




RESEARCH ARTICLE

**REVISED** **New chimeric RNAs in acute myeloid leukemia [version 2; referees: 2 approved]**

Florence Rufflé<sup>1,2</sup>, Jerome Audoux<sup>1,2</sup>, Anthony Boureux<sup>1,2</sup>, Sacha Beaumeunier<sup>1,2</sup>, Jean-Baptiste Gaillard<sup>3</sup>, Elias Bou Samra<sup>4,5</sup>, Andre Megarbane<sup>6</sup>, Bruno Cassinat <sup>7</sup>, Christine Chomienne<sup>7,8</sup>, Ronnie Alves <sup>1,9</sup>, Sebastien Riquier<sup>1,2</sup>, Nicolas Gilbert<sup>2</sup>, Jean-Marc Lemaitre<sup>2</sup>, Delphine Bacq-Daian<sup>10</sup>, Anne Laure Bougé<sup>1,2</sup>, Nicolas Philippe<sup>1,2</sup>, Therese Commes <sup>1,2</sup>

- <sup>1</sup>Institut de Biologie Computationnelle, Université Montpellier, Montpellier, France
- <sup>2</sup>Institut de Médecine Régénératrice et de Biothérapie, INSERM U1183, CHU Montpellier, Montpellier, France
- <sup>3</sup>Laboratoire de Cytologie et Cytogénétique, CHU Caremeau, Nîmes, France
- <sup>4</sup>Université Paris Sud, Université Paris-Saclay, Orsay, France
- <sup>5</sup>Institut Curie, PSL Research University, Paris, France
- <sup>6</sup>Institut Jérôme Lejeune, Paris, France
- <sup>7</sup>Laboratoire de Biologie Cellulaire, Hôpital Saint-Louis, Assistance publique - Hôpitaux de Paris (AP-HP), Paris, France
- <sup>8</sup>Hôpital Saint-Louis, Université Paris Diderot, INSERM UMRS 1131, Paris, France
- <sup>9</sup>Instituto Tecnológico Vale, Nazaré, Belém, PA, Brazil
- <sup>10</sup>CEA Institut de Génomique, Centre National de Génotypage, Evry, France




**v2** **First published:** 02 Aug 2017, 6(ISCB Comm J):1302 (doi: 10.12688/f1000research.11352.1)  
**Latest published:** 19 Dec 2017, 6(ISCB Comm J):1302 (doi: 10.12688/f1000research.11352.2)


**Abstract**

**Background:** High-throughput next generation sequencing (NGS) technologies enable the detection of biomarkers used for tumor classification, disease monitoring and cancer therapy. Whole-transcriptome analysis using RNA-seq is important, not only as a means of understanding the mechanisms responsible for complex diseases but also to efficiently identify novel genes/exons, splice isoforms, RNA editing, allele-specific mutations, differential gene expression and fusion-transcripts or chimeric RNA (chRNA).  
**Methods:** We used **Crac**, a tool that uses genomic locations and local coverage to classify biological events and directly infer splice and chimeric junctions within a single read. Crac’s algorithm extracts transcriptional chimeric events irrespective of annotation with a high sensitivity, and **CracTools** was used to aggregate, annotate and filter the chRNA reads. The selected chRNA candidates were validated by real time PCR and sequencing. In order to check the tumor specific expression of chRNA, we analyzed a publicly available dataset using a new tag search approach.  
**Results:** We present data related to acute myeloid leukemia (AML) RNA-seq analysis. We highlight novel biological cases of chRNA, in addition to previously well characterized leukemia chRNA. We have identified and validated 17 chRNAs among 3 AML patients: 10 from an AML patient with a translocation between chromosomes 15 and 17 (AML-t(15;17)), 4 from patient with normal karyotype (AML-NK) 3 from a patient with chromosomal 16

**Open Peer Review**

**Referee Status:**  

	Invited Referees	
	1	2
<b>REVISED</b>		
<b>version 2</b> published 19 Dec 2017		report
		
<b>version 1</b> published 02 Aug 2017	 report	 report

- 1 **Hui Li**, University of Virginia, USA
- 2 **Charles Gawad** , St. Jude Children's Research Hospital, USA

**Discuss this article**

Comments (0)

inversion (AML-inv16). The new fusion transcripts can be classified into four groups according to the exon organization.

**Conclusions:** All groups suggest complex but distinct synthesis mechanisms involving either collinear exons of different genes, non-collinear exons, or exons of different chromosomes. Finally, we check tumor-specific expression in a larger RNA-seq AML cohort and identify new AML biomarkers that could improve diagnosis and prognosis of AML.



This article is included in the [International Society for Computational Biology Community Journal gateway](#).

**Corresponding author:** Therese Commes ([therese.commes@inserm.fr](mailto:therese.commes@inserm.fr))

**Author roles:** **Rufflé F:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Audoux J:** Formal Analysis, Methodology, Software, Supervision; **Boureaux A:** Data Curation, Resources; **Beaumeunier S:** Methodology, Software; **Gaillard JB:** Methodology, Resources; **Bou Samra E:** Data Curation; **Megarbane A:** Resources; **Cassinat B:** Resources; **Chomienne C:** Resources; **Alves R:** Software; **Riquier S:** Data Curation, Formal Analysis, Methodology, Software; **Gilbert N:** Methodology; **Lemaitre JM:** Methodology; **Bacq-Daian D:** Methodology; **Bougé AL:** Data Curation, Formal Analysis; **Philippe N:** Conceptualization, Formal Analysis, Methodology, Software; **Commes T:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Rufflé F, Audoux J, Boureaux A *et al.* **New chimeric RNAs in acute myeloid leukemia [version 2; referees: 2 approved]** *F1000Research* 2017, **6**(ISCB Comm J):1302 (doi: [10.12688/f1000research.11352.2](https://doi.org/10.12688/f1000research.11352.2))

**Copyright:** © 2017 Rufflé F *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

**Grant information:** This work was supported by the French ANR IBC project "Investissement d'avenir en bioinformatique-projet-IBC" and FRM "Appel d'offres urgence pour la bioinformatique, projet DBI20131228566". This work was supported by the France Génomique National infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale pour la Recherche (ANR-10-INBS-09). We acknowledge la Ligue Nationale Contre le Cancer for financial support (EL2015.LNCC/JML) to JML's team, and Le Cancéropôle Grand Sud-Ouest (GSO).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 02 Aug 2017, **6**(ISCB Comm J):1302 (doi: [10.12688/f1000research.11352.1](https://doi.org/10.12688/f1000research.11352.1))

**REVISED Amendments from Version 1**

We have carefully read the manuscript and corrected the typos.

We replaced commas with periods in the “AML samples and cell lines” section, and the degree symbols with percentage signs, as required in the “FISH Experiments” section.

As requested by the referee, we have added the following statement in the “RNA-seq experiments” section: “The RNAseq was performed using polyA-selection with the TruSeq RNA Lib-Prep Kit (Illumina) adjusted with GATC specific procedure for strand specificity.”

We also substituted “MDR” with “MRD” for minimal residual disease.

We have also changed the resolution of [Figure 5A](#), increasing the font size.

**See referee reports**

## Introduction

High-throughput sequencing technologies (NGS) enable the detection of new biomarkers used for tumor classification and disease monitoring, including patient response to therapies. Whole-transcriptome analysis with RNA-seq is increasingly acquiring a key role, not only to learn about mechanisms responsible for complex disease, but also to identify novel genes/exons, splice isoforms, RNA editing, allele-specific mutation, differential gene expression, fusion-transcripts and chimeric RNA (chRNA)<sup>1,2</sup>.

For chimeric RNA, a group of fusion transcripts is increasingly used by geneticists in oncology diagnosis<sup>3,4</sup>. These cancer biomarkers are generated at DNA level from gene fusions by mechanisms such as translocations, inversions, or more complex chromosomal rearrangements. In some well-documented cases, gene fusions, in addition to contributing to neoplastic transformation, produce fusion RNA and proteins used as therapeutic targets (Mertens *et al.*, 2015<sup>5</sup>, Yoshihara *et al.*, 2015<sup>6</sup> and references therein). However recent RNA-seq analyses have revealed the existence of an enlarged “chimeric transcriptome”<sup>5,7–10</sup> generated by new RNA processing events such as cis- and trans-splicing, whose mechanisms and functional roles are poorly understood. It is crucial to determine whether these events represent artefacts of RNA-sequencing, transcriptional noise with little impact on cell functions or tissue-specific transcripts, or whether these events are important in tumor development. Several significant examples of chRNA without corresponding fusions at DNA level have been described in the context of neoplasia<sup>11,12</sup>.

Both from a methodological and biological perspective, profiling chimeric RNA is a challenging issue. chRNA's distinctive features must be identified to provide relevant biological information, including synthesis mechanisms and pertinence as a biomarker. Though RNA-Seq may enable new biomarker discovery, there is a lack of consensus on which analysis tool and algorithmic strategy should be used, especially in the detection of chRNA<sup>13</sup>. We recently proposed Crac<sup>14</sup>, a novel way of analyzing reads that combines genomic locations and local coverage to classify the biological events, to directly infer splice and

chimeric junctions within a single read. The Crac software is based on an innovative algorithm that allows extraction of transcriptional events, irrespective of annotation. The main advantage of using Crac to detect chimeras is its high sensitivity and specificity, allowing detection of rare events with confidence. We developed the complementary CracTools module to aggregate, annotate, and filter the chRNA reads. The procedure classifies reads into 4 classes, depending on the exon organization. The first class corresponds to fusion transcripts arising from two different chromosomes, the second class includes parts of two genes belonging to the same chromosome strand. The third and fourth classes involve non-collinear transcription on the same chromosome. This categorization suggests that each class represents distinct, complex synthesis mechanisms.

In order to assess Crac's potential to identify new biomarkers, we present data related to acute myeloid leukemia (AML) RNA-seq analysis. AML provides a practical cellular model to detect chRNA biomarkers in order to improve classification and patient follow-up in precision medicine<sup>2</sup>. Despite the fact that leukemia chRNAs are well characterized, our study highlights new biological cases of chRNA. We identify and validate 17 chRNAs initially detected in 3 AML patients that belong to the 4 classes. We then explore their specific expression in the publicly available LEUCEGENE cohort<sup>15</sup>, and propose new criteria for distinguishing chRNAs based on their recurrence, tumor, subgroup, or patient-specific expression. We also classify the chRNAs according to differences in expression of the genes linked to the chimera, in healthy donors vs AML patients. Finally, we identify new biomarkers that could improve diagnosis and prognosis of AML.

