Integrative Single-Cell Genomic Analysis of Mesenchymal Stem Cells

Joshua Sodicoff[1], Yuki Matsushita[2], Wanida Ono[2], Joshua Welch[3,4], Noriaki Ono[2]

[1]Department of Biomedical Engineering - University of Michigan College of Engineering, Ann Arbor, MI, [2]Department of Orthodontics and Pediatric Dentistry - University of Michigan School of Dentistry, Ann Arbor, MI, [3]Department of Computational Medicine and Bioinformatics - University of Michigan Medical School, Ann Arbor, MI, [4]Department of Computer Science and Engineering - University of Michigan College of Engineering, Ann Arbor, MI

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Introduction

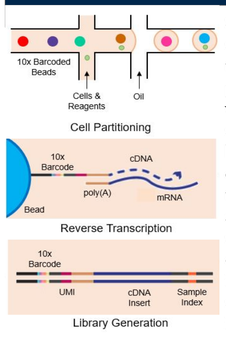
Defining the molecular features that identify cell types remains a challenge in biology. Whereas other genomic methods take average expression values across a sample of cells, single-cell analysis preserves cell-type differences, allowing unbiased cell type discovery. However, single-cell experiments can generally measure only one type of molecular modality per cell. Computational integration provides a way to leverage multiple single-cell modalities for cell type discovery and investigate the regulatory relationships between the transcriptome and epigenome.

This project applies integrative single cell genomics to mesenchymal stem cells (MSC), a multipotent cell found in bone marrow. The study of MSCs will advance understanding of the mechanisms of bone formation and regeneration. Because the bone marrow contains a heterogeneous mixture of cell types, including MSCs, single-cell analysis.

We used scRNA-seq and scATAC-seq data from mouse bones, collected and processed by the Ono laboratory, to jointly define cell types. We first compared multiple approaches for calculating expression-like features from scATAC-seq and performed data cleaning and imputation data to account for technical artifacts and data sparsity. We then used LIGER, an algorithm we previously developed, to integrate the RNA and ATAC data. This allowed us to define, for the first time, mouse skeletal stem cell types using both transcriptomic and epigenomic features. Our results indicate strong concordance between transcriptome and epigenome states and represent a first step toward understanding epigenomic regulation in MSCs.

Experimental Methods

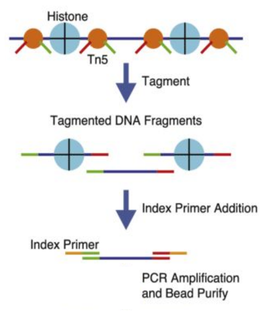
scRNA-seq



Single cell RNA sequencing enables the study of transcriptomic features of tissues and individual cells. In this process, microfluidics is used to pair cells with beads, covered in short DNA sequences, including barcodes that will be used to identify the cell of origin during sequencing. Base pairing between the bead's DNA strands and the mRNA and reverse transcription creates a DNA sequence that contains information that can be used to identify both genetic information and its source. The raw sequencing data obtained is processed into a cell by gene matrix.

Figure 1: scRNA-sequencing can be broken down into several steps, namely the partitioning of cells to create a one-to-one association between cell and barcode, reverse transcription for cDNA, and library generation, which prepares nucleic acid strands for sequencing.
Credit: Genewiz

scATAC-seq

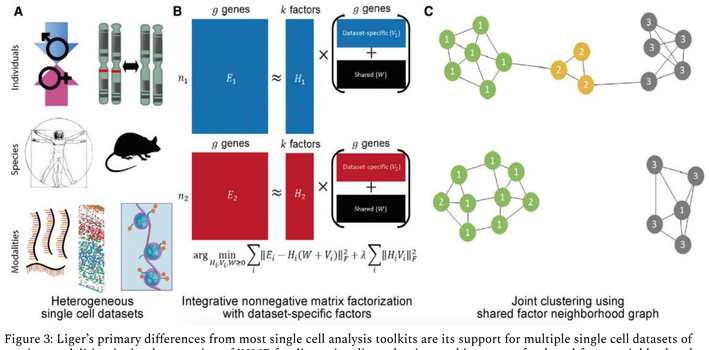


Single cell ATAC sequencing profiles open regions of chromatin to define epigenomic features. Although the method shares similarities with scRNA-seq, primarily in partitioning and sequencing, it relies on the fragmentation and tagging of chromatin, whereas scRNA-seq relies on previously transcribed mRNA. Because the exposed regions of chromatin may not correlate directly with genes, cells may be profiled by "peaks", or regions of accessible chromatin.

Figure 2: scATAC-sequencing relies on the hyperactive Tn5 transposase to bind to open regions of chromatin and inserts sequencing adapters. This process, also known as "tagmentation" is more efficient than other methods that require tagging after fragmentation.
Credit: Chaitankar et al.

Computational Methods

LIGER (linked inference of genomic experimental relationships) performs integrative nonnegative matrix factorization (iNMF) to jointly define "metagenes", from multiple single-cell datasets. iNMF yields interpretable dimensions that represent patterns of gene coexpression and often correspond to cell types, biological processes, or sources of technical variation. This factorization strategy differs from conventional NMF in that it learns both shared and dataset-specific metagenes and includes a penalty term to minimize the dataset-specific terms.



