AML: Supervised VS Unsupervised Paradigms

Noemi Bongiorni, Elisa Nordera, Sara Redaelli, Irene Rusconi, Rachele Zanin
Dr. Soumick Chatterjee, Dr. Sina Kanannejad

Introduction

Acute Myeloid Leukemia (**AML**) is a severe blood cancer that begins in the bone marrow and is characterized by the uncontrolled accumulation of immature myeloid cells, or blasts, which do not mature into functional blood cells.

Ongoing research aims to uncover the molecular mechanisms behind AML and to develop more effective, personalized treatments. A key focus is the **analysis of gene and protein expression** in bone marrow cell populations, as specific molecular changes can indicate the presence and progression of the disease.

Aims

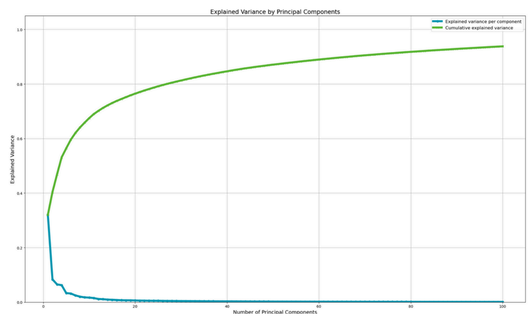
- Simplify complex molecular data while preserving key biological and clinical information.
- Identify a **subset of genes and proteins** that reliably indicate the presence of AML from a large molecular dataset.
- Detect **similarities between patients** based on gene expression profiles (unsupervised analysis) and build a **classifier** able to identify the presence of the illness (supervised analysis).
- Explore the adaptability of the methodology to other genetically influenced diseases.

Preprocessing

Two datasets collected from **23 individuals**: 7 AML adult patients, 10 AML pediatric patients and 6 healthy donors. The first dataset (sc-RNA seq data) contains the expression level of the **genes** in each cell (*211,969 x 36,601*). The second one (CITE-seq data) cointains the abundance of the **proteins** expressed in each cell (*211,969 x 61*).

Aggregating all rows belonging to the same patient and cell type by computing the **median expression value**, the dataset is reduced to 434 rows, which can reasonably considered as independent units.

Applying Principal Component Analysis (**PCA**) and retaining the first 30 components, which together account for approximately 80% of the total variance.

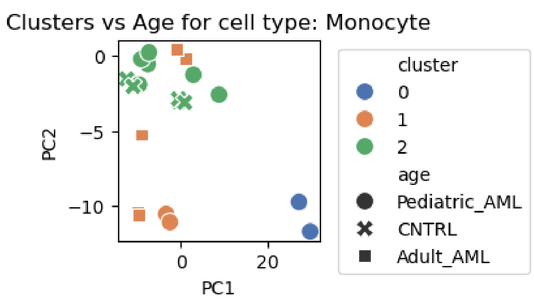


UMAPs of the cells of each patient, plotted in the space given by the principal components.

Unsupervised Analysis

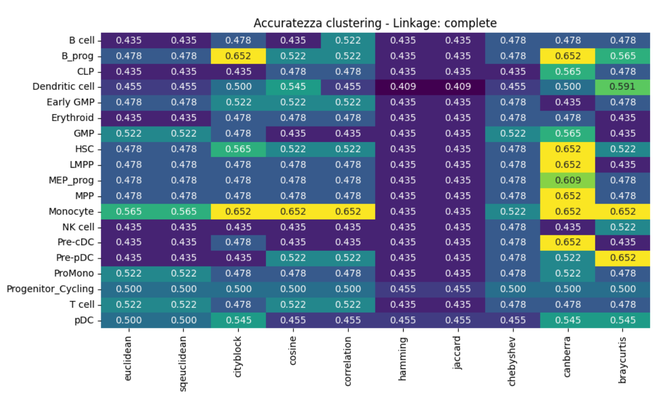
K-Means

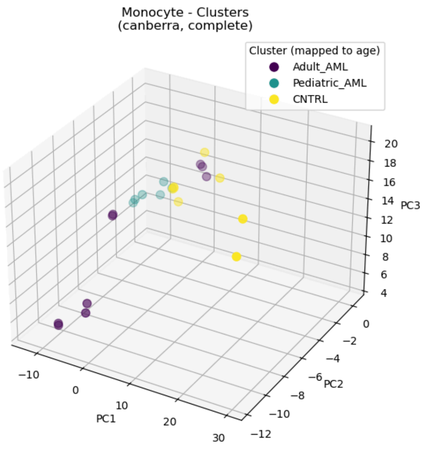
- Applied with **3 clusters** (reflecting known groups: adult AML, pediatric AML, and controls).
- Maximum accuracy of **0.652**. Clusters often mix individuals from different groups.

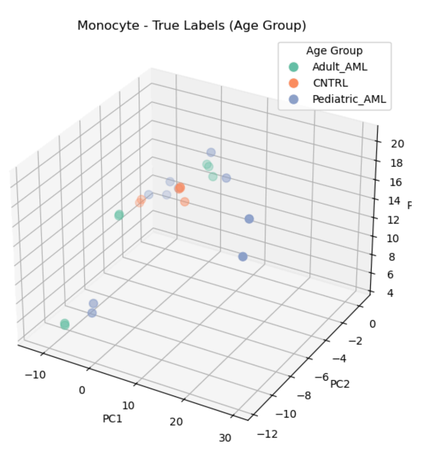


Hierarchical Clustering

- Tested 10 distance metrics with 3 linkage types.
- Best combination is **Canberra distance** + **complete linkage**. Accuracy above **0.545** for most cell types.
- Cophenetic Correlation Coefficients (CPCC) are high (up to **0.909**): the clustering preserves the original distance structure well.







