scViralQuant: A tool for efficient alignment and quantification of viral reads from 10x single cell sequencing data sets

Leanne S. Whitmore[1*], Jennifer Tisoncik-Go[1], Michael Gale, Jr.[1*]

[1]Center for Innate Immunity and Immune Disease, Department of Immunology, School of Medicine, University of Washington, Seattle, WA

*Corresponding Authors:

Leanne S. Whitmore lwhitmo@uw.edu

Michael Gale, Jr. mgale@uw.edu

Competing interests:

Funding Statement: This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201800008C (M. Gale Jr.), and by the Bill & Melinda Gates Foundation INV-002704 (J. Estes). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Abstract

Currently there is a lack of efficient computational pipelines/tools for conducting alignment of viral reads to a viral genome for specific viral genome sequence reads from single cell RNA sequencing (scRNAseq) experiments of virus-infected cells. Contemporary options utilize a time and labor-intensive process that includes integration of the viral genome and viral gene location information into the host reference genome data set, regeneration of genome index files, and then conducting read alignments. To address the need for new tools to directly map and quantify viral sequence reads from within an infected cell scRNAseq data sets, we have built a python package, called scViralQuant. scViralQuant extracts sequences that were not aligned to the primary host genome, maps them to a viral genome of interest, counts viral reads, and then reintegrates viral sequence counts into matrix files that are used by standard single cell pipelines for downstream analyses with only one command. scViralQuant provides a scRNAseq viral genome-wide sequence read abundance analysis.

Introduction

Defining the virus-host interactions of viral infections is reliant on understanding the effect of the virus on host cell gene expression concomitant with evaluating the dynamics of viral gene expression. The emergence of SARS-CoV-2 and the COVID-19 pandemic underscore this need for integrated analysis of host and viral gene expression in single cells during acute virus infection. Integrated analysis of host and viral gene expression allows for the discovery of molecular mechanisms of viral and/or host-mediated pathogenesis, and possible genetic targets for therapeutic intervention of infection and disease. Similarly, such analyses are paramount for understanding the outcome of chronic virus infection such as hepatitis B virus or human immunodeficiency virus (HIV), and further demonstrate the need for a rapid and efficient analysis pipeline for quantifying genome-wide viral and host gene expression in single cell RNA sequencing (scRNAseq) experiments. The current process for quantification of viral RNA sequence reads with 10x single cell technology is laborious, involving the following computational options: 1) integrating the viral genome into the host genome, regenerating the CellRanger(1) genome index files, and then running the read alignment process or 2) performing a series of post-processing steps to extract unmapped reads from CellRanger output and aligning the extracted reads to the viral genome of interest. Both options have shortcomings. Option 1 requires additional extensive computational run time, as the target genome must be regenerated followed by read mapping, whereas Option 2 is highly cumbersome to require the implementation of several additional bioinformatics tools while additionally increasing the analysis time line.

Here, we report a new tool for alignment and quantification of viral sequence reads from 10x single cell sequencing data of virus-infected cells. Our new tool, scViralQuant (SVQ), minimizes the workload on users by packaging the necessary bioinformatics tools into an easy to install and use python library that can be executed with a single command. SVQ is initiated

with a single command that executes several analysis steps, including: 1) isolation of unaligned reads, 2) mapping of the isolated reads to the viral genome of interest with either bbmap(2) or bowtie2(3), 3) quantification of the resulting unique molecular identifier (UMI) read counts with HTseq(4) and 4) integration of the data set output back into the matrix files from the 10x scRNAseq data set. This integration step then allows the user to proceed with the analysis using common tools, such as 10x Loupe Browser (https://www.10xgenomics.com/products/loupe-browser), Seurat(5) or Monocle(6), or any other standard scRNAseq analysis pipeline.

