Simplified point-of-care full SARS-CoV-2 genome sequencing using nanopore technology

Anton Pembaur[1,*], Erwan Sallard[2,*], Patrick Philipp Weil[1], Jennifer Ortelt[3], Parviz Ahmad-Nejad[3], Jan Postberg[1,†]

*both authors contributed equally

[†]corresponding author

[1]Clinical Molecular Genetics and Epigenetics, Faculty of Health, Centre for Biomedical Education & Research (ZBAF), Witten/Herdecke University, Alfred-Herrhausen-Str. 50, 58448 Witten, Germany

[2]Institute of Virology and Microbiology, Faculty of Health, Centre for Biomedical Education & Research (ZBAF), Witten/Herdecke University, Stockumer Str. 10, 58453 Witten, Germany

[3]HELIOS University Hospital Wuppertal, Institute of Medical Laboratory Diagnostics, Centre for Clinical & Translational Research (CCTR), Witten/Herdecke University, Heusnerstr. 40, 42283 Wuppertal, Germany

E-Mail addresses:

Anton Pembaur:	anton.pembaur@uni-wh.de
Erwan Sallard:	erwan.sallard@uni-wh.de
Patrick Philipp Weil:	patrick.weil@uni-wh.de
Jennifer Ortelt:	jennifer.ortelt@helios-gesundheit.de
Parviz Ahmad-Nejad:	parviz.ahmad-nejad@helios-gesundheit.de
Jan Postberg:	jan.postberg@uni-wh.de

Abstract

Background: The scale of the ongoing SARS-CoV-2 pandemic warrants the urgent establishment of a global decentralized surveillance system to recognize local outbreaks and the emergence of novel variants-of-concern. Among available deep-sequencing technologies, nanopore-sequencing could be an important cornerstone, since it is mobile, scalable and costs-effective. Therefore, streamlined nanopore-sequencing protocols need to be developed and optimized for SARS-CoV-2 variants identification.

Results: We adapted and simplified existing workflows using the 'midnight' 1,200 bp amplicon split primer sets for PCR, which produce tiled overlapping amplicons covering almost the entire SARS-CoV-2 genome. Subsequently, we applied Oxford Nanopore Rapid Barcoding and the portable MinION Mk1C sequencer combined with the interARTIC bioinformatics pipeline. We tested a simplified and less time-consuming workflow using SARS-CoV-2-positive specimens from clinical routine and identified pre-analytical parameters, which may help to decrease sequencing failures rates. Complete pipeline duration was approx. 7 hrs for one specimen and approx. 11 hrs for 12 multiplexed barcoded specimens.

Conclusions: The adapted protocol contains less processing steps and can be completely conducted within one working-day. Diagnostic CT values are principal criteria for specimen selection.

Keywords

(+)RNA genome sequencing, COVID-19 surveillance, variant-of-concern (VOC)

Background

To face the ongoing SARS-CoV-2 pandemic, a global decentralized warning system is being established to recognize local outbreaks and the emergence of novel variants-of-concern (VOC). A particular focus is given to the identification of VOCs with accelerated transmission rates, increased infectivity or immune escape mutations, since these variants would warrant adaptations in containment and vaccination strategies (Cobey, Larremore et al. 2021, Gupta 2021, Harvey, Carabelli et al. 2021, Winger and Caspari 2021). Moreover, the search for the zoonotic origin of SARS-CoV-2 from comparative analyses of genomic data is an ongoing issue with relevance for the early recognition of future outbreak scenarios (Andersen, Rambaut et al. 2020, V'Kovski, Kzel et al. 2021).

Among the available deep-sequencing technologies, nanopore-sequencing could be an important cornerstone, since it is mobile, scalable and acquisition investments are comparably low. Further, nanopore sequencing devices do not require large-scale IT infrastructure. In the past they were already involved actions of genome surveillance, e.g. the 2015 Ebola outbreak in Liberia, Guinea and sierra Leone (Quick, Loman et al. 2016), and since then there has been a substantial technological progress. However, at least for smaller hospital laboratories with lower throughput, it is desirable to develop protocols as streamlined as possible.

Nanopore sequencing allows the sequencing of either DNA or RNA (Garalde, Snell et al. 2018), and does not require PCR amplification. Furthermore the technique has the potential of producing very long, continuous reads, which theoretically allows to sequence in only one read the 29.903nt long (+)RNA genome of SARS-CoV-2 (Taiaroa, Rawlinson et al. 2020), or its deriving cDNA after reverse transcription (Figure 1).

With the aims of facilitating implementation in routine diagnostics with lower specimens throughput and of simplifying the workflow, we tested modifications of existing ARTIC protocols for SARS-CoV-2 full length (+)RNA genome sequencing (Freed, Vlkova et al. 2020, Li, Wang et al. 2020, Tyson, James et al. 2020). Furthermore, we tested the simplified and less time-consuming workflow on confirmed SARS-CoV-2-positive specimens from clinical routine and identified pre-analytical parameters, which may help to decrease the rate of sequencing failures.

Methods

Nucleic acids isolation

Specimens included nasopharyngeal swabs (Xebios Diagnostics), which underwent routine COVID-19 diagnostic testing. RNA purification was performed via magnetic beads (Seegene NIMBUS/Tanbead). Alternatively, total RNA was purified from 250 µL liquid specimen using 750 µL QIAzol lysis reagent (Qiagen, Cat. No. 158845) upon manufacturer's recommendations, or using silica columns (QiaAmp Viral Mini Kit, Qiagen).

RT-PCR, quality assessments and library preparation

The 'midnight' split primer set from the ARTIC protocol was used for SARS-CoV-2 cDNA amplification in 2 multiplex PCR reactions. To avoid overlaps during multiplex PCR, each single-tube PCR reaction generates consecutively tiled, non-overlapping 1,200 bp amplicons. Mixed together after PCR, both resulting complementary amplicon mixtures cover almost the entire SARS-CoV-2 genome (Freed, Vlkova et al. 2020).

