Universitat Oberta de Catalunya

uoc.edu

# scRNAseq Analysis of AML $CD34^{+}/CD38^{-}$ population with Diagnosis and Relapse paired data

**Joseba Elizazu Perez**

MSc in Bioinformatics and Biostatistics

Area 1 – Gene and transcript expressions in cancer

**Juan Luís Trincado Alonso**

**David Merino Arranz**

02/06/2022

# FINAL WORK CARD

| | |
|---|---|
| **Title:** | scRNAseq Analysis of AML $CD34^+/CD38^-$ population with Diagnosis and Relapse paired data |
| **Author:** | Joseba Elizazu Perez |
| **Tutor** | Juan Luis Trincado Alonso |
| **PRA:** | David Merino Arranz |
| **Date:** | 02/06/2022 |
| **Studies:** | MSc in Bioinformatics and Biostatistics |
| **Area** | Area 1 – Gene and transcript expressions in cancer |
| **Languaje:** | English |
| **NUMBER OF ECTS:** | 15 ECTS |
| **Keywords** | AML, LSC, Relapse |

## Abstract – Spanish

La Leucemia Mieloide Aguda (AML) is una afección maligna originada a partir de las células madre hematopoiéticas (HSC) u otros progenitores más maduros en la médula ósea (BM). Aunque generalmente se consigue la remisión con las terapias actuales, la alta incidencia de recaída con AML resistente es el mayor desafío en el campo. Se cree que la causa más importante de la recaída reside en las células madres leucémicas (LSC), es decir, células con características similares a las HSCs en cuanto a quiescencia y capacidad de auto renovación. Actualmente no existen maradores ni terapias específicas para LSCs. Éstas se encuentran principalmente en la población $CD34^+/CD38^-$, y estudios recientes han tratado de caracterizar las LSC por marcadores de membrana celular, estado metabólico o su integración en el nicho de la BM. Sin embargo, hasta ahora ningún estudio ha tratado de analizar la población $CD34^+/CD38^-$ en pacientes de AML con muestras en el diagnóstico (Dx) y a la recaída (REL) a nivel celular. En este projecto he analizado datos de scRNAseq de 8 poblaciones $CD34^+/CD38^-$ pareadas de pacientes de AML a Dx y REL, usando *pipelines* y herramientas estándar en R a nivel de muestra, paciente y poblacional. Siguiendo esta metodología, he identificado una subpoblación celular, presente en todos los niveles y pacientes, con enriquecimiento en funciones relacionadas con LSC, y que es detectable en Dx y es más abundante en REL. En general, unido a los altos valores de fimas genéticas relacionadas con quiescencia y resistencia a fármacos, los resultados sugieren que esta subpoblación está particularmente enriquecida en LSC dentro de la poblacion $CD34^+/CD38^-$.

## Abstract

Acute Myeloid Leukemia (AML) is a clonal malignancy originated from hematopoietic stem cell (HSC) or more mature progenitors in the bone marrow. Although generally remission can be achieved with current treatments, the high relapse rate with therapy resistant AML remains as the greatest challenge in the field. The main cause of relapse is thought to reside in the Leukemic Stem Cells (LSC), namely cells with similar characteristics to the HSCs in terms of quiescence and self-renewal capacity, but on account of the heterogeneity and complexity of the disease no marker nor LSC- targeted therapy has been discovered. Commonly, LSCs have been considered to reside within the $CD34^{+}/CD38^{-}$ cell population, and recent studied have tried to better characterize LSCs by cell surface markers, metabolic state or integration in the BM niche, among others. However, no study to date has analyzed the $CD34^{+}/CD38^{-}$ population in AML patients with samples at diagnosis (Dx) and at relapse (REL) in a single-cell precision level. In this project, I have analyzed scRNAseq data of 8 $CD34^{+}/CD38^{-}$ populations paired at Dx and REL, taking advantage of standard tools and pipelines in R at sample level, patient level and whole data level. Following this approach, I have been able to define a subpopulation of cells, present in all patients, enriched in genes involved in LSC- like functions, that is already detectable at Dx and increased in REL samples. Overall, together with the high scoring on quiescence- and drug resistance- related gene expression signatures in that subpopulation, these results suggest that the aforementioned subpopulation is particularly enriched for LSCs within the $CD34^{+}/CD38^{-}$ population.

# INDEX

# Figure list

# Index of Tables

# 1 Abstract

Acute Myeloid Leukemia (AML) is a clonal malignancy originated from hematopoietic stem cell (HSC) or more mature progenitors in the bone marrow. Although generally remission can be achieved with current treatments, the high relapse rate with therapy resistant AML remains as the greatest challenge in the field. The main cause of relapse is thought to reside in the Leukemic Stem Cells (LSC), namely cells with similar characteristics to the HSCs in terms of quiescence and self-renewal capacity, but on account of the heterogeneity and complexity of the disease no marker nor LSC- targeted therapy has been discovered. Commonly, LSCs have been considered to reside within the CD34+/CD38- cell population, and recent studied have tried to better characterize LSCs by cell surface markers, metabolic state or integration in the BM niche, among others. However, no study to date has analyzed the CD34+/CD38- population in AML patients with samples at diagnosis (Dx) and at relapse (REL) in a single-cell precision level. In this project, I have analyzed scRNAseq data of 8 CD34+/CD38- populations paired at Dx and REL, taking advantage of standard tools and pipelines in R at sample level, patient level and whole data level. Following this approach, I have been able to define a subpopulation of cells, present in all patients, enriched in genes involved in LSC- like functions, that is already detectable at Dx and increased in REL samples. Overall, together with the high scoring on quiescence- and drug resistance- related gene expression signatures in that subpopulation, these results suggest that the aforementioned subpopulation is particularly enriched for LSCs within the CD34+/CD38- population.

# 2 Introduction

## 2.1 Context and justification of the Project

Acute Myeloid Leukemia (AML) is a malignancy characterized by anomalous the proliferation and accumulation of immature clonal myeloid progenitors that impair normal hematopoiesis in the bone marrow (BM). Among other manifestations, AML patients suffer from recurrent infections, anemia and easy bleeding due to the absence of normal blood cells, contributing to the morbidity of the neoplasm (1). Accounting for its morphology, response to treatment and genetic and epigenetic signatures, AML is considered a heterogeneous group of disorders (2). Although it can be developed in patients with an underlying hematologic disorder or as a consequence of a prior therapy (e.g. exposition to radiation or to alkylating agents), on most cases it appears as a *de novo* malignancy on previously healthy individuals. Its diagnostic is based on the presence of >20% of blasts in the BM or peripheral blood, extramedullary tissue infiltration or observation of t(8;21), inv(16) or t(15;17) chromosomal alterations (3).

The first classification of AML, the French-American-British or FAB, was established in 1976, and defined eight AML types ranging from M0 to M7, based on leukemic cell's morphological and cyto-chemical characteristics. Later in 2001, the World Health Organization defined a new classification with the aim to unify the advances accomplished in the diagnostic and management of AML. On its 2016 update, it distinguishes six groups of AML: **(I)** AML with recurrent genetic abnormalities, **(II)** AML with myelodysplasia-related features, **(III)** therapy- related AML, **(IV)** AML not otherwise specified, **(V)** myeloid sarcoma; and **(VI)** myeloid proliferation related to Down syndrome (3,4). Among the first group, 11 subgroups are further defined according to 11 different chromosomal translocations (Table 1). Furthermore, AML is also classified in three prognostic risk groups following cytogenetics and molecular subsets, that have different responses to standard therapies: **(A)** favorable, **(B)** intermediate and **(C)** adverse (5,6).

The causes of AML are still unknown. Genetic mutations can be found in more than 97% of the cases, usually in the absence of mayor chromosomal abnormalities.

Studies performed on murine models have led to the development of the "two-hits" leukemogenesis model, in which conjoined class I mutations (that result in the activation of pro-proliferative pathways, like FLT3, K/NRAS and TP53) and class II mutations (that compromise the normal hematopoietic differentiation, like NPM1 and CEBPA) must occur in order to develop leukemia (7). Alterations in genes that regulate the epigenome have recently been described as a third class mutations, among which genes related to DNA methylation can be found. Mutations in DNMT3A, TET2, IDH-1 or IDH-2, for example, are present in at least 40% of AML cases.

Prognosis is of mayor importance in the management of AML. Patients are stratified according risk to therapy resistance or Treatment Related Mortality (TRM). Amongst the risk-related clinical variables age and performance status are of most importance. Currently, AML treatment includes a combination of chemotherapy, use of hypomethylating agents and/or hematopoietic stem cell (HSC) transplant (2). Despite the recent advances in the diagnosis, treatment and classification of AML, Overall Survival (OS) remains poor, mainly due to the high risk of relapse.

The standard treatment, usually used on patients with a favorable or intermediate prognosis and low TRM probability, consists of a combination of 7 days of continuous administration of cytarabine followed by 3 days of an anthracycline, usually referred to as '7+3' (3). The goal of this inductory therapy is to achieve Complete Remission (CR),

**Table 1:** WHO classification of AML and related neoplasms, from De kouchkovsky & Abdul-Hay (2016)

| *Types* | *Genetic abnormalites* |
|---|---|
| AML with recurrent genetic abnormalities | AML with t(8:21)(q22;q22); RUNX1-RUNX1T1 |
| | AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBFB-MYH11 |
| | APL with PML-RARA |
| | AML with t(9;11)(p21.3;q23.3); MLLT3-KMT2A |
| | ML with t(6;9)(p23;q34.1); DEK-NUP214 |
| | AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM |
| | AML (megakaryoblastic) with t(1;22)(p13.3;q13.3); RBM15-MKL1 |
| | AML with BCR-ABL1 (provisional entity) |
| | AML with mutated NPM1 |
| | AML with biallelic mutations of CEBPA |
| | AML with mutated RUNX1 (provisional entity) |
| AML with myelodysplasia-related changes | |
| Therapy-related myeloid neoplasms | |
| | AML with minimal differentiation |
| | AML without maturation |
| | AML with maturation |
| | Acute myelomonocytic leukemia |
| | Acute monoblastic/monocytic leukemia |
| | Acute erythroid leukemia |
| | Pure erythroid leukemia |
| | Acute megakaryoblastic leukemia |
| | Acute basophilic leukemia |
| | Acute panmyelosis with myelofibrosis |
| Myeloid sarcoma | |
| Myeloid proliferations related to Down syndrome | Transient abnormal myelopoiesis |
| | ML associated with Down syndrome |

Abbreviations: AML, acute myeloid leukemia; APL, acute promyelocytic leukemia; ML, myeloid leukemia; WHO, World Health Organization.

defined as a blast concentration <5% in the BM (among other variables). Up to 80% of the patients with favorable prognostic and 50-60% with intermediate-adverse prognostic achieve CR (8). Then the post-remission treatment is administered, usually consisting of cytarabine-based intensive chemotherapy followed by HSC transplant (5). However, the majority of patients advance from the inductory therapy with a Minimal Residual Disease, and approximately 66% suffer a relapse of the leukemia, most of those during the first 18 months (9). To date, a standard therapy to treat refractory AML has not been established. Given that all tested treatments have not been successful, the recommendation is to sign the patients into clinical trials (5).

Recently different approaches to deal with the disease have emerged, ranging from mutated genes target therapy (like FLT3, IDH1 and IDH2, NPM1 or TP53), genes involved in apoptotic pathways (BCL2 and MCL1) to immunotherapy (2,10).

In this context, studying the causes, and more, precisely, the cells in charge of the relapse is crucial to advance in our knowledge of leukemogenesis as well as in the management of AML. Looking at the normal BM, the hematopoiesis is organized hierarchically with a reduced number of HSCs that produce all blood cell types, in a tightly regulated (both intrinsically by the stem cells and the surrounding BM microenviroment) differentiation and maturation process consisting of a number of intermediate progenitors (Figure 1). Different studies suggest that AML is originated from these HSC or from early myeloid progenitors (11). Hence, AML would follow a similar model, that is: hierarchically organized, originating from cells located in the aforementioned BM niche and with similar characteristics to HSC: the leukemic stem cells (LSC, also named as leukemia initiating cells) (12).

## The Leukemic Stem Cells

The clonal nature of hematologic neoplasms, including leukemia, has been recognized since the 1970 decade (13), but it was not until 1994, with the development of murine xenograph models, that LSCs were functionally defined (14). These are leukemic cells, with **(A)** self-renewal, **(B)** sustained survival in *ex vivo* co-culture system and **(C)** capacity of engraftment into immunodepressed mice. According to the LSC theory, the leukemic clones are hierarchically organized, with **(I)** the more mature or differentiated cells entering apoptosis after a number of cell divisions, originated from **(II)** cells with limitless self-renewal capacity, the LSCs (15,16). Since the LSCs

(as well as the HSCs) are capable of maintaining quiescence, they are chemotherapy-resistant. Therefore, LSCs are pointed as the main relapse drivers in AML (1).

Literature suggest the origin of the LSCs lies on pre-leukemic cells that arise from sequential mutations on HSCs or early myeloid progenitors (Figure 1). Early alterations cause the acquisition of self-renewal potential and, in most cases, impairment of differentiation. Protein-coding genes that regulate the epigenome and TP53 are among the mos common examples of these mutations. Consequent mutations in signaling pathways would enhance the proliferation and ultimately resolve in the development of AML (12,15).

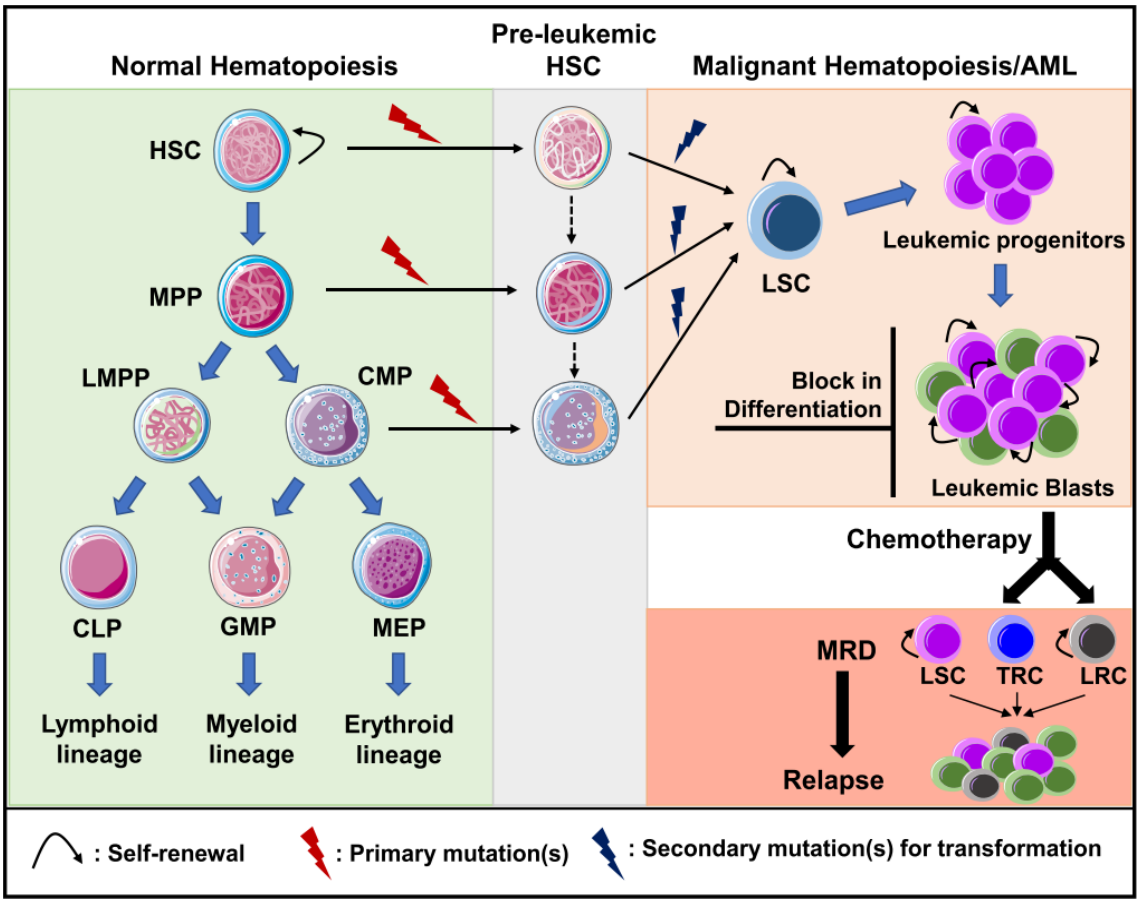


**Figure 1:** Schematic representation of normal hematopoiesis and leukemogenesis.
Initial mutations in HSCs and other more committed progenitors (multipotent progenitors (MPP), lymphoid primed multipotent progenitors (LMPP) and common lymphoid progenitors (CMP)) can give rise to pre-leukemic stem cells, that over time and with additional mutations can be further transformed into LSCs. While treatment with standard induction chemotherapy results in complete remission in the majority of AML patients, a population of (chemo)therapy-resistant cells (TRCs) constituting AML cells with leukemia-initiating potential survive the treatment. Instead of (chemo)therapeutic selection of pre-existing subpopulations of LSCs, AML cells might adaptively obtain a transient leukemia regenerating cell (LRC) phenotype upon exposure to treatment allowing for regeneration of the leukemia and clinical relapse. From Long et al., 2022.

