S7–S10.

Experimental validations and long read sequencing

Reverse transcription and real-time PCR validation were performed as previously described in Ruffle *et al.* (16). Primers are listed in Supplementary Table S6. CMML7 and CMML13 RNA samples were selected for long read sequencing experiments based on the high quality of the RNAs (RIN > 7). The required sequencing depth was set at 6 Gbp to provide appropriate conditions for comparing the fusion RNA detection capabilities with short-read RNA-seq. Library preparation and nanopore sequencing were performed at the core facility of the Ecole Normale Supérieure Genomique ENS (Paris, France). 10 ng of total RNA was amplified and converted to cDNA using the SMART-Seq v4 Ultra Low Input RNA kit (Clontech). An average of 9 fmol of amplified cDNA was then used for library preparation using the SQK-PBK004 kit (PCR Barcoding kit; ONT). After ligation of the PCR adapter, a 0,6X Agencourt Ampure XP beads clean-up was performed and 2 fmol of purified product was added to the PCR and barcode for an additional amplification of 18 cycle with 17 min elongation time. Sequencing was performed using the SQK-PBK004 72-hour sequencing protocol run on the MinION MkIC for each sample, using the MinKNOW software (version 20.06.15) and the FLO-MIN106 flowcell. 2.3 million reads passing the ONT quality filter were obtained for each of the two samples. Base-calling of read event data was performed using Guppy (v4.0.11). All barcoded fastq files from guppy output that passed the implemented filters were concatenated in one fastq file.

Indexes construction and request

Dataset indexes were built using REINDEER (https://github.com/kamimrcht/REINDEER) (version 1.02) and loaded onto the server to be queried via the transipedia web interface (https://transipedia.fr/). K-mers whith an abundance of <2 were excluded from the index construction. The queried chRNA-kmers were then submitted to Transipedia, which returns a count for each of them in all the samples making up the index, and displays a heatmap.

Results

Global chRNAs analyses in CMMLs

We analyzed RNA-seq data from 20 CMML samples (Supplementary Table S5) to identify chRNAs through a workflow that is summarized on Figure 1. ChRNA events were predicted using CRAC and annotated by the CracTools-chimCT module. Each chRNA was assigned to one of the four classes described previously, depending on the chromosomal and exonic organization of the chimeric read (see Material and Methods: RNA-sequencing analysis section and Supplementary Figure S3). It was observed that the number of chRNAs obtained per sample depended on the sequencing depth, as expressed by the read count (Figure 2A). Their distribution according to structure showed that class 2 and 3 chRNAs were the main categories, accounting for nearly 80% of the predicted chRNAs (Figure 2A). To improve chRNA prediction and reduce the effect of background noise due to ambiguous read alignments, we implemented standard filters based on chimeric read count and ChimValue, which take into account quality score, mapping profile and read coverage. These filters were applied to each detected chRNA ID (see Materials and methods : chimeric RNA filtering process). Thus, from the 26 569 chimeric RNAs found in the first steps, we selected a total number of 1787 chRNAs (Figures 1, 2B). Of these, 91 are chimeras involving genes on different chromosomes (Class 1, Supplementary Table S7). 806 belong to class 2 chimeras, involving genes on the same chromosome strand. They correspond to splicing events of readthrough transcripts that join two adjacent genes in a single RNA molecule (Supplementary Table S8). 821 chRNAs belong to class 3 chimeras and involve either a single gene for which we observe an inversion in the order of sequence segments, or an exon inversion of two adjacent genes or an unannotated region (Supplementary Table S9). Finally, class 4 chRNAs correspond to splicing junctions between exons of genes located on opposite strands of the same chromosome (Supplementary Table S10).

qPCR validations

In order to assess the accuracy of CRAC reinforced by the implemented filters to appropriately detect chRNAs, we performed complementary molecular biology methods to validate the biological existence of these fusion transcripts. Validations were performed by qPCR followed by qPCR product sequencing on a subset of 28 chRNAs (Figure 3 and Supplementary Table S10bis). The selection of this subset was based on the high ChimValue of the chimeras, and was distributed as follows: 1 from class 1, 10 from class 2, 11 from class 3 and 6 from class 4. 24 of these 28 chRNAs were identified by RNA-seq in at least three samples, two (*ACIN1::ACIN1* and *PLIN3::ARRDC5*) in two samples and two (*ZEB2::BCL2L11* and *NRIP1::MIR99AHG*) in a single sample. qPCR and Sanger sequencing validation was observed for $\frac{3}{4}$ of the chRNAs (21/28, Figure 3). Within the classes 1, 2 and 4, only one chRNA was not validated by Sanger sequencing (*CIITA::RP11-876N24.3*). However, it has recently been annotated as an alternative transcript of *CIITA-217* (ENST00000644232.1). Finally, 6 out of 11 class 3 chRNAs couldn't be validated by Sanger sequencing of the qPCR product. This poor efficiency, is most likely related to the structure of such chimeras. They consist of short fragment repetitions and, as primers recognize both the normal and the

