Temporal transcriptomic analysis using TrendCatcher identifies early and persistent neutrophil activation in severe COVID-19

Xinge Wang, … , Yang Dai, Jalees Rehman


Graphical abstract

Find the latest version:
https://jci.me/157255/pdf
Temporal transcriptomic analysis using TrendCatcher identifies early and persistent neutrophil activation in severe COVID-19

Xinge Wang¹,²,³, Mark A. Sanborn¹,²,³, Yang Dai¹ and Jalees Rehman¹,²,³*

¹Department of Biomedical Engineering, University of Illinois Colleges of Engineering and Medicine, Chicago, IL, USA
²Department of Pharmacology and Regenerative Medicine, University of Illinois College of Medicine, Chicago, IL, USA
³Division of Cardiology, Department of Medicine, University of Illinois College of Medicine, Chicago, IL, USA

*Please address correspondence to Jalees Rehman (jalees@uic.edu)

The University of Illinois at Chicago, College of Medicine (M/C 868)
835 S. Wolcott Ave. Rm. E403, Chicago, IL 60612, USA
Office phone: +1 312-996-5552; Fax: +1 312-996-1225

Conflict-of-interest statement
The authors have declared that no conflict of interest exists.
Abstract

Studying temporal gene expression shifts during disease progression provides important insights into the biological mechanisms that distinguish adaptive and maladaptive responses. Existing tools for the analysis of time course transcriptomic data are not designed to optimally identify distinct temporal patterns when analyzing dynamic differentially expressed genes (DDEGs). Moreover, there is a lack of methods to assess and visualize the temporal progression of biological pathways mapped from time course transcriptomic datasets. In this study, we developed an open-source R package TrendCatcher (https://github.com/jaleesr/TrendCatcher), which applies the smoothing spline ANOVA model and break point searching strategy to identify and visualize distinct dynamic transcriptional gene signatures and biological processes from longitudinal datasets. We used TrendCatcher to perform a systematic temporal analysis of COVID-19 peripheral blood transcriptomes, including bulk and single-cell RNA sequencing time course data. TrendCatcher uncovered the early and persistent activation of neutrophils and coagulation pathways as well as impaired type I interferon (IFN-I) signaling in circulating cells as a hallmark of patients who progressed to severe COVID-19, whereas no such patterns were identified in individuals receiving SARS-CoV-2 vaccinations or patients with mild COVID-19. These results underscore the importance of systematic temporal analysis to identify early biomarkers and possible pathogenic therapeutic targets.
Introduction

Time course transcriptomic profiling has been widely used to study and model dynamic biological processes in cells (1). By profiling mRNA levels during consecutive time points, researchers can infer dynamic responses to various external cues that cannot be observed by looking at only initial and terminal states. Recent improvements in high-throughput RNA sequencing (RNA-seq) technologies, including single-cell RNA-sequencing (scRNA-seq) provide viable approaches to study dynamic gene expression changes (2). Especially scRNA-seq allows for the in-depth analysis of temporal changes in distinct cell populations, thus providing insights into the heterogeneity and dynamics of responses to environmental cues or pathogenic stimuli. However, the analysis and visualization of longitudinal bulk RNA-seq data or scRNA-seq can be computationally challenging.

Currently, there are two predominant strategies for the analysis of sequential transcriptomic datasets. One strategy treats the sampling time points as categorical variables and is based on generalized linear models (GLMs). GLM-based packages include DESeq2 (3), edgeR (4) and limma (5). A complementary strategy is to treat time as a continuous variable and fit the time expression data into a spline-like model. These methods include DESeq2Splline (DESeq2 adopted with spline model for temporal RNA-seq datasets) fitting, ImpulseDE2 (6) and NextmaSigPro (7). The former strategies focus on the magnitude of change instead of the time order of gene expression, and may also suffer from a relative loss of statistical testing power, especially if many time points are assessed (6). The latter strategy increases the power of detecting dynamic genes, but is based on strong model assumptions which are not optimally suited for multiphasic responses, such as gene trajectory patterns that reflect initial acute stimuli followed by counter-regulatory compensatory responses. Furthermore, there is a lack of tools that leverage existing knowledge of functional pathway databases to infer and visualize pathway trajectories instead of individual gene trajectories only.
We found this is challenging in a complex disease, such as coronavirus disease 19 (COVID-19) caused by the SARS-CoV-2 virus infection (8). COVID-19 is characterized by distinct disease progression patterns that suggest diverse host immune responses (9). Patients with severe disease exhibit profound inflammatory responses and immunopathology (10). COVID-19 immunophenotyping studies involve a large number of time points from corresponding RNA-seq and scRNA-seq datasets (11, 12). Existing approaches for temporal transcriptomic analysis either do not take the temporal sequence into account when identifying dynamic differentially expressed genes (DDEGs) using pairwise comparison between time points, nor systematically analyze pathways that are dysregulated at defined time points.

In this study, we developed TrendCatcher, an open-source R package (https://github.com/jaleesr/TrendCatcher) tailored for longitudinal bulk RNA-seq and scRNA-seq analysis. TrendCatcher uses a framework that combines the smooth spline ANOVA model and break point searching strategy, which identifies inflection points when gene expression trends reverse. We show that TrendCatcher outperformed commonly used methods for longitudinal RNA-seq analysis when using simulated time course data for benchmarking. We also analyzed bulk RNA-seq and scRNA-seq gene expression profiles of peripheral blood cells in COVID-19 patients at various disease time points. TrendCatcher allowed us to identify and visualize dynamic gene expression signature profiles in peripheral blood that were associated with poor disease outcomes during the early phases of disease, and could thus serve as novel mechanistic targets as well as early biomarkers for patient prognostication.
Results

TrendCatcher accurately identifies DDEGs in simulated datasets

First, we tested the prediction performance of the TrendCatcher platform (Figure 1A) using a set of simulated time course RNA-seq datasets, because simulated data provides defined standards to assess the accuracy of novel analytical platforms. We considered a comprehensive collection of gene temporal trajectory patterns to simulate a set of realistic data with biological characteristics (1). We embedded 10,000 simulated trajectories with varied temporal patterns, including 90% non-dynamic trajectories, 2.5% monotonous transition trajectories (continuously increasing or continuously decreasing gene expression levels throughout the time course), 2.5% impulse shaped single break point trajectories (only one temporal inflection point, i.e. up-peak-down or down-trough-up), 2.5% two break point trajectories and 2.5% three break point trajectories (multimodal dynamic response, e.g., a combination of 2 or more basic types of trajectories). Compared to DESeq2, DESeq2Spline and ImpulseDE2 using receiver operator characteristic (ROC), TrendCatcher had the largest area under the curve (AUC) in a mixed simulated dataset for time course data with 7 time points (Figure 1B). We also tested each model’s performance on a varying number of time points, including 3, 5, 7, 9 and 11 time points. As shown in Figure 1C, TrendCatcher had the highest prediction AUC across all time points, with the AUC values range from 0.88 to 0.90.

