



Real-time SARS-CoV-2 diagnostic and variants tracking over multiple candidates using nanopore DNA sequencing

François Stüder, Jean-Louis Petit, Stefan Engelen, Marco Antonio Mendoza-Parra

► To cite this version:

François Stüder, Jean-Louis Petit, Stefan Engelen, Marco Antonio Mendoza-Parra. Real-time SARS-CoV-2 diagnostic and variants tracking over multiple candidates using nanopore DNA sequencing. Scientific Reports, 2021, 11 (1), 10.1038/s41598-021-95563-w . hal-03321396

HAL Id: hal-03321396

<https://hal.science/hal-03321396>

Submitted on 5 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



OPEN

Real-time SARS-CoV-2 diagnostic and variants tracking over multiple candidates using nanopore DNA sequencing

François Stüder^{1,3}, Jean-Louis Petit^{1,3}, Stefan Engelen² & Marco Antonio Mendoza-Parra¹✉

Since December 2019, a novel coronavirus responsible for a severe acute respiratory syndrome (SARS-CoV-2) is accountable for a major pandemic situation. The emergence of the B.1.1.7 strain, as a highly transmissible variant has accelerated the world-wide interest in tracking SARS-CoV-2 variants' occurrence. Similarly, other extremely infectious variants, were described and further others are expected to be discovered due to the long period of time on which the pandemic situation is lasting. All described SARS-CoV-2 variants present several mutations within the gene encoding the Spike protein, involved in host receptor recognition and entry into the cell. Hence, instead of sequencing the whole viral genome for variants' tracking, herein we propose to focus on the SPIKE region to increase the number of candidate samples to screen at once; an essential aspect to accelerate diagnostics, but also variants' emergence/progression surveillance. This proof of concept study accomplishes both at once, population-scale diagnostics and variants' tracking. This strategy relies on (1) the use of the portable MinION DNA sequencer; (2) a DNA barcoding and a SPIKE gene-centered variant's tracking, increasing the number of candidates per assay; and (3) a real-time diagnostics and variant's tracking monitoring thanks to our software RETIVAD. This strategy represents an optimal solution for addressing the current needs on SARS-CoV-2 progression surveillance, notably due to its affordable implementation, allowing its implantation even in remote places over the world.

Since December 2019, the emergence of a novel coronavirus responsible for a severe acute respiratory syndrome (SARS-CoV-2) is accountable for a major pandemic situation, leading to date to nearly 3 million deaths world-wide. Early on 2020 major viral diagnostic efforts were deployed, including the development of new strategies aiming at (1) reducing the diagnostics time, (2) decrease the costs per assay, (3) providing population-scale solutions, but also (4) presenting a high sensitivity and specificity, notably to discriminate asymptomatic cases. While costs and required time per assay were early on addressed by the development of immune-based tests¹, their sensitivity relative to viral nucleic acid-targeting diagnostics were initially questioned, but finally accepted in several countries notably due to the low throughput of diagnostics based on RT-qPCR assays².

With the aim of taking advantage of the viral nucleic acid targeting diagnostics but at the same time increasing the number of candidates' through-put, strategies based on the use of massive parallel DNA sequencing for population-scale diagnostics were proposed^{3–5}. These efforts rely on (1) the incorporation of DNA molecular barcodes during sample preparation; (2) pooling barcoded samples for an “in-one-run” diagnostics assay by the use of massive parallel DNA sequencing; (3) stratification of diagnostics' outcome with the help of bioinformatics processing. While such strategies demonstrated to up-scale the diagnostics power beyond several thousand candidates per assay, they require to count with a DNA sequencing platform where large instruments (Illumina Sequencers) are driven by specialized technical engineers. As consequence, this strategy remains applicable to highly developed countries (due to the prohibitive costs of the required DNA sequencers), which in addition requires a dedicated pipeline strategy for transporting collected samples towards sequencing centers.

Beyond the diagnostics requirements, the identification of the highly transmissible SARS-CoV-2 variant B.1.1.7—originated in the south of England—has strongly accelerated the worldwide interest in tracking SARS-CoV-2 variants' occurrence. More recently, two other extremely infectious variants, namely the P.1 issued on the

¹Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ of Évry, Paris-Saclay University, 91057 Evry, France. ²Genoscope, Institut de Biologie François-Jacob, Commissariat À L'Energie Atomique (CEA), Université Paris-Saclay, Evry, France. ³These authors contributed equally: François Stüder and Jean-Louis Petit. ✉email: mmendoza@genoscope.cns.fr

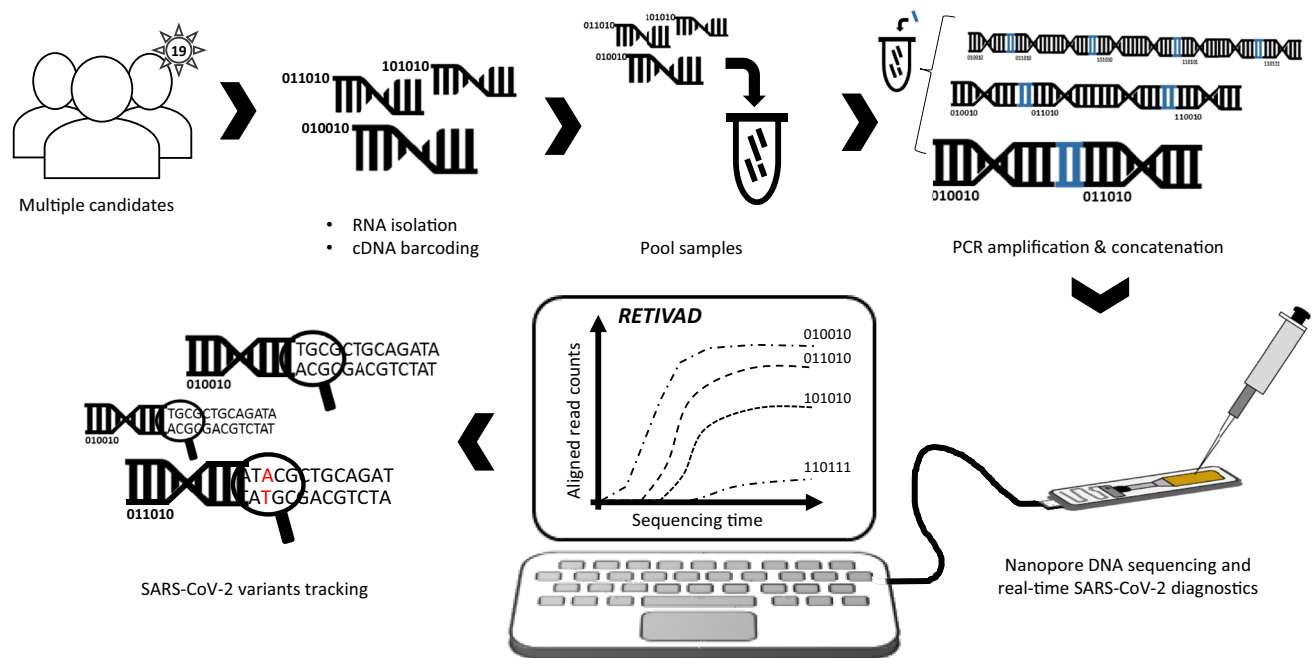


Figure 1. Scheme illustrating the proposed real-time SARS-CoV-2 diagnostics and variants tracking strategy. RNA samples isolated from multiple candidates are converted into complementary DNA (cDNA) such that samples associated to each candidate are labelled by molecular DNA barcodes appended to their extremities. Barcoded-cDNA samples are pooled together and PCR amplified in presence of a bridge sequence (blue) allowing to generate concatemers. Long DNA concatemers are loaded into the MinION nanopore sequencer for real-time SARS-CoV-2 diagnostics and variants tracking with the help of RETIVAD.

