Bridging genomic gaps: A versatile SARS-CoV-2 benchmark dataset for adaptive laboratory workflows

Sara E. Zufan[1,2*], Louise M. Judd[1,3], Calum J. Walsh[1,3], Michelle L. Sait[4], Susan A. Ballard[4], Jason C. Kwong[5,6], Timothy P. Stinear[1,2], Torsten Seemann[1,2], Benjamin P. Howden[1,2,4]

1 Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
2 The Center for Pathogen Genomics, The University of Melbourne, Melbourne, Victoria, Australia
3 Doherty Applied Microbial Genomics, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
4 Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
5 Department of Infectious Diseases, Austin Health, Heidelberg, Victoria, Australia
6 Department of Infectious Diseases, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

* s.zufan@unimelb.edu.au

Abstract

Genomic sequencing's adoption in public health laboratories (PHLs) for pathogen surveillance is innovative yet challenging, particularly in the realm of bioinformatics. Low- and middle-income countries (LMICs) face increased difficulties due to supply chain volatility, workforce training, and unreliable infrastructure such as electricity and internet services. These challenges also extend to high-income countries (HICs) where bioinformatics is nascent in PHLs and hampered by a lack of specialized skills and computational infrastructure. This underlines the urgency for flexible and resource-aware strategies in genomic sequencing to improve global pathogen surveillance. In response to these challenges, the present research was conducted to identify and analyse key variables influencing the quality and accuracy of amplicon sequence data. An extensive benchmark dataset was developed that encompassed a diverse collection of isolates, viral loads, primer schemes, library preparation methods, sequencing technologies, and basecalling models, totalling 750 sequences. This dataset was analysed with bioinformatic workflows selected for varying levels of technical capacity. The evaluation focused on quality metrics, consensus accuracy, and common genomic epidemiological indicators. The analysis uncovers complex interactions between multiple parameters in laboratory and bioinformatic processes. emphasising resource-constrained PHLs, practical guidelines are proposed. Insights from the benchmark dataset aim to guide the establishment of specific laboratory and bioinformatics protocols for amplicon sequencing in these settings. The findings can also be used to guide the creation of specialised training curricula, further advancing genomic equity. The benchmark dataset itself allows laboratories to customise and evaluate workflows, catering to their distinct requirements and capacities. Such a holistic approach is imperative to build the capacity to monitor pathogens worldwide.

Author summary

This study marks a step toward equity in the field of pathogen genomics, especially for resource-constrained PHLs. It develops and evaluates a comprehensive amplicon sequencing benchmark dataset, offering vital insights for PHLs engaged in genomic surveillance. In particular, the study finds that the choice of basecaller model has a minimal impact on the quality and accuracy of consensus sequences derived from ONT data, which is crucial for labs with limited computational resources. It also highlights the effectiveness of longer amplicons in ensuring consistent coverage and reducing amplicon dropouts at higher viral loads. While Illumina remains a gold standard for data quality, the combination of the Midnight primer scheme with ONT's Rapid library preparation is shown to be a viable alternative, reducing costs, procedural complexity, and hands-on time. The study synthesises these findings into practical guidelines to aid in the development of amplicon sequencing workflows for SARS-CoV-2 with implications for other pathogens.

Introduction

The COVID-19 pandemic has demonstrated the transformative role of genomics in public health, transitioning from a predominantly academic tool to an essential component in pathogen surveillance and public health strategy. The integration of genomic data into public health frameworks has profoundly influenced global and regional surveillance of SARS-CoV-2 [1, 2], improving our understanding of the dynamics of virus transmission [3], evolutionary patterns [4], and the emergence of novel variants [5]. This information has been instrumental in shaping informed and effective public health responses [6].

The World Health Organization's (WHO) call for widespread genomic sequencing to strengthen disease surveillance highlights its critical role in global health [7, 8]. However, there is a stark contrast in the sequencing capacity between high-income countries (HICs), which report more sequences per capita, and low- and middle-income countries (LMICs), despite the commendable efforts in some African nations [9–11]. Moreover, the urgent demand for genomic sequencing in public health from the pandemic has exposed several challenges in all economic contexts. In LMICs, limited access to genomic technologies, lab infrastructure, supply chain reliability, training of the workforce, and data quality assurance prevail [12, 13]. In contrast, HIC PHLs face the need to expand their bioinformatics personnel and improve their computational infrastructure to handle the surge in genomic data [14]. These opposing sets of challenges emphasise the need for versatile and resilient genomic surveillance strategies that are effective in varying economic and resource contexts.

In addressing these challenges, the potential of benchmark datasets in public health genomics, particularly for quality assurance, should be explored. These datasets, which offer high-quality genomic sequences and annotations, are crucial for the validation of bioinformatic tools and methods used in outbreak surveillance [15, 16]. By engaging with these datasets, laboratories can critically evaluate and adapt sequencing workflows to their resource availability, establishing nuanced quality control parameters. However, existing benchmark datasets for SARS-CoV-2 often cover only a limited range of variables (e.g. primer scheme, sequencing platform). For example, recent interlaboratory validation studies have highlighted discrepancies in variant calling at mixed allele sites, such as those arising from intrahost variation, attributed to platform-specific bioinformatic workflows [17, 18]. The multitude of variables inherent in laboratory and bioinformatic workflows underscores the need for more comprehensive datasets that can more effectively address the wide range of challenges encountered in diverse global contexts.

This study provides a benchmark dataset for SARS-CoV-2 amplicon sequencing, reflecting a wide range of isolates, primer schemes, sequencing technologies, and bioinformatic approaches. Our analysis determines how different laboratory and bioinformatic workflows affect data quality and consistency. The insights gained aim to guide the formulation of evidence-based practices for effective amplicon sequencing across all PHLs, regardless of resource availability. This work supports the advancement of training initiatives around the world, promoting equitable genomic technology proficiency for enhanced global health surveillance.

Materials and methods

Sample preparation

This benchmark dataset's contrived specimens were sourced from five isolates, courtesy of the Victorian Infectious Disease Reference Laboratory (VIDRL) (S1 Table). Their isolation and preparation adhered to established protocols previously used for SARS-CoV-2 interlaboratory assessments [19, 20]. Briefly, the negative sample matrix was prepared by pooling 1 mL of viral transport medium (VTM) from 300 throat and deep nasal swabs, all previously confirmed negative by the Aptima® SARS-CoV-2 assay. This pooled VTM was divided into 10 aliquots and retested to verify negativity. Subsequently, 3 ml of this confirmed negative matrix was mixed with 55 µL of heat-killed viral stock in a sterile 5 mL tube, followed by inverting five times. A 500 µL sample was then transferred to a Panther Fusion specimen lysis tube for analysis using the Hologic Panther Fusion SARS-CoV-2 PCR assay. To simulate varying viral concentrations, the samples were serially diluted 1:10 in triplicate using the remaining negative matrix and quantified with the SARS-CoV-2 PCR E-gene assay.

