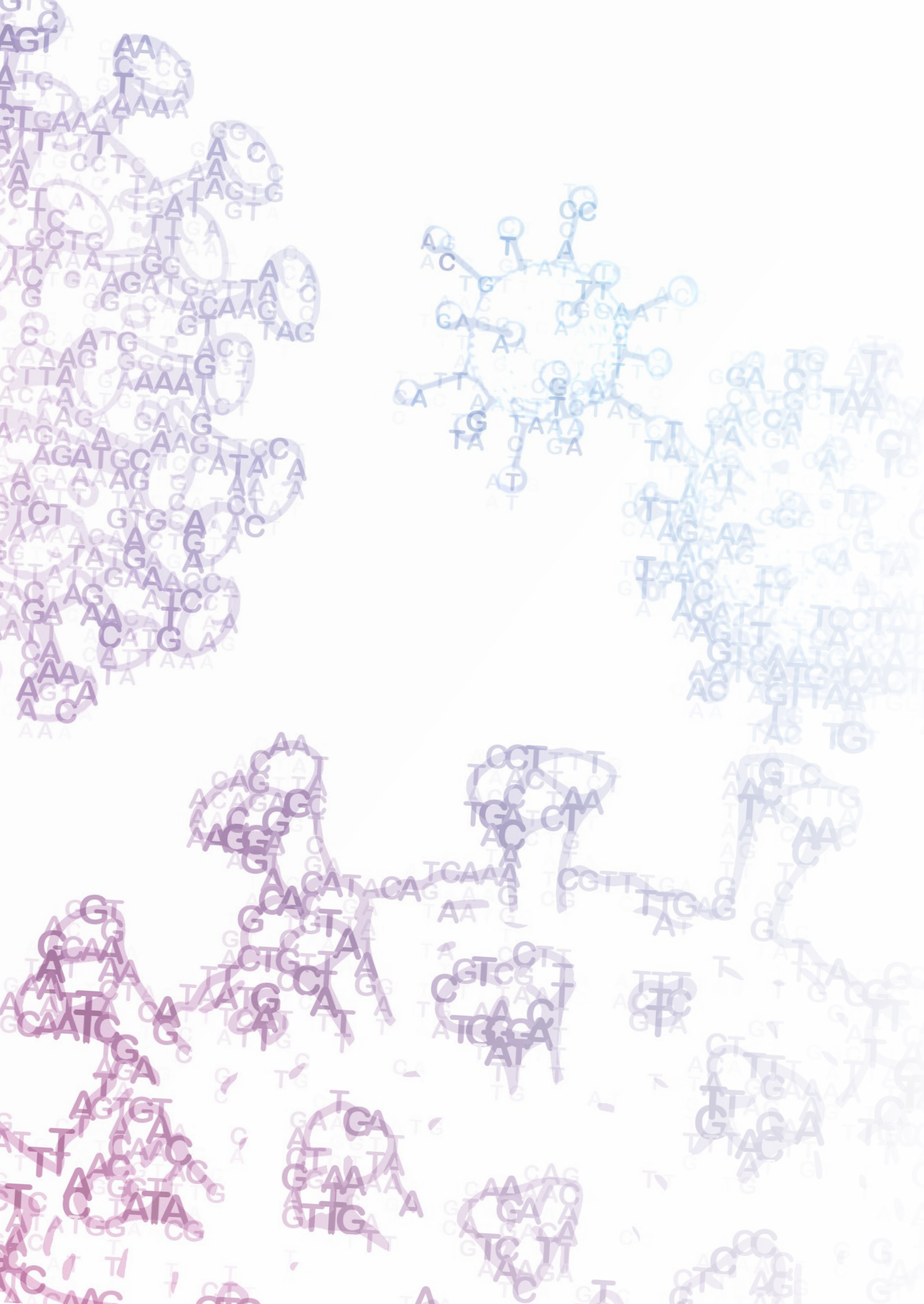
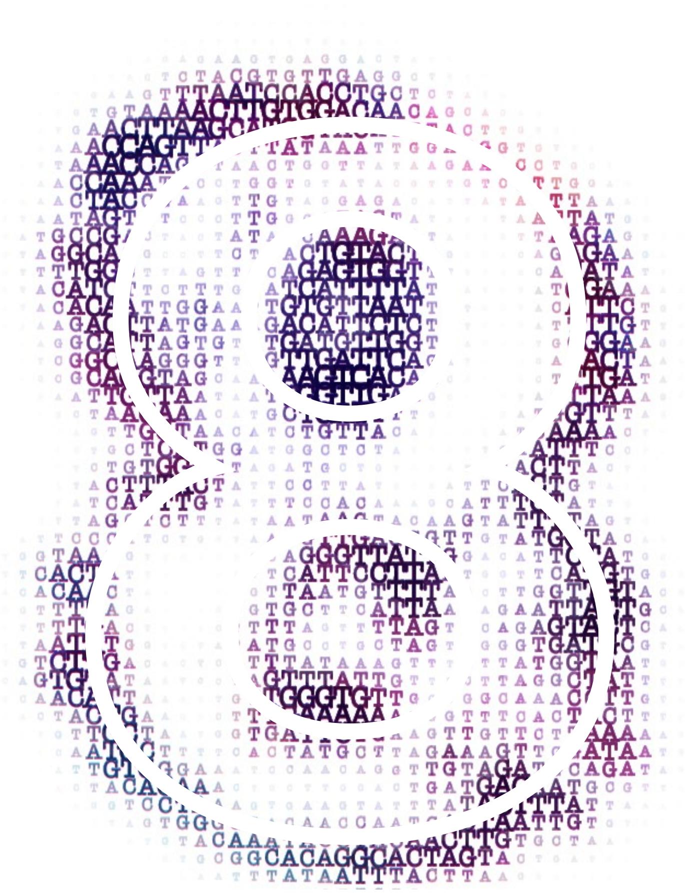
Metagenomic sequencing in clinical virology: advances in pathogen detection and future prospects

Carbo, E.C.





Chapter 8 A comparison of five Illumina, Ion torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2

Ellen C. Carbo [1], Kees Mourik [1], Stefan A. Boers [1], Bas Oude Munnink [2], David Nieuwenhuijse [2], Marcel Jonges [3], Matthijs R.A. Welkers [3], Sebastien Matamoros [3], Joost van Harinxma thoe Slooten [1], Margriet Kraakman [1], Evita Karelioti [4], David van der Meer [4], Karin Ellen Veldkamp [1], Aloys C.M. Kroes [1], Igor Sidorov [1], Jutte J.C. de Vries [1]

1 Clinical Microbiological Laboratory, Department of Medical Microbiology, Leiden University Medical Center; Leiden, the Netherlands
2 Department of Viroscience, Erasmus Medical Centre, Rotterdam, the Netherlands
3 Department of Medical Microbiology and Infection Prevention, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands
4 GenomeScan B.V., Leiden, the Netherlands

Adjusted version of article published in European Journal of Clinical Microbiology & Infectious Diseases

Abstract

Rapid identification of the rise and spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants of concern currently remains critical for monitoring of the efficacy of diagnostics, therapeutics, vaccines, and control strategies. A wide range of SARS-CoV-2 next-generation sequencing (NGS) methods have been developed over the last years, but cross-sequence technology benchmarking studies are scarce. In the current study, 26 clinical samples were sequenced using five protocols: AmpliSeq SARS-CoV-2 (Illumina), EasySeq RC-PCR SARS-CoV-2 (Illumina/NimaGen), Ion AmpliSeq SARS-CoV-2 (Thermo Fisher), custom primer sets (Oxford Nanopore), and capture probe-based viral metagenomics (Roche/Illumina). Studied parameters included genome coverage, depth of coverage, amplicon distribution, and variant calling.

The median SARS-CoV-2 genome coverage of samples with cycle threshold (Ct) values of 30 and lower ranged from 81.6 to 99.8 for, respectively, the Oxford Nanopore protocol and Illumina Ampliseq protocol. Correlation of coverage with PCR Ct-values varied and was dependent on the protocol. Amplicon distribution signatures differed across the methods, with peak differences of up to 4 log10 at disbalanced positions in samples with high viral loads (Ct-values ≤ 23). Phylogenetic analyses of consensus sequences showed clustering independent of the workflow used. The proportion of SARS-CoV-2 reads in relation to background sequences, as a (cost-)efficiency metric, was highest for the EasySeq protocol. The hands-on time was lowest when using EasySeq and ONT protocols, with the latter additionally having the shortest sequence runtime.

In conclusion, the studied protocols differed on a variety of the studied metrics. This study provides data that can assist laboratories when selecting protocols for their specific setting.

Keywords:

Whole genome sequencing; SARS-CoV-2; benchmark

Introduction

Genomic surveillance of severe acute respiratory syndrome coronavirus (SARS-CoV-2) has proven critical for early detection of the rise and spread of SARS-CoV-2 variants of concern, for monitoring and developing effective diagnostic, therapeutic, and preventive strategies [1-3]. In addition, genomic surveillance assists in contact tracing, transmission tracking at population level, and public-health decision making [4]. The widespread application of genomics for pandemic surveillance is exemplified by the more than 10 million SARS-CoV-2 sequences deposited in the GISAID repository as of April 2022 [5].

