Effects of scRNA hyperparameter selection on downstream results and biological implications of findings

Abdulrahman Al-Sharabati
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, MN
alsha054@umn.edu

Nicole Sullivan
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, MN
sull1120@umn.edu

Stephanie M. Holtorf
The Hormel Institute
Dept. of Bioinformatics and Computational Biology
University of Minnesota
Austin, MN
sholtorf@umn.edu
ORCID:0000-0002-3867-0611

Husheng Ding
Dept. of Bioinformatics and Computational Biology
University of Minnesota
Rochester, MN
ding0326@umn.edu

***Abstract*—Traditional sequencing methods for determining the gene expression of a tissue, such as bulk RNA sequencing, give an average of gene expression within a tissue sample. However, tissues are made up of many different cell types and bulk RNA sequencing masks the distinct expression profiles for individual cells. Single cell RNA sequencing (scRNA-seq) is a method first developed and described by Tang et al., 2009 [1]. It is a novel approach to study tissue heterogeneity at a single cell level.**

The development of this technology has increased rapidly, as well as computational methods to address the unique challenges that come with high-throughput sequencing and the analysis of large gene expression matrices across thousands of cells. Originally, bulk RNA sequencing analysis methods were applied to single cell RNA sequencing, but this unique technology has necessitated the development of new analysis techniques to address issues unique to scRNA-seq [2] [3].

Many of the current methods for scRNA-seq analysis are not reproducible across different tissue types. This has led to reproducibility issues within the field [4]. For our project, we sought to address challenges unique to scRNA-seq analysis as shown in Fig 7. Specifically, the analysis pipeline involves many pre-processing steps, such as quality control, normalization, and scaling, that have been adapted to scRNA-seq. Additionally, there are many computational techniques that have been addapted to single cell data such as dimensionality reduction, clustering cells based on gene expression, using marker genes within each cluster to identify cell type, and conducting differential gene expression to compare different cell types and clusters. We compared how different parameters within the scRNA-seq pipeline affect the downstream analysis and reproducibility issues within the scRNA-seq field in the context of a publicly available bone marrow dataset [5].

***Index Terms*—scRNA-seq, genomics, computational biology**

GitHub: https://github.umn.edu/alsha054/CSCI5481_SkinCancerSingleCell

I. Summary of Previous Findings

There is an interesting phenomenon in case reports where some human patients that have received bone marrow transplants have developed cancer of donor bone marrow origin years after the transplant [6]. Using this knowledge, prior work in the Rebecca Morris lab was able to induce skin tumors in mice using a bone marrow transplant with carcinogen exposed bone marrow cells. These bone marrow cells were recruited into skin tumors using a well-validated two stage skin carcinogenesis model that was modified to include a bone marrow transplant [7].

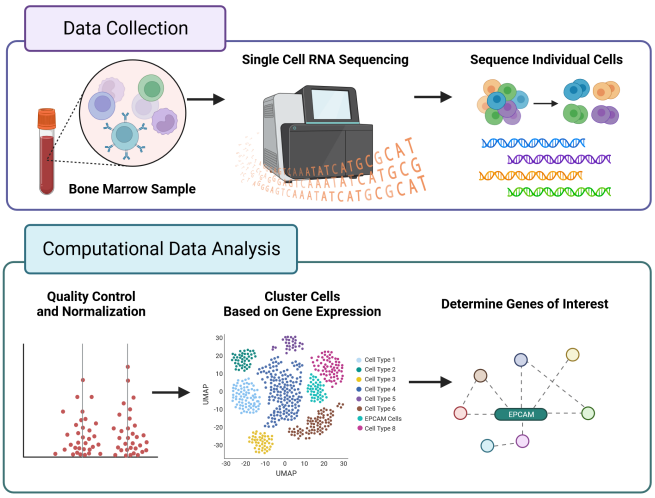

Fig. 1. Experimental Design showing the computational data analysis steps we assessed within a single cell RNA sequencing Bone Marrow Dataset.

This research led the groundwork for Ms. Holtorf's current project, which aims to identify what bone marrow cells were recruited into skin tumors in a mouse model with the goal of translating those results to humans. As skin is an epithelial tissue, traditional epithelial markers were used in biological techniques such as PCR, flow cytometry, and immunofluorescence microscopy to identify a potential subpopulation of epithelial cells in normal mouse and human bone marrow [8]. These bone marrow cells, positive for an epithelial specific marker, Epithelial Cell Adhesion Molecule (EpCAM), could be the bone marrow cells recruited during skin tumor formation. However, traditional methods have been limited in their ability to identify additional novel marker genes for this subpopulation of cells. Single cell RNA sequencing is a method that can analyze many bone marrow cells at once and determine other marker genes of interest for this subpopulation of cells. However, the technique is only just recently available commercially, so there are not many standard pipelines to perform the scRNA-seq analysis in a reproducible way across different sample types and sizes.