Combined reverse transcription and amplification of multiple 1,200 bp amplicons (RT-PCR) was performed in single tube 20 µL reactions using the Luna One-Step RT-qPCR Kit (NEB; E3005). For RT-PCR, 8 µL of purified template RNA were used for each reaction. From 100 µM primer pools, 1 µL was used in each reaction. Reverse transcription was performed at 55°C for 30 minutes, followed by incubation at 95°C for one minute. Then 34 cycles (pool 1) or 30 cycles (pool 2) of denaturation at 95°C for 20 seconds and annealing and extension in one step at 60°C for 210 seconds were performed. A final extension was performed at 65°C. During the implementation phase amplicon sizes and DNA concentrations were routinely checked by agarose gels or by microvolume electrophoresis (Agilent Bioanalyzer, Agilent DNA 12000 kit). Thereafter, amplicons from primer pools 1 and 2 were quantified by fluorimetry (Promega Quantus) and then mixed at equal concentrations. Library

108 preparation was done using the Rapid Barcoding Sequencing Kit (Oxford nanopore; SQK-RBK004)
109 upon manufacturers recommendations.
110
111 *Nanopore sequencing*
112 Sequencing was performed on a MinION Mk1C with the options 'basecalling' and 'demultiplexing'
113 being enabled. As output format, FAST5 and FASTQ files were chosen. Sequencing time was set for
114 72 hours as default. Sequencing was stopped after reaching at least 10 megabases for each barcode.
115
116 *Bioinformatics*
117 Consensus sequences were built from the barcode-sorted, quality-filtered FAST5 and FASTQ files
118 containing sequencing reads, using the interARTIC pipeline. Except otherwise stated, the
119 'Nanopolish' algorithm was routinely used.
120 Installation and usage of interARTIC pipeline was done following the developers' instructions:
121 https://psy-fer.github.io/interARTIC/installation/. For faster analysis, all available threads were
122 activated in the advanced settings. The interARTIC pipeline performs read filtering, alignments and
123 returns a consensus FASTA file as well as coverage charts for visualization (Figure 2B-H).
124 Consensus FASTA files were uploaded to the Nextstrain webapp (https://clades.nextstrain.org) to
125 perform phylogenetic analyses (Hadfield, Megill et al. 2018).
126
127 *Data availability*
128 FASTQ files and assembled FASTA formatted consensus sequences are available
129 (https://t1p.de/minion-seqdata)
130
131
132 **Results and Discussion**
133 We defined milestones to simplify the protocol and decrease hands-on sample time and working step
134 numbers. 1. We tested whether specimens can be directly taken from residual diagnostic specimens
135 extracted from 96-deepwell-plates using magnetic beads (Seegene NIMBUS/Tanbead). For
136 comparison we applied RNA purification protocols using silica columns (QiaAmp Viral Mini Kit,
137 Qiagen) or guanidinium isothiocyanate (GITC) for RNA extraction (QIAzol lysis reagent, Qiagen). 2.
138 We tested whether reverse transcription can be successfully primed by SARS-CoV-2-specific primers,
139 which are subsequently used for multiplex 1,200 bp amplicon amplification. Purified RNA from the
140 different extraction protocols of milestone 1 was used for cDNA synthesis and successive multiplex
141 PCR with the 2 'midnight' split 1,200 bp amplicon primer pools in single tube reactions 3. We aimed
142 at improving the efficiency of the nanopore sequencing workflow by using the onboard Guppy
143 basecalling capability of the Oxford Nanopore MinION Mk1C device. Moreover, implementation of

144 the interARTIC interface should help to avoid command line-based bioinformatics analyses as much
145 as possible in order to provide a user-friendly and efficient analysis pipeline.
146
147 *Protocol implementation considering differences in viral loads and influences of RNA extraction*
148 *protocols*
149 Sequencing was performed from serially diluted specimens of purified RNA from patients' samples,
150 which exhibited low cycle threshold (CT) value (CT=16-18) after routine diagnostics RT-qPCR (N
151 gene, RdRP gene; Seegene). Using the quantitative reference sample Ch07470 for calibration we
152 determined that a CT=25 (respectively 16 and 18) corresponded to a SARS-CoV-2 copy number of
153 $1.0x\ 10^6$ (respectively $5.12x\ 10^8$ and $1.28x\ 10^8$). Routinely used dilution factors were 2^0-2^{-10} in order to
154 cover several CT value magnitudes and to simulate different amounts of viral loads. In terms of RNA
155 yield and 1,200 bp amplicon PCR performance the magnetic beads-based RNA purification protocol
156 outperformed slightly the GITC method as well as the column-based protocol, but we did not observe
157 differences in read and coverage quality between these different isolation methods (Figure S1). Since
158 direct sampling from 96-deepwell plates allowed us to directly exploit residual specimens, which
159 remained after routine RT-qPCR diagnostics, we decided to focus on magnetic beads-based RNA
160 purification during further protocol development. Moreover, it shortened and simplified the workflow.
161 After semiquantitative or quantitative RT-PCR using multiplex primer pools 1 and 2 in separate single
162 tube reactions for combined reverse transcription of the SARS-CoV-2 (+)RNA and amplification of
163 1,200 bp amplicons, band intensities exhibited strong dependence on viral loads. Moreover, after 32
164 PCR cycles band intensities using primer pool 1 were weaker when compared with primer pool 2
165 (Figure S2). We determined that 34 PCR cycles for primer pool 1 and 30 cycles for primer pool 2 were
166 a good compromise. Therefore, reverse transcription can be successfully primed by the 'midnight'
167 primers.
168
169 *Surveillance of multiplex nanopore sequencing and multiple reuses of flow cells*
170 Using the Rapid Barcoding Kit (SQK-RBK004, Oxford Nanopore) enabled us to barcode the PCR
171 products without purification steps and sequence them immediatley. As a standard, we used 12
172 barcoded libraries for multiplex sequencing on R9 flow cells. In contrast to the Oxford Nanopore
173 MinION Mk1B, the MinION Mk1C device features onboard guppy basecalling. In combination with
174 the Rapid Barcoding Kit, we exploited this opportunity for real-time surveillance of basecalling and
175 demultiplexing for each of the 12 multiplexed samples per run. This enabled us to recognize the exact
176 time point at which a reading depth of approx. 10 Mbp per barcode was achieved. The time to reach
177 this threshold depended heavily on the viral load (simulated by serially diluted samples). Decreased
178 viral loads let to a considerable decrease of passed reads (Figure 2A). A manual stop of sequencing
179 followed by flow cell washing (Flow Cell Wash Kit, EXP-WSH004, Oxford Nanopore) allowed us to
180 reuse a single flow cell for a series of 3 sequencing runs, each using 12 multiplexed barcoded libraries.

