CellDrift: identifying cellular and temporal patterns of perturbation responses from single-cell data

Kang Jin [1,2] , Daniel Schnell [1], Guangyuan Li [1,2], Surya Prasath [1,2,3], Rhonda Szczesniak [4], Bruce J Aronow [1,2,3]

[1] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center; [2] Department of Biomedical Informatics, University of Cincinnati; [3] Department of Electrical Engineering and Computer Science, University of Cincinnati; [4] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center; Kang.Jin@cchmc.org; https://kang-bioinfo.github.io/

Abstract

Shifted gene programs are the key to understanding perturbation responses in single-cell RNA sequencing experiments. With the increasing complexity of perturbational experiments, generative models, such as scGen and CPA, have been used to interrogate perturbation latent features utilizing the power of deep neural networks. However, a lack of interpretability still prevents biologists from straightforwardly understanding perturbation responses. Here we present **CellDrift**, a generalized linear model (**GLM**) that accounts for major covariates, including perturbation groups, cell types, and their interactions in perturbational single-cell data. We applied Function Data Analysis (**FDA**) based on the results of GLM for perturbational studies with time series and identified temporal patterns of gene programs in perturbation responses. To illustrate CellDrift, we applied the model to a COVID-19 and sepsis blood single-cell dataset and identified disease-specific temporal patterns of interferon responses in sepsis and COVID-19 patients.

