

1 **Oligonucleotide Capture Sequencing of the SARS-CoV-2 Genome and**
2 **Subgenomic Fragments from COVID-19 Individuals**

3
4 Harsha Doddapaneni^{1*}, Sara Javornik Cregeen², Richard Sucgang², Qingchang Meng¹, Xiang
5 Qin¹, Vasanthi Avadhanula³, Hsu Chao¹, Vipin Menon¹, Erin Nicholson^{3,4}, David Henke³, Felipe-
6 Andres Piedra³, Anubama Rajan³, Zeineen Momin¹, Kavya Kottapalli¹, Kristi L. Hoffman², Fritz J.
7 Sedlazeck¹, Ginger Metcalf¹, Pedro A. Piedra^{3,4}, Donna M. Muzny¹, Joseph F. Petrosino²,
8 Richard A. Gibbs^{1*}

9
10 Corresponding authors*: doddapan@bcm.edu (HD) and agibbs@bcm.edu (RAG)

11 ¹Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United
12 States of America; ²Alkek Center for Metagenomics and Microbiome Research, Department of
13 Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, United
14 States of America; ³Department of Molecular Virology and Microbiology, and ⁴Pediatrics, Baylor
15 College of Medicine, Houston, Texas, United States of America, USA.

16
17
18
19
20
21
22
23

24 **Abstract**

25 The newly emerged and rapidly spreading SARS-CoV-2 causes coronavirus disease 2019
26 (COVID-19). To facilitate a deeper understanding of the viral biology we developed a capture
27 sequencing methodology to generate SARS-CoV-2 genomic and transcriptome sequences from
28 infected patients. We utilized an oligonucleotide probe-set representing the full-length genome
29 to obtain both genomic and transcriptome (subgenomic open reading frames [ORFs])
30 sequences from 45 SARS-CoV-2 clinical samples with varying viral titers. For samples with
31 higher viral loads (cycle threshold value under 33, based on the CDC qPCR assay) complete
32 genomes were generated. Analysis of junction reads revealed regions of differential
33 transcriptional activity and provided evidence of expression of ORF10. Heterogeneous allelic
34 frequencies along the 20kb ORF1ab gene suggested the presence of a defective interfering
35 viral RNA species subpopulation in one sample. The associated workflow is straightforward, and
36 hybridization-based capture offers an effective and scalable approach for sequencing SARS-
37 CoV-2 from patient samples.

38

39 **Introduction**

40 The COVID-19 pandemic has spread worldwide with alarming speed and has led to the worst
41 healthcare crisis in a century. The agent of COVID-19, the novel SARS-CoV-2 coronavirus
42 (family *Coronaviridae*), has a ~30 Kb positive-sense single-stranded RNA genome predicted to
43 encode ten open reading frames (ORFs) [1]. Similar to other RNA viruses, coronaviruses
44 undergo mutation and recombination [2, 3] that may be critical to understanding physiological
45 responses and disease sequelae, prompting the need for comprehensive characterization of
46 multiple and varied viral isolates.

47

48 To date, reports highlighting genomic variation of SARS-CoV-2 have primarily used amplicon-
49 based sequencing approaches (e.g., ARTIC) [4-7]. Attaining uniform target coverage is difficult
50 for amplicon-based methods, and is exacerbated by issues of poor sample quality [8]. Genome
51 variation in the amplicon primer region may also impact sequence assembly. Transcriptome
52 characterization can further contribute to our knowledge of mutation within the SARS-CoV-2
53 genome, and direct RNA long read sequencing, both alone and in combination with short read
54 sequencing, have been described [1, 9, 10]. Unfortunately, these analyses are equally
55 hampered by sample quality limitations and necessitate use of cultured cell lines.

56 Oligonucleotide capture ('capture') mitigates these issues as hybridization to specific probes not
57 only enriches for target sequences but enables the analysis of degraded source material [11-
58 14]. Capture enrichment has also been applied to viral sequencing, where a panvirome probe
59 design resulted in up to 10,000-fold enrichment of the target sequence and flanking regions [15-
60 17]. Direct RNA enrichment method has also been reported for viral genome sequencing, but
61 each sample was enriched separately followed by pooling for sequencing [18].

62 Hybridization-based enrichment of RNAs can also aid in the identification of gene fusions or
63 splice variants [13, 19, 20], which are particularly important for coronavirus biology. In addition
64 to encoding a polyprotein that undergoes autocatalyzed hydrolysis, coronaviruses employ
65 subgenomic RNA fragments generated by discontinuous transcription to translate proteins
66 required for viral replication and encapsidation. These subgenomic RNA fragments share a
67 common 62-bp leader sequence derived from the 5' end of the viral genome, detectable as a
68 fused junction to interior ORFs [1, 10]. Direct RNA sequencing of cultured cell lines infected with
69 SARS-CoV-2 revealed that the junctional sequences are not evenly distributed between the
70 ORFs, suggesting that individual proteins may be translated at different rates [1]. How virus
71 translation profiles from infected human patients differ from those from cultured cells is as yet
72 unknown.

73

74 Here we have utilized capture probes and a streamlined workflow for sequence analysis of both
75 the SARS-CoV-2 genomic sequences and of the junction reads contained within the genomic
76 subfragments generated by discontinuous transcription (Fig1). The method can be applied at
77 scale to analyze samples from clinical isolates. Enriching for genomic and transcriptional RNA,
78 followed by deep short-read sequencing, sheds light on variation in clinical SARS-CoV-2
79 genomic sequences and expression profiles.

80

81 **Fig 1. Schematic workflow.** Presented in the workflow are the different steps involved in the
82 SARS-CoV-2 capture and sequencing methodology.

83

84

85 **Material and methods**

86 **COVID-19 viral testing, Collection, RNA extraction and real-time reverse transcription**
87 **polymerase chain reaction (RT- PCR).** The CLIA Certified Respiratory Virus Diagnostic
88 Laboratory (ID#: 45D0919666) at Baylor College of Medicine performed real time reverse
89 transcriptase polymerase chain reaction (RT-PCR) tests for SARS-CoV-2 on mid-turbinate
90 nasal swab samples collected from adults presenting to the hospitals or clinics at the Texas
91 Medical Center from March 18 through April 25, 2020. Viral RNA was extracted from nasal
92 swab samples using the Qiagen Viral RNA Mini Kit (QIAGEN Sciences, Maryland, USA) with an
93 automated extraction platform QIAcube (QIAGEN, Hilden, Germany) according to the
94 manufacturer instructions. Starting with 140 ul of the collected sample, nucleic acids were
95 extracted and eluted to 100ul. All samples were tested by CDC 2019- Novel coronavirus (2019-
96 nCoV) Real-Time RT-PCR Diagnostic panel. Primers and probes targeting the SARS-CoV-2
97 nucleocapsid genes, N1 and N2, were used. Samples were also tested for Ribonuclease P
98 (RNase P) gene, to determine the quality of sample obtained. PCR reaction was set up using

99 TaqPath™ 1-Step RT-qPCR Master Mix, CG (Applied Biosystems, CA) and run on 7500 Fast
100 Dx Real-Time PCR Instrument with SDS 1.4 software. Samples with cycle threshold (Ct) values
101 below 40 for both SARS-CoV-2 N1 and N2 primers were necessary to determine positivity. For
102 seven samples with very low viral loads (N=7); Ct >37 and <40, the RNA was concentrated 4-
103 fold by doubling the extraction volume - 280 µl and halving the elution volume - (50 µl) and
104 submitted for sequencing.

105

106 **Library, capture, sequencing**

107 **Sequenced samples.** Forty-five mid-turbinate nasal swab samples were collected from 32
108 unique individuals (S1 Table). The RNA Integrity Number (RIN) values ranged from 2.3 and 5.2
109 with Ct values from 16-39. The amount of RNA used as input for cDNA varied from 13.6 ng to
110 120 ng (S1 Table). As positive controls, 1,500 (Ct=36.2) and 150,000 copies (Ct=29.6) of the
111 Synthetic SARS-CoV-2 RNA from Twist Biosciences (Cat# 102024) were spiked into two 50 ng
112 Universal Human Reference RNA samples. To generate the synthetic RNA, six non-
113 overlapping 5 Kb fragments of the SARS-CoV-2 reference genome (MN908947.3) sequence
114 were synthesized by Twist Inc. as double stranded DNA, and transcribed in vitro into RNA.
115 Three SARS-CoV-2 free mid-turbinate nasal swab samples which were negative for SARS-
116 CoV-2 by real-time RT-PCR, were sequenced as negative controls. Due to limited sample size
117 in this study, no other patient metadata was used to interpret results.