## Materials and methods

### AML samples and cell lines

Three sets of samples from patients with AML were used in this study. Each patient is designed by a specific ID including the OM or OS code corresponding respectively to bone marrow or blood sample ([Supplementary Table S1](#)). The first set of 25 AML samples consisted of 11 AML-NK samples, 4 AML-inv16, 5 AML-UK, 2 AML-t(15; 17) also called Acute Promyelocytic Leukemia (APL), and 3 other AML-AK samples ([Table S1](#) lines 3 to 27). They were supplied by JBG (Biological Resource Center CHU-Nîmes, France) and included both RNA and peripheral blood mononuclear cells (PBMC), stored in RNALater (Ambion, USA) according to the manufacturer instructions. The second set of 14 AML samples consisted of 10 AML-UK, 2 AML-NK, and 2 AML AK. They were supplied by AM (Medical Genetic Unit, University St Joseph, Lebanon) ([Table S1](#) lines 28 to 41). They included blood and bone marrow stored in TRIzol reagent (Life Technologies, USA). The third set of AML samples included 3 AML-t(15;17), and were provided by CC and BC (Cell Biology Unit, Hôpital St-Louis, France). For the latter, peripheral blood mononuclear cells were collected by ficoll-hypaque density gradient and cultured at a concentration of 1×10<sup>6</sup>/ml, with or without 0.1µM ATRA, for 3 days. PBMC samples from healthy donors were pooled and used as control sample. The patients and healthy donors provided written informed consent to participate in the study, in accordance with the Declaration of Helsinki. The U937 leukemia cell line (DSMZ, Braunschweig, Germany), NB4 promyelocytic cell line and

NB4-LR2 cell line (provided by CC) were cultured in RPMI 1640 (Invitrogen, USA) containing 10% decomplexed FBS (Dutscher, Brumath, France). For differentiation conditions, NB4 and U937 cells were cultured as previously described<sup>16,17</sup>. The chemical agents used for differentiation were 1 $\mu$ M all-trans retinoic acid (ATRA; Sigma-Aldrich, Gillingham, UK), or 0.1 $\mu$ M vitamin D (VD) associated with 500pg/ml transforming growth factor beta TGF $\beta$  (Promega Corporation, USA) for the NB4 cell line<sup>18</sup>. For the U937 cell line, 0.1 $\mu$ M TTNPB associated with 1 $\mu$ M Targretin (LGD1069) and 0.1 $\mu$ M 1 alpha, 25 dihydroxyvitamin D3 (VD) were used. TTNPB, LGD1069 and VD were kindly provided by Dr Klaus (Hoffman-La Roche, Switzerland), JHiernaux (Glaxo-welcome Laboratories, France), and L Binderup (Leopharmaceutical products, Denmark), respectively. We also used human neuroblastoma cancer SH-SY5Y cells, human breast cancer MCF7 cells and human prostate cancer MDA-PCa cells. Cell pellets were kindly provided by S. Marchal (University Montpellier, France), and by D.Noel (Institute of Regenerative Medicine and Biotherapies, France) for the latter.

### RNA extraction and reverse transcription

RNA was extracted with the RNeasy Qiagen kit (Qiagen, Germany), additional DNase treatment was performed in order to remove residual DNA (RNase free DNase set, Qiagen, Germany). Total RNA was quantified using a NanoDrop<sup>®</sup> ND-1000 spectrophotometer (NanoDrop ND-Thermo Fisher Scientific, USA). RNA quality and quantity were assessed using the 2100-Bioanalyzer (Agilent Technologies, Waldronn, Germany). Reverse transcription was performed with random primers and MultiScribe Reverse Transcriptase (High-capacity cDNA Archive kit; Applied Biosystems, USA), using 1  $\mu$ g of total RNA. To check for possible chRNA formation induced by transcriptional artefacts, part of the samples were double reverse transcribed. In this case, the second reverse transcription reaction was performed with ImProm-II<sup>™</sup> Reverse Transcriptase (Improm-II Reverse Transcription System, Promega, USA).

### RNA-seq experiments

Three AML samples (AML-NK, AML-t(15;17) and AML-inv16) taken from patients OM100011, OM110223 and OS110089, respectively (AML test group) from the Biological Resource Center, CHU-Nîmes, France, were selected for RNA-seq experiments. 4  $\mu$ g of total RNA taken from bone marrow (OM100011, OM110223) or blood PBMCs (OS110089) were sent to GATC biotech and analyzed on an Illumina HiSeq 2000 to generate 100 base pairs of stranded RNA-Seq paired-end reads. The RNA-Seq was performed using polyA-selection with the truSeq RNA Lib-Prep Kit (Illumina, San Diego, CA) adjusted with GATC specific procedure for strand specificity. The following publicly available datasets of the LEUCEGENE project dedicated to Acute myeloid and lymphoid leukemia studies<sup>15</sup>, have been used in this study (a detailed list is provided in [Supplementary Table S6](#)):

- GSE48846 (17 CD34 hematopoietic stem cell),
- GSE49642 part1 (43 AML-NK),
- GSE62190 part3 (82AML-AK with 24 inv16, 19 t(8;21), 5 inv3,7 t(9;11), 5 t(6;11), 6 t(11;19) and complex karyotypes)

- ENCODE publicly available datasets ([Supplementary Table S7](#)).

### RNA-seq analysis and chRNA extraction procedure using Crac and cracTools

RNA-seq analyses were performed serially using Crac (V2) and CracTools (V1.2)<sup>14</sup>. Crac is a software for analyzing reads when a reference genome is available, that is completely independent of annotations. It ignores the sequence quality of reads and classifies reads by detecting diverse biological events (mutations, splice junctions, and chRNAs) and sequencing errors from a RNA-seq read collection. In this analysis, we used two Crac versions in succession (V1.6 and V2) to extract and classify chRNAs, with the GRCh37/hg19 genome as reference genome. Crac extracts the chimeric reads supporting the chimeric junction (spanning junction) made of a non-collinear arrangement of genomic regions<sup>14</sup>. CracTools was then used to aggregate, annotate and filter the chRNA reads and extract the chimeric paired reads (spanning PE) ([Figure S1](#)). Reads were annotated according to a GFF file from ENSEMBL Genome Browser (link to GFF file in Data Availability section) by giving priority to location in exons of the annotated genes. The GFF file was built from ENSEMBL (Ensembl 84 annotations). When reads were located on a non-annotated, transcribed region, the corresponding "NONE" annotation was mentioned. The procedure included classifying chRNA into four categories depending on exon organization, as described in the introduction. This classification resembles the one depicted in [Gingeras, 2009](#)<sup>7</sup> and can be summarized as follows:

- Class 1, the exons are located on different chromosomes;
- Class 2, the exons are collinear but most likely belong to different genes, to be verified through the annotation;
- Class 3, the exons are on the same chromosome and same strand, but not in the order in which they are found on DNA;
- Class 4, the exons are on the same chromosome but on different strands.

For each analyzed chimera, the pipeline provides related information, including a unique read identifier annotated by the pipeline, class, number of spanning junctions and spanning reads:

1. ID: A unique read ID for each chimera, composed of 'sample name: chimera ID'
2. Fusion gene names (left-right)
3. Chr(left): Chromosome number of the 5' part of the chimera
4. Pos1: Genomic position of the 5' part of the chimera
5. Strand1: Genomic strand of the 5' part of the chimera (+1 or -1).
6. Chr(right): Chr number of the 3' part of the chimera
7. Pos2: Genomic position of the 3' part of the chimera
8. Strand2: Genomic strand of the 3' part of the chimera
9. ChimValue: the chim-value takes into account methodological parameters and ambiguities, including the read mapping (P\_loc) and the read coverage (P\_support).<sup>13</sup>

10. Spanning junction normalized: Spanning junction reads coverage (normalized per billion of reads). A spanning junction read is the read that contains the chimeric junction.

11. Spanning PE normalized: Coverage of paired-end reads (normalized per billion of reads) that contains the chimeric junction in the non-sequenced part

12. Class: Chimeric class from 1 to 4.

The filtering process with CracTools considered the following thresholds:

- a. Only candidate fusions (chRNA) with at least one read covering the fusion breakpoint,
- b. Spanning reads must be associated with pair-end reads,
- c. An annotation of the fusion junction matching almost a known expressed sequence (gene A or Gene B),
- d. A ChimValue of up to 60, allowing the removal of false positives corresponding to pseudogenes and splicing events detected by GSNAP.

Candidate fusion transcripts involving adjacent genes within a 3Kb distance region were discarded. To estimate the number of supporting reads for a chimeric candidate, CracTools extracted the count number of spanning junctions and spanning PE reads. All candidate fusion transcripts were validated using qPCR and Sanger sequencing except the class 3 overlap which requires systematic reconstruction of the fusion transcript for the design of primers.

### Manual annotation

The potential chimeras are listed in [Table S2](#) with the appropriate features. For each fusion transcript, the Crac software provided a reconstructed sequence comprising, on one hand, the chimeric junction sequence based on the most representative read, and on the other hand, the paired read sequence. The symbol (#) marks the segment of the read that was not sequenced and the (\*) symbol marks the junction point ([Table S2](#) and [Supplementary Figure S1](#)). The sequences of reads, both junction and paired, which supported the chimera were mapped (BLAT, UCSC) to the human genome GRCh37/hg19 in order to identify complex biological events (splicing, SNPs, insertions, deletions, repeats, polymorphisms, etc.). Complementary annotations were identified using the ENSEMBL genome browser to determine exons and spliced variants involved in the transcript. Protein sequences and functional annotations were also verified to identify affected protein domains and to evaluate potential protein damage in selected chRNAs.

### PCR validation

Reverse transcription was performed as described above. 1 µl of each cDNA sample (2ng/µl) was added to a 5 µl of reaction mix containing 3 µl of Master Mix (LightCycler®480 sybr green I Master, Roche Diagnostics, GERMANY) and 0.66 µM forward and reverse primers. Mix and cDNA were loaded onto the 384-well PCR plate using an epMotion 5070 automated pipetting system (Eppendorf, Germany). Primer sequences were designed

using the [Primer3Plus](#) web interface, with some constraints as described in the PCR strategy (see [Supplementary Figure S1](#)), and synthesized by Eurofins MWG Operon, Germany. The amplification area was centered on the junction, and primers were designed to tag each sides of the junction. Primers are listed in [Supplementary Table S3](#). PCRs were carried out in 384-well plates on a LightCycler®480 Real-Time PCR System (Roche Diagnostics, Germany). Amplifications were performed according to the following conditions: 95°C for 5 min, then 55 cycles as follows: 95°C for 10s, followed by T°C depending on T<sub>m</sub> for 10s, and 72°C for 10s. Ultimately, a melting curve analysis ranging from 60°C to 95°C was performed to control primer specificity. For each sample, the graph of the negative first derivative of the melting curve gave a specific peak corresponding to the amplified transcript. Samples with T<sub>m</sub> value peaks different from those found in the negative control were considered potential positive targets and retained for sequencing. PCR products were purified with the Minelute PCR purification kit (Qiagen, Germany) and sequenced on the ABI 3730XL (Eurofins MWG Operon, Germany).

For the newly discovered Class1 chRNA, identified in patient OS110089 and corresponding to Chr2 and Chr13 positions and to PAN3-NONE annotations, the presence of chRNA was checked in leukemia samples obtained during the patient follow-up ([Figure S3](#)). The NONE transcript expression was checked by qPCR in human embryonic stem cells HD129 (cDNA was kindly provided by J. De Vos, Institute of Regenerative Medicine and Biotherapies, France), in AML samples and in U937, NB4, SH-SY5Y, MDA-PCa and MCF7 cell lines. To this end, we designed forward and reverse primers on the 5'NONE sequence ([Figure S2](#)).

### FISH Experiments

Molecular cytogenetics were performed on metaphases from bone marrow aspirate collected from the samples using a synchronised protocol. A first step aging slide with the cytogenetic preparation was performed by immersing slides in 2xSSC solution (saline sodium citrate) for 30 minutes at 37°C, followed by dehydration in 3 baths of increasing ethanol concentration: 70%, 85% and 100%, each for 1 minute. Finally, slides were air dried at room temperature. To confirm the putative t(2;13) translocation related to the PAN3-NONE fusion gene, a fusion probe was designed using bacterial artificial chromosome (BAC) technology, framing the following regions of interest: RP11-239J16, RP11-339H12 in chromosome 2p21 (labeled in Cy3 Orange) and RP11-179F17, RP11-95G6 in chromosome 13q12 (labeled in Alexa 488 Green) (BlueGnome, Cambridge, UK). 1 µl of each of the BACs was mixed and diluted in 9 µl of hybridization BlueFISH Buffer and then applied to the slide. Chromosomes fixed on the slide and probe of interest were denatured in a single step using Thermobrite (Abbot Molecular, USA). Codenaturation was done at 75°C for 5 minutes, followed by hybridization at 37°C in a humid atmosphere overnight. To remove the probe that would not properly hybridize, two successive washes in stringent conditions were performed: the first one in 0.4xSSC and 0.1% Igepal at 73°C for 2 minutes, the second one in 2xSSC and 0.3% Igepal for 1 minute at room temperature. After complete

drying, 10 µl of counterstaining reagent (DAPI 125ng/ml, Abbott Molecular, Chicago, USA) was added. Slides were observed on an epifluorescence microscope (Olympus BX60). For both probes, positivity was defined as the presence of two fusion signals (co-location) in addition to a red signal and a green signal.