Background

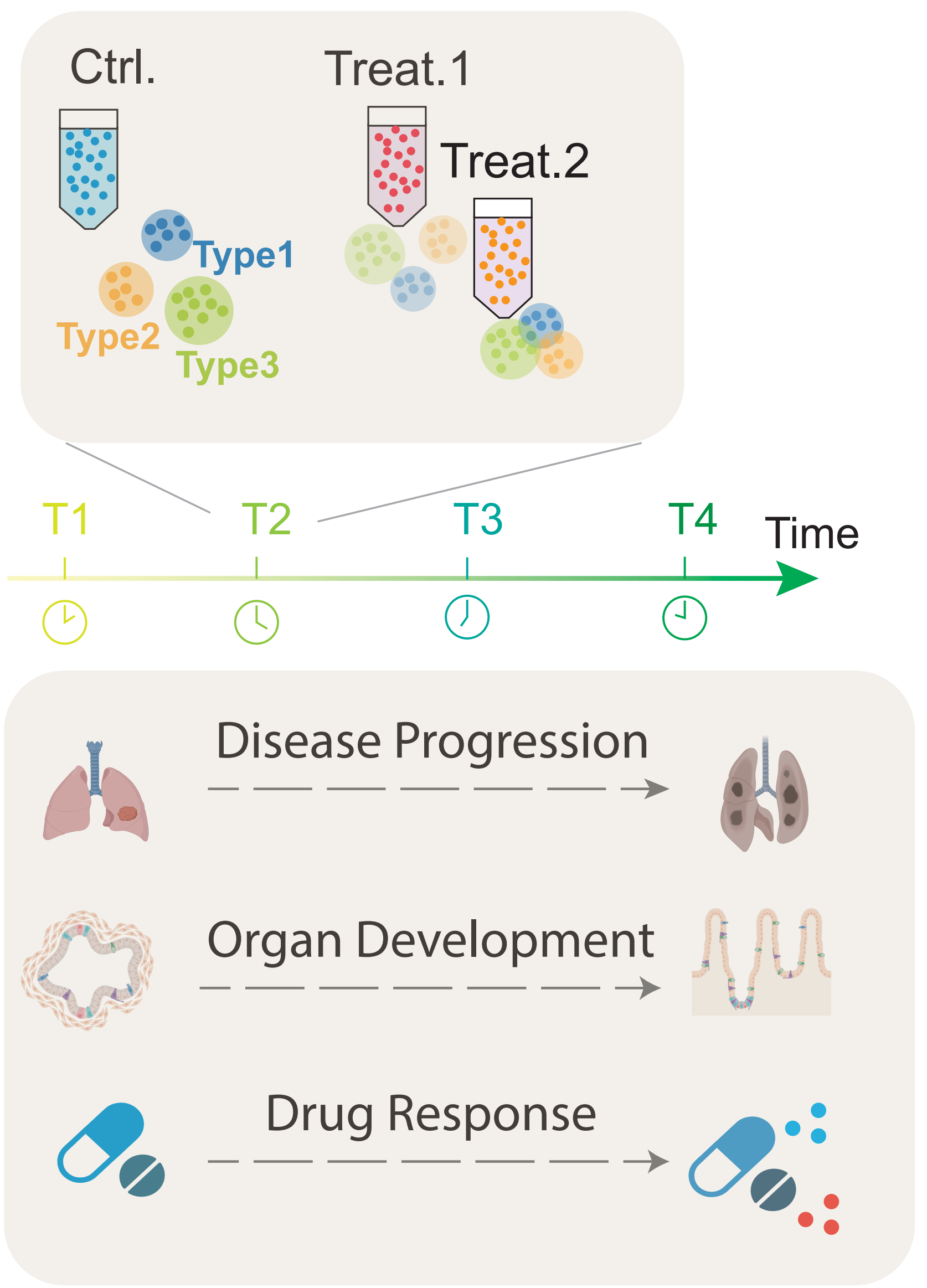

Figure1. Real world examples of perturbational single-cell data with time series.

GLM-based Function Data Analysis

Generalized Linear Model for cellular perturbation responses

We model gene expression data as nonlinear function of cell type and perturbation effects using a negative binomial (NB) generalized linear model (GLM). In the GLM, a reference model to disentangle raw count $\boldsymbol{\gamma_{ng}}$ in cell type c and perturbation p with their coefficients $\boldsymbol{\beta_{gc}}$and $\boldsymbol{\beta_{gp}}$ for each gene $\boldsymbol{g}$ and cell $\boldsymbol{n}$ (*Eq.1, Eq.2*). By adding the interaction term $\boldsymbol{\beta_{gcp}}$ to the major effects, including cell type and perturbation, we create a contrast model where we assume perturbation effects vary across cell types (*Eq.3*).

$$\gamma_{ng} \sim NB(\mu_{ng}, \phi_{ng}) \quad (Eq.1)$$

$$log\mu_{ng} = \eta_{ngcp}. \quad (Eq.2)$$

$$\eta_{ngcp} = logs_n + \beta_{g0} + \beta_{gc}\rho_{gc} + \beta_{gp}\rho_{gp}$$
$$+ \beta_{gcp}\rho_{gcp} + \sum_{j=1}^{b}\beta_{gb}\rho_{gb} \quad (Eq.3)$$

Function Data Analysis for temporal patterns

We utilized Function Data Analysis (FDA) to identify temporal perturbation patterns based on cell specific perturbation coefficients ($\boldsymbol{\beta_{gp} + \beta_{gcp}}$) from the GLM model. We applied dynamic time wrapping to align different time series from multiple perturbations groups. Then we utilized fuzzy K-Means clustering and identified temporal patterns, from where we applied various FDA approaches, such as one-way ANOVA tests and functional principal component analysis (FPCA).