Design and Implementation

SVQ, initiates by extracting unaligned sequence reads from the host genome using SAMtools(7), followed by alignment of read to the viral genome of interest using either bbmap (default)(2) or bowtie2(3) aligners (Fig 1). These two aligners were chosen for their speed and accuracy. Bbmap(2) and bowtie2(3) use different alignment algorithms, thus enabling the user to choose the aligner appropriate for their analysis (Table 1). Default parameters, such as rate of mismatches, insertions, and deletions (INDELs) vary across both alignment tools, resulting in slightly different outputs. SVQ uses bbmap and bowtie2 default settings but the user may change any parameter of the aligners by adjusting the --bbmap_params or --bowtie2_params options, giving the user more control over the viral alignment process compared to CellRanger where individual STAR alignment parameters cannot be altered. UMI counts are quantified using HTSeq(4) first, counting any reads mapping to the viral genome and then quantifying individual gene counts if a gene transfer format (gtf) file for the specific viral genome is provided by the user. Quantifying UMI counts both for the whole viral genome and individual viral genes provides output to determine if viral transcripts exist in the cell while providing specific information on viral gene expression levels in each cell. UMI counts per cell are then integrated into the CellRanger raw and filtered feature matrix files to allow for easy downstream analysis of viral genome and specific viral gene expression levels concomitant with host gene expression levels. Additionally, SVQ output provides viral count results in a csv file format such that the data can be visualized in violin plots that show the single cell analysis of viral reads and viral gene counts (Fig 1).

Fig 1. Software layout. Yellow boxes show the progressive steps implemented by SVQ analysis. White boxes with sold outline indicate required input from the user. White boxes with dotted outline are output from SVQ**.**

Table 1: **Alignment tools used in the analyses**

Aligners	Version	Sequencing Data Type	Alignment Algorithm	Splice Aware
SVQ/BBMAP	38.9	RNA/DNA	Multi kmer seed and extend (http://bib.irb.hr/datoteka/773708.Josip_Maric_diplomski.pdf)	Y
SVQ/Bowtie2	2.4.2	RNA/DNA	Burrows-wheeler transform https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25	N
CellRanger/STAR	6.1.2	RNA	Maximal Mappable Prefix (*MMP*) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/	Y

Results

Comparing SVQ to other methods of viral alignment

As a proof-of-principal, we assessed the functionality and accuracy of SVQ by quantifying viral read counts from two published 10x scRNAseq cell culture virus infection experiments (Table 2), including SARS-CoV-2 infection of human bronchial epithelial cells(8) and HIV infection in a CD4 T cell model of viral infection and latency(9). Each of these experiments reported detection of viral reads in cells within the infected cultures using CellRanger(8)-(9). We ran both CellRanger and SVQ (using bbmap and bowite2 with default alignment parameters) on all samples from both experiments, and directly compared the count output results from SVQ to the CellRanger output.

Table 2: **Sequencing sets used in analyses**.

Single cell data	Sample name	SRA ID	Reference
SARS Cov-2	1dpi CoV2 HBECs	SRR13711614	(8)
	2dpi CoV2 HBECs	SRR13711615	
	3dpi CoV2 HBECs	SRR13711616	
HIV	HIV infected donor 1	SRR6825024	(9)
	HIV infected donor 2	SRR6825025	

HIV comparison

The single HIV infection scRNAseq study evaluated viral infection and latency in primary CD4+ T cells isolated from two healthy human donors and infected ex vivo using the recombinant pNL4-3-Δ6-drEGFP HIV strain(9). We utilized the data set from this study because expression of enhanced green fluorescent protein (EGFP) from the engineered viral genome could be monitored to further confirm viral-mediated gene expression in cells as in internal control. We aligned reads from the donor samples infected with HIV to the human genome GRCh38 integrated with the EGFP gene using CellRanger v6.0.1. To obtain viral read alignments and compare SVQ to current methods, we performed 3 different analyses: 1) computational integration of the HIV genome into the human genome followed by aligning reads using CellRanger, which is the current method of aligning viral reads, 2) aligned reads to the human genome and then aligned viral reads to the HIV genome using SVQ with bbmap and 3) aligned viral reads to the HIV genome using SVQ with bowtie2. In both HIV-infected donor samples, we observed 100% overlap in the cells that had both EGFP and HIV reads (EGFP+HIV+) across all 3 alignment methods. The number of EGFP+HIV+ cells were 3,653 and 183 for donor samples 1 and 2, respectively. The read counts for HIV were correlated (Kendall correlation) in EGFP+HIV+ cells between methods to ensure SVQ was obtaining comparable counts to CellRanger (**Fig 2A)**. We observed high correlation of counts between CellRanger and SVQ in EGFP+HIV+ cells, showing the accuracy of both alignment methods in SVQ.

