LRT: T Cell Trajectory Inference by Integrative Analysis of Single Cell TCR-seq and RNA-seq data

Juan Xie[1], Gang Xin[2,3], Qin Ma[3,4] and Dongjun Chung[3,4,*]

[1]The Interdisciplinary Program in Biostatistics, The Ohio State University, Columbus, Ohio, USA.

[2]Department of Microbial Infection and Immunity, The Ohio State University, Columbus, Ohio, USA.

[3]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.

[4]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA.

*To whom correspondence should be addressed. Email: chung.911@osu.edu

Abstract

Single-cell RNA sequencing (scRNA-seq) data has been widely used for cell trajectory inference, with the assumption that cells with similar expression profiles share the same differentiation state. However, the inferred trajectory may not reflect true clonal relationships among cells. Single cell T cell receptor sequencing (scTCR-seq) data provides invaluable insights into the clonal relationship among cells, yet it lacks the functional characteristics. Therefore, scRNA-seq and scTCR-seq data complement each other in improving the trajectory inference. However, very limited computational efforts have been devoted to the integration of scRNA-seq and scTCR-seq data for cell trajectory inference. To address this gap, we developed LRT, an R package for the integrative analysis of scTCR-seq and scRNA-seq data for T cell trajectory inference. In addition, it provides two Shiny apps that allow users to interactively explore distributions of clonotypes and implement clustering of cell trajectories. We illustrated its utility using scRNA-seq and scTCR-seq data of $CD4^+$ T cells with acute lymphocytic choriomeningitis virus infection.

Availability: The R package 'LRT' is freely available at https://github.com/JuanXie19/LRT.

Contact: chung.911@osu.edu

Supplementary information: Supplementary data are available online.

Keywords: Immunoinformatics; Single cell; Cell trajectory inference; Single cell RNA-seq; Single cell TCR-seq

1. Introduction

Cell trajectory inference aims to understand how cells differentiate, which is a long-standing problem in developmental biology. Understanding how progenitor cells transform into specified functional cells can provide valuable insights into the molecular mechanisms underlying normal tissue formulation, as well as developmental disorders and pathologies (Wagner and Klein, 2020). Single-cell RNA-seq (scRNA-seq) data has been widely used to investigate cell trajectory. Assuming that cells in different states express different sets of marker genes, we may order cells along a differentiation trajectory via capturing differentiated transcriptional activities (Trapnell, et al., 2014). While this approach has been shown to be useful, they still suffer from intrinsic limitations. Specifically, it assumes that cells with similar expression profiles share the similar developmental stages and may come from the same lineage, which might not be always true (Wagner and Klein, 2020; Zafar, et al., 2020). Besides, in the absence of clonal information, the inferred trajectory may not reflect the true clonal relationship between cells (Wagner and Klein, 2020).

The recent advent of single-cell T cell receptor (TCR) sequencing (scTCR-seq) has enabled the use of TCR sequences as unique 'barcode' to identify clonally related cells. This is because the cells with identical paired TCR sequences usually arose from the same T cell clone due to the high dimensionality of TCR sequence space (Stubbington, et al., 2016). On the other hand, while TCR data provides invaluable insights about the breadth of the antigenic response and clonal relationships between cells, it does not offer information about the functional characteristics of the T cells. Integrating scTCR-seq and scRNA-seq may help elucidate the interplay between TCR sequences and T cell phenotypes. To the best of our knowledge, computational efforts to integrate scRNA-seq and scTCR-seq data for cell trajectory inference are limited in the literature, despite their potential to improve T cell lineage inference.

To address this limitation, here we propose LRT (**L**ineage inference by integrative analysis of sc**R**NA-seq and sc**T**CR-seq data), an R package for the integrative analysis of scTCR-seq and scRNA-seq data for cell trajectory inference. It provides functions for data processing and cell trajectory inference. In addition, it provides two Shiny apps that allow users to interactively explore distributions of clonotypes and implement and visualize clustering of cell trajectories.

