Novel SARS-CoV-2 Whole-genome sequencing technique using Reverse Complement PCR enables fast and accurate outbreak analysis

Femke Wolters[1*a], Jordy P.M. Coolen[1a], Alma Tostmann[1], Lenneke F.J. van Groningen[2], Chantal P. Bleeker-Rovers[3], Edward C.T.H. Tan[4,5], Nannet van der Geest-Blankert[6], Jeannine L.A. Hautvast[7], Joost Hopman[1], Heiman F.L. Wertheim[1], Janette C. Rahamat-Langendoen[1], Marko Storch[8], Willem J.G. Melchers[1].

[a]Shared first authorship

[1]Department of Medical Microbiology and Radboudumc Center for Infectious Diseases, Radboud university medical center, Nijmegen, The Netherlands

[2]Department of Haematology, Radboud university medical center, Nijmegen, The Netherlands

[3]Department of Internal Medicine, division of Infectious Diseases, Radboud university medical center, Nijmegen, The Netherlands

[4]Department of Emergency Medicine, Radboud university medical center, Nijmegen, The Netherlands

[5]Department of Surgery, Radboud University Medical Center, Nijmegen, The Netherlands

[6]Department of Occupational Health, Radboud university medical center, Nijmegen, The Netherlands

[7]Regional Public Health Service, Nijmegen, the Netherlands

[8]London Biofoundry, Imperial College Translation & Innovation Hub, White City Campus, 84 Wood Lane, London, W12 0BZ, UK;

***Corresponding author:**

Femke Wolters, MD

Radboud university medical center, Geert Grooteplein Zuid 10, 6525 GA, Nijmegen, The Netherlands

Femke.wolters@radboudumc.nl

Word count: 3761

Abstract

Current transmission rates of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) are still increasing and many countries are facing second waves of infections. Rapid SARS-CoV-2 whole genome sequencing (WGS) is often unavailable but could support public health organizations and hospitals in monitoring and determining transmission links. Here we report a novel reverse complement polymerase chain reaction (RC-PCR) technology for WGS of SARS-CoV-2. This technique is unique as it enables library preparation in a single PCR saving time, resources and enables high throughput screening. A total of 173 samples tested positive for SARS-CoV-2 between March and September 2020 were included. RC-PCR WGS applicability for outbreak analysis in public health service and hospital settings was tested on six predefined clusters containing samples of healthcare workers and patients. RC-PCR resulted in WGS data for 146 samples. It showed a genome coverage of up to 98,2% for samples with a maximum Ct value of 32. Three out of six suspected clusters were fully confirmed, while in other clusters four healthcare workers were not associated. Importantly, a previously unknown chain of transmission was confirmed in the public health service samples. These findings confirm the reliability and applicability of the RC-PCR technology for SARS-CoV-2 sequencing in outbreak analysis and surveillance.

Introduction

In December 2019 China reported a group of patients with a severe respiratory illness caused by a thus far unknown coronavirus. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified as the causative agent.[1] Since its outbreak, the virus evolved into a pandemic with almost 37 million infections and over a million deaths worldwide by October 2020.[2] Many countries are currently fighting second waves of infection whilst the healthcare systems are still under pressure from the first wave. To reduce spread and mitigate workforce depletion, large scale testing of healthcare workers (HCW) was implemented in the Netherlands early on.[3]

Current testing is based on RT-PCR detection of SARS-CoV-2 in nasopharynx or oropharyngeal swabs. If tested SARS-CoV-2 positive, HCW are instructed to self-isolate at home, and source finding and contact tracing is performed. These procedures enable us to identify patients and personnel at risk

of infection and to identify chains of transmission in the hospital. In the community setting, source finding and contact tracing is performed by public health staff upon a notification of a SARS-CoV-2 positive individual. It facilitates the implementation of quarantine measures for high risk contacts in the community. Contact tracing is time consuming and with rising numbers of infections as currently seen in the second wave, the public health capacity may reach the limits of feasibility of thorough source and contact tracing investigations.[4] Routine sequencing the SARS-CoV-2 genome from positive samples provides crucial insights into viral evolution and supports outbreak analysis.[5,6] Current whole-genome sequencing (WGS) workflows often require cumbersome preparation, are laborious to implement for high throughput screening or use less widely accessible sequencing platforms, preventing widespread implementation. Here we present a novel strategy for fast, simple and robust Next-Generation Sequencing (NGS) WGS library preparation. We show that the RC-PCR method, which integrates tiled target amplification with Illumina library preparation has a simple workflow with minimal hands-on time. We used this novel and practical method to I) validate and compare it with another sequence technology to demonstrate its reliability and capacity and II) apply it to a set of epidemiologically linked cases to illustrate its added value in detecting potential transmission events in public health and hospital settings.

Material and Methods

In this study we conducted a validation to assess the performance and reproducibility of the novel RC-PCR SARS-CoV-2 sequencing technology. Subsequently, we performed a clinical validation to assess the potential added value in identifying chains of transmission in a hospital and public health setting.

Sample collection

Nasopharyngeal and oropharyngeal swabs collected in UTM or GLY medium of patients, healthcare workers and samples for the local public health services that were tested for SARS-CoV-2 in our laboratory. Samples collected between March 2020 and September 2020 were included in this study and stored at -80°C. Detailed descriptions on included samples can be found in supplementary table 1. A total of 173 SARS-CoV-2 positive and fifteen SARS-CoV-2 negative samples were tested.

86

87 **Samples and selection of epidemiological clusters**

88 Nineteen out of 188 samples were previously sequenced using Oxford Nanopore Technologies (ONT).
89 These nineteen samples were collected at the beginning of the pandemic, between March 9th and
90 March 20th and ONT sequencing data of these samples has been deposited at GISAID, a global
91 initiative curating sequenced SARS-CoV-2 genomes for public access (https://www.gisaid.org/).[6]

92

93 *Hospital samples*

94 Six epidemiolocal hospital clusters that were identified by the infection prevention and control (IPC)
95 team were included in this study. These clusters involved patients admitted at and healthcare workers
96 (HCW) employed by the Radboud university medical center. Of the identified clusters, three were
97 clusters of healthcare workers with an epidemiological link, and three involved a patient and several
98 healthcare workers with a suspected epidemiological link. To determine whether other HCW could be
99 linked to one of the clusters, samples of sporadic HCW (all other HCW who tested positive for SARS-
100 CoV-2 in September 2020) were included in the selection, as the second wave of infections in the
101 Netherlands started late August 2020. These consist of Radboud university medical center HCW and
102 the majority work in direct or indirect patient care. A minority of positive samples include employees
103 working at the medical faculty or research departments. Additionally, twenty samples were included
104 from patients and HCW who were tested between March and September 2020 and who were not
105 associated with any of these predefined clusters.

106

107 *Community samples*

108 We also included an additional 64 community samples that tested positive for SARS-CoV-2 in March
109 and April 2020 and that were tested by the local public health service. These were samples of persons
110 living in the defined public health region surrounding our hospital. See Table 1 for an overview of the
111 groups and clusters.