## Cell surface phenotype of LSC

Initially, LSCs were immunophenotipically characterized as CD34+/CD38- in the peripheral blood of AML patients, This phenotype, however, made them indistinguishable from normal HSCs, and due to the heterogeneity of the disease it is expected for the LSCs to also be highly heterogeneous. Proving that the definition of LSC needs to be based on functional analyses (17), LSCs have been found in more mature cell populations (CD34+/CD38+) and in AML subtypes that do not express CD34. In the later group, associated with AML subtypes with better prognosis, it is suggested that LSCs are derived from more differentiated healthy hematopoietic cells, through mutations in genes that aberrantly grant the self-renewal capacity and stem-like properties (12). Even with the recent effort put into the characterization and targeting of LSCs, principally due to the rapid development of NGS and xenograph techniques, no LSC CD34+/CD38- markers have been found that are not also expressed on HSCs or bulk AML cells (1,12,18).

Defining LSC-specific surface markers is critical to set further investigations on the CD34+/CD38- population, to better follow the evolution of the AML patient and to develop new therapies against surface antigens in LSC. Here, the greatest feat is the heterogeneity both between and within patients. A wide variety of antigens currently associated with LSCs are involved in immunological processes, suggesting that LSC and AML bulk cells may have different immune-related interactions(1), like niche interaction signaling and immune response modulation. Many of those can be seen in Figure 2. Although most of the cell surface markers are specific to CD34 expressing AML, some have also been found in LSC regardless of CD34 expression(19,20).

## Metabolism of LSCs

Similarly to other cancer types, leukemia cells are highly glycolytic despite being in an environment with an abundance of oxygen (Warburg effect). Nonetheless, since LSCs comprise a small percentage of total AML cells (early studies estimated that they accounted for 1 per 1.2 to 5.3 x 106 cells), analysis on bulk AML is of limited value, and a better understanding of LSC's cell biology is necessary. Consistent with their similarities with HSCs, LSCs have a low production of reactive oxygen species (ROS) compared to normal AML blasts, which also extends to other cancer stem cells (17). Importantly, literature suggest that while HSCs rely primarily on glycolysis but can

maintain low ROS due to the hypoxic niche of the BM, LSC rely mostly on oxidative phosphorylation (OXPHOS) and mitochondrial function. Markedly, those quiescent, low ROS AML cells overexpress B-cell lymphoma 2 (BCL2) gene, an inhibitor of the mitochondrial-initiated apoptotic pathway and regulator of the oxidative state and mitochondrial metabolism. Inhibition of OXPHOS (either by inhibition BCL2, protein translation or directly targeting the electron transport chain) has shown to target LSCs across multiple AML subtypes. Nonetheless, despite the initial success patients with anti-BCL2 treatment (azacitidine) usually relapse with it being less efficacious in post-refractory AML, possibly indicating an adaptation to different energy production strategies (12,15,17,21).

Regarding that hypothesis of differences in LSC metabolism between *de novo* and refractory AML, a study reported that the former LSC group were more reliant to amino acid metabolism than AML blasts with high ROS levels. Treatment with azacitidine

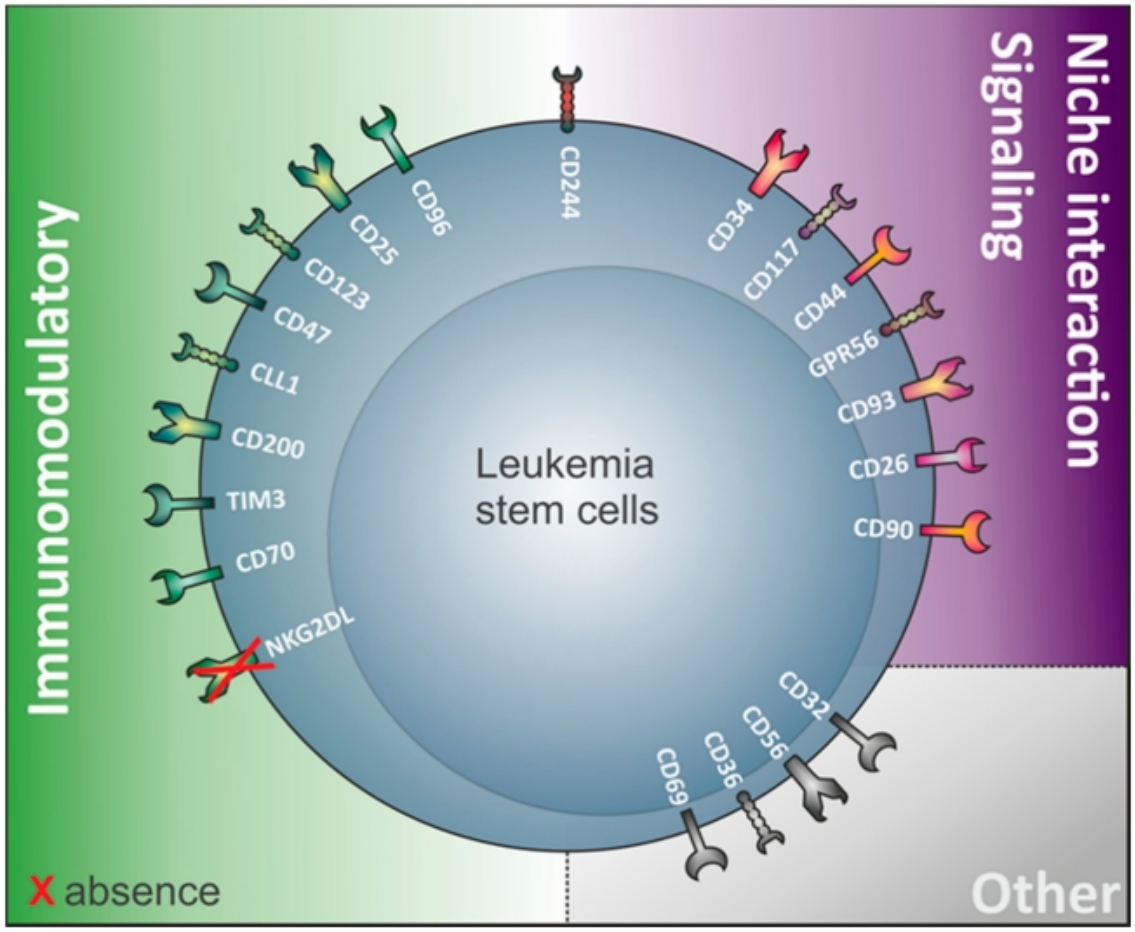


**Figure 2:** Surface antigens involved in leukemic stem cells (LSC) identification (from Arnone et al., 2020)

According to their biological function in healthy conditions, the markers are classified into 3 categories: Immune modulation (Green), BM niche interaction signaling (Purple) and other functions (Other).

reduced amino acid uptake, contributing to OXPHOS inhibition. In contrast, relapsed LSCs, on top of being resistant to azacitidine treatment, were not sensible to amino acid depletion, and likely the cells had adapted through an increase in fatty acid metabolism to maintain OXPHOS (12,22).

## The Stem Cell Niche

The normal HSC niche in the BM estroma is composed of different cell population (e.g. hematopoietic cells, osteoblasts, mesenchymal cells, adipocytes…) that tightly regulate HSC function. In AML, however, that niche is altered by the disease, making AML LSCs outcompete HSC. In the hypoxic BM compartment (Figure 12), of great importance for both HSCs and LSCs, mesenchymal cell are major regulators in maintaining quiescence and retention of HSCs by production of stem cell factor (SCF) and S-X-C motif chemokine 12 (CXCL12) (12,18).

In fact, HSC and LSC homing in the niche are regulated by CXCL12 interaction with its receptor (CXCR4) in those $CD34^+$ cells, and CXCR4 expression is associated with poor OS and relapse risk in AML patients. In line with this, CXCR4 expression is upregulated in some AML samples. Given the dependency of CXCL12/CXCR4 interaction in HSC maintenance, if this is reflected in LSCs, inhibiting it would release LSCs from the niche and force their differentiation, making them more sensitive to chemotherapy. Several CXCR4 inhibitors are already in clinical trials, with some promising results (18).

Besides these proteins, other molecules may influence cell adhesion and stemness in AML, as cytokines (probably through activation of NF-κB pathway as in other cancer types), integrins (VLA-4), cadherins and tyrosine kinase receptors (Figure 3). Recently GPR56, a G protein-coupled receptor, has been described as a novel target, being HSC and LSC specific (independently of CD34 expression), and its loss was associated with increased leukemic cell apoptosis and impairment of LSCs to adhere to the BM niche (12,18,20).

## Studying the $CD34^+/CD38^-$ population

Therefore, understanding the $CD34^+$ cell population in AML patients is crucial for the management of the disease, due to the enrichment of LSC that gives them an

important role in the development and relapse of the most aggressive types of AML (12,18). A better characterization of this population is still needed in order to develop more efficient therapies. Either by aiming at described markers with immunotherapies, metabolism or epigenomic landscape, the current strategies in clinical trials against LSCs have had severe toxicities. To solve that, it is of great interest that we find diagnostic markers, prognostic determinant and cell surface targets accessible to immunotherapy (16). Since those cells account for a small number of AML bulk, single-cell analysis is likely to give us better insight into the heterogeneity of AML within and between patients than bulk profiling methods like RNA-seq, microarrays or flow cytometry (23).

Additionally, with the metabolic state and cell surface markers in LSCs being altered after therapy, it is of major interest to assess the changes of CD34+ population at diagnostic time (Dx) and at relapse (REL) (24). Being initially underrepresented, the therapy resistant LSC population may grow out after therapy, becoming increasingly detectable, while clones sensitive to therapy may decrease or be eradicated. Consequently, currently described LSC markers may not be conserved in Dx and in REL samples(1).

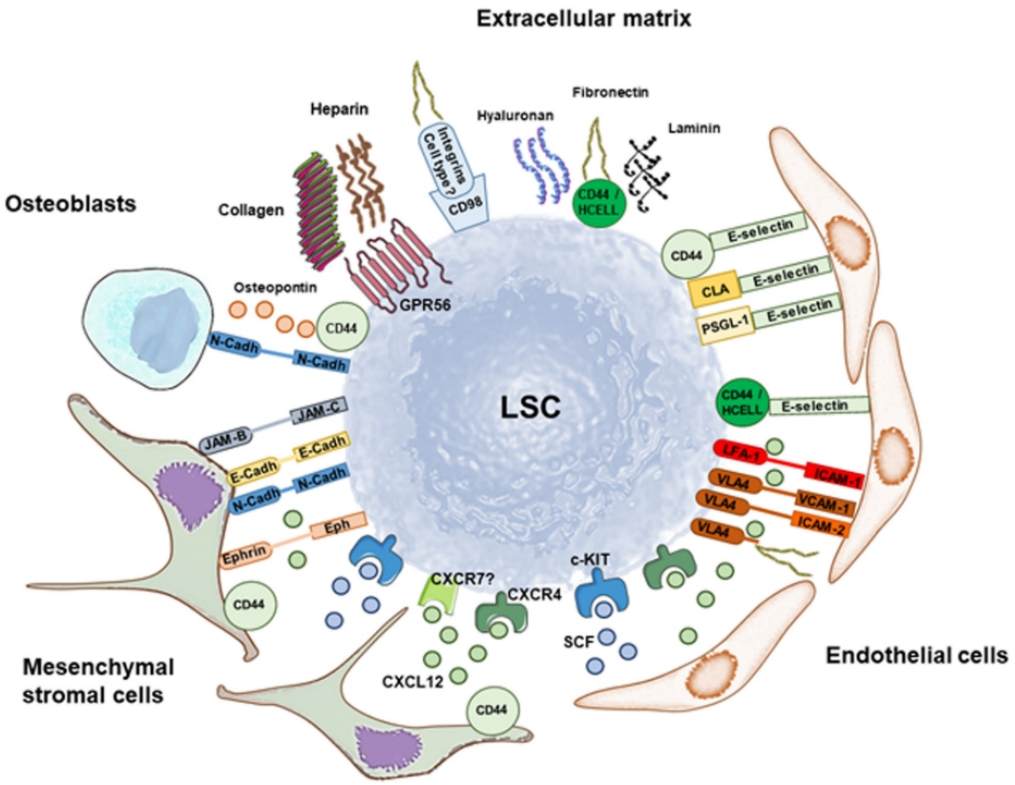


**Figure 3:** Depiction of some interactions in LSC retention, from Villatoro et al., 2020. Different molecules involved with various cell types and extracellular matrix are shown.

## 2.2 Objectives

Precisely, the main objective of this MTP is to elucidate the transcriptional landscape shift of the more immature AML $CD34^{+}/CD38^{-}$ population, in 8 patient-paired Dx and REL samples by single-cell RNAseq (scRNAseq) representing two of the most common chormosomic alterations subgroups: inv(16) (n = 2) and t(8;21) (n=2) (25). In order to do that, I have followed and used standard pipelines and tools to, **(A)** first, analyze each Dx and REL data separately to gain the first insight into the cell identity landscape by clustering. Then, **(B)** integrate the Dx and REL datasets of each patient to characterize the clusters and find possible differences between conditions, and finally, **(C)** integrate the four integrated datasets into a single one in an effort to identify coordinated changes among patients. This is achieved by completing the following specific objectives explained below: **(I)** Raw data processing and quality control (QC), **(II)** cell clustering and dimensionality reduction, **(III)** giving the clusters a biological meaning and **(IV)** data integration and final analysis.

## 2.3 Approach and methodology

To achieve the objectives, based on reviews of current available single-cell analysis tools (26,27) and on advise from my supervisor, I have chosen the following approach, considering the features of the data.

### Raw data processing and quality control

In this project I have taken advantage of the data generated by Dr Pablo Menendez's group (25). With the intent gain greater precision into the this potentially therapy resistant population, they isolated the $CD34^{+}/CD38^{-}$ by fluorescence activated cell sorting (FACS). As the data has already been processed, I have started from the CellRanger Single Cell Software Suite output (see Velasco-Hernandez et al., 2022 (25) for more information about raw data pre-processing and alignment), loaded it into an R (28) enviroment as a Seurat (29) object, and followed Seurat's vignettes. Following the standard pre-processing workflow, QC will be performed at cell level based on three commonly used covariates to ensure that all cellular barcode data correspond to viable cells: the number of counts per barcode, the number of genes per barcode, and the

fraction of counts of mitochondrial genes per barcode, with the goal of identifying dying cells, membrane-broken cells or cell doublets. Then, I have used the SCTransform based normalization (30), again following the pipeline described in its vignette.

### Cell clustering and dimensionality reduction

Principal Components Analysis (PCA) has been performed on scaled data for dimensionality reduction. The number of PC has been determined heuristically by the elbow plot method and by recognition of genes of interest among the first genes by PC loadings. Then unsupervised cell clustering has been performed based on community detection methods by constructing a K-Nearest Neighbor (KNN) graph from Euclidean distances in PCA space and the Louvain algorithm. Cluster selection has been optimized by approaches explained in methodology. Data visualization has been performed by applying Uniform Manifold Approximation and Projection (UMAP) algorithm.

### Giving the clustering a biological meaning

With the intent to give the clusters biological meaning, cluster markers were found by performing differential expression tests, based on non-parametric Wilcoxon rank sum test of a single cluster against all other cells in the dataset. Then, enrichment analysis was performed on Gene Ontology (31), KEGG (32) and Reactome (33).

