Topic: Single-cell RNA-seq analysis reveals immune cell heterogeneity in human peripheral blood mononuclear cells.

Introduction

A major objective in modern molecular biology is to understand how complex biological systems are organized at the cellular level. Many tissues are composed of multiple distinct cell types that differ in both function and gene expression. Initially, gene expression studies relied on bulk RNA sequencing (RNA-seq), which measures messenger RNA abundance across an entire sample. However, bulk RNA-seq produces an average signal derived from thousands or millions of cells. I recognized that these averaging masks biologically important variability and prevents the identification of rare cell populations or transient cellular states (Kiselev et al., 2019). Differences between individual cells are therefore lost, and the resulting expression profile reflects only the dominant cell types within the sample.

Single-cell RNA sequencing (scRNA-seq) addresses these limitations by measuring gene expression in individual cells (Tang et al., 2009). Recent advances in droplet-based sequencing technologies enable thousands of cells to be isolated, barcoded, and sequenced in parallel (Macosko et al., 2015; Zheng et al., 2017). Each cell is captured separately and assigned a unique molecular identifier, allowing its transcriptome to be reconstructed computationally. By analysing gene expression at cellular resolution, I can identify distinct cell populations, detect rare cell types, and study functional cellular states that would otherwise remain hidden in bulk measurements. Because cells with similar transcriptional profiles cluster together in high-dimensional gene expression space, computational approaches such as dimensionality reduction and graph-based clustering can be used to reconstruct tissue composition directly from sequencing data (Wolf et al., 2018).

For this project I focused on peripheral blood mononuclear cells (PBMCs), a heterogeneous population of immune cells circulating in human blood. PBMCs include several well-characterized cell types, such as T lymphocytes, B lymphocytes, natural killer (NK) cells, monocytes, and dendritic cells. Each population plays a distinct role in the immune response: T cells regulate adaptive immunity, B cells produce antibodies, NK cells mediate cytotoxic activity against infected or malignant cells, and monocytes and dendritic cells participate in antigen presentation and inflammatory signalling (Villani et al., 2017). Because these populations are biologically well understood and associated with specific marker genes, PBMCs represent a suitable model system for validating computational approaches in single-cell bioinformatics.

Beyond serving as a benchmark dataset, PBMCs are also clinically relevant. Alterations in immune cell composition and transcriptional activity are associated with infections,

autoimmune diseases, inflammatory conditions, and cancer. Analysing PBMCs at single-cell resolution therefore provides insight into immune activation and dysregulation. In this study, I analysed a publicly available PBMC single-cell RNA-seq dataset using a computational pipeline implemented in Python with the Scanpy framework (Wolf et al., 2018). My aim was to identify transcriptionally distinct immune cell populations, determine their marker genes, and demonstrate how single-cell transcriptomic analysis can reconstruct cellular composition from high-dimensional gene expression data.

Methods

Computational environment and software

All analyses were performed on a personal computer using Python (version 3.10) within a Conda virtual environment. Single-cell transcriptomic analysis was conducted using the Scanpy framework, which is specifically designed for large-scale single-cell gene expression analysis (Wolf et al., 2018). Gene expression matrices were stored and processed using the AnnData data structure, which allows efficient handling of high-dimensional single-cell datasets together with associated cell and gene metadata. Clustering was performed using the Leiden community detection algorithm, and differential gene expression analysis was carried out using the Wilcoxon rank-sum statistical test implemented in Scanpy.

Dataset acquisition

A publicly available peripheral blood mononuclear cell dataset was obtained using the built-in Scanpy function sc.datasets.pbmc3k(). This dataset contains gene expression measurements from approximately 2,700 human peripheral blood mononuclear cells generated using droplet-based single-cell RNA sequencing technology. The dataset consists of a gene-by-cell count matrix representing the number of detected transcripts per gene in each individual cell.

Quality control and filtering

Quality control filtering was applied to remove low-quality cells and technical artefacts prior to downstream analysis. First, mitochondrial genes were identified based on gene names beginning with the prefix “MT-”. The percentage of mitochondrial RNA counts per cell was calculated as a measure of cellular stress or apoptosis.

