2025 August

Novel Approaches for dimensionality reduction of single cell RNA-sequencing to gene expression programs using hematopoietic stem cell transplantation as a model

Albert Ho, Nick Chan, B.S., Hojun Li, M.D., Ph. D.

Abstract

Hematopoietic stem cell transplantation (HSCT) is a critical therapeutic approach to treating hematological malignancies, immune deficiencies, and other disorders affecting blood cell production. Single-cell RNA sequencing (scRNA-seq) provides a view of cellular heterogeneity, allowing us to characterize gene expression patterns. In this study, I employed novel analysis packages for scRNA-seq, specifically the starCAT analysis package, to analyze gene expression over time following HSCT, focusing on CD34+ early hematopoietic stem and progenitor cells. By using the starCAT package, we identified key signatures, lineage-specific gene expression patterns (GEPs), and transcriptional programs. Furthermore, with this analysis, we demonstrated the versatility of the starCAT package beyond its intended use in T-cell annotations, evaluating its accuracy in interpretation and compatibility across various applications.

Introduction

Hematopoietic stem cell transplantation (HSCT), also known as bone marrow transplantation, involves the infusion of healthy hematopoietic stem cells into patients with compromised or depleted bone marrow. Clinically, several HSCT procedures exist, with differences in the method of stem cell retrieval and transplantation method depending on the context of the disease and patient's condition. HSCT is a critical therapeutic strategy for the treatment of both malignant and non-malignant hematological disorders through the restoration of hematopoietic function. This restoration enables the destruction of malignant tumor cells and regeneration of functional immune and blood cells seen in leukemia, lymphoma, and other immunodeficiencies. Having a crucial role in immune reconstruction, HSCT serves as a pivotal role in enabling the recovery of the hematopoietic system and improving long-term patient outcomes[2].

Single-cell RNA sequencing (scRNA-seq) has emerged as a key tool in analyzing cellular heterogeneity through the analysis of lineage-specific gene expression patterns (GEPs). The contribution of GEPs within a hematopoietic system, especially with stem cells found in bone marrow, allows insight on hematopoiesis following HSCT. By analyzing cells that expressed the CD34 positive protein, known as a biomarker for HSCs, and hematopoietic

stem precursor cells, this marker would allow us to tag cells that have an influence within hematopoiesis. Moreover, we can follow the HSCs' lineage as it differentiates across time[8].

However, when analyzing these readings, hard clustering, which was the predominant analysis technique, cannot easily reflect the multiplicity of GEPs which are expressed[1]. To overcome these limitations, component-based models like non-negative matrix factorization (NMF), hierarchical Poisson factorization, and SPECTRA were introduced.

Although these new component-based models assist in analyzing and identifying novel GEPs, they do tend to struggle due to their computational methods, specifically in "averaging" GEPs. To combat this consensus non-negative matrix factorization (cNMF) was introduced to discover and identify GEPs through non-negative matrix factorization. cNMF would have "replicates" that would identify a specified number of GEPs, breaking the inputted cell by gene dataset into different clusters (GEPs). Following this calculation, the cNMF algorithm would generate associated GEP usage scores per cell matrix, and a gene contribution matrix for each GEP as well, a direct result of the non-negative matrix factorization (Fig 1A, 1C).

However, as convenient as it is to use cNMF to generate GEPs and their associated matrices, if you wanted to conduct an analysis between two different datasets, using the same GEPs, issues could arise. Since you are running an entirely new analysis on a new dataset, cNMF would create an entirely new list of GEPs as well as the associated matrices, possibly excluding GEPs that you wanted to analyze. Thus, this poses the question: Can we use a different computational method that could infer information about a particular dataset using prior known information?

starCAT, was introduced as a method to characterize T-cells, specifically using predefined GEPs to capture activation states and cellular subsets[1]. The authors proposed to utilize non-negative least squares regression (NNLS), to infer data from two existing matrices. Within the algorithm, the regression would "test" various replicates of the inferred matrix (usage scores) from the query datasets (cell by gene matrix and gene contribution score matrix). After compiling all the replicates, the algorithm would find a resulting GEP usage by cell matrix based on the query and inputted dataset.

In this study, we applied starCAT to an already GEP analyzed dataset to verify its transferability to other datasets outside of its original purpose, checking its correlation and visual outputs. Once we verified these results, we wanted to test starCAT's application on an entirely new dataset, a post-HSCT CD34 dataset, to view its ability to conduct inference on new data based on an already processed reference.