Brazilian city of Manaus and the B.1.351 initially detected in South Africa, were described, and unsurprisingly further others are expected to be discovered, notably due to the long period of time on which the pandemic situation is lasting⁶.

Herein, we describe a proof of concept study for population-scale diagnostics and variants' tracking using massive parallel DNA sequencing performed with the MinION Oxford Nanopore Technology (ONT). ONT MinION sequencers are well-known for their portability, their reduced cost, and their simplicity of use; thus, representing a suitable strategy for its implementation even in remote places over the world (reviewed⁷). Indeed, this technology is widely used for SARS-CoV-2 whole-genome sequencing, notably by the use of the amplicon sequencing protocol developed by the ARTIC Network (<https://artic.network/ncov-2019>)⁸. Furthermore, a Covid-19 diagnostics strategy combining Loop-mediated isothermal Amplification (LAMP) with Nanopore sequencing has been also described⁹.

In addition to a dedicated molecular biology protocol, we deliver a computational solution (RETIVAD: Real-Time Variants Detector) providing the outcome of the diagnostic in a real-time mode during sequencing; allowing to earn time for diagnostics report, essential in the current context of the pandemic situation. Finally, we have also demonstrated that our strategy is able to perform SARS-CoV-2 variants' detection across multiple candidates, notably by targeting the SPIKE gene instead of covering the whole viral genome (Fig. 1). This targeted variants' tracking strategy allows increasing ~10× the potential sequencing coverage delivered during the assay, thus allowing to redistribute this resource towards the variants' screening of a large number of candidates at once.

In summary, the proposed strategy satisfies the current need for counting with a population diagnostic assay, and discriminates between the various SARS-CoV-2 variants, which are recognized to potentially present variable infection properties as well as clinical outcomes^{6,10}.

Results

A combinatorial molecular barcode strategy for SARS-CoV-2 diagnostics over multiple candidates. With the aim of performing SARS-CoV-2 diagnostics over multiple candidates with the MinION sequencer, we combined a regular reverse-transcription assay, with a particular PCR amplification allowing the generation of long DNA concatemers. This strategy allows to target short amplicons, like those used on quantitative real-time PCR strategies, but at the same time take advantage of the long fragments sequencing capabilities of the Oxford Nanopore technology.

Reverse transcription (RT) is performed with a Covid-19 specific primer, presenting in addition, a unique molecular barcode (20nt length) and a common sequence adapter (Gibson sequence; 30nt length) (Fig. 2A). Hence, multiple candidate' samples are reverse-transcribed with a unique molecular barcode per candidate, allowing their identification after massive parallel DNA sequencing.

After RT assay, candidates' samples are pooled for PCR amplification into a single reaction. For it, a second Covid-19 specific primer, presenting, in addition, a unique molecular barcode (20nt length) and a

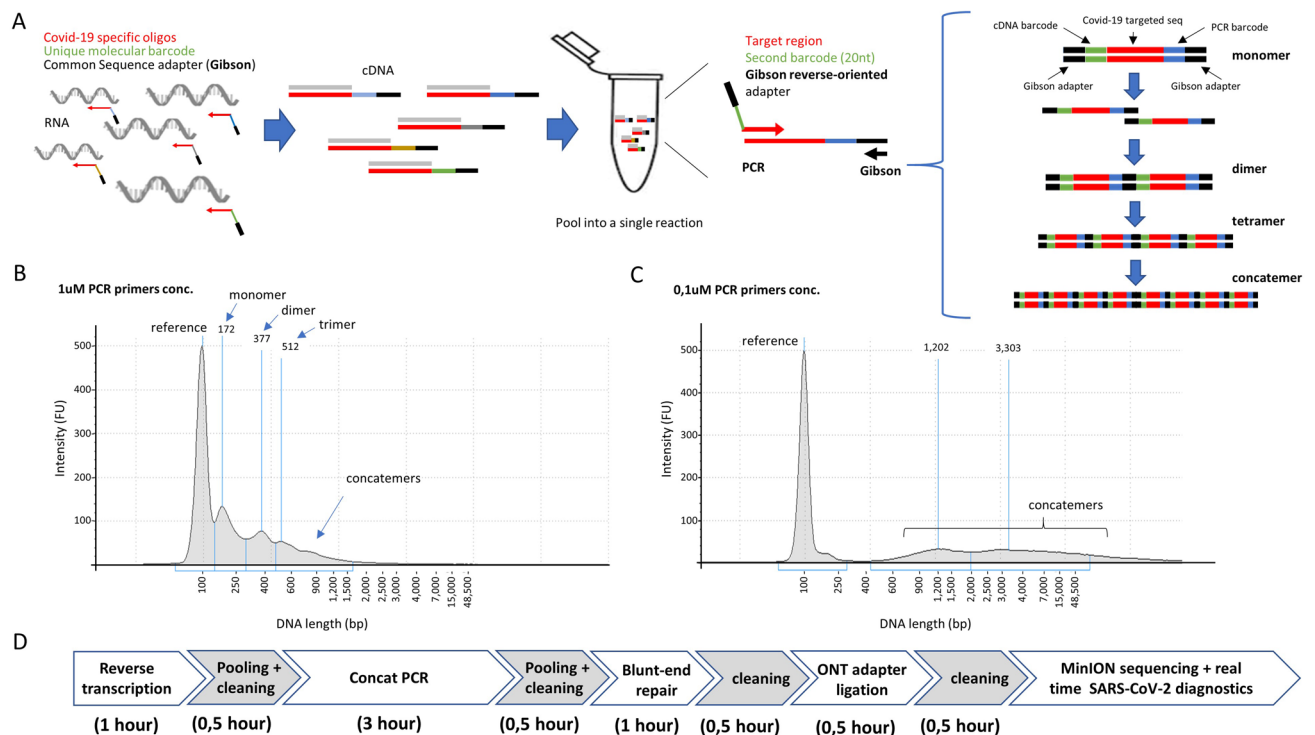


Figure 2. Dual molecular DNA barcoding and PCR-driven concatemerization applied to SARS-CoV-2 diagnostics. **(A)** Reverse transcription is performed with oligonucleotides targeting a Covid-19 region, but in addition containing a unique molecular barcode per sample and a common adapter sequence (Gibson). Reverse transcription assays (cDNA) performed for multiple candidates are pooled together for amplification. PCR amplification is performed with oligonucleotides targeting the common Gibson sequence and a Covid-19-specific primer hosting a second molecular barcode as well as a complementary Gibson sequence. During PCR cycles, amplified products are able to concatenate thanks to the Gibson flanking sequences. We have called this particular PCR strategy “Concat-PCR”. **(B)** DNA electropherogram (Agilent TapeStation system) illustrating concatemer products obtained with PCR primers at 1uM final concentration. **(C)** DNA electropherogram for a concat-PCR assay performed with PCR primers at 0.1uM final concentration, demonstrating a preferential production of long concatemers in contrast to those observed in **(B)**. **(D)** Timeline illustrating the various molecular biology steps required for samples preparation prior engaging samples into MinION DNA sequencing.

reverse-oriented sequence for the common Gibson sequence adapter (30nt length) is used together with a Gibson primer (Fig. 2A). This strategy allows to incorporate a second unique molecular barcode, thus providing a combinatorial barcoding strategy per candidate which is extremely useful for increasing the number of candidates to interrogate without having to linearly increase the number of primers to purchase; an aspect which is of major relevance due to their long length (> 70nt).