Cells with fewer than 200 detected genes were removed because they likely represented empty droplets or debris. Cells with more than 2,500 detected genes were excluded to reduce potential doublets, in which two cells are captured in the same droplet.

Additionally, cells with more than 5% mitochondrial RNA content were filtered out, as high mitochondrial expression is commonly associated with damaged or dying cells.

Normalization and transformation

To make gene expression values comparable across cells, library size normalization was performed using total count normalization to a target sum of 10,000 counts per cell. This step corrects for differences in sequencing depth between cells. Following normalization, gene expression values were log-transformed using a natural logarithm transformation (log1p), which stabilizes variance and improves the suitability of the data for downstream statistical analyses.

The full normalized dataset was stored as raw expression values (adata.raw) to allow visualization of marker genes that were not included among the highly variable genes.

Highly variable gene selection

Not all genes contribute equally to cell type identification. Therefore, highly variable genes were identified using Scanpy's dispersion-based method. Approximately 2,000 genes with the highest variability across cells were selected for downstream analysis. This step reduces noise and focuses the analysis on genes most informative for distinguishing cell populations.

Dimensionality reduction

Principal component analysis (PCA) was performed on the scaled expression matrix of highly variable genes to reduce dimensionality while preserving biological variation. The first 10 principal components were selected based on the variance explained and used for further analysis. Scaling was applied to ensure that highly expressed genes did not dominate the principal components.

Neighbourhood graph construction and UMAP visualization

A k-nearest neighbour graph was constructed using the first 10 principal components to model similarity relationships between cells. Uniform Manifold Approximation and Projection (UMAP) was then applied to embed the cells into a two-dimensional space for visualization. UMAP preserves local neighbourhood relationships and allows visualization of transcriptionally similar cells forming clusters in low-dimensional space.

Clustering

Unsupervised clustering was performed using the Leiden graph-based clustering algorithm with a resolution parameter of 0.5. This method identifies communities of cells with similar transcriptional profiles within the neighbourhood graph. Each cluster was interpreted as a putative cell population.

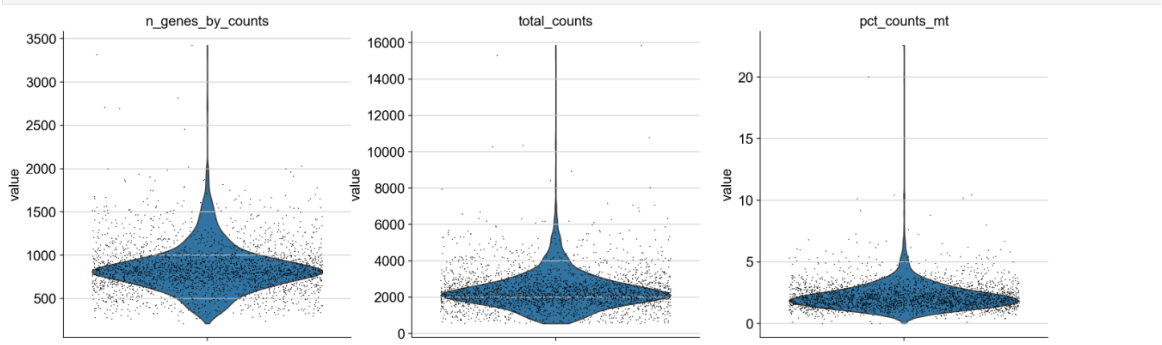
Differential expression and cell type annotation

Marker genes for each cluster were identified using the Wilcoxon rank-sum test, which compares gene expression between one cluster and all remaining cells. The top differentially expressed genes were used to assign biological identities to clusters based on known immune cell marker genes. Cell types were annotated by comparing cluster-specific gene expression with established markers for T cells, B cells, natural killer cells, monocytes, dendritic cells, and platelets.

