

Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq

Derek Tshiabuila (✉ derektshiabuila@gmail.com)

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences <https://orcid.org/0000-0001-5221-2126>

Jennifer Giandhari

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Sureshnee Pillay

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Upasana Ramphal

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Yajna Ramphal

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Arisha Maharaj

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Ugochukwu Jacob Anyaneji

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Yeshnee Naidoo

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Houriyyah Tegally

Stellenbosch University Faculty of Medicine and Health Sciences

Emmanuel James San

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Eduan Wilkinson

Stellenbosch University - Tygerberg Campus: Stellenbosch University Faculty of Medicine and Health Sciences

Richard J. Lessells

University of KwaZulu-Natal Nelson R Mandela School of Medicine: University of KwaZulu-Natal College of Health Sciences

Tulio de Oliveira


University of Stellenbosch Faculty of Science: Stellenbosch University Faculty of Science

Research Article

Keywords: SARS-CoV-2, Illumina MiSeq, Oxford Nanopore Technology GridION, Nanopore sequencing, Next Generation Sequencing, Bioinformatics

Posted Date: February 17th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1249711/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Abstract

Background: Over 4 million SARS-CoV-2 genomes have been sequenced globally in the past 2 years. This has been crucial in elucidating transmission chains within communities, the development of new diagnostic methods, vaccines, and antivirals. Although several sequencing technologies have been employed, Illumina and Oxford Nanopore remain the two most commonly used platforms. The sequence quality between these two platforms warrants a comparison of the genomes produced by the two technologies. Here, we compared the sequence quality produced by the Oxford Nanopore Technology GridION and the Illumina MiSeq for 28 sequencing runs.

Results: Our results show that the MiSeq had a significantly higher number of sequences classified by Nextclade as good and mediocre compared to the GridION. The MiSeq also had a significantly higher sequence coverage and mutation counts than the GridION.

Conclusion: Due to the low sequence coverage, high number of indels, and sensitivity to viral load noted with the GridION when compared to MiSeq, we can conclude that the MiSeq is more favourable for genomic surveillance, as successful genomic surveillance is dependent on high quality, near-whole genome sequences.

Background

December 2019 saw a novel viral pneumonia emerge from a seafood market in Wuhan China later found to be a new type of Coronavirus, now known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (1, 2). On 11 March 2020, after approximately 118 000 cases had been reported globally, the World Health Organization (WHO) declared SARS-CoV-2 a global pandemic (3, 4). SARS-CoV-2 is an ongoing pandemic that requires continuous surveillance with approximately 270 031 622 cases confirmed globally as of 14 December 2021 (3, 5).

Sequencing of SARS-CoV-2 allowed for the rapid identification of the virus and the development of diagnostic tests and other tools for a rapid response to the pandemic (6). Sequencing provides genotypic information about a patient's infection, which can be used to gain knowledge on the specific infecting strain, assist in identifying transmission within communities, and advance the development of new diagnostic methods, vaccines, and antivirals (7). Multiple next generation sequencing (NGS) technologies have been used for SARS-CoV-2 sequencing, including Sanger, Illumina, ION torrent, and Oxford Nanopore Technology (8). However, Illumina sequencing remains the most commonly used technology (9). As of 05 November 2021, 4 892 742 SARS-CoV-2 sequences had been deposited into the Global Initiative on Sharing all Influenza Data (GISAID) with over 65% from Illumina and approximately 25% from Oxford Nanopore Technology (ONT) (10).

A major challenge with whole-genome sequencing (WGS) is obtaining whole viral genomes from clinical samples promptly (11). Illumina SARS-CoV-2 sequencing is generally limited by long sequencing times and the high cost and labour associated with library preparation for high-throughput sequencing (12).

Another limitation is their relatively short reads (2 x 300 bp), as genomes generally contain multiple repeated sequences, known as tandem repeats, that may be longer than the NGS reads and may result in gaps and misassemblies (13). Due to the large footprint of most sequencers, portability can be a challenge which is unfortunate as there is generally a large distance between sample collection sites and sequencing laboratories (14). Nanopore sequencing overcomes these challenges as they sequence in real-time and are long-read sequencing technologies that allow for portability and have a relatively low initial investment on sequencing equipment with the MinION costing \$1000 (15). ONT sequencing is, however, limited by the high number of false negatives and low sensitivity (16).

Short-read sequencing technologies are useful for population-level genetic analysis and clinical variant discovery as they provide low-cost, high-accuracy data when done in large batches. Long-read sequencing approaches, however, are well suited for de novo genome assembly, sequencing of genomes with long repetitive regions, copy number alterations, and complex structural variations (17). Several studies have compared the sequencing of SARS-CoV-2 between Illumina and ONT platforms and have shown that despite the high error rates observed with ONT sequencing, highly-accurate SARS-CoV-2 consensus genomes can be achieved (18). ONT sequencing, however, failed to detect short indels identified by Illumina sequencing (18). There has also been a lower raw-read accuracy with nanopore sequencing when compared to Illumina sequencing (18, 19).

A comparison of SARS-CoV-2 WGS sequence coverage and variant detection between Illumina and Nanopore sequencing is necessary as it allows us to determine whether SARS-CoV-2 genomes produced by Nanopore sequencing can be reliably used for genomic surveillance and the development of diagnostic measures. This study aimed to determine whether Nanopore sequencing is a viable alternative to Illumina sequencing for SARS-CoV-2 whole-genome analysis. We hypothesize that Nanopore sequencing will produce consensus genomes that are comparable to consensus genomes produced by Illumina sequencing at a faster rate. SARS-CoV-2 sequencing results, for multiple runs, from the Illumina MiSeq and the ONT GridION were compared and although Nanopore sequencing was able to produce complete SARS-CoV-2 genomes, the sequence quality observed was not as good as those obtained with Illumina sequencing. The ONT GridION can sequence up to 5 flowcells with 96 samples in a single run and is cheaper than sequencing with the Illumina MiSeq. These advantages can allow for more clinical facilities to sequence SARS-CoV-2 allowing for a greater response to the COVID-19 pandemic.

