Front. Med. 2022, 16(2): 251–262
https://doi.org/10.1007/s11684-021-0915-9

RESEARCH ARTICLE

PathogenTrack and Yeskit: tools for identifying intracellular pathogens from single-cell RNA-sequencing datasets as illustrated by application to COVID-19

Wei Zhang[1,2], Xiaoguang Xu[1], Ziyu Fu[1], Jian Chen (✉)[3], Saijuan Chen (✉)[1], Yun Tan (✉)[1]

[1]*Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China;* [2]*School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China;* [3]*Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai 200433, China*

Abstract **Pathogenic microbes can induce cellular dysfunction, immune response, and cause infectious disease and other diseases including cancers. However, the cellular distributions of pathogens and their impact on host cells remain rarely explored due to the limited methods. Taking advantage of single-cell RNA-sequencing (scRNA-seq) analysis, we can assess the transcriptomic features at the single-cell level. Still, the tools used to interpret pathogens (such as viruses, bacteria, and fungi) at the single-cell level remain to be explored. Here, we introduced PathogenTrack, a python-based computational pipeline that uses unmapped scRNA-seq data to identify intracellular pathogens at the single-cell level. In addition, we established an R package named Yeskit to import, integrate, analyze, and interpret pathogen abundance and transcriptomic features in host cells. Robustness of these tools has been tested on various real and simulated scRNA-seq datasets. PathogenTrack is competitive to the state-of-the-art tools such as Viral-Track, and the first tools for identifying bacteria at the single-cell level. Using the raw data of bronchoalveolar lavage fluid samples (BALF) from COVID-19 patients in the SRA database, we found the SARS-CoV-2 virus exists in multiple cell types including epithelial cells and macrophages. SARS-CoV-2-positive neutrophils showed increased expression of genes related to type I interferon pathway and antigen presenting module. Additionally, we observed the *Haemophilus parahaemolyticus* in some macrophage and epithelial cells, indicating a co-infection of the bacterium in some severe cases of COVID-19. The PathogenTrack pipeline and the Yeskit package are publicly available at GitHub.**

Keywords scRNA-seq; intracellular pathogen; microbe; COVID-19; SARS-CoV-2

Introduction

Microbes are the most ubiquitous life forms with little known in terms of their diversity. Due to a deeper understanding of their role in regulating host immune response and causing host pathogenesis, studies on microbes have become a hotspot in the biological exploration of human health and disease. In addition, pathogenic microorganisms show regulatory roles on the machineries of genetic information flow in host cells, such as affecting the transcriptional program. This raises the question of how cell-to-cell variability predicts or alters the host relationship with microbial organisms in a given disease context. In recent years, much insight has been gained into host responses to microbial infections, but tools used for interpreting the distribution of pathogens in a single cell and the impact of pathogens on host cell homeostasis are not available.

Single-cell RNA-seq (scRNA-seq) has recently been engaged as the most useful tool for studying transcriptomic characteristics at the single-cell level [1]. It has been widely used for revealing the distribution of cells in the microenvironment of organs, tissues and tumors, tracking cell hierarchy, understanding tumor heterogeneity, and inferring intracellular communication and regulatory networks [2–7]. Since the RNA of intracellular pathogens

Received December 3, 2021; accepted December 20, 2021
Correspondence: Saijuan Chen, sjchen@stn.sh.cn;
Yun Tan, ty12260@rjh.com.cn;
Jian Chen, yufuchenjian@163.com

may also be captured when preparing libraries for transcriptome sequencing, the scRNA-seq data set can also be used to track pathogens at the single-cell level [8]. In fact, a recent study focusing on viral–host interaction revealed the SARS-CoV-2 sequencing reads in 3′ scRNA-seq data [9], highlighting the potential usage of scRNA-seq in identifying intracellular pathogens. Nevertheless, there has not been any computational tools that systematically explore the metagenomic features in host cells at the single-cell level.

Here, we established a computational framework for identifying and exploring intracellular pathogens (bacteria or viruses) at the single-cell level. The method includes a Python package (PathogenTrack) for intracellular pathogen identification and an R package (Yeskit) for integration, clustering, differential gene analysis, functional annotation, and visualization of single-cell data. Our algorithm has been tested on various simulated and real scRNA-seq data sets and performed robustly. Taking the scRNA-seq data of two severe COVID-19 patients as an example, we used PathogenTrack to identify microbial infected cells at the single-cell level and used Yeskit to explore the biological functions that may be related to SARS-CoV-2 infection.

Materials and methods

The PathogenTrack workflow

The first step of PathogenTrack is to extract cell barcodes (CBs) and unique molecular identifiers (UMIs) and add them to the header of the sequenced reads. Many existing tools for whitelisting CBs are available, such as cellranger [1], alevin [10], and UMI-tools [11]. Since cellranger and alevin are designed for scRNA-seq quantification and generate more reliable CBs, it is recommended to use cellranger or alevin to acquire valid CBs. The CBs and UMIs are extracted and added to the header of read2 with UMI-tools. Fastp [12] is an ultra-efficient tool for read quality control and is employed to remove low-quality or low-complexity reads. After quality control, reads not aligned to the host reference genome with STAR [13] are kept for taxonomy classification. Kraken2 [14] is a read-level taxonomy classification tool with high precision and speed and is employed in PathogenTrack. Since strains from the same species may have shared genomic sequences, the k-mer based method like kraken2 would not accurately classify all reads at the species level [14]. To solve this problem, PathogenTrack corrects the taxonomy IDs with less read support to those supported by the most abundant reads under the same species level. Reads with assigned taxonomy IDs are deduplicated and then quantified with UMIs. Finally, a quantification matrix of pathogen species with UMI counts is generated.

The Yeskit package

Yeskit is an R package designed for single-cell gene expression data importation, integration, clustering, differential analysis, functional analysis, and visualization. It consists of 17 functions, including data importation and integration (*scRead*, *scIntegrate*, and *scOne*), differential analysis (*scDGE* and *scPathogenDGE*), functional analysis (*scGO*, *scPathogenGO*, and *scMSigdbScoring*), visualization (*scDimPlot*, *scDensityPlot*, *scPopulationPlot*, *scVizMeta*, *scPathogenRatioPlot*, *scVolcanoPlot*, *scGOBarPlot*, *scGODotPlot*, and *scScoreDimPlot*). Yeskit obeys the default data structure of Seurat [15], and stores pathogen expression data and pathway enrichment scores in the obj@meta.data slot and stores differential gene analysis results and GO enrichment results in the obj@misc slot.

