Unambiguous detection of SARS-CoV-2 subgenomic mRNAs with single cell RNA sequencing

Phillip Cohen[1], Emma J DeGrace[1], Oded Danziger[1], Roosheel S Patel[1], Brad R Rosenberg[1]#

[1]Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY 10035

#Address correspondence to Brad R Rosenberg, brad.rosenberg@mssm.edu

Abstract

Single cell RNA sequencing (scRNAseq) studies have provided critical insight into the pathogenesis of Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2), the causative agent of COronaVIrus Disease 2019 (COVID-19). scRNAseq workflows are generally designed for the detection and quantification of eukaryotic host mRNAs and not viral RNAs. The performance of different scRNAseq methods to study SARS-CoV-2 RNAs has not been thoroughly evaluated. Here, we compare different scRNAseq methods for their ability to quantify and detect SARS-CoV-2 RNAs with a focus on subgenomic mRNAs (sgmRNAs), which are produced only during active viral replication and not present in viral particles. We present a data processing strategy, single cell CoronaVirus sequencing (scCoVseq), which quantifies reads unambiguously assigned to sgmRNAs or genomic RNA (gRNA). Compared to standard 10X Genomics Chromium Next GEM Single Cell 3′ (10X 3′) and Chromium Next GEM Single Cell V(D)J (10X 5′) sequencing, we find that 10X 5′ with an extended R1 sequencing strategy maximizes the unambiguous detection of sgmRNAs by increasing the number of reads spanning leader-sgmRNA junction sites. Differential gene expression testing and KEGG enrichment analysis of infected cells compared with bystander or mock cells showed an enrichment for COVID19-associated genes, supporting the ability of our method to accurately identify infected cells. Our method allows for quantification of coronavirus sgmRNA expression at single-cell resolution, and thereby supports high resolution studies of the dynamics of coronavirus RNA synthesis.

Importance

Single cell RNA sequencing (scRNAseq) has emerged as a valuable tool to study host-viral interactions particularly in the context of COronaVIrus Disease-2019 (COVID-19). scRNAseq has been developed and optimized for analyzing eukaryotic mRNAs, and the ability of scRNAseq to measure RNAs produced by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has not been fully characterized. Here we compare the performance of different scRNAseq methods to detect and quantify SARS-CoV-2 RNAs and develop an analysis workflow to specifically quantify unambiguous reads derived from SARS-CoV-2 genomic RNA and subgenomic mRNAs. Our work demonstrates the strengths and limitations of scRNAseq to measure SARS-CoV-2 RNA and identifies experimental and analytical approaches that allow for SARS-CoV-2 RNA detection and quantification. These developments will allow for studies of coronavirus RNA biogenesis at single-cell resolution to improve our understanding of viral pathogenesis.

Introduction

Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) is the causative agent of COronaVIrus Disease-2019 (COVID-19), which as of November 2021 has caused over 250 million cases and over 5 million deaths globally(1, 2). Global efforts to understand the pathogenesis of SARS-CoV-2 infection have led to the development of vaccines and antivirals, which have reduced morbidity and mortality(3). "Omics" methods have been instrumental in studying SARS-CoV-2 in part because they have generated large amounts of data regarding host-viral interactions at unprecedented speed(4–15). Single-cell RNA sequencing (scRNAseq) studies in particular have been used to study viral tropism(16–22), peripheral immune changes(23–33), transcriptional changes induced by infection(34, 35), and to develop cell atlases of COVID-19 pathology(23, 24, 36, 37). Of note, most scRNAseq workflows have been developed and optimized for studies of eukaryotic transcription but not viral, specifically SARS-CoV-2, transcription. The performance of different scRNAseq methods to detect and quantify viral RNAs may impact the analysis and interpretation of such studies.

SARS-CoV-2 is a betacoronavirus with a 29 kB positive-sense, single stranded RNA genome(38, 39). SARS-CoV-2 generates genomic RNA (gRNA), subgenomic mRNAs (sgmRNAs), and negative-sense antigenomic RNA during active infection(40, 41). Both gRNA and sgmRNAs are poly-adenylated, which enables detection by scRNAseq protocols that rely on poly-T primed reverse transcription(39–41). Translation of gRNA results in the production of one of two polyproteins, pp1a and pp1ab, which are subsequently cleaved into an array of non-structural proteins involved in pathogenesis and replication(39, 41). Translation of sgmRNAs generates structural and accessory viral proteins critical for virion production and pathogenesis(39, 41). sgmRNAs are produced only in cells with actively replicating virus, while gRNA is present in both infected cells and virions(40, 41). Therefore, specific detection of sgmRNAs can allow for: 1) specific identification of cells with actively replicating virus and 2) analysis of the dynamics of viral gene expression within and across cells and viruses.

sgmRNAs are generated by discontinuous transcription events during negative strand synthesis(40). Transcription Regulatory Sequences (TRS), present in the 5′ leader sequence of the virus (TRS-L) and upstream of each ORF body (TRS-B), regulate this process(40). Template switching of the viral polymerase from a TRS-B to a TRS-L generates sgmRNAs with the 5′ leader sequence fused to the sgmRNA ORF body (**Figure 1A**)(40). These “nested” sgmRNAs share the viral ORF sequence downstream of the junction site in addition to a common leader sequence upstream of the junction site(40). This redundancy poses a challenge for standard scRNAseq data processing pipelines because reads mapping to redundant sgmRNA sequences are categorized as “ambiguous” and typically excluded from quantification. This problem has been addressed in bulk RNAseq by quantifying SARS-CoV-2 reads spanning leader-ORF junctions, which unambiguously identify sgmRNAs(12, 15). However, many scRNAseq methods do not sequence this region of sgmRNAs at significant coverage due to differences in library format and configuration of sequencing reads.

We hypothesized that both experimental (i.e. scRNAseq method) and data processing decisions influence the ability to detect, resolve, and quantify SARS-CoV-2 RNA species with scRNAseq. We developed a data processing workflow, single cell coronavirus sequencing (scCoVseq), to quantify only unambiguous SARS-CoV-2 reads in scRNAseq data. We found that SARS-CoV-2 RNA detection differed by 10X Genomics Chromium scRNAseq method, due in part to ambiguity of the library fragments generated by each method. We show that 10X Chromium Next GEM Single Cell V(D)J (10X 5′) scRNAseq with an extended read 1 (R1) sequencing strategy maximized unambiguous SARS-CoV-2 reads and thereby increased detection of SARS-CoV-2 RNAs. Using this method, we identify infected and uninfected “bystander” cells within the same culture and determine differentially expressed genes between infected, bystander, and mock cells.

Materials and Methods

Cell lines and Viral Infection

Vero-E6 cells (ATCC, CRL-1586) were maintained in Dulbecco's Modified Eagle Medium (DMEM, Corning #10-017-CV) supplemented with 10% fetal-bovine serum (FBS) and 1% Penicillin Streptomycin (PSN, Fisher scientific #15-140-122), and routinely cultured at 37° C with 5% CO_2.

SARS-CoV-2 (isolate USA-WA1/2020, BEI resource NR-52281) and control media (mock infected) stocks were grown by inoculating a confluent T175 flask of Vero-E6 cells (passage 2). Mock and SARS-CoV-2 infected cultures were maintained in reduced serum DMEM (2% FBS) for 72 hours, after which culture media was collected and filtered by centrifugation (8000 x g, 15 minutes) using an Amicon Ultra-15 filter unit with a 100KDa cutoff filter (Millipore # UFC910024). Concentrated stocks in reduced-serum media (2% FBS), supplemented with 50mM HEPES buffer (Gibco #15630080) were stored at -80°C. Viral titers were determined by plaque assay as previously described(42). All SARS-CoV-2 propagations and experiments were performed in a Biosafety Level 3 facility in compliance with institutional protocols and federal guidelines.

scRNAseq

For scRNAseq experiments, Vero-E6 cells in 6 well plates were infected with SARS-CoV-2 at a MOI of 0.1, or with an equivalent volume of control media, in reduced-serum media (2% FBS) for 24 hours. To prepare cells for scRNAseq, mock and SARS-CoV-2 infected cultures were washed with calcium/magnesium-free PBS and disassociated using TrypLE (Gibco # 12605010, 5 minutes at 37° C), after which samples were centrifuged (200 x g, 5 minutes), resuspended in calcium/magnesium-free PBS supplemented with 1% BSA, and counted. Mock and SARS-CoV-2 infected cell culture samples were filtered through a 40μm FlowMi strainer (ScienceWare # H13680-0040) and counted prior to loading on the 10X Genomics Chromium Controller according to manufacturer's protocol. Mock and infected samples were loaded on separate lanes of a 10X Genomics Chromium Controller for either NextGEM Single Cell 3′ v3.1 (10X 3′), or NextGEM Single Cell V(D)J v1.1, (10X 5′).

Gene expression libraries were prepared for 10X 3′ and 5′ samples according to manufacturer's guidelines. Final 10X 3′ mock and infected gene expression libraries and the 10X 5′ infected gene expression library were PCR amplified for 16 cycles while the 10X 5′ mock gene expression library was amplified for 14 cycles. 10X 3′ gene expression libraries were pooled and sequenced by short-read sequencing on an Illumina NextSeq 500 using a high output 150 cycle reaction kit according to manufacturers' protocol with the following read lengths: read 1 28 nt; i7 index 8 nt; and read 2 130 nt. 10X 5′ gene expression libraries were also pooled and sequenced with 10X recommended read lengths (read 1 26 nt; i7 index 8 nt; and read 2 132 nt) or with extended R1 protocol (read 1 158 nt; i7 index 8 nt; no read 2).

scRNAseq Pre-Processing

Conversion of Illumina BCL files to fastq

Fastq files for standard sequencing 3′ and 5′ gene expression libraries were generated using the mkfastq command in cellranger v.3.1.0 (10X Genomics). Fastq files for 5′ libraries sequenced with the extended R1 strategy were generated using bcl2fastq v2.20.0 (Illumina, Inc). Extended R1 fastqs were then separated into pseudo R1 fastqs, containing the cell barcode and UMI, and pseudo read 2 (R2) fastqs, containing cDNA sequence, using a customized Python/3.7.3 script (available at github link pending) as follows. The cell barcode and UMI are selected from the first 26 bp of R1. The subsequent 13 bp derive from the template switch oligonucleotide and are ignored. The remaining nucleotides (and corresponding quality scores) are reverse complemented and stored as pseudo R2. The read header of the pseudo R2 fastqs are modified to reflect the format for standard R2 fastqs.