The presence of Gibson sequences on both extremities of the PCR amplified products allows generating concatemers during amplification (Fig. 2A). While this process takes place in an uncontrolled manner during PCR amplification, low primer concentrations were shown to give rise preferentially to long concatemers (Fig. 2B,C). As consequence, concatenated PCR products can be purified from long primers (> 70nt length) with SPRI-select reagent (Beckman; B23318), without the risk of losing the short amplicons. Furthermore, the use of long concatemers increased the average sequencing coverage of MinION by a factor of ~9 (i.e. 1 million concatemers sequenced corresponds to 9 million monomers; Figure S1 and Fig. 3E), such that a large number of candidates can be potentially screened like in previous population-scale diagnostics efforts based on the use of Illumina instruments^{3–5}.

Overall, the proposed pipeline is composed of a standard molecular biology strategy, requiring less than 8 h of samples preparation prior Nanopore sequencing (Fig. 2D and “Methods” section).

Real-time SARS-CoV-2 diagnostics over multiple pseudo-candidates presenting variable amounts of viral copies.

One of the major bottlenecks for population-scale Covid-19 diagnostics based on massive parallel DNA sequencing is the required time to deliver the outcome of the diagnostic. In addition to the molecular biology workflow required for samples’ processing till sequencing library preparation, DNA sequencing, performed on Illumina instruments, requires an incompressible processing time, which delays downstream analyses since they can only be performed at the end of the whole sequencing process. In contrary, Oxford Nanopore sequencers deliver sequenced reads on a rolling basis (4000 sequenced reads per generated fastq file). Hence, we have developed a computational pipeline dedicated to (1) collect available fastq files; (2)

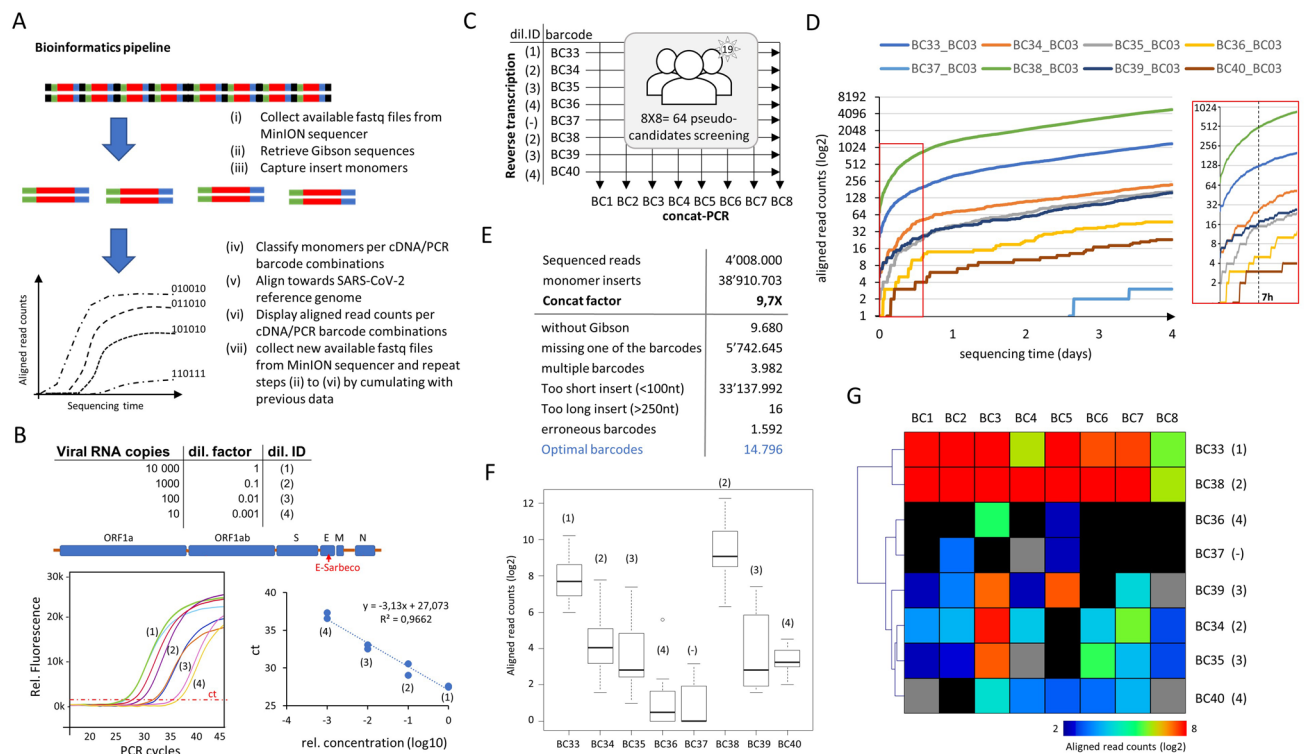


Figure 3. SARS-CoV-2 diagnostics assay with Nanopore DNA sequencing. **(A)** Scheme of the bioinformatics pipeline within RETIVAD. Long concatemer sequenced reads are computationally fragmented in monomers with the help of the Gibson sequence location. Monomers are stratified on the basis of flanking barcodes (cDNA and PCR incorporated barcodes) and aligned to the Covid-19 viral reference genome to verify its specificity. Considering that Oxford Nanopore instruments provide sequenced data during sequencing, the aforementioned steps are performed on real-time, such that the number of aligned read counts per candidate (defined by the combination of cDNA & PCR incorporated DNA barcodes) are assessed and displayed during the MinION sequencing process. **(B)** The synthetic SARS-CoV-2 RNA control sample (Twist Bioscience; Wuhan-Hu-1 strain) was subsequently diluted and used for reverse transcription and RT-qPCR assay targeting the E-Sarbeco viral region. **(C)** Reverse transcription assay scheme for 8 conditions corresponding to different viral RNA dilutions (dilution ID corresponding to that presented in **(B)**) and targeting the E-Sarbeco region. Each of these reverse transcription assays are identified by the indicated molecular barcodes (BC33–BC40). Note that among them a sample devoid of viral RNA is also included (BC37). All 8 samples are pooled at the end of the assay and concat-PCR amplified with a set of other eight barcoded primers targeting the E-Sarbeco gene (BC1–BC8), providing a combinatorial barcoding strategy corresponding to 64 pseudo-candidates. **(D)** Real-time display illustrating the number of aligned read counts associated to the E-Sarbeco region per dual barcode combination (BC3 vs all cDNA barcodes) assessed during Nanopore sequencing. Inset: Zoom view highlighting that the first seven hours of DNA sequencing are sufficient to reveal read count differences between samples. **(E)** Summary of the bioinformatics pipeline assessed at the end of 4 days of DNA sequencing. Concatenation factor is inferred from the ratio between the number of detected monomers and the initial number of sequenced concatemers. Optimal barcodes correspond to the number of sequences presenting an optimal monomer length (> 100nt; < 250nt) and presenting both cDNA & PCR barcode sequences. **(F)** Boxplots illustrating the total aligned read counts associated to each cDNA barcode (BC33–BC40) and the different PCR barcodes. Each of the boxplots are labelled with their corresponding viral RNA dilution (dil.ID indicated in **(B)**). **(G)** Heatmap illustrating the total aligned read counts associated to each cDNA barcode (BC33–BC40) and the different PCR barcodes (BC1–BC8). Note that **(F,G)** reveals a sensitivity of 100 viral copies, since the next RNA dilution condition appears in some cases at the same level of the negative samples.

retrieve Gibson adapter sequences within sequenced molecules; (3) split them into monomers; (4) evaluate the presence of cDNA & PCR molecular barcodes in their flanking regions; (5) align the inner monomer sequences towards the SARS-CoV-2 reference genome; and (6) display the aligned read counts per cDNA/PCR barcode combinations. This pipeline is setup to be reinitiated in an automated time interval (e.g. every 10 min) such that newly available sequences can be cumulated to the previous analyses, thus providing a real-time view of aligned read counts associated with cDNA/PCR barcode combinations (Fig. 3A).

