

RESEARCH

Open Access



DengueSeq: a pan-serotype whole genome amplicon sequencing protocol for dengue virus

Chantal B. F. Vogels^{1,2*†}, Verity Hill^{1†}, Mallery I. Breban¹, Chrispin Chaguza^{1,2}, Lauren M. Paul³, Afeez Sodeinde¹, Emma Taylor-Salmon^{1,4}, Isabel M. Ott¹, Mary E. Petrone^{1,5}, Dennis Dijk⁶, Marcel Jonges⁶, Matthijs R. A. Welkers^{6,7}, Timothy Locksmith⁸, Yibo Dong⁹, Namratha Tarigopula⁹, Omer Tekin⁹, Sarah Schmedes⁹, Sylvia Bunch⁸, Natalia Cano⁸, Rayah Jaber⁸, Charles Panzera⁸, Ian Stryker⁸, Julieta Vergara⁸, Rebecca Zimler¹⁰, Edgar Kopp⁸, Lea Heberlein⁸, Kaylee S. Herzog¹¹, Joseph R. Fauver¹¹, Andrea M. Morrison¹⁰, Scott F. Michael³ and Nathan D. Grubaugh^{1,2,12,13*}

Abstract

Background The increasing burden of dengue virus on public health due to more explosive and frequent outbreaks highlights the need for improved surveillance and control. Genomic surveillance of dengue virus not only provides important insights into the emergence and spread of genetically diverse serotypes and genotypes, but it is also critical to monitor the effectiveness of newly implemented control strategies. Here, we present DengueSeq, an amplicon sequencing protocol, which enables whole-genome sequencing of all four dengue virus serotypes.

Results We developed primer schemes for the four dengue virus serotypes, which can be combined into a pan-serotype approach. We validated both approaches using genetically diverse virus stocks and clinical specimens that contained a range of virus copies. High genome coverage (>95%) was achieved for all genotypes, except DENV2 (genotype VI) and DENV 4 (genotype IV) sylvatics, with similar performance of the serotype-specific and pan-serotype approaches. The limit of detection to reach 70% coverage was 10-100 RNA copies/μL for all four serotypes, which is similar to other commonly used primer schemes. DengueSeq facilitates the sequencing of samples without known serotypes, allows the detection of multiple serotypes in the same sample, and can be used with a variety of library prep kits and sequencing instruments.

Conclusions DengueSeq was systematically evaluated with virus stocks and clinical specimens spanning the genetic diversity within each of the four dengue virus serotypes. The primer schemes can be plugged into existing amplicon sequencing workflows to facilitate the global need for expanded dengue virus genomic surveillance.

Keywords Genomic surveillance, Next-generation sequencing, Amplicon sequencing, Whole-genome sequencing, Dengue virus

[†]Chantal B. F. Vogels and Verity Hill are co-first authors.

*Correspondence:

Chantal B. F. Vogels
chantal.vogels@yale.edu

Nathan D. Grubaugh
nathan.grubaugh@yale.edu

Full list of author information is available at the end of the article



Background

The global burden of the mosquito-borne dengue virus continues to increase with an estimated 3.9 billion people at risk of infection [1]. During the last two decades, the number of reported dengue cases has increased by more than eightfold to over 4 million in 2019, particularly in tropical and sub-tropical regions in the Americas and Asia [2]. As of July 2023, close to 3 million dengue cases have been reported from the Americas this year, which has already surpassed the total number of 2.8 million cases reported in 2022 [3]. With global dengue cases surging, there is a critical need for enhanced genomic surveillance to track the emergence and spread of dengue virus lineages [4]. These lineages consist of four antigenically distinct serotypes (DENV1-4) which can be further subdivided into genetically diverse genotypes [5, 6]. Additionally, the rollout of novel vaccines and other biological interventions, such as the release of mosquitoes carrying the virus-inhibiting *Wolbachia* bacterium, warrant close monitoring of changes in dengue virus genetic diversity that may impact or reflect the effectiveness of these strategies [7, 8].

Amplicon-based sequencing is a cost-effective, sensitive, and high-throughput method to conduct routine virus genomic surveillance. During the 2015-2016 Zika epidemic, challenges with sequencing the often low-titer virus using metagenomic protocols led to the development of PrimalSeq. PrimalSeq is a multiplexed tiled primer approach where overlapping amplicons are split into two PCR reactions [9, 10]. Later, this approach was adapted as the primary sequencing method for SARS-CoV-2 (i.e., the “ARTIC” protocol) [11]. During the COVID-19 pandemic, investments in genomics infrastructure have significantly increased sequencing capacity globally, leading to a record number of over 16 million publicly available SARS-CoV-2 genomes. In contrast, less than 13,000 near-complete dengue virus genomes are publicly available on GenBank, and of those, only ~20% were collected in the last five years [12, 13]. Whole genome sequencing, rather than sequencing of targeted genes, is required to provide sufficient resolution for phylogenetic analyses, especially with more slowly evolving viruses such as dengue [14]. Moreover, a pan-serotype approach is needed to facilitate the sequencing of any serotype without the need for prior serotyping. Thus,

designing a unified pan-serotype approach to sequence the global diversity of dengue virus, which fits within the established SARS-CoV-2 workflows, could rapidly accelerate the implementation of dengue virus genomic surveillance systems [15].

Here, we describe the development and validation of DengueSeq to sequence the known diversity of dengue virus from clinical samples. We designed primer schemes for all four dengue virus serotypes, and the serotype-specific primers can be used individually or combined as a universal pan-serotype amplicon-based sequencing approach. By adapting DengueSeq to currently existing amplicon-based sequencing workflows, our method can enhance the global capacity for whole-genome dengue virus sequencing.

Results

Primer scheme design

The PrimalSeq protocol uses a series of overlapping PCR amplicons generated with primers designed using PrimalScheme (<https://primalscheme.com>) to efficiently sequence virus genomes directly from clinical specimens [9, 10]. Here, we used PrimalScheme to develop primer schemes for all four serotypes of dengue virus, which can be used with the amplicon-based sequencing protocols widely established during the COVID-19 pandemic [15]. We selected genetically diverse genomes within each of the dengue virus serotypes and then used PrimalScheme to generate four serotype-specific primer schemes (Fig. 1A). We chose genomes from all of the defined genotypes spanning a range of years (DENV1: 2007-2017, DENV2: 1992-2018, DENV3: 1956-2016, DENV4: 2007-2021; Additional file 1) to ensure that the primer schemes captured the genetic diversity within each of the dengue virus serotypes (Fig. 1B). Specifically, we focused primarily on including viruses that have caused human outbreaks, and excluded genotypes with only incomplete genomes available and highly divergent genotypes (e.g. DENV1 genotype II, DENV2 genotype IV and VI (sylvatic), and DENV4 genotypes III and IV (sylvatic)). Each serotype primer scheme consists of 35-37 primer pairs with an average amplicon length of 400 bp (the default PrimalScheme amplicon length [9]). We also developed a bioinformatics pipeline that aligns sequencing reads against six default reference genomes (DENV1-4, and

(See figure on next page.)

Fig. 1 Overview of the DengueSeq primer design. **A** For each of the four serotypes, a selection of genetically diverse genomes was made as input for PrimalScheme. Using PrimalScheme, four serotype-specific primer schemes were generated. Each scheme consists of 35-37 primer pairs which are split into two pools. Pan-serotype primer pools can be created by combining the serotype-specific pools 1 and 2, with which all four serotypes can be sequenced. Figure created with BioRender.com. **B** Dengue virus genomes representing the genotypes within each of the four serotypes used to design the primer schemes. Red circles and stars indicate the genomes used to design the primer schemes for each serotype

