OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–11
https://doi.org/10.1093/bib/bbac324
Problem Solving Protocol

CellDrift: inferring perturbation responses in temporally sampled single-cell data

Kang Jin, Daniel Schnell, Guangyuan Li, Nathan Salomonis, V. B. Surya Prasath, Rhonda Szczesniak and Bruce J. Aronow

Corresponding author. Bruce J. Aronow, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA. Tel.: 513-636-4865; E-mail: bruce.aronow@cchmc.org

Abstract

Cells and tissues respond to perturbations in multiple ways that can be sensitively reflected in the alterations of gene expression. Current approaches to finding and quantifying the effects of perturbations on cell-level responses over time disregard the temporal consistency of identifiable gene programs. To leverage the occurrence of these patterns for perturbation analyses, we developed CellDrift (https://github.com/KANG-BIOINFO/CellDrift), a generalized linear model-based functional data analysis method that is capable of identifying covarying temporal patterns of various cell types in response to perturbations. As compared to several other approaches, CellDrift demonstrated superior performance in the identification of temporally varied perturbation patterns and the ability to impute missing time points. We applied CellDrift to multiple longitudinal datasets, including COVID-19 disease progression and gastrointestinal tract development, and demonstrated its ability to identify specific gene programs associated with sequential biological processes, trajectories and outcomes.

Keywords: single-cell RNA sequencing, perturbation effects, temporal patterns, functional data analysis, generalized linear model

Introduction

Single-cell transcriptomics sequencing has revolutionized discoveries in complex biological systems by identifying a wide variety of cell populations in high resolution [1–3]. Researchers have applied the technology in experiments with perturbation settings, such as diseases [4, 5], treatments [6, 7], genetic mutations [7, 8] and organ differentiation [9, 10], to explore transcriptional profiles across various biochemical states.

However, the response to perturbation can vary over time, which is overlooked in many single-cell studies. Nowadays, researchers are increasingly considering the impact of time when designing experiments. For example, the genetic effects of autism risk genes have been studied during the development of the nervous system using brain organoids [11, 12]. Additionally, influenza vaccination effects have been evaluated by monitoring immune responses over multiple follow-up periods [13]. Moreover, the impact of infections, such as human immunodeficiency virus (HIV), has been studied in patients at varying stages of their illness [14, 15]. By having access to single-cell profiles over time, researchers can accurately report perturbation effects during treatment procedures, disease progression and organ development.

There have been various approaches introduced to quantify transcriptional changes in single-cell ribonucleic acid (RNA)-seq (scRNA-seq) data from perturbation experiments [16] (Table 1). Although traditional methods, such as the Wilcoxon test or *t*-test, are commonly used in single-cell differential expression analysis, they are not sufficient to resolve batch effect and data sparsity issues [17]. More advanced algorithms, such as MAST [18] and muscat [19], have been developed. However, their flexibility in measuring perturbation effects is still limited, such as the decomposition of common and cell-type-specific perturbed genes. Meanwhile, machine learning approaches have been developed for the analysis of complex perturbation data. For example, scGen applied autoencoder models to learn perturbation responses in a latent space and to predict unseen scenarios [20]. Although it is powerful in analyzing high-dimensional data, interpretability in latent spaces remains a significant challenge. In addition,

Kang Jin is a PhD student at the University of Cincinnati. His research interests include developing computational methods in single-cell analysis.
Daniel Schnell is a biostatistician at the Cincinnati Children's Hospital Medical Center. His research interests include statistical modeling in bioinformatics.
Guangyuan Li is a PhD student at the University of Cincinnati. His research interests include developing computational tools in cancer immunotherapy.
Nathan Salomonis is an associate professor at the Cincinnati Children's Hospital Medical Center. His research interests are computational genomics and cancer informatics.
V. B. Surya Prasath is an assistant professor in biomedical informatics at the Cincinnati Children's Hospital Medical Center. His research interests are machine learning and biomedical imaging informatics.
Rhonda Szczesniak is an associate professor in biostatistics at the Cincinnati Children's Hospital Medical Center. Her research interests are predictive modeling and medical monitoring.
Bruce J. Aronow is a professor at the Cincinnati Children's Hospital Medical Center. His research interests are bioinformatics and system biology.
Received: April 19, 2022. **Revised:** June 27, 2022. **Accepted:** July 18, 2022
© The Author(s) 2022. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Table 1. Comparisons of perturbation or temporal evaluation methods in single-cell analysis

Method	Algorithm	Feature space	Temporal evaluation	Perturbation evaluation	Limitation	Reference
MAST	GLM	Genes	Manual comparison	Yes	Algorithm was not designed for temporal analysis	[18]
scGen	Autoencoder	Neural network latent space	Not available	Yes	No implementation of the time covariate; lack of interpretability of the latent space features	[20]
CellBox	Ordinary differential equation (ODE)	Protein and phenotypes	ODE	Yes	Algorithm was not optimally designed for single-cell analysis	[23]
MEFISTO	Factor analysis	Factors	GP	No	Lack of understanding of perturbation responses	[21]
CPA	Autoencoder + adversial network	Neural network latent space	Adversarial network	Yes	Lack of interpretability in the latent space features	[22]
CellDrift	GLM	Genes	FDA	Yes	Time covariate was not incorporated in GLM	This paper

these methods were not originally designed to measure perturbation effects in a temporal context.

Other methods took time as a covariate in the model [21]. For example, compositional perturbation autoencoder (CPA) utilized the combination of linear models and deep-learning approaches to interpret temporal impacts, but the interpretation of gene-level impacts is not straightforward [22]. Furthermore, CellBox analyzes the perturbation effects over time using ordinary differential equations [23]. However, the performance of the method is limited by the sparsity and stochasticity issues in single-cell data.

Generalized linear models (GLM) have been widely used in modeling single-cell transcriptomics data, outperforming linear regression by more accurately and efficiently capturing non-linear relations in count data through non-Gaussian distribution families [24]. For example, sctransform successfully removed technical effects by introducing cellular sequencing depth as a covariate [25].

Functional data analysis (FDA) has been widely used in longitudinal data analysis [26, 27]. A general form of FDA is the analysis of multiple curves varying over time, where each curve is a sample tracing with a series of time points, which can be characterized as a function. Such data are called functional data. One of the most popular tools in FDA is functional principal component analysis (FPCA), which identifies the dominant modes of variation of functional data [28]. It has been widely used in disease progression profiling and predictions. For example, FPCA has been used in the monitoring of glucose levels in hyperglycemic patients [29]. We utilized the flexibility of the FDA to identify temporal perturbation patterns.

To address the aforementioned issues, we developed CellDrift, a GLM-based FDA model, to disentangle temporal patterns in perturbation responses in scRNA-seq data. CellDrift first captures cell-type specific perturbation effects by adding an interaction term in the GLM and then utilizes predicted coefficients to calculate contrast coefficients, which represent perturbation effects in our study. Concatenated contrast coefficients over time are defined as functions, and Fuzzy C-mean clustering is used to identify temporal patterns, which is accompanied by FPCA to find the major components that account for the most temporal variance. We benchmarked CellDrift with multiple functional clustering methods with statistical results from differential expression approaches, such as Wilcoxon and *t*-test, and CellDrift achieved superior performance in the identification of temporal patterns and imputation of perturbation effects. We applied CellDrift in COVID-19 single-cell data and a gut development atlas and identified temporal patterns and functional principal components associated with varying immune responses and gut organ morphogenesis.