To validate the performance of the proposed diagnostics strategy, synthetic SARS-CoV-2 RNA control samples (Twist Biosciences) were used for mimicking variable viral RNA copies per assay. Specifically, subsequent dilutions of the synthetic Wuhan-Hu-1 (GeneBankID: MN908947.3) SARS-CoV-2 RNA control sample has been used for reverse transcription, followed by a quantitative PCR assay targeting a region on the E gene of

SARS-CoV-2, defined by primer sequences described previously (E-Sarbeco)¹¹. This quantitative PCR allowed to detect as few as 10 viral copies of the synthetic SARS-CoV-2 RNA control sample (Fig. 3B).

For Nanopore DNA sequencing-based diagnostic, as described on Fig. 2A, we have designed eight barcoded E-Sarbeco reverse primers (targeting the same amplicon region used for the quantitative PCR assay), and other eight barcoded E-Sarbeco forward primers (incorporating the required Gibson sequence), for concat-PCR amplification. This combinatorial strategy allows to mimic the screening of 64 pseudo-candidates (Fig. 3C).

Subsequent dilutions of the synthetic SARS-CoV-2 RNA control sample were used for reverse transcription, performed with defined barcoded E-Sarbeco primers (BC33-BC36; BC38-BC40), in addition to a negative control sample devoid of viral RNA material (BC37) (dilution ID in Fig. 3C). The real-time diagnostic assay revealed significant differences between samples presenting different viral RNA copies based on the number of aligned read counts towards the E-Sarbeco target region. Such differences, observed as early as 7 h after DNA sequencing, revealed that samples issued from high viral RNA titers presented systematically higher amounts of aligned read counts to the E-Sarbeco target region (Fig. 3D and Figure S2). Furthermore, after 4 days of DNA sequencing (72 h of DNA sequencing, and 1 day more for finalizing base-calling) a total of ~4 million reads were sequenced, corresponding to ~38 million monomers ($9.7 \times$ concatemerization factor) (Fig. 3E). Despite this large gain on monomer sequences, only 14 thousand optimal sequences (i.e. presenting both cDNA and PCR flanking barcodes and within the expected monomer amplicon insert size) were used for the diagnostic assay, all others presenting either too short inserts (< 100nt) or missing one or both of the required barcodes, essential for the identification of the samples. Noteworthy, the analysis of ~14 thousand optimal sequences was sufficient for discriminating samples with viral RNA titers ranging within 1 to 100 fold dilutions of the synthetic Covid-19 RNA material (Fig. 3F,G); equivalent to 100 copies detected with the quantitative PCR assay (Fig. 3B). Indeed, the sensitivity of the assay fails to discriminate in some cases between the negative control samples (BC37) and those issued from the least viral RNA titer (1/1000 dilution; BC36) (Fig. 3F,G).

Real-time SARS-CoV-2 variants' tracking over multiple pseudo-candidates. Considering the current needs on tracking SARS-CoV-2 variants' emergence and spreading over the population, we have evolved the aforementioned real-time SARS-CoV-2 diagnostic strategy into a variants' tracking system. Because all current described SARS-CoV-2 variants present an important number of mutations within the SPIKE gene, we have focused the real-time tracking system to this region. In fact, by concentrating the sequencing efforts towards the SPIKE gene (~3.8 kb) we can afford to combine diagnostic and variants' tracking into a single assay, which per se represents a major progress relative to the current strategies.

In contrast to the aforementioned diagnostics effort targeting a short amplicon (E-Sarbeco), variants' tracking over the whole SPIKE gene uses a random-hexamer sequence for the reverse transcription step, and four primers targeting the SPIKE gene at different regions (~1 kb distance among each primer) (Fig. 4A). Like in the case of the diagnostics assay, the random-hexamer primer used for the reverse transcription present a unique molecular barcode as well as the common Gibson sequence. In contrary, the PCR primers (P1-P4) present a unique molecular barcode but they are devoid of the Gibson sequence. In fact, PCR amplification over cDNA issued from random-hexamer priming leads to amplicons of variable length (including fragments > 1 kb; see Fig. 4B & Figure S3), thus the concatemerization step used on the short amplicon strategy does not appear essential. To simplify its use, PCR amplification is performed in presence of a pool of all four primers targeting the SPIKE region (P1-P4) and a Gibson primer sequence (in an equimolar concentration; i.e. $4 \times$ Gibson primer for $1 \times$ for each SPIKE targeting primer).

To better mimic a real situation, in which the diagnostics/variants' tracking assay is performed with samples collected from human candidates (e.g. nasopharyngeal swabs), human RNA—issued from cell lines in culture—has been combined with the synthetic viral RNA samples, such that the random-hexamer primers used for the reverse transcription assay could also provide means to count with a positive control for sample collection. Furthermore, for addressing variants tracking performance, in addition to the Wuhan-Hu-1 SARS-CoV-2 RNA control sample, we have used synthetic viral RNA samples corresponding to the highly transmissible SARS-CoV-2 variant B.1.1.7, first described in the UK in December 2020, as well as the strain HF2393 described as a particular clade circulating in France since January 2020¹² (Fig. 4C).

To perform diagnostics and variants' tracking assay over multiple pseudo-candidates, we have used eight barcoded random-hexamer primers for reverse transcription (BC37-BC44) in presence of different viral RNA titers (consecutive dilutions) and issued from the aforementioned synthetic viral RNA strains (Fig. 4D). To multiply the number of pseudo-candidates, we have amplified the eight barcoded cDNA products with four barcoded primers, either targeting the SPIKE gene (BC1-BC4 corresponding to pools of P1-P4) or targeting the E-Sarbeco region (BC5-BC8).

The cDNA products issued from barcoded-random-hexamer priming were first validated by quantitative PCR amplification, targeting either the human RNase-P gene or the viral E-Sarbeco region (Fig. 4E). Then, the PCR products generated with barcoded primers targeting SPIKE (BC1-BC4) and those targeting the E-Sarbeco region (BC5-BC8) were pooled together and analyzed by ONT sequencing. The real-time diagnostic assay for the SPIKE gene or the E-Sarbeco region, revealed significant differences between samples presenting different viral RNA copies based on the number of aligned read counts. Specifically, the negative control sample (BC44) presented systematically the least number of aligned counts for the SPIKE and the E-Sarbeco region, and in the opposite, the sample presenting the most concentrated viral RNA titers (BC43) presented the highest number of aligned reads per aforementioned target regions (Fig. 4F,G).