Figure 2: SQV comparison results from HIV-infected cells. Kendall correlation values between SVQ and CellRanger UMI counts for EGFP+HIV+ cells from **A)** HIV infected donor 1, and **B)** HIV infected donor 2. Upset plots of the overlap of EGFP-HIV+ cells detected by CellRanger and SVQ with bbmap or bowtie2 are shown for **C)** HIV infected donor 1, and **D)** HIV infected donor 2. Scatter plots showing the correlation of UMI counts in EGFP-HIV+ cells

between CellRanger and SVQ with either bbmap or bowtie2 are shown for **E)** HIV infected donor 1, and **F)** HIV infected donor 2. **G)** Kendall correlation values between SVQ bbmap and bowtie2 for viral gene UMI counts for HIV-infected donor samples 1 and 2 are shown**.**

In addition to EGFP+HIV+ cells, both CellRanger and SVQ detected other cells with HIV reads but no EGFP reads (EGFP-HIV+). A total of 8,133 and 712 cells in HIV infected donor samples 1 and 2, respectively, were detected to be EGFP-HIV+ by all 3 alignment methods. A smaller number of cells were detected to be HIV+ by CellRanger & SVQ bbmap, SVQ bowtie2 & bbmap, or uniquely by CellRanger or SVQ bbmap (Figs 2C and D). Variations in the cells detected with viral reads were due to different alignment algorithms with different acceptable default rates of mismatches and INDELs (Table 1). We note that bbmap and CellRanger are splice aware aligners in that splice aware aligners predict where transcripts are being post transcriptionally spliced together across introns by integrating known splice junction sequence information when aligning reads to a genome. In contrast, splice unaware aligners do not account for splicing when aligning reads to genomes. We found that the HIV UMI counts in EGFP-HIV+ cells between all 3 aligners were highly correlative (Figs 2E and F), with bbmap and CellRanger UMI counts being slightly more strongly correlated since both alignment algorithms are splice aware. Lastly, we compared UMI counts for individual HIV genes (gag, nef, pol, vif and vpr) between bowtie2 and bbmap in EGFP-HIV+ cells by performing a correlation of the counts for each gene. Significant and high correlation values were observed for HIV genes in each sample (Fig 2G), showing a consistent quantification of HIV gene counts between the two aligner options in SVQ. EGFP-HIV+ read output in the HIV latency model is appropriate and expected, as viral latency is linked with selective and minimal HIV gene expression(10).

SARS-CoV-2 comparison

The recent COVID-19 pandemic exposed a high need for viral quantification pipelines for scRNAseq data set analyses(11). Therefore, we also tested SVQ performance on SARS-CoV-2 infection data sets. We selected a 10x scRNAseq data set of human bronchial epithelial cells (HBECs) infected with SARS-CoV-2 with time point analysis of samples collected following 1, 2, or 3 days post-infection (dpi)(8). Using CellRanger we aligned sequence reads first to the human genome GRCh38 followed by alignment of remaining reads to the SARS-CoV-2 genome (GCF_009858895.2) using the 3 alignment methods described above for the HIV analyses. Whereas the SARS-CoV-2 dataset did not have an internal positive control for viral expression, all 3 alignment methods were able to identify SARS-CoV-2 sequence reads in the same 4,279, 13,998 and 26,935 cells for samples from 1,2 and 3 dpi, respectively (**Fig 3A-C**). The UMI counts in these cells were found to be highly correlated to the CellRanger UMI counts for both SVQ bbmap and bowtie2 wherein a slightly higher correlation was observed between SVQ bbmap and CellRanger with this analysis (**Figs 3D-I**). Similar to the findings with the HIV infection dataset, a smaller percentage of cells had SARS-CoV-2 detected by CellRanger & SVQ bbmap, SVQ bowtie2 & bbmap, or uniquely by CellRanger or SVQ bbmap. Lastly, we performed correlation between SVQ bowtie2 and bbmap on the quantification of UMIs for the ten SARS-CoV-2 genes. These analyses revealed significant and high correlations of the quantification of the SARS-CoV-2 genes in each sample between the two alignment methods for SVQ (**Fig 3J**).

Figure 3: SQV comparison results from SARS-CoV-2-infected cells. Upset plots depict the number of cells with SARS-CoV-2 RNA that overlap between analyses using CellRanger, SVQ bowtie2 and SVQ bbmap for **A)** 1dpi human bronchial epithelial cells (HBEC), **B)** 2dpi HBEC and **C)** 3dpi HBEC. Correlation of UMI counts of SARS-CoV-2 RNA+ cells using

CellRanger and SVQ bbmap for **D)** 1dpi HBEC, **E)** 2dpi HBEC and **F)** 3dpi HBEC samples. Correlation of UMI counts of SARS-CoV-2 RNA+ cells using CellRanger and SVQ bowtie2 for **G)** 1dpi HBEC, **H)** 2dpi HBEC and **I)** 3dpi HBEC samples. **J)** Correlation values for SARS-CoV-2 gene UMI counts between SVQ alignment methods bbmap and bowtie2 for SARS-CoV-2-infected samples collected at 1, 2, and 3 dpi.

