PNAS
www.pnas.org

Supplementary Information for

Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues

Liguo Zhang[1], Alexsia Richards[1], M. Inmaculada Barrasa[1], Stephen H. Hughes[2], Richard A. Young[1, 3] and Rudolf Jaenisch[1, 3, #]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA, USA.

[2] HIV Dynamics and Replication Program, Center for Cancer Research, National Cancer Institute, Frederick, MD, USA.

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.

[#]Correspondence: jaenisch@wi.mit.edu

This PDF file includes:

Figures S1 to S7
Tables S1 to S4
Legends for Datasets S1 to S4

Other supplementary materials for this manuscript include the following:

Datasets S1 to S4

Supplementary Figures

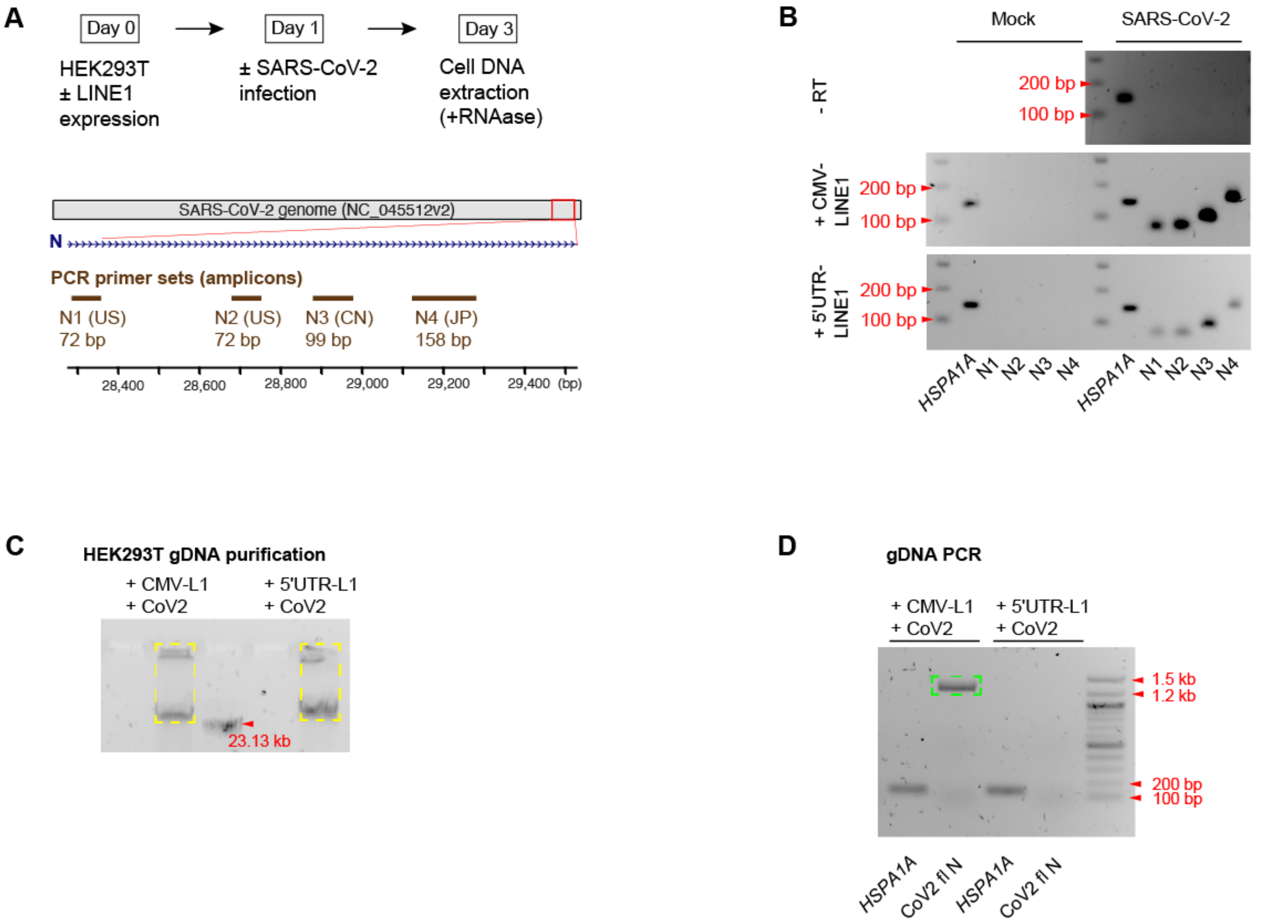
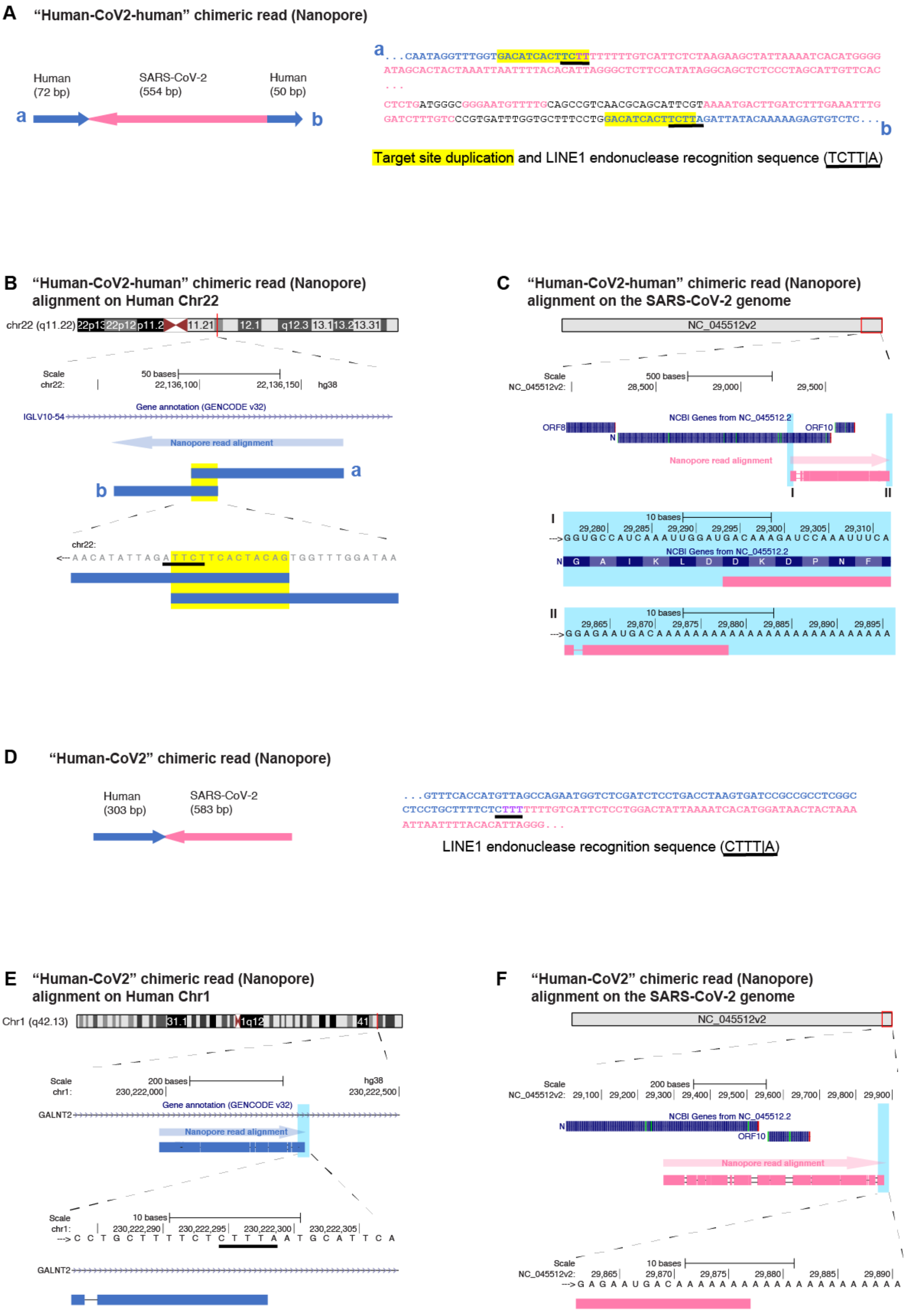