A wide range of SARS-CoV-2 next-generation sequencing (NGS) technologies and protocols have been developed and adapted since the first genome sequence was generated using a metagenomic approach [6-8]. SARS-CoV-2 whole genome sequencing (WGS) protocols have been improved to increase the technical performance, including sensitivity and genome coverage, and logistical aspects have also been addressed, such as scalability and hands-on time [9-12]. Studies have been published on SARS-CoV-2 WGS with innovative protocol adaptations in order to decrease the error rate and the turn-around-time by combining PCR and tagging steps [12]. However, these studies are typically focused on the technology developed by the authors, whereas comparison of a novel protocol with other methods is limited. Benchmark studies of SARS-CoV-2 genome sequencing technologies are scarce and generally restricted to comparison of protocols for the single type of sequencing technology available at the study site of the authors [13-15]. In contrast, cross-platform studies are still relatively scarce [16,17]. A recent external quality assessment (EQA) report assessed the outcome of complete workflows from nucleic acid extraction to the reported consensus sequence by testing SARS-CoV-2 cultured isolates; however, no detailed distinction between the different workflow components could be made [16].

Here, we describe a cross-platform benchmark study that includes Illumina, Ion torrent, and nanopore-based SARS-CoV-2 sequencing technologies in one study. Five protocols (Figure 1), employing a diversity of sequencers with a wide range of throughput, accuracy and runtime were compared using clinical samples. The performance was studied by comparing genome coverage, read depth, amplicon distribution, variant calling, and the proportion of on-target reads.

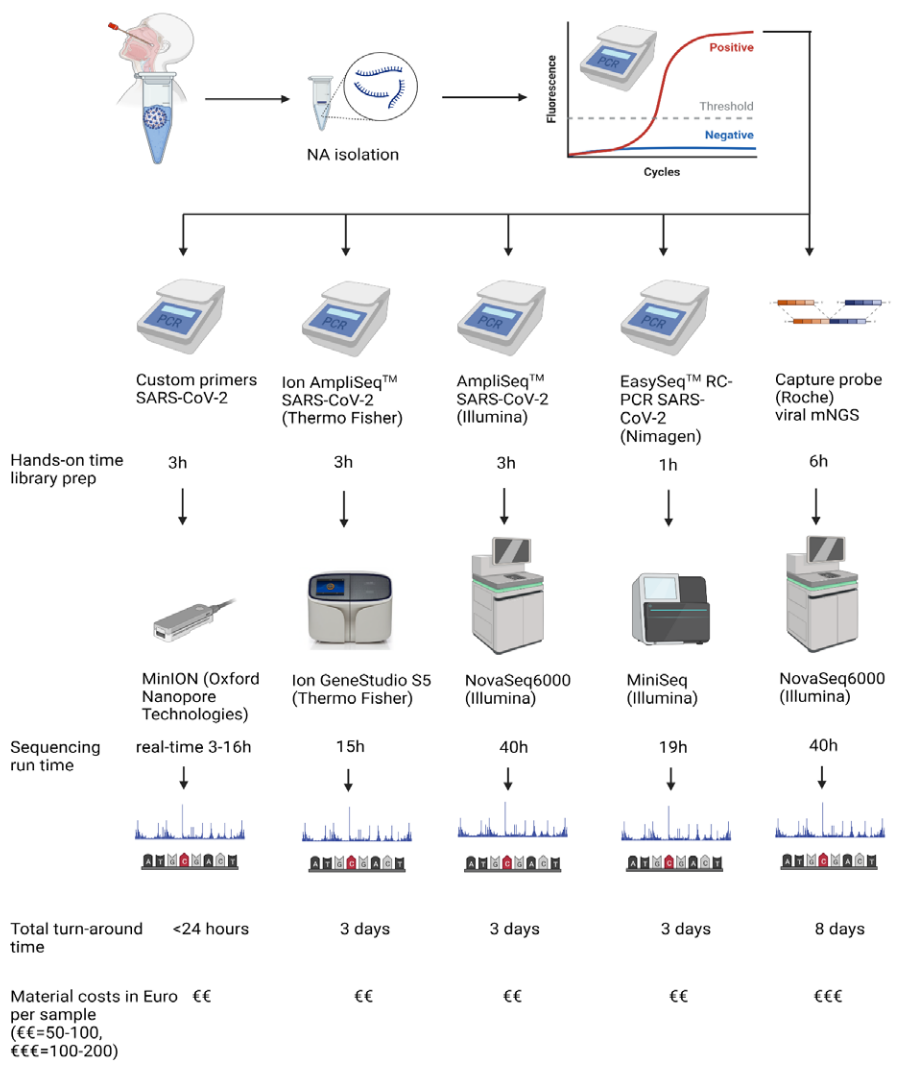

NA isolation
Fluorescence
Positive
Threshold
Negative
Cycles
PCR
Custom primers SARS-CoV-2
Ion AmpliSeq™ SARS-CoV-2 (Thermo Fisher)
AmpliSeq™ SARS-CoV-2 (Illumina)
EasySeq™ RC-PCR SARS-CoV-2 (Nimagen)
Capture probe (Roche) viral mNGS
Hands-on time library prep
3h
3h
3h
1h
6h
MinION (Oxford Nanopore Technologies)
Ion GeneStudio S5 (Thermo Fisher)
NovaSeq6000 (Illumina)
MiniSeq (Illumina)
NovaSeq6000 (Illumina)
Sequencing run time
real-time 3-16h
15h
40h
19h
40h
A T G C G A C T
Total turn-around time
<24 hours
3 days
3 days
3 days
8 days
Material costs in Euro per sample (€€=50-100, €€€=100-200)
€€
€€
€€
€€
€€€

Figure 1. Schematic overview of the design, workflow, and technologies adopted in this study. Twenty-six respiratory samples, mainly nasopharyngeal swabs and tracheal aspirates, were tested by five SARS-CoV-2 WGS protocols. PCR Ct-values ranged from 13.9-33.6. To exclude potential variability resulting from different nucleic acid extraction methodologies, the extraction method used was identical for all five protocols. Four protocols were tiled amplicon based, one protocol was capture probe based, targeting all viruses known to infect vertebrates. In order to minimize potential differences resulting from variation in bioinformatic analyses tools and settings, a uniform pipeline for sequence data from Illumina and Ion platforms, for ONT data, platform-specific tools handling higher error rates were used to gain optimal results from this type of dataset (Suppl. Figure 1). Created using Biorender.com.

Methods

Sample selection

In total, 26 SARS-CoV-2 PCR positive samples of 24 patients were selected: nine tracheal aspirates, 16 nasopharyngeal/throat swabs, and one lung lymph node biopsy. Fifteen of these samples were obtained for cluster identification. Samples were retrospectively included to be tested with five WGS protocols. Samples were previously sent to the Clinical Microbiological Laboratory of the Leiden University Medical Center (LUMC, the Netherlands) for SARS-CoV-2 PCR testing in the period March - October 2020 (Wuhan-like viruses circulating). As previously described [18], and stored at -80 °C until WGS analysis. In total 26 samples with a wide range of Ct-values (13.9-33.6, confirmed by re-testing) were included to assess the performance of each of the five WGS protocols. The range and distribution of PCR Ct-values was chosen based on relevance for routine clinical practice.

Ethical approval

Approval was obtained from the ethical committee of the LUMC (B20.002, Biobank Infectious Diseases 2020-03), and the Institutional Review Board of the LUMC for observational Covid-19 studies (CoCo 2021-006).

Extraction of nucleic acids

To exclude potential variability resulting from different nucleic acid extraction methodologies, the extraction method used was identical for all five protocols. Nucleic acids were extracted from 200 ul plasma using the MagNApure96 DNA and Viral NA small volume extraction kit on the MagNAPure 96 System (Roche Diagnostics, Almere, the Netherlands) with 100 ul output eluate.

SARS-CoV-2 sequencing protocols (see also Figure 1)

Ampliseq SARS-CoV-2 sequencing (Illumina)

Libraries were prepared using the AmpliSeqTM SARS-CoV-2 Research Panel for Illumina®, which is a targeted RNA/cDNA amplicon assay for epidemiological research of the SARS-CoV-2 virus. This panel contains a two pool design of 247 amplicons/primer pairs (pool 1: 125 amplicons, pool 2: 122 amplicons). In total, 237 amplicons were SARS-CoV-2 targets while the remaining amplicons mapped to five different regions of the human genome and were used as control. The amplicons' lengths ranged from 125 to 275 bp. From each sample, 15 ul of eluate was concentrated using the Speedvac vacuum concentrator (Eppendorf, Hamburg, Germany). Samples were then dissolved in 10 µl AmpliSeq cDNA synthesis master mix. Next, the AmpliSeq cDNA Synthesis for Illumina Kit (Illumina) was used to reverse transcribe RNA to cDNA. Amplicon primer pools of the AmpliSeqTM SARS-CoV-2 Research Panel for Illumina® were subsequently added to each sample. cDNA target amplification reaction was performed according to manufacturer's instructions, followed by partial digestion of primer dimers. AmpliSeq CD indexes were then ligated and further library PCR amplification was performed. The libraries were purified with the AgencourtTM AMPureTM XP Reagent (Beckman Coulter). The final quality and quantity of each barcoded cDNA library was determined using the Fragment Analyzer (Agilent). From all amplified libraries, 2 µl was pooled and loaded for a short sequencing run to indicate the size of the intact libraries. Based on the indicative read counts, equimolar amounts of each sample were pooled (1.1 nM) and submitted for DNA sequencing using the NovaSeq6000 system (Illumina, San Diego, CA, USA) according to manufacturer's protocols. Approximately 10 million 150 bp paired-end reads were obtained per sample. Data processing was performed in real-time by the NovaSeq Control Software v1.7.