Materials And Methods

Study Population

The study population consisted of positive COVID-19 male and female patients whose nasopharyngeal swabs were sent from routine PCR diagnostic services for genomic surveillance to the Kwazulu-Natal Research Innovation and Sequencing Platform (KRISP). A total of 2608 COVID-19 positive nasopharyngeal swabs were used for sequencing from 28 different runs split evenly between the GridION

and MiSeq. Samples were randomized and were from South Africa, Angola, Malawi, Mozambique, and Zimbabwe.

Total Nucleic Acid Extraction

RNA was extracted using the NA/gDNA kit on the automated Chemagic 360 system (Perkin Elmer) as per the manufacturer's instructions. Briefly, samples were lysed using lysis buffer and proteinase K, followed by binding to silica magnetic beads. The beads were then washed to remove unbound samples, and the RNA was eluted. Extracted RNA was stored at -80°C before use.

Tiling PCR

Complementary DNA synthesis was performed using SuperScript IV reverse transcriptase (Life Technologies) in combination with random hexamer primers. This was then followed by gene-specific multiplex PCR using the ARTIC protocol (20). Primers were designed on a primal scheme (<http://primal.zebraproject.org/>) to cover the SARS-CoV-2 whole genome. Primers generated were 400 base pair (bp) amplicons, with an overlap of 70 bp to cover the 30 kilobases (kb) SARS-CoV-2 genome. Purification of PCR products was performed using AmpureXP purification beads in a 1:1 ratio (Beckman Coulter, High Wycombe, UK) and quantification was performed using the Qubit double-strand DNA (dsDNA) High Sensitivity Assay Kit on a Qubit 4.0 instrument (Life Technologies).

Illumina MiSeq Library Preparation and Sequencing

Sequencing libraries were generated using the amplicons generated by tiling PCR as described above. Indexed paired-end libraries were prepared using the Nextera DNA Flex Library Prep Kits (Illumina) as per the manufacturer's instructions. Briefly, amplicons were tagmented to allow for unfragmented DNA to be cleaved and tagged. Each sample was barcoded with a unique barcode using the Nextera CD Indexes (Illumina) to enable downstream pooling of all libraries. Libraries were purified and normalized to 4 nM prior to pooling. The pooled library was denatured using 0.2 N sodium acetate and then diluted to a final concentration of 8 pM. The library was spiked with 1% PhiX Control v3 (adapter-ligated library used as a control), and the libraries were sequenced using a 500-cycle v2 MiSeq Reagent Kit on the Illumina MiSeq instrument (Illumina, San Diego, CA, USA). The full details of the amplification and sequencing have been previously published (21). Fastq files produced from Illumina MiSeq were assembled using Genome Detective (<https://www.genomedetective.com/>) and the coronavirus typing tool (22).

ONT GridION Library Preparation and Sequencing

Amplicons generated using the tiling PCR were prepared for nanopore sequencing using the ONT Native Barcoding Expansion Kits as per the manufacturer's guidelines. Libraries were multiplexed on FLO-MIN106 flowcells and run on the GridION X5. Furthermore, a no-template control from the PCR amplification step was added to each plate before running. Sequencing performance was monitored, in real-time, using the MinKNOW software app. Sequencing was terminated after 21hrs and the resulting reads were base-called using Guppy (4.0.14) and aligned to the Wuhan-Hu-1 reference genome (MN908947.3) using minimap2 (2.17-r941). Primer sequences were trimmed from the termini of read

alignments and sequencing depth was capped at a maximum of 400-fold coverage using the ARTIC tool `align_trim`. Variant candidates were identified using Nanopolish (23).

Sequence Analysis

Consensus sequences produced by both platforms were uploaded to Nextclade Online Tool (<https://clades.nextstrain.org/>) for sequence clade assignments, mutation calling, sequence quality checks, and to determine the sequence position on the SARS-CoV-2 phylogenetic tree. A Maximum-likelihood (ML) tree was constructed using IQ-TREE and was visualized using FigTree v1.4.4 (<https://github.com/rambaut/figtree/releases>) (24). Data visualization and statistical analysis were performed using ggplot2 v3.3.1 package and R v.4.1.1.

Statistical Considerations

The non-parametric nature of the data influenced the use of a Wilcoxon test to compare the number of sequences produced by the GridION and the MiSeq classified within each category of the online Nextclade sequence analysis tool. The Wilcoxon test was also used to compare the difference in sequence coverage, number, and type of mutations detected between the GridION and the MiSeq. Statistical correlations were performed between Ct score and sequence coverage and Ct score and the number of reads for both platforms.

Ethics

The University of KwaZulu-Natal Biomedical Research Ethics Committee waived the requirement for informed consent and approved the study (protocol reference no. BREC/00001195/2020; project title: COVID-19 transmission and natural history in KwaZulu-Natal, South Africa: epidemiological investigation to guide prevention and clinical care). All methods were performed in accordance with the relevant guidelines and regulations. We also used de-identified remnant nasopharyngeal and oropharyngeal swab samples from patients testing positive for SARS-CoV-2 by RT-qPCR from public health laboratories in South Africa. Informed consent for study participation was not applicable for this study because de-identified (anonymous) remnant samples, which would have been otherwise discarded, were used.

Results

Comparison of sequencing performance

To compare sequencing performance and runtime between the MiSeq and the GridION, Run116 was sequenced on both platforms (Table 1). A total of 93 samples were sequenced and 93 consensus genomes were produced after assembly using Genome Detective. The sequencing runtime for the MiSeq was 36 hrs, whilst the GridION had a runtime of 21 hrs. The MiSeq had an overall higher average coverage than the GridION, having coverages of 94.34% and 72.96%, respectively. There was also a higher number of sequences that passed the QC used for GISAID submissions (>80% sequence coverage) from the MiSeq, 83 (89.2%), than the GridION, 29 (27.9%).

