Single Cell RNA-Seq Analysis: Normal Hematopoietic Stem Cells and Leukemic Stem Cells in Chronic Myeloid Leukemia

Brief Summary

Chronic myeloid leukemia (CML) is a hematological malignancy driven by the BCR:ABL1 fusion gene, affecting various cell lineages and originating from myeloid progenitors within the bone marrow. This study leverages single-cell RNA sequencing (scRNA-seq) data from patients in the chronic phase of CML to elucidate the dynamics of leukemic stem cells (LSC) and hematopoietic stem cells (HSC). After comprehensive preprocessing, including quality control, normalization, and batch effect correction, 18,283 cells from the GSE218184 dataset were analyzed. Highly variable genes were identified, and principal component analysis (PCA) coupled with uniform manifold approximation and projection (UMAP) revealed nine distinct cell clusters. Differential gene expression analyses pinpointed top markers for each cluster, facilitating cluster annotation. The expression patterns of major markers representing different cell types were visualized, shedding light on the heterogeneity within these clusters. Diseased cells cluster negatively along PC1, while normal cells disperse positively, indicating substantial molecular differences. Specific genetic characteristics associated with leukemia are suggested by the distinct clustering of diseased cells. Figure 11 identifies top genes, including PDLIM4, S100A9, MIR3681HG, CCL2, and IFI44L. Processes like "response to bacterium," "leukocyte migration," "adaptive immune response," "T cell activation," "inflammatory response," and "chemotaxis" are notably enriched, signifying the gene set's strong association with immune-related functions. Additionally, terms related to cell adhesion, such as "cell-cell adhesion" and "leukocyte cell-cell adhesion," exhibit notable enrichment, implying a potential role in mediating cell interactions critical for physiological processes, including immune responses. This scRNA-seq analysis offers a comprehensive view of the CML landscape at the single-cell level, providing valuable insights into the potential roles of leukemic stem cells and avenues for targeted therapeutic interventions.

1. Introduction

Chronic myeloid leukemia (CML) is a malignant blood disorder caused by the presence of BCR:ABL1 in a cell that possesses inherent or acquired biological capabilities for leukemia development[1,2]. This ongoing and sometimes fatal malignancy is associated with various cell lineages, including erythroid, monocytic, myeloid, megakaryocytic-B, and occasionally T-lymphoid[3]. Notably, CML originates within the bone marrow, arising from myeloid CD34+/CD38-/CD90+ progenitors. In the chronic phase of CML, the proliferation of leukemia cells is regulated, allowing them to mature normally and respond appropriately to normal regulators, such as granulocyte-colony-stimulating and macrophage-colony-stimulating factors [1]. The root cause of CML lies in the incomplete differentiation of blood or hematopoietic stem cells into mature cells, leading to the accumulation of immature hematopoietic stem cells in both the bone marrow and peripheral blood. At the cytogenetic level, CML is associated with a reciprocal translocation event between chromosomes 9 and 22, resulting in the formation of a shortened "Philadelphia chromosome"[4].

TKIs, anti-BCR::ABL medications, have been instrumental in helping over 80% of CML patients achieve long-term remissions, even curing approximately one-third of patients. TKIs action minimizes CML proliferation, blocks cell signaling, and induces cell death in CML clones - effectively inducing deep molecular remission in the disease[5]. However, chronic therapy induces changes in CML cancer stem cells, leading to new resistance and a significant percentage of patients experiencing relapses upon discontinuation of TKI therapy. This phenomenon suggests the survival of CML stem cells (CML-SC) during the remission stage[6,7].

Studies have demonstrated that CML-SC in the chronic phase express identical surface markers (Lin−CD34+CD38−) as normal hematopoietic stem cells (HSC)[2]. To differentiate between CML-

SC and normal HSC, various cell surface markers, such as CD25, CD26, CD33, CD93, and IL1RAP, have been identified based on BCR::ABL1 enrichment[8–11]. This project intends to identify cell clusters from haemopoietic stem cells (HSC) and leukaemic stem cells (LSC) found in patients with chronic myeloid leukaemia (chronic phase).

2. Methodology

2.1 Data set

The scRNA-Seq data used for this analysis was gotten from NCBI-Gene Expression Omnibus Database. The dataset of interest was identified with the unique identifier (ID) GSE218184. GEO Accession viewer (nih.gov) The data was submitted by Stevens T et al., 2023. In the experiment, single-cell RNA sequencing (RNA-seq) was conducted on hematopoietic stem cells (HSC) extracted from CD34+ cell samples, obtained by mobilization with granulocyte-colony stimulating factor (G-CSF), derived from five unrelated individuals with no bone marrow disease. Additionally, single-cell RNA-seq was carried out on leukemic stem cells (LSC) isolated from CD34+ samples of five unrelated patients in the chronic phase of chronic myeloid leukemia (CML).

The data set contain 10 samples, of which 6 sample was used for the scRNA-seq analysis. The 6 samples are:

Samples	Label
GSM6736187	CML patient sample A scRNAseq
GSM6736188	CML patient sample B scRNAseq
GSM6736189	CML patient sample C scRNAseq
GSM6736193	Normal sample B scRNAseq
GSM6736195	Normal sample D scRNAseq
GSM6736196	Normal sample E scRNAseq

2.2 Processing of the scRNA-seq Data

The “Seurat” package was employed for the scRNA-seq data processing, including quality control, data exploration, statistical analysis, and result visualization[12]. First, low quality cells were excluded according to the following quality control criteria:

- genes detected in <1000 cells;
- cells with <1000 or >30,000 detected genes; and
- cells with >5% of mitochondrial expressed genes.

Then, the batch effects of the scRNA-seq data were corrected by “harmony” package[13].

Third, the scRNA-seq data were normalized through “LogNormalize” method and subsequently, top 1,000 highly variable genes were identified by “VST” package[14].

Next, principal component analysis (PCA) was used for dimensionality reduction of the cells to determine the significantly available dimensions ($P < 0.05$)[15]. Based on top 20 PCs, the uniform manifold approximation and projection (UMAP) algorithm was utilized for dimensionality reduction and clustering across all cells [16]. Genes with the cut of criteria of adjusted $P < 0.05$ and |log2 fold change (FC)| > 1 were regarded as the marker genes in each cluster through “seurat” package.[17]