Fig. S1. Detection of DNA copies of SARS-CoV-2 RNA in infected LINE1-overexpressing cells. A) Experimental workflow (top) and PCR primer sets (bottom) used to detect reverse-transcription and integration of SARS-CoV-2 RNA. **B)** PCR detection of SARS-CoV-2 NC sequences in DNA purified from mock (left) or SARS-CoV2 (right) infected HEK293T cells without or with transfection of human LINE1 (CMV-LINE1 or 5'UTR-LINE1) plasmids. *HSPA1A*: human *HSPA1A* gene as control; N1 – N4: SARS-CoV-2 NC amplicons as shown in **A)**. N1 – N4 PCR products were loaded on gel three times the amount of *HSPA1A* PCR product. Note that we didn't detect DNA copies of SARS-CoV-2 sequences in cells without LINE1 overexpression by this low-sensitive PCR assay. **C)** Gel purification of large fragments of genomic DNA (yellow boxes) from SARS-CoV-2 infected HEK293T cells that were transfected

with CMV-LINE1 or 5'UTR-LINE1. **D)** Cloning of a DNA copy of a complete SARS-CoV-2 NC gene sequence (CoV2 fl N, green box) from gel-purified HEK293T genomic DNA**.**



A "Human-CoV2-human" chimeric read (Nanopore)

a ...CAATAGGTTTGGTGACATCACTTCTTTTTTTTGTCATTCTCTAAGAAGCTATTAAAATCACATGGGG
ATAGCACTACTAAATTAATTTTACACATTAGGGCTCTTCCATATAGGCAGCTCTCCCTAGCATTGTTCAC
...
CTCTGATGGGCGGGAATGTTTTGCAGCCGTCAACGCAGCATTCGTAAAATGACTTGATCTTTGAAATTTG
GATCTTTGTCCCGTGATTTGGTGCTTTCCTGGACATCACTTCTTAGATTATACAAAAGAGTGTCTC... b

Target site duplication and LINE1 endonuclease recognition sequence (TCTT|A)

B "Human-CoV2-human" chimeric read (Nanopore) alignment on Human Chr22

C "Human-CoV2-human" chimeric read (Nanopore) alignment on the SARS-CoV-2 genome

D "Human-CoV2" chimeric read (Nanopore)

...GTTTCACCATGTTAGCCAGAATGGTCTCGATCTCCTGACCTAAGTGATCCGCCGCCTCGGC
CTCCTGCTTTTCTCTTTTTTTGTCATTCTCCTGGACTATTAAAATCACATGGATAACTACTAAA
ATTAATTTTACACATTAGGG...

