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.

45 **Introduction**

46 Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) is the
47 causative agent of COronaVIrus Disease-2019 (COVID-19), which as of November
48 2021 has caused over 250 million cases and over 5 million deaths globally(1, 2). Global
49 efforts to understand the pathogenesis of SARS-CoV-2 infection have led to the
50 development of vaccines and antivirals, which have reduced morbidity and mortality(3).
51 "Omics" methods have been instrumental in studying SARS-CoV-2 in part because they
52 have generated large amounts of data regarding host-viral interactions at
53 unprecedented speed(4–15). Single-cell RNA sequencing (scRNAseq) studies in
54 particular have been used to study viral tropism(16–22), peripheral immune
55 changes(23–33), transcriptional changes induced by infection(34, 35), and to develop
56 cell atlases of COVID-19 pathology(23, 24, 36, 37). Of note, most scRNAseq workflows
57 have been developed and optimized for studies of eukaryotic transcription but not viral,
58 specifically SARS-CoV-2, transcription. The performance of different scRNAseq
59 methods to detect and quantify viral RNAs may impact the analysis and interpretation of
60 such studies.

61 SARS-CoV-2 is a betacoronavirus with a 29 kB positive-sense, single stranded
62 RNA genome(38, 39). SARS-CoV-2 generates genomic RNA (gRNA), subgenomic
63 mRNAs (sgmRNAs), and negative-sense antigenomic RNA during active infection(40,
64 41). Both gRNA and sgmRNAs are poly-adenylated, which enables detection by
65 scRNAseq protocols that rely on poly-T primed reverse transcription(39–41). Translation
66 of gRNA results in the production of one of two polyproteins, pp1a and pp1ab, which are

67 subsequently cleaved into an array of non-structural proteins involved in pathogenesis
68 and replication(39, 41). Translation of sgmRNAs generates structural and accessory
69 viral proteins critical for virion production and pathogenesis(39, 41). sgmRNAs are
70 produced only in cells with actively replicating virus, while gRNA is present in both
71 infected cells and virions(40, 41). Therefore, specific detection of sgmRNAs can allow
72 for: 1) specific identification of cells with actively replicating virus and 2) analysis of the
73 dynamics of viral gene expression within and across cells and viruses.

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

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

100 **Materials and Methods**

101 *Cell lines and Viral Infection*

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

106 SARS-CoV-2 (isolate USA-WA1/2020, BEI resource NR-52281) and control media
107 (mock infected) stocks were grown by inoculating a confluent T175 flask of Vero-E6 cells
108 (passage 2). Mock and SARS-CoV-2 infected cultures were maintained in reduced serum
109 DMEM (2% FBS) for 72 hours, after which culture media was collected and filtered by

110 centrifugation (8000 x g, 15 minutes) using an Amicon Ultra-15 filter unit with a 100KDa
111 cutoff filter (Millipore # UFC910024). Concentrated stocks in reduced-serum media (2%
112 FBS), supplemented with 50mM HEPES buffer (Gibco #15630080) were stored at -80°C.
113 Viral titers were determined by plaque assay as previously described(42). All SARS-CoV-
114 2 propagations and experiments were performed in a Biosafety Level 3 facility in
115 compliance with institutional protocols and federal guidelines.

116 *scRNAseq*

117 For scRNAseq experiments, Vero-E6 cells in 6 well plates were infected with
118 SARS-CoV-2 at a MOI of 0.1, or with an equivalent volume of control media, in reduced-
119 serum media (2% FBS) for 24 hours. To prepare cells for scRNAseq, mock and SARS-
120 CoV-2 infected cultures were washed with calcium/magnesium-free PBS and
121 disassociated using TrypLE (Gibco # 12605010, 5 minutes at 37° C), after which
122 samples were centrifuged (200 x g, 5 minutes), resuspended in calcium/magnesium-
123 free PBS supplemented with 1% BSA, and counted. Mock and SARS-CoV-2 infected
124 cell culture samples were filtered through a 40μm FlowMi strainer (ScienceWare #
125 H13680-0040) and counted prior to loading on the 10X Genomics Chromium Controller
126 according to manufacturer's protocol. Mock and infected samples were loaded on
127 separate lanes of a 10X Genomics Chromium Controller for either NextGEM Single Cell
128 3′ v3.1 (10X 3′), or NextGEM Single Cell V(D)J v1.1, (10X 5′).

129 Gene expression libraries were prepared for 10X 3′ and 5′ samples according to
130 manufacturer's guidelines. Final 10X 3′ mock and infected gene expression libraries
131 and the 10X 5′ infected gene expression library were PCR amplified for 16 cycles while

132 the 10X 5′ mock gene expression library was amplified for 14 cycles. 10X 3′ gene
133 expression libraries were pooled and sequenced by short-read sequencing on an
134 Illumina NextSeq 500 using a high output 150 cycle reaction kit according to
135 manufacturers' protocol with the following read lengths: read 1 28 nt; i7 index 8 nt; and
136 read 2 130 nt. 10X 5′ gene expression libraries were also pooled and sequenced with
137 10X recommended read lengths (read 1 26 nt; i7 index 8 nt; and read 2 132 nt) or with
138 extended R1 protocol (read 1 158 nt; i7 index 8 nt; no read 2).

139 *scRNAseq Pre-Processing*

140 *Conversion of Illumina BCL files to fastq*

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

152 *Downsampling fastqs to control for sequencing depth*

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

167 *Preparation of empirically derived SARS-CoV-2 genome reference*

168 Downsampled fastq files were then mapped using cellranger count/3.1.0 (10X
169 Genomics, Inc) to an empirically defined reference of SARS-CoV-2 sgmRNAs derived
170 from previously reported SARS-CoV-2 (BetaCoV/Korea/KCDC03/2020) RNAs
171 sequenced with long-read direct RNA Nanopore sequencing(12). These were
172 downloaded from the UCSC Genome Browser Table Browser(47) after filtering for TRS-
173 dependent transcripts and score > 900 and exporting to gtf format. Transcripts for
174 previously unknown ORFs were excluded from the annotation. An additional annotation

175 for genomic RNA was included which covered the entire length of the SARS-CoV-2
176 genome. Aligning the BetaCoV/Korea/KCDC03/2020 genome with USA/WA-CDC-
177 WA1/2020 genome showed that the USA/WA-CDC-WA1/2020 had an additional 21 3′
178 adenosine nucleotides annotated. To account for this in our reference, we extended any
179 annotations from BetaCoV/Korea/KCDC03/2020 that ended at the 3′ end of the genome
180 by an additional 21 bases. This SARS-CoV-2 reference was appended to the host
181 ChlSab1.1 Ensembl reference.

182 *scCoVseq*

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

196 *scRNAseq Analysis*

197 *Sashimi Plots*

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

203 *Classification of SARS-CoV-2 Infected Cells*

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

217 compared for viral UMI expression per cell, and the cluster with more viral UMIs was
218 classified as infected and the other as uninfected.

