Single-cell analysis of human iPSCs and neural lineage entry to discover novel marker genes through a consensus pseudotime trajectory

Claire Malley, Pei-Hsuan Chu, Christopher P. Austin, Anton Simeonov, Ilyas Singeç

National Center for Advancing Translational Sciences (NCATS), Stem Cell Translation Laboratory (SCTL), NIH, Rockville, MD 20850

Introduction

Single-cell RNA sequencing (scRNA-Seq) combined with pseudotime trajectory inference can illuminate the cell differentiation process of pluripotent stem cells. Currently, competing methods exist for constructing pseudotime trajectories based on expected topology (i.e. linear, cyclic, branching) and prior parameters (i.e. start cell, number of expected clusters). The researcher is left to judge which approach to select and no strategy exists that incorporates a statistical confidence measure in the predicted cell pseudotimes.

We performed controlled neural induction of human **induced pluripotent stem cells (iPSCs)** using dual-SMAD inhibition over six different timepoints (day 0-7) and carried out scRNA-Seq using the ddSEQ platform. We ran multiple **trajectory inference (TI)** methods as recommended by the Dyno R package, which has benchmarked over fifty methods for accuracy against a gold standard, scaling, and quality control. A **consensus pseudotime** was created using machine learning methods to combine top-scoring TI methods, followed by **Gene Set Enrichment Analysis (GSEA)**.

Methods

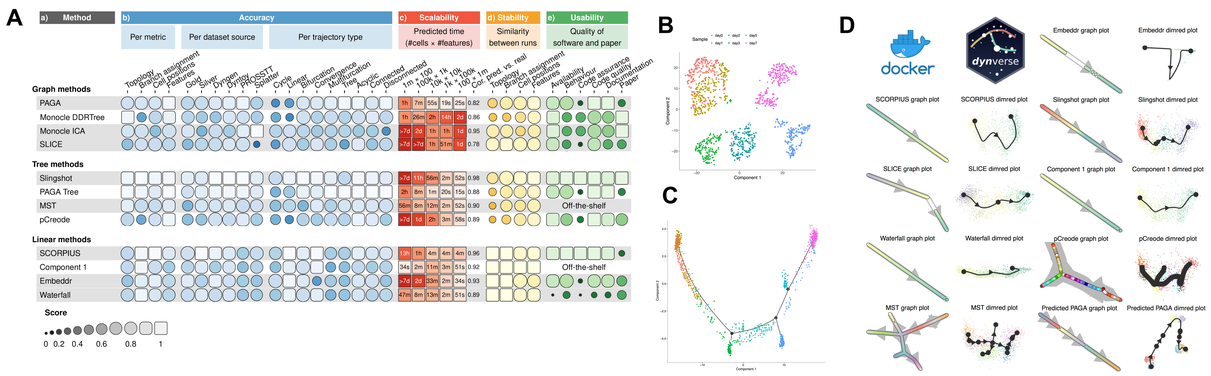


Fig. 1: Multiple pseudotime TI methods run in Dyno R Shiny.
A. In a benchmark study by Saelens et al. 2019 Nat. Biotech, TI methods are not all alike in accuracy, scalability, stability, and usability (subset of the authors' Fig. 3 to TIs used here). **B.** Seurat 3.0 tSNE of 1,385 cells from the time-course neural differentiation. **C.** Standalone Monocle 2 DDRTree was run separately for comparison. **D.** Network graph and dimensionality reduction plots from each method.