EasySeq RC-PCR SARS-CoV-2 sequencing (NimaGen/Illumina)

Libraries were prepared using the EasySeq RC-PCR SARS-CoV-2 kit version 4.02 (NimaGen) for Illumina as described by Coolen et al [12]. cDNA synthesis was performed using the iScriptTM Advanced cDNA Synthesis Kit (Bio-Rad) according to manufacturer's instructions using 10 ul of eluate. This version of the EasySeq RC-PCR SARS-CoV-2 kit uses 154 designed primer pairs (pool A and B) with a tiling strategy, resulting in approximately 435 bp size amplicons. The EasySeq protocol enables a one-step procedure for adding SARS-CoV-2 target specific PCR primers, sequence adapters and Unique Dual Indices (UDI's) by hybridization of the SARS-CoV-2 primers with universal primers that include adapters and UDI's. After the PCR with 5 µl cDNA as input, samples were pooled based on Ct value into pool A and B, which were individually cleaned using AmpliCleanTM Magnetic Bead PCR Clean-up Kit (NimaGen, Nijmegen, The Netherlands). Subsequently, quantification was performed using the Qubit double strand DNA (dsDNA) High Sensitivity assay kit on a Qubit 4.0 instrument (Life Technologies) and pool A and B were combined. Sequencing was performed on Illumina MiniSeq® using a Mid Output Kit (2 × 149 or 2 × 151-cycles) (Illumina, San Diego, CA, USA) by loading 0.8 pM on the flowcell, obtaining approximately 50,000 paired-end reads per sample. The sequence runs were conducted using a balanced library pooling strategy based on estimated cDNA input according to the manufacturer's protocol.

Ion AmpliSeq SARS-CoV-2 sequencing (Thermo Fisher)

The Ion AmpliSeq SARS-CoV-2 research panel supplied by Thermo Fisher Scientific contained 247 primer pairs designed to cover the SARS-CoV-2 genome with 125 to 275 bp overlapping amplicons. For cDNA synthesis, the SuperScipt VILO cDNA Synthesis Kit (11754050, ThermoFisher Scientific, The Netherlands) was used according to manufacturer's instructions using 7 µl of diluted nucleic acid solution to an estimated input of 100 copies/reaction using nuclease free water (AM9939, Ambion, Thermo Fisher Scientific, The Netherlands). SARS-CoV-2 whole genome amplification, adapter ligation and purification were performed using the Ion AmpliSeq SARS-CoV-2 Insight Research Assay (A51305, Thermo Fisher Scientific, The Netherlands) according to manufacturer's instruction. Libraries were quantified using the Ion Library TaqMan Quantitation Kit (4468802, Thermo Fisher Scientific, The Netherlands) according to manufacturer's instructions. Samples were then sequenced on an Ion GeneStudio S5 system (ThermoFisher Scientific, The Netherlands) using an Ion 540 chip (ThermoFisher Scientific, The Netherlands), obtaining approximately up to 1 million paired-end reads per sample.

Custom primers with MinION sequencing (ONT)

A SARS-CoV-2 specific multiplexed PCR for nanopore sequencing was performed using custom-made primers as previously described [4]. In short, primers for 89 overlapping amplicons spanning the whole SARS-CoV-2 genome were designed using primal [19]. The amplicon length was approximately 500 bp with a 75 bp overlap between the different amplicons. cDNA was transcribed using SuperScript III Reverse Transcriptase (Invitrogen, Darmstadt, Germany) [20]. Libraries were generated using the native barcode kits from Oxford Nanopore Technologies (EXP-NBD104, EXP-NBD114 and SQK-LSK109) using 5µl cDNA as input, and sequenced on a R9.4 flow cell multiplexing 96 samples per sequence run (Oude Munnink et al). On average, 68k reads with an average size of 423 bp were obtained per sample.

Capture probe (Roche) with viral metagenomic NGS (Illumina)

The viral metagenomic NGS protocol has previously been described [21-23]. After nucleic acid extraction, 50 µL of eluate was concentrated with the SpeedVac vacuum concentrator (Eppendorf, Hamburg, Germany) and dissolved in 10 µl fragmentation master mix (NEBNext). The NEBNext Ultra II Directional RNA Library prep kit (New England Biolabs, Ipswich, MA, USA) for Illumina was used for RNA library preparation, incorporating several alterations to the manufacturer's protocol to be able to detect both DNA and RNA in the sample. Specifically, poly-A mRNA capture isolation, rRNA depletion and DNase treatment steps were omitted and dual indexed adaptors were used. The SeqCap EZ Hypercap probes (Roche, Basel, Switzerland) were designed in 2015 to cover 207 taxa genomes of viruses known to infect vertebrates including humans [24]. Recently, it has been shown that the probes cover >99% of the SARS-CoV-2 genome [25] due to similarity with bat coronaviruses and the variability incorporated in the probe design. Viral DNA enrichment was performed using the SeqCap EZ HyperCap Workflow User's Guide in pools of four amplified DNA libraries with overnight probe incubation. Washing and recovering captured DNA was performed using the HyperCap Target Enrichment kit and HyberCap Bead kit. Lastly, post-capture PCR amplification was performed with KAPA HiFi HotStart ReadyMix (2X) and Illumina NGS primers following manufacturers' instructions, followed by AMPure bead purification. The quality and quantity of the post-capture multiplexed libraries were assessed by Fragment Analyzer (Agilent) or Bioanalyzer (Agilent, Santa Clara, CA, USA). Sequencing was performed on the NovaSeq6000 system (Illumina, San Diego, CA, USA) obtaining approximately 10 million 150 bp paired-end reads per samples.

Data analyses

In order to minimize potential differences resulting from variation in analysis tools and settings, a uniform pipeline for QC, trimming, mapping, and variant calling was used for sequence data from Illumina and Ion platforms (Supplementary Figure 1). For ONT data, platform-specific mapping and variant calling tools handling higher error rates were used to gain optimal results from this type of dataset.

Illumina data from AmpliSeq, EasySeq and viral metagenomic protocols

Demultiplexing was performed according to Illumina manufacturer protocol using bcl2fastq v2.20 (Illumina). Removal of duplicate reads was not performed since unique molecular identifiers (UMI's) in principle are not compatible with the non-random, tiled amplicon based WGS protocols in the current study, and were thus not incorporated in any of the wet lab procedures described here. Quality control and trimmings per read was performed utilizing Trimmomattic v0.36 [26]. To remove and count the number of sequence reads mapping to the human genome, reads were mapped to GRCh38 using Bowtie2 v2.1.0 [27]. Unmapped reads were subsequently mapped to the SARS-CoV2 genome NC_045512.2 [28]. Mapped reads were indexed in a genome sorted bam file by Samtools v1.7 [29,30]. Variant calling was done using Bcftools v.1.7 [31].

Ion AmpliSeq data

Primer-removed fastq-files were exported for further analysis using the Torrent Suite Software (ThermoFisher Scientific, The Netherlands). Per read quality control was performed using Trimmomatic v0.36 [26]. The resulting quality checked reads were first mapped to the human reference genome HG19 using BWA v0.7.17 [32] with default settings ("bwa bwasw") to remove all reads of potential human origin. Unmapped reads were subsequently mapped to the SARS-CoV-2 refence genome Wuhan-Hu-1 [33]. The resulting sequence alignment map (SAM) files were converted to BAM, sorted and indexed using SAMtools v1.14 [29,30]. Variant calling was performed using Bcftools v.1.7 [31].

ONT custom primers data

Demultiplexing was performed using Porechop v0.2.4 [34]. Primers were trimmed using Cutadapt v3.0 [35]. Reference-based alignment was carried out using Minimap2 v2.17-r941 [36] against both the human genome GRCH38 and SARS-CoV-2 genome NC_045512.2 [28]. Variant calling was performed by filtering of variants using the Python module Pysam v 0.16.0.1 [37].

