Single-cell Lineage Tracing by Integrating CRISPR-Cas9 Mutations with Transcriptomic Data

Hamim Zafar[1,*], Chieh Lin[2,*], and Ziv Bar-Joseph[1,2]
[1]Computational Biology Department, [2]Machine Learning Department
School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[*] Authors contributed equally

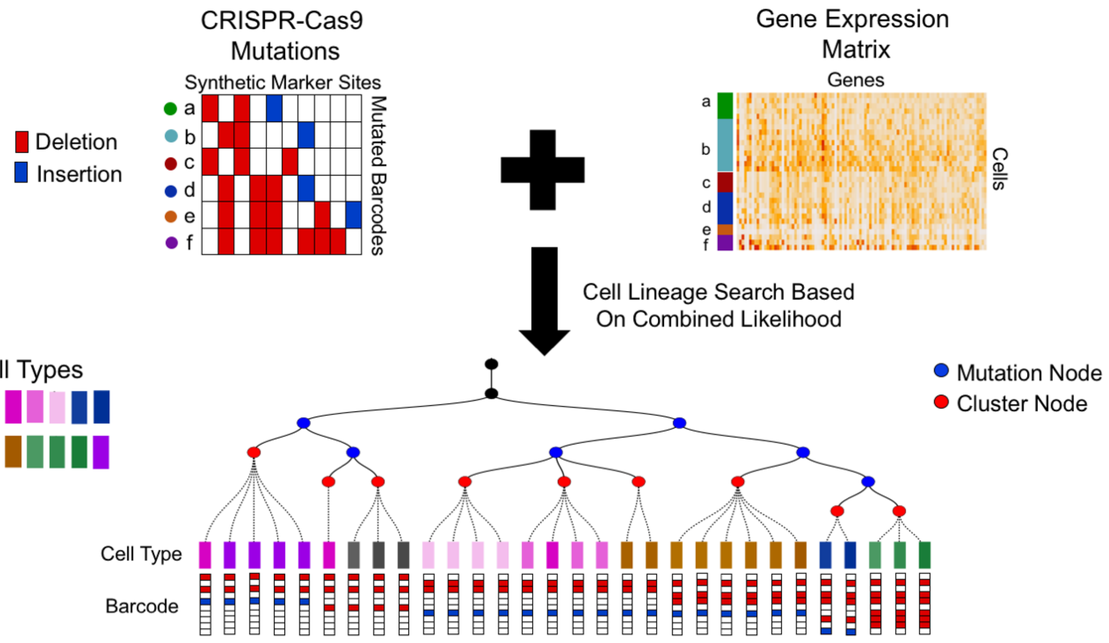
INTRODUCTION

Reconstructing cell lineages that lead to the formation of tissues and organs is of crucial importance in developmental biology. Recent studies combine two novel technologies, single-cell RNA-sequencing and CRISPR-Cas9 barcode editing for elucidating developmental lineages at the whole organism level. These studies raise several computational challenges.

- lineages are reconstructed based on noisy and often saturated random mutation data using Maximum Parsimony (multiple optimal trees)
- resulting lineage tree sometimes fails to separate different types of cells
- due to the randomness of the mutations, lineages from multiple experiments cannot be combined to reconstruct a consensus lineage tree

To address these issues we developed a novel method, **LinTIMaT**

- reconstructs cell lineages using a maximum-likelihood framework by integrating mutation and expression data
- enables the integration of different individual lineages for the reconstruction of a consensus lineage tree



METHODS

LinTIMaT reconstructs the lineage tree ($\mathscr{T}$) by maximizing a log-likelihood function ($\mathcal{L}_T$) that accounts for both mutations and expression data.

$$\mathcal{L}_T(\mathscr{T}) = \omega_1 \log \mathcal{L}_M(\mathscr{T}) + \omega_2 \log \mathcal{L}_E(\mathscr{T})$$

Mutation Likelihood ($\mathcal{L}_M$): By imposing a Camin-Sokal parsimony criterion for each synthetic marker (s), we first use Fitch's algorithm to assign ancestral states $\mathcal{A}_s$ for each marker to each internal node v (with children u and w) of the tree satisfying maximum parsimony.

$$\mathcal{L}_M(\mathscr{T}) = P(\mathcal{E}|\mathscr{T}, \mathcal{A}) = \prod_{s=1}^{S} P(\mathcal{E}_{*s}|\mathscr{T}, \mathcal{A}_s)$$

$$L_s^v(\mathcal{A}_s^v = x) = P(\mathcal{E}_s^v|\mathscr{T}, \mathcal{A}_s^v = x) = \left[P_{t_{\mathcal{A}_s^v \to \mathcal{A}_s^u}} L_s^u\right] \left[P_{t_{\mathcal{A}_s^v \to \mathcal{A}_s^w}} L_s^w\right]$$