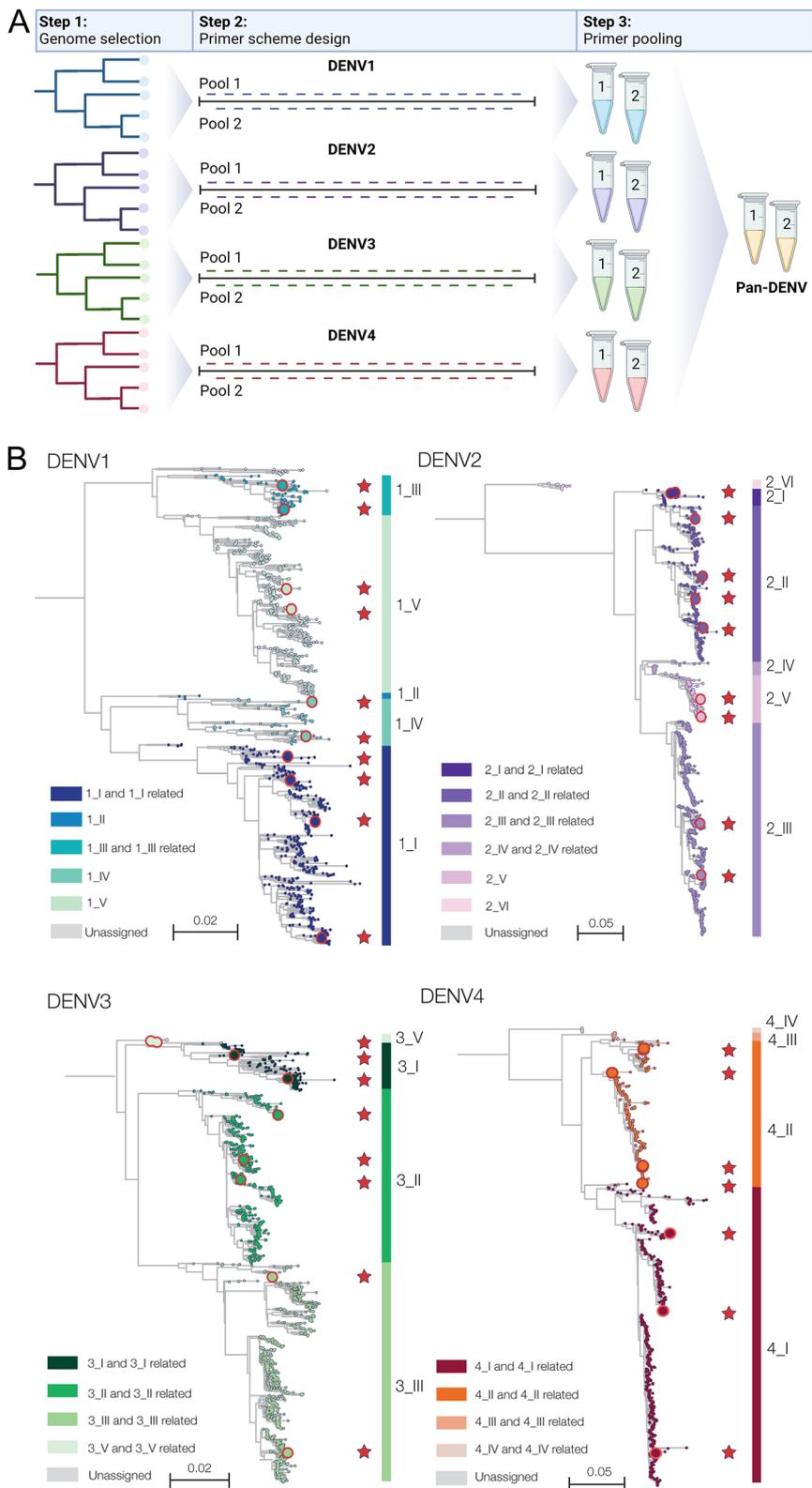


Fig. 1 (See legend on previous page.)

DENV2 and DENV4 sylvatics) through an iterative (non-competitive) loop to generate consensus genomes using iVar (Fig. S1; (https://github.com/grubaughlab/DENV_pipeline/tree/v1.0)). Our assay, called DengueSeq, can be used individually with any of the four serotype-specific primer sets or combined into a pan-serotype approach, as validated below.

To account for high virus genetic diversity, we tested whether including degenerate nucleotides in primers at positions with mismatches to some viruses would help amplify a wider range of genetic diversity. For example, if some viruses had an adenine (A) at a primer position, and others had a guanine (G), then designing a primer with an “R” code would direct for some primers to be randomly synthesized with either an A or G nucleotide at that position. To make the alternate design, we aligned the primer sequences to all of the publicly available dengue virus genomes, and we included degenerate nucleotides at positions with alternate nucleotide frequencies above 10%. After comparing the “non-degenerate (original)” and “degenerate” DENV2 primer schemes, we found similar genome coverage when sequencing undiluted and diluted virus stocks (Fig. S2). As we found minimal differences between the two designs, we proceeded with further validation of the “non-degenerate” DENV1-4 primer schemes.

Serotype-specific primer scheme validation

To perform our initial validation of the individual serotype-specific DengueSeq primer schemes, we sequenced 127 dengue virus stocks representing all of the defined genotypes that have caused human outbreaks across a broad temporal range (DENV1: 1944-2016, DENV2: 1959-2009, DENV3: 1966-2015, DENV4: 1961-2011; Fig. 2, Fig. S3, Additional file 1). Based on results from the CDC dengue multiplex RT-qPCR assay [16], we sequenced undiluted RNA using the corresponding serotype-specific primer scheme and the Illumina COVID-Seq test. We used our bioinformatics pipeline to generate consensus genomes at a minimum depth of coverage of 20x and Genome Detective to assign genotypes [17]. We generated genomes with high genome coverage (>95%) in the coding sequence for all genotypes (Fig. 2A, C, E, G), except for the sylvatic genotypes (DENV2 genotype

VI (60.0-91.2%) and DENV4 genotype IV (93.3-94.2%); Fig. 2C, G) that were not included in the primer design. These findings demonstrated that our serotype-specific primer schemes capture the genetic diversity present within each of the serotypes.

Next, we validated the DengueSeq primer schemes with clinical specimens to determine how they would work with various levels of sample quality and virus copies. We tested 600 sera samples of dengue-infected individuals from the Florida Department of Health, mostly from travelers, that also included many of the defined genotypes (DENV1: 2010-2022, DENV2: 2010-2022, DENV3: 2016-2023, DENV4: 2012-2022, Fig. 2, Fig. S4, Additional file 1). Before sequencing, we determined the serotype for each sample using the CDC dengue multiplex RT-qPCR assay. We then selected the corresponding serotype-specific primer scheme for sequencing. For all four serotypes, we generated near-complete genomes for samples with high virus copies ($\sim 10^4$ copies per μL or higher), with a clear drop-off in genome coverage with decreasing virus copies ($\sim 10^2$ copies per μL ; Fig. 2B, D, F, H). Importantly, we noticed that several dengue virus genomes had mismatches with the CDC dengue multiplex RT-qPCR primers and probes, resulting in decreased PCR efficiency and inaccurate measurements of virus copies. The impact of these mismatches was evident for DENV2 and DENV4, where we could often sequence viruses that were undetected by the RT-qPCR assay, and makes it difficult to set a threshold for including samples for sequencing using this assay. However, we show that our serotype-specific primer schemes can be used to successfully sequence dengue virus from clinical specimens.

Pan-serotype primer scheme validation

The majority of clinical diagnostics used for dengue virus do not provide information on the serotype [18, 19]. Thus, we developed a pan-serotype version of DengueSeq to facilitate whole-genome sequencing of clinical specimens of unknown serotypes for streamlined high-throughput sequencing. By combining each of the four serotype-specific primer pools 1 and 2 at an increased concentration of 20 μM to account for a dilution effect, we created the two pan-serotype primer pools 1 and 2 (Fig. 1A). We resequenced a subset of undiluted

(See figure on next page.)

Fig. 2 Percent genome coverage at a depth of coverage of 20x for dengue virus stocks and clinical specimens. A/C/E/G) Initial validation of the serotype-specific approach was validated by sequencing virus stocks representing the majority of genotypes within each of the four serotypes. Dengue virus 2 genotype VI and dengue virus 4 genotype IV comprise sylvatic viruses, which were the only genotypes for which genome coverage was consistently below 100%. B/D/F/H) Additional sequencing was done from clinical specimens of patients diagnosed with dengue. Specimens comprised a wider range of RNA titers, and lower genome coverage with lower RNA copies. Phylogenetic trees showing the genome sequences used for validation are shown in Fig. S3 (virus stocks) and Fig. S4 (clinical specimens)

