Acute myeloid leukemia risk stratification in younger and older patients through transcriptomic machine learning models

Raíssa Silva[1,2], Cédric Riedel[1,2], Maïlis Amico[1,3], Jerome Reboul[1,2], Benoit Guibert[1,2], Camelia Sennaoui[1,2], Florence Ruffle[1,2], Nicolas Gilbert[1,2], Anthony Boureux[1,2], and Thérèse Commes[1,2,*]

[1]Université de Montpellier, Montpellier, 34000, France
[2]Institute of Regenerative Medicine and Biotherapies, Montpellier, 34295, France
[3]Clinical Research and Epidemiology Unit, University of Montpellier Hospital Centre, Montpellier, 34090, France
[*]therese.commes@inserm.fr

ABSTRACT

Acute Myeloid Leukemia (AML) is a heterogeneous disease that may occur at any age. Although it has been shown that the incidence of AML increases with age and that different genetic alterations are present in younger versus older patients, the current AML classifications do not include age as a factor in classifying patients. In this work, we analyzed 404 RNA-seq samples with AML initial diagnosis to highlight the differences between younger and older patients in favorable and adverse risk using a k-mer-based approach for transcriptomic machine learning models. We predicted the risk of patients with more than 90% accuracy. We also provided lists of genes of interest for AML that were able to distinguish favorable from adverse ELN risks. From these lists, we selected prognostic biomarkers that have an impact on survival. Furthermore, we analyzed the biological context involved in the transcriptome complexity of younger and older patients. We observed differences in tumor profiles as well as in the presence of immune and stromal cell populations with specific features in older patients.

Introduction

Acute Myeloid Leukemia (AML) is a heterogeneous and complex disease with differences in morphology, immunophenotype, genetic, epigenetic signatures, leading to different responses to treatment [1]. Current classifications such as the European LeukemiaNet (ELN) base the AML stratification on cytogenetics and molecular biomarkers at the genomic DNA level, in which patients are classified by risk into favorable, intermediate, and adverse categories. A recent study [2] presents the revised 2022 ELN genetic-risk classification as suitable for prognostic stratification of patients with AML, but it encourages the separation of younger from older patients because the genetic alterations and outcomes differ between these age groups [3]. Although AML can occur at any age, studies have described age as an important factor in the prognosis of AML [4], and its management presents more challenges as age increases [5,6].

RNA-sequencing (RNA-seq) is an accurate method to analyze transcriptome profiles, allowing the identification of key molecular mechanisms and signatures that can drive disease pathogenesis and progression, and which makes it a method of choice to conduct a deep investigation of patient subgroups according to prognosis. As proposed by Docking [4], RNA-seq can provide a standalone assay for diagnosis and prognosis in AML.

However, classical RNA-seq analysis based on genome mapping has major limitations, it consumes an enormous amount of computer power that limits large-scale analysis and is guided by reference annotation that restrains the set of information taken into account. Machine Learning (ML) methods offer new opportunities in this field, the integration of ML and RNA-seq has recently shown its effectiveness for prognosis in cancer. But here again, they are mainly based on gene expression features counted from annotation [7,8].

To address these limitations, we demonstrated that RNA-seq data can be analyzed using k-mer based approaches, which, unlike traditional methods guided by reference annotations, are reference-free methods without pre-mapping or assembly [9,10]. All the reads can be divided into successive substrings of length k, the k-mers, that can be counted and indexed to provide a compressed representation of the data, useful for exact and approximate sequence search [11,12] in a reduced time [13]. The integration of ML algorithms and k-mers can be a good strategy to capture the complex biological content of transcriptome data and to manipulate data on a large scale.

Considering the need to improve the classification of AML patients, we explore in this work the difference between the

age groups in ELN classification by analyzing the corresponding RNA-seq data at the k-mer level. We predict favorable and adverse risks in AML using ML models trained with k-mers count matrices for younger and older patients with initial diagnosis. Then, we analyze the difference between favorable and adverse risk in the two age groups and we provide a list of genes of interest for AML and the impact on survival. We also observe biological features that differentiate the older group in link with the risk prediction.

Materials and methods

Transcriptome data and clinical information

We analyzed 4 transcriptome cohorts with AML patients. To investigate ELN risk stratification, we analyzed 404 samples with favorable or adverse risk: 212 samples from the Beat-AML [14]; 112 samples from the Beat-AML2.0 [15]; and 80 samples from Leucegene [16]. We also performed survival analysis using 242 samples from the Beat-AML and 37 from the GSE62852 cohort. Table 1 presents an overview of risk stratification and age groups from the AML cohorts.

ELN Analysis						
	Beat-AML (training)		Beat-AML2.0 (test)		Leucegene (test)	
	Favorable	Adverse	Favorable	Adverse	Favorable	Adverse
Younger	65	33	29	25	36	21
Older	41	73	19	39	12	11
Survival Analysis						
	Beat-AML		GSE62852			
Younger	119		12			
Older	123		25			

Table 1. Number of transcriptome samples for ELN and survival analysis. ELN analysis with 404 samples and survival analysis with 279 samples.

The cohorts for ELN analysis were used in the ML process. We used Beat-AML to train the models and Beat-AML2.0 to test them. To avoid data leakage and to confirm the effectiveness of the evaluation, we used the two datasets as different groups. Thus, samples of the Beat-AML2.0 cohort belonging to Beat-AML patients were removed from the analysis. Beat-AML and Beat-AML2.0 samples were obtained from the dbGAP database, accessions ID phs001657.v1.p1 and phs001657.v2.p1, respectively. Furthermore, we also used the Leucegene cohort (accessions ID GSE49642, GSE52656, and GSE62190) to test the models.

The cohorts were authorized for use and underwent a quality control process. We checked the quality of the raw data using fastQC version 0.11.9 [17] and MultiQC version 1.9 [18]. As a complementary quality control, we verified the sequencing protocol information and contamination with KmerExplor [19]. More information about quality control can be found in the supplementary material.