Expression Likelihood ($\mathcal{L}_E$): The lineage is modeled as a Bayesian hierarchical clustering (BHC) of the cells. We compute the marginal likelihoods of all the partitions consistent with the given lineage tree based on a Dirichlet process mixture model.

$$\mathscr{L}_g^v = P(\mathcal{Y}_g^v|\mathscr{T}^v) = \pi_v P(\mathcal{Y}_g^v|\mathscr{H}_1^v) + (1-\pi_v) P(\mathcal{Y}_g^u|\mathscr{T}^u) P(\mathcal{Y}_g^w|\mathscr{T}^w)$$

Hypothesis 1: each data point is independently generated from a mixture model and each cluster corresponds to a distribution component.

$$P(\mathcal{Y}_g^v|\mathscr{H}_1^v) = \int P(\mathcal{Y}_g^v|\theta) P(\theta|\beta) d\theta = \int \left[\prod_{\mathbf{y}^{(i)} \in \mathcal{Y}_g^v} P(\mathbf{y}^{(i)}|\theta)\right] P(\theta|\beta) d\theta$$

Hypothesis 2: Data under an internal node comes from two or more clusters

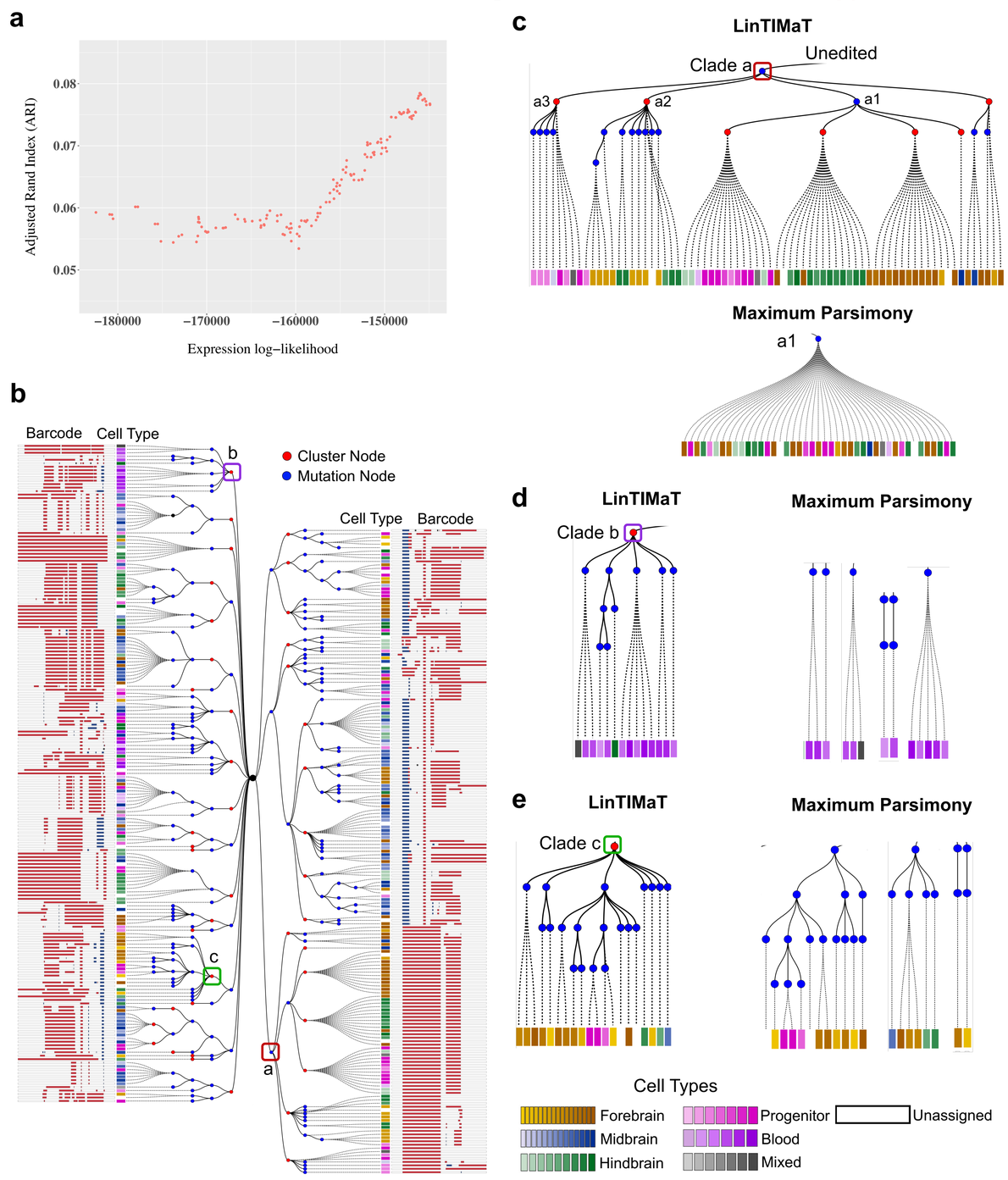
$$P(\mathcal{Y}_g^v|\mathscr{H}_2^v) = \mathscr{L}_g^u \mathscr{L}_g^w = P(\mathcal{Y}_g^u|\mathscr{T}^u) P(\mathcal{Y}_g^w|\mathscr{T}^w)$$