112 The Research Ethics Committee of the region Arnhem/Nijmegen reviewed the current study and
113 waived additional ethical approval. All personal data of patients, HCW and public health service
114 samples was anonymized. Cluster information was provided anonymously by the IPC team.

Table1: number of groups and clusters of samples that were sequenced for SARS-CoV-2.

Groups	**Samples (N)**	**Month SARS-CoV-2 PCR positive**	**IPC cluster information**
Oxford Nanopore Technology (ONT)	19	March 2020	None
Cluster 1 – External outbreak link	6	September 2020	HCW linked to a known community outbreak and who had either visited the venue or had close contact to people (with positive test) who had visited the venue
Cluster 2 – Department C	5	September 2020	All HCW working at the same department in close proximity and who tested positive in the same week.
Cluster 3 – Patient ward E	2^	September 2020	A patient and an HCW; the HCW had contact with the patient without adequate personal protective equipment (PPE).
Cluster 4 – Patient ward H	3*	May 2020	Two HCW and one patient tested positive at the same department in a short time period. An epidemiological link was suspected since the employees came in contact with the patient.
Cluster 5 – Laboratory R	9	April 2020	All HCW working at the same department, tested positive in the same week.
Cluster 6 – Patient ward S	6	September 2020	One patient and 5 HCW, the HCW tested positive 5 days after being in contact with the positive patient, the event included an unexpected aerosol generating procedure and HCW were not protected with PPE.
Sporadic HCW September 2020	39	September 2020	none
Public Health services samples	64	March & April 2020	none
Other *(patients/employees tested up to September 2020)*	20^	March – September 2020	none
Negative	15	n.a.	n.a.
Total	188		

115

*^1 sample was isolated and sequenced twice, *2 samples were isolated and sequenced twice*

116

117 **Real-Time Polymerase Chain Reaction**

118 SARS-CoV-2 RT-PCR was performed on all samples during routine diagnostics. RNA was isolated
119 using Roche COBAS 4800 (Roche Diagnostics Corporation) with a CT/NG extraction kit according to
120 the manufacturers protocol. RT-PCR with primers targeting the envelope (E-gene) was used as
121 described by Corman *et al.* and performed on a LightCycler 480 (Roche Diagnostics Corporation)
122 using Roche Multiplex RNA Virus Mastermix.[7]

123

124

125 **Reverse Complement Polymerase Chain Reaction**

126 For all 188 selected samples, RNA isolation was repeated on the MagnaPure 96 (Roche Diagnostics
127 Corporation) using Small Volume isolate protocol with 200µl of sample and eluting isolated RNA in
128 50µl. cDNA-synthesis was performed using Multiscribe RT (Applied Biosystems) with 10µl of RNA
129 input (supplementary table 2). Four samples were replicates, RNA was isolated twice and tested in two
130 separate sequencing runs. They were randomly selected for the first run, but were also part of an IPC
131 identified cluster and therefore included in the second run.

132 Whole genome sequencing (WGS) was performed in 3 independent runs (96 samples each) using the
133 novel EasySeq™ RC-PCR SARS-CoV-2 WGS kit (NimaGen BV, Nijmegen, The Netherlands).
134 Figure 1 and 2 show a detailed description of the technology in which two types of oligo's are used to
135 start the targeted amplification. The RC-probe and the universal barcoding primer hybridize and start
136 the formation of specific SARS-CoV-2 primers with Unique Dual Index (UDI) and adapter sequences
137 already included. In contrast to other techniques where multiple steps are needed to add sequence
138 adapters and UDI's. This means a regular PCR-system can be used to produce SARS-CoV-2 specific
139 amplicons ready for sequencing. The kit uses 155 newly designed probes with a tiling strategy
140 previously implemented in the ARTIC protocol.[8] The probes are divided in two pools, A and B. Pool
141 A contains 78 probes and Pool B contains 77 probes. This strategy requires two separate RC-PCR
142 reactions but ensures there is minimal chance of forming chimeric sequences or other PCR artifacts
143 (See Figure 2). After the PCR, samples of each plate are pooled into an Eppendorf tube, resulting in
144 two tubes, for pool A and B, respectively. These are individually cleaned using AmpliClean™
145 Magnetic Bead PCR Clean-up Kit (NimaGen, Nijmegen, The Netherlands). Afterwards, quantification
146 using the Qubit double strand DNA (dsDNA) High Sensitivity assay kit on a Qubit 4.0 instrument
147 (Life Technologies) is performed and pool A and B are combined. The amplicon fragment size in the
148 final library will be around 435 bp. Next Generation Sequencing (NGS) was performed on an Illumina
149 MiniSeq® using a Mid Output Kit (2x150-cycles) (Illumina, San Diego, CA, USA) by loading 0.8 pM
150 on the flowcell. The first two runs (Run1 and Run1_new) were conducted to test the performance of
151 the RC-PCR on a large variety of Ct-values (Ct 16 – 41) using the standard protocol provided by
152 NimaGen. For sequencing Run1_new the RC-PCR product from Run1 was re-used and sequenced

with the exception that the final sequencing library was created by using a balanced library pooling strategy based on estimated cDNA input (2 ul for Ct<20, 5 ul 20≤Ct<27 or 10ul Ct≥27). The final sequence run (Run2) contains samples with a Ct range from 16 – 32, using the same Ct dependent balanced library strategy.