Table 1
Comparison of sequencing Run116 on both the MiSeq and the GridION

	Run116	
	MiSeq	GridION
Runtime (hrs.)	36	21
No. of samples sequenced	93	93
Consensus genomes	93	93
Average coverage (X)	94.34%	72.96%
Passing GISAID QC (>+80%)	83 (89.2%)	29 (27.9%)

The table above summarizes the sequencing of Run116 on both the MiSeq and the GridION. The sequencing runtime for the MiSeq was 36 hrs, whilst that of the GridION was 21hrs. Of the 93 samples sequenced by both platforms, 93 consensus genomes were produced by each. Sequences from the MiSeq had an average coverage of 94.34% with 89.2% having a coverage of 80% and over. Sequences from the GridION had an average coverage of 72.96% with 27.9% having a coverage of 80% and over.

Comparison of sequence quality of Nanopore and Illumina sequencing

Consensus sequences produced by the GridION and the MiSeq were uploaded to Nextclade to determine the sequence quality (Figure 1). Both the GridION and the MiSeq had a total of 14 runs with 1255 and 1183 consensus genomes, respectively. The total number of consensus genomes produced by the GridION and the MiSeq was significantly different ($p = 0.0053$). Nextclade classifies sequences as either good, mediocre, or bad, based on the amount of missing data, and the number of mixed sites, private mutations, clustered mutations, frameshifts, and misplaced stop codons (Table 2). The number of sequences the two platforms classified as good ($p = 0.00280$), mediocre ($p = 0.00250$), and bad ($p = 0.00037$) also differed significantly (Table 3).

Table 2
QC score thresholds implemented by Nextclade for sequence quality analysis

Score	Meaning	Color Designation
0 – 29	Good quality	Green
30 – 99	Mediocre quality	Yellow
100 and above	Bad quality	Red

The table above shows the QC score thresholds used to classify sequences as either good, bad, or mediocre, and the color shown on the online Nextclade tool for each group. Sequencing scores ranging

between 0 – 29 are classified as good and shown as green, 30 – 99 are classified as mediocre and shown as yellow, whilst 100 and above are classified as bad and shown as red.

Table 3
Nextclade sequence quality analysis for the GridION and the MiSeq

Nextclade Sequence Quality	GridION	MiSeq
Good quality	17% (231)	52% (571)
Mediocre quality	12% (129)	27% (330)
Bad quality	71% (895)	21% (282)

The table above shows the sequence quality analysis for sequences produced using the GridION and the MiSeq. Nextclade classified 17% of sequences from the GridION as good quality, 12% as mediocre, and 71% as bad quality. Nextclade classified 52% of sequences from the MiSeq as good, 27% as mediocre, and 21% as bad quality sequences.

Comparison of sequence coverage generated by the GridION and MiSeq

Identical samples (RUN116) were sequenced on both the GridION and the MiSeq and the sequencing coverage was compared to determine the effect of sample quality on sequencing (Figure 2-A). All the runs for both platforms were then compared (Figure 2-B). A total of 86 sequences were used from RUN116 after removing sequences with more than 100 mutations. Samples run on the MiSeq had a significantly greater sequence coverage than the GridION ($p = 8.1 \times 10^{-16}$). GridION sequences ranged from 35 – 100%, whilst MiSeq sequences ranged from 80 – 100%. The sequencing coverage for all runs, 2351 sequences, was then compared. There was a significantly higher overall sequencing coverage observed with the MiSeq than with the GridION ($p < 2.2 \times 10^{-16}$).

Comparison of Orf1ab- and S-gene coverage for GridION and MiSeq sequencing

To compare the depth of coverage of the ORF1ab- and S-gene for the GridION and the MiSeq, fastq files produced from both platforms were assembled on Genome Detective to produce consensus genomes. The results for each sequence were obtained and the coverages for the ORF1ab-gene (Figure 3-A) and S-gene (Figure 3-B) were compared. All 14 runs for each platform were compared and Wilcoxon rank sum tests were performed. The ORF1ab-gene coverage ranged from 35 – 100% for the GridION and 80 – 100% for the MiSeq. The S-gene coverage ranged from 25 – 100% for the GridION and 80 – 100% for the MiSeq. There was a statistically significant difference in coverage for both genes on the GridION and the MiSeq with $p = 1.2 \times 10^{-15}$ (RUN116) and $p = 1.7 \times 10^{-15}$ (all sequences).

Effect of Ct score on sequencing using the GridION and MiSeq

A correlation was performed to determine the effect of Ct score on sequence coverage (Figure 4) and the number of reads produced by the GridION and the MiSeq during sequencing (Figure 5). Due to the availability of Ct scores, three runs were used for each platform. Run101 (35 samples), Run111 (91 samples), and Run123 (64 samples), represented by graphs A, B, and C, respectively, were used for the GridION. Run100 (68 samples), Run109 (54 samples), and Run122 (88 samples), represented by graphs D, E, and F, respectively, were used for the MiSeq. A negative correlation was observed between Ct Score and sequence coverage for all six runs. The GridION's Runs 101, 111, and 123 had correlation coefficients of $R = -0.88$ ($p = 4.5e-12$), $R = -0.45$ ($p = 7.2e-06$), and $R = -0.31$ ($p = 0.012$), respectively. The MiSeq's Runs 100, 109, and 122 had correlation coefficients of $R = -0.35$ ($p = 0.0039$), $R = -0.19$ ($p = 0.18$), and $R = -0.33$ ($p = 0.0017$), respectively. We note a significantly strong negative correlation between Ct score and number of reads for all GridION runs, whereas a significantly negative correlation was only noted for Run122 sequenced on the MiSeq. Run100 and Run109 showed non-significant correlations.

