ACoRE: Accurate SARS-CoV-2 genome reconstruction for the characterization of intra-host and inter-host viral diversity in clinical samples and for the evaluation of re-infections

Luca Marcolungo[1]*, Cristina Beltrami[1]*, Chiara Degli Esposti[1], Giulia Lopatriello[1], Chiara Piubelli[2], Antonio Mori[2], Elena Pomari[2], Michela Deiana[2], Salvatore Scarso[2], Zeno Bisoffi[2,3], Valentina Grosso[1], Emanuela Cosentino[1], Simone Maestri[1], Denise Lavezzari[1], Barbara Iadarola[1], Marta Paterno[1], Elena Segala[1], Barbara Giovannone[1], Martina Gallinaro[1], Marzia Rossato[1,4] and Massimo Delledonne[1,4 #].

[1]Department of Biotechnology, University of Verona, Strada le Grazie 15, 37134 Verona, Italy

[2]Department of Infectious and Tropical Diseases and Microbiology, IRCCS Sacro Cuore Don Calabria Hospital, Negrar di Valpolicella, 37024 Verona, Italy

[3]Department of Diagnostics and Public Health, University of Verona, 37134 Verona, Italy

[4]Genartis srl, via IV Novembre 24, 37126 Verona, Italy

*These authors contributed equally to this work

corresponding author: massimo.delledonne@univr.it

ABSTRACT

We report Accurate SARS-CoV-2 genome Reconstruction (ACoRE), an amplicon-based viral genome sequencing workflow for the complete and accurate reconstruction of SARS-CoV-2 sequences from clinical samples, including suboptimal ones that would usually be excluded even if unique and irreplaceable. We demonstrated the utility of the approach by achieving complete genome reconstruction and the identification of false-positive variants in >170 clinical samples, thus avoiding the generation of inaccurate and/or incomplete sequences. Most importantly, ACoRE was crucial to identify the correct viral strain responsible of a relapse case, that would be otherwise mis-classified as a re-infection due to missing or incorrect variant identification by a standard workflow.

KEYWORDS: SARS-CoV-2 genome sequencing, genetic variants, re-infection, suboptimal samples, low-viral titer

47 **BACKGROUND**

48 The coronavirus disease 2019 (COVID-19) pandemic has thus far resulted in the infection of more than
49 84 million people, causing at least 1.8 million deaths (Johns Hopkins University, 1/1/2021)[1] . The
50 agent responsible for COVID-19 is a β-coronavirus known as severe acute respiratory syndrome-
51 associated coronavirus 2 (SARS-CoV-2) with a compact single-stranded RNA genome of 29,903
52 nucleotides. The first SARS-CoV-2 genome sequence was published soon after the initial outbreak [2],
53 and more than 260,000 complete genome sequences have subsequently been deposited in the GISAID
54 database [3]. The phylogenetic analysis of genomic sequences provides a valuable tool to track viral
55 diversity during the course of a pandemic and to identify the emergence of prevalent strains
56 characterized by lineage-specific single nucleotide variants (SNVs), such as the D614G variant in the
57 SARS-CoV-2 spike protein gene (23403:A→G) [4–6]. As the virus propagates in human-to-human
58 transmission, changes in the reference genome sequence must be recorded to monitor correlations
59 between viral genotype and disease communicability, manifestation and severity [4,7–9]. The
60 combination of genomic analysis and epidemiological data can also reliably determine the extent of
61 SARS-CoV-2 transmission in different nations [10–12] and thus facilitates early decision-making to
62 control local transmission [13]. Finally, mutations that may be relevant to the fitness or antigenic
63 profile of the virus can be identified to ensure the efficacy of vaccines and immunotherapeutic
64 interventions in the clinic[4,14].

65 Consensus variations reflect the analysis of virus sequences that differ between patients, but the
66 analysis of intra-individual single nucleotide variations (iSNVs) is also important because it helps us to
67 understand more about virus–host interactions, as previously demonstrated for Ebola, Zika, influenza
68 and HIV [15–19]. The analysis of iSNVs during the COVID-19 pandemic may also provide data about
69 the potential of SARS-CoV-2 for immunological escape and resistance to therapy, as well as on the
70 sensitivity of molecular diagnostic assays [20–22]. However, the identification of iSNVs in clinical
71 samples can be challenging because current protocols often feature enrichment and amplification
72 steps that introduce technical errors indistinguishable from true biological variants [23].

73 The reconstruction of complete and accurate genomic sequences to detect both SNVs and iSNVs is
74 therefore necessary to produce reliable data, at all these aims. In addition, the accumulation of
75 meaningful data during pandemics requires the analysis of many samples, and the corresponding
76 methods must therefore be cost-effective, straightforward and suitable for high-multiplexing [24]. The
77 protocols must also be sensitive enough to detect low viral titers but applicable over a wide dynamic
78 range of virus concentrations to allow the analysis of clinical samples with different viral loads, ideally
79 including samples from early and late infection stages, that usually show a lower viral detection, or
80 from re-infection/relapse cases [25,26].

81 Among the many approaches available for SARS-CoV-2 whole-genome analysis, the amplicon-based
82 sequencing method developed by the ARTIC Network [27] is currently the most widely used [13,24,28–
83 32]. Based on the PrimalSeq protocol originally developed for Zika virus [23,33], the ARTIC Network
84 designed a set of 98 tiled amplicons in two PCR pools for the targeted whole-genome amplification of
85 SARS-CoV-2 [27]. This approach is simple and highly sensitive, but it suffers from technical biases
86 leading to uneven genome coverage, thus reducing the completeness and accuracy of genome
87 sequencing, especially for the identification iSNVs in samples with low viral titers [34–36]. Sequencing
88 technical replicates of multiple cDNAs generated from the same sample has been proposed as a
89 mitigation strategy to identify iSNVs more reliably [23]. However, whereas amplicon-based
90 sequencing has been widely used for the analysis of low-frequency variants [20–22,37,38] only a few
91 studies thus far have evaluated the confidence of such calls and have implemented the sequencing of
92 cDNA replicates to ensure accuracy [23] . False positives have also been reported among high-
93 frequency variants supported by good sequencing depth, indicating that the risks of inaccurate
94 sequencing are not limited to suboptimal samples [39].

95 To avoid the generation of incomplete genomic sequences typically associated with poor genome
96 coverage [40–42], the sequencing of samples with fewer than 1000 virus copies per RT-qPCR reaction
97 (Ct < 30) is currently discouraged [23,43]. However, the strict implementation of such

98 recommendations would lead to the exclusion of many clinical samples, which are often unavoidably
99 collected or stored under suboptimal conditions. Since specimens with these features may be unique
100 and irreplaceable -central to the investigation conducted-, numerous studies therefore report
101 sequencing data from samples with (very) low viral titers (Ct > 30) despite this advice [26,44,45] . To
102 address these challenges, we set out to develop an optimized workflow, ACoRE (Accurate SARS-CoV-
103 2 genome Reconstruction, for the reliable reconstruction of complete and accurate SARS-CoV-2
104 genomes from clinical samples with a broad range of Ct values, aiming to improve the flexibility,
105 accuracy and throughout of amplicon-based sequencing.

106 **RESULTS**

107 **Accuracy of SARS-CoV-2 genome reconstruction**

108 The original Primalseq protocol stipulates two independent reverse transcriptions per sample and the
109 subsequent amplification of the separate cDNAs in order to reduce technical errors. In this study, we
110 initially tested replicate amplifications from the same cDNA to investigate whether this alternative
111 approach could affect the reproducibility in the generation of SARS-CoV-2 consensus sequences and
112 in the identification of intra-host variants. At this aim, we selected five COVID-19-positive swabs
113 representing viral loads ranging from ~500 to ~2 million, based on Ct values (determined by RT-qPCR)
114 ranging from 15.07 to 28.5 (**Table S1**). For each sample, we generated three cDNAs and carried out
115 two separate amplifications, resulting in six replicates per starting RNA (**Figure 1A**). An individual KAPA
116 library was prepared from each replicate, and sequencing in 250PE mode produced an average of
117 1 million fragments. The dataset was normalized to ~800,000 fragments per library, corresponding to
118 ~7800× coverage per sample after alignment to the SARS-CoV-2 reference genome (**Table S3**).

119 The sequencing coverage was variable across the different amplicons of the ARTIC panel, particularly
120 in samples with a higher Ct value (**Figure 2 and Figure S1**). Interestingly, most amplicons showed either
121 high (>500×) or very low (≤10×) to zero coverage, and amplicons absent in one replicate could be

122 present in another, even when produced from the same cDNA. The concordance (R_c) in sequencing
123 coverage was high for replicates of four samples (R_c ~0.99–1) but lower in sample S5 (R_c ~0.95) with
124 the lowest viral load (**Figure 1B and Table S4**), but there was no significant difference between
125 replicates from the same or different cDNAs ($p = 0.25$, Wilcoxon test). Variations in coverage can affect
126 genotyping accuracy, so we evaluated reproducibility in terms of genotypability by calculating the
127 fraction of genomic positions where it is possible to call a genotype after aligning reads to the
128 reference genome. The genotypability R_c was optimal or slightly lower than 1 in all samples (R_c = 0.99–
129 1), but lower in sample S5, which also showed the lowest sequencing coverage R_c (**Figure 1C and Table**
130 **S5)**. Reproducibility was similar between inter-cDNA replicates and intra-cDNA replicates ($p > 0.99$,
131 Wilcoxon test). To assess how fluctuations in genotypability and coverage affect the final viral genome
132 sequences, we generated a consensus sequence for each replicate. The reproducibility among
133 consensus variants was optimal in the first four samples, but consistently dropped to ~0.3 for sample
134 S5 (**Figure 1D and Table S6**). Nevertheless, reproducibility was again similar between inter-cDNA
135 replicates and intra-cDNA replicates ($p > 0.99$, Wilcoxon test).

