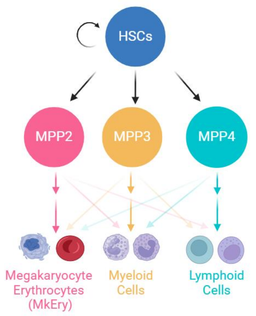
Comparative Analysis of scRNA-seq Annotation Pipelines for Hematopoietic Stem and Progenitor Cells

ARYAHI DEORUKHKAR[1,2], Jennifer J. Chia[3,4,5], Alexander Hoffmann[4,6,7]

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA, [2] Department of Biochemistry, School of Medicine, Case Western University, Cleveland, OH, [3] Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, [4] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, [5] Broad Stem Cell Research Center, University of California, Los Angeles, CA, [6] Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, CA, [7] Institute for Quantitative and Computational Biosciences (QCB), University of California, Los Angeles, CA

BACKGROUND



- Hematopoiesis is a continuous process that gives rise to all blood cells from the bone marrow (BM).
- Single-cell RNA sequencing (scRNA-seq) is useful to study this inherently single cell process.
- Hematopoietic stem cells (HSCs) are at the top of the differentiation hierarchy, followed by CD34+ progenitors with biases toward megakaryocyte/erythrocyte (MkEry), myeloid, and lymphoid lineages.
- The transcriptomic signatures hematopoietic stem and progenitor cells (HSPCs) are similar given their multipotent nature, making them challenging to classify in scRNA-seq datasets.

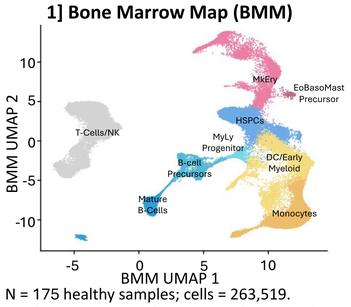
Research Question:

How similar are the HSPC label outputs of an in-lab scRNA-seq classifier designed for HSPCs compared to a publicly available classifier built for total bone marrow?

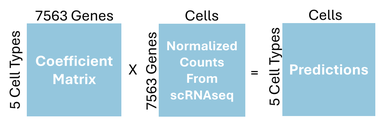
METHODS

Two classifiers for annotating bone marrow scRNA-seq data:

1] Bone Marrow Map (BMM)

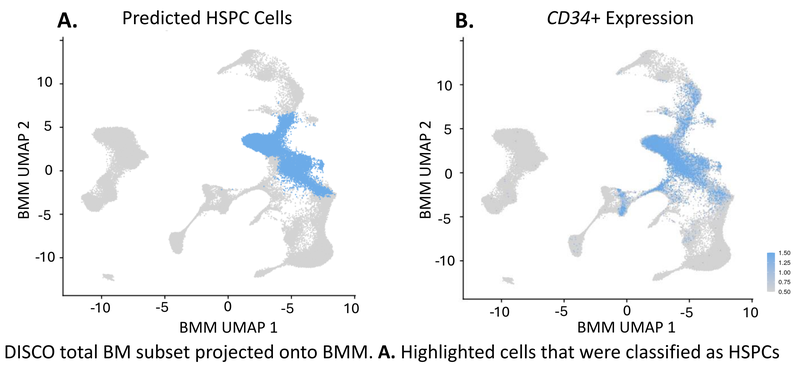


2] In-Lab Pipeline



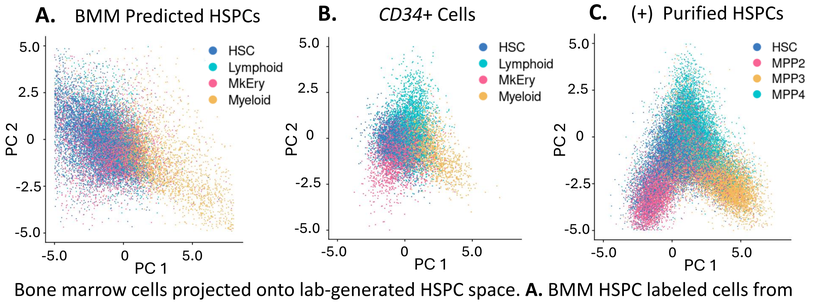
- N = 175 healthy samples; cells = 263,519.
- Batch corrected with harmony and projected using symphony.
- Annotates full bone marrow.
- Cell types were collapsed to lineages for fair comparisons.
- **Subset used**: DISCO BM ATLAS (Yee et al., 2023) 89,268 cells by downsampling T cells, monocytes, and stromal cells.
- Microarray data (Pietras et al., 2015) from purified HSPCs was used to create a coefficient matrix.
- Matrix multiplication can yield cell type predictions per cell.
- Annotates HSPCs only.

