Elucidation of B-cell Function in Aplastic Anemia Through Single-cell RNA-Seq of Bone Marrow Tissue

Kevin Liu

Introduction and Dataset Description

The present analysis is inspired by the large-scale single-cell RNA-seq (scRNA-seq) study originally conducted by Tonglin et al. (2022) on the elucidation of B cell destruction of hematopoiesis in aplastic anemia (AA) patients [1]. Using transcriptomic data from scRNA-seq of approximately 20,000 bone marrow cells isolated from samples of both AA cases and non-AA controls, Tonglin et al. identified 8 broad cell clusters, 17 refined cell clusters, their respective marker genes associated with immune function, and elevated expression of B cells that are specific to AA patients, thereby providing insights into hematopoiesis failure related to aberrance of B cells in AA patients. [1]
In this analysis, we will use the scRNA-seq dataset from Tonglin et al. to validate and extend their findings of B cells on bone marrow tissue samples that were collected from two adult AA patients and two non-AA donors. Based on Tonglin et al. (2022), their methodology follows the standard 10X Genomics workflow with mononuclear cell (MNC) and CD34+ cell enrichment, where single-cell isolates were obtained using a nanodroplet protocol [2], MNCs were isolated via Ficoll-Hypaque gradient separation, and CD34+ cell enrichment was achieved with an anti-CD34 microbead kit; the two cell-types were then mixed at a 4:1 ratio and the cell suspension was analyzed using 10X Genomics platform. Finally, cellular barcodes were demultiplexed, reads aligned, and the digital gene expression matrix was generated for downstream analysis. [1]

Methodology

In this study, we follow a standard analytical workflow for the analysis of scRNA-seq datasets generated via 10X Genomics protocols using Seurat [3] under an R environment. The digital gene expression (DGE) matrix and associated metadata was read into R and joined as a Seurat object for analysis.

Quality Control and Batch Correction
Quality control of the DGE matrix was performed to remove low quality cells, where cells with unique molecular identifier (UMI) counts fewer than 200 or greater than 6000 were filtered and cells with greater than 10% of mitochondrial UMIs were also removed. These thresholds were determined using marker genes for platelets (PBPP) and hematopoietic stem cells (HSCs, CDK6) by visualizing each of the cell-types on the two-dimensional projections following dimensionality reduction using uniform manifold approximation and projection (UMAP, see data processing for details) as well as scatter plots of percent mitochondrial UMI and number of UMIs plotted against their respective marker gene expression levels, with the aim to retain the two cell-types in our dataset, thereby preventing the loss of tissue-specific genes and cell-types; the two marker

genes were each selected from lists of 10 marker genes that were provided in the reference cell atlas used in the Azimuth annotation tool [3–5] and literature references for their specificity to bone marrow tissue [6]. After filtering, 9.23% of cells were removed and 21,499 cells remained.
Batch correction was performed using harmony [7] and samples as a variable was empirically determined as a batch variable by observing the amount of mixing between cells in their respective clusters on UMAP plots before and after correction along with considerations for its contribution to the potential of technical variation.

Data Processing
The UMI counts were log-normalized per cell and scaled to transcripts per 10,000. The top 2000 highly variable genes were selected using the 'vst' method via Seurat and the features were then centered and scaled. Dimensionality reduction was performed using principal components analysis (PCA) and the first 26 principal components (PCs) were used for computing the UMAP reductions; the determination of including the first 26 PCs was based on JackStraw analysis and plot through Seurat after observing a significant drop in p-value following the 25^{th} PC since the estimation via elbow plot was ambiguous and estimations of 35 PCs resulted in singletons during clustering.