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

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

227 *SARS-CoV-2 Read Distribution by scRNAseq Method*

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

232 *Differential Expression Analysis*

233 To explore expression differences between infected, bystander, and mock cells,
234 differential expression testing with edgeR (version 3.32.1) was performed with
235 modifications for scRNAseq as previously described(55, 56). Viral genes were excluded
236 from analysis, and only host genes expressed in at least 10% of cells were tested. To
237 account for differences in RNA content of infected cells due to virally-induced
238 transcriptional shutdown, all cells were downsampled to the 25th percentile of total UMIs

239 of infected cells. Cells with fewer UMIs than the threshold were excluded from analysis.
240 Differential gene expression was performed with edgeR using a generalized linear
241 model quasi-likelihood F test adapted with a term for gene detection rate(55, 56). Genes
242 with an absolute $\log_2$ fold change greater than or equal to 1 and false discovery rate
243 less than 0.05 were considered significant. For KEGG enrichment analysis, pairwise
244 tests between mock, bystander, and infected cells were performed. Differentially
245 expressed genes with an absolute $\log_2$ fold change greater than or equal to 1 and false
246 discovery rate less than 0.05 were considered significant and subject to KEGG
247 enrichment analysis using the KEGG annotations for African Green Monkey as
248 implemented in the edgeR function kegga.

249 *Quantification of SARS-CoV-2 sgmRNA Junction Sites*

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

254 *Flow Cytometry*

255 Vero E6 cells were fixed with 4% paraformaldehyde at room temperature for a
256 minimum of 24 hours, washed once with PBS and permeabilized with 1X perm-wash
257 buffer (BDBiosciences #554723) for 5 minutes. SARS-CoV nucleocapsid (N) antibody
258 (clone 1C7C7) (kindly provided by Thomas Moran, Icahn School of Medicine at Mount
259 Sinai, New York, NY), conjugated to AlexaFluor 647 was diluted 1:400 in perm-wash
260 buffer, and added directly to samples. Samples were then incubated at room

261 temperature for 40 minutes in the dark. After staining, samples were washed once with
262 1X perm-wash buffer, once with PBS, resuspended in FACS buffer (PBS supplemented
263 with 1% FBS), and acquired on a Gallios flow cytometer (Beckman-Coulter). For all viral
264 infections, analysis was performed with FlowJo software (version 10.7.1, Becton
265 Dickinson), excluding cell doublets and debris and gating according to mock infected
266 populations.

267 *Immunofluorescence microscopy*

268 Vero E6 were seeded in 6-well plates (Falcon REF-353046) with one coverslip (Fisher
269 Scientific 12-550-143) per well. After 24 hours post infection, cells were washed with
270 PBS and fixed with 4% paraformaldehyde (Fisher Scientific AA433689M) overnight.
271 Fixed cells were permeabilized using 0.1% Triton-X (Fisher Scientific AC327371000) in
272 PBS and blocked with 4% bovine serum albumin (BSA, Fisher Scientific BP1600-100) in
273 PBS. Blocked coverslips were incubated with mouse anti-SARS-CoV N antibody (clone
274 1C7, 1:500 in 4% BSA PBS) overnight at 4C, washed three times with PBS, and
275 incubated for 45 minutes with 1:500 AlexaFluor 488-conjugated anti-mouse (Invitrogen
276 A11001, 1:500 in 4% BSA PBS) plus DAPI (Thermo Fisher Scientific D1306, 1:1000 in
277 4% BSA PBS) at room temperature. Coverslips were then stained with phalloidin (1:400
278 in PBS) for 1 hour at room temperature and washed again three times with PBS.
279 Coverslips were mounted using Prolong Diamond (Life Technologies P36970). Confocal
280 laser scanning was performed using a Leica SP5 DMI (ISMMS Microscopy CoRE and
281 Advanced Bioimaging Center) with a ×40/1.25 oil objective. Images were collected at a

282 resolution of 512 × 512 pixels in triplicate per slide. Images were processed and
283 analyzed using LAS X and CellProfiler v4(58).

284 *Data Availability*

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

287 **Results**

288 SARS-CoV-2 generates gRNA and sgmRNAs during infection, which are highly
289 redundant in their sequences (**Figure 1A**). Reads mapping to redundant sequences are
290 assigned to all genes which contain that sequence and are typically excluded from
291 quantification steps in scRNAseq processing pipelines. We therefore identified read
292 structures which could unambiguously identify gRNA or different species of sgmRNAs
293 to allow for their quantification (**Figure 1B**). Reads derived from gRNA should be
294 contiguous and could map anywhere on the SARS-CoV-2 genome. Reads derived from
295 sgmRNA could be either gapped or contiguous and could map to the 5′ leader and/or
296 downstream of the start site of S, the most 5′ sgmRNA. Because contiguous reads
297 mapping downstream of S could derive either from gRNA or sgmRNAs, they were
298 excluded from quantification. Only reads aligning in part to the 5′ leader and in part
299 downstream of S could be confidently derived from sgmRNAs. We therefore defined
300 gRNA reads as contiguous reads aligning upstream of regions contained in sgmRNAs.
301 sgmRNA reads were defined as discontinuous reads spanning the leader region and
302 regions used by sgmRNAs. Reads that did not match either of these formats could not
303 unambiguously be assigned to gRNA or an sgmRNA and were therefore excluded from

304 quantification (**Figure 1B**). With this framework, we developed scCoVseq to quantify
305 unambiguous genomic and subgenomic viral reads (**Figure 1C**). Using scCoVseq, we
306 compared the abilities of different scRNAseq methods to quantify SARS-CoV-2 RNAs.
307 In the widely available Chromium scRNAseq method developed by 10X
308 Genomics, Inc, there are two formats for droplet-based scRNAseq: 10X 3′ and 10X 5′.
309 10X 3′ generates library fragments derived from the 3′ regions of poly-adenylated RNAs
310 within a cell (**Figure 2A**). Because sgmRNAs share all viral sequence 3′ of the leader-
311 body junction site, 10X 3′ library fragments derived from SARS-CoV-2 heavily cover the
312 3′ end of the viral genome and do not contain leader-ORF junctions (**Figure 2D**). These
313 reads cannot differentiate gRNA from sgmRNA or distinguish different sgmRNA
314 species. 10X 5′ generates library fragments from the 5′ termini of poly-adenylated
315 RNAs (**Figure 2B**). These fragments are on average approximately 500 bp long
316 (according to the manufacturer's documentation) and should contain leader-ORF
317 junctions of SARS-CoV-2 sgmRNAs. The transcript read (R2), however, derives from
318 the 3′ end of these fragments and at the recommended read length of 91 bases is not
319 long enough to consistently sequence into the leader-sgmRNA junction site (**Figure**
320 **2B**). 10X 5′ can therefore detect some but not all junctions (**Figure 2E**). We reasoned
321 that we could use 10X 5′ library fragments to detect junction-spanning reads by
322 sequencing from the 5′ end of the fragment. To do this we extended R1, which is
323 normally used to sequence the cell barcode and UMI, to sequence into the leader-body
324 junction site (**Figure 2C**). Using 10X 5′ with Extended R1, we were able to sequence
325 more leader-sgmRNA junction sites and increase our ability to unambiguously quantify