136 The number of iSNVs (frequency >3%) varied significantly between technical replicates, with a small
137 fraction of iSNVs shared by different replicates compared to the total number of iSNVs identified
138 (**Table S7**). The R_c was suboptimal (<0.95) for all samples and steadily decreased as the Ct value
139 increased (**Figure 1E and Table S8**), but there was no significant difference between replicates
140 generated from the same or different cDNAs ($p = 0.44$, Wilcoxon test). In summary, consensus
141 sequences and intra-host variants can be strongly affected by uneven amplicon representation and
142 PCR errors (**Figure 2**) confirming the need to sequence at least two replicates to achieve an accurate
143 characterization of the SARS-CoV-2 genome. However, the two amplifications can be generated from
144 the same starting cDNA, thus reducing sample consumption and costs.

145 **Improvement of genome reconstruction by merging technical replicates**

146 While addressing the reproducibility issues observed for both SNVs and iSNVs in samples with low viral
147 loads, we also tested whether merging two or more technical replicates could improve coverage and
148 genotypability. The rationale was the observation that amplicons with the lowest coverage varied
149 across different replicates, and amplicons missing in one replicate could have a coverage >100× or
150 >1000× in others (**Figure S1**). All possible combinations of two replicates for each sample were merged
151 and downsampled to 800,000 fragments (400,000 for each replicate) to obtain the same sequencing
152 input data as the initial analysis based on a single replicate (**Table S9**). When considering the merged
153 datasets rather than single-replicate data, the average coverage consistently increased in the sample
154 with the highest Ct value ($p < 0.0001$, Mann Whitney U-test), confirming that merging two
155 amplification replicates (intra-cDNA or inter-cDNA) could mitigate the technical variability in amplicon
156 coverage (**Figure 3A-C**) as well as significantly ($p < 0.0001$, Mann Whitney U-test) enhance the
157 genotypability (**Figure 3B**). Merging up to six replicates achieved a slight further improvement in both
158 coverage and genotypability (**Figure 3A-B**), indicating that both properties can be maximized by
159 analyzing replicates of samples with low viral loads. Indeed, merging all sequence data available for
160 sample S5 (with the lowest reproducibility) increased coverage sufficiently to achieve >96.98% non-
161 ambiguous bases in the consensus sequence (**Figure 3C-D**), which is the GISAID threshold for
162 classifying a SARS-CoV-2 genome as complete [3]. Similar improvement was achieved in a panel of 170
163 clinical samples analyzed in duplicate or quadruplicate (**Figure 3E-G** shows three representative
164 samples).

165 **Improvement of the technical workflow for viral genome sequencing**

166 One drawback of the ARTIC protocol on the Illumina platform is the need for 250PE sequencing to
167 cover the full length of the amplicons (400 bp). This type of sequencing is currently available only for
168 MiSeq and NovaSeq6000 SP flow cells, increasing the cost per sample and reducing the sample
169 throughput. We therefore generated shorter libraries using the NexteraFlex approach and tested the
170 use of alterative flow cells (NextSeq500/550 and NovaSeq6000 S1) and sequencing mode (150PE) on

171 the 30 samples originally tested using the KAPA library (**Figure 1A**). Despite skipping the laborious
172 input DNA and library quantification steps before sequencing, the variability in the number of
173 fragments analyzed per sample was lower (CV = 22.5%) than the full-amplicon approach (CV = 38.3%)
174 described above (**Figure 4A**). The sequencing data were mapped to the reference genome (**Table S10**)
175 and compared to the 250PE dataset (KAPA library) normalized with the same average-mapped
176 coverage as the 150PE dataset (NexteraFlex library) (**Table S11**). Sequencing coverage was evenly
177 distributed along the amplicons even when the NexteraFlex protocol was used, because the partial
178 overlap of ARTIC amplicons compensated for the expected loss of sequence representation at the
179 amplicon ends due to tagmentation (**Figure 4B**). The sequencing of fragmented amplicons had no
180 adverse impact on genome coverage and genotypability, which were significantly higher compared to
181 the full-length amplicon sequencing ($p < 0.001$ and $p = 0.024$, respectively, Friedman test; **Figure 4C-**
182 **D**). Despite the lower coverage, similar results were observed with 100PE sequencing simulated after
183 trimming the 150PE dataset (**Figure 4C-D**). The fragmented-amplicon approach was therefore
184 advantageous for multiple aspects of SARS-CoV-2 sequencing, by increasing coverage, genotypability
185 and throughput (allowing higher multiplexing) while reducing sequencing costs and eliminating
186 unnecessary protocol steps such as DNA quantification after PCR and library quantification before
187 pooling.

188 Although the NexteraFlex protocol saves on costs, this is offset by the requirement for multiple
189 sequencing replicates from the same sample to improve genome coverage. We therefore compared
190 the effect of sequencing a library generated from two replicates (each amplified from 5 µL of cDNA)
191 and a standard library prepared from a single amplification generated from double amount of cDNA
192 (10 µL). Because samples with a low viral load benefit the most from multiple replicates, we analyzed
193 20 samples with a Ct range of 25–35 (**Figure S2A**). Two samples showed a lower coverage in libraries
194 produced from a single cDNA, but overall there was little difference in coverage ($p = 0.1$) or
195 genotypability ($p = 0.09$) when comparing the two conditions (Wilcoxon test; **Figure S2B-C**). This result
196 confirmed that the reconstruction of SARS-CoV-2 genomes can also be maximized by increasing the

197 amount of template cDNA through the use of more complex samples. Although such adjustments can
198 improve coverage and genotypability, technical replicates are still required for the identification of
199 true-positive variants.

200 **Application of the optimized workflow to large sets of samples**

201 Next we applied the optimized workflow to a set of 170 clinical samples representing a wide range of
202 viral loads, with Ct values in the range 15–40 (**Figure S3**). Each sample was amplified in duplicate or
203 quadruplicate starting from 10 µL cDNA, and 100PE sequencing was carried on a NovaSeq6000 SP flow
204 cell using NexteraFlex libraries, generating an average of ~2.8 million fragments per replicate (**Table**
205 **S12**). After pooling data from the replicates, ~75% of the samples showed both coverage and
206 genotypability >96.98% (**Figure 5A-B**) which is a clear improvement over the sequencing of a single
207 cDNA (**Figure 5C-D**). Most (90.9%) of the samples that were not fully reconstructed were characterized
208 by a low viral load (Ct > 30), but almost half (45%) of the samples in this Ct value range were
209 nevertheless reconstructed optimally (**Figure 5E-F**). In particular, five of the seven viral genomes from
210 swabs with a Ct value ≥ 38 were completely reconstructed (>96.98%), indicating that the outcome is
211 not solely determined by the viral titer in the starting material. In order to generate accurate
212 consensus sequences, we applied the same approach used to identify true-positive iSNVs (only
213 variants in both replicates were included in the final consensus). This approach revealed that 22
214 samples (12.94%), with Ct 25.9-40, would have included at least one false-positive variant in the
215 consensus sequences based on single-cDNA analysis, but these were efficiently removed by
216 considering the concordance between replicates (**Table S13**).

217 **Impact of genome reconstruction accuracy on the evaluation of a potential re-infection case**

218 The identification of SARS-CoV-2 genetic variants at different time points can reveal whether recurrent
219 infections are relapses caused by the same strain or independent infections with a different strain.
220 We therefore evaluated our optimized workflow in a case-study of relapse/re-infection involving a 48-
221 year-old female patient who was hospitalized with mild COVID-19 symptoms following a positive

222 nasopharyngeal swab on 4/3/2020, discharged with no symptoms on 11/3/2020 followed by two
223 consecutive negative swab tests, but readmitted with mild COVID-19 symptoms 12 days later. During
224 the second hospital stay, the nasopharyngeal swab test results fluctuated, and the patient was finally
225 discharged on 21/4/2020 with no symptoms, and two consecutive negative molecular tests. Three
226 swab samples (one from the first and two from the second hospitalization period) were sequenced to
227 identify the viral strain responsible for infection (**Table 1**). All samples were sequenced in duplicate or
228 quadruplicate (**Table S14**), and consensus variants were called in order to identify the viral strains.
229 Depending on the replicate, some consensus variants identified in the first hospitalization period were
230 missing or could not be genotyped in the second hospitalization period, leading to the hypothesis that
231 different strains could be responsible for each infection (**Table 1**). In contrast, when merging
232 sequencing replicates, the same variants were identified in all three samples (**Table 1**) and a very high-
233 frequency (99.95%) false-positive variant could be identified at position 12890 (**Table S13**). Based on
234 this analysis, we concluded that the same viral strain was responsible of both the first and second
235 infection, and that the latter should therefore not be classed as a re-infection.

