BIOENG 290: Final Project

Brontë Kolar

Dataset Selection

This assignment was an opportunity to learn about an area of biology that I know little about – our immune system. Published in May 2022, the Cross-tissue Immune Cell Atlas was created to broaden our limited knowledge of immune cells, which play a crucial role in health and disease (Domínguez Conde et al. 2022). The immune compartment of 16 tissues was sampled from 12 deceased adult donors and single-cell RNA sequencing and VDJ sequencing were conducted to generate a dataset of 360,000 cells. While not covered in my analysis, they also developed CellTypist, a machine-learning tool for rapid cell type annotation, and tested it on their curated dataset. They also have an online site where you can interact with the data via the cellxgene API (atl 2022). I ultimately chose this dataset because I wanted to learn more about immune cells and this published dataset was both very comprehensive (lots of data, well-annotated) and transparent (it was straightforward to understand the processing that happened to produce the matrix that I worked with).

Dataset Background + Preprocessing

The original dataset contains reads from 360k cells and processing all of this data from scratch in Google Colab was not feasible (ENA 2022; EBI 2022). Instead of working with the raw sequencing data, I was able to locate the processed data at their Cross-tissue Immune Cell Atlas website. The global dataset was 5 GB, so I opted to use the processed data from B cell compartment cells (1.55 GB). I downloaded the the raw count data per gene (*B Cell Compartment > Download Raw*), which still included their processed matrix (*B Cell Compartment > Download*) that was already normalized with Scanpy (*scanpy.pp.normalize_per_cell*, scaling factor 10,000). The research team also already removed cells with fewer than 1,000 UMI counts and detected doublets using Scrublet.

Data Format

The file downloaded was an h5ad file that contained an AnnData object with 54,934 cells and 36,601 genes. After exploring the AnnData format documentation, it was interesting to learn that Scanpy is based on AnnData, which allows you to easily access different annotation information about the single cell data matrix. The data matrix (*dataset.X*), represented as a numpy array, stored the processed data per cell, with the cells represented as rows and the genes as columns. The gene names could be accessed via *dataset.var* as a pandas dataframe and the metadata for each cell could be accessed via *dataset.obs* as a pandas dataframe. The raw count data, if I was interested in using, could be accessed via *dataset.layers["counts"]*. For every cell, the following metadata was provided: Organ, Donor, Chemistry, Predicted_labels_CellTypist, Majority_voting_CellTypist, Manually_curated_celltype, Sex, Age_range. I only used the Manually_curated_celltype to assign cell type.

Research Question

The original paper investigated differentially expressed genes between each immune cell grouping, including B cells. They targeted their investigation on marker genes of the identified B cell populations. I wanted to run a similar search, but both corroborate some of the original results and investigate potential marker genes not mentioned in their report.

Analysis

1 - Data Processing

The original dataset was still too large to process in Colab, so I randomly sampled 15,000 cells out of the original 55k. The selected cells are saved each time in a text file that can be loaded. The instance of randomly sampled cells that I used for my analysis is submitted with the code.

2 - Filtering

Lowly expressed genes were first filtered from the gene expression matrix during pre-processing. I set my cut-off to genes that were present in at least 10 cells, which when plotted, kept the bulk of the data while removing genes with low expression. Following the removal of these genes, the number of genes dropped to 19,205. In most scRNA experiments, dropout occurs and many lowly expressed genes are not detected, even if they are expressed. These genes must be removed, as there is not sufficient information about these genes to make any strong inferences.

3 - Visualization with PCA and UMAP

The expression matrix was visualized using both PCA, out of personal interest, and UMAP, as the original paper produced UMAP graphs that could be used as a point of reference. While cell-type labels were manually curated, I wanted to further investigate which variables accounted for the most variation in the data.

PCA

PCA analysis was conducted using 50 principal components, as from experience in the past labs, this worked well on a similar sized dataset. The data was plotted on the first four principal components and coloured by cell-type and chemistry, which are visualized in Figure 1 and Figure 2, respectively. PCs are ranked by how much they describe the data. The first two principal components cluster cell type fairly distinctly, while it appeared that the second and the third cluster by chemistry. The format of the Supplementary Materials did not allow to search for keywords, but I assume chemistry is referring to different Chromium Single Cell 3′ Solutions. Different methods appear to have been used on different cell types, but I am unsure of whether this chemistry choice was based on tissue type or based on what different labs had available to use on different donors.