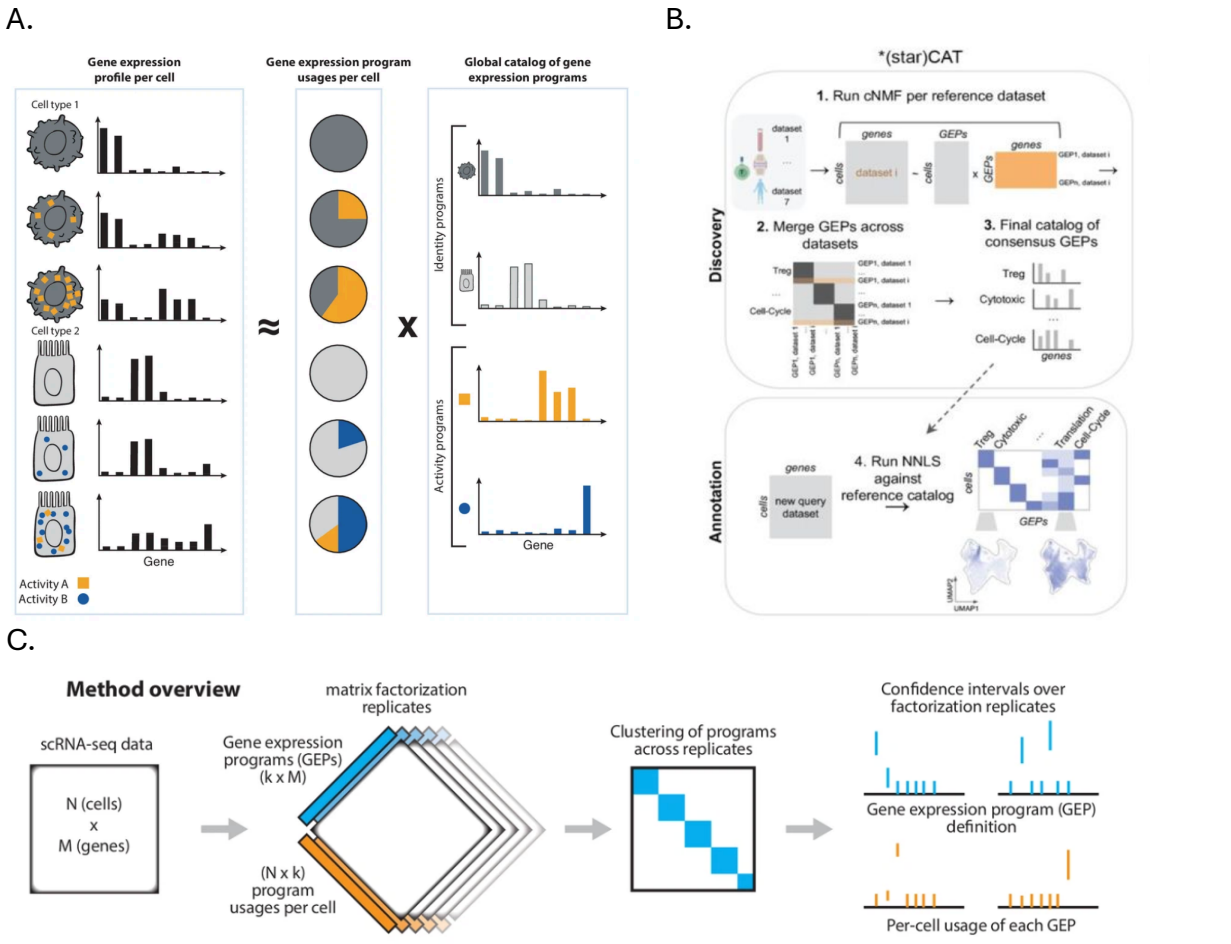


Fig 1. Algorithms and Models for Gene Expression Programs
(A) Schematical representation of how gene expression is modeled with matrix factorization. On the left side, we plug in the individual gene expression profiles for each cell. With this dataset, it can be split into a global catalog of gene expression programs for the entirety of our dataset, and into the gene expression program usages for each of the cells within the dataset. The different colorings for each of the individual portions signify the variability that each of these reads can have[5]. **(B)** A schematical representation of the starCAT (T-cell Annotator) pipeline and methodology[1]. **(C)** A schematical representation of the consensus matrix factorization pipeline and methodology[5].

Methods

Dataset Retrieval[3]

The data was retrieved from a report on the *Integrative and Longitudinal Analysis of Hematopoietic Reconstitution in Transplant Recipients*. The methods discussed below for the how the data was retrieved was referenced from the report.

Within the study, we monitored data from 12 pediatric patients who received hematopoietic stem cells (HSCs) through donor bone marrow after being diagnosed with a hematopoietic disorder (Fig. 2A)[3]. For this reason, we focused specifically on CD34 proteins as they are markers for hematopoietic stem and progenitor cells (HSPCs.) With the yielded samples, they were stained with hashtag oligos (HTOs) to differentiate between patients and time points. These transmembrane proteins which were first identified on HSCs and progenitor cells that are strongly enriched in hematopoietic progenitor populations[3,4]. Staining these proteins with a CD34 selection antibody, they were enriched to separate only CD34 positive and negative populations. After conducting quality control to remove erroneous cells and demultiplexing[3], we had the completed dataset, ready for analysis (Fig. 2B, 2C).

To use starCAT for our analysis, it is necessary to have a global reference of gene expression programs for the algorithm to infer the GEPs for each of our cells in the dataset. With a focus on hematopoietic stem cells following transplantation, we wanted to infer our information using a gene by GEP matrix that had an emphasis on the dynamics of hematopoiesis, especially one that dealt with differing age groups.

To do this, we utilized the 35 GEPs generated in *The dynamics of hematopoiesis over the human lifetime* (will be referenced as Human Lifetime paper throughout this report). Within the paper, the authors defined GEPs that drove lineage commitment over time, marking points in time where differentiating cells undergo irreversible changes that restrict further differentiating potential[6,7]. These GEPs were then compared to unilineage progenitor states and validated against signatures of committed hematopoietic cells from additional independent datasets using gene set enrichment analysis (GSEA)[6].

Utilizing the cNMF implementation to retrieve these usage scores per GEP, the paper used 35 components and a local-density-threshold of 0.15. As a result, cNMF generated 35 GEPs with associated scores for each gene, indicating their contribution to each GEP[6]. In addition, cNMF generated gene contribution scores, and GEP usage scores per cell, for each of these GEPs. With the usage scores for each GEP, we would use this as our baseline for gauging whether starCAT's method of inferring GEP usage scores were valid.