236 **Table 1. High-frequency variants identified in the COVID-19 relapse case study**

		1° Hospitalization					**2° Hospitalization**							
		05/03/2020					**22/03/2020**			**03/04/2020**				
		Ct 27					**Ct 34**			**Ct 35.7**				
Genome	**Reference allele**	**9075**	**9075**	**9075**	**9075**	**9075**	**9076**	**9076**	**9076**	**9078**	**9078**	**9078**	**9078**	**9078**
Position		**1.1**	**1.2**	**2.1**	**2.2**	**merged**	**1.1**	**1.2**	**merged**	**1.1**	**1.2**	**2.1**	**2.2**	**merged**
241	C	T	T	T	T	T	T	-	T	-	-	-	T	T
3037	C	T	T	T	T	T	-	-	-	-	T	-	-	T
13620	C	T	T	T	T	T	T	-	T	-	-	T	T	T
14408	C	T	T	T	T	T	T	T	T	-	-	-	T	T
23403	A	G	G	G	G	G	G	G	G	-	-	G	-	G
28881	G	A	A	A	A	A	-	A	A	-	A	-	-	A
28882	G	A	A	A	A	A	-	A	A	-	A	-	-	A
28883	G	C	C	C	C	C	-	C	C	-	C	-	-	C

237 The positions of high-frequency variants (>75%) are shown in the consensus sequence of a specimen
238 collected during the first hospitalization. For each of these positions, the genotypes identified in the
239 samples collected during the second hospitalization are also shown. Genotypes are reported for each
240 sequencing replicate independently or after merging all replicates from the same sample (merged).
241 Positions that could not be genotyped are indicated with a dash.

242 **DISCUSSION**

243 **Protocol optimization for simplicity, flexibility, throughput and cost-efficiency**

244 Amplicon-based sequencing (originally called PrimalSeq) is the most sensitive and widely-used
245 protocol for SARS-CoV-2 whole-genome analysis from clinical isolates, but its disadvantages include
246 uneven amplicon coverage and poor accuracy when the viral load is low [23]. We addressed these
247 limits by improving the accuracy and completeness of sequencing, as well as the cost-efficiency and
248 throughput, thus achieving the highly reliable analysis of SARS-CoV-2 genomes. This benchmarking
249 analysis established a robust workflow, ACoRE, that allowed the complete and accurate
250 characterization of SARS-CoV-2 genomes in 170 clinical samples, including a subset (42%) with very
251 low viral titers (Ct ≥ 30). We were also able to properly categories an infection-relapse case study.

252 The protocol optimized by the ARTIC Network for SARS-CoV-2 genome sequencing utilizes a tiling
253 primer scheme generating 400-bp viral amplicons for adaptor ligation and 250PE sequencing [33]. This
254 limits the sequencing options on Illumina platforms because this read type is compatible only with the
255 MiSeq v2 chemistry and NovaSeq6000 SP flow cells. To increase flexibility, we used the NexteraFlex
256 kit to prepare amplicon libraries with shorter inserts (170–200 bp) suitable for 150PE sequencing
257 without loss of performance. This also confers the ability to pool up to 384 samples in a single run
258 using unique dual indexes, reducing costs from €80 per sample to €3.5 on the NovaSeq6000 with S1
259 flow cell or €12 on the NextSeq500 with HighOutput flow cell. Even shorter sequencing reads (100PE)
260 resulted in shorter overlap of paired ends, reducing the number of sequencing fragments required per
261 sample and translating to even lower costs of €3 per sample. Because the NexteraFlex method does
262 not require the quantification of starting amplicons or final sequencing libraries, this further reduces
263 costs and processing time. Further savings could potentially be achieved by using half the volume of

264 tagmentase reagent, but testing is required to ensure that accuracy and coverage is maintained. The
265 generation of amplification replicates from a single starting cDNA (instead of multiple cDNAs, as
266 recommended by the original protocol[23]) would also save time and costs, while preserving the
267 sample for additional tests. The fragmented amplicon approach and other adjustments therefore
268 improved protocol simplicity, flexibility, multiplexing and economy, allowing the cost-effective and
269 timely processing of larger cohorts of samples by ACoRE.

270 **Sequencing multiple replicates to increase accuracy and completeness**

271 Clinical specimens with low viral loads reduce the accuracy of variant calling and the completeness of
272 genome reconstruction, both of which are inversely correlated with the quality and quantity of
273 starting material[23,30,43]. Current guidelines for viral genotyping recommend a lower limit of 1000
274 virus copies per reaction [23,43] but this would rule out a large proportion of clinical samples,
275 including ~53% of the samples in our cohort. A Ct value of ~25 was identified as the median for virus
276 detection in symptomatic patients, with a consistent proportion of samples (15–25%) falling above Ct
277 30 [25,46] . Low viral loads are often found in patients with prolonged COVID-19 infection [47–49],
278 and five of six reported cases of potential re-infection involved samples with Ct values >30 [50], but
279 whole-genome sequencing is nevertheless recommended to differentiate between relapse and new
280 infections caused by a different SARS-CoV-2 variants [50,51]. The ability to sequence SARS-CoV-2
281 genomes in low-titer samples is therefore necessary to track infections and correlate different strains
282 with disease communicability, manifestation and severity.

283 Increasing the depth of sequencing has been proposed as a strategy to achieve complete genome
284 reconstruction in low-titer samples, but this does not overcome limitations caused by missing
285 amplicons [43]. Similarly, improvement in ARTIC primer design and compatibility (currently version 3)
286 can also ameliorate genome coverage, but again cannot make up for missing amplicons [24,30] . We
287 found that only a few specific amplicons were reproducibly suboptimal (64, 70 and 91) whereas most
288 showed coverage variations limited to particular samples or replicates. We therefore merged the

289 sequencing data from two or more replicates as a simple solution to enhance coverage and
290 genotypability, achieving a more homogeneous representation of the viral genome and rescuing the
291 suboptimal samples. The random amplification observed in low-titer samples most likely reflects the
292 low sample complexity rather than poor assay sensitivity or performance. Accordingly, the sampled
293 RNA and corresponding cDNA fragments before amplification are unlikely to represent the complete
294 genome based on our observation that the coverage achieved by sequencing two amplification
295 replicates (each from 5 µL of cDNA) was similar to that achieved with a single amplification starting
296 from double the amount of cDNA (10 µL). Therefore, to optimize genome reconstruction, a single large
297 cDNA batch should be amplified in several parallel reactions, using as much sample volume as possible
298 to increase complexity. The multiple PCR products can then be pooled before library preparation and
299 sequenced as a single sample to avoid increasing costs.

300 As well as improving coverage and genotypability, at least two amplification reactions must be
301 analyzed to achieve accurate variant calling (SNVs and iSNVs). It is well established that the analysis
302 of viral iSNVs down to 3% frequency requires the generation of multiple replicates to distinguish true-
303 positive iSNVs from low-frequency PCR or sequencing errors [23]. In contrast, the generation of
304 consensus sequences for the analysis of SNVs in epidemiological studies requires the identification of
305 the most-frequent nucleotide at each position and is typically based on single replicates [12,45].
306 However, we discovered that consensus sequences also contain frequent SNV errors (>12% in our
307 cohort) and the comparison of technical replicates is required to ensure accuracy. This was not
308 confined to low-titer samples (Ct > 30) but also included some samples with moderate viral loads (Ct
309 = 25–30) potentially leading to the submission of inaccurate consensus sequences to public
310 repositories such as GISAID. These false-positive variants probably arose due to PCR errors because
311 they were not found in other amplification replicates (either from the same or different cDNA).
312 However, studies reporting SARS-CoV-2 consensus sequences thus far have not included the analysis
313 of technical replicates, even in the case of low-titer samples (Ct > 30)[26,52]. The accuracy of SARS-
314 CoV-2 consensus sequences deposited in GSAID has been called into question for documented

315 sequences with putative errors or a significant number of variants in one particular submission
316 (singletons) [35] and the use of stringent filters and bioinformatic tools has been proposed as a
317 solution [52,53]. Instead, with ACoRE we propose the use of replicates as a simple experimental
318 solution to avoid the generation of incorrect consensus sequences prior to database submission.

319 **The assessment of re-infections**

320 Reconstruction of highly accurate sequences from sub-optimal samples was crucial to identify the
321 correct viral strain responsible of a second hospitalization case, that was hypothesized to be a re-
322 infection. A standard workflow would have missed or included incorrect variants in support of such
323 hypothesis, while ACoRE properly recognized that the different time-point samples contained the
324 same viral strain.