Results

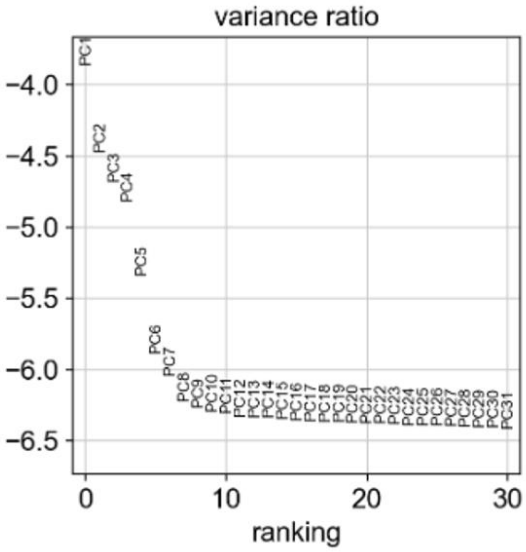
Quality control assessment

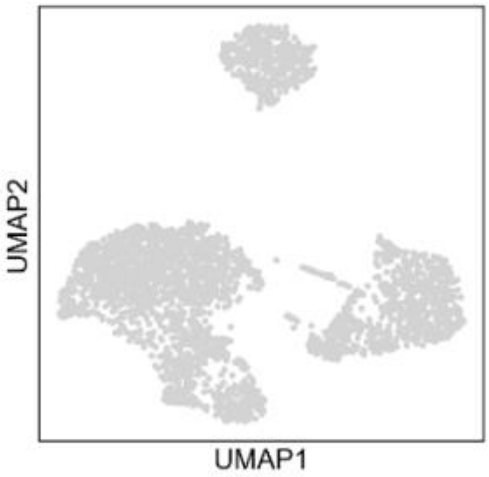
Quality control analysis was performed to evaluate the overall quality of the single-cell dataset and to remove low-quality cells prior to downstream analysis. The violin plots demonstrated the distribution of detected genes per cell, total transcript counts, and the percentage of mitochondrial gene expression. Most cells contained approximately 600–1200 detected genes and 1500–3000 total counts, indicating adequate sequencing coverage. A small subset of cells showed elevated mitochondrial RNA percentages above 5%, suggesting cellular stress or apoptosis. These cells were removed from further analysis. After filtering, 2638 high-quality cells remained, indicating that the majority of the dataset was suitable for transcriptomic analysis.



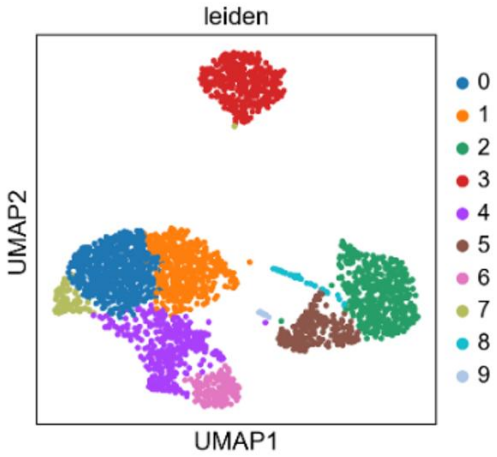
UMAP visualization and clustering

Dimensionality reduction using PCA followed by UMAP embedding revealed a non-random structure in the dataset. Instead of forming a continuous distribution, the cells organized into several clearly separated clusters. This indicates that the cells do not represent a homogeneous population but instead consist of multiple transcriptionally distinct groups. The presence of discrete clusters suggests that gene expression differences reflect underlying biological cell identities rather than technical variation.





Graph-based clustering using the Leiden algorithm further partitioned the dataset into ten transcriptionally distinct clusters.