LINE1 endonuclease recognition sequence (CTTT|A)

E "Human-CoV2" chimeric read (Nanopore) alignment on Human Chr1

F "Human-CoV2" chimeric read (Nanopore) alignment on the SARS-CoV-2 genome

Fig. S2. Nanopore sequencing reads provide evidence for integration of SARS-CoV-2 sequences. A) A Nanopore sequencing read showing integration of a SARS-CoV-2 NC sub-genomic RNA sequence (magenta) and human genomic sequences (blue) flanking both sides of the integrated viral sequence. Features indicative of LINE1 mediated "target-primed reverse transcription" include: the target site duplication (yellow highlight) and the LINE1 endonuclease recognition sequence (underlined). Sequences that could be aligned to both genomes are shown in purple and sequences that cannot be aligned are shown in black. Arrows indicate sequence orientations with regard to the human and SARS-CoV-2 genomes as shown in **B, C**), with the two sides labeled as "a" and "b" in blue. **B)** Alignment of the Nanopore read in **A**) with the human genome (chromosome 22) showing the integration site. The human sequences at the junction region show the target site which was duplicated when the SARS-CoV-2 cDNA was integrated (yellow highlight) and the LINE1 endonuclease recognition sequence (underlined). **C)** Alignment of the Nanopore read in **A**) with the SARS-CoV-2 genome showing the integrated viral DNA represents a DNA copy of a portion of the NC sub-genomic RNA. Light blue highlighted regions are enlarged to show the two ends of the read. **D)** A Nanopore sequencing read showing the integrated portion of a SARS-CoV-2 RNA sequence (magenta) and human genomic sequences (blue) from one side of the junction with a LINE1 endonuclease recognition sequence (underlined). Sequences that could be mapped to both genomes are shown in purple. Arrows indicate sequence orientations with regard to the human and SARS-CoV-2 genomes. **E)** Alignment of the Nanopore read in **D**) with the human genome (chromosome 1) showing the integration site. The light blue highlighted region is enlarged to show a LINE1 endonuclease recognition sequence (underlined). **F)** Alignment of the Nanopore read **D**) with the SARS-CoV-2 genome showing the integrated viral sequence. The light blue highlighted region is enlarged to show the 3' end of the viral sequence at the junction with human sequence.

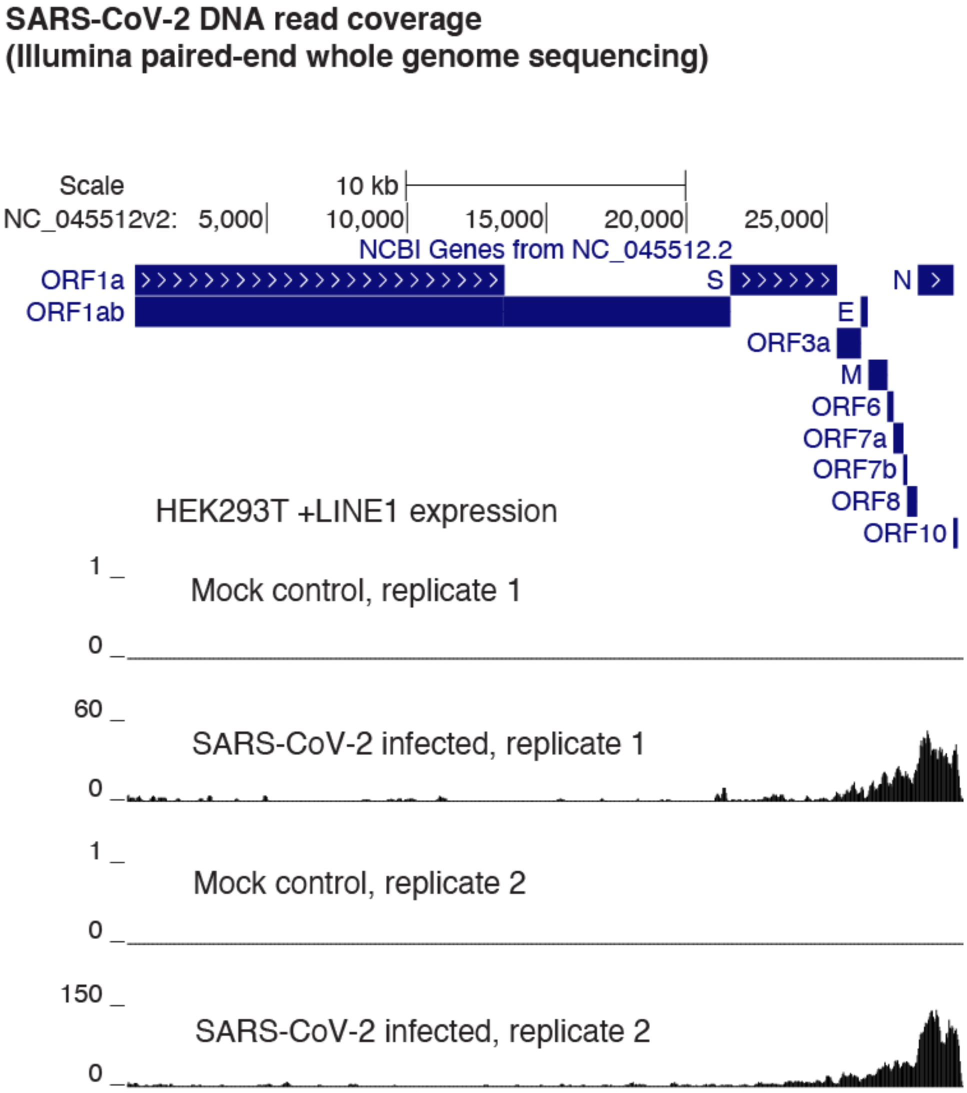


Fig. S3. Reverse transcribed viral sequences are predominant from the 3' end of SARS-CoV-2 genome. Viral reads were obtained using Illumina whole-genome paired-end sequencing of DNA from HEK293T cells that overexpressed LINE1. Genomic tracks showing the number of viral reads binned at 10 bp.