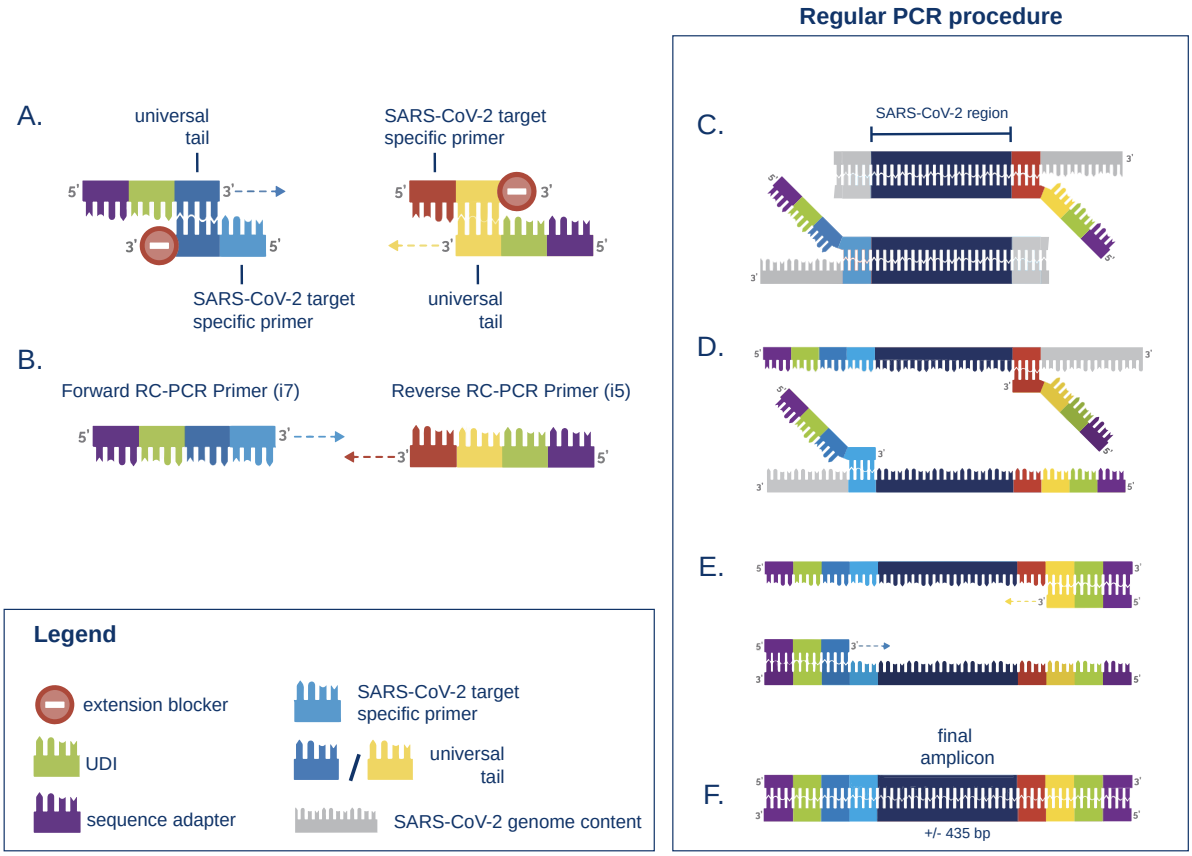

Figure 1. Schematic representation of the RC-PCR technology to WGS SARS-CoV-2. The protocol consists of one single PCR-like reaction consisting of 2 steps. The schematic is adapted from Kieser *et al.* (Kieser *et al.*, 2020) **A.** Two types of oligo's are present, 1) the universal barcoding primer which includes a Unique Dual Index (UDI), sequence adapter, and universal tail. 2) the RC probe which contains an extension blocker, universal sequence, and the reverse complement of the SARS-CoV-2 genomic target sequence. **B.** The universal tail sequences anneal and form a SARS-CoV-2 specific PCR primer. **C - E.** A regular PCR in which the SARS-CoV-2 specific amplicons are created. **F.** The final amplicons are ready to sequence on an Illumina sequencer.

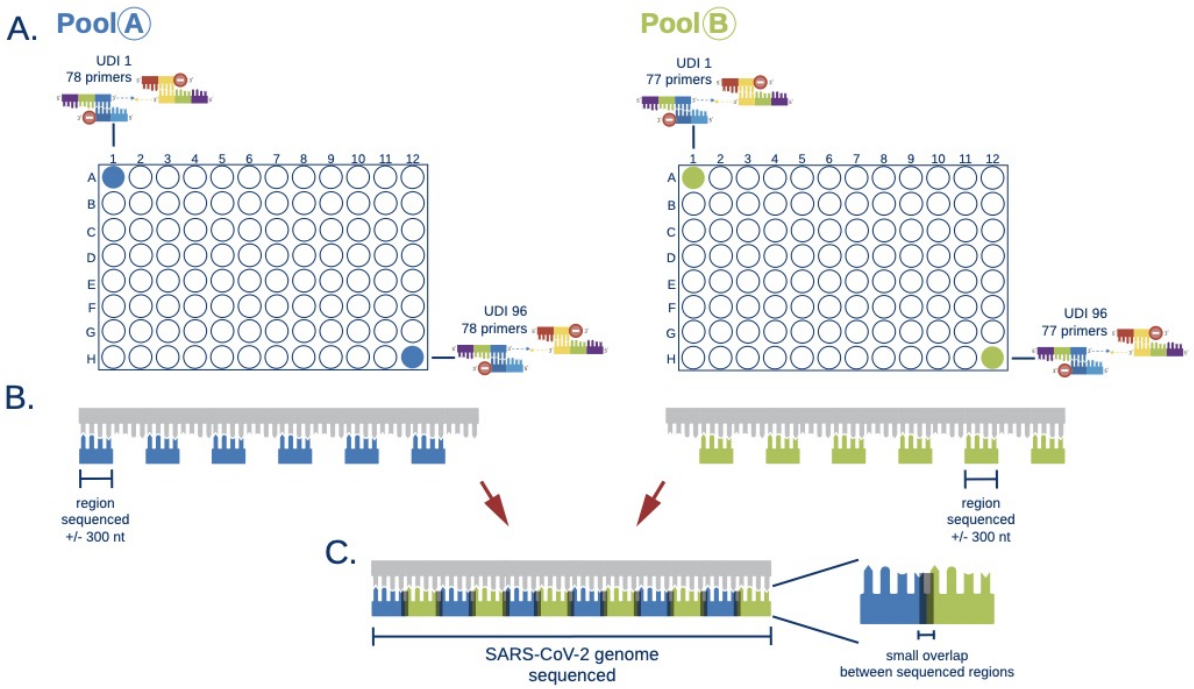

Figure 2. Schematic overview of the EasySeq SARS-CoV-2 WGS process. **A.** For 96 samples the kit provides a Pool A plate and a Pool B plate. These plates contain the RC-probe and the universal oligo's with an UDI per well. Identical UDI's between Pool A and B are used. **B.** For covering the full SARS-CoV-2 genome a tiling method similar to the ARTIC protocol is used.(DNA Pipelines R&D, 2020) The kit provides two distinct plates to separate into amplicon pools that do not overlap. This greatly enhances accurate sequencing output. **C.** The same wells of pool A and B share the same index and allow the combination of corresponding sequencing results to cover the full SARS-CoV-2 genome.

Data analysis

VirSEAK (JSI, Ettenheim, Germany) was used to map the Illumina paired-end reads to SARS-CoV-2 reference NC_045512.2. Consensus sequences were extracted for each sample using the virSEAK export option, settings used can be found in supplementary table 3. All consensus sequences and reference NC_045512.2 were aligned using MUSCLE (version 3.8.1551) using default settings.[9] Sequence statistics were calculated using faCount (version 377). Mean read depth (RD) was calculated using JSI/SEQUENCE PILOT (JSI, Ettenheim, Germany) to evaluate the amplicon depth of each of the 155 amplicons. For the validation samples (ONT group Table 1) the sequence starts and ends were trimmed to match RC-PCR region with Oxford Nanopore region. A maximum-likelihood phylogenetic tree was inferred using IQ-TREE (version 2.0.3) under the GTR + F + I + G4 model with the ultrafast bootstrap option set to 1,000. Phylogenetic tree visualization and annotation was performed using iTOL (version 5.6.3) or FigTree (version 1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/).[10] SNP distances between samples was calculated using snp-dists (version 0.7.0) (https://github.com/tseemann/snp-dists). From the genome alignments we calculated a minimum spanning tree (MST) by applying the MSTreeV2 algorithm using GrapeTree (version 1.5.0).[11] Visualization of the MST was performed using GrapeTree.