181 This specific scenario resulted in costs of approx. 40 USD per sample. Theoretically, the possible
182 number of reuses depends largely on the duration of the sequencing, which itself depends mainly on
183 the number of used barcodes. Further, if a MinION Mk1B is considered for this purpose, a similar
184 surveillance functionality could be achieved using RAMPART (https://github.com/artic-
185 network/rampart) on a dedicated LINUX environment.
186
187 *Assembly of full SARS-CoV-2 genomes and pathogen genome data analyses*
188 For mapping and full length SARS-CoV-2 genome assembly, we used the FASTQ files resulting from
189 Guppy basecalling in order to generate FASTA formatted consensus sequence files. We used the
190 ARTIC pipeline through a graphical user interface (interARTIC; https://github.com/Psy-
191 Fer/interARTIC). Once installed, this is an easy to use and relatively fast pipeline with only five
192 minutes hands-on time, which enables the use of the 'Nanopolish' or 'Medaka' algorithms for
193 simultaneous analyses of multiplexed barcoded samples. As an example, Figure 2B shows the
194 complete and deep coverage of the complete SARS-CoV-2 genome after the combination of pools 1
195 (light blue) and 2 (pink). This demonstrates that all contained 1,200 bp amplicon were specifically and
196 efficiently amplified during the combined RT-PCR reaction (Figure 2B).
197 To compare the interARTIC and Geneious Prime pipelines, we used exactly the same FASTQ file
198 from the same sample shown in Figure 2B (https://go.geneious.com/video/how-to-assemble-
199 coronavirus-genomes). Phylogenetic analyses with Nextstrain using the FASTA consensus files
200 obtained from the interARTIC or the Geneious Prime pipelines resulted in considerably different
201 phylogenetic distances in clade 20I, showing that the bioinformatics pipeline influences the result
202 (Figure S3). The interARTIC pipeline proved superior in terms of coverage and sequencing depth.
203 Notably, within the interARTIC pipeline both options, the 'Nanopolish' and 'Medaka' algorithms
204 performed equally well with respect to consensus sequence quality, but 'Medaka' was considerably
205 faster.
206 We observed that the consensus sequences returned by the 'Nanopolish' and 'Medaka' algorithms
207 contain numerous unsolved regions (for which only 'N's are indicated). Interestingly, the regions
208 unsolved by one algorithm were generally solved by the other, so we assumed that the two consensus
209 sequences could be combined to produce a 'super-consensus' with improved variant prediction value.
210 We developed a Python code that merges the 'Nanopolish' and 'Medaka' consensus sequences and
211 generates the corresponding variant calling file. As expected, the 'super-consensus' contained less
212 unsolved regions when compared with the 'Nanopolish' and 'Medaka' consensus sequences alone
213 (Figure 4B). In addition, it retained the high quality mutations, which were identified by both
214 algorithms, while removing most false-positives probably caused by sequencing and alignment errors
215 (Figure 4A). Consequently, our consensus-merging code improves the quality of variant calling and
216 highlights the complementarity of 'Nanopolish' and 'Medaka' for nanopore-sequencing of SARS-
217 CoV-2 and others. Nevertheless, this quality increase comes at the cost of information loss such as the

218 number of reads per variant or other metadata which were initially generated by 'Nanopolish' and
219 'Medaka' and are not transferred to the 'super-consensus'.
220 Serial input RNA dilutions or, respectively, viral load influenced the depth of sequencing (Figure 2C-
221 H). For the output of high-quality consensus sequences in the FASTA file format a coverage threshold
222 of 20 was used as default. We generally observed that this could be reached when a SARS-CoV-2 titer
223 of 4x 10^6 was given. Viral copy numbers lower than 4x 10^6 were associated with incompletely
224 assembled SARS-CoV-2 genomes. Thus, we provide here a convincing line of evidence that the copy
225 number-normalized CT values of diagnostic RT-qPCR can be used as the criterion of sequencing
226 success.
227 Single or batch high-quality consensus FASTA formatted sequences were used for phylogenetic tree
228 visualization and variant calling using the Nextstrain webapp (https://clades.nextstrain.org) (Hadfield,
229 Megill et al. 2018). In our hands, the obtained sequences could faithfully be assigned to specific clades
230 in the reference tree (Figure 3). Again, an influence of viral load was observed. However, despite
231 incomplete coverage in those cases enough informative sequence data could be obtained for
232 phylogenetic analyses from several low copy number samples. As a result of serial sample dilutions,
233 we observed deviating phylogenetic distances within the clade, wherein specimens classification
234 occurred (Figure 3A), which eventually could lead to incorrect clade association. The use of
235 specimens from diagnostic routine with viral copy numbers higher than approx. 4x 10^6 apparently led
236 to their faithful association with different clades, which were clades 20I and 19A in the shown
237 example (Figure 3B). The 'Nanopolish' and 'Medaka' consensus as well as the merged super-
238 consensus could be reliably associated to the corresponding clade (20I in the shown example). The
239 'Medaka' consensus mapped at a greater distance than the 'Nanopolish' consensus, probably because
240 'Medaka' algorithm does not correct frameshifts, while the super-consensus had an intermediary
241 distance between the two other consensus (Figure 4C).
242
243 Taken together, the main achievements of an optimized workflow are: 1. Purified RNA from SARS-
244 CoV-2-positive patients can be directly taken from residual diagnostic specimens in 96-deepwell-
245 plates; 2. cDNA synthesis and successive multiplex PCR with 2 split primer pools can be performed in
246 single tube reactions. Since cDNA synthesis is primed by SARS-CoV-2-specific primers for 1,200 bp
247 amplicon amplification, there is no need for use of unspecific hexanucleotide priming. 3. Onboard
248 Guppy basecalling with the Oxford Nanopore MinION Mk1C device and implementation of the
249 interARTIC led to a further reduction of working steps and hands-on time (Figure 5). Implementation
250 in smaller hospital laboratories with lower specimens' throughput can be easily done at moderate
251 costs.
252
253 We provide a detailed protocol for all steps here, which includes the Python code (Supplementary
254 Information 1) and corresponding command line: dx.doi.org/10.17504/protocols.io.bwhppb5n

Conclusions

The adapted protocol contains less processing steps. Diagnostic CT values are the principal criteria for specimen selection. After diagnostic qRT-PCR, multiplex library preparation, quality controls, nanopore sequencing and the bioinformatic pipeline can be completely conducted within one working-day.

List of abbreviations

(+)RNA (sense polarity [single-stranded] ribonucleic acid), RdRP (RNA-dependent RNA polymerase), RT-qPCR (reverse transcription-quantitative polymerase chain reaction, SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2)

Declarations

Ethics approval and consent to participate

For the collection and use of specimens at Helios University Hospital Wuppertal (North Rhine-Westphalia, Western Germany) we obtained approval of the Witten/Herdecke University Ethics board (No. 160/2020). All work has been conducted according to the principles expressed in the Declaration of Helsinki.

Competing interests

There are no conflicting interests for none of the authors, which need declaration.

Funding

The study was financed by own institutional means.

References

Andersen, K. G., A. Rambaut, W. I. Lipkin, E. C. Holmes and R. F. Garry (2020). "The proximal origin of SARS-CoV-2." Nat Med **26**(4): 450-452.

Cobey, S., D. B. Larremore, Y. H. Grad and M. Lipsitch (2021). "Concerns about SARS-CoV-2 evolution should not hold back efforts to expand vaccination." Nat Rev Immunol **21**(5): 330-335.

Freed, N. E., M. Vlkova, M. B. Faisal and O. K. Silander (2020). "Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding." Biol Methods Protoc **5**(1): bpaa014.

Garalde, D. R., E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, M. Jordan, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, E. J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, S. Young, D. Brocklebank, S. Juul, J. Clarke, A. J. Heron and D. J. Turner (2018). "Highly parallel direct RNA sequencing on an array of nanopores." Nat Methods **15**(3): 201-206.