2. Materials and Methods

R package 'LRT' provides a workflow for integrative analysis of scTCR-seq and scRNA-seq data (Fig S1). LRT takes a Seurat object as the input for scRNA-seq so that LRT can be easily engineered with the existing pipelines. In the case of scTCR-seq, users can provide amino acid or nucleotide sequences of TCR α, β, or paired α-β chain CDR3 region. Taking scRNA-seq and scTCR-seq data as input, LRT implements the lineage inference. Specifically, a weighted graph is constructed for each group of cells sharing the same TCR sequence. Here nodes are cell-cluster-specific centers of these cells while edge weight represents the Euclidean distance between nodes on the UMAP space. A minimum spanning tree (MST) (Graham and Hell, 1985) is then constructed, where paths through the MST across cell clusters are interpreted as lineages. Finally, clustering analysis is conducted to group similar lineages. Specifically, dynamic time warping distances (Berndt and Clifford, 1994) between trajectories are calculated and a hierarchical or partitional clustering algorithm is applied to cluster these trajectories. One key challenge in such integrative analysis is the high dimensionality of clonotypes. To address this challenge and help users intuitively implement the lineage analysis, LRT provides two Shiny apps. First, *shinyClone* allows users to implement exploratory analysis to understand the distribution of clonotypes on the UMAP space generated using scRNA-seq data (Fig S2). Second, *shinyClust* allows users to interactively implement the lineage clustering analysis (Fig S3).

3.Example

To demonstrate the utility of LRT, we analyzed the scTCR-seq and scRNA-seq data generated from the antigen specific $CD4^+$ T cells during acute lymphocytic choriomeningitis virus infection (Khatun, et al., 2021). Here Seurat was used to preprocess scRNA-seq data and implement UMAP dimension reduction and cell clustering. We obtained 11 cell clusters and annotated them based on the expert knowledge (Fig S4), including T central memory precursor (Tcmp) and type 1 helper T cell (Th1) clusters. Here Th1 cells were further divided into pre-Th1, Th1-intermediate, and Th1 cells. We used CDR3 β chain amino acid sequence for scTCR-seq data. After preprocessing and integration, 7,395 cells were used for the LRT analysis.