Also, taking advantage of *in silico* healthy BM cell type predictions (based on Triana et al, 2021) previously projected onto this data (25,34), I was able to give the clusters obtained an additional biological meaning beyond classical differential expression or enrichment analysis.

### Data integration

Once independently analyzed, the sample datasets have been integrated, following instructions of the Seurat's integration vignette, to create four Dx-REL paired datasets. After performing the same steps explained above and ensuring that the integration was correct, those 4 datasets were merged into a final dataset.

### Analysis

For the analysis, I have focused on clusters that were more represented in REL samples with respect to the Dx samples. On those clusters, I was interested in enriched

pathways related with LSC function and projected healthy BM cell type identity, and several gene expression signatures related with LSC function were analyzed.

## 2.4 Work program

The proposed work plan is depicted in the Gantt diagram Figure 4. the main tasks carried out in the project are:

1. Raw data processing
    - 1.1. Cell-level QC
    - 1.2. Data normalization
2. Sample data dimensionality reduction and clustering
    - 2.1. PCA and finding optimal number of dimensions
    - 2.2. Clustering optimization
3. Data integration and analysis
    - 3.1. Patient paired data integration
        - 3.1.1. Integration diagnosis
        - 3.1.2. Dimensionality reduction
        - 3.1.3. Clustering optimization
    - 3.2. Full data integration
        - 3.2.1. Integration and diagnosis
        - 3.2.2. Dimensionality reduction
        - 3.2.3. Clustering optimization
4. Analysis
    - 4.1. Enrichment analysis
    - 4.2. Dx vs REL comparison

### 4.3. Signature analysis

Furthermore, multiple milestones have been set, defined by the PAC deadlines:

- PAC0: Definition of the project contents (23/02/2022)
- PAC1: Work plan definition (07/03/2022)
- PAC2: First report (11/04/2022)
- PAC3: Second report (16/05/2022)
- PAC4: Final report (02/06/2022)
- PAC5a: Presentation (06/06/2022)
- PAC5b: Public defense (24/06/2022)

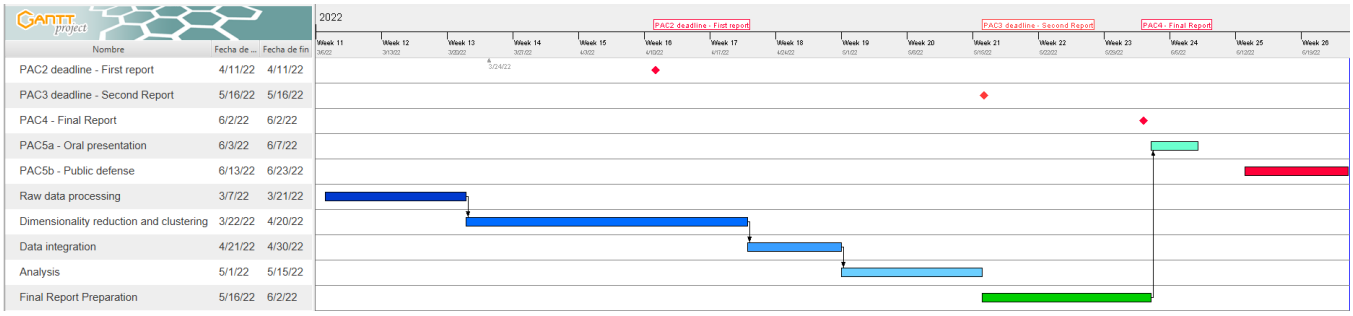


**Figure 4:** Gantt diagram of the proposed work plan for the project.
Created with the software GanttProject (https://www.ganttproject.biz/)

# 3 Methods

## 3.1 Single sample analysis

### Data information

Cells were sorted in FACS to isolate the CD34+/CD38- population and partitioned into Gel Bead-In-Emulsions using the Chromium Controller system (10X Genomics), with a target recovery of 5,000 total cells of each population. Sequencing of cDNA libraries was carried out on an Illumina NovaSeq 6000 with the goal to obtain approximately 25-30,000 reads per cell. Reads were previously aligned to Hg38 (*Homo sapiens*) reference genome and quantified by CellRanger Single-Cell Software Suite (v6.1.2). First, the cellranger mkfastq pipeline was run for demultiplexing Illumina base call files (BCL), and then the cellranger count pipeline was used to align the FASTQ files to the reference transcriptome. That output from CellRanger has been the starting point of this work. For more information about patients, samples and raw data processing carried out before, see Velasco-Hernander et al., 2022 (25). All analyses have been performed with R (versions 4.1.3 and 4.2.0) in RStudio IDE.

**Table 2:** Sample data summary

| Patient | Subtype | Dx Sample | REL Sample |
|---|---|---|---|
| AML7.10 | inv(16) | AML7 | AML10 |
| AML9.13 | t(8;21) | AML9 | AML13 |
| AML14.15 | t(8;21) | AML14 | AML15 |
| AML16.17 | inv(16) | AML16 | AML17 |

## Quality control

The main tool used has been Seurat v4.1 (29), and most procedures used have been chosen following Seurat's introductory vignette (35). All samples have been processed and analyzed following the same criteria. First, following standard procedures, mitochondrial gene percentages have been obtained using 'PercentageFeatureSet' function ins Seurat. Then, cell-level QC has been performed in three covariates: Number of counts per barcode or count depth, number of genes per barcode and counts from mitochondrial genes per barcode (27). The most usual cutoffs for each of these factors have been: below 30000 counts per barcode to avoid doublets, from 200-1,000 to 4,500-5,500 genes per barcode to filter out empty droplets and a mitochondrial gene percentage below 20% to remove dying cells or cells with broken membranes. Although, due to some quality problems with some samples (principally AML16, but also AML17, AML15 and AML10) that would be more evident in downstream steps, more strict values were chosen for those samples, thereby reducing the number of cells available for the analysis (Table 3).

## Data normalization

In this project I have chosen to move from the most standard workflow in Seurat to perform the SCTransform normalization, based on regularized negative binomial regression. According to Satija and Hafemeister (30,36), this new normalization method

**Table 3:** Cells left for analysis after cell-level QC

| Sample | Cells after QC |
|---|---|
| AML10 | 3058 |
| AML13 | 1755 |
| AML14 | 2556 |
| AML15 | 3722 |
| AML16 | 658 |
| AML17 | 2179 |
| AML7 | 2792 |
| AML9 | 1420 |

applicable to UMI based scRNAseq datasets omits the need for pseudocount addition or log-normalization, improving downstream analysis such as variable gene selection, dimensional reduction and differential expression. Given the uncertainty of the heterogeneity and LSC representation of these populations, a normalization method sharper at recovering biological distinction could be better suited for this project than other commonly used methods as log-normalization. Following the vignette, normalization was performed regressing mitochondrial gene percentage and with 3000 highly variable genes. Adding cell cycle scoring to regression was also considered, but I considered that eliminating that variability from the analysis could be misleading, given that one of the objectives of this project could be to find cells with different cell cycle phases.

### Dimensionality reduction

Linear dimensionality reduction was achieved by PCA. The number of dimensions used for downstream analysis was established with the elbow plot method. When different dumber of dimensions could have been chosen, the larger number of those was selected (except when mostly pseudogenes or ribosomal genes accounted for most of the variability of that component), following advice from Seurat vignettes and Luecken & Theis (27). Usually 10 to 20 components were selected depending on dataset complexity. The number of dimensions chosen in the previous step were chosen as input to the 'runUMAP' function of Seurat.

### Unsupervised clustering

As explained before, the unsupervised clustering has been achieved using community detection methods by constructing a KNN graph from Euclidean distances in PCA space and the Louvain algorithm. Multiple strategies have been picked in order to optimize the clustering resolution parameter of the Louvain algorithm with the intention to get a clustering that:

A) The clustering is stable to perturbations of input data. Based on the example given in the fifth chapter of the "Advanced Single-Cell Analysis with Bioconductor" online book (37), the PC with the selected dimensions was extracted and a wide range of resolutions (14 different values ranging from 0,2 to 3.2) were set. Then, using the scran package (38), the stability of the clustering

obtained with each resolution was compared. First, clustering was performed for each resolution. Due to an incompatibility between Seurat and scran, the stability could not be assessed with Seurat's functions. Therefore, the clustering was performed following the example of the book: constructing a KNN graph from Euclidean distances in PCA space and extracting the clustering with the Louvain algorithm, with the exception of the use of the Annoy algorithm for nearest neighbor identification with the intent to resemble the default clustering parameters in Seurat. Each of the obtained clusterings was stored as metadata in the Seurat object for downstream analysis. Then, the stability of that clustering was calculated with the function 'boostrapStability', with 25 bootstrap interactions. In more detail, the adjusted Rand index (ARI) is computed, defined as the proportion of pairs of cells that retain the same status in both clusterings and the number of concordant pairs expected under random permutations of the clustering is subtracted (37). Lastly, the clustering with the best stability (the one with the highest ratios in the diagonal and the lower ratios off the diagonal) was selected.

B) The clusters had biological meaning in terms of cell identity. Healthy BM cell type predictions from Triana et al. (34) were previously calculated for the data used in this project (see Velasco-Hernandez et al. (25)). In that work, a workflow based on scmap was used, and the sample code for reference atlas projection is avaliable at https://git.embl.de/triana/nrn//tree/master/Projection_ Vignette (25). With those projections, the clustering that better reflected the projected cell type heterogeneity was marked as optimal.

C) The obtained clusters had approximately 100 differentially expressed genes (DEGs) on average. If the number of clusters was too high (for example, with all clusters having about 500 DEGs), it could mean that an important part of the diversity could have been lost, given that it was not expected for the clusters to be that different among them and that the true diversity of the data should be in clusters with more subtle differences, due to the nature of the samples. However, clusters with a scarce number of DEGs could mean that the clustering highlighted differences without biological meaning.

D) There were not a large number of most DEGs repeated among clusters. In line

with the last point, many repeated cluster marker genes could mean that those repeated clusters are redundant.

E) The clusters had biological meaning in terms of DEGs. First, I extracted the DEGs by comparing each cluster versus all others making use of 'FindAllMarkers' function of Seurat. Only genes detected on at least 25% of the cells between the two conditions were tested (the specific cluster and others) with a Wilcoxon Rank Sum test (the default of the function and used on the vignettes), and the top 100 genes (increasingly ordered by p-value) with a logFC > 0.25 and a p-value adjusted (by False Discovery Rate) < 0.05 were selected as input to the enrichment analysis. Then, taking advantage of clusterProfiler v4.2 (39), Over Representation (ORA) analysis was performed using the 'compareCluster' function of the package on GO, KEGG and Reactome (via 'ReactomePA' (40) package). I chose the clustering in which the top 5 pathways/sets enriched on each cluster were not redundant with another cluster in the same setting. Overlapping clusters in terms of differentially enriched pathways could mean that those clusters are redundant.

## 3.2 Data integration

### Paired data integration

Following Seurat's vignette 'Introduction to sc-RNAseq integration' (30,41), patient paired data previously analyzed was integrated, setting 3000 integration features. Then, all steps previously outlined were repeated, with the intent of assuring that the integration was correct and that the clusters independently generated were, to a extent, conserved after integration.

### Integration into a single dataset

Then, the resulting 4 datasets were integrated. Here, 6000 features were selected for integration. With that, I expected to compensate for the expected heterogeneity of the data and get information about more subtle changes in gene expression across patients. The same methodology was followed for the clustering, but due to the size of the data and the hardware requirements to analyze the cluster stability less resolution

values were used, ranging from 0.2 to 1.6.

In order to select the increased clusters at REL, association plots were created using vcd package in R. Clusters with a Pearson residual > 2 were marked as increased.

Next, I obtained the conserved marker genes for each cluster using 'FindConservedMarkers' of Seurat. Specifically, after defining a categorical variable (Dx and REL) this function computes the differential expression analysis of that cluster's subset of the condition against all other cells of the same condition. It outputs a table with the marker genes, specifying the results of that gene for each condition tested. Since I was searching for genes conserved through both conditions, I chose the genes significantly enriched in both conditions and kept the 100 most enriched in REL.

Finally, based on the terms or pathways obtained in the enrichment analyses, different gene expression signatures from MsigDB (42,43) and recently published LSC signatures (LSC17 (44) and LSC6 (45)) were analyzed using 'AddModuleScore' from Seurat.

## 3.3 Data Visualization

UMAP plots were constructed with Seurat's functions 'RunUMAP' and 'DimPlot', For cluster, condition, patient and cell type projection coloring discrete palettes have been used consistently across all plots, with different palettes between independently analyzed sample clusters, patient-paired data clusters and whole data clusters.

ORA enrichment results have been displayed in a dotplot format with the package enrichplot (39). Association plots with Pearson residual shading have been created using vcd package in R (46–48). Signature enrichment has been presented with the functions DotPlot and RidgePlot in Seurat.

## 3.4 Data availability

All scripts used in this project, results generated and supplementary files are available at the GitHub repository: https://github.com/elizazuperez/scRNAseq-analysis-of-AML-CD34posCD38neg-population-with-Dx-and-REL-paired-data .

# 4 Results

## 4.1 Sample characteristics

In this project, I started processing and analyzing the samples independently. The obtained clustering and healthy BM predictions can be seen in Figure 5. From the UMAP plots, it is clear that the most abundant cell type are myeloblasts. Other cell types, such as lymphoid-primed multipotent progenitors (LPMP) promyelocites are also prominent, and in most samples are enriched in a small number of samples, indicating that cell type predictions distribution is, at least partially, a good predictor of cluster identity, reinforcing the approach taken in the clustering step. The promyelocites also tend to co-localize with the more committed or mature cell types. In some samples (AML16 principally, but also AML13, AML15 and AML10), ribosomal genes, mitochondrial genes and pseudogenes accounted for an important part of the variance and the number of reads was lower than in expected. In the case of AML16, the more strict QC resulted in an important decrease of cells (Table 3). Also, in those samples, the clusters located further in the UMAP (Figure 5) showed an important enrichment in ribosomal markers, and consequently, an enrichment in ribosome-related functions. Overall, except for the AML16.17 patient (Figure 5D), the number of clusters between Dx and REL samples showed no differences, and the REL samples (AML10, AML13, AML15, AML17) had a lower number of reads, higher number of cells (Table 3) and an increased importance of ribosomal genes.

## 4.2 Patient-paired integration

In order to get insight into the complexity of the data set and the main objective, it is necessary to integrate patient samples. In Figure 6, the distribution of cells of both conditions are homogeneously distributed, meaning that the datasets are correctly integrated. Importantly, in all patients regions in which either Dx or REL cells are most enriched can be found, suggesting changes in the population across conditions. Besides, the number of clusters obtained after integration (Figure 7) was unchanged. The most predominant cell types projected are also myeloblasts, with noticeable enrichment of

LPMPs, promyelocytes and eosinophil-basophil-mast cell progenitors (EBMCPs) and erytro-myeloid progenitors (EMPs) in individual clusters. Some examples are the clusters 7, 4 (in LPMPs) and 2 (promyelocytes) in the patient AML7.10; clusters 4 (LMPMs), 2 (promyelocytes) and 3 (EBMCPs and EMPs) in the patient AML9.13; clusters 5 (LPMPs), 6 (EBMCPs and EMPs) and 8 (promyelocites) in patient AML14.15; and clusters 7 (LPMPs) and 6 (promyelocytes) in patient AML16.17 (Figure 7).

Searching for clusters that are bigger in REL versus Dx (Figure 8), I found that the clusters enriched in LMPMs consistently increase at REL, while the clusters enriched in promyelocites or other more mature cell types tend to decrease. Additionally, the LPMP-enriched cluster consistently forms a loop-like pattern with another cluster that is commonly enriched in REL (cluster 7,4 and 1 in AML7.10, clusters 4 and 6 in AML9.13, clusters 5 and 7 in AML14.15 and clusters 5 and 7 in AML16.17). Interestingly, these mentioned clusters, by ORA, are consistently enriched in pathways related with **(I)** cytoskeleton function and movement, **(II)** cell cycle checkpoints and regulation (including the Reactome pathway 'GO and early G1' in AML9.13, AML14.15 and AML16.17 that suggests a quiescencent state), **(III)** DNA binding and repair, **(IV)** drug resistance and/or **(V)** TP53-related pathways across the three databases (GO, KEGG and Reactome) in which the analysis was performed (Supp. Files 1, 'patient paired results' directory).