176 The clinical validation consisted of a comparison of the epidemiological information of the
177 community and hospital samples and the WGS findings to see whether sequencing confirmed or
178 dismissed the suspected links between the samples.
179
180 **Results**
181 **Technical results RC-PCR**
182 In this study we performed three Illumina MiniSeq Mid Output (2x150 bp) runs containing 96 samples
183 each that were prepared using the EasySeq^{TM} RC-PCR SARS-CoV-2 WGS kit. It has a turnaround
184 time of about 8.5 hours, consisting of 1-hour hands-on time for preparing 96 samples, 6.5 hours for
185 performing the RC-PCR, and 1-hour of hands-on time for pooling, sample clean-up. Run 2 had the
186 highest number of positive SARS-CoV-2 VirSEAK consensus retrievals (100%). Of Run1 65% was
187 retrieved, Run1_new 67%. Run2, containing samples with higher viral loads (Ct values 16-32),
188 reached an average coverage of 96.69%. Genome coverage for Run1_new was 88%. (Figure 3B)
189 Supplementary table 4 provides a detailed overview of the technical results of the three sequence runs.
190
191 *Amplicon depth plots*
192 The amplicon depth distribution highlights which parts of the SARS-CoV-2 genome are represented
193 and the number of reads for each of the amplicons. In essence this shows how well the individual parts
194 of the SARS-CoV-2 genome are represented in the results. To illustrate the amplicon distribution on
195 the SARS-CoV-2 genome, for each of the 155 amplicons a sequencing depth was calculated and
196 plotted per run (See Figure 3A). Most amplicons are centered around a Mean read depth (RD) of 100-
197 1000. While some amplicons show less depth, in most cases they still result in a consensus sequence.
198 Additionally, for Run2 the interquartile range of the Mean RD is smaller compared to the two other
199 runs. When comparing the amplicon depth obtained per probe, boxplots are made for each run divided
200 in three Ct groups (Ct<20, 20≤Ct<27, and Ct≥27) (see Figure3C). We see a decline in depth for
201 samples with Ct above 27. For Run1_new and Run2 samples with a Ct between 20 and 27 perform
202 slightly better than the Ct<20 group this is probably an effect of the balanced library input strategy
203 applied for these runs. To evaluate if the impact of amplicon sequencing depth affects SARS-CoV-2

genome completeness, boxplots with the Ct groups are displayed to the effect on genome coverage (see Figure 3D). Here we notice a decline in genome coverage with increasing Ct values for Run1 and Run1_new. Run2 maintains high genome coverages, however does not contain samples with Ct values above 32.

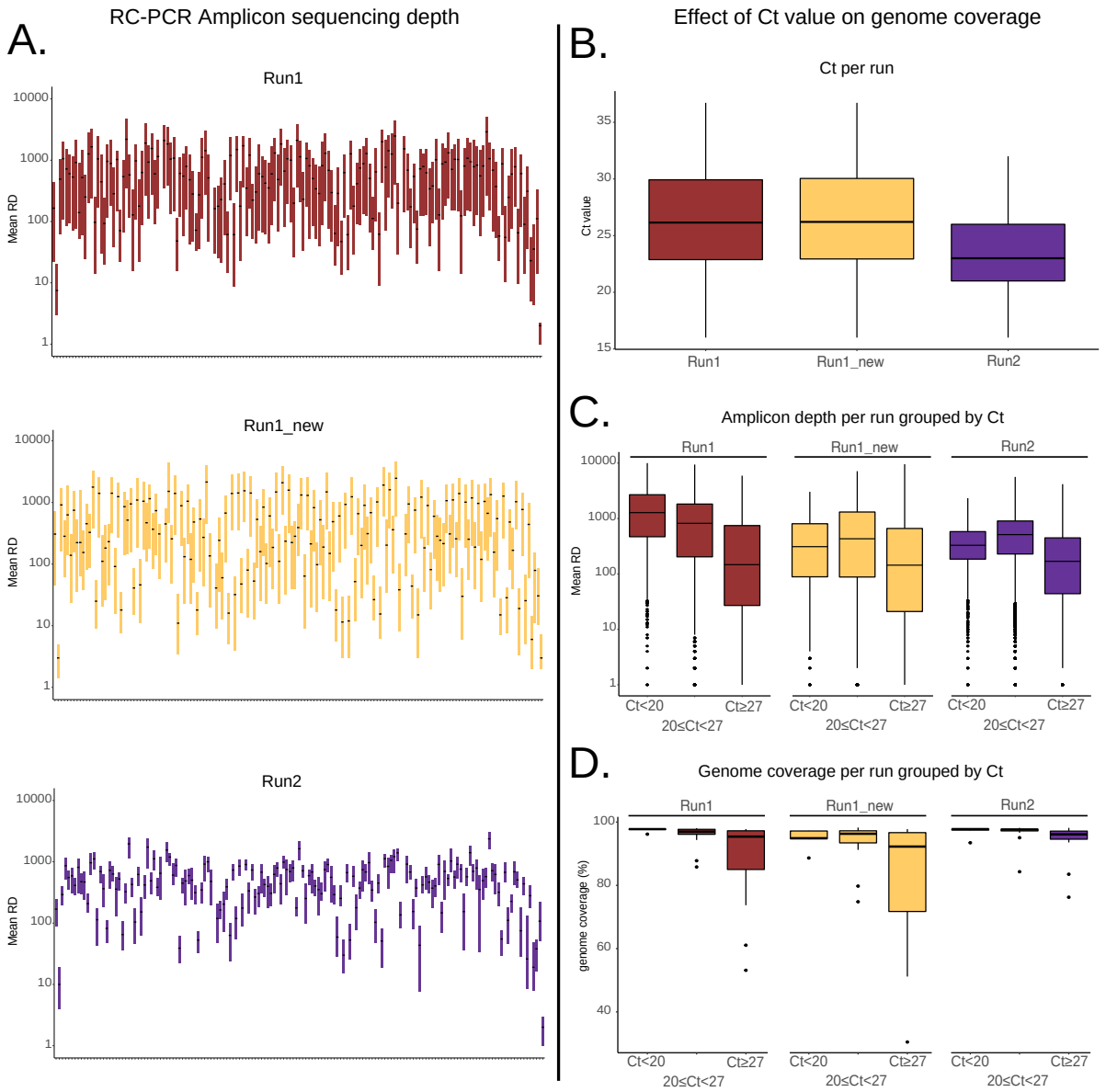

Figure 3. Graphical representation of performance of the RC-PCR Illumina sequence runs. **A.** Boxplots of the interquartile range of the Mean read depth (RD) of the Amplicons on a $\log_{10}$ scale for all 155 probes sorted on the SARS-CoV-2 genome. **B.** Boxplot of the Ct value as determined by RT-PCR to illustrate the differences of viral load of the sample per run. **C.** Boxplots of the Mean RD of the amplicons ($\log_{10}$ scale used) grouped per run and by Ct value. **D.** Boxplots of the SARS-CoV-2 genome coverage as achieved by RC-PCR grouped per run and Ct value.