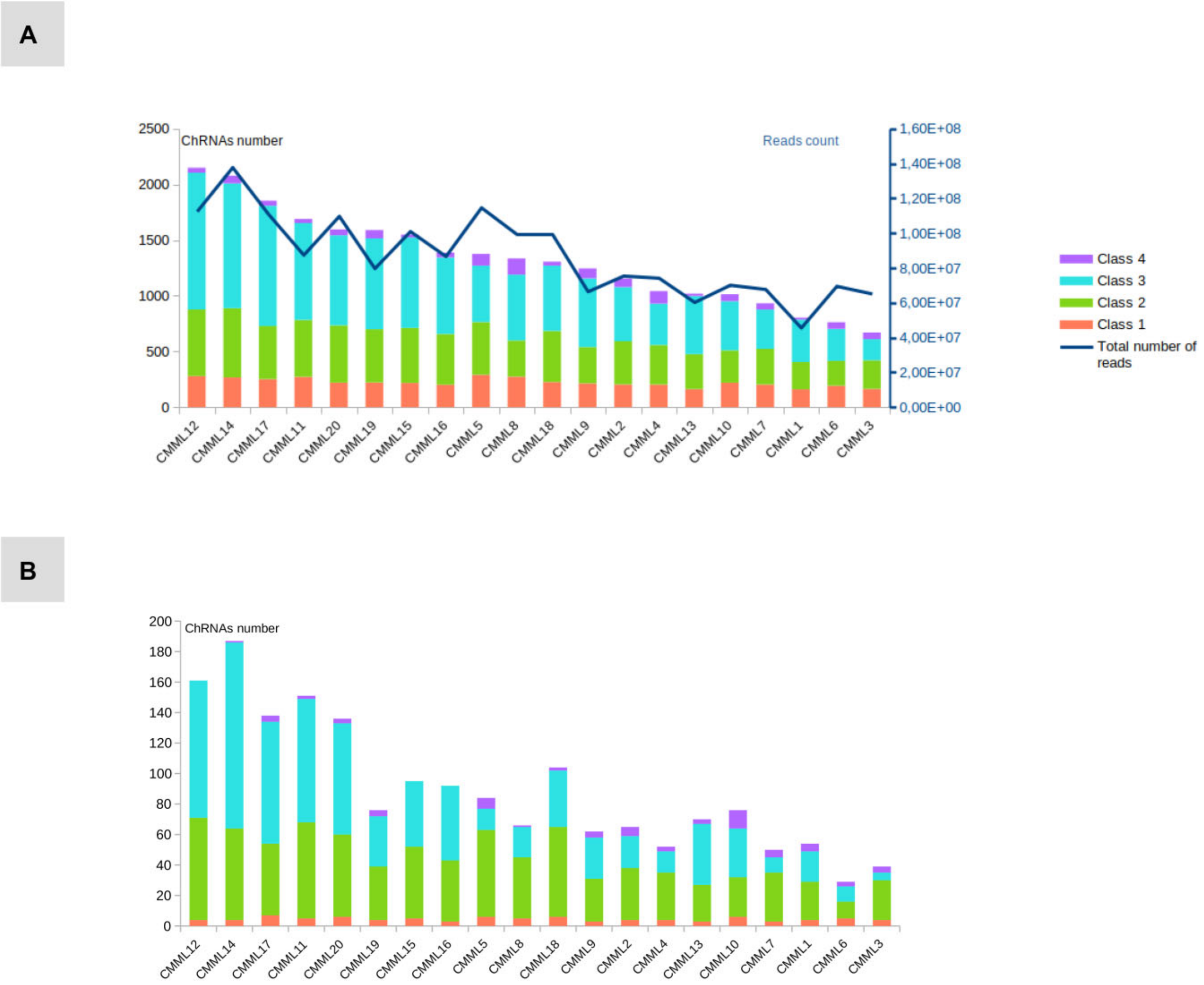


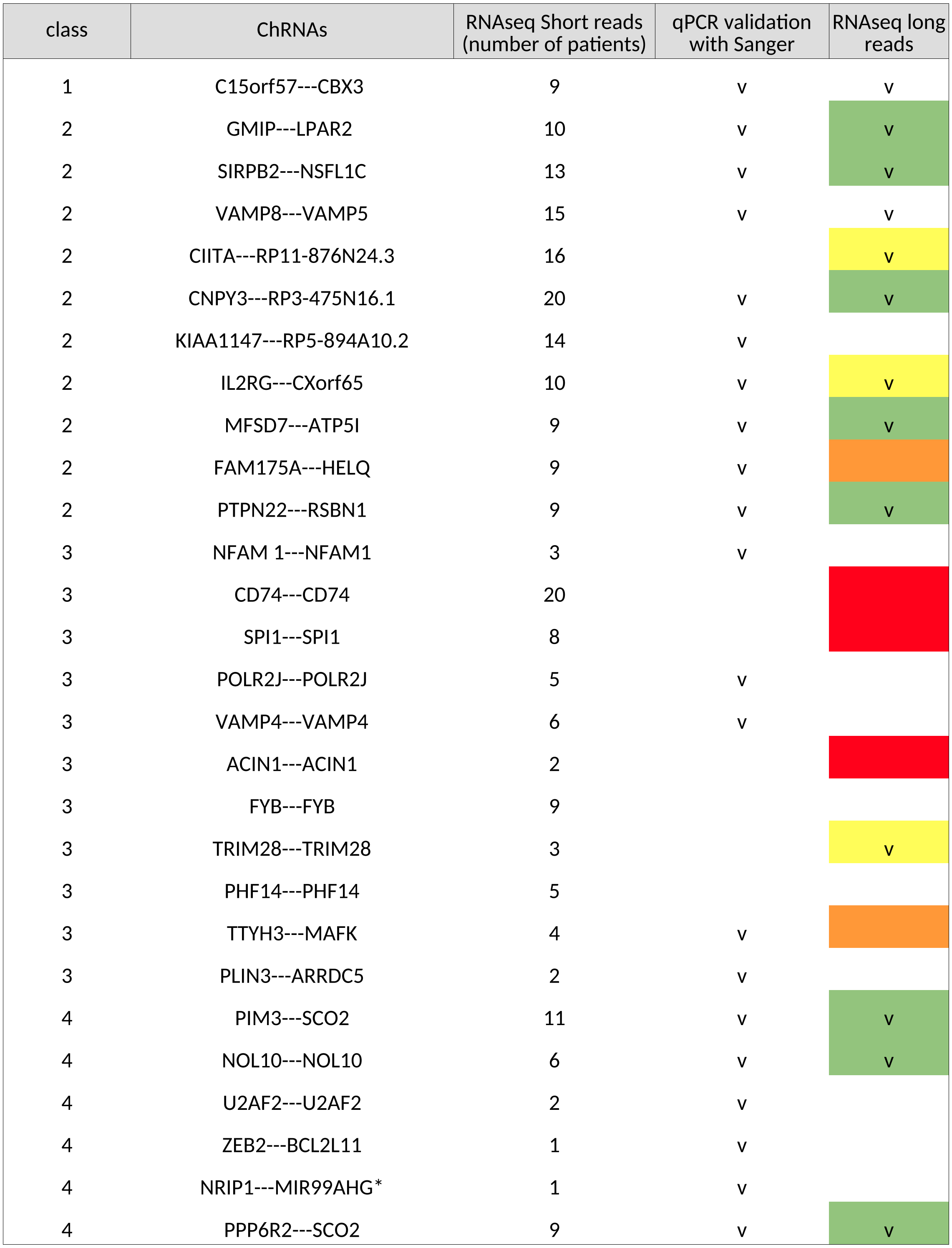
Figure 2. ChRNAs distribution among the 20 CMML samples. (**A**) Repartition of chRNAs classes and read count per sample. (**B**) Repartition of chRNAs classes per sample after applying standard filters.

chimeric transcript, the amplification may favor shorter and more abundant targets. Consequently, such chimeras need to be considered with caution and further validated. We therefore decided to perform long-read sequencing to better characterize them (see below). Nevertheless, we validated the biological presence in samples of most of selected chimeras and, by extension, the biological presence of the 1787 fusion transcripts initially filtered, without presuming tissue or tumor specificity. Indeed, many fusion transcripts are known to be present in normal cells (24), and the next challenge is to identify those that are tumor specific.

Long read sequencing

To further confirm the presence of the predicted chRNAs, we sequenced two samples (CMML7 and CMML13, in which 16 of the 28 chRNAs subset were detected by short read RNA-seq) using Oxford Nanopore long read sequencing technology (ONT). Both samples have, among others, class 3 chRNAs that were identified by short read sequencing but not validated by qPCR (2 and 5 respectively). To detect rare chRNA events, a coverage comparable to that originally obtained with the Illumina technology was used (Supplementary Table S5). As the Minion technology generates sequencing errors, we built 20nt specific sequences (k-mers with $k = 20$) covering chimeric junctions and searched each fastq file for an exact match. This length was optimized to maintain specificity and overcome the higher sequencing error compared to short reads (25). The reverse complement of each k-mer was also searched as sorted sequencing reads were not stranded. Matching sequences were retrieved and verified by sequence homology search against the reference genome in order to assess the origin of the two separate segments.