Downsampling fastqs to control for sequencing depth

To control for differences in sequencing depth for each library, read depth per library was downsampled to approximately 50,000 reads per cell. To generate a whitelist of cell barcodes for downsampling while accounting for transcriptional shutdown in SARS-CoV-2 infected cells (35), we generated preliminary gene x cell matrices for our dataset using cellranger/3.1.0 count (10X Genomics, Inc) to quantify and align reads to a host reference (African Green Monkey, ChlSab1.1) combined with SARS-CoV-2 transcripts as annotated by the NCBI SARS-CoV-2 reference (NC_045512.2) with modifications for USA/WA01 strain for each dataset. The resulting output was analyzed in R/4.0.4 with Seurat/4.0.1(43–45) to filter out putative doublets and empty droplets according to total UMIs/cell, number of genes/cell, and percent of mitochondrial gene expression. After filtering, putative cell-containing cell barcodes were output to a whitelist per library. Based on the these whitelists, the initial fastq files were downsampled using seqtk (version 1.2)(46) to a total read depth of 50,000 multiplied by the number of cells in the library.

Preparation of empirically derived SARS-CoV-2 genome reference

Downsampled fastq files were then mapped using cellranger count/3.1.0 (10X Genomics, Inc) to an empirically defined reference of SARS-CoV-2 sgmRNAs derived from previously reported SARS-CoV-2 (BetaCoV/Korea/KCDC03/2020) RNAs sequenced with long-read direct RNA Nanopore sequencing(12). These were downloaded from the UCSC Genome Browser Table Browser(47) after filtering for TRS-dependent transcripts and score > 900 and exporting to gtf format. Transcripts for previously unknown ORFs were excluded from the annotation. An additional annotation for genomic RNA was included which covered the entire length of the SARS-CoV-2 genome. Aligning the BetaCoV/Korea/KCDC03/2020 genome with USA/WA-CDC-WA1/2020 genome showed that the USA/WA-CDC-WA1/2020 had an additional 21 3′ adenosine nucleotides annotated. To account for this in our reference, we extended any annotations from BetaCoV/Korea/KCDC03/2020 that ended at the 3′ end of the genome by an additional 21 bases. This SARS-CoV-2 reference was appended to the host ChlSab1.1 Ensembl reference.

scCoVseq

To unambiguously assign and quantify scRNAseq reads to SARS-CoV-2 RNAs, the cellranger output BAM was filtered for reads mapping to SARS-CoV-2 or ChlSab1.1 references using samtools (version 1.11)(48). SARS-CoV-2 aligned reads were then subset to likely genomic RNA reads or sgmRNA reads. Genomic reads were defined as those containing no gaps in their alignment and mapping upstream of the start of the most 5′ sgmRNA, S. sgmRNA reads were defined as SARS-CoV-2 reads containing a gap and mapping in part to the 5′ leader sequence, defined as the 5′ proximal 80 nucleotides of the SARS-CoV-2 genome, and in part 3′ to the start of S. All other reads mapping to the SARS-CoV-2 genome were discarded. Reads passing these filtering steps were quantified with umi_tools (version 1.0.0)(49). An R/3.5.3 script using the Matrix (version 1.2-18)(50) and readr (version 1.3.1)(51) packages was used to convert this to a sparse matrix and save as an rds file to decrease file size. UMIs that were assigned to multiple genes were removed from the resulting matrix during analysis.

scRNAseq Analysis

Sashimi Plots

Reads from 10X 3′, 10X 5′, and 10X 5′ extended R1 data that were aligned to the SARS-CoV-2 reference by cellranger were subset from the cellranger output BAM file. Each BAM file was downsampled to approximately 1 x 10^6 reads to control for differences in sequencing depth across libraries. Sashimi plots were generated with ggsashimi (version 1.0.0)(52).

Classification of SARS-CoV-2 Infected Cells

scCoVseq-derived gene by cell matrices were loaded into R/4.0.4 and analyzed with the Seurat/4.0.1(43–45) package. For each 10X method, mock and infected gene x cell matrices were merged with the Seurat merge command. Scaled SARS-CoV-2 UMI expression of 600 sampled cells were clustered with five methods (k means clustering, hierarchical/Ward clustering, DIANA, mixture model-based clustering, and k medoids clustering) using the clValid (version 0.7) package(53). Based on optimal performance as measured by average distance, average distance between means, average proportion of non-overlap, connectivity, Dunn index, figure of merit, and silhouette width, k-medoids clustering implemented with the PAM algorithm and k set to 2 optimally separated infected from uninfected cells. Therefore to identify infected and bystander cells within SARS-CoV-2 treated cultures, euclidean distance between the z-scaled expression of SARS-CoV-2 sgmRNA UMIs per cell was clustered using pam (k = 2) implemented in the cluster (version 2.1.2) package(54). Output clusters were then compared for viral UMI expression per cell, and the cluster with more viral UMIs was classified as infected and the other as uninfected.

Comparison of SARS-CoV-2 RNA UMIs per scRNAseq Method

To examine the distribution of SARS-CoV-2 UMIs per cell by scRNAseq method, the 25th percentile of total UMIs was quantified for all infected cells from each 10X method. Any cells with fewer UMIs than the minimum 25th percentile of all samples were discarded, and all cells were subsequently downsampled to this same number of total UMIs/cell using the Seurat SampleUMI command. Each dataset was randomly downsampled to the same number of infected cells to equalize for differences in cell numbers, and viral sgmRNA UMIs/cell were plotted by scRNAseq method.

SARS-CoV-2 Read Distribution by scRNAseq Method

SARS-CoV-2 reads were defined as genomic or subgenomic using scCoVseq. Reads aligning to the SARS-CoV-2 reference that were excluded from scCoVseq were classified as ambiguous. The number of genomic, subgenomic, or ambiguous reads per million SARS-CoV-2 reads was calculated and plotted for each scRNAseq method.

Differential Expression Analysis

To explore expression differences between infected, bystander, and mock cells, differential expression testing with edgeR (version 3.32.1) was performed with modifications for scRNAseq as previously described(55, 56). Viral genes were excluded from analysis, and only host genes expressed in at least 10% of cells were tested. To account for differences in RNA content of infected cells due to virally-induced transcriptional shutdown, all cells were downsampled to the 25th percentile of total UMIs of infected cells. Cells with fewer UMIs than the threshold were excluded from analysis. Differential gene expression was performed with edgeR using a generalized linear model quasi-likelihood F test adapted with a term for gene detection rate(55, 56). Genes with an absolute $\log_2$ fold change greater than or equal to 1 and false discovery rate less than 0.05 were considered significant. For KEGG enrichment analysis, pairwise tests between mock, bystander, and infected cells were performed. Differentially expressed genes with an absolute $\log_2$ fold change greater than or equal to 1 and false discovery rate less than 0.05 were considered significant and subject to KEGG enrichment analysis using the KEGG annotations for African Green Monkey as implemented in the edgeR function kegga.

Quantification of SARS-CoV-2 sgmRNA Junction Sites

We explored the ability of our extended R1 sequencing to detect SARS-CoV-2 sgmRNA junctions using STARsolo (version 2.7.8a)(57). Aligned reads were re-mapped to the empirical SARS-CoV-2 annotation described above and junction sites per cell were quantified. The resulting junction per cell matrix was plotted in R/v4.0.4.

Flow Cytometry

Vero E6 cells were fixed with 4% paraformaldehyde at room temperature for a minimum of 24 hours, washed once with PBS and permeabilized with 1X perm-wash buffer (BDBiosciences #554723) for 5 minutes. SARS-CoV nucleocapsid (N) antibody (clone 1C7C7) (kindly provided by Thomas Moran, Icahn School of Medicine at Mount Sinai, New York, NY), conjugated to AlexaFluor 647 was diluted 1:400 in perm-wash buffer, and added directly to samples. Samples were then incubated at room temperature for 40 minutes in the dark. After staining, samples were washed once with 1X perm-wash buffer, once with PBS, resuspended in FACS buffer (PBS supplemented with 1% FBS), and acquired on a Gallios flow cytometer (Beckman-Coulter). For all viral infections, analysis was performed with FlowJo software (version 10.7.1, Becton Dickinson), excluding cell doublets and debris and gating according to mock infected populations.

Immunofluorescence microscopy

Vero E6 were seeded in 6-well plates (Falcon REF-353046) with one coverslip (Fisher Scientific 12-550-143) per well. After 24 hours post infection, cells were washed with PBS and fixed with 4% paraformaldehyde (Fisher Scientific AA433689M) overnight. Fixed cells were permeabilized using 0.1% Triton-X (Fisher Scientific AC327371000) in PBS and blocked with 4% bovine serum albumin (BSA, Fisher Scientific BP1600-100) in PBS. Blocked coverslips were incubated with mouse anti-SARS-CoV N antibody (clone 1C7, 1:500 in 4% BSA PBS) overnight at 4C, washed three times with PBS, and incubated for 45 minutes with 1:500 AlexaFluor 488-conjugated anti-mouse (Invitrogen A11001, 1:500 in 4% BSA PBS) plus DAPI (Thermo Fisher Scientific D1306, 1:1000 in 4% BSA PBS) at room temperature. Coverslips were then stained with phalloidin (1:400 in PBS) for 1 hour at room temperature and washed again three times with PBS. Coverslips were mounted using Prolong Diamond (Life Technologies P36970). Confocal laser scanning was performed using a Leica SP5 DMI (ISMMS Microscopy CoRE and Advanced Bioimaging Center) with a ×40/1.25 oil objective. Images were collected at a resolution of 512 × 512 pixels in triplicate per slide. Images were processed and analyzed using LAS X and CellProfiler v4(58).

Data Availability

Raw and processed scRNAseq data are available at (*GEO accession number pending*) and code is available at (*github pending*).

Results

SARS-CoV-2 generates gRNA and sgmRNAs during infection, which are highly redundant in their sequences (**Figure 1A**). Reads mapping to redundant sequences are assigned to all genes which contain that sequence and are typically excluded from quantification steps in scRNAseq processing pipelines. We therefore identified read structures which could unambiguously identify gRNA or different species of sgmRNAs to allow for their quantification (**Figure 1B**). Reads derived from gRNA should be contiguous and could map anywhere on the SARS-CoV-2 genome. Reads derived from sgmRNA could be either gapped or contiguous and could map to the 5′ leader and/or downstream of the start site of S, the most 5′ sgmRNA. Because contiguous reads mapping downstream of S could derive either from gRNA or sgmRNAs, they were excluded from quantification. Only reads aligning in part to the 5′ leader and in part downstream of S could be confidently derived from sgmRNAs. We therefore defined gRNA reads as contiguous reads aligning upstream of regions contained in sgmRNAs. sgmRNA reads were defined as discontinuous reads spanning the leader region and regions used by sgmRNAs. Reads that did not match either of these formats could not unambiguously be assigned to gRNA or an sgmRNA and were therefore excluded from quantification (**Figure 1B**). With this framework, we developed scCoVseq to quantify unambiguous genomic and subgenomic viral reads (**Figure 1C**). Using scCoVseq, we compared the abilities of different scRNAseq methods to quantify SARS-CoV-2 RNAs.