Amplicon sequencing

Specimens were processed using multiplexed amplicon sequencing with the ARTIC V4 [21] and Midnight [22] primer schemes. Libraries for Illumina sequencing were prepared using the Nextera XT kit. For Nanopore sequencing, libraries were prepared using the ONT Native Barcoding Kit (EXP-NBD104) (hereafter referred to as Ligation) and the ONT Rapid Barcoding Kit (SQK-RBK004) (hereafter referred to as Rapid). Sequencing was performed on the Illumina iSeq and Nanopore GridION platforms, using MinION flow cells (S2 Table). A breakdown of the entire workflow is depicted in Fig. 1.

Bioinformatic workflows

Established methods commonly utilised in PHLs were employed for the generation of consensus sequences. Short-read data were processed using `iVar` software (version 1.4.2) [23], where primer trimming and consensus sequence generation were conducted. Parameters were set at a quality threshold of 20 (-q 20), a variant calling threshold of zero (-t 0), and a minimum depth of 10 (-m 10), conforming to standards widely accepted in the field [16]. This approach aligns with web-based pipelines [24] and the methods prevalent in PHLs [1], ensuring compatibility with current practices.

Nanopore reads were basecalled using `Guppy` (v6.2.11) super high accuracy (hereinafter referred to as 'Super') and fast (hereinafter referred to as Fast) models, representing variability in computational infrastructure. Basecalled, demultiplexed reads were processed using the `wf-artic` (v0.3.30) workflow [25]. This workflow, available through ONT's EPI2ME bioinformatics platform, offers both a graphical user interface

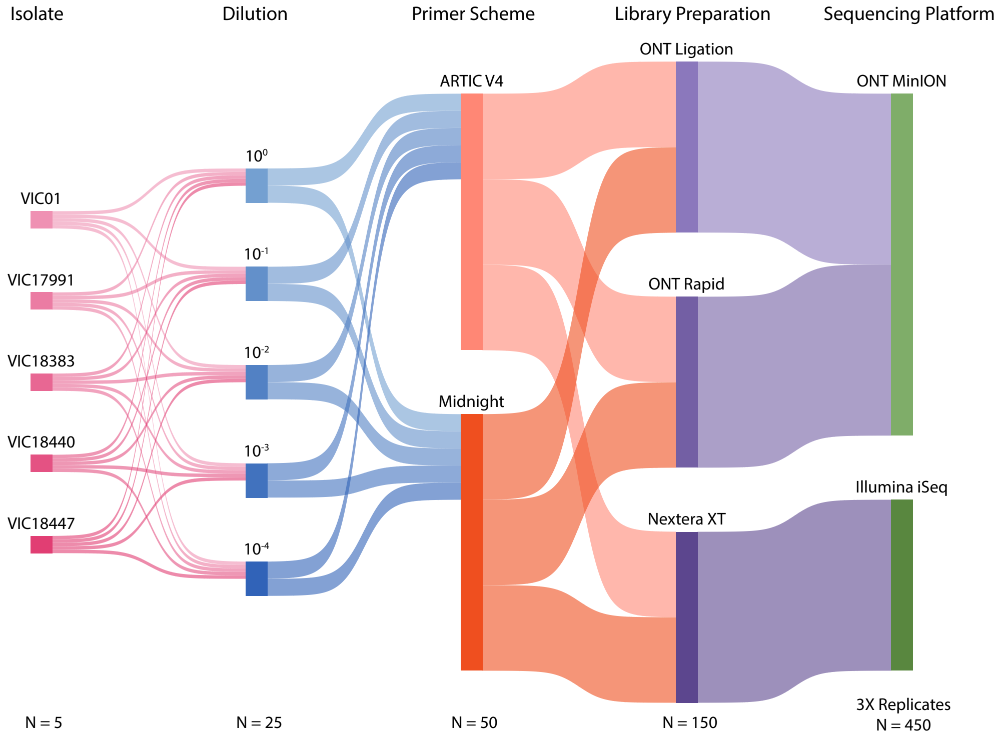

Fig 1. Sample preparation workflow. Sankey diagram depicting the sample preparation workflow for the SARS-CoV-2 amplicon sequencing benchmark dataset. Beginning with five distinct isolates, each undergoes serial dilution before proceeding through two primer schemes: ARTIC V4 and Midnight. Subsequent library preparation is conducted using ONT Ligation and ONT Rapid library preparation for sequencing on ONT MinION, and Nextera XT for the Illumina iSeq platform. Technical replicates were performed for all samples totaling 450 sequences. Additionally, positive and negative controls were processed concurrently.

(GUI) and command-line functionality. `wf-artic` is a derivative of the widely utilised `artic` pipeline [26], with notable modifications. These include optimised primer trimming for fragmented reads, a common occurrence in transposase-based rapid barcoding systems (`https://github.com/artic-network/fieldbioinformatics/issues/99`).

Performance evaluation

Quality metrics were collected using `ncov-tools` (v1.9.1) (`https://github.com/jts/ncov-tools`), `QualiMap` (v2.3) [27], and `nextclade` (v2.14.0) [28]. Metrics included read count, error rates, base quality, amplicon depth, breadth of coverage, lineage assignment, phylogenetic clustering, and consensus accuracy. Accuracy was evaluated by comparing the variants identified in specimen sequences with those in the primary isolate sequence (Table **??**). Within this framework, the variants detected in both the specimen and the primary isolate were classified as true positives (TP). Variants identified in the specimen but not in the primary isolate were considered false positives (FP). In contrast, variants identified in the primary isolate but absent in the specimen were deemed false negatives (FN). The following measures were used for statistical evaluation:

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{1}$$

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}. \tag{2}$$

$$\text{F1 score} = 2 \times \frac{(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}. \tag{3}$$

All subsequent statistical analyses were performed in `R` (v4.3.0) and `SciPy` (v1.11.4).