Consensus Lineage Reconstruction: Infers cell clusters from individual lineages, performs a greedy matching to pair the clusters from different individual lineages based on the similarity of gene expression data, optimizes two distance functions, the first one minimizes the disagreement between consensus lineage topology and the individual lineages, the second one is minimized for improving the cluster matching.

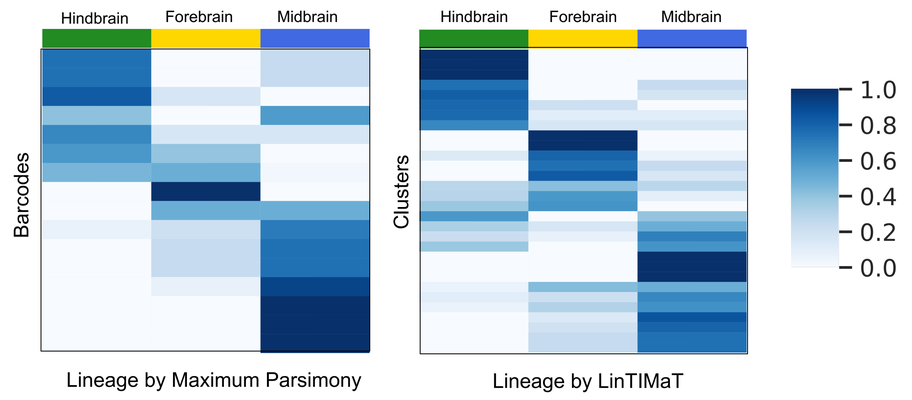
$$\mathscr{T}_{cons}, \mathcal{M}_{cons} = \underset{\mathscr{T}^*, \mathcal{M}^*}{\operatorname{argmin}}\ \omega_1 \sum_{i=1}^{I} \mathcal{S}(\mathscr{T}^*, \mathscr{T}_i^c) + \omega_2 \sum_{j=1}^{K} \mathscr{E}(c_j)$$

RESULTS

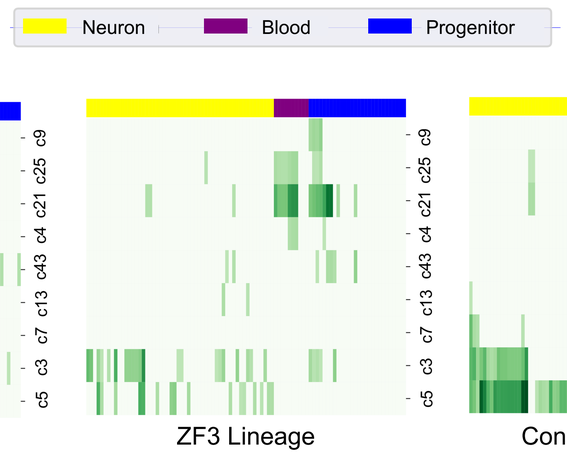
LinTIMaT can refine subtrees in which all cells share the same barcode, can cluster cells with different barcodes together based on their cell types.



Clusters in LinTIMaT reconstructed lineages display better spatial enrichment.



Consensus lineage improves on the individual lineages by uncovering more functionally significant neuronal, blood and progenitor cell clusters



CONCLUSION & FUTURE WORK

- Incorporating complementary data types into a likelihood-based framework improves cell lineage reconstruction.
- In-silico validation study on *c. elegans* cell lineage
- Analyzing other CRISPR-Cas9 lineage datasets
- LinTIMaT is freely available at https://github.com/jessica1338/LinTIMaT

** This work was partially funded by the National Institutes of Health (NIH) [grants 1R01GM122096 and OT2OD026682 to Z.B.J.].