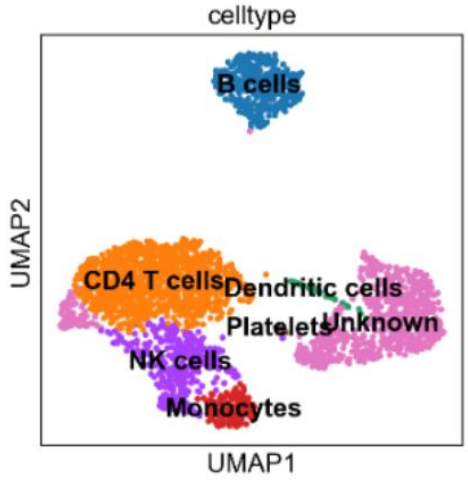


Each cluster contained cells with similar expression profiles, suggesting that each group corresponds to a specific immune cell population.

Annotated cell atlas

To determine the biological identity of each cluster, I assigned cell types based on canonical immune cell marker genes. Mapping these annotations onto the UMAP embedding produced a cell atlas of peripheral blood mononuclear cells. The largest populations corresponded to CD4 T lymphocytes, indicating that T cells dominate the peripheral blood immune compartment. Additional clusters were identified as B cells, natural killer (NK) cells, monocytes, dendritic cells, and platelets.

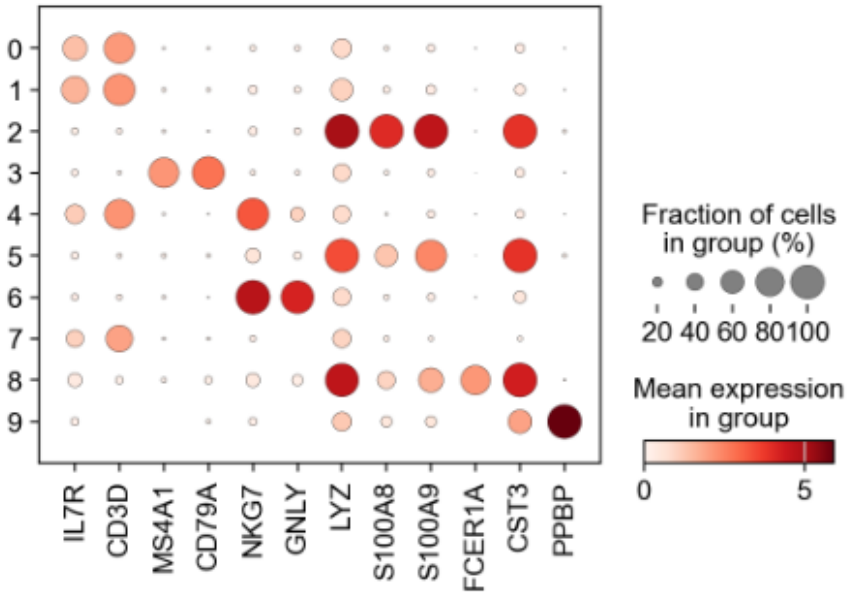
The spatial separation of annotated cell types on the UMAP plot confirmed that clustering was biologically meaningful. Cells of the same type localized to the same region of the embedding, demonstrating that transcriptional similarity reflects shared biological function.



Marker gene dotplot analysis

To further validate the cell type assignments, expression of known immune cell marker genes was visualized using a dotplot. T cell populations showed strong expression of CD3D and IL7R, B cells expressed MS4A1 and CD79A, and NK cells expressed cytotoxicity-associated genes NKG7 and GNLY. Monocytes demonstrated high expression of LYZ and inflammatory genes S100A8 and S100A9. Dendritic cells expressed FCER1A and CST3, while platelets showed specific expression of PPBP.