3. Results

Identification of Cell Clusters Using scRNA-seq Data

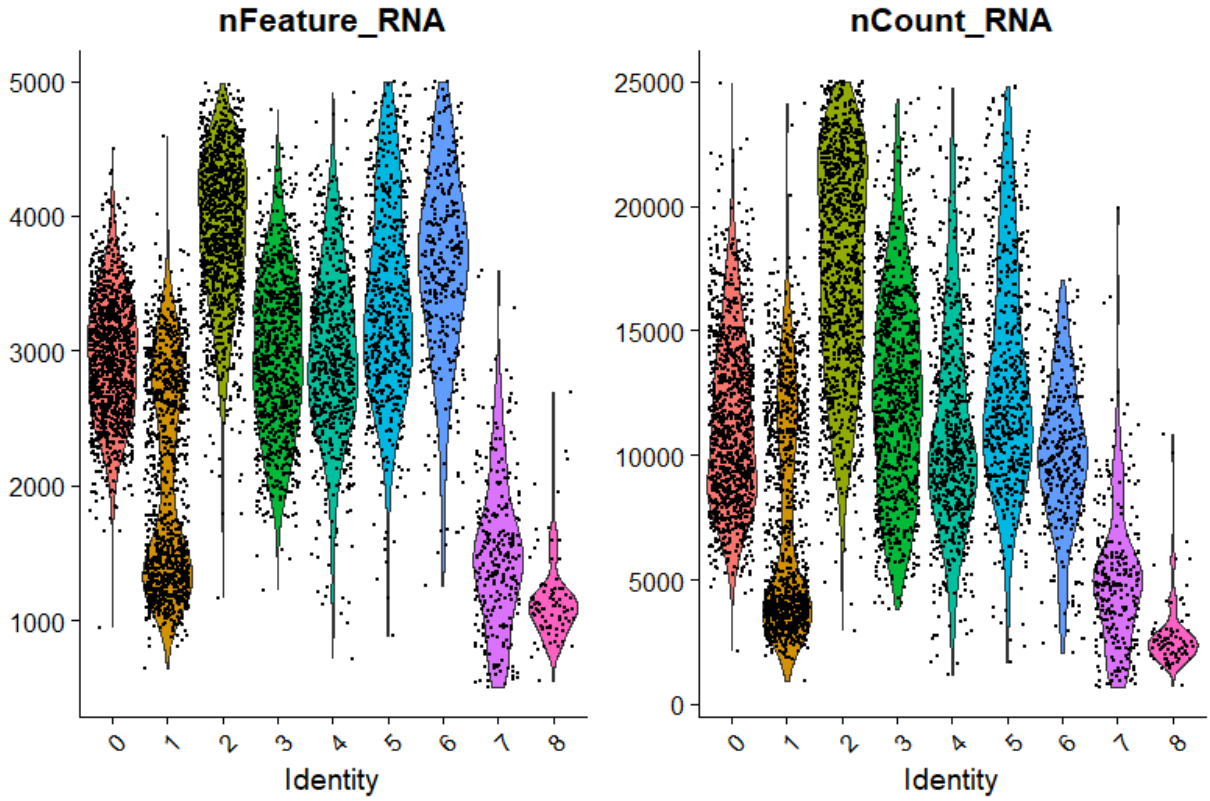

Figure 1: Violin plots of the RNA information of processed scRNA-seq data

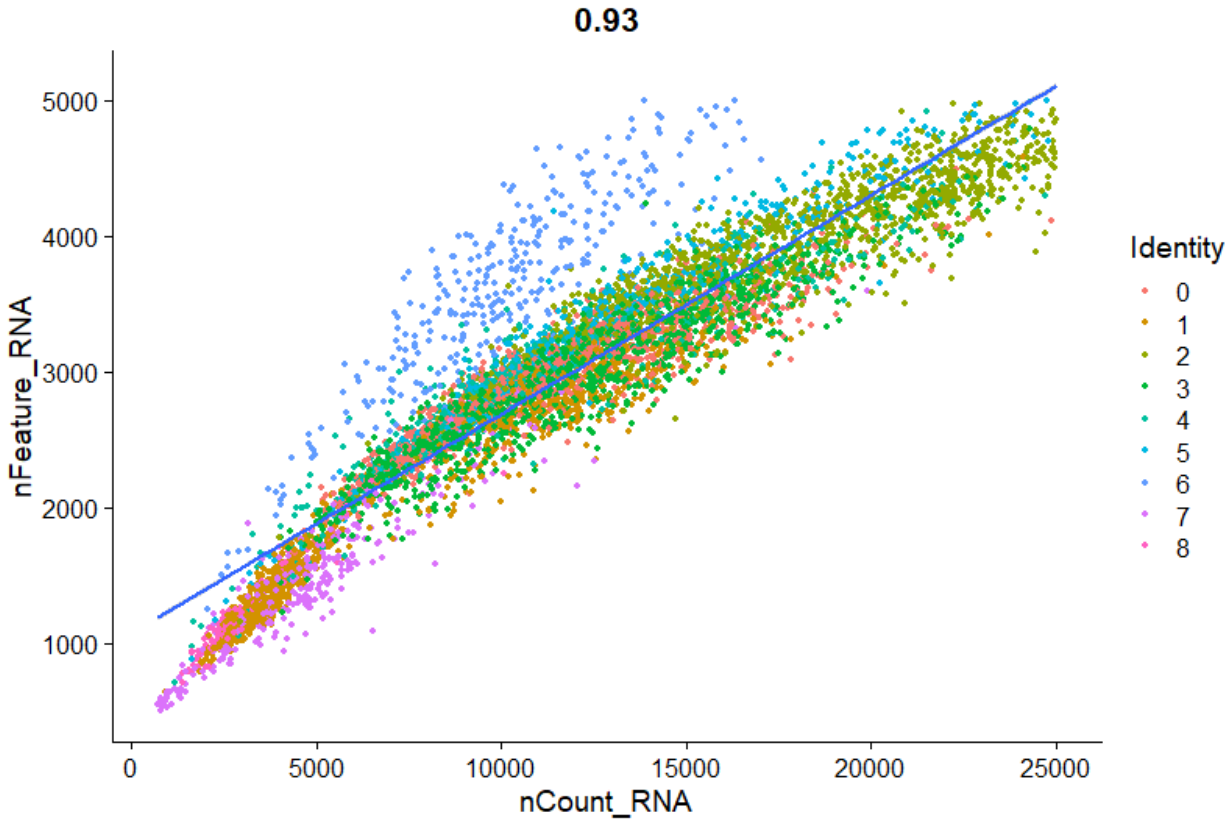

Figure 2: Scatter plot of the correlation between the numbers of detected genes and sequencing depth