BMM HSPC Labels do not Distinguish Mature and Immature Cells Well



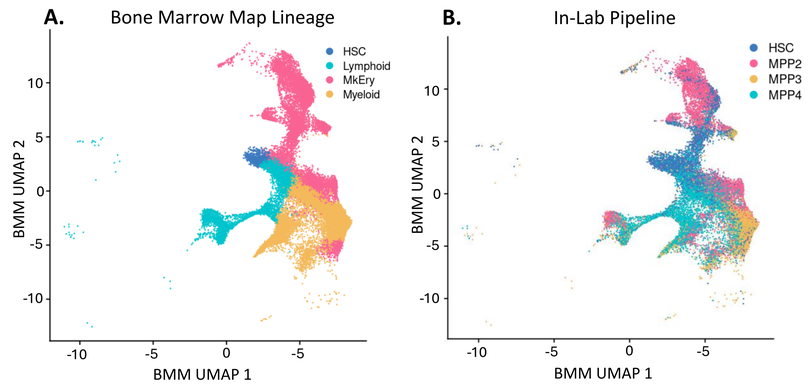
DISCO total BM subset projected onto BMM. **A.** Highlighted cells that were classified as HSPCs using BMM. **B.** *CD34*+ expression across subset.

CD34+ Cells Align with HSPC PCA Space Despite BMM Labeling as Mature

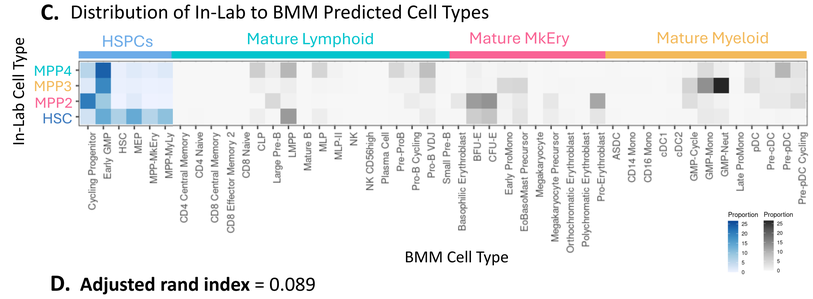


Bone marrow cells projected onto lab-generated HSPC space. **A.** BMM HSPC labeled cells from DISCO subset are dispersed. **B.** *CD34*+ cells in the DISCO subset follow the expected curvature of HSPC cells. **C.** Positive control of purified HSPC cells from Ainciburu et al., 2023.

Purified HSPCs are Classified as Mature Cell Types by BMM and Discrepancies Between Classifiers Lineage Labeling is Observed



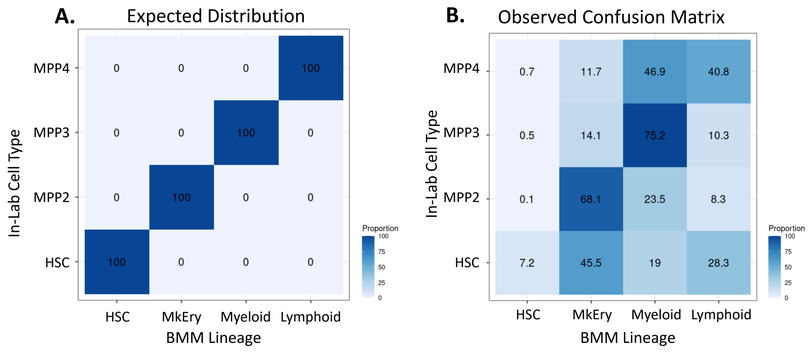
C. Distribution of In-Lab to BMM Predicted Cell Types



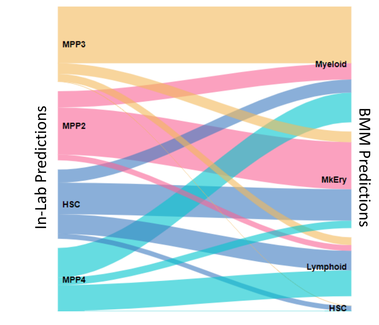
D. Adjusted rand index = 0.089

Labeling of Ainciburu et al., 2023 purified HSPCs using BMM and In-Lab classifiers. The data was projected onto the BMM UMAP space with **A.** BMM lineage labels and **B.** In-Lab classifier labels. Discrepancies in cell type assignment were observed. **C.** Heatmap of In-Lab cell type assignment across BMM full-diversity cell type labels. **D.** Adjusted rand index computed for cell type comparison.

Classifiers Recognize Lineages Similarly Among Purified HSPCs, but have Lower Correspondence with HSCs



C. Sankey Diagram of Observed Distribution



D. Precision Across Cell Types

Measure	Positive Prediction Value
HSC/HSC	0.072
MPP2/MkEry	0.681
MPP3/Myeloid	0.752
MPP4/Lymphoid	0.408

E. Computed overall statistics

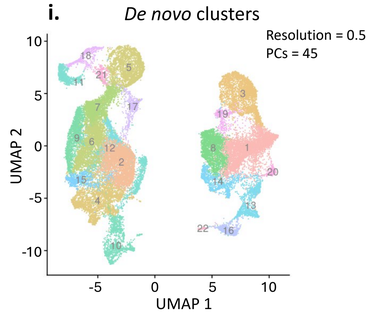
Measure	Value	Excluding HSCs
Accuracy (95% CI)	(0.481,0.491)	(0.620, 0.632)
Adjusted Rand Index	0.154	0.206
Cohen's Kappa	0.310	0.431

Expected correspondence of lineage labels. Observed correspondence through **B.** Confusion matrix and, **C.** Sankey diagrams. Statistics computed using the confusion matrix were calculated: **D.** Precision across cell types and, **E.** Overall computed statistics.