Performance and statistical analyses

Mapping coverage was analysed using a threshold of 10x depth per base for all platform data except for ONT data, where a 20x depth per base was considered as threshold to ensure reliable variant calling [38]. Coverages per base were calculated using Samtools v1.7 [29,30] with the corresponding depth option. Correlation between genome coverage percentage and Ct-values was calculated using Spearmans' rho [39]. Read mapping quality and base quality (phred) were computed using Samtools v.11 [29,30] with the coverage option. High mapping quality represents a more unique alignment and low mapping quality represents a marginal difference between the alignment and the best secondary alignment option within the reference. High phred scores represent accurate base calling.

Phylogenetic trees

Maximum likelihood trees of the consensus genomes from all methods was generated using the Samtools consensus option [29], Clustal Omega v1.2.4 [39], FastTree v2.1.11 [40,41], and IQTree [42]. Consensus genomes with ≥98% genome coverage were included, genome coverages based on minimal 10x read depth for all methods, and 20x read depth for ONT sequencing. Variant frequencies of >50% were implemented in the consensus genome, though error profiles, like those of ONT, and short insertions/deletions (indels) not consistently called by Samtools can lead to an inaccuracy of the consensus.

Results

In total 26 clinical samples from 24 patients were sequenced using the five SARS-CoV-2 sequencing protocols included in the current comparison: AmpliSeq SARS-CoV-2 (Illumina), EasySeq RC-PCR SARS-CoV-2 (Nimagen/Illumina), Ion AmpliSeq SARS-CoV-2 (Thermo Fisher), custom SARS-CoV-2 primers-based (Oxford Nanopore), and capture probe (Roche) viral mNGS (Figure 1). Additional protocol characteristics, such as hands-on time and sequence runtime are listed in Suppl. Table 1. The breadth of genome coverage, depth of genome coverage, proportion of SARS-CoV-2 reads, and performance of variant calling were compared.

Genome coverage

SARS-CoV-2 genome coverages were generated using a 10x read depth threshold per base for Illumina and Ion Torrent data, and 20x for ONT sequence data (Figure 2, and Suppl. Table 2, incl. normalised read depth per 100,000 total reads.) (Baker et al). As anticipated, amplicon-based protocols generally resulted in higher genome coverage rates compared to the probe hybridization-based metagenomics protocol, though median genome coverages using the custom primer ONT protocol were within the same range for samples with Ct-values of ≤30 (81.2% for ONT and 86.7% for mNGS, Suppl. Table 2). The median genome coverage across the other three amplicon-based protocols was comparable for samples with Ct-values of ≤30: respectively 99.7% and 99.8% when using the Ion AmpliSeq and the Illumina AmpliSeq protocol, followed by the EasySeq protocol for Illumina (98.05%). An increase in Ct-values resulted in only limited reduction of genome coverage when using the Ion AmpliSeq (R = -0.327) and Illumina AmpliSeq (R = -0.523) protocols. When considering all samples, including high Ct values the genome coverage differed greatly between the amplicon-based protocols.

The median read depth of coverage per position ranged from 316 when using the Illumina EasySeq protocol to 860 when using ONT, and >2000 for the Ion AmpliSeq and the probe hybridization-based metagenomics protocol. This depended on the throughput of the platform and kit, the total number of reads requested, and the number of samples multiplexed.

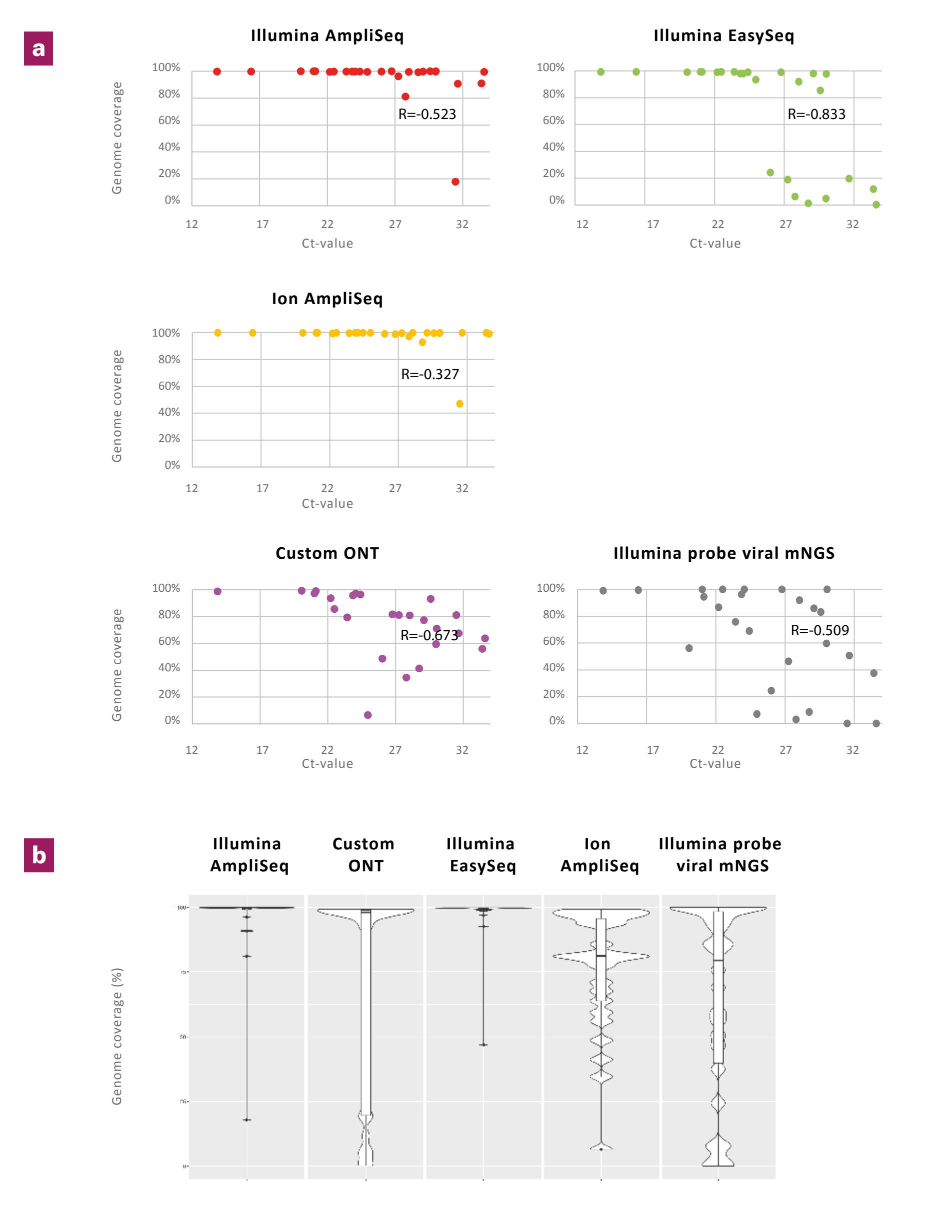


Figure 2. Proportion of SARS-CoV-2 genome coverage of sequencing reads using the five protocols compared. The scatter plots (a) indicate the SARS-CoV-2 genome (NC_045512.2) coverage per PCR Ct-values, each dot represents a single sample. A threshold of 10x depth per base was considered for all platform data except for ONT data, were a 20x depth per base was considered as threshold ensuring reliable variant calling. R values represent Spearmans' correlation coefficient (rho). The violin plots (b) indicate the distribution of the proportion covered per protocol, horizontal markers indicate the median, and the interquartile range.

SARS-CoV-2 amplicon balance

The SARS-CoV-2 amplicon balance was assessed by evaluating the distribution of sequence reads across the SARS-CoV-2 genome. The average read depth per genome position was computed for a selection of nine samples with the highest viral loads (Ct-values ranging from 13-23) (Figure 3). When comparing the genome coverage profiles across the five protocols, distinct signatures were observed for each method. The read depth was most even when using the Illumina AmpliSeq protocol, in contrast to the uneven depth obtained using the probe hybridization-based protocol. The difference in depth between depth of coverage peaks and dips varied generally 2 log10-fold when using the Illumina AmpliSeq protocol, up to 4 log10-fold for the probe-based viral metagenomics protocol. When examining the differences in read depths in more detail, certain positions had protocol dependent, structural lower read depth for multiple samples. An example of a protocol with a structural drop of depth (to 0-11X read depth per sample) was observed at genome position 4,117- 4,149 (ORF1a) when using the Illumina AmpliSeq and Ion Ampliseq protocols. These findings were indicative of a primer failure caused by a specific SNV. The custom ONT protocol resulted in several samples with a low read depth in the amplicons spanning the regions 2,690-2,715 and 6,260-6,490 (ORF1a). Hybridisation probe viral mNGS resulted in the largest regions with low coverage, especially regions 1,000-10,000 (ORF1a) and 22,250-23,000 (Spike), with the last one at risk for missing mutations in the spike protein.