### Tag search approach and gene expression clustering

The tag search approach consisted of extracting representative sequences (Tags) 30 nucleotides in length, centered on the chimeric junctions. The latter were then searched in LEUCEGENE and ENCODE publicly available RNA-seq datasets. A FASTA file, listing these tags of interest, designed for each chRNA, was submitted to a specific pipeline (countTags, <https://github.com/jaudoux/countTags>) that searched for sequences and their reverse complement with an exact match in the FASTQ files. For each FASTQ file and each tag, the total number of tags (total k-mers) is counted. Final value is given in a delimited table, with a tag count normalized per 5 billion of k-mers.

For the gene expression clustering, chRNA were selected as below. Chimeric genes were extracted using cracTools predictions for samples OS110089, OM100011 and OM110223. Among all chimeric junctions detected, only those having a “chimValue” greater than 75 and with at least 3 spanning reads were conserved. Read-through chimeras were further selected based on three criteria:

- i. Chimera annotated as "Class 2" (collinear transcription),
- ii. Short fusion distance (max 300kb),
- iii. Short exon-end distance (max 20bp).

Tandem repeat chimeras were further selected based on three criteria:

- i. Chimera annotated as "Class 3"
- ii. Overlapping chimeric fragments (the two parts of the chimeric read correspond to overlapping sequences on the genome)
- iii. Both chimeric fragments are located on the same exon of the same gene.

For each chimera type (read-through and tandem-repeat), only genes involved in at least 2 different chimeric events (either from the same or different samples) were finally selected as candidates for the clustering of gene expression. LEUCEGENE data were downloaded from SRA using the fastq-dump utility (version 2.5.4) and converted to FASTQ. Using Kallisto 0.42.4 software and Ensembl 84 annotations, we determined transcript expression. Transcript counts computed by Kallisto were merged at gene-level<sup>19</sup>. The normalization of counts was performed with DESeq2 (version 1.14.1) (design ~ 1), so values used in the clustering were normalized counts transformed with the variance stabilization method provided in DESeq2 package. Heatmaps were produced with heatmap.2 function from the gplots package (R version number 3.3.2), using default parameters (i.e. complete-linkage clustering and Euclidean distance).

## Results

### Validation of chRNA candidates

We performed RNA-seq on samples from 3 AML patients. One presented with a normal karyotype (NK), while the other two presented with an abnormal karyotype (AK), one with an Inv16 and the other with a t(15,17) translocation (sample names 1–3, [Supplementary Table S1](#)). The sequences were analyzed using Crac and CracTools. The selected chRNA candidates were tested by qPCR, and sequenced when qPCR displayed a positive signal. Some candidates could not be validated by qPCR due to the difficulty in designing suitable primers. The CBFβ-MYH11 and PML-RARA fusion transcripts expressed in the AML-inv16 and AML-t(15;17) samples were identified using both RNA-seq and qPCR analysis, confirming the reliability of RNA-seq and the Crac suite in this type of analysis.

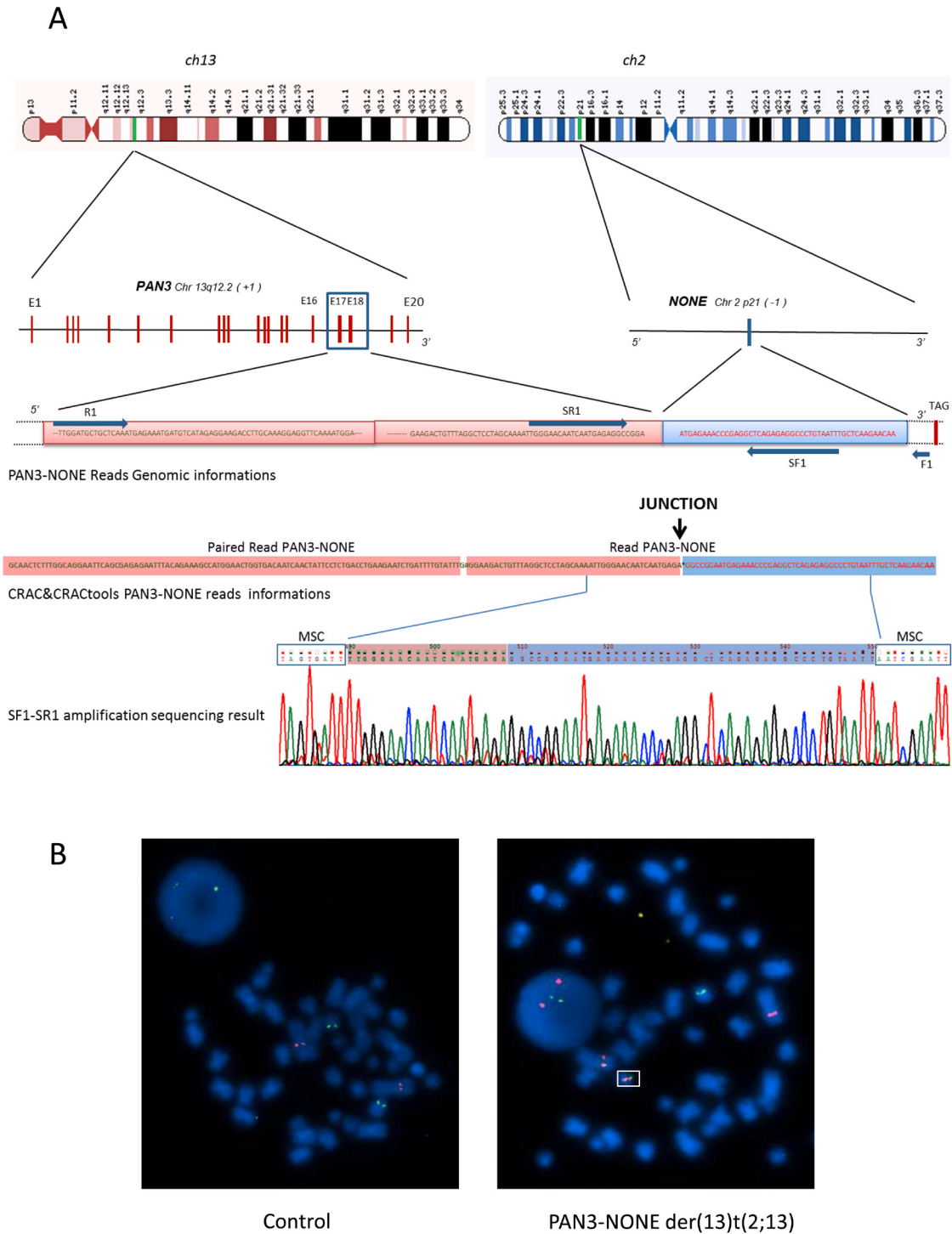
We identified 17 chRNAs among 44 candidates; 10/23 from AML-t(15,17), 4/18 from AML-NK, and 3/3 from AML-inv16 ([Supplementary Table S2](#) and [Supplementary Table S4](#)). The validated chRNAs were distributed into 4 classes (definitions of these are in the materials and methods section) as follows:

- Class 1: 5
- Class 2: 3
- Class 3: 5
- Class 4: 4.

Among the Class 1 chRNAs more frequently associated with genomic translocation, we identified 4 chRNAs associated with PML and RARA genes in patient OM110223 suffering from AML-t(15,17). The remaining chRNA of this class was associated with the PAN3 gene and a non-annotated region (NONE), and was found in patient OS110089 suffering from AML-inv16. We highlighted two types of Class 2 chRNAs that depend on the genomic distance between the two parts of the read. A short distance was consistent with read-through, whereas a long distance would be associated with other mechanisms. We found two kinds of Class 3 chRNA - in the first, the chRNA processes the 3' exon before the 5' start of the same gene. In the second, the chimera involve distant exons from different genes. Finally we also validated several Class 4 chRNAs, among which the CBFβ-MYH11 chimera associated with Inv16.

### New Class 1 PAN3-NONE chRNAs associated with a genomic translocation

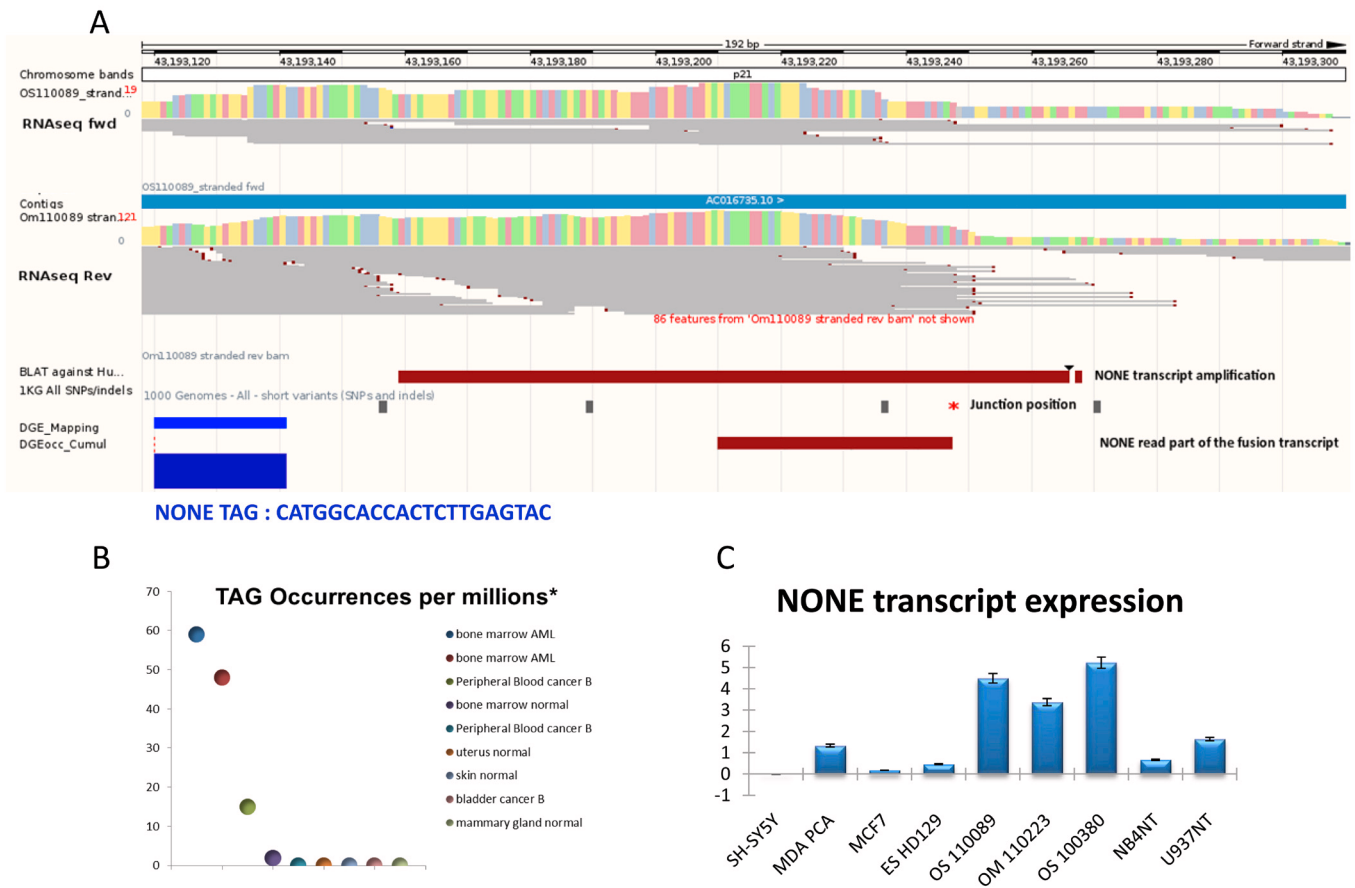
A new Class 1 chRNA was identified in patient OS110089 presenting the inv16 associated fusion transcript CBFβ-MYH11. The new fusion junction read (n° 14888 PAN3-NONE; [Supplementary Table S2](#)) corresponds to 13q12.2 and 2p21 chromosomal locations at the 5' and 3' of the RNA, respectively. The read chimeric annotation indicates a fusion between the 3' end of PAN3 exon18 (chr13+), with a non-annotated transcribed region “NONE” (chr2-pos 43193199-43193247) as shown in [Figure 1A](#). The PAN3-NONE fusion transcript was validated by qPCR, and was only detected in patient OS110089 as shown in results of the AML cohort ([Figure 5A](#)). We compared its qPCR Ct value