DB Scan

- Implemented using **Canberra distance** with various parameter settings.
- Poor performance: best-case accuracy of only **0.545**.
- Adding MDS for dimensionality reduction does not significantly improve the outcome.

Aggregating Clustering Results

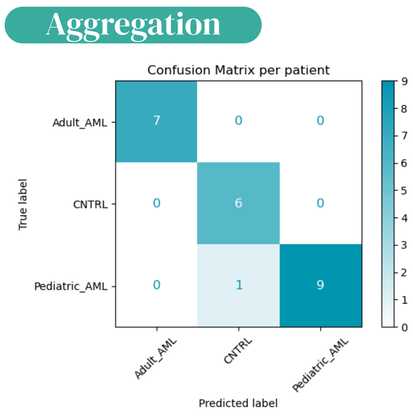
- 19 separate clustering results, one for each cell type.
- A **similarity matrix** between patients is built by comparing their clustering across all cell types.
- Cell types are **weighted** based on their importance in distinguishing AML presence.
- A final clustering is performed using two methods:
  a. Direct clustering from the similarity matrix (accuracy = **0.478**).
  b. **MDS + K-Means** (accuracy = **0.739**), correctly identifying all adult AML patients.

Supervised Analysis

Methods

- Linear Discriminant Analysis (**LDA**): not feasible due to unequal covariances among groups (Box'M test).
- Quadratic Discriminant Analysis (**QDA**): accuracy of **0.908**, chosen for the rest of the analysis.
- Support Vector Machines (**SVM**): accuracy of **0.863**.

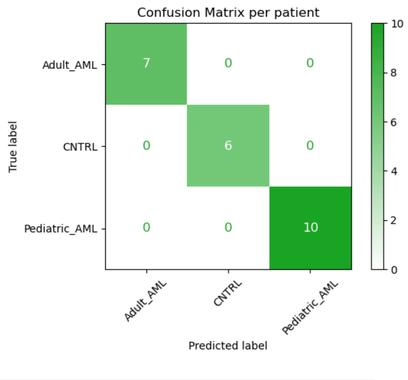
Aggregation

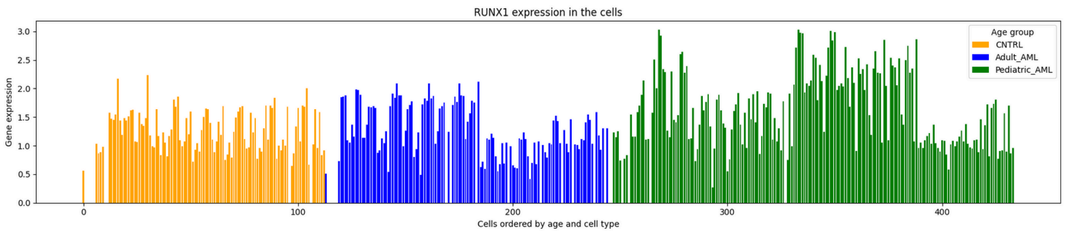


- Results obtained for every cell type using the first **30 PCs**. Aggregated by weighting the cell types based on their importance.
- Highly accurate classification. **Only one mistake**: patient 10 (pediatric AML group), is predicted as healthy (CNTRL). This is explained by UMAPs, as the cellular profile of patient 10 is nearly identical to those of healthy patients 22 and 23.

Gene Selection

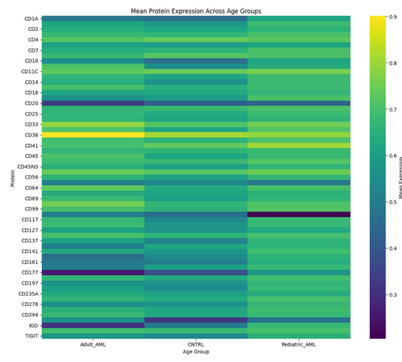
- Goal: identify a smaller set of highly informative genes. Selection of the 9 genes with the highest loadings per component → **179 most relevant genes**.
- QDA accuracy on the reduced dataset: **0.713.**
- Perfect patient-level classification: reduced noise and improved generalization.
- Selected genes: RUNX1 (known marker gene) and many other genes involved in immune response, proliferation and inflammation.

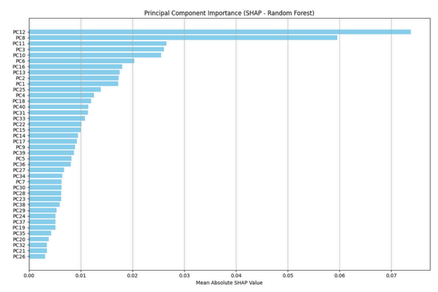




Protein Analysis

- Perform **PCA** and retain **40** principal components.
- Supervised learning approach using a **random forest** model, achieving an accuracy of **0.938** on a 30% test set.
- Compute **Shapley values** and extract the top five proteins with the highest loadings from each principal component, resulting in a set of **21 unique proteins**.





- Accuracy on the reduced dataset is **0.94**.
- These proteins are linked to relevant genes and biological functions commonly disrupted in AML.

Conclusions

While unsupervised methods are less powerful than supervised ones, this approach is valuable for **generalization**, especially when no labeled data are available.
It has been shown that is possible to detect AML accurately without relying on the full gene and protein set, by focusing on a **smaller, biologically meaningful subset**. This **improves interpretability** and **reduces complexity**. The methodology is potentially **extendable** to other genetically driven diseases involving multiple cell types, where disease signals may not be uniformly distributed.