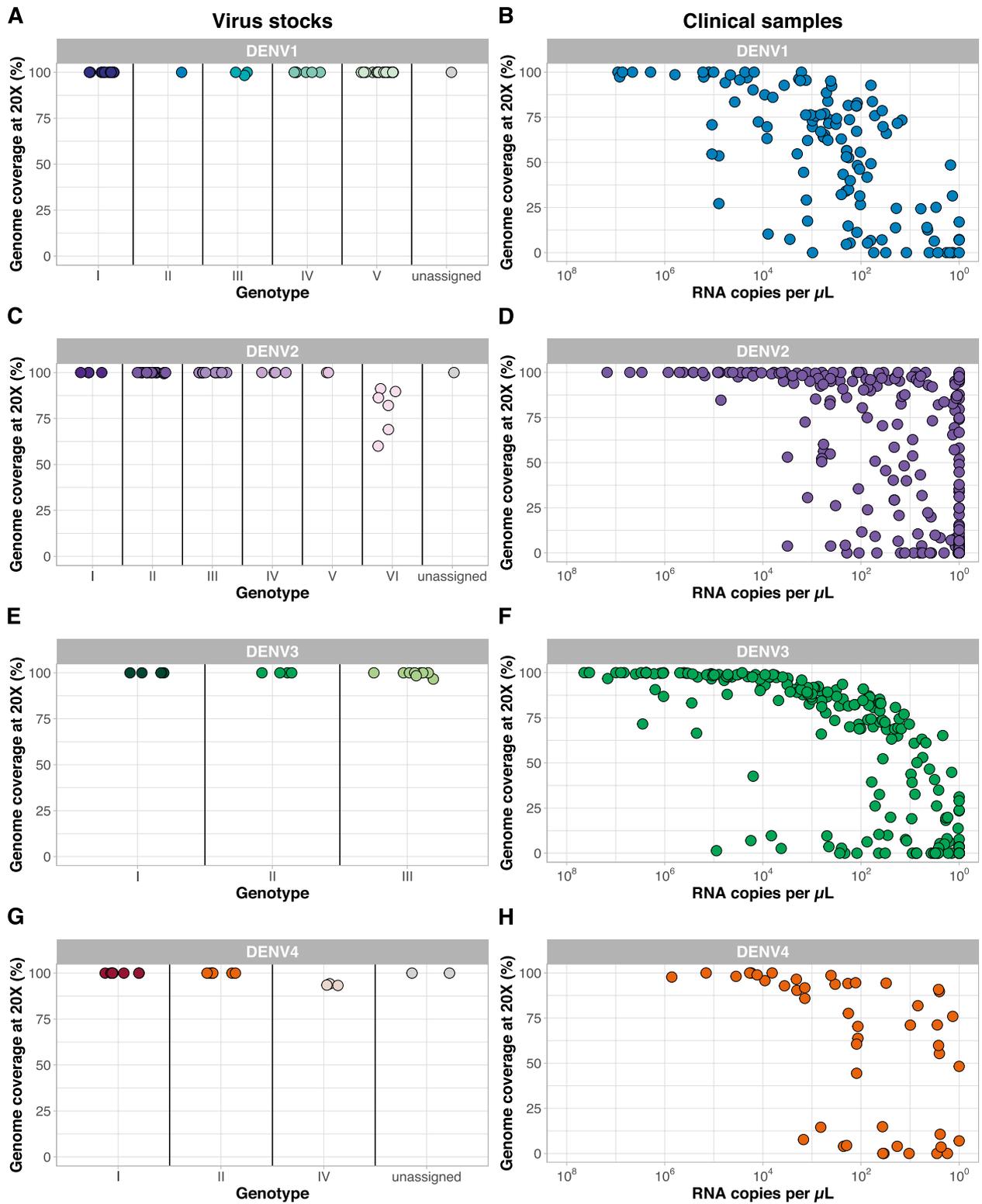


Fig. 2 (See legend on previous page.)

and diluted virus stocks as well as clinical specimens with the pan-serotype approach to compare genome coverage with the serotype-specific approach. We used a threshold of 70% genome coverage to constitute “successful sequencing”, which is a balance between obtaining as many genomes as possible while maintaining the accuracy required for correct phylogenetic placement [14]. This threshold may be lowered when sequencing data is only used to determine serotypes and genotypes, without further phylogenetic analyses. Of the genomes with >70% coverage with the serotype-specific primer scheme (Fig. 2), we found that 84/90 (93.3%) of DENV1, 152/155 (98.1%) of DENV2, 37/38 (97.4%) of DENV3, and 47/47 (100%) of DENV4 met the coverage threshold with the pan-serotype approach (Fig. 3). When further comparing the coverage of these genomes, we found no significant differences in median genome coverage between the serotype-specific and pan-serotype approaches for either virus stocks or clinical specimens (Kruskal-Wallis tests, $P > 0.05$ for all pairwise comparisons after Bonferroni correction; Fig. S5). These findings suggest that both the serotype-specific and pan-serotype approaches perform similarly when sequencing virus stocks or clinical specimens across a range of virus copies. For high-throughput and streamlined sequencing of dengue virus, we recommend using the pan-serotype approach for sequencing any dengue virus serotype.

Limits of detection

We determined the limits of detection to evaluate how DengueSeq performs compared to schemes designed for other viruses. As we described above, mismatches with some of our tested viruses to the CDC dengue multiplex RT-qPCR assay led to inaccurate measurements of virus copies. To account for this, we selected viruses that do not contain impactful mismatches to determine the limits of detection for DengueSeq. For all four serotypes, we estimate that the limit of detection to achieve at least 70% genome coverage ranged between 10^1 - 10^2 RNA copies/ μ L (Fig. 4A, C, E, G). When comparing the limits of detection with other viruses, we found that the primer schemes for SARS-CoV-2 (Fig. 4B) and Powassan virus (Fig. 4D) had similar thresholds to achieve >70% genome coverage of approximately 10^2 RNA copies/ μ L, when

sequencing clinical or field samples. The threshold for Zika virus (Fig. 4F) and West Nile virus (Fig. 4H) based on sequencing diluted virus stocks was slightly higher at 10^3 - 10^5 RNA copies/ μ L. This shows that DengueSeq is at least as sensitive as other commonly used primer schemes for whole-genome sequencing of viruses of public health importance.

Contrived co-infections

With multiple dengue virus serotypes concurrently circulating in the same areas, there is a risk for co-infections with different serotypes [20–22]. The added advantage of using the pan-serotype primer scheme with DengueSeq is the ability to detect co-infections and sequence multiple serotypes in the same sample. When sequencing clinical specimens (Fig. 3), we simultaneously generated near-complete genome sequences of both DENV1 (92.9%) and DENV3 (100%) from the same specimen. To further explore the ability to sequence multiple serotypes during co-infections, we (1) created co-infections by mixing two clinical specimens containing two different serotypes and (2) mixed different concentrations of DENV1 and DENV2 stocks. With the former, we successfully sequenced two different serotypes across a range of virus copies, and the genome coverages were similar to what we generated from single infections (Fig. 5A). By sequencing the stock co-infections at different dilutions, we found that genome coverage was not affected by the presence of another dengue serotype across the tested range of dilutions (Fig. 5B). Thus, by sequencing contrived clinical and stock virus co-infections we showed that the pan-serotype DengueSeq approach has an added benefit of reliably detecting multiple serotypes in the same sample.

Other library prep kits

Previous studies have shown that amplicon primer schemes can be used with a variety of library prep kits [9, 10, 23]. Here, we validated DengueSeq primarily with the Illumina COVIDSeq test kit because it enables high-throughput sequencing without manual normalization steps, and many labs around the world have implemented this kit during the COVID-19 pandemic. However, as some labs prefer different amplicon

(See figure on next page.)

Fig. 3 Comparison of serotype-specific and pan-serotype sequencing approaches. Dengue virus stock and clinical specimens were sequenced with both the serotype-specific and pan-serotype virus approaches to compare genome coverage between both methods. The pan-serotype approach was created by mixing the serotype-specific primer schemes into two pools (pool 1 and pool 2 for each of the combined schemes). All samples were tested with the CDC DENV1-4 multiplex qPCR assay, and libraries were prepared using the Illumina COVIDSeq test kit with serotype-specific or pan-serotype primer schemes. Connecting lines between dots indicate the same samples sequenced with both approaches and the dashed line indicates our threshold of 70% genome coverage