Availability and Future Directions

SVQ is a new bioinformatics analysis tool that provides an alternative and efficient method for quantifying viral genome-specific sequence reads in scRNAseq experiments. In addition to the two experiment examples this analysis, SVQ is widely applicable to analysis of any scRNAseq experiment of virus-infected cells. SVQ also provides viral gene expression information allowing the user more specific data on viral gene expression and replication dynamics. These features enable more control to the user over the alignment parameters compared to current methods. The SVQ package is available software accessible at https://github.com/galelab/scViralQuant to serve as a resource for the community.

References

1. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Jan 16;8(1):14049.

2. BBMap: A Fast, Accurate, Splice-Aware Aligner (Conference) | OSTI.GOV [Internet]. [cited 2022 Dec 27]. Available from: https://www.osti.gov/biblio/1241166

3. Fast gapped-read alignment with Bowtie 2 - PMC [Internet]. [cited 2022 Dec 27]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/

4. HTSeq—a Python framework to work with high-throughput sequencing data - PMC [Internet]. [cited 2022 Dec 27]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287950/

5. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.

6. Reversed graph embedding resolves complex single-cell trajectories | Nature Methods [Internet]. [cited 2023 Jan 19]. Available from: https://www.nature.com/articles/nmeth.4402

7. Sequence Alignment/Map format and SAMtools | Bioinformatics | Oxford Academic [Internet]. [cited 2022 Dec 27]. Available from: https://academic.oup.com/bioinformatics/article/25/16/2078/204688

8. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes | PLOS Biology [Internet]. [cited 2022 Dec 27]. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001143

9. Bradley T, Ferrari G, Haynes BF, Margolis DM, Browne EP. Single-Cell Analysis of Quiescent HIV Infection Reveals Host Transcriptional Profiles that Regulate Proviral Latency. Cell Reports. 2018 Oct 2;25(1):107-117.e3.

10. Battistini A, Sgarbanti M. HIV-1 Latency: An Update of Molecular Mechanisms and Therapeutic Strategies. Viruses. 2014 Apr 14;6(4):1715–58.

11. Liu W, Jia J, Dai Y, Chen W, Pei G, Yan Q, et al. Delineating COVID-19 immunological features using single-cell RNA sequencing. Innovation (Camb). 2022 Jul 21;3(5):100289.