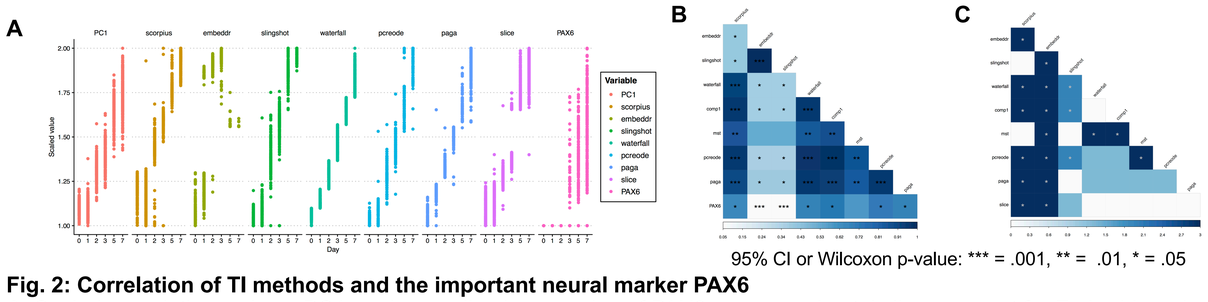


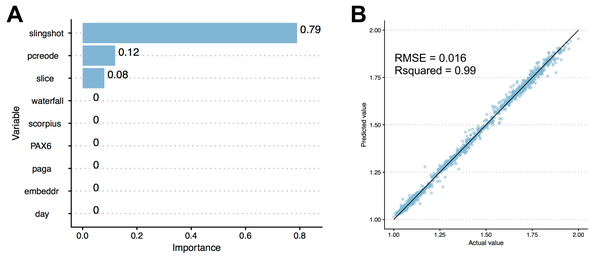
Fig. 2: Correlation of TI methods and the important neural marker PAX6
A. Scaled pseudotime values, PC1, and normalized expression of PAX6 are approximately linear except for Embeddr and SLICE, which are disjoint between days 1 - 3. **B.** Correlation of methods and PAX6. MST was then excluded due to nonsignificant correlation. **C.** Paired Wilcoxon rank sum tests. The methods have at least 1 significantly different alternative method included.

Results

Box 1

Type of machine: random forest regressor
Predictor: PC1
Features: pseudotimes, day, and PAX6 normalized expression
Number of trees: 100
Maximum depth: 10
Minimum samples per split: 2

Mean absolute error: 0.02 pseudotime "hours"
% Variance explained: 98.6



Box 1 and Fig. 3: Machine learning results
1. Summary of random forest (RF) regression. **A.** Variable importances from the model. Only three TI methods were needed to predict pseudotime: Slingshot, pCreode, and SLICE. **B.** Actual PC1 vs. predicted pseudotime values. The resulting consensus pseudotime had high accuracy, high correlation, and low RMSE.

Results, continued

GSEA Gene Set Enrichment Analysis

Method	Cells with pseudotime	Positively upregulated gene sets	Sig. at FDR <25%	Sig. at p <0.01	Sig. at p <0.05
RF Consensus	1385 (100%)	4398/4566	499	393	730
Monocle 2 DDRTree	1320 (95%)	1269/3811	170	86	194
Overlap	1320	4430	502	9	22

Table 1: GSEA results.
GSEA using the consensus pseudotime phenotype found more enriched GO terms than when using Monocle 2. The overlapping significant terms (N = 502) had overall higher enrichment scores, p value, and lower FDR for RF consensus.

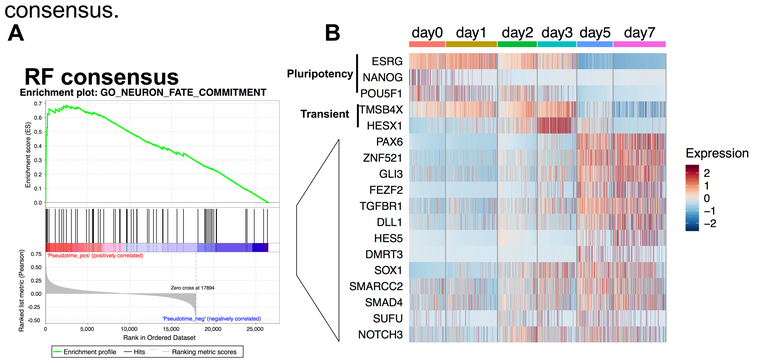


Fig. 4: Day 7 neural precursors are enriched for the Neuron Fate Commitment GO Biological Process
A. Running enrichment score and positions of gene set members in the rank ordered list. **B.** Heatmap of normalized expression. The core gene set has increased expression towards day 7 cells. For comparison, pluripotency (day 0) markers and transient markers (days 2-3) are shown, which are significantly differentially expressed as per Seurat Wilcoxon Rank Sum tests.