118

119 **cDNA Preparation.** cDNA was generated utilizing NEBNext® RNA First Strand Synthesis
120 Module (E7525L; New England Biolabs Inc.) and NEBNext® Ultra™ II Directional RNA Second
121 Strand Synthesis Module (E7550L; NEB). Total RNA in a 15 µl mixture containing random
122 primers and 2X 1st strand cDNA synthesis buffer were incubated at 94°C for 10 min to fragment
123 the RNA. RNA were converted to cDNA by adding a 5-µl enzyme mix containing 500ng
124 Actinomycin D (A7592, Thermo Fisher Scientific), 0.5µl RNase inhibitor, and 1 µl of Protoscript

125 II reverse transcriptase, then incubated at 25°C for 10 minutes, 42°C for 50 minutes, 70°C 15
126 minutes, before being cooled to 4°C on a thermocycler. Second strand cDNA were synthesized
127 by adding a 60 µl of mix containing 48 µl H₂O, 8 µl of 10X reaction buffer, and 4 µl of 2nd strand
128 synthesis enzyme, and incubated at 16°C for 1 hour on a thermocycler. The double strand (ds)
129 cDNA were purified with 1.8X volume of AMPure XP Beads (A63882, Beckman) and eluted into
130 42 µl 10 mM Tris buffer (Cat#A33566, Thermo Fisher Scientific). Because these libraries were
131 prepared primarily for sequence capture, rRNA depletion or Ploy A+ isolation steps were not
132 performed.

133

134 **Library preparation.** The double-stranded cDNA was blunt-ended using NEBNext[®] End Repair
135 Module (E6050L, NEB). Five µl 10X ER reaction buffer and 5 µl ER enzyme were added to the
136 ds cDNA. The ER reaction was incubated for 30 minutes at 20°C on a thermocycler. After the
137 ER reaction, cDNA were purified with 1.8X volume AMPure XP Beads and eluted into 42 µl
138 nuclease free water (129114, Qiagen). Next, 5 µl of 10X AT buffer and 3 µl of Klenow enzyme
139 from NEBNext[®] dA-Tailing Module (E6053L, NEB) was added to the sample. The AT reaction
140 was incubated at 37°C for 30 minutes. After incubation, samples were purified with 1.8X volume
141 AMPure XP Beads and eluted into 33 µl nuclease free water (129114, Qiagen). Illumina unique
142 dual barcodes adapters (Cat# 20022370) were ligated onto samples by adding 2 µl of 5uM
143 adapter, 10 µl 5X ligation buffer and 5 µl of Expresslink Ligase (A13726101, Thermo Fisher),
144 and incubated at 20°C for 15 minutes. After adapter ligation, libraries were purified twice with
145 1.4X AMPure XP and eluted into 20 µl H₂O. Libraries were amplified in 50 µl reactions
146 containing 150 pmol of P1.1 and P3 primer and Kapa HiFi HotStart Library Amplification kit
147 (Cat# kk2612, Roche Sequencing and Life Science). The amplification was incubated at 95°C
148 for 45 seconds, followed by 15 cycles of 95°C for 15 sec, 60°C 30 seconds, and 72°C 1 minutes,
149 and 1 cycle at 72°C for 5 minutes. The amplified libraries were purified with 1.4X AMPure XP
150 Beads and eluted into 50 µl H₂O. The libraries were quality controlled on Fragment Analyzer

151 [using DNA7500 kit (5067-1506, Agilent Technologies). The library yields were determined
152 based on 200-800-bp range.

153

154 **Capture enrichment and sequencing.** cDNA libraries with Illumina adaptors constructed from
155 SARS-CoV-2 positive individuals were pooled into six groups (S1 Table). Pools 1 and 2 were
156 from batch 1 and pools 3-6 are from batch 2. The RT-qPCR Ct value of virus N gene varied in
157 these pools as follows: Pool 1 with 6 samples (Ct 20.4 - 28.34); Pool 2 with 5 samples (Ct 29.75
158 - 37.95; Pool 3 with 5 samples (Ct 17.3 – 38); Pool 4 with 6 samples (Ct 27.8 - 39.3); Pool 5 with
159 11 samples (Ct 33 - 38.9) and Pool 6 with 12 samples (Ct 32.9 - 39.5). Pooled cDNA pre-
160 capture libraries were hybridized with probes from the SARS-CoV-2 Panel (Twist, Inc) at 70°C
161 for 16 hours. Total probe length is 120 Kb. Post-capture insert molecules were further amplified
162 (12-16 cycles) to obtain the final libraries that were sequenced on Illumina NovaSeq S4 flow
163 cell, to generate 2x150 bp paired-end reads. To evaluate the effect of hybridization-based
164 enrichment 9 samples were sequenced before and after capture enrichment. Ribosomal RNA
165 was removed computationally.

166

167 **Data analysis**

168 **Sequence Mapping, genome reconstruction and variant calling:** Raw fastq sequences were
169 processed using BBDuk (<https://sourceforge.net/projects/bbmap/> ; BBDuk version 38.82) to
170 quality trim, remove Illumina adaptors and filter PhiX reads. Trimming parameters were set to a
171 k-mer length of 19 and a minimum Phred quality score of 25. Reads with a minimum average
172 Phred quality score below 23 and length shorter than 50 bp after trimming were discarded. The
173 trimmed fastqs were mapped to a combined PhiX (standard Illumina spike in) and human
174 reference genome (GRCh38.p13; GCF_000001405.39) database using a two-step BBTools
175 approach (BBMap version 38.82). Briefly, the trimmed reads were first processed through the
176 bloomfilter script, with a strict k=31 to remove reads identified as human. The remaining reads

177 were mapped to the reference genome with BMap using a k-mer length of 15, the bloom filter
178 enabled, and fast search settings in order to determine and remove hg38/PhiX reads. Trimmed
179 and human-filtered reads were then processed through VirMAP [21] to obtain full length
180 reconstruction of the SARS-CoV-2 genomes. SPAdes assembler [22] was also used for genome
181 reconstruction. The resulting assemblies were compared to those from VirMAP. A
182 reconstructed genome with >99% the length of the SARS-CoV-2 reference genome,
183 NC_045512.2, was considered a fully reconstructed genome. Plots were generated using R
184 (version 3.6.1) and the tidyverse (version 1.3.0) and ggplot2 (version 3.2.1) packages.
185 Alignments and reference mapping were done using mafft [23] (version 1.4.0) and BMap
186 (version 38.82). Sequence variation compared to SARS-CoV-2 reference genome was
187 performed using the genome alignment from mafft with in-house scripts. For heterozygous
188 variant analysis, the sequence reads were aligned to the reference genome using BWA-mem
189 [24] with default parameters, realigned using GATK [25], and variants were called using Atlas-
190 SNP2 [26]. Variant annotation was performed with SnpEff [27] Lineage assignment of SARS-
191 COV-2 following Rambaut et al (2020) used the Pangolin COVID-19 Lineage Assigner
192 (<https://pangolin.cog-uk.io>).

193

194 **Subgenomic mRNA and junction reads analysis:** Illumina sequence reads were aligned to
195 SARS-CoV-2 reference genome NC_045512.2 using STAR aligner v2.7.3a [28] with penalties
196 for non-canonical splicing turned off as described by Kim et al¹. Alignment bam files were
197 parsed using an in-house script to obtain junction-spanning reads that contained the leader
198 sequence (5' end of the junction falls within 34 to 85 bp of the reference genome). Sub genomic
199 RNAs were categorized by junction reads according to the genes of the immediate start codon
200 downstream of the 3' of the junctions. Junction read counts were normalized to the total number
201 of mapped reads.

202

203 **Results**

204 A total of 45 samples collected from 32 patients between March 18 and April 25, 2020 in
205 Houston, Tx, USA were analyzed. These were a subset of individuals tested for the presence of
206 SARS-CoV-2 early during the pandemic. RNA fractions were isolated from viral transport media
207 and converted to cDNA. SARS-CoV-2 cDNA libraries were pooled into six groups (S1 Table). All
208 45 capture-enriched and nine of the pre-capture libraries were sequenced on an Illumina
209 platform based on details provided in the online methods. A schematic workflow is shown in (Fig
210 1).

211

212 **Sequencing results and capture enrichment efficiency**

213 A total of 7.15 billion raw reads were generated for the 45 SARS-CoV-2 positive samples
214 sequenced (S1 Table). Since this study was to optimize the methodology, samples were
215 sequenced deeper to ensure that results among samples were not biased. Sequences
216 were trimmed to filter low quality reads and subsequently mapped to the GRCh38 reference
217 genome to identify human reads (Fig 2A). Trimmed non-human sequence reads were analyzed
218 using the VirMAP [21] pipeline where between 7- 86.4% of total reads from post-capture
219 libraries mapped to the SARS-CoV-2 reference. One sample (192000446B), which had only
220 6.37 ng total RNA starting material, did not generate any SARS-CoV-2 reads. Overall, the
221 percentage of reads represented by SARS-CoV-2 was higher in samples with CDC protocol-
222 based RT-qPCR Ct values <33 (Fig. 2A).

223

224 **Fig 2. Sequence data.** Ct value vs percent raw sequencing reads mapped to SARS-CoV-2 in
225 (a) Capture enriched samples; (b) Pre-capture samples; (c) Positive and negative controls.
226 Percentage of reads mapped to the ‘SARS-CoV-2’ genome, to the ‘human’ reference genome
227 and a third category called the ‘reads others’, which is the combined total of trimmed reads and
228 reads that do not fall under the two other categories are plotted in this figure. CT values in bold
229 indicate samples that provided full-length genome assemblies.