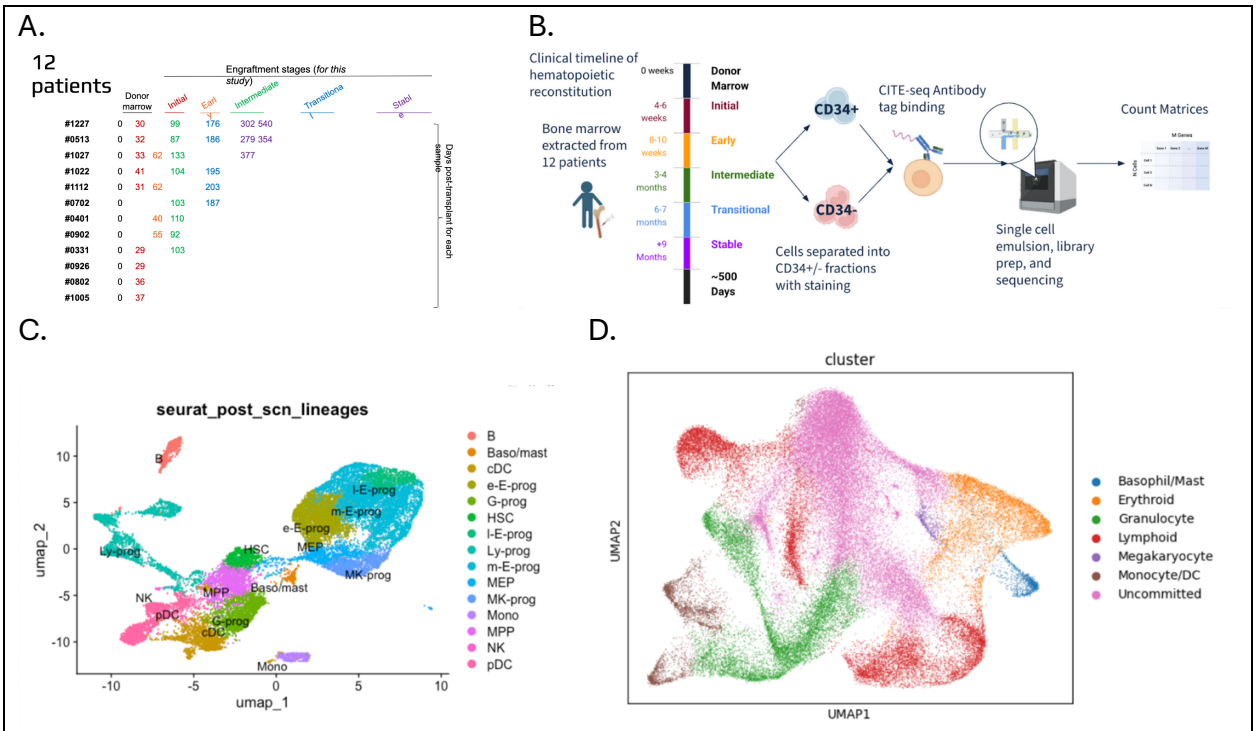


Fig 2. Process and Data Monitoring for Patients at Single Cell Resolution
(A) Chart of patients and bone marrows sampling time points. **(B)** Experimental designed of the study. The bone marrow was taken from 12 patients across the timeline of engraftment. Samples were then sequenced at single-cell resolution yielding transcriptome and surface protein profiles.[3] **(C)** UMAP of the cell type and lineage clusters from the Post Transplant dataset. **(D)** UMAP of the cell type and lineage clusters from the Human Lifetime dataset.

Matrix Preprocessing

When utilizing the starCAT package, adjustments to the datasets were required to correctly map and infer information through the package's algorithm. To do this, we first had to change the datasets into formats that were interactable with the algorithm. To retrieve and modify these datasets, which were both Seurat Objects, we converted them into Seurat object variables within the R text editor to allow for these modifications. When working with the starCAT package, our primary issue was within the cells by genes and reference genes by GEPs matrices as they had a dimensional mismatch which non-matrix multiplication did not accept as required by the package. To address this, we utilized matrix transformations to invert the orientation, matching the associated matrix format. In addition to this, the package required specific file formats that our original dataset did not come in. As a result, we converted these formats through the "readr" R package to match these requirements.

Dataset Visualization

To visualize our data, we utilized uniform manifold approximation and projection for dimension reduction (UMAP). By utilizing this visualization technique, we can visually recognize GEP clustering, giving us visual representation of how many cells utilize a specific GEP (shown numerically through starCAT's cell by GEP usage matrix.) Although we utilized UMAP visualizations for both datasets, the Human Lifetime paper's dataset utilized the built in Seurat UMAP projection while the transplantation dataset had to utilize coordinates extracted from the Seurat Object. Despite these differences, both methods rendered correctly formatted UMAP projections and clustering.

Correlation calculations

To quantify the validity of the starCAT implementation to infer GEP usage, we calculated correlation values for the resulting GEP usage per cell matrix. Following the methods mentioned in the paper that starCAT was proposed in, the authors mentioned that for starCAT to accurately infer the usage of GEPs, there must be an overlap of query and reference dataset of Pearson $R>0.7$[1]. Following this metric, we calculated the Pearson correlation coefficient of the inferred GEP usage with the simulated ground truth usage (in our case, the cNMF usage score) through a visualization and calculation Python script.

starCAT and cNMF Usage Score Plotting

To visualize the UMAP clusters, I developed a python script to grab coordinates from each cell location (UMAP 1 and UMAP 2 coordinates), as well as the associated GEP usage scores, and plotted them on an x-y axis coordinate plane. Each point would have a heatmap-style coloring, with the opacity of the color changing based on its usage score—higher score with higher opacity and vice versa. In addition to this heatmap style plotting, only the top 5% of the cells found in a specific GEP's usage score were kept. This was done so that only the most confident cells were plotted to reduce noise that could interfere with the visualization and cluster labeling. Within the Human Lifetime dataset, this was done for both the starCAT and cNMF GEP usage scores, while the Post Transplant dataset was for only starCAT (due to no cNMF runs to compare it to).