Fig 1

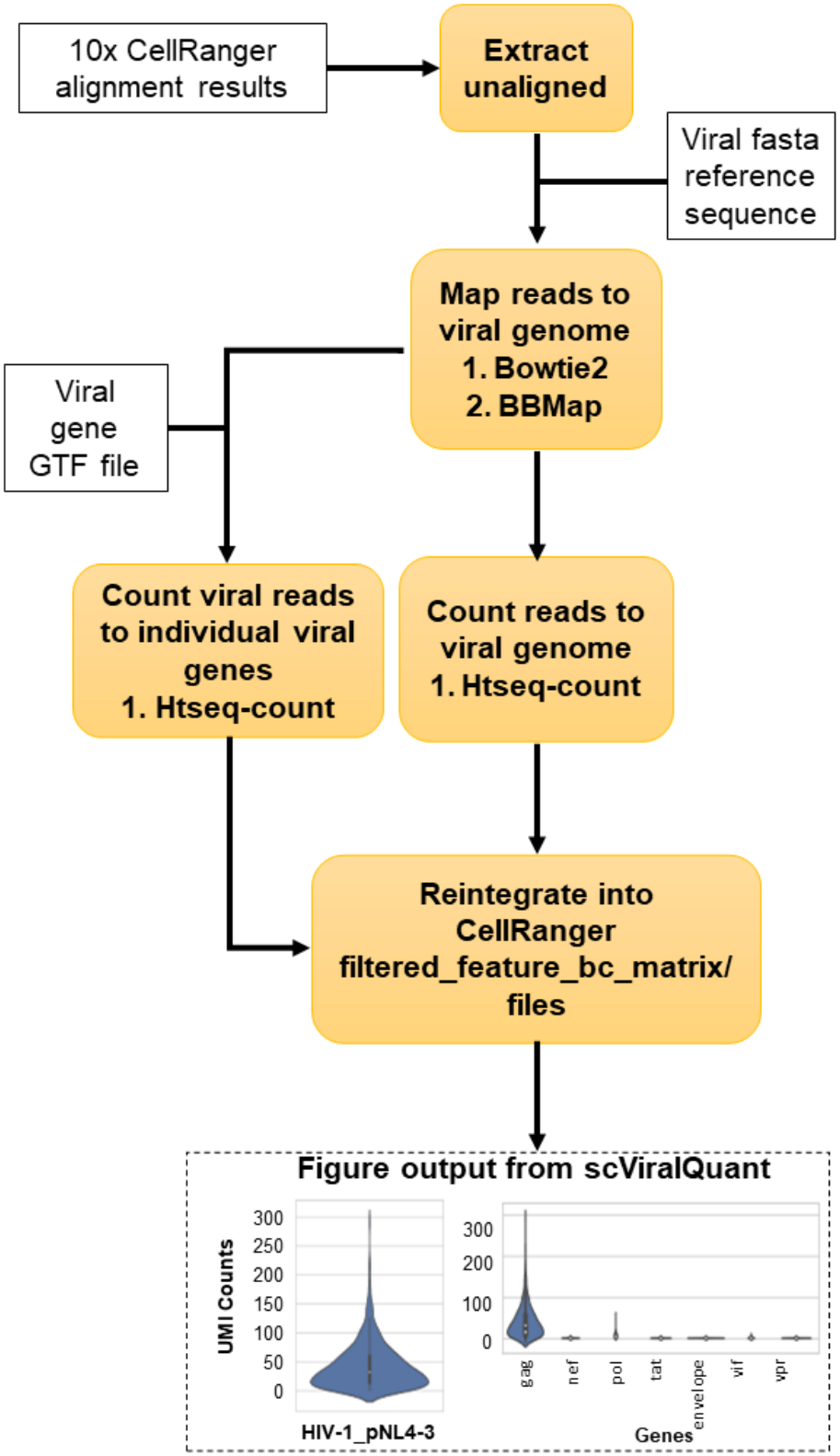

10x CellRanger alignment results
Extract unaligned
Viral fasta reference sequence
Map reads to viral genome
1. Bowtie2
2. BBMap
Viral gene GTF file
Count viral reads to individual viral genes
1. Htseq-count
Count reads to viral genome
1. Htseq-count
Reintegrate into CellRanger filtered_feature_bc_matrix/ files
Figure output from scViralQuant
UMI Counts
300
250
200
150
100
50
0
HIV-1_pNL4-3
300
200
100
0
gag
nef
pol
tat
envelope
vif
vpr
Genes