Gupta, R. K. (2021). "Will SARS-CoV-2 variants of concern affect the promise of vaccines?" Nat Rev Immunol **21**(6): 340-341.

Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford and R. A. Neher (2018). "Nextstrain: real-time tracking of pathogen evolution." Bioinformatics **34**(23): 4121-4123.

Harvey, W. T., A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden, R. Reeve, A. Rambaut, C.-G. U. Consortium, S. J. Peacock and D. L. Robertson (2021). "SARS-CoV-2 variants, spike mutations and immune escape." Nat Rev Microbiol **19**(7): 409-424.

Li, J., H. Wang, L. Mao, H. Yu, X. Yu, Z. Sun, X. Qian, S. Cheng, S. Chen, J. Chen, J. Pan, J. Shi and X. Wang (2020). "Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing." Sci Rep **10**(1): 17492.

Quick, J., N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouedraogo, B. Afrough, A. Bah, J. H. Baum, B. Becker-Ziaja, J. P. Boettcher, M. Cabeza-Cabrerizo, A. Camino-Sanchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. G. G. Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui, A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzarelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallash, L. V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, R. Trina, A. Bello, A. A. Sall, O. Faye, O. Faye, N. Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune, K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O'Shea, A. M. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Woelfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keita, A. Rambaut, P. Formenty, S. Gunther and M. W. Carroll (2016). "Real-time, portable genome sequencing for Ebola surveillance." Nature **530**(7589): 228-232.

Taiaroa, G., D. Rawlinson, L. Featherstone, M. Pitt, L. Caly, J. Druce, D. Purcell, L. Harty, T. Tran, J. Roberts, M. Catton, D. Williamson, L. Coin and S. Duchene (2020). "Direct RNA sequencing and early evolution of SARS-CoV-2." bioRxiv: 2020.2003.2005.976167.

Tyson, J. R., P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, N. Prystajecky, I. Goodfellow, S. J. Wilson, R. Harrigan, T. P. Snutch, N. J. Loman and J. Quick (2020). "Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore." bioRxiv: 2020.2009.2004.283077.

V'Kovski, P., A. Kratzel, S. Steiner, H. Stalder and V. Thiel (2021). "Coronavirus biology and replication: implications for SARS-CoV-2." Nat Rev Microbiol **19**(3): 155-170.

Weil, P. P., J. Hentschel, F. Schult, A. Pembaur, B. Ghebremedhin, O. Mboma, A. Heusch, A. C. Reuter, D. Muller, S. Wirth, M. Aydin, A. C. W. Jenke and J. Postberg (2021). "Combined RT-qPCR and pyrosequencing of a Spike glycoprotein polybasic cleavage motif can uncover pediatric SARS-CoV-2 infections associated with heterogeneous presentation." Mol Cell Pediatr **8**(1): 4.

Winger, A. and T. Caspari (2021). "The Spike of Concern-The Novel Variants of SARS-CoV-2." Viruses **13**(6).

Figures and Figure legends

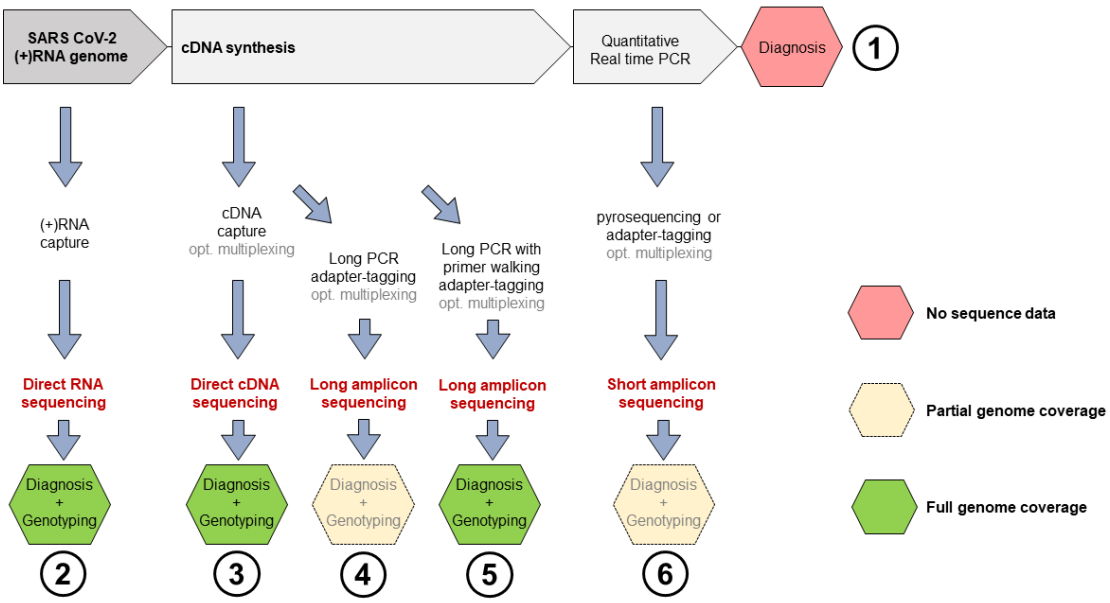


Figure 1. Nanopore sequencing options (2-6) compatible with diagnostic qPCR pipelines (1). The pyrosequencing option in path (6) requires biotin-tagged primers for qPCR (Weil, Hentschel et al. 2021).