We next evaluated the prediction performance for each type of temporal trajectories. As shown in Supplemental Figure 1, A-C, although other three methods achieved slightly higher accuracy than TrendCatcher in monotonic trajectories, their AUCs dropped markedly when more complicated trajectories were embedded. DESeq2 only achieved AUC values of 0.49 to 0.62 for both bi-phasic trajectory and multimodal trajectory. The DESeq2Spline approach using a spline curve fitting model also dropped to an AUC of approximately 0.7 once multiphasic trajectories were introduced. These results suggests that existing approaches for longitudinal or time course analyses are well-suited for monotonic trajectories (continuously
up or continuously down) but that TrendCatcher maybe more broadly applicable, because it identifies monotonic, biphasic (up-down, down-up) and multi-phasic shifts in gene expression, which are especially important in complex pathological setting when initial biological responses are followed by counter-regulatory adaptive or maladaptive responses.

**TrendCatcher identifies rapid but transient upregulation of interferon signaling in peripheral blood following SARS-CoV-2 infection in a non-human primate model**

To define the key dynamic gene signatures associated with SARS-CoV-2 infection in peripheral blood, we first analyzed the global transcriptomics profiles from a non-human primate dataset (13), in which samples were collected on day 0 (uninfected controls), 1, 2, 4, 7, 10 and 14 following the live SARS-CoV-2 inoculation. This experiment in non-human primates had the advantage of clearly defining the timing of inoculation and following the time course of dynamic genes. TrendCatcher identified 962 DDEGs out of 12,754 total genes, accounting for 7.6% of total expression, suggesting that over 90% of the expressed genes in the peripheral blood remain close to the baseline expression levels even in the setting of the SARS-CoV-2 infection. We observed two major types of dynamic trajectories: (a) 167 genes followed a biphasic “0D-2D up, 2D-14D down” pattern, with their expression level peaking at day 2 and gradually returning close to baseline levels at day 14 (Figure 2A). These dynamic genes were primarily associated with host defense biological pathways annotated by gene ontology (GO), such as defense response to viruses, regulation of viral life cycle and type I interferon signaling pathways (Figure 2B). (b) 263 genes follow a “0D-14D down” pattern (Figure 2A). This set of genes followed a monotonous trajectory, with their expression gradually decreasing until day 14. Interestingly, we found these genes were primarily associated with mitochondrial ATP synthesis and oxidative phosphorylation, suggesting that disruption of mitochondrial respiration and ATP generation may be an important feature of the circulating cells transcriptomic shift in a SARS-CoV-2 infection (Figure 2C). TrendCatcher
assigned trajectory pattern types to all dynamic genes, and provided hierarchical pie charts to visualize the composition of trajectory patterns (Supplemental Figure 2A).

To systematically assess and visualize the dynamic programming of the top biological pathways associated with SARS-CoV-2 infection, TrendCatcher generated a TimeHeatmap. The TimeHeatmap function of TrendCatcher visualizes shifts in pathways by displaying the mean-fold change of individual DDEGs in a given pathway at defined time points while also depicting the number of dynamic genes and the percentage of dynamic genes within that pathway (Figure 2D). This quantifies the pathway level shifts over time, and allows for the magnitude of pathway change to be visualized. Together with the number and fraction of dynamic genes in a given pathway, to gauge the relative importance of the pathway during the time course.

During initial infection (day 0 to day 2), pathways related to innate immune response and interferon pathways were highly upregulated. Examples are upregulation of pathways such as defense response to virus and regulation of innate immune response increased with an average log2 fold change of 2.57 and 1.76 within the first day. Mucosal immune response, antimicrobial humoral response and killing of cells of other organisms were activated during the later stage of infection (day 4 to day 7), with an average log2 fold change around 2. On the other hand, mitochondrial ATP synthesis coupled electron transport and protein targeting to ER, on the other hand, were gradually down-regulated until day 14. Dynamic gene signatures from type I interferon signaling pathway and mitochondrial ATP synthesis coupled electron transport were shown using traditional heatmaps in Supplemental Figure 2, B and C.

The temporal analysis of this non-human primate dataset highlights the rapidity of interferon signaling activation but also underscores that this initial burst of immune activation is transient and followed by a gradual downregulation of the antiviral interferon responses during the first week following infection.
Increased generation of immunoglobulin synthesis in plasma B cells as early as day 1 of symptom onset in patients diagnosed with SARS-CoV-2 infection

Next, we analyzed the longitudinal gene expression profiles of peripheral blood mononuclear cells (PBMCs) obtained from patients diagnosed with a SARS-CoV-2 infection who were admitted to the hospital, but predominantly had uncomplicated disease progression with 4 out of 5 patients showing only mild symptoms (14). The study performed single cell RNA-seq analysis on PBMCs (Figure 3A), thus allowing for a cell-type specific analysis of gene expression shifts in distinct B cell subtypes, T cell subtypes and NK cells. We adopted the cell-type labels from the original study and generated pseudo-bulk RNA-seq datasets for each cell type in order to quantify changes in DDEGs for a specific cell type. We also adopted the time annotation using stages, by binning the disease processes of COVID-19 patients from symptom onset to discharge into stages 0, 1, 2, 3 and 4. We defined stage 0 as the time point of samples obtained from healthy controls and stages 1, 2, 3 and 4 represented day 1 to day 16 from symptom onset. We only applied TrendCatcher to cell types containing more than 1,000 cells at any given stage to ensure the robustness of the results.

TrendCatcher identified 400 DDEGs in memory B cells, 213 DDEGs in naive B cells, 1,413 DDEGs in plasma B cells, 398 DDEGs in CD4+ T cells, 645 DDEGs in CD8+ T cells, 423 DDEGs in MAIT, 1,161 DDEGs in naive T cells and 667 DDEGs in NK cells. The TimeHeatmap of plasma B cells visualized the most dynamic biological pathways. Importantly, this temporal analysis of single cell RNA-seq data showed how rapidly plasma B cells ramp up the upregulation of Fc-gamma receptor signaling and immunoglobulin synthesis as early as stage 1 (which corresponds to day 1 of symptom onset) (Figure 3B). However, not all immunoglobulin genes are upregulated during the same temporal phase. As shown in Supplemental Figure 3, genes involved in immunoglobulin synthesis show distinct temporal patterns. Due to the comparatively lower number of plasma B cells, the prominence of such increased immunoglobulin changes may be diluted in peripheral blood bulk RNA-seq analysis.
or PBMC bulk RNA-seq analysis. However, temporal analysis of all peripheral blood mononuclear cell types, across the whole time course demonstrated increases in type I interferon signaling and defense responses to viruses as the most prominent changes over time (Figure 3C), thus mirroring the responses to SARS-CoV-2 we observed in the peripheral blood of non-human primates (Figure 2). To define cell-type specific temporal dynamics, we generated TimeHeatmaps for individual PBMC cell types and found significant upregulation of type I interferon signaling in T cells, NK cells and memory B cells during stage 1 but subsequent downregulation by stage 2 of the disease in patients with mild symptoms (Supplemental Figure 4, A-C and Supplemental Figure 5, A-D).

