RNAmut: robust identification of somatic mutations in acute myeloid leukemia using RNA-sequencing

Acute myeloid leukemia (AML) is an aggressive malignancy of haematopoietic stem cells driven by a well-defined set of somatic mutations.[1,2] Identifying the mutations driving individual cases is important for assigning the patient to a recognized World Health Organisation category, establishing prognostic risk and tailoring post-consolidation therapy.[3] As a result, AML research and diagnostic laboratories apply diverse methodologies to detect important mutations and many are introducing next-generation sequencing (NGS) approaches to study extended panels of genes in order to refine genomic classification and prognostic category.[1] Besides the implications of these developments on costs, expertise and reliance on commercial providers, they also do not capture gene expression data, which have independent prognostic value that cannot be inferred from somatic mutation profiles. The ability to detect AML gene mutations as well as gene expression profiles from a single assay, could provide a holistic tool that accelerates research, simplifies diagnostic work-up and helps develop integrated algorithms to refine individual patient prognosis. Here, we show that AML RNA sequencing (RNA-seq) data can be used to reliably detect all types of clinically important mutations and develop a bespoke fast and easy-to-use software (RNAmut) for this purpose that can be readily used by teams/laboratories without in-house bioinformatic expertise.

We focused on detection of mutations in 33 genes that are relevant to AML classification and prognosis (Table 1)

Table 1. Genes and types of mutations detected by RNAmut. For acute myeloid leukemia fusions the eight most common partners were searched for.

Genes	Hotspots	Indel & SNV	Tandem duplication	Gene Fusion
NPM1	W288fs	Yes		
FLT3	D835-D839	Yes	*FLT3*-ITD	
IDH1	R132	Yes		
IDH2	R140, R172	Yes		
CEBPA		Yes		
TET2		Yes		
DNMT3A	R882	Yes		
RUNX1		Yes		
TP53		Yes		
ASXL1		Yes		
WT1		Yes		
BCOR		Yes		
SRSF2		Yes		
SF3B1		Yes		
U2AF1		Yes		
KMT2A (MLL)			*MLL*-PTD	*MLL-partners*
PML				*PML-RARA*
RARA				*PML-RARA*
MYH11				*MYH11-CBFB*
CBFB				*MYH11-CBFB*
RUNX1T1				*RUNX1-RUNX1T1*
BCR				*BCR-ABL*
ABL1				*BCR-ABL*
NUP98				*NUP98-NSD1*
NSD1				*NUP98-NSD1*
MLLT1				*KMT2A-MLLT1*
AFF1 (MLLT2)				*KMT2A-AFF1*
MLLT3				*KMT2A-MLLT3*
AFDN (MLLT4)				*KMT2A-AFDN*
EPS15 (MLLT5)				*KMT2A-EPS15*
ELL				*KMT2A-ELL*
MLLT10				*KMT2A-MLLT10*
MLLT11				*KMT2A-MLLT11*

SNV: single-nucleotide variant.