326 sgmRNAs (**Figure 2F**). Indeed 10X 5′ Extended R1 increased the number of leader-
327 sgmRNA spanning reads over 10X 5′ and 10X 3′ (**Figure 2G**). When quantified with
328 scCoVseq, we found that 10X 5′ Extended R1 quantifies more UMIs per sgmRNA per
329 cell compared to 10X 5′ or 10X 3′ (**Figure 2H**). Importantly, the average host gene
330 expression per sample was significantly correlated across methods, suggesting that
331 host gene measurements were minimally affected by 10X 5′ Extended R1
332 (**Supplemental Figure 1**). Taken together, 10X 5′ libraries sequenced with extended
333 R1 sequencing results in a greater number of unambiguous reads derived from
334 sgmRNAs over 10X 3′ or 10X 5′, and consequently recovers more sgmRNA UMIs/cell.
335 Using this method, we analyzed Vero E6 cells 24 hours post infection with
336 SARS-CoV-2 at an MOI of 0.1 (**Figure 3A**). We were able to quantify sgmRNAs and
337 gRNA at single-cell resolution (**Figure 3B**). Using expression values for sgmRNAs, we
338 compared multiple unsupervised methods to identify infected cells. We found that a k-
339 medoid clustering approach implemented with the pam algorithm performed best as
340 indicated by multiple metrics to separate infected from uninfected cells (**Supplemental**
341 **Figure 2A-C**). We found that this classification method detected a similar percentage of
342 infected cells as detected using flow cytometry and immunofluorescence microscopy of
343 the same cultures (**Supplemental Figure 2D**). Using our infection classification, we
344 performed differential expression testing of infected cells compared to bystander cells
345 within the same culture as well as to cells from a mock culture. As previously
346 described(35), we observed downregulation of many host genes in infected cells
347 accompanied by an upregulation of cellular stress response genes (**Figure 3D, E**). We

348 further observed that, while bystander and mock cells had similar gene expression, a
349 small number of genes were upregulated in bystander cells compared to mock cells
350 (**Figure 3D, E**). This is especially notable given the inability of Vero E6 cells to produce
351 interferons in response to viral infection (59). KEGG enrichment analysis of differentially
352 expressed genes in pairwise comparisons of infected, mock, and bystander cells
353 showed that genes related to COVID19 were enriched in our infected cells supporting
354 our method for infection classification (**Figure 3F**).

355 **Discussion**

356 In this study, we examined the ability of two commonly used scRNAseq methods,
357 10X 3′ and 10X 5′, to detect and quantify SARS-CoV-2 derived RNAs with a focus on
358 sgmRNAs. Because of the redundant nature of coronavirus sgmRNA sequences, we
359 developed scCoVseq, which unambiguously quantifies both sgmRNAs and gRNAs in
360 10X data. We found that 10X methods detect unambiguous leader-sgmRNA junction-
361 spanning reads to different degrees. We were able to increase the detection of leader-
362 sgmRNA junction-spanning reads by extending the length of R1 during sequencing of
363 10X 5′ libraries, an approach we term 10X 5′ Extended R1 sequencing. Combining 10X
364 5′ Extended R1 with scCoVseq maximized quantification of sgmRNA UMIs compared to
365 10X 5′ or 10X 3′.

366 The ability to use sgmRNA expression to identify cells with actively replicating
367 virus may improve the utility of scRNAseq in studies of coronavirus tropism. A challenge
368 in many scRNAseq studies, particularly studies of primary tissues, has been identifying
369 cells with active infection as opposed to cells with large amounts of ambient or

370 extracellular viral RNA (such as phagocytic cells and/or cells not supporting active
371 infection)(16). Because sgmRNAs are produced only during viral replication and are
372 absent from virions, our method allows us to distinguish between infected and
373 uninfected cells associated with “background” viral RNA (**Supplemental Figure 2C**). In
374 downstream analyses of host transcriptomic changes induced by infection, accurate
375 classification of infected cells is important for robust analyses of transcriptional
376 differences between infected and uninfected cells because incorrect classifications may
377 dilute effect sizes and resultant significance values. This method also enables the
378 comparison of sgmRNA expression dynamics at single cell resolution. Such analyses
379 may be particularly relevant for comparing viral gene expression between different cell
380 types, coronaviruses, or between SARS-CoV-2 variants of interest, which have been
381 described to have different kinetics of sgmRNA expression(60). This approach could be
382 extended to any coronavirus or nidovirus, including potentially novel emerging
383 coronaviruses. Finally, scCoVseq can be used to examine differential junction site
384 usage within single cells (**Supplemental Figure 3**). Several groups have identified
385 TRS-independent SARS-CoV-2 sgmRNAs(12, 13, 15), the significance of which
386 remains unknown. It is possible that changes in junction site usage between cell types
387 or during the course of infection may play a role in pathogenesis.
388 It should be noted that there are some limitations to our study. With our dataset,
389 we are unable to know the true infection state of a cell processed for scRNAseq, and
390 therefore we cannot assess the true accuracy of our method to classify infected cells.
391 An additional limitation of our method is that quantification of viral genes with scCoVseq

392 is dependent on accurate annotation of viral RNAs. We derived our annotation based on
393 published empirically-defined TRS-dependent RNAs(12), but this does not preclude the
394 existence of other viral RNAs at time points or in cell types not studied. Importantly, we
395 explicitly exclude TRS-independent RNAs from our analyses. Methods such as
396 STARsolo(57) or sequencing 10X libraries with long-read sequencing(61) may allow for
397 detection and quantification of viral RNAs without reference annotation and irrespective
398 of TRSs.

399 **Acknowledgments**

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

404 We thank Randy Albrecht for BSL3 facility management and support. We also
405 thank Thomas Moran, Center for Therapeutic Antibody Discovery at the Icahn School of
406 Medicine at Mount Sinai, for kindly providing anti-SARS-CoV N antibody. We thank
407 Michael A Schotsaert for flow cytometry support. This work was supported in part
408 through the computational resources and staff expertise provided by Scientific
409 Computing at the Icahn School of Medicine at Mount Sinai. Research reported in this
410 paper was supported by the Office of Research Infrastructure of the National Institutes
411 of Health under award number S10OD026880. The content is solely the responsibility of
412 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.

430 **References**

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

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

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

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

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

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

465 6. Bojkova D, Klann K, Koch B, Widera M, Krause D, Ciesek S, Cinatl J, Münch C.
466 2020. SARS-CoV-2 infected host cell proteomics reveal potential therapy targets
467 https://doi.org/10.21203/rs.3.rs-17218/v1.