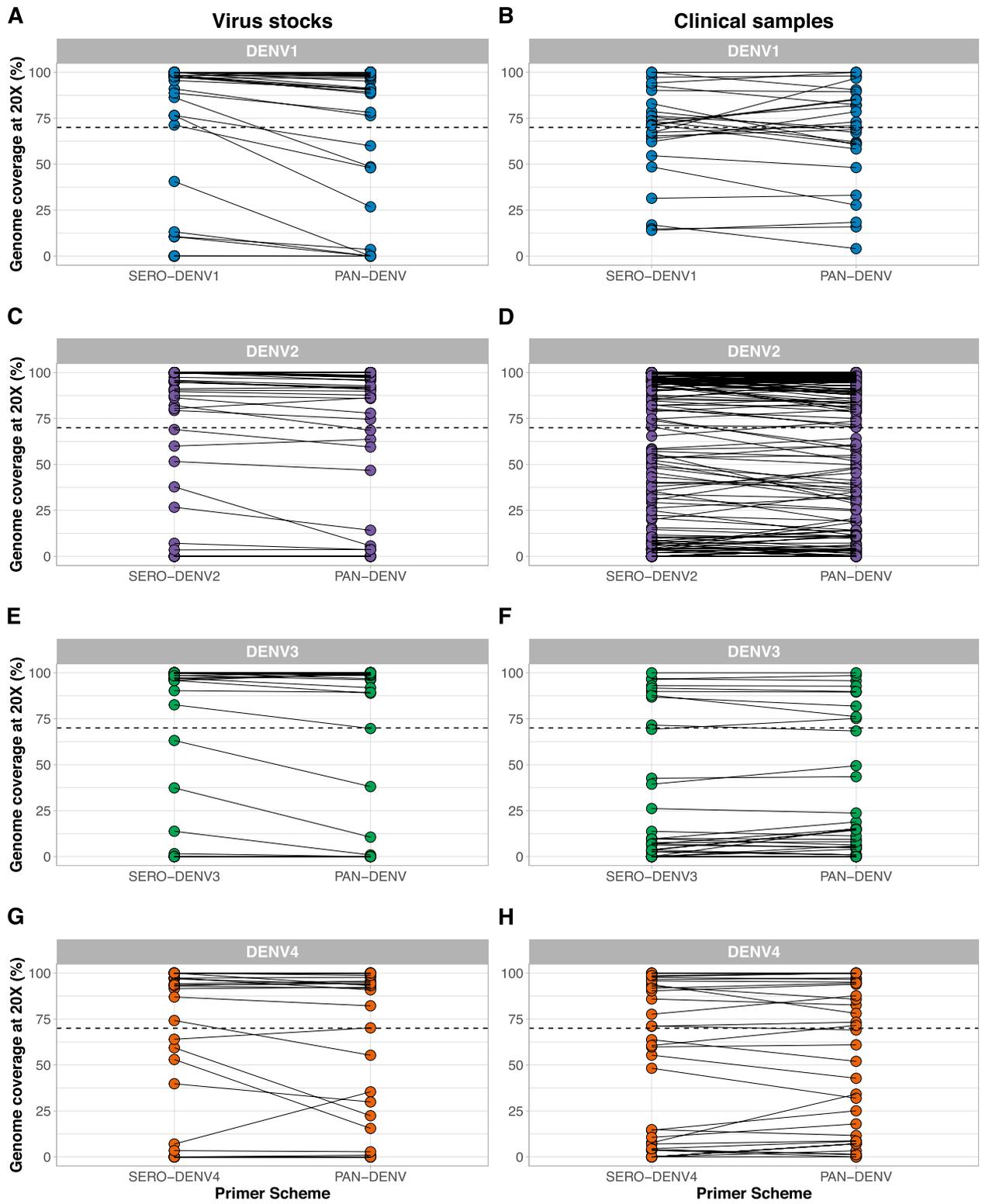


Fig. 3 (See legend on previous page.)

sequencing workflows, we tested DengueSeq with different library prep kits and sequencing instruments. We validated the pan-serotype primer scheme with four additional workflows in different labs: (1) Illumina Nextera XT library prep kit and the Illumina MiSeq at the Florida Department of Health, (2) New England Biolabs NEBNext and the Illumina NovaSeq at the Yale School of Public Health, (3) “Midnight” protocol using the Oxford Nanopore Technologies rapid barcoding kit and the GridION at the Amsterdam University Medical Centers, and (4) Oxford Nanopore Technologies native barcoding kit and the MinION at the University of Nebraska Medical Center (Fig. 6). Although we cannot make direct comparisons, we showed that the pan-serotype primers can be used across a variety of amplicon sequencing workflows and in different labs. Some of the samples that were sequenced with the NEBNext library prep kit had lower-than-expected genome coverage. An imbalance in the final concentrations of individual libraries in the final pool due to manual normalization may explain the higher variation observed and can be further optimized. Testing the pan-serotype primer scheme across library prep kits, sequencing instruments, and in different labs shows that DengueSeq can be integrated into existing amplicon sequencing workflows with minimal change to the overall protocol.

Discussion

Expanded genomic surveillance of dengue virus is critically needed to gain insight into currently circulating genetic diversity, emergence and spread of new lineages, and to evaluate the effectiveness of newly implemented control tools. We developed DengueSeq, a whole-genome pan-serotype sequencing protocol, and associated bioinformatics pipelines to facilitate expanded genomic surveillance of dengue virus by utilizing widely implemented amplicon sequencing workflows. We designed the multiplexed primers to amplify viruses comprising the known diversity of dengue virus causing human outbreaks (Fig. 1), and then validated the approach with a panel of genetically diverse virus stocks and clinical specimens from all four serotypes (Fig. 2). We showed that DengueSeq can be used with any of the

four serotype-specific primers or as a combined pan-serotype approach (Fig. 3). We estimate that 10^1 - 10^2 virus copies per μL is required to generate at least 70% genome coverage for all four serotypes, which is similar to the ARTIC SARS-CoV-2 assay (Fig. 4). Furthermore, we demonstrate that the pan-serotype protocol can be used for samples with unknown serotypes and is able to sequence multiple serotypes in the same sample (e.g. detection of coinfections; Fig. 5). DengueSeq provides a unified approach for whole genome amplicon sequencing of dengue virus that can be used with a variety of library prep methods and sequencing instruments, with minimal change needed to currently established amplicon sequencing workflows (Fig. 6). The ability to combine the four serotype-specific primer schemes into a unified pan-serotype approach demonstrates the proof of principle that amplicon sequencing schemes can be scaled to broad-spectrum virus sequencing workflows.

The development of PrimalScheme has enabled researchers to develop custom primer schemes for amplicon-based sequencing. DengueSeq is not the first nor the only primer scheme available for dengue virus, and several other custom dengue virus primer schemes have recently been developed [24–30]. What makes DengueSeq different from other custom primer schemes is that we (1) designed our primer schemes to include global genetic diversity within each of the serotypes, (2) performed comprehensive validation with samples representing genetic diversity within each serotype, different sample types, and a range of virus copies, and (3) validated a pan-serotype approach that can be used to perform whole genome sequencing of all four serotypes without prior serotyping. Additionally, several commercial virus target enrichment panels are currently available for broad virus surveillance. The main advantage of these panels is that they capture a large number of viruses, but the disadvantages are the relatively high cost and lower sensitivity as compared to amplicon sequencing. Thus, several solutions are available that can potentially be used for dengue virus sequencing. DengueSeq is particularly useful for labs that have already implemented amplicon sequencing workflows and/or are looking for a relatively low-cost, sensitive, and high-throughput sequencing

(See figure on next page.)

Fig. 4 Limits of detection of pan-serotype approaches and primer schemes for other viruses. A/C/E/G) Virus stocks were diluted to represent a wide range of RNA titers and were sequenced with the pan-serotype virus approach using the Illumina CovidSeq test kit. B) Clinical specimens testing positive for SARS-CoV-2 were sequenced using the ARTIC v4.1 primer scheme using the Illumina COVIDSeq test kit. D) *Ixodes scapularis* ticks testing positive for Powassan virus were sequenced with the Powassan virus primer scheme using the Illumina COVIDSeq test kit. F) Two Zika virus stocks were sequenced with the Zika virus primer scheme using the Illumina CovidSeq test kit. H) A West Nile virus stock was sequenced with the West Nile virus primer scheme using the Illumina CovidSeq test kit. Dots represent individual samples, and a line with confidence intervals was fitted using the `geom_smooth` function with the `glm` method from `ggplot2`

