Identifying Meta-Clonal Trajectories in Lineage Tracing Datasets

Arya Kaul, Huidong Chen, Emily Butka, Samantha Morris, Luca Pinello

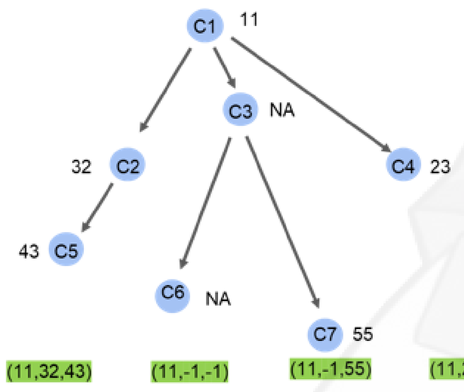
Background

Lineage tracing technologies allow biologists to observe developmental trajectories of discrete populations of cells. Exact protocol may differ, but all follow the same general technique: (1) Tag a population of cells with a heritable marker, (2) scRNA-seq a sample of these cells, (3) wait some interval of time, (4) repeat 1-3 as desired. The purpose of Megatron is to take this collection of data, and computationally determine recurrent trajectories in the data (meta-clones). Intuitively, these meta-clones describe common developmental pathways undertaken by cells within the initial population. Unbiased determination of these pathways can yield novel biological insight into the mechanisms of differentiation.

Methods

Megatron encompasses the entire computational pipeline:

1. **Experiment Agnostic File Format**: We have defined the specifications for an AnnData input file storing all relevant information for lineage tracing datasets. This universal format may be applied to any data, while conserving memory and resources.
2. **Distance Calculations**:
   i. Wasserstein – calculates 1D earth mover's distance between clones in every dimension
   ii. MNN – calculates ratio of neighbors belonging between clones
   iii. Shortest Paths of Directed Graphs – defines graph for each clone and finds shortest paths between clones
3. **Clustering**: Using pairwise distances, clones are clustered into meta-clones
4. **Evaluation**: To determine correctness, validation was performed by computing the adjusted rand score of predicted meta-clones against observed biological truth in Weinreb et al. (see figure to right)
5. **Visualization**: Constitutive cells within meta-clones are visualized, and ELPIgraph is drawn to describe overall trajectories.