II. Results

A. Cluster quality

We evaluated cluster quality across the 7 experiments involving feature selection and clustering hyperparameters given in Table I. The quantity of clusters and cluster homogeneity across experiments is given in Fig. 2; cluster homogeneity results are shown in Fig. 3.

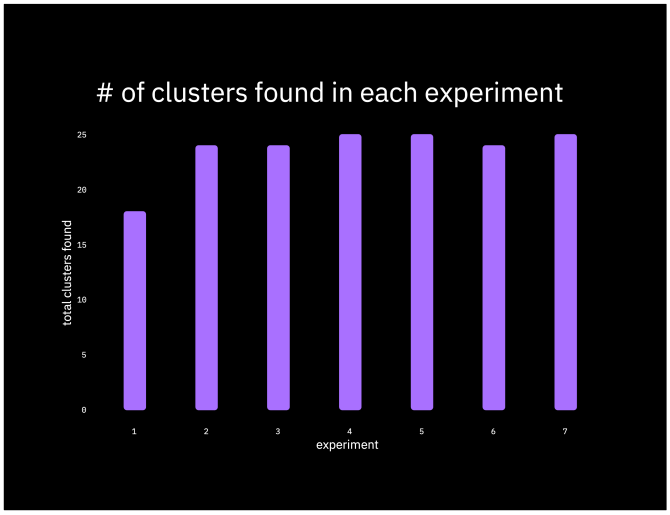

Fig. 2. # of clusters across feature selection and clustering algorithm experiments.

We note that we obtained similar numbers of clusters in each experiment (range: 18-25), regardless of the upstream hyperparameters. Interestingly, we see that cluster homogeneity appears to improve when selecting *fewer* genes as input to the dimensionality reduction algorithm (PCA), with experiment 4 (500 highest-variation genes selected) boasting the lowest WCSS. However, we note that further exploration into this finding is needed using metrics more fine-grained than WCSS.

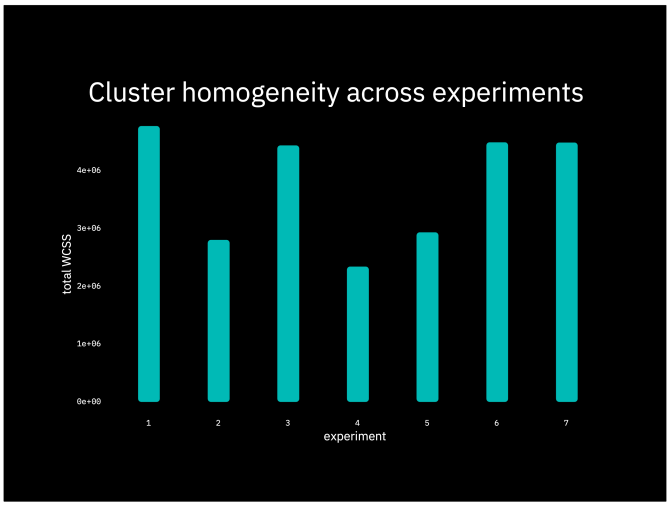

Fig. 3. Cluster homogeneity across experiments.

B. Cell type annotation with GPT-3.5

We give an example of the cell type annotations generated for the experiment 2 clusters in Fig. 4; for demonstrative purposes, examples of the textual cluster summary and predicted cell types are shown for clusters 4 and 11 in Fig 5.

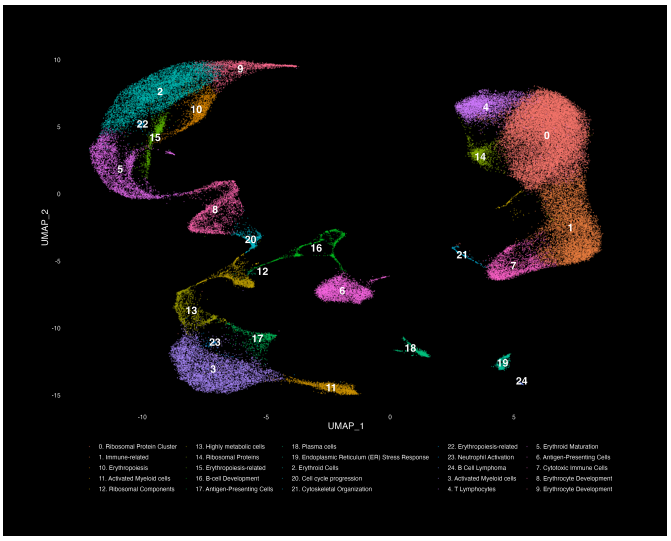

Fig. 4. UMAP 2D projection of clusters for experiment 2 with cell type annotations provided by GPT-3.5.