In the widely available Chromium scRNAseq method developed by 10X Genomics, Inc, there are two formats for droplet-based scRNAseq: 10X 3′ and 10X 5′. 10X 3′ generates library fragments derived from the 3′ regions of poly-adenylated RNAs within a cell (**Figure 2A**). Because sgmRNAs share all viral sequence 3′ of the leader-body junction site, 10X 3′ library fragments derived from SARS-CoV-2 heavily cover the 3′ end of the viral genome and do not contain leader-ORF junctions (**Figure 2D**). These reads cannot differentiate gRNA from sgmRNA or distinguish different sgmRNA species. 10X 5′ generates library fragments from the 5′ termini of poly-adenylated RNAs (**Figure 2B**). These fragments are on average approximately 500 bp long (according to the manufacturer's documentation) and should contain leader-ORF junctions of SARS-CoV-2 sgmRNAs. The transcript read (R2), however, derives from the 3′ end of these fragments and at the recommended read length of 91 bases is not long enough to consistently sequence into the leader-sgmRNA junction site (**Figure 2B**). 10X 5′ can therefore detect some but not all junctions (**Figure 2E**). We reasoned that we could use 10X 5′ library fragments to detect junction-spanning reads by sequencing from the 5′ end of the fragment. To do this we extended R1, which is normally used to sequence the cell barcode and UMI, to sequence into the leader-body junction site (**Figure 2C**). Using 10X 5′ with Extended R1, we were able to sequence more leader-sgmRNA junction sites and increase our ability to unambiguously quantify sgmRNAs (**Figure 2F**). Indeed 10X 5′ Extended R1 increased the number of leader-sgmRNA spanning reads over 10X 5′ and 10X 3′ (**Figure 2G**). When quantified with scCoVseq, we found that 10X 5′ Extended R1 quantifies more UMIs per sgmRNA per cell compared to 10X 5′ or 10X 3′ (**Figure 2H**). Importantly, the average host gene expression per sample was significantly correlated across methods, suggesting that host gene measurements were minimally affected by 10X 5′ Extended R1 (**Supplemental Figure 1**). Taken together, 10X 5′ libraries sequenced with extended R1 sequencing results in a greater number of unambiguous reads derived from sgmRNAs over 10X 3′ or 10X 5′, and consequently recovers more sgmRNA UMIs/cell.

Using this method, we analyzed Vero E6 cells 24 hours post infection with SARS-CoV-2 at an MOI of 0.1 (**Figure 3A**). We were able to quantify sgmRNAs and gRNA at single-cell resolution (**Figure 3B**). Using expression values for sgmRNAs, we compared multiple unsupervised methods to identify infected cells. We found that a k-medoid clustering approach implemented with the pam algorithm performed best as indicated by multiple metrics to separate infected from uninfected cells (**Supplemental Figure 2A-C**). We found that this classification method detected a similar percentage of infected cells as detected using flow cytometry and immunofluorescence microscopy of the same cultures (**Supplemental Figure 2D**). Using our infection classification, we performed differential expression testing of infected cells compared to bystander cells within the same culture as well as to cells from a mock culture. As previously described(35), we observed downregulation of many host genes in infected cells accompanied by an upregulation of cellular stress response genes (**Figure 3D, E**). We further observed that, while bystander and mock cells had similar gene expression, a small number of genes were upregulated in bystander cells compared to mock cells (**Figure 3D, E**). This is especially notable given the inability of Vero E6 cells to produce interferons in response to viral infection (59). KEGG enrichment analysis of differentially expressed genes in pairwise comparisons of infected, mock, and bystander cells showed that genes related to COVID19 were enriched in our infected cells supporting our method for infection classification (**Figure 3F**).

Discussion

In this study, we examined the ability of two commonly used scRNAseq methods, 10X 3′ and 10X 5′, to detect and quantify SARS-CoV-2 derived RNAs with a focus on sgmRNAs. Because of the redundant nature of coronavirus sgmRNA sequences, we developed scCoVseq, which unambiguously quantifies both sgmRNAs and gRNAs in 10X data. We found that 10X methods detect unambiguous leader-sgmRNA junction-spanning reads to different degrees. We were able to increase the detection of leader-sgmRNA junction-spanning reads by extending the length of R1 during sequencing of 10X 5′ libraries, an approach we term 10X 5′ Extended R1 sequencing. Combining 10X 5′ Extended R1 with scCoVseq maximized quantification of sgmRNA UMIs compared to 10X 5′ or 10X 3′.

The ability to use sgmRNA expression to identify cells with actively replicating virus may improve the utility of scRNAseq in studies of coronavirus tropism. A challenge in many scRNAseq studies, particularly studies of primary tissues, has been identifying cells with active infection as opposed to cells with large amounts of ambient or extracellular viral RNA (such as phagocytic cells and/or cells not supporting active infection)(16). Because sgmRNAs are produced only during viral replication and are absent from virions, our method allows us to distinguish between infected and uninfected cells associated with "background" viral RNA (**Supplemental Figure 2C**). In downstream analyses of host transcriptomic changes induced by infection, accurate classification of infected cells is important for robust analyses of transcriptional differences between infected and uninfected cells because incorrect classifications may dilute effect sizes and resultant significance values. This method also enables the comparison of sgmRNA expression dynamics at single cell resolution. Such analyses may be particularly relevant for comparing viral gene expression between different cell types, coronaviruses, or between SARS-CoV-2 variants of interest, which have been described to have different kinetics of sgmRNA expression(60). This approach could be extended to any coronavirus or nidovirus, including potentially novel emerging coronaviruses. Finally, scCoVseq can be used to examine differential junction site usage within single cells (**Supplemental Figure 3**). Several groups have identified TRS-independent SARS-CoV-2 sgmRNAs(12, 13, 15), the significance of which remains unknown. It is possible that changes in junction site usage between cell types or during the course of infection may play a role in pathogenesis.

It should be noted that there are some limitations to our study. With our dataset, we are unable to know the true infection state of a cell processed for scRNAseq, and therefore we cannot assess the true accuracy of our method to classify infected cells. An additional limitation of our method is that quantification of viral genes with scCoVseq is dependent on accurate annotation of viral RNAs. We derived our annotation based on published empirically-defined TRS-dependent RNAs(12), but this does not preclude the existence of other viral RNAs at time points or in cell types not studied. Importantly, we explicitly exclude TRS-independent RNAs from our analyses. Methods such as STARsolo(57) or sequencing 10X libraries with long-read sequencing(61) may allow for detection and quantification of viral RNAs without reference annotation and irrespective of TRSs.

Acknowledgments

This work was supported in part by NIH grants R21 AI149180, R01 AI151029, and U01 AI150748. P.C. was supported by the Mount Sinai Medical Scientist Training Program T32 GM007280. P.C, R.S.P., and E.J.D. were supported by the Viral Host Pathogenesis Training Grant T32 AI07647.

We thank Randy Albrecht for BSL3 facility management and support. We also thank Thomas Moran, Center for Therapeutic Antibody Discovery at the Icahn School of Medicine at Mount Sinai, for kindly providing anti-SARS-CoV N antibody. We thank Michael A Schotsaert for flow cytometry support. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National

Institutes of Health. Microscopy was performed at the Microscopy CoRE at the Icahn School of Medicine at Mount Sinai.

Author Contributions

Conceptualization: P.C., B.R.R.

Data Curation: P.C., B.R.R.

Formal Analysis: P.C., B.R.R.

Funding Acquisition: B.R.R.

Investigation: P.C., E.J.D., O.D.

Methodology: P.C., B.R.R.

Project Administration: B.R.R.

Software: P.C., R.S.P., B.R.R.

Supervision: B.R.R.

Validation: P.C., R.S.P., E.J.D., O.D.

Visualization: P.C., B.R.R.

Writing – original draft: P.C., B.R.R.

Writing – review & editing: P.C., E.J.D., O.D., R.S.P., B.R.R.

References

1. Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20:533–534.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, Investigating CNC, Team R. 2020. A novel coronavirus from patients with pneumonia in china, 2019. N Engl J Med 382:727–733.

3. Carvalho T, Krammer F, Iwasaki A. 2021. The first 12 months of COVID-19: a timeline of immunological insights. 4. Nat Rev Immunol 21:245–256.

4. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, Tummino TA, Huettenhain R, Kaake RM, Richards AL, Tutuncuoglu B, Foussard H, Batra J, Haas K, Modak M, Kim M, Haas P, Polacco BJ, Braberg H, Fabius JM, Eckhardt M, Soucheray M, Bennett MJ, Cakir M, McGregor MJ, Li Q, Meyer B, Roesch F, Vallet T, Mac Kain A, Miorin L, Moreno E, Naing ZZC, Zhou Y, Peng S, Shi Y, Zhang Z, Shen W, Kirby IT, Melnyk JE, Chorba JS, Lou K, Dai SA, Barrio-Hernandez I, Memon D, Hernandez-Armenta C, Lyu J, Mathy CJP, Perica T, Pilla KB, Ganesan SJ, Saltzberg DJ, Rakesh R, Liu X, Rosenthal SB, Calviello L, Venkataramanan S, Liboy-Lugo J, Lin Y, Huang X-P, Liu Y, Wankowicz SA, Bohn M, Safari M, Ugur FS, Koh C, Savar NS, Tran QD, Shengjuler D, Fletcher SJ, O'Neal MC, Cai Y, Chang JCJ, Broadhurst DJ, Klippsten S, Sharp PP, Wenzell NA, Kuzuoglu D, Wang H-Y, Trenker R, Young

JM, Cavero DA, Hiatt J, Roth TL, Rathore U, Subramanian A, Noack J, Hubert M, Stroud RM, Frankel AD, Rosenberg OS, Verba KA, Agard DA, Ott M, Emerman M, Jura N, von Zastrow M, Verdin E, Ashworth A, Schwartz O, d'Enfert C, Mukherjee S, Jacobson M, Malik HS, Fujimori DG, Ideker T, Craik CS, Floor SN, Fraser JS, Gross JD, Sali A, Roth BL, Ruggero D, Taunton J, Kortemme T, Beltrao P, Vignuzzi M, García-Sastre A, Shokat KM, Shoichet BK, Krogan NJ. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature https://doi.org/10.1038/s41586-020-2286-9.

5. Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, Quan S, Zhang F, Sun R, Qian L, Ge W, Liu W, Liang S, Chen H, Zhang Y, Li J, Xu J, He Z, Chen B, Wang J, Yan H, Zheng Y, Wang D, Zhu J, Kong Z, Kang Z, Liang X, Ding X, Ruan G, Xiang N, Cai X, Gao H, Li L, Li S, Xiao Q, Lu T, Zhu Y, Liu H, Chen H, Guo T. 2020. Proteomic and metabolomic characterization of COVID-19 patient sera. Cell https://doi.org/10.1016/j.cell.2020.05.032.

6. Bojkova D, Klann K, Koch B, Widera M, Krause D, Ciesek S, Cinatl J, Münch C. 2020. SARS-CoV-2 infected host cell proteomics reveal potential therapy targets https://doi.org/10.21203/rs.3.rs-17218/v1.

7. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, Tamir H, Achdout H, Stein D, Israeli O, Beth-Din A, Melamed S, Weiss S, Israely T, Paran N, Schwartz M, Stern-Ginossar N. 2020. The coding capacity of SARS-CoV-2. Nature 1–9.

8. Bouhaddou M, Memon D, Meyer B, White KM, Rezelj VV, Marrero MC, Polacco BJ, Melnyk JE, Ulferts S, Kaake RM, Batra J, Richards AL, Stevenson E, Gordon DE, Rojc A, Obernier K, Fabius JM, Soucheray M, Miorin L, Moreno E, Koh C, Tran QD, Hardy A, Robinot R, Vallet T, Nilsson-Payant BE, Hernandez-Armenta C, Dunham A, Weigang S, Knerr J, Modak M, Quintero D, Zhou Y, Dugourd A, Valdeolivas A, Patil T, Li Q, Hüttenhain R, Cakir M, Muralidharan M, Kim M, Jang G, Tutuncuoglu B, Hiatt J, Guo JZ, Xu J, Bouhaddou S, Mathy CJP, Gaulton A, Manners EJ, Félix E, Shi Y, Goff M, Lim JK, McBride T, O'Neal MC, Cai Y, Chang JCJ, Broadhurst DJ, Klippsten S, Wit ED, Leach AR, Kortemme T, Shoichet B, Ott M, Saez-Rodriguez J, tenOever BR, Mullins D, Fischer ER, Kochs G, Grosse R, García-Sastre A, Vignuzzi M, Johnson JR, Shokat KM, Swaney DL, Beltrao P, Krogan NJ. 2020. The Global Phosphorylation Landscape of SARS-CoV-2 Infection. Cell 0.

9. Schmidt N, Lareau CA, Keshishian H, Ganskih S, Schneider C, Hennig T, Melanson R, Werner S, Wei Y, Zimmer M, Ade J, Kirschner L, Zielinski S, Dölken L, Lander ES, Caliskan N, Fischer U, Vogel J, Carr SA, Bodem J, Munschauer M. 2021. The SARS-CoV-2 RNA–protein interactome in infected human cells. 3. Nat Microbiol 6:339–353.

10. Flynn RA, Belk JA, Qi Y, Yasumoto Y, Wei J, Alfajaro MM, Shi Q, Mumbach MR, Limaye A, DeWeirdt PC, Schmitz CO, Parker KR, Woo E, Chang HY, Horvath TL, Carette JE, Bertozzi CR, Wilen CB, Satpathy AT. 2021. Discovery and functional

interrogation of SARS-CoV-2 RNA-host protein interactions. Cell 184:2394-2411.e16.

11. Blanco - Melo D. 2020. Imbalanced host response to SARS - CoV - 2 drives development of COVID - 19. Cell.

12. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. Cell 181:914-921.e10.

13. Wang D, Jiang A, Feng J, Li G, Guo D, Sajid M, Wu K, Zhang Q, Ponty Y, Will S, Liu F, Yu X, Li S, Liu Q, Yang X-L, Guo M, Li X, Chen M, Shi Z-L, Lan K, Chen Y, Zhou Y. 2021. The SARS-CoV-2 Subgenome Landscape and its Novel Regulatory Features. Mol Cell 0.

14. Ziv O, Price J, Shalamova L, Kamenova T, Goodfellow I, Weber F, Miska EA. 2020. The short- and long-range RNA-RNA Interactome of SARS-CoV-2. Mol Cell https://doi.org/10.1016/j.molcel.2020.11.004.

15. Chang JJ-Y, Rawlinson D, Pitt ME, Taiaroa G, Gleeson J, Zhou C, Mordant FL, Paoli-Iseppi RD, Caly L, Purcell DFJ, Stinear TP, Londrigan SL, Clark MB, Williamson DA, Subbarao K, Coin LJM. 2021. Transcriptional and epi-transcriptional dynamics of SARS-CoV-2 during cellular infection. Cell Rep 0.

16. 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. 2020.

Host-viral infection maps reveal signatures of severe COVID-19 patients. Cell https://doi.org/10.1016/j.cell.2020.05.006.

17. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, Talavera-López C, Maatz H, Reichart D, Sampaziotis F, Worlock KB, Yoshida M, Barnes JL, Network HLB. 2020. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med 26:681–687.

18. Lukassen S, Chua RL, Trefzer T, Kahn NC, Schneider MA, Muley T, Winter H, Meister M, Veith C, Boots AW, Hennig BP, Kreuter M, Conrad C, Eils R. 2020. SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. EMBO J 39:e105114.

19. Fodoulian L, Tuberosa J, Rossier D, Landis B, Carleton A, Rodriguez I. 2020. SARS-CoV-2 receptor and entry genes are expressed by sustentacular cells in the human olfactory neuroepithelium. BioRxiv https://doi.org/10.1101/2020.03.31.013268.

20. Qi F, Qian S, Zhang S, Zhang Z. 2020. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. Biochem Biophys Res Commun 526:135–140.

21. Ravindra NG, Alfajaro MM, Gasque V, Wei J, Filler RB, Huston NC, Wan H, Szigeti-Buck K, Wang B, Montgomery RR, Eisenbarth SC, Williams A, Pyle AM, Iwasaki A, Horvath TL, Foxman EF, van Dijk D, Wilen CB. 2020. Single-cell

longitudinal analysis of SARS-CoV-2 infection in human bronchial epithelial cells. BioRxiv https://doi.org/10.1101/2020.05.06.081695.

22. Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics I Nature Medicine.

23. Bieberich F, Vazquez-Lombardi R, Yermanos A, Ehling RA, Mason DM, Wagner B, Kapetanovic E, Roberto RBD, Weber CR, Savic M, Rudolf F, Reddy ST. 2021. A single-cell atlas of lymphocyte adaptive immune repertoires and transcriptomes reveals age-related differences in convalescent COVID-19 patients. bioRxiv 2021.02.12.430907.

24. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martínez-Colón GJ, McKechnie JL, Ivison GT, Ranganath T, Vergara R, Hollis T, Simpson LJ, Grant P, Subramanian A, Rogers AJ, Blish CA. 2020. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. 7. Nat Med 26:1070–1076.

25. Yao C, Bora SA, Parimon T, Zaman T, Friedman OA, Palatinus JA, Surapaneni NS, Matusov YP, Chiang GC, Kassar AG, Patel N, Green CER, Aziz AW, Suri H, Suda J, Lopez AA, Martins GA, Stripp BR, Gharib SA, Goodridge HS, Chen P. 2021. Cell-Type-Specific Immune Dysregulation in Severely Ill COVID-19 Patients. Cell Rep 34.

26. MacDonald L, Otto TD, Elmesmari A, Tolusso B, Somma D, McSharry C, Gremese E, McInnes IB, Alivernini S, Kurowska-Stolarska M. 2020. COVID-19 and

Rheumatoid Arthritis share myeloid pathogenic and resolving pathways. bioRxiv 2020.07.26.221572.

27. Schreibing F, Hannani M, Ticconi F, Fewings E, Nagai JS, Begemann M, Kuppe C, Kurth I, Kranz J, Frank D, Anslinger TM, Ziegler P, Kraus T, Enczmann J, Balz V, Windhofer F, Balfanz P, Kurts C, Marx G, Marx N, Dreher M, Schneider RK, Saez-Rodriguez J, Filho IGC, Kramann R. 2021. Dissecting CD8+ T cell pathology of severe SARS-CoV-2 infection by single-cell epitope mapping. bioRxiv 2021.03.03.432690.

28. Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, Liu X, Xie L, Li J, Ye J, Dong L, Cui X, Miao Y, Wang D, Dong J, Xiao C, Chen W, Wang H. 2020. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discov 6:31.

29. Lee JS, Park S, Jeong HW, Ahn JY, Choi SJ, Lee H, Choi B, Nam SK, Sa M, Kwon J-S, Jeong SJ, Lee HK, Park SH, Park S-H, Choi JY, Kim S-H, Jung I, Shin E-C. 2020. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. Sci Immunol 5.

30. Wang F-S, Zhang J-Y, Wang X, Xing X, Xu Z, Zhang C, Song J-W, Fan X, Xia P, Fu J-L, Wang S-Y, Xu R-N, Dai X-P, Shi L, Huang L, Jiang T-J, Shi M, Zhang Y, Zumla A, Maeurer M, Bai F. 2020. Single-cell landscape of immunological responses in COVID-19 patients. bioRxiv 2020.07.23.217703.

31. Zhang J-Y, Wang X-M, Xing X, Xu Z, Zhang C, Song J-W, Fan X, Xia P, Fu J-L, Wang S-Y, Xu R-N, Dai X-P, Shi L, Huang L, Jiang T-J, Shi M, Zhang Y, Zumla A, Maeurer M, Bai F, Wang F-S. 2020. Single-cell landscape of immunological responses in patients with COVID-19. 9. Nat Immunol 21:1107–1118.

32. Kalfaoglu B, Almeida-Santos J, Tye CA, Satou Y, Ono M. 2020. T-cell hyperactivation and paralysis in severe COVID-19 infection revealed by single-cell analysis. BioRxiv https://doi.org/10.1101/2020.05.26.115923.

