UNIVERSITY of WASHINGTON

LCL: Lineage-aware Contrastive learning for scRNA-seq data

Shizhao Joshua Yang, Yixin Wang, Kevin Lin

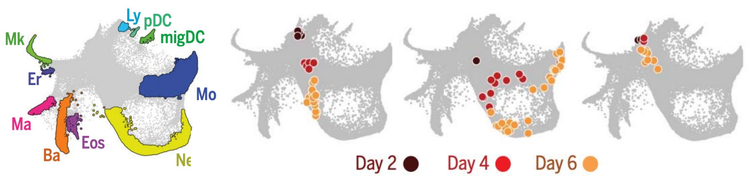
BIOSTATISTICS
SCHOOL of PUBLIC HEALTH

Motivation

Understanding the history, state, and fate of cells is crucial in biology. **Single-cell lineage tracing (scLT)** technology, using inheritable barcodes, enables the tracking of progenitor and stem cell development, providing insights into gene differential expression (DE) and cell lineage relationships.

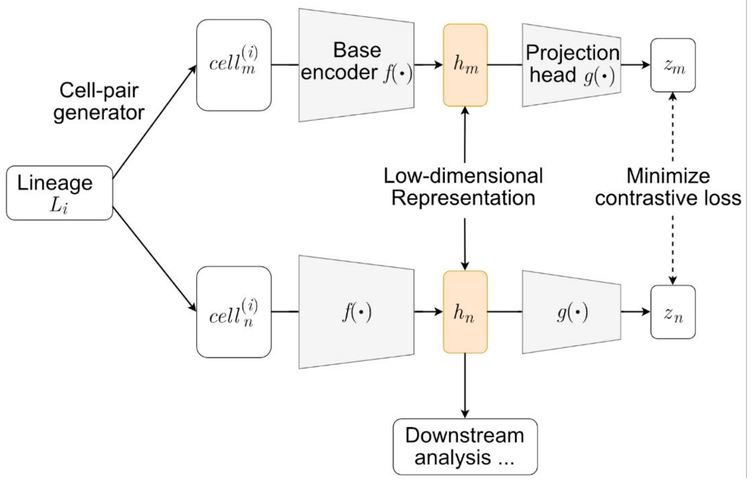
Weinreb et al. (2020) applied lineage barcodes to learn about the differentiation of hematopoietic progenitor cells (hpc). This dataset revealed distinct differences in cell type differentiation behavior across various lineages, including single-fate and multiple-fate (pluripotent) differentiation, as well as the self-renewal of early progenitors. Based on these observations, **a further question arises: Can we learn fate commitment in single-cell lineage-tracing data and uncover lineage-specific gene pathways?**

The challenge lies in isolating lineage-specific signals, which are often overshadowed by more dominant signals such as cell type variations. Traditional methods like variational autoencoders (VAEs) excel at reconstructing gene expression but often miss these crucial lineage signals. To address this, we employ a contrastive learning framework tailored to lineage-barcoded scRNA-seq data. This approach leverages the inherent supervision provided by lineage barcodes, making contrastive learning a powerful tool for capturing and analyzing lineage-specific signals across different time points and states.



Framework

Lineage-aware Contrastive Learning (LCL) framework consists of three main components:**cell-pair generator**, **base encoder**, and **projection head**.



B) scLT: Gene expression, lineage, and time

	Genes
Cell 1	Log-normalized gene expressions
Cell 2	
Cell 3	
Cell 4	
Cell 5	

Lineage	Time point	Cell type
Lineage 1	Day 4	Stem
Lineage 1	Day 12	Terminal 1
Lineage 2	Day 4	Stem
Lineage 2	Day 12	Terminal 1
Lineage 2	Day 12	Terminal 2

Cell-pair Generator: Define a positive pair if two cells are from the same lineage. Given a single-cell RNA-seq dataset with *p*=2000 genes with corresponding cell lineage information , each batch select *N* cell pairs from *N* different lineages. Cells from different lineages are considered as negative pairs.

Base Encoder f(): a 3-layer Multilayer Perceptron (MLP) that generates cell representations h from gene expression vectors. Each MLP layer uses ReLU activation and Batch Normalization, which helps stabilize and accelerate training. The input dimension is p (corresponding to the 2000 genes), and the output dimension is 64.

Projection Head g(): consists of a two-layer MLP with ReLU that transforms representations *h* into the space utilized for contrastive loss, denoted by *z*.