Clustering and Cell-type Annotation
We clustered the cells into eight cell subsets by first generating the shared nearest-neighbor graph based on the first 26 PCs post-batch correction and the cells were then partitioned into eight clusters using a clustering resolution of 0.01 via the Louvain algorithm; the clustering resolution was determined empirically by iteratively tuning the resolution and plotting the UMAP reductions until platelets and HSCs, as defined using the marker genes described in Quality Control and Batch Correction, were occupying distinct clusters within the plot.
Gene markers for each cluster were then identified based on differential gene expression analysis (DGEA) using Wilcoxon rank-sum tests of genes that display a statistically significant difference in the averaged log2-fold change through a one-against-all comparison strategy by cluster assignment, and only those with positive averaged log2-fold changes and magnitudes greater than 0.5 were considered for a given cluster. The top 10 gene markers were then used for cell-type annotation and are listed in Table 1.
For each of the eight cell-type clusters, known cell-type markers of bone marrow tissue cell-types annotated by Tonglin et al. [1] were used to generate a series of UMAP plots and each plot was visually inspected for consistency and alignment with the labeled clusters; this process allowed us to successfully annotate three of the eight clusters, including the HSC (later generalized to hematopoietic stem and progenitor cell, HSPC), B-cell, and platelet clusters. Subsequently, we utilized a reference-based mapping tool, Azimuth, to generate cell-type annotations using their compiled human bone marrow reference atlas [3–5] and the annotated clusters were exported and added back to our Seurat object in R. Using the annotations from Azimuth, we then completed and refined our previous annotations by overlaying the broad cell-type annotations followed by the detailed cell-type annotations from Azimuth on our clusters and changing our cluster labels accordingly each time. The fully annotated and labeled UMAP plot of cell-type

clusters is shown in Figure 1, which included the T and NK cell cluster, HSPC cluster, monocyte cluster, B-cell cluster, plasma cell cluster, DC cluster, platelets cluster, and stromal cell cluster.

Further Analyses

In our formal analyses, we conducted differential abundance analysis between AA and non-AA cell clusters and differential gene expression analysis (DGEA) with gene set enrichment analysis to validate existing findings regarding the differential abundance of cell-types in AA bone marrow tissue and the role of B-cells in the human immune system.

Differential Abundance Analysis
We conducted differential abundance analysis between AA and non-AA controls for each of the annotated clusters to assess the differences in cell-type abundances between AA and non-AA samples of bone marrow tissue. Group-wise cell-type abundances were calculated and tested for differences in abundances between AA cases and controls (Figure 2). Using Wilcoxon rank-sum tests and the Benjamini-Hochberg method to adjust for multiplicity, we did not identify any statistically significant differences in cell-type abundance using a significance threshold of 0.05; however, we plotted side-by-side bar charts of cell-types between the two groups and show that the AA group had markedly lower proportions of HSPCs and monocytes and higher proportions of plasma cells and T and NK cells relative to the non-AA group (Figure 3). Based on existing literature, we have verified that AA individuals tend to have lower proportions of HSPCs and higher proportions of T and NK cells, which is likely due to abnormal immune cell activation [1].
Based on these findings, we attribute the lack of statistically significant findings to our overall analytical strategy, which is limited by sample size and thus low statistical power through the application of a series of Wilcoxon rank-sum tests. Further analysis would benefit from using alternatives approaches that are more robust to sample size limitations or take on an alternative analytical approach. Nonetheless, our qualitative findings suffice as a validation of existing findings in AA pathophysiology.

*Differential Gene Expression Analysis and Gene Set Enrichment Analysis**
Subsequently, we performed DGEA and gene set enrichment analysis using the B-cell cluster among AA samples to identify AA-associated B-cell gene markers and their associated biomolecular functions.
We first performed DGEA and identified a set of 10 differentially expressed genes within B-cells of AA bone marrow tissue samples (Figure 4), which included genes related to B-cell surface antigen receptors (immunoglobulins), such as CD79A and CD79B [8]; other genes, such as CD74, HLA-DQA1, HLA-DRA, HLA-DQB1, regulates immune cell development [9] and encodes peptides that are associated with major histocompatibility complex (MHC) class II molecules and are also often associated with autoimmune diseases [10]. Therefore, it is expected that our findings from gene set enrichment analysis revealed coinciding evidence of these gene and their involvement in immune

function. Specifically, we found that among the top 5 annotations for the enriched genes, two are molecular functions associated with MHC class II protein complex binding and one biological pathway associated with the production of molecular mediator of immune response; the remaining were two human phenotypes for cough and meningitis, which we did not attempt to interpret. Nonetheless, our findings of B-cell differentially expressed genes used in gene set enrichment analysis validated the role of MHC class II protein complexes and their role in immune function, as they are directly related to the function of B-cells are the major regulators of B-cells during an immune response [11]; this also explains the biological pathway of these genes participating in the production of molecular mediators of immune response.
While our findings from the present study did not contribute any novel findings to the greater field of immunology nor cell biology, we were able to validate several existing findings in literature, which suggests that single cell-based studies can provide a wide variety of information that is also accurate and reproducible.