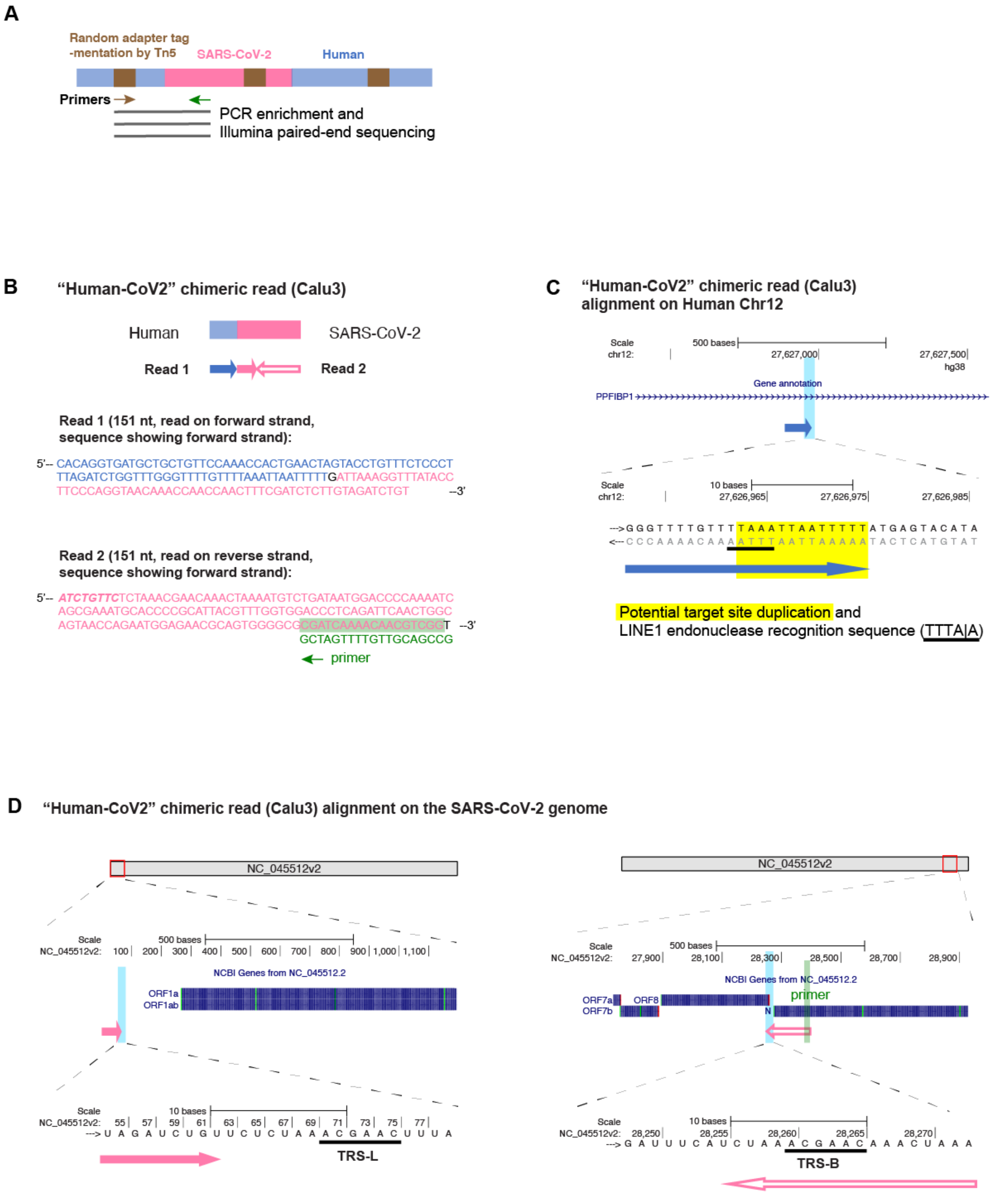
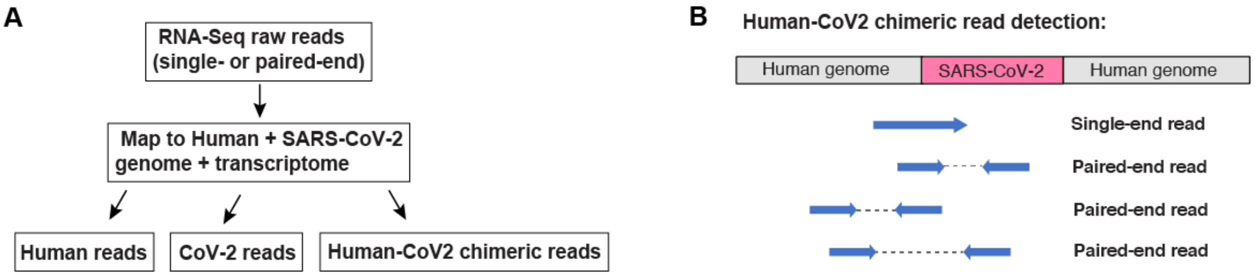