Results and discussion

Benchmark sample characteristics

The benchmark dataset was made up of contrived specimens $N = 75$, including two biological replicates in five lineages and dilutions, and one set of technical replicates performed in triplicate. TCID50 values ranged from 0.01 to 1000 (mean ($\bar{x}$=129.10); median ($\tilde{x}$=0.05); inter-quartile range (IQR)=9.97) (Fig. 2). Amplicon sequencing was performed using the ARTIC V4 and Midnight schemes, with library preparations from Nextera XT, ONT Ligation, and ONT Rapid kits, resulting in 450 samples for sequencing. Furthermore, ONT sequences were basecalled using Super and Fast models, totalling 750 sequences for comprehensive analysis.

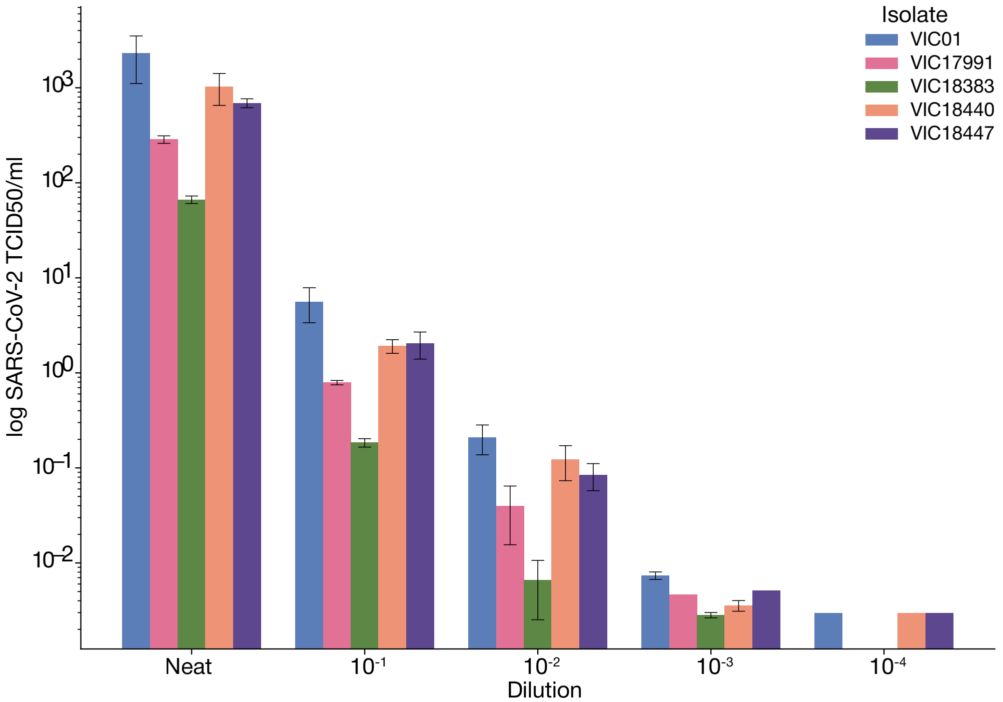

Fig 2. TCID50 estimates for contrived specimen. Quantitative assessment of SARS-CoV-2 viral titers across various dilutions and isolates prepared for this study. TCID50 values were extrapolated from Ct values obtained via the Panther Hologic fusion assay, guided by a predefined standard curve (S1 Fig.

Basecaller performance

Nanopore run summaries

Each scheme and library preparation pair were run on an individual MinION flow cell. Initial pore scans ranged from 371 (18.12%) to 1714 (83.70%) active pores (Fig. 3A). Among the Midnight scheme runs, Ligation achieved a basecall pass rate of 74.22% with a total output of 65.92 Mb (Fig. 3B). In contrast, Rapid obtained a pass rate of 64.00% and a total output of 886.17 Mb, reflecting the low initial pore activity (Fig. 3). With respect to ARTIC V4, Ligation showed a notable basecall pass rate of 85.87% with a total output of 712.11 Mb. Rapid had the lowest pass rate of 28.02%. Considering the high initial pore activity, the low pass rate may be indicative of poor library purity.

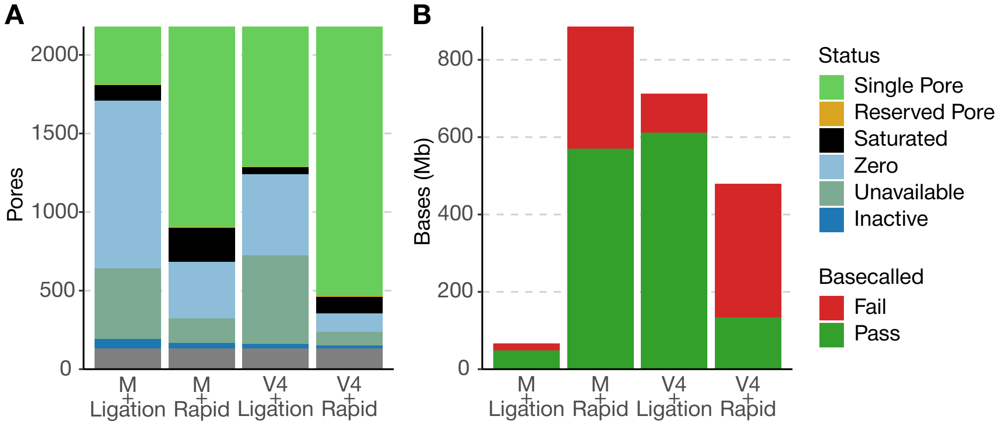

Fig 3. Comparative characteristics of Nanopore sequencing runs. A) Distribution of initial pore statuses across multiple runs, reflecting the operational efficiency of the flowcells used. B) Aggregate basecalling output for each run, categorized into bases that passed or failed the quality threshold (Q $\geq$ 7).

Basecaller performance

The ONT base caller models were evaluated for average base quality (avgBQ), mismatch count, and consensus accuracy. The Fast model showed a mean avgBQ of 25.09 with a median of 27.90 and an IQR of 10.90 (S2 FigC), while Super had a mean of 24.75, a median of 24.45, and a smaller IQR of 8.00, indicating a more consistent avgBQ. No significant difference in avgBQ was observed between Super and Fast (Wilcoxon rank sum, p = 9.654e-01). Stratification by library preparation (S2 FigB) revealed Rapid preparation improved avgBQ with Super ($\bar{x}$=20.75) over Fast ($\bar{x}$=17.78; Wilcoxon rank-sum, p=3.313e-34), unlike Ligation. Primer scheme analysis (S2 FigA) showed that with Midnight, Super's avgBQ ($\bar{x}$=25.12) exceeded Fast's ($\bar{x}$=23.41; Wilcoxon rank-sum, p=1.160e-04), whereas with ARTIC V4, Fast outperformed Super ($\bar{x}$=28.43 vs. $\bar{x}$=25.70; Wilcoxon rank-sum, p=6.709e-10). In particular, Midnight with Rapid was the only significant differential, favouring Super (Wilcoxon rank sum, p = 4.308e-26), aligning with the trend that Rapid avgBQ improves when using the Super model. Sparse Midnight with Ligation data likely influenced these scheme-specific divergences.