Generating training k-mer count matrices

A k-mer is a substring extracted from a biological sequence (read) of fastq raw data. To extract and count k-mers from the fastq files, we used Kmtricks [13], a tool to count k-mers in large datasets and produce a k-mer count matrix across multiple samples. To generate the k-mer count matrices, we used only samples classified as favorable or adverse risk with initial diagnosis. Samples from intermediate patients were not used due to the difficulty to define this group in real-life [20]: some patients fall between classes and are classified as favorable/intermediate or intermediate/adverse due to the difficulty to define risk stratification.

To avoid noisy data, we used filters from Kmtricks, we counted k-mers with a minimum abundance of 4 (which means, a k-mer has to be found at least 4 times in a sample) and present in at least 5% of all samples of the cohort analyzed. We generated a matrix from younger patients and another from older patients using the Beat-AML cohort.

Selecting k-mers

Due to the high dimensionality of the data, we applied a feature selection step in the k-mer count matrices from the training cohort. As shown in Fig. 1.A, from the k-mers count generated by Kmtricks, we applied three filters to each k-mer individually. (1) Expressed k-mer: we selected a k-mer when the count of at least 70% of samples in favorable or adverse risk was different from zero. (2) Highly expressed outliers removal: we removed k-mer values considered as outliers. We considered as outliers a

k-mer count higher than the third quartile for all samples in the k-mer. (3) Differential expression: we applied a coefficient of variation in the k-mer count between the favorable and adverse samples.

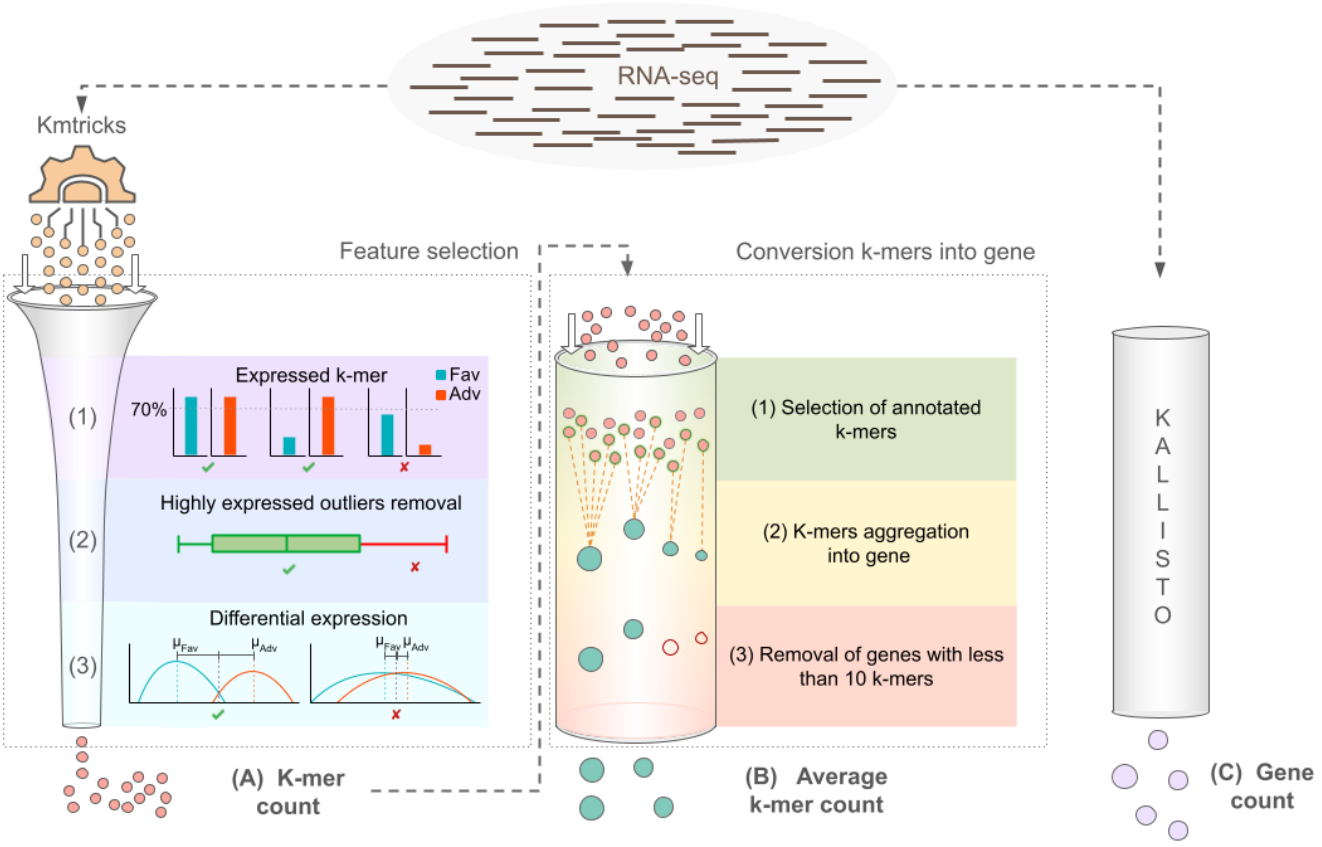

Figure 1. Overview to generate different counting methods. (A) K-mer count generated in the feature selection step. (B) Average k-mer count generated by the conversion from k-mers into gene. (C) Gene count generated by Kallisto.

The coefficient of variation is defined by Equation 1.

$$\theta = \frac{\sigma}{\mu}$$

$$\sigma = \sqrt{\frac{(\mu_{Fav} - \mu)^2 + (\mu_{Adv} - \mu)^2}{2}} \tag{1}$$

$$\mu = \frac{\mu_{Fav} + \mu_{Adv}}{2}$$

Where σ is the standard deviation of the distances between the average k-mer count of favorable and adverse, and the average k-mer count for all samples. μ is the sum of the average k-mer count from favorable samples and adverse samples by the number of prognostics. A k-mer is selected if the coefficient of variation value is higher than or equal to 1. This process produced new k-mer count matrices, one from younger patients and one from older patients, used to train the ML models.