Variant calling and phylogenetic analysis

To assess the performance of variant calling across the protocols, consensus sequences were aligned to the SARS-CoV-2 reference NC_045512.2; SNVs detected per protocol are depicted in Suppl. Table 3. Consensus sequences used to build a phylogenetic tree for samples in which ≥4 protocols had a genome coverage of 98% and higher (n=14 samples). In the phylogenetic tree where gaps in the sequence (uncovered positions and indels) were considered a match with the reference sequence (Figure 4a), consensus genomes of specific samples clustered independent of the used protocol and analysis pipeline. However, when gaps were simply masked in the pairwise comparison (affecting solely the denominator, the total number of positions counted), for highly identical sequences (lower part of the tree) some per protocol clustering was also observed across Illumina, Ion, ONT and probe-based technologies, up to 0.005 substitutions/site distances between methods (Figure 4b). These findings indicate the effect of gaps in sequences in relation to the type of cluster analyses in case of highly identical sequences.

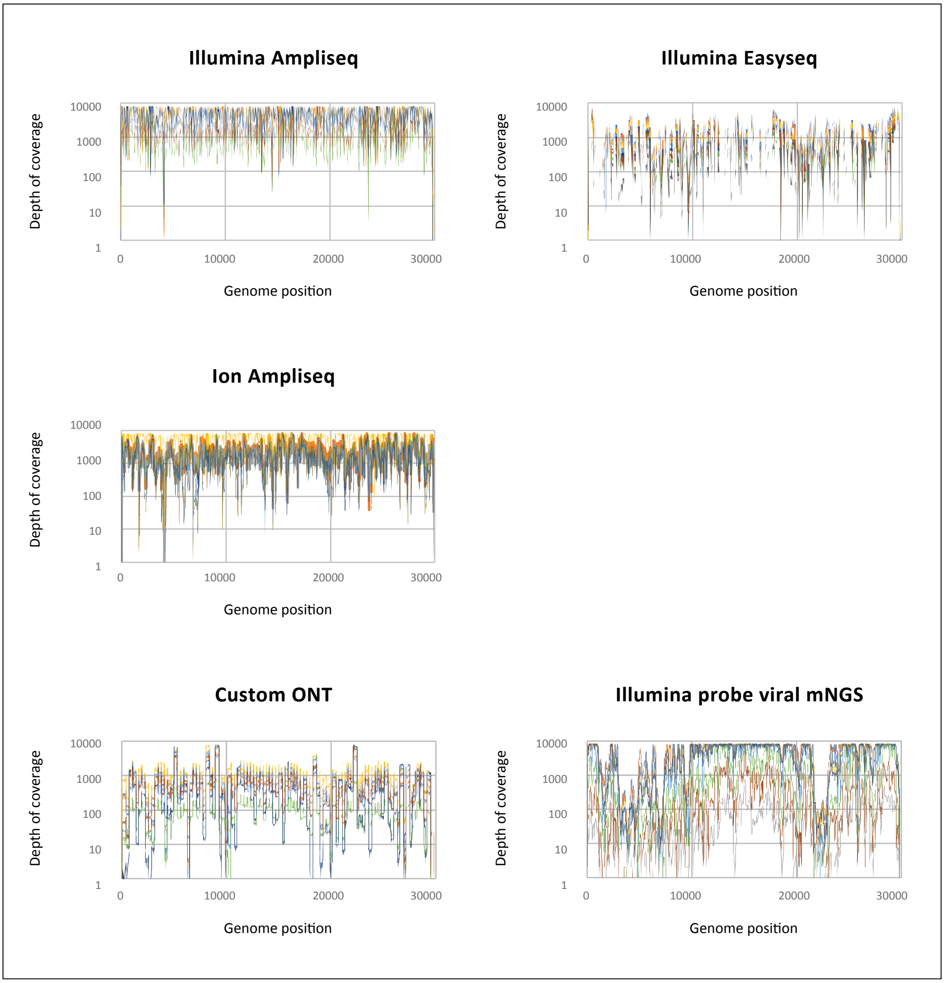


Figure 3. Distribution of sequence read depth over the SARS-CoV-2 genome using the five protocols compared.

The number of sequence reads (logarithmic scale) per SARS-CoV-2 genome (NC_045512.2) position, using the five protocols compared. A selection of nine samples with higher viral loads (Ct-values ranging from 13-23) is visualized. Each color represents an individual sample.