Despite variations in avgBQ across multiple strata, a regression analysis found no significant correlations between avgBQ and mismatches (S2 FigD) or accuracy (S2 FigE) for both Fast and Super models. Pearson's correlation coefficients (r) for Fast showed negligible relationships ((r = -0.03)) for mismatches and percent consensus accuracy ((r = -0.03)). Similarly, Super only exhibited weak correlations ((r = 0.21)) for mismatches and precision ((r = -0.08)).

Furthermore, no statistically significant differences were observed in the number of mismatches (Mann-Whitney U test, p = 0.899) or in the percentage of accuracy (Mann-Whitney U test, p=0.896) between the Super and Fast models. In particular, the Fast model, while maintaining comparable performance, offers computational efficiency, making it suitable for resource-limited environments. Subsequent analyses will primarily focus on the Fast model data, aligning with the study's objectives unless otherwise specified.

Benchmark dataset characteristics

Taking into account the importance of achieving a breadth of coverage $\geq$ 90% as a key QC threshold for the subsequent genomic epidemiological analysis [16, 18], the study evaluated consensus sequences against this standard. A total of 30 sequences met this criteria (6.67%). The Midnight primer scheme paired with Nextera XT library preparation attained the greatest genome completeness rate of 19.74% (n=15), closely followed by Midnight paired with Rapid, with 17.33% (n=13) (Fig. 4A). ARTIC V4 paired with Nextera XT showed comparable results, while ARTIC V4 paired with Ligation lag significantly, with only 2. 67% (n = 2) achieving the required coverage. Both Midnight paired with Ligation and ARTIC V4 paired with Rapid did not meet the $\geq$ 90% coverage threshold. Consequently, further analyses concentrated on the 10^0 dilution series, as only one sample met the breadth threshold in subsequent dilutions (S3 Table).

In this narrowed scope of the 10^0 dilution series, the Pass rate in `nextclade` QC status became a focal point. Here, Pass was defined as receiving an overall QC status of "good" or "mediocre". Both Midnight paired with Nextera XT and ARTIC V4 paired with Nextera XT demonstrated a Pass rate of 73.33%, indicating their efficiency. In contrast, Midnight paired with Rapid outperformed slightly with a Pass rate of 78.57%, whereas ARTIC V4 paired with Ligation exhibited a significantly lower rate of 13.33% (Fig. 4B).

Benchmark dataset performance

Breadth of coverage

The performance of various sequencing schemes and library preparation methods was evaluated, with a focus on the breadth of coverage at different dilution levels. The analyses were concentrated in dilutions 10^0 and 10^1 due to the low viral concentrations in the prepared sample.

In the 10^1 dilution analysis, the Midnight scheme paired with Nextera XT showed a high mean coverage of 76.06% ($\tilde{x}$=99.33%; IQR=48.02%), while Midnight with Rapid had a $\bar{x}$ coverage of 68.87% ($\tilde{x}$=85.12%; IQR=65.14%). ARTIC V4 paired with Nextera XT reported a $\bar{x}$ coverage of 65.38% ($\tilde{x}$=68.72%; IQR=62.74%), and with Ligation, it had the lowest $\bar{x}$ coverage of 47.27% ($\tilde{x}$=36.81%; IQR=71.04%).

At the 10^0 dilution, the results differed significantly. Midnight with Nextera XT showed a high and consistent coverage, with a $\bar{x}$ of 99.51% ($\tilde{x}$=99.52%; IQR=0.04%). Similarly, Midnight with Rapid also demonstrated high coverage ($\bar{x}$=98.63%; $\tilde{x}$=99.17%; IQR=0.18%) . In ARTIC V4 sequencing, Ligation exhibited a $\bar{x}$ coverage of 62.85% ($\tilde{x}$=74.40%, IQR=55.46%), while Nextera XT had a mean coverage of 88.69% ($\tilde{x}$=99.03%, IQR=1.23%).

The stratified analysis revealed significant variations in the breadth of coverage $\geq$ 90% between different primer schemes and library preparation methods. For example,

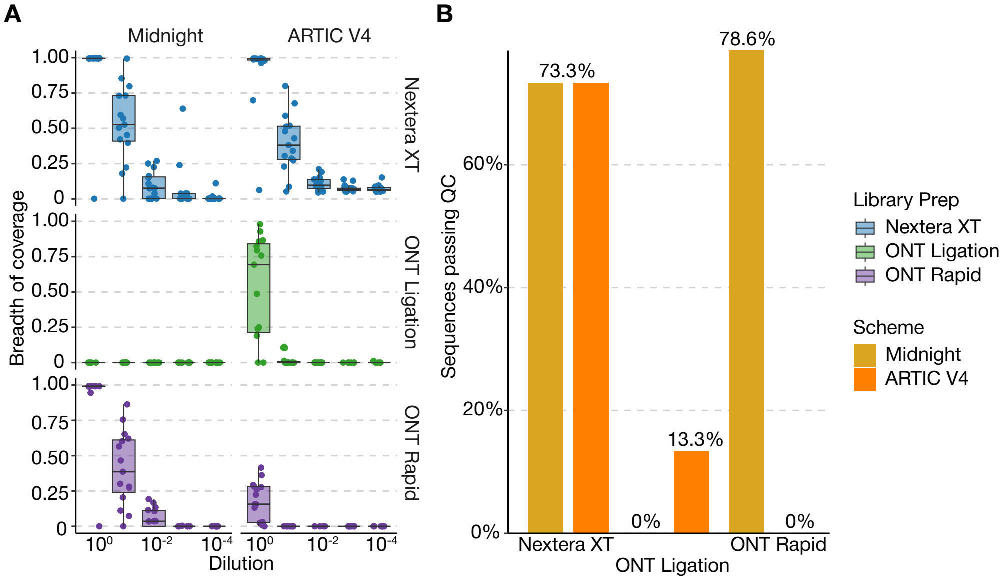

Fig 4. Characteristics of consensus sequences. Excluding samples processed with the the Super-basecalling model, which showed negligible influence on quality control (QC) metrics, panel A illustrates the breadth of coverage across serial dilutions. Here, individual data points denote the breadth of coverage for each sample, with boxplots providing a summary of the overall data distribution. Panel B examines the proportion of sequences that met the `nextclade` QC criteria, segregated by library preparation method and primer scheme, specifically focusing on the 10^0 dilution subset.