Loss Function: Suppose for each batch we have *N* cell pairs (*2N* cells in total), we define the loss function for a positive cell pair (*m*, *n*) as:

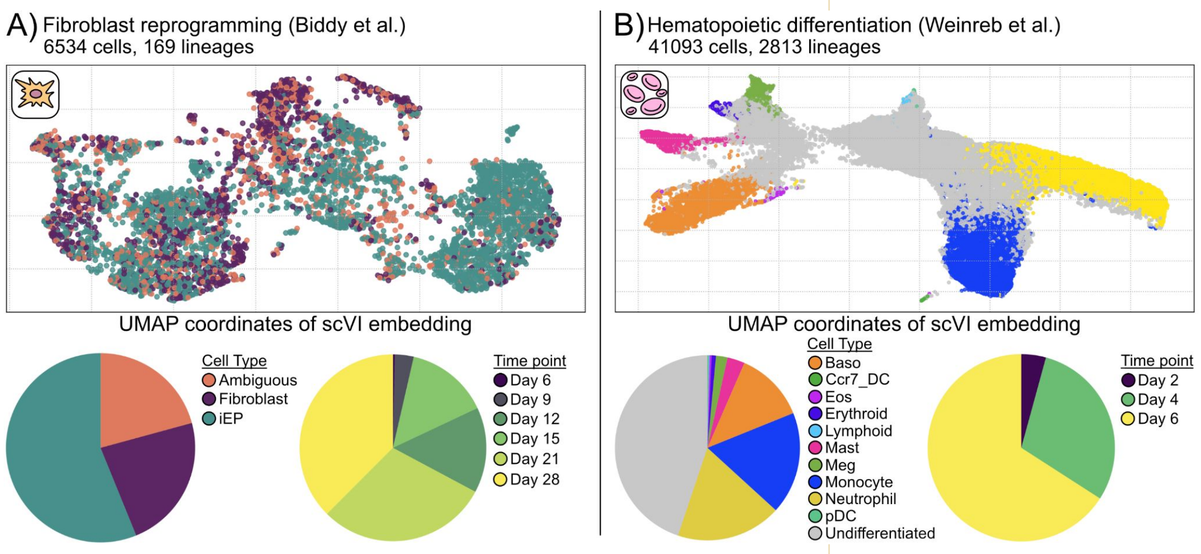
$$loss_{m,n} = -\log\left(\frac{\exp(\mathrm{sim}(z_m, z_n)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq m]} \exp(\mathrm{sim}(z_m, z_k)/\tau)}\right)$$

where $\mathrm{sim}(z_m, z_n)$ is the cosine similarity calculated as $(z_m \cdot z_n)/(\|z_m\|_2 \cdot \|z_n\|_2)$, τ is the temperature parameter. Here, $\mathbb{1}_{[k\neq m]}$ in the denominator part is an indicator function that denotes if two cells come from the same lineage. The total loss function then su$loss_{m,n}$ over all *2N* positive pairs in the batch. **The algorithm therefore aims to maximize the agreement between cells from the same lineage and minimize the agreement between negative pairs.**

Dataset

Biddy dataset, generated from experiments on reprogramming mouse embryonic fibroblasts (MEFs) into induced endoderm progenitors (iEPs)

Weinreb dataset, generated from experiments on the differentiation of mouse hematopoietic progenitor cells (HPCs)