33. Wei L, Ming S, Zou B, Wu Y, Hong Z, Li Z, Zheng X, Huang M, Luo L, Liang J, Wen X, Chen T, Liang Q, Kuang L, Shan H, Huang X. 2020. Viral invasion and type I interferon response characterize the immunophenotypes during COVID-19 infection. Electron J https://doi.org/10.2139/ssrn.3555695.

34. Wyler E, Mösbauer K, Franke V, Diag A, Gottula LT, Arsie R, Klironomos F, Koppstein D, Ayoub S, Buccitelli C, Richter A, Legnini I, Ivanov A, Mari T, Del Giudice S, Papies JP, Müller MA, Niemeyer D, Selbach M, Akalin A, Rajewsky N, Drosten C, Landthaler M. 2020. Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention. BioRxiv https://doi.org/10.1101/2020.05.05.079194.

35. Miorin L, Kehrer T, Sanchez-Aparicio MT, Zhang K, Cohen P, Patel RS, Cupic A, Makio T, Mei M, Moreno E, Danziger O, White KM, Rathnasinghe R, Uccellini M, Gao S, Aydillo T, Mena I, Yin X, Martin-Sancho L, Krogan NJ, Chanda SK, Schotsaert M, Wozniak RW, Ren Y, Rosenberg BR, Fontoura BMA, García-Sastre

A. 2020. SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. Proc Natl Acad Sci 117:28344–28354.

36. Melms JC, Biermann J, Huang H, Wang Y, Nair A, Tagore S, Katsyv I, Rendeiro AF, Amin AD, Schapiro D, Frangieh CJ, Luoma AM, Filliol A, Fang Y, Ravichandran H, Clausi MG, Alba GA, Rogava M, Chen SW, Ho P, Montoro DT, Kornberg AE, Han AS, Bakhoum MF, Anandasabapathy N, Suárez-Fariñas M, Bakhoum SF, Bram Y, Borczuk A, Guo XV, Lefkowitch JH, Marboe C, Lagana SM, Del Portillo A, Zorn E, Markowitz GS, Schwabe RF, Schwartz RE, Elemento O, Saqi A, Hibshoosh H, Que J, Izar B. 2021. A molecular single-cell lung atlas of lethal COVID-19. Nature 1–6.

37. Delorey TM, Ziegler CGK, Heimberg G, Normand R, Yang Y, Segerstolpe Å, Abbondanza D, Fleming SJ, Subramanian A, Montoro DT, Jagadeesh KA, Dey KK, Sen P, Slyper M, Pita-Juárez YH, Phillips D, Biermann J, Bloom-Ackermann Z, Barkas N, Ganna A, Gomez J, Melms JC, Katsyv I, Normandin E, Naderi P, Popov YV, Raju SS, Niezen S, Tsai LT-Y, Siddle KJ, Sud M, Tran VM, Vellarikkal SK, Wang Y, Amir-Zilberstein L, Atri DS, Beechem J, Brook OR, Chen J, Divakar P, Dorceus P, Engreitz JM, Essene A, Fitzgerald DM, Fropf R, Gazal S, Gould J, Grzyb J, Harvey T, Hecht J, Hether T, Jané-Valbuena J, Leney-Greene M, Ma H, McCabe C, McLoughlin DE, Miller EM, Muus C, Niemi M, Padera R, Pan L, Pant D, Pe’er C, Pfiffner-Borges J, Pinto CJ, Plaisted J, Reeves J, Ross M, Rudy M, Rueckert EH, Siciliano M, Sturm A, Todres E, Waghray A, Warren S, Zhang S, Zollinger DR, Cosimi L, Gupta RM, Hacohen N, Hibshoosh H, Hide W, Price AL,

Rajagopal J, Tata PR, Riedel S, Szabo G, Tickle TL, Ellinor PT, Hung D, Sabeti PC, Novak R, Rogers R, Ingber DE, Jiang ZG, Juric D, Babadi M, Farhi SL, Izar B, Stone JR, Vlachos IS, Solomon IH, Ashenberg O, Porter CBM, Li B, Shalek AK, Villani A-C, Rozenblatt-Rosen O, Regev A. 2021. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature 1–8.

38. Mechanisms of SARS-CoV-2 Transmission and Pathogenesis: Trends in Immunology.

39. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. 2020. Coronavirus biology and replication: implications for SARS-CoV-2. 3. Nat Rev Microbiol 1–16.

40. Sola I, Almazán F, Zúñiga S, Enjuanes L. 2015. Continuous and discontinuous RNA synthesis in coronaviruses. Annu Rev Virol 2:265–288.

41. Perlman S, Masters PS. 2020. Coronaviridae: The Viruses and Their Replication, p. 411–442. *In* Howley, PM, Knipe, DM, Whelan, S (eds.), Fields Virology: Emerging Viruses, 7th ed. Wolters Kluwer Health/lippincott Williams & Wilkins, Philadelphia, PA.

42. Daniloski Z, Jordan TX, Wessels H-H, Hoagland DA, Kasela S, Legut M, Maniatis S, Mimitou EP, Lu L, Geller E, Danziger O, Rosenberg BR, Phatnani H, Smibert P, Lappalainen T, tenOever BR, Sanjana NE. 2021. Identification of Required Host Factors for SARS-CoV-2 Infection in Human Cells. Cell 184:92-105.e16.

43. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214.

44. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36:411–420.

45. Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20:296.

46. Li H. 2021. lh3/seqtk. C.

47. Fernandes JD, Hinrichs AS, Clawson H, Navarro Gonzalez J, Lee BT, Nassar LR, Raney BJ, Rosenbloom KR, Nerli S, Rao A, Schmelter D, Zweig AS, Lowe TM, Ares M, Corbet-Detig R, Kent WJ, Haussler D, Haeussler M. 2020. The UCSC SARS-CoV-2 genome browser. BioRxiv https://doi.org/10.1101/2020.05.04.075945.

48. Twelve years of SAMtools and BCFtools | GigaScience | Oxford Academic.

49. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.

50. 2021. Sparse and Dense Matrix Classes and Methods [R package Matrix version 1.3-4]. Comprehensive R Archive Network (CRAN).

51. 2020. Read Rectangular Text Data [R package readr version 1.4.0]. Comprehensive R Archive Network (CRAN).

52. Garrido-Martín D, Palumbo E, Guigó R, Breschi A. 2018. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLOS Comput Biol 14:e1006360.

53. 2008. clValid: An R Package for Cluster Validation by Guy Brock, Vasyl Pihur, Susmita Datta, Somnath Datta.

54. Maechler M, original) PR (Fortran, original) AS (S, original) MH (S, Hornik [trl K, maintenance(1999-2000)] ctb] (port to R, Studer M, Roudier P, Gonzalez J, Kozlowski K, pam()) ES (fastpam options for, Murphy (volume.ellipsoid({d >= 3})) K. 2021. cluster: "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.

55. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

56. Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. 4. Nat Methods 15:255–261.

57. Kaminow B, Yunusov D, Dobin A. 2021. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv 2021.05.05.442755.

58. McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, Doan M, Ding L, Rafelski SM, Thirstrup D, Wiegraebe W, Singh S, Becker T, Caicedo JC, Carpenter AE. 2018. CellProfiler 3.0: Next-generation image processing for biology. PLOS Biol 16:e2005970.

59. Emeny JM, Morgan MJY 1979. Regulation of the Interferon System: Evidence that Vero Cells have a Genetic Defect in Interferon Production. J Gen Virol 43:247–252.

60. Parker MD, Lindsey BB, Shah DR, Hsu S, Keeley AJ, Partridge DG, Leary S, Cope A, State A, Johnson K, Ali N, Raghei R, Heffer J, Smith N, Zhang P, Gallis M, Louka SF, Whiteley M, Foulkes BH, Christou S, Wolverson P, Pohare M, Hansford SE, Green LR, Evans C, Raza M, Wang D, Gaudieri S, Mallal S, Consortium TC-19 GU (COG-U, Silva TI de. 2021. Altered Sub-Genomic RNA Expression in SARS-CoV-2 B.1.1.7 Infections. bioRxiv 2021.03.02.433156.

61. Russell AB, Elshina E, Kowalsky JR, Te Velthuis AJW, Bloom JD. 2019. Single-cell virus sequencing of influenza infections that trigger innate immunity. J Virol 93.

Figure Legends

Figure 1: A. Illustration of SARS-CoV-2 genomic RNA, gRNA, and subgenomic RNAs, sgmRNAs. **B.** *Top:* Reads included for analysis by scCoVseq. Either: 1) contiguous reads mapping to ORF1a/b and therefore derived from gRNA or 2) discontinuous reads spanning the leader region and ORFS transcribed by sgmRNAs *Bottom*: Reads excluded from analysis by scCoVseq. Either: 1) discontinuous reads that do not include sequence mapping to the leader region and downstream of S or 2) contiguous reads that map to ORFs other than ORF1a/b, which are ambiguous. **C.** Activity diagram of scCoVseq pipeline. Blue rectangles indicate inputs/outputs for each stage. Orange rounded rectangles indicate a process in bold with software indicated.

Figure 2: A-C. Illustration of gRNA and S and ORF3a sgmRNAs. Red box indicates regions contained in final 10X library. *Lower:* Example illustration of 10X library fragments derived from gRNA and S and ORF3a sgmRNAs with sequencing read 1 and read 2 indicated. 10X 3′ (**A**), 10X 5′ (**B**), and 10X 5′ Extended R1 (**C**) libraries are illustrated. **D-F.** Sashimi plot of 10X 3′ (**D**), 10X 5′ (**E**), and 10X 5′ Extended R1 (**F**) reads mapped to the SARS-CoV-2 genome filtered to show only junctions supported by at least 1,000 reads. Total number of reads visualized is indicated in the bottom right. **G.** Reads per million reads mapped to SARS-CoV-2 reads mapping to a single viral gene in 10X 3′, 10X 5′, or 10X 5′ with Extended R1 data. Reads are colored by their mapping with contiguous reads mapping to ORF1a/b in yellow, leader-sgmRNA junction-spanning reads in blue, and ambiguously mapped reads in grey. **H.** UMIs per cell for all sgmRNAs in infected cells in each dataset. Each dataset was downsampled to an equal

number of infected cells and each cells' total UMIs were downsampled to the same value to control for differences in sequencing depth. The leader region is enlarged in illustrations of the genome for visibility. L = Leader.