Most interestingly, looking at the clusters obtained in the sample data sets that are most present in those clusters after integration, the same pattern of enrichment can be found (Figure 9, Supp. Files 2, 'sample results' directory). In addition, in all patients an enrichment in hypoxia and/or ROS-related gene sets was detected in at least one of the sample clusters, and in 3 out of the 4 patients (AML7.10, AML9.13 and AML14.15) it was already visible in the Dx sample clusters.

Given the clonal nature of AML, I hypothesized that while the clusters that decrease in number at relapse reflected therapy-reactive cells that died during therapy, clusters that increase at REL could reflect cells that are resistant and are able to increase their representation in relapse samples. The consistent enrichment of cell cycle regulatory gene sets, together with other functions described as important for LSC function like oxidative stress regulators, migration, DNA maintenance and quiescence

could make these clusters candidates to be enriched in LSCs (1,12,18).

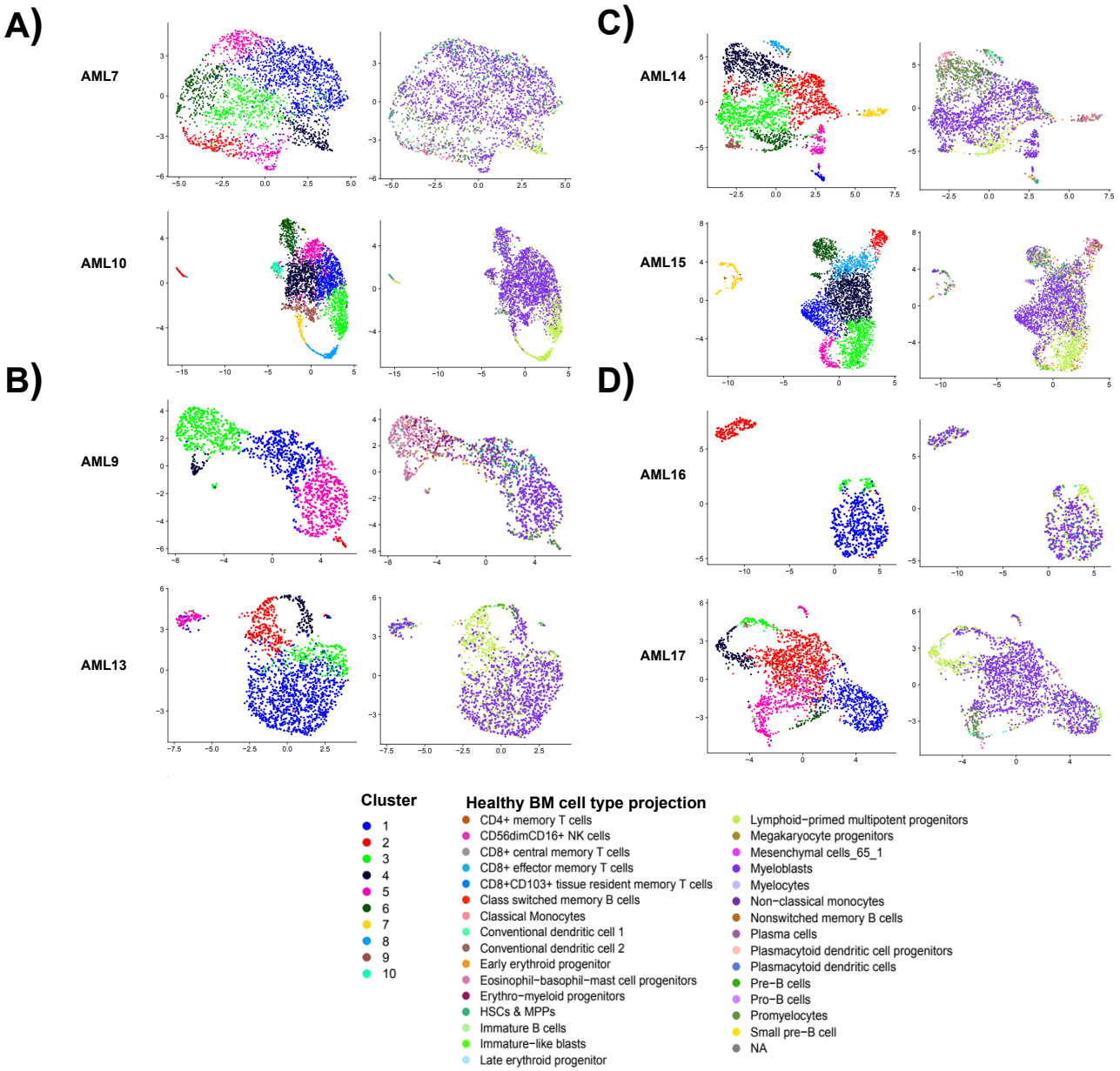


**Figure 5:** Sample clustering (Left) and healthy BM cell type (Right) in UMAP.

Up to 10 clusters have been established in the samples, without clear differences between Dx (AML7, AML9, AML14, AML16) and REL (AML10, AML13, AML15, AML17) samples. Healthy BM cell type projections show that the main projected cell type are the myeloblasts. Distinctive clusters of LPMP, myeloblasts and/or EBMCP and EMP cells can be found in all samples. Sample organization: **(A)** Samples from patient AML7.10, **(B)** Samples from patient AML9.13, **(C)** Samples from patient AML14.15, **(D)** Samples from patient AML16.17.

## 4.3 Whole data integration

Then, I wanted to understand the general overview of the data. For that, the previously obtained patient-paired data sets were integrated into a single Seurat object. With that, I expected to get a better look into the general trends between the two condition, the differences between inv(16) and t(8;21) subtypes, and the patient-specific characteristics and changes. Although there are some clusters that almost specifically include a single patient cells, as clusters 13 and 15 almost entirely consist of cells from

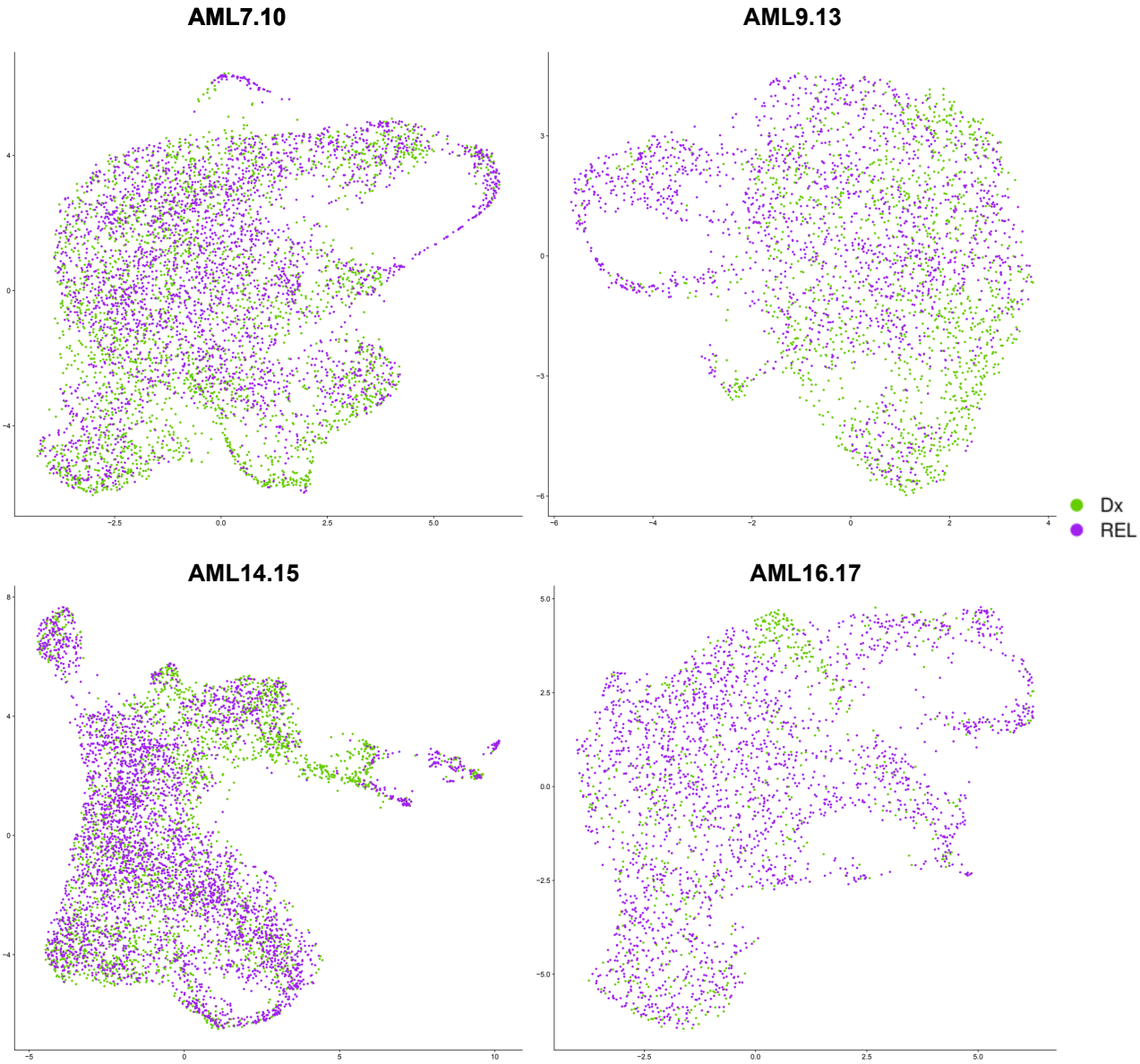


**Figure 6:** Patient data integration, distribution of Dx and REL cells in UMAP.
The distribution of Dx and REL data is homogeneous, implying that no evident batch effect can be found. However, areas with a higher quantity of Dx or REL cell can be found, suggesting shifts in the population between the two conditions.

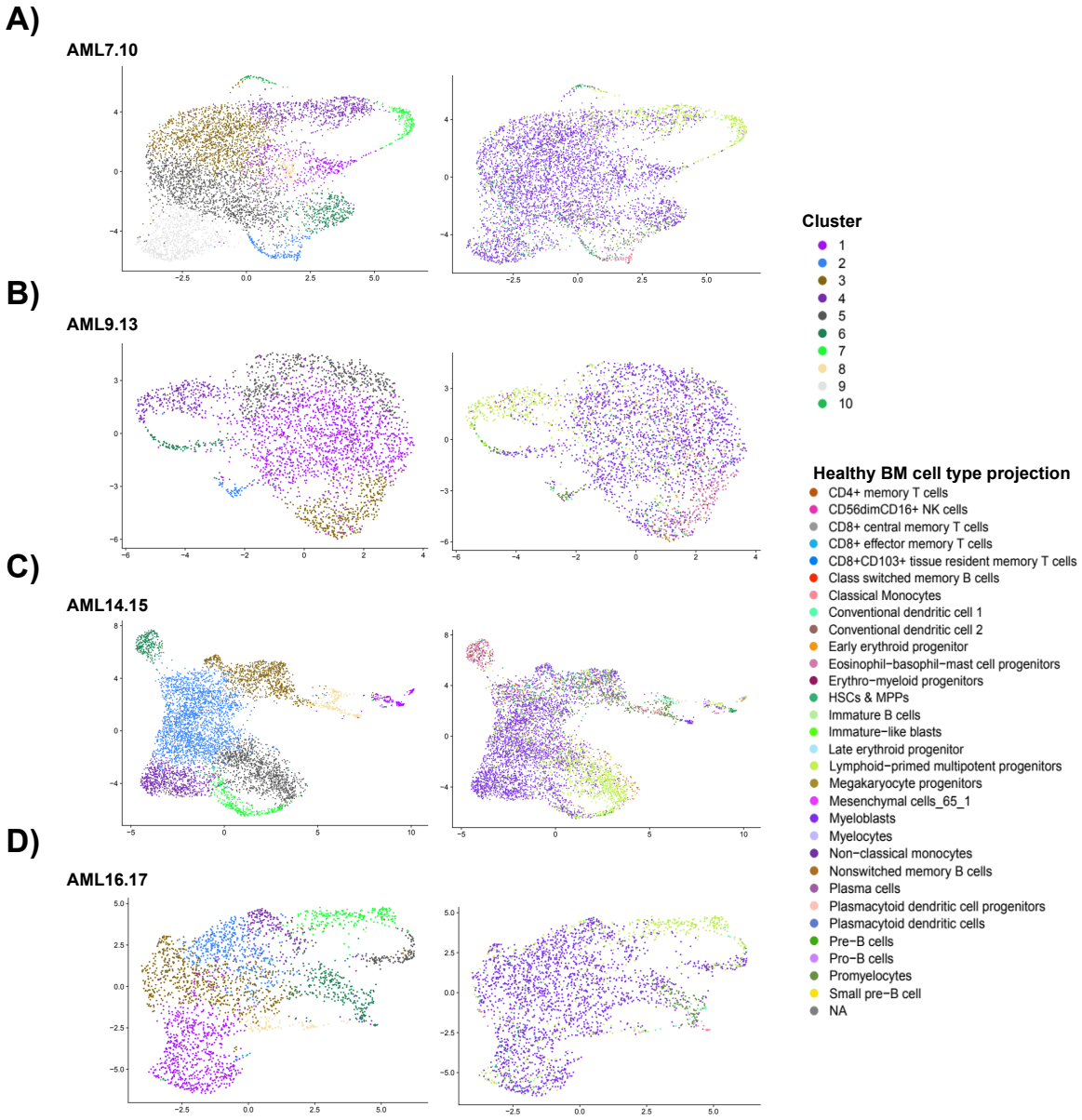


**Figure 7:** Patient-paired data clustering after integration (Left) and healthy BM cell types projections (Right) in UMAP plots.

Up to 10 clusters were defined after patient data integration, and the projections follow a similar distribution seen in the samples, with myeloblasts being the most dominant cell type, and LPMP, EMPs, EBMCPs and promyelocites being enriched in a smaller clusters.

AML14.15 patient (Figure 10A), overall the integration of the patient-paired data sets has performed similarly to that of the patient-paired data. On the one hand, different patient cells (Figure 10A) and Dx and REL cells (Figure 10B) showed no concerning distribution in clusters (Figure 10C) and UMAP space. In line with that, cell type projections (Figure 10D) displayed a similar pattern to what has been described: Myeloblasts forming the 'main body' of the UMAP plot, LPMPs, promyelocytes and other more rare and/or more mature cell types defining clusters (clusters 2, 5, 12, 13,

15).