Reference Matrix Mismatch

When using the starCAT package, we came across an issue for negative reference values. Since starCAT is an implementation of non-negative matrix factorization, it is required that the matrices that we passed in were non-negative. However, on our first run, we passed in non-normalized scores for our reference gene by GEP matrix; as a result, the implementation failed. Taking this into account, we switched to a z-scored matrix that

could be passed into the implementation. However, by doing this, we limited the size of the genes within the dataset. Rather than using a reference that contained GEPs for around 40,000 genes, we were limited to 2,000 genes. Despite these changes, our downstream analysis still proved to be valid in the context of the datasets.

starCAT preprocessing

When using the starCAT package, you are required to retrieve the query dataset's count matrix (cell by gene matrix), barcodes, features, and gene names. As a given, the count matrix is the dataset that you want the package to carry out an inference on, so this is already provided. On the other hand, the barcodes, features, and gene names must be retrieved from the dataset. In our case, our datasets came in the form of a Seurat objects, so we retrieved information through Seurat object operations. There is a reference on starCAT's GitHub repo on how to retrieve this information (see Data Availability).

Results

Dynamics of Hematopoiesis CD34 Dataset StarCAT Inference

Prior to running starCAT on a novel dataset, we wanted to ensure that the data was viable for our analysis. To do this, we used the Human Lifetime paper's dataset for our initial runs of starCAT. With the already completed dataset and corresponding matrices (GEP usage scores per cell and a reference for GEPs by genes), we could test out whether running starCAT would output similar results to the paper's cNMF runs. With these runs, we can gauge how accurate and reputable the inference package would be on pre-existing data.

Within our analysis, there are 35 total GEPs that we could analyze. However, for the starCAT runs for both the Human Lifetime paper's dataset and the post-transplant dataset, we focused only on 9 GEPs: Monocyte Lineage (GEP 1), Elderly-biased HSC (GEP 4), Late-Erythroid Lineage (GEP 6), Basophil/Mast Cell Lineage (GEP 14), Dendritic Cell Lineage (GEP 25), Lymphoid Lineage (GEP 26), Fetal-biased HSC (GEP 28), Granulocyte Lineage (GEP 30), and Erythroid Lineage (GEP 32). Since we are analyzing HSCs across lineages and post-transplantations where they can potentially differentiate, we can view these GEPs as fingerprints for fate decisions of HSCs and aid in identifying age-related changes within patients. In addition to this, since these GEPs are known to exist within cells in bone marrow, they should be present in both datasets, allowing them to be transferrable between datasets.

Using these GEPs of interest, we preprocessed the data and matched data types to match the criteria required by the starCAT package (see Methods). With this, we ran starCAT on the associated files to retrieve the resulting GEP usage score matrix. To run this analysis, I created a Python script that would run a Pearson correlation calculation between this matrix and the Human Lifetime's cNMF GEP usage score matrix. When looking at the

matrix, we noticed that the correlation value averaged above starCAT's recommended correlation value of R>0.7 (Sup. Fig. 1A, B, C, D, E). As a result, we found that starCAT's analysis is accurate in inferring GEP usage scores based off a "reference" dataset (GEP contribution scores).

With our confirmed findings, we wanted to ensure that starCAT's cell inference was qualitatively sound. To do this, I developed a Python script that UMAP plotted the top 5% of cells within a particular GEP's usage score for both starCAT and cNMF's inference and explorations (see Methods). After running the script, we noticed that, visually, the results from both the starCAT and cNMF runs were relatively similar; verifying that the correlation results were sound (Fig. 3A, 3B, 3C).

Since starCAT demonstrated accuracy both through its implementation[1] and through the visual validation of its output, we are confident that it can make a reliable inference on datasets beyond T-cell annotation. In our runs on the dataset, using the GEPs of interest (which we expected to yield meaningful patters within the UMAP visualization) starCAT successfully inferred the usage scores consistent with our reference dataset (cNMF runs)[6].