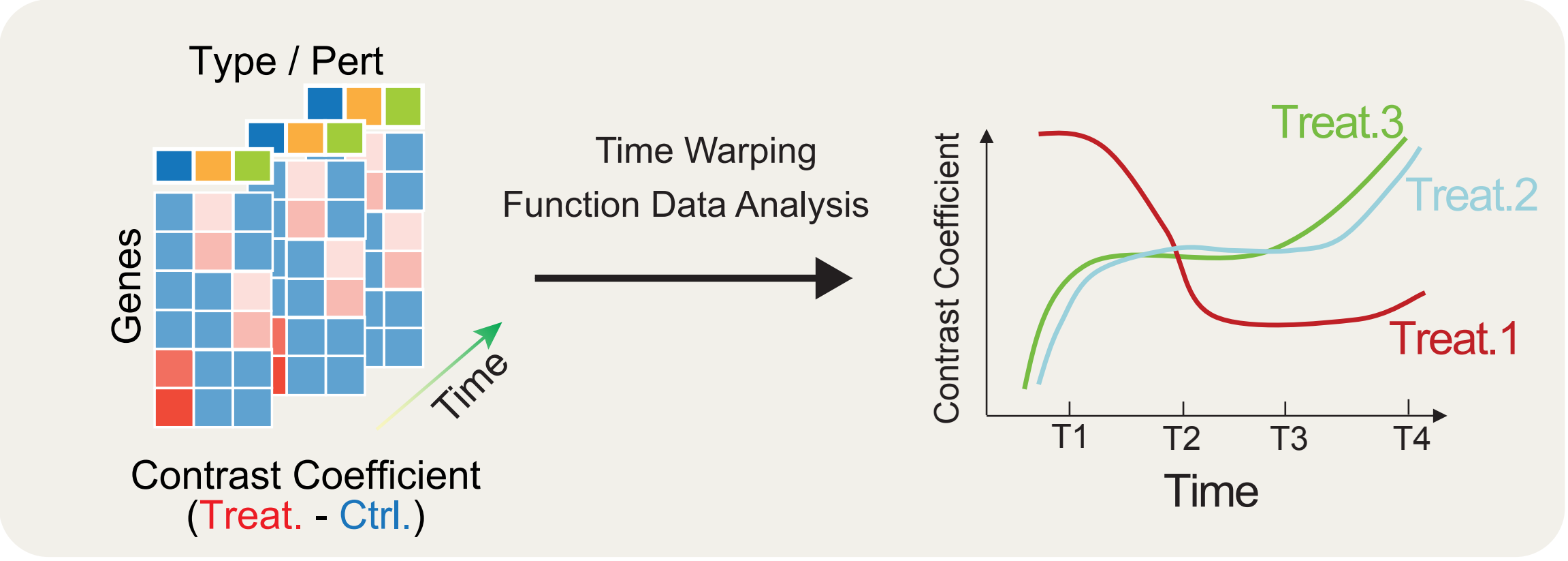

Figure2. Workflow of CellDrift. The output of GLM is used for function data analysis.

Simulation & Performance

We first simulated data without time series and benchmarked our GLM model with other popularly used differential expression (DE) approaches in single-cell analysis. Notably, we observed higher True Positive Rate (TPR) across datasets with different batch effect levels and differential expression patterns. As we expected, the false discovery rates (FDR) are insensitive to batch effects. However, CellDrift doesn't perform well in datasets with very little proportion differentially expression genes. (Fig.3)