Generating test k-mer count matrices

To evaluate the ML models we needed to test k-mer count matrices with the same k-mers used in the training. To generate these k-mer count matrices, we applied "Back_to_sequences" [21]. This tool indexes a set of k-mers of interest and computes the number of occurrences of k-mers in the sequences (k-mer count). We provide to "Back_to_sequences" the k-mer list selected in the feature selection step, and the fastq files from Beat-AML2.0 and Leucegene. "Back_to_sequences" generated the k-mer

count for each sample, allowing us to build test k-mer count matrices for younger and older patients. Then, the values less than 4 in the k-mer count matrices were replaced by zero to avoid noisy data.

Machine Learning methods

Using the k-mer count matrices with selected k-mers, we built machine learning models to predict whether the patient has favorable or adverse risk in younger and older patients. We selected six Machine Learning algorithms used in other works to predict cancer [22,23], including AML [24,25]. We used three complex models: Neural Network (NN), Random Forest (RF), and eXtreme Gradient Boosting (XGB); and three less complex models: Decision Tree (DT), K-nearest neighbors (KNN), and Logistic Regression (LR).

We implemented the algorithms using the Scikit-Learn version 1.2.2 [26] and XGBoost version 1.7.4 [27] packages in Python, applying for each model a grid search with different parameters and a stratified cross-validation with 10-folds. The models were used to predict the favorable or adverse risk in the test k-mer count matrices.

The models were evaluated by accuracy, sensitivity, specificity, and Matthew's correlation coefficient (MCC), metrics that express the relations between True/False Positives and True/False Negatives. True Positives (TP) are favorable samples that were correctly predicted as favorable; True Negatives (TN) are adverse samples that were predicted as adverse; False Positives (FN) are adverse samples that were wrongly predicted as favorable; and False Negatives (FN) are favorable samples that were predicted as adverse. The metrics can be defined by Equations 2 to 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{5}$$

Additionally, we used AUC (Area Under the ROC Curve), which summarizes the True Positive Rate vs False Positive Rate into a single value. All scripts, models, Jupyter notebooks, and data used for machine learning methods can be found in the supplementary material.

Mapping and annotation

We identified and associated the k-mers from the feature selection step with the genes they belong to. To this end, we applied STAR 2.7.8a [28] to map the k-mers to a reference human genome, the GRCh38 assembly, and we used SAMtools 1.11 [29] to generate flexible alignment formats, SAM and BAM files, with mapping positions. Then, we implemented a script in R using the Ensembl REST API [30] to request the gene annotation for each k-mer using the SAM and BAM files. This process gave us the corresponding gene for each k-mer when the k-mer was mapped only in one position in the genome and with 100% alignment.

Expression counting methods

We used different methods to quantify the information from reads, as presented in Fig. 1: (A) k-mer count; (B) average k-mer count; and (C) gene count. Method A uses directly the k-mer count given by Kmtricks from the k-mers retained after the high dimensionality reduction step (process described in Selecting k-mers section). The method B is a conversion from k-mers into genes that includes three steps: (1) a selection of annotated k-mers exclusive for each age group; (2) the aggregation from k-mers into the genes that they belong to by average count; (3) a selection of genes with more than 10 k-mers, which was indispensable considering that genes with few k-mers can be poorly representative. Method C gives classic "gene count". We used a widely used gene method computed by Kallisto [31] to compare with the A and B methods.

Survival statistical analysis

Identification of genes impacting patients' survival, defined as the time from diagnosis to death, was performed based on survival analysis. From the genes identified in the counting method B, we applied a correlation test from Caret package, where if two genes have more than 95% correlation, we removed the gene with the largest mean absolute correlation. Then , gene expression was defined as "high" if gene expression value was greater or equal to the mean, or "low" if it was lower. Age was included in the analysis as it is known to impact survival. Given the large number of genes, a two-step approach was performed. First, we performed dimension reduction by selecting genes based on Cox LASSO method[32] using Glmnet package. The penalization parameter was determined using cross-validation and was chosen such that the deviance of the model was minimal. Then, genes with a non-zero coefficient were included in a multivariate Cox model[33] using Survival package. Proportional hazards assumption was assessed for genes and age individually using statistical tests[34] based on Schoenfeld residuals. The effect size for each gene and age was estimated using the hazard ratio together with its 95% confidence interval. The analysis was performed in R and can be found in supplementary material.

Biological context

We investigated the biological context of the samples by looking for the percentage of blast cells, mutation profile, fusion gene presence, and ratio of immune and stromal cells. We analyzed bone marrow (BM) and peripheral blood (PB) samples from the Beat-AML cohort. The blast percentage was provided by the Vizome website. The mutation and fusion gene information was obtained from metadata in cbioportal. For the mutation profile, we considered the mutations in DNMT3, TET2, IDH1, IDH2, and ASLX1 genes frequently associated with clonal hematopoiesis as described by[6]. For the fusion genes, CBFB-MYH11, DEK-NUP214, GATA2-MECOM, MLLT3-KMT2A, PML-RARA, and RUNX1-RUNX1T1 fusions were taken into account. In order to count the immune and stromal cells, we used a previously reported method based also on k-mer count. We designed specific k-mers from the gene list of MCP-counter [35] using Kmerator [19]. Kmerator is a tool able to generate k-mers that are present only in the gene requested (specific k-mers). We regrouped the genes (average of k-mers) by cell type, including B lineage, CD8 T cells, T cells, cytotoxic lymphocytes, natural killer (NK) cells, dendritic, monocytic, neutrophils, fibroblasts, and endothelial cells. Additionally, we used g:Profiler [36] to provide the biological interaction from the identified genes. g:Profile finds statistically significant Gene Ontology (GO) terms, pathways, and other gene function-related terms.

Results

Predicting favorable and adverse outcome patients

We generated using Kmtricks a matrix from younger patients with 364,846,009 k-mers in 98 samples and a matrix from older patients with 399,034,012 k-mers in 114 samples. After the feature selection, new k-mer count matrices were generated, one from younger patients and one from older patients, with 35,098 and 63,929 selected k-mers, respectively. These new matrices using the count method A were used to train 6 ML models to predict favorable or adverse risk in younger and older patients. Then, the models were used to predict the prognosis in the test cohorts and they were evaluated by metrics. Figure 2.A1 shows Logistic Regression (LR) had the best performance in predicting in the younger group, with an accuracy of 92%. In the older group, we achieved an accuracy of 92% with eXtreme Gradient Boosting (XGB). To see if the k-mer selection was consistent with the results, we computed UMAP projections (Fig. 2.A2), where we confirmed that the prognosis favorable and adverse was effectively separated by k-mers.