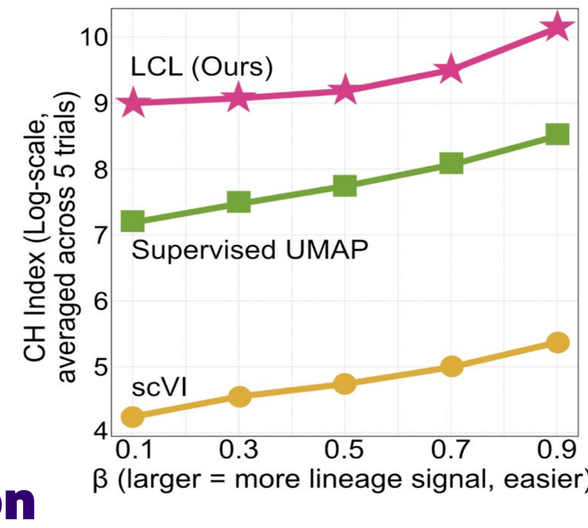


Simulation study

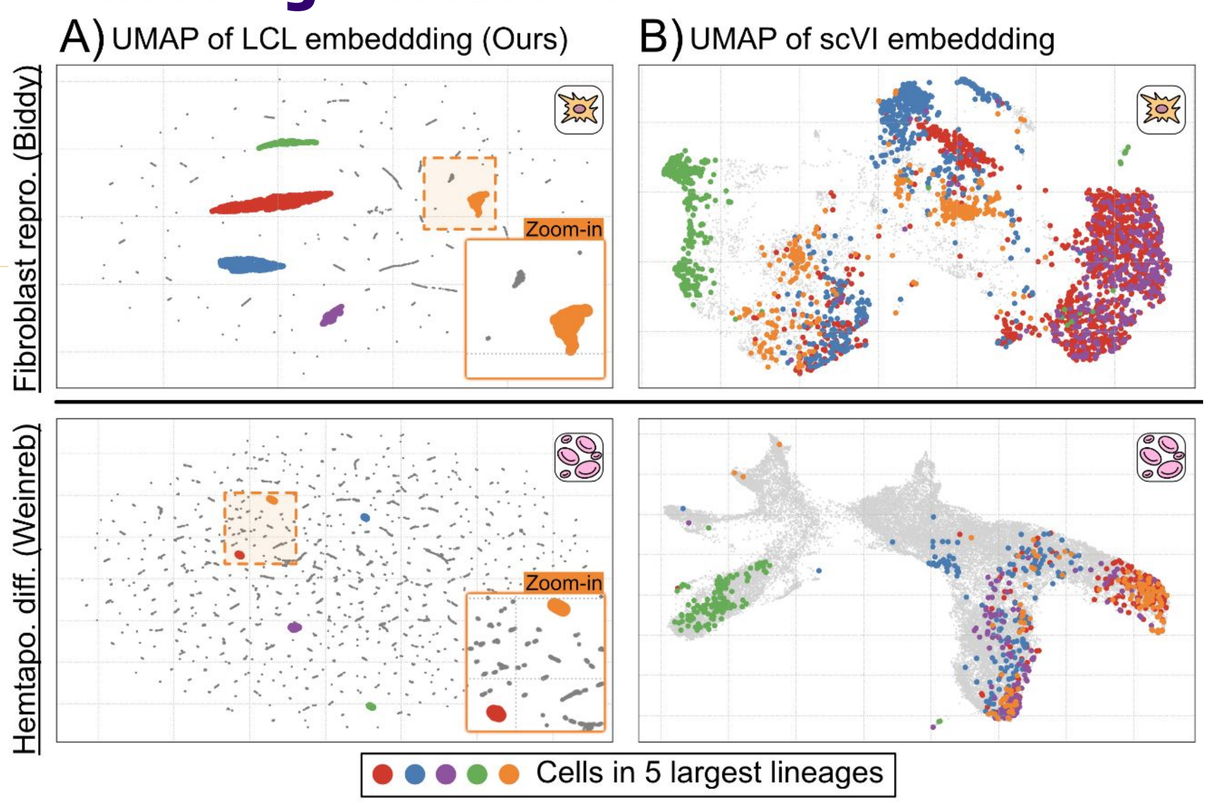
Comparison on pseudo-real data

To test our method's consistency across datasets of varying difficulty, we generated simulation datasets by reassigning lineage information to cells from the Weinreb dataset.

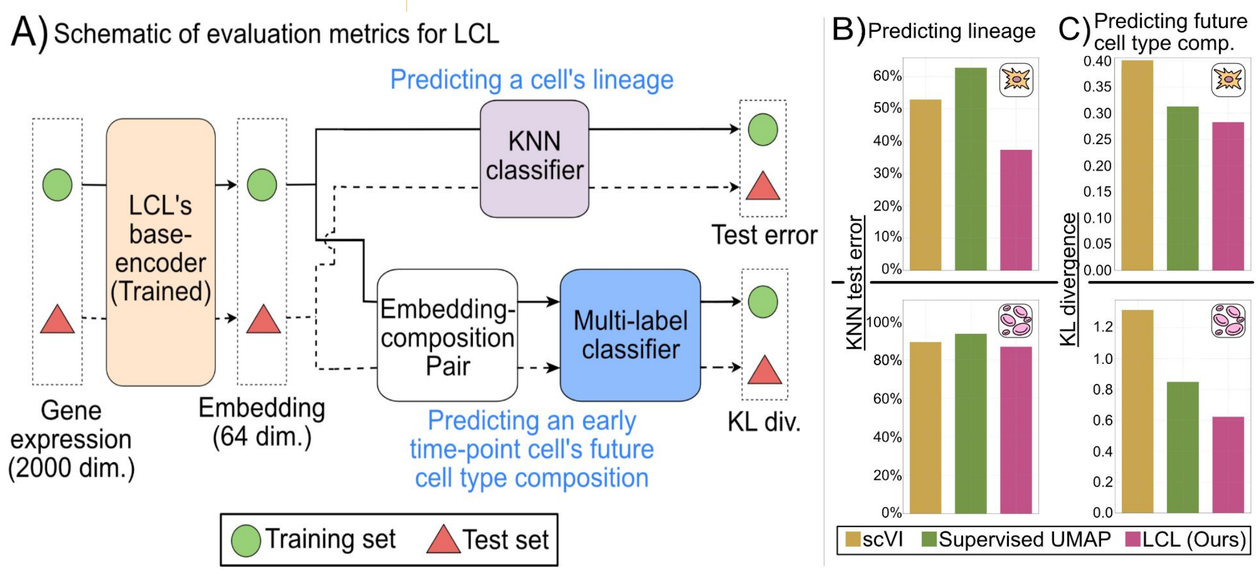
Calinski-Harabasz Index (CHI) of different betas comparing how LCL compares with scVI and Supervised UMAP, where a higher index means the embedding better separates the different lineages.