Here, I have focused on the clusters that were significantly increased at REL, following the hypothesis described in the previous section. For that, I built an association plot (Figure 11 left) that depicts the changes in frequencies between the main condition (Dx and REL) and the clustering, and selected all clusters that had a

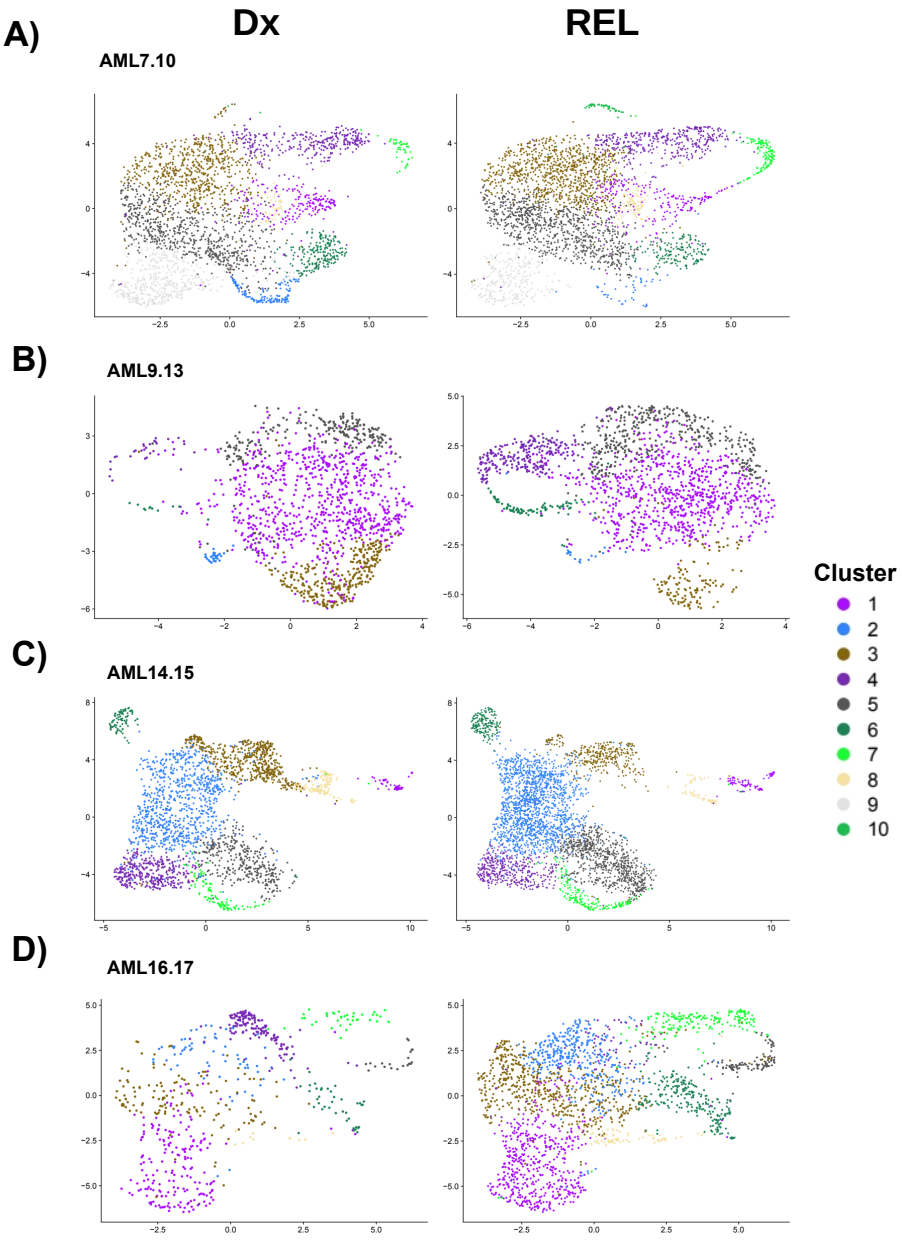


**Figure 8:** Changes in cluster size between Dx (Left) and REL (Right).
For each condition, its set of cells was plotted in UMAP maintaining the newly obtained clusters. A similar pattern of cluster size shift can be seen: in patient AML7.10 **(A)**, clusters 7 and 4 are increased at REL, equally for clusters 4 and 6 for patient AML9.13 **(B)**, clusters 5 and 7 in patient AML14.15 **(C)** and clusters 5 and 7 in patient AML16.17.

Pearson residual > 2 in REL (clusters 5, 7, 10, 12, 13, 15). However, due to the AML14.15 patient overrepresentation in clusters 13 and 15 (Figure 11 Right, Figure 10A), I opted for discarding these two clusters from the selection and continuing with clusters 5, 7, 10, and 12.

In those, clusters 5 and 12 were the most enriched in LPMPs among all clusters in the data set and cluster 10 was enriched in HSCs and multi-potent progenitors (MPPs), while cluster 7 consisted primarily of myeloblasts, without other significant enrichment among the cell types that accounted for at least 0.5% of the total projected cells in the dataset (Figure 12, Figure 13).

The enrichment analysis, however, did not replicate the pattern described in the previous section. While the selected clusters were enriched in gene sets related with LSC function, they were not specific for those clusters (Supp files 3 - 'Full data results/Normal ORA' directory). Additionally, clusters 7 and 10 showed an enrichment of ribosomal genes and pathways, suggesting that its increase in REL may be primarily due to the observed importance of ribosomal genes in relapse samples. Cluster 5 and 12 are enriched in DNA replication and damage repair, cell cycle checkpoints, TP53 function and regulation and detoxification of ROS.

Therefore, in order to get insight into the markers and processes present at Dx and REL, I performed an ORA analysis on the output from the function 'FindConservedMarkers' (Figure 14, Supp. Files 4 – 'Full data results/Dx REL conserved markers ORA'). Here, despite terms involved in DNA damage repair (clusters 3, 5, 6, 11, ,12, 14, ), cytoskeletal function (clusters 4, 8, 11, 12), oxidative stress regulators (clusters 2, 3, 10, 11) or OXPHOS (clusters 3, 6, 7, 10 and 11) that can be widely detected, others are in line with the clusters identified in the previous section: TP53 function in clusters 5, 11 and 12; Cell cycle checkpoints (including the Reactome term 'G0 and early G1') in clusters 5 and 12 (also in cluster 11, with the exception of the otlined term) and drug resistance exclusively in clusters 5, 10 and 12.

The clusters obtained in the patient-paired data are mostly well defined and distinguishable in the new UMAP coordinates (Figure 15). Unsurprisingly, the clusters underlined in the previous section are the main contributors to the clusters 5 and 12 of the full data. Despite not being significantly increased in REL, the cluster 11 is also represented in the clusters selected of the patient-paired data, and indeed, most of the

functions mentioned previous paragraph are also enriched in this pathway. The 10, however, is not represented by any of the clusters marked as interesting in the previous data sets. Additionally, the clusters that are represented in the cluster 10 did not show an enrichment in those pathways and terms, suggesting that integrating the four patient paired data sets allowed the identification of a new cluster of cells with similar characteristics to those attributed to LSCs.

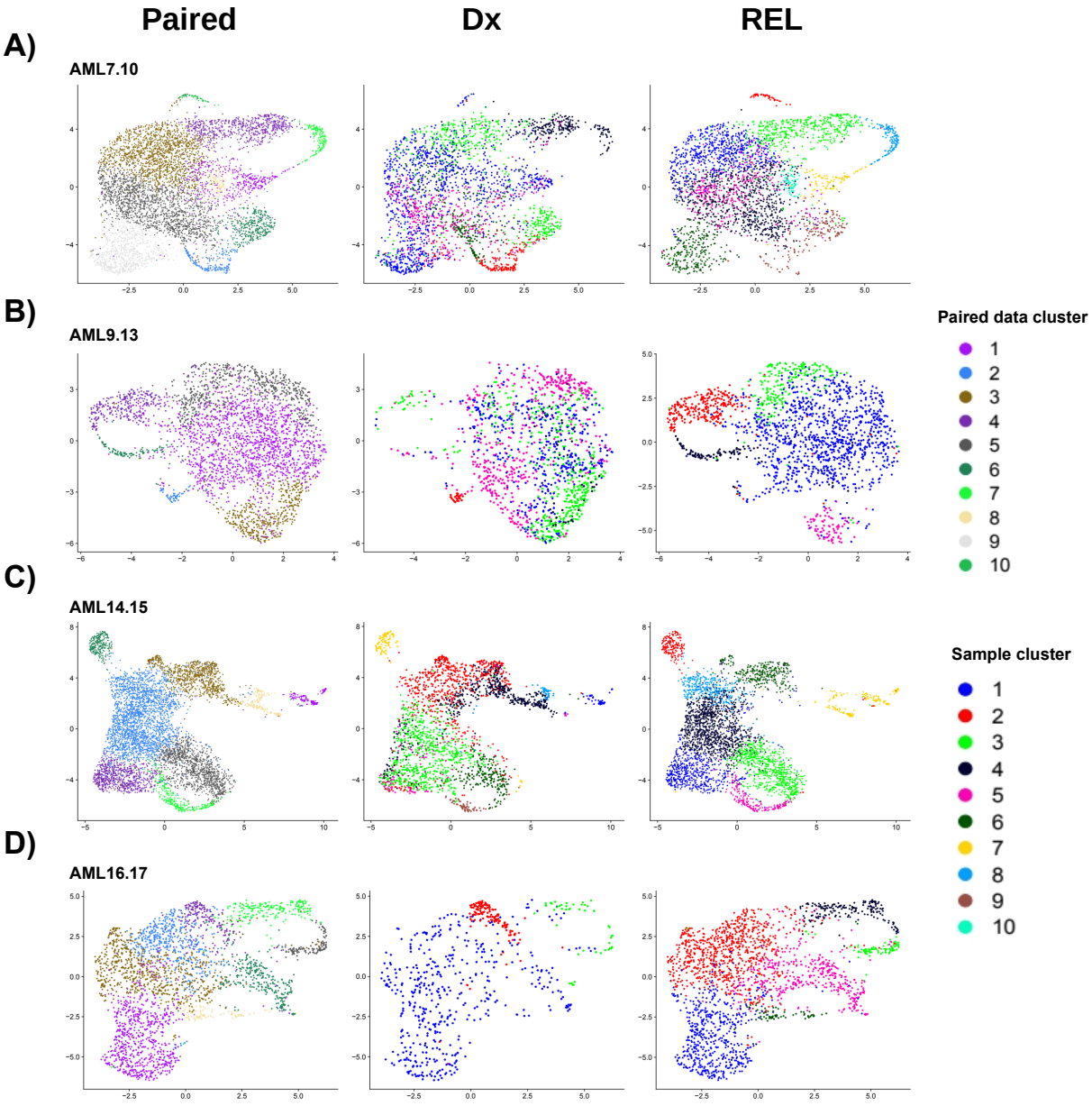


**Figure 9:** Integrated data clustering (Left) and independently obtained clusters in Dx (Center) and REL (Right) cells projected in the integrated data set with UMAP.

The clusters outlined in Figure 8 are among the best conserved in the Dx and REL clusters. Concerningly, the REL sample clustering is better defined in the integration clusters for all samples than the Dx clusters, proposing that the REL samples dominate over the Dx samples for clustering , meaning that some variation in the Dx samples could be lost after integration.

Given that, multiple gene expression signatures have been analyzed (Figure 16). Taken from MSigDB (42,43), some of the enriched terms have been added to the Seurat object via 'AddModuleScore' function in Seurat: The Reactome pathway **G0 and early G1**, the WikiPathways **Aerobic Glycolysis**, the Reactome pathway **Glycolysis**, the GO term **G0 to G1 transition**, the KEGG pathway **Drug Metabolism - Other Enzymes** (*KEGG: Drug Metabolism* arrest in the plot), the GO term **DNA Damage Response Signal Transduction by P53 Class Mediator Resulting in Cell Cycle Arrest** (*GO: p53 DNA dam. resp.* in the plot) and the Reactome pathway **TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest** (*PA: TP53 G1 arrest* in the plot) (Figure 16, Supp. Files 5 – 'Full data results/Signatures' directory). Regarding G0

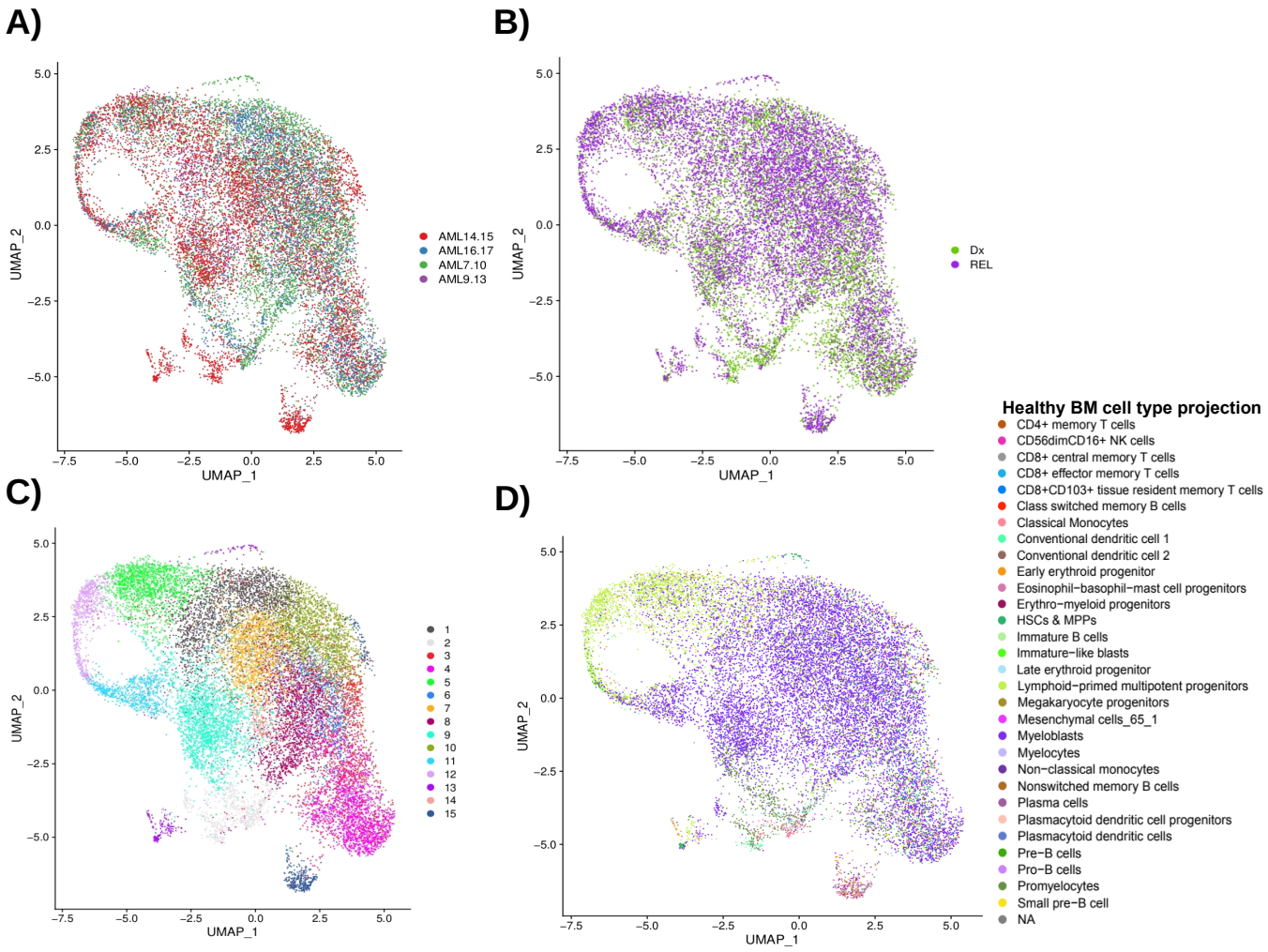


**Figure 10:** Whole data integration in UMAP plots.
**(A)** Patient data distribution. With the exception of a small subset of cells in the AML14.15 patient, the integrated data is evenly distributed. **(B)** Dx and REL cells distribution is homogeneous in the first two dimensions of UMAP, with a trend of enrichment of Dx cells in the positive values of the X axis and enrichment of REL cells in the most negative values. **(C)** Clustering after whole data integration, with 15 clusters defined. **(D)** Healthy BM cell type projections, replicating the distribution of the patient paired data and the sample data.

transition to G1 and drug metabolism, the most enriched clusters both at Dx and REL are clusters 5 and 12, while TP53 mediated cell cycle arrest is specific for cluster 12. Expectedly, glycolysis can be found homogeneously in the whole dataset, without major differences between Dx and REL. Small but green or purple points, for example in Drug Metabolism in cluster 5 suggests that a small subset of cells account for the enrichment of this function in this cluster (Figure 16B).

Moreover, two recently defined LSC signatures have been tested. The LSC17 signature (44) score values were high and homogeneous for most clusters, while the LSC6 score (45) allowed showed enrichment in clusters more associated with ribosomal function by ORA (Figure 17) and decreased in size at REL. Nonetheless, both signature scores were increased globally at REL.