Methods

CellDrift takes the input of multiple scRNA-seq UMI count matrices across diverse captures (batches, b), conditions (perturbations,p) and time points (t). The main goal of the algorithm is to disentangle the major effects of different cell types (c) and perturbations (p). The derived contrast coefficients associated with each gene, cell type and condition after implementing the GLM model across time points are used for FDA to identify the temporal patterns of perturbation responses (Figure 1).

Perturbation coefficient model

To begin, we introduce a model and notation for a single time point. We model the raw count of single-cell data γ_{ng} for cell n and gene g using a GLM with a negative binomial (NB) distribution. z_n and x_n represent the cell type and perturbation group of cell n, which are c and p here:

$$\gamma_{ng} \mid (z_n = c, x_n = p) \sim \mathrm{NB}\left(\mu_{ngcp}, \phi_{ngcp}\right), \tag{1}$$

where μ_{ngcp} and ϕ_{ngcp} represent the mean and inverse dispersion of the NB distribution for cell n and gene g.

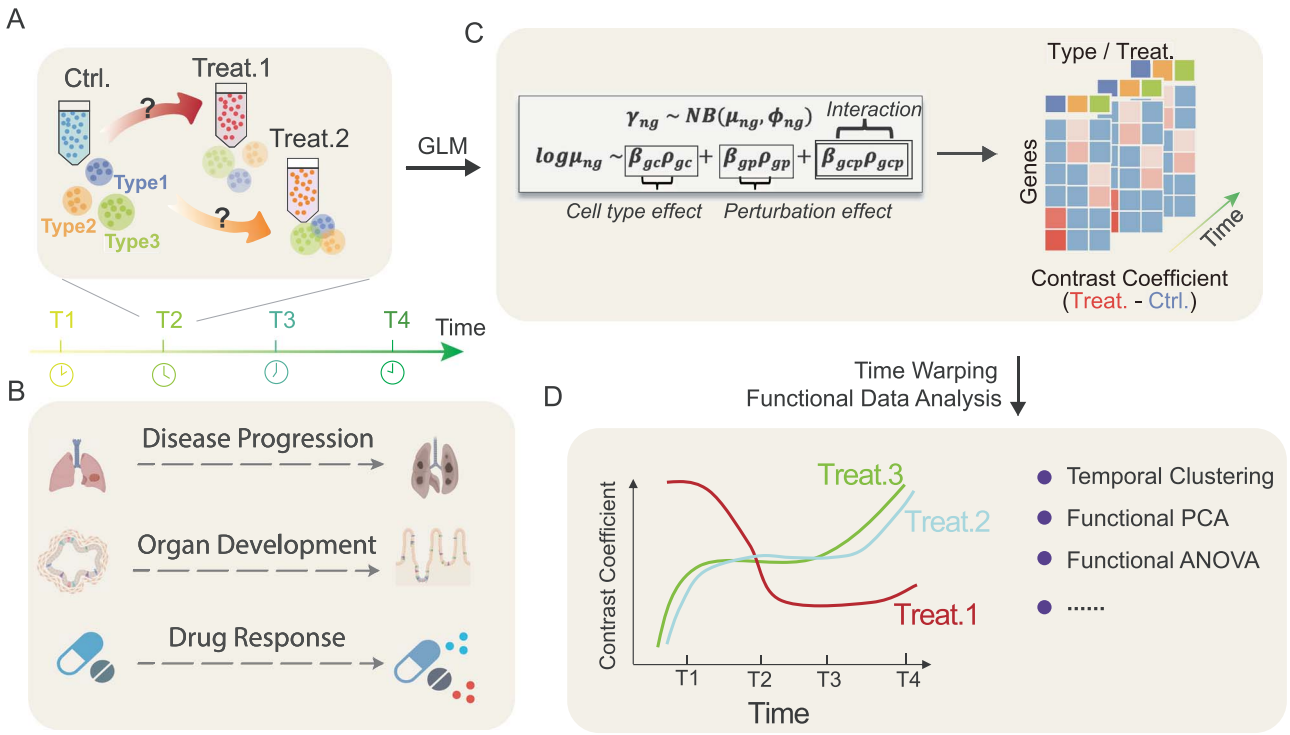

Figure 1. Workflow of CellDrift. (**A**) An example of a perturbational single-cell experiment with multiple time points. (**B**) Real scenarios of single-cell experiments with varying perturbation effects over time. (**C**) GLM with the interaction of cell-type-perturbation applied separately at each time point, and contrast coefficients are derived as the representation of perturbation responses. The complete model can be found in the Methods. (**D**) Contrast coefficients are used as input for various applications in FDA.

For example, z_n and x_n could be the user-defined cell type and perturbation group for cell n, such as CD4+ T cell and Drug A treatment.

We use a log link function (ln) for μ_{ngcp} and disentangle η_{ngcp} with a linear model with cell-type coefficients β_{gc} and perturbation coefficient β_{gp} for each gene g:

$$\log \mu_{ngcp} = \eta_{ngcp}, \tag{2}$$

$$\eta_{ngcp} = \log s_n + \beta_{g0} + \beta_{gc}\rho_{nc} + \beta_{gp}\rho_{np} + \sum_{b=1}^{B}\beta_{gb}\rho_{nb}, \tag{3}$$

where ρ_{nc} and ρ_{np} indicate whether cell n belongs to cell-type c and perturbation group p, as represented by one-hot matrices. The intercept β_{g0} represents the base expression level of gene g. In addition, we account for the library size s_n and batch effects β_{gb} in the model, and ρ_{nb} is the one-hot matrix for cell n and batch b (e.g. donor, sequencing platform). B is the total number of batch types, commonly >1 in complicated datasets. Batch types are incorporated into the model as fixed effects. For simplicity, we omit the b (batch) subscript in the mean count (μ_{ngcp}) and linear predictor symbols (η_{ngcp}).

In Equation (3), cell-type effect and perturbation effect are independent covariates. In real cases, however, different cell types usually have distinct responses toward the same perturbation. Thus, we add an interaction term β_{gcp} for the cell-type and perturbation covariates, representing cell-type-specific perturbation effects. In contrast, β_{gp} represents common perturbation effects across cell types:

$$\eta_{ngcp} = \log s_n + \beta_{g0} + \beta_{gc}\rho_{nc} + \beta_{gp}\rho_{np} + \beta_{gcp}\rho_{ncp} + \sum_{b=1}^{B}\beta_{gb}\rho_{nb}, \tag{4}$$

where ρ_{ncp} is a one-hot matrix of the cell type and perturbation group, indicating whether cell n belongs to cell-type c and perturbation group p at the same time. Likelihood ratio test is used to do the model selection (Supplementary Methods, Figure S1 available online at https://academic.oup.com/bib).