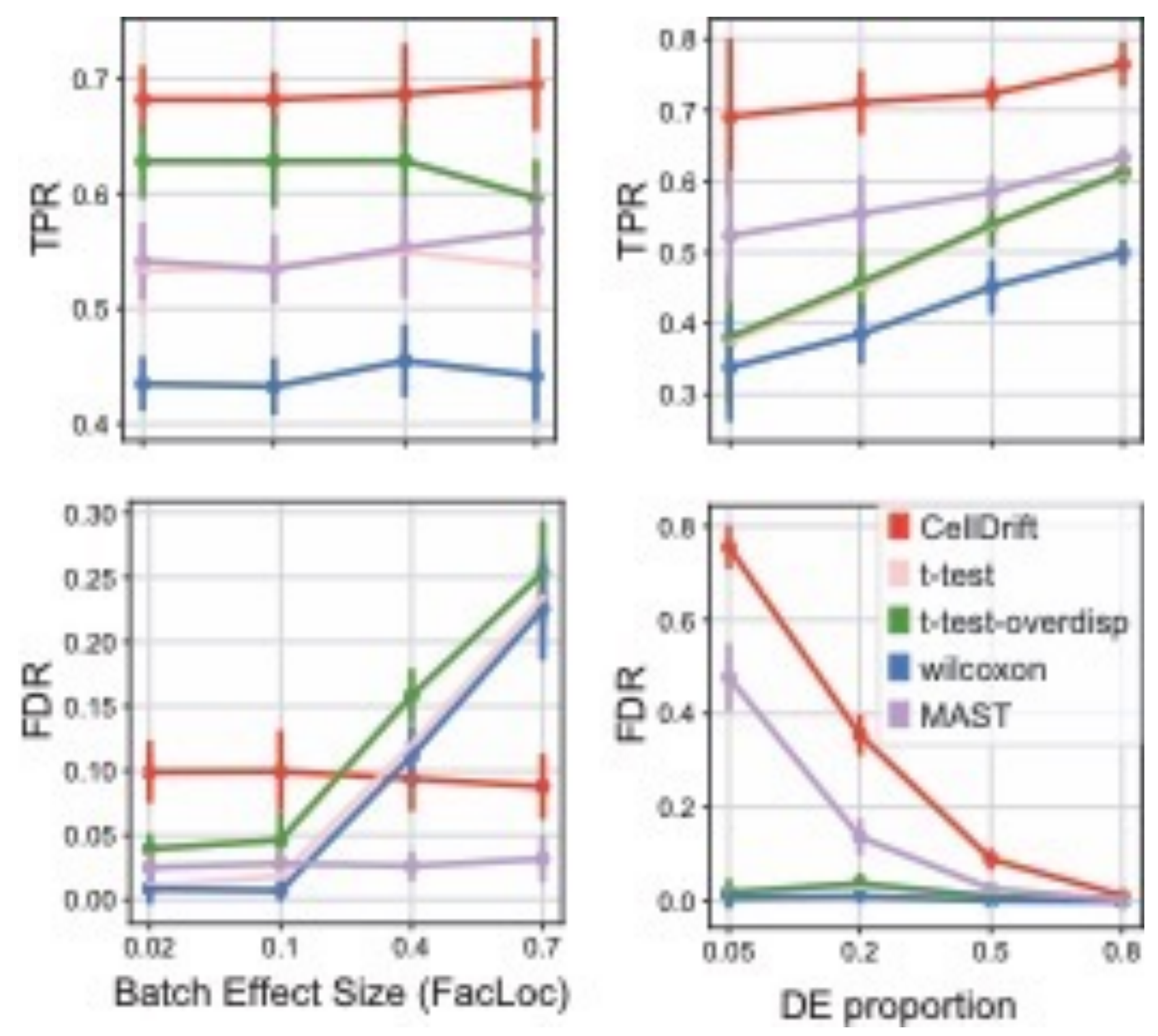

Figure3. Performance of CellDrift GLM model in simulated data with different levels technical and biological variations.

We simulated perturbational single-cell data, and benchmarked different approaches in recovering temporal patterns in the perturbation data. GLM results with fuzzy kmeans turns out to perform best. (Fig.4)

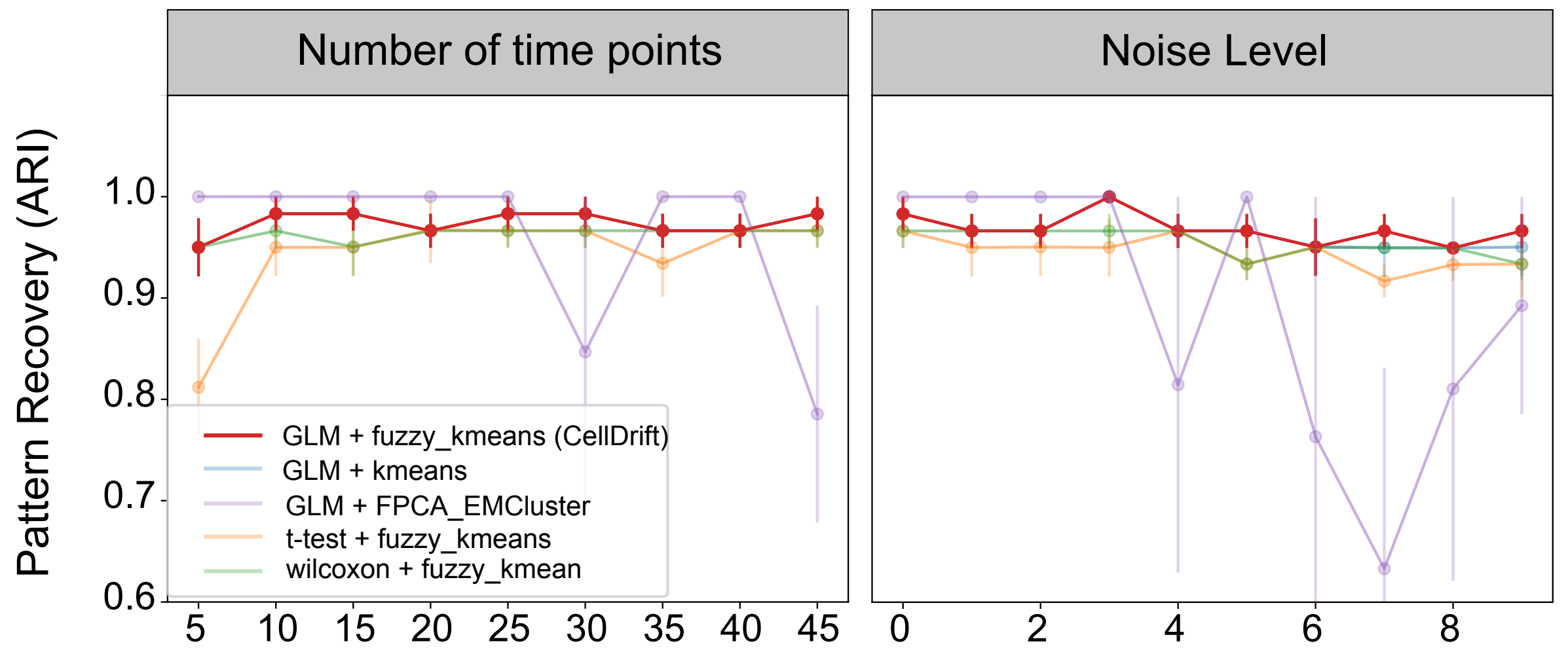

Figure4. Comparison of CellDrift with other approaches in the identification of temporal perturbational clusters.