The *scRead* function is designed for reading 10x Genomics single-cell count matrix and filtering out low quality cells. Besides, two features (reading pathogen count matrix and handling PDX model) were included to fulfill the corresponding circumstances. If the pathogen count matrix was specified, *scRead* will read and store the pathogen-by-cell count matrix into the obj@meta.data slot. If the input file was a scRNA-seq count matrix from xenografts samples (PDX model with human and mouse genomes), *scRead* can distinguish human cells and mouse cells by the threshold fraction (a minimum of 90% host-specific reads by default) of reads to separate human and mouse cells. In the quality control step, soft thresholds of the number of expressed genes (99-th quantile), and the percentage of mitochondrial genes (99-th quantile) were used instead of hard thresholds.

The *scIntegrate* function is a wrapper for the Seurat standard workflow. It can be used to merge two or more Seurat object together, normalize the data, select features, scale the data, perform linear dimensional reduction (PCA), cluster the cells, run nonlinear dimensional reduction (UMAP/tSNE), and return an integrated Seurat object. Since the heterogeneity among clinical samples is always very large, the robust and time-efficient batch effect removing method Harmony [16,17] is included in the *scIntegrate* function. If the project has only one sample, *scOne* function can be used instead of *scIntegrate* to complete the Seurat standard workflow.

The *scDensityPlot* function is used to visualize the cell density between various samples. The dark red area indicates that the cell density in this area is high, and the white area indicates that the cell density in this area is low.

It is routine to calculate marker genes in each cluster or differentially expressed genes between two conditions. Seurat's *FindMarkers* function is very useful for finding markers (differentially expressed genes) for identity classes. The *scDGE* function is a wrapper of *FindMarkers*, which is used to detect differentially expressed genes in

each cluster between two groups. The *scPathogenDGE* function is also a wrapper for *FindMarkers*, used to detect genes differentially expressed between pathogen-positive and pathogen-negative (or pathogen-bystander) cell groups.

The *scGO* function is designed for annotating the Gene Ontology functions of each cluster with the topGO [18] package. For cell markers generated by *FindAllMarkers* from the Seurat package, only the function of upregulated genes can be annotated, while for results generated by *scDGE*, both the functions of up- or downregulated genes can be annotated.

The *scVizMeta* function can be used to display any numerical column stored in the obj@meta.data slot in UMAP/tSNE/PCA embeddings.

The *scMSigdbScoring* function is used to calculate the pathway scores from MSigDB [19] and store them in the obj@meta.data slot. It uses the *AddModuleScore* function of Seurat to calculate and stores gene module scores in the obj@meta.data slot.

Single-cell RNA-seq data sets simulation

Simulation is a compelling benchmarking strategy since the ground truth is known when the data are generated, making it possible to evaluate the performance of various methods. We need a scRNA-seq data set that mimics host cells infected with pathogens. There are two main processes in the scRNA-seq read simulation stage: single-cell count matrix preparation and single-cell sequencing read generation.

During the preparation of single-cell expression data, the host single-cell count matrix and the pathogen count matrix must be generated separately. To make the simulated host scRNA-seq data closer to the real data set, the publicly peripheral blood mononuclear cells data set (pbmc 4k) from 10x Genomics website [1] was used as the host single-cell count matrix. The pathogen single-cell count matrix was generated by Splatter [20]. Splatter is a powerful count-level simulation method that can generate scRNA-seq count data robustly. We obtained 20 clinically common pathogenic species with complete sequenced core genomes from the list of human infectious pathogens at Wikipedia website and retrieved their gene sequences from NCBI. Since prokaryotic genomes vary greatly in gene size and number of genes, to ensure that each species is fully simulated, we randomly selected up to 50 genes for each species. To simulate host cells infected with pathogens at different levels, Splatter's default parameters were used except the library size parameter (lib.loc) was set between 1 and 5, with an increment of 1. After the pathogen single-cell count matrix is generated, the host and pathogen single-cell count matrices were combined into one count matrix for the single-cell reads generation process.

In the single-cell reads generation process, minnow [21] was chosen to simulate scRNA-seq reads. Since numerous simulation methods have been introduced for scRNA-seq data [21–25], minnow is a powerful simulator that can currently be used for read-level simulation of single-cell experiments. It can mimic the single-cell sequencing process, such as randomly assigning UMIs to molecules, simulating PCR duplicates based on real reads distribution, imputing sequencing errors, and generating random start positions from transcripts. Therefore, we employed minnow to simulate single-cell reads guided by the single-cell count matrix. Since the start position of the reads simulated from each transcript obeys the truncated normal distribution $N(\mu,\sigma)$, to avoid repeated sampling reads from the same start position, we run minnow with the default parameters, except that the standard deviation σ was changed [26].

To systematically simulate host cells infected with pathogens under various conditions, technical features including UMI length, read length, PCR cycles, and reads coverage were considered. In more detail, we repeated simulations with two UMI lengths (10 and 12 bp, characteristic of the 10x Genomics Single Cell 3′ Version 2 and Version 3, respectively), three read lengths (from 50 to 150 bp, at 50 bp increments), three σs (the standard deviation of start position from 25 to 75 in increments of 25), three PCR cycles (from 4 to 6 in increments of 1), five incremental pathogen infection levels (Splatter's library size location parameter from 1 to 5 in increments of 1, to indicate the infection level of bacterial or viral reads) and three replicates per simulation. In total, this represents 810 simulations (2 UMI lengths × 3 read lengths × 3 σs × 3 PCR cycles × 5 infection levels × 3 replicates). We have limited our assessment to smaller simulated data sets consisting of 100 cells by down sampling the PBMC using geosketch [27] to save computational resources.

In the "time and memory evaluation" step, to save calculation time, we randomly sampled 18 data sets. The detailed simulation parameters are as follows: (1) UMI length was set to 10; (2) Read length was set to 100; (3) PCR cycle was set to 5; (4) Pathogen infection level was set to 3; (5) The σ was set to 50; (6) The number of cells was set to 100, 500, 1000, 2000, 3000, and 4000 each time, and each simulation was repeated three times.

Performance evaluation on simulated data

The simulated data sets were processed with Viral-Track and PathogenTrack. Viral-Track uses UMI-tools to detect valid barcodes, while PathogenTrack uses barcodes generated by alevin or cellranger. Alevin is an accurate and fast end-to-end tool for processing droplet-based scRNA-seq data from fastq to count matrix. It performs better in CB detection and UMI deduplication. To make a fair comparison between Viral-Track and PathogenTrack,

we replaced the default barcode file of Viral-Track with the barcode file generated by alevin.

The accuracy of cells infected by a particular pathogen is evaluated by converting the detection results into a binary matrix. Each column represents a specific pathogen, and each row represents a cell. Then we record whether each cell is classified as pathogen infection (1) or not (0). Since we know the actual cells infected by specific pathogens in the simulation data set, we can evaluate each pathogen species' sensitivity and specificity.