**TrendCatcher identifies early and persistent neutrophil activation as a hallmark of severe COVID-19**

We then assessed whether a systematic temporal analysis of gene expression trends could be used to distinguish disease severity and prognosis of COVID-19 patients. We thus applied TrendCatcher to a time course human whole blood bulk RNA-seq dataset (11), which contained longitudinal whole blood transcriptomes from COVID-19 patients with mild, moderate and severe clinical outcomes. TrendCatcher identified 77 DDEGs from the mild group, 226 DDEGs from the moderate group, and 1,205 DDEGs from the severe group. Only 42 DDEGs were shared among these three groups (Figure 4A), and these were primarily associated with B cell mediated immunity, Fc-gamma receptor signaling pathways, humoral immune response and lymphocyte mediated immunity (Figure 4B). However, TrendCatcher identified 978 DDEGs uniquely shown in the severe COVID-19 patient group. Importantly, these genes were strongly enriched for neutrophil-related biological pathways, including neutrophil activation and neutrophil mediated immunity. These severe-disease associated genes also included genes found in pathways such as myeloid cell differentiation, reactive oxygen species metabolic process, and positive regulation of cytokine production (Figure 4B). To compare how DDEGs and pathways identified by TrendCatcher compared to those
identified by other platforms, we also applied three additional computational platforms (DESeq2, DESeq2Spline and ImpulseDE2) to RNA-seq data obtained from different COVID-19 severity groups and set the adjusted P value to 0.05 as a threshold to compare the DDEGs identified by different methods. As shown in the Venn diagram in Supplemental Figure 6, A-C, all methods showed that the number of identified DDEGs increased with disease severity and that the neutrophil activation pathway was highly enriched in the severe COVID-19 group when compared to the mild and moderate Supplemental Figure 6, D-F. While there was broad consistency across platforms in the enriched pathways, the central advantage of the TrendCatcher approach is that it also identifies the time interval when the gene expression enrichment between groups begins to diverge.

To systematically characterize which biological processes in whole blood RNA-seq samples were most dynamic in severe COVID-19 patients and how they progress over time, we applied the TrendCatcher’s TimeHeatmap function. As seen in the non-human primate PBMC and the human COVID-19 single cell RNA-seq datasets, we again observed the upregulation of genes involved in the response to virus, humoral immune response and type I interferon signaling pathway within week 1 of disease onset (Figure 4C). Importantly, some dynamic biological responses were only enriched in the group of patients which subsequently progressed to severe COVID-19, including neutrophil activation (117 DDEGs, 23.4% of the corresponding GO term), blood coagulation (36 DDEGs, 10.5% of the corresponding GO term), and regulation of response to cytokine stimulus and respiratory burst (Figure 4C and Supplemental Figure 7, A and B). For instance, neutrophil activation gene expression increased by a mean of 1.3 log2 units (approximately 2.5-fold increase in mean gene expression) within week 1, increased continuously until week 4, and only very gradually decreased in surviving patients by week 7. However, the summation of the averaged log2 fold change from the TimeHeatmap was larger than zero, which indicates that the neutrophil activation may not have returned fully to baseline levels by 7 weeks. We next applied LOESS smooth curve fitting to all the neutrophil activation DDEGs identified from the three severity groups, and used a permutation test
approach to quantify when and how the gene signatures differed significantly between groups. LOESS fitting confirmed that severe COVID-19 patients showed markedly high neutrophil activation at the early stage of infection (week 1 to week 2), and also remained highly activated even after 7 weeks (Figure 4D). The permutation test function module in the TrendCatcher package identified the time interval when the differences between groups became apparent. We found that persistent neutrophil activation significantly separated the mild COVID-19 group from the severe COVID-19 group as early as the first time interval (day 0 to day 7, Supplemental Figure 8A), whereas the difference in neutrophil activation between the moderate and severe group become apparent only at the beginning of week 2 (Supplemental Figure 8B).

Next, we combined TrendCatcher with the in silico cellular deconvolution tool MuSiC (15) to estimate cellular composition change over time, using a single-cell whole blood dataset (16) as a reference. As shown in Supplemental Figure 9, A and B, each stacked bar chart represents the cellular composition of a given sample. After quantifying estimated neutrophil percentages over time (Supplemental Figure 9C), we found an increase in the neutrophil percentage starting in week 1 in the severe COVID-19 group. We also applied TrendCatcher to the deconvoluted neutrophil gene expression profile using weighted cellular composition, and observed persistent upregulation of neutrophil activation in the severe COVID-19 group (Supplemental Figure 9D), suggesting that the increase in neutrophil activation gene expression is driven by changes intrinsic to neutrophils.

From our bulk RNA analysis, we also observed severe COVID-19 patients showed evidence of greater humoral immune response gene upregulation (Figure 4E), as well as markedly higher upregulation blood coagulation genes (Figure 4F) and respiratory burst genes (Figure 4G), which remained upregulated even after week 6. These were not found to be upregulated in patients with mild or moderate COVID-19 disease. These DDEGs and dynamic biological pathways may thus be suited to serve as early biomarkers that distinguish severe COVID-19
from mild and moderate COVID. All these dynamic gene signatures from neutrophil activation, blood coagulation and respiratory burst pathways were listed using traditional heatmaps (Supplemental Figure 10 and Supplemental Figure 11, A and B).

**Early impaired IFN-I signaling in PBMC provides a hallmark of severe COVID-19**

Next, to define the cell-type specific dynamic gene signatures and biological processes, we used TrendCatcher to analyze a human PBMC single cell RNA-seq dataset time course dataset in which patients were categorized as having either moderate or severe COVID-19 (17). Importantly, as this dataset only contained mRNA from peripheral blood mononuclear cells, it lacked mRNA from neutrophils. TrendCatcher generated “pseudo-bulk” mRNA profiles for each cell type in order to perform the analysis of gene expression dynamics in a cell-type specific manner. TrendCatcher identified more dynamic shifts in almost all cell types from severe COVID-19 patients than moderate groups (Supplemental Table 1 and Supplemental Table 2). To identify dynamic responsive processes unique to the severe group, we compared GO enrichment for each cell type between severe and moderate. As shown in Figure 5A, all innate immune cells and some adaptive immune cells (including B cells, CD8+ T cells) from moderate COVID-19 were highly enriched in type I interferon signaling, negative regulation of viral process and defense to virus, whereas this was not observed significantly enriched in severe COVID-19. Severe COVID-19 patients were highly dynamic in MAPK cascade, NF-κB signaling, T cell receptor signaling and positive regulation of cytokine production for both NK cells and CD4+ T cells. For monocytes and DC (dendritic cells), no uniquely enriched dynamic biological processes were observed in severe COVID-19 patients versus moderate COVID-19 patients.