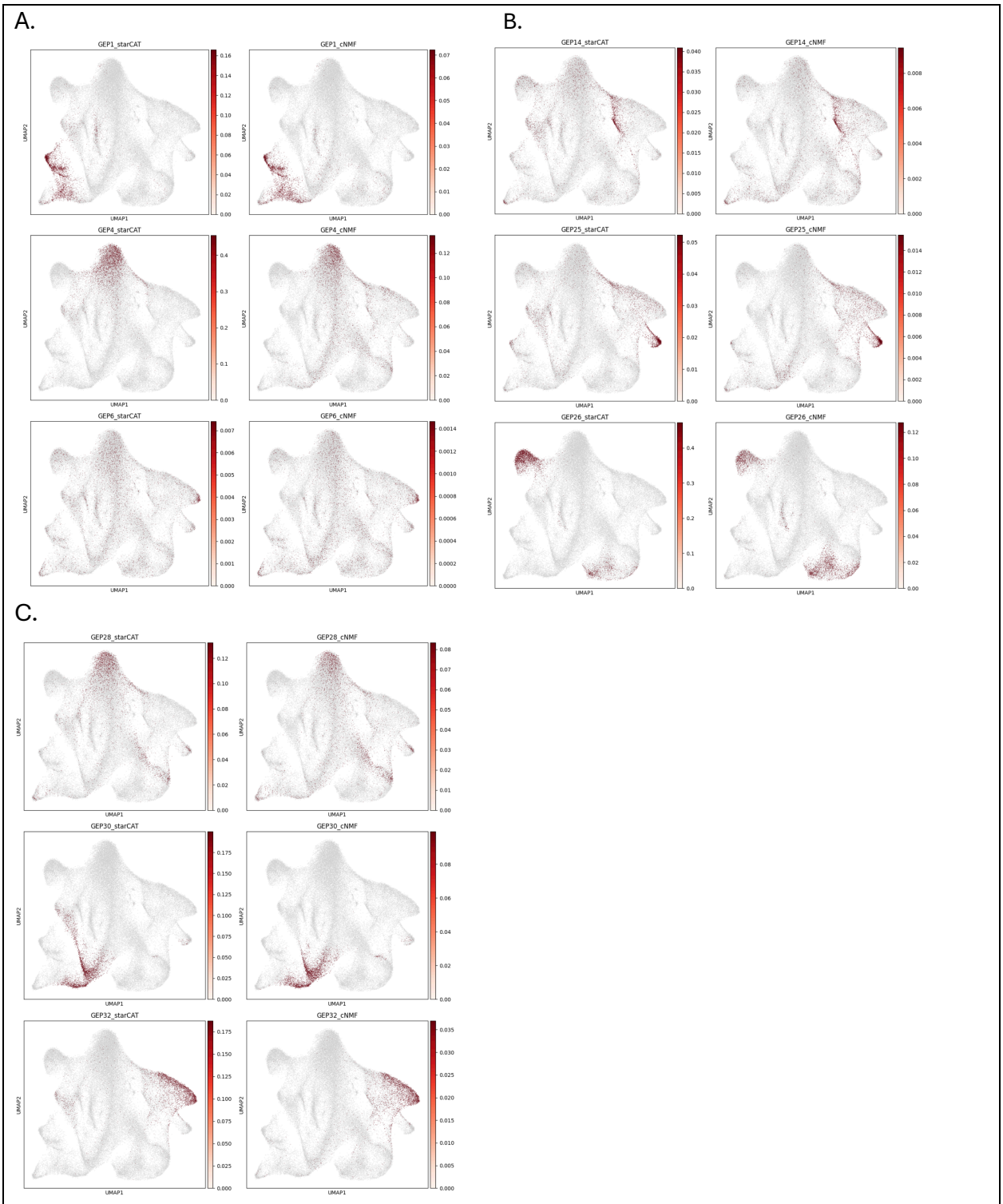


Fig 3. Human Lifetime UMAP of Top 5% Cells with GEP usage after starCAT inference with cNMF exploration comparison

(A) UMAPs for the top 5% cells of GEP usages for GEPs 1, 4, and 6. **(B)** UMAPs for the top 5% cells of GEP usages for GEPs 14, 25, and 26. **(C)** UMAPs for the top 5% cells of GEP usages for GEPs 28, 30, and 32.

Post Transplant CD34 Dataset StarCAT Inference

Using what we learned from the Human Lifetime dataset run of starCAT, with validation and inference, we wanted to test starCAT's applicability on an entirely new dataset. Using the same GEPs as the Human Lifetime dataset runs, we wanted to see if starCAT could interpret GEP scores on an entirely different dataset, knowing that the GEPs have an impact on HSC readings (additionally, they both come from the bone marrow; therefore, should have similar inference ability). To test this, we utilized the Post Transplant dataset to test starCAT's inference ability.

Using the cleaned dataset, we preprocessed it to fit within the required parameters of the starCAT implementation (see Methods). Moreover, since we are using a known catalog of GEP contribution scores (reference dataset) from our known runs on an HSC dataset, we can carry over the same dataset for this analysis. With both required matrices, we ran starCAT on our Post Transplant dataset, inferring our GEP usage scores from the inputted cell by gene dataset and reference dataset.

From our starCAT runs, we retrieved the corresponding GEP usage scores from the inference algorithm. With this information, I utilized the same script for visualization to see the inferred cell usage (see Methods). Looking at the inference, there is clearly GEP usages for the same GEPs as the Human Lifetime dataset displayed on our Post Transplant dataset (Fig. 4A, 4B, 4C).

Unlike the previous runs, since we are running starCAT on an entirely new dataset, there are no comparable visualizations or runs for validation of starCAT's analysis. However, we do plan on verifying these results post analyses, with more biological depth (see Discussion).

After retrieving these results, we see that starCAT is a viable option for inferring GEP usage scores from already known datasets (GEP contribution scores, reference dataset) by "projecting" known GEPs onto an unknown dataset (assuming biological tangibility). By utilizing starCAT for inferences, we can minimize the loss of GEPs and information from *de novo* explorations, like cNMF, when doing analyses comparing the usage of set GEPs. Moreover, as an analytical method, starCAT is an accurate method (given a baseline reference dataset)[1] for the inference of GEPs usage scores, given the inputted datasets.