For each pathogen detection method, we calculated the number of true positive (TP; pathogen-infected cells were correctly classified), false positive (FP; non-infected cells were classified as pathogen-infected cells), true negative (TN; non-infected cells were classified as non-infected cells) and false negative (FN; pathogen-infected cells were classified as non-infected cells). We then calculated the sensitivity of each method as TP/(TP + FN) and specificity as TN/(TN + FP). The sensitivity of each classified pathogen was calculated and recorded. To make a fair comparison of these two methods, we only used pathogens in the simulated data for the benchmark.

Computation time and memory-usage estimation

To benchmark the methods' performance, we implemented these detection tools in multiple high-performance computing platforms. "/usr/bin/seff" was used to record the run time and the maximum memory consumption.

Data and code availability

Any relevant data are available from the authors upon reasonable request. The scRNA-seq data used in this manuscript are all publicly available, and they are summarized in Table S1. The PBMC data sets are available at 10x Genomics's official website. The PathogenTrack pipeline and the Yeskit package are publicly available at GitHub websites. For simple installation, PathogenTrack has been deposited in the Bioconda channel and the PyPI repository.

Results

PathogenTrack: unsupervised characterization of the intracellular microbiome from scRNA-seq data

PathogenTrack is an unsupervised computational pipeline that uses unmapped reads to characterize intracellular pathogens at the single-cell level (Fig. 1A). PathogenTrack includes the following steps: (1) Pre-processing scRNA-seq reads with single-cell quantification software (such as cellranger or alevin) to obtain the gene quantification matrix and CB file (Fig. 1B). The CB file is taken as the whitelist file for UMI-tools to extract the CB and UMI from Read1 and is added to the header of Read2 (barcoded-read2). (2) Removing low quality or low complexity reads in barcoded-read2 using fastp. (3) Aligning the remaining reads, which passed the quality control, to the host reference genome (such as hg38) using STAR algorithm. The unmapped reads are reserved for further use. (4) Metagenomic classification of the unmapped reads using Kraken2 algorithm. The taxonomy identifiers (IDs) are appended to the header of the corresponding reads. (5) Reads assigned with taxonomy IDs are subject to de-duplication, taxonomic correction, and quantification with UMIs. The output pathogen species-by-cell quantification matrix with UMI counts is then ready for downstream analysis (Fig. 1A).

Yeskit: an R-based package for interpreting the scRNA-seq data

Next, we come up to a method named Yeskit (Yet another single-cell analysis toolkit) to integrate and interpret the host gene expression data and the intracellular pathogen quantification data at the single-cell level (Fig. 1C). Yeskit is an R package designed for single-cell gene expression matrix importation, data integration, clustering, differential analysis, functional analysis, and visualization. Since Yeskit does not change the default data structure of Seurat, it can be easily integrated into most existing scRNA-seq analysis workflows. Yeskit can be used to read other information (such as gene mutation-by-cell matrix, pathogen count-by-cell matrix) and store them as additional data in the Seurat obj@meta.data slot. Moreover, it calculates MSigDB pathway enrichment scores and stores them in the obj@meta.data slot. In addition, it performs differential gene analysis between groups or between pathogen-infected (Pos) and bystander (Neg) cells in each cluster. Furthermore, it performs GO enrichment analysis and stores their results in the obj@misc slot. Besides, when there are many points in the vector diagram, editing becomes difficult. To deal with the challenge, most visualization functions in Yeskit have the option to rasterize the point layer and keep all axes, labels, and text in vector format.

Decoding SARS-CoV-2 infection in COVID-19 patients with PathogenTrack and Yeskit approach

To evaluate the applicability of our workflow for detecting and decoding intracellular pathogens in human single-cell data, we took the BALF data sequenced by 10x Genomics technology from two severe COVID-19 patients (SRA accession number: SRP250732) as an example [28]. Gene expression data were obtained by cellranger, and pathogen quantification matrices were generated by PathogenTrack. We then used Yeskit to integrate the host and the pathogen

quantification matrices and explored the lesions of biological functions related to SARS-CoV-2 infection (Fig. 2). A total of 13 138 high-quality single cells were ultimately obtained. Four major cell lineages were identified: macrophages, neutrophils, lymphocytes, and epithelial cells (Fig. 2A). The cell distribution can be visualized by density plot, and the distribution of cell populations per samples were shown in Fig. 2B and 2C. We visualized the distribution of SARS-CoV-2 positive cells in each sample. In both samples, the sequence of SARS-CoV-2 could be found in epithelial cells and immune cells (Fig. 2D and Fig. 2E).

Secondary bacterial infections were reported to cause serious complications associated with worse outcomes in COVID-19 patients. PathogenTrack is optimally designed to systematically profile the source of infection or co-infections in human clinical samples. Interestingly, a small fraction of cells from one of the COVID-19 patients (patient C145) revealed the presence of a co-infected bacterium, *Hemophilus parahemolyticus*. The bacterium