Contrast coefficients

Both main effects (β_{gc} and β_{gp}) and interaction coefficients (β_{gcp}) are estimated using GLM after fitting the single-cell data. Then, we retrieve pairwise contrast coefficients $\Delta\beta_{gcp}$ based on estimated β_{gp} and β_{gcp}, which are used to quantify the difference between the perturbed state and baseline in specific cell types (Supplementary Methods). Briefly, they represent the perturbation effects of cell-type c in perturbation group p. Contrast coefficients are the basic representation of perturbation effects in this study. They are also the input data for FDA (Figure 1).

Representation of functional data using contrast coefficients

In the general form of FDA, each curve is a sample with a series of time points, which is commonly referred to as a function. In our study, in addition to cell types and

perturbation groups in single-cell perturbation data, we added another dimension of complexity, time covariate t, into our model, which is continuous and usually a sparse covariate in single-cell experiments. Real-life examples of time covariates include the elapsed time since disease onset, drug treatment (pre-post) and patient age.

In CellDrift, we estimate contrast coefficients $\Delta\beta_{gcp}$ for genes across cell types and perturbation groups at each time point t from the time series $\{1, 2, \ldots, T\}$. Then, for each combination of cell-type c and perturbation group p, we get a series of $\Delta\beta_{gcp}$ across available time points, represented as $\{\Delta\beta_{gcpt}\}_{t\in[0,1,\ldots,T]}$. Each series $\{\Delta\beta_{gcpt}\}_{t\in[0,1,\ldots,T]}$ denotes perturbation coefficients of gene g across time points for the selected cell type and perturbation group, which is the representation of a function or a sample in the following FDA framework.

Temporal pattern identification

There are two general input formats for FDA in our context:

(i) Functional data of various genes in a fixed cell type and perturbation group:

$$F_G = \Big\{\{\Delta\beta_{g_0cpt}\}_{t\in[0,1,\ldots,T]}, \{\Delta\beta_{g_1cpt}\}_{t\in[0,1,\ldots,T]}, \{\Delta\beta_{g_2cpt}\}_{t\in[0,1,\ldots,T]}, \{\Delta\beta_{g_3cpt}\}_{t\in[0,1,\ldots,T]}, \ldots\Big\}.$$

(ii) Functional data of combinations of various cell types and perturbations for a specific gene:

$$F_{CP} = \Big\{\{\Delta\beta_{gc_0p_0t}\}_{t\in[0,1,\ldots,T]}, \{\Delta\beta_{gc_0p_1t}\}_{t\in[0,1,\ldots,T]}, \ldots, \{\Delta\beta_{gc_1p_0t}\}_{t\in[0,1,\ldots,T]}, \{\Delta\beta_{gc_1p_1t}\}_{t\in[0,1,\ldots,T]}, \ldots\Big\},$$

where functional data are denoted as a set of functions.

Our goal is to identify genes (or cell type-perturbation combinations) that show similar dynamic changes in perturbation responses over time, which we refer to as temporal patterns. To find such patterns, we investigated functional clustering algorithms such as KMeans, Fuzzy C-means and EMCluster [30, 31]. Based on the benchmark results, we chose the Fuzzy C-means functional clustering algorithm to identify the temporal patterns of perturbation effects [32] (Supplementary Methods). For example, when clustering on F_G, the resultant cluster can be interpreted as a group of genes with a similar pattern of perturbation response over time.

Additional algorithm, evaluation, simulation and FDA details are provided in Supplemental Methods.

Results

CellDrift GLM improves perturbed gene detection

We first demonstrated the performance of CellDrift using simulated datasets (Figures S2 and S3 available online at https://academic.oup.com/bib, Supplementary Methods). True positive rates (TPR or sensitivity) and false discovery rates (FDR) were derived by comparing the ground truth and estimated results (Supplementary Methods). Compared with other commonly used differential expression methods, including t-test, Wilcoxon test and MAST, CellDrift achieved improved sensitivity in the diverse levels of batch effect sizes and differential expression sizes (Figure 2**A** and **B**), which indicates a stronger detection power for perturbed genes using CellDrift.

We observed that CellDrift has a higher FDR in experiments with small batch effects (<0.1). However, it has a stable and controlled FDR at larger batch effect sizes (0.4 and 0.7), where higher FDR was observed in other methods, such as t-test and Wilcoxon (Figure 2**A**). Similarly, MAST has a stable and small FDR of <0.05 across different batch effects, showing the best performance of controlling FDR among all methods, which may be due to the removal of technical covariates, such as batch effects, by the hurdle model [18]. However, its TPR is much lower than CellDrift.

CellDrift outperformed other methods with significantly higher TPR across various differential expression sizes (Figure 2**B**). Meanwhile, we observed a high FDR of CellDrift in experiments with small differential expression sizes (0.05 and 0.2), indicating a relatively inferior performance in controlling false discoveries. MAST had similar results. We argue, however, that it is more important to identify as many DE genes as possible than to avoid false discoveries in datasets with few perturbed genes (Figure 2**B**). FDR of CellDrift decreased with increasing differential expression sizes and achieved a low level (<0.15) in large DE sizes (0.5, 0.8).

Additionally, the CellDrift GLM model also achieved good performances in single-cell pseudo-time data (Figure S4 available online at https://academic.oup.com/bib). More details can be found in the Supplementary Note.

Fuzzy C-means clustering and CellDrift contrast coefficients improve temporal pattern identification and imputation performance

We simulated both linear and nonlinear time patterns of gene perturbation effects (Figure S5 available online at https://academic.oup.com/bib, Supplementary Methods). The temporal pattern recovery performance of three functional clustering algorithms (KMeans, Fuzzy C-means and FPCA-based EMCluster) [33] were examined with CellDrift GLM contrast coefficients as the input (Supplementary Methods). We used adjusted Rand index (ARI) as the metric to measure the accuracy of prediction for simulated temporal patterns (Supplementary Methods). The influence of different parameters, including the number of time points, the ratio of missing time points and noise levels, were evaluated (Figure 2**C** and **D**, Figure S6 available online at https://academic.oup.com/bib).

From the performance of clustering algorithms that used GLM-based contrast coefficients as the input, we observed that FPCA-based EMCluster achieved higher accuracy for a certain number of time points. However,