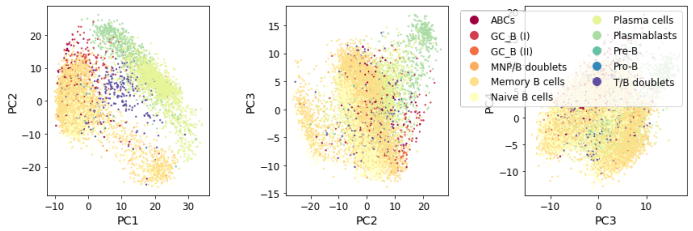


Figure 1: PCA Plot - First four components, labelled by manually curated cell type

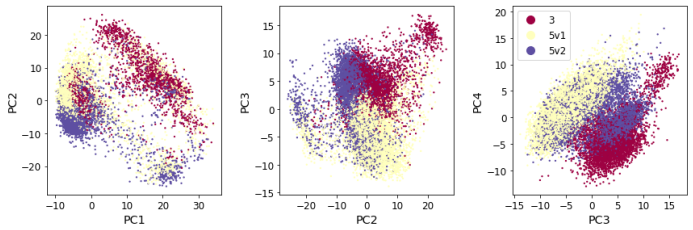


Figure 2: PCA Plot - First four components, labelled by Chemistry

UMAP

The UMAP analysis conducted in the previous labs did not appear to group by cell type, but after looking into the code from the paper, I used the scanpy.pl.umap function, which utilizes AnnData directly. The UMAP plot, show in Figure 3, clustered very strongly by cell-type. It was clear to see that the majority of the cell data is from Naive B cells and Memory B cells.

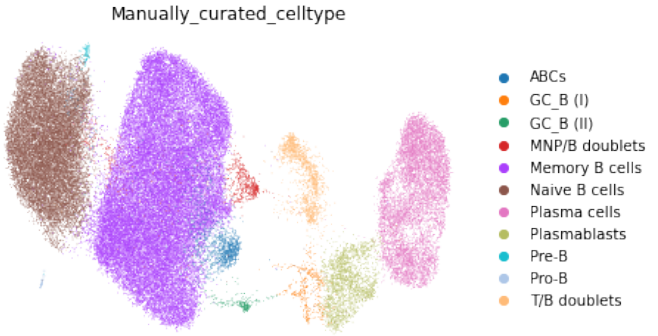


Figure 3: UMAP Plot - Coloured by cell type (*Manually_curated_celltype* metadata variable)

4 - Clustering

Clusters were plotted using both KMeans and Louvain. To evaluate the strength of a clustering for KMeans, prediction strength was used as a metric. After plotting an elbow curve for prediction strength, 6 clusters were selected, as this was the largest amount of clusters that still had a strong prediction strength. The clustering results are visualised in Figure 4. One important limitation to note is that the clusters agreed very little. Using the *adjusted rand score* as a metric to quantify how similar the clustering results were, KMeans and Louvain had a joint similarity score of 0.25. This could be as a result of the selection of 6 clusters for the Kmeans algorithm, but in a previous lab it was also found that these two methods did not agree.

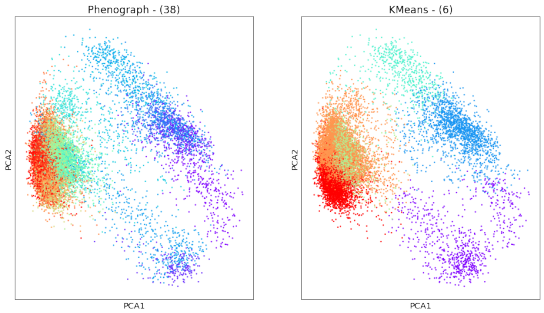


Figure 4: Clustering algorithm results plotted on PCA1 and PCA2, coloured by cluster. Left: Louvain (Phenograph package) method. Right: KMeans method.