Fig 2

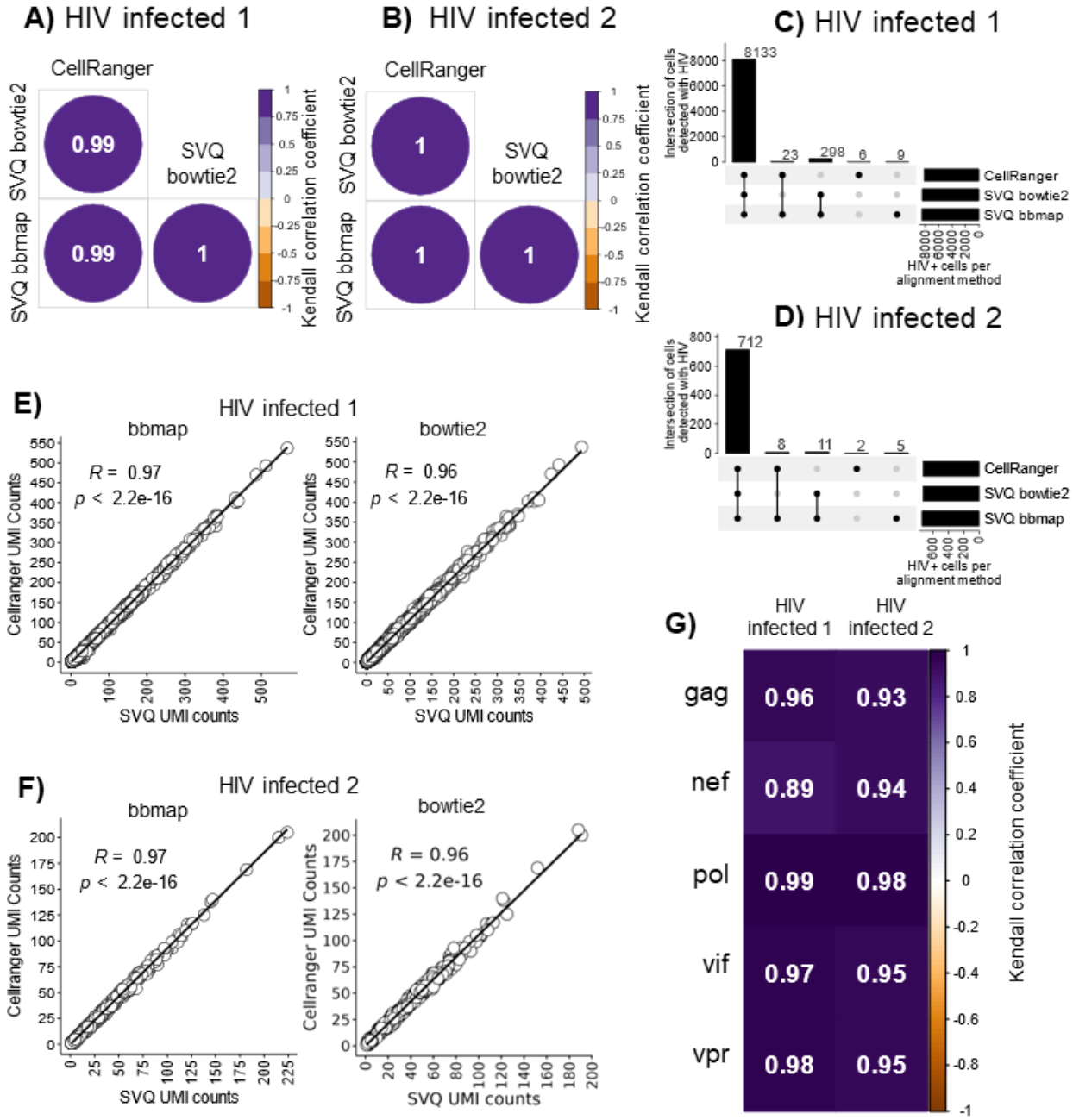
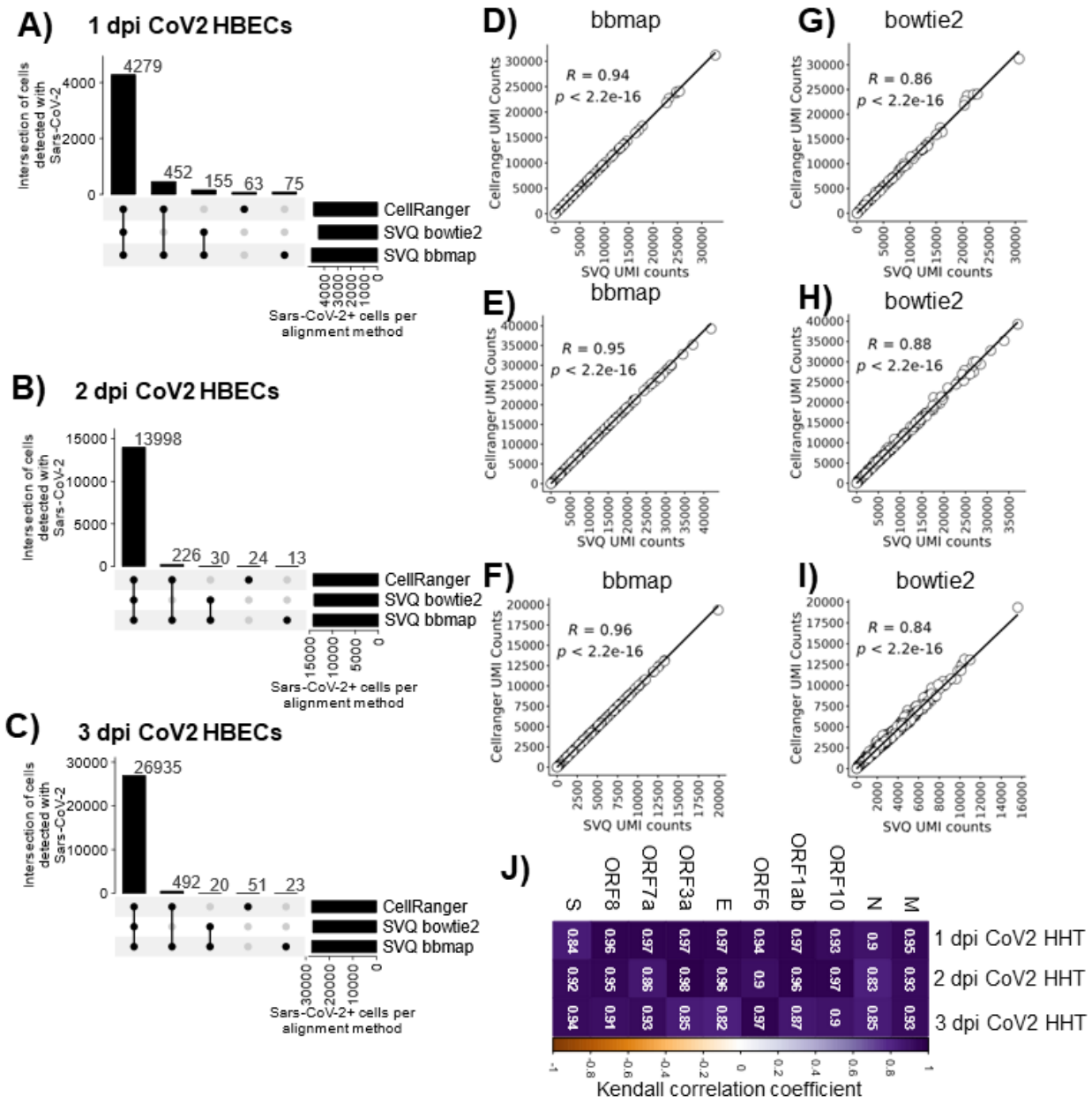