Overall, GPT-3.5 provides rich textual summaries for each cluster as well as predictions of the cell-type sub-populations represented therein, given just 5 of the top-differentiating genes for that cluster. In contrast, current methods for cell-type annotation are buggy and, unfortunately, we were unable to successfully run them in the given timeframe to produce side-by-side comparisons with which to evaluate GPT-3.5 results. This alone demonstrates the extreme usefulness of GPT-3.5: with just several lines of code (provided in our GitHub), we were able to call the API and use natural language to return

the necessary response - no databases or manual intervention needed.

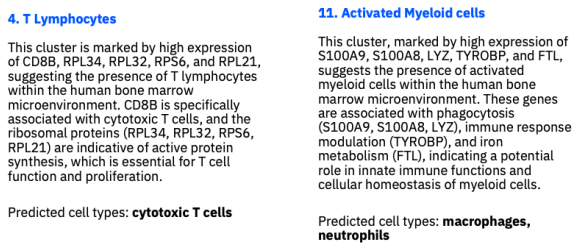

Fig. 5. GPT-3.5-generated cell summaries and predicted cell types for two clusters in experiment 2.

C. Sensitivity of downstream results

We also evaluated the sensitivity of downstream results to upstream hyperparameters. The distribution of epithelial marker gene AUROCs is given in Fig. 6. We found that, for the most part, the predictive power of the epithelial markers selected were unaffected by upstream hyperparameters - CD36, CD59 and KRT1, especially, show high ROCs for at least one cluster across *all* the hyperparameter experiments. We also conducted a qualitative analysis (comparing UMAP plots) of the clusters in which these markers showed highest expression, which revealed that cells with highest expression of these genes tended to show up in the same cluster in each experiment, demonstrating robustness to changes in feature selection and clustering hyperparameters.

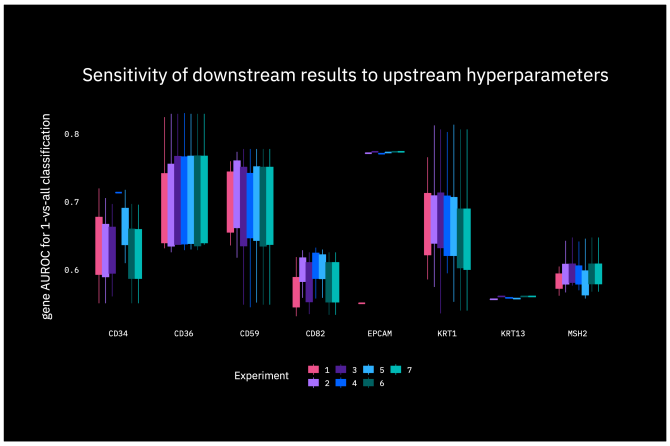

Fig. 6. Distribution of AUROCs for epithelial marker genes across clusters in each experiment.

D. Biological variation amongst sub-populations

Next, we conducted a biological analysis of sub-populations (gender, age). The UMAP plots are similar between the male group and the female group in Fig 7a. We saw almost equal expression of some representative genes in the two groups in Fig 7b.