A summary of the results is presented in Figure 3 and Table 1, and the validated long read sequences are listed in Supplementary Table S11. Class 3 *CD74::CD74*, *SPI1::SPI1* and *ACIN1::ACIN1* chRNAs, which were found in at least one of the two patients by short read RNA-seq but not validated by sequencing of the qPCR product, were also not validated by long read sequencing either. The lack of detection of long reads for *FAM175A::HELQ* and *TTYH3::MAFK*, which had been validated by qPCR and product sequencing (Supplementary Table S12), could be explained by either the low expression of the chimeric transcript or the high error rate of long read sequencing, which would confound the search for perfect matches with the k-mer. In contrast, 11/16 chRNAs



class	ChRNAs	RNAseq Short reads (number of patients)	qPCR validation with Sanger	RNAseq long reads
1	C15orf57---CBX3	9	v	v
2	GMIP---LPAR2	10	v	🟩 v
2	SIRPB2---NSFL1C	13	v	🟩 v
2	VAMP8---VAMP5	15	v	v
2	CIITA---RP11-876N24.3	16		🟨 v
2	CNPY3---RP3-475N16.1	20	v	🟩 v
2	KIAA1147---RP5-894A10.2	14	v	
2	IL2RG---CXorf65	10	v	🟨 v
2	MFSD7---ATP5I	9	v	🟩 v
2	FAM175A---HELQ	9	v	🟧
2	PTPN22---RSBN1	9	v	🟩 v
3	NFAM 1---NFAM1	3	v	
3	CD74---CD74	20		🟥
3	SPI1---SPI1	8		🟥
3	POLR2J---POLR2J	5	v	
3	VAMP4---VAMP4	6	v	
3	ACIN1---ACIN1	2		🟥
3	FYB---FYB	9		
3	TRIM28---TRIM28	3		🟨 v
3	PHF14---PHF14	5		
3	TTYH3---MAFK	4	v	🟧
3	PLIN3---ARRDC5	2	v	
4	PIM3---SCO2	11	v	🟩 v
4	NOL10---NOL10	6	v	🟩 v
4	U2AF2---U2AF2	2	v	
4	ZEB2---BCL2L11	1	v	
4	NRIP1---MIR99AHG*	1	v	
4	PPP6R2---SCO2	9	v	🟩 v

Figure 3. Overview of chRNAs biological validations. A subset of 28 chRNAs (column 2), from different classes (column 1) found by RNA-seq short reads in CMML samples (column 3) were subjected to qPCR followed by Sanger sequencing for biological validation (column 4). RNA-seq long read sequencing (column 5) was performed on 2 samples (CMML 7 and 13) to validate 16 chRNAs (highlighted by color) identified by short read sequencing in the corresponding samples. v; detected. Green; detected by both short and long reads. Yellow; detected by short reads in one sample and confirmed by long reads in the other sample. Orange; validated by qPCR and Sanger sequencing but not by long read. Red; not validated by qPCR and Sanger or long read. No color with v; not detected by short reads but detected by long reads.

Table 1. State of chRNA detection in CMML7 and CMML13 (chRNA classes are indicated in brackets)

Found in short and long reads	Found only in short reads
C15orf57—CBX3 (1)	FAM175—HELQ (2)
GMIP—LPAR2 (2)	CD74—CD74 (3)
IL2RG—CXorf65 (2)	SPI1—SPI1 (3)
MFSD7—ATP5I (2)	ACIN1—ACIN1 (3)
PTPN22—RSBN1 (2)	TTYH3—MAFK (3)
CNPY3—RP3-475N16.1 (2)	
CIITA—RP11-876N24.3 (2)	
VAMP8—VAMP5 (2)	
SIRPB2—NSFL1C (2)	
TRIM28—TRIM28 (3)	
PIM3—SCO2 (4)	
NOL10—NOL10 (4)	
PPP6R2—SCO2 (4)	

were found in both short and long reads, validating their biological expression in CMML patient samples. Interestingly, the class 3 chimera *TRIM28::TRIM28* which was neither validated by qPCR product sequencing nor found by ONT in CMML13, was detected in CMML7 with ONT without being detected with CRAC and CracTools-chimCT in RNA-seq short reads. In addition, long read sequencing is sensitive enough to reveal the presence of numerous chimeras that were not detected by short-read sequencing, for example 5 out of 10 and 2 out of 6 chimeras detected by long read sequencing were not detected by short read sequencing in CMML7 and 13 samples, respectively (Supplementary Table S10bis). Overall, the fusion transcripts identified by our selection process can be used for validation as potential RNA biomarkers.

Large scale exploration of chRNA expression with k-mers

In order to determine the recurrence as well as the tumor specificity of all chRNAs detected in the first instance, we used a counting procedure with specific k-mers representing the chRNAs, called 'chRNA-kmers'. The 1787 chRNAs from the 20 samples correspond to 1046 different chRNAs and therefore to 1046 chRNA-kmers. In this strategy, the chRNA-kmer design consisted of extracting a sequence of length $k = 31$nt centered on the chimeric junction of the reads given by the CRAC mapper. Then, for the exploration procedure, in a first step, we indexed all the k-mers from raw fastq files of selected RNA-seq datasets and determined their abundances using the reindeer data-structure (26). In a second step, chRNA-kmers were queried in these indexes to obtain their counts per sample using the transipedia web interface (https://transipedia.org/).

To begin with, we investigated the indexes built with 132 samples of normal hematopoietic cells (CD34, PBMCs, CD14 and normal hematopoietic progenitors). Of the 1046 chRNAs-kmers, we retained only those that were found at most once in the corresponding samples and excluded all the others. 44 chRNA-kmers were selected by this filtering and used for the following step (Figure 1).

We next examined the indexes constructed with our 20 CMML samples and selected from the 44 chRNA-kmer only those with a count of at least 9 in a CMML sample, as well as those that were weakly expressed but observed with ONT in CMML7 and 13 samples (Supplementary Table S13). Eighteen chRNA-kmers met these criteria. To be even more stringent in our selection, we also analyzed a subset of 1023 normal samples from 24 different tissue types from the Genotype-Tissue Expression (GTEX) project to determine whether these 18 chRNA-kmers were also present in healthy non-haematopoietic tissues and therefore also needed to be removed. We obtained a final list of 12 chRNA-kmers (Figure 1 and 4A).