Mutation analysis

To determine whether the number of mutations detected by GridION and MiSeq differed significantly, the number of mutations detected for each sample was compared for Run116 (Figure 6-A) and all the runs (Figure 6-B). The total number of insertions, deletions, and substitutions detected by both platforms were also compared for Run116 (Figure 6-C) and all the runs (Figure 6-D). A total of 181 sequences obtained from the GridION and the MiSeq for Run116 were analyzed and a significant difference was noted in the number of mutations detected by each platform (Wilcoxon, $p = 3.7e-08$) with a greater number of mutations detected by the MiSeq (8 – 96 mutations) than the GridION (6 – 56 mutations). We also noted a significant difference (Wilcoxon, $p = 1.5e-09$) between the number of mutations detected from the sequences analyzed on the MiSeq (1183 sequences) and the GridION (1255 sequences). There was a significant difference in the number of insertions (Wilcoxon, $p = 8.2e-04$) and substitutions (Wilcoxon, $p = 5.3e-06$) detected by both platforms for RUN116. However, when all runs were analyzed; only the number of insertions were significantly different between the two platforms (Wilcoxon, $p = 7.5e-15$).

Phylogenetic analysis

To determine whether there was a difference in the phylogenetic inference between sequences generated by the GridION and the MiSeq, Run116 samples were sequenced on both platforms. A total of 93 samples sequenced on both the GridION and the MiSeq were uploaded to Nextclade and the results were compared. Of the 93 samples, 27 samples were classified within different clades (Table 4). A phylogenetic tree of the 27 samples was then created using IQTREE and visualized using FigTree (Figure 7). Of the 27 samples, only one sample, highlighted in blue, was grouped on the same branch.

Table 4
Comparison of the sequence coverage and assigned clade for run116
samples on Nextclade

Sample	Coverage (%)		Clade	
	GridION	MiSeq	GridION	MiSeq
K013400	72	92	20C	20A
K013408	57	86	20H (Beta, V2)	20C
K013410	70	90	20C	20A
K013411	76	93	20H (Beta, V2)	20A
K013415	63	91	20C	20A
K013417	63	94	20C	20A
K013418	62	94	20C	20A
K013423	63	91	20C	20A
K013425	60	93	20C	20A
K013426	57	89	20C	20A
K013429	64	95	20H (Beta, V2)	20C
K013432	65	93	20C	20A
K013433	50	94	20C	20A
K013434	68	94	20C	20A
K013437	35	92	20H (Beta, V2)	20C
K013445	65	91	20C	20A
K013447	92	98	20A	20D
K013449	49	94	20C	20A
K013450	68	97	20C	20A
K013452	68	94	20C	20A
K013454	56	90	20C	20A
K013462	51	92	20C	20A
K013465	60	94	20C	20A
K013467	50	92	20C	20A
K013470	72	89	20C	20H (Beta, V2)

Coverage (%)			Clade	
K013476	69	91	20C	20H (Beta, V2)
Total	20A	1	20	
	20C	22	3	
	20D	0	1	
	20H (Beta, V2)	4	3	

The table above highlights the 27 samples which were sequenced on both the MiSeq and the GridION but were classified in different clades by Nexclade. Clades identified by the GridION include 20A (n = 1), 20C (n = 22), and 20H (Beta, V2) (n = 4). Clades identified by the MiSeq include 20A (n = 20), 20C (n = 3), 20D (n = 1), and 20H (Beta, V2) (n = 3). There was also an overall higher sequence coverage for consensus genomes from the MiSeq when compared to the GridION.

Discussion

SARS-CoV-2 has caused a global health crisis as it is highly infectious and risks mutations that could result in more lethal variants (1, 25). A major factor in helping curb the spread of the virus and decreasing the infection rate is rapidly sequencing the virus to detect new strains and identify transmission chains (7). The sequencing runtime on the MiSeq for Run116 was 36 hours, whilst on the GridION it was 21 hours. This 10-hour decrease in sequencing time allows for 480 samples to be sequenced each day on the GridION in comparison to the 96 that can be sequenced on the MiSeq every 36 hours. This is in agreement with reports that nanopore sequencing takes approximately 20 hours as a rapid library prep kit supplied by ONT can be used (26, 27). The lack of an image analysis step during nanopore sequencing facilitates real-time base-calling, which allows for the rapid detection of DNA for pathogen screening from clinical samples (28).

Studies have shown that Illumina sequencing may still be the most accurate way to sequence viruses (29). The majority of errors noted between Nanopore and Illumina consensus genomes have been attributed to Nanopore sequencing errors (30). Run116 samples were sequenced on both platforms to determine whether there was a significant difference in the sequencing coverage regardless of the sample. Sequencing coverage was significantly greater with the MiSeq when compared to the GridION and this result was also observed when comparing all sequence runs. Sequence coverage can be affected by sequencing time and thus GridION coverage may have increased if left to sequence for longer. We also note a statistically significant higher sequencing coverage for the S-gene and ORF1ab-gene with the MiSeq than with the GridION. Nanopore technology has been shown to provide lower per-read sequencing coverage when compared to short-read sequencing (31). Coverage biases seen with ONT's sequencing protocol can be a result of truncated reads caused by pore blocking or fragmentation

during library prep as transcripts are sequenced from the 3' to 5' end (32). ONT has made error correction tools such as Nanopolish available to try and reduce the error rate observed with Nanopore sequencing (33). In this study, variant calling was achieved using Nanopolish but we still note a significantly lower sequence quality obtained from the GridION than the MiSeq. These low-quality sequences cannot be used to confidently acquire information on the infecting viral strain and are generally removed through a series of quality control checks (34). Although more sequences can be produced using the GridION than the MiSeq, the low-quality sequences which are removed would eliminate the advantage of having a large number of consensus genomes produced.