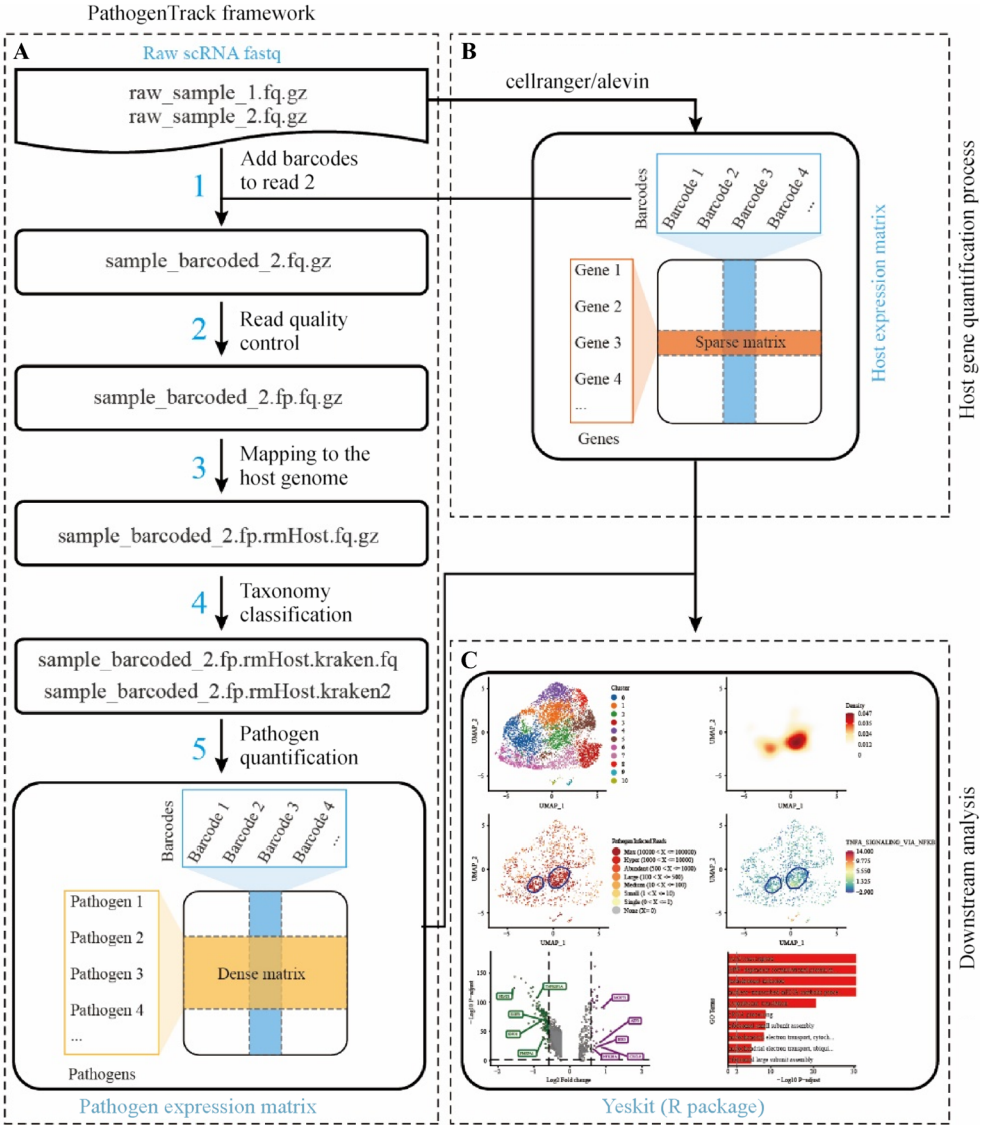

Fig. 1 An overview of the PathogenTrack workflow and the downstream analysis package Yeskit. (A) The input to the PathogenTrack pipeline are sample-demultiplexed FASTQ files, and there are several steps required to process this data and obtain per-cell pathogen species level quantification estimates. The output count matrix is a dense matrix, where rows stand for pathogens and columns stand for barcodes of cells. (B) scRNA-seq reads are processed by cellranger or alevin, and the output barcode file is used as input to the PathogenTrack workflow. (C) The host gene-by-cell count matrix and the pathogen species-by-cell count matrix are taken as input of Yeskit for single-cell integration analysis. Yeskit contains 17 functions for importing, integrating, analyzing, and visualizing the host-pathogen interactions at the single-cell level.

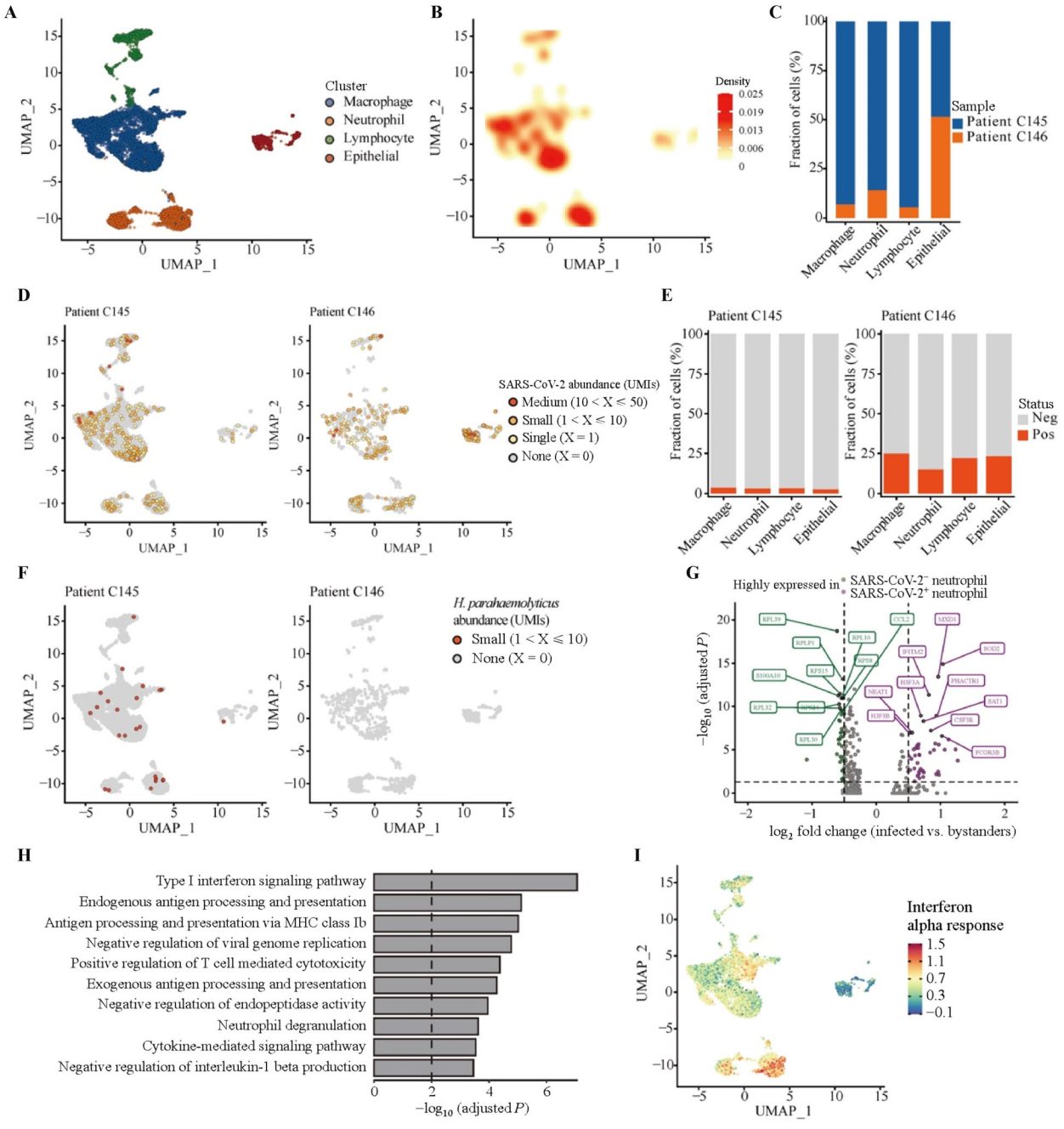

Fig. 2 Application of the PathogenTrack and Yeskit in the scRNA-seq analysis of COVID-19. (A) Overview of the cell clusters of 13 138 single cells derived from two severe COVID-19 patients. Clusters were named based on the cluster-specific gene expression patterns. (B) Density plot depicting projection of cells on the 2D map shown in (A). (C) Proportion of subpopulations in each sample. (D) Viral load of SARS-CoV-2 in each cell quantified by PathogenTrack. (E) Proportion of SARS-CoV-2 infected (Pos) and bystander (Neg) cells in each cluster. (F) Bacterial load of *Hemophilus parahemolyticus* in each cell quantified by PathogenTrack. (G) Volcano plot showing differentially expressed genes between neutrophil cells with or without SARS-CoV-2 RNA detected. Differentially expressed (> 0.5 absolute $\log_2$ fold change) and statistically significant (adjusted P value < 0.05) are colored in green (downregulated) or purple (upregulated). (H) Enriched Gene Ontology (GO) terms in genes highly expressed in SARS-CoV-2-positive neutrophil cells shown in (G). (I) Scores of the interferon alpha response gene module across all cells, projected on the 2D map shown in (A). Color scale represents the average expression level of the gene module subtracted by the aggregated expression of control feature sets.