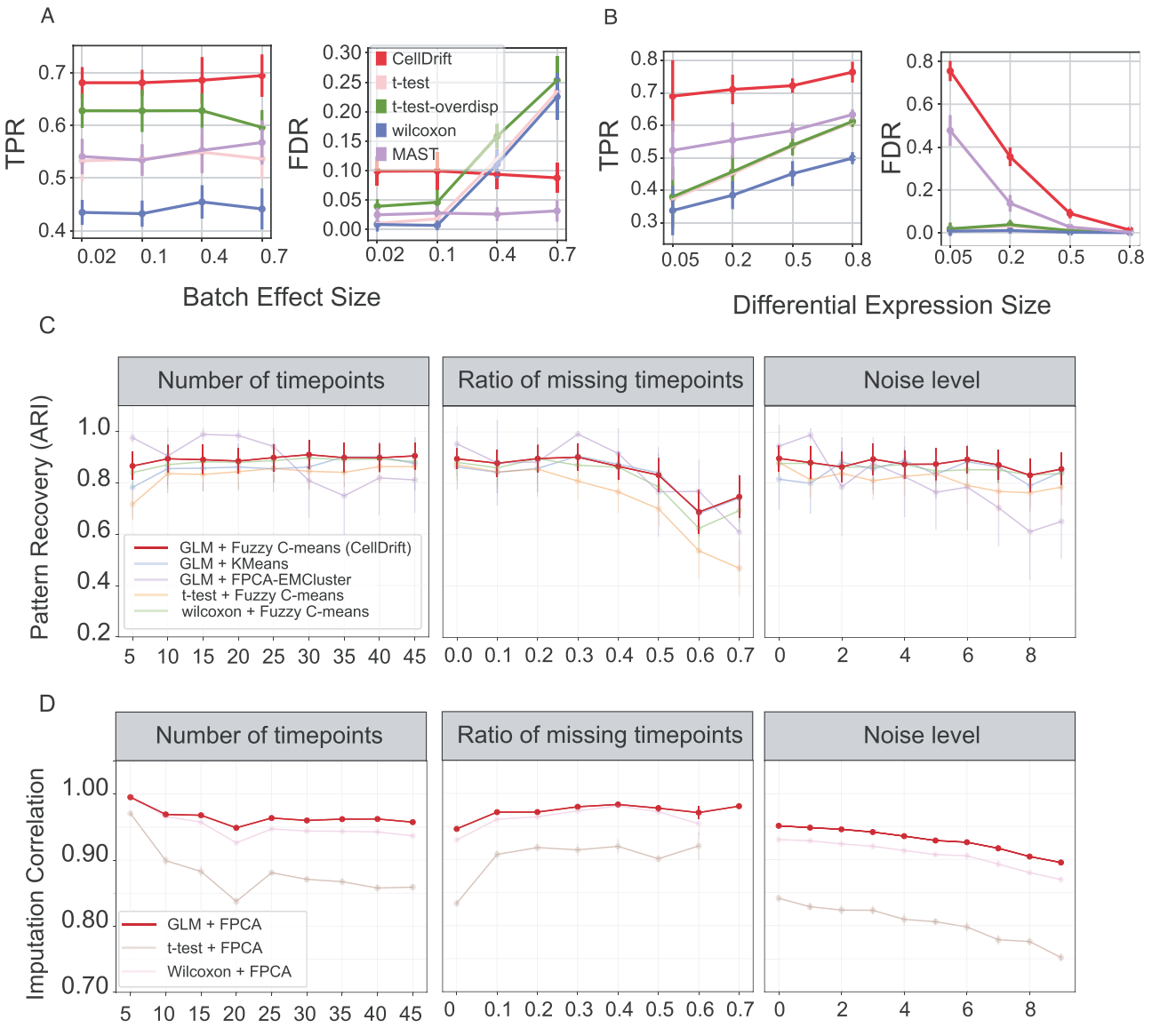

Figure 2. Performance of CellDrift in the identification of perturbed genes and temporal perturbation patterns. (**A**, **B**) Benchmark results of the GLM of CellDrift and other commonly used differential expression approaches in the simulated data. Simulated data with multiple batch effect sizes (**A**) and differential expression sizes (**B**) were used for benchmarking. TPR and FDR were used as metrics. The mean and SD derived from 10 replicates for each test are shown in the figure. (**C**) Benchmark results of temporal pattern recovery for CellDrift strategy (GLM + Fuzzy C-mean) and other approaches. Temporal pattern recovery was evaluated with ARI and measured across varying parameters, including numbers of time points, ratios of missing time points and noise levels. (**D**) Benchmark results of imputation performance using Pearson correlations between imputed contrast coefficients from FPCA and real simulated perturbation coefficients. The same parameters were used as (**C**).

Fuzzy C-means achieved a stable and better performance than most other methods at varying numbers of time points, with ARI reaching 0.9. Additionally, the ARI of Fuzzy C-means exceeded 0.8 at ratios of missing time points <0.5. ARI decreased at greater ratios of missing time points but remained at the 0.7 level and outperformed most other clustering algorithms, such as FPCA-based EMCluster. Note that the ARI of Fuzzy C-mean remained stable at high noise levels with relatively higher accuracy than normal KMeans and other methods (Figure 2**C**). Furthermore, Fuzzy C-mean has shown an improved performance for a variety of non-linear time patterns, pattern coefficient gaps and sequencing depth (Figure S6**A** available online at https://academic.oup.com/bib). In summary, our findings suggest that Fuzzy C-mean has the most stable and relatively better temporal pattern recognition performance.

We also compared GLM-based contrast coefficients with other statistical scores, such as scores from *t*-test and Wilcoxon test, as the input for FDA (Figure 2**C**). The *t*-test scores had an inferior performance in most parameter settings. The performance of Wilcoxon score at various time points is comparable with GLM-based functional clustering. However, its performance in large fractions of missing time points is inferior to GLM-based clustering (Figure 2**C**).

Additionally, we investigated the imputation performance for missing time points, which is a commonly seen situation in real temporal single-cell data.

Incorporated in our pipeline, FPCA provides smoothing and interpolation functions. We compared GLM-based input with statistical scores from *t*-test and Wilcoxon test, where we observed significant improvement in the imputation performance using GLM-based input (Figure 2**D** and Figure S6**B** available online at https://academic.oup.com/bib).

Furthermore, we demonstrated that CellDrift can identify temporal patterns in complex datasets with a variety of cell types that display distinctive perturbation responses (Figure S7 available online at https://academic.oup.com/bib). Moreover, CellDrift achieves a reasonable performance of temporal pattern identification in datasets with a small number of cells or time points (Figure S8 available online at https://academic.oup.com/bib, Supplementary Notes).

CellDrift identified temporal patterns of COVID-19 immune responses

We next demonstrated CellDrift by identifying the temporal patterns of immune responses in a large-scale COVID-19 PBMC single-cell dataset [1]. Samples were derived from COVID-19 patients with different conditions as well as from severe influenza and sepsis patients. The six most common cell types were extracted for our analysis. Disease progression time ranged from 0 to 25 days from disease onset (Figure 3**A**).

Compared with recent neural network-based autoencoder methods, such as scGen and CPA, we are able to investigate the gene-level perturbation effects across the time covariate. We first focused on the classical monocytes in severe COVID-19 patients and applied CellDrift for all genes after the feature selection (Supplementary Table S1 available online at https://academic.oup.com/bib). Genes with similar temporal patterns of perturbation responses clustered together, indicating that multiple gene patterns of dynamic changes were occurring upon virus infection during disease progression. The genes responding to perturbations in clusters 11, 13 and 17 showed three distinct temporal patterns, where the contrast coefficients of clusters 11 and 17 showed a positive and negative correlation with time, while cluster 13 showed an insensitive pattern to time (Figure 3**B**). Based on gene enrichment results, cluster 11 is highly associated with catabolic and biosynthesis processes, while cluster 17 appears to be involved in immune responses, indicating a rapid activation of immune response activities and a suppression of house-keeping activities in the early disease stage (d1~d15), with a reduced level of such changes in later stages (after d15). Additionally, the functional PCA also showcased distinct patterns of perturbation responses during disease progression (Figures S9–S11 available online at https://academic.oup.com/bib, Supplementary Notes).