**Figure 1. A new specific Class 1 PAN3-NONE chimera. A)** Fusion junction description and primer design. The sequences of spanning junctions and paired reads supporting the chrRNA and the corresponding designed primers are indicated. The orange and blue squares represent respectively the E17 and E18 of PAN3 gene (E16 of ENST00000399613.1 transcript), and the NONE part of the fusion transcript. The sequencing result of the PAN3-NONE SF1-SR1 PCR product is indicated. The symbol (#) denotes the part of the read which was not sequenced, (\*) denotes the junction, and MCS indicates the multiple cloning sites used to clone the PCR product. **(B)** Translocation checking by FISH. The left panel (Control) represents normal blood cells and the right panel represents cells from patient OS110089. The presence of the NONE sequence on chr2 is denoted by the red signal and the presence of the PAN3 sequence on chr13 mentioned by the green signal. The right panel presents the NONE-PAN3 fusion at the chromosomal level (white box).

with the Ct value obtained with the amplification of the CBFβ-MYH11 transcripts, and noticed that its value was higher (32 cycles vs 28, [Supplementary Figure S3](#)). This difference could be correlated either to a lower transcript expression level or to a heterogeneous expression in the tumor cell population. In order to determine whether this fusion transcript is associated with chromosomal rearrangement, FISH experiments were performed with a custom fusion probe. A corresponding translocation was observed in only 31% of the leukemia cells, demonstrating that the PAN3-NONE transcript belongs to a subclone, which could explain the lower expression level observed ([Figure 1B](#)). The analysis of the PAN3-NONE transcript during patient follow-up revealed its disappearance after the first induction of chemotherapy, without reappearance during the relapse period ([Figure S3](#)).

We next investigated the transcription of the non-annotated region “NONE” (chr2-pos 43193199-43193247) in normal and tumor tissues using the approach described previously combining digital Gene expression (DGE) and RNA-seq data<sup>20</sup>. Querying tissue expression profiles with DGE tags, we observed a tag in the NONE chromosomal area showing a specific expression in AML samples ([Figure 2B](#)). The RNA-seq read coverage and the DGE tag ([Figure 2A](#)) confirmed a new transcribed region, validated by qPCR in tumor cell lines ([Figure 2C](#)). We confirmed the presence of the NONE specific expression in AML and normal CD34+ hematopoietic stem cells (HSC) using a tag search approach in a largest RNA-seq collection of normal and tumor tissues ([Table S5](#) and data not shown). Together, these results revealed a new lincRNA specifically expressed in



**Figure 2. NONE transcript sequencing and expression.** **A)** Display of the ENSEMBL genome browser viewer for the NONE transcribed region. The blue horizontal bars represent the genomic sequence. The histogram of the RNA-Seq coverage in the chromosomal region is displayed on both strands (OS110089 stranded and RNA-seq fwd/rev tracks). Public and personal DGE data ('DGE tag location' track: blue rectangle for occurrence>2) are displayed on both strands of the chromosome, with their relative occurrences (histogram of 'DGE expression level' track) using a private DAS server. **(B)** Relative expression of the NONE tag per million reads in the DGE datasets (Philippe, *et al.* 2013). **(C)** Relative expression of the NONE transcript in different cell lines and AML samples (NT refers to not treated cells). The relative gene expression was determined using the 2- $\Delta\Delta$ CT method. Transcriptional modulation was calculated by comparing various lineages with SH-SY5Y (subline of the neuroblastoma cell line SK-N-SH). For normalization, RPS19 was selected as a reference transcript. Standard deviation was measured using duplicate.

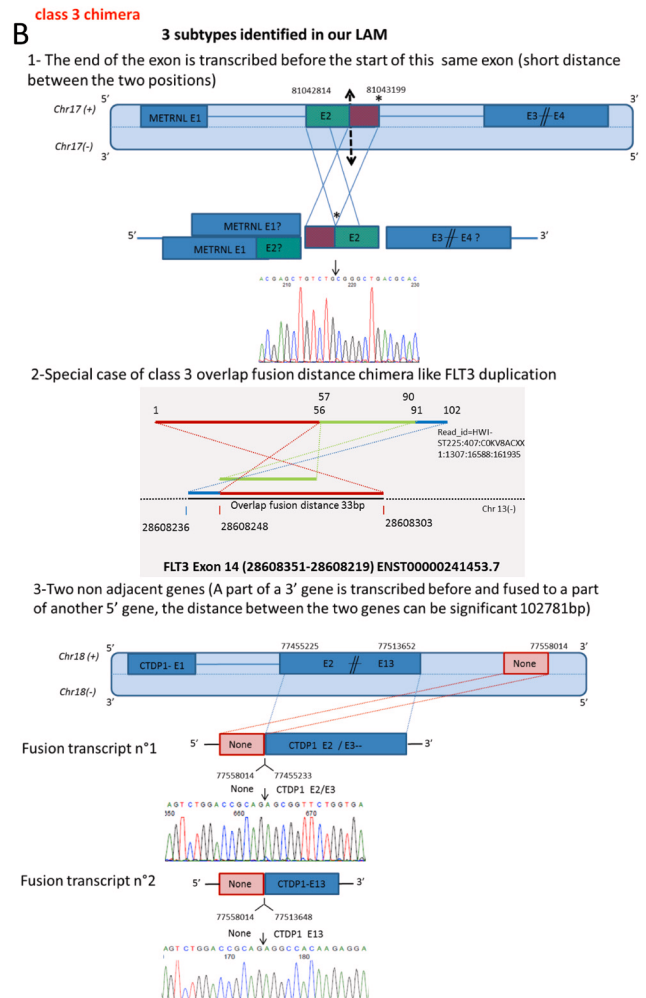
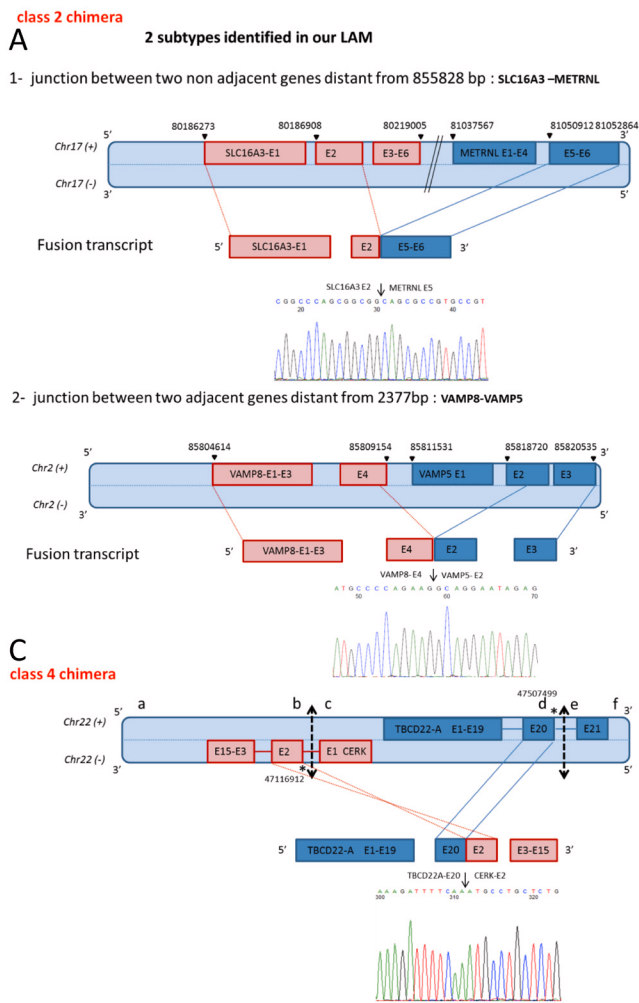




previously described fusion proteins (Figure 3D). Moreover, different fusion transcripts linked to the translocation between the chr15 and 17 can coexist in a tumor sample, with their expression changing with time or treatment (Figure 3C). The second new fusion transcript corresponds to a chRNA that joined the exon 3 of a known RARA antisense transcript with the antisense part of PML intronic region (Figure 3B). The primers were designed in the corresponding PML intron free from known transcription, in order to amplify the antisense chRNA. This transcript was only detected in the sequenced sample. The sequencing results of PML-RARA chRNA qPCR amplifications revealed another fusion transcript joining PML exon 3 and RARA exon 10 present in OM110223 leukemia cells, without corresponding spanning junction read.

**New Class 2, 3 and 4 chRNA**

Four classes of chimera were identified using the Crac and CracTools pipeline. Associated with specific features and annotations, subcategories could be defined and linked to potential biological mechanisms (Figure 4). Among the Class 2 chimera, we observed two distinct categories, depending on the vicinity of the two relevant genes (Figure 4A). The first category involves a junction between non-adjacent exons separated by thousands of base pairs (SLC16A3-METRNL and UBR5-AZIN1, Figure 4A and Table S2). The validated fusion transcript joins the SLC16A3 exon 2 with the METRNL start exon 5 at a distance of 855828bp. As shown from RNA-seq data, the region between the two genes is transcribed (data not shown), which calls into question the hypothesis of an intra-chromosomal deletion.



**Figure 4. Different classes of validated chRNA.** The position and orientation of the genes on the chromosomes have been schematically arranged. The start and end positions of the exons have been indicated. The representations of the chRNA junction supported by the Sanger sequencing results are also shown. **(A)** Two subtypes illustrating Class 2 chimera junctions. The first subtype<sup>1</sup> is illustrated by the junction of two non-adjacent genes SLC16A3 and METRNL, separated by 855828 bp. The second subtype<sup>2</sup> is illustrated by the junction between adjacent genes VAMP8 and VAMP5, separated by 2377 bp. **(B)** Three subtypes illustrating Class 3 chimera junction. They all involve the inversion of exon order, first two concern an inversion within a same exon of a gene (METRNL and FLT3 cases), and the third subtype involves an inversion of a transcribed sequence, separated by about 103 000 bp (NONE-CTDP1). **(C)** A Class 4 chimeric construction joining an exon from (+) strand gene TBCCD22A with an exon from (-) strand gene CERK, arising from the same chromosome 22.

The second subcategory concerns two adjacent genes, fused with a loss of some exons. In the illustrated example of VAMP8-VAMP5, VAMP8-001 transcript exon 4 joined VAMP5-001 transcript exon 2 (Figure 4A). The subgroup most likely defines a read-through transcript category with an alternative splicing.