We next found interesting to query these 12 chRNA-kmers against the indexes of the Biomarker-Based-Treatment of Acute Myeloid Leukemia (Beat AML) cohort, a closely related myeloid malignant haemopathy, as they may share common markers. Among them, 2 were not found and 6 had chRNA-kmer counts <10 per sample (Figure 4B). Despite these low counts, the chRNA-kmers GSE1-KLHL36_38, GSE1-KLHL36_40 and YWHAZ-AZIN1 were found present in 54, 63 and 29 out of 474 samples, respectively. The ANXA6-TNIP1 chRNA-kmer was detected in 57/474 samples with a low count. In contrast, NRIP1-MIR99AHG$_3$, which identify a class 4 chRNA *NRIP1::MIR99AHG*, was present in only 5/474 samples but with an expression comparable to the well-known *CBFB::MYH11* chimera (mean expression value of 21.5 for *NRIP1::MIR99AHG* and 13.5 for *CBFB::MYH11* Supplementary Figure S4). Notably, these 5 AML samples are associated with myelodysplasia related changes (Supplementary Figure S5 and Supplementary Table S15), which is consistent with its detection in the CMML sample.

Characterization of selected chRNAs and cross validation with long read sequencing

Based on the 12 chRNA-kmers selected above, we retrieved the corresponding long reads, if present, in the fastq files of CMML7 and CMML13. Three identified long reads were submitted to Local Basic Alignment Tool (BLAT from UCSC, (27)) for full length RNA characterization. As expected, *GSE1::KLHL36* and *YWHAZ::AZIN1* were detected in CMML7 and CMML13 samples. We noticed that chRNA-kmers GSE1-KLHL36_38 and GSE1-KLHL36_40 recognized the same full length chimeric sequence and correspond to the same chimeric junction (the two chRNA-kmer sequences have an offset of 8 nucleotides).

In *YWHAZ::AZIN1*, the full length RNA starts with the *YWHAZ* gene, then links the splicing acceptor site of *YWHAZ* exon 1 (*YWHAZ*-201 ENST00000353245.7) to the splicing donor site of *AZIN1* exon 2 (*AZIN1*-201 ENST00000337198.10) and continues until the end of the *AZIN1 gene* transcript (Supplementary Table S13 and Figure 5A). Similarly, the *GSE1::KLHL36* chimeric transcript links the splicing acceptor site of exon 1 (pos 85556363) of the *GSE1* gene (ENST00000635906.1) to the splicing donor site of *KLHL36* exon 2 (pos 84650850, ENST00000564996.6) (Figure 5B).

We also identified a long read for *ANXA6::TNIP1* class 2 chRNA in both samples, although its presence was not detected with short read sequencing in the corresponding samples. This chRNA can be described as a readthrough transcript that links the penultimate exon (exon 25, pos 151103570) of the *ANXA6* gene (ENST00000354546.10) with exon 3 (pos 151063747) of the *TNIP1* gene (ENST00000521591.6) (Figure 5C).

Finally, although no *NRIP1::MIR99AHG* chRNA was found in CMML7 and CMML13 samples using either short read or long read sequencing technology, we decided to

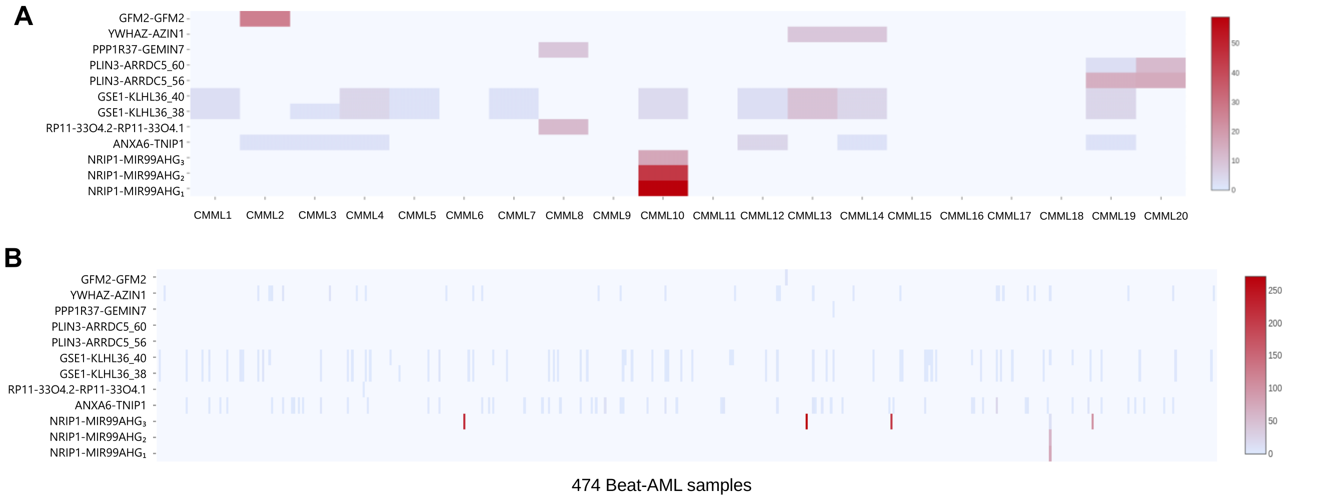


Figure 4. Expression of the 12 selected chRNA-kmers. A- chRNA-kmers count in the 20 CMML cohort. B- chRNA-kmers count in the BEAT-AML cohort. X axis correspond to 474 samples in the beatAML cohort, IDs were removed because of lack of readability (corresponding table available in Supplementary Table S14).

focus on this chRNA again. One of the main reasons is that this fusion transcript showed a similar level of expression in CMML 10 as the *CBFB::MYH11* fusion transcript, a class 4 chimera detected and used as a biomarker in acute myeloid leukemia. In addition, the *NRIP1::MIR99AHG* chRNA was recently identified in AML samples by Kerbs et al. and correlated with genomic rearrangements (15). By analyzing the short RNA-seq data, we pointed out three different junctions in the CMML10 sample. Two of them associate exon 3 of *NRIP1* gene (ENST00000318948.7) with either exon 7 or 8 of *MIR99AHG* LncRNA (ENST00000619222.5). The third associates exon 2 of *NRIP1* with exon 8 of *MIR99AHG* (Figure 6A). Consistent with these observations, short-read genomic sequencing covering the *NRIP1::MIR99AHG* region detected a chromosomal inversion in this region in sample CMML10 (Figure 6B and C). The breakpoints joining intron 6–7 of *MIR99AHG* with intron 3–4 of *NRIP1* could generate transcripts whose alternative splicing would be consistent with the three fusion transcripts observed.

Discussion

The present report uses in-depth analysis of RNA-seq data to analyze transcript diversity and detect transcripts that are not fully accounted for, even by the best reference databases (28). This approach was applied to the detection of novel fusion RNAs or chRNAs in CMML, a disease in which personalized prediction of overall survival and AML transformation remains challenging despite recent improvements in CMML classification (7,29). The identification of novel somatically acquired molecular abnormalities that may contribute to driving disease progression from chronic to acute phase would refine these predictions.

Our strategy can identify transcripts whose appearance in malignant cells is associated with both genomic and/or transcriptional abnormalities. CRAC and CracTools-chimCT enable the detection of 4 classes of chimeric RNAs revealing potential translocations, readthrough, deletions, repeats and inversions, making it an original tool suite (16,23). Nevertheless, it is worth noting that difficulties in chRNAs prediction due to technical or algorithmic false positives make it difficult to consider them as cancer biomarkers from a biological point of view. Therefore, we used common strategies to filter false positives with annotations (pseudogenes, superfamily genes, repeats) and metrics associated with mapping information (average mapping quality, read coverage, split-reads counts supporting spanning junction, supporting paired-end read count or spanning paired-reads) (24,30). This step, based on the information provided by CracTools-chimCT, helps to exclude some false positives. Still, this approach was not sufficient to classify fusion transcripts as cancer biomarkers.