230
231 To estimate the capture enrichment efficiency, pre-capture libraries for nine samples, ranging in
232 Ct values of 20.4 to 37.95 (i.e. high to low titer in the original samples), were also sequenced,
233 generating 152.1 – 322.9 million reads per sample. Samples 192000106B and 192000090B,
234 with Ct > 37 produced zero reads mapping to the SARS-CoV-2 reference genome. In the
235 remaining seven samples, less than 0.022% of reads were deemed SARS-CoV-2 (Fig 2B).
236 Collectively, post-capture enrichment increased the SARS-CoV-2 mapping rate to 50.9%, a
237 9,243-fold enrichment.

238 Spiked synthetic SARS-CoV-2 RNA, encompassing six fragments of 5 Kb each, served as a
239 positive control and were enriched successfully at both 1,500 and 150k copies per sample (Fig
240 2C). In the 1,500 copy libraries (n=2), 3-5% of reads mapped to the SARS-CoV-2 genome,
241 while approximately 65% of reads from the 150k copy libraries (n=2) did the same (S1 Table).
242 This translates to an approximate 91,858-fold enrichment in the 1,500 copy libraries and
243 13,778-fold enrichment in the 150k copy libraries compared to their starting amounts in the
244 RNA. Three SARS-CoV-2 PCR negative samples were also sequenced, where <0.5% of reads
245 mapped to the SARS-CoV-2 reference genome at 3-5 locations that are not conserved in the
246 SARS-CoV-2 genome (S1 Table; S1 Fig).

247

248 **Genome reconstruction and genomic variations**

249 In order to assess the ability of the capture methodology to assemble full-length genomes, both
250 the nine pre-capture and 45 post capture libraries were assembled using both the VirMAP
251 pipeline and the SPAdes *de novo* assembler [22]²⁰.

252
253 Full-length SARS-CoV-2 genomes were obtained from 17 of the 45 capture-enriched samples.
254 Genome coverage in these 17 samples varied from 1071x to 3.19x million (S1 Table).
255 Successful full-length genome assembly was correlated with Ct values below 33 (Fig 3),
256 regardless of the total reads generated during sequencing. No variability between samples due
257 to random priming of the cDNA synthesis or no gaps in genome coverage were noticed
258 using this method (S2 Fig) Two samples with Ct values above 33, 192000296 (Ct 33.9) and
259 192000354 (Ct 35.5), obtained from a single patient, also yielded full-length genome
260 reconstructions with acceptable quality ($N \leq 0.5\%$). Partial genome reconstructions were
261 achieved for the remaining samples although somewhat surprisingly, the correlation between
262 percentage of the genome that was reconstructed and the Ct value of that sample was not
263 tightly correlated when Ct values were above 33 (Fig 3). Full-length genome sizes of the 17
264 capture-enriched and assembled sequences varied from 29.68 Kb to 30.15 Kb (S3 Fig).
265 Variants relative to the SARS-CoV-2 reference genome sequence NC_045512.2, including
266 single nucleotide polymorphisms and a single indel, ranged from 5 to 15 per sample, with a
267 mean of nine.

268
269 **Fig 3. Scatter plot showing genome completeness as a function of Ct value.** Pink circles
270 represent post-capture samples and black asterisks represent pre-capture samples.

271

272

273 Out of the nine pre-capture samples, three (192000072B, 192000021B, 1920000003B), all with
274 Ct values ≤ 27.4 , yielded full-length genomes with 28x – 265x genome coverage, while in the
275 other four samples, genome reconstructions were partial and also had a poor genome coverage
276 of 1-6x. SARS-CoV-2 reads were not detected in the two remaining samples.

277 Alignment of DNA sequence reads from one sample (192000051B) to the reference SARS-
278 CoV-2 genome sequence NC_045512.2 that is based on the first published isolate from Wuhan
279 SARS-CoV-2 reference genome, revealed multiple heterogenous alleles (Fig 4; S4 Fig). Most
280 isolates spreading into Europe derive from the 'B' lineage (based on the Wuhan sequence), but
281 three samples including this sample contained an additional fraction of reads representing the A
282 lineage [29] (S2 Table). Further investigation of the clinical correlates of this observation are
283 underway. The genomic position 23,403 in the Wuhan reference strain had good coverage in 28
284 of the capture enriched samples. The A-to-G nucleotide mutation at this location that results in
285 the Spike protein D614G amino acid change was noticed in 23 of the 28 samples [30].

286

287 **Fig 4. Schematic representation of 192000051B assembly.** Black bars represent loci where
288 the assembly called alleles different from the NCBI reference sequence NC_045512. Green
289 bars represent mixed loci where both reference and alternative alleles were called. All mixed
290 loci are in the ORF1ab gene, and are listed in the table, along with the frequency of the
291 alternate allele at the position, and the predicted effect in translation.

292

293 **Characterization of SARS-CoV-2 subgenomic mRNAs**

294 To identify and quantitate subgenome-length mRNAs, reads were aligned to the SARS-CoV-2
295 reference genome NC_045512.2. Only samples with full-length genomes (N=17 capture and
296 N=5 pre-capture) were analyzed for junction reads to avoid introduction of any bias in identifying
297 subgenomic RNA due to gaps in sequence coverage (Fig 5A and S3 Table). While full-length

298 genomes were reconstructed from three pre-capture samples, an additional two samples with
299 >95% genomes reconstructed, 192000135B (with 97.4%) and 192000088B (95.3%), were also
300 included in this comparison (Fig 5A and in S4 Table). To characterize ORF expression in the
301 capture and pre-capture libraries, the number of junction reads/million were calculated and
302 plotted in Fig 5A (see details in S3 Table). Among the five pre- and post-capture comparison
303 pairs, junction reads were identified in more ORFs after capture, and in instances where
304 junction reads were found before and after capture, the expression trend agreed between the
305 two groups.

306

307 **Fig 5. SARS-CoV-2 subgenomic mRNAs. (a)** Junction read quantification per gene estimated
308 as number of junction reads per million (log transformed) showing values generated from five
309 pre-capture and 17 capture samples. Samples chosen for this analysis have above 95%
310 genome completeness. The coverage level per sample is shown below the gene heatmap.
311 Samples in bold denote same sample sequenced as pre-capture and capture. **(b)** ORF read
312 coverage shown as normalized read counts (RPKM) per gene for 17 capture samples.

313

314 In the capture libraries, junction reads were identified in all 17 samples in the S gene, followed
315 by ORF8 in 16 samples, ORF3a and ORFa in 14 sample samples, N gene in 13 samples, M
316 gene in 11 and ORF6 in 10 samples with remaining ORFs seen in between 3 to 10 samples.
317 Junction reads containing canonical leader sequences were not identified in ORF1ab in any
318 sample, suggesting the translation of ORF1ab from genomic RNA is independent of the
319 canonical leader sequence. The average number of junction reads/million was highest for
320 ORF3a (176.3), followed by ORF8 (104.3) and S gene (10.8). The N gene junction reads/million
321 average was skewed due to its high presence in sample 192000052B. Log transformed values
322 are shown in Fig 5A. For the remaining genes, the average was less than 10 junction

323 reads/million. The expression of ORF10 gene was detected in three of the 17 samples
324 (192000052B, 192000251B, and 192000440B) with expression values of 0.13, 0.13 and 0.02
325 reads/million (S5 Fig). Among the 17 libraries with full-length genomes, there is only one pair
326 192000296B (Ct 33.8) and 192000354B (Ct 35.5), sampled twice from the same subject
327 (Patient #12) and the junction read expression was lower but detectable in both of these
328 samples (S3 Table).

329

330 There were no gaps in the ORF read coverage in any of the 17 capture samples (Fig 5B). From
331 5' to 3' of the genome, there was a gradual increase in the read coverage as expected, for the
332 genomic and subgenomic (transcriptomic) RNA reads. Across the genes in these 17 samples,
333 ORF1ab and ORF3a had the lowest reads per kilobase million (RPKM) values (average 32509
334 and 27957 RPKM, respectively) while the highest values were seen for ORF10 with a count of
335 121,643 (Fig 5B).

336

337 **Discussion**

338 We employed a hybridization-based oligonucleotide capture methodology, combined with short
339 DNA read sequencing, for culture-free genome reconstruction and transcriptome
340 characterization of the SARS-CoV-2 virus. The approach provided complete viral genome
341 sequences and identified sub genomic fragments containing ORFs, shedding light on SARS-
342 CoV-2 transcription in clinical samples. This method uses routine cDNA and library preparation
343 along with Illumina sequencing, employing 96 or more barcodes. Patient samples can be pooled
344 for capture and sequencing, to generate sequence data in large numbers.

345 The capture method provided considerable enrichment of SARS-CoV-2 in all samples tested.

346 The enrichment efficiency was calibrated using two spike-in synthetic SARS-CoV-2 RNA
347 controls in the background of human UHR, and yielded a 91,858-fold enrichment in the 1,500

348 copy (Ct=36.2), libraries and 13,778-fold enrichment in the 150k copy (Ct=29.6) reconstructed
349 samples. For nine patient samples, where sequence data from pre and post capture libraries
350 were compared, a 9,243-fold enrichment was observed. Some human sequences were
351 observed in the data generated from low viral load samples (CT>33) and these were removed *in*
352 *silico*¹⁵, and did not effectively interfere with the enrichment. Some unevenness in SARS-CoV-2
353 sequence representation was initially observed when pooling samples within a range of Ct.
354 values. This was managed by pooling groups of samples based upon their range of CT values
355 before capture enrichment.