The Class 3 chimera could reflect three different mechanisms associated with: genomic duplication, polymorphism, or intrinsic transcriptional mechanisms (Figure 4B). Two cases concern intra-exonic transcripts, with one exon end being present before the start of the same exon. This subcategory is illustrated by two fusion transcripts (METRNL; Figure 4B and RNF220 data not shown). In the transcript METRNL-001, the 3' part of exon 2 (chr17 (+)/pos 81043015-81043199) is present upstream of the 5' part of exon 2 (chr17 (+)/pos 81042813-81042850). Such transcript could be generated either by transcriptional event or by the transcription of a rearranged allele. Indeed, this subgroup, intragenic chRNA, could highlight the presence of tandem duplication. As an example, we also identified the well-known FLT3 tandem repeat involved in acute myeloid leukemia. This type of tandem duplication is characterized by an overlap in the read's genomic positioning, which can be easily detected using our workflow (Figure 4B).

The last case involves two distant genes. The genomic location of both transcripts is on the same chromosome but at a distance of 102781bp. We identified the chimera NONE-CTDP1 in the OM110223 patient sample. The transcript of a non-annotated chr18 (+)/pos 77557943-77558014 segment is fused with exon2-CTDP1-001 or with exon13-CTDP1-001 (Figure 4B).

The Class 4 chRNA corresponds to reads whose 5' and 3' parts match on the chromosome's opposite strands (Figure 4C). This kind of chimera reflects a genomic inversion like CFBF-MYH11 in the absence of overlapping elements in the read. Besides the widely described inv16 (CBFB-MYH11), we validated three chRNA linked with a possible chromosomal inversion (TBC1D22A-CERK, MAEA-CTBP1, DHRS7B-TMEM11). In the TBCD22A-CERK fusion transcript, the TBCD22A exon 20 end is fused with the CERK exon 2 start (Figure 4C).

### Recurrence of chRNA in normal hematopoietic stem cells and AML

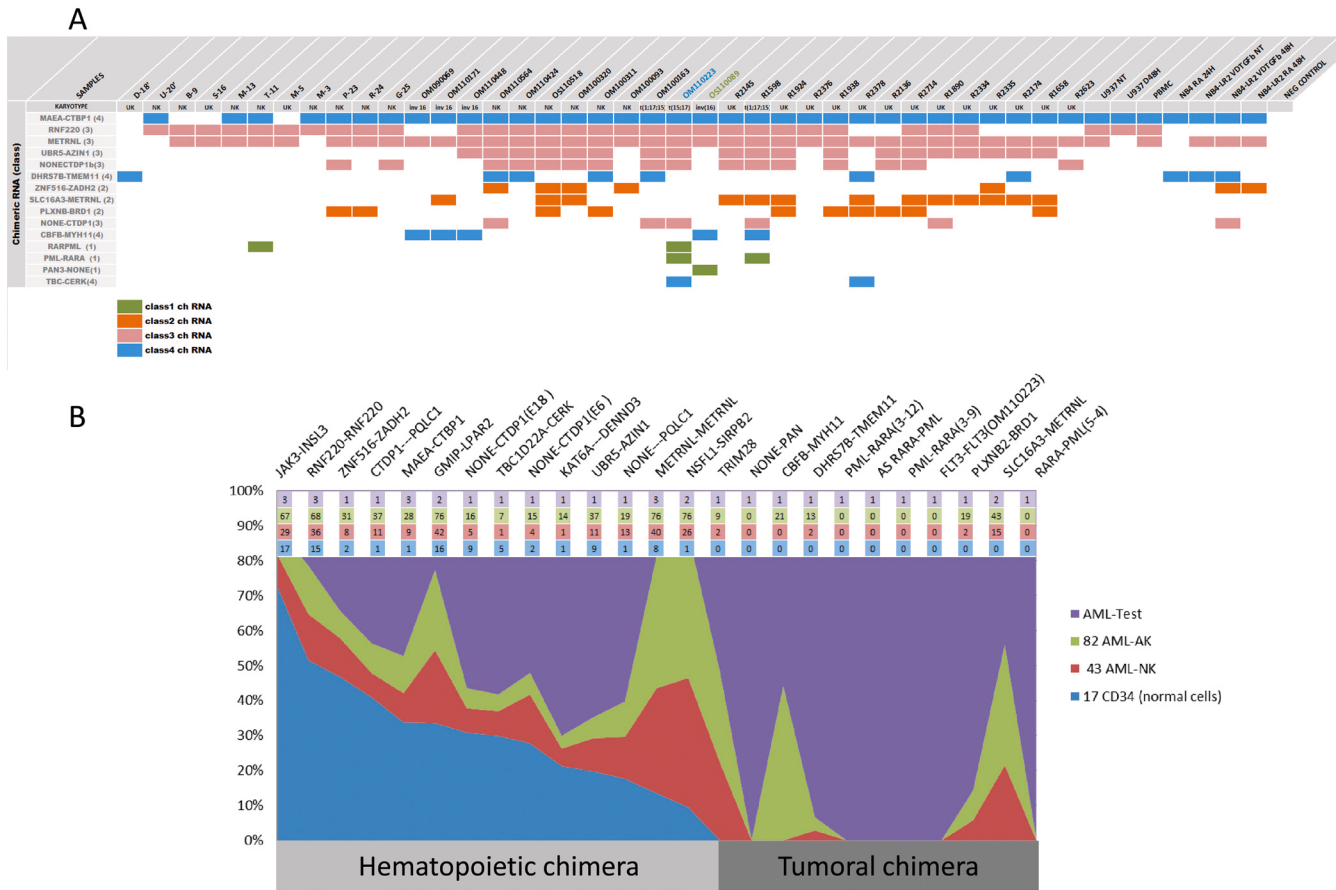
To extend our analysis, we tested the expression of the validated chRNA in a large cohort of AML samples with different karyotypes (Figure 5A, Table S1) using qPCR. We then classified the chRNA into AML subtypes or tumor-specific chRNA subgroups, taking into account the frequency and tissue-specific expression. We also distinguished the non-tumor chRNA by examining their expression in normal PBMCs. Among the 15 chRNA tested, some of them were widely expressed in all AML subtypes (RNF220-RNF220, METRNL-METRNL and MEA-CTDP1).

The frequency and tissue-specific expression did not depend on the chRNA classes.

In order to validate our strategy on a large cohort, we analyzed a publicly available dataset of 125 AML and 17 normal CD34+ HSC RNA-seq using a tag search approach (see Materials and methods). For this purpose, we selected qPCR validated chRNA (Table S4) and candidates previously untested due to difficulty in designing primers (Table S2). The chRNAs were classified by their relative expression in normal CD34+ HSCs, in order to distinguish non-tumor and tumor-specific ones (Figure 5B). Among the chRNAs expressed in CD34+ HSCs, we observed different profiles: those more highly expressed in normal CD34+ cells than in AML, and those with low expression in CD34+ HSCs (see NSFL1-SIRPB2, METRNL-METRNL).

We identified four new types of tumor-specific chRNA; TRIM28-TRIM28, DHRS7B-TMEM11, PLXNB-BLRD1 and SLC16A3-METRNL, expressed in all AML groups (Figure 5B). DHRS7B-TMEM11 and PLXNB-BLRD1 transcripts are most abundant in the AML test group, whereas TRIM28-TRIM28 and SLC16A3-METRNL are equally expressed in the three AML groups (AML-test, 82 AML-AK, 43 AML-NK). It is worth noticing the involvement of the METRNL gene in two identified chimeras. TRIM28-TRIM28, FLT3-FLT3, PML-RARA (with (3-9) or (5-4) junction) and CFBF-MYH11 tag counts showed high mean expression levels (see Figure S4). FLT3, PML-RARA and CFBF-MYH11 are strong markers in AML, and also useful for prognostic and MRD monitoring. The FLT3 tag identified in OM110223 is not found in other samples, yet the pipeline reveals other FLT3 fusion transcripts, with different sequences in other AMLs (data not shown), indicating several variations at this fusion point. Among the new, TRIM28-TRIM28 chRNA is present at low frequency in normal and abnormal AML karyotype (2/43 AML-NK and 9/82 AML-AK). The high expression level of this chRNA in positive samples (comparable to previously described known markers CFBF-MYH11, FLT3-FLT3 and PML-RARA), suggests it could have a key role in such tumors.

In order to verify whether genes involved in chRNA have an aberrant expression profile<sup>25</sup> that could influence tumorigenesis, we analyzed the impact of gene expression on AML with unsupervised clustering. As described for the read-through chRNA subcategory, we compared RNA-seq data from normal CD34+ HSC and AML-AK subtypes (LEUCEGENE, part3). Figure 6 shows a quantitative analysis obtained with "read-through related" genes of the input cohort. Known AK-AML subgroups could be distinguished from CD34+ HSC by their expression profile. We performed a similar study with the Class 3 tandem duplication subgroup, showing interesting differential profiles with "tandem duplication related" genes, mostly comprised of the newly identified TRIM28, CEPBD and FLT3 (Supplementary Figure S5).