Furthermore, we also found the extent of the interferon signaling response to be the key distinguishing feature between moderate and severe COVID-19. To quantitatively compare the trajectory differentiation of the type I interferon signaling pathway, we performed the LOESS smooth fitting to the DDEGs identified from this pathway. As shown in Figure 5B, there
is a profound separation between these two groups, with early strong activation of IFN-I signaling in NK cells, monocytes, B cells and CD8* T cells in moderate COVID-19 patients but not in severe COVID-19 patients. Patients with moderate COVID-19 showed activation of type I interferon signaling pathways whereas patients with severe COVID-19 had a blunted activation of type I interferon signaling. This is also shown in the cell-type specific TimeHeatmap. As shown in Figure 5, C and D, although NK cells from both moderate and severe COVID-19 patients demonstrated activation of type I interferon response within the first week, moderate COVID-19 exhibited a stronger activation than severe group, with average 1.62 log2 fold change compared to 0.87 log2 fold change. In CD8* T cells, only moderate COVID-19 groups were observed to have a strong type I interferon response within the first week. On the other hand, CD8* T cells in patients who would go on to develop severe COVID-19 showed upregulation of cell proliferation and cell differentiation genes instead (Figure 5, E and F). Together, these data suggest that a robust early type I interferon response in PBMCs is associated with reduced severity of COVID-19.

**TrendCatcher identifies metabolic gene expression shifts in NK cells as a hallmark response to COVID-19 vaccination**

We next applied TrendCatcher to a longitudinal human PBMC single cell RNA-seq vaccination dataset (18), to provide insights into how the immune system physiologically responds to mRNA vaccines over time. The vaccination study collected single-cell PBMCs from 56 healthy volunteers vaccinated with the Pfizer-BioNtech on day 0, 1, 2, 7, 21, 22, 28, and 42 after the vaccination. We processed each patient’s scRNA-seq individually. We clustered cells using the Seurat algorithm (19) and annotated the cell types using SingleR (20). Then we remove cell types containing less than 1,000 cells for each time point across all samples. As Figure 6A shows, it is the UMAP of one patient’s PBMCs single cell data on day 0.
TrendCatcher identified 650 DDEGs in NK cells, 450 in B cells, 23 in CD8+ T cells, 62 in monocytes and only 6 in CD4+ T cells. This indicates NK cells exhibit a strong dynamic gene expression shift after vaccination. After comparing the GO enrichment analysis across cell types, we found NK cells gene expression shifts were enriched in metabolic processes in response to the Pfizer-BioNtech SARS-CoV-2 mRNA vaccine, such as regulation of cellular amino acid metabolism and ATP metabolism (Figure 6B). As shown in Figure 6C, a TimeHeatmap of NK cells shows 54 DDEGs, which account for nearly half of the ATP metabolic process pathway genes and exhibited a mean of 0.44 log2 fold upregulation after the first dose of vaccination. These findings indicate the reprogramming of metabolism in NK cells may be an indicator of an intact vaccine response. Another hallmark of the vaccine response was the activation of interferon pathways, including type I interferon signaling pathway and response to interferon-gamma (Supplemental Figure 12A). Importantly, the TimeHeatmap demonstrates that these gene expression shifts were very transient, usually decreasing by day 2 or day 7, and were again upregulated when the subjects received the booster vaccine at day 21. Interferon pathways showed an average increase of 0.71 and 1.42 log2 fold change. In adaptive immune cells, we also observed strong type I interferon signaling pathways activated in both B cells and CD8+ T cells (Supplemental Figure 12, B and C). Additionally, B cells demonstrated upregulation of antigen processing and protein synthesis (Supplemental Figure 12B), which were upregulated only for the first day, and then began decreasing. It is noteworthy that early upregulation of interferon pathways in the PBMCs of vaccine recipients mirrors that seen in patients with moderate COVID-19 infection, consistent with the notion that early and transient interferon upregulation is a hallmark of a healthy immune response to the SARS-CoV-2 infection.
Discussion

Temporal analysis of gene expression is gaining importance in the analysis of complex dynamic processes such as disease progression. Besides gene dynamic pattern characterization, time course gene expression data are also used to infer regulatory and signaling relationships among genes (21, 22). Integrating with other different types of measurements, such as pathology and infection over time helps disentangle the complex dynamic processes and possible underlying mediators (23). Thus, accurate identification of DDEGs in a time course RNA-seq or single cell RNA-seq study can help identify time-dependent disease mechanisms, adaptive and maladaptive molecular signatures as well as potential biomarkers that may be associated with disease severity.

In recent years, tools have been implemented to characterize time course RNA-seq data; however, these tools were focused on either bulk RNA-seq datasets (3, 6, 7) and few infer trajectories from scRNA-seq (24, 25). Compared to methods that analyze time course data without considering the sequential nature of time points, modeling time as a continuous function avoids a relative loss of statistical testing power, especially when many multiple time points were studied. There is also a need for methods which combine DDEGs identification with the visualization of dynamic pathway shifts.

In this study, we developed an open-source R software package to perform temporal analysis in longitudinal bulk RNA-seq and scRNA-seq datasets. TrendCatcher can identify DDEGs and infer the trajectories for each DDEG. By combining a time window sliding strategy on inferred gene trajectories and leveraging annotated biological pathway databases, TrendCatcher can infer and visualize the most dynamic biological pathways in response to the external stimuli. Using simulated datasets for benchmarking, TrendCatcher achieved higher accuracy (AUC = 0.9) compared to other commonly used methods for the analysis of temporal gene expression datasets, when analyzing three or more time points. When comparing our results with other methods using the whole blood RNA-seq data, we found that DESeq2 identified the fewest
number of DDEGs in all patient groups, likely caused by a relative loss of statistical testing power by treating time as factorial variable. ImpulseDE2, identified the largest number of DDEGs but around 50% of the DDEGs had only a 0.5 log2 fold change. This oversensitivity to gene expression changes of limited magnitude may be due to the strong model assumption, which is that the dynamic gene expression trajectory must either follow a monotonic curve or an impulse shape. However, none of the other three methods provided the output of gene expression estimation from replicates at each time point and also did not infer gene expression trajectories. As a result, these approaches cannot center a given gene’s expression level to its baseline, let alone compare their trajectories. In contrast, TrendCatcher provides an output of gene expression estimation from multiple replicates for each time point, which makes it possible for further trajectory comparison and further analysis, such as timing of the up-regulation and down-regulation of specific biological pathways. A key advantage of using TrendCatcher is identifying the time intervals when differences emerge. This is especially apparent in the setting of biphasic or multiphasic temporal trajectories, in which gene expression levels can change their trend of upregulation or downregulation during a time course. Importantly, by utilizing the TimeHeatmap function, TrendCatcher can help researchers identify the magnitude and dynamic nature of pathways shifts.