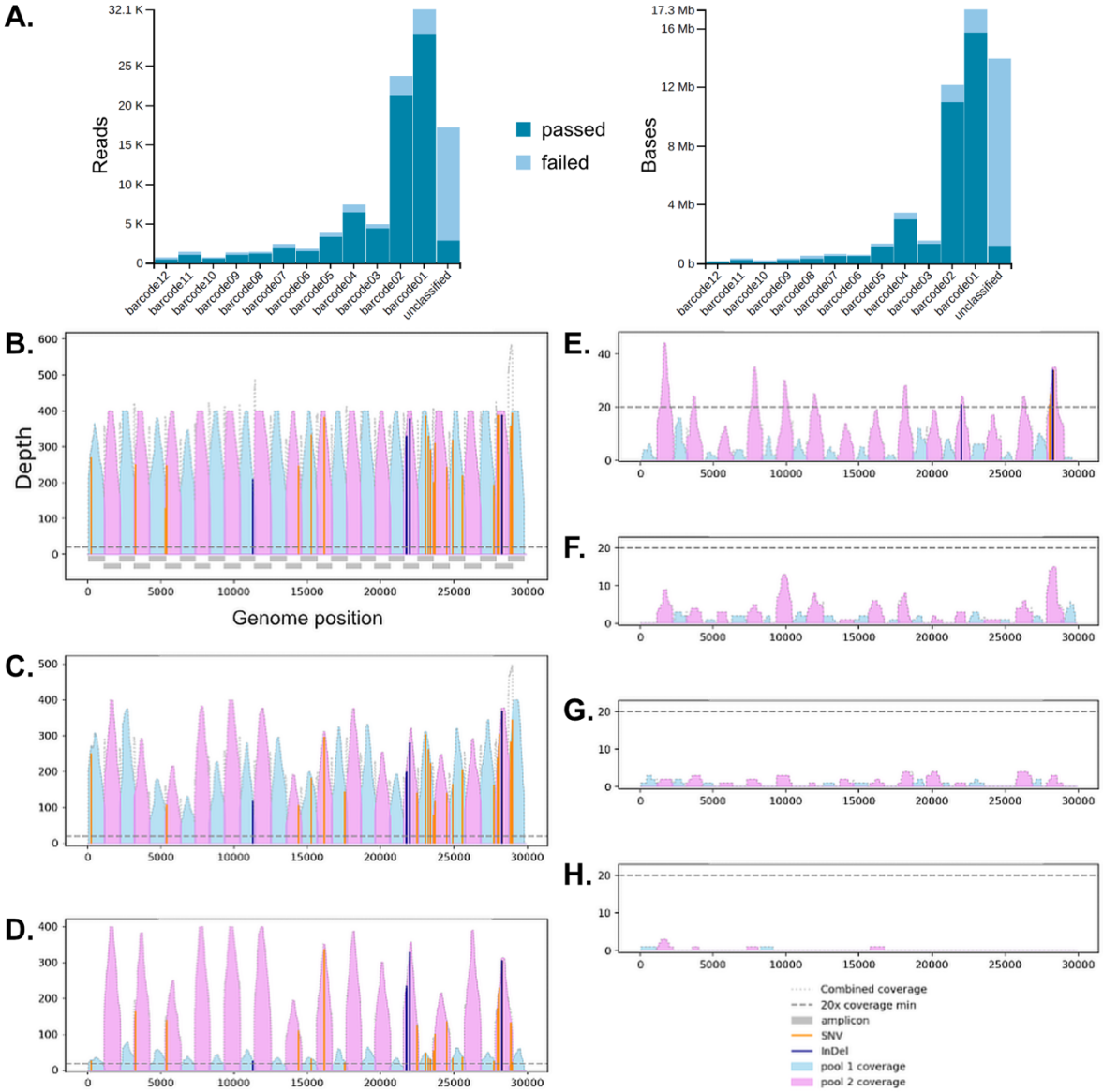


Figure 2. Surveillance using nanopore sequencing (A.), and effects of specimens dilution on the SARS-CoV-2 genome coverage (B.), whereby viral copy numbers were 1.28x 10^6 (B.); 5.12x 10^8 (C.); 2.56x 10^8 (D.); 4x 10^6 (E.); 2x 10^6 (F.); 1x 10^6 (G.); 5x 10^5 (H.). In (A.) increasing barcode numbers (X-axis) correspond to decreasing viral titers.

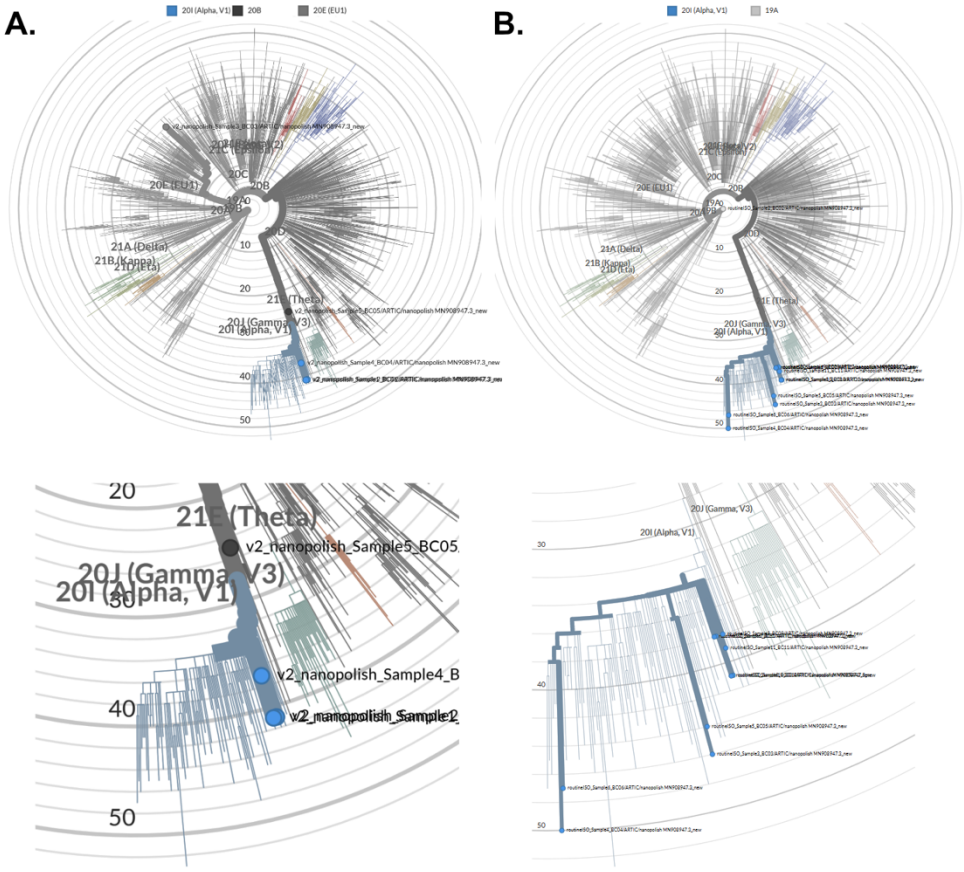
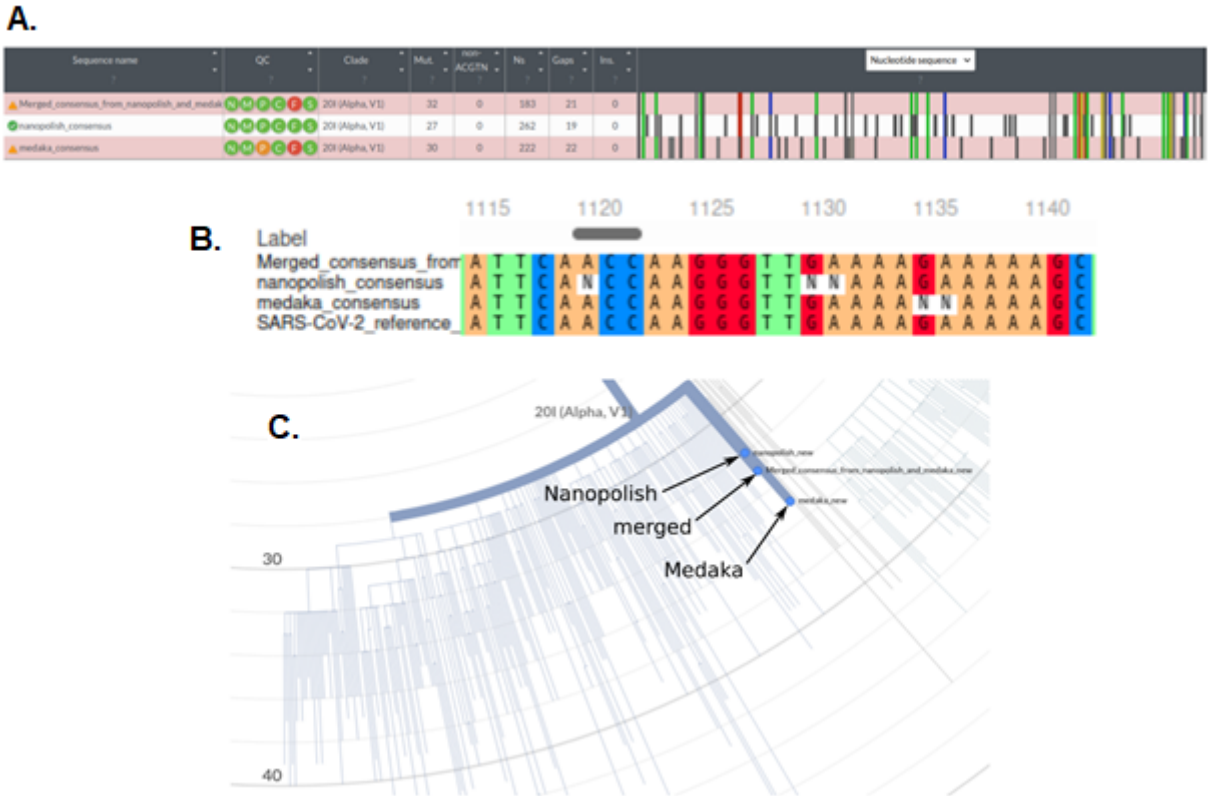