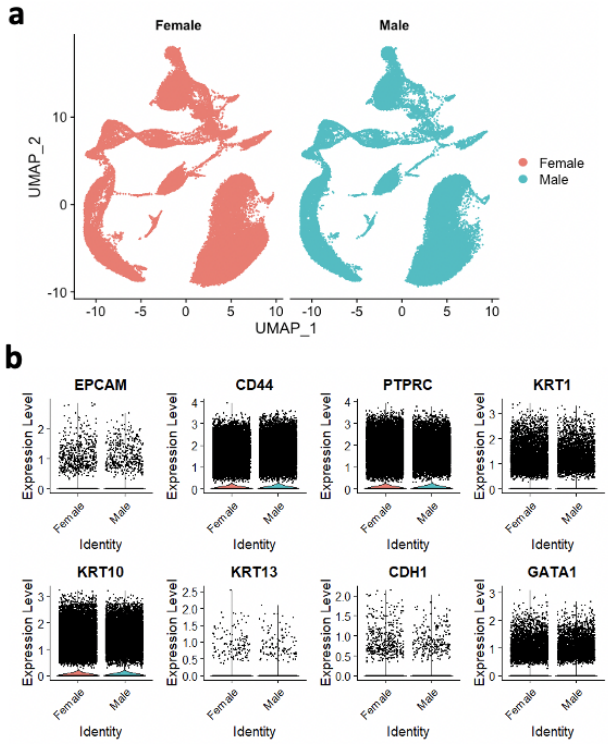

Fig. 7. Gender-associated variations in the two groups. UMAP plots in (a) show that the male and female subsets look similar. Violin plots in (b) show that cells express common epithelial and stem cell markers equally.

We observed age-associated changes in the bone marrow samples when we superimposed the UMAP plots from the younger group (< 50) to the older group ($>= 50$) in Fig 8.

III. Methods

A. Dataset

We used a single cell RNA sequencing dataset from 20 human bone marrow donors, including 10 females and 10 males with ages ranging from 24 to 84 years old. The dataset is publicly available, and there are 3 files for each donor (barcodes, genes, matrix). We pre-processed and integrated the donor's age and gender so that they could be used for identifying age and gender associated changes.

B. Technical: hyperparameter search

Within the standard Seurat scRNA-seq analysis pipeline, we experimented with the following hyperparameters:

1) Data pre-processing
   - Filters on mitochondrial-DNA
   - Filters on # of genes expressed
   - Filters on # of cells expressing a gene
   - Normalization method
2) Feature selection
   - # of highly variable genes selected
3) Dimensionality reduction
   - # of principal components kept

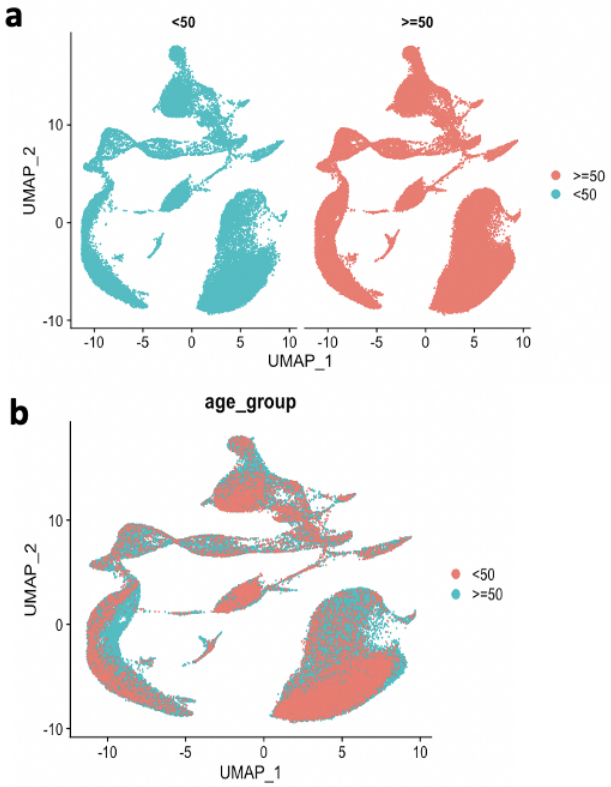

Fig. 8. Age-associated variations in the two groups. UMAP plots in (a) show that groups under 50 and equal to or over 50 look similar. The superimposed UMAP in (b) shows that age subsets look similar.

4) Clustering

- Clustering algorithm

As running a random grid search of all the possible hyperparameter combinations in the intervals of interest would be computationally prohibitive with the given time constraints (one run in Seurat on average takes 20-30 minutes), we worked together with our biological subject matter experts (Husheng and Stephanie) to develop a set of hyperparameters to test that represented combinations of greatest interest to counterparts in the biological research community. Altogether, we ran 25 different experiments; the hyperparameters toggled in the feature selection and clustering experiments are given in Table I.