356 Full length SARS-CoV-2 genomes were able to be assembled from 17 of the 45 samples
357 analyzed. High quality, full-length reconstructions from capture enrichment appears to be
358 reliably achieved with a viral Ct \leq 33. Between a Ct of 33 and 36, the full-length genome is
359 recovered in some samples while partial genomes, consisting of >50% of the genome length,
360 were reconstructed for the majority (Fig 3). For Sars-Cov-2 genome sequencing, multiplex
361 amplicon sequencing has been used the most to date which includes the primer pools designed
362 by ARTIC consortium (V1 V2 and latest is V3) as well as a third version NIID-1 (Quick J) [31]
363 [4]. ARTIC V1 primer set, worked well for full-length genome recovery with relatively high viral
364 load (Ct < 25) in clinical qPCR tests, as certain primer pairs were under performing. The
365 updated V3 and NIID-1 primer sets addressed this problem and were shown to work well with Ct
366 values in clinical qPCR from 25 to 30 [4]. A multiplex amplicon-based approaches by CDC²³
367 where the effectively generating full length genome sequences \leq Ct of 33 although Ct. values
368 between 30 and 33, genome recovery varied between samples. In another report, ARTIC
369 primers were used initially for amplification of SAR-COV-2 clinical samples and the full-length
370 genome recovery from sequencing these amplicons were compared by different library
371 preparation methods for Illumina sequencing [32]. They reported that samples below Ct. <27
372 produced near full-length genomes, although from samples with Ct. <30, longer and higher
373 quality genomes were reported. In comparison to several of these studies, using the capture

374 enrichment methodology, full-length genomes were obtained consistently from clinical samples
375 up to Ct. 33, which is the ability to enrich 8-fold lower genome equivalents. However, as shown
376 from the data in (Table S1), generating more sequence data for low titer samples does not lead
377 to full-length genome recovery. There is supporting information now based on the success rate
378 of the culture of the Sars-Cov-2 at different Ct. Values, where the probability of culturing virus
379 declines to 8% in samples with Ct > 35 and to 6%, 10 days after symptom onset [33]. Putting
380 this information together with our own observation of partial genome recovery from samples with
381 Ct >33, suggests these individuals may likely be carrying only genomic fragments in them at the
382 time of the sampling.

383

384 Capture enrichment enabled identification of a mixed population of SARS-CoV-2 virus in sample
385 192000051B, including a putative defective interfering viral RNA species that likely is incapable
386 of translating the viral polyprotein encoded in ORF1ab alongside a replication competent strain.
387 All heterogeneously called alleles are in ORF1ab, the 20 kb gene encoding the polyprotein
388 essential to the replication of the viral genome. Only one of these alleles (T20520C) is expected
389 to produce a synonymous change in the coding sequence. All the other loci are predicted to
390 change the amino acid sequence of the polyprotein. Most notable is T1783A, which introduced
391 a stop codon early in the translation of ORF1ab. Introduced stop codons are rare among the
392 submitted genome assemblies tracking the evolution of SARS-CoV2 (nextstrain.org), but are
393 distributed all along the genome (S4 Fig). In some regions, these introduced stop codon alleles
394 occur in multiple loci along multiple lineages, one of which at a significant enough frequency to
395 be scored with high homoplasy [34]. The low phylogenetic signal disqualifies these loci from
396 much further analysis. A stop codon early in the ORF1ab gene should prevent propagation of
397 the virus, but it can possibly be complemented by the presence of a functional copy of the gene
398 from a co-infecting replication competent virus.

399 Defective interfering viral RNA can be replicated and packaged in the presence of replicating
400 viruses, and have been detected in other coronaviruses [35]. If the requirement for translational
401 fidelity of the ORF1ab gene were lost, it would remove any selective pressure on the remainder
402 of the gene and could explain the accumulation of additional mutations observed in the
403 defective species. It would not interfere with the generation of sub genomic segments of the rest
404 of the genome for translation of the proteins necessary to package the virus. Thus, the defective
405 virus can only be maintained in a heterogeneous population with a replication competent virus.
406 Engineering defective interfering viruses have the potential to modulate the replication of
407 functional viruses during the infection cycle.

408

409 Our capture approach enabled simultaneous detection and quantitation of the sub genomic
410 fragments. RPKM values plotted in Fig 5B were for reads originating from both genomes and
411 sub-genomes. Plotting of this data shows that capture is not biased in enrichment and that the
412 increase in coverage of the reads from 5'-3' is in agreement with the transcription pattern of the
413 sub-genomes as described by Kim et al [1]. Kim et al. [1], reported SARS-CoV-2 quantitative
414 expression in SARS-CoV-2 infected Vero cells (ATCC, CCL-81) based on junction reads
415 obtained from Nanopore based direct RNA sequencing. In their study, the N gene mRNA was
416 the most abundantly expressed, but they also identified expression in eight other ORFs with
417 least expression noted in the 7b gene. They did not detect sub genomic fragments enabling
418 translation of ORF10. Here, we searched for junctions reads in our data and used them to
419 quantitate ORF expression patterns in the 17 samples with full length genome reconstructions
420 (Fig 5 and S3 and S4 Tables). Differences in expression were noted among these 17 samples
421 suggesting that ORF expression is patient-specific and interestingly, this patient group
422 expression pattern also differed from the profiles reported by Kim et al.[1]. Further, evidence of
423 the expression of ORF10 was supported by multiple junction reads in three of our 17 samples
424 (192000052B, 192000251B, and 192000440B). The SARS-CoV-2 genome coverage in these

425 three samples was among the highest (3,192,285x, 1,196,745x, and 793,028x), which might
426 have contributed to their discovery (S1 Table). ORF10 was also undetected in the other
427 transcriptome study by Tairaoa et al., 2020 using ONT and SARS-CoV-2 infected Vero/hSLAM
428 cells. ORF10 is 117 bases in length so it may have been missed by these studies due to its low
429 or absent expression in cultured cells. We note however that the capture methodology is limited
430 in its ability to identify the RNA modifications that were reported by the above two direct RNA-
431 Seq methods.

432 In summary, this capture enrichment and sequencing method provides an effective approach to
433 generate SARS-CoV-2 genome and transcriptome data directly from clinical samples. Samples
434 with Ct values ≤ 33 , when sequenced to a depth of approximately 2 million reads (higher than
435 1000x coverage of the SARS-CoV-2 genome), appear to be sufficient for both full genome
436 reconstruction and identification and quantitation of junction-reads to measure differential ORF
437 expression. This article was posted on Bioarchive on July 27th, 2020. As a follow up to this
438 study, an additional 95 patient samples with Sars-Cov-2 Ct. values of 9.3-31.3 Ct. were
439 sequenced. For all 95 samples, SARS-CoV-2, full-length genomes were reconstructed
440 (unpublished data). This method has a straightforward work-flow and is scalable for sequencing
441 large numbers of patient samples.

442

443 **Accession numbers**

444 All the 17 full-length reconstructed SARS-CoV-2 genomes are available at GISAID
445 (www.gisaid.org) under the accession numbers EPI_ISL_444022, EPI_ISL_445078 -
446 EPI_ISL_445084, EPI_ISL_501168 – EPI_ISL_501174 and EPI_ISL_513294.

447

448 **Acknowledgements**

449 Part of this work was supported by the National Institute of Allergy and Infectious Diseases
450 (Grant#1U19AI144297). The authors are grateful to the production teams at HGSC for data
451 generation.

452

453 **Author contributions**

454 **Conceived and designed the experiments:** Harsha Doddapaneni, and Richard A. Gibbs
455 **Data Generation** - Harsha Doddapaneni, Qingchang Meng, Hsu Chao, Vipin Menon, Vasanthi
456 Avadhanula, Erin Nicholson, Felipe-Andres Piedra, Anubama Rajan, Zeineen Momin, Kavya
457 Kottapalli, Kristi L. Hoffman, Ginger Metcalf, Pedro A. Piedra, Donna M. Muzny, Joseph F.
458 Petrosino,
459 **Data analysis** - Sara Javornik Cregeen, Richard Sugang, Xiang Qin, David Henke, Fritz J.
460 Sedlazeck, Joseph F. Petrosino