In addition to the diagnostic outcome, RETIVAD performs a variants' detection relative to the SARS-CoV-2 reference genome. In fact, as illustrated on Fig. 4H, RETIVAD detected the presence of 7 significant variants within the SPIKE gene on all four samples issued from the BC37-barcoded cDNA (BC1-BC4 $\times$ BC37). Among

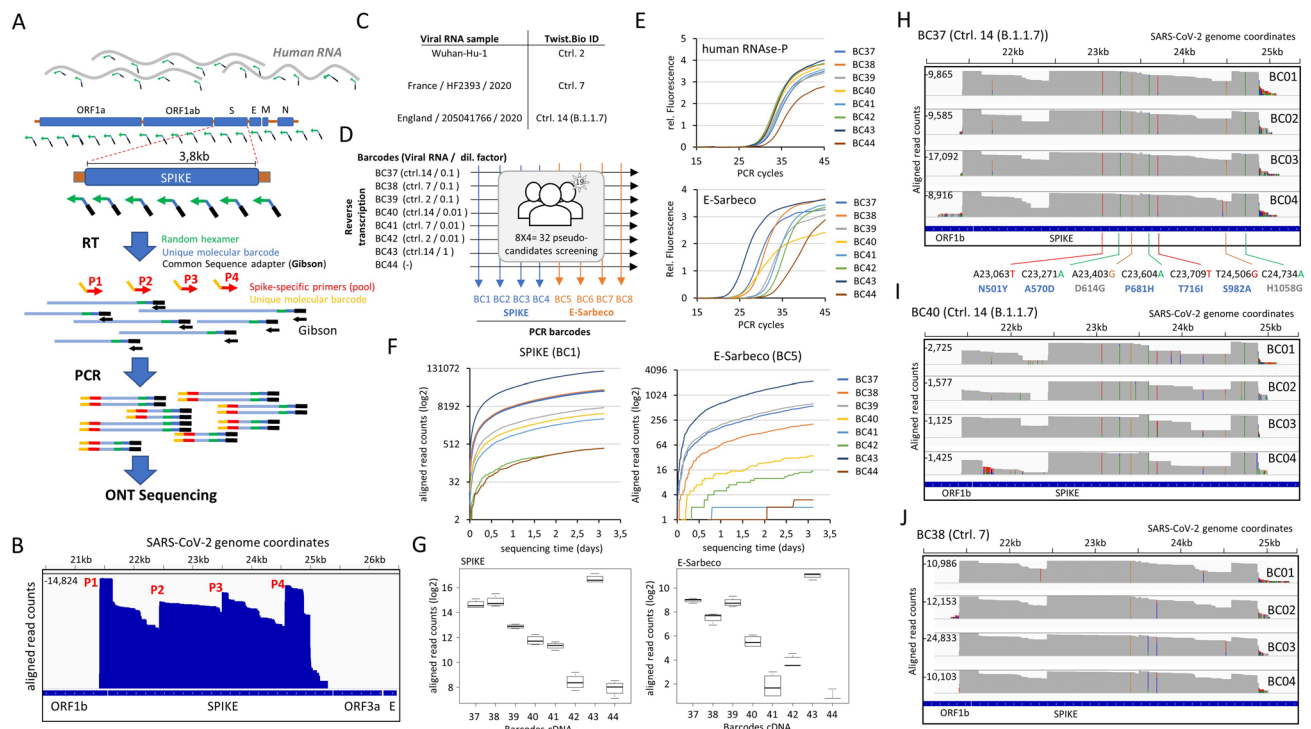


Figure 4. SARS-CoV-2 variants tracking within the SPIKE gene. **(A)** Scheme illustrating the SPIKE targeted strategy for variants tracking with Oxford Nanopore technology (ONT) sequencing. Reverse transcription is performed with random hexamer primers presenting unique molecular barcodes and the Gibson adapter sequence. This approach, allows to generate cDNA products of variable length covering the whole viral but also the human genome. Barcoded cDNA samples are PCR amplified with four SPIKE-specific primers (each of them including a common molecular DNA barcode and spaced by ~1 kb distance) and one Gibson primer targeting the adapter sequence introduced during reverse transcription. For facilitating the assay, all four SPIKE and Gibson primers are used within an equimolar pool. PCR amplified samples are analyzed by Oxford Nanopore MinION DNA sequencing. **(B)** Genome browser view displaying the aligned read counts obtained after ONT sequencing. Location of each of the SPIKE-targeting primers (P1–P4) is displayed and corresponds to regions presenting the highest read counts within the SPIKE gene. **(C)** Summary of the used synthetic SARS-CoV-2 viral RNA samples, corresponding to the Wuhan-Hu-1 control sample (reference strain) and two described variant strains (HF2393 and B.1.1.7). **(D)** Reverse transcription assay scheme for 7 conditions corresponding to different dilutions of viral RNA issued from the strains illustrated in (C) (BC37–BC43) and a control sample devoid of viral RNA (BC44). In all cases, 15 ng of human RNA material—issued from cell lines in culture—has been also introduced. All 8 samples are pooled at the end of the reverse transcription assay and PCR amplified with either four pools of barcoded primers targeting the SPIKE gene (BC1–BC4) or four barcoded primers targeting the E-Sarbeco region (BC5–BC8), providing a combinatorial barcoding strategy corresponding to 32 pseudo-candidates. **(E)** Quantitative PCR targeting either the human RNase-P gene or the viral E-Sarbeco region, performed over the reverse transcription assays described in (D). **(F)** Real-time Covid-19 diagnostic performed by RETIVAD displaying the number of aligned read counts per barcoded cDNA sample (described in (D)) for the gene SPIKE (BC1) or the E-Sarbeco region (BC5). **(G)** Boxplots illustrating the total aligned read counts associated to the gene SPIKE or the E-Sarbeco region for each of the conditions described in (D). **(H)** Variant analysis performed by RETIVAD over the barcoded sample 37 (BC37) associated to the corresponding PCR barcodes (BC1–BC4). The displayed genome browser view reveals the presence of multiple nucleotide variants (colored bars), among them seven being retrieved on all four samples (BC1–BC4) and with a significant read coverage. Five of them (A23,063T; C23,271A; C23,604A; C23,709T; T24,506G) correspond to missense mutations that were described within the B.1.1.7 UK variant strain (blue). In addition, two other missense mutations were also detected; the D614G (previously described within SARS-CoV-2 European strains) and the H1058G. **(I)** Similar to (H) but for the sample BC40, corresponding to a tenfold lower RNA titer than that used for sample BC37. Note that the same missense mutations retrieved in (H) are observed (colored bars), and this in despite of the lower aligned read counts. **(J)** Variants detection assay over the sampled BC38, corresponding to the SARS-CoV-2 variant strain HF2393, described in France in 2020. In this case, only the missense mutation D614G is retrieved among all four samples (BC1–BC4).

them, five correspond to those previously described within the strain B.1.1.7 (N501Y, A570D, P681H, T716I and S982A); coherent with the fact that samples associated to the BC37 were generated with the synthetic viral

RNA corresponding to such SARS-CoV-2 variant (Fig. 4D). Similarly, variants detection performed on samples issued from the BC40-barcode cDNA (BC1-BC4 x BC40) revealed a similar signature, in spite of their lower coverage (Fig. 4I). In fact, samples associated to the BC40 were also generated with the synthetic viral RNA corresponding to the B.1.1.7 strain, but with a tenfold lower titer than in the case of the BC37 (Fig. 4D).

Finally, variants' tracking performed on samples issued from the BC38-barcode cDNA (BC1-BC4 x BC38; French strain HF2393) revealed a single common mutation, D614G, previously described within the French strain HF2393¹², but also among other European strains, including the B.1.1.7, as observed within the variants detected on samples issued from BC37 and BC40 (Fig. 4H,I).

Discussion

Despite the current vaccination efforts, the pandemic situation anticipates the propagation of the recently described variants, as well as the emergence of new others, notably due to the major discrepancies between countries to access to optimal vaccination programs⁶. As consequence, counting with methodologies for Covid-19 diagnostics but also variants' tracking on a population scale remains essential.

Currently, diagnostics and variants' tracking are performed in two steps, which per se delays the time for obtaining a complete diagnostic. Furthermore, variants' tracking relies on full genome sequencing of SARS-CoV-2, which represent an expensive and labor-intensive effort, incompatible with population-scale assays. Finally, most of the variants' tracking efforts, relying on the use of large and expensive sequencing instruments, requires to centralize samples processing to dedicated institutions, which are rather absent on poor countries.