was enriched in neutrophils and macrophages (Fig. 2F). *Hemophilus parahemolyticus* has been reported to cause acute respiratory distress syndrome and septic shock [29]. Differential gene expression analysis between SARS-CoV-2-RNA-positive and negative neutrophil cells indicated that SARS-CoV-2 positive cells exhibited elevated expression of a diverse set of genes required for monocyte activation, such as G-CSF receptor (*CSF3R*), CD16 (*FCGR3B*), and interferon-induced transmembrane protein 2 (*IFITM2*) (Fig. 2G). These genes were enriched in pathways such as "type I interferon signaling pathway," "negative regulation of viral genome replication," and "neutrophil degranulation" (Fig. 2H). Additionally, we scored the potential enriched pathways of all cells. Interestingly, the type I interferon response gene module was enriched in the neutrophil cell cluster (Fig. 2I). These results suggest that the SARS-CoV-2 activates the IFN response pathway in neutrophil cells.

Altogether, our analysis depicted the distribution of SARS-CoV-2-infected cells and *Hemophilus parahemolyticus*-infected cells in BALF samples and revealed the activated IFN response by SARS-CoV-2.

Performance of PathogenTrack under various conditions

Next, we systematically evaluated various factors that may affect the detection performance of our method. Since there was no "gold standard" data to assess the accuracy of pathogen detection at the single-cell level, we evaluated our method on simulated data sets. We employed minnow to simulate 810 data sets of host cells infected with pathogens under various simulation parameters. 20 microbes, including 10 bacteria (the Gram-positive *Clostridioides difficile*, *Clostridium perfringens*, *Corynebacterium diphtheriae*, *Listeria monocytogenes*, and *Staphylococcus aureus*; the Gram-negative *Chlamydia trachomatis*, *Helicobacter pylori*, *Legionella pneumophila*, *Salmonella enterica*, and *Vibrio cholerae*) and 10 viruses (EBV, HIV, *Human metapneumovirus* (hMPV), *Human papillomaviruses* (HPV), *Herpes simplex virus* (HSV), *Molluscum contagiosum virus* (MCV), *Middle East respiratory syndrome coronavirus* (MERS), *Rabies virus*, SARS-CoV-2, and *Varicella-zoster virus* (VZV)), were involved in the stimulation. The technological features including UMI length, read length, PCR cycles, pathogen infection levels, and σ (the standard deviation of start positions) were considered. As shown in Fig. 3A–3C, UMI length, PCR cycles, and σ have almost no effect on the performance of our method. The performance increases with the augmentation of read length or pathogen infection level (Fig. 3D and 3E). Read length over 100 bp showed good performance. In addition, we analyzed the accuracy of our method. A good agreement between the number of pathogen-infected cells predicted by our method and the expected genuine was shown in Fig. 3F.

Comparison with existed tools on simulated scRNA-seq data sets

Since Viral-Track [9] was the only single-cell intracellular virus detection tool, we compared the sensitivity of our algorithm with Viral-Track on the 810 simulated data sets. The performance statistics of Viral-Track are illustrated in Fig. S1. As is depicted in Fig. S1A, PathogenTrack performs as good as Viral-Track in virus detection. More importantly, PathogenTrack has the ability in detecting bacteria with high sensitivities (Fig. S1A).

To reduce the computational time when evaluating the time and memory consumptions of both methods, we randomly sampled 100 to 4000 cells. Benchmarking showed a roughly linear relationship between the number of cells and the processing time required, and the maximum memory consumption of PathogenTrack is less than that of Viral-Track (Fig. S1B).

Performance of PathogenTrack and Yeskit in real scRNA-seq data sets

To compare the performance of Viral-Track and PathogenTrack for detecting pathogenic reads in real human samples, we next compared the results on several real human scRNA-seq data sets (Table S1). These data sets consist of a variety of tissues and cell lines (blood, lung, intestine, stomach, and lymphoblastoid cell lines) and various well-known viruses and bacteria: *influenza A* (H1N1 and H3N2), *human immunodeficiency virus 1* (HIV-1), *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), *Epstein-Barr virus* (EBV), and *Helicobacter pylori* (*H. pylori*). Since cellranger or alevin provides more reliable CBs than UMI-tools, we used the default parameters to run these two methods, except that the input whitelist file was changed to barcode file generated by cellranger (applicable to the 10x Genomics platform). Because the pathogenic microbes and their infected cells were not aware in real data sets, only the well-known pathogens in each data set were evaluated.

PathogenTrack is robust in detecting pathogens at the single-cell level

We observed that the number of virus-infected cells predicted by PathogenTrack was close to that of Viral-Track's, but there were a few differences (Fig. 4). In the *in vitro* influenza A data (Cal07), PathogenTrack obtained nearly equal virus-infected cells to Viral-Track in infected samples, but fewer or none in controls (Fig. 4A). It may imply PathogenTrack has higher specificity in detecting