A) HIV infected 1
CellRanger
SVQ bowtie2
SVQ bbmap
0.99
0.99
1
1
Kendall correlation coefficient
B) HIV infected 2
CellRanger
SVQ bowtie2
SVQ bbmap
1
1
1
Kendall correlation coefficient
C) HIV infected 1
Intersection of cells detected with HIV
8133
23
298
6
9
CellRanger
SVQ bowtie2
SVQ bbmap
HIV+ cells per alignment method
D) HIV infected 2
Intersection of cells detected with HIV
712
8
11
2
5
CellRanger
SVQ bowtie2
SVQ bbmap
HIV+ cells per alignment method
E)
HIV infected 1
bbmap
R = 0.97
p < 2.2e-16
Cellranger UMI Counts
SVQ UMI counts
bowtie2
R = 0.96
p < 2.2e-16
Cellranger UMI Counts
SVQ UMI counts
F)
HIV infected 2
bbmap
R = 0.97
p < 2.2e-16
Cellranger UMI Counts
SVQ UMI counts
bowtie2
R = 0.96
p < 2.2e-16
Cellranger UMI Counts
SVQ UMI counts
G)
HIV infected 1
HIV infected 2
gag
nef
pol
vif
vpr
0.96
0.93
0.89
0.94
0.99
0.98
0.97
0.95
0.98
0.95
Kendall correlation coefficient

Fig 3


A) 1 dpi CoV2 HBECs
Intersection of cells detected with Sars-CoV-2
4279
452
155
63
75
CellRanger
SVQ bowtie2
SVQ bbmap
Sars-CoV-2+ cells per alignment method
B) 2 dpi CoV2 HBECs
13998
226
30
24
13
C) 3 dpi CoV2 HBECs
26935
492
20
51
23
D) bbmap
R = 0.94
p < 2.2e-16
E) bbmap
R = 0.95
p < 2.2e-16
F) bbmap
R = 0.96
p < 2.2e-16
G) bowtie2
R = 0.86
p < 2.2e-16
H) bowtie2
R = 0.88
p < 2.2e-16
I) bowtie2
R = 0.84
p < 2.2e-16
Cellranger UMI Counts
SVQ UMI counts
J)
S
ORF8
ORF7a
ORF3a
E
ORF6
ORF1ab
ORF10
N
M
1 dpi CoV2 HHT
2 dpi CoV2 HHT
3 dpi CoV2 HHT
Kendall correlation coefficient