There are several procedures based on negative filters, together with blacklists based on chromosomal positions to select only the most specific cancer-specific fusion transcript candidates. However, such procedures, which carry a risk of information loss (30), require additional positive filters to rescue targets eliminated by stringent filters. We propose an alternative method, based on the k-mer approach to select potential biomarker chimeras.

In a previous study, we demonstrated the performance of specific k-mers for chRNAs discovery and tumor specificity selection (16). Rehn *et al.* recently published RaScALL, a similar approach based on the jellyfish structure (31). Here we provide a significant improvement of the k-mer process based on a fast query to Reindeer indexes constructed from raw fastq files (26), and the potential to query large RNA-seq datasets (up to 1000) via the transipedia web interface.

Our procedure does not require stringency in the initial filters and therefore includes a larger number of candidates without a priori. We preferred to base the stringency of the selection on the tissue specificity of expression. Indeed, by comparing the presence or absence of chimeric k-mers in multiple RNA-seq datasets of interest, we are able to select tumor specific fusion transcripts. Querying for specific chRNA k-mers provides a k-mer count that can be compared across indexed datasets of interest in a quantitative manner as previously described (21,32).

Applying the suite of CRAC tools with basic filters to the CMML dataset enables the identification of 1787 chRNAs in 4 classes. Before proceeding further, we had to validate the biological existence of the chRNAs. Therefore, we selected a

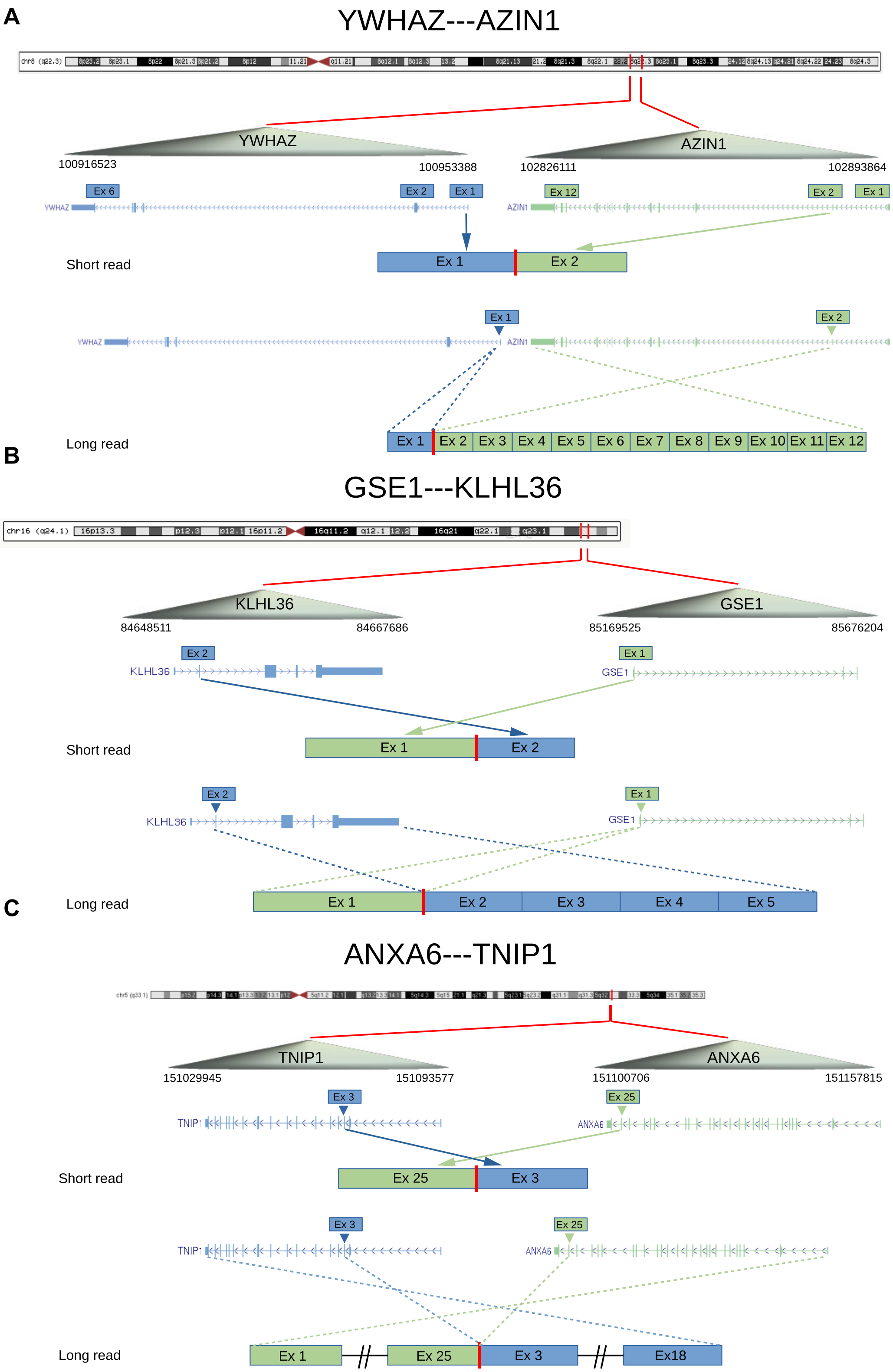


Figure 5. Molecular structure of chRNAs. Representation of the chRNA structure of class 3 *YWHAZ::AZIN1* (**A**) and *GSE1::KLHL36* (**B**), class 2 *ANXA6::TNIP1* (**C**). A graphical representation of the genomic origin of the chimeras is shown at the top of each panel. The short-read sequencing results are shown below (blue and green colors indicate the different gene origins). The long read sequencing results are displayed and compared to the expected transcript of the implicated genes.