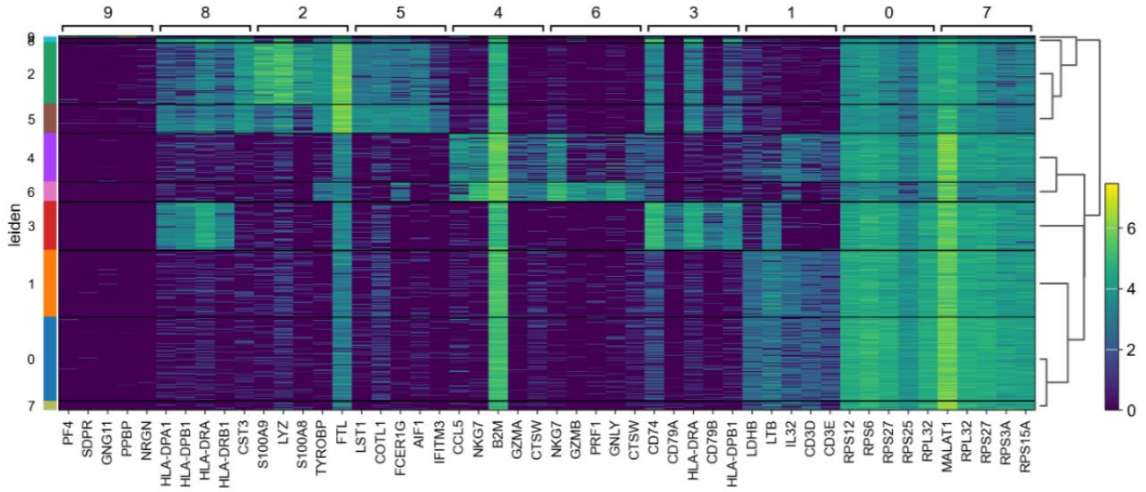
The selective expression of these markers in distinct clusters confirmed that the computational clustering accurately identified known immune cell populations.



Differential expression heatmap

Differential gene expression analysis identified genes that were significantly enriched in each cluster compared to all other cells. The heatmap demonstrated that each cluster possesses a unique transcriptional signature. For example, T cell clusters were enriched for T cell receptor-associated genes, while monocyte clusters displayed inflammatory and innate immune response genes. Platelets showed a distinct expression pattern dominated by platelet activation genes.

These findings confirm that the clusters represent biologically distinct cell types rather than arbitrary computational groupings. Together, the clustering and differential expression results demonstrate that single-cell RNA sequencing can successfully reconstruct immune cell composition from transcriptomic data.



Biological interpretation

The single-cell transcriptomic analysis revealed that peripheral blood mononuclear cells are highly heterogeneous and consist of multiple transcriptionally distinct immune populations. The largest group identified in the dataset corresponded to CD4 T lymphocytes. This finding is biologically expected because T cells represent the dominant circulating immune population in human peripheral blood. Their high abundance in the dataset confirms that the sequencing and clustering pipeline accurately captured the composition of the immune system.

B lymphocytes were identified by strong expression of MS4A1 and CD79A, genes associated with B cell receptor signalling and antibody production. Although B cells were less abundant than T cells, they formed a clearly separated cluster in the UMAP embedding, indicating a distinct transcriptional identity. This separation reflects their specialized role in adaptive immunity, particularly in humoral immune responses.

Natural killer (NK) cells were characterized by high expression of NKG7 and GNLY, genes involved in cytotoxic activity. NK cells play a critical role in the early defines against viral infection and tumour cells by directly killing abnormal cells without prior antigen exposure. The presence of a well-defined NK cell cluster demonstrates that single-cell RNA sequencing can distinguish immune populations based on functional gene expression programs rather than only lineage markers.

Monocytes showed high expression of LYZ, S100A8, and S100A9, which are associated with innate immune activation and inflammatory signalling. These genes are commonly upregulated during infection and inflammatory responses. The identification of monocytes highlights the ability of single-cell transcriptomics to capture cells involved in rapid innate immune responses and suggests that transcriptional profiling could be used to monitor inflammatory states in clinical samples.

A smaller cluster corresponding to dendritic cells was identified based on FCER1A and CST3 expression. Dendritic cells act as antigen-presenting cells and form a link between innate and adaptive immunity by activating T lymphocytes. Their detection, despite relatively low abundance, demonstrates the sensitivity of single-cell RNA sequencing in identifying rare but functionally important immune populations.