461

462 **Conflict of interest**

463 None declared.

464 **References**

- 465 1. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome.
466 Cell. 2020;181(4):914-21 e10. Epub 2020/04/25. doi: 10.1016/j.cell.2020.04.011. PubMed PMID:
467 32330414; PubMed Central PMCID: PMC7179501.
- 468 2. Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol. 2018;16(8):e3000003. Epub
469 2018/08/14. doi: 10.1371/journal.pbio.3000003. PubMed PMID: 30102691; PubMed Central PMCID:
470 PMC6107253.
- 471 3. Wu HY, Brian DA. Subgenomic messenger RNA amplification in coronaviruses. Proc Natl Acad Sci
472 U S A. 2010;107(27):12257-62. Epub 2010/06/22. doi: 10.1073/pnas.1000378107. PubMed PMID:
473 20562343; PubMed Central PMCID: PMC2901459.
- 474 4. Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M. Disentangling primer interactions
475 improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PLoS One. 2020;15(9):e0239403. Epub
476 2020/09/19. doi: 10.1371/journal.pone.0239403. PubMed PMID: 32946527; PubMed Central PMCID:
477 PMC7500614.
- 478 5. Long SW, Olsen RJ, Christensen PA, Bernard DW, Davis JR, Shukla M, et al. Molecular Architecture
479 of Early Dissemination and Evolution of the SARS-CoV-2 Virus in Metropolitan Houston, Texas. bioRxiv.
480 2020:2020.05.01.072652. doi: 10.1101/2020.05.01.072652.
- 481 6. Resende PC, Motta FC, Roy S, Appolinario L, Fabri A, Xavier J, et al. SARS-CoV-2 genomes recovered
482 by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing
483 platforms. bioRxiv. 2020:2020.04.30.069039. doi: 10.1101/2020.04.30.069039.
- 484 7. St Hilaire BG, Durand NC, Mitra N, Pulido SG, Mahajan R, Blackburn A, et al. A rapid, low cost, and
485 highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. bioRxiv.
486 2020:2020.04.25.061499. doi: 10.1101/2020.04.25.061499.

- 487 8. Zakrzewski F, Gieldon L, Rump A, Seifert M, Grutzmann K, Kruger A, et al. Targeted capture-based
488 NGS is superior to multiplex PCR-based NGS for hereditary BRCA1 and BRCA2 gene analysis in FFPE tumor
489 samples. *BMC Cancer*. 2019;19(1):396. Epub 2019/04/29. doi: 10.1186/s12885-019-5584-6. PubMed
490 PMID: 31029168; PubMed Central PMCID: PMC6487025.
- 491 9. Alexandersen S, Chamings A, Bhatta TR. SARS-CoV-2 genomic and subgenomic RNAs in diagnostic
492 samples are not an indicator of active replication. *medRxiv*. 2020:2020.06.01.20119750. doi:
493 10.1101/2020.06.01.20119750.
- 494 10. Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L, Druce J, et al. Direct RNA sequencing and
495 early evolution of SARS-CoV-2. *bioRxiv*. 2020:2020.03.05.976167. doi: 10.1101/2020.03.05.976167.
- 496 11. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human
497 genomic loci by microarray hybridization. *Nat Methods*. 2007;4(11):903-5. Epub 2007/10/16. doi:
498 10.1038/nmeth1111. PubMed PMID: 17934467.
- 499 12. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, et al. Whole exome
500 capture in solution with 3 Gbp of data. *Genome Biol*. 2010;11(6):R62. Epub 2010/06/23. doi: 10.1186/gb-
501 2010-11-6-r62. PubMed PMID: 20565776; PubMed Central PMCID: PMC2911110.
- 502 13. Cieslik M, Chugh R, Wu YM, Wu M, Brennan C, Lonigro R, et al. The use of exome capture RNA-
503 seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res*.
504 2015;25(9):1372-81. Epub 2015/08/09. doi: 10.1101/gr.189621.115. PubMed PMID: 26253700; PubMed
505 Central PMCID: PMC4561495.
- 506 14. Schuierer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, et al. A comprehensive
507 assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics*.
508 2017;18(1):442. Epub 2017/06/07. doi: 10.1186/s12864-017-3827-y. PubMed PMID: 28583074; PubMed
509 Central PMCID: PMC5460543.
- 510 15. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, et al. Virome Capture Sequencing
511 Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *mBio*. 2015;6(5):e01491-15. Epub
512 2015/09/24. doi: 10.1128/mBio.01491-15. PubMed PMID: 26396248; PubMed Central PMCID:
513 PMC4611031.
- 514 16. O'Flaherty BM, Li Y, Tao Y, Paden CR, Queen K, Zhang J, et al. Comprehensive viral enrichment
515 enables sensitive respiratory virus genomic identification and analysis by next generation sequencing.
516 *Genome Res*. 2018;28(6):869-77. Epub 2018/04/29. doi: 10.1101/gr.226316.117. PubMed PMID:
517 29703817; PubMed Central PMCID: PMC5991510.
- 518 17. Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing using targeted sequence
519 capture. *Genome Res*. 2015;25(12):1910-20. Epub 2015/09/24. doi: 10.1101/gr.191049.115. PubMed
520 PMID: 26395152; PubMed Central PMCID: PMC4665012.
- 521 18. Tan CCS, Maurer-Stroh S, Wan Y, Sessions OM, de Sessions PF. A novel method for the capture-
522 based purification of whole viral native RNA genomes. *AMB Express*. 2019;9(1):45. Epub 2019/04/10. doi:
523 10.1186/s13568-019-0772-y. PubMed PMID: 30963294; PubMed Central PMCID: PMC6453989.
- 524 19. Heyer EE, Deveson IW, Wooi D, Selinger CI, Lyons RJ, Hayes VM, et al. Diagnosis of fusion genes
525 using targeted RNA sequencing. *Nat Commun*. 2019;10(1):1388. Epub 2019/03/29. doi: 10.1038/s41467-
526 019-09374-9. PubMed PMID: 30918253; PubMed Central PMCID: PMC6437215.
- 527 20. Schroder J, Kumar A, Wong SQ. Overview of Fusion Detection Strategies Using Next-Generation
528 Sequencing. *Methods Mol Biol*. 2019;1908:125-38. Epub 2019/01/17. doi: 10.1007/978-1-4939-9004-7_9.
529 PubMed PMID: 30649725.
- 530 21. Ajami NJ, Wong MC, Ross MC, Lloyd RE, Petrosino JF. Maximal viral information recovery from
531 sequence data using VirMAP. *Nat Commun*. 2018;9(1):3205. Epub 2018/08/12. doi: 10.1038/s41467-018-
532 05658-8. PubMed PMID: 30097567; PubMed Central PMCID: PMC6086868.
- 533 22. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome
534 assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77. Epub

- 535 2012/04/18. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PubMed Central PMCID:
536 PMCPMC3342519.
- 537 23. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence
538 alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30(14):3059-66. Epub 2002/07/24.
539 doi: 10.1093/nar/gkf436. PubMed PMID: 12136088; PubMed Central PMCID: PMCPMC135756.
- 540 24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
541 *Bioinformatics.* 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324. PubMed
542 PMID: 19451168; PubMed Central PMCID: PMCPMC2705234.
- 543 25. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling
544 accurate genetic variant discovery to tens of thousands of samples. *bioRxiv.* 2018:201178. doi:
545 10.1101/201178.
- 546 26. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, et al. A SNP discovery method to assess
547 variant allele probability from next-generation resequencing data. *Genome Res.* 2010;20(2):273-80. Epub
548 2009/12/19. doi: 10.1101/gr.096388.109. PubMed PMID: 20019143; PubMed Central PMCID:
549 PMCPMC2813483.
- 550 27. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and
551 predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*
552 *melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6(2):80-92. Epub 2012/06/26. doi:
553 10.4161/fly.19695. PubMed PMID: 22728672; PubMed Central PMCID: PMCPMC3679285.
- 554 28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-
555 seq aligner. *Bioinformatics.* 2013;29(1):15-21. Epub 2012/10/30. doi: 10.1093/bioinformatics/bts635.
556 PubMed PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.
- 557 29. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature
558 proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol.* 2020. Epub 2020/07/17.
559 doi: 10.1038/s41564-020-0770-5. PubMed PMID: 32669681.
- 560 30. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in
561 SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell.* 2020. Epub
562 2020/07/23. doi: 10.1016/j.cell.2020.06.043. PubMed PMID: 32697968.
- 563 31. Quick J. nCoV-2019 sequencing protocol: @protocolsIO; 2020. Available from:
564 <https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w.pdf>.
- 565 32. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole Genome
566 Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation
567 during a Pandemic. *Genes (Basel).* 2020;11(8). Epub 2020/08/23. doi: 10.3390/genes11080949. PubMed
568 PMID: 32824573; PubMed Central PMCID: PMCPMC7464704.
- 569 33. Singanayagam A, Patel M, Charlett A, Lopez Bernal J, Saliba V, Ellis J, et al. Duration of
570 infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January
571 to May 2020. *Euro Surveill.* 2020;25(32). Epub 2020/08/15. doi: 10.2807/1560-
572 7917.ES.2020.25.32.2001483. PubMed PMID: 32794447; PubMed Central PMCID: PMCPMC7427302.
- 573 34. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic
574 diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol.* 2020;83:104351. Epub 2020/05/11.
575 doi: 10.1016/j.meegid.2020.104351. PubMed PMID: 32387564; PubMed Central PMCID:
576 PMCPMC7199730.
- 577 35. Banerjee S, Repass JF, Makino S. Enhanced accumulation of coronavirus defective interfering RNA
578 from expressed negative-strand transcripts by coexpressed positive-strand RNA transcripts. *Virology.*
579 2001;287(2):286-300. Epub 2001/09/05. doi: 10.1006/viro.2001.1047. PubMed PMID: 11531407;
580 PubMed Central PMCID: PMCPMC7133719.