325 Another interesting example, that would certainly benefit of ACoRE, comes from a publication that
326 reported the first individual in North America to have symptomatic reinfection with SARS-CoV-2 [26],
327 for whom "…genomic analysis of SARS-CoV-2 showed genetically significant differences between each
328 variant associated with each instance of infection…" suggesting that "…the patient was infected by
329 SARS-CoV-2 on two separate occasions by a genetically distinct virus…" [45]. The viral load of the swab
330 samples analyzed in that study was very low (Ct > 35) based on 14–22 PCR cycles-protocol without
331 amplification replicates, therefore potential false-positive variants and/or regions with low
332 genotypability may have influenced the results. We reanalyzed the data and noted that two of the
333 four variants specifically associated with the first infection had insufficient sequencing coverage to
334 achieve confident variant calling in the sample from the second infection (**Table S15**). In particular,
335 our bioinformatic pipeline revealed that position 539 was covered by only five reads, thus a genotype
336 could not be properly called; while variant 16741G→T (supported by 10 reads) was only just above
337 the genotypability threshold of 8 (**Table S15**). These positions were genotyped using the bioinformatic
338 pipeline utilized by the authors because the limit was set to five reads. Furthermore, variant 4113C→T
339 showed frequency of 67.82% in the first infection, suggesting that two viral strains were already

340 present: a predominant strain carrying the identified variant and a less-abundant strain lacking the
341 variant that became prevalent in the second infection (**Table S15**). However, the absence of replicate
342 analysis makes it impossible to confirm this hypothesis. Similarly, although the final variant
343 (7921A→G) was abundant, the absence of replication makes it impossible to rule out the possibility
344 of an amplification error, as frequently observed in our low-titer samples. These questions could be
345 resolved by sequencing two technical replicates rather than analyzing data from one sequencing
346 library using two different pipelines (as reported by the authors). The conclusions put forward by the
347 authors therefore appear to be only weakly supported by the raw data, but would nevertheless have
348 a major impact on future research by highlighting the possibility of re-infection and thus possibly
349 questioning the efficacy of vaccines. The analysis of such critical samples would greatly benefit from
350 the use of technical replicates, and robust evaluation is particularly important due to the ramifications
351 of the conclusions for the global research and biomedical communities.

352 **CONCLUSIONS**

353 We have optimized ACoRE, a workflow for SARS-COV-2 sequencing to improve flexibility and
354 throughout, thus reducing assay time and costs and facilitating the robust analysis of suboptimal
355 samples that would normally be excluded from sequencing even if they are central and irreplaceable
356 specimens. The sequencing of such low-titer samples without replication risks the generation of
357 consensus sequences containing false-positive SNVs and iSNVs, but we found that the inclusion of
358 technical replicates improves both the accuracy and completeness of viral genome analysis. This
359 reduces the risk of generating inaccurate and incomplete genomic sequences, favoring the submission
360 of robust sequences to public databases and enhancing the downstream analysis of SARS-CoV-2
361 genotyping data.

362 **METHODS**

363 **Clinical samples**

364 178 Nasopharyngeal swabs (eSwab, Copan, Italy) were obtained from 172 COVID-19 patients
365 diagnosed at the Department of Infectious, Tropical Diseases and Microbiology of the IRCCS Sacro
366 Cuore Don Calabria Hospital, qualified for SARS-CoV-2 molecular diagnosis by the regional reference
367 laboratory (Department of Microbiology, University Hospital of Padua). After collection, swabs were
368 stored at 4 °C for a maximum of 48 h and analysed by the routine-used molecular diagnostic method
369 (RT-qPCR as indicated in the following paragraph). The remaining quantity of swab was then aliquoted
370 and preserved at –80 °C. The study was approved by the competent Ethical Committee for Clinical
371 Research of Verona and Rovigo Provinces (Prot N° 39528/2020).

372 **RNA extraction and RT-qPCR analysis**

373 The routine RT-qPCR protocol was based on a recommended test (emergency use authorization)
374 standardized according to I asked what I'm reading WHO guidelines. Briefly, RNA was extracted from
375 200µL of swabs using the automated Microlab Nimbus workstation (Hamilton, Reno, NV, USA) coupled
376 to a Kingfisher Presto system (Thermo Fisher Scientific, Waltham, MA, USA). We also used a
377 MagnaMax Viral/Pathogen extraction kit (Thermo Fisher Scientific) according to the manufacturer's
378 instructions. RT-qPCR was carried out using the CDC 2019-nCoV rRT-PCR Diagnostic Panel assay and
379 protocol [54], targeting the nucleocapsid protein gene regions N1 and N2 (with the human RNAse P
380 gene as the internal control) on a CFX96 Touch system (Bio-Rad Laboratories, Milan, Italy) with white
381 plates. The amplification cycle threshold (Ct) was determined using CFX Maestro (Bio-Rad
382 Laboratories), setting a baseline threshold at 200 relative fluorescence units (RFU). A standard curve
383 from 5 to 500 genome copies per reaction was performed with serial dilution of the CDC control
384 plasmid containing the complete nucleocapsid gene of SARS-CoV-2 (**Table S1**).

385 **Reverse transcription and amplification of the SARS-CoV-2 genome**

386 Samples with Ct values of 15–18 were diluted 10-fold as suggested by the ARTIC Network [27]. RNA
387 from swab samples (5 µL) was first incubated with 1 µL of 60 µM Random Primer Mix (New England
388 Biolabs, Ipswich, MA, USA) and 1 µL of 10 mM dNTPs (New England Biolabs) at 65 °C for 5 min followed

389 by 1 min on ice to anneal the primers. We then added 4 µL of 5× SSIV buffer, 1 µL of 100 mM DTT, 1
390 µL of 40 U/µL RNaseOUT, 1 µL of 200 U/µL SSIV enzyme (Thermo Fisher Scientific) and 6 µL nuclease-
391 free water (total reaction volume = 20 µL) and heated the reaction to 23 °C for 10 min, 52 °C for 10
392 min and 80 °C for 10 min. We generated two or three cDNAs from each sample (depending on the
393 experiment), each of which was amplified 2–3 times using the ARTIC protocol. In each case, we mixed
394 2.5 or 5 µL cDNA (depending on the experiment) with 3.7 µL of 10 µM primer pools A and B from the
395 ARTIC nCoV-2019 V3 panel (IDT, Coralville, IA, USA), 12.5 µL Q5 high-fidelity DNA polymerase 2× (New
396 England Biolabs) for each of the primer pools, and nuclease-free water to a final volume of 25 µL. The
397 reaction was heated to 98 °C for 30 s, followed by 25 cycles (sample Ct ≤ 21) or 35 cycles (sample Ct
398 > 21) of 98 °C for 15 s and 65 °C for 5 min. The PCR products were then combined in a single tube,
399 cleaned up using 1× AMPure XP beads (Beckman Coulter, Brea, CA, USA) and eluted in 15 µL of water.
400 The resulting amplicons were analyzed on the 4150 TapeStation System (Agilent Technologies, Santa
401 Clara, CA, USA) and quantified using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific).

402 **Full-length amplicon sequencing**

403 Libraries were prepared from 50 ng of virus amplicons using the KAPA Hyper prep kit and unique dual-
404 indexed adapters (5 µL of a 15 µM stock) according to the supplier's protocol (Roche, Basel,
405 Switzerland). The post-ligation products were cleaned up using 0.8× AMPure XP beads followed by
406 library amplification (six cycles) with the KAPA Library Amplification Primer Mix (Roche). After a clean-
407 up with 1× AMPure XP beads, the libraries were analyzed on the 4150 TapeStation System (average
408 size 526–573 bp) and quantified using the Qubit dsDNA BR Assay kit (Thermo Fisher Scientific).
409 Barcoded libraries were pooled at equimolar concentrations and sequenced on the MiSeq platform
410 (Illumina, San Diego, CA, USA) with Miseq Reagent kit v2 to generate 250-bp paired-end (250PE) reads.

411 **Fragmented amplicon sequencing**

412 Libraries were prepared from 10 µL of purified viral amplicons using the NexteraFlex kit (Illumina)
413 according to the manufacturer's recommendations, and combinatorial dual indexes were added in six

414 cycles of PCR. We cleaned up 10-µL aliquots of each amplified library using a 1:1 ratio of sample
415 purification beads (Illumina) and eluted the purified library in 20 µL of resuspension buffer (Illumina).
416 The resulting libraries were analyzed on the 4150 TapeStation System (average size 335–369 bp),
417 pooled and quantified using the Qubit dsDNA BR Assay kit. The libraries were sequenced on a Novaseq
418 6000 device (Illumina) using an SP flow cell to generate 100-bp paired-end (100PE) reads, or on a
419 NextSeq500 (Illumina) to generate 150-bp paired-end (150PE) reads.