As a further test, we applied CellDrift to multiple cell types in the COVID-19 dataset to investigate the cell-type-dependent perturbation responses (Figure S12**A** available online at https://academic.oup.com/bib). Notably, distinct biological activities were highlighted in the temporal perturbation patterns across cell types (Figure S12**B** available online at https://academic.oup.com/bib). For example, lymphocytes, including CD4+ T cells, CD8+ T cells and NK cells, showcased the up-regulated patterns of cell cycle and division. Myeloid cells, including classical monocytes and non-classical monocytes, presented the up-regulation of vesicle and membrane fusion, indicating the synthesis of proteins against the virus infection. These activities have been reported in the literature [34, 35]. In addition to the cell-type resolution used above, we also applied CellDrift on CD4+ T cell subpopulations and identified the subpopulation with the strongest temporal perturbation responses (Figure S13 available online at https://academic.oup.com/bib, Supplementary Notes).

After we obtained temporal patterns in severe COVID-19 patients, we further examined whether the patterns vary across multiple perturbation groups, such as mild and severe COVID-19 patients. To achieve it, we applied functional analysis in classical monocytes of multiple disease conditions, where dynamic time warping was used to align multiple time series into a comparable time scale (Supplementary Methods). Next, we applied a one-way functional analysis of variance (ANOVA) test and calculated the ANOVA scores for each gene, representing the consistency of perturbation responses between disease conditions over time (Supplementary Methods). Based on ANOVA results, a number of genes from cluster 17 were identified as severe-prominent genes, including S100A8, S100A9, CTSD and others (Figure 3**C**). In agreement with our findings, elevated levels of calprotectin (S100A8/S100A9) have been found in severe COVID-19 patients with poor clinical outcomes [35, 36]. Apart from severe-prominent genes, we also prioritized mild-prominent genes and condition-irrelevant genes, representing distinct gene programs of temporal perturbation responses across disease conditions (Figure S14 available online at https://academic.oup.com/bib).

To validate our discoveries, we also applied CellDrift to the data from another large-scale COVID-19 single-cell experiment [1]. We observed similar temporal patterns between the mild and severe COVID-19 patients as shown in Figure 3**D**, which shows the reproducibility of CellDrift approach. Moreover, we retrieved the control and critical COVID-19 gene modules from the literature as the external knowledge [37] and evaluated their associations with CellDrift-derived temporal patterns. Notably, temporally up-regulated patterns are strongly associated with critical gene modules, while temporally down-regulated patterns are strongly related to control modules (Figure S15 available online at https://academic.oup.com/bib, Supplementary Methods).

Similarly, CellDrift also identified temporal immune response patterns in an HIV post-infection study (Figure S16 available online at https://academic.oup.com/bib)

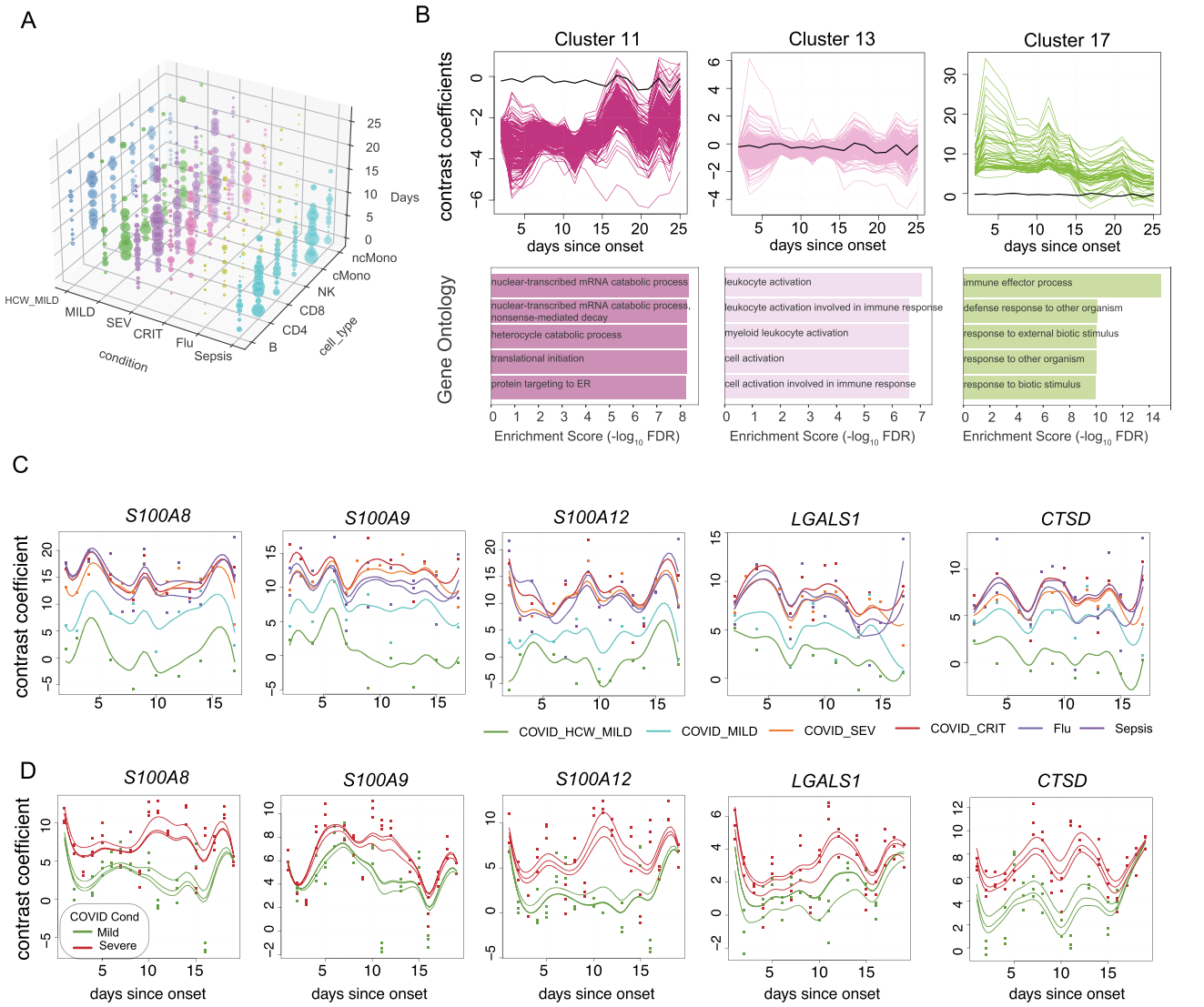

Figure 3. Temporal perturbation effects in COVID-19 atlas. (**A**) Overview of the number of cells in each cell group of the dataset, which contains multiple disease conditions, cell types and time points from days 1 to 25 since the disease onset. The size of dots represents the number of cells. HCW_MILD: healthcare workers with mild COVID-19; MILD, SEV, CRIT: mild, severe and critical COVID-19; CD4, CD8: CD4 T cell, CD8 T cell; cMono, ncMono: classical and non-classical monocyte. (**B**) Three distinct temporal patterns of contrast coefficients from classical monocytes of severe COVID-19 patients. The top row shows curves of genes with similar contrast coefficients in each cluster over time, and the bottom row shows the gene set enrichment analysis of genes in each temporal cluster. The black line represents the average time curve for all genes, which is the same across three plots. Gene enrichment scores are defined as $-\log_{10}$FDR-adjusted *P*-values of enrichment significance. (**C**) Five genes from cluster 17 were prioritized by the functional ANOVA test, which have significantly higher temporal curves in severe conditions than mild symptoms. Contrast coefficients for classical monocytes across disease conditions are shown on smoothed curves computed by FPCA, and time curves were aligned using dynamic time warping. (**D**) Validation of genes from (**C**) with another large-scale COVID-19 PBMC data [1]. FPCA smoothed curves for genes in three replicates of mild and severe patients are shown, which display similar temporal patterns as (**C**).