5 - Differentially Expressed Genes

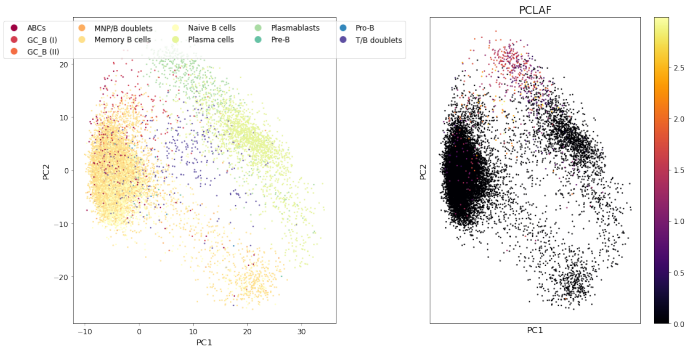
To investigate the differential expression of genes across different B Cells, a dotplot was generated. In the original paper, a dotplot was generated based on set of pre-selected marker genes from the identified B cell populations. Instead, I generated a dotplot that was based on the most differentially expressed genes for each B cell type. The top 8 differentially expressed genes for each lineage were selected and, after removing duplicates, 71 differentially expressed genes remained.

Differential expression between B cellntypes was calculated using the mean-difference and the rank-sum statistics. This was done by making use of the *scprep.stats.differential_expression_by_cluster function*. The rest of this section outlines some interesting findings when looking further into three genes of interest, two of which were not mentioned in the original paper: *Pclaf*, *Sdc1*, and *Ybx3*.

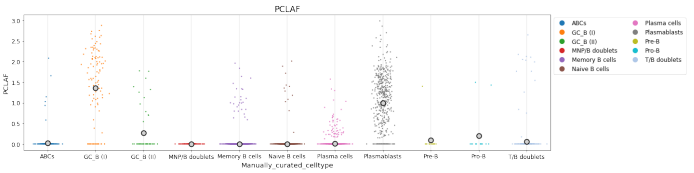
Figure 5: Dot plot for expression of top 8 differentially expressed genes from each identified B cell populations. Color represents maximum-normalized mean expression of cells expressing marker genes and size represents the percentage of cells expressing these genes.

Pclaf - plasmablasts, GC_B(I)

From the DE results, *Pclaf* was a gene that was highly expressed in both plasmablast and germinal center (GC-I, marked by expression of MK167) B cell samples and very little in other B cell types, which can be seen in Figure 6. *Pclaf* enables chromatin binding activity and is involved in several cellular processes. In a different study in mice, *Pclaf* has been noted as a proliferation marker of germinal center-like (GC-like) B cells (Nguyen et al. 2021). Also, *Pclaf* is over-expressed in human cancers and is a regulator of tumor progression, and has been recommended as a therapeutic target in neuroblastoma citepliu2022pclaf. Another paper also found that the expression of *Pclaf* was found to be positively correlated with activated CD4 T cells and type 2 T helper cells in cancer patients, which suggests that *Pclaf* may play a role in tumor immune infiltration (Liu et al. 2022). Upon revisiting the Supplementary Materials, the cause of death for all tissue donors was either intracranial haemorrhage or hypoxic brain injury and all donors were reported

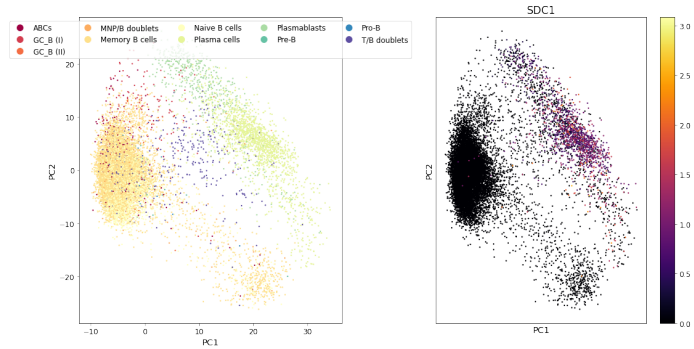


(a) Left: B cells plotted on PC1 and PC2, coloured by cell type. Right: B cells plotted on PC1 and PC2, coloured by *Pclaf* expression.