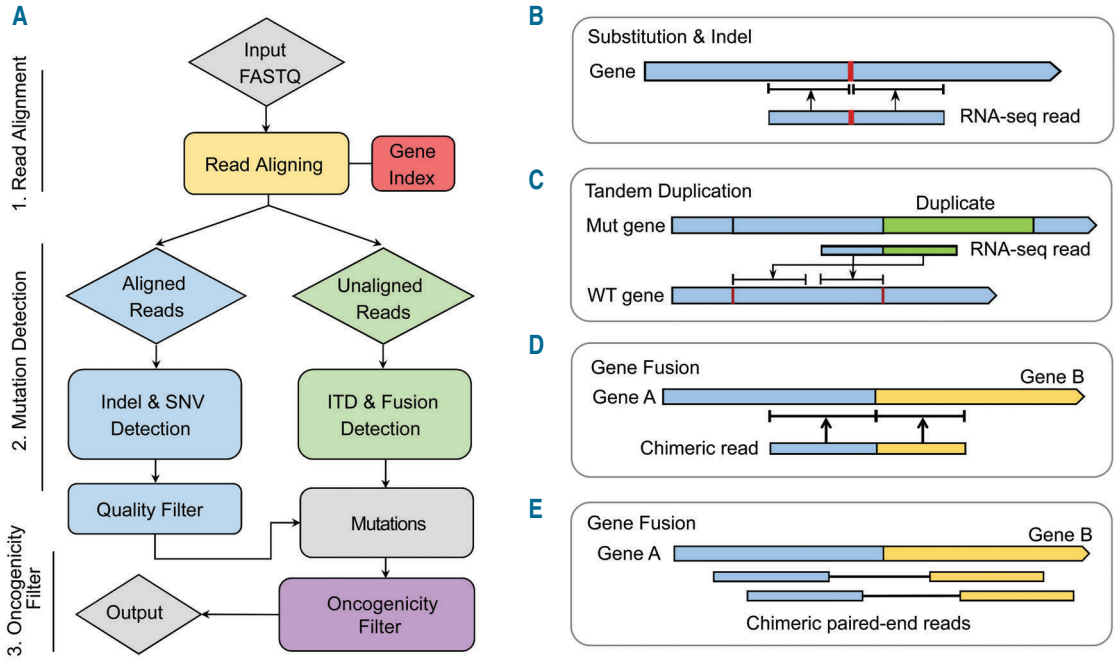

Figure 1. Schematic depiction of the RNAmut pipeline. (A) Pipeline flowchart. (B-E) Summarized explanation of detection strategies for (C) tandem duplications, (D) gene fusions using chimeric reads that capture the breakpoint and (E) gene fusions using chimeric paired-end reads. Detailed explanations are given in the *Online Supplementary Materials and Methods*.

and designed the software pipeline to operate in three stages: read alignment, mutation detection and an additional oncogenicity filter (Figure 1A). To ensure fast alignment of RNA-seq reads, we indexed all possible 10-mer sequences from our target genes into a look-up table (hash function) that maps the 10-mers to their locations on the 33 genes (*Online Supplementary Figure S1A*). We used 10-mers (instead of 9-mers or 11-mers *etc.*) for optimal balance between speed and memory requirements (*Online Supplementary Table S1*). To align RNA-seq reads, the sequence of each read is divided into consecutive 10-mers and each 10-mer is mapped to genic locus/loci using the pre-built look-up table. By examining all 10-mers in a read, RNAmut computes whether the read is perfectly aligned (Type A), aligned with mismatches (Type M) or not aligned (Type N; *Online Supplementary Figure S1B*). M-type reads are used to detect substitutions and small indels (Figure 1B). To detect tandem duplications, RNAmut uses the subset of N-type reads for which the 5' end is mapped downstream of their 3' end, and computes the location of the duplicated region (Figure 1C). Gene fusions are detected through two independent pieces of evidence: first, reads spanning the breakpoint (*i.e.* chimeric reads) are extracted from the N-type reads and used to report the precise location of the breakpoint (Figure 1D). Secondly, fusion genes can also be identified from paired-end RNA-seq reads when each of the two paired reads aligned to a different fusion partner (Figure 1E). All mutations covered by ≥3 unique reads are reported and these are then optionally parsed through an oncogenicity filter applying the criteria used by the largest AML sequencing study published to date[1] (*Online Supplementary Table S2*), which could be especially useful for diagnosticians. Full details of the RNAmut pipeline are given in the *Online Supplementary Materials and Methods*. To benchmark read mapping, we compared RNAmut's alignment with commonly used read aligners.[4-6] Our alignment showed very good agreement with panel-restricted alignments by BWA (*Online Supplementary Figure S12A-B*) and Salmon (*Online Supplementary Figure S14*), and global alignment by STAR (*Online Supplementary Figure S13*), for all of which both Pearson correlation and gradient were very close to 1.