[14]. More information can be accessed in the Supplementary Notes.

CellDrift discovered differential temporal gene patterns during fetal gut development

We further applied CellDrift to a single-cell fetal gut cell atlas to identify differential gene programs during organ development [38]. Researchers examined gut development in three compartments, including duo-jejunum, ileum and colon, at nine time points during development from the embryonic stage (week 6) to the fetal stage (week 11) (Figure 4**A**). We selected the epithelial and mesenchymal cells from all three compartments and used duo-jejunum as a reference compartment in the GLM model. Differential gene programs during development between the colon (or ileum) and duo-jejunum were identified by the GLM across time points (Supplementary Table S2 available online at https://academic.oup.com/bib).

The top two eigenfunctions from the subsequent FPCA step explain >99% of the temporal variance of mesenchymal cells between the colon and duo-jejunum, where the first eigenfunction (ϕ_1) shows reverse temporal patterns during the development (Figure 4**B**, Figure S17 available online at https://academic.oup.com/bib). CellDrift identified 20 temporal clusters (Figure S18 available

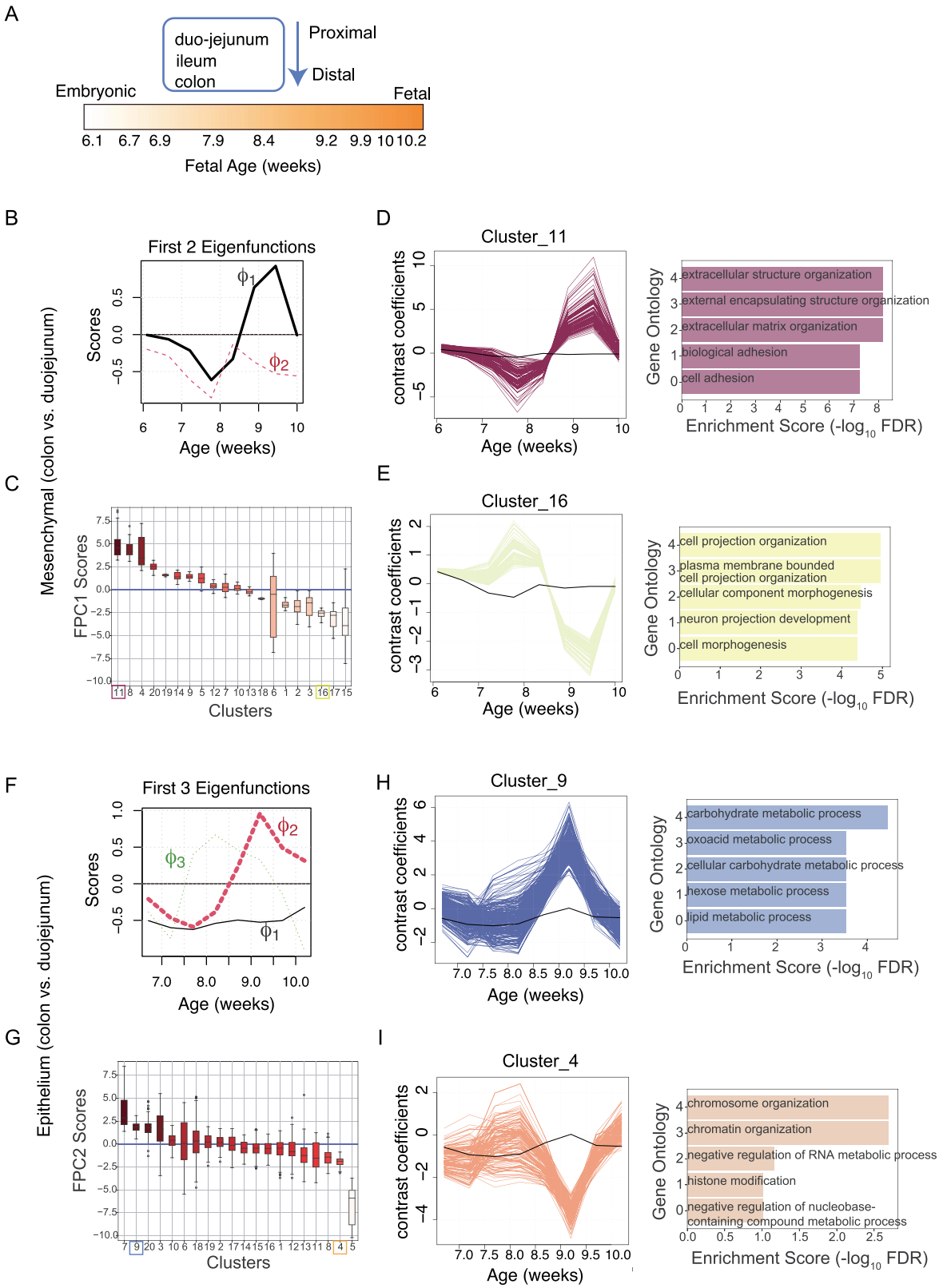

Figure 4. Differential temporal trajectories of gut development. (**A**) Overview of fetal gut cell atlas. The study contains single-cell sequencing data for three compartments in gut development, including duo-jejunum, ileum and colon, whose transcriptomics profiles were retrieved from weeks 6 to 11. (**B**) The top 2 FPCA eigenfunctions and their scores over time for functional data of comparisons of colon and duo-jejunum in mesenchymal cells. First eigenfunction ϕ_1 presents a strong temporal pattern as is highlighted. (**C**) Functional principal component 1 (FPC1) scores were computed for genes in each cluster, displayed in decreasing order. High positive scores indicate genes in the cluster may have similar temporal patterns as ϕ_1, while large negative scores indicate opposite temporal patterns as ϕ_1. Clusters 11 and 16 are highlighted for the following analysis. (**D**, **E**) FPCA smoothed curves for genes in cluster 11 (**D**) and 16 (**E**) and their gene enrichment results. (**F**) The top three FPCA eigenfunctions and their scores over time for functional data of comparisons of colon and duo-jejunum in epithelium cells. The second eigenfunction ϕ_2 presents a strong temporal pattern as is highlighted. (**G**) Functional principal component 2 (FPC2) scores were computed for genes in each cluster, which was ranked in a decreasing order. Clusters 4 and 9 are highlighted for the following analysis. (**H**, **I**) FPCA smoothed curves for genes in clusters 9 (**H**) and 4 (**I**) and their gene enrichment results.