Figure 3: A. Experimental design. Vero E6 cells were infected or mock infected with SARS-CoV-2 (USA-WA1/2020) at an MOI of 0.1. At 24 hours post-infection, cells were analyzed by scRNAseq using 10X 5′ with Extended R1 sequencing. **B-C.** 3,047 mock and infected cells embedded in tSNE space derived from euclidean distance of scaled viral sgmRNA expression. Cells are colored by (**B**) indicated viral RNA expression, or (**C**) experimental condition and assigned infection status of cells. **D.** Heatmap of genes differentially expressed in infected, bystander, or mock cells. Differential expression testing was performed on host gene expression downsampled to an equal number of UMIs/cell across cells to account for infection-induced transcriptional shutdown. Genes were selected for visualization based on false discovery rate of less than 0.05 and absolute $\log_2$ fold change of at least 1. Non-downsampled gene expression data is shown. Along the top, infection status, total viral UMIs and genomic RNA as quantified by CellRanger and scCoVseq are indicated. Cells and genes are clustered with ward d2 clustering based on euclidean distance. **E.** Expression of selected host genes per cell by infection status. Data shown is not downsampled. *Top:* genes induced in infected cells. *Middle:* genes repressed in infected cells. *Bottom:* genes upregulated in bystander cells compared to mock. **F.** KEGG pathway enrichment in genes differentially expressed in pairwise comparisons of downsampled infected, bystander, and mock cells. Dot size

and fill indicates the -$\log_{10}$ p value of enrichment with red dots indicating enrichment in the first infection state and blue in the second infection state noted above each panel.

Supplemental Figure 1: Average counts of host gene expression of cells analyzed by 10X 3′, 10X 5′, and 10X 5′ with Extended R1 sequencing. Each point represents the average UMIs/cell of a single gene assayed in the indicated assays. At the top left, the Pearson correlation coefficient and resulting p value are indicated.

Supplemental Figure 2: A. Comparison of performance metrics (average distance, AD; average distance between means, ADM; average proportion of non-overlap, APN; connectivity; Dunn Index; figure of merit, FOM; and silhouette index) by several clustering methods (diana, model-based, hierarchical, kmeans, and pam) run on sgmRNA expression of 600 randomly sampled cells analyzed with 10X 5′ Extended R1 and scCoVseq. *Left*: Performance metrics for each method across k values from 2 to 5. *Right:* performance metrics for each method with k = 2. **B.** Visualization of infection classification by different methods. **C.** Viral gene expression of cells by infection status, determined by pam clustering method. **D.** Percent of infected cells per sample as measured by flow cytometry, immunofluorescence, and infection classification with unsupervised (pam) method or supervised infection classification by classifying infected cells as those with at least 375 total viral UMIs. Because the same sample was sequenced with 10X 5′ and 10X 5′ Extended R1, flow cytometry and immunofluorescence results are duplicated for ease of visualization. Error bars for immunofluorescence indicate mean ± one standard deviation of percent infected cells based on three fields per sample.

Supplemental Figure 3: Detection of junction sites in SARS-CoV-2 reads with 10X 5′ Extended R1. Junction sites are represented by the 5′ start site and 3′ end site on the y and x-axis, respectively. The color indicates the $\log_2$ total UMIs/junction across all cells in the SARS-CoV-2 infected sample. Below each axis, the number of UMIs supporting a position as a junction start or end site is indicated with a density plot.

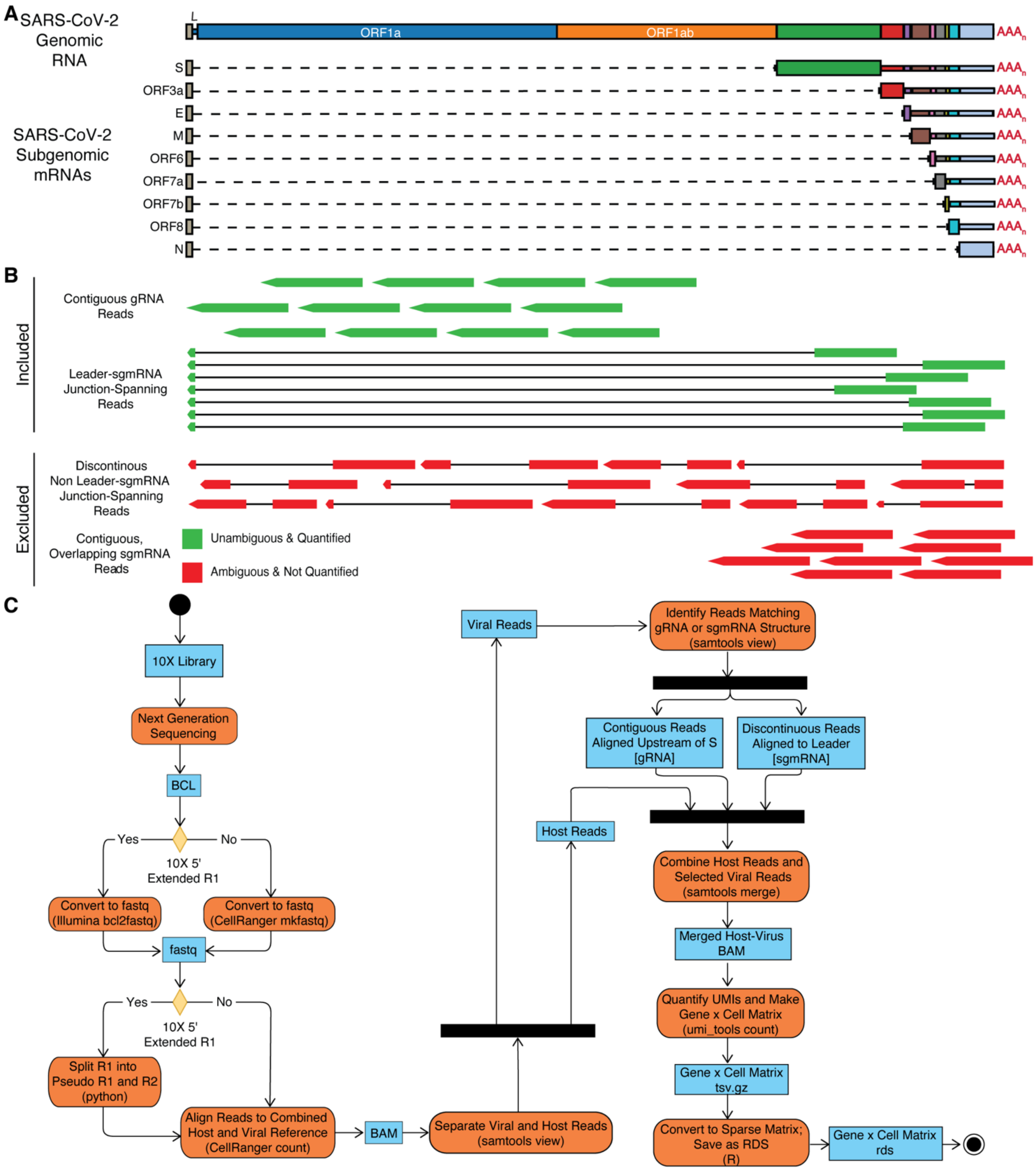

A
SARS-CoV-2 Genomic RNA
L
ORF1a
ORF1ab
AAAn
SARS-CoV-2 Subgenomic mRNAs
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
B
Included
Contiguous gRNA Reads
Leader-sgmRNA Junction-Spanning Reads
Excluded
Discontinous Non Leader-sgmRNA Junction-Spanning Reads
Contiguous, Overlapping sgmRNA Reads
🟩 Unambiguous & Quantified
🟥 Ambiguous & Not Quantified
C
10X Library
Next Generation Sequencing
BCL
Yes
No
10X 5' Extended R1
Convert to fastq (Illumina bcl2fastq)
Convert to fastq (CellRanger mkfastq)
fastq
Yes
No
10X 5' Extended R1
Split R1 into Pseudo R1 and R2 (python)
Align Reads to Combined Host and Viral Reference (CellRanger count)
BAM
Separate Viral and Host Reads (samtools view)
Viral Reads
Host Reads
Identify Reads Matching gRNA or sgmRNA Structure (samtools view)
Contiguous Reads Aligned Upstream of S [gRNA]
Discontinuous Reads Aligned to Leader [sgmRNA]
Combine Host Reads and Selected Viral Reads (samtools merge)
Merged Host-Virus BAM
Quantify UMIs and Make Gene x Cell Matrix (umi_tools count)
Gene x Cell Matrix tsv.gz
Convert to Sparse Matrix; Save as RDS (R)
Gene x Cell Matrix rds

Figure 1: A. Illustration of SARS-CoV-2 genomic RNA, gRNA, and subgenomic RNAs, sgmRNAs. **B.** *Top:* Reads included for analysis by scCoVseq. Either: 1) contiguous reads mapping to ORF1a/b and therefore derived from gRNA or 2) discontinuous reads spanning the leader region and ORFS transcribed by sgmRNAs *Bottom*: Reads

excluded from analysis by scCoVseq. Either: 1) discontinuous reads that do not include sequence mapping to the leader region and downstream of S or 2) contiguous reads that map to ORFs other than ORF1a/b, which are ambiguous. **C.** Activity diagram of scCoVseq pipeline. Blue rectangles indicate inputs/outputs for each stage. Orange rounded rectangles indicate a process in bold with software indicated.