420 **Data filtering and reference genome alignment**

421 Full-length amplicon sequencing data were randomly downsampled using *seqtk sample v1.3*
422 (https://github.com/lh3/seqtk). To compare sequencing data from the full-length and fragmented
423 amplicons, KAPA library reads were downsampled at the same mean mapped coverage as the
424 corresponding NexteraFlex replicates using *sambamba v0.6.7* [55]. To simulate sequencing using
425 100PE reads, data from the fragmented amplicon libraries were trimmed using a custom script. All
426 sequencing datasets were trimmed for quality and adapters were removed using *Trimmomatic v0.39*
427 [56] with the following parameters: *ILLUMINACLIP:adapters_file:2:30:10 LEADING:5 TRAILING:5*
428 *SLIDINGWINDOW:4:20*. Filtered reads were aligned to the SARS-CoV-2 reference genome (GenBank
429 ID: *MN908947.3*) using *BWA MEM v0.7.17* [57] with default parameters and the relative alignment
430 file was converted to a binary alignment map (BAM) file using *SAMtools v1.9* [58]. For the fragmented
431 libraries, duplicate reads were identified and discarded using *Picard v2.21.1*
432 (http://broadinstitute.github.io/picard). Subsequently, *iVar v1.2.2 trim* [23] was used to remove ARTIC
433 v3 primer sequences from the BAM files. For the fragmented libraries, the *-e* parameter was used to
434 include reads without primers. Finally, overlapping portions of reads were clipped using *fgbio ClipBam*
435 *v1.1.0* (https://github.com/fulcrumgenomics/fgbio) with the following parameters: *--clip-overlapping-*
436 *reads -c Hard* to avoid counting multiple reads representing the same fragment. Coverage and
437 genotypability statistics were calculated from the BAM files using *bedtools genomecov v2.19.1* [59]

438 and *GATK CallableLoci v3.8* [60], respectively. Raw genomic sequencing data were deposited in NCBI
439 GenBank (BioProject no PRJNA690890).

440 **Consensus variant calling and generation of the consensus sequence**

441 A pileup was calculated for each position in the BAM file of each replicate using the *SAMtools v1.9*
442 *mpileup* option with parameters *-aa -A -d 0 -Q 0*. The resulting files were used as input for *iVar*
443 *consensus v1.2.2* [23] to generate consensus sequences, considering those positions covered by at
444 least three reads (parameters: *-t* 0 -m 3). The most abundant nucleotide for each position was
445 reported in the consensus sequence, whereas positions covered by fewer than three reads or
446 reporting an equal proportion of nucleotides were represented by the ambiguous character N.

447 To call variants present in the consensus sequences (consensus variants), sequences were aligned to
448 the SARS-CoV-2 reference genome using *Minimap v2.17* [61] and the alignment file was converted to
449 the BAM format using *SAMtools v1.9*. Consensus variants were then called using *bcftools call v1.10.2*
450 [58] with the following parameters: *--ploidy 1 -A -m -P 0.05 -M -Oz*.

451 Final consensus sequences from the cohort of 170 samples and the relapse case were called after
452 merging sequencing data for each individual replicate. False-positive variants in the consensus
453 sequence were identified manually by comparing the presence of discordant iSNVs at the same
454 genomic position between replicates of the same sample and considering only positions genotyped in
455 both replicates. False-positive variants were removed from consensus sequences and replaced with
456 the reference allele.

457 **iSNV variant calling**

458 Alignment BAM files were used to call iSNVs present in each replicate with a minimum minor allele
459 frequency (MAF) threshold of 3%. Joint variant calling of the 30 entire amplicon libraries, and between
460 replicates of the same sample for fragmented amplicon libraries, was achieved by generating a pileup
461 using S*AMtools mpileup v1.9* [58] with the following parameters: *-A -d 600000 -B -Q 0*. The output file

462 was used to detect iSNVs with *VarScan mpileup2cns v2.3.9* [62] and the following parameters: *--min-*
463 *var-freq 0.03 --min-avg-qual 20*.

464 For each sample, inter-replicate discordant variants were identified by iSNV variant calling after
465 merging sequencing data from all replicates, considering only genotyped positions. A discordant
466 variant was defined as a variant called in one replicate, whereas the same position in the other
467 replicate reported the reference allele.

468 **Calculation of the concordance rate**

469 The concordance rate (R_c) between replicates samples was calculated as follows:

470

$$\mathrm{R_C} = \frac{N_c}{Mean(N_1, N_2)}$$

471

472 *Nc* represents (i) the number of shared variants (consensus variants or iSNVs) excluding positions that
473 could not be genotyped in at least one replicate, or (ii) the number of shared genotypable bases,
474 excluding positions marked N in at least one replicate, or (iii) the number of shared amplicons with
475 coverage higher than three reads in all replicates. *N1* and *N2* represent the total number of iSNVs,
476 consensus variants, genotypable bases or covered amplicons detected in each of the two samples in
477 the analysis. R_c was calculated by comparing couples of replicates generated from the same cDNA
478 (intra-cDNA concordance) and triplets of replicates generated from different cDNAs (inter-cDNA
479 concordance) as shown in **Table S2**.

480 **Statistical analysis**

481 The non-parametric Wilcoxon signed rank test and the Mann Whitney U-test were used to compare
482 matched pairs and non-matched data, respectively. The non-parametric Friedman test was used to
483 compare multiple paired groups. Significance of pairing was confirmed by calculating Spearman's rho.
484 We used GraphPad Prism 6.0 (GraphPad Software, San Diego, CA, USA) for all statistical analysis, with
485 a significance threshold of $p < 0.05$.

486 **DECLARATIONS**

487 **Ethics approval and consent to participate**

488 The study was approved by the competent Ethical Committee for Clinical Research of Verona and
489 Rovigo Provinces (Prot N° 39528/2020)

490 **Consent for publication**

491 Not applicable.

492 **Availability of data and materials**

493 The raw reads dataset supporting the conclusions of this article is available at the NCBI SRA repository
494 under BioProject ID PRJNA690890.

495 **Competing interests**

496 The authors declare that they have no competing interests

497 **Funding**

498 The work performed at IRCCS Sacro Cuore Don Calabria Hospital was supported by the Italian Ministry
499 of Health “Fondi Ricerca corrente—L1P5”.

500 **Authors' contributions**

501 The study was conceived and coordinated by MD and MR. The samples and RNA extraction and RT-
502 qPCR analysis were provided and performed by CP, AM, EP, MD, SS, ZB. Reverse transcription,
503 amplification of the SARS-CoV-2 genome, library preparation and sequencing were performed by CB,
504 CDE, VG, EM, MP, ES, BG. The data filtering and reference genome alignment, the consensus variant
505 calling and generation of the consensus sequence, the iSNV variant calling, the calculation of the
506 concordance rate and the statistical analysis were performed by LM, GL, SM, DL, BI, MG. The

507 manuscript was written by MR with input from all co-authors. All authors read and approved the final
508 manuscript

509 **Acknowledgements**

510 We gratefully acknowledge the Centro Piattaforme Tecnologiche (CPT) for granting access to the
511 genomic facility of University of Verona for sequencing on a MiSeq and NextSeq500 Illumina platform,
512 and Dr. Richard M Twyman (www.twymanrm.com) for editing the manuscript text.

513 **Authors' information**

514 1Department of Biotechnology, University of Verona, Strada le Grazie 15, 37134 Verona, Italy

515 2Department of Infectious and Tropical Diseases and Microbiology, IRCCS Sacro Cuore Don Calabria
516 Hospital, Negrar di Valpolicella, 37024 Verona, Italy

517 3Department of Diagnostics and Public Health, University of Verona, 37134 Verona, Italy

518 4Genartis srl, via IV Novembre 24, 37126 Verona, Italy

519

520 **REFERENCES**

521 1. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins
522 University (JHU) [Internet]. Available from: https://coronavirus.jhu.edu/map.html

523 2. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human
524 respiratory disease in China. Nature. 2020;579:265–9.

525 3. GISAID Initiative [Internet]. Available from: https://www.gisaid.org/

526 4. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-
527 CoV-2 fitness. Nature [Internet]. Springer US; 2020; Available from:
528 http://dx.doi.org/10.1038/s41586-020-2895-3

529 5. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in
530 SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell.
531 2020;182:812-827.e19.

532 6. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic
533 diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol [Internet]. Elsevier;
534 2020;83:104351. Available from: https://doi.org/10.1016/j.meegid.2020.104351

535 7. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral

536 Infectivity and Antigenicity. Cell [Internet]. Elsevier; 2020;182:1284-1294.e9. Available from:
537 http://dx.doi.org/10.1016/j.cell.2020.07.012

538 8. Yao H, Lu X, Chen Q, Xu K, Chen Y, Cheng M, et al. Patient-derived SARS-CoV-2 mutations impact
539 viral replication dynamics and infectivity in vitro and with clinical implications in vivo. Cell Discov
540 [Internet]. Springer US; 2020;6:1–16. Available from: http://dx.doi.org/10.1038/s41421-020-00226-1

541 9. Rahman MS, Islam MR, Alam ASMRU, Islam I, Hoque MN, Akter S, et al. Evolutionary dynamics of
542 SARS-CoV-2 nucleocapsid protein and its consequences. J Med Virol. 2020;

543 10. Geoghegan JL, Ren X, Storey M, Hadfield J, Jelley L, Jefferies S, et al. Genomic epidemiology
544 reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. medRxiv
545 [Internet]. 2020;2020.08.05.20168930. Available from:
546 https://www.medrxiv.org/content/10.1101/2020.08.05.20168930v3

547 11. Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, Gray KA, et al. Revealing COVID-19 transmission
548 in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med. 2020;

549 12. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread
550 of SARS-CoV-2 in the Icelandic population. N Engl J Med. 2020;382:2302–15.