Regions of low sequencing coverage

In a detailed analysis of the coverage of the SARS-CoV-2 genome obtained by RC-PCR 5 missing genomic regions were observed (Table 2). The largest missing region has a length of 186 bp and is

213 part of the Open Reading Frame 1a (ORF1a). A further two regions are the start (1-54 bp) and the end
214 (46-165bp) of the genome. We observed that region 14585-14725 is missing in the VirSEAK
215 consensus output but not in the JSI/SEQUENCE PILOT and at the time of writing the manuscript the
216 VirSEAK algorithm was updated to improve the consensus output. Overall, without this update, the
217 maximum SARS-CoV-2 genome coverage that can be achieved using RC-PCR is between 97,8% and
218 98,2%. In version 1 of the EasySeq™ RC-PCR SARS-CoV-2 WGS kit three probe pairs do not
219 produce amplicons, 6258_6426, 9504_9752, and 21241_21420, respectively. No data on these
220 genomic regions will be obtained (Table 2).
221

Table 2: missing regions in VirSEAK consensus output

VirSEAK consensus output		**JSI/SEQUENCE PILOT**		
Genomic location	**Length (bp)**	**Probes**	**Genomic location**	**Length (bp)**
1 - 54	54	No Probe		
6309 - 6407	99	6258_6426	6204 - 6372	169
9554 – 9739	186	9504_9753	9450 - 9699	250
14585 – 14725	141			
21322 - 21331	10	21241_21420	21187 - 21366	180
29739/29756/29858 – 29903	165/148/46	29630_29857	29576 - 29803	228
Total base-pairs	655/638/536			

222
223
224
225 **Validation of RC-PCR reproducibility**
226 All samples from Run1 that obtained a consensus (n=57) were compared to the same 57 samples from
227 Run1_new to determine whether results are reproducible when repeating sequencing with the RC-PCR

228 product. Results in supplemental figure 1 show that 50 of the 57 clusters fully align between Run1 and
229 Run1_new. There are 7 samples in which the phylogenetic distance is larger. For those samples in
230 which the phylogenetic distance is larger than expected, alignments were analyzed. The samples from
231 Run1_new show a lower genome coverage, explaining larger phylogenetic distances in these cases.
232 This is in line with the results observed in table 3 with average genome coverage of 88% in Run1_new
233 versus 93% in Run1. Which is either caused by RC-PCR product storage or the influence of the
234 balanced library pooling strategy based on Ct values of the samples.
235 Four sample pairs were tested in both Run1 and Run2 to serve as biological replicates. The entire
236 process from RNA isolation to sequence analysis was performed twice on these four samples.
237 Phylogenetic analysis depicted in Figure 4 (Illumina biological replicates) shows perfect agreement
238 between these repeats and confirms the specificity and reproducibility of RC-PCR.
239
240

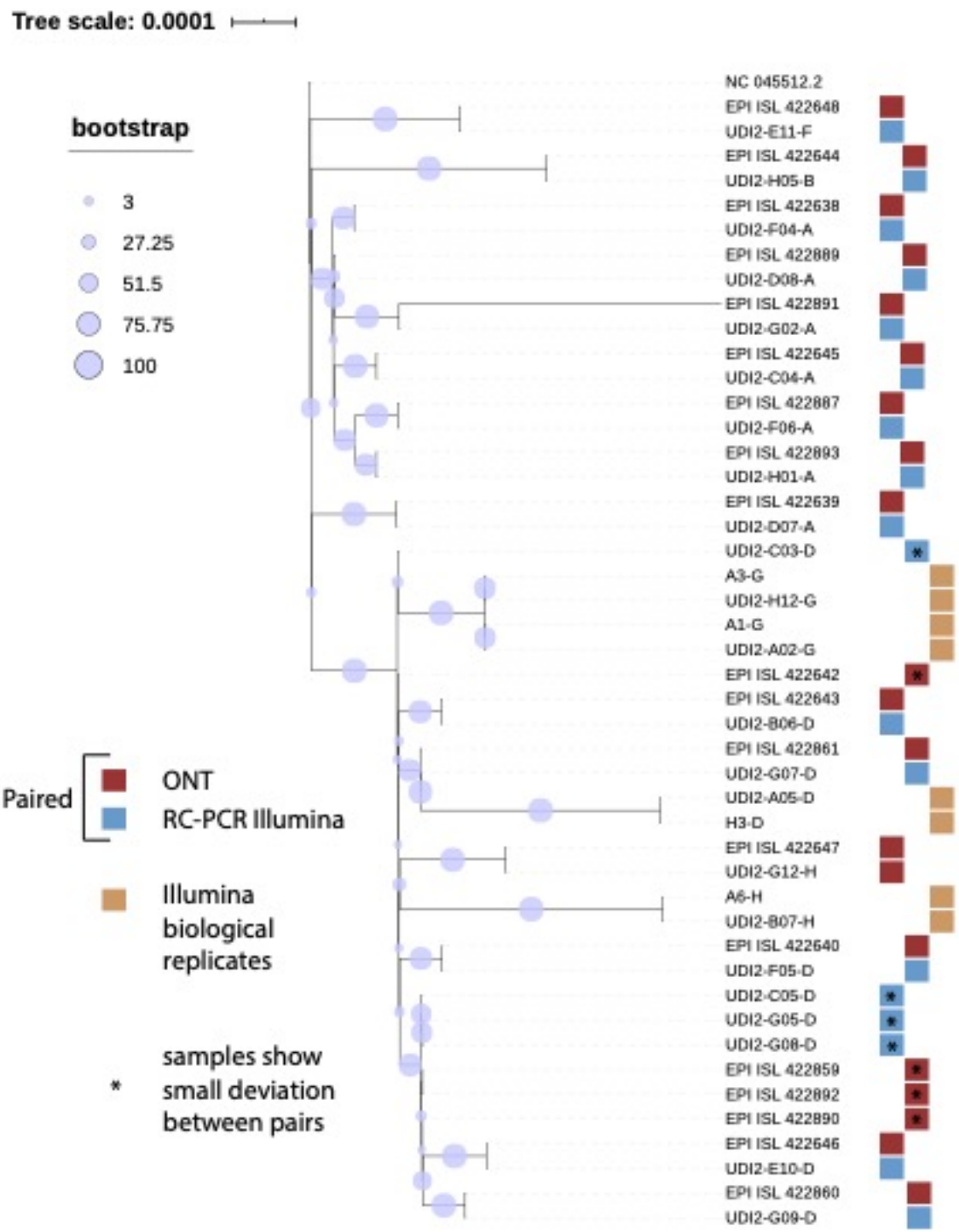

Figure 4. Validation of RC-PCR by comparing the results for the same 19 samples sequenced previously by ONT.(Munnink *et al.*, 2020) Additionally, reproducibility was tested by applying RC-PCR on 4 samples as biological replicates (beige).