Figure 3. Comparison on the evolutionary distance calculation, when data from the same specimen was processed by 2 different bioinformatics pipelines. Phylogenetic tree visualization was done using the Nextstrain open-source platform for pathogen genome data analyses (Hadfield, Megill et al. 2018). We used serial dilutions of identical specimens (A.), or a selection of samples from different individuals from clinical routine (B.) for phylogenetic analyses. Below the trees a magnification of clade 20I is shown, wherein most specimens grouped (A.,B.).

Figure 4. Comparison of the consensus sequences returned by 'Nanopolish' and 'Medaka' algorithm and the merged consensus for a nanopore sequencing run corresponding to a viral load of 1.28x10**6.



(A.) Alignment of the 'Nanopolish' and 'Medaka' consensus as well as the merged consensus on the SARS-CoV-2 reference genome by the Nextstrain program. The merged consensus conserved all of the high quality mutations that mapped to the known variant 20I (shown in colour or light grey), while most non-matching mutations (in dark grey, likely sequencing errors) of the 'Nanopolish' or 'Medaka' consensus were lost. **(B.)** Comparison of variant solving by 'Nanopolish', 'Medaka' and our code. The multiple alignment was performed by MAFFT online tool with the three consensus sequences and the SARS-CoV-2 reference genome (MN908947.3). Contrary to the 'Nanopolish' and 'Medaka' consensus, the merged consensus solved the entire region and led to accurate variant calling. **(C.)** Clade mapping and phylogenetic distances calculated by Nextstrain for the three consensus sequences.

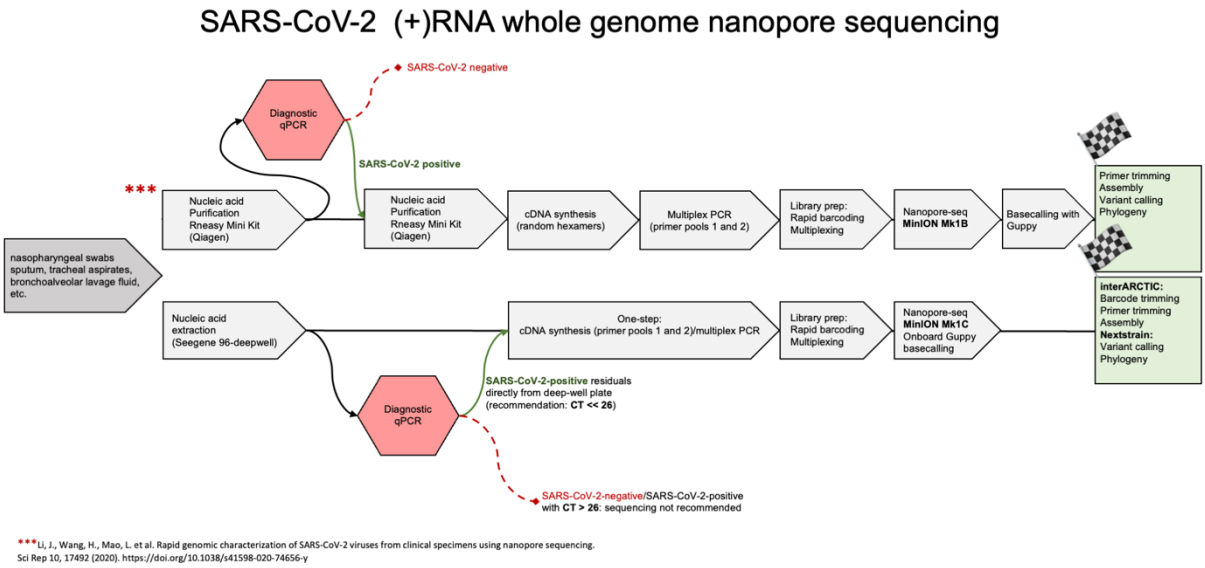


Figure 5. Comparison of a rapid SARS-CoV-2 whole (+)RNA genome nanopore sequencing pipeline (Freed, Vlkova et al. 2020) with the novel simplified workflow, whose main achievements are: 1. Purified RNA from SARS-CoV-2-positive patients can be directly taken from residual diagnostic specimens in 96-deepwell-plates; 2. cDNA synthesis and successive multiplex PCR with 2 primer pools can be performed in single tube reactions. Since cDNA synthesis is primed by SARS-CoV-2-specific primers for 1,200 bp amplicon amplification, there is no need for use of unspecific hexanucleotide priming. 3. Onboard Guppy basecalling with the Oxford Nanopore MinION Mk1C device and implementation of the interARTIC led to a further reduction of working steps and hands-on time.