Higher sequencing coverage for the Illumina MiSeq has been associated with lower Ct scores (21). Ct score is a value that refers to the number of cycles required to amplify viral RNA to a detectable level. There is therefore an inverse relationship between Ct score and viral load (35). In this investigation, we also noted an inverse relationship between Ct score and sequence coverage for both GridION and MiSeq sequencing. There is, however, a significantly stronger negative correlation seen with the GridION than the MiSeq, which may imply that the MiSeq's sequencing capabilities are less affected by sample Ct score and as a result, can be used for sequencing of samples within the early stages of infection when viral load is still low. This was, however, limited by not having the same runs to compare between the GridION and the MiSeq. Further analysis is required as the number of samples analyzed for each run was low and inconsistent due to the availability of Ct scores received with sample metadata. Additional analyses should be conducted to understand characteristics such as coverage bias, sequence biases, and reproducibility for the GridION sequencing platform (31). Sample quality may also have an effect on sequencing and thus it is very important to maintain a cold chain during storage of swabs and RNA.

Identifying sequence mutations involves aligning a sequence to a reference genome and identifying changes within the sequence. This is important, as it allows us to identify gene variants that may play a major role in the diagnosis of diseases (36). It has been shown that long-read sequencing platforms have a high error rate, which is mostly indels that are assumed to be randomly distributed within each read (37, 38). Prediction and interpretation of protein sequences may, therefore, be critically affected due to frameshifts and premature stop codons that may be introduced by the indels (39).

There was a significantly greater number of mutations detected by the MiSeq than the GridION for identical samples sequenced on both platforms. Although Nanopore platforms have been shown to make a large number of indel errors, in this study the MiSeq had a significantly higher number of insertions than the GridION. Paired-end sequencing, utilized by Illumina MiSeq, produces twice the number of reads, for the same sample and library preparation efforts, as single-end sequencing. This allows for a more accurate read alignment and detection of indel variants (40). Short read lengths have been shown to hinder the assignment of reads to parts of the genome that are complex, phasing of variants, resolving regions that are repeated, and the introduction of gaps and ambiguous regions in de novo assemblies. Longer reads can be used for sequencing of extended repetitive regions, allowing for the identification of mutations that are generally associated with disease (41). The higher number of indels noted with

GridION sequencing highlights that genomic surveillance using Nanopore sequencing should be conducted cautiously as incorrect information on a viral strain can be obtained.

The rapid increase in COVID-19 cases has been linked to different SARS-CoV-2 viral lineages (42). Viral lineages are separated based on the number and type of mutations they contain that differ from the parent strain (43). From the 93 sequences analyzed from both platforms, 27 sequences were classified within different clades. These sequences had unique mutations and the clade differences noted between the two platforms were 20A – 20C and 20C – 20H(Beta, V2). As the number of indels and substitutions produced by the MiSeq and the GridION were significantly different, we can expect there to be differences in clade classifications as viral clades are subject to viral-defining mutations (25). Table 3 shows that the GridION sequences have lower coverages than the MiSeq sequences. This may be one of the factors causing a difference in the clade assignment as errors arising from the amplification and sequencing process may result in incomplete genome coverage, which affects phylogenetic inference (44). Rambaut et al., 2020 suggests that new lineages should only be proposed if the genome coverage exceeds 70% of the coding region. Degradation of RNA can result in the introduction of mutations, which may cause a variant change (45). The GridION library for RUN116 was prepared simultaneously with that of the MiSeq and the amount of RNA used is also lower. Therefore, we can eliminate the possibility of RNA degradation and RNA input amount as factors that may have caused a difference in the variants called by each instrument. Lineages identified by the GridION need to be further analyzed to determine whether the mutations are valid or are a result of sequencing errors. Accurate identification of lineages can assist in identifying transmission chains and allow for the development of diagnostic methods and treatments (42).

Conclusions

The results of this study show that the ONT GridION is less ideal for SARS-CoV-2 genomic surveillance than the Illumina MiSeq but can be used to produce consensus genomes from samples of high quality and low CT scores. The GridION can, however, be used as a diagnostic method for SARS-CoV-2 as it can sequence up to 480 samples every 21 hours and .. This can allow healthcare facilities to identify and isolate infected individuals, thus aiding in stopping the spread of the disease.

Abbreviations

1. SARS-CoV-2
2. WHO
3. NGS
4. GISAID
5. ONT
6. WGS
7. COVID-19

8. KRISP
9. RNA
10. DNA
11. PCR
12. dsDNA
13. ML
14. Ct
15. QC
16. ns

Declarations

Ethics approval and consent to participate

This study was approved by the Biomedical Research Ethics Committee (BREC)-UKZN (BREC/00002764/2021, 04 October 2021). Consent to participate was waived as the study made use of de-identified remnant nasopharyngeal and oropharyngeal swab samples from patients testing positive for SARS-CoV-2.

Consent for publication

Not applicable

Availability of data and material

The datasets generated and analysed during the current study are available in the SRA (<https://www.ncbi.nlm.nih.gov/sra>) and GISAID (<https://www.gisaid.org/>) data repositories.

Competing interests

The authors declare that they have no competing interests

Funding

This study was funded by the University of Kwazulu-Natal and the Kwazulu-Natal Research Innovation and Sequencing Platform.

Author Contributions

Conceptualization, DT; methodology, DT, JG, and SP; formal analysis, DT; resources, JG, and SP; data curation, DT; writing—original draft preparation, DT; writing—review and editing, JG, SP, UR, YR, AM, UJA, YN, HT, EJS, EW, RJL, TdO; visualization, DT and EJS; supervision, TdO; project administration, TdO. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

I would like to thank Mrs Lavanya Singh for assisting in validating the GridION X5 for SARS-CoV-2 sequencing.