582 **Supporting information**

583

584 **S1 Fig. Genome coverage plots for the three SARS-CoV-2 negative samples.** Coverage is

585 localized despite the 45-91 M reads that these samples obtained post-capture.

586

587 **S2 Fig. Genome coverage plots.** Genome coordinates on X-axis and coverage in log scale of

588 Y-axis for the 17 samples with full length SARS-CoV-2 genome reconstructions

589 **S3 Fig. A multiple sequence alignment (using MAFFT) of 17 reconstructed SARS-CoV-2**

590 **genomes and Wuhan-Hu-1 reference genome (NC_045512).** Grey indicates agreement with

591 the reference, black is a disagreement, and pink marks areas in the reconstruction with an

592 ambiguous nucleotide, "N". The pangolin lineage assignment is listed next to the sample name.

593 The extra length of the 192000251B seen here is an assembly artifact and was excluded from

594 analysis.

595 **S4 Fig. Stop codon variants in sampled SARS-CoV-2 genomic assemblies.** A snapshot of

596 full length SARS-CoV-2 genome assemblies from GISAID and NCBI on 27 May 2020 was

597 downloaded (comprising 39246 entries), and processed to detect single nucleotide variant

598 alleles that introduced a stop codon. Introduced stop codons were detected in 270 entries, and

599 the frequency of these alleles are plotted along the SARS-CoV-2 reference genome position.

600 Introduced stop codons are rare but are distributed throughout the genomic sequences. Multiple

601 loci harbor stop codons in unrelated assemblies.

602

603 **S5 Fig. Junctions reads to support expression of ORF10 192000052B, 192000251B and**

604 **192000440B.** Expression values were calculated as 0.13, 0.13 and 0.02 reads/million. Few

605 examples of those junction reads are shown in the figure (purple arrows).

606

607 **S1 Table.** Sample information, capture pools and sequencing metrics details.

608 **S2 Table.** Lineage analysis of the 17 full-length genomes.

609 **S3 Table.** Junction read counts is reads/million identified in the post capture data of 17 samples
610 with full-length genomes.

611 **S4 Table.** Junction read counts in reads/million identified in the nine samples sequenced before
612 (IDxxxxB-2) and after capture (IDxxxxB) enrichment.

613

614

615

616

617

618

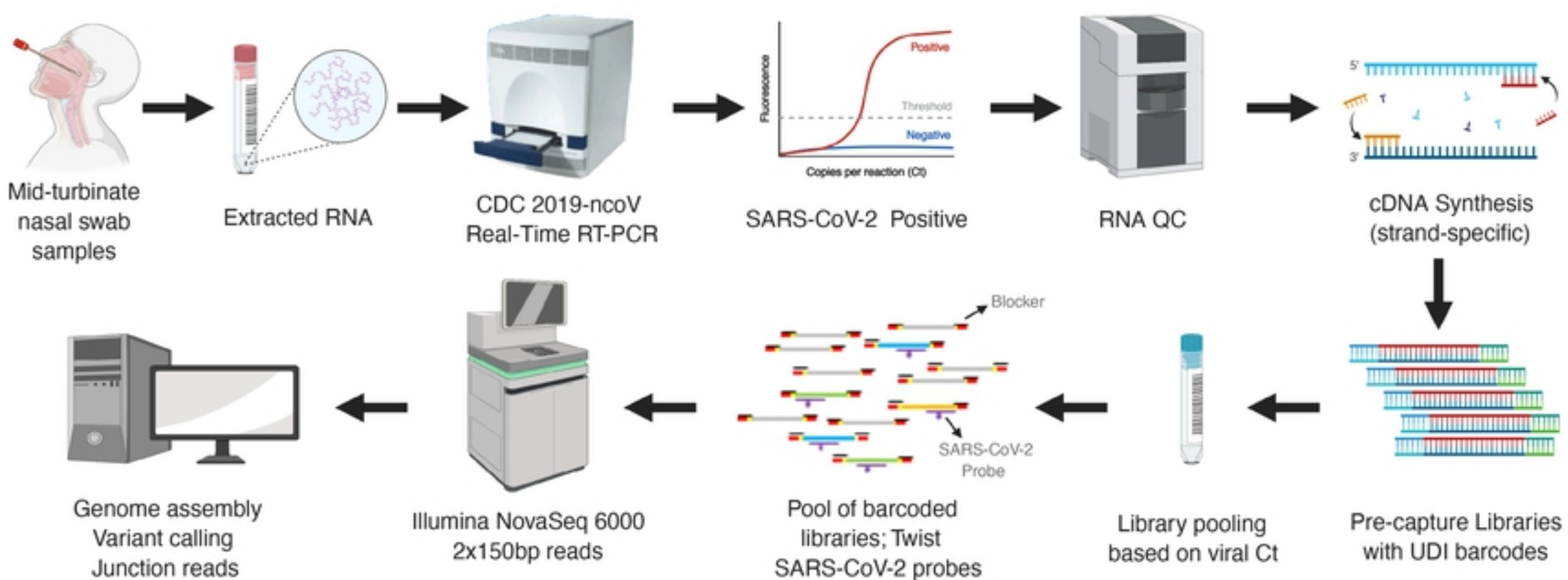
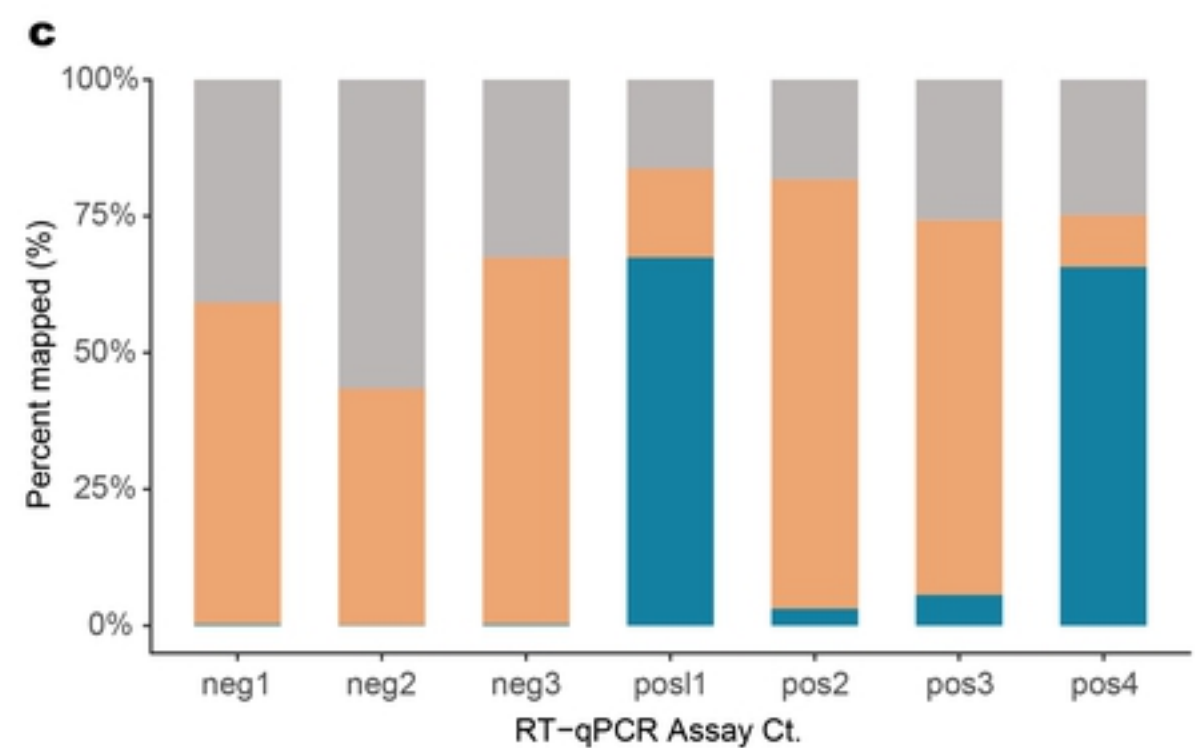
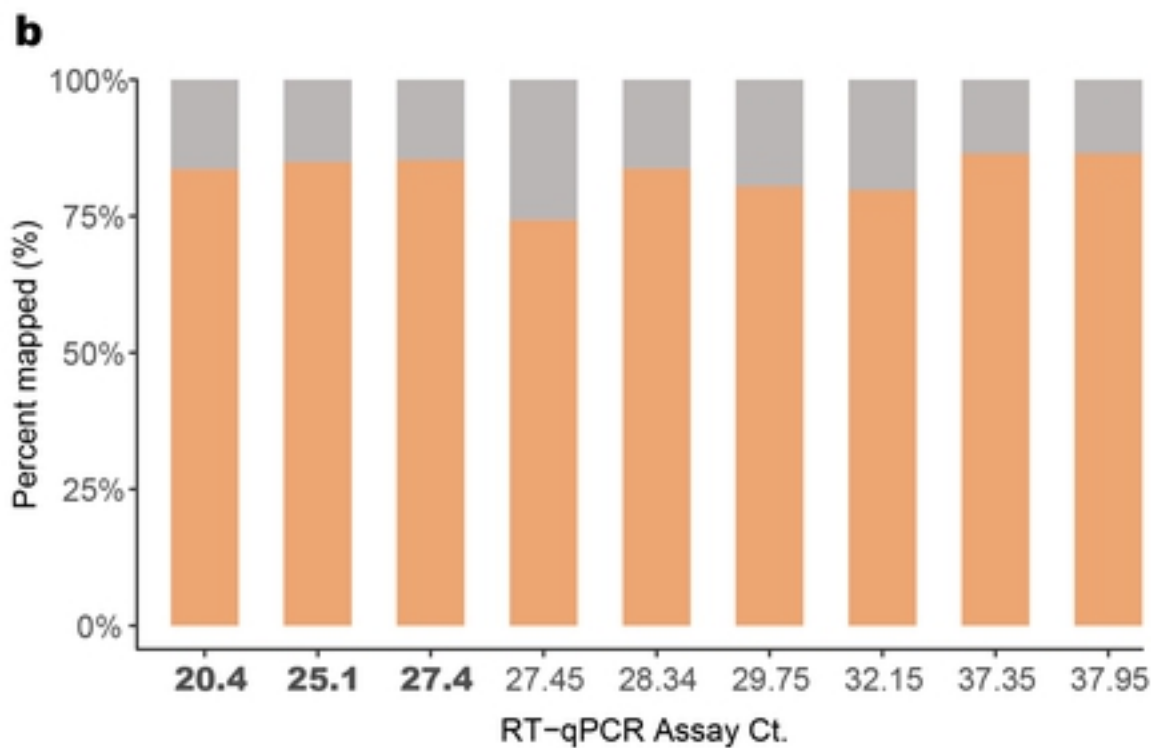
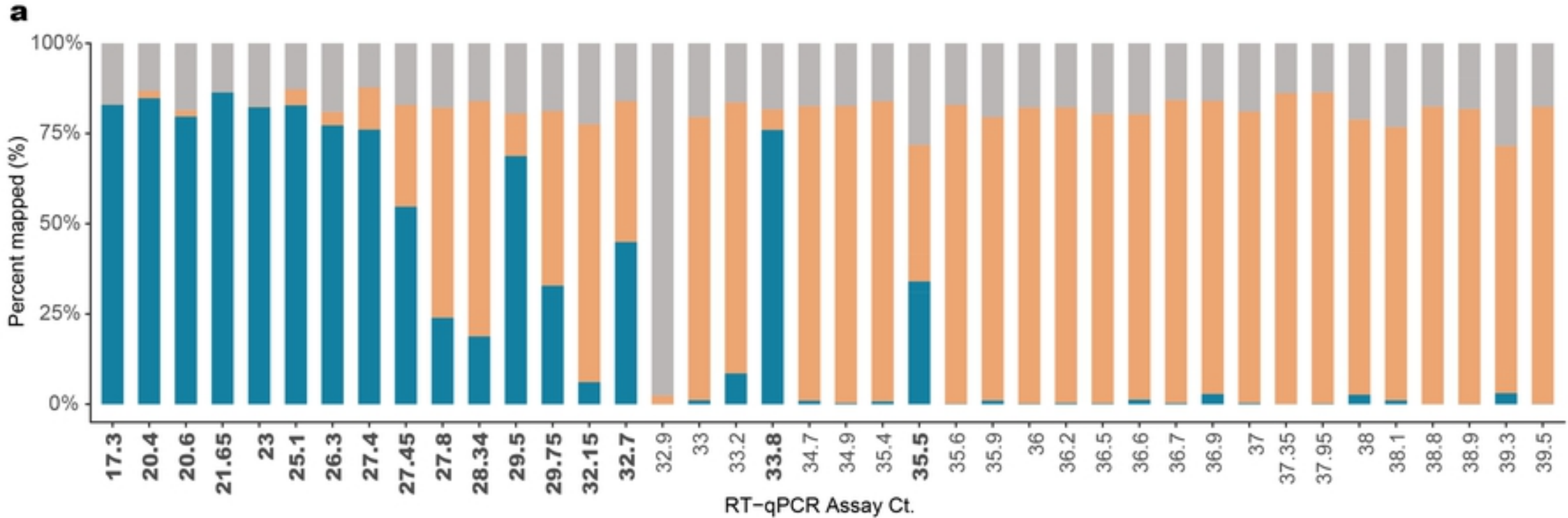