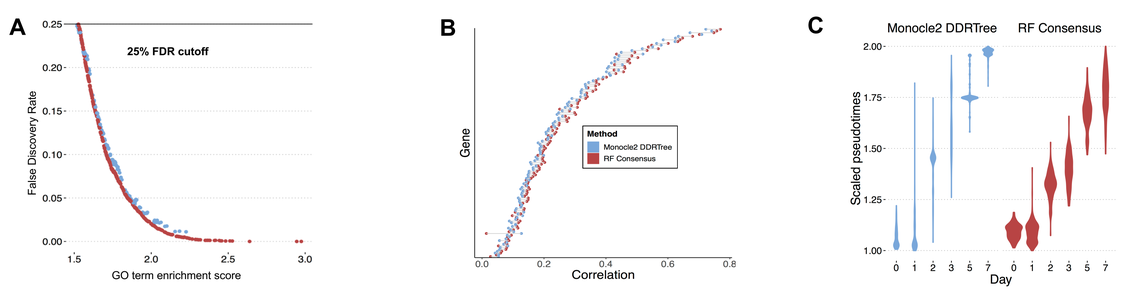


Fig. 5: Consensus pseudotime GSEA led to discovery of more and higher-scoring gene sets.
A. Scores vs. false discovery rate (FDR). The consensus (red) has both higher scores, lower FDR, and included sets not found in the Monocle 2 method (blue). **B.** Correlation of each pseudotime method with average expression of genes per set. Note RF consensus points are higher than Monocle 2. **C.** Overall distribution of cell pseudotime was more smoothly distributed in the consensus than Monocle, seen in large ranges for days 1-3, but Monocle defined a smaller range for day 7 cells.

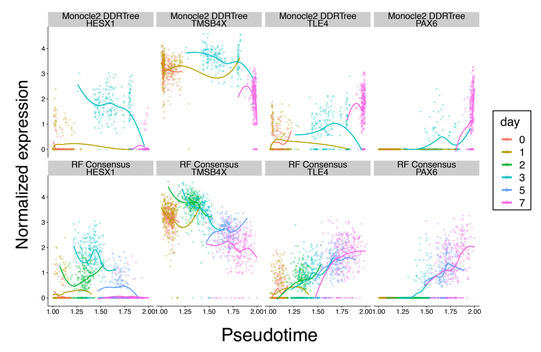


Fig. 6: Patterns of pseudotime vs. expression
Two transient and two day-7 neural markers show different trends when plotted against the two pseudotime methods.

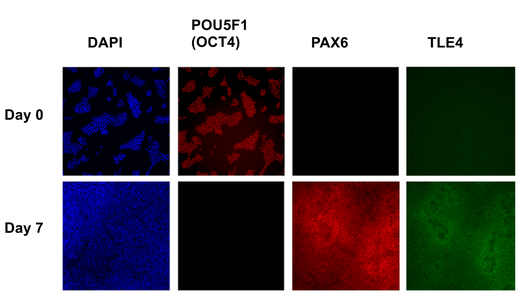


Fig. 7: Immunocytochemical analysis of iPSCs and neural precursors.
Follow-up of bioinformatic analyses with antibody staining confirmed protein levels of differential expression markers.

Summary

Trajectory method choice impacts resulting marker gene analysis, which will be critical to measure as scRNA-Seq use may eventually supersede bulk RNA-Seq. An aggregation framework based on comprehensive benchmarking may prove superior to single method choice as demonstrated here. We identified precisely regulated early marker genes of neuralization (HESX1, TMSB4X) at days 2-3 that led to definitive neural lineage commitment by day 7 (PAX6, TLE4). Pseudotime trajectory construction followed by GSEA uncovered a suite of related contributing transcription factors.

Funding sources: NIH Common Fund and NCATS Intramural Research