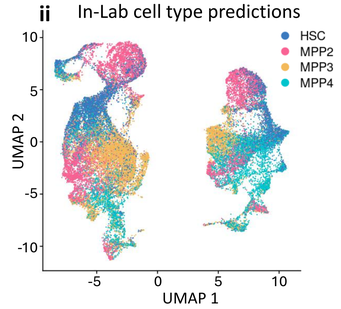
HSPC Labels Correspond to *de novo* Clusters Similarly for Both In-Lab and BMM Classifiers

A. All Purified HSPCs Projected onto *de novo* UMAP Space

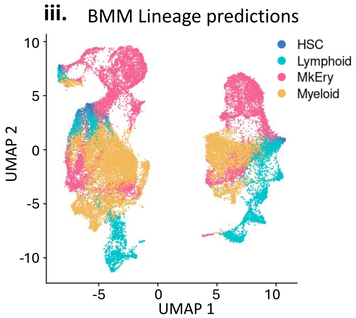
i. *De novo* clusters



ii In-Lab cell type predictions



iii. BMM Lineage predictions

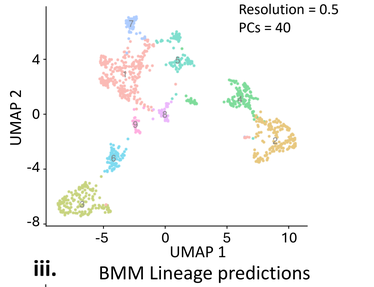


iv. Adjusted Rand Index (ARI)

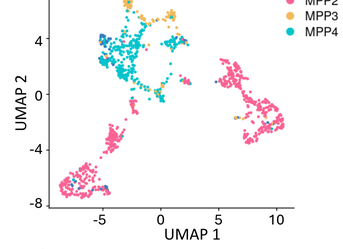
Classifier	ARI
Bone Marrow Map Lineage	0.111
In-Lab Cell Type	0.093

B. Most Easily Classifiable Purified HSPCs (selected by *CD34*+, high confidence in-lab labeling) Projected onto *de novo* UMAP Space

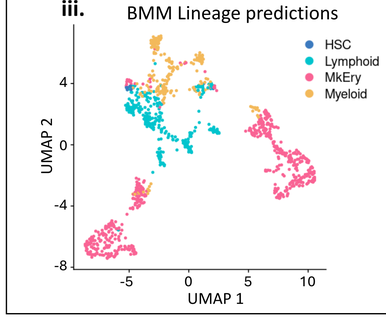
i. *De novo* clusters



ii. In-Lab cell type predictions



iii. BMM Lineage predictions



iv. Adjusted Rand Index (ARI)

Classifier	ARI
Bone Marrow Map Lineage	0.312
In-Lab Cell Type	0.450

CONCLUSION & DISCUSSION

- Bone Marrow Map does not entirely distinguish immature cells from mature cells.
- Even in purified HSPCs, BMM calls several mature cell types.
- Analyzing purified HSPC data reveals some correspondence between lineage assignment between classifiers, particularly for MPP2s and MPP3s.
- HSC classifications are less well conserved between classifiers.
- Matching data diversity to the cell label classifier is likely to yield better cell type predictions.
- Resolving HSPCs is more challenging than mature cell type labels and benefits from a specialized labeling classifier.

References

Ainciburu, M., et al. (2023). Uncovering perturbations in human hematopoiesis associated with healthy aging and myeloid malignancies at single-cell resolution. *ELife*, 12, e79363. https://doi.org/10.7554/eLife.79363

Andy G.X. Zeng, et al. (2023). Precise single-cell transcriptomic mapping of normal and leukemic cell states reveals unconventional lineage priming in acute myeloid leukemia. *BioRxiv* (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2023.12.26.573390

Pietras, Eric M et al. (2015). Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell*, 17(1), 35–46. https://doi.org/10.1016/j.stem.2015.05.003

Yee, N., et al. (2023). Establishing a human bone marrow single cell reference atlas to study ageing and diseases. *Frontiers in Immunology*, 14. https://doi.org/10.3389/fimmu.2023.1127879

Acknowledgements

I am deeply grateful to Dr. Alexander Hoffmann and Dr. Jennifer Chia for welcoming me into the Hoffmann Lab this summer, helping me find a meaningful project, and providing great mentorship. Thank you, David Mastro, for orienting me to the lab and the project, answering my questions, and being a good friend. I would also like to acknowledge Dr. Apeksha Singh, Dr. Chia, and all the other lab members who contributed to developing the in-lab HSPC labeling pipeline.