Midnight with Nextera XT achieved 14. 67% high-coverage samples (n = 11/75) in the Nextera XT and Rapid preparations, compared to only 2.67% (n=2/75) high-coverage samples for ARTIC V4 with Ligation.

In general, these findings emphasise the significant impact of primer schemes and library preparation methods in achieving the high breadth of coverage needed for accurate genomic epidemiological analysis of SARS-CoV-2. The Midnight scheme consistently showed high coverage across different library preparations, while ARTIC V4 displayed more variability. This is in agreement with Freed *et al.*, who found the length of the 1200 bp amplicon of Midnight to be optimal for uniform coverage across viral concentrations [22]. However, it should be noted that Midnight outperformed ARTIC V4 with higher viral loads ($\sim$TCID50 ≥ 5, or, $C\tau \leq 30$) (S3 Fig). Furthermore, dilution 10^0 generally led to higher and more consistent coverage, especially with Nextera XT, underscoring the importance of considering viral load when prioritising candidate samples for genomic sequencing.

Consensus sequence evaluation

Comparison of consensus sequence accuracy metrics across various primer schemes and library preparation methods indicated significant performance differences. The Midnight scheme, when used with Rapid library preparation, exhibited the highest metrics, demonstrating a recall of 59. 27%, precision of 92. 95%, and an F1 score of 65. 44%. In contrast, the combination of Midnight with Nextera XT displayed moderate performance, with a recall of 55. 36%, precision of 78. 48%, and an F1 score of 59. 03%.

The ARTIC V4 scheme showed more variability in its performance. Together with Ligation, it achieved a recall of 49. 83%, precision of 93. 33%, and an F1 score of 57.

57%. However, when ARTIC V4 was combined with Nextera XT, it resulted in the lowest performance metrics among the evaluated groups, with a recall of 29. 33%, precision of 43. 74%, and a F1 score of 31. 97%. This relatively lower performance of the ARTIC V4 scheme with Nextera XT, particularly in samples with higher Cvalues τ, aligns with the findings of Mboowa *et al.*, who reported that the ARTIC SARS-CoV-2 sequencing protocol on the Illumina MiSeq platform was sensitive and accurate to Cτ of 24 in Ugandan settings [29]. This context highlights the importance of selecting appropriate sequencing protocols based on the characteristics of the samples being analysed.

At the 10^0 dilution level, the Midnight scheme with Nextera XT achieved the highest scores, achieving perfect recall, precision, and F1-Score. The Midnight scheme with Rapid also showed high efficiency, with recall at 98.91%, precision at 100%, and F1-Score at 99.44%. ARTIC V4 with Ligation showed moderate results (recall = 66. 00%, precision = 100%, F1-Score = 72. 75%), with slight improvement when paired with Nextera XT (recall=88.73%, precision=91.67%, F1-Score=90.02%).

Combined, these results indicate that the Midnight scheme yields the greatest coverage and accuracy overall. Illumina continues to produce more accurate results, as has been continuously reported [30, 31]. However, read accuracy continues to improve with advances in ONT technology and basecalling models [32, 33].

Examining lineage assignment and phylogenetic clustering in our study reinforced the importance of stringent QC thresholds in genomic epidemiology. Despite variations in sequencing accuracy between different methods, 100% concordance was observed for both lineage assignment and phylogenetic clustering in all combinations with a coverage breadth greater than 80% (Fig. 5). Similarly, concordance was observed when `nextclade` designated overall QC status as ”good” or ”mediocre”. All pairs showing a steep decline in concordance when overall QC status was designated as ”bad” (S4 Table).

Limitations

Although it provides valuable information, this study is not without limitations. A primary challenge was the high Cτ values of the specimens, indicative of low viral concentrations. This aspect posed difficulties across all sequencing methods, impacting their efficiency and consequently limiting the number of sequences available for thorough analysis.

The study's challenges with two failed ONT runs, resulting in low yield and quality, limited the comprehensive evaluation of all primer scheme and library preparation combinations. This shortfall could bias interpretations due to missing data points. These issues, characteristic of ONT sequencing's variable quality and yield [34, 35], are particularly pertinent for resource-constrained PHLs in developing effective sequencing workflows. Despite these constraints, the study's findings are valuable, emphasising the need for adaptable sequencing methodologies to accommodate such inherent uncertainties in genomic sequencing.

This benchmark dataset, designed primarily for SARS-CoV-2, presents a flexible model suitable for various pathogens. The increasing use of multiplexed amplicon sequencing in pathogen surveillance, including Dengue [36], Mpox [37], and *Plasmodium falciparum* drug resistance and analysis of vaccine targets [38], underscores its importance. Our findings reveal a complex interplay of variables in the sequencing process, highlighting the need for meticulous quality assurance. Developing comprehensive benchmark datasets is crucial to ensure data quality in various settings, particularly in resource-constrained PHLs, and promotes genomic equity, thereby significantly improving global health surveillance.