Despite the extraordinary success of rapidly developed and deployed mRNA vaccines against SARS-CoV-2, the ongoing COVID-19 pandemic remains a major global health problem in part due to the emergence of newer highly contagious SARS-CoV-2 variants of concern as well as vaccine hesitancy. This requires the identification of novel mechanistic targets, especially in vulnerable patients who have a high risk of developing severe COVID-19. One of the key pathogenic factors driving COVID-19 severity is the profound immune dysregulation observed in patients with severe COVID-19 that can result in respiratory failure (26-29). Human immune responses to SARS-CoV-2 infection are highly dynamic and time dependent, requiring upregulation as well as downregulation of distinct immune signaling pathways at the appropriate times to ensure optimal host defense (26-29). Understanding the dynamics of the
COVID-19 immune response could form the basis of developing therapies that are appropriate for a given time window (30). Personalized medicine or precision medicine are gaining traction by developing therapies tailored to patients based on their genotype and phenotype, but personalization likely also requires tailoring therapies based on the temporal phase of a disease. These studies highlight the need for time course or longitudinal analyses of the host responses to the SARS-CoV-2 infection.

In this study, we applied TrendCatcher to systematically analyze sequential blood samples from either non-human primates infected with SARS-CoV-2, human patients with COVID-19 of varying severity or human subjects who received a SARS-CoV-2 mRNA vaccine. TrendCatcher identified dynamic gene expression and biological pathway signatures for (a) SARS-CoV-2 infection progression over time, (b) severe COVID-19 vs. moderate or mild COVID-19 and (c) COVID-19 vaccine recipients versus control. TrendCatcher established response to virus, humoral immune response and type I interferon signaling pathway activation across peripheral blood cell types in mild, moderate and severe COVID-19. However, we found temporal patterns of gene expression shifts that were unique in the severe COVID-19 patients. Severe COVID-19 was associated with marked activation of neutrophils and upregulation of coagulation pathways as well as blunted type I interferon signaling as early as week 1 in the peripheral blood of patients. Importantly, severe COVID-19 was associated with persistent activation of neutrophils and genes regulating coagulation for as long as 6 weeks, underscoring that the importance of the temporal analysis by TrendCatcher which identified hallmarks of severe COVID-19 in peripheral blood samples. It is important to note that while peripheral blood samples are convenient to obtain for longitudinal studies because they only involve minimally invasive blood draws, they only give a limited picture of the gene expression and activation states because gene expression patterns of immune cells in a tissue may be different from their counterparts in the circulating blood.
Our findings complement recent studies which implicate neutrophils in the excessive inflammation, coagulopathy, immunothrombosis and organ damage which characterize severe COVID-19 (31-35). Neutrophils are particularly active in highly vascularized organs, such as lungs and kidneys, which are prime targets of SARS-CoV-2 induced injury in COVID-19 (36). Dysregulated neutrophil responses, such as prolonged activation may cause damage to vessels and underlying parenchymal (37-39). Studies also show that activated neutrophils through Toll-like receptors, chemokine- and cytokine receptors can stimulate the neutrophil extracellular trap (NET) formation (31). And recent studies showed the disbalance between NET formation and degradation can trigger immunothrombosis and tissue injury (40, 41). Excessive activation of macrophages and adaptive immune cells in severe COVID-19, which can form vicious cycles of positive feedback circuits has also been demonstrated (42, 43), but less is known about how early neutrophils are activated in disease progression because many of the studies on neutrophils in clinical COVID-19 or animal models of severe SARS-CoV-2 focused on the late stages of the disease as well as the neutrophils in the lung tissue. In our study, we used a systematic temporal analysis and observed that profound neutrophil activation in the peripheral blood, which was predominant in the severe COVID-19 group, occurs as early as week one after diagnosis or symptoms persists even after six or seven weeks in surviving patients.

Our observation of early upregulation of coagulation genes in the whole blood transcriptomes of severe COVID-19 patients also points to another feature that is associated with severe COVID-19, thrombotic and embolic complications such as strokes (13, 44, 45). Recent studies have found the procoagulant changes in endothelial cells underly the coagulopathy in severe COVID-19 (46). Endothelial dysfunction in COVID-19 patients may be exacerbated through inflammatory cytokines and neutrophil extracellular traps, thus pointing to interactions between circulating or recently recruited neutrophils and the vessel wall endothelial cells that are in contact with circulating immune cells as drivers of such coagulopathic manifestations (47-49). We believe the analysis of the RNA-seq and scRNA-seq data support neutrophil
activation and upregulation of coagulation genes as defining features of severe COVID-19, highlighting their role as early biomarkers to provide prognostic information and thus optimally treat patients who have a higher likelihood of progressing to severe disease. Importantly, our results also raise questions about the role of neutrophils and of coagulation in what is referred to as “long COVID” or “post-acute sequelae of COVID” (50-52), because we observed that the activation persists for as long as six to seven weeks after the initial infection.

Our temporal analysis showed that severe COVID-19 single-cell PBMCs were characterized by impaired type I interferon signaling at the onset of infection (week 1), compared to the moderate COVID-19 group. This impaired type I interferon signaling was identified in innate cells, such as NK cells and monocytes, and adaptive cells, such as B cell and CD8+ T cells. Type I interferons, which are essential for antiviral immunity (53, 54), have been shown to be upregulated in COVID-19 (55, 56). Other studies also suggested that impaired type I interferon signaling may promote severe COVID-19 and that interferon therapy could be used as therapy in severe COVID-19 (57). However, since these studies were not designed for longitudinal analyses, the timing of when to intervene on IFN signaling was not clear. Traditional analyses of gene expression data are not suited for the identification of temporal windows and thus do not address whether impaired IFN production is present at the onset of infection, whether it was delayed, or whether IFN production is exhausted after an initial activation (56). TrendCatcher addressed this question by showing the dynamic timing of type I IFN from both the moderate group and severe group.

In our COVID-19 vaccination single cell PBMC temporal data analysis, we also identified prominent metabolic shifts in NK cells. Cellular metabolism is recognized as an important factor that can determine the fate and function of lymphocytes (58, 59). Certain metabolic pathways have been shown to have direct immunoregulatory roles (58). Activated NK cells engage in a robust metabolic response, which is required for normal effector functions (60), and our data suggests that assessing NK metabolic shifts may be an intriguing alternative to
assessing vaccine responsiveness, which does not rely on antibody levels which can be highly variable in laboratory assays.

In conclusion, we have developed the novel TrendCatcher R package platform designed for time course RNA-seq data analysis which identifies and visualizes distinct dynamic transcriptional programs. When applied to real whole blood bulk RNA-seq time course datasets from COVID-19 patients, we observed that patients with severe COVID-19 showed gene expression profiles consistent with profound neutrophil activation and coagulopathy early during the progression of the disease (starting from the first week of symptom onset). Even though the application of TrendCatcher in this manuscript focused on COVID-19 datasets, it has been designed to be used for the analysis of a broad range of dynamic biological processes and diseases.
Methods

TrendCatcher framework

The main components of the TrendCatcher framework are shown in Figure 1. TrendCatcher requires two main inputs: the raw count table $C$ of a temporal study with a dimension of $m \times n$, where $m$ denotes the number of genes and $n$ denotes the number of samples, and a user-defined baseline time variable $T$, such as "0 hour". Since samples may have different sequencing depths and batch effect, TrendCatcher integrates with limma (5) and provides preprocessing steps, such as batch correction and normalization. For single-cell RNA sequencing datasets, TrendCatcher extracts cells for each cell type annotated in the meta data slot of Seurat object and convert it into cell-type specific “pseudobulk” time course RNA library. Based on a user-specified threshold, relatively low abundant genes are removed from the count table, and reads are normalized and batch effects are removed. TrendCatcher’s core algorithm is composed of five main steps: (a) baseline fluctuation confidence interval estimation, (b) model dynamic longitudinal count, (c) time point dynamic $P$ value calculation, (d) gene-wise dynamic $P$ value calculation, and (e) break point screening and gene-wise dynamic pattern assignment. Mathematical details will be expanded in the following sections.