To address all these pitfalls, we propose herein a methodology able to (1) perform SARS-CoV-2 diagnostic and variants' tracking at the same time; (2) shorten the time for accessing to the outcome of the diagnostic/variants' detection by the use of a real-time tracking system; (3) take advantage of the compact and affordable Oxford Nanopore MinION DNA sequencing instrument, notably for deploying the aforementioned diagnostics and variants' tracking platform anywhere over the world.

Within this study we have validated the proposed Nanopore DNA sequencing-based diagnostic, by targeting a region on the E gene of SARS-Cov-2 (E-Sarbeco) on a total of 64 pseudo-candidate samples (Fig. 3). Importantly, the few number of optimal sequences required for performing this diagnostic assay suggests that larger candidate screenings (several hundreds) are possible within the scale of the sequencing-depth afforded by the MinION sequencer.

The proposed variants' tracking focus on the analysis of the SPIKE gene instead of the whole SARS-CoV-2 genome. In fact, the SPIKE gene has shown to concentrate the majority of the currently discovered variants, which can be rationalized by the fact that it codes for the protein required for the interaction with the human cells. Noteworthy, most of the current vaccines against the Covid-19 are also based on the SPIKE region. Hence, concentrating variants' tracking towards this gene appears an optimal compromise to increase the number of candidates to screen at once. This philosophy is shared by a recent work describing the HiSpike method, performing variants' detection within the SPIKE gene with the help of the small Illumina MiSeq instrument¹³. Illumina instruments being a short fragments sequencing approach, HiSpike requires 42 primer pairs generating amplicons of 400 nucleotides of length for covering the SPIKE region. In contrary, our proposed strategy, based on the use of long DNA fragments, compatible with the Nanopore DNA sequencing technology, relies on 4 primers targeting the SPIKE region, one random primer for the cDNA step and a common Gibson sequence. In fact, a strategy for SARS-CoV-2 diagnostic and variants' surveillance, dedicated to be deployed anywhere on the world does not only requires to count with low cost instruments, but also involve a reduced number of reagents for allowing its use on a long-standing manner.

During the time of the final revision of this article, new variants derived from the lineage of B.1.1.7 has gained major attention notably due to the fact that they are vastly spreading in the UK, USA, and other locations (B.1.617.1, also known as Kappa; B.1.617.2 also known as "Delta" and B.1.618 also known as "triple mutant," as suspected to have evolved from B.1.617)¹⁴. In all cases, new mutations on the gene SPIKE were observed, further supporting our strategy for targeting this genomic region for variants surveillance.

For all these reasons, we are convinced that our proposed strategy, accompanied by the dedicated computational solution RETIVAD, can represent a game-changer approach for tracing variants' progression in the following months.

Methods

SARS-CoV-2 synthetic RNA. The following Synthetic SARS-CoV-2 RNA control samples, delivered at a concentration of 1 million copies per microliter has been used in this study:

- Wuhan-Hu-1 (GeneBankID: MN908947.3); Twist Biosc. ID: Ctrl2.
- France/HF2393/2020 (EPI_ISL_418227); Twist Biosc. ID: Ctrl7.
- England/205,041,766/2020 (EPI_ISL_710528); Twist Biosc. ID: Ctrl14 (B.1.1.7).

Barcode primer design. Primers used for reverse transcription are composed by (1) a fragment of the Gibson sequenced (redGibs: GAAGAACCTGTAGATAACTCGCTGT), (2) a barcode sequence issued from the ONT PCR Barcoding Expansion 1–96 kit (EXP-PBC096) and (3) the target sequence. For targeting the E-Sarbeco region within the SARS-CoV-2 genome, the reverse primer sequence provided by Corman et al.¹¹, has been used:

E_Sarbeco_R: ATATTGCAGCAGTACGCACACA

For SPIKE-centered variants tracing, the target sequence has been replaced by a random hexamer, notably to decrease the number of required primers during the assay.

The following barcoded primers targeting the E-Sarbeco region were used for reverse transcription:

redGibsBC33_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTCAGACTTGGTACGGTTGGGTAAC
TATATTGCAGCAGTACGCACACA
redGibsBC34_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTGGACGAAGAACTCAAGTCAA
AGGCATATTGCAGCAGTACGCACACA
redGibsBC35_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTCTACTTACGAAGCTGAGGGACTG
CATATTGCAGCAGTACGCACACA
redGibsBC36_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTATGTCCCAGTTAGAGGAGGAAAC
AATATTGCAGCAGTACGCACACA
redGibsBC37_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTGCTTGGCATTGATGCTTAGTATC
AATATTGCAGCAGTACGCACACA
redGibsBC38_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTACCACAGGAGGACGATACAGAGA
AATATTGCAGCAGTACGCACACA
redGibsBC39_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTCCACAGTGTCAACTAGAGCCTCT
CATATTGCAGCAGTACGCACACA
redGibsBC40_E_Sarbeco_R: GAAGAACCTGTAGATAACTCGCTGTAGTTTGGATGACCAAGGATAGC
CATATTGCAGCAGTACGCACACA

The following barcoded random-hexamer primers were used for reverse transcription:

redGibsBC37_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTGCTTGGCATTGATGCTTAGTATC
ANNNNNN
redGibsBC38_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTACCACAGGAGGACGATACAGAGA
ANNNNNN
redGibsBC39_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTCCACAGTGTCAACTAGAGCCTCT
CNNNNNN
redGibsBC40_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTAGTTTGGATGACCAAGGATAGC
CNNNNNN
redGibsBC41_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTGGAGTTCGTCCAGAGAAGTACAC
GNNNNNN
redGibsBC42_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTCTACGTGTAAGGCATACCTGCCA
GNNNNNN
redGibsBC43_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTCTTTCGTTGTTGACTCGACGGTA
GNNNNNN
redGibsBC44_Rhexamer_R: GAAGAACCTGTAGATAACTCGCTGTAGTAGAAAGGGTTTCTTCCCACT
CNNNNNN

For PCR amplification a second set of barcoded primers were designed presenting the following structure: (1) A reverse-complementary Gibson fragment (revGibs: ACAGCGAGTTATCTACAGGTTCTTCAATGT), (2) a barcode sequence issued from the ONT PCR Barcoding Expansion 1–96 kit (EXP-PBC096), and (3) the target sequence. For targeting the E-Sarbeco region within the SARS-CoV-2 genome, the forward primer sequence provided by Corman et al.¹¹, has been used:

E_Sarbeco_F: ACAGGTACGTTAATAGTTAATAGCGT

The following barcoded primers targeting the E-Sarbeco region were used in this study:

revGibs_BC01_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTAAAGAAAGTTGTCCGT
GTCTTTGTGACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC02_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTTTCGATTCCGTTTGTAGTC
GTCTGTACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC03_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTGAGTCTTGTGTCCCA
GTTACCAGGACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC04_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTTTCGATTCTATCGTGT
TCCCTAACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC05_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTCTTGTCCAGGGTTTG
TGTAACCTTACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC06_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTTTCTCGCAAAGGCAG
AAAGTAGTCACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC07_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTGTGTTACCGTGGGAA
TGAATCCTTACAGGTACGTTAATAGTTAATAGCGT
revGibs_BC08_E_Sarbeco_F: ACAGCGAGTTATCTACAGGTTCTTCAATGTTTCAGGGAACAAACC
AAGTTACGTACAGGTACGTTAATAGTTAATAGCGT

For SPIKE-centered variants tracing, the reverse-complementary Gibson sequence has been excluded, notably because the PCR amplification performed over random-hexamer generated cDNA templates generate large fragments (> 1 kb). In contrary, the amplicon product issued from E-Sarbeco targeting oligonucleotides gives rise to short amplicons—compatible with quantitative PCR assays, thus requiring concatemerization prior nanopore sequencing.