To test the performance of our RNAmut, we analyzed 151 RNA-seq datasets from AML bone marrow samples generated by the Cancer Genome Atlas (TCGA)[2] and detected 40 *NPM1*, 37 *FLT3*-ITD, 35 *DNMT3A*, 17 *IDH2*, 13 *IDH1*, 17 *RUNX1*, 17 *CEBPA*, 13 *TP53*, 13 *TET2*, 10 *FLT3* TKD, 7 *MLL*-PTD, 11 *WT1*, 3 *ASXL1*, 1 *BCOR*, 12 *SRSF2*, 3 *SF3B1* and 7 *U2AF1* mutations, along with 15 *PML-RARA*, 10 *MYH11-CBFB*, 7 *RUNX1-RUNX1T1*, 3 *BCR-ABL1*, 3 *NUP98-NSD1* fusions and 8 *MLL (KMT2A)* fusions with various partners (*Online Supplementary Figure S4A* and *Online Supplementary Materials and Methods*). Notably RNAmut accurately detects the lengths and positions of duplicated regions of *FLT3*-ITD (*Online Supplementary Materials and Methods* and *Online Supplementary Figure S8A-B*) while also reporting the number of mutated and WT reads and allelic frequencies (*Online Supplementary Figure S8C*). To assess the accuracy of our software, we compared our results with the mutations detected in these samples by Ley *et al.*[2] Our software detected 289 of the 291 reported mutations (Figure 2). The two cases that we failed to detect were an *IDH1* R132C in TCGA-AB-2984 and an *MLL*-PTD exon2-8 in TCGA-AB-2977. *IDH1* R132C was missed due to the gene's low expression in this sample: only five good quality reads covered R132 of which only one was mutated and thus does not meet the minimum of three mutant reads required by RNAmut (*Online Supplementary Figure*

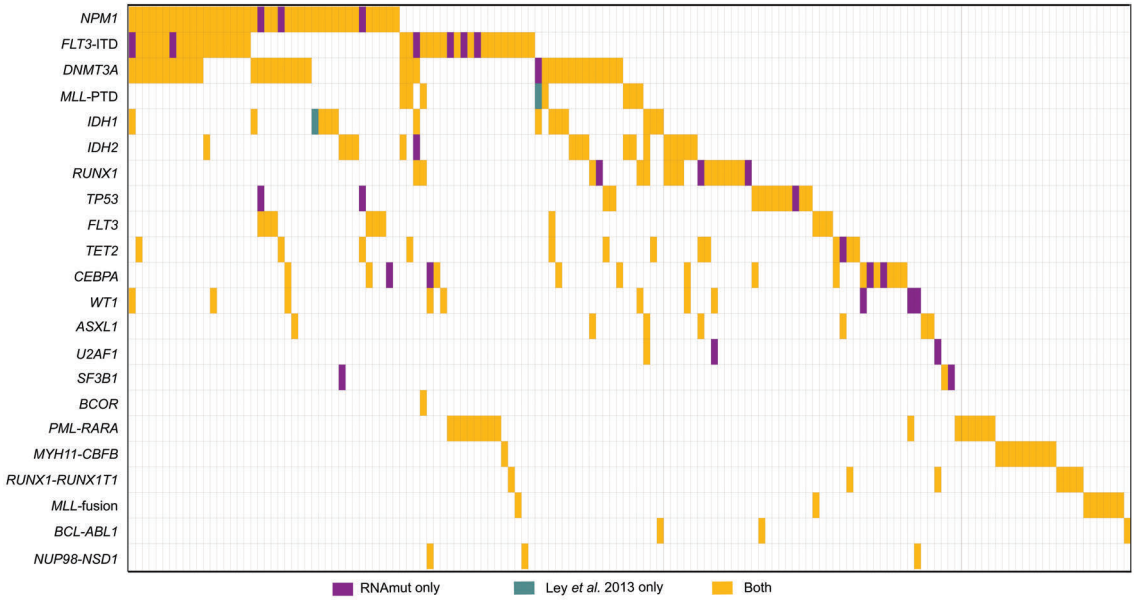

Figure 2. Assessment of our software's accuracy against previously published annotations. Results of testing RNAmut against the 151 RNA-seq datasets from the acute myeloid leukemia (AML) cohort of TCGA. Mutations detected by both our software and Ley *et al.* 2013 are depicted in yellow, additional mutations detected only by our software in purple and the two mutations missed by our software but detected by Ley *et al.* 2013 in green. Mutations in *SRSF2* (not called by Ley *et al.*) are omitted from the plot and are shown in the *Online Supplementary Appendix (Excel File)*. The three samples with no mutations detected are not shown. Details of the specific mutations in individual samples are provided in the *Online Supplementary Figures*.