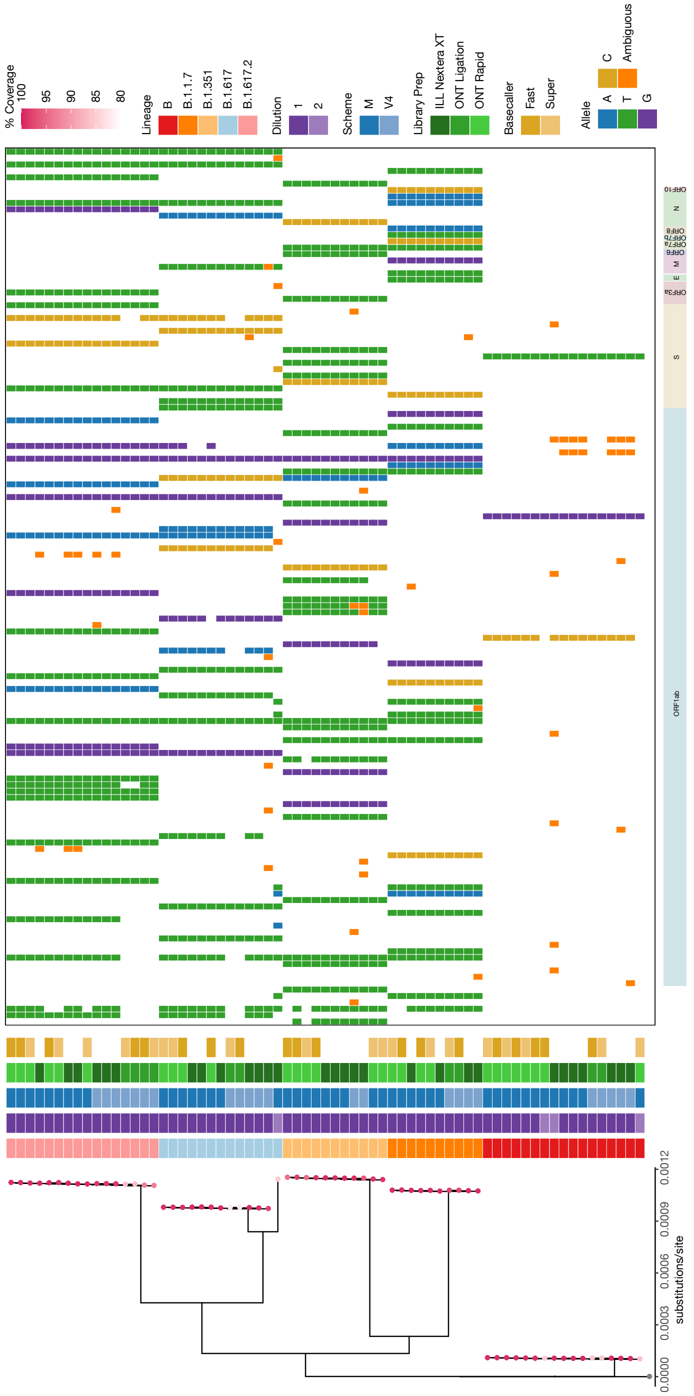

Fig 5. Clustering and characteristics of consensus sequences. Consensus sequences with breadth of coverage ≥ 80% were selected for phylogenetic analysis in with `ncov-tools`. Tip colors denote breadth of coverage, followed by columns of sequence characteristics. The internal table show variants called colored by allele along the genome.

Conclusion

The comprehensive evaluation of primer schemes and library preparation methods in this study, using a robust benchmark dataset, offers essential guidance for resource-limited PHLs. Key findings include the minimal impact of basecaller model selection on the quality and accuracy of consensus sequences, which is advantageous for laboratories with limited computational resources. The research further reveals the benefits of longer amplicons at elevated viral concentrations, such as more uniform coverage and reduced susceptibility to dropouts, thus decreasing the need for frequent monitoring and replacement. Although Illumina remains the gold standard for data quality, the combination of Midnight primer schemes with ONT Rapid library preparation achieved comparable metrics. A synthesis of the findings, along with the existing literature, is presented in Table 1. This collated guide is designed to support the development and evaluation of SARS-CoV-2 amplicon sequencing workflows and can be adopted for other pathogens. In general, this study significantly contributes to the advancement of global genomic surveillance equity and the strengthening of public health strategies in the rapidly evolving domain of genomic sequencing.

Table 1. Guidelines for developing amplicon sequencing workflows in resource-constrained settings. The resource intensity is denoted as high 🔴, medium 🟡, and low 🟢.

	Cost	**Infrastructure**	**Workforce**	**Summary**
Primer Scheme				
ARTIC V4	🟡	🟢	🟡	Short amplicons are more susceptible to dropouts, affecting sequencing consistency [39]. However, it shows greater sensitivity at high Cτ values (S3 Fig), ideal for degraded samples or low viral loads.
Midnight	🟢	🟢	🟢	Amplicons of this length yield uniform coverage, even at high Cτ values [22, 40]. Longer amplicons are less prone to dropouts, reducing the need for frequent monitoring and replacements.
Library Preparation				
Nextera XT	🔴	🟡	🟡	Despite its complexity, Nextera XT, an Illumina-based library preparation method, provides efficient end-to-end workflows like COVIDSeq [41]. This feature aids in simplifying processing from sample to data analysis, although it demands advanced equipment and infrastructure.
ONT Ligation	🔴	🟡	🟡	Offers the advantage of longer read lengths, crucial for structural variation analysis [30], and increased data yield (Fig 3B) [42, 43]. However, its complex procedure demands higher sample input and technical expertise.
ONT Rapid	🟢	🟢	🟢	Reduces turnaround time and costs [22], but may compromise library purity and read quality (Fig 3B), affecting accuracy, especially for low-frequency variants in low-yield samples.
Platform				
Illumina	🔴	🟡	🟢	Gold-standard for accurate, high-throughput viral sequencing [30], but entails high initial and operational costs, complex library preparation, and substantial infrastructure.

Continued on next page

Table 1 continued from previous page

	Cost	Infrastructure	Workforce	Summary
Nanopore	🟡	🟢	🟢	Offers portability, cost-effectiveness, and quick turnaround [30], suitable for remote testing [44–46]. Real-time sequencing capability aids rapid outbreak responses. However, sequencing data quality varies based on sample and preparation quality, impacting consistency (Fig 3B).
Basecalling				
Illumina	🟢	🔴	🟢	Conducted on-board, eliminating the need for external computational power and ensuring accuracy, but delays data availability until run completion.
Fast	🟢	🟢	🟢	Provides rapid, near-real-time basecalling even on standard laptops, producing lower quality reads that can still yield accurate consensus genomes with appropriate QC thresholds.
Super	🔴	🔴	🟡	Requires GPU for efficient processing, delivering higher quality reads. Can be performed on-board Mk1C, GridION, and PromethION devices, otherwise needs to be performed post-hoc using High Performance Computing (HPC) resources.
Analysis				
On-board	🟡	🟢	🟢	Illumina's Dragen platform offers cloud-based workflows, needing internet and incurring service fees. In contrast, ONT's EPI2ME, free and locally runnable, offers an alternative.
GUI	🟡	🟡	🟢	Sequence QC and assembly often rely on licensed or fee-based workflows like CLC Genomics and TheiCov, with varying access requirements. Free options like Galaxy exist but need internet access for remote servers.
CLI	🟢	🟡	🔴	Provides detailed control over parameters, demanding high technical expertise. While recent developments focus on reducing computational intensity, CLI typically operates on high-performance systems and is generally free.