Finally, the expression of previously described LSC markers has been analyzed. In contrast with the enrichment in LSC functions, the expression of markers involved in immune modulation (CD244, CD96, IL2RA, ILR3A, CD47, CLEC12A, CD200, HAVCR2, CD70, CD27, THY1), BM niche integration ( CXCR4, ITGB2, CDH2, ITGAL, ACKR3, CD9) and cell signaling (PTPRC, CD44) was lower in the selected

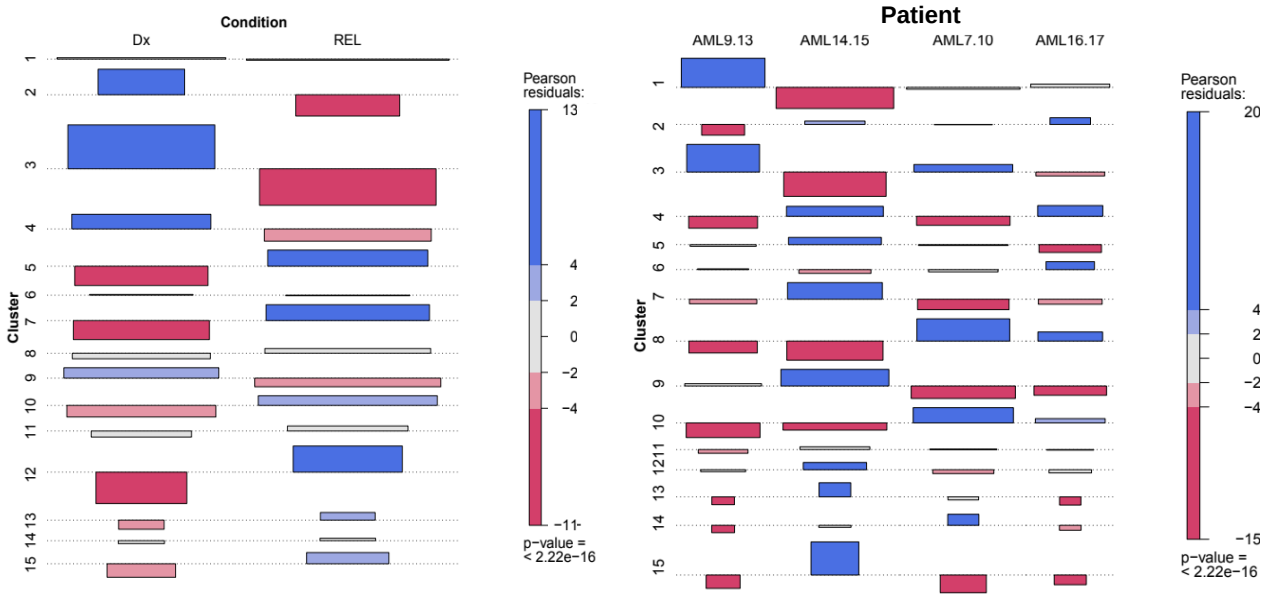


**Figure 11:** Association plot between the clusters and condition (Left) and patient (Right).
Red color indicates negative Pearson residuals smaller than 2, while blue colors indicate positive Pearson residuals bigger than 2. Below or above that threshold, the Pearson residual suggests a lack of fit in the standard normal distribution of residuals N(0,1), while values below or above 4 is considered an statistically significant ($p < 0.5$) shift. As a visualization method for contingency tables, in this plot the rectangle area is proportional to the difference in observed and expected frequencies.

clusters than in the clusters 2, 3, 4 and 8 (Figure 18). Surprisingly, cluster 2 has the lowest scores in both LSC signatures. However, clusters 5, 7 and 12 are the CD200 most expressing clusters, and have a high expression of CD96.

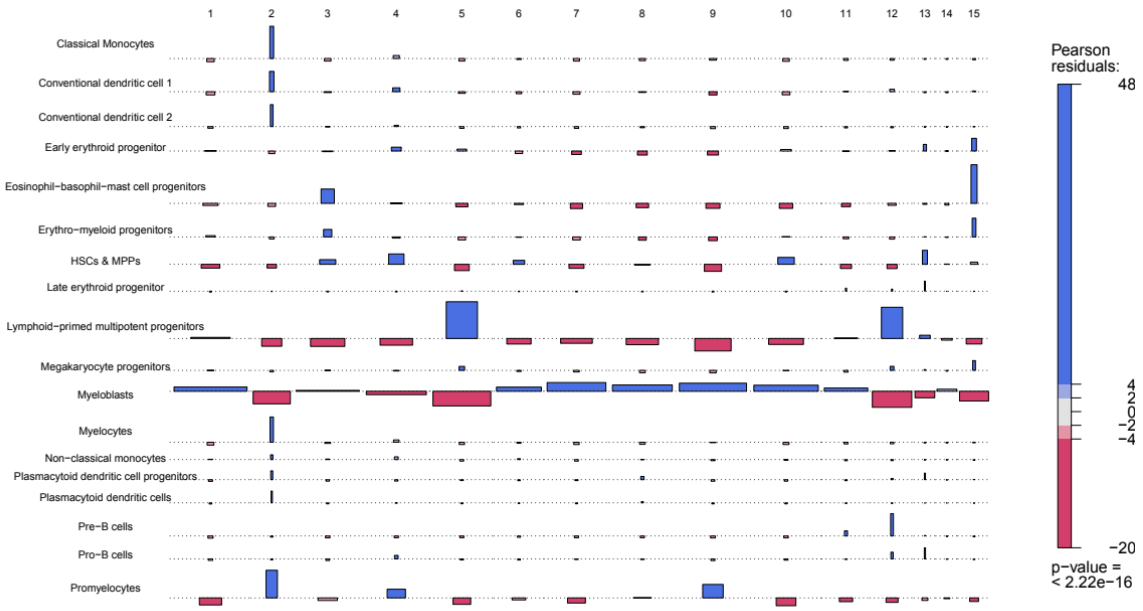


**Figure 12:** Association plot of clusters and healthy BM cell type projections.
Only the cell types accounting for at least 0.5% of the total projected cells in the dataset have been analyzed. Selected clusters are enriched for LPMPs (clusters 5 and 12), HSCs (cluster10) and myeloblasts (cluster 7). See Figure 1 for more information about the plot.

# 5 Discussion

In this project I have analyzed scRNA-seq data from 4 patients, with samples at diagnostic time (Dx) and relapse (REL). The $CD34^{+}/CD38^{-}$ cell population from those samples was isolated with FACS, with the objective to isolate the LSC-enriched population. In order to characterize the complexity of the data, first the samples were analyzed separately (Figure 5), and the clustering for each sample was optimized independently adjusting for **(A)** the stability of the clustering, **(B)** the number of DEGs for each cluster, **(C)** the number of the most 4 DEGs repeated among the clusters, **(D)** the cluster distribution in the UMAP space, **(E)** the variety of the most significantly enriched pathways after ORA across clusters, and **(F)** the distribution of projected

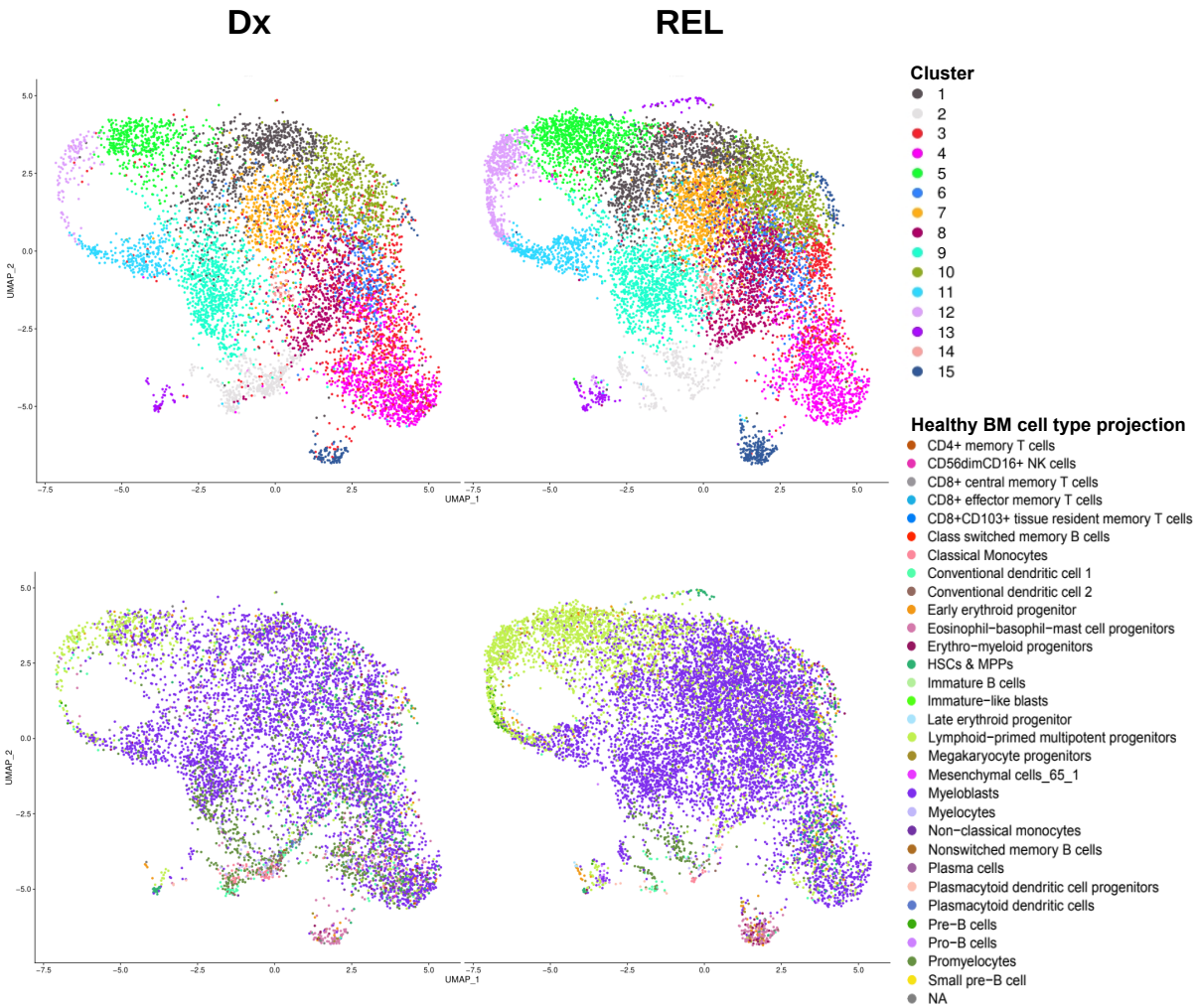


**Figure 13:** Dx vs REL cluster (Top) and healthy BM cell type projections (Bottom).
In line with the patient-paired data, the LPMPs are notably enriched in the REL samples, and are particularly concentrated in the loop-like pattern, consisting of clusters 5, 11 and 12.

healthy BM cell types across clusters. A similar number of clusters were obtained within patients in Dx and REL, and cell type projections distribution was heterogeneous within sample clusters, indicating that cell type predictions are accountable for variation detected in the data. In most cases, three group of cells were easily distinguishable in the UMAP plot (Figure 5): a centered and largest group of cells projected as myeloblasts, and two more variable and more isolated groups of promyelocites, EMPs and/or EBMCPs and LPMPs.

Due to the heterogeneity between patients, it is convenient to analyze each patient independently to better understand the landscape of the $CD34^{+}/CD38^{-}$ population. For that, I integrated the 8 samples into 4 patient-paired data sets. Here, following the approach explained previously, I performed a new clustering for each patient-paired data set, and identified 2-3 neighboring clusters commonly increased after REL with a high concentration of LPMPs in at least one of them, and forming a loop-like structure in UMAP (Figure 6, Figure 7, Figure 8). The enrichment analysis on the patient paired data showed that those clusters were enriched for pathways related to LSC function: **(I)** Drug resistance (cluster 4 in AML7.10, cluster 4 in AML9.13, cluster 5 in AML14.15, cluster 7 in AML16.17), **(II)** cell cycle checkpoints (clusters 4 and 7 in AML7.10, clusters 4 and 6 in AML9.13, clusters 5 and 7 in AML14.15, clusters 5 and 7 in AML16.17), **(III)** cytoskeleton assembly and movement (cluster 7 in AML7.10, cluster 6 in AML9.13, cluster 7 in AML14.15, cluster 5 in AML16.17), **(IV)** TP53-related functions (clusters 4 and 7 in AML7.10, cluster 4 in AML9.13, cluster 7 in AML14.15, clusters 5 and 7 in AML16.17), **(V)** DNA damage repair (cluster 7 in AML7, cluster 7 in AML14.15, clusters 5 and 7 in AML16.17) and **(VI)** telomere maintenance (cluster 4 in AML7.10, cluster 4 in AML9.13) (1,12,18,49,50). Although some of these can be occasionally found in other clusters, the drug metabolism pathways are specific for these clusters.

Interestingly, that enrichment pattern is reproduced on the sample data. The ORA of the clusters from the sample-independent data most predominant in those clusters increased at REL, in addition to the mentioned pathways, also show enrichment in functions related with oxidative phosphorylation (cluster 7 in AML10, cluster 17 in AML17, both from the inv(16) AML subgroup, possibly suggesting a subtype-specific feature) and oxidative stress (cluster 4 in AML7 and cluster 7 in AML10, cluster 5 in AML9 and cluster 4 in AML13, cluster 6 and 9 in AML14, cluster 3 in AML17)

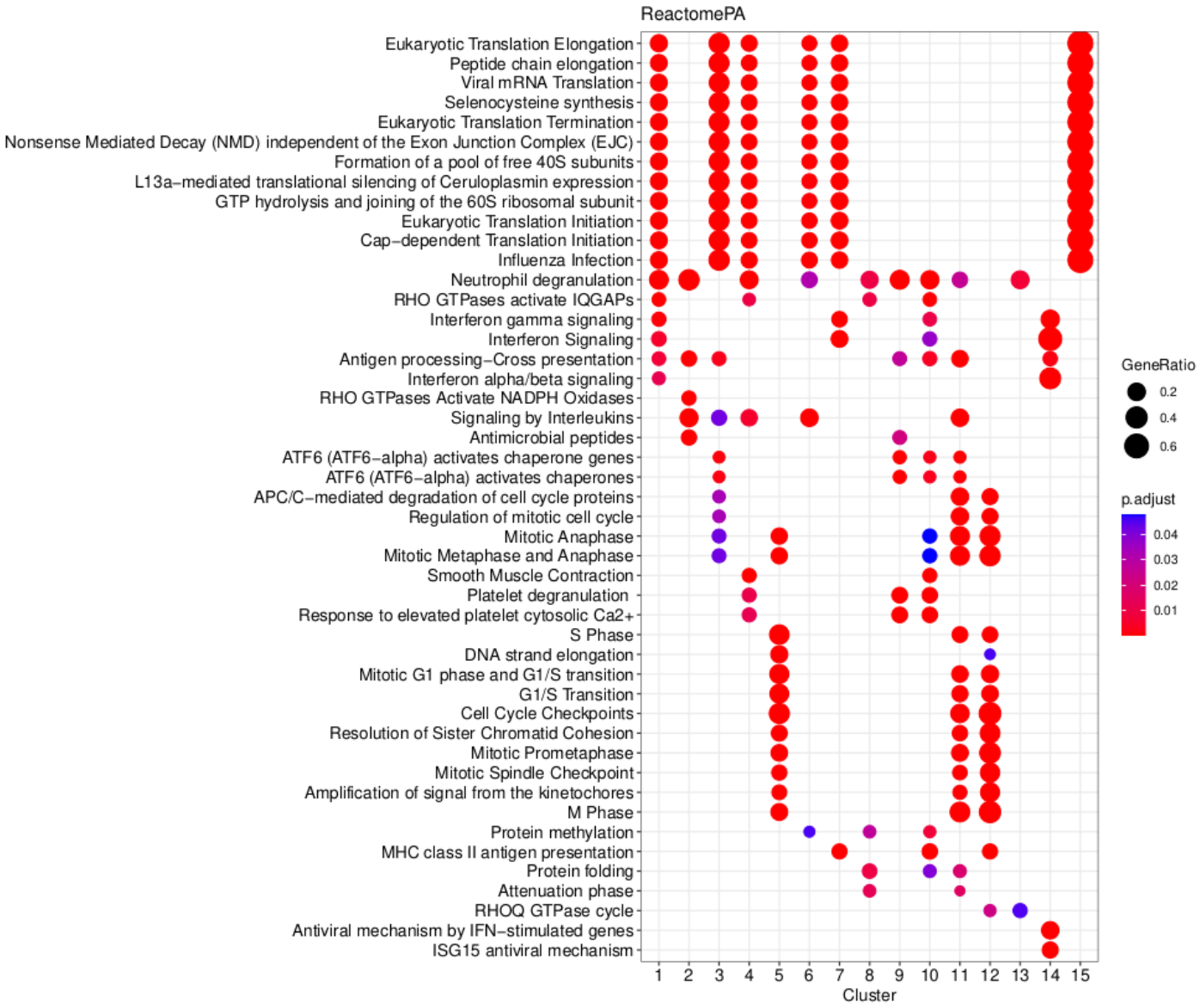


**Figure 14:** ORA analysis in Reactome database in the whole integration data set.
Here, using FindConservedMarkers function in Seurat, the top 100 DEGs for each cluster were used that were differentially expressed in Dx and REL cells. The first 5 enriched pathways (ordered by p-value) for each cluster are shown.