Since the predictions with the count method A were well-performed, we trained 6 models with method B using the average k-mer count. The process of conversion from k-mer count into average k-mer count generated 99 genes for younger and 250 genes for older groups (see Supplementary Table S5 and S6) . Using count method B (Fig. 2.B1), the LR and RandomForest (RF) models had the best performances in younger patients, with 87% accuracy. In older patients, an LR model remained with the best performance, predicting with 87% accuracy. We also confirmed that the favorable and adverse patients continued to be well separated in the average k-mer count by the UMAP projection (Fig. 2.B2).

When we used Kallisto to generate gene count method C, we produced the count for the same genes identified in the average k-mer count, 99 for younger and 250 for older. Thus, we trained the ML models with the method C (Fig. 2.C1). The best model in the younger group was RF, losing 6% of accuracy when compared with method B. The models for older patients had more loss of performance, LR was the best model with only 67% accuracy. Following the negative impact on prediction, the UMAP projection (Fig. 2.C2) showed that the information at the gene count was not enough to separate the favorable from adverse patients.

All the prediction performances, including Beat-AML2.0 and Leucegene cohorts separately, can be seen in the Supplementary Tables S1 to S4.

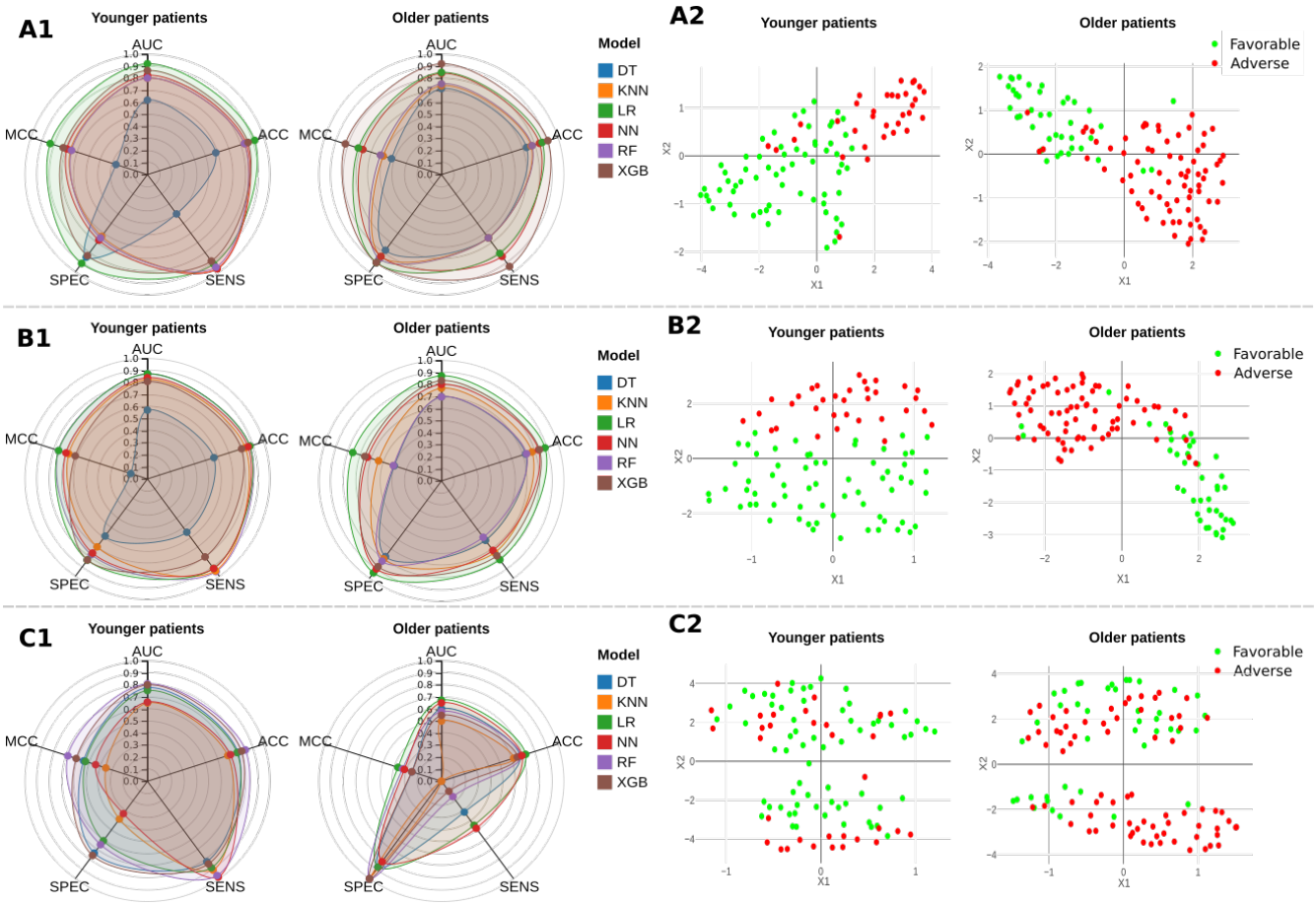

Figure 2. Predicting favorable and adverse risk with k-mers (A1), average k-mer (B1), and genes (C1) counts using Decision Tree (DT), K-nearest neighbors (KNN), and Logistic Regression (LR), Neural Network (NN), Random Forest (RF), and eXtreme Gradient Boosting (XGB) models. Metrics Area under the curve (AUC), accuracy (ACC), sensitivity (SENS), specificity (SPEC), and Matthew's correlation coefficient (MCC) for evaluating models in younger and older patients. UMAP projection with k-mer (A2), average k-mer (B2), and gene (C2) count.