Fig 1



Legend ■ Percent reads mapped to SARS-CoV-2 ■ Percent reads mapped to human reference ■ Percent reads other

Fig 2

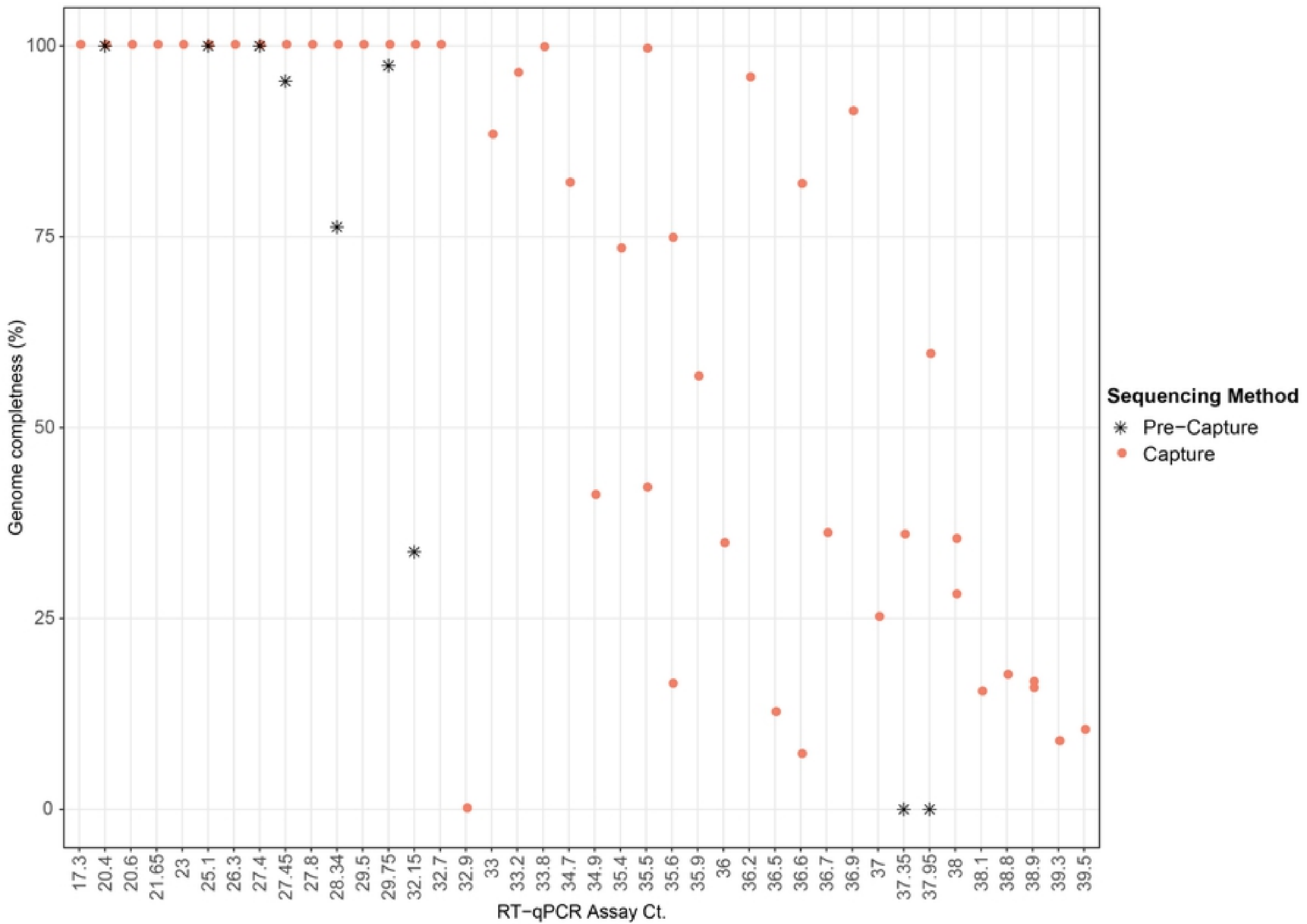
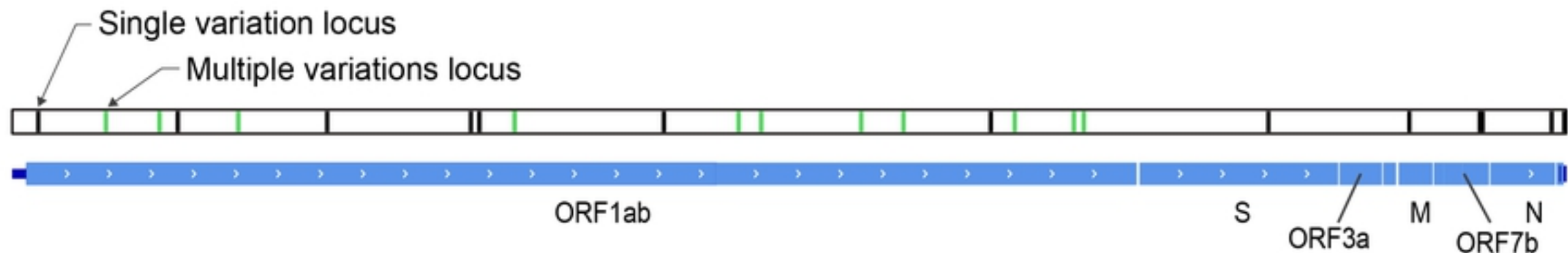