For the output of TrendCatcher, there are mainly two components: a master table and a set of functions for versatile visualization purposes. The master table contains all the dynamic details of each single gene, including its dynamic $P$ value, its break point location time, and its dynamic trajectory pattern. In addition to the master table, TrendCatcher produces five main types of visualizations: (a) a figure showing the observed counts and fitted splines of each gene, (b) genes trajectories from each trajectory pattern group (c) a hierarchical pie chart that represents trajectory pattern composition, (d) a TimeHeatmap to infer trajectory dynamics of top dynamic biological pathways and (e) a two-sided bar plot to show the top most positively and negatively changed (averaged accumulative log2FC) biological pathways.
Baseline fluctuation confidence interval estimation

We assumed that the observed number of RNA-seq reads count from the baseline time (e.g., t=0 hour) $X_{t,t_{\text{baseline}}}$ was generated from a negative binomial distribution (Equation 1):

$$X_{t,t_{\text{baseline}}} \sim NB(\mu_{t,t_{\text{baseline}}}, \varphi_{t})$$

(1)

Where gene $i = 1, ..., n$ and $\mu_{t,t_{\text{baseline}}}$ is the mean count of gene $i$ at baseline time, and $\varphi_{t}$ is the dispersion factor. First, the dispersion factor $\varphi_{t}$ was pre-estimated as a constant hyperparameter for each gene with DESeq2 (3), as shown in Equation 2. Here, $\sigma_{i}(t)$ is the variance at time $t$.

$$\sigma_{i}(t)^2 = \mu_{i}(t) + \varphi_{i} \cdot \mu_{i}(t)^2$$

(2)

Then, $\mu_{t,t_{\text{baseline}}}$ was estimated using maximum likelihood from Equation 3.

$$\hat{\mu}_{i,t_{\text{baseline}}} = \arg\max L_{NB}(u, \varphi) = \arg\max \frac{\Gamma(\varphi + x)}{x! \Gamma(\varphi)} \left( \frac{\mu}{\mu + \varphi} \right)^x \left( \frac{\varphi}{\varphi + \mu} \right)^\varphi$$

(3)

Based on the negative binomial distribution and the estimated mean count for baseline time, we constructed the 90% confidence interval $[X_{i,t_{\text{baseline}},0.05}, X_{i,t_{\text{baseline}},0.95}]$ as the baseline count fluctuation interval (Equation 4(a) and Equation 4(b)).

$$P(X_{i,t_{\text{baseline}}} \leq X_{i,t_{\text{baseline}},0.05} | \mu_{t,t_{\text{baseline}}}, \varphi_{t}) = 0.05$$

(4a)

$$P(X_{i,t_{\text{baseline}}} \geq X_{i,t_{\text{baseline}},0.95} | \mu_{t,t_{\text{baseline}}}, \varphi_{t}) = 0.95$$

(4b)

Model dynamic longitudinal count estimation

To model the time-dependent gene expression value, we applied a smoothing spline ANOVA model (61, 62) with a negative binomial family constraint to fit the reads from samples across non-baseline multiple time points. The random variable $X_{i,t}(t \neq t_{\text{baseline}})$ is assumed to follow the negative binomial (NB) distribution in Equation 5, with a positive integer $\alpha$ represents number of failures before the $\alpha th$ success in a sequential of Bernoulli trials, and $p(t) \in (0,1)$ represents the success probability.
\[ X_{i,t(t \neq t_{baseline})} \sim NB(\alpha, p(t)) \]  

(5)

The log-likelihood given a time course observed count \( x = \{x_{i,t(t \neq t_{baseline})}\}_{i=1,...,n; t=1,...,T} \) is calculated as Equation 6.

\[
L = \log L(\alpha, p(t) \mid X = x) = \sum_{t=1}^{T} \sum_{i=1}^{n} \left[ x_{i,t(t \neq t_{baseline})} \log (1 - p(t)) + a \log p(t) + \log \Gamma(\alpha + x_{i,t(t \neq t_{baseline})}) - \log \log \Gamma(\alpha) \right] + C
\]

(6)

Taking the logit link and model time effect, we define the logit link \( \eta \)

\[
\eta = \log \frac{p(t)}{1 - p(t)}
\]

(7)

To allow for flexibility in the estimation of \( \eta \), and find the best trade-off between goodness of fit and the smoothness of the spline curve, soft constraints of the form \( J(\eta) \) is added to the minus log-likelihood, with the smoothing parameter \( \lambda > 0 \).

\[-L + \lambda \ J(\eta) \]

(8)

The solution to the optimization of Equation 8 leads to a smoothing fitting to the reads from samples across different non-baseline time points. The estimated mean of \( \mu_{i,t(t \neq t_{baseline})} \) can be estimated using Equation 9.

\[
\mu_{i,t(t \neq t_{baseline})} = \hat{\alpha} e^{\hat{\beta}(t)} = \frac{\hat{\alpha} \hat{p}(t)}{1 - \hat{p}(t)}
\]

(9)

**Gene’s dynamic P value calculation**

To calculate gene’s non-baseline dynamic signal significance, each gene’s non-baseline estimated mean count \( \mu_{i,t(t \neq t_{baseline})} \) was tested against the baseline fluctuation interval. Based on Equation 10(a) and Equation 10(b), for each gene at each single non-baseline time point, a dynamic time P value was calculated.
Then, we applied Fisher's combined probability test method to calculate a gene-wise dynamic P value (Equation 11):

\[
\chi^2_{L2t} \sim -2 \sum_{t=0}^{T} \ln(p_{i,t})
\]  

Trajectory pattern assignment

First, we connect all the significant dynamic signal time points, with a P value threshold less than 0.05. Then we applied a break point searching strategy to capture the gene expression change trend. The definition of break point is defined using Equation 12(a) and Equation 12(b).

\[
\text{If } \mu_i,t(\neq t_\text{baseline}) \geq \mu_i,t_{\text{baseline}}, \text{ then } p_{i,t} = \int P(x \geq \mu_i,t(\neq t_\text{baseline}) | \mu_i,t_{\text{baseline}}, \varphi) 
\]  

\[
\text{If } \mu_i,t(\neq t_\text{baseline}) \leq \mu_i,t_{\text{baseline}}, \text{ then } p_{i,t} = \int P(x \leq \mu_i,t(\neq t_\text{baseline}) | \mu_i,t_{\text{baseline}}, \varphi) 
\]

There are two types of break points: type I means a gene’s expression level is up-regulated followed by a down-regulation, and type II means a gene’s expression level is down-regulated and then followed by an up-regulation. By screening along the break point, the master-pattern and sub-pattern were assigned to each gene.