(Figure 9). These results point out that those clusters, or at least a small group of cells within those clusters, are already present in Dx and increased at REL, may be therapy resistant, have strict control over cell cycle and are capable of maintaining a ROS reduced state.

Then, I integrated the four patient-paired data into a single dataset. The initial diagnostic of the integration performance (Figure 10) showed that while some are specific for a single patient, most of the clusters obtained are present in all patients, hinting that while the heterogeneity between patients is an important factor, at least part

of the intrinsic variability of the $CD34^{+}/CD38^{-}$ population is shared across conditions and patients. Indeed, a greater number of clusters has been obtained in this data set than with the patient-paired data, suggesting that this approach has made it possible to identify some subpopulations of cells that were not previously defined.

Following the same approach as with the patient paired data, I was able to identify clusters that were significantly increased at REL taking advantage of association plots (Figure 11). There, the clusters 5, 7, 10 and 12 were increased and present in all patients. Those clusters, with a significant enrichment in LPMPs (clusters 5 and 12), myeloblasts (cluster 7) and HSCs and MPPs (cluster 10) (Figure 12, Figure 13), did not replicate the the ORA pattern previously explained after directly running the differential expression analysis against all other clusters (Supp files 3 - 'Full data results/Normal ORA' directory). However, this was overcome identifying the genes that were differentially expressed both at Dx and REL in each cluster and then performing the ORA with the intent to find functions that were conserved in Dx and REL in those clusters (Figure 14, Supp. Files 4 – 'Full data results/Dx REL conserved markers ORA'). Across the three used databases for ORA (GO, KEGG and Reactome), **(I)** drug resistance (clusters 5, 10 and 12), **(II)** cell cycle checkpoints (cluster 5 and 12), **(III)** cytoskeleton function (cluster 12), **(IV)** DNA damage repair (clusters 5 and 12), **(V)** TP53-related functions (cluster 5 and 12), **(VI)** oxidative stress response (cluster 7 and 10) and **(VII)** telomere maintenance functions (cluster 5) were identified in the selected clusters.

Finally, different gene expression signatures were analyzed on this data set. Overall, cluster 5 and 12 were the most commonly significantly enriched for the signatures 'G0 and early G1' and 'G0 to G1 transition' from Reactome and Gene Ontolgy, the KEGG pathway 'Drug Metabolism - Other Enzymes' and both TP53-mediated G1 arrest pathways from GO and Reactome (Figure 16). Due to the global high levels of glycolysis in AML blasts, the glycolysis signatures were not useful when distinguishing differences in metabolism in the data. Therefore, those clusters, and specially clusters 5 and 12, may be the most enriched for LSCs according to ORA and those gene signatures, being already present in the patient-paired data and sample data (Figure 15).

However, there may be methodological problems that led to these conclusions.

First of all, the consistent enrichment in cell cycle genes, function, DNA repair and cytoskeleton function may be indicating that those are clusters of cells in mitosis, which is a common source of variation in scRNAseq studies that produces systematic bias (51). Since it is expected for the cells to be differentiating, in this project I opted out of regressing for cell cycle score during normalization. A solution could be to follow the

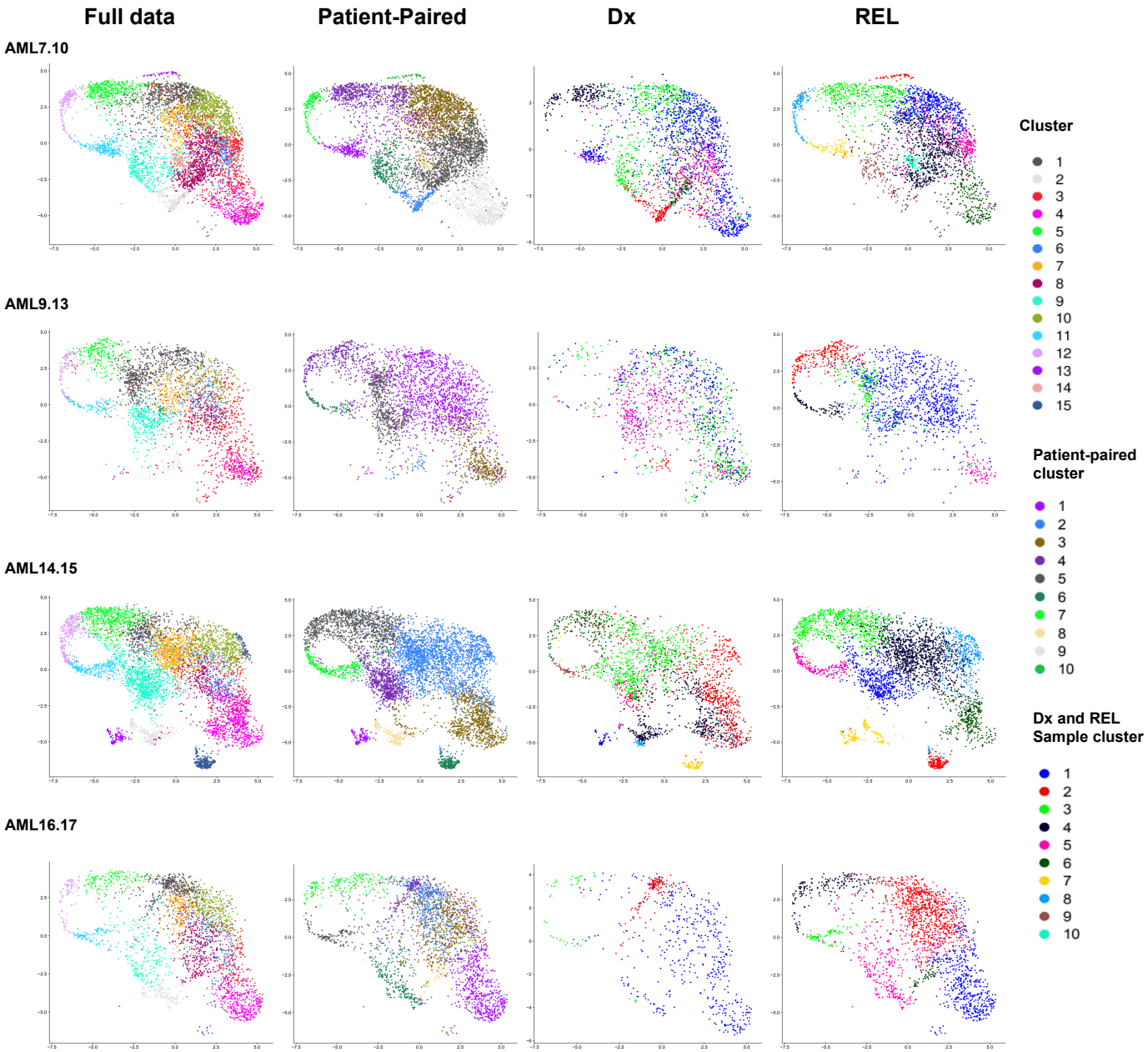


**Figure 15:** Clustering across different datasets.

The clusters obtained in the patient-paired data can be defined in most cases, implying that the integration has been correctly performed. The loop-shape forming clusters are present in all patients and in the whole integration data. However, as explained in Figure 9, the Dx clusters are more stable after integration than the Dx clusters, implying that some variability of the Dx sample may be lost after integration.

alternate workflow of Seurat's vignette on cell cycle, in which only the differences between the G2M and S phase scores are regressed out. Despite these concerns, the concentration of LPMP projection on these clusters and the unique significant enrichment in drug resistance and metabolism in those pathways indicate that the importance of these subpopulations in this project are at most partially due to the cell cycle noise.

Other concerns have arisen during dimensionality reduction and normalization. Generally, the REL samples have lower number of reads and greater variance of ribosomal genes. On top of that, the clusters corresponding to the REL samples are better conserved in the integrated data, both in the patient-paired data set (Figure 9) and in the fully integrated data set (Figure 15), meaning that the REL samples have more importance than the Dx sample for cell clustering, thus possibly masking the Dx sample population after the integration. In line with this, the enrichment analysis of REL versus Dx cells showed a constant pattern of results both comparing the total number of cells in the data and after subsetting for each cluster (Supp. Files 6 – 'Full data results/Dx vs REL ORA' directory). A possible explanation could involve the intrinsic nature of the relapse sample. At diagnosis, the samples usually contain a greater proportion of AML blasts, while after the initial diagnosis the refractory AML is usually detected earlier and the amount of blasts in the samples is lower, which results in an increased contamination of healthy cells. Therapy can also lead to more destabilized AML cells, particularly if the relapse is diagnosed shortly after treatment. This leads to a greater heterogeneity within the sample, and ultimately in a dominance of the REL data over Dx data during integrated data normalization and clustering.

Recently, various studies have tried to characterize LSCs based on gene signatures. The LSC17 signature is a 17-gene stemness scoring system, generated from a bulk RNAseq and microarray study. There, the authors first isolated the $CD34^{+}/CD38^{-}$ population form the rest of AML cells from 78 patients, and the signature performed better at predicting prognosis than other gene expression or clinical parameters (44). Then, following a similar approach, a six gene LSC signature (LSC6) was developed in pediatric patients of AML in a microarray study (45). In this data, however, those signatures have not performed as expected (Figure 17). The LSC17 signature was homogeneously expressed in most of the clusters, thereby not being useful at discerning potentially LSC enriched clusters. On the contrary, the LSC6 signature showed a

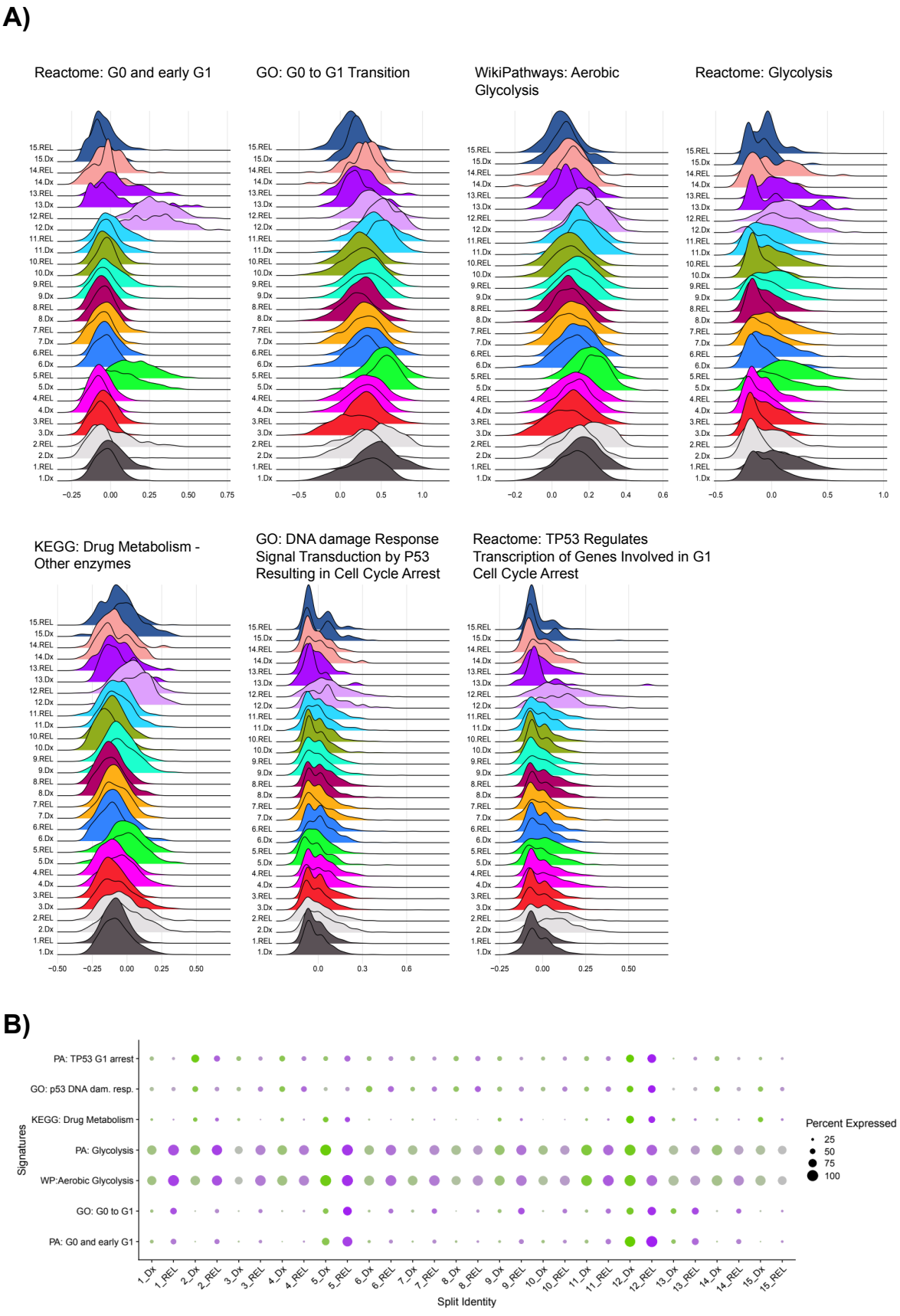


**Figure 16:** Signature enrichment.

**(A)** Ridge plots of the 7 signatures analyzed in all clusters by condition and **(B)** dot plot representation of the signatures. Overall, greater enrichment can be seen in REL, and clusters 5 and 12 are consistently among the top enriched clusters (both at Dx and REL) in all signatures.

constant increase in the REL subpopulation of each cluster than the Dx subpopulation, and with higher values in clusters most enriched by Ora in ribosomal function. The type of analysis performed in those publications (bulk transcriptome methods), as well as the definition of the signatures by risk prognosis, lead to the conclusion that those two signatures are not useful analyzing scRNAseq data of an already LSC-enriched population and trying to discern subpopulations of cells for their stemness. Besides the 'AddModuleScore' function for analyzing LSC17 and LSC6 signatures, in a previous work performed by Dr. Pablo Menendez's group they were also analyzed by assigning to the genes the original regression weights described in the paper, accounting for negative weight assigned to some genes. However, that method did not perform better than the former, probably due to the gene dropout in the scRNAseq data.