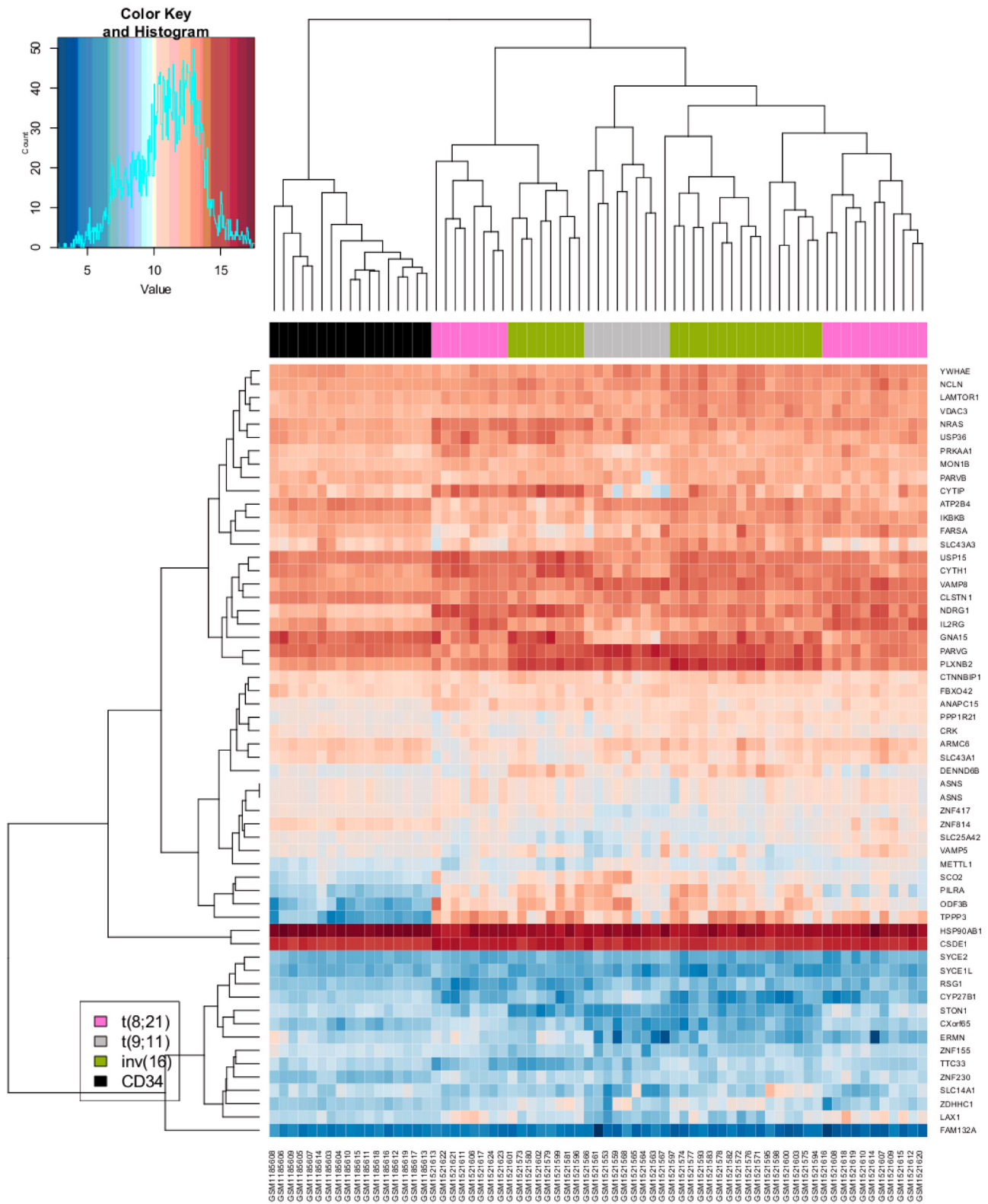


**Figure 5. Chimeric RNA recurrence screening.** (A) Expression of validated chRNA in a large AML cohort. ChRNA classes are identified by colours (green for Class 1, orange for Class 2, red for Class 3 and blue for Class 4). Karyotypes are indicated below the sample name (UK for unknown karyotype, NK for normal karyotype). For abnormal karyotype, chromosomal rearrangement is indicated. The screening was also performed on not treated (NT) or differentiated cell lines. For PML-RARA and CTDP1 exons involved in the chRNA are indicated. (B) Classification of chRNA depends on normal or tumor level expression. ChRNA expression, in normal CD34+ HSCs (CD34) and in AML (LEUCEGENE data), is presented using the tag search approach. For each group (AML-test, AML-AK, AML-NK, CD34), samples displaying a positive chRNA tag count are selected. The average tag expression is calculated with the selected samples and chRNA classified by their relative expression in the CD34 group. The table with highlighted colour indicates for each chimera, within the different cohort, the number of samples presenting the TAG.

**Discussion**

The use of RNA-seq to provide a detailed view of the transcriptome and to detect new RNA transcripts, opens up new opportunities for improving diagnosis and treatment of human diseases. The characterization of new chRNA presents as a great opportunity, as it could reveal new transcriptomic biomarkers for cancer and therefore could be useful in personalized medicine. ChRNAs, also known as “fusion RNA” or “canonical chimeras”<sup>5,26</sup>, are already used in diagnosis, but many other chimeric fusion products generated by transcriptional mechanisms such as read-throughs, cis or trans-splicing<sup>5,7,9</sup>, also have the potential to be used in diagnosis if correctly categorized. In this study, we successfully developed new tools to classify these fusion transcripts methodologically and biologically into chRNA categories and subcategories, and associated them with biological mechanisms.

One of the major challenges in chRNA detection is to distinguish true candidates from false positives during RNA-seq analysis. False positives can result from technical artefacts that occur during sample preparation, mainly produced by reverse transcription or downstream PCR errors<sup>27</sup>. Bioinformatics processes also generate artefacts associated with the algorithm’s approach for mapping raw reads to the complex reference genome. Many attempts have been made to improve the bioinformatics analysis of chRNAs by proposing multistep filtering pipelines including gene annotation, lists of known fusion genes or machine learning approaches to improve prediction<sup>13</sup>. However, pipeline choice remains a difficult task for biologists and bioinformaticians. We have developed a benchmarking system that enables the calibration and selection of pipelines optimised for the detection of fusion RNAs. Our work also entailed development, within Crac, a machine learning model used to optimise the



**Figure 6. Expression profile of genes involved in read-through Class 2 chRNAs.** The gene expression profile is analyzed by an unsupervised clustering method. Normal CD34+ HSC (CD34) and AML subgroups (LEUCEGENE RNA-seq data) were compared by analyzing the expression of a set of genes involved in read-through chRNAs. Parameters for gene selection are described in the Materials and methods section. Samples are identified by colours (black for normal CD34, pink for AML t(8;21), grey for AML t(9;11), green for AML-inv16). Gene names are indicated on the right.

selection of fusion RNAs (personal data, submitted for publication). Crac offers precise prediction of all chRNA categories, and the CracTools pipeline helps the biologist increase biological rate validation, producing a ChimValue that takes into account methodology and annotation.

As mentioned above, what is at stake is the possibility of exploring the “whole chimeric transcriptome” to classify chRNA, and hence identify cancer biomarkers. High throughput studies, like the Cancer Genome Project<sup>6,28,29</sup>, performed on large cohorts with thousands of experiments, have focused on “canonical fusion genes”. Most of the molecular cancer markers are based on fusion genes because of their ease of detection. It is more difficult to identify features that distinguish chRNAs into new categories, and to elucidate their biological mechanisms and functions. Integrating DNA-seq and RNA-seq data to improve classification is a possible solution<sup>5,30</sup>, but it is time and cost consuming. In this work, we confirm that RNA-seq is a good solution for canonical and new chRNA extraction, and propose a classification system based on subcategories and specific expression profiles.

The present study also reveals new chRNA candidates among the well characterized subcategories. We identified a Class 1 PAN3-NONE chRNA transcript associated with a new translocation in a tumor subclone of a characterized Inv(16) AML, that could be used in patient follow-up. We also identified novel PML-RARA isoforms, shorter than the isoforms currently used in diagnosis, which could again be used in patient follow-up. A recent publication revealed that several isoforms can coexist in leukemia cells from the same patient. The authors showed that the ATRA cell response is isoform-dependent, as the short isoform lacks sensitivity to ATRA<sup>31</sup>. MRD and patient follow-up in APL is usually performed by PML-RARA transcript QPCR, and relapse is associated with an increase of bcr1, bcr2 or bcr3 fusion transcripts<sup>32</sup>. Then, it would be useful to have a picture of the complete isoforms, to best address treatment in this context. Though many patients with AML-inv16 or AML-t(15;17) can benefit from effective treatment, some may develop resistance, leading to adverse outcomes. The appearance or increase of fusion transcripts during treatment could be an indicator of such resistance.

Besides the translocation and inversion mechanisms, our pipeline highlights other events that correlate with chromosomal rearrangement and cancer diagnosis and prognosis. We find Class 3 overlap fusion transcripts like FLT3, corresponding to tandem duplication that could be of use in the prognosis and MRD of AML<sup>33</sup>. The real advantages of RNA-seq in highlighting tandem repeat sequences are its open nature and its capacity to detect outside “hotspots”. Furthermore, we also detected fusion transcripts resulting from “read-through transcription”, described in chRNA studies on cancer.

Most newly identified chRNAs used canonical splice sites and were detected in normal hematopoietic tissues. This observation confirms previously published works concerning the recurrence

of chimeric fusion RNAs in healthy cells<sup>34</sup>. It is unlikely that they are the products of genomic abnormalities, since they are expressed in healthy samples. However, the pathogenetic impact of these chimeric fusions remains unclear. Recent findings have demonstrated the role of read-through chRNA in renal carcinoma and breast cancer<sup>26,35</sup>, and our data demonstrates that genes involved in these events are differently expressed in AML. More studies are needed to elucidate the physiopathological impact of these chRNAs.

The potential of NGS technologies, particularly of RNA-seq, in increasing the capabilities of personalized medicine is clear<sup>2</sup>. However, to achieve this, efforts must be made to facilitate the interpretation of complex high throughput data. For chRNA, this is feasible only if the fusion transcripts are well classified and characterized. New technologies are available to simplify disease follow-up at reduced costs. Here, we propose a robust, open method based on a single process to identify different classes of chRNA. This approach provides a chRNA transcriptome map of biomarkers for disease characterization and monitoring including known canonical gene fusions and new chRNA. In combination with a tag-based approach and gene expression profile, this map can give a global picture of the complex physiological processes and could correlate with current leukemia classification.

**Abbreviations:** chRNA, chimeric RNA; AML, acute myeloid leukemia; NK, normal karyotype; UK, unknown karyotype; AK, abnormal karyotype; APL, Acute Promyelocytic Leukemia; PBMCs, peripheral blood mononuclear cells; Inv16, chromosome 16 inversion; t(15;17), translocation of chromosomes 15 and 17; qPCR, quantitative polymerase chain reaction; NONE, non-annotated region; LincRNA, Long intergenic noncoding RNAs; Bcr, break chromosomal region; FISH, Fluorescence *in situ* hybridization; ATRA, all-trans retinoic acid; VD, vitamin D; TGFβ, transforming growth factor beta; MRD, minimal residual disease.

### Data availability

The Crac and CracTools software is hosted on <http://crac.gforge.inria.fr/>.

Raw data (FASTQ files) for OM100011, OM110223 and OS110089 patients are available under accession number [E-MTAB-5767](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=E-MTAB-5767) in the ArrayExpress database at EBI.

The publicly available ENCODE datasets (a detailed list is provided in Supplementary Table S7) used in the study are available on <https://www.ENCODeproject.org/>.

The following LEUCEGENE datasets: GSE62190 (82AML-AK), GSE48846 (17 CD34), GSE49642 (43AML-NK) (a detailed list is provided in [Supplementary Table S6](#)) are available on <https://www.ncbi.nlm.nih.gov/geo/>

GRCh37/Hg19 genome sequences are available at [ftp://ftp.ensembl.org/pub/grch37/release-88/fasta/homo\\_sapiens/dna/](ftp://ftp.ensembl.org/pub/grch37/release-88/fasta/homo_sapiens/dna/).

The annotations GFF file from the ENSEMBL genome browser is available at [ftp://ftp.ensembl.org/pub/grch37/release-84/gff3/homo\\_sapiens/Homo\\_sapiens.GRCh37.82.gff3.gz](ftp://ftp.ensembl.org/pub/grch37/release-84/gff3/homo_sapiens/Homo_sapiens.GRCh37.82.gff3.gz).

### Consent

The patients and healthy donors provided written informed consent to participate in the study, in accordance with the Declaration of Helsinki. The study was approved by the ethics board of Nîmes University (CCPPRB 2002/1103).

### Author contributions

FR, NP and TC initiated the project. JA and NP developed the pipelines. JA, AB and SB performed the analytic work. JA performed clustering analysis. JBG performed fish analysis. EBS, SB and FR performed biological validation. AM, BC, CC provided biological samples, diagnostic tests and specific help for chRNA validation. RA, SR, performed tag RNA-seq data sets selection and tag search approach. NG performed control RT experiments. JML and ALB contributed to scientific discussion. FR, TC wrote the paper. ALB contributed to manuscript revisions. All authors read and approved the final manuscript.

### Competing interests

No competing interests were disclosed.

### Grant information

This work was supported by the French ANR IBC project “Investissement d’avenir en bioinformatique-projet-IBC” and FRM “Appel d’offres urgence pour la bioinformatique, projet DBI20131228566”. This work was supported by the France Génomique National infrastructure, funded as part of the “Investissements d’Avenir” program managed by the Agence Nationale pour la Recherche (ANR-10-INBS-09). We acknowledge la Ligue Nationale Contre le Cancer for financial support (EL2015.LNCC/JML) to JML’s team, and Le Cancéropôle Grand Sud-Ouest (GSO).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

### Acknowledgements

Jérôme Audoux is fellowship of the FRM (Appel d’offres urgence pour la bioinformatique, projet DBI20131228566) and Ronnie Alves are fellowships of the IBC (ANR “Investissement d’avenir en bioinformatique-projet-IBC”); We thank Philippe Clair (Engineer, University of Montpellier, France) for advice on qPCR (qPHD platform, Montpellier GenomiX) and Jean-Marc Holder (CMO SeqOne) for text corrections.