A
Random adapter tag
-mentation by Tn5
SARS-CoV-2
Human
Primers
PCR enrichment and
Illumina paired-end sequencing
B
"Human-CoV2" chimeric read (Calu3)
Human
SARS-CoV-2
Read 1
Read 2
Read 1 (151 nt, read on forward strand,
sequence showing forward strand):
5'-- CACAGGTGATGCTGCTGTTCCAAACCACTGAACTAGTACCTGTTTCTCCCT
TTAGATCTGGTTTGGGTTTTGTTTTAAATTAATTTTTGATTAAAGGTTTATACC
TTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGT --3'
Read 2 (151 nt, read on reverse strand,
sequence showing forward strand):
5'-- ATCTGTTCTCTAAACGAACAAACTAAAATGTCTGATAATGGACCCCAAAATC
AGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGC
AGTAACCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGT --3'
GCTAGTTTTGTTGCAGCCG
primer
C
"Human-CoV2" chimeric read (Calu3)
alignment on Human Chr12
Scale
chr12:
500 bases
27,627,000
27,627,500
hg38
Gene annotation
PPFIBP1
Scale
chr12:
10 bases
27,626,965
27,626,975
27,626,985
--->G G G T T T T G T T T T A A A T T A A T T T T T A T G A G T A C A T A
<---C C C A A A A C A A A A T T T A A T T A A A A A T A C T C A T G T A T
Potential target site duplication and
LINE1 endonuclease recognition sequence (TTTA|A)
D
"Human-CoV2" chimeric read (Calu3) alignment on the SARS-CoV-2 genome
NC_045512v2
Scale
NC_045512v2:
500 bases
100
200
300
400
500
600
700
800
900
1,000
1,100
NCBI Genes from NC_045512.2
ORF1a
ORF1ab
Scale
NC_045512v2:
10 bases
55
57
59
61
63
65
67
69
71
73
75
77
---> U A G A U C U G U U C U C U A A A C G A A C U U U A
TRS-L
NC_045512v2
Scale
NC_045512v2:
500 bases
27,900
28,100
28,300
28,500
28,700
28,900
NCBI Genes from NC_045512.2
primer
ORF7a
ORF8
ORF7b
N
Scale
NC_045512v2:
10 bases
28,250
28,255
28,260
28,265
28,270
---> G A U U U C A U C U A A A C G A A C A A A C U A A A
TRS-B

Fig. S4. Evidence for integration of SARS-CoV-2 cDNA in Calu3 cells. A) Experimental design for the Tn5 tagmentation mediated enrichment sequencing method used to map

integration sites in the host cell genome. The viral primer (reverse) was designed to target near-5' end of the viral NC gene (green arrow). **B)** A human-viral chimeric read pair supporting viral integration. The reads are aligned with the human (blue) and SARS-CoV-2 (magenta) genomic sequences. Arrows indicate read orientations relative to the human and SARS-CoV-2 genomes as shown in **C, D**). Closed arrows show read 1 in the pair that was mapped to both human (blue) and SARS-CoV-2 (magenta) sequences. The open arrow (magenta) shows read 2 in the pair that was mapped to the SARS-CoV-2 genome. The sequence corresponding to the viral primer in read 2 is shown with green highlight (corresponding to the green arrow illustrated in **A**). **C)** Alignment of the read pair in **B**) with the human genome (chromosome 12, blue arrow). The highlighted (light blue) region of the human sequence is enlarged to show the LINE1 recognition sequence (underlined) and the potential target site duplication (highlighted in yellow) that would be generated by LINE1 mediated retroposition. **D)** Alignment of the read pair in **B**) with the SARS-CoV-2 genome (magenta arrows). The closed arrow (left) corresponds to read 1 in **B**), aligned to the viral leader sequence. The open arrow (right) corresponds to read 2 in **B**), aligned to the NC gene body (the beginning 8 bases in this read is aligned to the viral leader sequence, shown in italics in **B**). The highlighted (light blue) regions of the SARS-CoV-2 sequences are enlarged to show the TRS-L (left) and TRS-B (right) sequences (underlined, these are the sequences where the viral polymerase jumps to generate the sub-genomic RNA). The viral primer sequence is shown with green highlight.


A
RNA-Seq raw reads (single- or paired-end)
Map to Human + SARS-CoV-2 genome + transcriptome
Human reads
CoV-2 reads
Human-CoV2 chimeric reads
B
Human-CoV2 chimeric read detection:
Human genome
SARS-CoV-2
Human genome
Single-end read
Paired-end read
Paired-end read
Paired-end read

C

Data ID (GEO)	Sample	SARS-CoV-2 infection	RNA-Seq read length (nt)	Min overhang (threshold)
GSE147507	Calu3, A549-ACE2, NHBE	MOI 2, 24 hrs	150	50
GSE153277	iPSC-derived type II alveolar epithelial-like cells	MOI 5, 4 days	76+76	50
GSE156754	iPSC-derived cardiomyocytes	MOI 0.1, 48 hrs	76+76	50
GSE157852	iPSC-derived choroid plexus organoids	MOI 0.1-0.05, 72 hrs	72	20
GSE153684	Cultured organoid from pediatric stomach tissue (age 11 year)	MOI 1	51+51	20
GSE154998	Blood buffy coat cells	COVID19+ patient (ICU)	72	20

Fig. S5. Analysis of published data for human-viral chimeric transcripts. A) The pipeline used to identify human-CoV2 chimeric RNA-Seq reads. **B)** Schema showing human- SARS-CoV2 chimeric RNA-seq reads mapped to potential SARS-CoV-2 integration sites. **C)** Published data used to identify human-viral chimeric reads: Data ID (GEO accession number), sample type, infection method/type (MOI: Multiplicity of Infection), RNA-Seq format (single or paired-end with read length), and threshold to call chimeric reads (Min overhang: minimum number of bases mapped to either human or SARS-CoV-2 genome/transcriptome to call a chimeric reads).