TABLE I
EXPERIMENTS WITH FEAT. SELECTION AND CLUSTERING HYPERPARAMETERS IN SEURAT

experiment	*n variable features*	*n PCs*	*clustering algo*
1	3,000	20	original Louvain
2	1,000	20	original Louvain
3	2,000	20	original Louvain
4	500	25	original Louvain
5	800	25	original Louvain
6	2,000	25	Louvain w/ MLR
7	2,000	25	SLM

The most straightforward approach to compare the robustness of results across experiments is to evaluate cluster homogeneity and quantity of clusters found. To measure cluster homogeneity, we used within-cluster sum of squares (WCSS), which can be formalized as:

$$WCSS = \frac{1}{2|S_i|} \sum_{x,y \in S_i} \|x - y\|^2$$

A lower total WCSS (across all the cluster in the experiment) indicates higher within-cluster homogeneity, while higher WCSS indicates greater heterogeneity and mixing of less closely related observations (cells).

C. Technical: annotating clusters with cell-type and descriptions

Based on the top 5 most predictive genes in each cluster, we used GPT-3.5 to annotate the clusters found. We experimented with several prompts utilizing different prompt engineering approaches (0- vs. 1- vs. 2-shot learning, asking for a structured response in JSON vs. Python dictionary). Ultimately, we used 2-shot learning and asked for 3 outputs structured in a Python dictionary. The final prompt is available in the project's GitHub repo (https://github.umn.edu/alsha054/CSCI5481_SkinCancerSingleCell).

To access GPT-3.5, we used openai's Python sdk, version 1.3.7 (importantly, note that this version has breaking changes from versions < 1.0). We utilized the asyncio library to parallelize batch requests to the API and exponential backoff between batches to ensure rate limits weren't exceeded.

D. Technical: downstream epithelial analysis

To determine sensitivity of biological results to the hyperparameters set, we investigated the expression patterns of 9 different epithelial marker genes significant in skin cancer research: EPCAM, CDH1, KRT1, KRT13, CD34, CD36, CD59, CD82, MSH2. For each cluster, area under the receiver operator curve (AUROC) of each gene in discriminating between the cluster of interest and all the other clusters was calculated (one of the standard ways of determining marker genes for each cluster). AUROC is a metric used to determine a model's balance between true and false positive rates and can be obtained by plotting the a model's true positive rate vs. false positive rate at varying thresholds and integrating the area underneath the resulting curve (in practice, an approximation of integrating is usually used).

AUROC ranges from 0 to 1, with 1 (and 0, consequently) representing the best possible discrimination, and 0.5 representing discrimination no better than chance. We then looked at the distribution of AUROCs of the 9 epithelial markers selected across all the clusters in each experiment; we note that not all the epithelial markers achieve AUROC values that would warrant consideration as "markers" for that particular cluster, but investigate this metric as a way of understanding how stable the "texture" or sub-population of clusters are (using marker genes discovered in this way to identify the cell type of the cluster is established practice in the scRNA-seq

research). Put another way, this is a way of understanding the variability of cell-type assignment depending on the upstream hyperparameters used.

E. Technical: code improvements

The original code we started working off of was provided by one of our team members, Ms Holtorf. This code is originally a part of her ongoing research efforts. If the code was to enjoy lasting utility in the research, there were several changes needed to make the code more flexible and performant.

First off, since the single cell pipeline is highly sensitive to the results of the previous steps, it is a worthwhile effort to make sure we are passing optimal configurations to those previous steps. We also are not sure if the choice of one parameter is linked to the choice of another parameter. This is similar to cost optimization problems in computer science, where we need to seek a balance between multiple linked parameters, such as resource cost and budget. Thus, the parameters need to be tested in permutations with other parameters, and not in isolation.

To facilitate this, much of our initial work was de-hardcoding the configurations and allowing them to be passed in to the script arguments. We then wrote a Bash script to run these permutations in parallel, since R is single-threaded. However, the large datasets in the memory of each instance ended up overwhelming our machines and crashing them. We ended up running the tests serially overnight.

Last, but not least, R is single-threaded, which is a shame because all modern CPUs have four cores at least. However, Seurat does expose parallel computation support for some of its API, which is possible to explore in future work. Another idea for the future is improving the memory management in the script so that memory load is not quite as heavy and we may be able to run multiple instances at a time.

F. Biological: subgroup variation

We also explored biological patterns within various subpopulations. To do so, the top 10 variable genes were identified, and most of them were coding immunoglobulin proteins. The variable counts were found in 2,500 genes and the other 31,194 genes are non-variable. We were curious about the variation between healthy donors as well as age-associated and gender-associated changes in cell population frequencies. We split the 20-bone marrow donors into two groups: the age of one group was less than 50, and the age of the other group was equal to or older than 50. There did not seem to be too many gender or age specific patterns within the initial dataset analysis, but further computational analysis is needed to see if more hidden changes exist within the dataset between these variables as it is expected that biologically there would be gender and age associated changes.