## Supplementary material

### Figure S1: PCR strategy

PCR primers were designed on both sides of the junction point (\*) using spanning junction read or a sequence reconstructed by combining paired-end and chimeric spanning reads (meta-read). SF and SR indicated respectively short forward and short reverse primers, LF and LR, long forward and long reverse primers. Forward (F) and reverse (R) indicated primers pairs designed on spanning meta-reads.

[Click here to access the data.](#)

### Figure S2: NONE transcript amplification strategy

The PAN3-NONE read sequence and its corresponding genomic position on GrCh37 is shown. The junction point indicated the chr2: 43193247 (-) positions for NONE transcript. From this position, we selected an upstream 100 base pairs sequence in order to amplify the putative NONE transcript. Primers were designed on each part of the chr2: 43193247 position to specifically amplify the non-chimeric transcript.

[Click here to access the data.](#)

### Figure S3: PAN3-NONE qPCR amplification

Both PAN3-NONE and CFBF-MYH11 chRNAs were amplified from each sample. CFBF-MYH11 chRNA, was used as a positive control. For each amplicon, the Ct value and the melting temperature (Tm) are shown (upper table).

The second table shows PAN3-NONE and CFBF-MYH11 qPCR results for the time points corresponding to the 2 years follow-up of the OS110089 patient. Clinical information and QPRC results for both fusion transcripts are mentioned. The molecular biology for CFBF-MYH11 (Inv16) performed by the diagnostic laboratory is also indicated as positive or percentage values; DNQ= detected-not quantified; ND= not detected.

[Click here to access the data.](#)

### Figure S4: Expression of ChRNA in different AML groups and hematopoietic cells by tag counting approach.

For each group (AML-test, AML-AK, AML-NK and CD34), the average Tags mean expression is indicated. Tags were designed for the

indicated chRNA as described in materials and methods. Only positive samples were retained to calculate the average tag expression. The tag counts are normalized per 5 billion of total k-mers.

[Click here to access the data.](#)

#### Figure S5: Expression profile of genes involved in Class3 tandem repeat chRNAs.

Heat map representing colour-coded expression levels of class3 chimeras (chRNAs normalized count) across patient samples. Names of genes involved in class3 chRNAs are indicated on the right. Black, pink, green and grey bars at the top represent respectively normal CD34, AML-t(8;21), AML-inv16, AML-t(9;11).

[Click here to access the data.](#)

#### Table S1: Patient data.

ID and patient information for the AML cohort.

[Click here to access the data.](#)

#### Table S2: CRAC & Cractools pipeline output obtained from OS110089, OM100011 and OM110223 RNAseq data analysis.

[Click here to access the data.](#)

#### Table S3: Primers table for chRNA qPCR validation.

[Click here to access the data.](#)

#### Table S4: Validated chRNAs information

The table gives the list of TAGs used for tag search approach. For the chRNA, exons involved in the chimeric junction and tag sequences used for the tag search approach are indicated.

[Click here to access the data.](#)

#### Table S5: NONE and PAN3 tags expression within a set of libraries

Specific tags for PAN3 and non-chimeric NONE transcript corresponding respectively to Chr13(+) : 28813770-28813799 and Chr2(-) : 43193287-43193316 position on GRCh37/Hg19 genome were designed and counted in FASTQ files of 52 libraries from various tissues. Expression was normalized per 5 billion of total k-mers.

[Click here to access the data.](#)

#### Table S6: List of LEUCEGENE and ENCODE datasets used in the study

Table S6 describes the different parts of the LEUCEGENE cohort used in the study. Samples indicated by a GSM number are classified by groups of different Karyotype. Table S6 also indicates the ENCODE dataset used in the study.

[Click here to access the data.](#)

## References

- Maher CA, Palanisamy N, Brenner JC, *et al.*: **Chimeric transcript discovery by paired-end transcriptome sequencing.** *Proc Natl Acad Sci U S A.* 2009; **106**(30): 12353–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Byron SA, Van Keuren-Jensen KR, Engelthaler DM, *et al.*: **Translating RNA sequencing into clinical diagnostics: opportunities and challenges.** *Nat Rev Genet.* 2016; **17**(5): 257–71.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Atlas of Genetics and Cytogenetics in Oncology and Haematology [Internet].** [cité 28 janv 2017].  
[Reference Source](#)
- Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer [Internet].** [cité 28 janv 2017].  
[Reference Source](#)
- Mertens F, Johansson B, Fioretos T, *et al.*: **The emerging complexity of gene fusions in cancer.** *Nat Rev Cancer.* 2015; **15**(6): 371–81.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Yoshihara K, Wang Q, Torres-Garcia W, *et al.*: **The landscape and therapeutic relevance of cancer-associated transcript fusions.** *Oncogene.* 2015; **34**(37): 4845–54.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gingeras TR: **Implications of chimaeric non-co-linear transcripts.** *Nature.* 2009; **461**(7261): 206–11.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jia Y, Xie Z, Li H: **Intergenicallly Spliced Chimeric RNAs in Cancer.** *Trends Cancer.* 2016; **2**(9): 475–84.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Latysheva NS, Babu MM: **Discovering and understanding oncogenic gene fusions through data intensive computational approaches.** *Nucl Acids Res.* 2016; **44**(10): 4487–503.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Djebali S, Lagarde J, Kapranov P, *et al.*: **Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells.** *PLoS One.* 2012; **7**(1): e28213.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Maher CA, Kumar-Sinha C, Cao X, *et al.*: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature.* 2009; **458**(7234): 97–101.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rickman DS, Pflueger D, Moss B, *et al.*: **SLC45A3-ELK4 Is a Novel and Frequent Erythroblast Transformation-Specific Fusion Transcript in Prostate Cancer.** *Cancer Res.* 2009; **69**(7): 2734–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Beaumeunier S, Audoux J, Boureux A, *et al.*: **On the evaluation of the fidelity of**



- supervised classifiers in the prediction of chimeric RNAs. *BioData Min.* 2016; **9**: 34.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Philippe N, Salson M, Commes T, *et al.*: **CRAC: an integrated approach to the analysis of RNA-seq reads.** *Genome Biol.* 2013; **14**(3): R30.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  15. **LEUCEGENE Project [Internet].** [cité 28 janv 2017].  
[Reference Source](#)
  16. Piquemal D, Commes T, Manchon L, *et al.*: **Transcriptome analysis of monocytic leukemia cell differentiation.** *Genomics.* 2002; **80**(3): 361–71.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  17. Quere R, Baudet A, Cassinat B, *et al.*: **Pharmacogenomic analysis of acute promyelocytic leukemia cells highlights CYP26 cytochrome metabolism in differential all-trans retinoic acid sensitivity.** *Blood.* 2007; **109**(10): 4450–60.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  18. Defacque H, Piquemal D, Basset A, *et al.*: **Transforming growth factor-beta1 is an autocrine mediator of U937 cell growth arrest and differentiation induced by vitamin D3 and retinoids.** *J Cell Physiol.* 1999; **178**(1): 109–19.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  19. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–7.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  20. Philippe N, Bou Samra E, Boureux A, *et al.*: **Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome.** *Nucleic Acids Res.* 2014; **42**(5): 2820–32.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  21. van Dongen JJ, Macintyre EA, Gabert JA, *et al.*: **Standardized RT-PCR analysis of fusion gene transcripts from chromosome aberrations in acute leukemia for detection of minimal residual disease. Report of the BIOMED-1 Concerted Action: investigation of minimal residual disease in acute leukemia.** *Leukemia.* 1999; **13**(12): 1901–28.  
[PubMed Abstract](#)
  22. Gabert J, Beillard E, van der Velden VH, *et al.*: **Standardization and quality control studies of 'real-time' quantitative reverse transcriptase polymerase chain reaction of fusion gene transcripts for residual disease detection in leukemia – A Europe Against Cancer Program.** *Leukemia.* 2003; **17**(12): 2318–57.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  23. Walz C, Grimwade D, Saussele S, *et al.*: **Atypical mRNA fusions in PML-RARA positive, RARA-PML negative acute promyelocytic leukemia.** *Genes Chromosomes Cancer.* 2010; **49**(5): 471–9.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  24. Pandolfi PP, Alcalay M, Fagioli M, *et al.*: **Genomic variability and alternative splicing generate multiple PML/RAR alpha transcripts that ENCODE aberrant PML proteins and PML/RAR alpha isoforms in acute promyelocytic leukaemia.** *EMBO J.* 1992; **11**(4): 1397–407.  
[PubMed Abstract](#) | [Free Full Text](#)
  25. Grosso AR, Leite AP, Carvalho S, *et al.*: **Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma.** *eLife.* 2015; **4**: pii: e09214.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  26. Jividen K, Li H: **Chimeric RNAs generated by intergenic splicing in normal and cancer cells.** *Genes Chromosomes Cancer.* 2014; **53**(12): 963–71.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  27. Peng Z, Yuan C, Zellmer L, *et al.*: **Hypothesis: Artifacts, Including Spurious Chimeric RNAs with a Short Homologous Sequence, Caused by Consecutive Reverse Transcriptions and Endogenous Random Primers.** *J Cancer.* 2015; **6**(6): 555–67.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  28. International Cancer Genome Consortium, Hudson TJ, Anderson W, *et al.*: **International network of cancer genome projects.** *Nature.* 2010; **464**(7291): 993–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  29. Cancer Genome Atlas Research Network, Ley TJ, Miller C, *et al.*: **Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia.** *N Engl J Med.* 2013; **368**(22): 2059–74.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  30. Zhang J, White NM, Schmidt HK, *et al.*: **INTEGRATE: Gene fusion discovery using whole genome and transcriptome data.** *Genome Res.* 2016; **26**(1): 108–18.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  31. Tan Y, Bian S, Xu Z, *et al.*: **The short isoform of the long-type PML-RARA fusion gene in acute promyelocytic leukaemia lacks sensitivity to all-trans-retinoic acid.** *Br J Haematol.* 2013; **162**(1): 93–7.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  32. Cassinat B, de Botton S, Kelaidi C, *et al.*: **When can real-time quantitative RT-PCR effectively define molecular relapse in acute promyelocytic leukemia patients? (Results of the French Belgian Swiss APL Group).** *Leuk Res.* 2009; **33**(9): 1178–82.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  33. Bibault JE, Figeac M, Hélevaut N, *et al.*: **Next-generation sequencing of FLT3 internal tandem duplications for minimal residual disease monitoring in acute myeloid leukemia.** *Oncotarget.* 2015; **6**(26): 22812–21.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  34. Babiceanu M, Qin F, Xie Z, *et al.*: **Recurrent chimeric fusion RNAs in non-cancer tissues and cells.** *Nucl Acids Res.* 2016; **44**(6): 2859–72.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  35. Varley KE, Gertz J, Roberts BS, *et al.*: **Recurrent read-through fusion transcripts in breast cancer.** *Breast Cancer Res Treat.* 2014; **146**(2): 287–97.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

# Open Peer Review

Current Referee Status:  

---

## Version 2

Referee Report 16 March 2018

doi:10.5256/f1000research.14532.r29147



**Charles Gawad** 

Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA

I am satisfied with the authors responses and feel the report is now suitable for indexing.

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

## Version 1

Referee Report 03 October 2017

doi:10.5256/f1000research.12254.r25306



**Charles Gawad** 

Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA

In this report, Ruffle et al. present a new approach for identifying novel chimeric transcripts using CracTools. They classify the transcripts into four categories: 1) different chromosomes (interchromosomal translocations), 2) co-linear but from different genes (putative transcriptional read through or complex intrachromosomal structural variation), 3) exons from same chromosome but are not in the expected order based on the reference (tandem duplication, complex intrachromosomal rearrangement), or 4) exons on same chromosome but different strands. The authors then go on to validate a subset of putative new chimeric transcripts using RT-PCR and FISH. These types of studies have been previously performed. One new aspect of this study is the stranded library preparation which allows for the identification of the class 4 chimeras. In addition, they tried to minimize the bias in their analysis pipeline by not relying on a reference genome, which enabled the discovery of new transcripts. Overall, their approach provides a validated new strategy for identifying novel transcripts in RNA-seq data.