**Please note that the knitted .html file contains slightly different outcomes regarding get set enrichment analysis; here, we are using the original results obtained on our first attempt, which is consistent with the text in the knitted file.*

References

1. Tonglin H, Yanna Z, Xiaoling Y, Ruilan G, Liming Y. Single-Cell RNA-Seq of Bone Marrow Cells in Aplastic Anemia. *Frontiers Genetics*. 2022;12:745483. doi:10.3389/fgene.2021.745483

2. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/j.cell.2015.05.002

3. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048

4. Oetjen KA, Lindblad KE, Goswami M, et al. Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *Jci Insight*. 2018;3(23):e124928. doi:10.1172/jci.insight.124928

5. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. 2019;37(12):1458-1465. doi:10.1038/s41587-019-0332-7

6. Fagerberg L, Hallström BM, Oksvold P, et al. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics*. *Mol Cell Proteomics*. 2014;13(2):397-406. doi:10.1074/mcp.m113.035600

7. Korsunsky I, Fan J, Slowikowski K, et al. Fast, sensitive, and accurate integration of single cell data with Harmony. *Biorxiv*. Published online 2018:461954. doi:10.1101/461954

8. Huse K, Bai B, Hilden VI, et al. Mechanism of CD79A and CD79B Support for IgM+ B Cell Fitness through B Cell Receptor Surface Expression. *J Immunol*. 2022;209(10):2042-2053. doi:10.4049/jimmunol.2200144

9. Su H, Na N, Zhang X, Zhao Y. The biological function and significance of CD74 in immune diseases. *Inflamm Res*. 2017;66(3):209-216. doi:10.1007/s00011-016-0995-1

10. Rock KL, Reits E, Neefjes J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol*. 2016;37(11):724-737. doi:10.1016/j.it.2016.08.010

11. Katikaneni DS, Jin L. B cell MHC class II signaling: A story of life and death. *Hum Immunol*. 2019;80(1):37-43. doi:10.1016/j.humimm.2018.04.013

Figures and Tables

Table 1. Cell-type clustering and associated top 10 marker genes identified using differential gene expression analysis.

T and NK Cells (Cluster 0)	HSPCs (Cluster 1)	Monocytes (Cluster 2)	B-cells (Cluster 3)	Plasma Cells (Cluster 4)	DCs (Cluster 5)	Platelets (Cluster 6)	Stromal Cells (Cluster 7)
GNLY	HBB	S100A8	CD79A	IGLV2-14	PTGDS	PPBP	CXCL12
IL32	HBA1	S100A9	MS4A1	IGKV3-20	IRF7	TUBB1	TF
NKG7	HBA2	S100A12	CD74	IGHA1	PLD4	PF4	CFD
CCL5	HBD	LYZ	CD79B	IGHG1	IRF8	CAVIN2	LEPR
CD3E	HBG2	DEFA3	HLA-DQA1	IGKC	ITM2C	GNG11	IGFBP5
IFITM1	CA1	CST3	HLA-DRA	IGLV3-25	GZMB	RGS18	APOE
IL7R	HBM	FCER1G	HLA-DQB1	IGKV1-5	LILRA4	CLU	OLFML3
CD3D	AHSP	PRTN3	CD83	IGLV1-51	APP	GP9	IFITM3
GZMA	ALAS2	TYROBP	CD37	IGKV1-39	C12orf75	SPARC	VCAM1
SARAF	SLC25A37	S100A11	TCL1A	IGKV3-15	CCDC50	NRGN	MDK