241

242

243

244

245 **Validation of RC-PCR with Oxford Nanopore Technologies® (ONT)**

246 Nineteen out of the 188 samples were tested using both ONT and Illumina® sequencing. The ONT
247 sequences were available in the GISAID database and compared to the results of RC-PCR sequencing.
248 All nineteen samples provided sequencing results on both platforms (Figure 4, ONT in red and RC-
249 PCR in blue). Fourteen out of nineteen samples provided perfect pairs, four samples show a small
250 divergence in the phylogenetic tree. Single nucleotide polymorphism (SNP) distance was calculated to
251 identify the number of nucleotides discrepant between samples. This in combination with manual
252 inspection showed that they have identical sequences but RC-PCR samples miss certain genomic
253 regions compared to ONT which results in the phylogenetic differences. One pair does not match, the
254 ONT sample shows a large distance (EPI ISL 422891). Manual inspection of the alignment revealed a
255 wrongly placed ambiguous region in the ONT sample.

256

257 **Clinical validation**

258 Of the 188 tested samples, 173 were SARS-CoV-2 positive of which sequencing results were obtained
259 for 146 (57 in Run1 and 89 in Run2). All samples, excluding nineteen ONT and four duplicate
260 samples used for validation, are depicted in the phylogenetic tree of Figure 5. Only HCW and patients
261 are included in the minimum spanning tree of Figure 6. Figure 5 shows the genetic diversity of the
262 samples at different time points during the pandemic. Those collected during the first months of March
263 and April (community samples from public health service) are clearly separated from the other
264 samples, especially compared to the samples from September 2020 (Cluster 1,2,3,6, and the HCW).
265 In Figure 6 it is clear the epidemiological link between the samples three of the six clusters was
266 completely confirmed by the sequencing results. Clusters two, five, and six contained HCW that were
267 not related. In cluster one, linked to a venue outside the hospital, five samples group together with no
268 SNP distances, one sample has a distance of a single SNP suggesting the possibility of linked cases.
269 However, multiple "sporadic HCW" tested in September and two HCW previously linked to cluster
270 two and five also group within cluster one.
271 In cluster two only two samples group together, two others are genetically unrelated samples and one
272 samples has a SNP distance of 2 which could still be within the transmission chain. Cluster three, a

patient and HCW show a distance of only 1 SNP. Sample collection was performed on one occasion, twelve days apart, which could account for the SNP difference. In cluster four two HCW and a patient group together, confirming the suspected link. Cluster five, an outbreak at a laboratory, eight HCW samples have identical SARS-CoV-2 genomes, only one sample is phylogenetically linked. Cluster six originated from a SARS-CoV-2 positive patient seen at a department, where at that time multiple HCW had close contact to the patient. At the time of presentation, no symptoms were present that were indicative of SARS-CoV-2 and screening using a questionnaire was negative. Five of the HCW tested SARS-CoV-2 positive in the following weeks. In four HCW a genetically similar SARS-CoV-2 virus was detected. Surprisingly multiple other HCW group in this same cluster with minimal differences (0-3 SNPs), which could mean the outbreak was larger than anticipated or the patient was not the source of the infection.

Even though no new clusters were identified among the “sporadic HCW”, they do group with previously identified clusters. Additional information about these HCW revealed that many of them had a direct or indirect link to the community source that was known by the public health services, Cluster one.

Sequencing of the 64 community samples showed seven people clustered together in the phylogenetic tree of Figure 5. There was no prior information available on these tested persons, but additional information provided by the Local Public Health Service indicated that two of the seven worked at the same location, two were their partners, the others lived in the same neighbourhood at the initial four people, although they had no known epidemiological link to these people other than the area of residence. Of other public health service samples no contact tracing information was available and other samples clustering could not be confirmed with an epidemiological link.