Supplementary Information

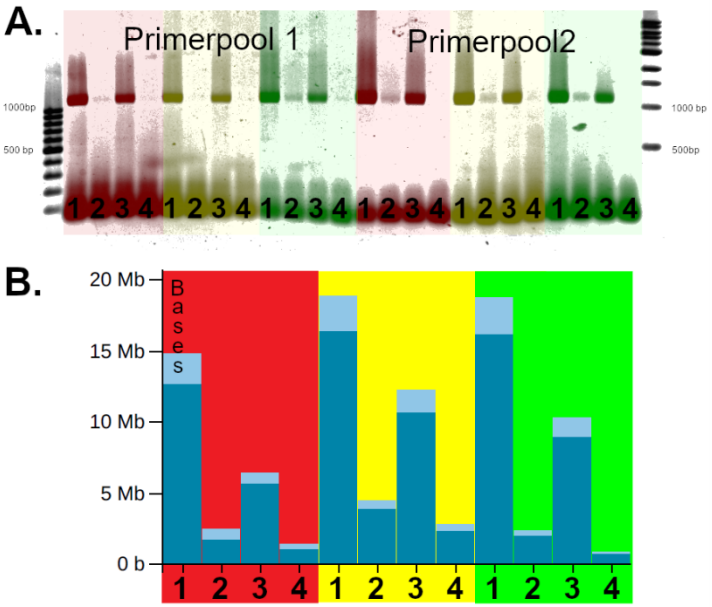


Figure S1. Comparison of different isolation methods: Silica columns (Red), magnetic beads (Yellow) and GITC for RNA extraction (Green). Four samples with different Ct-Values were used (1: Ct 19, 2: Ct 25, 3: Ct 19, 4: Ct 24).

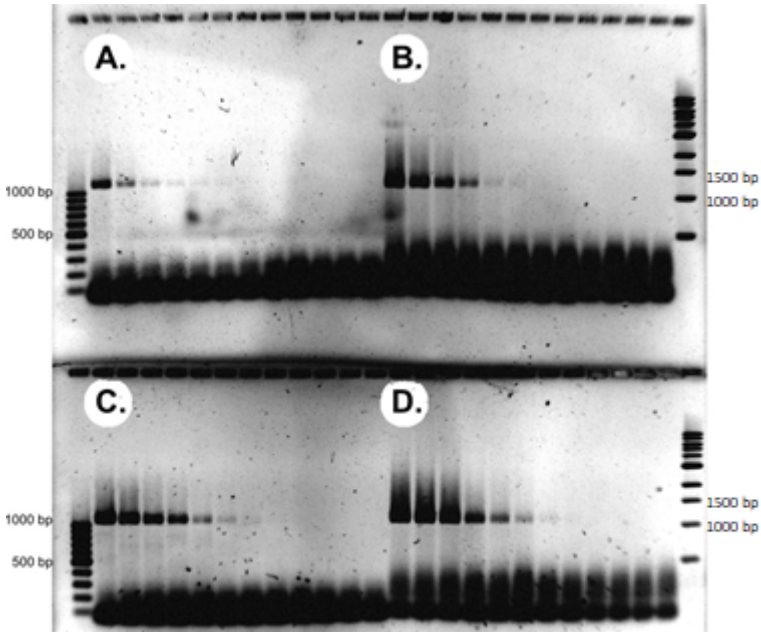


Figure S2. The influence of viral load on the amplification of 1,200 bp amplicon was analysed by agarose gel electrophoresis using samples, where RNA was serially diluted prior to cDNA synthesis and semiquantitative multiplex PCR (A, B) or quantitative multiplex PCR (C, D). Multiplex primer pool 1 (A, C) and pool 2 (B, D) were used in separate single tube reactions. (A-D.) From left to right: Decreasing viral loads (dilution factors 2^{0}-2^{-10} plus no template control).

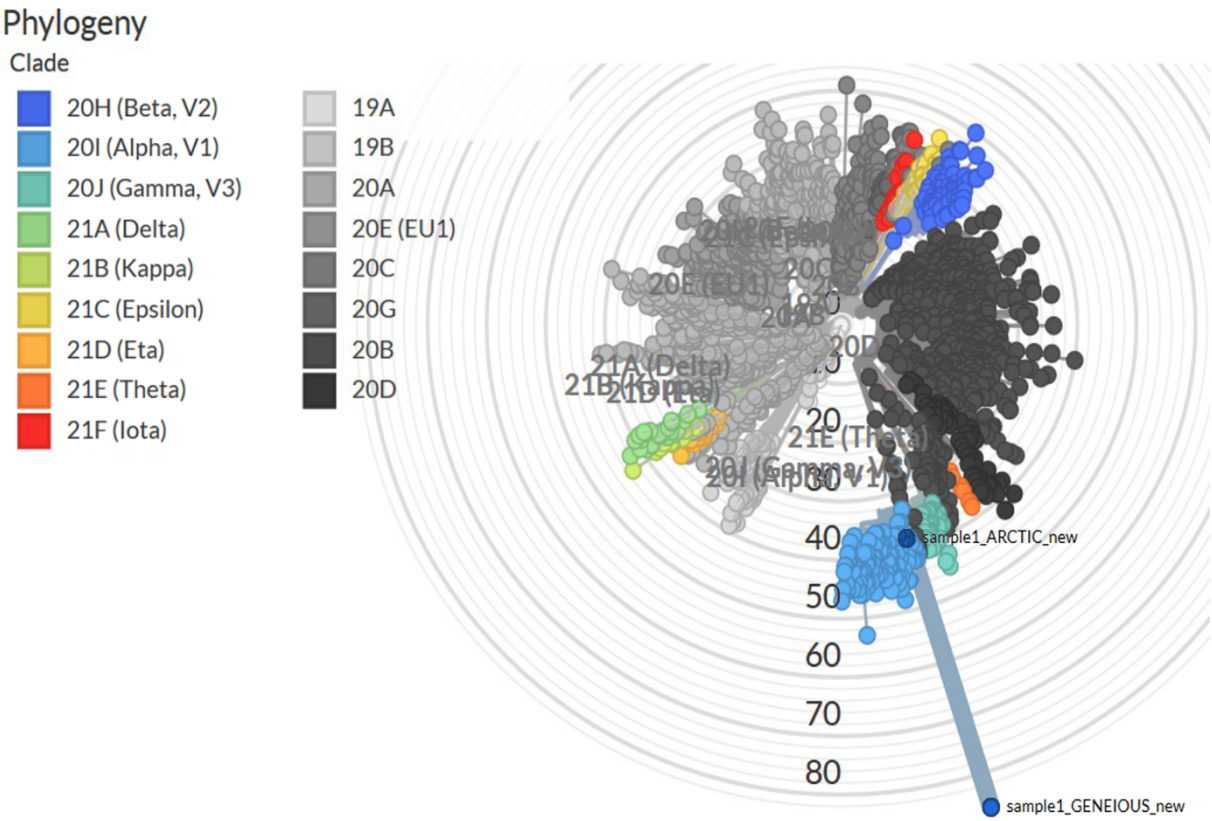


Figure S3. Comparison on the phylogenetic distance, when data from the same specimen was processed by 2 different bioinformatics pipelines Phylogenetic tree visualization was done using the Nextstrain web app (https://clades.nextstrain.org) for pathogen genome data analyses (Hadfield, Megill et al. 2018).