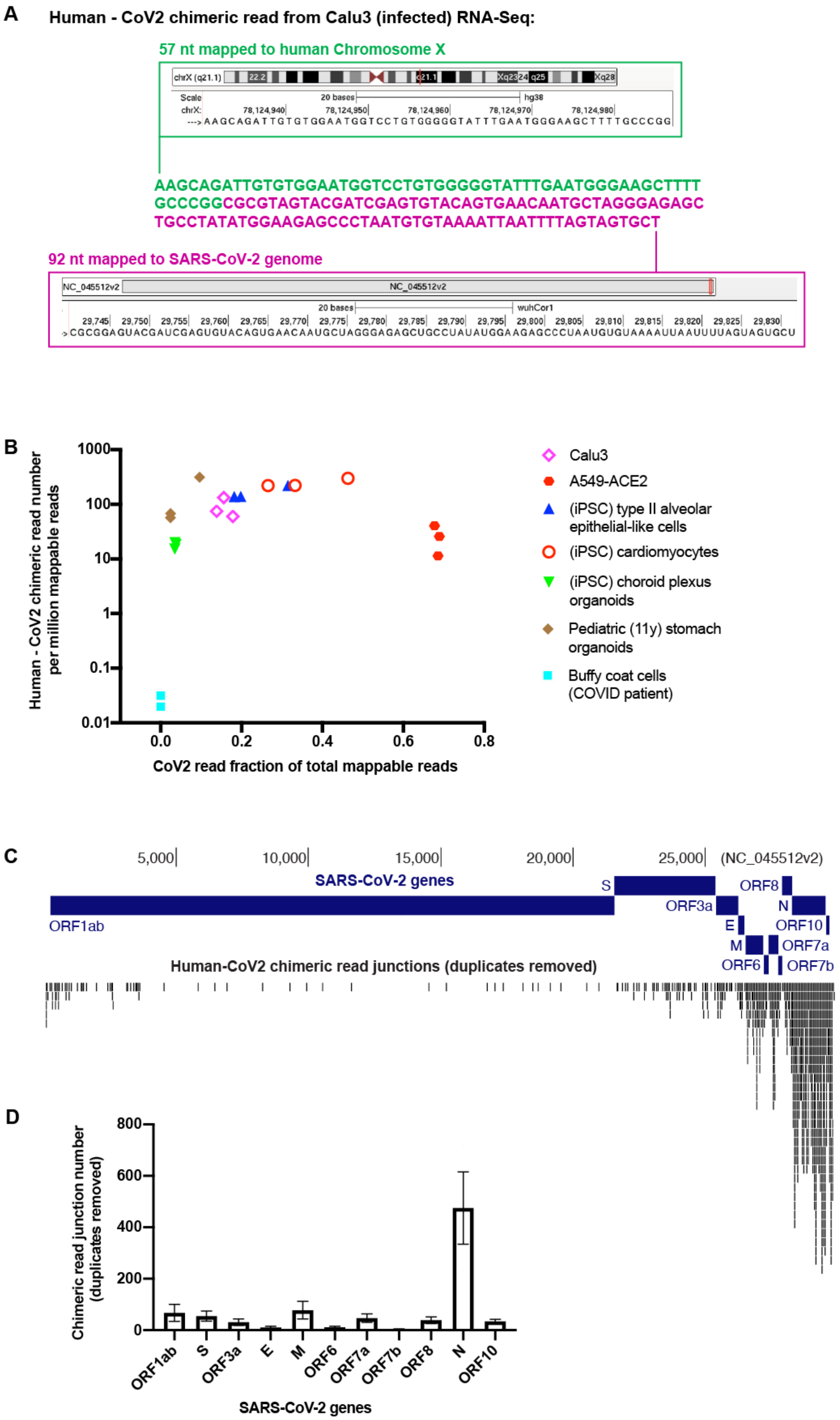

A
Human - CoV2 chimeric read from Calu3 (infected) RNA-Seq:
57 nt mapped to human Chromosome X
chrX (q21.1)
22.2
q21.1
Xq23
24
q25
Xq28
Scale
20 bases
hg38
chrX:
78,124,940
78,124,950
78,124,960
78,124,970
78,124,980
---> AAGCAGATTGTGTGGAATGGTCCTGTGGGGGTATTTGAATGGGAAGCTTTTGCCCGG
AAGCAGATTGTGTGGAATGGTCCTGTGGGGGTATTTGAATGGGAAGCTTTT
GCCCGGCGCGTAGTACGATCGAGTGTACAGTGAACAATGCTAGGGAGAGC
TGCCTATATGGAAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCT
92 nt mapped to SARS-CoV-2 genome
NC_045512v2
NC_045512v2
20 bases
wuhCor1
29,745
29,750
29,755
29,760
29,765
29,770
29,775
29,780
29,785
29,790
29,795
29,800
29,805
29,810
29,815
29,820
29,825
29,830
-> CGCGGAGUACGAUCGAGUGUACAGUGAACAAUGCUAGGGAGAGCUGCCUAUAUGGAAGAGCCCUAAUGUGUAAAAUUAAUUUUAGUAGUGCU
B
Human - CoV2 chimeric read number per million mappable reads
1000
100
10
1
0.1
0.01
0.0
0.2
0.4
0.6
0.8
CoV2 read fraction of total mappable reads
◇ Calu3
⬢ A549-ACE2
▲ (iPSC) type II alveolar epithelial-like cells
○ (iPSC) cardiomyocytes
▼ (iPSC) choroid plexus organoids
◆ Pediatric (11y) stomach organoids
■ Buffy coat cells (COVID patient)
C
5,000
10,000
15,000
20,000
25,000
(NC_045512v2)
SARS-CoV-2 genes
S
ORF8
ORF3a
N
ORF1ab
E
ORF10
M
ORF7a
ORF6
ORF7b
Human-CoV2 chimeric read junctions (duplicates removed)
D
Chimeric read junction number (duplicates removed)
800
600
400
200
0
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
SARS-CoV-2 genes

Fig. S6. Human-viral chimeric reads from published RNA-seq data. A) A chimeric RNA read (149 nt) from (SARS-CoV-2) infected Calu3 RNA-Seq with 57 nt mapped to human Chromosome X (green) and 92 nt (magenta) mapped to the SARS-CoV-2 genome. **B)** Scatter plot showing the number of human-CoV2 chimeric reads (per million total mappable reads, y-axis) versus the fraction of SARS-CoV-2 reads in total mappable reads (x-axis) in published RNA-Seq datasets from different SARS-CoV-2 infected samples. **C-D)** Human-CoV2 chimeric read junctions (duplicates removed) mapped to the SARS-CoV-2 genome (**C**) and distribution among SARS-CoV-2 genes (**D**, three biological replicates; mean ± s.e.m.). RNA-Seq data is from SARS-CoV-2 infected Calu3 cells (GSE147507). A chimeric read junction is defined by the "break point" of the sequences mapped to human or SARS-CoV-2 genome/transcriptome in a given RNA-Seq read.