Supporting information

S1 Table. Strains used in this study and their PANGO Lineage with WHO designations, if applicable. Primary specimens were previously sequenced and are available on GISAID. The dilution values represent the mean cycle threshold (Cτ) of the replicates of the specimen.

S2 Table. Flow cells used in ONT sequencing experiments.

S3 Table. Proportion of samples with coverage ≥ 0.9 stratified by primer scheme, library preparation method, and dilution. Each strata consisted of 15 contrived specimens.

S4 Table. Proportion of observations that meet the expected lineage (n = count of meets) by primer scheme, library preparation method, and QC status.

S1 Fig. Standard curve used to approximate TCID50 using the Hologic Panther Fusion SARS-CoV-2 PCR assay.

S2 Fig. Effect of ONT basecalling models across variables. A) Overall mean avgBQ and average mapping quality (avgMQ) by schemes; B) avgBQ and avgMQ stratified by ONT library preparation and basecalling model; C) Distribution of avgBQ by ONT basecalling model; D) Linear relationship between avgBQ and mismatches by ONT basecalling model; E) Linear relationship between avgBQ and consensus sequence accuracy by ONT basecalling model.

S3 Fig. Linear relationship between breadth of coverage and $C\tau$ for two high performing variables. This plot illustrates the relationship between breadth of coverage and $C\tau$ values for the Midnight paired with Rapid (in purple) and ARTIC V4 with Nextera XT (in blue). The regression line for the Midnight with Rapid scheme exhibits a slope of -0.0887, indicating a substantial negative correlation where coverage significantly decreases with increasing $C\tau$ values. In contrast, the ARTIC V4 with Nextera XT scheme shows a less steep slope of -0.0575, suggesting a more moderate negative correlation. Notably, the Midnight with Rapid scheme demonstrates greater breadth of coverage when $C\tau \geq 30$, whereas ARTIC V4 with Nextera XT is more effective in achieving higher coverage when $C\tau < 30$. Samples with 'Undetermined' $C\tau$ values are excluded from this analysis.

Acknowledgments

We acknowledge the support of the National Health and Medical Research Council, Australia, through their Partnership Grant (GNT1149991), and the Australian Government's Medical Research Future Fund under the Genomics Health Futures Mission Flagships—Pathogen Genomics Grant (FSPGN000045) for this work. Additionally, S.E.Z. is a recipient of the Melbourne Research Scholarship from the University of Melbourne. Our gratitude extends to the Victorian Infectious Disease Reference Laboratory for supplying the viral isolates used in this study.

References

1. Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. Nature communications. 2020;11(1):4376. doi:10.1038/s41467-020-18314-x.
2. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global SARS-CoV-2 genomic surveillance: What we have learned (so far). Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2023;108. doi:10.1016/j.meegid.2023.105405.
3. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 2022;603(7902):679–+. doi:10.1038/s41586-022-04411-y.
4. Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell. 2021;184(20):5189–5200.e7. doi:10.1016/j.cell.2021.09.003.
5. 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–+. doi:10.1038/s41586-021-03402-9.
6. Saravanan KA, Panigrahi M, Kumar H, Rajawat D, Nayak SS, Bhushan B, et al. Role of genomics in combating COVID-19 pandemic. Gene. 2022;823:146387. doi:10.1016/j.gene.2022.146387.
7. Carter LL, Yu MA, Sacks JA, Barnadas C, Pereyaslov D, Cognat S, et al. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022-2032. Bulletin of the World Health Organization. 2022;100(4):239–+. doi:10.2471/blt.22.288220.
8. Akande OW, Carter LL, Abubakar A, Achilla R, Barakat A, Gumede N, et al. Strengthening pathogen genomic surveillance for health emergencies: insights from the World Health Organization's regional initiatives. Frontiers in Public Health. 2023;11. doi:10.3389/fpubh.2023.1146730.
9. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. Nature communications. 2022;13(1):7003. doi:10.1038/s41467-022-33713-y.
10. Sawadogo Y, Galal L, Belarbi E, Zongo A, Schubert G, Leendertz F, et al. Genomic Epidemiology of SARS-CoV-2 in Western Burkina Faso, West Africa. Viruses-Basel. 2022;14(12). doi:10.3390/v14122788.
11. Sisay A, Tshiabuila D, van Wyk S, Tesfaye A, Mboowa G, Oyola SO, et al. Molecular Epidemiology and Diversity of SARS-CoV-2 in Ethiopia, 2020-2022. Genes. 2023;14(3). doi:10.3390/genes14030705.
12. Inzaule SC, Tessema SK, Kebede Y, Ogwell Ouma AE, Nkengasong JN. Genomic-informed pathogen surveillance in Africa: opportunities and challenges. The Lancet infectious diseases. 2021;21(9):e281–e289. doi:10.1016/S1473-3099(20)30939-7.
13. Yek C, Pacheco AR, Vanaerschot M, Bohl JA, Fahsbender E, Aranda-Díaz A, et al. Metagenomic Pathogen Sequencing in Resource-Scarce Settings: Lessons Learned and the Road Ahead. Frontiers in epidemiology. 2022;2. doi:10.3389/fepid.2022.926695.

14. Nadon C, Croxen M, Knox N, Tanner J, Zetner A, Yoshida C, et al. Public health genomics capacity assessment: readiness for large-scale pathogen genomic surveillance in Canada's public health laboratories. BMC public health. 2022;22(1):1817. doi:10.1186/s12889-022-14210-9.

15. Timme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. PeerJ. 2017;5:e3893. doi:10.7717/peerj.3893.

16. Xiaoli L, Hagey JV, Park DJ, Gulvik CA, Young EL, Alikhan NF, et al. Benchmark datasets for SARS-CoV-2 surveillance bioinformatics. PeerJ. 2022;10:e13821. doi:10.7717/peerj.13821.

17. Foster MA, Hofmeister MG, Yin S, Montgomery MP, Weng MK, Eckert M, et al. Widespread Hepatitis A Outbreaks Associated with Person-to-Person Transmission - United States, 2016-2020. MMWR Morbidity and mortality weekly report. 2022;71(39):1229–1234. doi:10.15585/mmwr.mm7139a1.