Survival in younger and older patients

Identification of genes impacting patients' survival was performed for 99 genes for younger and 250 genes for older groups from Beat-AML and the GSE62852 cohorts. We first analyzed the correlation between genes, where we removed 11 genes highly correlated in the younger group. In the older group, we do not find highly correlated genes. After we applied a dimension reduction step, 13 variables were retained by the Cox LASSO for the younger group, including age. For the older group, age was the only variable selected confirming the difference observed in the transcriptome in this group compared to younger patients. The proportional hazards assumption was verified for the final Cox models including the selected variables (results in supplementary material). The results of the models are shown in Table 2. Age had a significant effect on survival (p-value = 0.03). It appeared that survival was negatively impacted for every additional year of age (HR (95% CI) = 1.03 (1.003 – 1.058)). Four genes had a significant impact on survival in the younger group: GLCCI1 (p-value < 0.01), SLC29A2 (p-value < 0.01), RACK1 (p-value = 0.04), and LINGO3 (p-value = 0.04). For GLCCI1, SLC29A2, and LINGO3, survival was lower for high expression compared to low expression (HR (95%CI): 3.32 (1.44 – 7.67), 4.46 (2.04 – 9.75), 2.15 (1.03 – 4.46), respectively). On the contrary, survival was greater for patients with a high RACK1 gene expression compared to low expression (HR (95%CI): 0.40 (0.17 – 0.96). For the older group, age had a significant impact on survival (p-value < 0.01). As for young patients, as age increases, the survival decreases (HR (95% CI): 1.07 (1.04 – 1.10). More detail can be found in Supplementary Table S5 to S10.

Younger group		
	Hazard ratio (95% CI)	**P-value**
age	1.03 (1.00 - 1.05)	0.03
MPO	0.38 (0.10 - 1.41)	0.15
CLCN5	1.41 (0.67 - 2.98)	0.36
TPM1	1.56 (0.76 - 3.20)	0.22
CLEC4OP	0.48 (0.05 - 4.59)	0.52
TIAM1	0.97 (0.43 - 2.18)	0.94
GLCCI1	3.32 (1.43 - 7.67)	< 0.01
SLC29A2	4.45 (2.03 - 9.75)	< 0.01
RACK1	0.39 (0.16 - 0.95)	0.03
NEIL1	0.87 (0.38 - 1.98)	0.75
LINGO3	2.14 (1.03 - 4.45)	0.04
IGKV2-24	0.63 (0.22 - 1.80)	0.38
BMI1	1.47 (0.74 - 2.91	0.26
Older group		
	Hazard ratio (95% CI)	**P-value**
age	1.06 (1.03 - 1.09)	< 0.01

Table 2. Cox regression survival analysis in younger and older groups. Proportional hazard ratio with 95% confidence interval (CI).

The complexity in older patients

Using the count method B, we set an additional test (Fig. 3.1), we inverted the trained models to predict: the 6 ML models from younger patients were used to predict risk in older patients, and the 6 ML models from older patients to predict risk in younger patients.

The models from younger patients had a good performance in older patients, achieving 81% accuracy with the RF model to predict favorable or adverse risk. The models trained in older patients had more difficulty predicting risk in younger patients, the best performance was the LR model with 70% accuracy. These results gave us evidence that the model trained in older patients learned different information from the transcriptome of younger patients.

As the transcriptome complexity in older patients is different than in younger patients, we made some hypotheses about the biological and cellular content to explain it. The differences could originate from inherent tumor cell profiles, external factors like tumor microenvironment cell types, or physiological parameters linked to aging. Thus, in order to evaluate tumor differences, we first checked the percentage of blast cells in the BM and PB samples (Fig. 3.2). We observed a significant difference in blast percentage in the BM of older patients from favorable versus adverse risk. This difference was not significant in the younger group and, as expected, in peripheral blood. We then characterized the molecular profile of tumor cells in the different groups comparing the presence of mutations in genes frequently associated with clonal hematopoiesis in older patients. We also characterized the presence of fusion genes observed usually with a higher rate in young patients. As expected, significant differences were observed in older and younger groups (Fig. 3.3). The presence of mutations or fusions depends on the risk with the highest rate of mutations found in the older patients group with an adverse risk. On the contrary, the highest level of fusion genes is observed in the younger patients group with a favorable risk. These results confirm a different tumor behavior with aging.

Then, we searched for the presence of immune and stromal cell markers, in the cell population of BM samples, to see if they can account for differences according to ELN risk. Figure 4 confirmed that only dendritic cells are significantly different between favorable and adverse risks in younger patients. In contrast, a more pronounced presence of B and T cell lineages as well as endothelial and fibroblast cell subtypes were significantly measured in adverse versus favorable risk in older patients.

Figure 5 shows the analysis of the enrichment in gene ontology (GO) molecular function (MF), biological process (BP), or cellular component (CC) in the set of genes from younger and older patients. Interestingly, specific functions were only observed in the transcriptome data of older patients in link with the immune response (peptide antigen binding, antigen processing, and MHC protein complex) as well as stromal cell features (cell adhesion, cardiac fibroblast cell development) and correlated with observed cell populations. The genes associated with each GO can be seen in Supplementary Table S11 and S12.