Finally, a platelet cluster was detected through strong expression of PPBP. Platelets are not classical nucleated immune cells; however, they contain RNA and contribute to inflammatory and coagulation processes. Their presence in the dataset further illustrates the capacity of single-cell RNA sequencing to identify diverse cellular components within a biological sample.

Overall, the transcriptional differences observed between clusters reflect known immune cell functions. Cells grouped together because they expressed genes required for their biological roles, such as antigen recognition, cytotoxic activity, or inflammatory signalling. This confirms that unsupervised clustering of gene expression profiles can reconstruct cellular composition without prior labelling or microscopy. The results demonstrate how single-cell RNA sequencing provides a detailed map of immune system organization and enables the study of cellular specialization within complex tissues.

The ability to recover known immune cell populations using unsupervised clustering indicates that transcriptional identity is strongly correlated with cellular function.

Limitations

Despite successfully identifying major immune cell populations, several technical and biological limitations should be considered when interpreting the results of this analysis.

One important limitation of single-cell RNA sequencing is the presence of dropout events. Because mRNA capture efficiency is limited, transcripts that are truly expressed in a cell may not be detected during sequencing. As a result, gene expression matrices contain many zero values that do not necessarily represent true biological absence. This sparsity can affect clustering accuracy and may lead to underestimation of marker gene expression in certain cells.

In addition, single-cell data are affected by technical noise introduced during cell capture, library preparation, and sequencing. Variability in amplification efficiency and sequencing depth can produce artificial differences between cells that are not biologically meaningful. Although normalization procedures reduce these effects, technical variation cannot be completely eliminated.

Another potential issue is the presence of doublets, in which two cells are captured within the same droplet and sequenced together. Doublets can produce hybrid gene expression profiles that resemble a new cell type and may lead to incorrect cluster

identification. While filtering based on gene counts reduces this risk, some doublets may remain undetected.

Batch effects also represent a common challenge in single-cell transcriptomics. Differences between experimental runs, reagent lots, or sequencing conditions can introduce systematic variation unrelated to biology. In this project a single dataset was analysed, minimizing batch-related variability; however, integration of multiple datasets would require additional correction methods.

Finally, single-cell RNA sequencing lacks spatial information. The technique measures gene expression but does not preserve the physical location of cells within tissues. Consequently, interactions between neighbouring cells and tissue architecture cannot be directly inferred. Spatial transcriptomics methods would be required to determine how identified cell populations are organized within the biological environment.

Overall, while single-cell RNA sequencing provides powerful insight into cellular heterogeneity, the results should be interpreted with awareness of these technical and methodological constraints.

Conclusion

This study demonstrates that single-cell RNA sequencing combined with computational analysis can successfully reconstruct the cellular composition of human peripheral blood. Using an unsupervised clustering approach, multiple immune cell populations were identified and validated using known marker genes. The agreement between transcriptional clustering and established biological cell types confirms that gene expression profiles contain sufficient information to infer cell identity without prior labeling.

Overall, the analysis highlights the power of single-cell transcriptomics for studying cellular heterogeneity and provides a reproducible workflow for analysing high-dimensional sequencing data. These methods are broadly applicable to other tissues and disease contexts, where understanding cell-specific gene expression is essential for interpreting biological mechanisms and identifying therapeutic targets.

References

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. & Surani, M.A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), 377–382. https://www.nature.com/articles/nmeth.1315

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A. & McCarroll, S.A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5), 1202–1214. linkinghub.elsevier.com/retrieve/pii/S0092867415005498

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J. & Bielas, J.H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8, 14049. https://www.nature.com/articles/ncomms14049

Wolf, F.A., Angerer, P. & Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19, 15. https://link.springer.com/article/10.1186/s13059-017-1382-0

Kiselev, V.Y., Andrews, T.S. & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20, 273–282. https://www.nature.com/articles/s41576-018-0088-9

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P.L., Rozenblatt-Rosen, O., Lane, A.A., Haniffa, M., Regev, A. & Hacohen, N. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335). https://www.science.org/doi/10.1126/science.aah4573