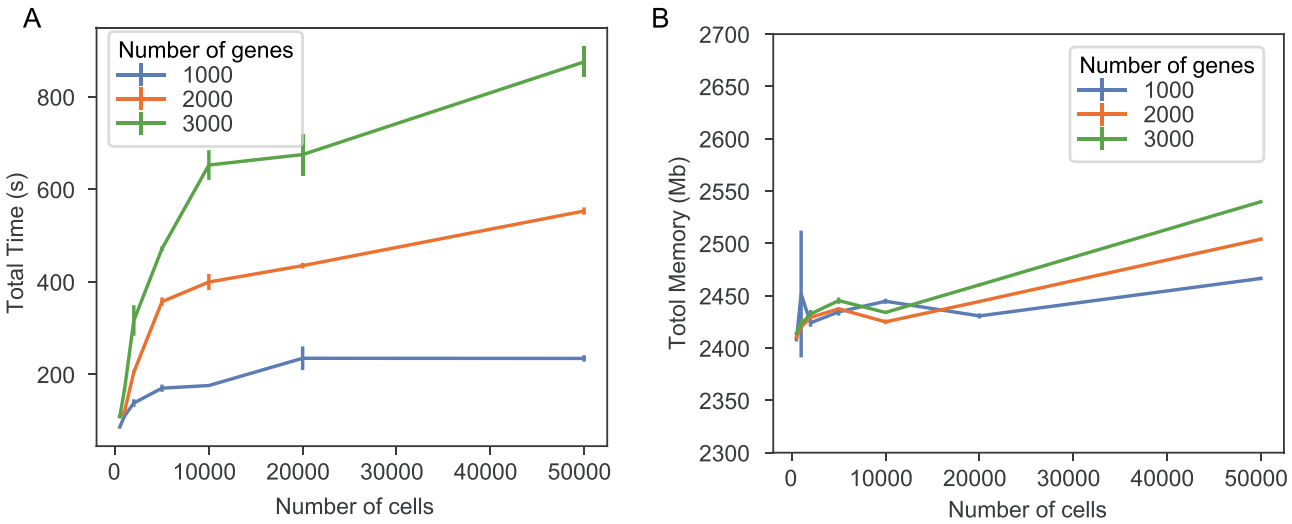

Figure 5. Computational performance of CellDrift. (**A**, **B**) Total amount of running time (**A**) and memory usage (**B**) of the CellDrift application in single-cell datasets with various numbers of cells and genes. The mean and standard error values were calculated for three runs in each condition.

online at https://academic.oup.com/bib) and ranked them by FPC1 scores, in which clusters 11 and 16 had high positive and negative correlations with FPC1 scores (Figure 4**C**). Follow-up gene enrichment analyses indicate that extracellular matrix organization genes are more active in early stages (weeks 7–8) in the duo-jejunum and then in later stages (weeks 9–10) in the colon, whereas morphogenesis genes are more prominent in the distal tissues (colon) at the beginning and proximal (duo-jejunum) later on.

Similarly, such time-dependent differentiation gene programs were also identified in the epithelium cells by comparing the colon and duo-jejunum (Figure 4**F**–**I**, Figures S17 and S18 available online at https://academic.oup.com/bib), revealing temporal patterns that appear like waves from the proximal to distal compartments throughout the gut development.

Additionally, we investigated genetic effects on the neuron development trajectory using the pseudotime as the time covariate of CellDrift. We identified temporal patterns that indicated the immaturity of projection neurons caused by the mutation of an autism-risk gene (Figures S19 available online at https://academic.oup.com/bib, Supplementary Notes).

Computational performance of CellDrift

We evaluated the speed and memory usage of CellDrift in single-cell datasets with different sizes (Figure 5). Tests were conducted on the 8 GB MacBook Pro (2.3 GHz Intel Core i5) using eight cores of CPU. The running time and memory usage were positively correlated with the number of cells and genes. With datasets of 50000 cells and 3000 features, CellDrift completed the task within 15 minutes (Figure 5**A**). The memory usage of CellDrift depends on the number of cells and features. For large-scale datasets, it took <3 GB of memory (Figure 5**B**).

Discussion

In this study, we presented a framework to identify the temporal patterns of perturbation responses. As far as we know, CellDrift is the first method to use FDA in the evaluation of longitudinal perturbation effects in single-cell data, with the advantages of investigating gene-level temporal perturbation effects. Using GLMs, we modeled perturbational single-cell data and introduced the new concept of cell type-perturbation interaction, which improves the sensitivity of detecting both common and cell-type-specific perturbation effects in real-life single-cell experiments. As a result of allowing for batch covariates, we address a significant barrier to finding real perturbed genes. Unlike currently available single-cell methods, which either focus on temporal analysis or perturbation investigation, we utilized the flexibility of GLM and FDA to combine these two areas together and gained insights into complicated longitudinal perturbation responses.

In our study, we successfully improved TPR in multiple settings of batch effects and perturbation effect size compared with the popular methods in differential expression analysis, enabling the capture of more perturbed features. The FDR is insensitive to varying batch sizes, indicating the successful repression of batch effects by CellDrift. Notably, although the performance of a GLM with the Fuzzy C-mean was not uniformly superior in identifying temporal patterns, it was the most stable approach and performed well in the majority of benchmark experiments. Gaussian process (GP) has been found to be effective in inferring temporal patterns from single-cell data [21]. FDA was selected over the option of GPs in this study because of its flexibility and versatile functions, including smoothing curves, FPCA, one-way ANOVA tests and newly implemented deep learning methods [39, 40]. Nevertheless, we are interested in exploring the application of GP in the analysis of

temporal perturbational data in the future. Compared with the autoencoder network approach in CPA, CellDrift is able to investigate the gene-level perturbation effects across the temporal covariate. Despite the dramatically reduced dimensions in the latent space of autoencoder models, it can be difficult or impossible to understand the shift in the latent space in perturbation settings. Instead, we utilized the power of FPCA to reduce the complexity of temporal patterns of perturbation responses and to interpret the contribution of genes in the reduced dimensions.

The cost of sample collection and single-cell sequencing technology is still one of the major obstacles to collecting more longitudinal single-cell data. Yet, we are beginning to see more large-scale longitudinal single-cell experiments due to the popularity of single-cell sequencing technology and the progress of organoid research [11]. We have demonstrated the effective performance of CellDrift in the identification of temporal patterns of gene perturbation effects. These temporal changes could be used in conjunction with the clinical events of patients and to facilitate the application of machine learning methods, such as *k*-nearest neighbors, to predict the possibility of certain clinical events of patients before they happen. Notably, other applications of FDA, such as extrapolation and kernel regression, can greatly enhance our ability to evaluate temporal perturbation effects.

There are several important areas that CellDrift and this evaluation do not address. First, we did not establish the effectiveness of CellDrift in studies with complicated temporal patterns. More sophisticated longitudinal data should be incorporated from real-life experiments and simulations in the future. Additionally, we did not include time as a covariate in the GLM. Instead, the contrast coefficient information was combined from GLM runs of separate time points, which might result in lower statistical power or in increased probability of a type 1 error as well as making the CellDrift procedure cumbersome. This may be an area for future improvement. Additionally, we did not introduce covariance between genes, which would reduce the power of detecting gene correlations of perturbation effects.