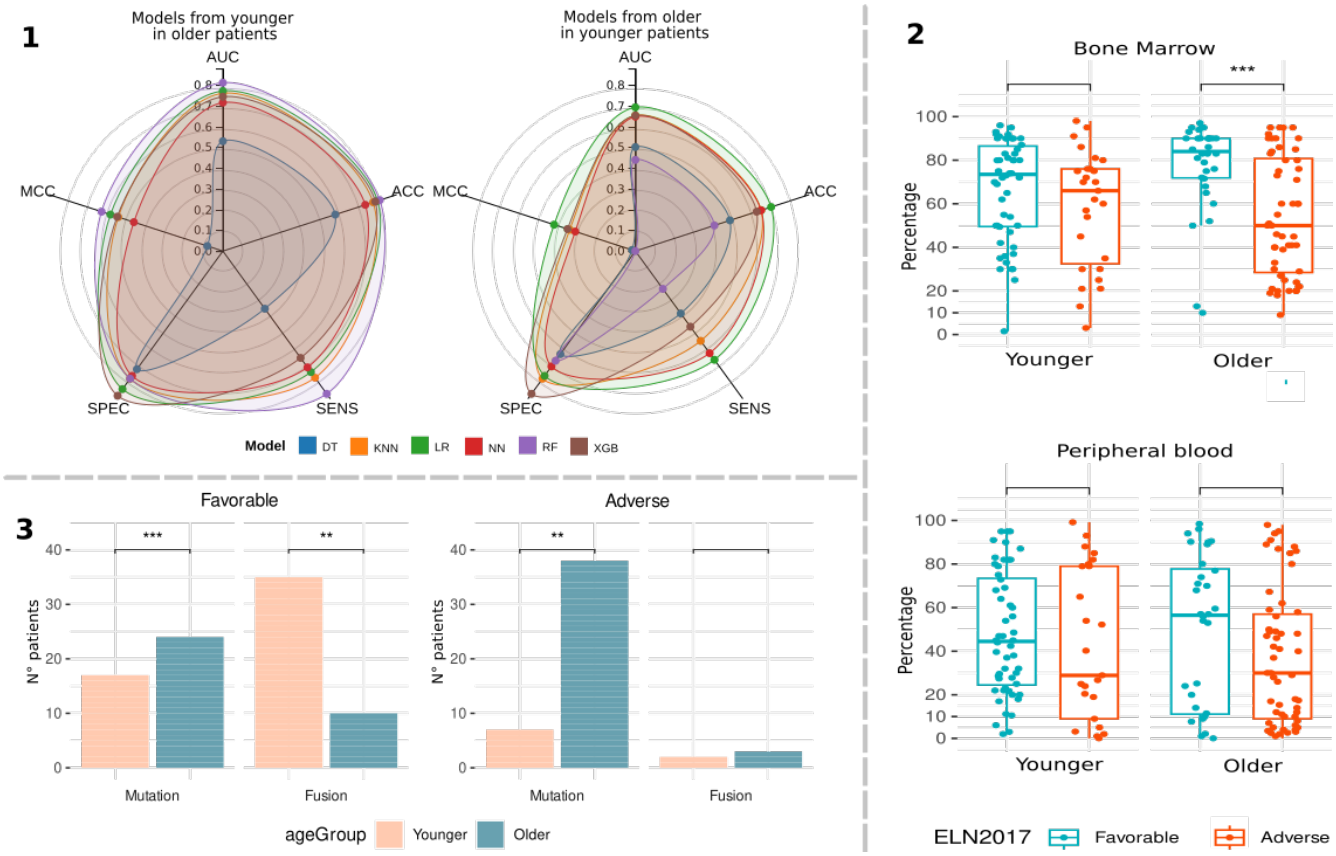

Figure 3. (1) Performance for models from younger patients to predict in older patients and models from older patients to predict in younger patients. (2) Percentage of blast cells in bone marrow and peripheral blood samples. (3) Number of younger and older patients with mutations and fusion genes in favorable and adverse risk. '****': $p \leq 0.0001$; '***': $p \leq 0.001$; '**': $p \leq 0.01$; '*': $p \leq 0.05$; ' ': $p > 0.05$.

Discussion

In this study, we applied machine learning models to transcriptome data using a k-mer based approach to investigate the difference in the risk stratification of younger and older AML patients. Through the RNA-seq data, we analyzed the expression level of transcriptomes by different counting methods to detect qualitative and quantitative changes in specific conditions. We observed that k-mer proved to be an valuable approach to investigate RNA-seq data due to two main reasons. First, it allowed us to analyze the data on a large scale without pre-mapping or assembly. Second, it captures the biological information more precisely, even a single mutation[12].

Our study results showed genes (counting method B) able to distinguish favorable from adverse risk. We found 99 genes for younger patients and 250 genes for older patients. The genes were identified based on k-mers selected with the ELN risk information, without any other previous knowledge. The good performance of the ML prediction confirmed the efficiency of both, the feature selection step (counting method A) and the gene to k-mer conversion step (counting method B) when predicting favorable and adverse risks. When comparing our prediction of ELN risk based on k-mers and the prediction performed with gene quantification taking into account the whole gene reads (counting method C), we found the best results with our method showing that a k-mer contains enough information and that full transcript-scale information is not required.

When performing survival analysis to analyze the time to death, 5 variables (including age) had a significant impact on it. For GLCCI1, SLC29A2, and LINGO3 genes, high expression was associated with a bad prognosis compared to low expression, while for RACK1 high expression was associated with a good prognosis. SCL29A2 gene was recently described in a prognostic risk-scoring model in AML[37], and despite the MPO gene not being shown to be significant in our model, this gene is already known as highly expressed in favorable risk for pediatric AML patients[38]. Furthermore, as expected, in both age groups the survival rate decreases when the age increases, indicating a crucial role of age in AML, mainly in older patients. We also investigated the biological context of the patients. Through the percentage of blast cells, we noticed that other cells were filling up the BM of older adverse patients which led us to investigate the immune and stromal cells content. Stromal cells, as