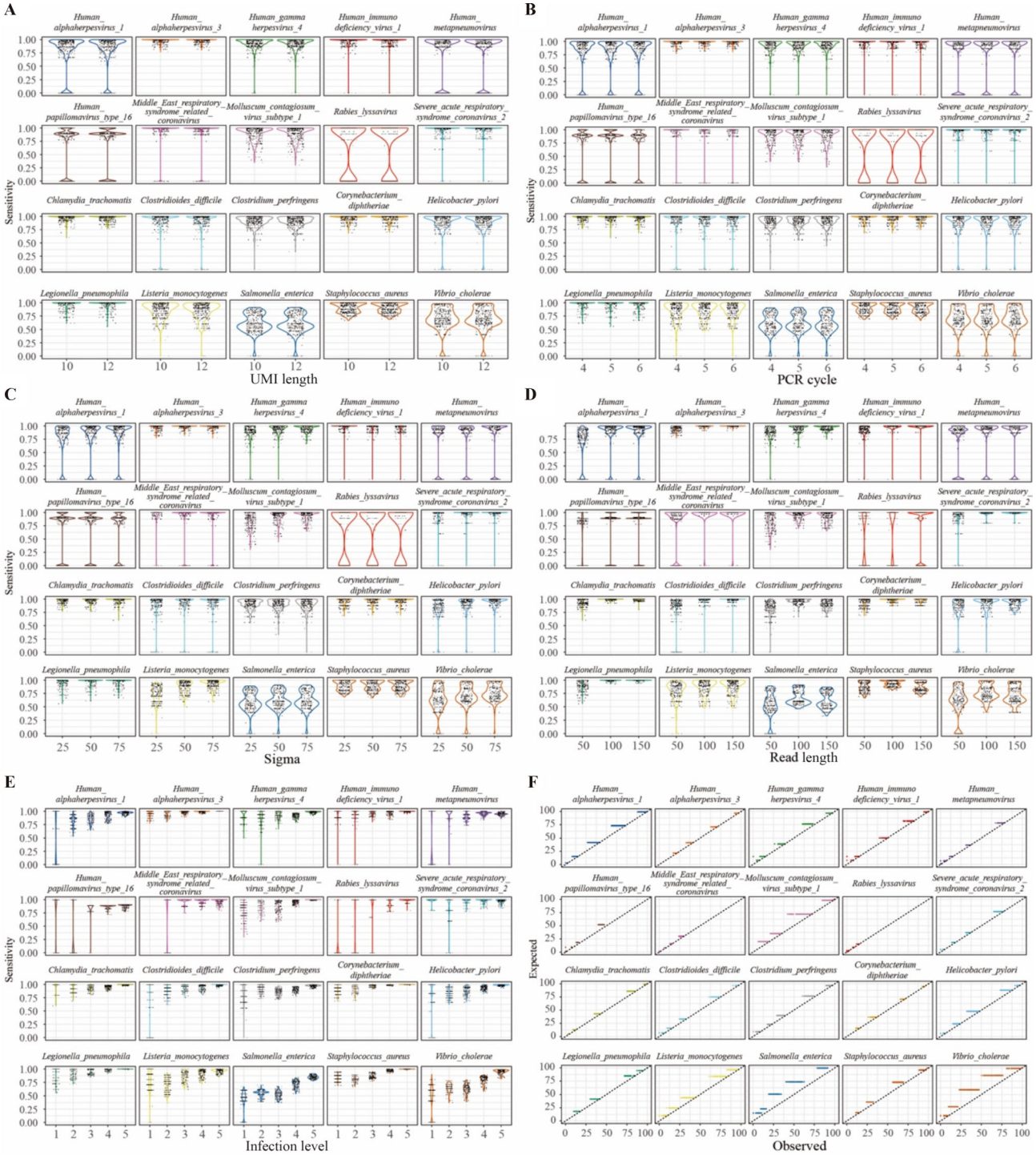

Fig. 3 Performance of PathogenTrack under various simulation parameters. (A–E) The impact of UMI length, PCR cycles, Sigma, read length, and infection level on the performance of the PathogenTrack. In each panel, only one simulation parameter is varied, as shown on the *x*-axis. (A) UMI length showed no impact on the performance. 10 bp and 12 bp UMI length were used for simulation. (B) PCR cycles showed no impact on the performance. 4–6 PCR cycles were used for simulation. (C) Sigma showed no impact on the performance. Three different Sigma, the standard deviation of start positions, were used simulation. (D) Longer read length showed better performance. The 50 bp, 100 bp, 150 bp read lengths were used for simulation. (E) Higher infection level showed better performance. Five infection levels (levels 1–5) were used for simulation. (F) Accuracy of the PathogenTrack in pathogen detection. Scatter plot shows the relation of the number of pathogen-infected cells predicted by PathogenTrack and the expected genuine.