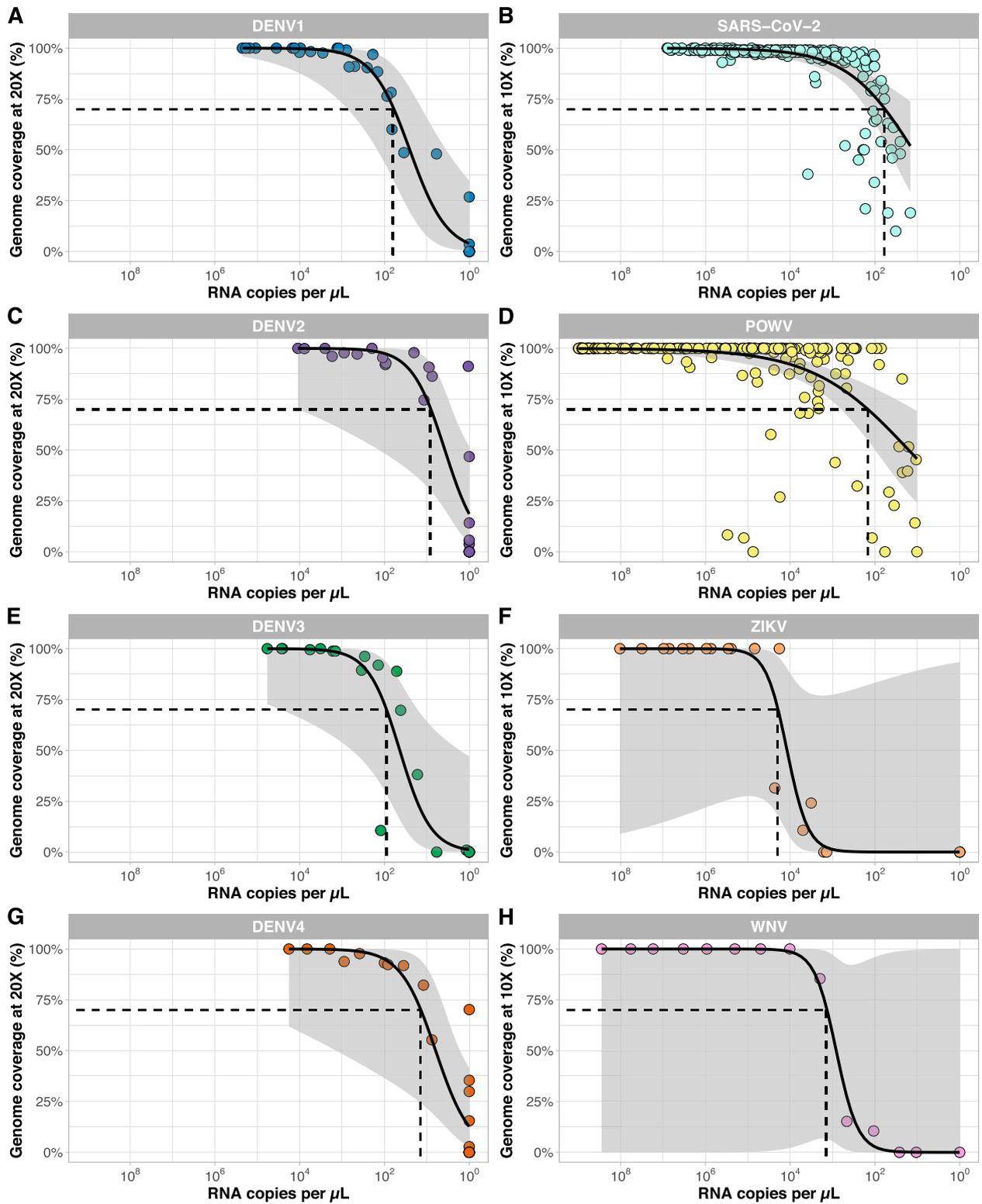


Fig. 4 (See legend on previous page.)

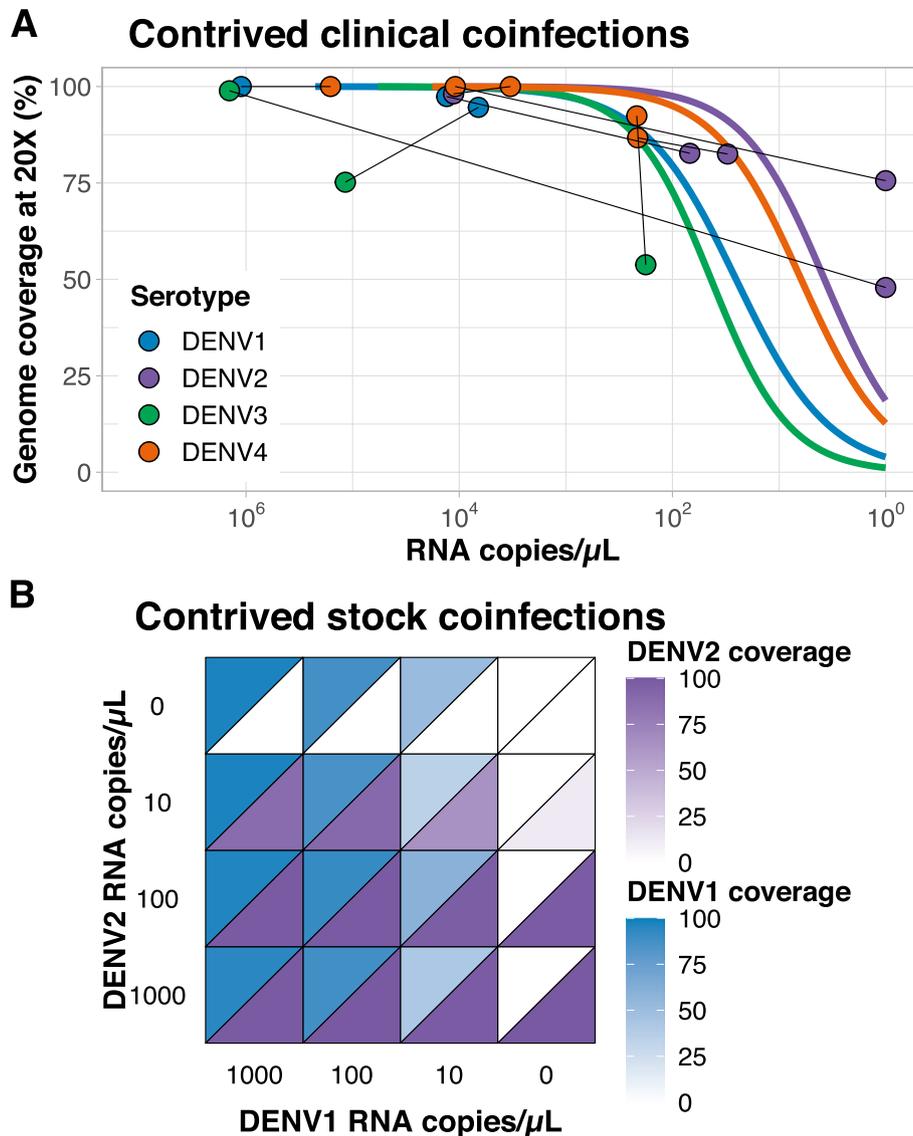


Fig. 5 Detection of dengue virus coinfections using the pan-serotype primer scheme. **A** Clinical specimens containing different dengue virus serotypes were mixed to create contrived coinfections. A total of 8 contrived samples were created by mixing two different clinical specimens. Dots represent individual samples connected with lines to indicate mixed pairs and colored lines are expected genome coverage estimated from fitted logistic regression lines determined based on coverage data from Fig. 4. **B** Two virus stocks (serotypes 1 and 2) were serially diluted and mixed at different concentrations to represent a range of virus titers in contrived stock coinfections

system to conduct genomic surveillance of dengue virus from clinical samples that typically have low virus copies. These features of the DengueSeq approach would also make it easily deployable for dengue genomic surveillance, especially in low-income settings with scant genomic data.

High dengue virus genetic diversity poses challenges to successful whole genome sequencing. Our approach of combining the four serotype-specific primer schemes into the pan-serotype approach allows us to capture the

current genetic diversity present within distinct genotypes in the four serotypes. However, we cannot predict whether the current primer schemes will capture newly emerging lineages. This scenario is exemplified by the lower coverage of sylvatic genomes (Fig. 2C, G), which enables identification but not whole genome sequencing. However, updating the DengueSeq scheme to include newly characterized lineages using the open source PrimalScheme tool can improve the sequencing coverage of these emerging lineages. Several other factors may also