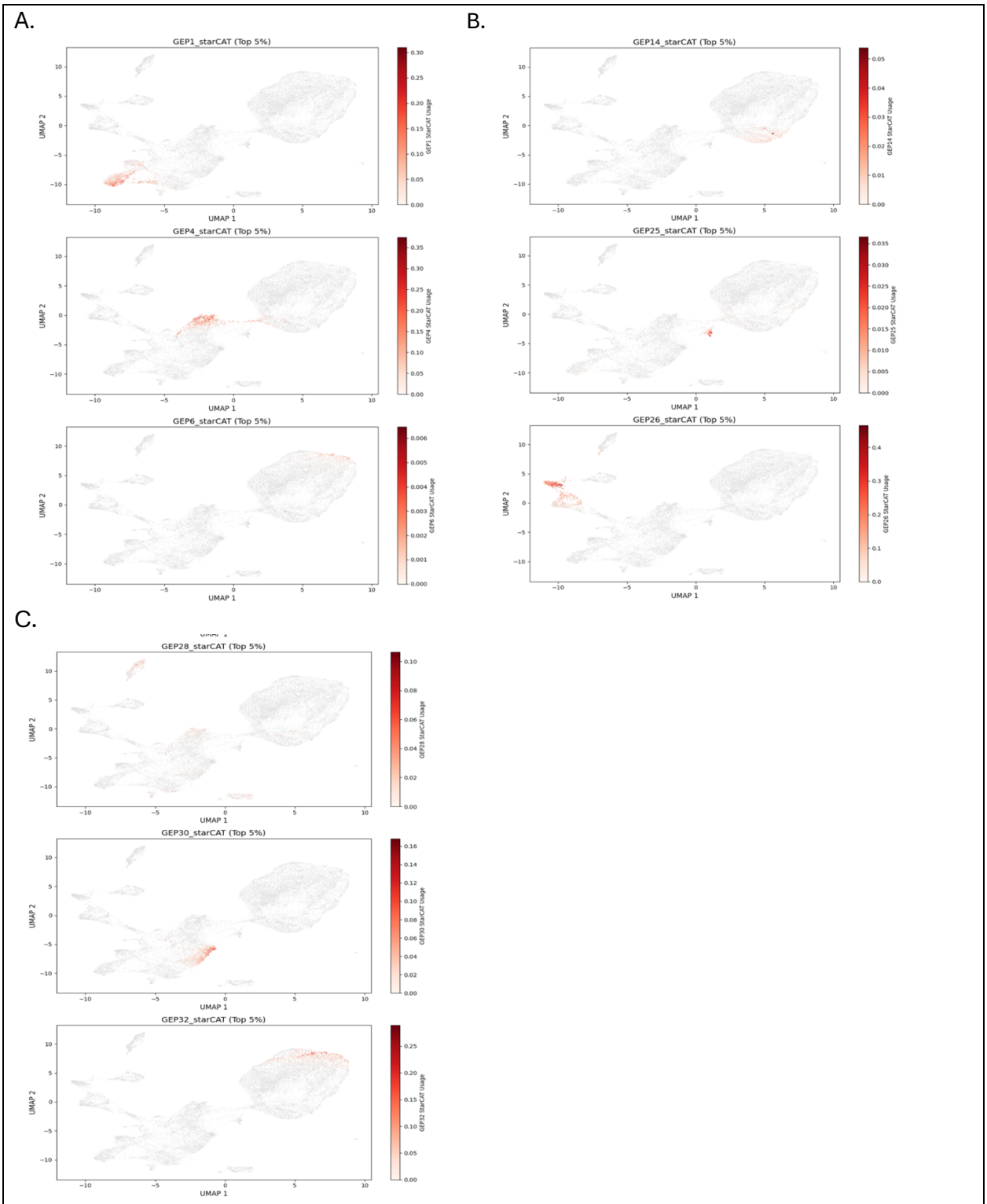


Fig 4. Post Transplant UMAP of Top 5% Cells with GEP usage after starCAT inference **(A)** UMAPs for the top 5% cells of GEP usages for GEPs 1, 4, and 6. **(B)** UMAPs for the top 5% cells of GEP usages for GEPs 14, 25, and 26. **(C)** UMAPs for the top 5% cells of GEP usages for GEPs 28, 30, and 32.

Discussion

starCAT Future Direction with Dataset Analysis

With our analysis, we have seen that starCAT is applicable to other datasets (particularly HSC analysis), aside from its intended purpose of T-cell expression. With this information, it can be inferred that the starCAT implementation can be used on other datasets outside of the HSC and T-cell analysis field and applied more broadly. Similarly to machine learning transfer learning, starCAT can use previously known data (reference GEP contribution scores) to analyze various datasets, assuming that the GEPs are present and is conducted on a tangible analysis.

Further Exploring starCAT results with GSEA

When looking at the results of starCAT's inference on the Human Lifetime dataset and the Post Transplant dataset, a numerical analysis shows that they are valid results (as shown in the correlation plots). However, it is still undetermined whether the GEP usage scores for the Post Transplant dataset are biologically sound. Using our usage scores, we hope to further analyze and confirm our findings with additional processing and analytical steps, particularly with gene set enrichment analyses.

Exploring Activity Programs with starCAT Analysis

As mentioned in cNMF's implementation, cells should be able to express both activity and identity programs[1]. As a result, cNMF's implementation reflects this through its ability to explore these GEPs as well as generate GEP usage scores per cell and gene contribution scores for these GEPs accordingly. As a direct result of this, starCAT's implementation should do the same due to its usage of the cNMF algorithm to help with its inference when left with the lack of contribution scores.

Despite knowing this, our analysis only included identity programs, rather than a mixture of both activity and identity programs. As a result, we could not test this implementation, as it was outside of our initial tests for inference with known identity programs. However, after testing and seeing starCAT's ability to infer contribution and usage scores from a catalog of GEPs accurately (according to our correlation plots based on starCAT's implementation[1]), we will investigate starCAT's ability to infer activity programs, as well a mixture of both GEP types.

Issues with Non-Negative Matrices in starCAT and Future Steps

Focusing on the computational side of the starCAT implementation, there was a major issue within our preprocessing that affected the entire analysis, related directly to starCAT's main framework. When using starCAT, there are strict matrix input requirements, particularly a non-negative matrix. This seemed to be a particular hindrance in our inference runs of the GEP usage scores (especially within our analysis), especially with our primary runs of starCAT on the Human Lifetime Dataset. Specifically, this was an issue with the gene contribution score by GEP matrix.