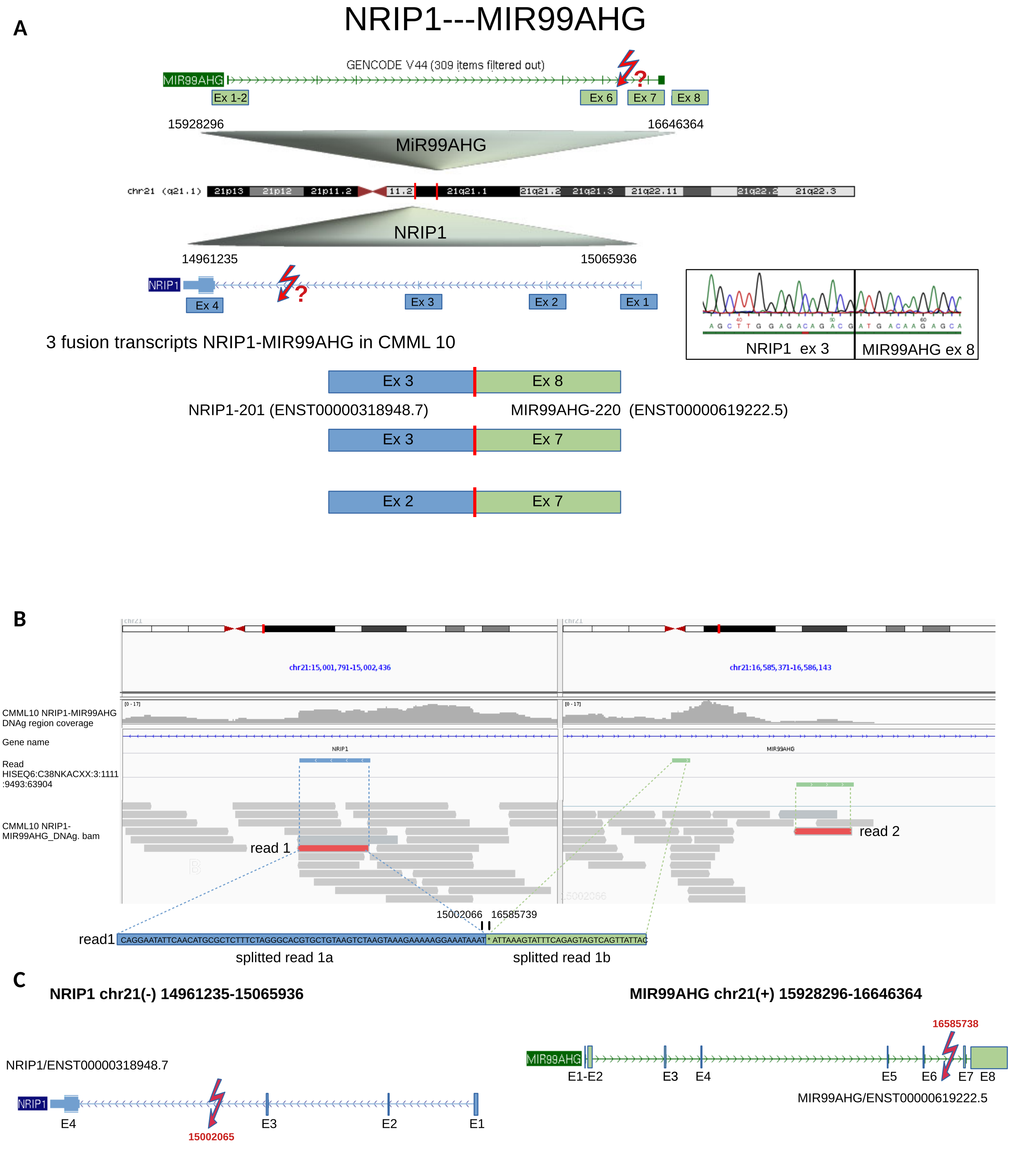

A
NRIP1---MIR99AHG
GENCODE V44 (309 items filtered out)
MIR99AHG
Ex 1-2
Ex 6
Ex 7
Ex 8
?
15928296
16646364
MiR99AHG
chr21 (q21.1)
21p13
21p12
21p11.2
11.2
21q21.1
21q21.2
21q21.3
21q22.11
21q22.2
21q22.3
NRIP1
14961235
15065936
NRIP1
Ex 4
?
Ex 3
Ex 2
Ex 1
NRIP1 ex 3
MIR99AHG ex 8
3 fusion transcripts NRIP1-MIR99AHG in CMML 10
Ex 3
Ex 8
NRIP1-201 (ENST00000318948.7)
MIR99AHG-220 (ENST00000619222.5)
Ex 3
Ex 7
Ex 2
Ex 7
B
chr21:15,001,791-15,002,436
chr21:16,585,371-16,586,143
CMML10 NRIP1-MIR99AHG DNAg region coverage
[0 - 17]
Gene name
NRIP1
MIR99AHG
Read HISEQ6:C38NKACXX:3:1111:9493:63904
CMML10 NRIP1-MIR99AHG_DNAg. bam
read 1
read 2
15002066
16585739
read1
CAGGAATATTCAACATGCGCTCTTTCTAGGGCACGTGCTGTAAGTCTAAGTAAAGAAAAAGGAAATAAAT
* ATTAAAGTATTTCAGAGTAGTCAGTTATTAC
splitted read 1a
splitted read 1b
C
NRIP1 chr21(-) 14961235-15065936
MIR99AHG chr21(+) 15928296-16646364
NRIP1/ENST00000318948.7
NRIP1
E4
15002065
E3
E2
E1
16585738
MIR99AHG
E1-E2
E3
E4
E5
E6
E7
E8
MIR99AHG/ENST00000619222.5

Figure 6. Visualization of NRIP1-MIR99AHG genomic DNA rearrangement. (**A**) Representation of the class 4 *NRIP1::MIR99AHG* chRNA structure. A graphical representation of the genomic origin of the chimeras is shown with the results of the short-read sequencing (blue and green colors indicate the different gene origins). The three observed junctions are shown as well as the qPCR sequencing result for one of them. (**B**) Visualization of the CMML10 *NRIP1*::*MIR99AHG* DNA region with DNAg short read sequencing. Reads are displayed using the Integrative Genomics Viewer (IGV) interface. (**C**) Schematic view of DNA breakpoint based on illumina DNAg sequencing and alignment. Exons are indicated by vertical bars.

subset of 28 chRNAs and tested their presence in CMML RNA samples by qPCR and Sanger sequencing. To further validate the biological presence of chRNAs in CMML, we used Oxford Nanopore long-read sequencing data from two patients with a high coverage. The complementary long read information provides not only additional biological validation compared to qPCR for low expressed chRNA, but also full length sequences of the new candidates. The biological validation of 23/28 chRNAs thus reflected the relevance of our standard filters for the 1787 chRNAs extracted in the first instance.

From the 1787 chRNAs, we generated 1046 chRNA-kmers covering the fusion junctions. They were directly searched in indexed RNA-seq datasets relevant to the biological question. Indeed, to answer the question of the clinical relevance of a new fusion transcript we must first verify the tumor specificity of the biomarker. Here, our filtering strategy allowed us to search for them in normal samples of the haematopoietic lineage with defined criteria to retain only those present in CMML patients (found less than twice in the group of 132 healthy samples and present with a chRNA-kmer count greater than or equal to 9 in at least one sample of the CMML cohort). Another advantage of this k-mer based filtering step is to overcome technological biases that should be equally present in tumor and normal samples. We even added an additional filter by searching in the indexed RNA-seq of the GTEX cohort and excluding chRNAs that would be found (Supplementary Table S13). We ended up with 12 chRNA-kmers corresponding to 11 chRNA candidates.