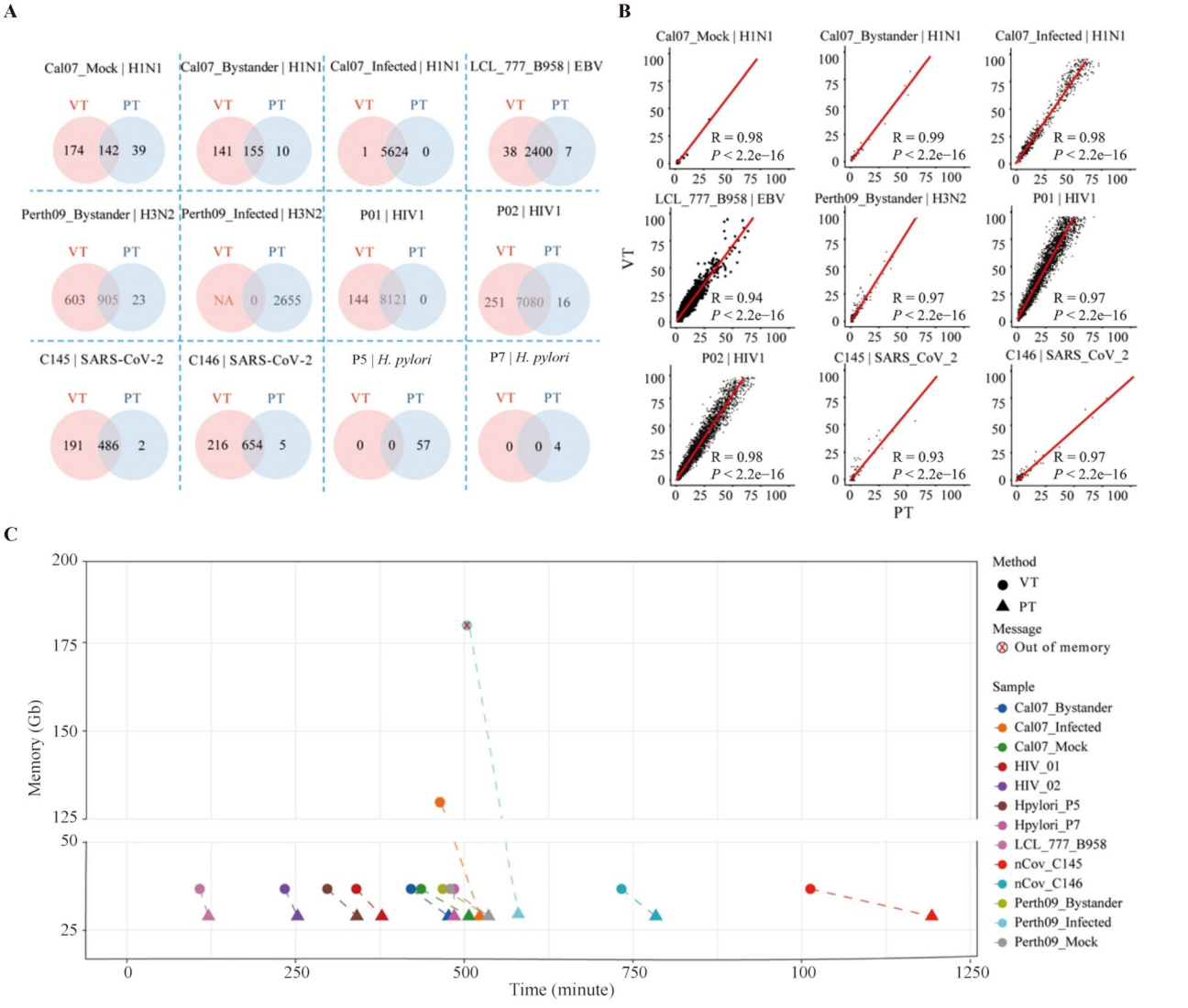

Fig. 4 Performance evaluation on 13 real scRNA-seq data sets (VT, Viral-Track; PT, PathogenTrack). (A) Venn diagram shows the logical relation between the number of pathogen-infected cells detected by Viral-Track and PathogenTrack. (B) Scatter plot shows the correlation between the UMI counts of pathogen-infected cells generated by Viral-Track and PathogenTrack (UMIs $\leqslant$ 100). (C) Time and memory performance of Viral-Track and PathogenTrack on 13 real data sets. Note that sample Perth09_Infected ran out of memory when running on a computer with 180 Gb RAM.

virus-infected cells. Besides, the Viral-Track failed to detect any bacteria, such as *H. pylori* (Fig. 4A). Next, we calculated the correlation of the UMI counts of the intracellular pathogens predicted by both methods for each sample. The results show that pathogen UMI counts are in good agreement between these two methods (Fig. 4B).

We further tested the applicability of PathogenTrack in tracing bacterial reads in human clinical samples. We ran PathogenTrack on 13 gastric antral mucosa biopsy scRNA-seq data sets (SRA accession number: SRP215370). *Helicobacter pylori* was detected only in two samples (Fig. 4A, last two venn diagram), which was consistent with the clinical information in the original paper [30].

Subsequently, we asked if the pathogen-infected cells only predicted by PathogenTrack are reliable. Since it seems impossible to conduct experiments at the single-cell level to verify pathogen infection events, we used bioinformatics tool to address this question. As PathogenTrack provides pathogenic sequences that support each cell infection event, we randomly selected a few pathogenic reads and verified the predictions by performing BLAST searches on the Nucleotide database.

The time and memory consumption of these two methods are illustrated in Fig. 4C. PathogenTrack consumed almost constant memory (~30 Gb) and reasonable time on all real data sets, while Viral-Track consumed up to 128 Gb of memory in Cal07_Infected data and ran

out of memory in Perth09_Infected data (> 180 Gb RAM). One reason could be that PathogenTrack detects pathogens at read-level while Viral-Track consumes more memory when assembling a virus genome.

Discussion

The last few years have witnessed emerging of many computational methods on detecting microbe-derived sequences in human clinical samples. However, methods for thoroughly identifying and interpreting intracellular pathogens at the single-cell level are rarely documented. In this study, we propose a computational method for identifying (PathogenTrack) and exploring (Yeskit) intracellular pathogens (viruses and bacteria) at the single-cell level from unmapped scRNA-seq data. PathogenTrack performed robustly on simulated and clinical data sets; Yeskit provided valuable information about pathogen infection status and pathogen-induced pathways. The currently accessible human clinical single-cell transcriptome data provide us with an excellent opportunity to identify these pathogen species and explore their functions in disease progression. These pathogens may indicate disease states and shed new lights on drug targets.

Using COVID-19 as an example, we showed the SARS-CoV-2 existed in neutrophils, lymphocytes, macrophages, and epithelial cells. In addition, SARS-CoV-2 positive cells exhibited distinct gene expression patterns compared to those bystander cells. Although the SARS-CoV-2 mainly enter the host cell via the ACE2 protein, which is primarily expressed in type II alveolar epithelial cells, the SARS-CoV-2 might also enter immune cells owing to the following reasons: (1) viral overload; (2) the SARS-CoV-2 might be swallowed via endocytosis, trogocytosis, and phagocytosis. SARS-CoV-2 positive neutrophils showed an enhanced expression of genes related to interferon response and antigen-presenting; (3) ACE2 was found to be expressed on the surface of pulmonary macrophages [31], which might serve as an entry path for SARS-CoV-2 into these cells. As a result, there could be viral proliferation in the pulmonary macrophage, leading to a vicious cycle of severe pulmonary viral infection and/or cytokine release syndrome. Alternately, pulmonary macrophage may play a role in antigen processing/presentation to other immune cells [31,32]. Analysis of the BALF data suggests the SARS-CoV-2 may trigger distinct transcriptional programs related to aberrant immune response in COVID-19.

In addition to infectious disease, emerging studies have also unveiled the intracellular pathogens, including viruses and bacteria, in the progression of malignant diseases [33–35]. Generally, the pathogenic microbiome is present in the digestive tract and respiratory tract [36–40]. Studies have also found that certain microbiomes may exist in cells and contribute to the occurrence and development of malignant diseases [37,41]. For instance, many pathogens have been reported to be key regulators in malignant diseases, such as EBV, HBV, and HPV. However, how these viruses engaged in tumorigenesis remains unclear. We have shown the PathogenTrack can detect these pathogens at single-cell levels. In the current study, SARS-CoV-2 was enrolled as an example to illustrate the capacity of the PathogenTrack and Yeskit. Future studies focusing on other pathogen-related disease, such as malignant disease, might be conducted to unveil the potential regulatory role of pathogen in disease progression at the single-cell level. We might establish an online database for pathogens at single-cell level in multiple pathogen-related disease, such as EBV-related lymphoma, HBV-related hepatocellular carcinoma, HPV-related cervical cancer, and bacteria in multiple cancers. Currently, our method might only be applicable to the 10x Genomics and microwell-based scRNA-seq data sets, which are the most commonly used method to conduct scRNA-seq, and we might expand the application of our tools in future.

Acknowledgements

We thank the support from Prof. Gang Lv and the ASTRA computing platform in the National Research Center for Translational Medicine (Shanghai) and the Pi computing platform in the Center for High Performance Computing at Shanghai Jiao Tong University. This work was supported by grants from National Natural Science Foundation of China (Nos. 8210010124 and 81890994), Double First-Class Project (No. WF510162602), National Key R&D Program of China (No. 2019YFA0905902), Natural Science Foundation of Shanghai (Nos. 21ZR1480900 and 21YF1427900), Shanghai Jiao Tong University (No. YG2021-QN19), and the Shanghai Guangci Translational Medical Research Development Foundation.