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

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

485 9. Schmidt N, Lareau CA, Keshishian H, Ganskih S, Schneider C, Hennig T, Melanson
486 R, Werner S, Wei Y, Zimmer M, Ade J, Kirschner L, Zielinski S, Dölken L, Lander
487 ES, Caliskan N, Fischer U, Vogel J, Carr SA, Bodem J, Munschauer M. 2021. The
488 SARS-CoV-2 RNA–protein interactome in infected human cells. 3. Nat Microbiol
489 6:339–353.

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

493 interrogation of SARS-CoV-2 RNA-host protein interactions. Cell 184:2394-
494 2411.e16.

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

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

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

503 14. Ziv O, Price J, Shalamova L, Kamenova T, Goodfellow I, Weber F, Miska EA.
504 2020. The short- and long-range RNA-RNA Interactome of SARS-CoV-2. Mol Cell
505 https://doi.org/10.1016/j.molcel.2020.11.004.

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

510 16. Bost P, Giladi A, Liu Y, Bendjelal Y, Xu G, David E, Blecher-Gonen R, Cohen M,
511 Medaglia C, Li H, Deczkowska A, Zhang S, Schwikowski B, Zhang Z, Amit I. 2020.

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

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

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

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

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

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

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

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

536 23. Bieberich F, Vazquez-Lombardi R, Yermanos A, Ehling RA, Mason DM, Wagner B,
537 Kapetanovic E, Roberto RBD, Weber CR, Savic M, Rudolf F, Reddy ST. 2021. A
538 single-cell atlas of lymphocyte adaptive immune repertoires and transcriptomes
539 reveals age-related differences in convalescent COVID-19 patients. bioRxiv
540 2021.02.12.430907.

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

545 25. Yao C, Bora SA, Parimon T, Zaman T, Friedman OA, Palatinus JA, Surapaneni
546 NS, Matusov YP, Chiang GC, Kassar AG, Patel N, Green CER, Aziz AW, Suri H,
547 Suda J, Lopez AA, Martins GA, Stripp BR, Gharib SA, Goodridge HS, Chen P.
548 2021. Cell-Type-Specific Immune Dysregulation in Severely Ill COVID-19 Patients.
549 Cell Rep 34.

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

552 Rheumatoid Arthritis share myeloid pathogenic and resolving pathways. bioRxiv
553 2020.07.26.221572.

554 27. Schreibing F, Hannani M, Ticconi F, Fewings E, Nagai JS, Begemann M, Kuppe C,
555 Kurth I, Kranz J, Frank D, Anslinger TM, Ziegler P, Kraus T, Enczmann J, Balz V,
556 Windhofer F, Balfanz P, Kurts C, Marx G, Marx N, Dreher M, Schneider RK, Saez-
557 Rodriguez J, Filho IGC, Kramann R. 2021. Dissecting CD8+ T cell pathology of
558 severe SARS-CoV-2 infection by single-cell epitope mapping. bioRxiv
559 2021.03.03.432690.

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

564 29. Lee JS, Park S, Jeong HW, Ahn JY, Choi SJ, Lee H, Choi B, Nam SK, Sa M, Kwon
565 J-S, Jeong SJ, Lee HK, Park SH, Park S-H, Choi JY, Kim S-H, Jung I, Shin E-C.
566 2020. Immunophenotyping of COVID-19 and influenza highlights the role of type I
567 interferons in development of severe COVID-19. Sci Immunol 5.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

641 45. Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-
642 cell RNA-seq data using regularized negative binomial regression. Genome Biol
643 20:296.

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

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

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

651 49. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve
652 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.

672 57. Kaminow B, Yunusov D, Dobin A. 2021. STARsolo: accurate, fast and versatile
673 mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv
674 2021.05.05.442755.

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

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

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

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

689

690 **Figure Legends**

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

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

712 number of infected cells and each cells' total UMIs were downsampled to the same
713 value to control for differences in sequencing depth. The leader region is enlarged in
714 illustrations of the genome for visibility. L = Leader.
715
716 **Figure 3: A.** Experimental design. Vero E6 cells were infected or mock infected with
717 SARS-CoV-2 (USA-WA1/2020) at an MOI of 0.1. At 24 hours post-infection, cells were
718 analyzed by scRNAseq using 10X 5′ with Extended R1 sequencing. **B-C.** 3,047 mock
719 and infected cells embedded in tSNE space derived from euclidean distance of scaled
720 viral sgmRNA expression. Cells are colored by (**B**) indicated viral RNA expression, or
721 (**C**) experimental condition and assigned infection status of cells. **D.** Heatmap of genes
722 differentially expressed in infected, bystander, or mock cells. Differential expression
723 testing was performed on host gene expression downsampled to an equal number of
724 UMIs/cell across cells to account for infection-induced transcriptional shutdown. Genes
725 were selected for visualization based on false discovery rate of less than 0.05 and
726 absolute $\log_2$ fold change of at least 1. Non-downsampled gene expression data is
727 shown. Along the top, infection status, total viral UMIs and genomic RNA as quantified
728 by CellRanger and scCoVseq are indicated. Cells and genes are clustered with ward d2
729 clustering based on euclidean distance. **E.** Expression of selected host genes per cell
730 by infection status. Data shown is not downsampled. *Top:* genes induced in infected
731 cells. *Middle:* genes repressed in infected cells. *Bottom:* genes upregulated in bystander
732 cells compared to mock. **F.** KEGG pathway enrichment in genes differentially expressed
733 in pairwise comparisons of downsampled infected, bystander, and mock cells. Dot size

734 and fill indicates the -$\log_{10}$ p value of enrichment with red dots indicating enrichment in
735 the first infection state and blue in the second infection state noted above each panel.
736
737 **Supplemental Figure 1:** Average counts of host gene expression of cells analyzed by
738 10X 3′, 10X 5′, and 10X 5′ with Extended R1 sequencing. Each point represents the
739 average UMIs/cell of a single gene assayed in the indicated assays. At the top left, the
740 Pearson correlation coefficient and resulting p value are indicated.
741
742 **Supplemental Figure 2: A.** Comparison of performance metrics (average distance, AD;
743 average distance between means, ADM; average proportion of non-overlap, APN;
744 connectivity; Dunn Index; figure of merit, FOM; and silhouette index) by several
745 clustering methods (diana, model-based, hierarchical, kmeans, and pam) run on
746 sgmRNA expression of 600 randomly sampled cells analyzed with 10X 5′ Extended R1
747 and scCoVseq. *Left*: Performance metrics for each method across k values from 2 to 5.
748 *Right:* performance metrics for each method with k = 2. **B.** Visualization of infection
749 classification by different methods. **C.** Viral gene expression of cells by infection status,
750 determined by pam clustering method. **D.** Percent of infected cells per sample as
751 measured by flow cytometry, immunofluorescence, and infection classification with
752 unsupervised (pam) method or supervised infection classification by classifying infected
753 cells as those with at least 375 total viral UMIs. Because the same sample was
754 sequenced with 10X 5′ and 10X 5′ Extended R1, flow cytometry and
755 immunofluorescence results are duplicated for ease of visualization. Error bars for

756 immunofluorescence indicate mean ± one standard deviation of percent infected cells

757 based on three fields per sample.