We did not identify any class 1 chRNA corresponding to translocation, which is not surprising since these clonal cytogenetic events most often characterize AML, CML or lymphoid neoplasms. We detected two class 2 chRNAs corresponding to readthrough: *RP11-3304::1-RP11-3304.2* (now reclassified as alternative transcript of the NHEJ1 gene) found in one CMML and *ANXA6::TNIP1* seen in 6/20 CMMLs. We found 5 class 3 chRNAs, 4 of them are found in one or two CMML samples (*GFM2::GFM2*; *YWHAZ::AZIN1*; *PPP1R37::GEMIN7, PLIN3::ARRDC5)*, the fifth, *GSE1*::*KLHL36*, is present in 50% of the CMML cohort. Finally, we identified 3 class 4 chRNAs representing 3 different isoforms (see below).

Interestingly, the class 4 chRNA *NRIP1::MIR99AHG* corresponding to an inversion in the chr21, already identified in AML (15), is also detected in one CMML patient (CMML10). We detected three different isoforms for *NRIP1::MIR99AHG* in this patient and constructed the corresponding specific chRNA-kmers to search them in the Beat-AML cohort and found 5 positive cases. The predominant isoform in CMML10, also described in the Rjun database, contains the exon 3 of *NRIP1* joined with the exon 8 of *MIR99AHG* and was also found with the chRNA-kmer in 1/5 of the Beat-AML samples. The fusion RNA reported by Kerbs *et al.* (15) in AML patients combines the exon 3 of *NRIP1* and the exon 7 of *MIR99AHG*. This is the only isoform observed in 4/5 AML patients while it's the less expressed in the CMML sample. In addition to these two isoforms, we highlighted a third novel chimeric isoform joining the exon 2 of *NRIP1* and the exon 7 of *MIR99AHG*, which is absent in all AML patients and expressed with an intermediate level in CMML10. Whole genome sequencing analysis of the CMML10 patient reveals two breakpoints positions on chromosome 21, the first in *NRIP1* gene at position 15002065 and the second in the *MIR99AHG* gene at position 16585738. The breakpoint localizations are consistent with the 3 *NRIP1::MIR99AHG* chRNA isoforms and confirm the inversion. In addition, among the most important common patient characteristics, all AML patients carrying the *NRIP1::MIR99AHG* chRNA are diagnosed with AML with myelodysplasia related changes in the beat AML cohort, indicating a potential clinical history of MDS, MDS/MPN, such as CMML, and consistent with transformation towards AML.

Altogether, we demonstrate that k-mers can be used to detect any type of fusion transcript in large datasets. This flexible approach can be applied to any tumor type with RNA-seq data. Here, we detect 11 fusion transcripts in a cohort of 20 CMMLs, 4 of which are validated by long read or whole genome sequencing. Further prospective studies will determine whether *YWHAZ*::*AZIN1*, *GSE1*::*KLHL36* and *ANXA6*::*TNIP1* are passenger or recurrent genetic events directly involved in this disease. This question is regularly raised for chRNAs as well as their clinical relevance with regard to their low expression levels (33,34). Nevertheless, given their tissue specificity, such transcripts are promising biomarkers that could be used for diagnostic, prognostic or minimal residual disease follow-up where tumor markers are needed, especially for patient with a normal karyotype.

Finally, the *NRIP1::MIR99AHG* chRNA deserves special attention in this disease because this fusion transcript detected in a CMML patient has previously been described in AML with dysplastic features, suggesting that this specific fusion could potentially play a role in disease evolution and thus contribute to refining disease stratification.

Data availability

Sequencing data are available on the European Bioinformatics Institute (EBI) website in the arrayexpress repository under project E-MTAB-13763. Transipedia web interface is available at https://transipedia.org. KmerExploR, Crac and CracTools-ChimCT can be downloaded at https://github.com/Transipedia/kmerexplor; https://www.bio2m/crac and https://github.com/Bio2M/cractools-chimct

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We would like to thank the MGX-qPCR HD (DBS-B3ESTE) platform and in particular Mr Philippe Clair for his advice on the use of the Labcyte Echo525 Acoustic Liquid Handler

Funding

France Génomique National infrastructure, funded as part of the 'Investissements d'Avenir' program managed by the Agence Nationale pour la Recherche (ANR) projet HEMASEQ - France Génomique [ANR-10-INSB-09]; le Cancéropôle GrandSud-Ouest (GSO) [2017-EM24]; ANR : Transipedia [ANR-18-CE45-0020], full-RNA [ANR-22-CE45-0007]; R.S. was recipient of a fellowship from La

Ligue Nationale Contre le Cancer; C.B. was supported by La Fondation de France.

Conflict of interest statement

None declared.