Application of COVID-19 Atlas

We applied CellDrift in COVID-19 PBMC data and identified 20 clusters of genes with distinct perturbation responses across time. Among them, cluster 9 is associated with immune effector pathways with a strong peak in the early stage.

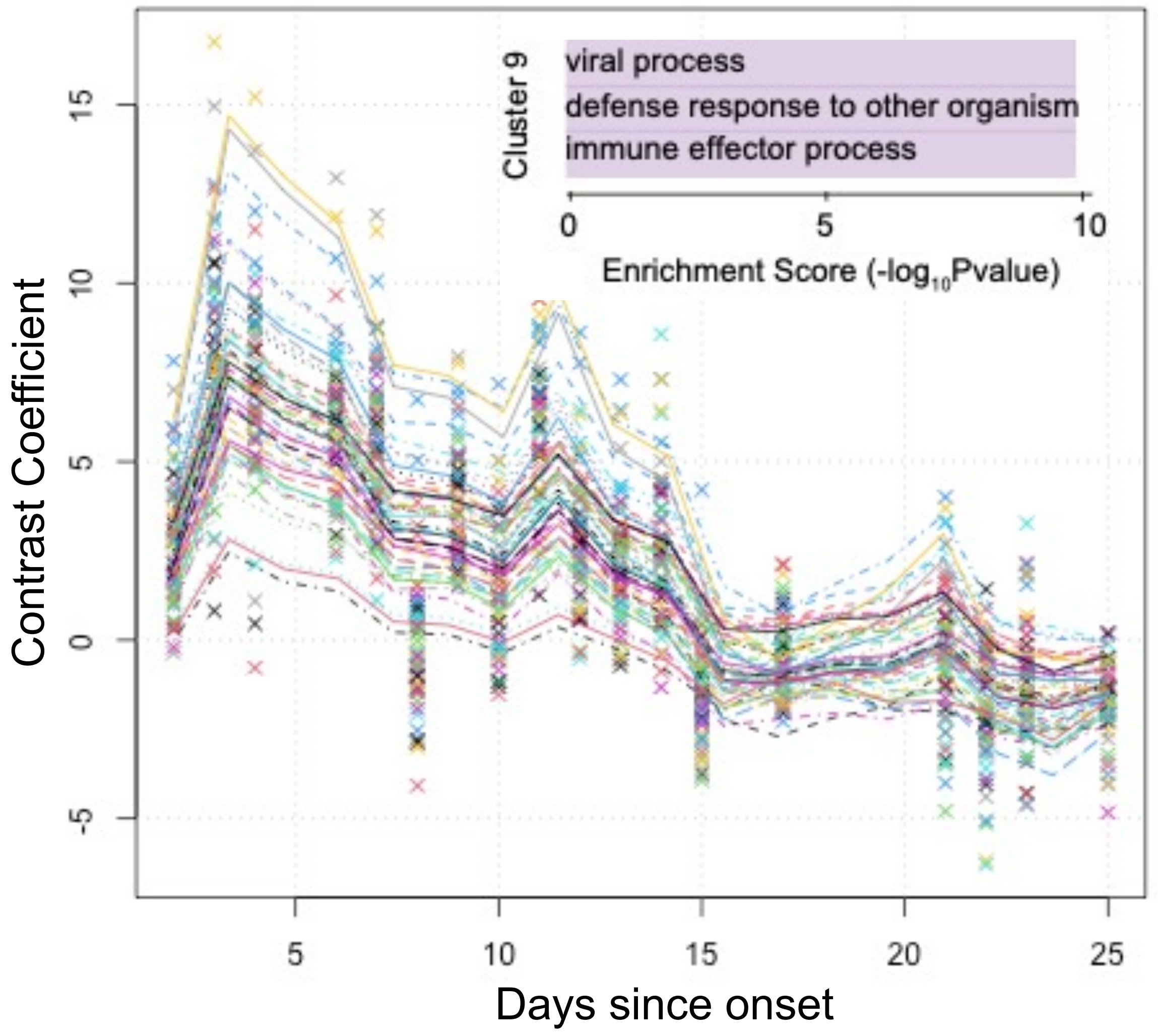

Figure5. Cluster 9 is a temporal perturbation pattern associated with immune effector pathways in monocytes of severe COVID-19 patients.