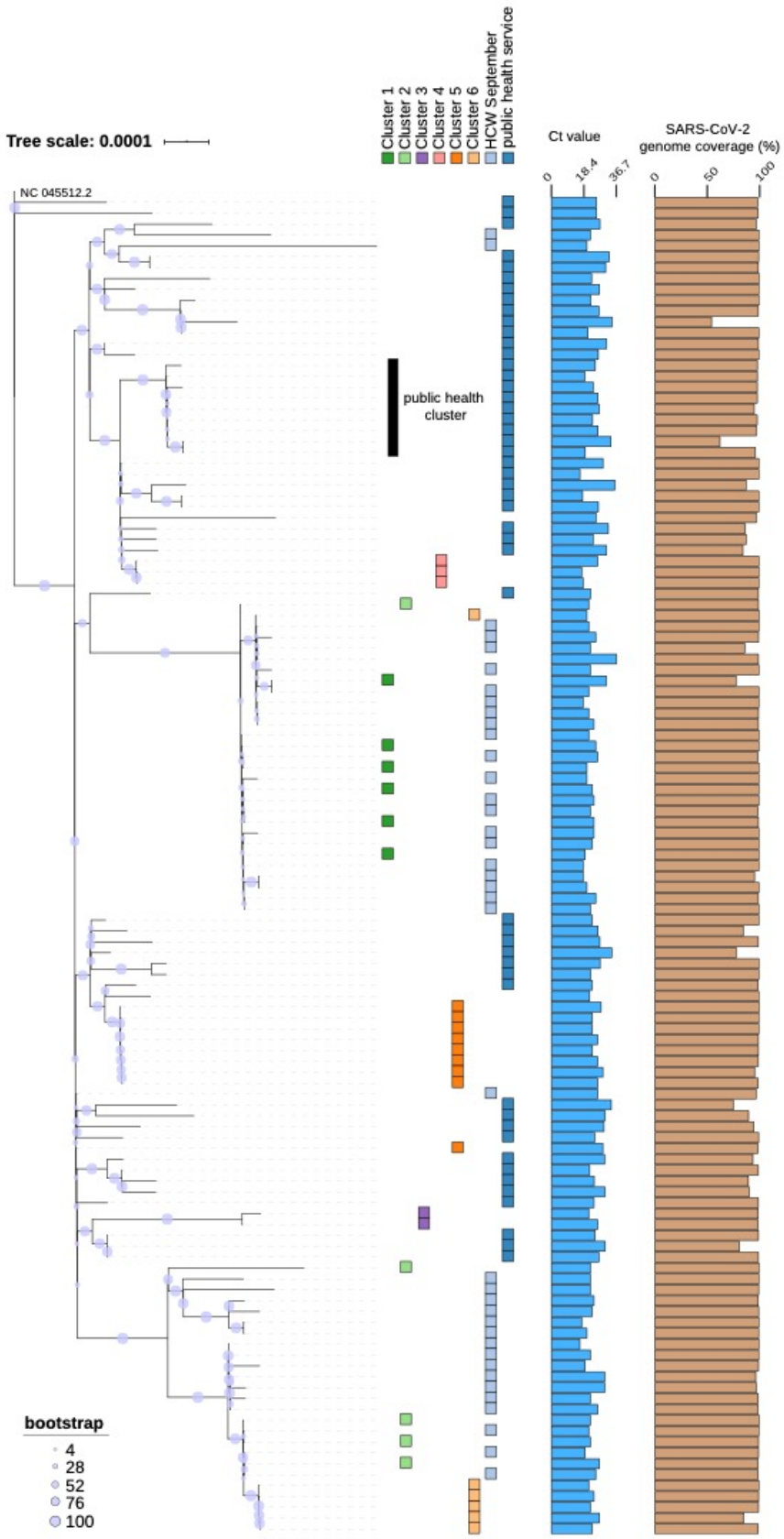

Figure 5. Phylogenetic tree of 123 RC-PCR WGS SARS-CoV-2 genomes rooted at the NC 045512.2 reference genome. These samples are obtained from SARS-CoV-2 positive tested patients, HCW, and samples provided by public health services. Six clusters of samples were identified by the hospital Infection prevention control team. Sample groups are indicated by the colored blocks. Additional Ct values and genome coverage are plotted in barplots to illustrate the diversity in viral load between the samples and the high genome coverage that can be achieved by RC-PCR, respectively.

296

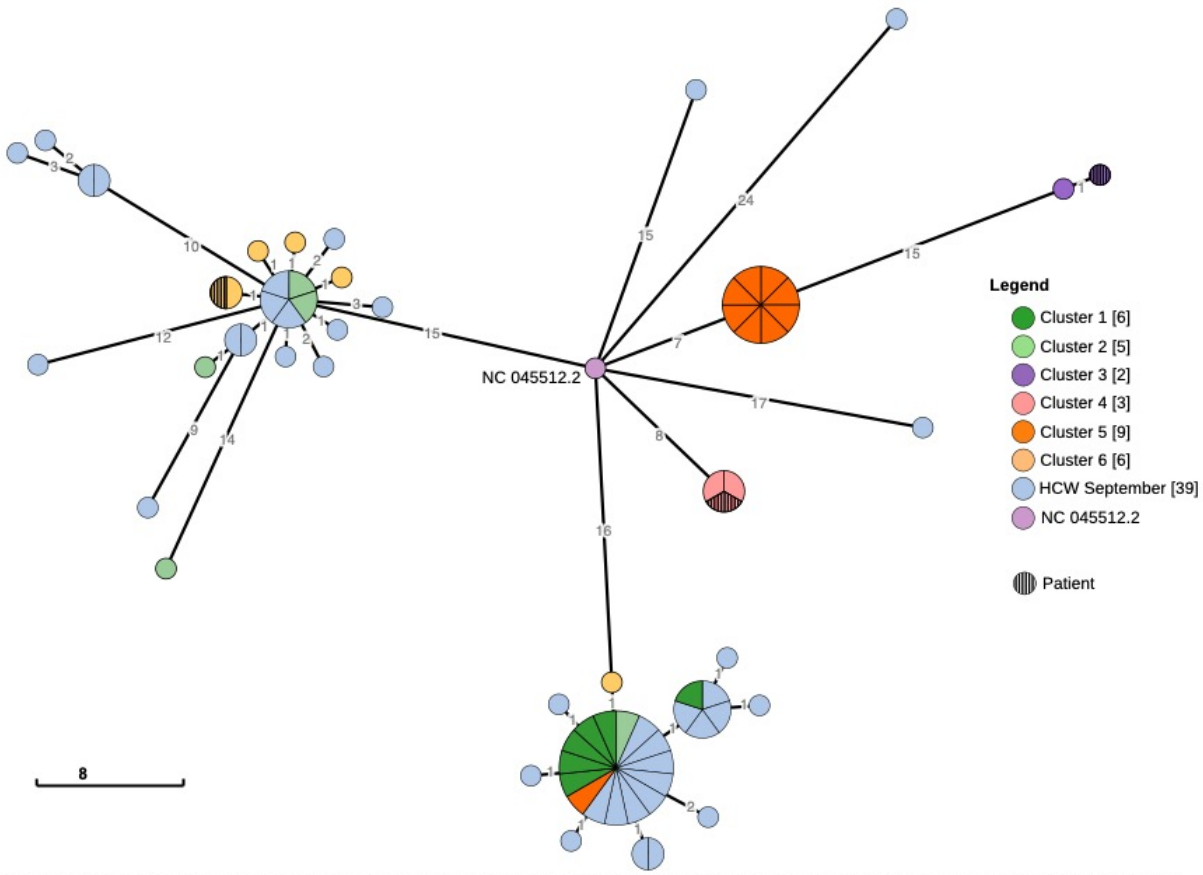

Figure 6. Minimum spanning tree of all samples being part of a cluster as defined by the infection prevention team and all HCW September samples. The tree was calculated using GrapeTree with the MSTreeV2 algorithm. This figure clearly illustrates the relationship between the samples and the clusters.

Discussion

In this study we present the application of a novel method called Reverse Complement-PCR to sequence the SARS-CoV-2 genome which combines target amplification and indexing in a single procedure, directly creating a sequencing ready Illumina library. We applied this method to 173 hospital and community samples that tested positive for SARS-CoV-2 with RT-PCR. Most epidemiological clusters from the hospital and the community were confirmed by phylogenetic clustering. Based on our data, RC-PCR is a reproducible technology, it correlates well with Oxford Nanopore sequencing, is able to sequence samples with Ct values up to 32 determined by RT-PCR and within these samples retrieves a high SARS-CoV-2 genome coverage. Optimization of the protocols is expected to increase coverage in samples with lower viral loads even further.

Previous studies showed the benefit of using WGS of SARS-CoV-2 for outbreak investigation purposes and to study transmission routes.[6,12-16] Several methods have been optimized for this purpose. The ARTIC Illumina method, a tiling multiplex PCR approach, was the first that enabled WGS of