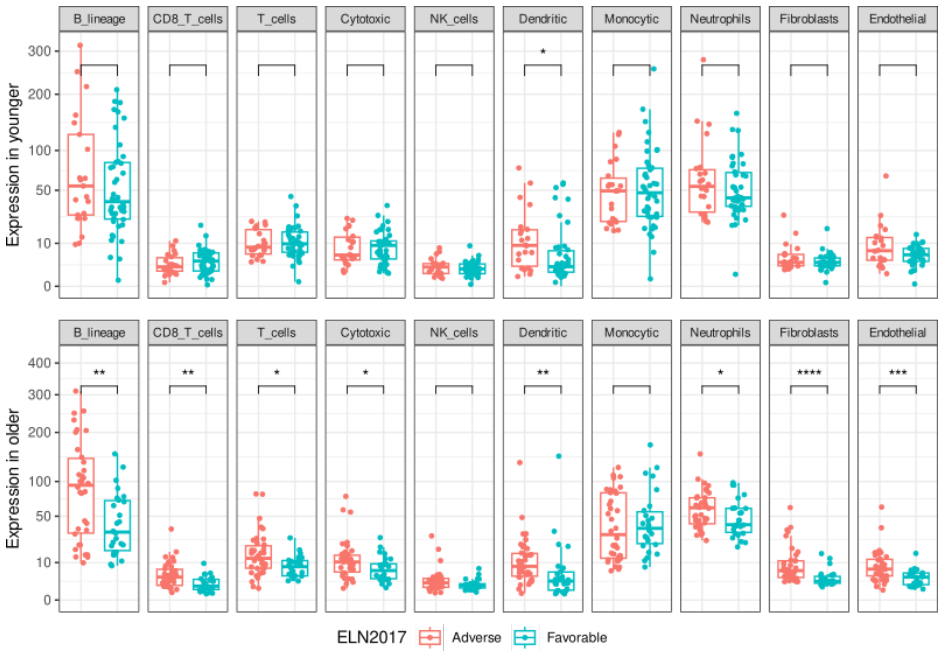

Figure 4. Expression of immune and stromal cells in bone marrow samples and comparison of favorable and adverse risk. '****': $p \leq 0.0001$; '***': $p \leq 0.001$; '**': $p \leq 0.01$; '*': $p \leq 0.05$; ' ': $p > 0.05$.

endothelial and fibroblasts, are components of the BM microenvironment and are fundamental in hematopoiesis regulation. In a tumor microenvironment, endothelial cells acquire specific functions to support the survival of cancer cells and act as interactions mediatorsbetween tumor and immune cells[39]. Similarly, fibroblast cells have also been reported to contribute to leukemia cell survival[40]. In line with this, the age-related AML progression and the dependence on the BM microenvironment for progression to cancer[41] may explain the high prominence of endothelial and fibroblasts cells in BM samples of adverse older patients.

Additionally, our results showed that immune cells are also highly expressed in adverse older patients. The impact on these patients can be explained by the fact that the immune system undergoes profound changes with aging and immune cells are shown to support leukemogenesis and resistance to therapy[3]. A high proportion of regulatory T cells was already reported as of poor prognosis by interfering with immunologic synapse formation[42]. Also, regulatory B cells were reported with high expression in patients with poor prognosis[43]. In summary, our results point in the same direction as the recent literature showing that leukemic cells influence the BM microenvironment to support their survival[4445]. Moreover, our results highlight the influence of AML cells in the BM microenvironment mainly in older patients when the risk increases, which can implicate leukemia cell survival, and also resistance to therapy in this age group.

Data availability

The data analyzed in this study are publicly available in the dbGAP database (https://www.ncbi.nlm.nih.gov/gap/) with the accessions ID phs001657.v1.p1 and phs001657.v2.p1, and the GEO database (https://www.ncbi.nlm.nih.gov/geo/) accessions ID GSE49642, GSE52656, GSE62190, and GSE62852. The data generated during the study are available from the corresponding author upon reasonable request.

References

1. Döhner, H., Wei, A. H. & Löwenberg, B. Towards precision medicine for aml. *Nat. reviews Clin. oncology* **18**, 577–590 (2021).

2. Mrózek, K. *et al.* Outcome prediction by the 2022 european leukemianet genetic-risk classification for adults with acute myeloid leukemia: an alliance study. *Leukemia* **37**, 788–798 (2023).

3. Perzolli, A., Koedijk, J. B., Zwaan, C. M. & Heidenreich, O. Targeting the innate immune system in pediatric and adult aml. *Leukemia* 1–11 (2024).

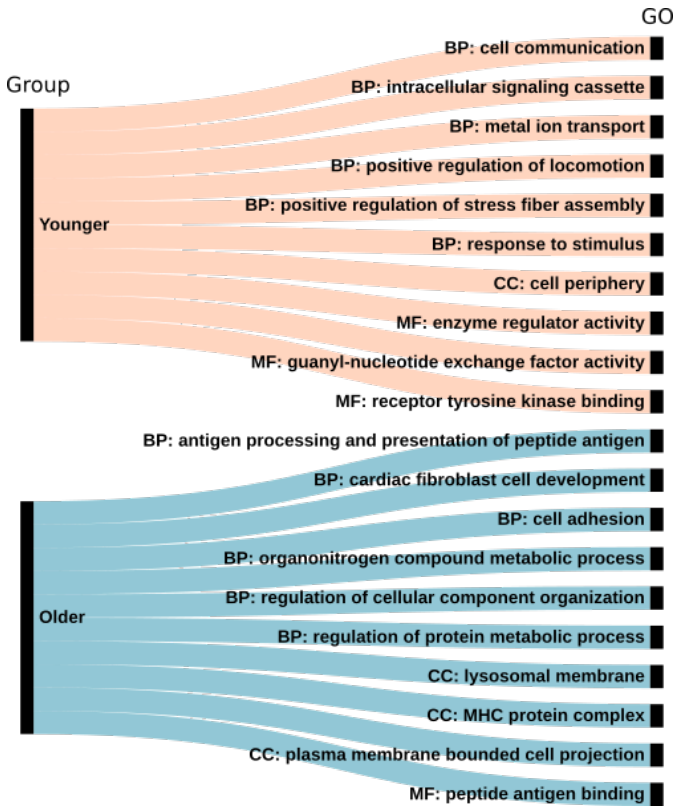

Figure 5. Gene Ontology (GO) for younger and older groups.

4. Docking, T. R. *et al.* A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.* **12**, 2474 (2021).

5. LeBlanc, T. W. & Erba, H. P. Shifting paradigms in the treatment of older adults with aml. In *Seminars in Hematology*, vol. 56, 110–117 (Elsevier, 2019).

6. Li, J.-F. *et al.* Aging and comprehensive molecular profiling in acute myeloid leukemia. *Proc. Natl. Acad. Sci.* **121**, e2319366121 (2024).

7. Tran, T. H. *et al.* Whole-transcriptome analysis in acute lymphoblastic leukemia: a report from the dfci all consortium protocol 16-001. *Blood advances* **6**, 1329–1341 (2022).