551 13. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid
552 SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in
553 the Netherlands. Nat Med. 2020;26:1405–10.

554 14. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing.
555 Nat Rev Microbiol. Nature Publishing Group; 2017;15:183–92.

556 15. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance
557 elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (80-).
558 2014;345:1369–72.

559 16. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic processes
560 constrain the within and between host evolution of influenza virus. Elife. 2018;7:1–19.

561 17. Gardy J, Loman NJ, Rambaut A. Real-time digital pathogen surveillance - the time is now.
562 Genome Biol [Internet]. Genome Biology; 2015;16:15–7. Available from:
563 http://dx.doi.org/10.1186/s13059-015-0726-x

564 18. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola Virus Epidemiology,
565 Transmission, and Evolution during Seven Months in Sierra Leone. Cell. 2015;161:1516–26.

566 19. Dube Mandishora RS, Gjøtterud KS, Lagström S, Stray-Pedersen B, Duri K, Chin'ombe N, et al.
567 Intra-host sequence variability in human papillomavirus. Papillomavirus Res [Internet]. Elsevier B.V.;
568 2018;5:180–91. Available from: https://doi.org/10.1016/j.pvr.2018.04.006

569 20. Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, Mentis A. SARS-CoV-2 exhibits
570 intra-host genomic plasticity and low-frequency polymorphic quasispecies. J Clin Virol [Internet].
571 Elsevier B.V.; 2020;131:104585. Available from: https://doi.org/10.1016/j.jcv.2020.104585

572 21. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic diversity of SARS-CoV-2 in COVID-19
573 patients. 2019;1–27.

574 22. Sashittal P, Luo Y, Peng J, El-Kebir M. Characterization of SARS-CoV-2 viral diversity within and
575 across hosts. 2020;

576 23. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-
577 based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and

iVar. Genome Biol. Genome Biology; 2019;20:1–19.

24. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv Prepr Serv Biol [Internet]. 2020; Available from: http://www.ncbi.nlm.nih.gov/pubmed/32908977%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7480024

25. Walsh KA, Jordan K, Clyne B, Rohde D, Drummond L, Byrne P, et al. SARS-CoV-2 detection, viral load and infectivity over the course of an infection. J Infect. Elsevier Ltd; 2020;81:357–71.

26. Lescure FX, Bouadma L, Nguyen D, Parisey M, Wicky PH, Behillil S, et al. Clinical and virological data of the first cases of COVID-19 in Europe: a case series. Lancet Infect Dis. 2020;20:697–706.

27. ARTIC Network. Available from: https://artic.network/ncov-2019

28. Li C, Debruyne D, Spencer J, Kapoor V, Liu L, Zhou B, et al. Highly sensitive and full-genome interrogation of SARS-CoV-2 using multiplexed PCR enrichment followed by next-generation sequencing. 2020;

29. Resende PC, Motta FC, Roy S, Appolinario L, Fabri A, Xavier J, et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. 2020;1–11.

30. Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PLoS One [Internet]. 2020;15:1–11. Available from: http://dx.doi.org/10.1371/journal.pone.0239403

31. McNamara RP, Caro-Vegas C, Landis JT, Moorad R, Pluta LJ, Eason AB, et al. High-Density Amplicon Sequencing Identifies Community Spread and Ongoing Evolution of SARS-CoV-2 in the Southern United States. Cell Rep [Internet]. ElsevierCompany.; 2020;33:108352. Available from: https://doi.org/10.1016/j.celrep.2020.108352

32. Klempt P, Brož P, Kašný M, Novotný A, Kvapilová K, Kvapil P. Performance of targeted library preparation solutions for SARS-CoV-2 whole genome analysis. Diagnostics. 2020;10:1–12.

33. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12:1261–6.

34. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, et al. Stability of SARS-CoV-2 phylogenies [Internet]. PLOS Genet. 2020. Available from: http://dx.doi.org/10.1371/journal.pgen.1009175

35. Rayko M, Komissarov A. Quality control of low-frequency variants in SARS-CoV-2 genomes. 2020;

36. Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front Microbiol. 2020;11:1–13.

37. Moreno G, Braun K, Halfmann P, Prall T, Riemersma K, Haj A, et al. Limited SARS-CoV-2 diversity within hosts and following passage in cell culture. 2020;

38. Andrés C, Garcia-Cehic D, Gregori J, Piñana M, Rodriguez-Frias F, Guerrero-Murillo M, et al. Naturally occurring SARS-CoV-2 gene deletions close to the spike S1/S2 cleavage site in the viral quasispecies of COVID19 patients. Emerg Microbes Infect. 2020;9:1900–11.

39. Liu T, Chen Z, Chen W, Chen X, Hosseini M, Yang Z, et al. A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples. bioRxiv [Internet].

620 2020;2020.11.10.375022. Available from: https://doi.org/10.1101/2020.11.10.375022

621 40. Doddapaneni H, Cregeen SJ, Sucgang R, Meng Q, Qin X, Avadhanula V, et al. Oligonucleotide
622 capture sequencing of the SARS-CoV-2 genome and subgenomic fragments from COVID-19
623 individuals. bioRxiv [Internet]. 2020;2020.07.27.223495. Available from:
624 https://doi.org/10.1101/2020.07.27.223495

625 41. Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, et al. Genomic Epidemiology of SARS-CoV-2 in
626 Guangdong Province, China. Cell. 2020;181:997-1003.e9.

627 42. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole genome
628 sequencing of sars-cov-2: Adapting illumina protocols for quick and accurate outbreak investigation
629 during a pandemic. Genes (Basel). 2020;11:1–13.

630 43. Kubik S, Marques AC, Xing X, Silvery J, Bertelli C, De Maio F, et al. Guidelines for accurate
631 genotyping of SARS-CoV-2 using amplicon-based sequencing of clinical samples. bioRxiv [Internet].
632 2020;2020.12.01.405738. Available from:
633 http://biorxiv.org/content/early/2020/12/01/2020.12.01.405738.abstract

634 44. Torres D de A, Ribeiro L do CB, Riello AP de FL, Horovitz DDG, Pinto LFR, Croda J. Reinfection of
635 COVID-19 after 3 months with a distinct and more aggressive clinical presentation: Case report. J
636 Med Virol. 2020;

637 45. Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, et al. Genomic evidence for
638 reinfection with SARS-CoV-2: a case study. Lancet Infect Dis [Internet]. Elsevier Ltd; 2020;21:52–8.
639 Available from: http://dx.doi.org/10.1016/S1473-3099(20)30764-7

640 46. Buchan BW, Hoff JS, Gmehlin CG, Perez A, Faron ML, Munoz-Price LS, et al. Distribution of SARS-
641 CoV-2 PCR cycle threshold values provide practical insight into overall and target-Specific sensitivity
642 among symptomatic patients. Am J Clin Pathol. 2020;154:479–85.

643 47. Zhang RZ, Deng W, He J, Song YY, Qian CF, Yu Q, et al. Case Report: Recurrence of Positive SARS-
644 CoV-2 Results in Patients Recovered From COVID-19. Front Med. 2020;7:1–5.

645 48. Li Q, Zheng XS, Shen XR, Si HR, Wang X, Wang Q, et al. Prolonged shedding of severe acute
646 respiratory syndrome coronavirus 2 in patients with COVID-19. Emerg Microbes Infect. 2020;9:2571–
647 7.

648 49. Zapor M. Persistent Detection and Infectious Potential of SARS-CoV-2 Virus in Clinical Specimens
649 from COVID-19 Patients. Viruses. 2020;12:1–17.

650 50. Brief TA. European Centre for Disease Prevention and Control. Reinfection with SARS-CoV:
651 considerations for public health response: ECDC; 2020. 2020; Available from:
652 https://www.ecdc.europa.eu/en/publications-data/threat-assessment-brief-reinfection-sars-cov-2

653 51. Lu J, Tillett R, Long Q, Kong H, Kong H, Kong H, et al. COVID-19 reinfection: are we ready for
654 winter? EBioMedicine. 2020;62.

655 52. Voloch CM, da Silva Jr RF, P de Almeida LG, Brustolini OJ, Cardoso CC, Gerber AL, et al. Intra-host
656 evolution during SARS-CoV-2 persistent infection. medRxiv [Internet]. 2020;2020.11.13.20231217.
657 Available from: https://doi.org/10.1101/2020.11.13.20231217

658 53. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2
659 genomes. Proc Natl Acad Sci U S A. 2020;117:9241–3.

660 54. CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel [Internet].
661 Available from: https://www.fda.gov/media/134922/download

662 55. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment
663 formats. Bioinformatics. 2015;31:2032–4.

664 56. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data.
665 Bioinformatics. 2014;30:2114–20.

666 57. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
667 2013;00:1–3. Available from: http://arxiv.org/abs/1303.3997

668 58. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
669 population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.

670 59. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features.
671 Bioinformatics. 2010;26:841–2.