IV. Conclusion

We have just described our findings when looking at the computational aspects of the single cell RNA sequencing data analysis workflow. Overall, we saw that scRNA-seq was able to identify cells with specific markers for epithelial cells within a human bone marrow dataset, confirming results seen in Ms. Holtorf's previous biological experiments and scRNA-seq analysis of other datasets. We were able to optimize and streamline the code to be able to use different parameters within our analysis and determine if they had a noticeable effect on downstream analysis. We were also able to look within the dataset to see if there were any noticeable gender or age associated changes across samples. Overall, we found that the most commonly used data analysis package, Seurat, is quite robust to changes in parameters. However, since the Seurat package is only in R, we were limited when analyzing such a large dataset and Seurat did not implement well when trying to perform subsequent analysis within Python.

Going forward, Ms. Holtorf will be able to use changes implemented within our group project to analyze the age and gender associated changes within the dataset, as well as use additional computational and biological methods to further elucidate the subpopulation of epithelial cells within the bone marrow.

Acknowledgments

Ms. Holtorf:

- Experimental design and bone marrow dataset to use
- Original Seurat workflow code in R
- Instructions on utilizing human bone marrow dataset
- Authored Abstract, Summary of Previous Findings, and Conclusions
- Created Figure 1 of Experimental Design using Biorender Subscription
- Extensive consultation on technical work and interpreting results of each experiment

Mr. Abdulrahman:

- Hyperparameter experiments with pre-processing and normalization
- Code improvements and optimization
- Provided access to Overleaf Subscription for Final Write-Up
- Set up GitHub for group to use to share code
- Writing up parts of the Results and Methods sections

Ms. Sullivan:

- Hyperparameter experiments with feature selection, dimensionality reduction and clustering
- Annotating clusters with cell type using GPT-3.5 via openai sdk
- Evaluating robustness of downstream epithelial analysis across feat selection, dim reduction, and clustering experiments
- Making figures for Results Section
- Wrote up most of the Methods section

Mr. Ding:

- Implementing outside data into the metadata of Seurat object
- Sub variation analysis of age and gender associated changes within the dataset
- Consultation as needed as a biological SME to technical contributors
- Making figures for Results section
- Writing parts of the Methods and Results sections

REFERENCES

[1] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377-82. doi: 10.1038/nmeth.1315. Epub 2009 Apr 6. PMID: 19349980.

[2] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun;36(5):411-420. doi: 10.1038/nbt.4096. Epub 2018 Apr 2. PMID: 29608179; PMCID: PMC6700744.

[3] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6. PMID: 31178118; PMCID: PMC6687398.

[4] Grün D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. Cell. 2015 Nov 5;163(4):799-810. doi: 10.1016/j.cell.2015.10.039. PMID: 26544934.

[5] Oetjen KA, Lindblad KE, Goswami M, Gui G, Dagur PK, Lai C, Dillon LW, McCoy JP, Hourigan CS. Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. JCI Insight. 2018 Dec 6;3(23):e124928. doi: 10.1172/jci.insight.124928. PMID: 30518681; PMCID: PMC6328018.

[6] Kolb HJ, Guenther W, Duell T, Socie G, Schaeffer E, Holler E, Schumm M, Horowitz MM, Gale RP, Fliedner TM. Cancer after bone marrow transplantation. IBMTR and EBMT/EULEP Study Group on Late Effects. Bone Marrow Transplant. 1992;10 Suppl 1:135-8. PMID: 1521085.

[7] Park H, Lad S, Boland K, Johnson K, Readio N, Jin G, Asfaha S, Patterson KS, Singh A, Yang X, Londono D, Singh A, Trempus C, Gordon D, Wang TC, Morris RJ. Bone marrow-derived epithelial cells and hair follicle stem cells contribute to development of chronic cutaneous neoplasms. Nat Commun. 2018 Dec 13;9(1):5293. doi: 10.1038/s41467-018-07688-8. PMID: 30546048; PMCID: PMC6294255.

[8] Holtorf SM, Boyle J, Morris R. Evidence for EpCAM and Cytokeratin Expressing Epithelial Cells in Normal Human and Murine Blood and Bone Marrow. J Vis Exp. 2023 Apr 21;(194):10.3791/65118. doi: 10.3791/65118. PMID: 37154563; PMCID: PMC10653199.