Compliance with ethics guidelines

Wei Zhang, Xiaoguang Xu, Ziyu Fu, Jian Chen, Saijuan Chen, and Yun Tan declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This article does not contain any studies with human or animal subjects.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11684-021-0915-9 and is accessible for authorized users.

References

1. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM,

Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017; 8(1): 14049

2. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 2017; 357(6352): 661–667
3. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017; 14(11): 1083–1086
4. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 2018; 18(1): 35–45
5. Zhang Q, He Y, Luo N, Patel SJ, Han Y, Gao R, Modak M, Carotta S, Haslinger C, Kind D, Peet GW, Zhong G, Lu S, Zhu W, Mao Y, Xiao M, Bergmann M, Hu X, Kerkar SP, Vogt AB, Pflanz S, Liu K, Peng J, Ren X, Zhang Z. Landscape and dynamics of single immune cells in hepatocellular carcinoma. Cell 2019; 179(4): 829–845.e20
6. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. Inference and analysis of cell–cell communication using CellChat. Nat Commun 2021; 12(1): 1088
7. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018; 50(8): 1–14
8. Westermann AJ, Barquist L, Vogel J. Resolving host-pathogen interactions by dual RNA-seq. PLoS Pathog 2017; 13(2): e1006033
9. Bost P, Giladi A, Liu Y, Bendjelal Y, Xu G, David E, Blecher-Gonen R, Cohen M, Medaglia C, Li H, Deczkowska A, Zhang S, Schwikowski B, Zhang Z, Amit I. Host-viral infection maps reveal signatures of severe COVID-19 patients. Cell 2020; 181(7): 1475–1488.e12
10. Srivastava A, Malik L, Smith T, Sudbery I, Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol 2019; 20(1): 65
11. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 2017; 27(3): 491–499
12. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018; 34(17): i884–i890
13. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29(1): 15–21
14. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019; 20(1): 257
15. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell 2019; 177(7): 1888–1902.e21
16. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019; 16(12): 1289–1296
17. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 2020; 21(1): 12
18. Alexa A, Rahnenführer J. Gene set enrichment analysis with topGO. Bioconductor Improv 2009; 27: 1–26
19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011; 27(12): 1739–1740
20. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 2017; 18(1): 174
21. Sarkar H, Srivastava A, Patro R. Minnow: a principled framework for rapid simulation of dscRNA-seq data at the read level. Bioinformatics 2019; 35(14): i136–i144
22. Li WV, Li JJ. A statistical simulator scDesign for rational scRNA-seq experimental design. Bioinformatics 2019; 35(14): i41–i50
23. Zhang X, Xu C, Yosef N. Simulating multiple faceted variability in single cell RNA sequencing. Nat Commun 2019; 10(1): 2611
24. Dibaeinia P, Sinha S. SERGIO: a single-cell expression simulator guided by gene regulatory networks. Cell Syst 2020; 11(3): 252–271.e11
25. Tian J, Wang J, Roeder K. ESCO: single cell expression simulation incorporating gene co-expression. Bioinformatics 2021; 37(16): 2374–2381
26. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics 2015; 31(17): 2778–2784
27. Hie B, Cho H, DeMeo B, Bryson B, Berger B. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell Syst 2019; 8(6): 483–493.e7
28. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, Liu L, Amit I, Zhang S, Zhang Z. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med 2020; 26(6): 842–844
29. Le Floch AS, Cassir N, Hraiech S, Guervilly C, Papazian L, Rolain JM. Haemophilus parahaemolyticus septic shock after aspiration pneumonia, France. Emerg Infect Dis 2013; 19(10): 1694–1695
30. Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, Du S, Li S. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Rep 2019; 27(6): 1934–1947.e5
31. Wang C, Xie J, Zhao L, Fei X, Zhang H, Tan Y, Nie X, Zhou L, Liu Z, Ren Y, Yuan L, Zhang Y, Zhang J, Liang L, Chen X, Liu X, Wang P, Han X, Weng X, Chen Y, Yu T, Zhang X, Cai J, Chen R, Shi ZL, Bian XW. Alveolar macrophage dysfunction and cytokine storm in the pathogenesis of two severe COVID-19 patients. EBioMedicine 2020; 57: 102833
32. Tan Y, Zhang W, Zhu Z, Qiao N, Ling Y, Guo M, Yin T, Fang H, Xu X, Lu G, Zhang P, Yang S, Fu Z, Liang D, Xie Y, Zhang R, Jiang L, Yu S, Lu J, Jiang F, Chen J, Xiao C, Wang S, Chen S, Bian XW, Lu H, Liu F, Chen S. Integrating longitudinal clinical laboratory tests with targeted proteomic and transcriptomic analyses reveal the landscape of host responses in COVID-19. Cell Discov 2021; 7(1): 42
33. Rodriguez RM, Hernandez BY, Menor M, Deng Y, Khadka VS. The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing. Comput Struct Biotechnol J 2020; 18: 631–641
34. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-

Montgomery S, Heaton R, Mckay R, Patel SP, Swafford AD, Knight R. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 2020; 579(7800): 567–574

35. Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, Rotter-Maskowitz A, Weiser R, Mallel G, Gigi E, Meltser A, Douglas GM, Kamer I, Gopalakrishnan V, Dadosh T, Levin-Zaidman S, Avnet S, Atlan T, Cooper ZA, Arora R, Cogdill AP, Khan MAW, Ologun G, Bussi Y, Weinberger A, Lotan-Pompan M, Golani O, Perry G, Rokah M, Bahar-Shany K, Rozeman EA, Blank CU, Ronai A, Shaoul R, Amit A, Dorfman T, Kremer R, Cohen ZR, Harnof S, Siegal T, Yehuda-Shnaidman E, Gal-Yam EN, Shapira H, Baldini N, Langille MGI, Ben-Nun A, Kaufman B, Nissan A, Golan T, Dadiani M, Levanon K, Bar J, Yust-Katz S, Barshack I, Peeper DS, Raz DJ, Segal E, Wargo JA, Sandbank J, Shental N, Straussman R. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. Science 2020; 368(6494): 973–980
36. Gareau MG, Sherman PM, Walker WA. Probiotics and the gut microbiota in intestinal health and disease. Nat Rev Gastroenterol Hepatol 2010; 7(9): 503–514
37. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet 2012; 13(4): 260–270
38. Sommer F, Anderson JM, Bharti R, Raes J, Rosenstiel P. The resilience of the intestinal microbiota influences health and disease. Nat Rev Microbiol 2017; 15(10): 630–638
39. Sanders ME, Merenstein DJ, Reid G, Gibson GR, Rastall RA. Probiotics and prebiotics in intestinal health and disease: from biology to the clinic. Nat Rev Gastroenterol Hepatol 2019; 16(10): 605–616
40. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. Cell Res 2020; 30(6): 492–506
41. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol 2009; 9(5): 313–323