With the aim of covering the SPIKE gene (~ 3.8 kb), the following four primers targeting the SPIKE region with ~ 1 kb interval within them were designed:

Spike1_F: AGGGGTACTGCTGTTATGTCT
Spike2_F: TGCACTTGACCCTCTCTCAG
Spike3_F: GCAGGCTGTTTAATAGGGGC
Spike4_F: TGCAGACATATGTGACTCAACA

The following Spike targeting primers including the corresponding barcode sequences were used for this study:

BC01_Spike_1_F: AAGAAAGTTGTCGGTGTCTTTGTGAGGGGACTGCTGTTATGTCT
 BC02_Spike_1_F: TCGATTCCGTTTGTAGTCGTCTGTAGGGGACTGCTGTTATGTCT
 BC03_Spike_1_F: GAGTCTTGTGTCCAGTTACCAGGAGGGGACTGCTGTTATGTCT
 BC04_Spike_1_F: TTCGGATTCTATCGTGTTCCTAAGGGGACTGCTGTTATGTCT
 BC01_Spike_2_F: AAGAAAGTTGTCGGTGTCTTTGTGTGCACTTGACCCTCTCTCAG
 BC02_Spike_2_F: TCGATTCCGTTTGTAGTCGTCTGTTGCACTTGACCCTCTCTCAG
 BC03_Spike_2_F: GAGTCTTGTGTCCAGTTACCAGGTGCACTTGACCCTCTCTCAG
 BC04_Spike_2_F: TTCGGATTCTATCGTGTTCCTATGCACTTGACCCTCTCTCAG
 BC01_Spike_3_F: AAGAAAGTTGTCGGTGTCTTTGTGGCAGGCTGTTAATAGGGGC
 BC02_Spike_3_F: TCGATTCCGTTTGTAGTCGTCTGTGCAGGCTGTTAATAGGGGC
 BC03_Spike_3_F: GAGTCTTGTGTCCAGTTACCAGGGCAGGCTGTTAATAGGGGC
 BC04_Spike_3_F: TTCGGATTCTATCGTGTTCCTAGCAGGCTGTTAATAGGGGC
 BC01_Spike_4_F: AAGAAAGTTGTCGGTGTCTTTGTGTGCAGACATATGTGACTCAACA
 BC02_Spike_4_F: TCGATTCCGTTTGTAGTCGTCTGTTGCAGACATATGTGACTCAACA
 BC03_Spike_4_F: GAGTCTTGTGTCCAGTTACCAGGTGCAGACATATGTGACTCAACA
 BC04_Spike_4_F: TTCGGATTCTATCGTGTTCCTATGCAGACATATGTGACTCAACA

Reverse transcription. Reverse transcription assays were performed using SuperScript™ IV Reverse Transcriptase (ThermoFisher Scientific; 18090200; 0.5 ul superscript IV (200U/ul)) with a standard reaction volume of 20ul including 0.5ul of super RNaseIn (ThermoFisher Scientific; AM2694; 20U/ul), 1ul DTT (0.1 M), 4 ul SSIV buffer (5X), 0.5 ul dNTP (10 mM each), 1 ul of synthetic SARS-CoV-2 RNA, 1 ul of 2uM gene-specific-barcoded oligonucleotide (e.g. E-Sarbeco barcodes bc33-bc48; 0.1uM final concentration) or 1 ul of 50uM random hexamer-barcoded oligonucleotide (final concentration: 2.5uM). For assays including random hexamer-barcoded primers, 1ul of human RNA (15 ng/ul; collected from cell lines in culture) were added to the assay together to the synthetic SARS-CoV-2 RNA and prior adding the random hexamers.

Synthetic SARS-CoV-2 RNA was added at a copy number of 1 million (dilution factor, d.f: 1), 100 thousand (d.f: 0.1), 10 thousand (d.f: 0.01) and 1 thousand (d.f: 0.001) per reaction respectively. Note that these consecutive dilutions give rise after downstream dilutions to the number of copies of 10 thousand, 1 thousand, 100 and 10 respectively, analyzed by quantitative PCR (Fig. 3B).

Multiple reverse transcription assays—each of them performed in presence of a defined barcoded primer—were incubated at 50 °C during 30 min, then at 80 °C 10 min. Barcoded cDNA samples were pooled (8 samples at once) and cleaned with SPRIselect reagent (Beckman; B23318) with a ratio of 0.5x. Cleaned cDNA is recovered on 80ul (half of the total initial cleaned volume; i.e. 2 × concentration). 40ul of this pooled & cleaned cDNA will be used for PCR amplification (5ul × 8 PCR reactions; see below), while the remaining material is used for quantitative PCR evaluation.

SARS-CoV-2 quantitative PCR validation. Barcoded-cDNA material has been validated by quantitative PCR assay with the following primer sequences:

Primers targeting region on E gene of SARS-Cov-2 (annealing temp: 58 °C):

E_Sarbeco_F: ACAGGTACGTTAATAGTTAATAGCGT

E_Sarbeco_R: ATATTGCAGCAGTACGCACACA

Primers targeting the human gene RnaseP (annealing temp: 60 °C):

Human_RNASE_P-F: AGATTTGGACCTGCGAGCG

Human_RNASE_P-R: GAGCGGCTGTCTCCACAAGT

Quantitative PCR assays were performed with QuantiTect SYBR Green PCR Kit (ref: 204,145) over a 1/10 diluted cDNA material.

Concat PCR amplification and nanopore sequencing. PCR amplification targeting the E-Sarbeco amplicon is performed on 25ul final volume including 12.5ul of Phusion Hot Start II High-Fidelity PCR Master Mix (ThermoFisher ref: F565), 2.5ul of 1uM barcoded-E-Sarbeco targeting primer (BC1-8), 2.5ul of 1uM short Gibson primer (AGAACCTGTAGATACTCGCTGT) and 5ul of the pooled & cleaned cDNA.

For short-amplicon E-Sarbeco PCR amplification the following thermal cycling program is used:

1. 98 °C 30 s
2. 98 °C 10 s
3. 61 °C 20 s
4. 72 °C 30 s
5. Repeat 2–4 steps 25×
6. 98 °C 10 s
7. 61 °C 20 s
8. 72 °C 2 min
9. Repeat 6–8 steps 20×
10. 72 °C 10 min
11. Keep at 12 °C

For PCR amplification targeting SPIKE and E-Sarbeco on random-hexamers generated cDNA material the assay is performed with a pool of barcoded SPIKE primers and shortGibson. For 100ul mix 10ul of barcoded primer Spike1 (10uM), 10ul of Spike2 (10uM), 10ul of Spike (10uM), 10ul of Spike4 (10uM) and 40ul of Short-Gibson (10uM) completed with distilled water is prepared. Then the PCR amplification is performed on 25ul final volume including 12.5ul of Phusion Hot Start II High-Fidelity PCR Master Mix (ThermoFisher ref: F565), 1.25ul of pooled barcoded-SPIKE & shortGibson primers (0.05uM final concentration per SPIKE targeting primers and 0.2uM final concentration for ShortGibson) and 5ul of pooled & cleaned cDNA.

The following thermal cycling program is used:

1. 98 °C 30 s
2. 98 °C 10 s
3. 61 °C 45 s
4. 72 °C 2 min
5. Repeat 2–4 steps 39×
6. 72 °C 10 min
7. Keep at 12 °C

In both cases (E-Sarbeco Concat-PCR or SPIKE PCR amplification), PCR products were verified by TapeStation automated electrophoresis (Genomic DNA ScreenTape; Agilent ref: 5067–5365).