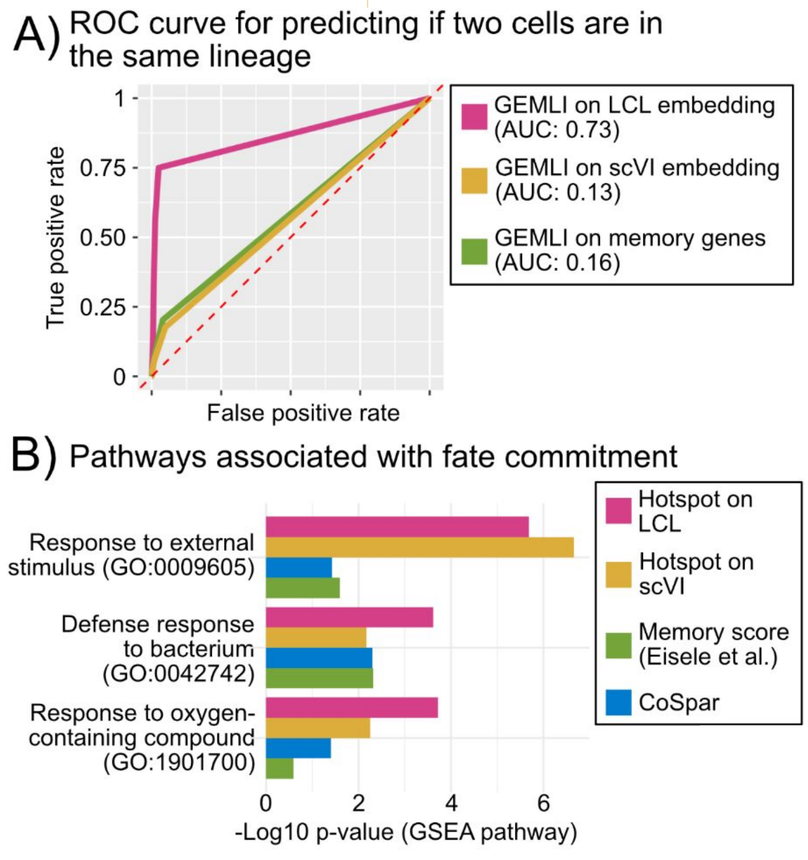
Embedding visualization



Lineage Predicting and Compositional Analysis



Biological Insight: LCL enables identification of fate-determining genes



GEMLI analysis:
ROC of three different inputs, each originating from a different method, for GEMLI to predict if two cells are from the same lineage, with the AUC denoted for each method.
memory genes (generated by GEMLI), which are genes defined to have a small coefficient of variation within a lineage compared to between different lineages

Hotspot analysis:
Gene Pathway that associated with fate commitment: Three example pathways, each scored using results from four different methods via Gene Set Enrichment Analysis (GSEA). All three pathways shown are statistically significant for LCL after multiple testing, but only scVI's pathway for external stimulus is statistically significant.

Main reference

- Yang, S. J., Wang, Y., & Lin, K. Z. (2024). LCL: Contrastive learning for lineage barcoded scRNA-seq data. *bioRxiv*. https://doi.org/10.1101/2024.10.28.620670
- Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science (New York, N.Y.), 367(6479), eaaw3381. https://doi.org/10.1126/science.aaw3381
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. International Conference on Machine Learning pp. 1597–1607 (2020)
- Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., Morris, S.A.: Single-cell mapping of lineage and identity in direct reprogramming. Nature 564(7735), 219–224 (2018)