S18). To examine the missed *MLL*-PTD, we constructed the nucleotide sequence of the reported exon2-8 junction and found no RNA-seq reads reporting such a junction, indicating this may have been an annotation error. Moreover, our software identified 29 samples with mutations that were not reported by Ley *et al.* (Figure 2). For all these samples, we found evidence at the level of RNA and, where available, also DNA (eight samples with whole exome sequencing data) to show that they are indeed true positives (*Online Supplementary Figure S19-20,* and *Online Supplementary Table S4*). To further demonstrate the robustness of RNAmut, we tested its performance on the RNA-seq data from two other sources: i) a set of 164 myelodysplastic syndrome (MDS) patients[7,8] and ii) 437 AML patients studied by the Leucegene consortium,[9-12] both derived from bone marrow samples. For the MDS samples, RNAmut detected all panel-gene mutations identified through targeted DNA sequencing by the authors (*Online Supplementary Figure S5*), as well as 34 mutations that were not (*Online Supplementary Figure S7*). For Leucegene where mutation data are not available on a per sample basis, RNAmut detected similar landscapes of mutations overall (*Online Supplementary Figure S4B* and *S6*) including all 27 instances of an NPM1 mutation reported in one of the consortium's publications.[11]

To validate our method, we first checked and confirmed that all exon sequences of the 33 panel genes are unique in the transcriptome, ruling out the possibility that RNA fragments from non-panel genes are mistakenly aligned to the panel (*Online Supplementary Figures S9-11*). To benchmark mutation calling, we compared RNAmut with commonly used mutation callers. Our variant allele frequency (VAF) calculation agreed very closely with Samtools[13] for substitutions (*Online Supplementary Figure S15A*) and with Varscan[14] for both substitutions (*Online Supplementary Figure S15B*) and indels (*Online Supplementary Figure S15C*). Furthermore, RNAmut detected all gene fusions identified by Fuseq[15] and displayed better sensitivity for detection of MLL fusions (*Online Supplementary Table S3*). Finally, we also compared the VAF detected in RNA-seq with the ones detected in whole exome DNA sequencing data and observed a good correlation for most substitutions (*Online Supplementary Figure S16*). Nonsense mutations in *DNMT3A* (n=3) and *TP53* (n=1) had lower RNA than DNA VAF, possibly due to nonsense-mediated decay (NMD). Nevertheless, a scan of all gain-of-stop codon mutations showed that transcripts potentially subjected to NMD were within detectable levels in AML RNA-seq datasets (*Online Supplementary Figure S17*).

Whilst somatic mutation detection from RNA-seq data is not a novel concept,[16] existing software packages are designed for whole-transcriptome detection, which requires significantly larger memory and long computation time.[17] Also, the lack of integrated pipelines demands intensive scripting and manual adjustment of parameters. Moreover, most existing packages are restricted to the UNIX system, which excludes the Windows user base and with it, most laboratories without in-house bioinformatic expertise. In this study, we present RNAmut, a fast, memory efficient and platform-independent software, which can run on personal computers including laptops and takes less than 30 minutes to detect all types of mutations affecting the selected 33 AML genes, from a typical RNA-seq dataset of 100 million paired-end reads (*Online Supplementary Table S1*). RNAmut can be easily extended to other malignancies by adding or removing genes from its gene index. In addition, it has the option to operate through a graphical user interface, making it accessible to users without any programming knowledge and is freely available in GitHub as a Java application. (*https://github.com/muxingu/rnamut*). Users of RNAmut

should be aware of its limitations such as the fact that it is not designed to detect copy number variations or indels longer than 30 bp other than *FLT3*-ITD or *MLL*-PTD, and as it relies on transcribed RNA it cannot identify intronic or intergenic single-nucleotide variants (SNV). Furthermore, users of customized gene panels will need to ensure that their genes of interest are expressed sufficiently for RNAmut to detect any mutation, which is a general limitation of all RNA-seq based mutation callers.

The current molecular diagnosis of AML relies on multidisciplinary workflows including cytogenetic, molecular and NGS tests in order to detect different types of mutations. In this study, we demonstrate that all diagnostically important somatic mutations in AML can be reliably detected from RNA-seq within one single workflow. Our bespoke software, RNAmut, greatly reduces the difficulty and time required to analyse NGS data with results that match or even out-perform current methods. As our approach can be readily combined with information such as gene expression and splicing from the same RNA-seq dataset, it can be used to generate integrated algorithms that enhance prognostication and patient treatment. Furthermore, as RNA sequencing is a relatively straightforward procedure, our approach can readily be taken up by the AML research community and also by clinical laboratories, for whom it can significantly reduce experimental costs and accelerate AML genomic diagnosis and classification.