References

1. World Health O. Clinical management of severe acute respiratory infection when novel coronavirus (2019-nCoV) infection is suspected: interim guidance, 28 January 2020. Geneva: World Health Organization; 2020. Contract No.: WHO/nCoV/Clinical/2020.3.
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727–33.
3. Esbin MN, Whitney ON, Chong S, Maurer A, Darzacq X, Tjian R. Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for COVID-19 detection. *RNA*. 2020;26(7):771–83.
4. World Health O. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 2020 [Available from: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020>].
5. Phucharoen C, Sangkaew N, Stosic K. The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data. *EClinicalMedicine*. 2020;27.
6. Seth-Smith H, Bonfiglio F, Cuénod A, Reist J, Egli A, Wüthrich D. Evaluation of rapid library preparation protocols for whole genome sequencing based outbreak investigation. *Frontiers in public health*. 2019;7:241.
7. St Hilaire BG, Durand NC, Mitra N, Pulido SG, Mahajan R, Blackburn A, et al. A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. *bioRxiv*. 2020:2020.04.25.061499.
8. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine Biotechnology*. 2012;2012:251364.
9. GISAID. Pandemic coronavirus causing COVID-19 2021 [Available from: <https://www.gisaid.org/>].
10. Chantal Babb de Villiers LB, Cook S. Joanna Janus, Emma Johnson, Mark Kroese. Next generation sequencing for SARS-CoV-2. PHG Foundation; 2021.
11. Resende PC, Motta FC, Roy S, Appolinario L, Fabri A, Xavier J, et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. *bioRxiv*. 2020:2020.04.30.069039.
12. Gohl DM, Garbe J, Grady P, Daniel J, Watson RHB, Auch B, et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. *BMC Genom*. 2020;21(1):863.
13. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet*. 2018;34(9):666–81.

14. Xu Y, Lewandowski K, Jeffery K, Downs LO, Foster D, Sanderson ND, et al. Nanopore metagenomic sequencing to investigate nosocomial transmission of human metapneumovirus from a unique genetic group among haematology patients in the United Kingdom. *J Infect.* 2020;80(5):571–7.
15. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet.* 2010;19(R2):R227-R40.
16. Wang M, Fu A, Hu B, Tong Y, Liu R, Gu J, et al. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. *medRxiv.* 2020:2020.03.04.20029538.
17. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* 2016;17(6):333–51.
18. Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun.* 2020;11(1):6272.
19. Jayamohan H, Lambert CJ, Sant HJ, Jafek A, Patel D, Feng H, et al. SARS-CoV-2 pandemic: a review of molecular diagnostic tools including sample collection and commercial response with associated advantages and limitations. *Anal Bioanal Chem.* 2021;413(1):49–71.
20. Quick J. nCoV-2019 sequencing protocol. *Protocols* io[Google Scholar]. 2020.
21. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes (Basel).* 2020;11(8).
22. Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics.* 2020;36(11):3552–5.
23. Nick Loman WR, Andrew Rambaut. nCoV-2019 novel coronavirus bioinformatics protocol 2020-01-23 [Available from: <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>].
24. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol.* 2014;32(1):268–74.
25. Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nature Medicine.* 2021.
26. James P, Stoddart D, Harrington ED, Beaulaurier J, Ly L, Reid SW, et al. LamPORE: rapid, accurate and highly scalable molecular screening for SARS-CoV-2 infection, based on nanopore sequencing. *medRxiv.* 2020:2020.08.07.20161737.
27. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. *Dev Growth Differ.* 2019;61(5):316–26.
28. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology.* 2018;36(4):338–45.
29. Hourdel V, Kwasiborski A, Balière C, Matheus S, Batéjat CF, Manuguerra J-C, et al. Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing Through Comparison of

- MinION and Illumina iSeq100(TM) System. *Frontiers in microbiology*. 2020;11:571328-.
30. McNaughton AL, Roberts HE, Bonsall D, de Cesare M, Mokaya J, Lumley SF, et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Sci Rep*. 2019;9(1):7081.
31. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*. 2020;21(1):30.
32. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*. 2013;31(11):1009–14.
33. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods*. 2015;12(8):733–5.
34. Gleizes A, Laubscher F, Guex N, Iseli C, Junier T, Cordey S, et al. Virosaurus A Reference to Explore and Capture Virus Genetic Diversity. *Viruses*. 2020;12(11):1248.
35. Tom MR, Mina MJ. To Interpret the SARS-CoV-2 Test, Consider the Cycle Threshold Value. *Clin Infect Dis*. 2020;71(16):2252–4.
36. Wei X, Ju X, Yi X, Zhu Q, Qu N, Liu T, et al. Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. *PLoS One*. 2011;6(12):e29500.
37. Ardui S, Ameer A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*. 2018;46(5):2159–68.
38. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*. 2017;6.
39. Weston S, Frieman MB. COVID-19: Knowns, Unknowns, and Questions. *mSphere*. 2020;5(2).
40. illumina. Advantages of paired-end and single-read sequencing 2021 [updated 2021. Available from: <https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html>.
41. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*. 2015;13(5):278–89.
42. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. 2021;592(7854):438–43.
43. Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. *bioRxiv*. 2020:2020.08.05.239046.
44. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*. 2020;5(11):1403–7.
45. Abernathy E, Glaunsinger B. Emerging roles for RNA degradation in viral replication and antiviral defense. *Virology*. 2015;479-480:600–8.