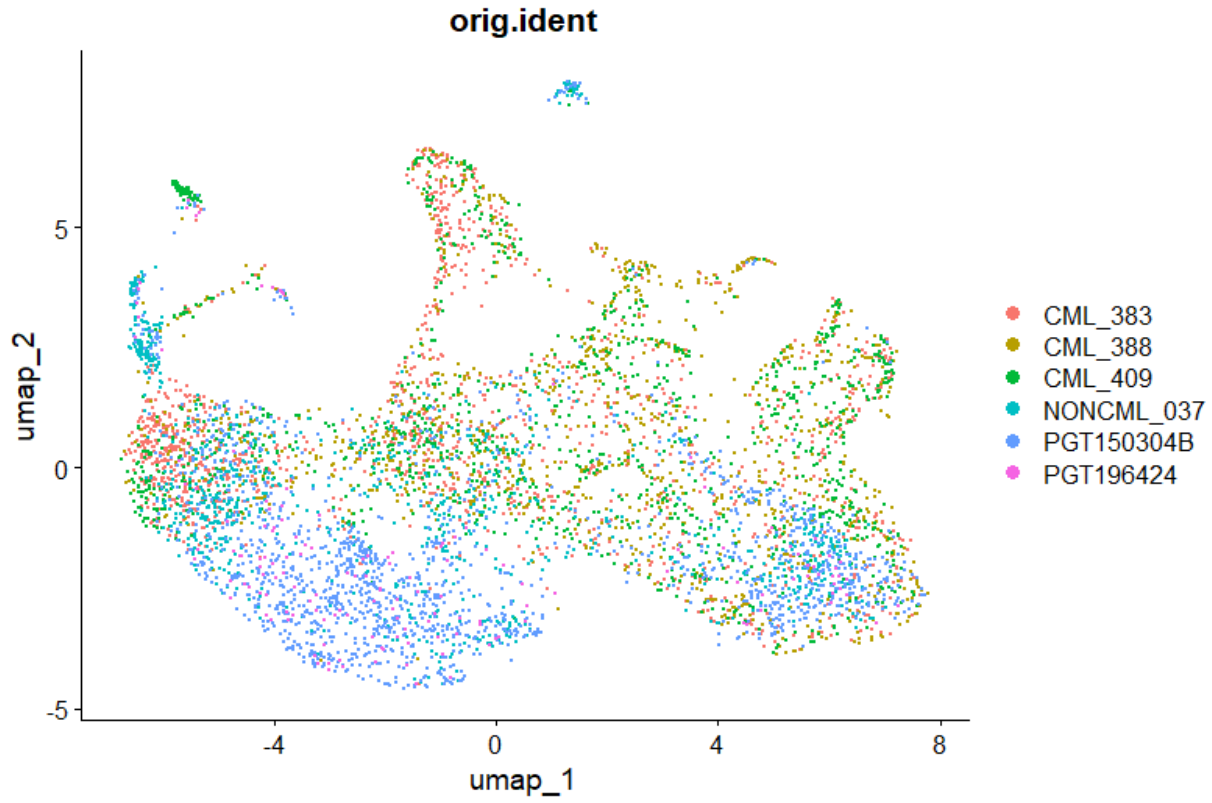

Figure 3: Scatter plot of the batch effect after correction

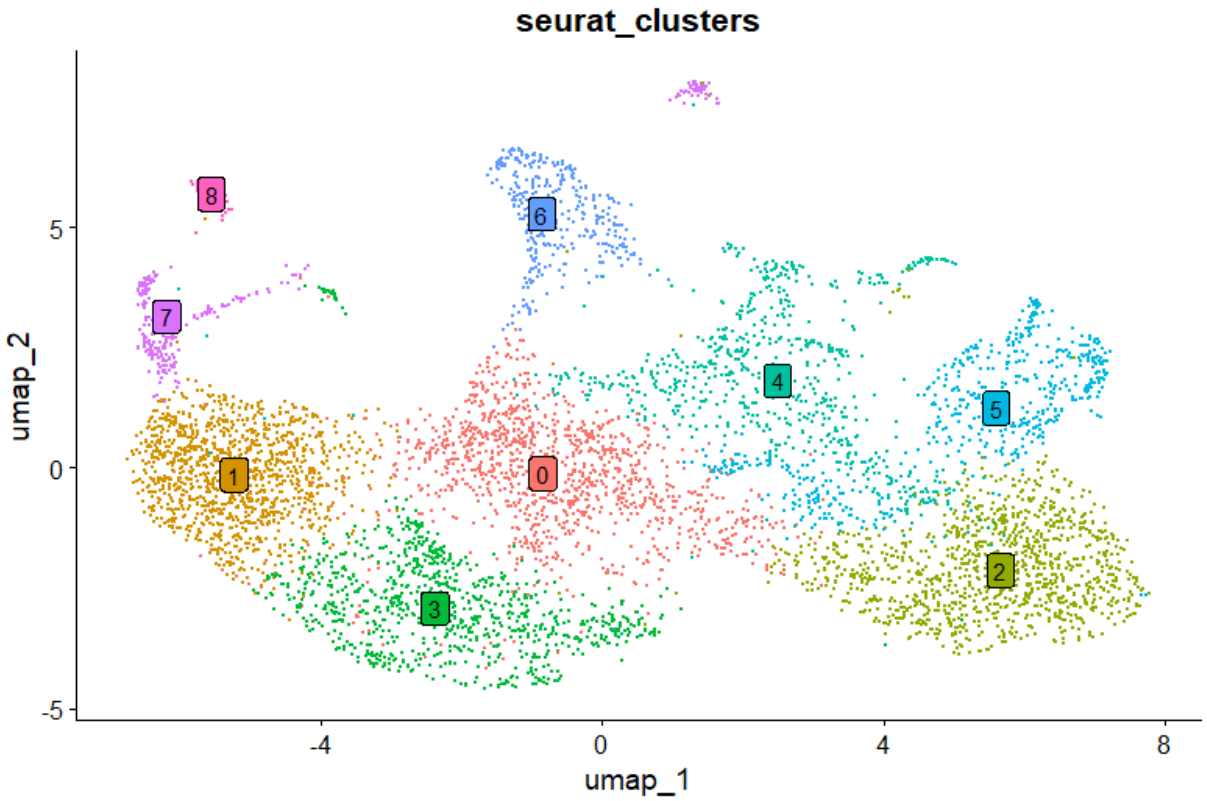

Figure 4: Scatter plot of 9 clusters processed by the UMAP algorithm

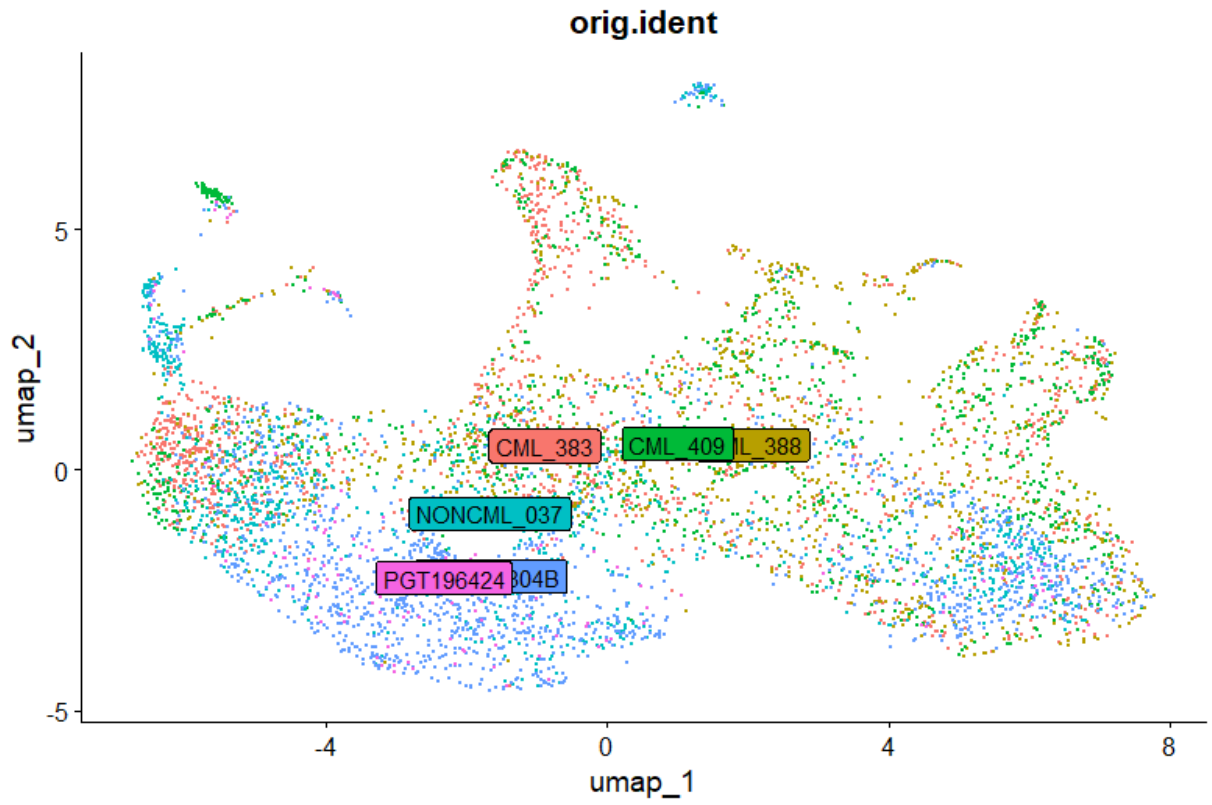

Figure 5: Scatter plot of 6 cell types obtained through annotation

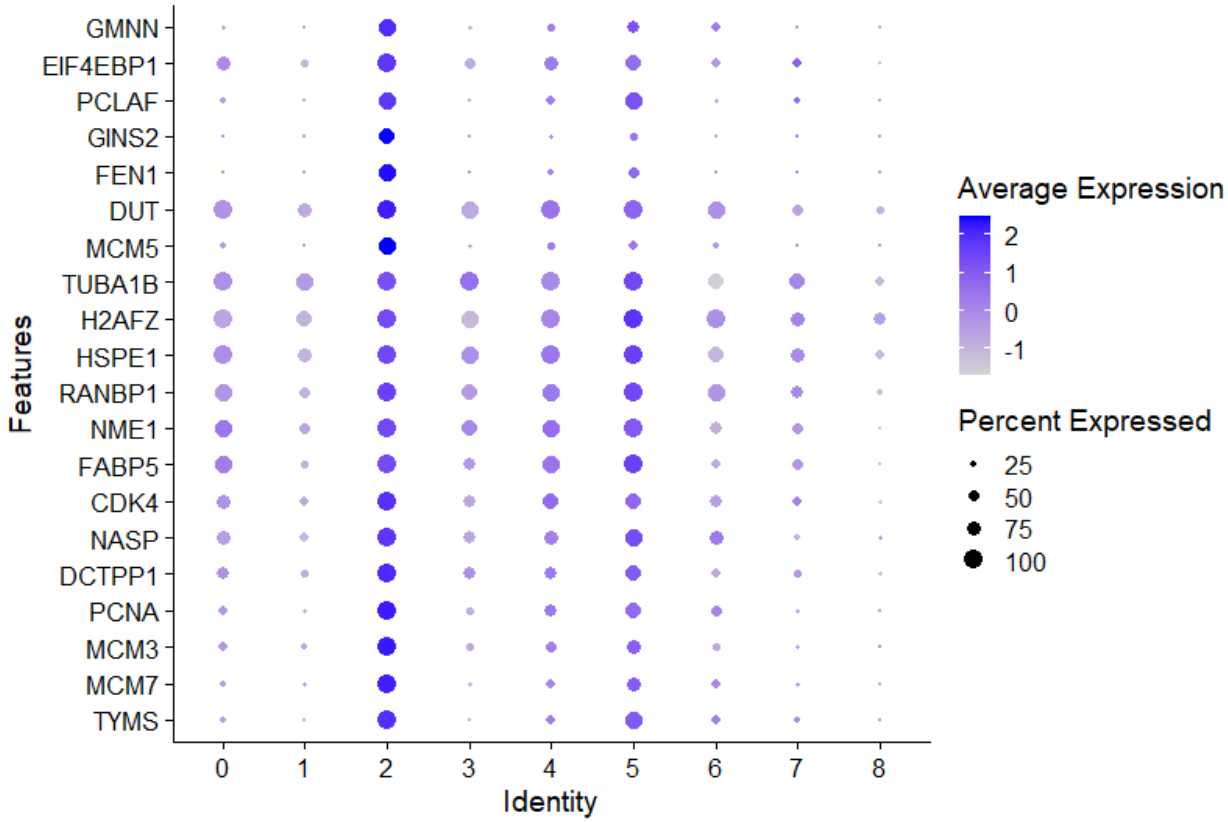

Figure 6: Dot plot of the expression of major marker genes in different clusters

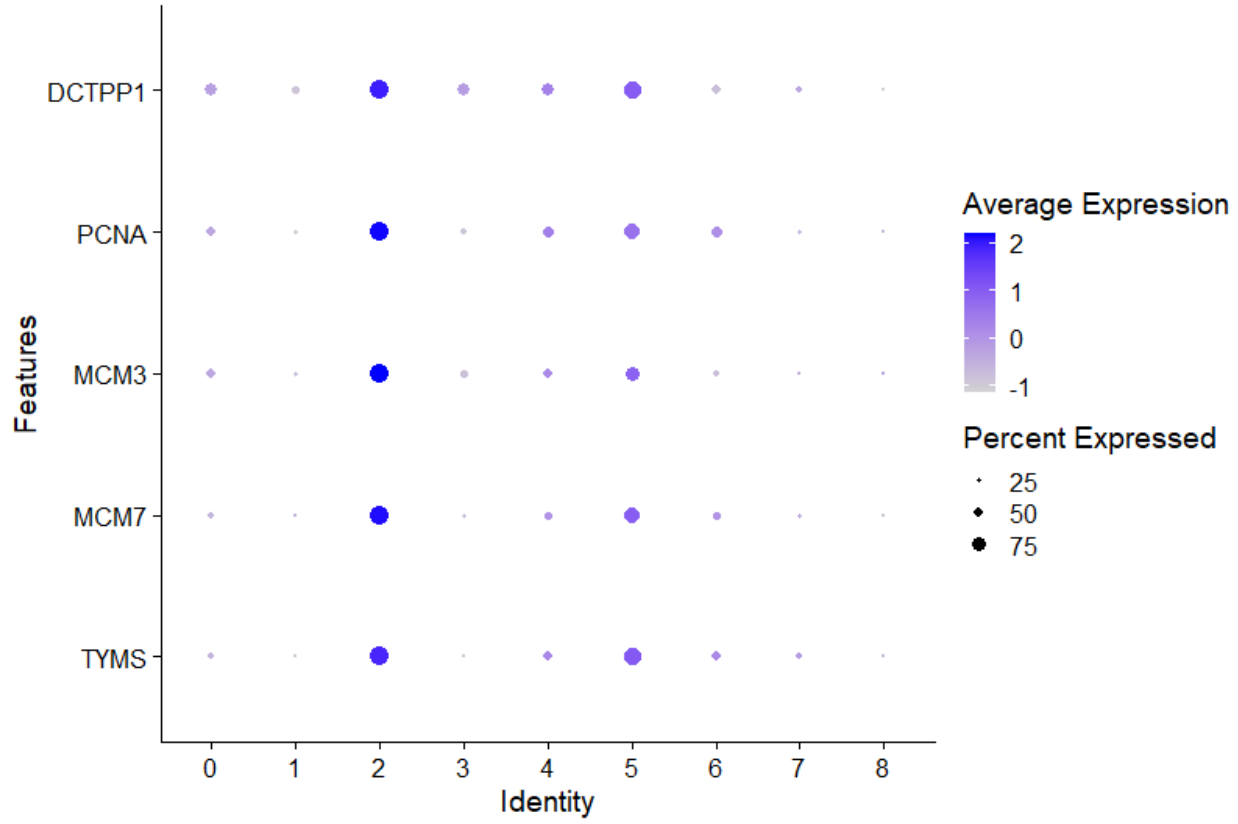

Figure 7: Top 5 differentially expressed marker genes of each cluster

After the preprocessing of scRNA-seq data, including quality control, normalization, and batch effect correction, 18,283 cells from the GSE218184 dataset were included in the analysis (Figure 1). The number of genes detected was significantly correlated with the sequencing depth (R=0.93, (Figure 2). The dimensional reduction plot displayed the batch effect after correction (Figure 3). Then, 18,283 genes were identified and top 1,000 genes were recognized as highly variable genes through variance analysis. Available dimensions were determined through a principal component analysis (PCA), and subsequently, related genes were identified in each principal component (PC). The dot plots showed the expression levels of 30 significantly related genes. (Figures 6). Cell cluster analysis was performed on 20 PCs with a P value <0.05 (Figures S1c and S1d). Afterward, the UMAP algorithm was applied to classify 18,283 cells into 9

clusters (Figure 4). The top 5 differentially expressed marker genes of each cluster were visualized as a dot plot (Figure 7).