Supplementary Information 1

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue Jun 29 21:10:17 2021

@author: Erwan Sallard
"""
import sys
from Bio import Align
from Bio import SeqIO

medaka_consensus = sys.argv[1]
nanopolish_consensus = sys.argv[2]
output_folder = sys.argv[4]
coronavirus_reference_sequence = sys.argv[3]

output_suffix=medaka_consensus.split('/')[-1].split('.')[-3].split('_')[-1]

### Merging the consensus sequences

# when a sequence contains 'N's and the other solved the variant, we will take the solved
sequence.
# In the very rare cases where both algorithms disagree and none of them give 'N's, we give
priority to nanopolish

## perform the alignment
n_consensus = SeqIO.read(nanopolish_consensus,'fasta').seq
m_consensus = SeqIO.read(medaka_consensus,'fasta').seq
aligner = Align.PairwiseAligner()
aligner.mismatch_score = -1
```

```
aligner.target_open_gap_score = -1.53
aligner.target_extend_gap_score = -1
aligner.query_extend_gap_score = -1
aligner.query_open_gap_score = -1.53
alignment = str(next(aligner.align(n_consensus, m_consensus))).split('\n')

## replace 'N's in the nanopolish consensus by the medaka sequence if it is solved
'''
gérer blocs modifiés, ex 'NN-' vs '--T'
'''
aligned_nanopolish=alignment[0]
aligned_medaka=alignment[2]
merged_consensus=''
position=0
alignment_length=len(aligned_nanopolish)
while position<alignment_length:
    if aligned_nanopolish[position]=='N':
        # there is an unsolved sequence in the nanopolish alignments
        # first step: identify the extent of the unsolved region ('N's and '-'s)
        unsolved_begin = position
        unsolved_end = position
        while (unsolved_begin>1) and (aligned_nanopolish[unsolved_begin-1] in ['N','-']):
            unsolved_begin-=1
        while (unsolved_end<alignment_length-1) and (aligned_nanopolish[unsolved_end+1] in
['N','-']):
            unsolved_end+=1
        # second step: see if medaka solved the region
        medaka_segment=aligned_medaka[unsolved_begin:unsolved_end+1]
        medaka_solved=''
        for base in medaka_segment:
            if base in ['A','G','C','T']:
                medaka_solved+=base
        # third step: if it exists, append the solved sequence to the merged consensus
        if medaka_solved!='':
            merged_consensus+=medaka_solved
        else:
            # medaka too didn't solve the sequence, so we keep nanopolish's sequence
            merged_consensus+=aligned_nanopolish[unsolved_begin:unsolved_end+1]
        position=unsolved_end+1
    else:
        merged_consensus+=aligned_nanopolish[position]
        position+=1

## remove '-'
contiguous_sequences=merged_consensus.split('-')
merged_consensus=''
for segment in contiguous_sequences:
    merged_consensus+=segment

## save the consensus sequence in fasta format
fasta = open(output_folder+'/merged_consensus_'+output_suffix+'.fa','w')
fasta.writelines('>Merged_consensus_from_nanopolish_and_medaka_'+output_suffix+'\n')
consensus_length = len(merged_consensus)
line_number = int(consensus_length/60)+1
for line in range(line_number-1):
    fasta.writelines(merged_consensus[60*line:60*(line+1)]+'\n')
fasta.writelines(merged_consensus[60*(line_number)-1:])
fasta.close()


### Identifying all solved variants in the combined algorithms output

## align the merged consensus with the reference SARS-CoV-2
reference = SeqIO.read(coronavirus_reference_sequence,'fasta').seq
consensus = SeqIO.read(output_folder+'/merged_consensus_'+output_suffix+'.fa','fasta').seq
alignment = str(next(aligner.align(consensus, reference))).split('\n')
aligned_consensus=alignment[0]
aligned_reference=alignment[2]

## identify the solved variants in the consensus
variants_summary = []
# this list will record for all variants their position in the reference genome,
# and the reference and variant alleles
index, position = 0, 1
reference_length = len(reference)
alignment_length = len(aligned_reference)
while index < alignment_length:
    if aligned_reference[index]=='-' and aligned_consensus[index-1]!='N':
```

```
        # the consensus contains an insertion (and not an unsolved region)
        mismatch_end = index+1
        while mismatch_end<alignment_length and aligned_reference[mismatch_end]=='-':
            mismatch_end+=1
        variants_summary.append([position-1,aligned_reference[index-
1],aligned_consensus[index-1:mismatch_end]])
        index=mismatch_end
    elif aligned_consensus[index]=='-' and aligned_consensus[index-1]!='N':
        # the consensus contains an deletion (and not an unsolved region)
        mismatch_end = index+1
        while mismatch_end<alignment_length and aligned_consensus[mismatch_end]=='-':
            mismatch_end+=1
        variants_summary.append([position-1,aligned_reference[index-
1:mismatch_end],aligned_consensus[index-1]])
        position+=mismatch_end-index
        index=mismatch_end
    elif aligned_consensus[index]!=aligned_reference[index] and aligned_consensus[index]!='N'
and aligned_consensus[index-1]!='N':
        # the consensus contains a substitution (and not an unsolved region)
        mismatch_end = index+1
        while mismatch_end<alignment_length and
aligned_consensus[mismatch_end]!=aligned_reference[mismatch_end]:
            mismatch_end+=1
        reference_allele = aligned_reference[index:mismatch_end]
        while (reference_allele[-1]=='-'):
            # in case the substitution is followed by an insertion, there is no need to record
the '-' sign
            reference_allele = reference_allele[:-1]
            position-=1
        consensus_allele = aligned_consensus[index:mismatch_end]
        while (consensus_allele[-1]=='-'):
            consensus_allele = consensus_allele[:-1]
        variants_summary.append([position,reference_allele,consensus_allele])
        position+=mismatch_end-index
        index=mismatch_end
    else:
        index+=1
        position+=1

## create a .vcf file to save the variants and write the header
output_vcf = output_folder+'/merged_variants_'+output_suffix+'.vcf'
variants=open(output_vcf,'w')
variants.writelines('##fileformat=VCFv4.2\n')
variants.writelines('#CHROM\tPOS\tID\tREF\tALT\tQUAL\tFILTER\tINFO\n')

## fill the .vcf file
for variant in variants_summary:
    variants.writelines('SARS-CoV-
2_genome\t'+str(variant[0])+'\t.\t'+variant[1]+'\t'+variant[2]+'\t.\t.\t.\n')

variants.close()
```