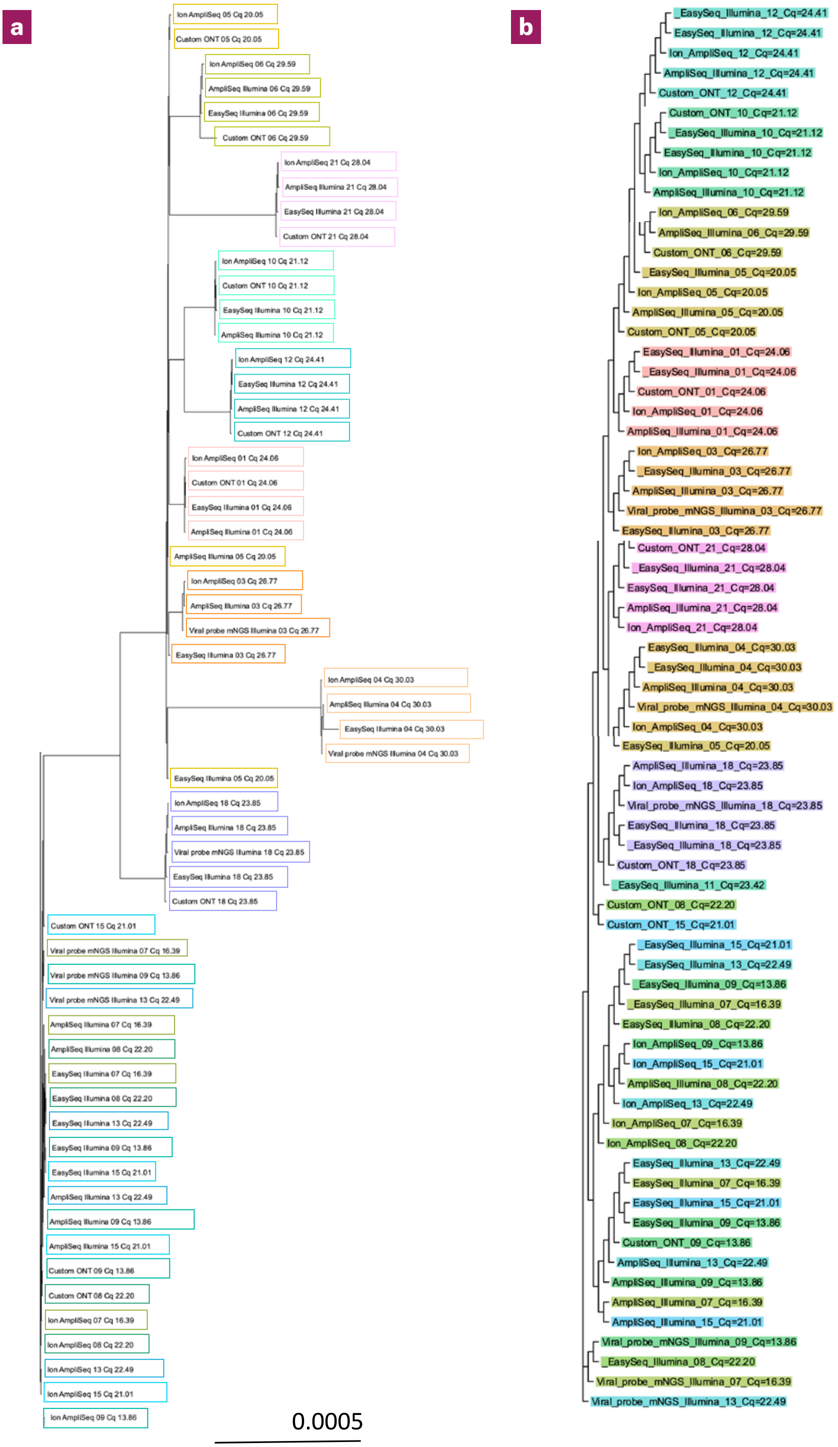

a
Ion AmpliSeq 05 Cq 20.05
Custom ONT 05 Cq 20.05
Ion AmpliSeq 06 Cq 29.59
AmpliSeq Illumina 06 Cq 29.59
EasySeq Illumina 06 Cq 29.59
Custom ONT 06 Cq 29.59
Ion AmpliSeq 21 Cq 28.04
AmpliSeq Illumina 21 Cq 28.04
EasySeq Illumina 21 Cq 28.04
Custom ONT 21 Cq 28.04
Ion AmpliSeq 10 Cq 21.12
Custom ONT 10 Cq 21.12
EasySeq Illumina 10 Cq 21.12
AmpliSeq Illumina 10 Cq 21.12
Ion AmpliSeq 12 Cq 24.41
EasySeq Illumina 12 Cq 24.41
AmpliSeq Illumina 12 Cq 24.41
Custom ONT 12 Cq 24.41
Ion AmpliSeq 01 Cq 24.06
Custom ONT 01 Cq 24.06
EasySeq Illumina 01 Cq 24.06
AmpliSeq Illumina 01 Cq 24.06
AmpliSeq Illumina 05 Cq 20.05
Ion AmpliSeq 03 Cq 26.77
AmpliSeq Illumina 03 Cq 26.77
Viral probe mNGS Illumina 03 Cq 26.77
EasySeq Illumina 03 Cq 26.77
Ion AmpliSeq 04 Cq 30.03
AmpliSeq Illumina 04 Cq 30.03
EasySeq Illumina 04 Cq 30.03
Viral probe mNGS Illumina 04 Cq 30.03
EasySeq Illumina 05 Cq 20.05
Ion AmpliSeq 18 Cq 23.85
AmpliSeq Illumina 18 Cq 23.85
Viral probe mNGS Illumina 18 Cq 23.85
EasySeq Illumina 18 Cq 23.85
Custom ONT 18 Cq 23.85
Custom ONT 15 Cq 21.01
Viral probe mNGS Illumina 07 Cq 16.39
Viral probe mNGS Illumina 09 Cq 13.86
Viral probe mNGS Illumina 13 Cq 22.49
AmpliSeq Illumina 07 Cq 16.39
AmpliSeq Illumina 08 Cq 22.20
EasySeq Illumina 07 Cq 16.39
EasySeq Illumina 08 Cq 22.20
EasySeq Illumina 13 Cq 22.49
EasySeq Illumina 09 Cq 13.86
EasySeq Illumina 15 Cq 21.01
AmpliSeq Illumina 13 Cq 22.49
AmpliSeq Illumina 09 Cq 13.86
AmpliSeq Illumina 15 Cq 21.01
Custom ONT 09 Cq 13.86
Custom ONT 08 Cq 22.20
Ion AmpliSeq 07 Cq 16.39
Ion AmpliSeq 08 Cq 22.20
Ion AmpliSeq 13 Cq 22.49
Ion AmpliSeq 15 Cq 21.01
Ion AmpliSeq 09 Cq 13.86
0.0005
b
_EasySeq_Illumina_12_Cq=24.41
EasySeq_Illumina_12_Cq=24.41
Ion_AmpliSeq_12_Cq=24.41
AmpliSeq_Illumina_12_Cq=24.41
Custom_ONT_12_Cq=24.41
Custom_ONT_10_Cq=21.12
_EasySeq_Illumina_10_Cq=21.12
EasySeq_Illumina_10_Cq=21.12
Ion_AmpliSeq_10_Cq=21.12
AmpliSeq_Illumina_10_Cq=21.12
Ion_AmpliSeq_06_Cq=29.59
AmpliSeq_Illumina_06_Cq=29.59
Custom_ONT_06_Cq=29.59
_EasySeq_Illumina_05_Cq=20.05
Ion_AmpliSeq_05_Cq=20.05
AmpliSeq_Illumina_05_Cq=20.05
Custom_ONT_05_Cq=20.05
EasySeq_Illumina_01_Cq=24.06
_EasySeq_Illumina_01_Cq=24.06
Custom_ONT_01_Cq=24.06
Ion_AmpliSeq_01_Cq=24.06
AmpliSeq_Illumina_01_Cq=24.06
Ion_AmpliSeq_03_Cq=26.77
_EasySeq_Illumina_03_Cq=26.77
AmpliSeq_Illumina_03_Cq=26.77
Viral_probe_mNGS_Illumina_03_Cq=26.77
EasySeq_Illumina_03_Cq=26.77
Custom_ONT_21_Cq=28.04
_EasySeq_Illumina_21_Cq=28.04
EasySeq_Illumina_21_Cq=28.04
AmpliSeq_Illumina_21_Cq=28.04
Ion_AmpliSeq_21_Cq=28.04
EasySeq_Illumina_04_Cq=30.03
_EasySeq_Illumina_04_Cq=30.03
AmpliSeq_Illumina_04_Cq=30.03
Viral_probe_mNGS_Illumina_04_Cq=30.03
Ion_AmpliSeq_04_Cq=30.03
EasySeq_Illumina_05_Cq=20.05
AmpliSeq_Illumina_18_Cq=23.85
Ion_AmpliSeq_18_Cq=23.85
Viral_probe_mNGS_Illumina_18_Cq=23.85
EasySeq_Illumina_18_Cq=23.85
_EasySeq_Illumina_18_Cq=23.85
Custom_ONT_18_Cq=23.85
_EasySeq_Illumina_11_Cq=23.42
Custom_ONT_08_Cq=22.20
Custom_ONT_15_Cq=21.01
_EasySeq_Illumina_15_Cq=21.01
_EasySeq_Illumina_13_Cq=22.49
_EasySeq_Illumina_09_Cq=13.86
_EasySeq_Illumina_07_Cq=16.39
EasySeq_Illumina_08_Cq=22.20
Ion_AmpliSeq_09_Cq=13.86
Ion_AmpliSeq_15_Cq=21.01
AmpliSeq_Illumina_08_Cq=22.20
Ion_AmpliSeq_13_Cq=22.49
Ion_AmpliSeq_07_Cq=16.39
Ion_AmpliSeq_08_Cq=22.20
EasySeq_Illumina_13_Cq=22.49
EasySeq_Illumina_07_Cq=16.39
EasySeq_Illumina_15_Cq=21.01
EasySeq_Illumina_09_Cq=13.86
Custom_ONT_09_Cq=13.86
AmpliSeq_Illumina_13_Cq=22.49
AmpliSeq_Illumina_09_Cq=13.86
AmpliSeq_Illumina_07_Cq=16.39
AmpliSeq_Illumina_15_Cq=21.01
Viral_probe_mNGS_Illumina_09_Cq=13.86
_EasySeq_Illumina_08_Cq=22.20
Viral_probe_mNGS_Illumina_07_Cq=16.39
Viral_probe_mNGS_Illumina_13_Cq=22.49
0.005

Figure 4. Tree of likelihood ratios based on consensus sequences of samples with genome coverages of ≥98% for each of the protocols.

Phylogenetic trees were build base on consensus sequences resulting from each of the protocols (FastTree [41,41] and IQTree [42]). For readability, a magnification is shown that includes samples with ≥98% genome coverage for four or more of the protocols (14 samples). A threshold of 10x depth per base was considered for all platform data except for ONT data, were a 20x depth per base was considered. Each color represents an individual sample. Clustering was independent of the protocol (a) IQTree, gtr [42], (b), however when gaps in the sequences (deletions and uncovered positions) were masked instead of considered as matches, in cases of closely related sequences (lower part of the tree) also clustering per protocol was detected.

SARS-CoV-2 sequencing efficiency: proportion of SARS-CoV-2 reads

To assess the efficiency of the protocols for sequencing SARS-CoV-2 genome in relation to background sequences, the proportion of SARS-CoV-2 read counts per sample, as opposed to human and other (bacterial) read counts, were computed (Figure 5). As anticipated, the proportion of SARS-CoV-2 sequences was higher for amplicon-based protocols in comparison to the hybrid capture-based protocol, but differed considerably among the last. The proportion of SARS-CoV-2 specific reads varied from 73.72% on average when using the Illumina EasySeq protocol, down to 8.19% on average when using the Illumina probe viral mNGS protocol. Mapping percentages of human reads ranged from 0.03%-99.87% for Illumina and Ion torrent amplicon-based protocols up to 69.98% on average for the Illumina probe viral mNGS protocol, with the long read ONT workflow resulting in the lowest number of human reads. Samples with an inefficient amplification, resulting in a low percentage of SARS-CoV-2 reads, showed a reverse pattern in the percentage of human reads (Figure 5). As can be deduced from these findings combined with Figure 2, some protocols with lower SARS-CoV-2 sequence efficiency compensated for these results by deeper sequencing.

Quality performance

To assess the mapping quality scores, representing the probability that a read is misaligned, median mapping quality scores were assessed (Suppl. Table 2). The mapping quality for all protocols was higher than 40, which equals a mapping accuracy of 99.99%. The median base quality (Phred) scores reflecting the estimates of errors emitted by the sequencing machines ranged from Q23.8 (ONT, P_{error} 0.004%) and Q26.6 (Ion, P_{error} 0.002%) to Q36 for Illumina protocols (P_{error} 0.0003%).