Fig 3



Nucleotide Position	Reference Allele	Alternate Allele	% Alternate	Reference amino acid	Predicted change
1783	T	A	89.0%	C	[Stop]
2822	C	T	89.8%	L	F
4320	C	T	10.7%	A	V
9611	C	T	77.3%	L	F
9620	G	T	79.2%	D	Y
13892	G	T	82.4%	C	F
14342	G	T	81.0%	C	F
16244	G	T	87.5%	G	V
17068	T	C	84.0%	S	P
19181	T	C	87.1%	V	A
20320	C	T	87.0%	H	Y
20520	T	C	84.2%	D	D

Fig 4

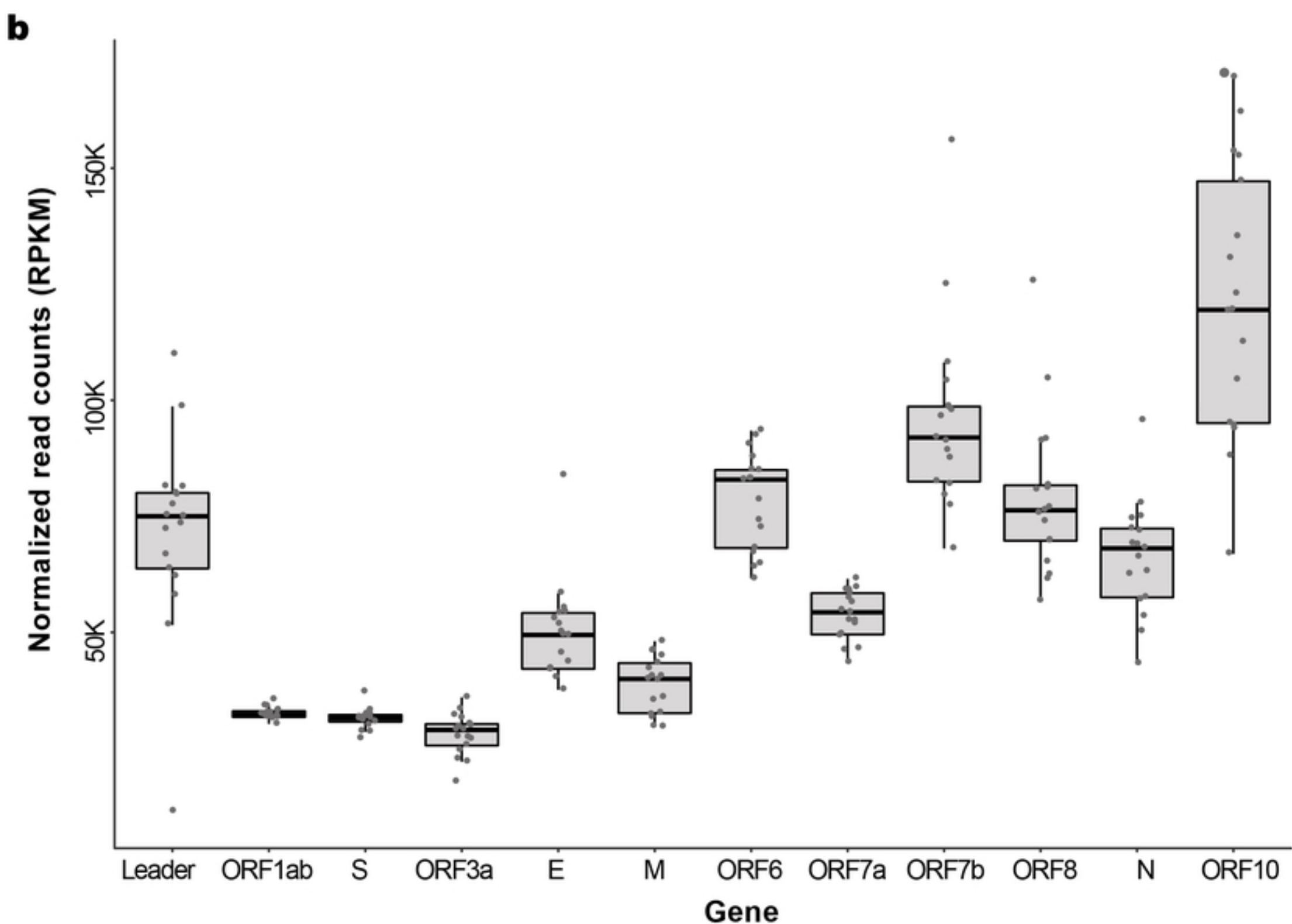
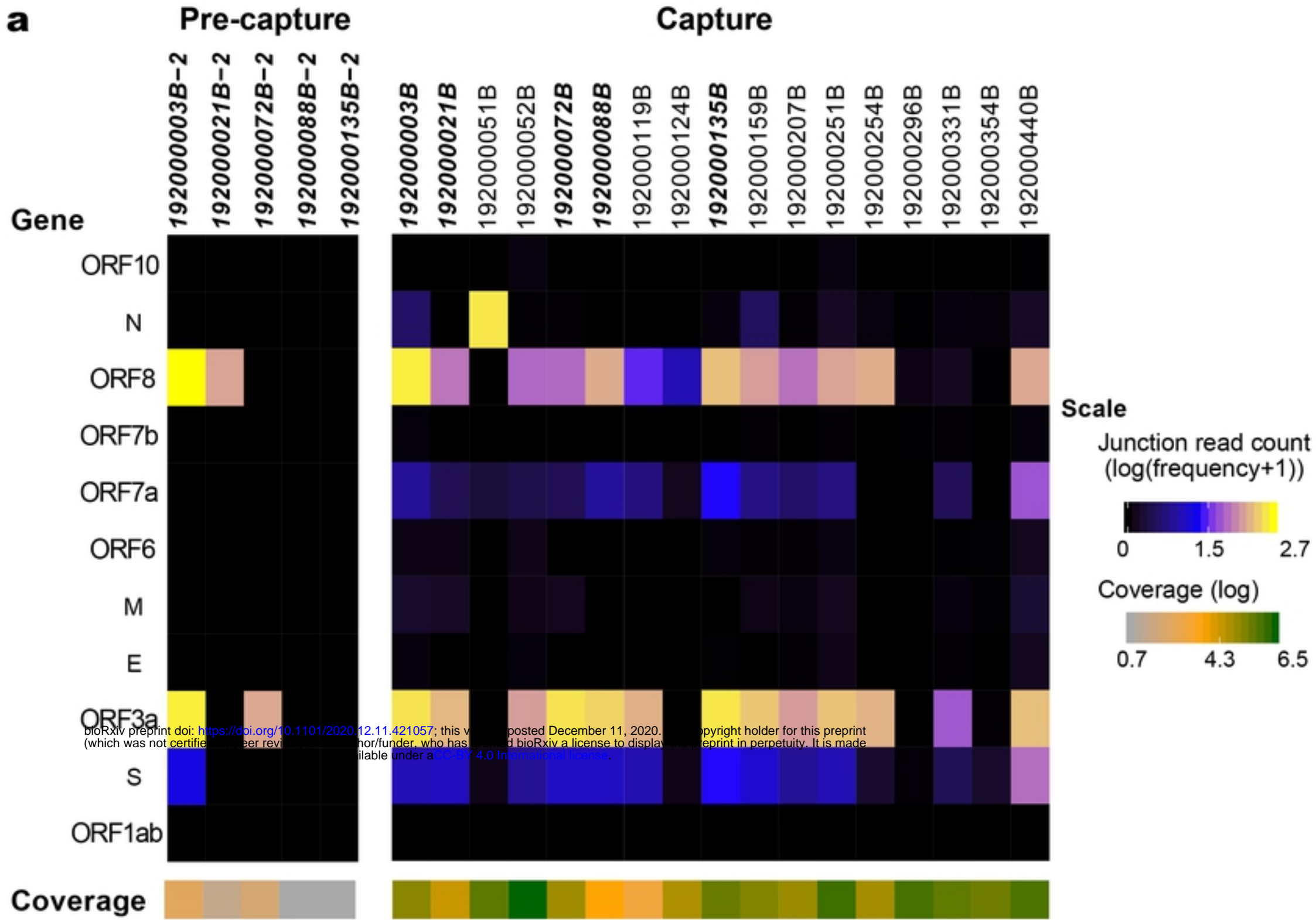


Fig 5