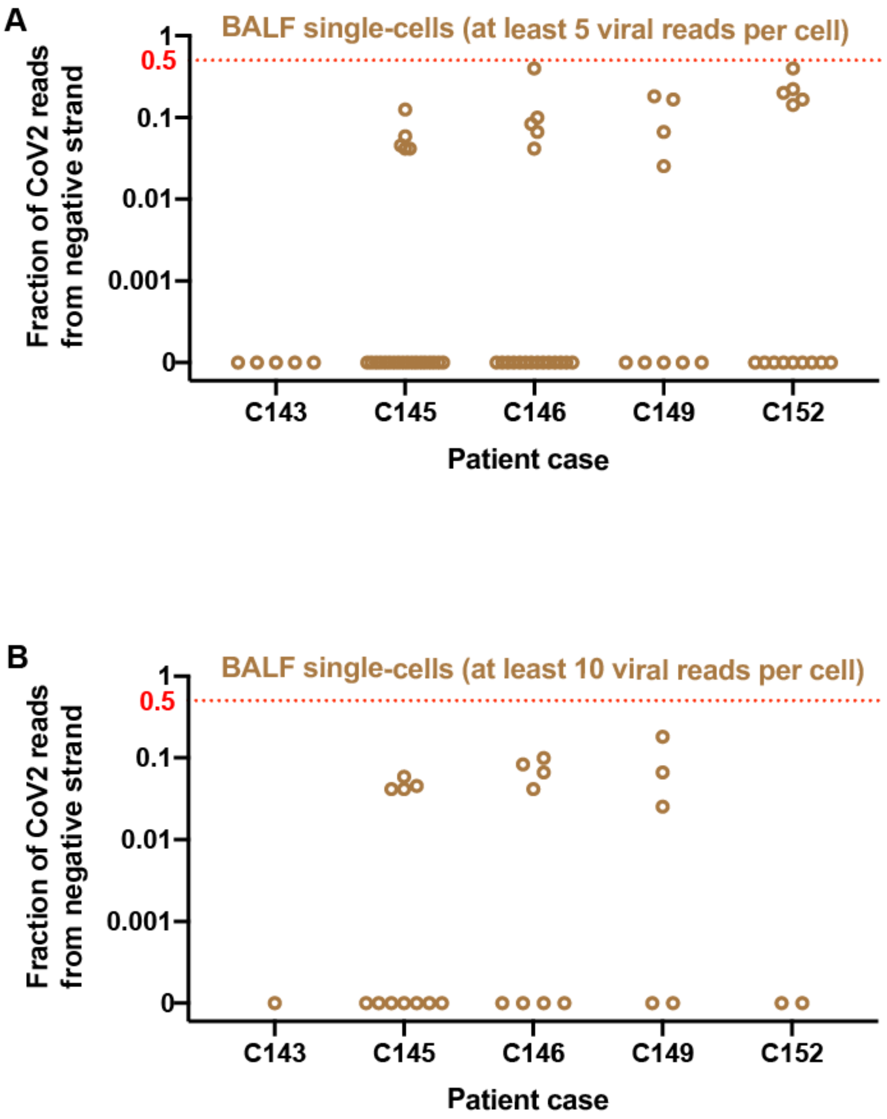


Fig. S7. Negative strand viral RNA-seq reads from patient single-cells suggest that integrated SARS-CoV-2 sequences are expressed. A) Fraction of viral reads that are derived from negative strand viral RNA in single BALF cells from patients with at least 5 viral reads per cell (published RNA-seq data, GSE145926). **B**) Fraction of viral reads that are derived from negative strand viral RNA in patient single BALF cells with at least 10 viral reads per cell (published RNA-seq data, GSE145926). Red dashed lines indicate the level at which 50% of all viral reads were from negative strand viral RNAs, a level expected if all the viral sequences were derived from integrated sequences.

Supplementary Tables

Table S1. Summary of negative strand viral and human-viral chimeric RNA-seq reads from acutely infected lung cells or organoids

Sample	CoV2 reads	Negative strand CoV2 read fraction	Human-CoV2 chimeric reads	Negative strand CoV2 (in chimeric RNAs) read fraction
Calu3, rep1	52,962,587	0.08%	8,170	0.81%
Calu3, rep2	61,256,542	0.10%	6,218	0.76%
Calu3, rep1 (Blanco-Melo et al.)	32,10,542	0.01%	4,430	0
Calu3, rep2 (Blanco-Melo et al.)	2,378,641	0.01%	1,859	0
Calu3, rep3 (Blanco-Melo et al.)	4,320,681	0.01%	9,702	0.01%
Lung organoid, rep1 (Han et al.)	615	0	1	0
Lung organoid, rep2 (Han et al.)	12,752	0.04%	15	0
Lung organoid, rep3 (Han et al.)	1,320	0.08%	2	0

Table S2. Summary of negative strand viral and human-viral chimeric RNA-seq reads from tissues of deceased COVID-19 patients (published RNA-seq data, Desai et al., GSE150316)