TimeHeatmap enrichment analysis and two-sided bar plot

To build the TimeHeatmap for visualizing the biological pathway enrichment change over time, we designed a window-sliding strategy to capture all the up-regulated or down-regulated genes within each time interval. If we denote time vector as \( t_j \) and \( j \in 1, ..., T \), each time interval is denoted as \([t_{j-1}, t_j]\). We found \( N_{up} \) up-regulated genes and \( N_{down} \) down-regulated
genes within the time window \([t_{j-1}, t_j]\), then Fisher’s exact test was performed to obtain the GO term enrichment with the corresponding time interval for up-regulated genes and down-regulated genes separately. Users can select the top most enriched biological pathways (ranked by enrichment P value) for each time interval (by default is 10). Then for each selected GO term within the corresponding time window, we calculated the averaged log 2 fold change (Avg_log2FC) of all the DDEGs from this GO term. A series of Avg_log2FC values over time characterize the trajectory dynamics of the corresponding biological pathway, it is defined as biological pathway trajectory inference in this study. The summation of the series of Avg_log2FC estimates the averaged accumulative log 2 fold change (GO_mean_logFC) for the corresponding GO term. TrendCatcher ranks biological pathways based on their dynamic magnitude inferred from the GO_mean_logFC value. Users are free to choose the top most positively and negatively changed (averaged accumulative log2FC) biological pathways. Besides GO enrichment analysis (63), TrendCatcher also packages Enrichr (64) biological pathway databases. Visualization is constructed using the ComplexHeatmap (65) package.

**Simulated dataset**

To mimic the real biological RNA-seq dataset, we only allowed ~10% of the genes to be dynamic responsive genes. In this study, we embedded 5 different types of trajectories into the temporal RNA-seq simulated datasets, including non-dynamic trajectory (~90%), monotonic trajectory (~2.5%), bi-phase trajectory (~2.5%), and multimodal trajectory (~5%) including 2 break-point and 3 break-point trajectory. Each type of trajectory was constructed by adding negative binomial distribution noise to the embedded trajectory count. For example, for monotonic trajectory, we defined the first and last time points’ RNA expression level change to be 0.5-2 log2 fold change, then we added negative binomial distribution noise to the embedded monotonic trajectory. To check how the AUC changes as time course study extends longer, we sampled 3, 5, 7, 9 and 11 different number of time points with even time intervals, and randomly sampled 3 replicates for each time point. To validate how the
prediction AUC varies across different types of trajectories, we embedded ~10% of the genes to have non-dynamic trajectories, and ~90% of the genes have a specific dynamic trajectory (monotonic, bi-phase or multimodal).

**Pseudobulk RNA library construction**

To construct "pseudo-bulk" RNA library from single cell RNA-seq datasets, all cells for each cell type in a given sample were computationally "pooled" by summing all counts for a given gene. Since pseudobulk libraries composed of few cells are not likely modeled properly, we removed cell types containing less than 1k cells in this study. Lowly expressed genes were removed for each cell type as well, using the filterByExpr function from edgeR R package (4). Gene counts were transformed using the log of the counts per million (CPM) and library size was normalized using calcNormFactors function with the method = "RLE".

**Gene set enrichment of DDEGs**

Gene sets enrichment analysis in this study was performed using clusterProfiler (63) R package, and enrichment comparison from multiple groups were visualized using the compareCluster function.

**Permutation testing for assessing differences between groups over time**

After fitting gene expression longitudinal profiles from each severity group with a LOESS smoothing spline, we binned the time variable into 100 time intervals and calculated the observed area ratio between two curves within each time interval. Next, we shuffled the severity group label on the gene expression longitudinal profiles and repeated the previous step to calculate the shuffled area ratio for each time interval. We iterated the shuffling step 1,000 times. In this way, for each time interval, we calculated the P value using the empirical distribution from the permutation test. This permutation test module in TrendCatcher allows
users to assess between group differences of dynamic gene expression pathways in a time interval dependent manner.

**Source code**
The R package source code of TrendCatcher is available on GitHub ([https://github.com/jaleesr/TrendCatcher](https://github.com/jaleesr/TrendCatcher)).

**Statistics**
The statistical tests for each computational approach model are described in-depth in the relevant Methods sections. Generally, P values less than 0.05 were considered significant and when multiple comparisons were performed, we applied Holm-Bonferroni method for P value adjustment.

**Study approval**
All the datasets used for the analyses were publicly available (as indicated in Table 1) and no human subjects were recruited.
Author contributions

X.W. and J.R. designed the study, X.W. performed the analyses and wrote the first draft of the manuscript, J.R. revised the initial draft of the manuscript. M.S. and Y.D. provided critical input for the analyses, visualizations and revisions of the manuscript.

Acknowledgments

The studies were supported by NIH grants R01-HL149300 (to JR), P01-HL151327 (to JR), and T32-HL139439 (to MS). The graphical abstract was generated with BioRender.com.
References


