Title: Identification of Cell-type Marker Genes from Plant Single-cell RNA-seq Data Using Machine Learning

Authors: Haidong Yan[1](yanhd@vt.edu), Qi Song[1,2,4], Jiyoung Lee[1,2], **John Schiefelbein**[3], **Song Li**[1,2*]

Institutions: [1]School of Plant and Environmental Sciences (SPES). [2]Graduate program in Genetics, Bioinformatics and Computational Biology (GBCB), [3]Department of Molecular, Cellular, and Developmental Biology, University of Michigan. Ann Arbor, MI 48109. *Corresponding author (songli@vt.edu).

Website: https://github.com/LiLabAtVT/SingleCellClassification

Project Goals:

Objective 1: Discover how extremophytes and stress sensitive species differ in the cell-type functions of roots and those triggered downstream of ABA.

Objective 2: Define how changes in the wiring of gene regulatory networks produce innovations in transcriptional regulation in extremophytes and how bioenergy crops have diverged.

Objective 3: Establish a data driven, predictive framework for accelerating functional testing of stress resilience genes using Arabidopsis and Camelina as a chassis for engineering.

Abstract text: An essential step of single-cell RNA sequencing analysis is to classify specific cell types with marker genes in order to dissect the biological functions of each individual cell. In this study, we integrated five published scRNA-seq datasets from the Arabidopsis root containing over 25,000 cells and 17 cell clusters. We have compared the performance of seven machine learning methods in classifying these cell types, and determined that the random forest and support vector machine methods performed best. Using feature selection with these two methods and a correlation method, we have identified 600 new marker genes for 10 root cell types, and more than 70% of these machine learning-derived marker genes were not identified before. We found that these new markers not only can assign cell types consistently as the previously known cell markers, but also performed better than existing markers in several evaluation metrics including accuracy and sensitivity. Markers derived by the random forest method, in particular, were expressed in 89-98% of cells in endodermis, trichoblast, and cortex clusters, which is a 29-67% improvement over

known markers. Finally, we have found 111 new orthologous marker genes for the trichoblast in five plant species, which expands the number of marker genes by 58-170% in non-Arabidopsis plants. Our results represent a new approach to identify cell-type marker genes from scRNA-seq data and pave the way for cross-species mapping of scRNA-seq data in plants.

References/Publications

1. Identification of cell-type marker genes from plant single-cell RNA-seq data using machine learning. Haidong Yan, Qi Song, Jiyoung Lee, John Schiefelbein, Song Li doi: https://doi.org/10.1101/2020.11.22.393165

Funding statement: This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0020358.

Title: Identification of cell-type marker genes from plant single-cell RNA-seq data using machine learning

Haidong Yan[1], Qi Song[1,2,4], Jiyoung Lee[1,2], John Schiefelbein[3], Song Li[1,2*]

[1]School of Plant and Environmental Sciences (SPES). [2]Graduate program in Genetics, Bioinformatics and Computational Biology (GBCB), [3]Department of Molecular, Cellular, and Developmental Biology, University of Michigan. Ann Arbor, MI 48109. *Corresponding author.

Abstract

An essential step of single-cell RNA sequencing analysis is to classify specific cell types with marker genes in order to dissect the biological functions of each individual cell. In this study, we integrated five published scRNA-seq datasets from the *Arabidopsis* root containing over 25,000 cells and 17 cell clusters. We have compared the performance of seven machine learning methods in classifying these cell types, and determined that the random forest and support vector machine methods performed best. Using feature selection with these two methods and a correlation method, we have identified 600 new marker genes for 10 root cell types, and more than 70% of these machine learning-derived marker genes were not identified before. We found that these new markers not only can assign cell types consistently as the previously known cell markers, but also performed better than existing markers in several evaluation metrics including accuracy and sensitivity. Markers derived by the random forest method, in particular, were expressed in 89-98% of cells in endodermis, trichoblast, and cortex clusters, which is a 29-67% improvement over known markers. Finally, we have found 111 new orthologous marker genes for the trichoblast in five plant species, which expands the number of marker genes by 58-170% in non-Arabidopsis plants. Our results represent a new approach to identify cell-type marker genes from scRNA-seq data and pave the way for cross-species mapping of scRNA-seq data in plants.