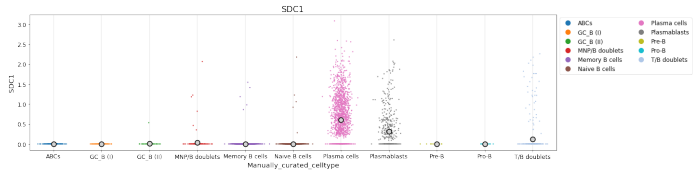


(b) *Pclaf* expression expression plotted by B cell type

Figure 6: Over-expression of *Pclaf* in plasmablasts and GC-I B cells, compared to other B cell types



(a) Left: B cells plotted on PC1 and PC2, coloured by cell type. Right: B cells plotted on PC1 and PC2, coloured by *Sdc1* expression.



(b) *Sdc1* expression expression plotted by B cell type

Figure 7: Over-expression of *Sdc1* in plasma cells, compared to other B cell types

to be free of cancer. While it is still unclear to me what specific function *Pclaf* carries out disproportionately in plasmablasts and GC-I cells, it is interesting to see the very recent literature has identified this gene as a marker for GC-like B-cells.

Sdc1 - plasma cells

The original paper mentioned that two populations of plasmablasts and plasma cells are marked by expression of *CD38*, *XBP1* and *Sdc1*. *Sdc1* was very prominently over-expressed in plasma cells, as seen in Figure 7. Syndecan 1 is a member of the syndecan family, which mediate cell binding, cell signaling, and cytoskeletal organization. It has also been previously reported as a marker for plasma cells (Rawstron 2006). Upon further research, plasma cells and less mature plasmablasts are classified as antibody secreting cells (ASCs), which produce high quantities of antibodies. These cells have the highest surface expression of *Sdc1*, which is also named in literature as CD138 (McCarron et al. 2017).

The function of *Sdc1* was unknown until 2017, when it was found that its surface expression gives a survival advantage by increasing IL-6 and APRIL signaling, which prevents cell death of plasma cells (McCarron et al. 2017). It has been known for a long time that IL-6 is a key factor in plasma cell survival and *Sdc1* is central to the mechanism by which IL-6 and APRIL are able to exert their survival-promoting function.

YBX3 - Pre-B and Pro-B

In this analysis, it was found that *Ybx3* was overexpressed in both Pre-B and Pro-B cell types, as depicted in Figure 8. This gene was also expressed largely in other cell types, like Naive B cells and Memory cells. For context, Pre-B cells are bone marrow lymphoid cells that lack surface immunoglobulin but contain intracytoplasmic IgM heavy chains (Pearl 1983). Pro-B cells become Pre-B cells after they assemble a Pre-B cell receptor complex (in addition to a few other steps), which signals their differentiation into Pre-B cells (Lutz et al. 2011).

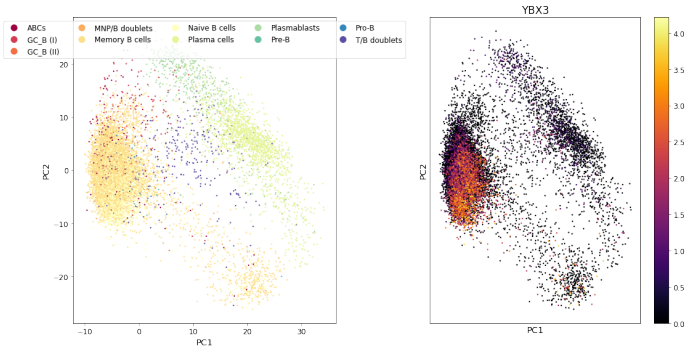
Ybx3 has a number of important functions within cells. It binds and regulates distinct sets of RNAs, stabilizes certain amino acid transporter transcripts for translation, and is a regulator of large neutral amino acid homeostasis (Cooke et al. 2019). In a recent study that researched B cell development and transformation, it was also found that *Ybx3* is uniquely upregulated in Pre-B cells. This paper identified the RNA-binding protein *Ybx3* as a marker of Pre-B cell differentiation (Lee et al. 2021). Given that Pro-B cells differentiate into Pre-B cells, the overexpression of *Ybx3* in both cell types makes sense and shows that *Ybx3* overexpression is also a marker for Pro-B cells.