### Major Concerns

1. The authors do not discuss circular RNA, which are likely to make up a large portion of their class 3 chimeras as found in many recent studies. The low numbers of class 3 chimeras also raises

concerns about the sensitivity of the approach, as most recent studies have found thousands of circular RNA isoforms per sample. The total RNA underwent RT with random primers, which would retain circular RNA. It is not clear if there was a polyA-selection step after that point or if the total RNA was ribosomal depleted. If it was the former, the circular transcripts would not be present.

2. If the RNA was not polyA-selected, the authors should specifically discuss the PML-RARA circular transcripts recently discovered in acute promyelocytic leukemia. I am not aware of any independent validation of that work.

#### Minor Concerns

1. The authors should read the manuscript closely for typos. For example in the AML samples and cells lines section there are commas where there should be periods, in the FISH methods section there are degree signs instead of percent, and two paragraphs before the discussion there is MDR instead of MRD.
2. The authors should include the kits used for ribosomal RNA-depletion/polyA-selection, as well as stranded library preparation.
3. What mechanisms do the authors have in mind for a structural variant and/or alternative splicing event that would result in class 4 chimeras, as they would not occur as a result of transcriptional read through of the same strand?
4. The resolution is too low for Figure 5A.

#### **References**

1. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO: Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One*. 2012; **7** (2): e30733 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE: Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*. 2013; **19** (2): 141-57 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Guarnerio J, Bezzi M, Jeong JC, Paffenholz SV, Berry K, Naldini MM, Lo-Coco F, Tay Y, Beck AH, Pandolfi PP: Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations. *Cell*. 2016; **165** (2): 289-302 [PubMed Abstract](#) | [Publisher Full Text](#)

#### **Is the work clearly and accurately presented and does it cite the current literature?**

Partly

#### **Is the study design appropriate and is the work technically sound?**

Yes

#### **Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

#### **If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

#### **Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 05 Dec 2017

**Therese Commes**, Universite de Montpellier, France

We thank the referee for his careful reading of our manuscript and for his remarks which led us to clarify the protocol used in this work.

### Major Concerns

- *The authors do not discuss circular RNA, which are likely to make up a large portion of their class 3 chimeras as found in many recent studies. The low numbers of class 3 chimeras also raises concerns about the sensitivity of the approach, as most recent studies have found thousands of circular RNA isoforms per sample. The total RNA underwent RT with random primers, which would retain circular RNA. It is not clear if there was a polyA-selection step after that point or if the total RNA was ribosomal depleted. If it was the former, the circular transcripts would not be present.*

Response: The remark is relevant and we agree with the referee that it was not clearly stated that we performed polyA selection for the RNAseq experiment. This point is only described in online data availability (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5767/>) and doesn't appear explicitly in mat&methods section. We added the following comment on page 4: "*The RNAseq was performed using polyA-selection with the TruSeq RNA Lib-Prep Kit (Illumina) adjusted with GATC specific procedure for strand specificity*".

We also agree with the referee that class3 fusion transcripts could arise from circular RNA (circRNA) when performing ribosomal depleted RNA-seq. However, as our polyA+ RNA-seq study doesn't enable the identification of this RNA subtype, we do not discuss about circRNA in the manuscript. Concerning class3 chimeric RNA, our pipeline detects a great number of candidates before applying specific filters. This category, as well as the class2, is the most represented in all datasets we analyzed. As described in mat&methods section, we used stringent criteria to minimize bias and greatly reduced the number of candidates.

For example, from a typical AML stranded and paired- end RNA-seq experiment (50 Millions reads; 100pb length), starting from the raw data (CRAC and CracTools process) we extracted 1406 chRNAs including 35% of class3chRNA reduced by 16 fold with the filtering process (26 class3chRNA).

- *If the RNA was not polyA-selected, the authors should specifically discuss the PML-RARA circular transcripts recently discovered in acute promyelocytic leukemia. I am not aware of any independent validation of that work.*

Response: As we choose polyA RNA-seq, we can only characterize new linear fusion transcripts.

From ribo-depleted RNAseq data analysis, the characterization of circRNA arising from fusion gene requires the identification of linear junction (lin-J) and circular junction (circ-J) described as spliced and back spliced junction respectively by Guarnerio et al (*Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations*, 2016; *Cell* 165 (2): 289-302). The linear and circular fusion transcripts share the same linear junction. The new (3-12) PML-RARA junction we discover corresponds necessarily to a linear junction because of its fusion junction sequence (see figure at <https://bio2m.montp.inserm.fr/papers/AMLchimera2017>). Moreover, in our polyA+ dataset it surely corresponds to a linear transcript. It is most probably an alternative splicing product transcribed from the PML-RARA bcr3 translocation since it coexists with the well-known (3-9) PML-RARA bcr3 transcript. However we could not exclude in the biological sample the presence of circRNA emerging from the PML-RARA fusion gene. However polyA RNAseq protocol doesn't enable to reveal them.

To answer to the referee remark, we performed complementary experiments and analyzed the Guarnerio dataset (*BioProjectID: PRJNA315254*) with our pipeline to detect fusion junctions (lin-J and Circ-J) specific to circRNA arising from the PML-RARA translocated genes. We found 3 PML-RARA and 1 RARA-PML linear fusion junctions but we did not detect the circRNA ones. We performed a tag search approach specific to the PML-RARA circ-junction (F-circ1) described in Fig S1B supplemental information of Guarnerio et al manuscript. We did not find it in their fastQ files. To conclude we were unable to detect PML-RARA circRNA in Guarnerio J dataset but we confirmed the data recently described by You, X. and Conrad, T. OF with the specific circRNA Acfs pipeline (*Acfs: accurate circRNA identification and quantification from RNA-Seq data*, *Scientific Report* 6,38820; doi: 10.1038/srep38820; 2016). The circRNA search for low abundance will certainly require more deepness in RNAseq.

#### Minor Concerns

- *The authors should read the manuscript closely for typos. For example in the AML samples and cells lines section there are commas where there should be periods, in the FISH methods section there are degree signs instead of percent, and two paragraphs before the discussion there is MDR instead of MRD.*

Response: We read carefully the manuscript and corrected the typos.

- *The authors should include the kits used for ribosomal RNA-depletion/polyA-selection, as well as stranded library preparation.*

Response: We added the following comment on page 4: "The RNAseq was performed using polyA-selection with the TruSeq RNA Lib-Prep Kit (Illumina, San Diego, CA) adjusted with GATC specific procedure for strand specificity".

- *What mechanisms do the authors have in mind for a structural variant and/or alternative splicing event that would result in class 4 chimeras, as they would not occur as a result of transcriptional read through of the same strand ?*

Response: Two mechanisms would result in class 4 chimeras. The first one could involve chromosomal duplication and inversion as described in Newman et al (*Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints*, *Am J Hum Genet.* 2015, Feb 5; 96(2):208-20. doi:10.1016/j.ajhg. 2014.12.017. Epub 2015 Jan 29). The second one could involve splicing events as described by Gingeras et al (*Implications of chimaeric non-co-linear transcripts*, *Nature.* 2009 Sep 10; 461(7261):206-11 doi: 10.1038/nature08452.)

- *The resolution is too low for Figure 5A.*

Response: We changed the resolution of the figure 5A, increasing the font size.

**Competing Interests**: No competing interests were disclosed.

Referee Report 29 August 2017

doi:10.5256/f1000research.12254.r25101



**Hui Li**

Department of Pathology, School of Medicine, University of Virginia, Charlottesville, VA, USA

Ruffle et al., performed RNA-Seq on three AML patients, and identified a number of new chimeric RNAs. They grouped them into reasonable categories, and discussed their generating mechanisms. For some, they performed tag search in larger sets of RNA-Seq data of AML and CD34+ HSCs. In general, the analysis is solid. Multiple levels of evidence were provided for some fusion RNAs (for instance FISH for a Class I fusion). The conclusion is justified.

No major issues were noted.

Minor suggestions:

1. It would be nice to perform the same RNA-Seq pipeline analysis on bigger datasets, such as the data from LEUCEGENE.
2. It is well known that different software tools behave differently and false positive/negative is a big issue. Would be nice to use another independent software for crosschecking.
3. The METRNL001 Class 3 chimera may be a product of backsplicing if the sequencing protocol is not restricted to polyA RNAs.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 17 Nov 2017

**Therese Commes**, Universite de Montpellier, France

**Response to referee Report ( Hui LI, Department of Pathology, School of Medicine, University of Virginia, Charlottesville, VA, USA)**

1. *It would be nice to perform the same RNAseq pipeline analysis on bigger datasets, such as the data from leucegene.*
2. *It is well known that different software tools behave differently and false positive /negative is a big issue. Would be nice to use another independent software for crosschecking.*
3. *The METRNL001 class3 chimera may be a product of backsplicing if the sequencing protocol is not restricted to polyA RNAs*

We want to thank the referee for taking time to read the manuscript and provide constructive feedback.

1. The first remark concerning LEUCEGENE analysis is relevant as this cohort is a unique dataset of 450 RNA-seq from AML patients and normal hematopoietic cells. This cohort includes diverse types of AML and clinical data with a deepness of 100-300x10<sup>6</sup> reads/per sample. However, our first aim in the present manuscript was to test our pipeline and to propose a new way to analyze, select and classify chRNA. We thus choose to analyze a small dataset and used the Leucegene cohort only to explore specific expression of the new chRNAs (revealed with our pipeline). This Expression with a tag search approach was based on their recurrence, tumor, subgroup, or patient-specific expression. The complete leucegene chRNA research represents a big task and constitutes a full-fledged work, we are now performing our RNAseq pipeline on this dataset. In this new task, we will check at a large scale the pipeline performance, in term of memory computing and time consuming, and also validate its capacity to detect all the well-known AML fusion genes. We then should be able to discover new chimeric RNA in AML.

2. We totally agree with the referee and are well aware of potential bias generated with different pipelines. Our group is particularly concerned about software comparison for chimeric RNA detection. CRAC was developed in the Lab and compared with other software before we choose it for this study (*Philippe et al, 2013, Genome Biology; an integrated approach to the analysis of RNA-seq reads genome Biology*) and we recently published in September 2017 a benchmark method dealing with this topic (*Audoux et al, 2017, BMC Bioinformatics; SimBA: A methodology and tools for evaluating the performance of RNA-Seq bioinformatic pipelines*). CRAC tools has been developed in the Lab in response to biologists requests to better characterize, filter and classify predicted fusion transcripts. Moreover, we recently compared the pipeline CRAC and CRACtools with Fusioncatcher in our AML dataset, as it provides interesting and complementary annotation (overlapping databases, protein impact) but the process is time consuming and the filtering step is more stringent, many of the new fusion transcripts we validated are ruled out.

Whatever the bioinformatics pipeline used, bias will subsist. In order to improve chRNA discovery,

we undertook to couple the bioinformatics approach with biological information and propose a tag search approach targeting the chimeric junctions in normal and tumoral large data set. We demonstrate that this complementary approach works well and will help the biologist to determine the biological relevance of chimeric events.

3. Since we use polyA+ RNA to perform RNAseq, the METRNL001 class3 chimera, which involves a single exon (end of exon 2 METRNL transcribed before the start of the same exon) more probably corresponds to a linear polyA+ transcript involving a short repeat sequence of 5' part of exon 2 as described in Fig4 rather to a circularization of a single exon formed by backsplicing.

**Competing Interests:** No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

F1000Research