Multiple PCR assays are performed in presence of different barcoded primers, such that the combination with the barcoded cDNA material allows to generate a large complexity. Specifically, for E-Sarbeco targeting assay, 8 barcoded contact-PCR assays are pooled ($8 \times 25 = 200$ ul) and cleaned with SPRIselect reagent (Beckman; B23318) with a ratio of 0.5×. Similarly, for SPIKE and E-Sarbeco PCR assays issued from random hexamer cDNA material, 4 barcoded PCR assays for each target regions (SPIKE: BC1–4; E-Sarbeco: BC5–8) were pooled ($4 \times 2 \times 25 = 200$ ul) and cleaned with SPRIselect reagent (Beckman; B23318) with a ratio of 0.5×.

Pooled & cleaned material is recovered on 50ul and used for Nanopore sequencing library preparation (Nanopore Ligation Sequencing Kit: SQK-LSK109). DNA libraries were sequenced on Nanopore FLO-MIN106D flowcells (R9) with the MinION Mk1C instrument (72 h sequencing; high-accuracy base calling).

Real-time SARS-CoV-2 diagnostics and variants tracking. Real-time processing of the fastq files generated by the Mk1C instrument during sequencing is performed with our in-house developed RETIVAD (Real-Time Variants Detector) software. RETIVAD collects each ten minutes the fastq files generated during sequencing from the Mk1C instrument and (1) search for Gibson sequences within the sequenced reads (regex query (<https://pypi.org/project/regex>) with maximum 4 mismatches accepted by default); (2) split sequenced reads on monomers—delimited by the presence of the identified Gibson sequences; (3) search for barcode sequences introduced during the reverse transcription (cDNA barcodes: BC33-BC48; regex query); (4) search for barcodes introduced during the PCR amplification (PCR barcodes: BC1-BC8; regex query); (5) collect the sequences between the cDNA and PCR-associated barcodes for their alignment towards the SARS-CoV-2 reference genome (Wuhan-Hu-1 genome: NC_045512; BWA aligner following nanopore parameters). Considering that steps (3–5) are performed within the retrieved monomers in step (2), RETIVAD matches cDNA/PCR barcode combinations to aligned outcomes, thus providing a diagnostic outcome for the associated pseudocandidate. Due to the low number of sequences retrieved at intervals of 10 min, and the relative short size of the SARS-CoV-2 genome (~29 kb), RETIVAD is able to provide such diagnostic outcome within such 10 min interval. Hence, at the next round of data collection, RETIVAD reiterates on the aforementioned steps over the newly collected fastq files and appends the outcome to the previous results. The continuous gain on aligned read-counts per region of interest (e.g. E-Sarbeco amplicon sequence, or SPIKE gene; defined as entry parameters) is visualized within scatterplot view (png file) updated during the processing (RETIVAD produces a png file per detected PCR barcode monomers; thus allowing to generate scatterplots displaying as many lines as detected cDNA barcode monomers; Figure S4).

In addition to the diagnostic outcome, RETIVAD has been designed for performing real-time variants detection. For it, the racon package (version 1.4.20) is used for raw de novo DNA assembly, followed by the alignment with Minimap2 (version 2.17). Medaka (version 1.2.5) is used to create a consensus sequence and variant calls from the aligned data (> 20 aligned reads required as default for variants calling). RETIVAD generates BAM files per cDNA/PCR barcode combination harboring base conversion information and their associated coverage (for visualization with Genome browsers like IGV). Furthermore, summary files per cDNA/PCR barcode combination are also generated, providing the genomic position in which the base conversion has been detected, the identity of the converted bases as well as their associated coverage. Like in the case of the diagnostics processing, variants detection is performed in real time, thus providing the possibility to the user to visualize the results and decide whether sequencing requires still to be performed or whether the collected information is sufficient to conclude the diagnostics/variants tracing for the corresponding candidates.

RETIVAD generates a final summary report for the diagnostics outcome when no new fastq files are available after 40 min from the last data collection and processing.

Data availability

All genome sequences included in this study are available under request. RETIVAD is freely available at <https://github.com/SysFate/retivad>.

Received: 18 May 2021; Accepted: 21 July 2021

Published online: 05 August 2021

References

1. Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A. & Kozlakidis, Z. Considerations for diagnostic COVID-19 tests. *Nat. Rev. Microbiol.* **19**, 171–183 (2021).
2. La Marca, A. *et al.* Testing for SARS-CoV-2 (COVID-19): A systematic review and clinical guide to molecular and serological in-vitro diagnostic assays. *Reprod. Biomed. Online* **41**, 483–499 (2020).
3. Palmieri, D., Siddiqui, J., Gardner, A., Fishel, R. & Miles, W. O. REMBRANDT: A high-throughput barcoded sequencing approach for COVID-19 screening. *bioRxiv* <https://doi.org/10.1101/2020.05.16.099747> (2020).
4. Schmid-Burgk, J. L. *et al.* LAMP-Seq: Population-scale COVID-19 diagnostics using combinatorial barcoding. *bioRxiv* <https://doi.org/10.1101/2020.04.06.025635> (2020).
5. Bloom, J. S. *et al.* Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. *MedRxiv* <https://doi.org/10.1101/2020.08.04.20167874> (2021).
6. Burki, T. Understanding variants of SARS-CoV-2. *Lancet* **397**, 462 (2021).
7. Runtuwene, L. R., Tuda, J. S. B., Mongan, A. E. & Suzuki, Y. On-site MinION sequencing. In *Single Molecule and Single Cell Sequencing* (ed. Suzuki, Y.) 143–150 (Springer, 2019). https://doi.org/10.1007/978-981-13-6037-4_10.
8. Charre, C. *et al.* Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* **6**, 2 (2020).
9. James, P. *et al.* LamPore: Rapid, accurate and highly scalable molecular screening for SARS-CoV-2 infection, based on nanopore sequencing. *MedRxiv* <https://doi.org/10.1101/2020.08.07.20161737> (2020).
10. Nakamichi, K. *et al.* Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. *Sci. Rep.* **11**, 4802 (2021).
11. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surv.* **25**, 2 (2020).
12. Gámbaro, F. *et al.* Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. *Euro Surv.* **25**, 2 (2020).
13. Fass, E. *et al.* HiSpike: A high-throughput cost effective sequencing method for the SARS-CoV-2 spike gene. *medRxiv* <https://doi.org/10.1101/2021.03.02.21252290> (2021).
14. Sanyaolu, A. *et al.* The emerging SARS-CoV-2 variants of concern. *Ther. Adv. Infect. Dis.* **8**, 20499361211024372 (2021).

Acknowledgements

We gratefully acknowledge Corinne Cruaud and her team at Genoscope, for the technical training on Oxford Nanopore sequencing. We also thank Veronique de Berardinis and Patrick Wincker for the logistics support required for this project.

Author contributions

Conceptualization: M.A.M.P. Methodology: J.L.P. and M.A.M.P. Software development: F.S. and S.E. Scientific evaluation: M.A.M.P. Writing, review, and editing: J.L.P., F.S., and M.A.M.P. Funding acquisition: M.A.M.P.

Funding

This work was supported by DIM ELICIT's Grants from Région Ile-de-France and by the institutional bodies, the CEA, CNRS, and Université d'Evry-Val d'Essonne. François Stüder was supported by the funding 2019-L22 from the "Institut National du Cancer (INCa).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-95563-w>.

Correspondence and requests for materials should be addressed to M.A.M.-P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021