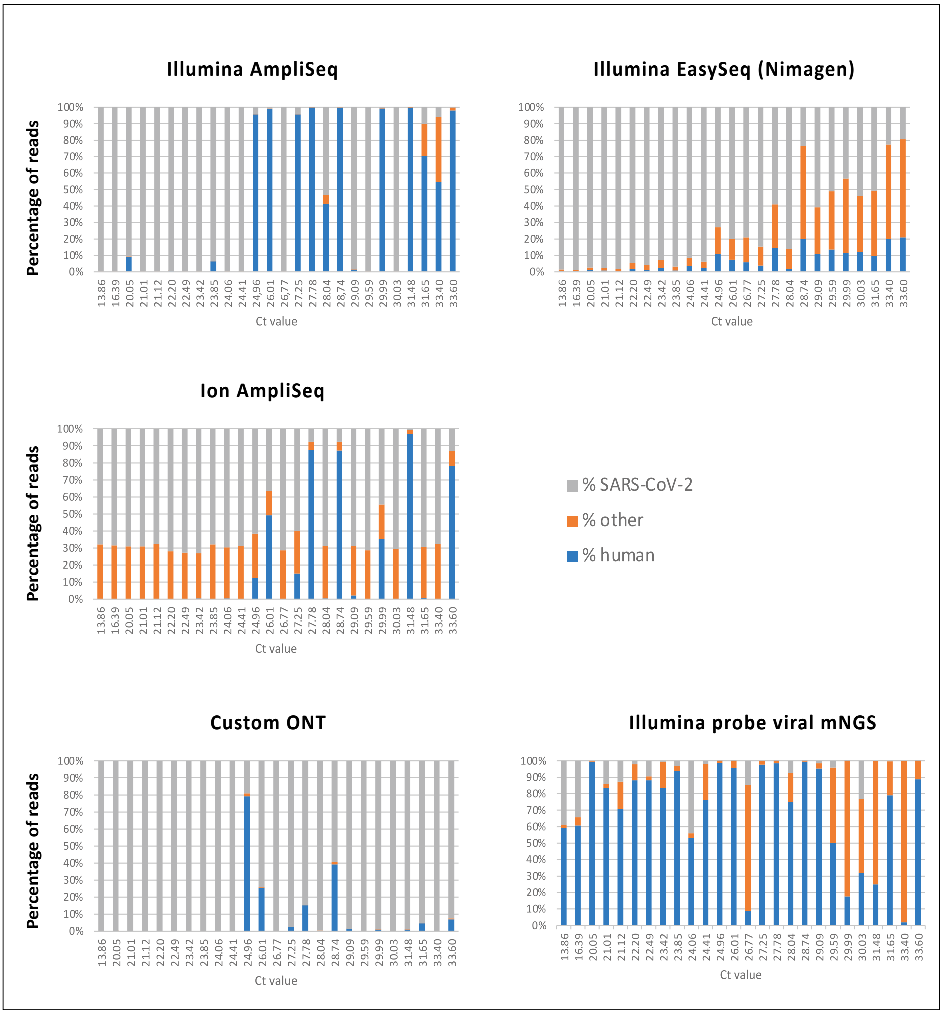


Figure 5. Proportion of SARS-CoV-2 read counts, compared to human and other (bacterial) read counts.

The proportion of SARS-CoV-2, human, and other read counts is shown for each of the five protocols. Each bar (PCR-Ct value) represents an individual sample.

Discussion

In this cross-platform benchmarking using clinical samples, the protocols differed with regard to the varying metrics studied. Each protocol had their own characteristics, advantages and disadvantages. When considering genome coverage, the Illumina and Ion Torrent amplicon-based protocols were in favor. However, amplicon balance was not always even and showed protocol specific drops. Protocols with uneven distribution of sequencing depth among amplicons may benefit from primer redesign or rebalancing of the primer pool to obtain a more even coverage threshold in difficult regions of the genome [37]. Phylogenetic analysis indicated the effect of gaps in sequences in relation to the type of cluster analyses in case of highly identical sequences, possibly resulting from platform-associated effects such as deletion artefacts. This is in contrast to the setting of cluster analyses using sequences obtained using a single platform, since the likelihood of technology-associated characteristics in the sequences may be approximately evenly distributed over the samples. The SARS-CoV-2 sequence efficiency in relation to background sequences was highest for the Illumina EasySeq protocol, comparable with the Ion Ampliseq protocol while the ONT protocol proportionally had the lowest number of human reads. Illumina EasySeq and the ONT protocol had the shortest hands-on time, with the latter additionally having the shortest sequence runtime and real-time data analysis.

As the pandemic continues worldwide and novel variants of interest and variants of concern continue to emerge [43,44], genomic surveillance remains a critical component of the sustained management approach adhered to by the WHO [45]. Accordingly, the need for rapid SARS-CoV-2 genome sequencing protocols that can be easily adopted, automated and that are flexible and scalable remains crucial. Innovative protocol adaptations aiming at high quality sequencing of low viral load samples (Ct-values >30) [11], inherent part of the diagnostic practice, have recently been reported, and such contributions may benefit the worldwide sequence community dedicated to surveillance. Implementation and compatibility of sequence regimes are influenced by characteristics of the local laboratory settings such as the availability of local resources and sequencing platforms with high or low-throughput nature. Reduction of the hands-on time needed for library preparation and overall turnaround time, scalability, and increased cost-efficiency of protocols would be beneficial in broader settings. Here, we aimed to provide data that can assist laboratories when selecting protocols for their local setting by comparing five platforms.

Drops in read depth of certain amplicons were detected in this study using different protocols. Regions with low read depth can result from i) low amplicon coverage by design. High coverage regions have been correlated by coverage of multiple amplicons, whereas genome regions with coverage by only one amplicon resulted in low coverage [13]. Low read depth can also result from ii) a SARS-CoV-2 variant resulting in primer mismatch in that particular amplicon, iii) low efficiency of matching primers in multiplex reactions, or iiii) an imbalance of the primer concentrations present in the multiplex. In our study, the length in bp of the drop in read depth assisted the distinction between single nucleotide variants resulting in a primer mismatch and low coverage by design as underlying cause. Besides low coverage, another factor that can compromise SNV detection are primer-originated "contaminated" sequences that are PCR-amplified [13]. Wet lab methods, and similarly bioinformatic tools can influence the performance of variant detection. Inaccurate trimming of primer sequences can mask or introduce SNVs located in the primer binding site, however our study was not designed to detect such a phenomenon. Also, for example, Minimap2 [35], designed for analyses of sequences from relatively high error-rate platforms, allows considerable mismatches in the alignment with the reference sequence, whereas more stringent mapping tools can result in an absence of coverage in the mutated region. Differentiation of these type of effects resulting from analyses would require a design with cross-comparison of bioinformatic tools, which was not part of the current study. Finally, the current study was restricted by our sample collection time frame (2020), thus our analyses did not contain the later emerged mutants.

Viral (DNA/RNA) metagenomic sequencing has increasingly been adopted for pathogen diagnostics, microbiome analyses, and transcriptome analyses. The focus of the current study specifically was based on SARS-CoV-2 sequencing and specific protocols to enrich for SARS-CoV-2. Metagenomic methods work well for high-throughput sequencing of samples with high viral loads but did not perform the most stable and accurate for low viral load samples, however they were the original clinical request at a time where commercial kits had not been developed yet. This exemplifies the benefit of the approach in earlier stages of pandemics. In later stages of the pandemic it appeared beneficial to have protocols available which also work for lower viral load samples.

Importantly, with the above described pursuing emergence of variants, there is a vital need for sequencing-based approaches that tolerate mutations [46]. Probe capture-based approaches can tolerate large target sequence differences of ~10%

or more from probe sequences [47,48] in comparison with primer-based approaches . These characteristics have resulted in FDA emergency-use-authorization for hybridization-based SARS-CoV-2 genome sequencing in September 2021, in order to improve genomic surveillance of SARS-CoV-2 variants, for tracking viral evolution and guiding vaccine updates [49].

In summary, in this study five cross-platform protocols for SARS-CoV-2 genome sequencing were benchmarked and evaluated on both technical performance and practicality. The results of our study build upon previous reports by providing additional comparison data testing Illumina, Ion Torrent and ONT sequencing in parallel, incorporating technically innovative protocol steps including several analysis workflows. These data will be specifically of assistance for the sequence laboratories dedicated to ongoing surveillance efforts.

Funding

None

Conflicts of interest

The authors to have no conflicts of interest to declare.

Supplementary File Information

Suppl. Figure 1. Overview of bioinformatic analyses tools used in the current study (created using Biorender.com).

Suppl. Table 1. Protocol characteristics of the five SARS-CoV-2 sequence methods compared in the current study. NA; not applicable

Suppl. Table 2. Overview of SARS-CoV-2 PCR Ct-values per sample, genome coverage, mean depth, normalised depth, mean base quality, and mean mapping quality, per sequencing protocol. Normalised depth was calculated per 100,000 total reads.

Suppl. Table 3. Overview of SNPs and indels called by the different protocols (Q13 threshold). A threshold of 10x depth per base was considered for all platform data except for ONT data, were a 20x depth per base was considered. Per genome position, read depths are shown for respectively reference and alternate calls (DP4; ref forward, ref backward, alternate forward, alternate backward counts). *; no variant called or no coverage of position

References

[1] W. T. Harvey et al., ‘SARS-CoV-2 variants, spike mutations and immune escape’, Nat Rev Microbiol, vol. 19, no. 7, pp. 409–424, Jul. 2021, doi: 10.1038/s41579-021-00573-0.

[2] K. Tao et al., ‘The biological and clinical significance of emerging SARS-CoV-2 variants’, Nat Rev Genet, vol. 22, no. 12, pp. 757–773, Dec. 2021, doi: 10.1038/s41576-021-00408-x.