758

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

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

761 and x-axis, respectively. The color indicates the $\log_2$ total UMIs/junction across all cells

762 in the SARS-CoV-2 infected sample. Below each axis, the number of UMIs supporting a

763 position as a junction start or end site is indicated with a density plot.

764

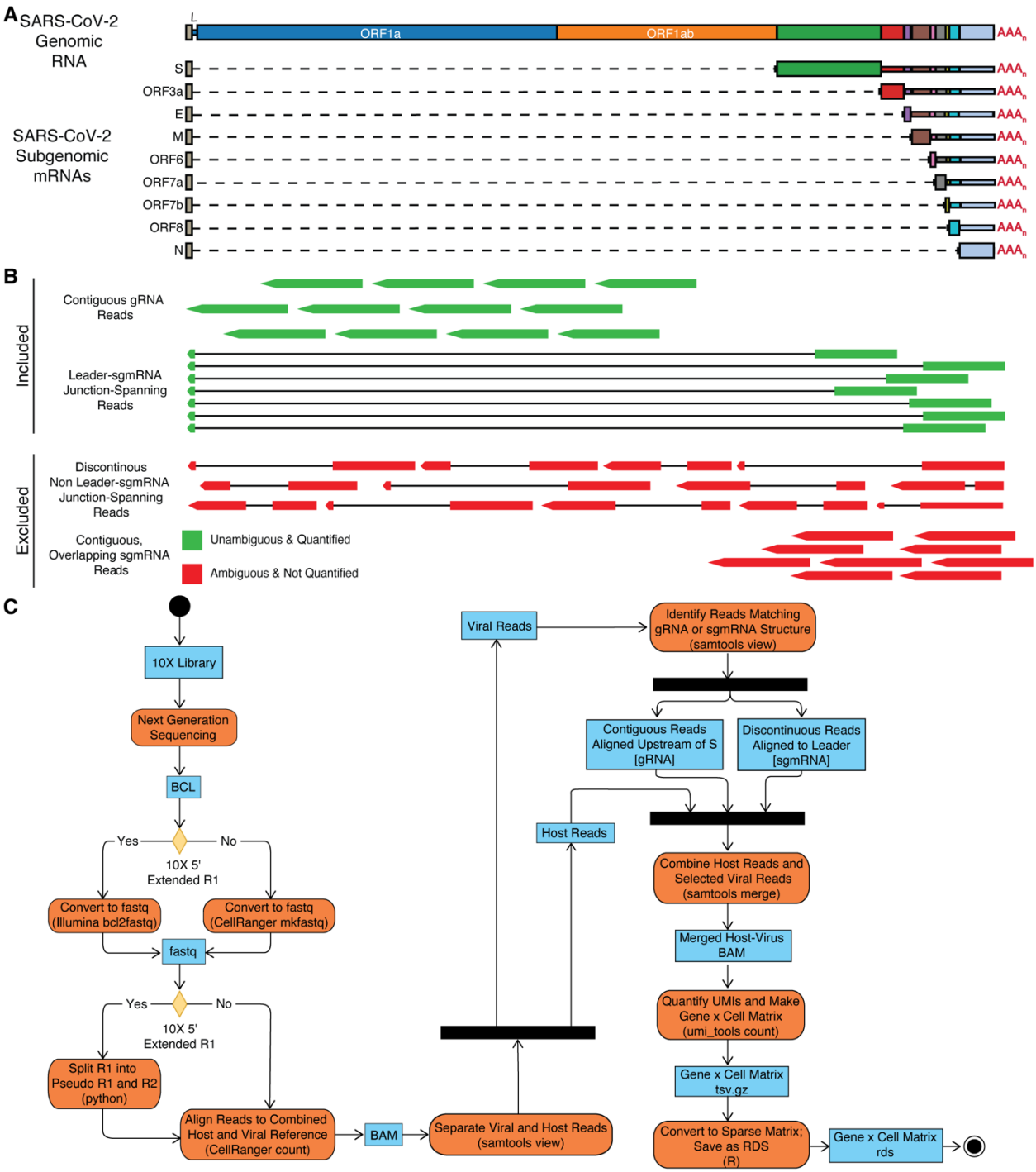


Figure 1: A. Illustration of SARS-CoV-2 genomic RNA, gRNA, and subgenomic RNAs, sgmRNAs. **B.** *Top:* Reads included for analysis by scCoVseq. Either: 1) contiguous reads mapping to ORF1a/b and therefore derived from gRNA or 2) discontinuous reads spanning the leader region and ORFS transcribed by sgmRNAs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

769 excluded from analysis by scCoVseq. Either: 1) discontinuous reads that do not include
770 sequence mapping to the leader region and downstream of S or 2) contiguous reads
771 that map to ORFs other than ORF1a/b, which are ambiguous. **C.** Activity diagram of
772 scCoVseq pipeline. Blue rectangles indicate inputs/outputs for each stage. Orange
773 rounded rectangles indicate a process in bold with software indicated.
774