Figures

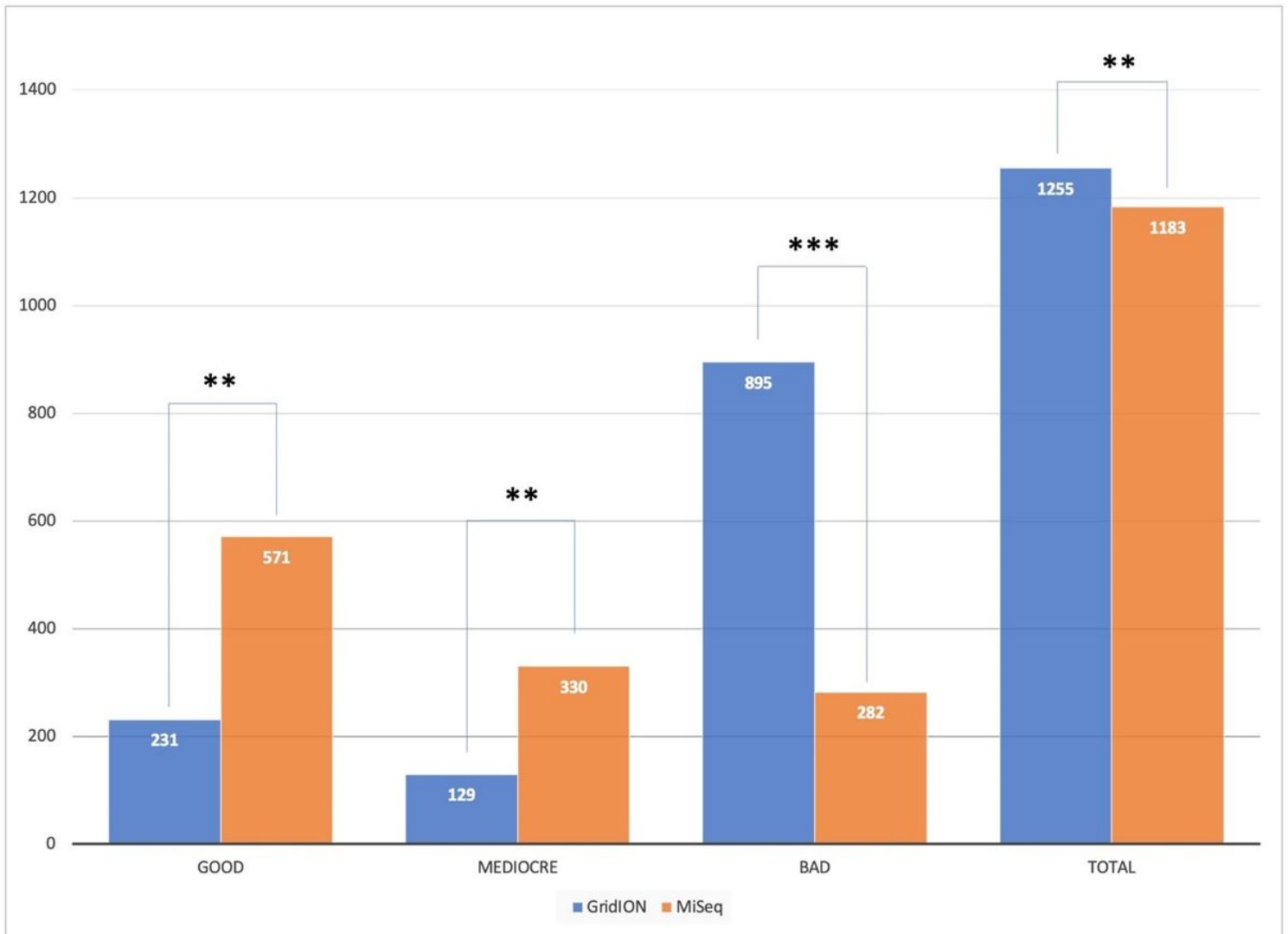


Figure 1

Comparison of sequence quality produced by the GridION and the MiSeq and analyzed on Nextclade: To compare the quality of sequences produced by the GridION and the MiSeq, consensus sequences produced by both platforms were uploaded to Nextclade and the results plotted on a double bar graph. Sequence quality was broken down into three groups; good, mediocre, and bad, with the GridION represented in blue and the MiSeq represented in orange. Statistical significance (Wilcoxon rank sum tests) is represented by “*” (**: $p < 0.01$, ***: $p < 0.001$).

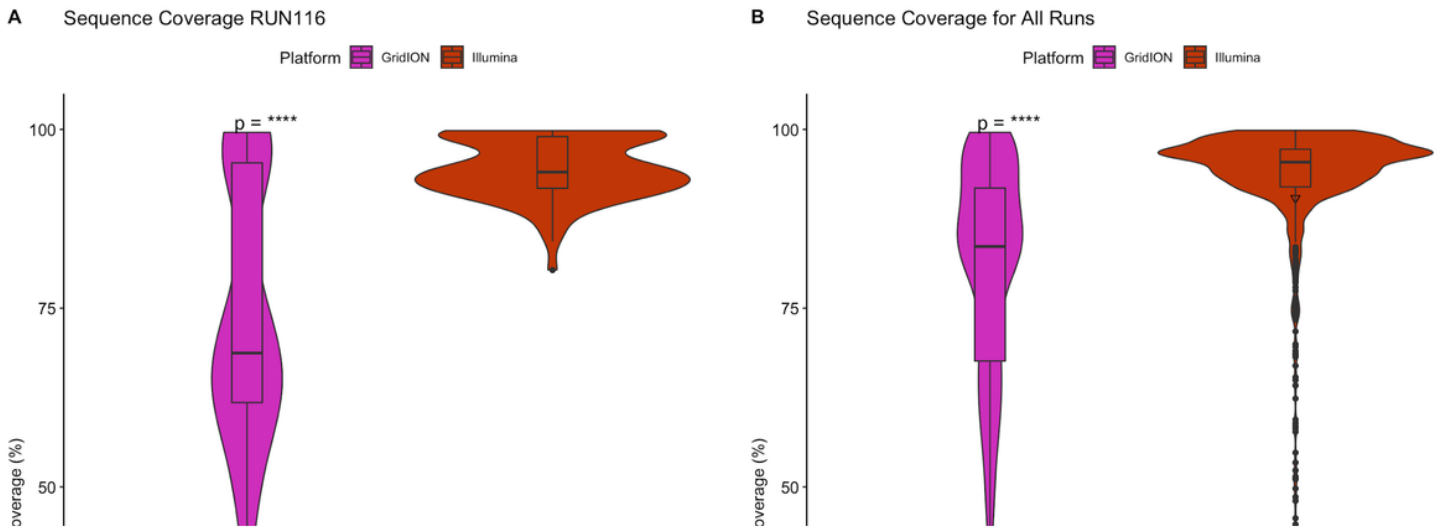


Figure 2

Comparison of GridION and MiSeq sequence coverage: Fastq files for RUN116 from both the MiSeq and the GridION were assembled using Genome Detective and the sequencing coverage was compared (A). The same was done for all sequences for both platforms (B). GridION samples are presented in purple, whilst Illumina MiSeq samples are presented in red. Statistical significance (Wilcoxon rank sum tests) is represented by “*” (****: $p < 0.0001$).