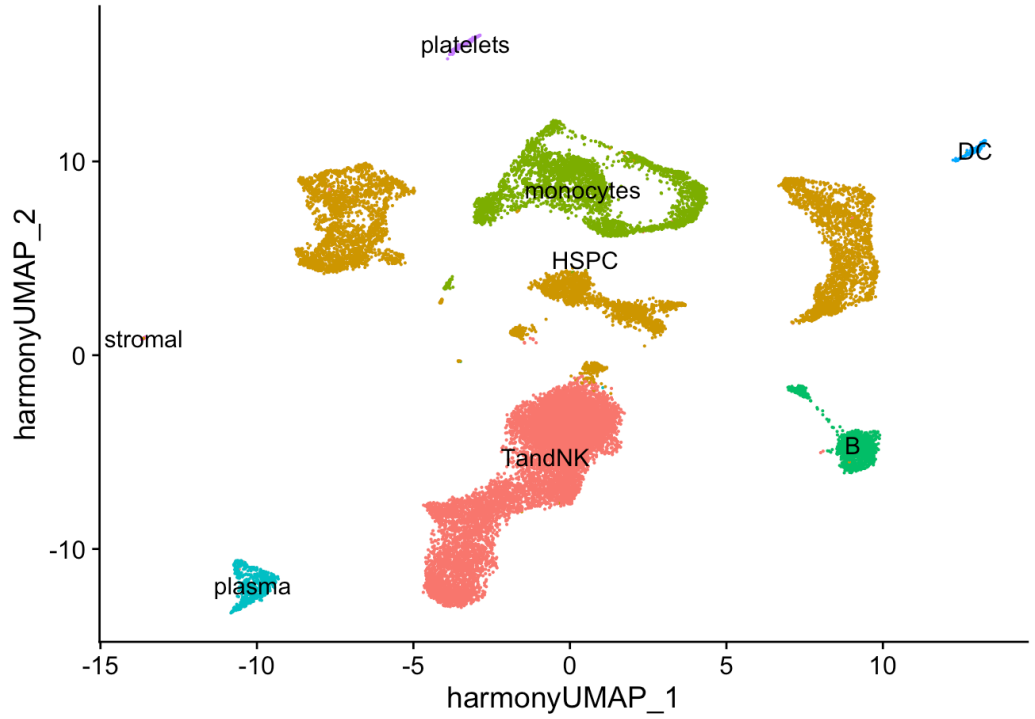

Figure 1. Annotated and labeled cell-type clusters from scRNA-seq of aplastic anemia and normal samples.

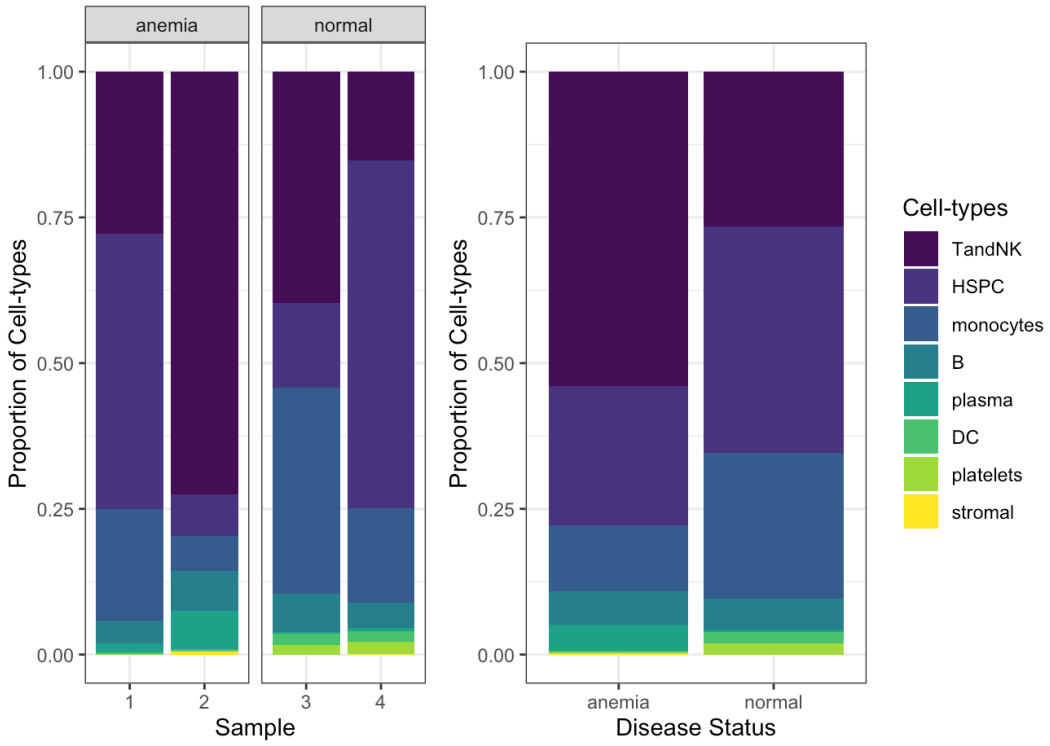

Figure 2. Sample- and group-level proportions of annotated cell-types from scRNA-seq of normal and aplastic anemia bone marrow tissue.