311 SARS-CoV-2 using Illumina sequencers.[17] The technique has subsequently been optimized and
312 analysis, albeit in small sample numbers, concluded that it delivers sufficient quality to perform
313 phylogenetic analysis.[18-20] It had been used as targeted and random RT-PCR screening with
314 subsequent sequencing of the population in order to study the spread through the community.[12] More
315 recently Sikkema *et al.* were the first to describe the use of SARS-CoV-2 sequencing in healthcare
316 associated infections and identify multiple introductions into Dutch hospitals through community-
317 acquired infections.[5]
318
319 SARS-CoV-2 has an estimated mutation rate of $1.12 \times 10{-}3$ substitutions per site per year, which
320 results in 2.8 mutations every month.[21] The minimum spanning tree of Figure 6 shows several samples
321 with a genetic distance of only a single SNP. With the mutation rate in mind it is unclear how to relate
322 these clusters since extensive contact tracing information is lacking and interpretation on SNP
323 regarding outbreak management is unknown. Since community samples of September were
324 unavailable, we are unable to determine whether the genetic diversity in the community was low
325 resulting in genetically similar SARS-CoV-2 strains in a hospital setting. However, since sequencing
326 of samples in March and April 2020 clearly resulted in a larger diversity of SARS-CoV-2, and this
327 was early on in the pandemic, it seems more likely that a common source of infection, in- or outside
328 the hospital is the cause. Further research is needed to determine the accepted SNP distance for the use
329 in outbreak analysis.[22] Although we know minimum spanning trees are often used in outbreak
330 analysis.[5] It is a simplification of the phylogeny which could result in erroneous conclusions in
331 outbreak analysis. Care should be taken in interpreting these results.
332
333 It should be noted that some of the amplicons result in lower coverage than others (See Figure 3).
334 Currently, developments are under way in which a better distribution of the amplicon depth will be
335 achieved resulting in genome coverage that could increase to almost 100%. The difference in genome
336 coverage between Run1 and Run1_new is most likely caused by storage of the library and subsequent
337 pooling on the basis of Ct value of the individual samples, nonetheless, repeated testing at higher Ct
338 values will be needed to confirm this.

With current increase in infections in many countries including the Netherlands and additional measures being put in place to reduce SARS-CoV-2 spreading, real-time sequencing of public health service samples could be used to target infection prevention measures nationwide and locally.[23] Its application can range from incidental cluster analysis in the case of uncertain epidemiological links to real-time surveillance in the community or health care institutes. Additionally, correlation between specific SARS-CoV-2 strains or mutations and clinical outcome could be identified, supporting clinical decision making to improve outcomes for patients.[24,25]

In conclusion, here we implemented for the first time, RC-PCR in the field of medical microbiology and infectious diseases thereby showing it to be a robust method which requires only minimal hands-on time compared to current sequencing methods and can be used for high throughput sequencing of SARS-CoV-2. Moreover, RC-PCR and sequence analysis can support epidemiological data with genomic data to identify, monitor, and screen clusters of samples to help identify chains of transmission of SARS-CoV-2, enabling a rapid, targeted and adaptive response to an ongoing outbreak that has great impact on public health and society.

Author contributions

F.W. and J.P.M.C. conducted the research, performed analysis, wrote manuscript and created the figures. L.F.J.vG., C.P.B-R., E.C.T.H.T., N.vdG-B., J.L.A.H. proofreading and provided clinical information and samples of patients and HCWs. W.A.vdV., S.V.vR., and J.P.G.T. provided experimental design and support. A.T.and J.H. conducted the contact tracing and proofreading of the manuscript. H.F.L.W., J.C.R-L., M.S., and W.J.G.M supervised the study and drafted the manuscript.

Conflict of interest disclosures:

The authors have no conflict of interest to disclose.

Funding/support:

The EasySeq™ RC-PCR SARS-CoV-2 WGS kit was supplied by NimaGen B.V and sequencing of the Illumina libraries was performed by NimaGen B.V.. Validation was performed by the Department of Medical Microbioly at the Radboud university medical center for the purpose of using the technology in routine diagnostics. Therefore no other funding was applied for.

Role of funder/sponsor:

NimaGen B.V. had no role in the design and conduct of the study; collection, management, data analysis; preparation or approval of the manuscript.

References

1. Zhu, N., *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* (2020).
2. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* **20**, 533-534 (2020).
3. Reusken, C.B., *et al.* Rapid assessment of regional SARS-CoV-2 community transmission through a convenience sample of healthcare workers, the Netherlands, March 2020. *Eurosurveillance* **25**, 2000334 (2020).
4. McLachlan, S., *et al.* The fundamental limitations of COVID-19 contact tracing methods and how to resolve them with a Bayesian network approach. (2020).
5. Sikkema, R.S., *et al.* COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *Lancet Infect Dis* **20**, 1273-1280 (2020).
6. Munnink, B.B.O., *et al.* Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nature medicine* **26**, 1405-1410 (2020).
7. Corman, V.M., *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **25**, 2000045 (2020).

389 8. DNA Pipelines R&D, B.F., Diana Rajan, Emma Betteridge, Lesley Shirley, Michael Quail,
390 Naomi Park, Nicholas Redshaw, Iraad F Bronner, Louise Aigrain, Scott Goodwin, Scott
391 Thurston, Stefanie Lensing, Charlotte Beaver, Ian Johnston. COVID-19 ARTIC v3 Illumina
392 library construction and sequencing protocol V.1. (2020).
393 9. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
394 *Nucleic Acids Res* **32**, 1792-1797 (2004).
395 10. Minh, B.Q.*, et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference
396 in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
397 11. Zhou, Z.*, et al.* GrapeTree: visualization of core genomic relationships among 100,000
398 bacterial pathogens. *Genome research* **28**, 1395-1404 (2018).
399 12. Gudbjartsson, D.F.*, et al.* Spread of SARS-CoV-2 in the Icelandic Population. *N Engl J Med*
400 **382**, 2302-2315 (2020).
401 13. Meredith, L.W.*, et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases
402 of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect*
403 *Dis* **20**, 1263-1272 (2020).
404 14. Quick, J.*, et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and
405 other virus genomes directly from clinical samples. *Nat Protoc* **12**, 1261-1276 (2017).
406 15. Richard, M.*, et al.* SARS-CoV-2 is transmitted via contact and via the air between ferrets.
407 *bioRxiv* (2020).
408 16. Stefanelli, P.*, et al.* Whole genome and phylogenetic analysis of two SARS-CoV-2 strains
409 isolated in Italy in January and February 2020: additional clues on multiple introductions and
410 further circulation in Europe. *Euro Surveill* **25**(2020).
411 17. Quick, J. nCoV-2019 sequencing protocol 2020. *Publisher Full Text* (2020).
412 18. Batty, E.M.*, et al.* Comparing library preparation methods for SARS-CoV-2 multiplex
413 amplicon sequencing on the Illumina MiSeq platform. *BioRxiv* (2020).
414 19. Pillay, S. Illumina Nextera DNA Flex library construction and sequencing for SARS-CoV-2:
415 Adapting COVID-19 ARTIC protocol. (2020).

416 20. Pillay, S., *et al.* Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for
417 Quick and Accurate Outbreak Investigation during a Pandemic. *Genes (Basel)* **11**(2020).
418 21. Koyama, T., Platt, D. & Parida, L. Variant analysis of SARS-CoV-2 genomes. *Bull World*
419 *Health Organ* **98**, 495-504 (2020).
420 22. Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* (2020).
421 23. RIVM. Current information about COVID-19 (novel coronavirus). (2020).
422 24. Gong, Y.N., *et al.* SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-
423 deletion mutant and clade possibly associated with infections in Middle East. *Emerg Microbes*
424 *Infect* **9**, 1457-1466 (2020).
425 25. Wang, C., *et al.* The establishment of reference sequence for SARS-CoV-2 and variation
426 analysis. *J Med Virol* **92**, 667-674 (2020).
427