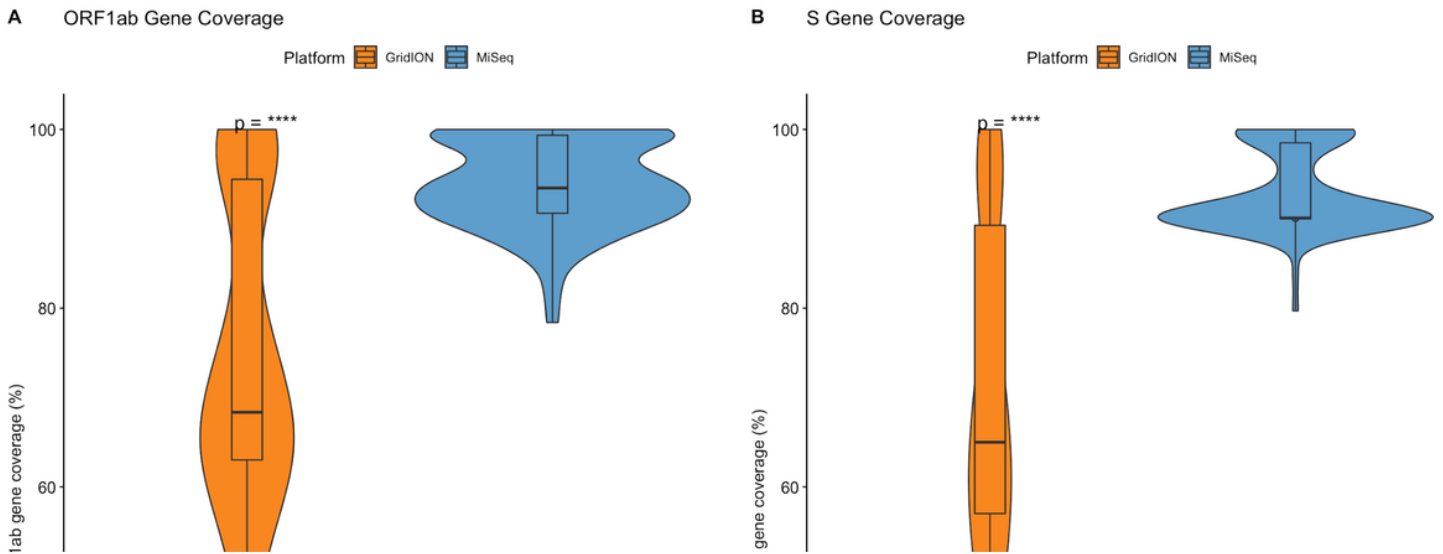


Figure 3

Comparison of ORF1ab- and S-gene coverage on the GridION and the MiSeq: Fastq files produced by both platforms were assembled on Genome Detective and the coverage for the ORF1ab- (A) and S-gene (B) was compared. Sequences from the GridION are represented in orange and sequences from the MiSeq are represented in blue. Statistical significance (Wilcoxon rank sum tests) is represented by “*” (****: $p < 0.0001$).

Figure 4

Correlation between sequence coverage and Ct score for samples sequenced on the GridION and MiSeq: A correlation was performed to determine the effect of Ct score on the sequence coverage obtained from the GridION and the MiSeq. Sequence coverage was plotted on the y-axis, whilst the sample’s average Ct score was plotted on the X-axis. GridION runs are represented by graphs A (Run101), B (Run111), and C (Run123), which are represented as green, blue, and red, respectively. MiSeq runs are represented by graphs D (Run100), E (Run109), and F (Run122) and are represented as black, purple, and gold, respectively. Statistical significance (Spearman’s rank correlation test) is represented by “*” (ns: non-

significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$). For both platforms, as the Ct score increased, there was a decrease in sequence coverage.

Figure 5

Correlation between the number of reads produced during sequencing and sample Ct Score: A correlation was performed for the number of reads produced by the GridION and the MiSeq and Ct score for SARS-CoV-2 samples. The number of reads was plotted on the Y-axis, whilst each sample's average Ct score was plotted on the X-axis. GridION runs are represented by graphs A (Run101), B (Run111), and C (Run123) and are shown as green, blue, and red, respectively. MiSeq runs are represented by graphs D (Run100), E (Run109), and F (Run122) and are shown as black, purple, and gold, respectively. Statistical significance (Spearman's rank correlation test) is represented by "*" (ns: non-significant, ****: $p < 0.0001$). An increase in Ct score resulted in a decrease in the number of reads produced for all GridION runs and 1 Illumina MiSeq run (Run122).

Figure 6

Analysis of mutations in samples sequenced on the GridION and the MiSeq: Consensus genomes produced by Genome Detective were uploaded to Nextclade and the results were analyzed. RUN116 was run on both platforms and the number and type of mutations detected by each platform was compared using a Wilcoxon rank sum test (**Figure 6-A and -C**). A consensus file for all runs, for each platform, was produced and uploaded to Nextclade and a Wilcoxon rank sum test was performed to compare the number and type of mutations detected by both platforms (**Figure 6-B and -D**). GridION samples are represented in yellow, whilst MiSeq samples are presented in green. Deletions, insertions, and substitutions are represented in pink, green, and blue, respectively.

Figure 7

Phylogenetic comparison between identical samples sequenced using both the GridION and MiSeq: A phylogenetic tree was created using IQTREE and visualized using FigTree for samples from Run116 sequenced on both the GridION and the MiSeq but classified in different clades by Nextclade. Only one of the 27 samples, represented in blue, clustered on the same branch. GridION sequences are annotated as 'barcode*', whilst MiSeq sequences are annotated as 'K0*'.