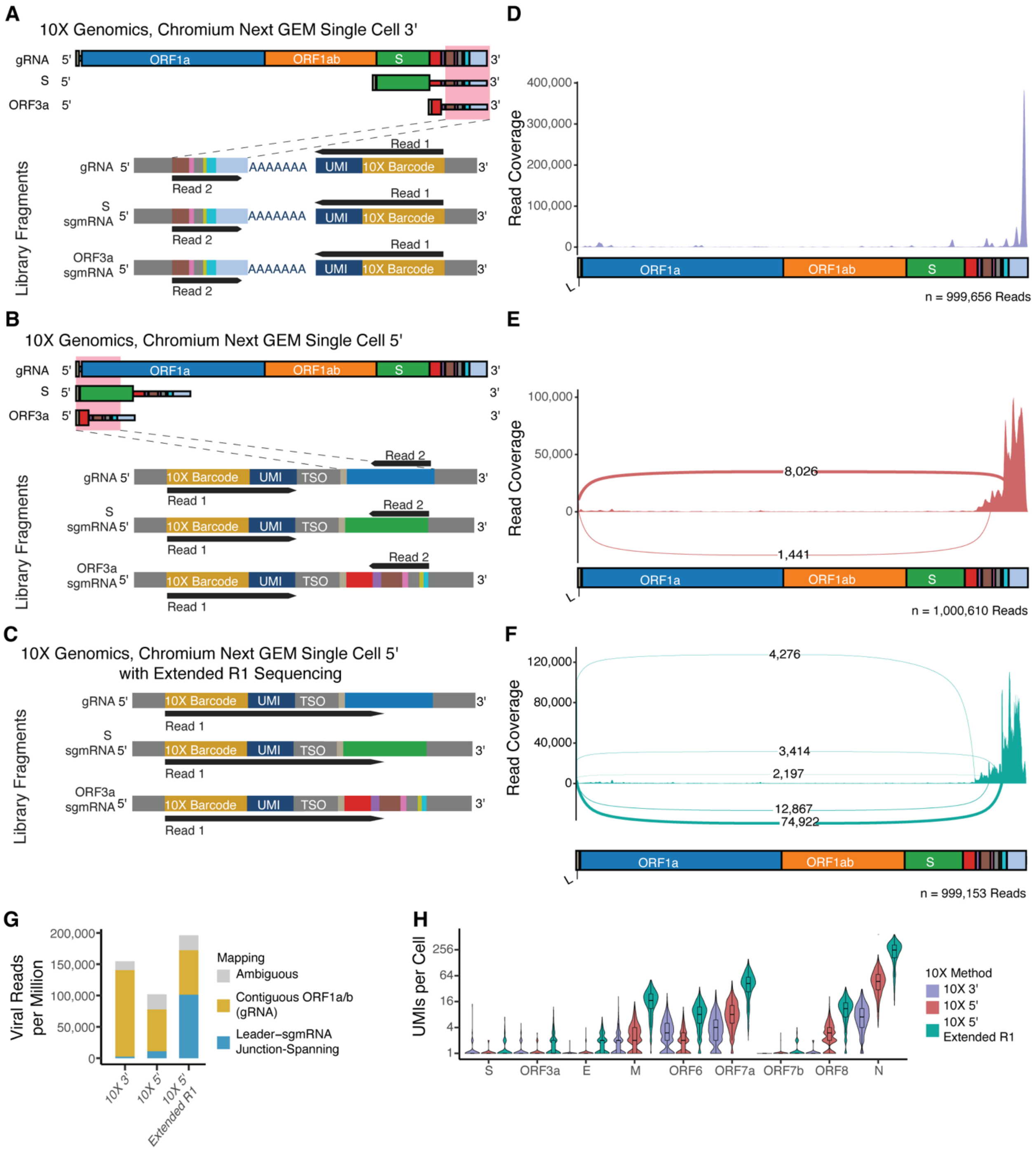
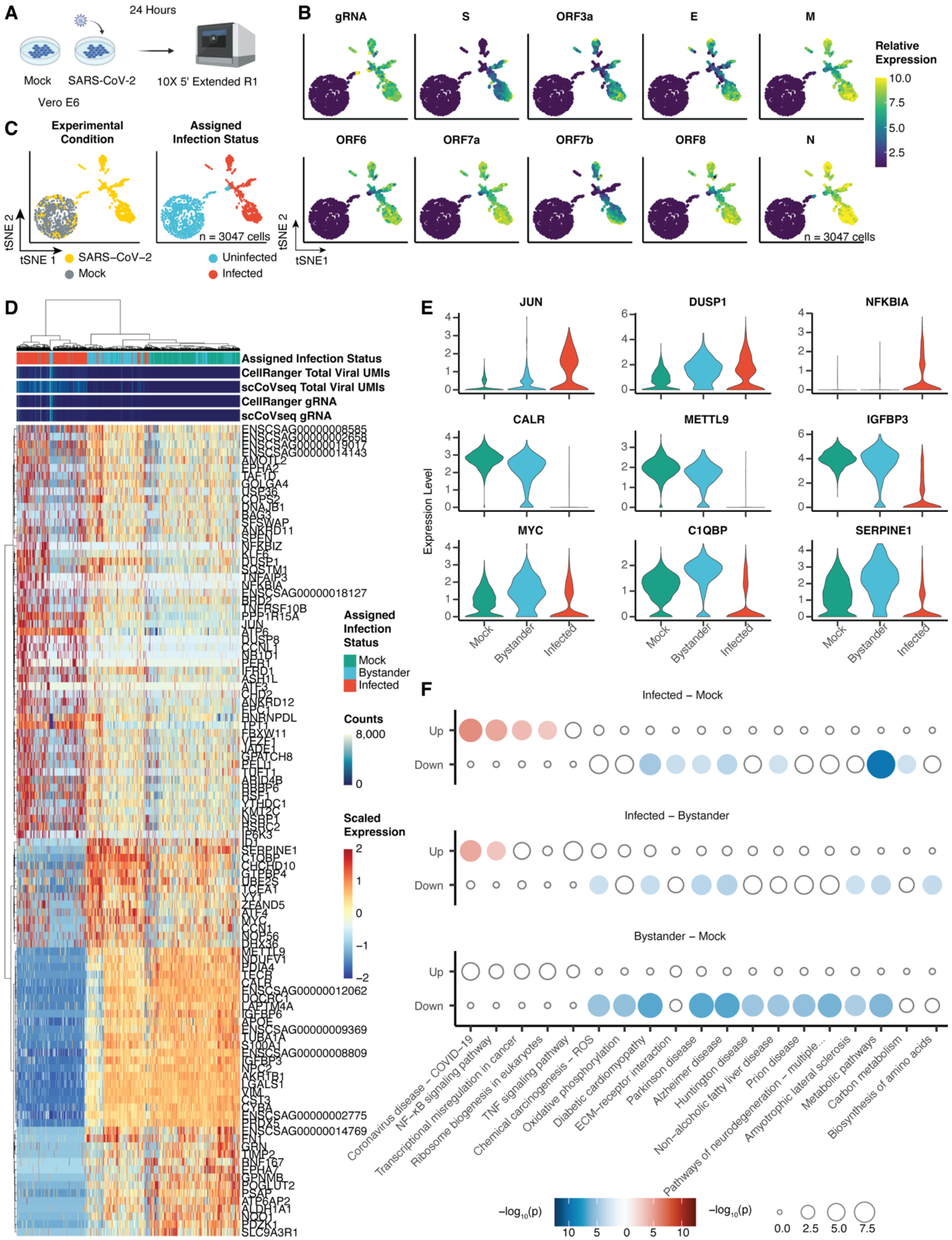


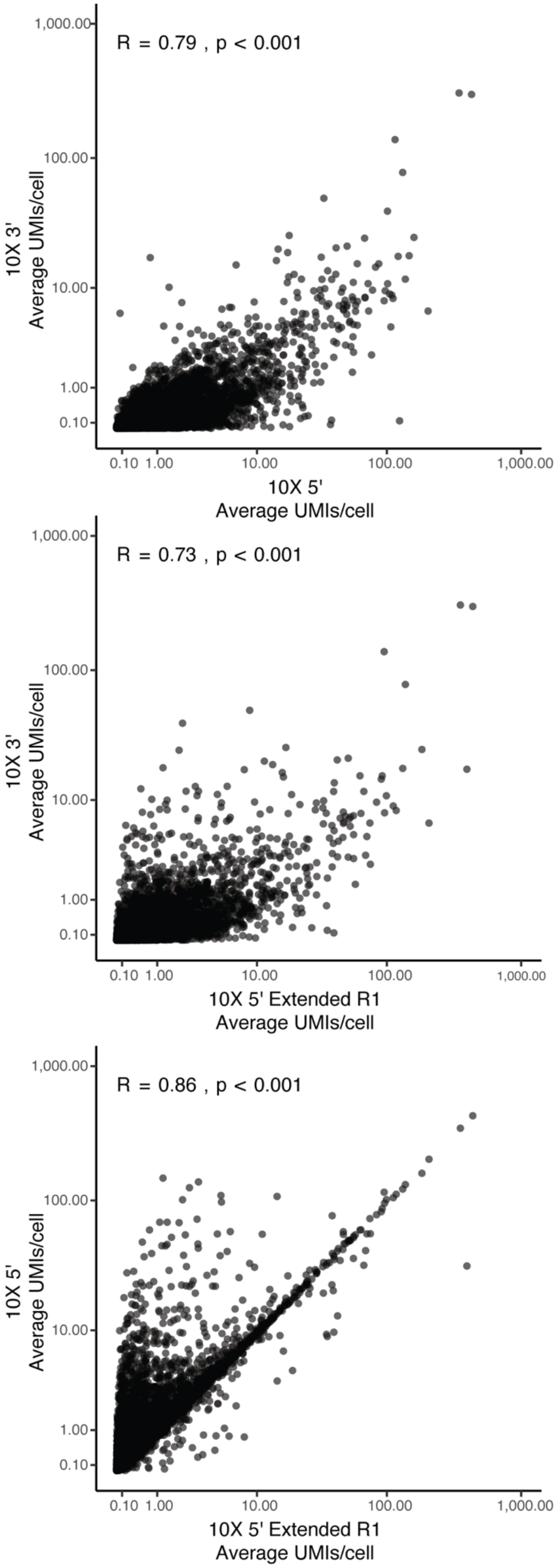
Figure 2: A-C. Illustration of gRNA and S and ORF3a sgmRNAs. Red box indicates regions contained in final 10X library. *Lower:* Illustration of 10X library fragments derived from gRNA and S and ORF3a sgmRNAs with sequencing read 1 and read 2 indicated. 10X 3′ (**A**), 10X 5′ (**B**), and 10X 5′ Extended R1 (**C**) libraries are illustrated. **D-F.**

Sashimi plot of 10X 3′ (**D**), 10X 5′ (**E**), and 10X 5′ Extended R1 (**F**) reads mapped to the SARS-CoV-2 genome filtered to show only junctions supported by at least 1,000 reads. Total number of reads visualized is indicated in the bottom right. **G.** Reads per million reads mapped to SARS-CoV-2 reads mapping to a single viral gene in 10X 3′, 10X 5′, or 10X 5′ with Extended R1 data. Reads are colored by their mapping with contiguous reads mapping to ORF1a/b in yellow, leader-sgmRNA junction-spanning reads in blue, and ambiguously mapped reads in grey. **H.** UMIs per cell for all sgmRNAs in infected cells in each dataset. Each dataset was downsampled to an equal number of infected cells and each cells' total UMIs were downsampled to the same value to control for differences in sequencing depth. The leader region is enlarged in illustrations of the genome for visibility. L = Leader.