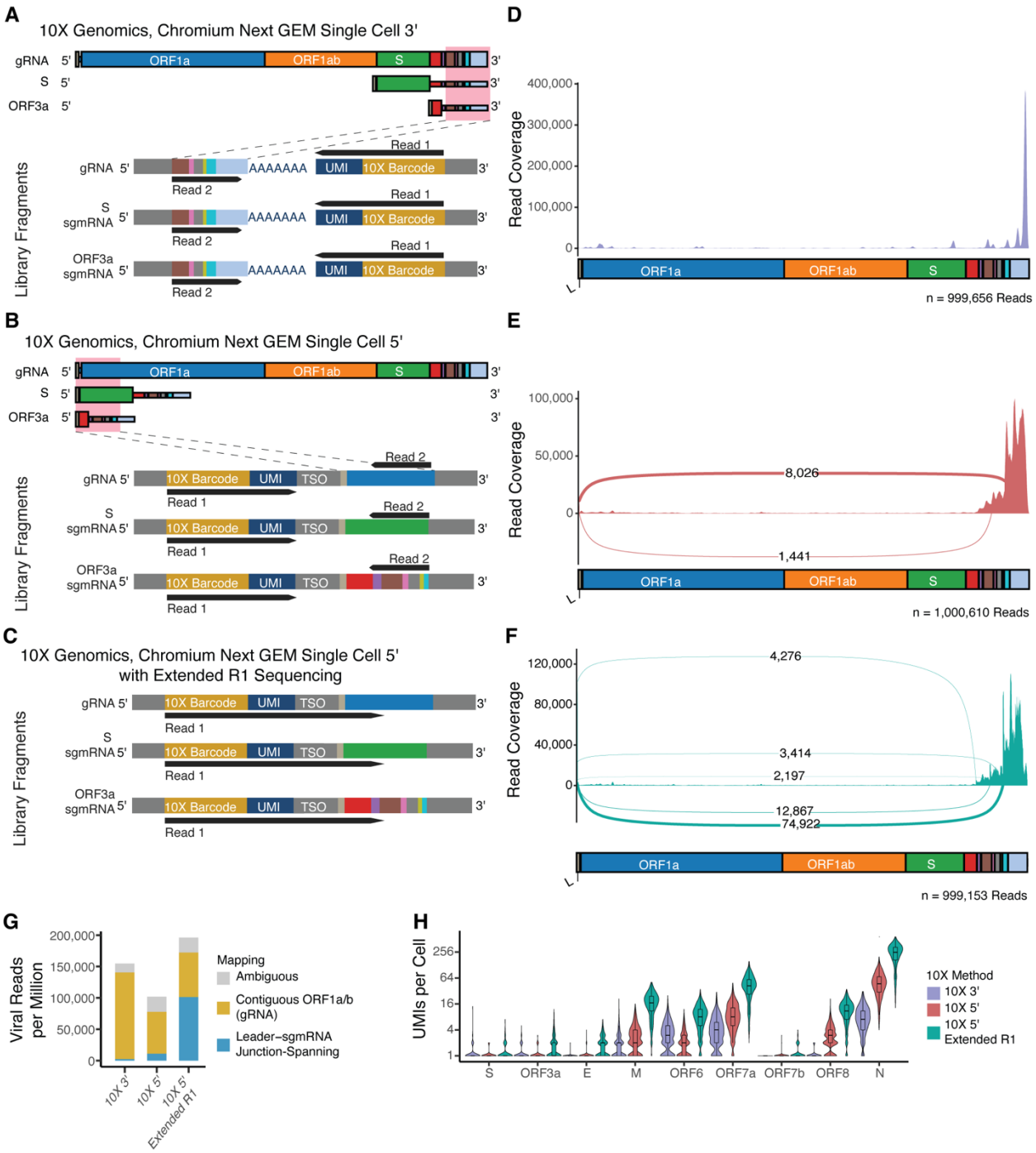


Figure 2: A-C. Illustration of gRNA and S and ORF3a sgmRNAs. Red box indicates regions contained in final 10X library. *Lower:* Illustration of 10X library fragments derived from gRNA and S and ORF3a sgmRNAs with sequencing read 1 and read 2 indicated. 10X 3′ (**A**), 10X 5′ (**B**), and 10X 5′ Extended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

779 Sashimi plot of 10X 3′ (**D**), 10X 5′ (**E**), and 10X 5′ Extended R1 (**F**) reads mapped to
780 the SARS-CoV-2 genome filtered to show only junctions supported by at least 1,000
781 reads. Total number of reads visualized is indicated in the bottom right. **G.** Reads per
782 million reads mapped to SARS-CoV-2 reads mapping to a single viral gene in 10X 3′,
783 10X 5′, or 10X 5′ with Extended R1 data. Reads are colored by their mapping with
784 contiguous reads mapping to ORF1a/b in yellow, leader-sgmRNA junction-spanning
785 reads in blue, and ambiguously mapped reads in grey. **H.** UMIs per cell for all sgmRNAs
786 in infected cells in each dataset. Each dataset was downsampled to an equal number of
787 infected cells and each cells' total UMIs were downsampled to the same value to control
788 for differences in sequencing depth. The leader region is enlarged in illustrations of the
789 genome for visibility. L = Leader.
790

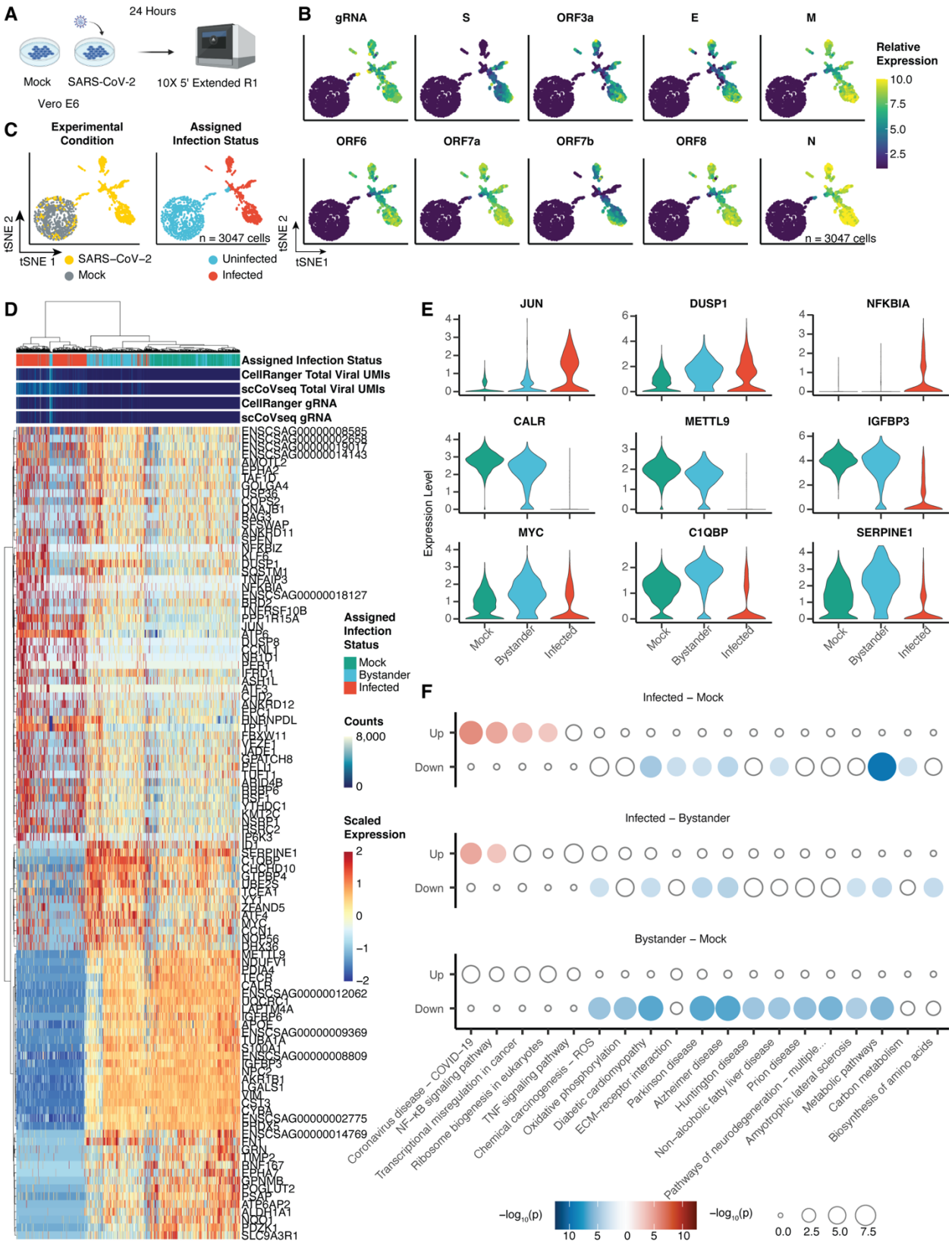
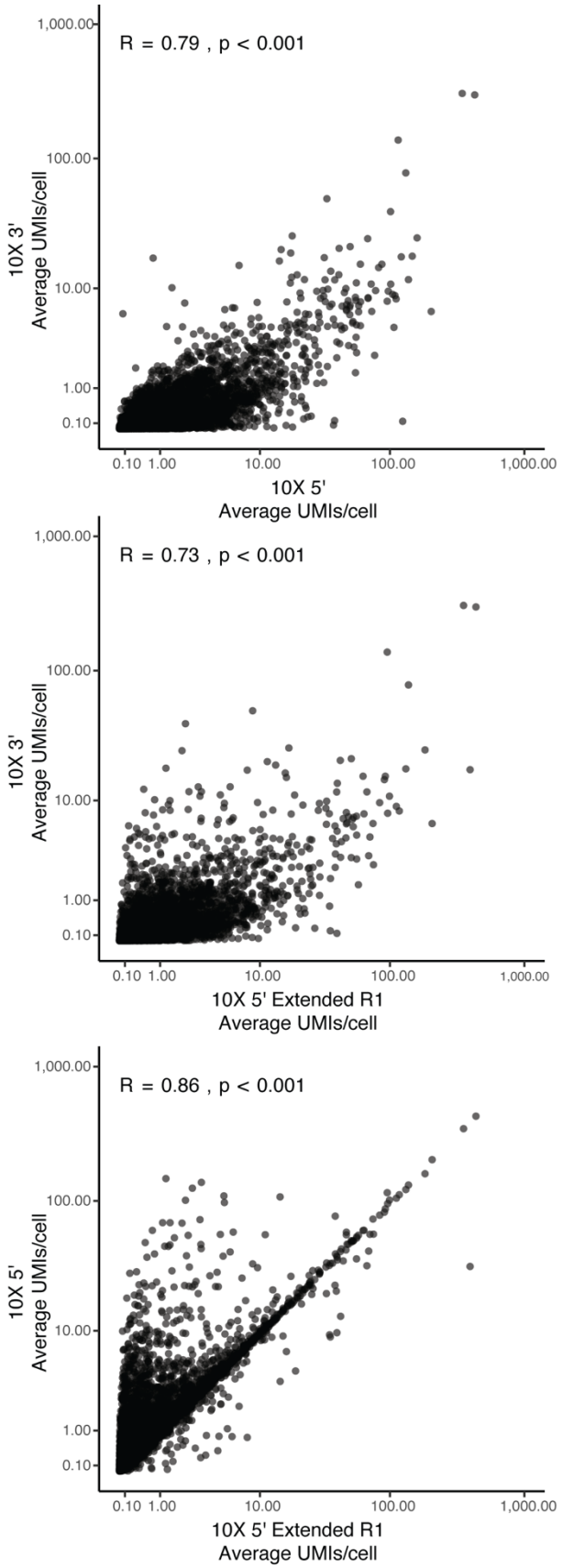