Clone Representation

	(11,32,43)	(11,-1,-1)	(11,-1,55)	(11,23,-2)
C1	1	1	1	1
C2	1	0	0	0
C3	0	1	1	0
C4	0	0	0	1
C5	1	0	0	0
C6	0	1	0	0
C7	0	0	1	0

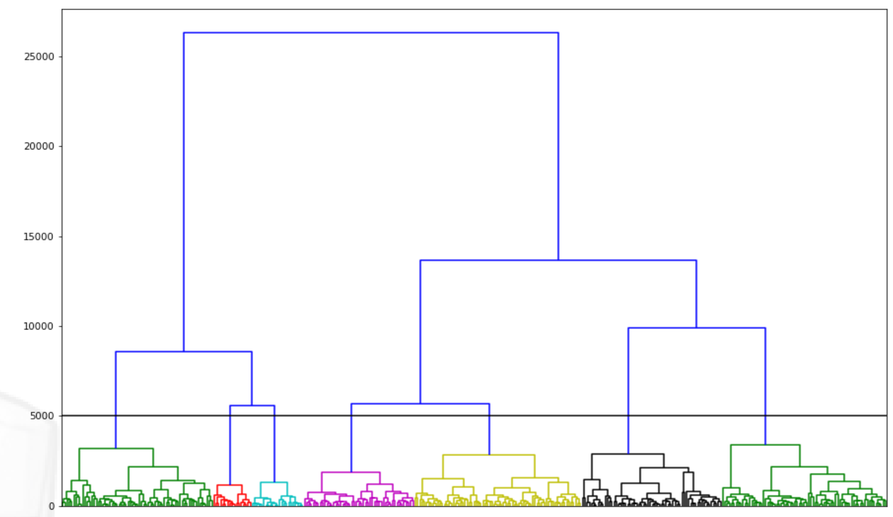
Metadata

	Timepoint
C1	Day0
C2	Day9
C3	Day6
C4	Day9
C5	Day15
C6	Day21
C7	Day28

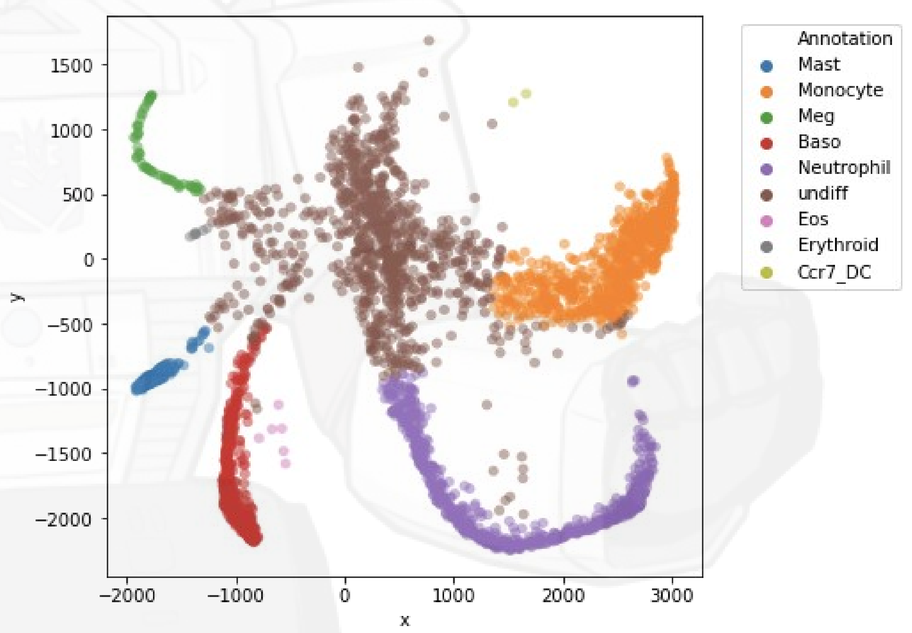
Coordinates

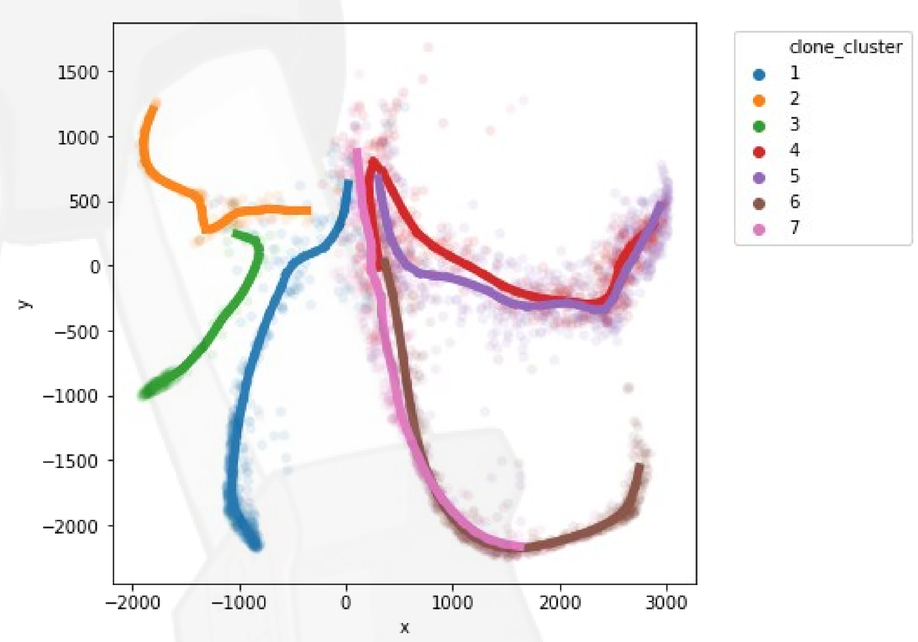
	X	Y	Z
C1	2.3	3.4	4.5
C2	1.4	5.6	6.5
C3	-0.6	0.71	6.71
C4	3.4	3.12	6.3
C5	4.5	6.8	1.22
C6	4.32	-1.33	7.2
C7	8.9	9.54	8.0

File Format: The AnnData file contains 3 linked fields. The clone representation is a sparse binary matrix with cells in rows and clonal ids in columns. Metadata includes time information, and the coordinates file shows cell location in the reduced space.



Clustering: After pairwise distances are calculated, users may specify a clustering algorithm and visualize meta-clones on a dendrogram





Evaluation: Top graph highlights annotated cell types as described in Weinreb et al. *Science* 2020. Bottom graph shows 7 meta-clones as described by Megatron's Wasserstein distance.

Results and Next Steps

We present the concept of meta-clones and describe Megatron, one of the first computational pipelines to identify meta-clones in lineage tracing datasets. Validation has yielded a maximum accuracy of 0.93 in capturing known cell types from Weinreb et al. Currently, we are applying Megatron to analyze lineage tracing data from Biddy et al. *Nature* 2018 and Bowling et al. *Cell* 2020. Future work will go into understanding the biological mechanisms underpinning meta-clones by including gene expression information in the AnnData object. In addition, we are interested in better characterizing the pros and cons of Megatron's distance metrics on different forms of the reduced expression space (i.e. PCA, tSNE, UMAP).

Contact: arya_kaul@g.harvard.edu