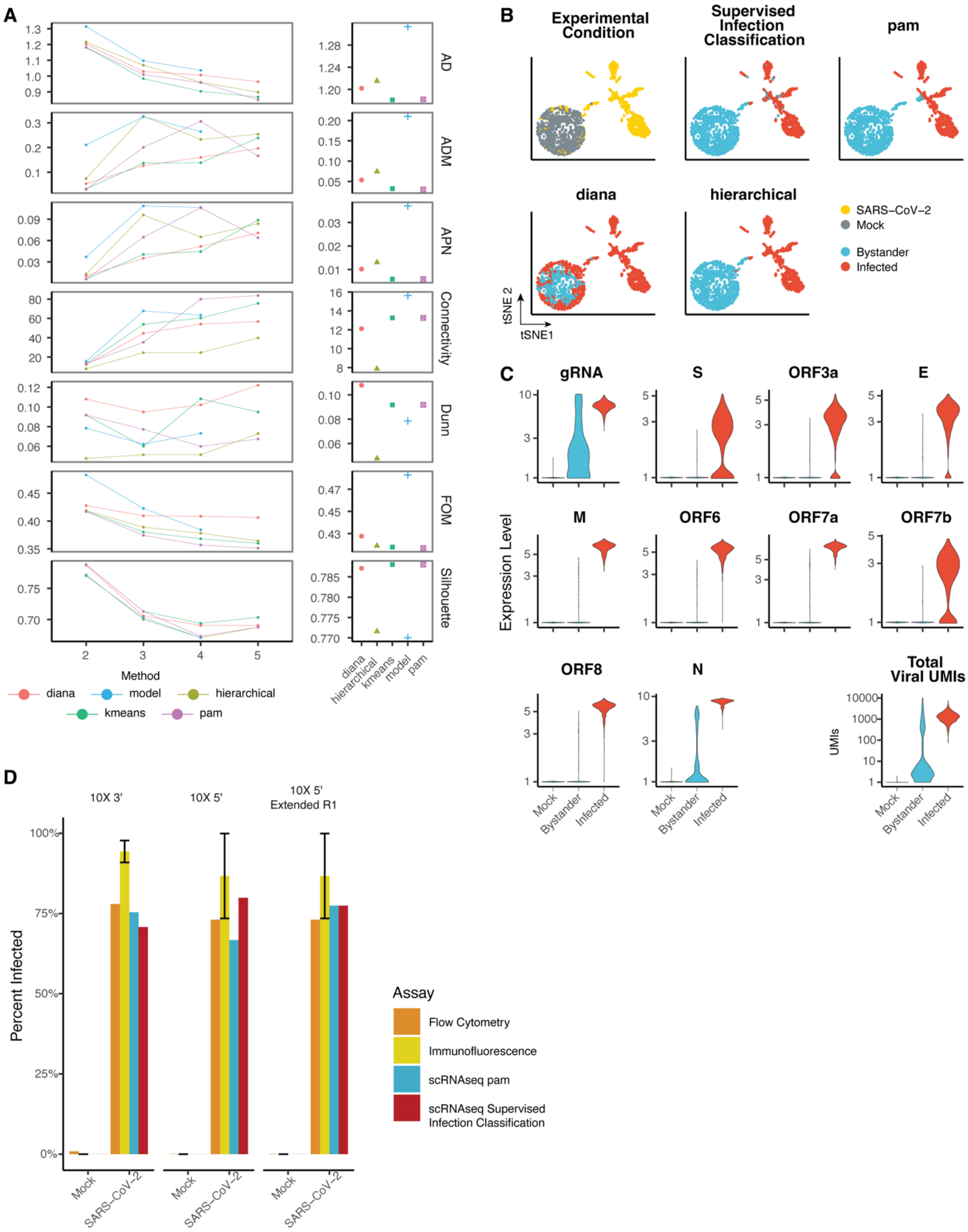
791 **Figure 3: A.** Experimental design. Vero E6 cells were infected or mock infected with

SARS-CoV-2 (USA-WA1/2020) at an MOI of 0.1. At 24 hours post-infection, cells were analyzed by scRNAseq using 10X 5′ with Extended R1 sequencing. **B-C.** 3,047 mock and infected cells embedded in tSNE space derived from euclidean distance of scaled viral sgmRNA expression. Cells are colored by (**B**) indicated viral RNA expression, or (**C**) experimental condition and assigned infection status of cells. **D.** Heatmap of genes differentially expressed in infected, bystander, or mock cells. Differential expression testing was performed on host gene expression downsampled to an equal number of UMIs/cell across cells to account for infection-induced transcriptional shutdown. Genes were selected for visualization based on false discovery rate of less than 0.05 and absolute $\log_2$ fold change of at least 1. Non-downsampled gene expression data is shown. Along the top, infection status, total viral UMIs and genomic RNA as quantified by CellRanger and scCoVseq are indicated. Cells and genes are clustered with ward d2 clustering based on euclidean distance. **E.** Expression of selected host genes per cell by infection status. Data shown is not downsampled. *Top:* genes induced in infected cells. *Middle:* genes repressed in infected cells. *Bottom:* genes upregulated in bystander cells compared to mock. **F.** KEGG pathway enrichment in genes differentially expressed in pairwise comparisons of downsampled infected, bystander, and mock cells. Dot size and fill indicates the $-\log_{10}$ p value of enrichment with red dots indicating enrichment in the first infection state and blue in the second infection state noted above each panel.



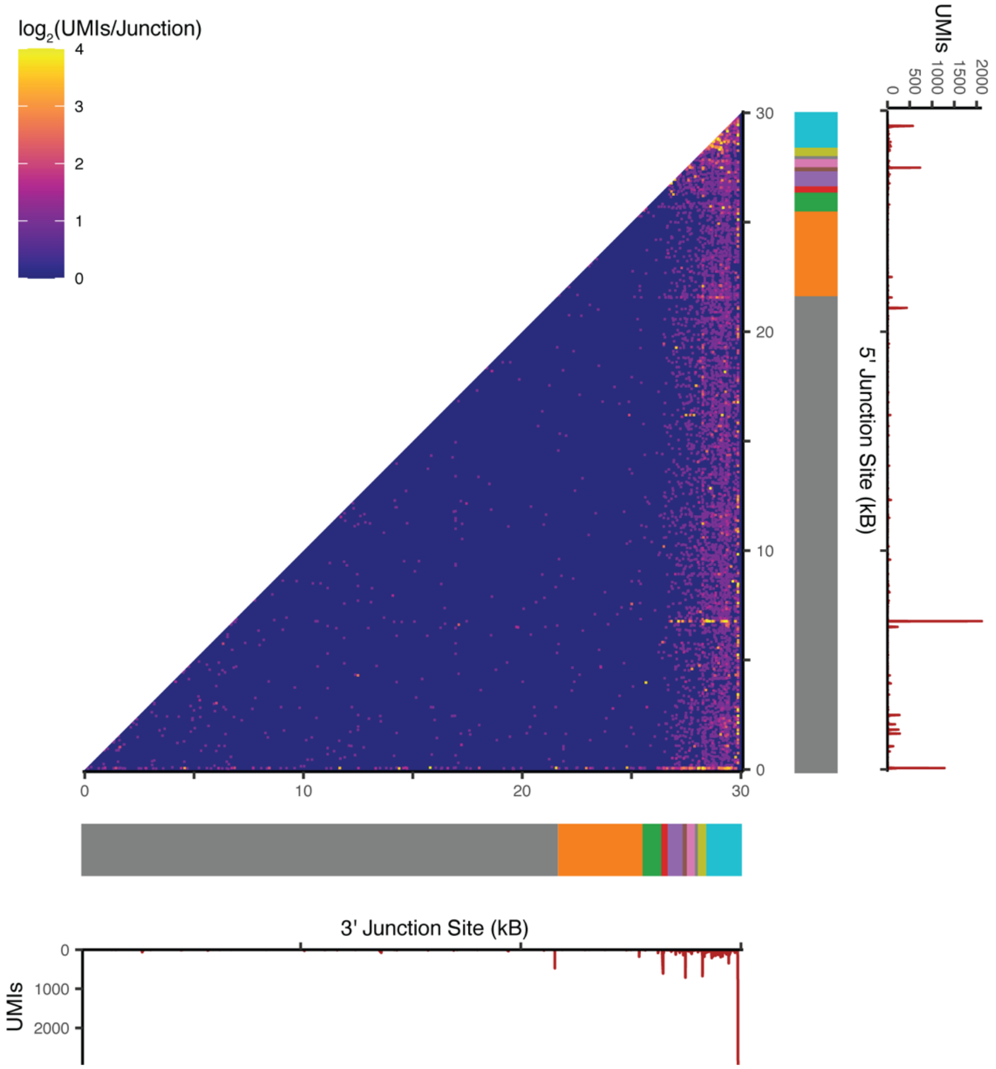
813 **Supplemental Figure 1:** Average counts of host gene expression of cells analyzed by

10X 3′, 10X 5′, and 10X 5′ with Extended R1 sequencing. Each point represents the average expression of a single gene assayed in the indicated assays. At the top left, the Pearson correlation coefficient and resulting p value are indicated.



Supplemental Figure 2: A. Comparison of performance metrics (average distance, AD; average distance between means, ADM; average proportion of non-overlap, APN;

connectivity; Dunn Index; figure of merit, FOM; and silhouette index) by several clustering methods (diana, model-based, hierarchical, kmeans, and pam) run on sgmRNA expression of 600 randomly sampled cells analyzed with 10X 5′ Extended R1 and scCoVseq. *Left*: Performance metrics for each method across k values from 2 to 5. *Right:* performance metrics for each method with k = 2. **B.** Visualization of infection classification by different methods. **C.** Viral gene expression of cells by infection status, determined by pam clustering method. **D.** Percent of infected cells per sample as measured by flow cytometry, immunofluorescence, and infection classification with unsupervised (pam) method or supervised infection classification by classifying infected cells as those with at least 375 total viral UMIs. Because the same sample was sequenced with 10X 5′ and 10X 5′ Extended R1, flow cytometry and immunofluorescence results are duplicated for ease of visualization. Error bars for immunofluorescence indicate mean ± one standard deviation of percent infected cells based on three fields per sample.



835 **Supplemental Figure 3:** Detection of junction sites in SARS-CoV-2 reads with 10X 5′

836 Extended R1. Junction sites are represented by the 5′ start site and 3′ end site on the y

837 and x-axis, respectively. Below each axis, the number of UMIs supporting a position as

838 a junction start or end site is indicated with a density plot.

839