Differential Gene Expression

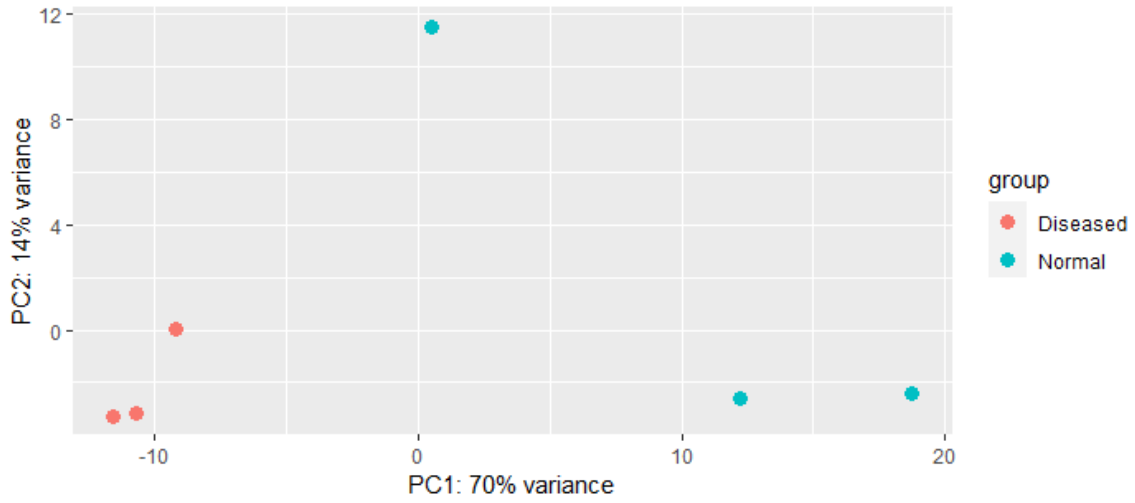

Figure 8: Principal component analysis for Diseased and Normal

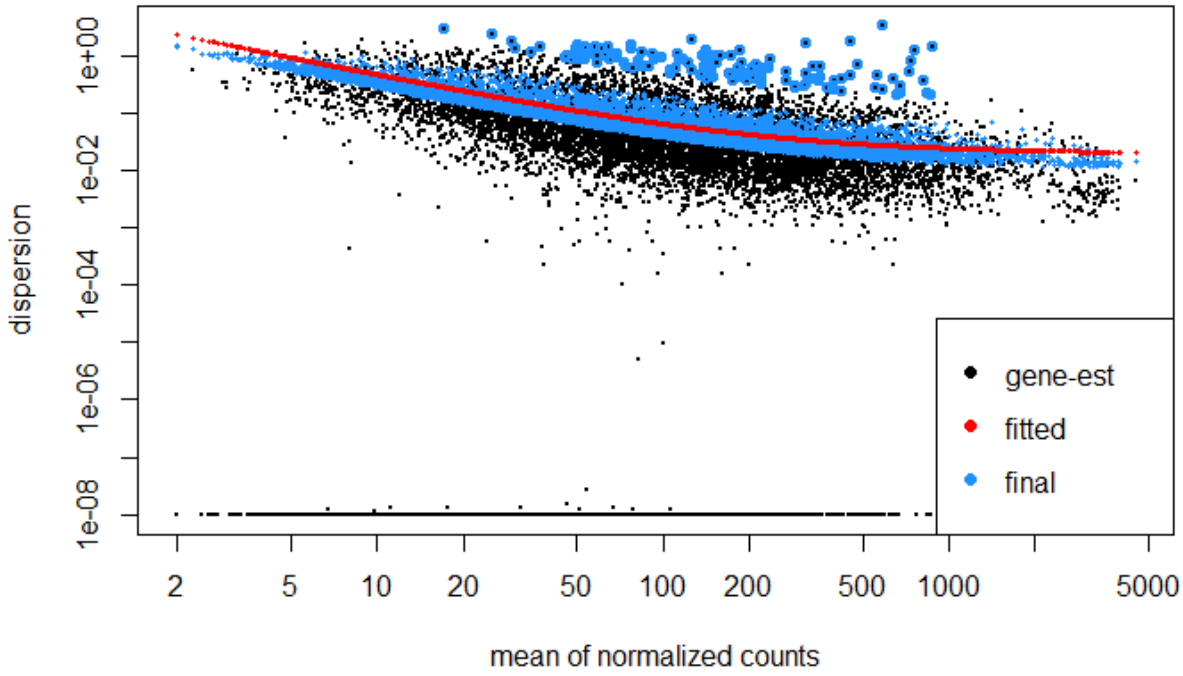

Figure 9: Gene dispersion and normalized mean counts

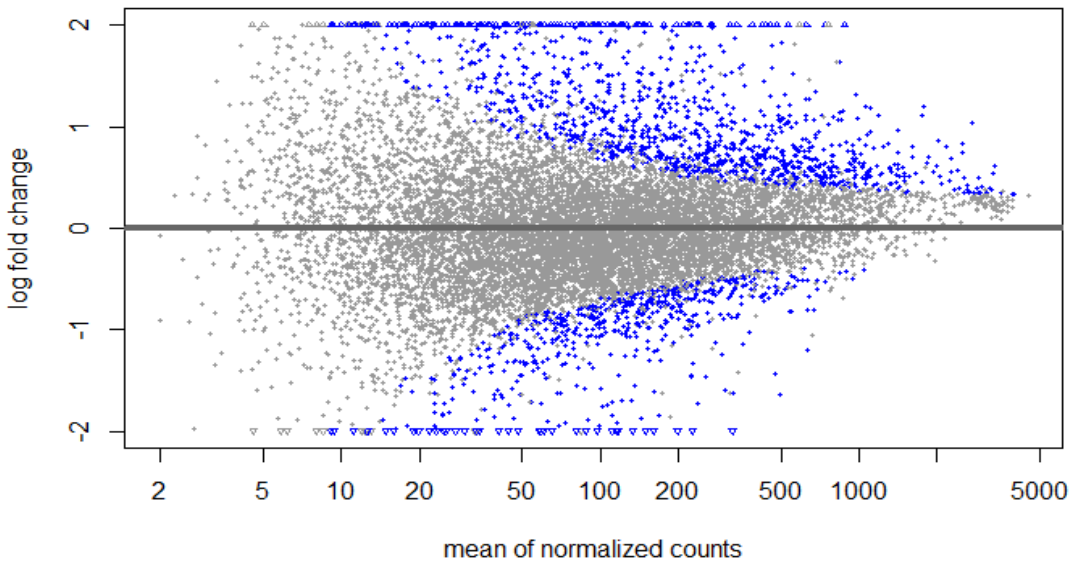

Figure 10: Log2FoldChange vs mean of normalized counts

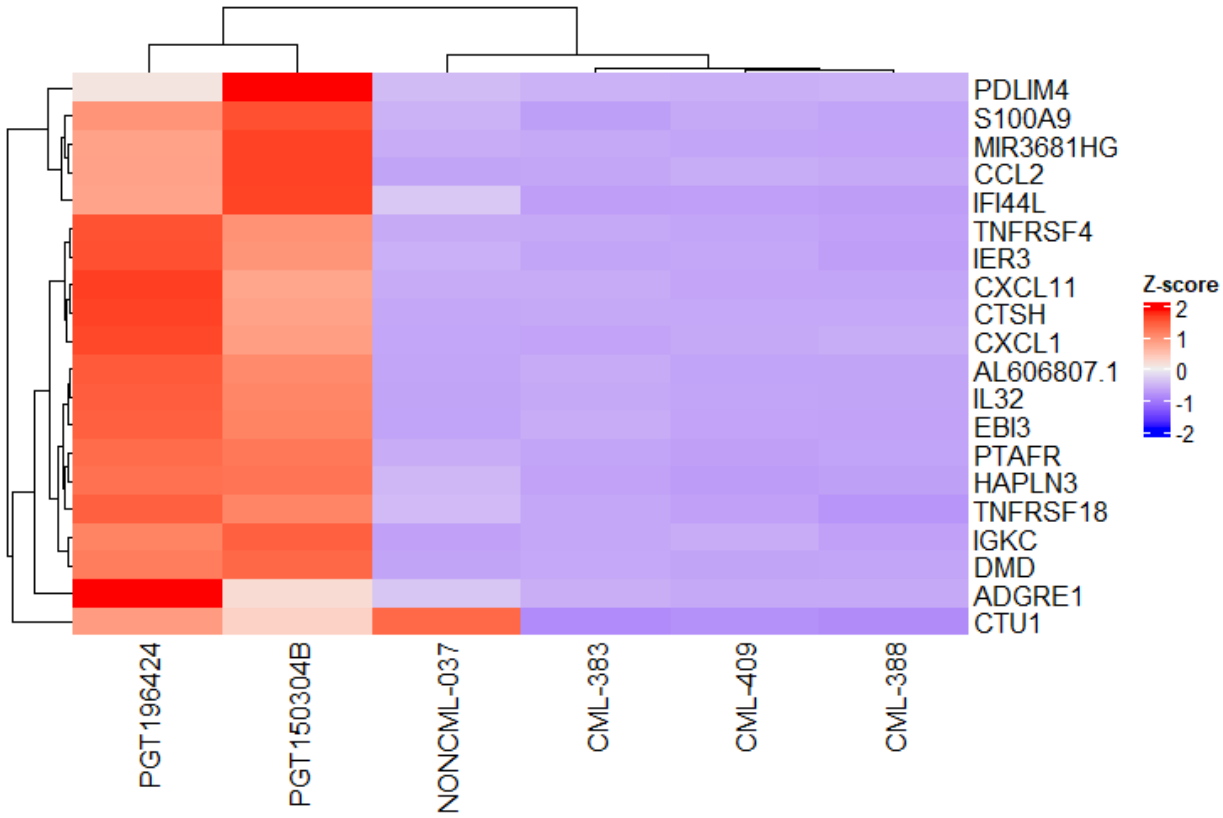

Figure 11: Heatmap of top 20 genes

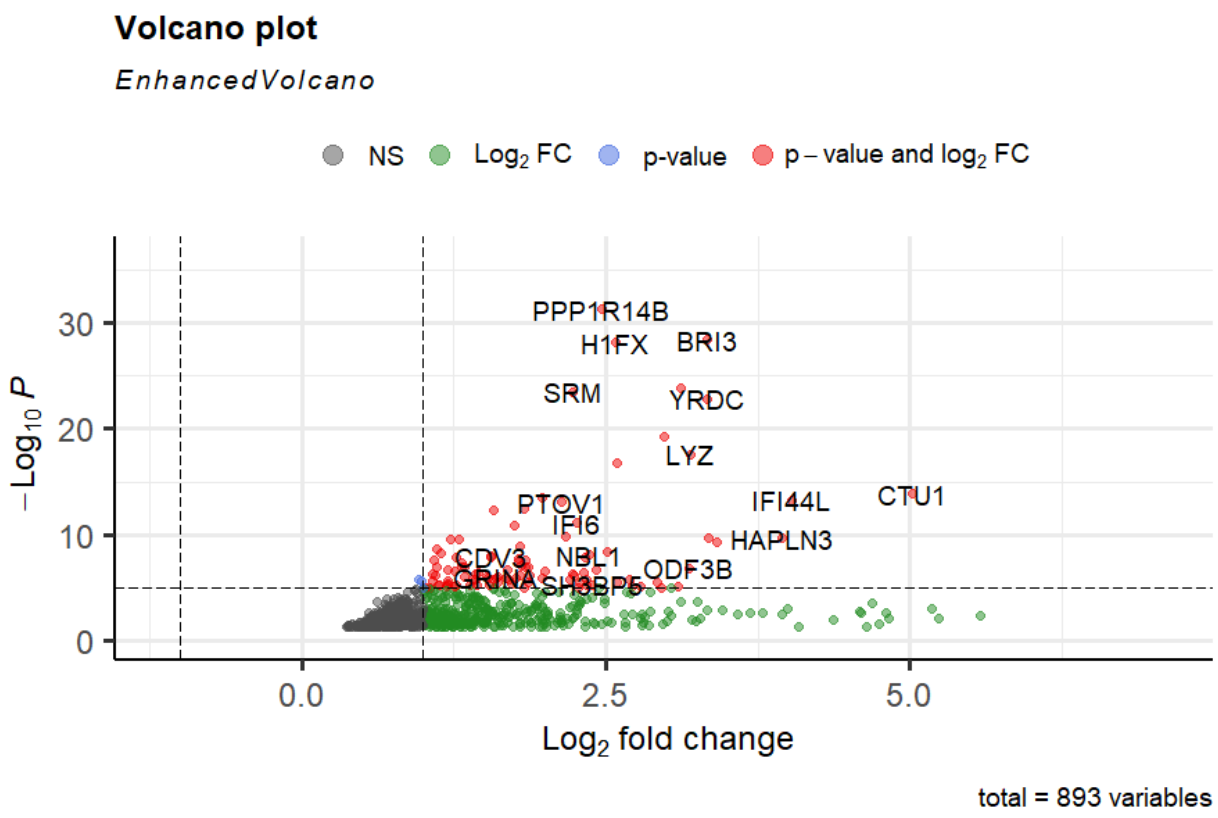

Figure 12: Volcano plot of significant differential expressed genes

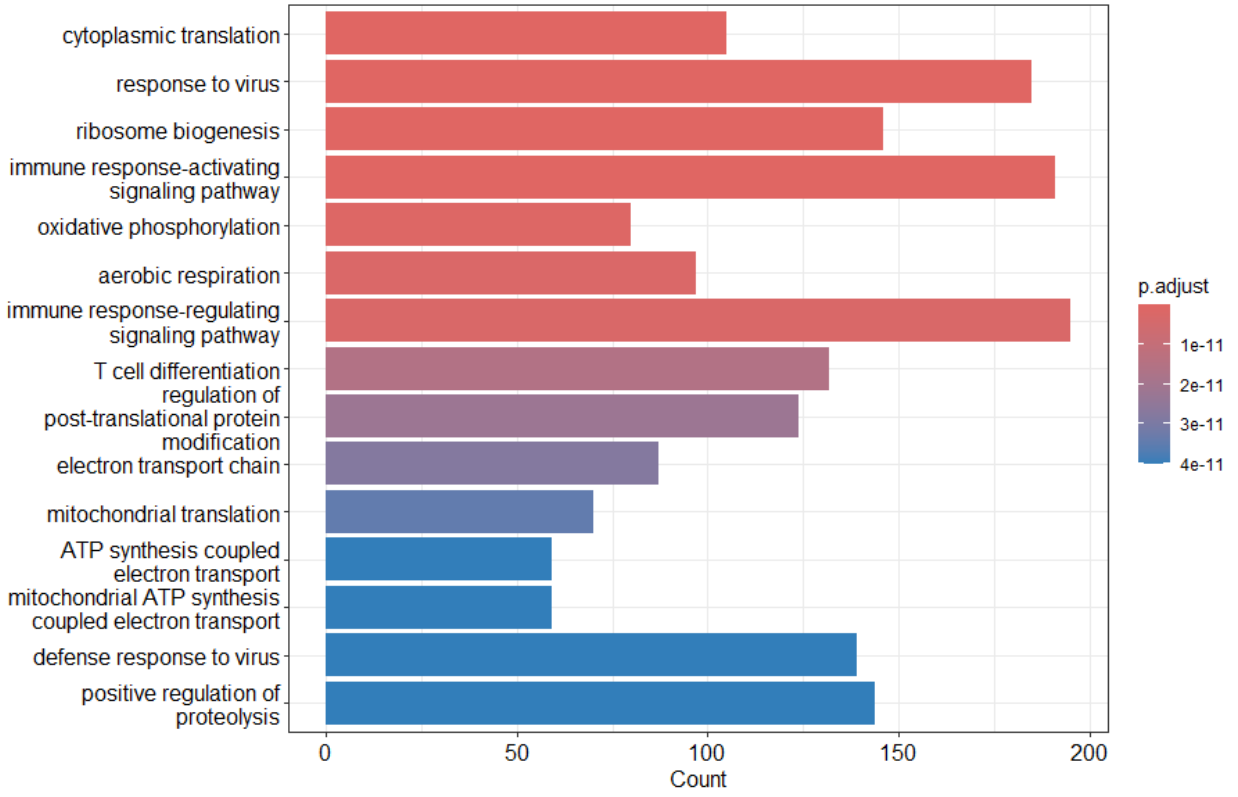

Figure 13: Top 15 GO Ontology biological processes

Figure 8 presents the data's principal components, with PC1 contributing to 70% of the variance and PC2 to 14%. Two distinct groups, represented by colors red and blue, signify diseased (Leukemia stem cells) and normal (Hematopoietic stem cells) cells, respectively. Notably, diseased cells cluster predominantly towards the negative side of the PC1 axis, while normal cells exhibit dispersion along positive values on both axes, particularly along PC1. This spatial separation suggests substantial differences in gene expression or other molecular features between the two groups. The distinct clustering of diseased cells indicates the presence of specific genetic or molecular characteristics associated with leukemia. The top 20 genes identified figure 11 were PDLIM4, S100A9, MIR3681HG, CCL2, IFI44L amongst others.