Figure 3: A. Experimental design. Vero E6 cells were infected or mock infected with

792 SARS-CoV-2 (USA-WA1/2020) at an MOI of 0.1. At 24 hours post-infection, cells were
793 analyzed by scRNAseq using 10X 5′ with Extended R1 sequencing. **B-C.** 3,047 mock
794 and infected cells embedded in tSNE space derived from euclidean distance of scaled
795 viral sgmRNA expression. Cells are colored by (**B**) indicated viral RNA expression, or
796 (**C**) experimental condition and assigned infection status of cells. **D.** Heatmap of genes
797 differentially expressed in infected, bystander, or mock cells. Differential expression
798 testing was performed on host gene expression downsampled to an equal number of
799 UMIs/cell across cells to account for infection-induced transcriptional shutdown. Genes
800 were selected for visualization based on false discovery rate of less than 0.05 and
801 absolute $\log_2$ fold change of at least 1. Non-downsampled gene expression data is
802 shown. Along the top, infection status, total viral UMIs and genomic RNA as quantified
803 by CellRanger and scCoVseq are indicated. Cells and genes are clustered with ward d2
804 clustering based on euclidean distance. **E.** Expression of selected host genes per cell
805 by infection status. Data shown is not downsampled. *Top:* genes induced in infected
806 cells. *Middle:* genes repressed in infected cells. *Bottom:* genes upregulated in bystander
807 cells compared to mock. **F.** KEGG pathway enrichment in genes differentially expressed
808 in pairwise comparisons of downsampled infected, bystander, and mock cells. Dot size
809 and fill indicates the $-\log_{10}$ p value of enrichment with red dots indicating enrichment in
810 the first infection state and blue in the second infection state noted above each panel.
811
812