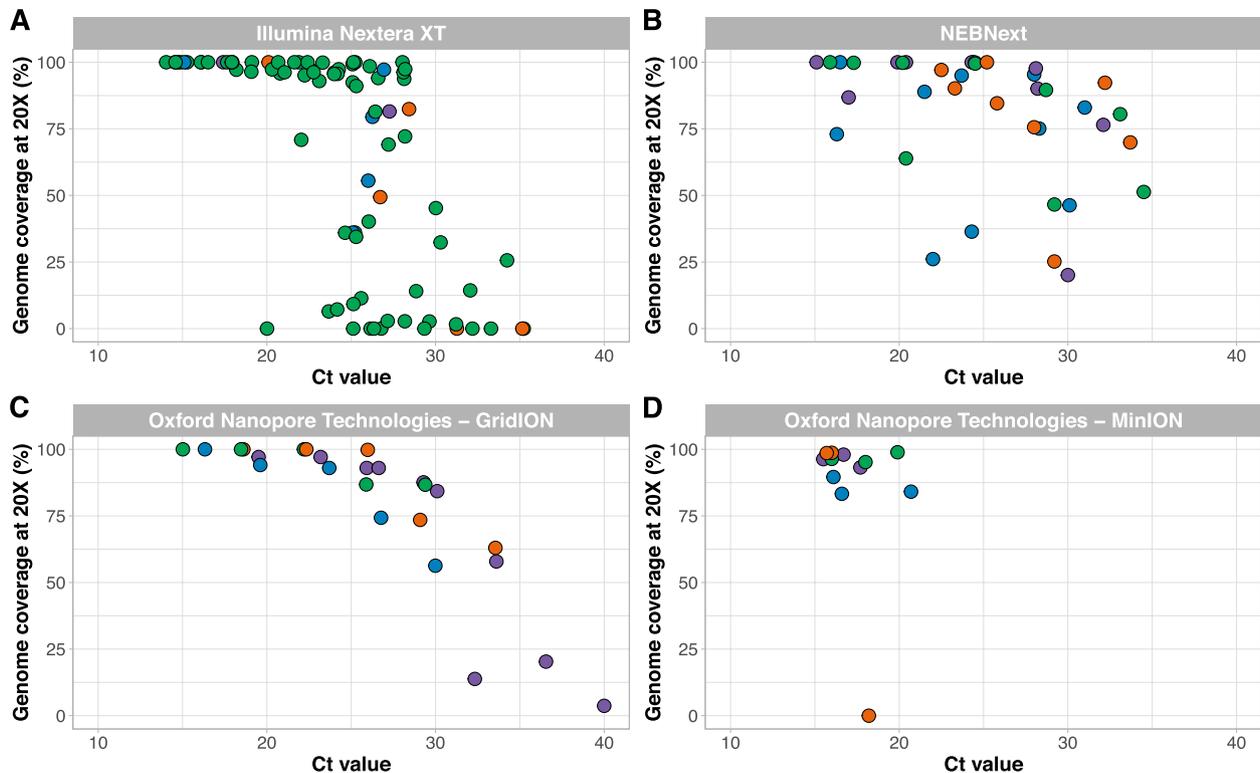


Fig. 6 Alternative library prep methods used with the pan-serotype primer scheme. **A** The Florida Department of Health sequenced clinical specimens using the pan-serotype primer scheme with the Illumina Nextera XT kit. **B** Clinical specimens were sequenced using the pan-serotype primer scheme with the NEB NEBNext library prep kit for Illumina. **C** Amsterdam UMC sequenced virus stocks using the pan-serotype primer scheme with the “Midnight” protocol and sequenced on the Oxford Nanopore Technologies GridION. **D** The University of Nebraska Medical Centers sequenced dengue virus stocks using the Oxford Nanopore Technologies native barcoding kit and sequenced on the Oxford Nanopore Technologies MinION

affect the future performance of DengueSeq. To account for the high genetic diversity of dengue virus, we initially designed “degenerate” primer schemes that included ambiguous nucleotides at divergent primer-binding positions. The inclusion of ambiguous nucleotides did not improve the sensitivity of the primer schemes, and, therefore, we proceeded with validation of the non-degenerate (original) primer schemes. We determined the limits of detection ranging between 10^1 - 10^2 RNA copies/ μ L to achieve at least 70% genome coverage. However, we should interpret these thresholds carefully as the threshold for DengueSeq when sequencing clinical specimens may need to be more conservative depending on sample quality, specimen type, available resources, potential mismatches with the qPCR primers and probe etc. We recommend that laboratories quantify the relationship between genome coverage and RNA copies to determine their threshold, as this would allow them to prioritize specimens for sequencing. Particularly when resources are limited, we recommend to select samples with higher RNA copies to increase the chances of generating near-complete genomes. The choice of reference

genomes for reference-based alignment of sequencing reads had a significant impact on genome coverage of highly divergent genomes. By using DENV2 and DENV4 sylvatic references in the bioinformatic analysis, we generated higher genome coverage than when aligning reads to the non-sylvatic reference genomes. Future development of refined virus-agnostic bioinformatics pipelines from amplicon-based sequencing would be beneficial to ensure proper alignment of divergent sequencing reads as well as coinfections. The choice of library prep methods is another consideration for labs wanting to implement DengueSeq. We showed that the pan-serotype approach can be used with a variety of library prep methods for sequencing on Illumina and Oxford Nanopore Technologies instruments. For some of the methods that require manual normalization and pooling, we found higher variation in genome coverage, which is likely the result of an imbalance in the individual libraries that were pooled. We recommend that laboratories use DengueSeq to expand currently-implemented amplicon sequencing workflows, e.g., for SARS-CoV-2, to include dengue virus.

Conclusions

We demonstrated that DengueSeq is a unified, sensitive, and high-throughput whole genome amplicon sequencing protocol for all four dengue virus serotypes. We designed and validated DengueSeq with a broad panel of samples to represent the breadth of genetic diversity, different sample types, and a range of virus copies. Our findings show that DengueSeq can be used with currently established amplicon sequencing workflows and provide a proof of principle that multiple primer schemes can be combined into unified approaches to improve the sequencing of multiple dengue serotypes. The use of DengueSeq will facilitate genomic surveillance of dengue virus, which is critically needed to understand the increasing scale and frequency of outbreaks and the effectiveness of control measures.

Methods

Primer scheme design

Four separate serotype-specific primer schemes were designed using PrimalScheme. For each serotype, genetically diverse genomes were selected to represent the distinct genotypes (Fig 1B). Genotypes with only incomplete genomes available and highly divergent genotypes were excluded (e.g. DENV1 genotype II, DENV2 genotype IV and VI (sylvatic), and DENV4 genotypes III and IV (sylvatic)), until a primer scheme with high coverage across the genome was found. Each primer scheme consists of 35-37 primer pairs with an average amplicon length of 400 bp. The primer schemes can be used as a serotype-specific approach when the serotype is known (working concentration of 10 μ M), or the same primers can be combined at equal concentrations as a pan-serotype approach (working concentration of 20 μ M; Pool 1 = 146 primers, Pool 2 = 138 primers). Primer sequences, reference files, and bed files are available on protocols.io [31].

Virus stocks for validation

Genetically diverse virus stocks with representatives for all dengue virus serotypes and genotypes from the Yale Arbovirus Research Unit and World Reference Center for Emerging Viruses and Arboviruses collections were kindly shared by the Connecticut Agricultural Experiment Station and the University of Texas Medical Branch. The total collection of virus stocks used for validation consisted of 46 DENV1, 46 DENV2, 17 DENV3, and 18 DENV4 isolates. Viruses were sequenced with the serotype-specific and pan-serotype approaches. The dengue multiplex RT-qPCR assay as developed by the CDC was used to determine serotype and PCR Ct values for each stock [16]. Virus stocks were serially diluted to represent a range of Ct values. Serial ten-fold dilutions of quantified

synthetic dengue virus 1-4 controls obtained from ATCC were used to convert Ct values to RNA copies per μ L.

Clinical specimens for validation

De-identified remnant clinical specimens were provided by the Florida Department of Health and nucleic acid was extracted by Florida Gulf Coast University. Clinical specimens were collected between 2010-2023 and comprised locally acquired cases as well as travel-related cases. Nucleic acid was extracted using the Qiagen QIAamp viral RNA mini kit. The dengue multiplex RT-qPCR assay, as developed by the CDC, was used to determine serotype and PCR Ct values for each specimen [16]. The total collection of clinical specimens used for validation consisted of 121 DENV1, 214 DENV2, 219 DENV3, and 46 DENV4 samples. The serotype-specific approach was used to sequence all specimens, and a smaller subset was re-sequenced using the pan-serotype approach, as described below.

Amplicon sequencing

Initial amplicon sequencing validation was performed at the Yale School of Public Health. To test the primer scheme in diverse lab settings and with different library prep kits, additional testing was performed by the Florida Department of Health, the Amsterdam University Medical Centers, and the University of Nebraska Medical Center. A detailed protocol is available on protocols.io [31].

At the Yale School of Public Health, initial validation of the primer schemes was done with (diluted) virus stocks and de-identified remnant clinical specimens. Samples were sequenced using the Illumina COVIDSeq test (RUO version) by replacing primers as part of the kit for the serotype-specific or pan-serotype primer pools. A negative template control was included during cDNA synthesis and amplicon generation for each sequencing run. Pooled libraries were sequenced on the Illumina NovaSeq 6000 (paired-end 150) at the Yale Center for Genome Analysis, targeting 1 million reads per sample. Consensus genomes were generated using the newly developed bioinformatics pipeline for Illumina data as described below. Sequencing data and consensus genomes are available in BioProject PRJNA951702 and PRJNA1001374.

At the Florida Department of Health, validation of the primer schemes was done with 84 clinical specimens consisting of serum from individuals diagnosed with dengue. Nucleic acid was extracted from clinical specimens using the Qiagen Qiacube DSP Viral RNA Mini and tested with the CDC Dengue Virus Typing (DENV-1/2/3/4) FDA IVD assay on the Thermo Fisher ABI 7500. Libraries were prepared for sequencing using the Illumina Nextera XT library prep kit, and sequenced on the Illumina Miseq

instrument (251x2), targeting 250,000 reads (range of 150,000–400,000) per sample. Consensus genomes were generated with the newly developed bioinformatics pipeline. Sequencing data and consensus genomes are available in BioProject PRJNA973096.

At Amsterdam University Medical Centers, the primer scheme was validated using five cultured dengue strains passaged on Vero E6 cells. Nucleic acids were extracted from 200 μ L of culture supernatant using the Qiagen QIAamp viral RNA mini kit with an elution volume of 50 μ L. Viral loads were determined using the Roche LightMix modular dengue virus kit on the Roche Lightcycler 480, according to the manufacturer's protocol. Sequencing libraries were prepared for sequencing using the Oxford Nanopore Technologies (ONT) rapid barcoding kit 96 following the "midnight" protocol [32], and sequenced on the ONT GridION, using an R.9.4.1 flowcell, targeting 30,000 reads per sample. Consensus genomes were generated at a depth of 20x with the ARTIC bioinformatics pipeline [33], by using the dengue virus 1–4 reference genomes (dengue virus 1: NC_001477, dengue virus 2: NC_001474, dengue virus 3: NC_001475, dengue virus 4: NC_002640) and corresponding BED files.