References

1. Bacher,U., Shumilov,E., Flach,J., Porret,N., Joncourt,R., Wiedemann,G., Fiedler,M., Novak,U., Amstutz,U. and Pabst,T. (2018) Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. *Blood Cancer J.*, **8**, 113.
2. Arindrarto,W., Borràs,D.M., de Groen,R.A.L., van den Berg,R.R., Locher,I.J., van Diessen,S.A.M.E., van der Holst,R., van der Meijden,E.D., Honders,M.W., de Leeuw,R.H., *et al.* (2021) Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. *Leukemia*, **35**, 47–61.
3. Docking,T.R., Parker,J.D.K., Jädersten,M., Duns,G., Chang,L., Jiang,J., Pilsworth,J.A., Swanson,L.A., Chan,S.K., Chiu,R., *et al.* (2021) A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.*, **12**, 2474.
4. Wang,Y., Zou,Q., Li,F., Zhao,W., Xu,H., Zhang,W., Deng,H. and Yang,X. (2021) Identification of the cross-strand chimeric RNAs generated by fusions of bi-directional transcripts. *Nat. Commun.*, **12**, 4645.
5. Shi,X., Singh,S., Lin,E. and Li,H. (2021) Chapter one - Chimeric RNAs in cancer. In: Makowski,G.S. (ed.) *Advances in Clinical Chemistry*. Elsevier, Vol. **100**, pp. 1–35.
6. Sun,Y. and Li,H. (2022) Chimeric RNAs discovered by RNA sequencing and their roles in cancer and rare genetic diseases. *Genes (Basel)*, **13**, 741.
7. Khoury,J.D., Solary,E., Abla,O., Akkari,Y., Alaggio,R., Apperley,J.F., Bejar,R., Berti,E., Busque,L., Chan,J.K.C., *et al.* (2022) The 5th edition of the World Health Organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *Leukemia*, **36**, 1703–1719.
8. Selimoglu-Buet,D., Wagner-Ballon,O., Saada,V., Bardet,V., Itzykson,R., Bencheikh,L., Morabito,M., Met,E., Debord,C., Benayoun,E., *et al.* (2015) Characteristic repartition of monocyte subsets as a diagnostic signature of chronic myelomonocytic leukemia. *Blood*, **125**, 3618–3626.
9. Jestin,M., Tarfi,S., Duchmann,M., Badaoui,B., Freynet,N., Tran Quang,V., Sloma,I., Droin,N., Morabito,M., Leclerc,M., *et al.* (2021) Prognostic value of monocyte subset distribution in chronic myelomonocytic leukemia: results of a multicenter study. *Leukemia*, **35**, 893–896.
10. Solary,E. and Itzykson,R. (2021) Chronic myelomonocytic leukemia gold jubilee. *Hemato*, **2**, 403–428.
11. Merlevede,J., Droin,N., Qin,T., Meldi,K., Yoshida,K., Morabito,M., Chautard,E., Auboeuf,D., Fenaux,P., Braun,T., *et al.* (2016) Mutation allele burden remains unchanged in chronic myelomonocytic leukaemia responding to hypomethylating agents. *Nat. Commun.*, **7**, 10767.
12. Gelsi-Boyer,V., Trouplin,V., Roquain,J., Adélaïde,J., Carbuccia,N., Esterni,B., Finetti,P., Murati,A., Arnoulet,C., Zerazhi,H., *et al.* (2010) ASXL1 mutation is associated with poor prognosis and acute transformation in chronic myelomonocytic leukaemia. *Br. J. Haematol.*, **151**, 365–375.
13. Itzykson,R., Kosmider,O., Renneville,A., Gelsi-Boyer,V., Meggendorfer,M., Morabito,M., Berthon,C., Adès,L., Fenaux,P., Beyne-Rauzy,O., *et al.* (2013) Prognostic score including gene mutations in chronic myelomonocytic leukemia. *JCO*, **31**, 2428–2436.
14. Patnaik,M.M. and Tefferi,A. (2016) Cytogenetic and molecular abnormalities in chronic myelomonocytic leukemia. *Blood Cancer J.*, **6**, e393.
15. Kerbs,P., Vosberg,S., Krebs,S., Graf,A., Blum,H., Swoboda,A., Batcha,A.M.N., Mansmann,U., Metzler,D., Heckman,C.A., *et al.* (2022) Fusion gene detection by RNA-sequencing complements diagnostics of acute myeloid leukemia and identifies recurring *NRIP1-MIR99AHG* rearrangements. *Haematologica*, **107**, 100–111.
16. Rufflé,F., Audoux,J., Boureux,A., Beaumeunier,S., Gaillard,J.-B., Bou Samra,E., Megarbane,A., Cassinat,B., Chomienne,C., Alves,R., *et al.* (2017) New chimeric RNAs in acute myeloid leukemia. *F1000Res*, **6**, 1302.
17. Tyner,J.W., Tognon,C.E., Bottomly,D., Wilmot,B., Kurtz,S.E., Savage,S.L., Long,N., Schultz,A.R., Traer,E., Abel,M., *et al.* (2018) Functional genomic landscape of acute myeloid leukaemia. *Nature*, **562**, 526–531.
18. Cassetta,L., Fragkogianni,S., Sims,A.H., Swierczak,A., Forrester,L.M., Zhang,H., Soong,D.Y.H., Cotechini,T., Anur,P., Lin,E.Y., *et al.* (2019) Human tumor-associated macrophage and monocyte transcriptional landscapes reveal cancer-specific reprogramming, biomarkers, and therapeutic targets. *Cancer Cell*, **35**, 588–602.
19. Franzini,A., Pomicter,A.D., Yan,D., Khorashad,J.S., Tantravahi,S.K., Than,H., Ahmann,J.M., O'Hare,T. and Deininger,M.W. (2019) The transcriptome of CMML monocytes is highly inflammatory and reflects leukemia-specific and age-related alterations. *Blood Adv.*, **3**, 2949–2961.
20. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
21. Riquier,S., Bessiere,C., Guibert,B., Bouge,A.-L., Boureux,A., Ruffle,F., Audoux,J., Gilbert,N., Xue,H., Gautheret,D., *et al.* (2021) Kmerator Suite: design of specific *k* -mer signatures and automatic metadata discovery in large RNA-seq datasets. *NAR Genomics Bioinform.*, **3**, lqab058.
22. Philippe,N., Salson,M., Commes,T. and Rivals,É. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.*, **14**, R30
23. Bougé,A.-L., Rufflé,F., Riquier,S., Guibert,B., Audoux,J. and Commes,T. (2018) RNA-Seq Analysis to Detect Abnormal Fusion Transcripts Linked to Chromothripsis. In: Pellestor,F. (ed.) *Chromothripsis, Methods in Molecular Biology*. Springer New York, NY, Vol. **1769**, pp. 133–156.
24. Babiceanu,M., Qin,F., Xie,Z., Jia,Y., Lopez,K., Janus,N., Facemire,L., Kumar,S., Pang,Y., Qi,Y., *et al.* (2016) Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res.*, **44**, 2859–2872.
25. Philippe,N., Boureux,A., Bréhélin,L., Tarhio,J., Commes,T. and Rivals,É. (2009) Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res.*, **37**, e104.
26. Marchet,C., Iqbal,Z., Gautheret,D., Salson,M. and Chikhi,R. (2020) REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics*, **36**, i177–i185.
27. Kent,W.J. (2002) BLAT—the BLAST-Like Alignment Tool. *Genome Res.*, **12**, 656–664.
28. Morillon,A. and Gautheret,D. (2019) Bridging the gap between reference and real transcriptomes. *Genome Biol.*, **20**, 112.
29. Arber,D.A., Orazi,A., Hasserjian,R.P., Borowitz,M.J., Calvo,K.R., Kvasnicka,H.-M., Wang,S.A., Bagg,A., Barbui,T., Branford,S., *et al.* (2022) International consensus classification of myeloid neoplasms and acute leukemias: integrating morphologic, clinical, and genomic data. *Blood*, **140**, 1200–1228.
30. Uhrig,S., Ellermann,J., Walther,T., Burkhardt,P., Fröhlich,M., Hutter,B., Toprak,U.H., Neumann,O., Stenzinger,A., Scholl,C., *et al.* (2021) Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.*, **31**, 448–460.
31. Rehn,J., Mayoh,C., Heatley,S.L., McClure,B.J., Eadie,L.N., Schutz,C., Yeung,D.T., Cowley,M.J., Breen,J. and White,D.L.

(2022) RaScALL: rapid (Ra) screening (Sc) of RNA-seq data for prognostically significant genomic alterations in acute lymphoblastic leukaemia (ALL). *PLoS Genet.*, **18**, e1010300.

32. Bessière,C., Xue,H., Guibert,B., Boureux,A., Rufflé,F., Viot,J., Chikhi,R., Salson,M., Marchet,C., Commes,T., *et al.* (2024) Exploring a large cancer cell line RNA-sequencing dataset with k-mers. bioRxiv doi: https://doi.org/10.1101/2024.02.27.581927, 01 March 2024, preprint: not peer reviewed.

33. Dorney,R., Dhungel,B.P., Rasko,J.E.J., Hebbard,L. and Schmitz,U. (2022) Recent advances in cancer fusion transcript detection. *Brief. Bioinform.*, **24**, bbac519.

34. Mertens,F., Johansson,B., Fioretos,T. and Mitelman,F. (2015) The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, **15**, 371–381.

Received: January 29, 2024. **Revised:** August 4, 2024. **Editorial Decision:** August 6, 2024. **Accepted:** August 28, 2024