8. Gélard, M., Richard, G., Pierrot, T. & Cournède, P.-H. Bulkrnabert: Cancer prognosis from bulk rna-seq based language models. *bioRxiv* 2024–06 (2024).

9. Morillon, A. & Gautheret, D. Bridging the gap between reference and real transcriptomes. *Genome biology* **20**, 112 (2019).

10. Audoux, J. *et al.* De-kupl: exhaustive capture of biological variation in rna-seq data through k-mer decomposition. *Genome biology* **18**, 1–15 (2017).

11. Marchet, C., Iqbal, Z., Gautheret, D., Salson, M. & Chikhi, R. Reindeer: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics* **36**, i177–i185 (2020).

12. Bessière, C. *et al.* Exploring a large cancer cell line rna-sequencing dataset with k-mers. *bioRxiv* 2024–02 (2024).

13. Lemane, T., Medvedev, P., Chikhi, R. & Peterlongo, P. kmtricks: Efficient and flexible construction of bloom filters for large sequencing data collections. *Bioinforma. Adv.* (2022).

14. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

15. Bottomly, D. *et al.* Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer cell* **40**, 850–864 (2022).

16. BCLQ, C., Montreal. Leucegene project (2019).

17. Andrews, S. *et al.* Fastqc: a quality control tool for high throughput sequence data (2010).

18. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

19. Riquier, S. *et al.* Kmerator suite: design of specific k-mer signatures and automatic metadata discovery in large rna-seq datasets. *NAR genomics bioinformatics* **3**, lqab058 (2021).

20. Sargas, C. *et al.* Comparison of the 2022 and 2017 european leukemianet risk classifications in a real-life cohort of the pethema group. *Blood Cancer J.* **13**, 77 (2023).

21. Baire, A. & Peterlongo, P. Back to sequences: find the origin of kmers. *bioRxiv* DOI: 10.1101/2023.10.26.564040 (2023).

22. Naji, M. A. *et al.* Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Comput. Sci.* **191**, 487–492 (2021).

23. Erdem, E. & Bozkurt, F. A comparison of various supervised machine learning techniques for prostate cancer prediction. *Avrupa Bilim ve Teknoloji Dergisi* **21**, 610–620 (2021).

24. Karami, K., Akbari, M., Moradi, M.-T., Soleymani, B. & Fallahi, H. Survival prognostic factors in patients with acute myeloid leukemia using machine learning techniques. *PloS one* **16**, e0254976 (2021).

25. Shanbehzadeh, M., Afrash, M. R., Mirani, N. & Kazemi-Arpanahi, H. Comparing machine learning algorithms to predict 5-year survival in patients with chronic myeloid leukemia. *BMC Med. Informatics Decis. Mak.* **22**, 236 (2022).

26. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

27. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

28. Dobin, A. & Gingeras, T. R. Mapping rna-seq reads with star. *Curr. protocols bioinformatics* **51**, 11–14 (2015).

29. Homer, N., Marth, G., Abecasis, G. & Durbin, R. The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

30. Yates, A. *et al.* The ensembl rest api: Ensembl data for any language. *Bioinformatics* **31**, 143–145 (2015).

31. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic rna-seq quantification. *Nat. biotechnology* **34**, 525–527 (2016).

32. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat. medicine* **16**, 385–395 (1997).

33. David, C. R. *et al.* Regression models and life tables (with discussion). *J. Royal Stat. Soc.* **34**, 187–220 (1972).

34. Grambsch, P. M. & Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994).

35. Meylan, M. *et al.* webmcp-counter: a web interface for transcriptomics-based quantification of immune and stromal cells in heterogeneous human or murine samples. *BioRxiv* 2020–12 (2020).

36. Reimand, J. *et al.* g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research* **44**, W83–W89 (2016).

37. Zhang, C. *et al.* Identification and validation of a prognostic risk-scoring model for aml based on m7g-associated gene clustering. *Front. Oncol.* **13**, 1301236 (2024).

38. Roberson, J. R. *et al.* Prognostic significance of myeloperoxidase expression in childhood acute myeloid leukemia. *Pediatr. blood & cancer* **50**, 542–548 (2008).

39. Leone, P. *et al.* Endothelial cells in tumor microenvironment: insights and perspectives. *Front. Immunol.* **15**, 1367875 (2024).

40. Ding, Z. *et al.* Cancer-associated fibroblasts in hematologic malignancies: elucidating roles and spotlighting therapeutic targets. *Front. Oncol.* **13** (2023).

41. Plakhova, N., Panagopoulos, V., Vandyke, K., Zannettino, A. C. & Mrozik, K. M. Mesenchymal stromal cell senescence in haematological malignancies. *Cancer Metastasis Rev.* **42**, 277–296 (2023).

42. Brück, O. *et al.* Immune profiles in acute myeloid leukemia bone marrow associate with patient age, t-cell receptor clonality, and survival. *Blood advances* **4**, 274–286 (2020).

43. Shi, Y., Liu, Z. & Wang, H. Expression of pd-l1 on regulatory b cells in patients with acute myeloid leukaemia and its effect on prognosis. *J. Cell. Mol. Medicine* **26**, 3506–3512 (2022).

44. Bassani, B. *et al.* Zeb1 shapes aml immunological niches, suppressing cd8 t cell activity while fostering th17 cell expansion. *Cell Reports* **43** (2024).

45. Bakhtiyari, M. *et al.* The role of bone marrow microenvironment (bmm) cells in acute myeloid leukemia (aml) progression: immune checkpoints, metabolic checkpoints, and signaling pathways. *Cell Commun. Signal.* **21**, 252 (2023).

Acknowledgements

This work has been supported by La Ligue Contre le Cancer and the Agence Nationale de la Recherche (TranSipedia and FullRNA projects).

Author contributions statement

RS wrote the manuscript, analyzed the data, and developed the methodology. CR interpreted the data and prepared the Figure 1. MA performed the survival analysis. JR revised the manuscript. CS are generated the data for immune and stromal cells analysis. BG and AB worked to manage the AML cohorts. FR and NG interpreted the annotation. TC designed the study.

Additional information

Supplementary data can be found at https://osf.io/kthvb/.

Competing interests

The authors declare no competing interests.