LIGER takes one or more cell by gene matrices, sharing a common set of gene features, as input and performs the following steps:

1. Normalization of gene expression counts across cells
2. Variable gene selection
3. Scaling of gene expression counts
4. Use of alternating least squares optimization for integrative non-negative matrix factorization
5. Creation of shared factor neighborhood graph across datasets, followed by factor normalization
6. Implementation of t-SNE or UMAP to map high-dimensional data to two dimensions

Figure 3: Liger's primary differences from most single cell analysis toolkits are its support for multiple single cell datasets of varying modalities, its implementation of iNMF for dimensionality reduction, and its usage of a shared factor neighborhood graph to jointly cluster samples.
Credit: Welch et al. Cell 2019

Results

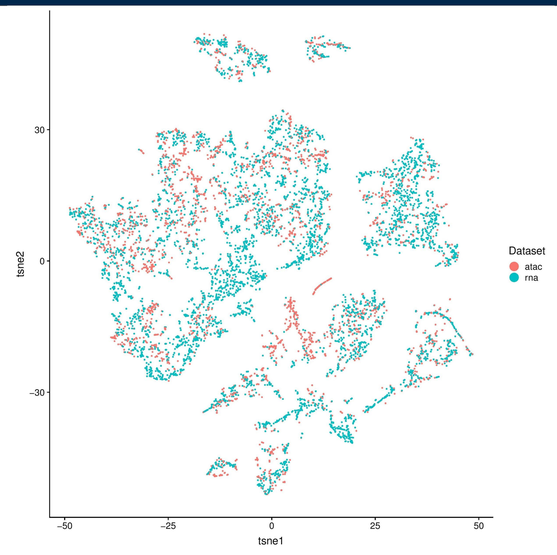


Figure 4: t-SNE plot of cells colored by modality. The near-uniform mixing of red and blue dots indicates that LIGER has successfully aligned the two datasets.

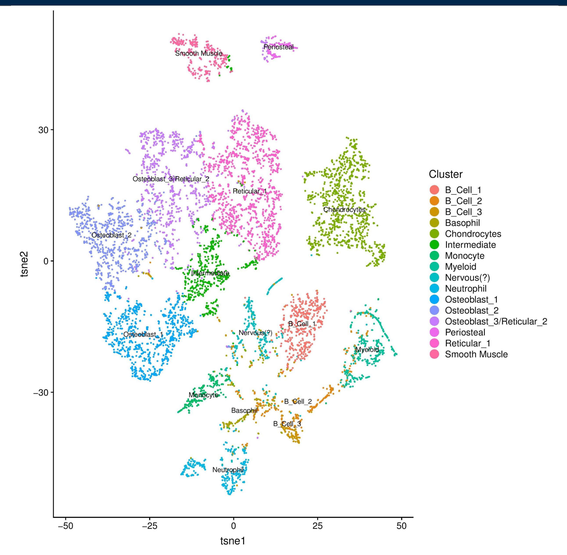
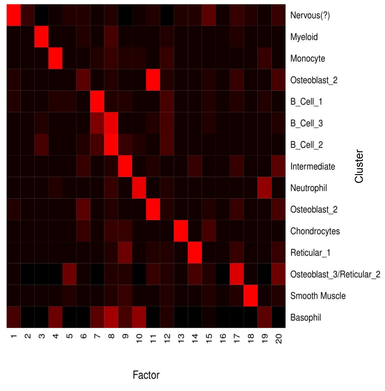


Figure 5: t-SNE plot of cells colored by cluster. Clusters were defined jointly using the LIGER joint space and labelled by a combination of known markers and gene ontology enrichment analysis.



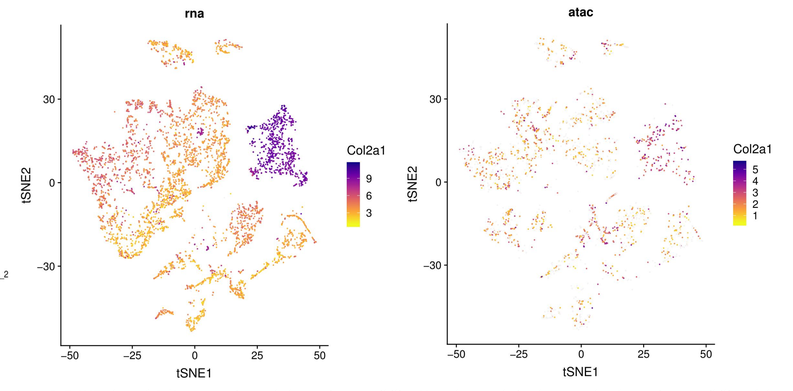


Figure 6: a heatmap showing which factors load on each cluster. Note that most factors have high loading on only a single cluster, and most clusters are primarily defined by a single factor.

Figure 7: T-SNE plots of *Col2a1* expression and chromatin accessibility. Note the correlation between the chromatin accessibility of the gene and its expression. *Col2a1* is a known marker of MSC differentiation into chondrocytes and is necessary for collagen formation.

Conclusion

Our results indicate that LIGER successfully integrates scRNA and scATAC data. This means that the jointly defined cell types may be a useful tool for connecting transcriptomic and epigenomic features. Now that we have integrated the datasets successfully and identified cell types, we will investigate the connections between the RNA and ATAC data within cell types. Further research into the molecular determinants of growth, differentiation, and regeneration within the environment of the bone promises to both expand our understanding of the subject as well as unlock new possibilities for the healing of bone injuries.

Bibliography

Fang, Rongxin, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau et al. "Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types." *bioRxiv*(2019): 615179.

Ge, Steven, and Dongmin Jung. "ShinyGO: a graphical enrichment tool for animals and plants." *bioRxiv* (2018): 315150.

Ono, Noriaki, Deepak H. Balani, and Henry M. Kronenberg. "Stem and progenitor cells in skeletal development." *Current topics in developmental biology* 133 (2019): 1.

van Dijk, David, Juozas Nainys, Roshan Sharma, Pooja Kathail, Ambrose J. Carr, Kevin R. Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. "MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data." *BioRxiv*(2017): 111591.

Welch, Joshua D., Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity." *Cell* (2019).

Special thanks to Dr. Joshua Welch and the MCubed Scholars program!