From Table 2, Several GO terms related to immune responses and cell migration exhibit high enrichment scores and significant p-values. Notably, "response to bacterium," "leukocyte migration," "adaptive immune response," "T cell activation," "inflammatory response," and "chemotaxis" are among the top enriched processes. These findings suggest a strong association between the analyzed gene set and immune-related functions, indicating potential involvement in responses to pathogens and cellular communication. Furthermore, terms associated with cell adhesion, such as "cell-cell adhesion" and "leukocyte cell-cell adhesion," demonstrate notable enrichment. Diseased cells cluster negatively along PC1, indicating distinct molecular features associated with leukemia. Top genes included PDLIM4 and S100A9. Immune-related processes with high enrichment scores, were "response to bacterium" and "leukocyte migration," underscoring potential biomarkers for leukemia. Enriched terms related to cell adhesion suggest a role in mediating crucial interactions for immune responses.

Table 2: Gene Set Enrichment Analysis

ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
GO:0009617	response to bacterium	310	0.529472	2.239956	6.72E-17	1.61E-13	1.26E-13	1978
GO:0050900	leukocyte migration	208	0.587301	2.387103	4.85E-17	1.61E-13	1.26E-13	1900
GO:0002250	adaptive immune response	258	0.555774	2.30964	2.07E-16	3.31E-13	2.58E-13	1474
GO:0042110	T cell activation	361	0.510057	2.191142	3.89E-16	4.66E-13	3.64E-13	1981
GO:0006954	inflammatory response	421	0.485653	2.113316	7.98E-16	7.67E-13	5.98E-13	1707
GO:0006935	chemotaxis	199	0.583928	2.356173	1.20E-15	8.23E-13	6.41E-13	1900
GO:0042330	taxis	199	0.583928	2.356173	1.20E-15	8.23E-13	6.41E-13	1900
GO:0060326	cell chemotaxis	139	0.63248	2.467243	2.27E-15	1.36E-12	1.06E-12	1954
GO:0030595	leukocyte chemotaxis	114	0.671543	2.541833	2.91E-15	1.55E-12	1.21E-12	1868
GO:0098609	cell-cell adhesion	439	0.471788	2.062186	4.53E-15	2.18E-12	1.70E-12	2131
GO:0007159	leukocyte cell-cell adhesion	254	0.529099	2.193427	1.99E-14	7.35E-12	5.73E-12	1792
GO:0071345	cellular response to cytokine stimulus	491	0.455888	2.002875	1.91E-14	7.35E-12	5.73E-12	1755
GO:0097529	myeloid leukocyte migration	116	0.651703	2.480364	1.90E-14	7.35E-12	5.73E-12	1900
GO:0002694	regulation of leukocyte activation	375	0.485322	2.088008	2.47E-14	8.47E-12	6.60E-12	1759
GO:0051249	regulation of lymphocyte activation	324	0.502927	2.140625	2.65E-14	8.48E-12	6.61E-12	1898
GO:0022407	regulation of cell-cell adhesion	281	0.521468	2.185524	4.59E-14	1.38E-11	1.07E-11	2150
GO:0007059	chromosome segregation	349	-0.42082	-2.08923	5.54E-14	1.56E-11	1.22E-11	2512
GO:0031347	regulation of defense response	460	0.453635	1.989411	2.03E-13	5.42E-11	4.23E-11	2131