[3] Z. Chen et al., ‘Global landscape of SARS-CoV-2 genomic surveillance and data sharing’, Nat Genet, vol. 54, no. 4, pp. 499–507, Apr. 2022, doi: 10.1038/s41588-022-01033-y.

[4] B. B. Oude Munnink et al., ‘Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands’, Nat Med, vol. 26, no. 9, pp. 1405–1410, Sep. 2020, doi: 10.1038/s41591-020-0997-y.

[5] ‘https://www.gisaid.org/’, Apr. 2022.

[6] F. Wu et al., ‘A new coronavirus associated with human respiratory disease in China’, Nature, vol. 579, no. 7798, pp. 265–269, Mar. 2020, doi: 10.1038/s41586-020-2008-3.

[7] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, ‘Shotgun metagenomics, from sampling to analysis’, Nat Biotechnol, vol. 35, no. 9, pp. 833–844, Sep. 2017, doi: 10.1038/nbt.3935.

[8] M. Chiara et al., ‘Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities’, Briefings in Bioinformatics, vol. 22, no. 2, pp. 616–630, Mar. 2021, doi: 10.1093/bib/bbaa297.

[9] M. Simonetti et al., ‘COVseq is a cost-effective workflow for mass-scale SARS-CoV-2 genomic surveillance’, Nat Commun, vol. 12, no. 1, p. 3903, Dec. 2021, doi: 10.1038/s41467-021-24078-9.

[10] S. H. Rosenthal et al., ‘Development and validation of a high throughput SARS-CoV-2 whole genome sequencing workflow in a clinical laboratory’, Sci Rep, vol. 12, no. 1, p. 2054, Dec. 2022, doi: 10.1038/s41598-022-06091-0.

[11] H. Choi, M. Hwang, D. H. Navarathna, J. Xu, J. Lukey, and C. Jinadatha, ‘Performance of COVIDSeq and Swift Normalase Amplicon SARS-CoV-2 Panels for SARS-CoV-2 Genome Sequencing: Practical Guide and Combining FASTQ Strategy’, J Clin Microbiol, vol. 60, no. 4, pp. e00025-22, Apr. 2022, doi: 10.1128/jcm.00025-22.

[12] J. P. M. Coolen et al., ‘SARS-CoV-2 whole-genome sequencing using reverse complement PCR: For easy, fast and accurate outbreak and variant analysis.’, Journal of Clinical Virology, vol. 144, p. 104993, Nov. 2021, doi: 10.1016/j.jcv.2021.104993.

[13] T. Liu et al., ‘A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples’, iScience, vol. 24, no. 8, p. 102892, Aug. 2021, doi: 10.1016/j.isci.2021.102892.

[14] J. A. Nasir et al., ‘A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture’, Viruses, vol. 12, no. 8, p. 895, Aug. 2020, doi: 10.3390/v12080895.

[15] M. Xiao et al., 'Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples', Genome Med, vol. 12, no. 1, p. 57, Dec. 2020, doi: 10.1186/s13073-020-00751-4.

[16] F. Wegner et al., 'External Quality Assessment of SARS-CoV-2 Sequencing: an ESGMD-SSM Pilot Trial across 15 European Laboratories', J Clin Microbiol, vol. 60, no. 1, pp. e01698-21, Jan. 2022, doi: 10.1128/JCM.01698-21.

[17] J. Plitnick, S. Griesemer, E. Lasek-Nesselquist, N. Singh, D. M. Lamson, and K. St. George, 'Whole-Genome Sequencing of SARS-CoV-2: Assessment of the Ion Torrent AmpliSeq Panel and Comparison with the Illumina MiSeq ARTIC Protocol', J Clin Microbiol, vol. 59, no. 12, pp. e00649-21, Nov. 2021, doi: 10.1128/JCM.00649-21.

[18] M. Zlei et al., 'Absence of rapid T cell control corresponds with delayed viral clearance in hospitalised COVID-19 patients', In Review, preprint, Aug. 2021. doi: 10.21203/rs.3.rs-783703/v1.

[19] J. Quick et al., 'Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples', Nat Protoc, vol. 12, no. 6, pp. 1261–1276, Jun. 2017, doi: 10.1038/nprot.2017.066.

[20] V. M. Corman et al., 'Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR', Euro Surveill., vol. 25, no. 3, 2020, doi: 10.2807/1560-7917.ES.2020.25.3.2000045.

[21] E. C. Carbo et al., 'Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics', Journal of Clinical Virology, p. 104566, Jul. 2020, doi: 10.1016/j.jcv.2020.104566.

[22] S. van Boheemen et al., 'Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients', The Journal of Molecular Diagnostics, vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[23] A. L. van Rijn et al., 'The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease', PLoS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[24] T. Briese et al., 'Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis', mBio, vol. 6, no. 5, pp. e01491-15, Sep. 2015, doi: 10.1128/mBio.01491-15.

[25] E. C. Carbo et al., 'Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness', Journal of Clinical Virology, vol. 131, p. 104594, Oct. 2020, doi: 10.1016/j.jcv.2020.104594.

[26] A. M. Bolger, M. Lohse, and B. Usadel, 'Trimmomatic: a flexible trimmer for Illumina sequence data', Bioinformatics, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[27] B. Langmead, 'Aligning Short Sequencing Reads with Bowtie', Curr. Protoc. Bioinform., vol. 32, no. 1, Dec. 2010, doi: 10.1002/0471250953.bi1107s32.

[28] 'https://www.ncbi.nlm.nih.gov/nuccore/1798174254', Apr. 2021.

[29] H. Li et al., 'The Sequence Alignment/Map format and SAMtools', Bioinformatics, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[30] H. Li, 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', Bioinformatics, vol. 27, no. 21, pp. 2987–2993, Nov. 2011, doi: 10.1093/bioinformatics/btr509.

[31] P. Danecek et al., 'Twelve years of SAMtools and BCFtools', GigaScience, vol. 10, no. 2, p. giab008, Jan. 2021, doi: 10.1093/gigascience/giab008.

[32] H. Li and R. Durbin, 'Fast and accurate long-read alignment with Burrows-Wheeler transform', Bioinformatics, vol. 26, no. 5, pp. 589–595, Mar. 2010, doi: 10.1093/bioinformatics/btp698.

[33] 'https://www.ncbi.nlm.nih.gov/nuccore/MN908947'.

[34] GitHub - rrwick/Porechop: adapter trimmer for Oxford Nanopore reads.

[35] M. Martin, 'Cutadapt removes adapter sequences from high-throughput sequencing reads', EMBnet j., vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.

[36] H. Li, 'Minimap2: pairwise alignment for nucleotide sequences', Bioinformatics, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191..

[37] 'https://github.com/pysam-developers/pysam'.

[38] D. J. Baker et al., 'CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes', Genome Med, vol. 13, no. 1, p. 21, Dec. 2021, doi: 10.1186/s13073-021-00839-5.

[39] C. Spearman, 'The Proof and Measurement of Association between Two Things', The American Journal of Psychology, vol. 15, no. 1, p. 72, Jan. 1904, doi: 10.2307/1412159.

[40] F. Sievers and D. G. Higgins, 'Clustal Omega', Current Protocols in Bioinformatics, vol. 48, no. 1, Dec. 2014, doi: 10.1002/0471250953.bi0313s48.

[41] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix', Molecular Biology and Evolution, vol. 26, no. 7, pp. 1641–1650, Jul. 2009, doi: 10.1093/molbev/msp077.

[42] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', PLoS ONE, vol. 5, no. 3, p. e9490, Mar. 2010, doi: 10.1371/journal.pone.0009490.

[43] A. Maxmen, 'Are new Omicron subvariants a threat? Here's how scientists are keeping watch', Nature, vol. 604, no. 7907, pp. 605–606, Apr. 2022, doi: 10.1038/d41586-022-01069-4.

[44] E. Callaway, 'Are COVID surges becoming more predictable? New Omicron variants offer a hint', Nature, vol. 605, no. 7909, pp. 204–206, May 2022, doi: 10.1038/d41586-022-01240-x.

[45] 'https://www.who.int/westernpacific/news-room/feature-stories/item/moving-from-pandemic-response-to-sustained-management-of-covid-19-in-the-western-pacific-region'.

[46] N. M. Butler, P. A. Atkins, D. F. Voytas, and D. S. Douches, 'Generation and Inheritance of Targeted Mutations in Potato (Solanum tuberosum L.) Using the CRISPR/Cas System', PLoS ONE, vol. 10, no. 12, p. e0144591, Dec. 2015, doi: 10.1371/journal.pone.0144591.

[47] 'https://www.twistbioscience.com/resources/white-paper/effects-mismatches-dna-capture-hybridization', Apr. 2022.

[48] 'https://apps.who.int/iris/handle/10665/338480', Apr. 2022.

[49] D. Nagy-Szakal et al., 'Targeted Hybridization Capture of SARS-CoV-2 and Metagenomics Enables Genetic Variant Discovery and Nasal Microbiome Insights', Microbiol Spectr, vol. 9, no. 2, pp. e00197-21, Oct. 2021, doi: 10.1128/Spectrum.00197-21.