Figure 1. Overview and benchmarking of TrendCatcher. (A) TrendCatcher’s framework. TrendCatcher preprocesses input data, which includes creating cell-type specific “pseudobulk” datasets for temporal analysis when single-cell sequencing data is used. TrendCatcher’s core algorithm is composed of five main steps. TrendCatcher’s output includes four main types of visualizations and DDEGs identification (numbered from 1 to 5). (B) TrendCatcher’s prediction ROC for a 7 time point simulated dataset compared to DESeq2, DESeq2Spline and ImpulseDE2, with mixed trajectories. (C) TrendCatcher’s prediction performance (AUC) across different number of time points, from 3 to 11 time points. TrendCatcher’s AUC values across time points from 3 to 11 are, 0.90, 0.92, 0.90, 0.89 and 0.88.
Figure 2. Dynamic gene expression in peripheral blood following SARS-CoV-2 inoculation in a non-human primate model. (A) Analysis of the two predominant trajectory patterns in the non-human primate peripheral blood RNA-seq data from day 0 to day 14. The top-left figure represents 167 DDEGs following an up-down expression pattern, which peaked at day 2 then slowly decreased until day 14. The top-right figure represents their expression using a traditional z-score normalized heatmap. The bottom-left figure represents 263 DDEGs following a monotonic down-regulated trajectory pattern and their gene expression values were represented in the corresponding heatmap on the right. Gene expression values have been normalized and log2 transformed. (B and C) Top 3 GO enrichment analysis pathways using 167 DDEGs from trajectory pattern "0D-2D Up, 2D-14D Down" and 263 DDEGs from trajectory pattern "0D-14D Down". X-axis represents the number of genes enriched in GO terms; Y-axis represents the enriched GO terms; p.adjust represents adjusted P values using Holm-Bonferroni methods; P value is generated by Fisher’s exact test. (D) TimeHeatmap of the top 15 dynamic pathways and their dynamic time windows visualizes the temporal patterns. Each column represents a time window. "0D-1D" represents day 0 to day 1. The “%GO” column represents the percentage of DDEGs found in the corresponding pathway. The “nDDEG” column represents number of DDEGs found in the corresponding pathway. Number in each grid presents the averaged log2 fold change of gene expressions compared to the previous time window. Color represents the averaged log2 fold change of the DDEGs within each time window for the corresponding pathway.
Figure 3. Cell-type specific dynamic gene expression in peripheral blood mononuclear cells following SARS-CoV-2 infection in patients. (A) UMAP visualization of single-cell RNA-seq PBMC dataset (Zhu et al.) with annotated cell types from the original study. (B) TimeHeatmap of top dynamic biological pathway from plasma B cells. Each column represents a time window. Stage 0 represents uninfected baseline. The "%GO" column represents the percentage of DDEGs found in the corresponding pathway. The "nDDEG" column represents number of DDEGs found in the corresponding pathway. Number in each grid presents the averaged log2 fold change of gene expressions compared to the previous time window. Color represents the averaged log2 fold change of the DDEGs within each time window for the corresponding pathway. (C) Top GO enrichment comparison analysis using DDEGs from each cell type. X-axis represents cell types with the number of DDEGs shown in the brackets; Y-axis represents the enriched GO terms; p.adjust represents adjusted P values using Holm-Bonferroni methods; P value is generated by Fisher's exact test. Dot size represents gene ratio.
Figure 4. Temporal analysis of whole blood RNA-seq data in patients grouped according to disease severity. (A) Venn diagram of DDEGs identified from three COVID-19 severity groups, including mild, moderate and severe. (B) Top GO enrichment from shared DDEGs across three groups compared with top GO enrichment from DDEGs only identified in severe group. X-axis represents comparison groups with the number of DDEGs shown in the brackets; Y-axis represents the enriched GO terms; p.adjust represents adjusted P values using Holm-Bonferroni methods; P value is generated by Fisher’s exact test. Dot size represents gene ratio. (C) TimeHeatmap of the top dynamic pathways from the severe group. Each column represents a time window. “0W-1W” represents week 0 (healthy control) to week 1. The “%GO” column represents the percentage of DDEGs found in the corresponding pathway. The “nDDEG” column represents number of DDEGs found in the corresponding pathway. Number in each grid presents the averaged log2 fold change of gene expressions compared to the previous time window. Color represents the averaged log2 fold change of the DDEGs within each time window for the corresponding pathway. (D-G) LOESS curve fitting of DDEGs identified in the severe COVID-19 group of the (D) neutrophil activation pathway, (E) humoral immune response pathway, (F) blood coagulation pathway and (G) respiratory burst pathway. Red curves represent the severe group, blue curves represent the moderate group, and green curves represent the mild group. X-axis represents time in weeks; Y-axis represents the averaged log2 fold change of gene expressions compared to the baseline.
Figure 5. Temporal analysis of single cell RNA-seq data of PBMCs from patients with either moderate or severe COVID-19. (A) Dot plot showing GO enrichment comparison between severe COVID-19 and moderate COVID-19 for each cell type. Each panel represents one cell type. X-axis represents severity group; Y-axis represents the enriched GO terms; p.adjust represents adjusted P values using Holm-Bonferroni methods; P value is generated by Fisher’s exact test. Dot size represents gene count. (B) LOESS curve fitting on DDEGs identified from the type I interferon pathway using TrendCatcher from moderate COVID-19 and severe COVID-19. Blue color indicates moderate group and red color indicates severe group. X-axis represents time in weeks; Y-axis presents the averaged log2 fold change of gene expressions compared to the baseline. (C-F) TimeHeatmap of NK cells from moderate and severe COVID-19, CD8+T cells from moderate and severe COVID-19. Each column represents a time window. “0W-1W” represents week 0 (healthy control) to week 1. The “%GO” column represents the percentage of DDEGs found in the corresponding pathway. The “nDDEG” column represents number of DDEGs found in the corresponding pathway. Number in each grid presents the averaged log2 fold change of gene expressions compared to the previous time window. Color represents the averaged log2 fold change of the DDEGs within each time window for the corresponding pathway compared to the previous time window.
Figure 6. Temporal analysis of PBMC single cell RNA-seq data from human subjects receiving the SARS-CoV-2 mRNA vaccine. (A) UMAP of the single-cell transcriptional profile of one patient (id=2047) on day 0. Cell types were auto-annotated by SingleR. (B) Dot plot of comparison of the top GO terms enriched from cell-type specific DDEGs. X-axis represents cell type with the number of DDEGs shown in the brackets; Y-axis represents the enriched GO terms; p.adjust represents adjusted P values using Holm-Bonferroni methods; P value is generated by Fisher’s exact test. Dot size represents gene ratio. (C) TimeHeatmap of NK cells. Each column represents a time window. “0D-1D” represents day 0 (healthy control) to day 1. The “%GO” column represents the percentage of DDEGs found in the corresponding pathway. The “nDDEG” column represents number of DDEGs found in the corresponding pathway. Number in each grid presents the averaged log2 fold change of gene expressions compared to the previous time window. Color represents the averaged log2 fold change of the DDEGs within each time window for the corresponding pathway. The first dose was administered on day 1 and the second dose was administered on day 21.
### Tables

Table 1. Datasets analyzed in this study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Organism</th>
<th>Number of time points</th>
<th>Study</th>
<th>Figure</th>
<th>Dataset Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulk Peripheral Blood (Aid et al.)</td>
<td>Macaca mulatta</td>
<td>7</td>
<td>Covid vs. Normal</td>
<td>Figure 2</td>
<td>GSE156701</td>
<td>(13)</td>
</tr>
<tr>
<td>scRNA PBMC (Zhu et al.)</td>
<td>Homo sapien</td>
<td>5</td>
<td>Covid vs. Normal</td>
<td>Figure 3</td>
<td><a href="https://db.cngb.org/search/project/CNP0001102/">https://db.cngb.org/search/project/CNP0001102/</a></td>
<td>(14)</td>
</tr>
<tr>
<td>Bulk Peripheral Blood (Bergamaschi et al.)</td>
<td>Homo sapien</td>
<td>8</td>
<td>Severe vs. Moderate</td>
<td>Figure 4</td>
<td><a href="https://www.covid19cellatlasc.org/patient/citid/">https://www.covid19cellatlasc.org/patient/citid/</a></td>
<td>(11)</td>
</tr>
<tr>
<td>scRNA PBMC (Liu et al.)</td>
<td>Homo sapien</td>
<td>4</td>
<td>Severe vs. Moderate</td>
<td>Figure 5</td>
<td>GSE161918</td>
<td>(17)</td>
</tr>
<tr>
<td>scRNA PBMC (Arunachalam et al.)</td>
<td>Homo sapien</td>
<td>8</td>
<td>Vaccination</td>
<td>Figure 6</td>
<td>GSE171964 and GSE169159</td>
<td>(18)</td>
</tr>
</tbody>
</table>