When initially running starCAT, we had a matrix of around 40,000 gene contribution raw scores by 35 GEPs, providing a sizeable reference for inferring our GEP usage score matrix. However, when running this, we found out that some of the values were negative, possibly meaning that there was an inverse relationship. This information would have been useful within our analysis; however, it would not be viable within the starCAT analysis due to containing negative values.

As a result, we had to switch to a much smaller matrix of around 2,000 gene contribution normalized scores by 35 GEPs for the algorithm to work. When we switched to this dataset, the analysis worked as intended, as starCAT outputted the according GEP usage scores matrix. Despite the algorithm working, the inference may have been limited in a reference to properly generate a GEP usage matrix, resulting in a much smaller matrix. In addition to having a smaller matrix, we lose vital information due to normalization of the dataset. As mentioned earlier, one of these factors was the removal of negative values which may have been significant within the resulting matrix. Knowing this, although our analysis shows the proof of concept that starCAT may accurately infer usage scores (according to correlation values[1]), a further analysis with an entirely non-negative, raw, contribution score matrix is required.

Data Availability

1. Post-Transplant CD34 Analysis Repo (https://github.com/Tofulati/postTrans_cd34)
2. Human Lifetime CD34 Analysis Repo (https://github.com/Tofulati/cd34hult)
3. starCAT implementation Repo (https://github.com/immunogenomics/starCAT)
4. Findings Presentation (https://docs.google.com/presentation/d/1lji7b9QSI5MWvcF9yKHkQUo6SpxomV7ICWVtfGnZvJA/edit?usp=sharing)

Acknowledgements

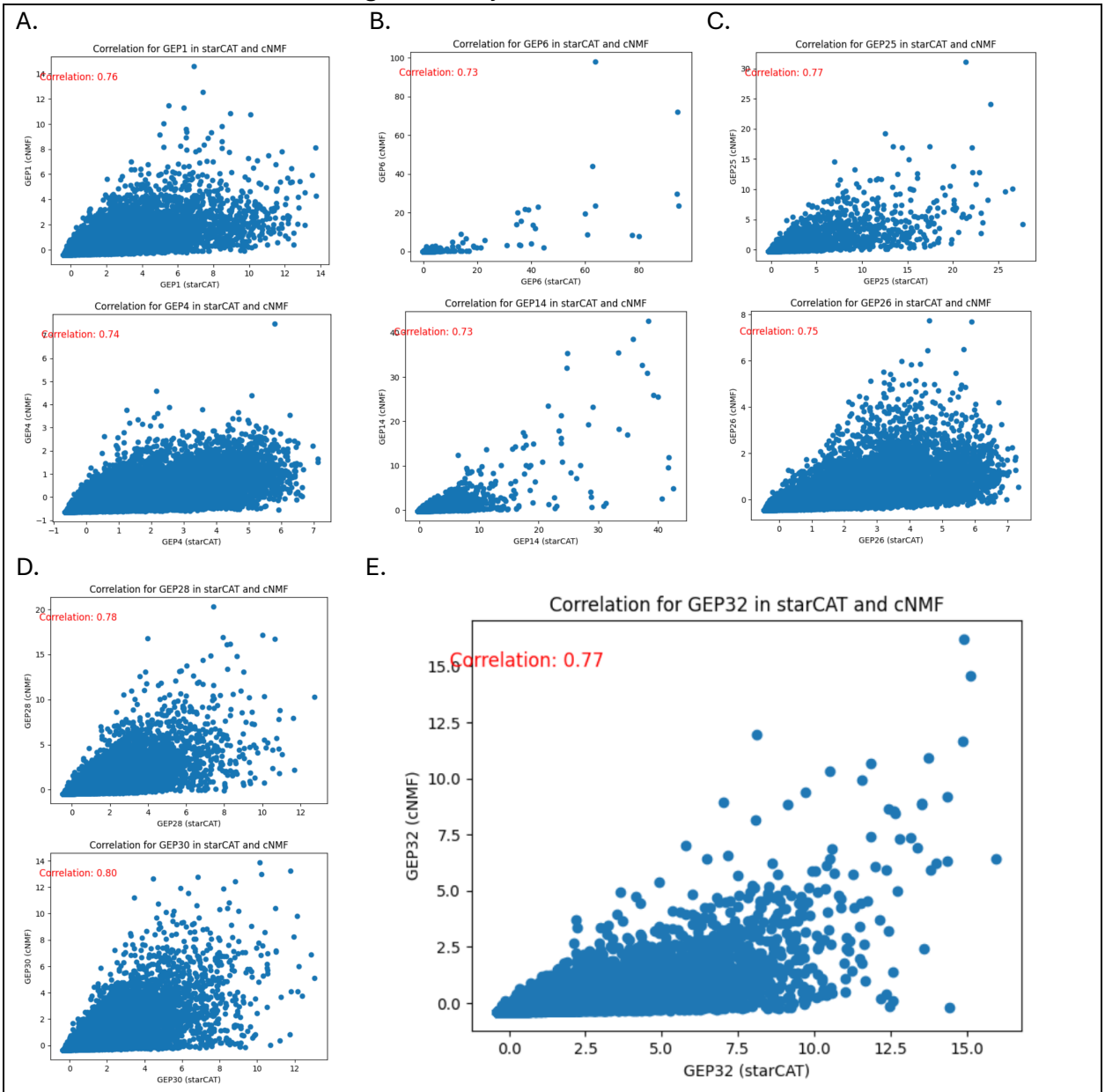
I would like to thank the URS Philip and Elizabeth Hiestand Scholarship for Engineering and/or SIO majors for providing me with the opportunity to continue my research this summer. In addition, I would like to thank the URS program for facilitating this year's summer programs and networking events. Finally, I would like to thank Dr. Hojun Li and Nick Chan from the Hojun Li Lab for supporting and helping with the research direction for my project, as well providing a space for me to conduct my research.