At the University of Nebraska Medical Center, the primer scheme was validated with 12 dengue virus stocks provided by the Yale School of Public Health. A New England Biolabs LunaScript RT SuperMix Kit was used to synthesize cDNA from each sample following the manufacturer's protocol modified for half-volume (10 μ L) reactions. Following amplification, sequencing libraries were prepared using the ONT Native Barcoding Kit V14 (SQK-NBD114.24) and the manufacturer's protocol. After end preparation, 50 ng of each sample was input into native barcode ligation reactions. Following ligation of unique barcodes, 2 μ L of each sample was combined to generate a pooled library. Two 21 ng aliquots of the final library were sequenced on separate ONT MinION R10.4.1 flow cells for 4 hr each using ONT MinKNOW software v.23.07.12 with pore scans every 1.5 hr. The sequencing runs generated an estimated 1.51 Gb and 2.27 Gb of data, respectively. Guppy v.6.5.7 (ONT) was used to demultiplex and basecall .pod5 data by requiring barcodes at both ends of reads. Consensus genomes were generated from basecalled .fastq data using the newly developed bioinformatics pipeline for ONT data as described below.

Bioinformatics pipeline

Pipeline for Illumina data

Raw demultiplexed read data generated from sequencing amplicons were processed using a pip-installable, Snakemake-based pipeline to generate consensus

genome sequences of the dengue serotypes in the clinical sample (Fig. S1). The bioinformatics pipeline used for this analysis is available on Github (https://github.com/grubaughlab/DENV_pipeline/tree/v1.0), and is based on Python v3.9 and bash scripts. The input of the pipeline are the forward and reverse FASTQ read files for each sample, and a set of BED and FASTA files for the primer positions and corresponding reference sequences. The default for the pipeline for the latter is the reference sequences and primers for DENV1 (GenBank accession: NC_001477.1), DENV2 (GenBank accession: NC_001474.2), DENV3 (NC_001475.2), DENV4 (GenBank accession: NC_002640.1), DENV2 sylvatic (GenBank accession: EF105380.1) and DENV4 sylvatic (GenBank accession: JF262780.1) for the protocol described in this paper. The reference sequences are selected to represent the diversity of all known dengue serotypes so that the best match can be found with the virus in the clinical sample.

Each set of reads is mapped against each reference sequence separately (non-competitive alignment) using BWA v0.7.17-r1188 [34], and primers corresponding to each reference are trimmed using iVar v.1.4 [10]. If no reads map successfully, then the rest of the algorithm is skipped. BAM files are then sorted and indexed using samtools v1.17 [35], and the consensus nucleotide sequence is generated using samtools and iVar. A frequency threshold of 0.75 for the single nucleotide variants included in the consensus sequence and a per-base depth of at least 10 mapped reads is used by default, although both can be changed by the user. We investigated these two parameters and found that these values had the best balance of accuracy and reliability.

Once the consensus genome sequence is generated, it is aligned to the same reference sequence using Nextalign v2.1 [36]. If there are too few reads, Nextalign may fail to generate the alignment; MAFFT v7.520 is used in these circumstances [37]. The genome coverage or genome completeness, i.e., the percentage of bases in the aligned sequence not containing ambiguous bases, is then calculated using a custom Python script utilizing the BioPython module [38]. Finally, variants and per-base depth at each position in the generated mapping output (sorted BAM file) are identified using samtools, iVar and BEDTools v2.31.0 [39].

Outputs of the pipeline include three summary tab-separated text files containing information on coverage against each possible reference sequence: one with all information, one with any "called" serotype (defined as above 50% coverage), and one with whatever the reference with the highest coverage is. Alignments in FASTA format, BAM files, and consensus sequences of called serotypes are also available, as are text files containing

depth at each position of each sample for their called serotype. Finally, text files containing information on nucleotide variants against called serotypes are output.

Pipeline for Oxford Nanopore Technologies (ONT) data

This bash wrapper script is designed to be run within an ARTIC conda environment [40, 41]. It is specifically designed to be serotype agnostic, and will choose the most likely serotype for each sample based on aligning reads to reference genomes. The “pass” basecalled read datasets output by Guppy for each sample and a comma-separated text file associating barcode names and sample names are used as input. Reads in each dataset are decompressed (if applicable) and the first 15 .fastq files (which contain variable numbers of reads depending on the sample) are then aligned to each of the four DENV serotype-specific reference sequences using minimap2 v2.26-r1175. Then, samtools v1.17 is used to generate and parse BAM files for each alignment. The total number of reads mapped to each reference genome are reported and compared, and the genome with the greatest number of reads mapped is assigned as the reference genome for that sample. If no reads or equal numbers of reads are mapped across the four reference genomes, the DENV1 reference genome is assigned to that sample by default. Reads for the sample are then concatenated into a single fastq file, which is used for input into the ARTIC guppy-plex command to filter by the values for minimum and maximum read length set by the user. The ARTIC pipeline is then run to generate a consensus genome for each sample using the sample-specific filtered read dataset, the serotype-specific reference genome assigned to that sample earlier in the pipeline and its associated bed file, and a user-specified medaka model.

Data analysis and visualizations

All data analysis and plotting was performed using R statistical software v4.3.1 [42] using the ggplot2 v3.4.2 [43], dplyr v1.1.2 [44], tidyverse v2.0.0 [45], cowplot v1.1.1 [46], and scales v1.2.1 packages [47]. Differences in coverage between the serotype-specific and pan-serotype approaches were tested with Kruskal-Wallis tests with Bonferroni correction for multiple comparisons. For coverage plots, lines with confidence intervals were fitted using the geom_smooth function with the 'glm' method from ggplot2.

For the phylogenetic trees displayed in Fig. 1, S3, and S4, we began by collating a background set of sequences for each serotype, which were the same background set used for primer selection. We combined these with each set of sequences of interest - i.e., the primers, the virus stocks, and the clinical sequences - using BioPython in a custom script [38], aligned them

using MAFFT v7.490 [37], and manually curated the resulting alignment. We then inferred a Maximum Likelihood tree using IQTree [48] with the HKY substitution model [49]. Finally, we visualized these using the Python package baltic [50].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10350-x>.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Acknowledgements

We thank S. Weaver, J. Plante, and K. Plante from the University of Texas Medical Branch for sharing dengue virus stocks from the World Reference Center for Emerging Viruses and Arboviruses; D. Brackney, A. Bransfield, and P. Armstrong from the Connecticut Agricultural Experiment Station for sharing dengue virus stocks from the Yale Arbovirus Research Unit Collection; New England Biolabs for providing reagents; A. Porzucek for help with proofreading the protocols.io; and P. Jack and S. Taylor for technical advice.

Authors' contributions

CBFV, VH, and NDG conceptualized and designed the study; CBFV, MIB, LMP, AS, ET-S, IMO, MEP, DD, MJ, MRAW, TL, SB, NC, RJ, CP, IS, JV, RZ, EK, LH, KSH, JRF, AMM, and SFM contributed to sample collection and data generation; CBFV, VH, CC, YD, NT, OT, SS, KSH, and JRF analyzed the data; CBFV, VH, and NDG wrote the manuscript; all authors reviewed and approved the final manuscript.

Funding

This publication was made possible by CTSA Grant Number UL1 TR001863 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (CBFV), the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number DP2AI176740 (NDG), the National Institutes of Health T32AI055403 (AS), the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139841 (AS), and the National Institute of General Medical Sciences R21GM142011 (SFM). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Validation of the sequencing protocol at the Amsterdam UMC was funded by the department of Medical Microbiology & Infection Prevention.

Availability of data and materials

All genomic data are available in the manuscript, supporting information, GitHub (https://github.com/grubaughlab/DENV_pipeline/tree/v1.0; https://github.com/josephfauver/DENV_MinION_Script), and BioProjects PRJNA951702, PRJNA1001374, and PRJNA973096.