Sample	GEO accession number	CoV2 reads	Negative strand CoV2 read fraction	Human-CoV2 chimeric reads	Negative strand CoV2 (in chimeric RNAs) read fraction
Case1-lung1 LUL	GSM4546576	51,4418	6.7%	108	1.9%
Case1-lung2 RML	GSM4546577	54,598	8.4%	9	0.0%
Case1-lung3 RUL	GSM4546578	20,746	13.0%	6	0.0%
Case1-lung4 LLL	GSM4546579	37,232	2.4%	4	0.0%
Case2-lung1 RLL	GSM4546581	483	12.4%	0	
Case2-lung2 LUL	GSM4546582	42	0.0%	0	
Case2-jejunum1	GSM4546583	10	0.0%	0	
Case2-lung3 RUL	GSM4546584	10	0.0%	0	
Case3-lung1 LUL	GSM4546586	16	25.0%	0	
Case3-lung2 RLL	GSM4546588	4	0.0%	0	
Case5-lung1 LLL	GSM4546596	220	0.0%	0	
Case5-lung2 RML	GSM4546597	38	47.4%	0	
Case5-lung3 LUL	GSM4546598	648	0.9%	0	
Case5-lung4 RML	GSM4546599	514	0.4%	0	
Case5-lung5 RUL	GSM4546601	722	4.4%	0	
Case5-liver1	GSM4546604	6	0.0%	0	
Case6-lung1 LUL	GSM4698531	14	0.0%	0	
Case7-lung5 LUL	GSM4698540	1,284	0.6%	0	
Case8- bowel1	GSM4698541	24	0.0%	0	
Case8- heart1	GSM4698542	78	0.0%	0	
Case8-lung1 RLL	GSM4698544	3,150	5.0%	0	
Case8-lung2 RUL	GSM4698545	200	51.0%	0	
Case8-lung3 LLL	GSM4698546	24,527	1.8%	46	0.0%
Case8-lung4 RML	GSM4698547	5,820	2.1%	0	
Case8-lung5 LUL	GSM4698548	102	0.0%	0	
Case9- lung1RLL	GSM4698549	45,539	6.4%	56	5.4%
Case9- lung2 RML	GSM4698550	154,157	9.6%	405	42.5%
Case9- lung3RUL	GSM4698551	138,578	8.1%	173	3.5%
Case9-lung4 LUL	GSM4698552	361,535	4.8%	652	5.8%
Case9- lung5LLL	GSM4698553	179,729	14.5%	145	35.2%
Case10- lung2 LLL	GSM4698522	112	0.0%	0	
Case10- lung3 RLL	GSM4698523	22	0.0%	0	

Case11 -bowel1	GSM4698524	92	0.0%	0	
Case11-lung1RML	GSM4698526	72,029	3.2%	141	2.8%
Case11-lung32RUL	GSM4698527	17,256	12.7%	13	0.0%
Case11-lung3RLL	GSM4698528	1,328	3.5%	0	
CaseA-lung	GSM4698554	1,202	4.0%	0	
CaseB-lung	GSM4698555	6	33.3%	0	
CaseC-lung	GSM4698556	168,935	1.5%	47	0.0%
CaseD-lung	GSM4698557	43,750	6.8%	9	0.0%

Table S3. Summary of negative strand viral and human-viral chimeric RNA-seq reads from BALF cells of COVID-19 patients (bulk analysis of published single-cell RNA-seq data, Liao et al., GSE145926)

Patient	GEO accession number	CoV2 reads	Negative strand CoV2 read fraction	Human-CoV2 chimeric reads	Negative strand CoV2 (in chimeric RNAs) read fraction
C143 (severe)	GSM4339771	1,525	18.6%	1	0.0%
C145 (severe)	GSM4339773	34,439	12.7%	16	0.0%
C146 (severe)	GSM4339774	893,327	15.3%	223	1.4%
C148 (severe)	GSM4475051	1,251	22.0%	0	
C149 (severe)	GSM4475052	89,928	19.6%	23	0.0%
C152 (severe)	GSM4475053	32,665	18.5%	5	0.0%

Table S4. PCR primers used in this study

Name	Sequences
N1	Forward: GACCCCAAAATCAGCGAAAT Reverse: TCTGGTTACTGCCAGTTGAATCTG
N2	Forward: GGGAGCCTTGAATACACCAAAA Reverse: TGTAGCACGATTGCAGCATTG
N3	Forward: GGGGAACTTCTCCTGCTAGAAT Reverse: CAGACATTTTGCTCTCAAGCTG
N4	Forward: AAATTTTGGGGACCAGGAAC Reverse: TGGCACCTGTGTAGGTCAAC
N (for cloning complete NC gene)	Forward: ATGTCTGATAATGGACCCCAAAAT Reverse: TTAGGCCTGAGTTGAGTCAGC
HSPA1A	Forward: ATCTCCACCTTGCCGTGTT Reverse: ATCCAGTGTTCCGTTTCCAG

Dataset S1 (separate file). Sanger sequencing results of the complete SARS-CoV-2 NC gene cloned from large-fragment cell genomic DNA from LINE1-overexpressing, SARS-CoV-2-infected HEK293T cells.

Dataset S2 (separate file). Summary of chimeric read sequences from Nanopore sequencing of DNA from LINE1-overexpressing, SARS-CoV-2-infected HEK293T cells.

Dataset S3 (separate file). Summary of chimeric sequences from Illumina paired-end whole genome sequencing of DNA from LINE1-overexpressing, SARS-CoV-2-infected HEK293T cells.

Dataset S4 (separate file). Summary of chimeric sequences from Tn5 tagmentation-mediated DNA integration site enrichment sequencing of DNA from SARS-CoV-2 infected HEK293T or Calu3 cells.