References

1. Kotliar, D. Reproducible single cell annotation of programs underlying T-cell subsets, activation states, and functions (2024).
2. Khaddour, K. Hematopoietic Stem Cell Transplantation (2023).
3. Chan, N. Integrative and Longitudinal Analysis of Hematopoietic Reconstitution in Transplant Recipients (2025).
4. Sidney, L. Concise Review: Evidence for CD34 as a Common Marker for Diverse Progenitors. Stem Cells (2014).
5. Kotliar, D. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq (2019).
6. Li, H. The dynamics of hematopoiesis over the human lifespan (2024).
7. Loughran, S. Lineage commitment of hematopoietic stem cells and progenitors: insights from recent single cell and lineage tracing technologies (2020).
8. Radu, Petru. CD34—Structure, Functions and Relationship with Cancer Stem Cells (2023).
9. Ma, S. Principal component analysis based methods in bioinformatics studies (2011).

Supplementary Figures

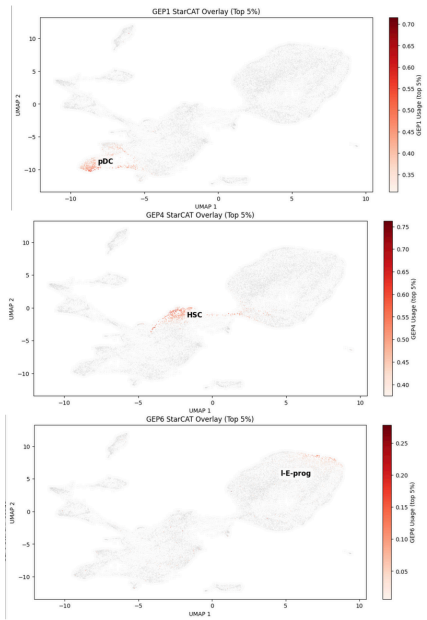
1. Human Lifetime GEP usage scores by cells starCAT and cNMF correlation chart



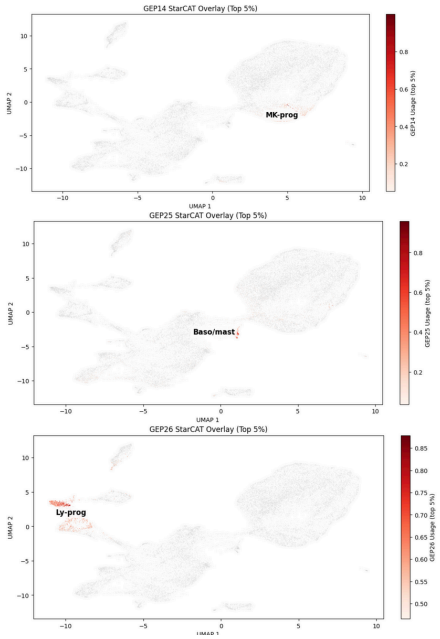
(A) Pearson correlation plot and value for GEP 1 and 4 between starCAT inference and cNMF exploration GEP usage scores. **(B)** Pearson correlation plot and value for GEP 6 and 14 between starCAT inference and cNMF exploration GEP usage scores. **(C)** Pearson correlation plot and value for GEP 25 and 26 between starCAT inference and cNMF exploration GEP usage scores. **(D)** Pearson correlation plot and value for GEP 28 and 30 between starCAT inference and cNMF exploration GEP usage scores. **(E)** Pearson correlation plot and value for GEP 32 between starCAT inference and cNMF exploration GEP usage scores.

2. Post Transplant dataset UMAP plots with cell type and lineages marked

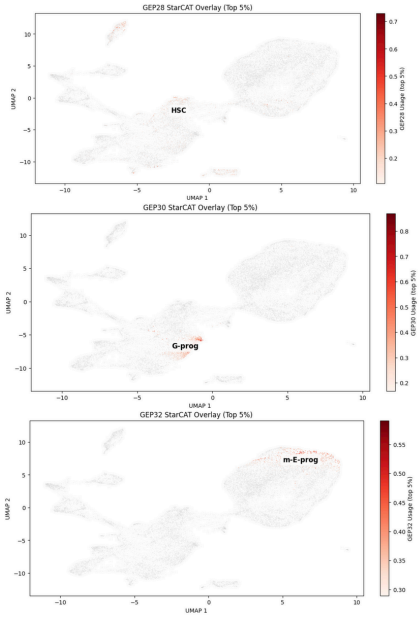
A.



B.



C.



(A) UMAP plot of top 5% cells in GEP usage scoring from starCAT inference. Each of the cell clusters are labeled with the cell type and lineage they are a part of. This chart shows GEPs 1, 4, and 6 accordingly. **(B)** UMAP plot of top 5% cells in GEP usage scoring from starCAT inference. Each of the cell clusters are labeled with the cell type and lineage they are a part of. This chart shows GEPs 14, 25, and 26 accordingly. **(C)** UMAP plot of top 5% cells in GEP usage scoring from starCAT inference. Each of the cell clusters are labeled with the cell type and lineage they are a part of. This chart shows GEPs 28, 30, and 32 accordingly.