Multiple previously described as LSC marker genes have been tested in the fully integrated data set (Figure 18). Surprisingly, after clustering for the expression of those

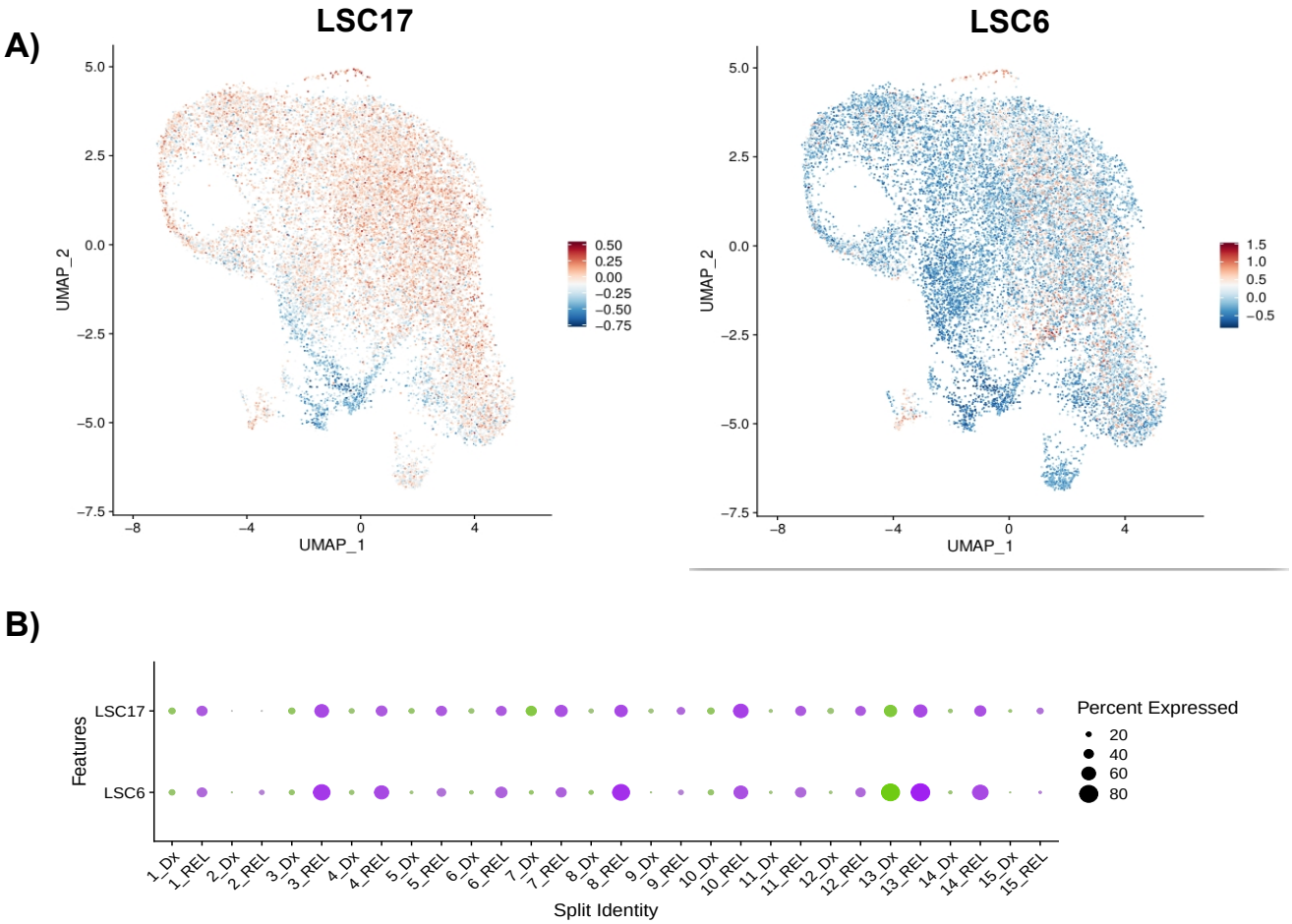


**Figure 17:** LCS17 and LSC6 signatures. Distribution in UMAP **(A)** and expression by clusters and condition in dotplot format **(B)**.
**(A)** LSC17 score is homogeneously in all cells, while LSC6 can be localized to a smaller number of clusters. **(B)** Additionally, both scores are higher for most clusters in REL (purple) than in DX (green).

markers, the clusters outlined in this project (5, 7, 10 and 12) were among the least expressing clusters. However, those were also among the clusters with the highest expression of CD200 and CD96, excluding cluster 10. CD200 has been reported to be expressed on normal HSCs and primitive AML blasts and in lymphoid lineages (which could explain the high number of LPMP projected onto these clusters), associated with poor clinical outcomes and involved in regulating immune responses (19). On the contrary, cluster 2, 3, 4 and 8 had the overall highest expression of these markers. Although more intensive analysis is needed in order to elucidate this expression pattern, those clusters have a higher enrichment in ribosomal related gene sets, but also show enrichment in other LSC-related functions such as cell-cell cadherin binding, oxidative stress and oxidative phosphorylation (Supp. Files 3 – 'Full data results/Normal ORA').

Taken together, the results in this MTP indicate that this approach was adequate for the completion of the objectives. Here, I have analyzed the transcriptional landscape changes of the more immature AML $CD34^{+}/CD38^{-}$ cell population between diagnosis (Dx) and relapse (REL), taking advantage of paired scRNAseq data obtained by Dr. Pablo Menendez's group (25). Using standard pipelines and tools, I have identified a subset of cells with distinctive characteristics at Dx that are more abundant at REL in all patients and that show an enrichment of functions related with LSCs, such as quiescence, BM niche integration, cell cycle regulation, DNA damage repair and oxidative stress response.

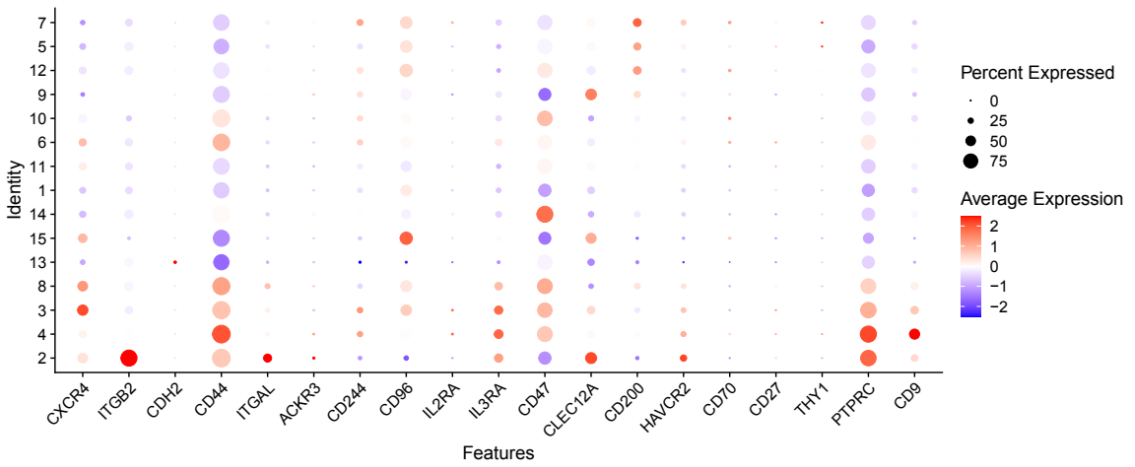


**Figure 18:** LSC marker expression.
Markers include functions in BM niche integration (CXCR4, ITGB2, CDH2, ITGAL, ACKR3, CD9), immune modulation (CD244, CD96, IL2RA, ILR3A, CD47, CLEC12A, CD200, HAVCR2, CD70, CD27, THY1) and cell signaling (PTPRC, CD44). Clusters are hierarchically ordered.

# 6 Conclusion

Due to initial doubts about the methodology proposed during the first months, the initial planning could not be strictly followed. At first, the clustering was performed only taking into account the number of DEGs per cluster, the distribution of these clusters in the UMAP plot, the ORA and the identification of LSC marker genes among the most DEGs. Then, with the intent to find a less biased method and find an optimal clustering, the healthy BM cell type projections previously calculated for this data (25) were used in order to give a biological meaning to the clusters, and the cluster stability was calculated to add a statistical significance. This process of finding a trustworthy clustering, together with the variability of the sample quality during the QC, have been the most challenging methodological issue faced in this project.

Consequently, the objectives had to be modified, and the features of the AML subgroups inv(16) and t(8;21) could not be studied. In line with this, no clear evidence of distinction between the two groups were seen, probably due to the low number of biological replicates that exacerbates the high heterogeneity of the disease.

Despite the concerns, the methodology used in this project has been decided according to standard pipelines and the characteristics of the sample. The main focus of analysis here has been the wholly integrated data set, aiming to detect more general changes between condition that could be attributable to as many patients as possible, while the patient-paired data analysis could yield better results at identifying patient-specific responses to therapy.

Additional work could include trying different QC approaches (e.g. cell cycle scoring or ribosomal gene regression to adjust for the more variable ribosomal gene expression observed in REL samples), different normalization methods (for example, the most common log-normalization and scaling before dimensionality reduction) or integration techniques. Furthermore, analyzing the total sum of results obtained in this work is beyond the scope of this project. Furthermore, validation of the results and the conclusions of this project by new techniques (e.g. trajectory analysis and subclustering) should also be considered in future work.

In conclusion, this project has been substantiated in three steps. Firstly, the sample processing and first characterization of the samples. Secondly, integration of paired data

and identification of clusters with abundance increase at REL and enrichment in LSC-like functions. Finally, using a similar approach, integration of the patient-paired data into a final data set and identification of 4 clusters increased at relapse that conserved LSC-related functions between the two conditions.

# 7 Glossary

- AML: Acute Myeloid Leukemia
- BM: Bone Marrow
- CR: Complete Remission
- Dx: Diagnosis
- EBMPC: Eosinophyl-basophil-mast cell progenitors
- EMP: Erythro-myeloid progenitors
- HSC: Hematopoietic Stem Cell
- KNN: K-Nearest Neighbor
- LPMP: Lymphoid-Primed Multipotent Progenitor
- MPP: Multipotent Progenitor
- ORA: OverRepresentation Analysis
- OS: Overall Survival
- OXPHOS: Oxidative phosphorylation
- PCA: Principal Component Analysis
- QC: Quality Control
- REL: Relapse
- ROS: Reactive Oxygen Species
- TRM: Treatment Related Mortality
- UMAP: Uniform Manyfold Aproximation and Projection

# 8 Bibliography


1. Arnone M, Konantz M, Hanns P, Stanger AMP, Bertels S, Godavarthy PS, et al. Acute Myeloid Leukemia Stem Cells: The Challenges of Phenotypic Heterogeneity. Cancers (Basel). 2020 Dec 1;12(12):1–21.
2. Döhner H, Wei AH, Löwenberg B. Towards precision medicine for AML. Nat Rev Clin Oncol. 2021;18(9):577–90.
3. De Kouchkovsky I, Abdul-Hay M. 'Acute myeloid leukemia: a comprehensive review and 2016 update''.' Blood Cancer J. 2016;6(7):e441.
4. ¿Cómo se clasifica la leucemia mieloide aguda? [Internet]. [cited 2022 May 13]. Available from: https://www.cancer.org/es/cancer/leucemia-mieloide-aguda/deteccion-diagnostico-clasificacion-por-etapas/como-se-clasifica.html
5. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017 Jan 26;129(4):424–47.
6. Pelcovits A, Niroula R. Acute Myeloid Leukemia: A Review. R I Med J (2013). 2020;103(3):38–40.
7. Takahashi S. Current findings for recurring mutations in acute myeloid leukemia. J Hematol Oncol. 2011;4:36.
8. Fernandez HF, Sun Z, Yao X, Litzow MR, Luger SM, Paietta EM, et al. Anthracycline Dose Intensification in Acute Myeloid Leukemia. N Engl J Med. 2009 Sep 24;361(13):1249.
9. Yilmaz M, Wang F, Loghavi S, Bueso-Ramos C, Gumbs C, Little L, et al. Late relapse in acute myeloid leukemia (AML): clonal evolution or therapy-related leukemia? Blood Cancer J. 2019 Feb 1;9(2):7.
10. Estey EH. Acute myeloid leukemia: 2019 update on risk-stratification and management. Am J Hematol. 2018 Oct 1;93(10):1267–91.
11. Bonnet D, Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat Med 1997 37. 1997 Jul;3(7):730–7.
12. Long NA, Golla U, Sharma A, Claxton DF. Acute Myeloid Leukemia Stem Cells: Origin, Characteristics, and Clinical Implications. Stem Cell Rev Reports. 2022;1211–26.

13. Fialkow PJ. Clonal origin of human tumors. Vol. 30, Annual review of medicine. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA; 1979. p. 135–43.

14. Lapidot T, Sirard C, Vormoor J, Murdoch B, Hoang T, Caceres-Cortes J, et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. Nat 1994 3676464. 1994;367(6464):645–8.

15. Vetrie D, Helgason GV, Copland M. The leukaemia stem cell: similarities, differences and clinical prospects in CML and AML. Nat Rev Cancer. 2020;20(3):158–73.

16. Herrmann H, Sadovnik I, Eisenwort G, Rülicke T, Blatt K, Herndlhofer S, et al. Delineation of target expression profiles in CD34+/CD38− and CD34+/CD38+ stem and progenitor cells in AML and CML. Blood Adv. 2020 Oct 21;4(20):5118.

17. Kreso A, Dick JE. Evolution of the cancer stem cell model. Vol. 14, Cell Stem Cell. Elsevier; 2014. p. 275–91.

18. Villatoro A, Konieczny J, Cuminetti V, Arranz L. Leukemia Stem Cell Release From the Stem Cell Niche to Treat Acute Myeloid Leukemia. Vol. 8, Frontiers in Cell and Developmental Biology. Frontiers Media S.A.; 2020. p. 607.

19. Ho JM, Dobson SM, Voisin V, McLeod J, Kennedy JA, Mitchell A, et al. CD200 expression marks leukemia stem cells in human AML. Blood Adv. 2020 Nov 10;4(21):5402.

20. Pabst C, Bergeron A, Lavallée VP, Yeh J, Gendron P, Norddahl GL, et al. GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. Blood. 2016 Apr 21;127(16):2018–27.

21. Morrison SJ, Scadden DT. The bone marrow niche for haematopoietic stem cells. Nature. 2014;505(7483):327.

22. Jones CL, Stevens BM, D'Alessandro A, Reisz JA, Culp-Hill R, Nemkov T, et al. Inhibition of Amino Acid Metabolism Selectively Targets Human Leukemia Stem Cells. Cancer Cell. 2018 Nov 12;34(5):724-740.e4.

23. Stetson LC, Balasubramanian D, Ribeiro SP, Stefan T, Gupta K, Xu X, et al. Single cell RNA sequencing of AML initiating cells reveals RNA-based evolution during disease progression. Leukemia. 2021;35(10):2799–812.

24. Boyd AL, Aslostovar L, Reid J, Ye W, Tanasijevic B, Porras DP, et al. Identification of Chemotherapy-Induced Leukemic-Regenerating Cells Reveals a Transient Vulnerability of Human AML Recurrence. Cancer Cell. 2018;34(3):483-498.e5.

25. Velasco-Hernandez T, Trincado JL, Vinyoles M, Closa A, Gutiérrez-Agüera F, Molina O, et al. A comprehensive single-cell expression atlas of human AML leukemia-initiating cells unravels the contribution of HIF pathway and its therapeutic potential. bioRxiv. 2022 Mar 4;2022.03.02.482638.

26. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc. 2021;16(1).

27. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6).

28. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2022.

29. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.

30. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019 Dec 23;20(1):1–15.

31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25.

32. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010 Oct 30;38(Database issue).

33. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D687–92.

34. Triana S, Vonficht D, Jopp-Saile L, Raffel S, Lutz R, Leonce D, et al. Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. Nat Immunol 2021 2212. 2021 Nov 22;22(12):1577–89.

35. Seurat - Guided Clustering Tutorial • Seurat [Internet]. [cited 2022 May 23]. Available from: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

36. Using sctransform in Seurat • Seurat [Internet]. [cited 2022 May 23]. Available from: https://satijalab.org/seurat/articles/sctransform_vignette.html

37. Advanced Single-Cell Analysis with Bioconductor [Internet]. [cited 2022 May 22]. Available from: http://bioconductor.org/books/3.15/OSCA.advanced/

38. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 2016 52122. 2016 Oct 31;5:2122.

39. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov. 2021 Aug 28;2(3).

40. Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016 Jan 26;12(2):477–9.

41. Introduction to scRNA-seq integration • Seurat [Internet]. [cited 2022 May 23]. Available from: https://satijalab.org/seurat/articles/integration_introduction.html

42. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 12;1(6):417.

43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.

44. Ng SWK, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature. 2016;540(7633):433–7.

45. Elsayed AH, Rafiee R, Cao X, Raimondi S, Downing JR, Ribeiro R, et al. A 6-gene leukemic stem cell score identifies high risk pediatric acute myeloid leukemia. Leukemia. 2020 Mar 1;34(3):735.

46. Meyer D, Zeileis A, Hornik K. vcd: Visualizing Categorical Data. 2021.

47. Meyer D, Zeileis A, Hornik K. The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd. J Stat Softw. 2006;17(3):1–48.

48. Zeileis A, Meyer D, Hornik K. Residual-based Shadings for Visualizing (Conditional) Independence. J Comput Graph Stat. 2007;16(3):507–25.

49. Zarou MM, Vazquez A, Vignir Helgason G. Folate metabolism: a re-emerging therapeutic target in haematological cancers. Leuk 2021 356. 2021 Mar 11;35(6):1539–51.

50. Ghatak D, Das Ghosh D, Roychoudhury S. Cancer Stemness: p53 at the Wheel. Front Oncol. 2021 Jan 11;10:2910.

51. Barron M, Li J. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. Sci Reports 2016 61. 2016 Sep 27;6(1):1–10.

# 9 Supplementary files

All scripts, results, and supplementary files are available at the GitHub repository https://github.com/elizazuperez/scRNAseq-analysis-of-AML-CD34posCD38neg-population-with-Dx-and-REL-paired-data.