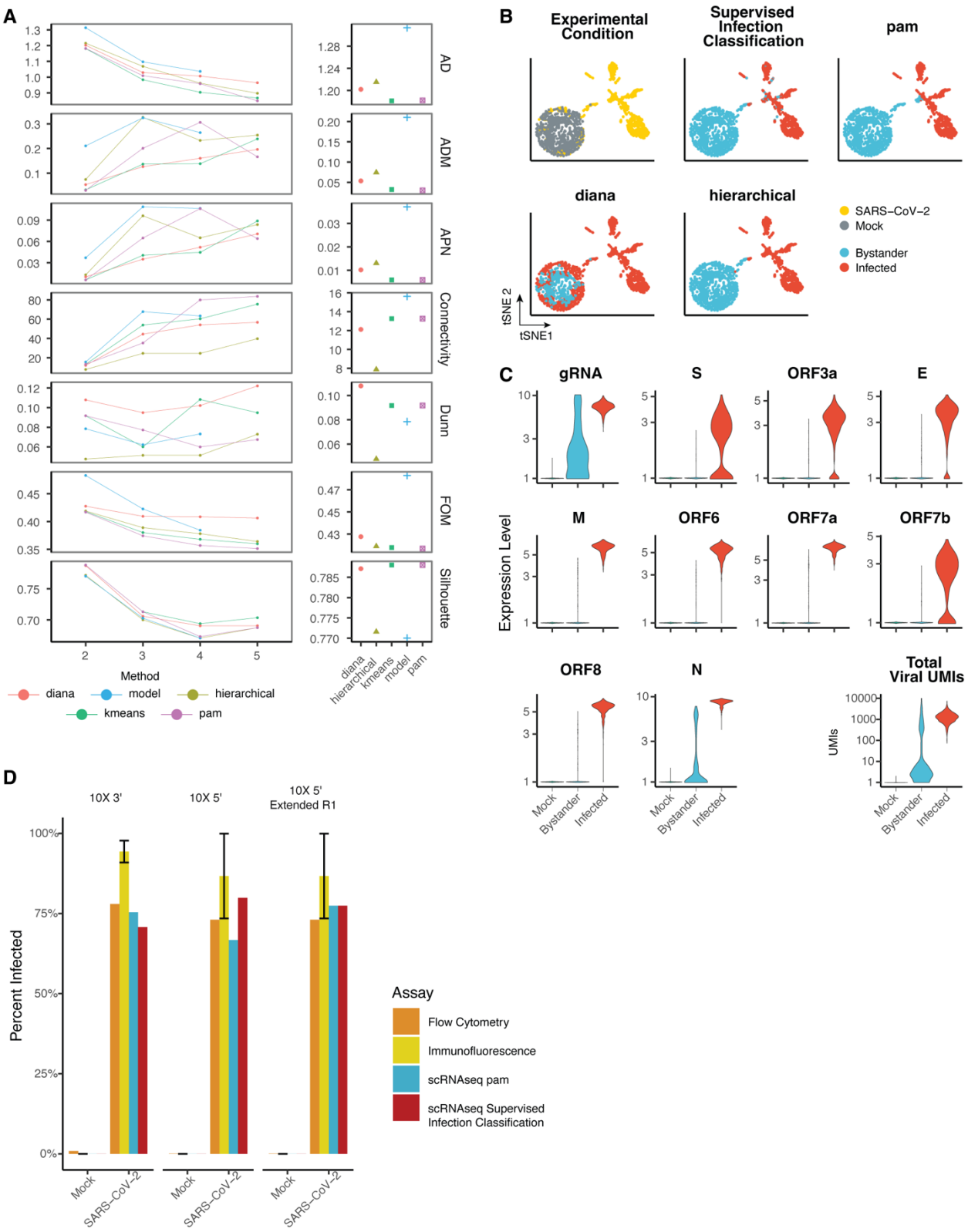
Supplemental Figure 1: Average counts of host gene expression of cells analyzed by

814 10X 3′, 10X 5′, and 10X 5′ with Extended R1 sequencing. Each point represents the

815 average expression of a single gene assayed in the indicated assays. At the top left, the

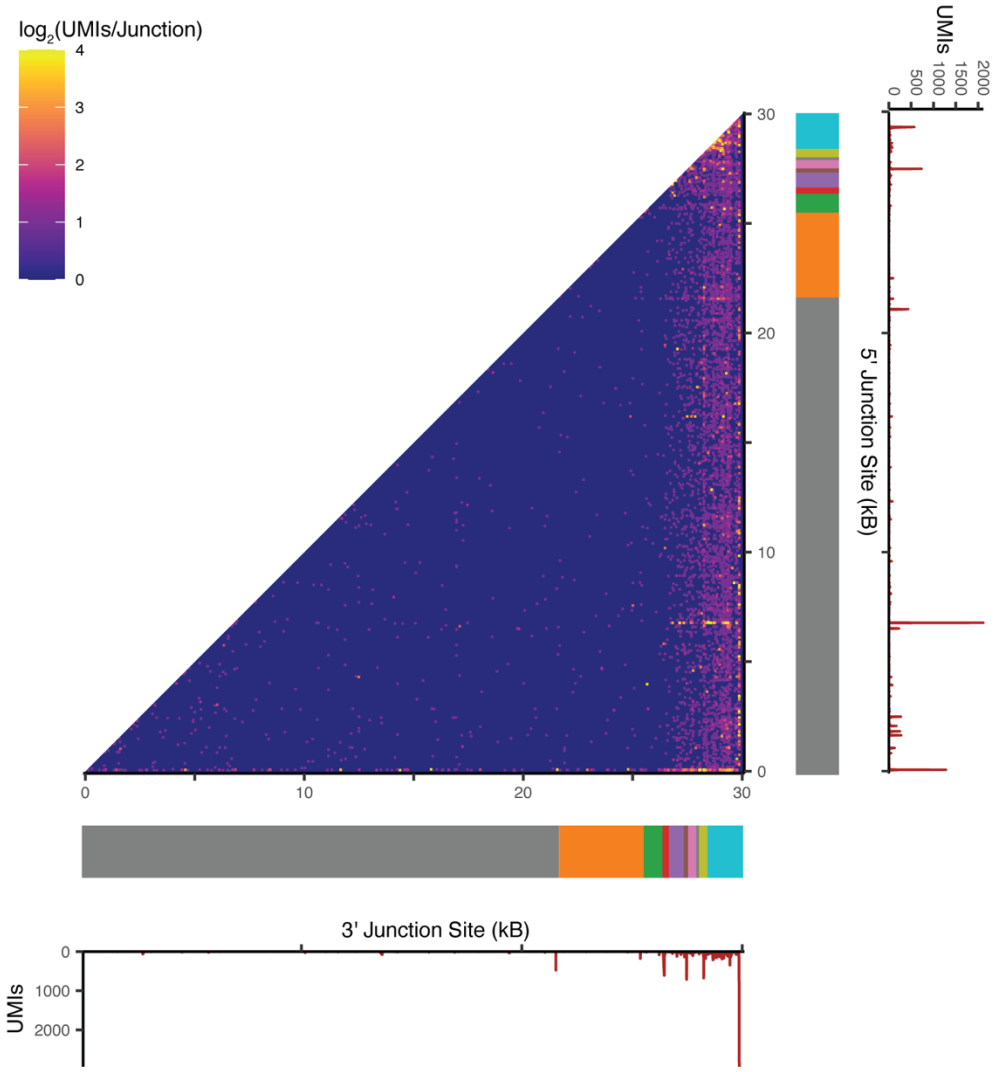
816 Pearson correlation coefficient and resulting p value are indicated.

817



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

820 connectivity; Dunn Index; figure of merit, FOM; and silhouette index) by several
821 clustering methods (diana, model-based, hierarchical, kmeans, and pam) run on
822 sgmRNA expression of 600 randomly sampled cells analyzed with 10X 5′ Extended R1
823 and scCoVseq. *Left*: Performance metrics for each method across k values from 2 to 5.
824 *Right:* performance metrics for each method with k = 2. **B.** Visualization of infection
825 classification by different methods. **C.** Viral gene expression of cells by infection status,
826 determined by pam clustering method. **D.** Percent of infected cells per sample as
827 measured by flow cytometry, immunofluorescence, and infection classification with
828 unsupervised (pam) method or supervised infection classification by classifying infected
829 cells as those with at least 375 total viral UMIs. Because the same sample was
830 sequenced with 10X 5′ and 10X 5′ Extended R1, flow cytometry and
831 immunofluorescence results are duplicated for ease of visualization. Error bars for
832 immunofluorescence indicate mean ± one standard deviation of percent infected cells
833 based on three fields per sample.
834



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. Below each axis, the number of UMIs supporting a position as a junction start or end site is indicated with a density plot.