Muxin Gu,[1,2] Maximillian Zwiebel,[1,3] Swee Hoe Ong,[4] Nick Boughton,[5] Josep Nomdedeu,[1,2,6] Faisal Basheer,[1,2,7] Yasuhito Nannya,[8] Pedro M. Quiros,[1,2] Seishi Ogawa,[8] Mario Cazzola,[9] Roland Rad,[3] Adam P. Butler,[4] MS Vijayabaskar[1,2] and George S. Vassiliou[1,2,7]

[1]Haematological Cancer Genetics, Wellcome Sanger Institute, Hinxton, Cambridge, UK; [2]Wellcome Trust–MRC Stem Cell Institute, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK; [3]German Consortium for Translational Cancer Research (DKTK), Partnering Site, Munich, Germany; [4]Cancer Ageing and Somatic Mutation, Wellcome Sanger Institute, Hinxton, Cambridge, UK; [5]Core Software Services, Wellcome Sanger Institute, Hinxton, Cambridge, UK; [6]Hospital de la Santa Creu I Sant Pau, Barcelona, Spain; [7]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge, UK; [8]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan and [9]Fondazione IRCCS Policlinico San Matteo and University of Pavia, Pavia, Italy

Correspondence: GEORGE VASSILIOU - gsv20@sanger.ac.uk MS VIJAYABASKAR - vm11@sanger.ac.uk

doi:10.3324/haematol.2019.230821

Funding: this project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116026. This Joint Undertaking receives support from the European Union´s Horizon 2020 research and innovation and EFPIA. MG is funded by Horizon 2020 (No. 116026) and Cancer Research UK (C22324/A23015). GSV is funded by a Cancer Research UK Senior Cancer Research Fellowship (C22324/A23015) and work in his laboratory is also funded by the Wellcome Trust, European Research Council, Kay Kendall Leukaemia Fund, Bloodwise, The Leukemia Lymphoma Society and the Rising Tide Foundation for Clinical Cancer Research. MSV is funded by the Wellcome Trust (WT098051). The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Information on authorship, contributions, and financial & other disclosures was provided by the authors and is available with the online version of this article at www.haematologica.org.

References

1. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med. 2016;374(23):2209-2221.
2. Cancer Genome Atlas Research N, Ley TJ, Miller C, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013;368(22):2059-2074.
3. Bullinger L, Dohner K, Dohner H. Genomics of acute myeloid leukemia diagnosis and pathways. J Clin Oncol. 2017;35(9):934-946.
4. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760.
6. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-419.
7. Shiozawa Y, Malcovati L, Galli A, et al. Gene expression and risk of leukemic transformation in myelodysplasia. Blood. 2017;130(24):2642-2653.
8. Shiozawa Y, Malcovati L, Galli A, et al. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. Nat Commun. 2018;9(1):3649.
9. Lavallee VP, Lemieux S, Boucher G, et al. RNA-sequencing analysis of core binding factor AML identifies recurrent ZBTB7A mutations and defines RUNX1-CBFA2T3 fusion signature. Blood. 2016;127(20):2498-2501.
10. Macrae T, Sargeant T, Lemieux S, Hebert J, Deneault E, Sauvageau G. RNA-Seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. PLoS One. 2013;8(9):e72884.
11. Pabst C, Bergeron A, Lavallee VP, et al. GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. Blood. 2016;127(16):2018-2027.
12. Simon C, Chagraoui J, Krosl J, et al. A key role for EZH2 and associated genes in mouse and human adult T-cell acute leukemia. Genes Dev. 2012;26(7):651-656.
13. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079.
14. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-576.
15. Vu TN, Deng W, Trac QT, Calza S, Hwang W, Pawitan Y. A fast detection of fusion genes from paired-end RNA-seq data. BMC Genomics. 2018;19(1):786.
16. Neums L, Suenaga S, Beyerlein P, et al. VaDiR: an integrated approach to Variant Detection in RNA. Gigascience. 2018;7(2).
17. Coudray A, Battenhouse AM, Bucher P, Iyer VR. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. PeerJ. 2018;6:e5362.