18. Zufan SE, Lau KA, Donald A, Hoang T, Foster CSP, Sikazwe C, et al. Bioinformatic investigation of discordant sequence data for SARS-CoV-2: insights for robust genomic analysis during pandemic surveillance. Microbial Genomics. 2023;9(11). doi:10.1099/mgen.0.001146.

19. Caly L, Druce J, Roberts J, Bond K, Tran T, Kostecki R, et al. Isolation and rapid sharing of the 2019 novel coronavirus (SARS-CoV-2) from the first patient diagnosed with COVID-19 in Australia. The Medical journal of Australia. 2020;212(10):459–462. doi:10.5694/mja2.50569.

20. Lau KA, Horan K, da Silva AG, Kaufer A, Theis T, Ballard SA, et al. Proficiency testing for SARS-CoV-2 whole genome sequencing. Pathology. 2022;54(5):615–622. doi:10.1016/j.pathol.2022.04.002.

21. Quick J. nCoV-2019 sequencing protocol (LoCost); 2020. Available from: https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bp2l6n26rgqe/v3.

22. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biology methods & protocols. 2020;5(1):bpaa014. doi:10.1093/biomethods/bpaa014.

23. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome biology. 2019;20(1):8. doi:10.1186/s13059-018-1618-7.

24. Maier W. sars-cov-2-pe-illumina-artic-variant-calling/COVID-19-PE-ARTIC-ILLUMINA;. https://workflowhub.eu/workflows/110. Available from: https://workflowhub.eu/workflows/110.

25. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nature biotechnology. 2020;38(3):276–278. doi:10.1038/s41587-020-0439-x.

26. fieldbioinformatics: The ARTIC field bioinformatics pipeline;. Available from: https://github.com/artic-network/fieldbioinformatics.

27. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics (Oxford, England). 2016;32(2):292–294. doi:10.1093/bioinformatics/btv566.

28. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of open source software. 2021;6(67):3773. doi:10.21105/joss.03773.

29. Mboowa G, Mwesigwa S, Kateete D, Wayengera M, Nasinghe E, Katagirya E, et al. Whole-genome sequencing of SARS-CoV-2 in Uganda: implementation of the low-cost ARTIC protocol in resource-limited settings. F1000Research. 2021;10:598. doi:10.12688/f1000research.53567.1.

30. 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. Nature communications. 2020;11(1). doi:10.1038/s41467-020-20075-6.

31. Tshiabuila D, Giandhari J, Pillay S, Ramphal U, Ramphal Y, Maharaj A, et al. Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq. BMC genomics. 2022;23(1). doi:10.1186/s12864-022-08541-5.

32. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome biology. 2018;19(1):90. doi:10.1186/s13059-018-1462-9.

33. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Computational and structural biotechnology journal. 2023;21:2352–2364. doi:10.1016/j.csbj.2023.03.038.

34. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature protocols. 2017;12(6):1261–1276. doi:10.1038/nprot.2017.066.

35. Paden CR, Tao Y, Queen K, Zhang J, Li Y, Uehara A, et al. Rapid, Sensitive, Full-Genome Sequencing of Severe Acute Respiratory Syndrome Coronavirus 2. Emerging infectious diseases. 2020;26(10):2401–2405. doi:10.3201/eid2610.20.1800.

36. Vogels CBF, Hill V, Breban MI, Chaguza C, Paul LM, Sodeinde A, et al. DengueSeq: A pan-serotype whole genome amplicon sequencing protocol for dengue virus. medRxiv : the preprint server for health sciences. 2023;doi:10.1101/2023.10.13.23296997.

37. Chen NFG, Chaguza C, Gagne L, Doucette M, Smole S, Buzby E, et al. Development of an amplicon-based sequencing approach in response to the global emergence of mpox. PLoS biology. 2023;21(6):e3002151. doi:10.1371/journal.pbio.3002151.

38. Girgis ST, Adika E, Nenyewodey FE, Senoo Jnr DK, Ngoi JM, Bandoh K, et al. Drug resistance and vaccine target surveillance of Plasmodium falciparum using nanopore sequencing in Ghana. Nature microbiology. 2023;8(12):2365–2377. doi:10.1038/s41564-023-01516-6.

39. Davis JJ, Long SW, Christensen PA, Olsen RJ, Olson R, Shukla M, et al. Analysis of the ARTIC Version 3 and Version 4 SARS-CoV-2 Primers and Their Impact on the Detection of the G142D Amino Acid Substitution in the Spike Protein. Microbiology spectrum. 2021;9(3):e0180321. doi:10.1128/Spectrum.01803-21.

40. Brejová B, Boršová K, Hodorová V, Čabanová V, Gafurov A, Fričová D, et al. Nanopore sequencing of SARS-CoV-2: Comparison of short and long PCR-tiling amplicon protocols; 2021. Available from: `http://medrxiv.org/lookup/doi/10.1101/2021.05.12.21256693`.

41. Pillay S, San JE, Tshiabuila D, Naidoo Y, Pillay Y, Maharaj A, et al. Evaluation of miniaturized Illumina DNA preparation protocols for SARS-CoV-2 whole genome sequencing. PloS one. 2023;18(4):e0283219. doi:10.1371/journal.pone.0283219.

42. González-Recio O, Gutiérrez-Rivas M, Peiró-Pastor R, Aguilera-Sepúlveda P, Cano-Gómez C, Jiménez-Clavero MÁ, et al. Sequencing of SARS-CoV-2 genome using different nanopore chemistries. Applied microbiology and biotechnology. 2021;105(8):3225–3234. doi:10.1007/s00253-021-11250-w.

43. Wick RR, Judd LM, Wyres KL, Holt KE. Recovery of small plasmid sequences via Oxford Nanopore sequencing. Microbial genomics. 2021;7(8). doi:10.1099/mgen.0.000631.

44. Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome medicine. 2016;8(1):97. doi:10.1186/s13073-016-0356-2.

45. Steinig E, Duchêne S, Aglua I, Greenhill A, Ford R, Yoannes M, et al. Phylodynamic Inference of Bacterial Outbreak Parameters Using Nanopore Sequencing. Molecular biology and evolution. 2022;39(3). doi:10.1093/molbev/msac040.

46. Lambisia AW, Mudhune GH, Morobe JM, Mohammed KS, Makori TO, Ndwiga L, et al. Temporal distribution and clinical characteristics of the Alpha, Delta and Omicron SARS-CoV-2 variants of concern in Laikipia, Kenya: institutional and community-based genomic surveillance. Wellcome Open Research. 2023;7(235):235.