672 60. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
673 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
674 Genome Res [Internet]. 2010/07/19. Cold Spring Harbor Laboratory Press; 2010;20:1297–303.
675 Available from: https://pubmed.ncbi.nlm.nih.gov/20644199

676 61. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

677 62. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation
678 and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–
679 76.

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695 **FIGURES**

696

697

698

699

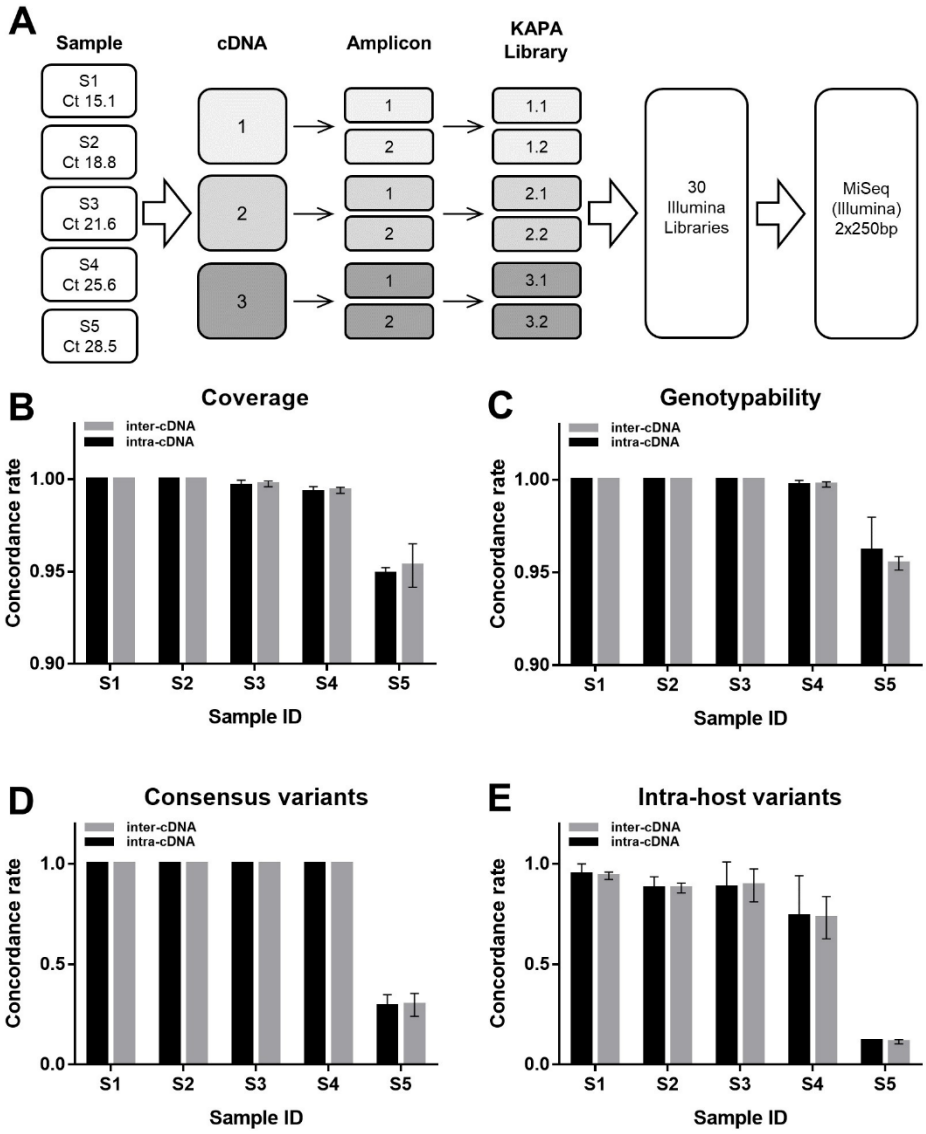
700

701

702

703

704



705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720 **Figure 1. Comparison of intra-cDNA and inter-cDNA replicates of SARS-CoV-2 genome amplification**
721 **and sequencing. (A)** Schematic diagram showing the five clinical samples obtained from COVID-19
722 patients, their RT-qPCR Ct values and the experimental workflow. For each sample, we generated
723 three independent cDNAs and each cDNA was amplified in duplicate using the ARTIC nCoV-2019 V3
724 Panel. Amplicons used as the input for library preparation were sequenced in 250PE mode on the
725 Illumina MiSeq platform. The bar charts show mean concordance rates (± standard deviations) for **(B)**
726 genome coverage, **(C)** genotypability, **(D)** consensus variants and **(E)** iSNV between amplification
727 replicates generated from different cDNAs (inter-cDNA) or the same cDNA (intra-cDNA).

728

729

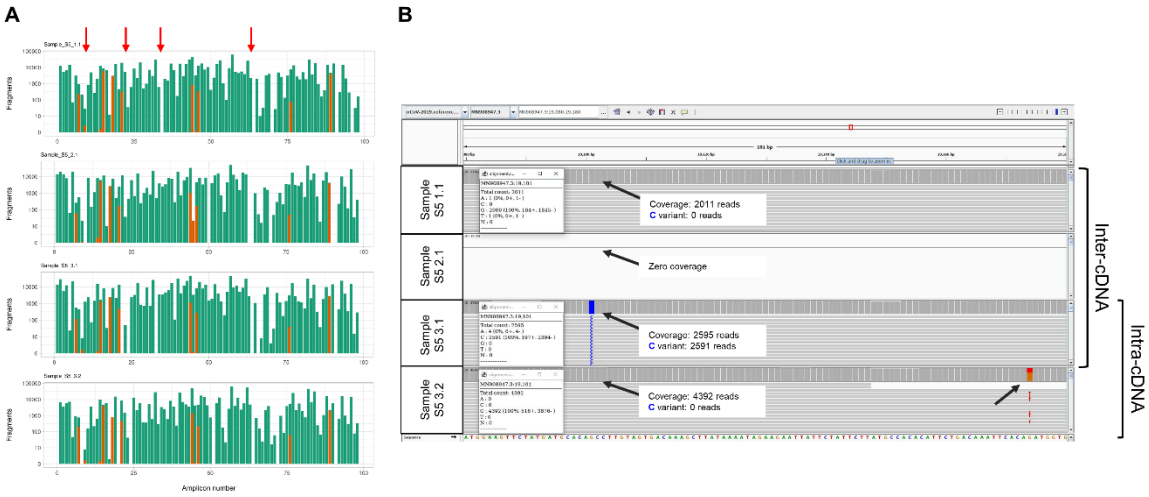


Figure 2. Coverage and variant calling between intra-cDNA and inter-cDNA replicates. (A) Sequencing coverage of the 98 amplicons of ARTIC V3 panel from four representative replicates of sample S5. Green bars represent the amplicons generated using the ARTIC original primer set, and orange bars represent the amplicons generated using the alternative V3 primers. Red arrows point at representative amplicons missing in only one replicate. **(B)** Integrative Genomics Viewer (IGV) visualization of four representative sequencing replicates of sample S5 in the region 19,080–19,180 of the SARS-Cov-19 genome. Black arrows indicate variants called only in one replicate. The amplicon was not amplified in replicate S5 2.1.