Using functional one-way ANOVE test, we identified genes with distinct temporal pattern in patients with mild and severe infection diseases. For example, responses of S100A8 and S100A9 are lower in mild patients than severe patients. This phenomenon indicates these two genes are not as strongly activated in mild patients versus severe patients.

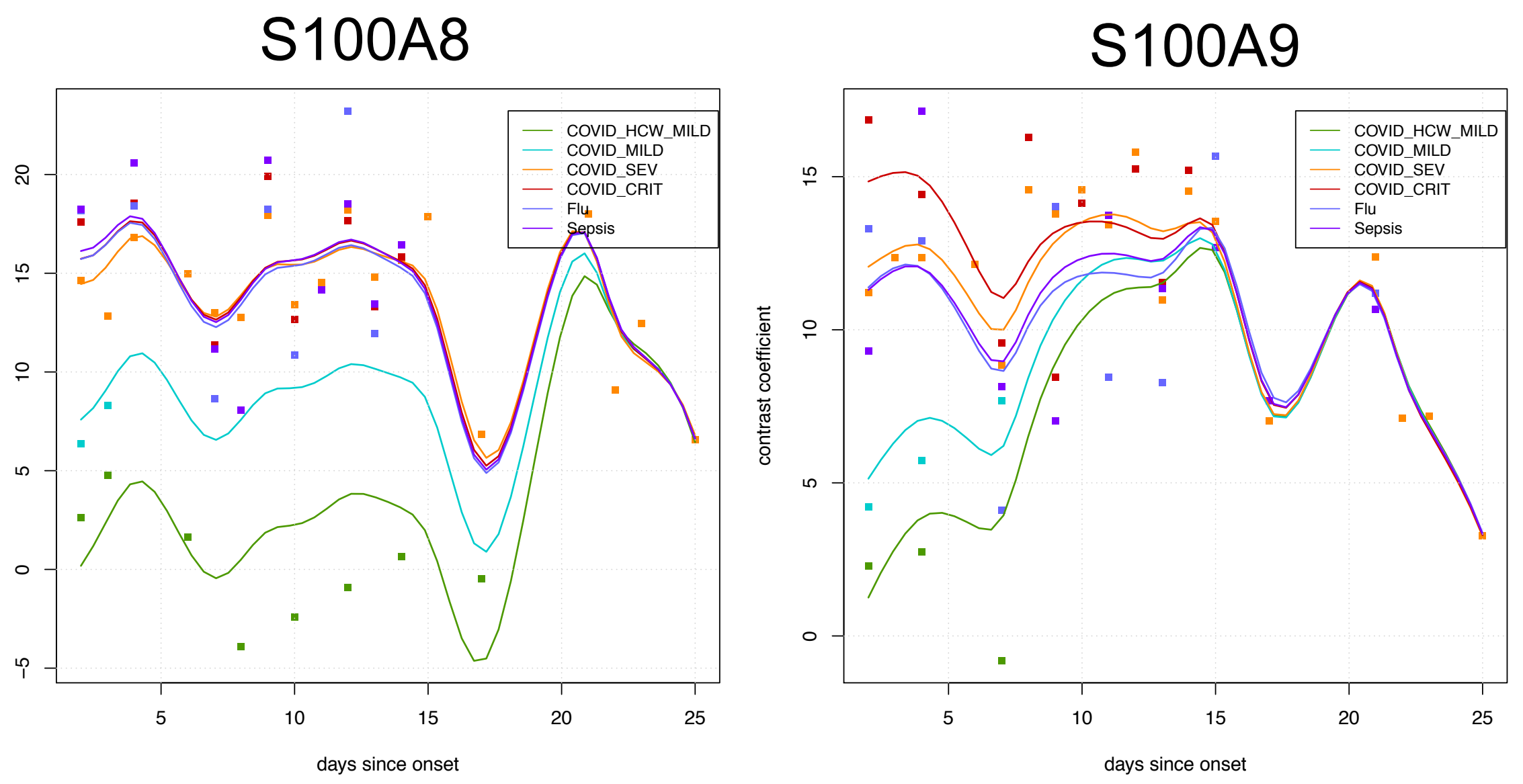

Figure6. Weaker perturbation effects of S100A8 and S100A9 across time in mild COVID-19 patients versus severe patients.