Authors' contributions

Conceptualization was by K.J, N.S. and B.J.A.; methodology was taken care of by K.J., R.S., D.S. and V.B.S.P.; investigation was done by K.J., G.L. and D.S.; writing was the responsibility of K.J., D.S., G.L., R.S. and B.J.A.; editing was done by D.S., G.L., R.S. and B.J.A.; funding acquisition was done by B.J.A.; resources and data curation were the responsibility of K.J. and R.S.; visualization was by K.J.; and supervision was done by B.J.A.

Key Points

- GLM with the interaction term enabled the investigation of cell-type-specific perturbation effects.
- FDA vastly enhanced the flexibility of temporal analysis of perturbation effects.
- GLM-based Fuzzy C-means clustering in CellDrift outperforms other methods in temporal pattern identification of perturbation effects.
- Predicted temporal patterns of immune cell responses toward COVID-19 represented time-dependent gene programs.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

We express our appreciation to Dr Kieran R. Campbell and Dr Catherine Blish for their valuable suggestions on the method design. We appreciate suggestions from Dr Emrah Gecili for the manuscript revision. Some figures in Figure 1 were generated from BioRender.

Funding

National Institutes of Health LungMap (HL148865); Digestive Health Center (DK078392); National Cooperative Reprogrammed Cell Research Program (MH104172); Center of Excellence in Molecular Hematology (DK126108).

Data availability

The source code and Python package are freely available at https://github.com/KANG-BIOINFO/CellDrift. The analysis performed in this paper can be found at https://github.com/KANG-BIOINFO/CellDrift_analysis. The source data can be found in Supplementary Table S3.

References

1. Ren X, Wen W, Fan X, *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 2021;**184**(23):5838.
2. Tucker NR, Chaffin M, Fleming SJ, *et al.* Transcriptional and cellular diversity of the human heart. *Circulation* 2020;**142**(5):466–82.
3. Habermann AC, Gutierrez AJ, Bui LT, *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* 2020;**6**:eaba1972.
4. Reyfman PA, Walter JM, Joshi N, *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;**199**(12):1517–36.
5. Reyes M, Filbin MR, Bhattacharyya RP, *et al.* An immune-cell signature of bacterial sepsis. *Nat Med* 2020;**26**(3):333–40.
6. Guo C, Li B, Ma H, *et al.* Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat Commun* 2020;**11**(1):3924.

7. Bassez A, Vos H, Van Dyck L, *et al.* A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nat Med* 2021;**27**:820–32.
8. Paulsen B, Velasco S, Kedaigle AJ, *et al.* Human brain organoids reveal accelerated development of cortical neuron classes as a shared feature of autism risk genes. *bioRxiv* 2020. https://doi.org/10.1101/2020.11.10.376509.
9. Angelidis I, Simon LM, Fernandez IE, *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat Commun* 2019;**10**(1):963.
10. Zheng Y, Liu X, Le W, *et al.* A human circulating immune cell landscape in aging and COVID-19. *Protein Cell* 2020;**11**:740–70.
11. Ungricht R, Guibbal L, Lasbennes M-C, *et al.* Genome-wide screening in human kidney organoids identifies developmental and disease-related aspects of nephrogenesis. *Cell Stem Cell* 2022;**29**(1):160–75.e7.
12. Paulsen B, Velasco S, Kedaigle AJ, *et al.* Autism genes converge on asynchronous development of shared neuron classes. *Nature* 2022;**602**(7896):268–73.
13. Wimmers F, Donato M, Kuo A, *et al.* The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination. *Cell* 2021;**184**(15):3915–35.e21.
14. Kazer SW, Aicher TP, Muema DM, *et al.* Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. *Nat Med* 2020;**26**(4):511–8.
15. Liu C, Martins AJ, Lau WW, *et al.* Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* 2021;**184**(7):1836–57.e22.
16. Ji Y, Lotfollahi M, Wolf FA, *et al.* Machine learning for perturbational single-cell omics. *Cell Syst* 2021;**12**(6):522–37.
17. Squair JW, Gautier M, Kathe C, *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**(1):5692.
18. Finak G, McDavid A, Yajima M, *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015;**16**(1):278.
19. Crowell HL, Soneson C, Germain P-L, *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun* 2020;**11**(1):6077.
20. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods* 2019;**16**(8):715–21.
21. Velten B, Braunger JM, Argelaguet R, *et al.* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat Methods* 2022;**19**(2):179–86.
22. Lotfollahi M, Susmelj AK, De Donno C, *et al.* Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv* 2021. https://doi.org/10.1101/2021.04.14.439903.
23. Yuan B, Shen C, Luna A, *et al.* CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell Syst* 2021;**12**(2):128–140.e4.
24. Agresti. Categorical data analysis. John Wiley & Sons, 2003.
25. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**(1):296.
26. Wang J-L, Chiou J-M, Müller H-G. Functional data analysis. *Annu Rev Stat Appl* 2016;**3**(1):257–95.
27. Kokoszka P, Reimherr M. *Introduction to Functional Data Analysis*. Chapman and Hall/CRC. 2017.
28. James GM, Hastie TJ, Sugar CA. Principal component models for sparse functional data. *Biometrika* 2000;**87**:587–602.
29. Gecili E, Huang R, Khoury JC, *et al.* Functional data analysis and prediction tools for continuous glucose-monitoring studies. *J Clin Transl Sci* 2020;**5**:e51.
30. Tokushige S, Yadohisa H, Inada K. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Comput Stat* 2007;**22**(1):1–16.
31. Luz López García M, García-Ródenas R, González GA. K-means algorithms for functional data. *Neurocomputing* 2015;**151**:231–45.
32. Ramos-Carreño C, Torrecilla JL, Suárez A. scikit-fda: a Python package for functional data analysis. Different varimax rotation approaches of functional PCA for the evolution of COVID-19 pandemic in Spain. 2019;**55**(1).
33. Chen WC, Maitra R, Melnykov V. EMCluster: EM algorithm for model-based clustering of finite mixture gaussian distribution. R Package, URL http://cran. r-project. org/package= EMCluster 2012.
34. Cao X. COVID-19: immunopathology and its implications for therapy. *Nat Rev Immunol* 2020;**20**(5):269–70.
35. Silvin A, Chapuis N, Dunsmore G, *et al.* Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. *Cell* 2020;**182**(6):1401–18.e18.
36. Chen L, Long X, Xu Q, *et al.* Elevated serum levels of S100A8/A9 and HMGB1 at hospital admission are correlated with inferior clinical outcomes in COVID-19 patients. *Cell Mol Immunol* 2020;**17**(9):992–4.
37. Bernardes JP, Mishra N, Tran F, *et al.* Longitudinal multi-omics analyses identify responses of megakaryocytes, erythroid cells, and plasmablasts as hallmarks of severe COVID-19. *Immunity* 2020;**53**(6):1296–314.e9.
38. Elmentaite R, Ross A, James KR, *et al.* Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. Dev. Cell 2020; **55**:771–783.e5.
39. Li K, Daniels J, Liu C, *et al.* Convolutional recurrent neural networks for glucose prediction. *IEEE J Biomed Health Inform* 2020;**24**(2):603–13.
40. Pérez-Gandía C, Facchinetti A, Sparacino G, *et al.* Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technol Ther* 2010;**12**(1):81–8.