Conclusion

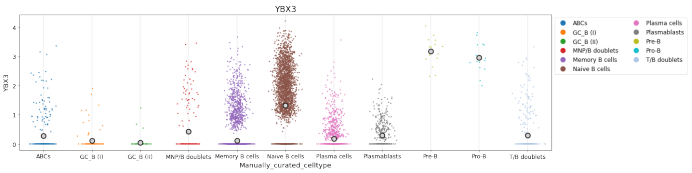
While *Ybx3* and *Pclaf* were not identified as marker genes for identified B cell type populations in the original paper, in this report I was able to identify them from the gene expression data and find previous literature to support their cell-type specific expression. Furthermore, it remains unclear what the function of *Pclaf* could be in the plasmablast cell types and it appears that research on this gene is ongoing.

A main limitation of this analysis is the sampling of 15,000 cells from the original dataset. A large amount of B cells were not utilized and it was not investigated further how the distribution of sampled cells across cell types compares to the distribution of the original set. Some cell populations were very sparse compared to others and it would be interesting to see if these expression results translate to the larger dataset. Given more time, it would be interesting to compare the expression between B cells and other immune cell populations, such as cells from the myeloid compartment or T innate lymphoid cells. This could have been done by making use of two down-sampled pre-processed matrices from the published Cross-tissue Immune Cell Atlas.

Another limitation of this work is that cell location by organ was not investigated. Further patterns or inferences could potentially be drawn by looking at cell-type specific expression combined with tissue-specific expression.



(a) Left: B cells plotted on PC1 and PC2, coloured by cell type. Right: B cells plotted on PC1 and PC2, coloured by *Ybx3* expression.



(b) *Ybx3* expression expression plotted by B cell type

Figure 8: Over-expression of *Ybx3* in Pre-B and Pro-B, compared to other B cell types

Literature cited

(2022). Cross-tissue immune cell atlas. https://www.tissueimmunecellatlas.org/.

(2022). Ebi. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/, Accession: E-MTAB-11536.

(2022). European nucleotide archive. https://www.ebi.ac.uk/ena/browser/view/ERR9249339-ERR9249748.

Cooke, A., Schwarzl, T., Huppertz, I., Kramer, G., Mantas, P., Alleaume, A.-M., Huber, W., Krijgsveld, J., and Hentze, M. W. (2019). The rna-binding protein ybx3 controls amino acid levels by regulating slc mrna abundance. *Cell reports*, 27(11):3097–3106.

Domínguez Conde, C., Xu, C., Jarvis, L., Rainbow, D., Wells, S., Gomes, T., Howlett, S., Suchanek, O., Polanski, K., King, H., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Lee, R. D., Munro, S. A., Knutson, T. P., LaRue, R. S., Heltemes-Harris, L. M., and Farrar, M. A. (2021). Single-cell analysis identifies dynamic gene expression networks that govern b cell development and transformation. *Nature communications*, 12(1):1–16.

Liu, X., Cheng, C., Cai, Y., Gu, Y., Wu, Y., Chen, K., and Wu, Z. (2022). Pan-cancer analyses reveal the regulation and clinical outcome association of pclaf in human tumors. *International journal of oncology*, 60(6):1–13.

Lutz, J., Heideman, M. R., Roth, E., van den Berk, P., Müller, W., Raman, C., Wabl, M., Jacobs, H., and Jäck, H.-M. (2011). Pro-b cells sense productive immunoglobulin heavy chain rearrangement irrespective of polypeptide production. *Proceedings of the National Academy of Sciences*, 108(26):10644–10649.

McCarron, M. J., Park, P. W., and Fooksman, D. R. (2017). Cd138 mediates selection of mature plasma cells by regulating their survival. *Blood, The Journal of the American Society of Hematology*, 129(20):2749–2759.

Nguyen, H. T. T., Guevarra, R. B., Magez, S., and Radwanska, M. (2021). Single-cell transcriptome profiling and the use of aid deficient mice reveal that b cell activation combined with antibody class switch recombination and somatic hypermutation do not benefit the control of experimental trypanosomosis. *PLoS pathogens*, 17(11):e1010026.

Pearl, E. R. (1983). Pre-b-cells in normal human bone marrow and in bone marrow from patients with leukemia in remission: persistent quantitative differences and possible expression of cell surface igm in vitro.

Rawstron, A. C. (2006). Immunophenotyping of plasma cells. *Current protocols in cytometry*, 36(1):6–23.