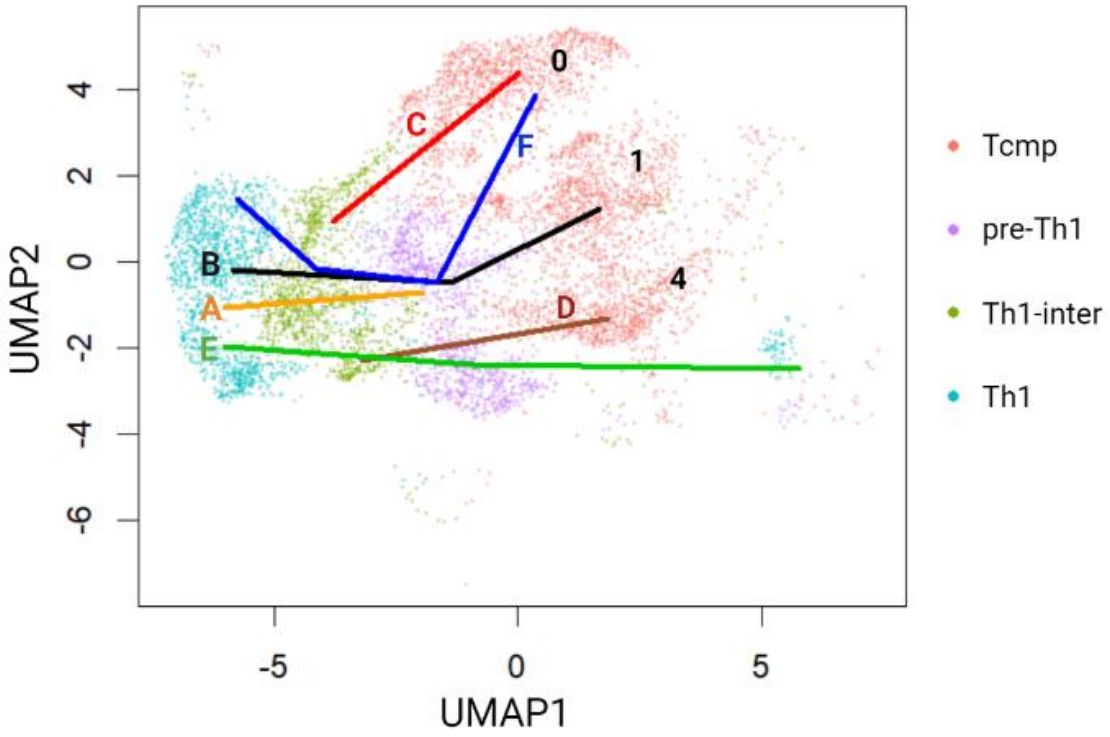

Fig.1. Clonotype lineage clusters obtained from LRT. Dots represent cells, with colors denoting different cell types. Numbers without circles denote cell cluster indices (only shown for Tcmp sub-clusters). Each line indicates the representative trajectory for each lineage cluster, and letters denote lineage cluster indices.

Trajectory analysis based only on scRNA-seq data using Slingshot (Street, et al., 2018) (Fig S5) suggested a single trajectory of Tcmp → pre-Th1 → Th1-intermediate → Th1, which coincides with the known lineage of Th1 cells. When we applied LRT, we further discovered six sub-groups of trajectories constituting this trajectory (Fig 1, Fig S6-11). Specifically, five out of these six lineage groups started from the Tcmp cluster and four of them started from different Tcmp sub-clusters. For instance, lineage clusters B, C, and D started from Seurat cell clusters 1, 0, and 4, respectively (Fig 1). Repertoire analysis indicates very low overlap of clonotypes among these Tcmp sub-clusters (Fig S12). Differential expression analysis and pathway enrichment analysis further indicated that these Tcmp sub-clusters play disparate functional roles (Fig S13). Specifically, cell cluster 4 was enriched for cell cycle mototic and cell cycle checkpoints genes; cell cluster 1 was enriched for IL2 activation and signaling pathways; and cell cluster 0 was enriched for immune response and T cell activation. The above analysis suggested that integration of scTCR-seq and scRNA-seq data using LRT can discover complexity of trajectories that can be missed by the trajectory analysis based only on scRNA-seq data.

Funding

This work was supported in part by NIH/NHGRI grant R21 HG012482, NIH/NIA grant U54-AG075931, NIH/NIGMS grant R01-GM122078 and R01-GM131399, NIH/NIDA grant U01-DA045300, NSF grant NSF-1945971, and the Pelotonia Institute of Immuno-Oncology (PIIO). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

Berndt, D.J. and Clifford, J. Using dynamic time warping to find patterns in time series. In, *KDD workshop*. Seattle, WA, USA:; 1994. p. 359-370.

Graham, R.L. and Hell, P. On the history of the minimum spanning tree problem. *Annals of the History of Computing* 1985;7(1):43-57.

Khatun, A., *et al.* Single-cell lineage mapping of a diverse virus-specific naive CD4 T cell repertoire. *Journal of Experimental Medicine* 2021;218(3).

Street, K., *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* 2018;19(1):1-16.

Stubbington, M.J., *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nature methods* 2016;13(4):329-332.

Trapnell, C., *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* 2014;32(4):381-386.

Wagner, D.E. and Klein, A.M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics* 2020;21(7):410-427.

Zafar, H., Lin, C. and Bar-Joseph, Z. Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data. *Nature communications* 2020;11(1):1-14.