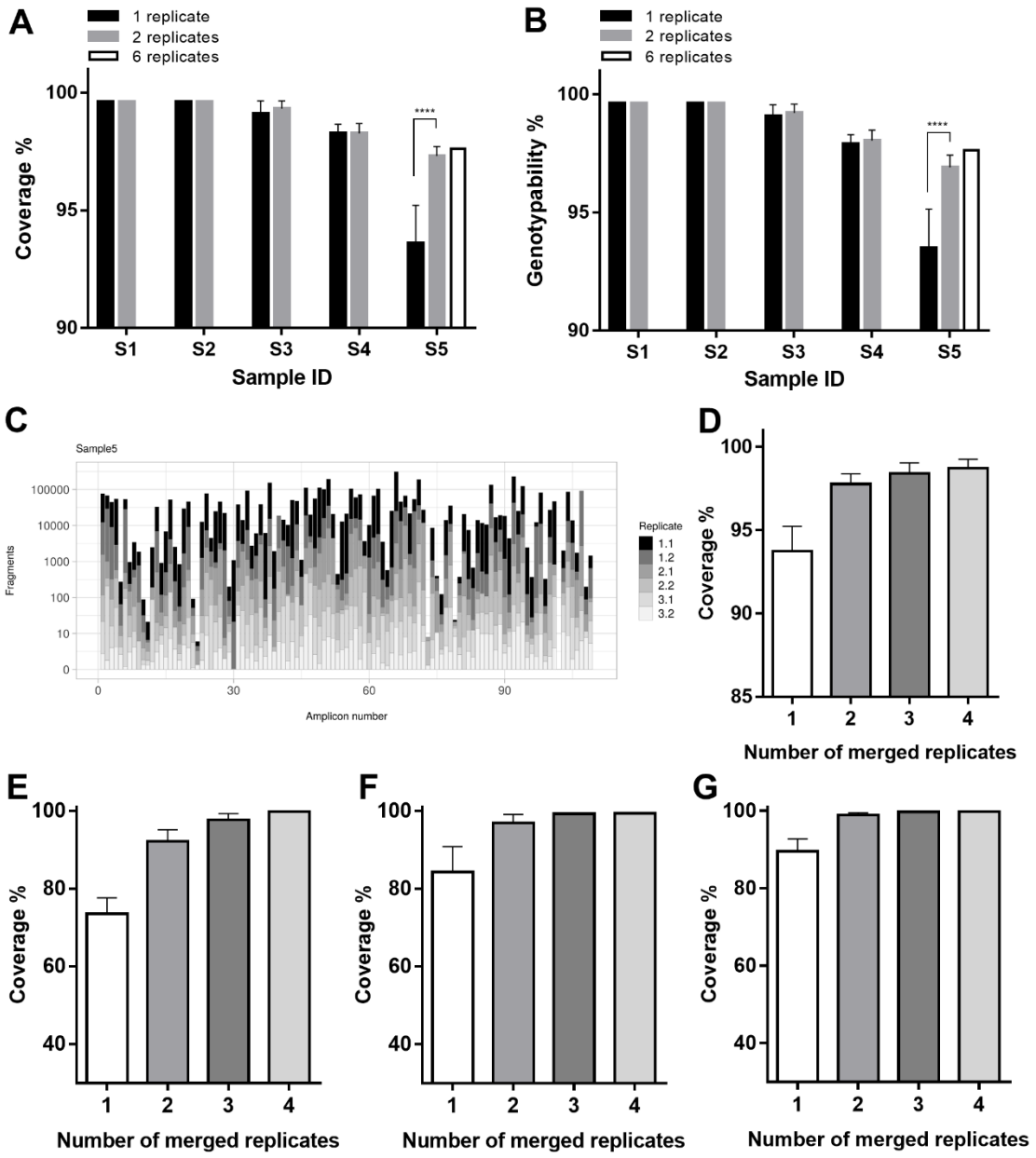


Figure 3. Merging sequencing replicates can improve coverage and genotypability. (A) Mean percentage genome coverage (± standard deviations). **(B)** Mean percentage genotypability (± standard deviations). Both genome coverage and genotypability were calculated for single replicates or after merging all possible combinations of two or six replicates, starting from the same total sequencing reads (****$p < 0.0001$, Mann Whitney U-test). **(C)** The coverage fraction contributed by each of the six replicates generated from sample S5. **(D)** Percentage of genome coverage after merging different numbers of replicates from sample S5, and from three other COVID-19-positive swab samples, namely samples 3270 **(E)**, 4572 **(F)**, 4173 **(E)**, whose sequencing results are reported in **Table S12**.

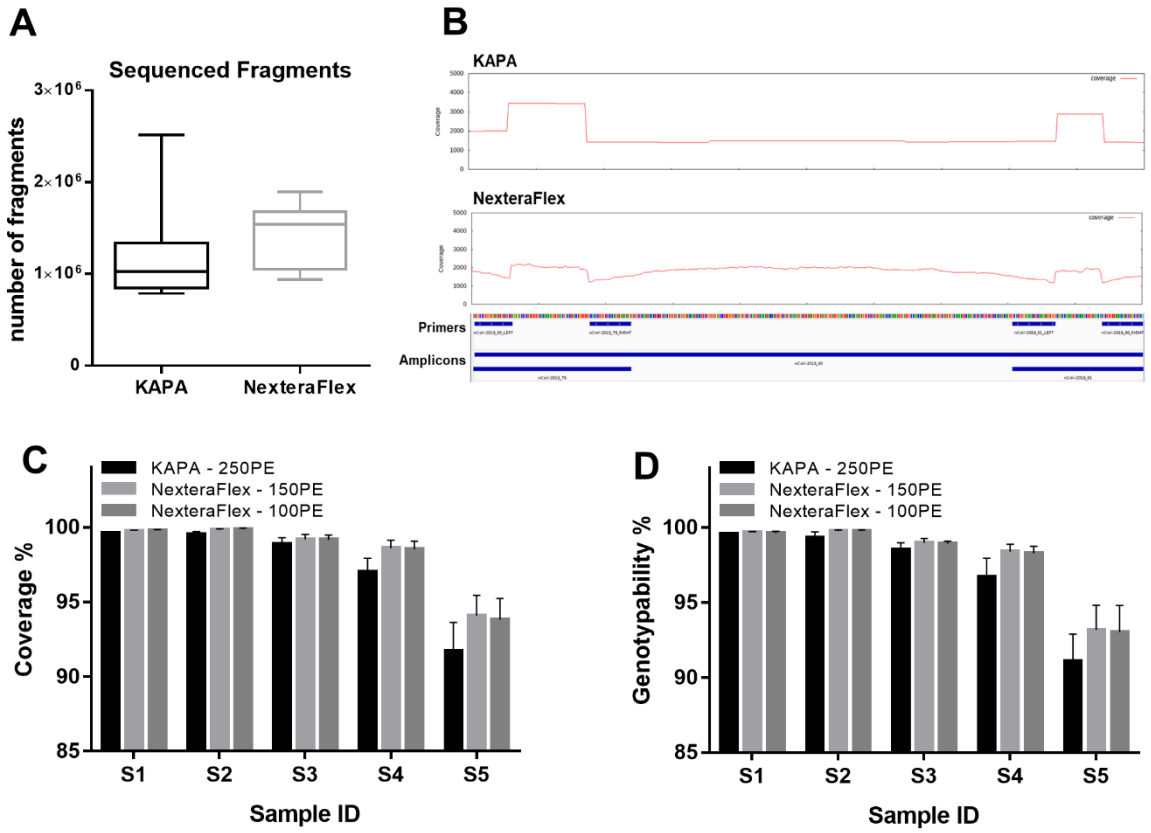


Figure 4. Comparison of SARS-CoV-2 sequencing and mapping results obtained using the KAPA and NexteraFlex library preparation kits. (A) Distribution of the number of fragments generated using the KAPA and NexteraFlex kits for the same set of 30 replicates. **(B)** Visualization of mean sequencing coverage on a representative ARTIC amplicon using the KAPA and NexteraFlex library kits. Given the overlap with adjacent amplicons, the 5′ and 3′ ends show increased coverage. **(C)** Mean coverage (± standard deviations) and **(D)** mean genotypability (± standard deviations) of sequencing libraries prepared from the 30 replicates using either the KAPA or NexteraFlex kits. The 100PE results were obtained from the 150PE dataset by *in silico* trimming.

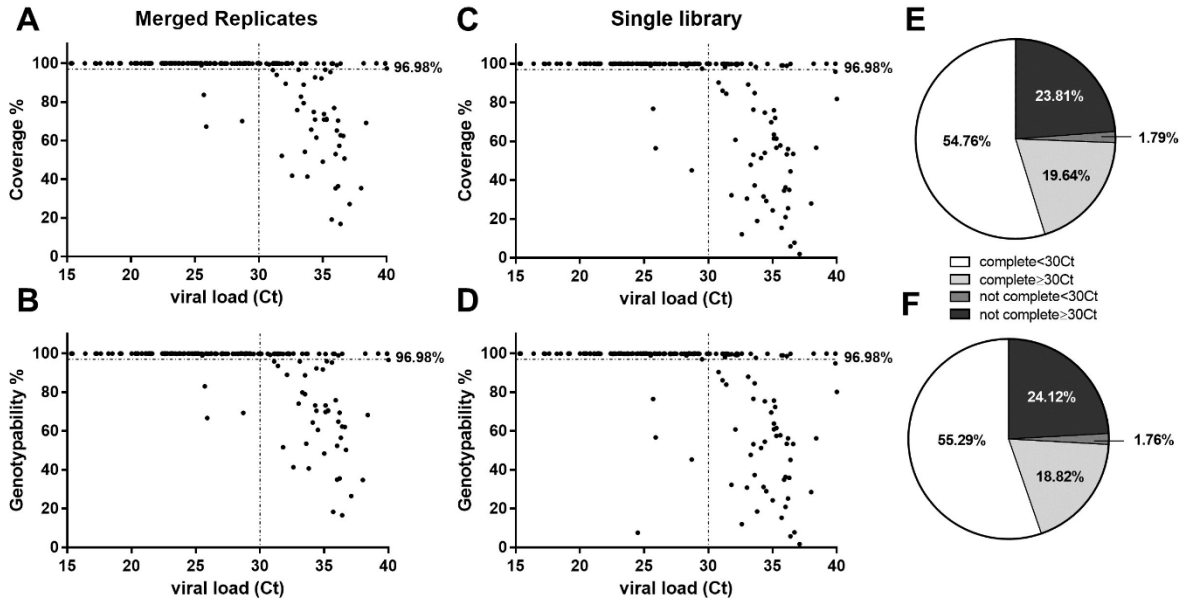


Figure 5. SARS-CoV-2 sequencing in a cohort of clinical samples with wide range of viral titers. (A-C) Percentage of genome coverage and **(B-D)** genotypability for each sample (N = 170) considering a single replicate (selected randomly) or after merging two sequencing replicates. The pie charts show the fraction of the complete SARS-CoV-2 (>96.98%) genome in terms of **(E)** coverage or **(F)** genotypability for samples with Ct < or ≥ 30.