GO:0002252	immune effector process	360	0.483125	2.073628	2.67E-13	6.74E-11	5.26E-11	1474
GO:0019221	cytokine-mediated signaling pathway	283	0.511239	2.142649	4.11E-13	9.86E-11	7.69E-11	1559
GO:0050778	positive regulation of immune response	445	0.452768	1.981911	7.68E-13	1.76E-10	1.37E-10	1712
GO:0050865	regulation of cell activation	402	0.463999	2.009892	8.49E-13	1.78E-10	1.39E-10	1898
GO:1903037	regulation of leukocyte cell-cell adhesion	231	0.530082	2.180969	8.53E-13	1.78E-10	1.39E-10	2011
GO:0050863	regulation of T cell activation	236	0.530305	2.191015	2.12E-12	4.24E-10	3.30E-10	1792
GO:0032103	positive regulation of response to external stimulus	339	0.480846	2.055819	2.90E-12	5.44E-10	4.24E-10	1993
GO:0051276	chromosome organization	495	-0.36287	-1.87214	2.95E-12	5.44E-10	4.24E-10	3431
GO:0030155	regulation of cell adhesion	441	0.446939	1.9544	5.46E-12	9.72E-10	7.58E-10	2150
GO:0001817	regulation of cytokine production	451	0.437191	1.914435	1.42E-11	2.44E-09	1.90E-09	1848
GO:0002696	positive regulation of leukocyte activation	246	0.508141	2.105403	1.51E-11	2.46E-09	1.92E-09	1519
GO:0050867	positive regulation of cell activation	254	0.496919	2.060021	1.54E-11	2.46E-09	1.92E-09	1519
GO:0019882	antigen processing and presentation	75	0.685951	2.443018	2.41E-11	3.74E-09	2.91E-09	1893
GO:0001816	cytokine production	458	0.433753	1.900985	3.06E-11	4.60E-09	3.59E-09	1848
GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	179	0.54181	2.162046	4.21E-11	6.13E-09	4.78E-09	1707
GO:0022409	positive regulation of cell-cell adhesion	200	0.529024	2.135339	4.41E-11	6.23E-09	4.86E-09	2011
GO:0045785	positive regulation of cell adhesion	283	0.484869	2.032129	5.78E-11	7.94E-09	6.19E-09	2011
GO:0051251	positive regulation of lymphocyte activation	223	0.511022	2.094663	6.03E-11	8.05E-09	6.27E-09	1519
GO:0006959	humoral immune response	67	0.689603	2.418913	6.77E-11	8.79E-09	6.85E-09	1385
GO:0098813	nuclear chromosome segregation	249	-0.42962	-2.07732	9.21E-	1.16E-	9.07E-	2295

					11	08	09	
GO:0097530	granulocyte migration	73	0.676098	2.391037	1.17E-10	1.44E-08	1.13E-08	1900
GO:0046631	alpha-beta T cell activation	117	0.600661	2.282691	1.23E-10	1.47E-08	1.15E-08	1533
GO:0000070	mitotic sister chromatid segregation	166	-0.47689	-2.2052	1.47E-10	1.72E-08	1.34E-08	2551
GO:0000280	nuclear division	302	-0.39093	-1.92714	1.69E-10	1.93E-08	1.50E-08	1958
GO:0001819	positive regulation of cytokine production	290	0.475321	1.997514	1.95E-10	2.18E-08	1.70E-08	1746
GO:0002683	negative regulation of immune system process	292	0.472236	1.982308	2.03E-10	2.22E-08	1.73E-08	2450
GO:0071674	mononuclear cell migration	101	0.618459	2.303222	2.28E-10	2.43E-08	1.90E-08	1519
GO:1903039	positive regulation of leukocyte cell-cell adhesion	183	0.536022	2.146689	2.90E-10	3.03E-08	2.36E-08	1519
GO:0001525	angiogenesis	253	0.489387	2.026682	3.12E-10	3.19E-08	2.49E-08	2131
GO:0051983	regulation of chromosome segregation	115	-0.52619	-2.26927	6.44E-10	6.44E-08	5.03E-08	2295
GO:0031349	positive regulation of defense response	282	0.472532	1.981506	6.74E-10	6.61E-08	5.15E-08	1980
GO:0002253	activation of immune response	331	0.453464	1.932753	8.27E-10	7.94E-08	6.19E-08	1980

REFERENCES

1. Houshmand, M. *et al.* Chronic myeloid leukemia stem cells. **33**, 490–0 (2019).
2. Vetrie, D., Helgason, G. V. & Copland, M. The leukaemia stem cell: similarities, differences and clinical prospects in CML and AML. *Nature Reviews Cancer* **20**, 158–73 (2020).
3. Khemka, R., Gupta, M. & Cml, J. N. CML with megakaryocytic blast crisis: Report of 3 cases. *Pathology & Oncology Research* **25**, 1253–8 (2019).
4. Iqbal, Z. *et al.* Integrated Genomic Analysis Identifies ANKRD36 Gene as a Novel and Common Biomarker of Disease Progression in Chronic Myeloid Leukemia. *Biology* **10**, 11 (2021).
5. Alves, R., Gonzalves, A. C., Rutella, S. & Almeida, A. M. Resistance to tyrosine kinase inhibitors in chronic myeloid leukemia?from molecular mechanisms to clinical relevance. **13**, 19 (2021).
6. Bhatia, R. *et al.* Persistence of malignant hematopoietic progenitors in chronic myelogenous leukemia patients in complete cytogenetic remission following imatinib mesylate treatment. *Blood* **101**, 4701–4707 (2003).
7. Corbin, A. S. *et al.* Human chronic myeloid leukemia stem cells are insensitive to imatinib despite inhibition of BCR-ABL activity. *The Journal of clinical investigation* **121**, 396–409 (2011).
8. Herrmann, H. *et al.* Dipeptidylpeptidase IV (CD26) defines leukemic stem cells (LSC) in chronic myeloid leukemia. *Blood* **123**, 3951–3962 (2014).

9. Järås, M. *et al.* Isolation and killing of candidate chronic myeloid leukemia stem cells by antibody targeting of IL-1 receptor accessory protein. *Proceedings of the National Academy of Sciences* **107**, 16280–16285 (2010).
10. Kinstrie, R. *et al.* Correction: CD93 is expressed on chronic myeloid leukemia stem cells and identifies a quiescent population which persists after tyrosine kinase inhibitor therapy. *Leukemia* **34**, 1975 (2020).
11. Sadovnik, I. *et al.* Identification of CD25 as STAT5-Dependent Growth Regulator of Leukemic Stem Cells in Ph+ CML. *Clinical Cancer Research* **22**, 2051–2061 (2016).
12. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
13. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21**, 12 (2020).
14. Lin, S. M., Du, P., Huber, W. & Kibbe, W. A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Research* **36**, e11 (2008).
15. Lall, S., Sinha, D., Bandyopadhyay, S. & Sengupta, D. Structure-Aware Principal Component Analysis for Single-Cell RNA-seq Data. *Journal of Computational Biology* **25**, 1365–1373 (2018).
16. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37**, 38–44 (2019).
17. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).