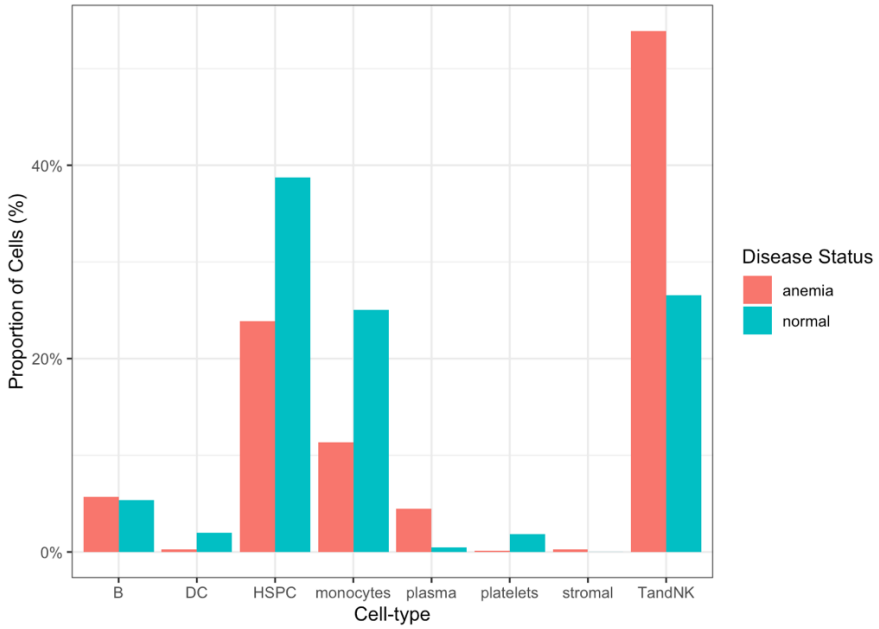

Figure 3. Comparison of annotated cell-type abundances between aplastic anemia and normal samples.

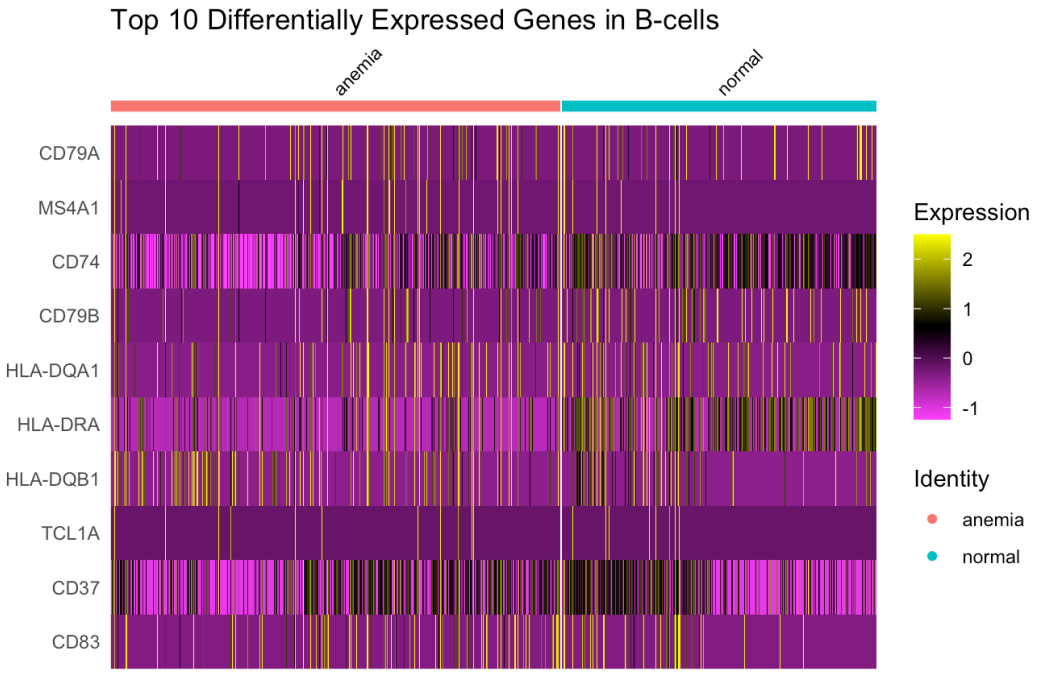

Figure 4. The top 10 differentially expressed genes in the B-cell cluster in bone marrow tissue samples from individuals with aplastic anemia.