Declarations

Ethics approval and consent to participate

The Florida Department of Health shared de-identified remnant clinical specimens that tested positive for dengue virus for sequencing at the Yale School of Public Health. The Institutional Review Boards (IRB) from the Yale University Human Research Protection Program, Florida Gulf Coast University, and Florida Department of Health determined that pathogen genomic sequencing of de-identified remnant diagnostic samples as conducted in this study is not research involving human subjects (Yale IRB Protocol ID: 2000033281 and 2000028599).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA. ²Yale Institute for Global Health, Yale University, New Haven, Connecticut, USA. ³Department of Biological Sciences, College of Arts and Sciences, Florida Gulf Coast University, Fort Myers, Florida, USA. ⁴Department of Pediatrics, Yale School of Medicine, New Haven, Connecticut, USA. ⁵Sydney Institute for Infectious Diseases, School of Medical Sciences, University of Sydney, Sydney, NSW, Australia. ⁶Department of Medical Microbiology & Infection Prevention, Amsterdam UMC Location AMC, Amsterdam, The Netherlands. ⁷Department of Infectious Diseases, Public Health Service of Amsterdam, Amsterdam, The Netherlands. ⁸Bureau of Public Health Laboratories, Division of Disease Control and Health Protection, Florida Department of Health, Tampa, FL, USA. ⁹Bureau of Public Health Laboratories, Division of Disease Control and Health Protection, Florida Department of Health, Jacksonville, FL, USA. ¹⁰Bureau of Epidemiology, Division of Disease Control and Health Protection, Florida Department of Health, Tallahassee, FL, USA. ¹¹Department of Epidemiology, University of Nebraska Medical Center, Omaha, Nebraska, USA. ¹²Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA. ¹³Public Health Modeling Unit, Yale School of Public Health, New Haven, Connecticut, USA.

Received: 16 November 2023 Accepted: 25 April 2024

Published online: 01 May 2024

References

- Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl Trop Dis*. 2012;6:e1760.
- Zeng Z, Zhan J, Chen L, Chen H, Cheng S. Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017. *EClinicalMedicine*. 2021;32:100712.
- Dengue – the Region of the Americas [Internet]. [cited 2023 Jul 26]. Available from: <https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON475>
- Wallau GL, Global Arbovirus Researchers United. Arbovirus researchers unite: expanding genomic surveillance for an urgent global need. *Lancet Glob Health*. 2023;11:e1501–2.
- Chen R, Vasilakis N. Dengue—quo tu et quo vadis? *Viruses*. 2011;3:1562–608.
- Katzelnick LC, Coello Escoto A, Huang AT, Garcia-Carreras B, Chowdhury N, Maljkovic Berry I, et al. Antigenic evolution of dengue viruses over 20 years. *Science*. 2021;374:999–1004.
- Walker T, Johnson PH, Moreira LA, Iturbe-Ormaetxe I, Frentiu FD, McMeniman CJ, et al. The wMel Wolbachia strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature*. 2011;476:450–3.
- Pinheiro-Michelsen JR, Souza R da SO, Santana IVR, da Silva P de S, Mendez EC, Luiz WB, et al. Anti-dengue Vaccines: From Development to Clinical Trials. *Front Immunol*. 2020;11:1252.
- 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. *Nat Protoc*. 2017;12:1261–76.
- 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 Biol*. 2019;20:8.
- Artic Network [Internet]. [cited 2022 Jul 25]. Available from: <https://artic.network/ncov-2019>
- Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res*. 2015;43:D571–7.
- NCBI Virus [Internet]. [cited 2023 Aug 1]. Available from: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Dengue%20virus,%20taxid:12637
- Dudas G, Bedford T. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC Evol Biol*. 2019;19:232.
- Hill V, Githinji G, Vogels CBF, Bento AI, Chaguza C, Carrington CVF, et al. Toward a global virus genomic surveillance network. *Cell Host Microbe*. 2023;31(6):861–73.
- Santiago GA, Vergne E, Quiles Y, Cosme J, Vazquez J, Medina JF, et al. Analytical and clinical performance of the CDC real time RT-PCR assay for detection and typing of dengue virus. *PLoS Negl Trop Dis*. 2013;7:e2311.
- Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics*. 2019;35:871–3.
- Muller DA, Depelsenair ACI, Young PR. Clinical and Laboratory Diagnosis of Dengue Virus Infection. *J Infect Dis*. 2017;215:S89–95.
- Pei-Yun Shu, Jyh-Hsiung Huang. Current Advances in Dengue Diagnosis. *Clin Vaccine Immunol*. 2004;11:642–50.
- Andrade EHP, Figueiredo LB, Vilela APP, Rosa JCC, Oliveira JG, Zibaoui HM, et al. Spatial-Temporal Co-Circulation of Dengue Virus 1, 2, 3, and 4 Associated with Coinfection Cases in a Hyperendemic Area of Brazil: A 4-Week Survey. *Am J Trop Med Hyg*. 2016;94:1080–4.
- Tazeen A, Afreen N, Abdullah M, Deeba F, Haider SH, Kazim SN, et al. Occurrence of co-infection with dengue viruses during 2014 in New Delhi. *India Epidemiol Infect*. 2017;145:67–77.
- Dhanoa A, Hassan SS, Ngim CF, Lau CF, Chan TS, Adnan NAA, et al. Impact of dengue virus (DENV) co-infection on clinical manifestations, disease severity and laboratory parameters. *BMC Infect Dis*. 2016;16:406.
- 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 Biol*. 2023;21:e3002151.
- Stubbs SCB, Blacklaws BA, Yohan B, Yudhaputri FA, Hayati RF, Schwem B, et al. Assessment of a multiplex PCR and Nanopore-based method for dengue virus sequencing in Indonesia. *Virology*. 2020;17:24.
- Su W, Jiang L, Lu W, Xie H, Cao Y, Di B, et al. A Serotype-Specific and Multiplex PCR Method for Whole-Genome Sequencing of Dengue Virus Directly from Clinical Samples. *Microbiol Spectr*. 2022;10: e0121022.
- Adelino TÊR, Giovanetti M, Fonseca V, Xavier J, de Abreu AS, do Nascimento VA, et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. *Nat Commun*. 2021;12:2296.
- Cruz CD, Torre A, Troncos G, Lambrechts L, Leguia M. Targeted full-genome amplification and sequencing of dengue virus types 1–4 from South America. *J Virol Methods*. 2016;235:158–67.
- Aw PPK, de Sessions PF, Wilm A, Hoang LT, Nagarajan N, Sessions OM, et al. Next-Generation Whole Genome Sequencing of Dengue Virus. In: Padmanabhan R, Vasudevan SG, editors., et al., *Dengue: Methods and Protocols*. Springer, New York, NY; 2014;175–95.
- Christenbury JG, Aw PPK, Ong SH, Schreiber MJ, Chow A, Gubler DJ, et al. A method for full genome sequencing of all four serotypes of the dengue virus. *J Virol Methods*. 2010;169:202–6.
- Ashall N, Shah S, Biggs JR, Chang JNR, Jafari Y, Brady OJ, et al. A phylogenetic study of dengue virus in urban Vietnam shows long-term persistence of endemic strains. *Virus Evol*. 2023;9:vead012.
- Vogels C, Chaguza C, Breban MI, Sodeinde A, Porzucek AJ, Grubaugh ND, et al. DengueSeq: A pan-serotype whole genome amplicon sequencing protocol for dengue virus. 2023 [cited 2023 Jul 28]; Available from: <https://www.protocols.io/view/dengueseq-a-pan-serotype-whole-genome-amplicon-seq-kqdg39xeg25/v2>
- Freed N, Murphy L, Schwessinger B, Silander O. "Midnight" SARS-CoV2 genome sequencing protocol using 1200bp amplicon primer set v2 and the Nanopore. 2023 [cited 2023 Aug 28]; Available from: <https://www.protocols.io/view/34-midnight-34-sars-cov2-genome-sequencing-protocol-14egn2q2yg5d/v1>
- Rambaut NLWR. ARTIC Network nCoV-2019 novel coronavirus bioinformatics protocol. 2020 [cited 2022 Jan 17]. Available from: <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*. 2013. Available from: <http://arxiv.org/abs/1303.3997>

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
36. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J Open Source Softw*. 2021;6:3773.
37. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–80.
38. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422–3.
39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
40. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 2016;530:228–32.
41. fieldbioinformatics: The ARTIC field bioinformatics pipeline. Github; [cited 2024 Jan 11]. Available from: <https://github.com/artic-network/fieldbioinformatics>
42. R Core Team. The R Project for Statistical Computing [Internet]. R: A language and environment for statistical computing. 2022 [cited 2022 Sep 12]. Available from: <https://www.r-project.org/>
43. Wickham H. ggplot2. New York: Springer; 2009.
44. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation [Internet]. Comprehensive R Archive Network (CRAN); 2022 [cited 2022 Sep 12]. Available from: <https://cran.r-project.org/web/packages/dplyr/index.html>
45. Wickham H, Girlich M. tidyr: Tidy Messy Data [Internet]. Comprehensive R Archive Network (CRAN); 2022 [cited 2022 Sep 12]. Available from: <https://cran.r-project.org/web/packages/tidyr/index.html>
46. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2” [Internet]. 2020 [cited 2022 Sep 12]. Available from: <https://CRAN.R-project.org/package=cowplot>
47. Scale Functions for Visualization [R package scales version 1.2.1]. 2022 [cited 2023 Jul 28]; Available from: <https://cran.r-project.org/web/packages/scales/index.html>
48. 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*. 2015;32:268–74.
49. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 1985;22:160–74.
50. Dudas G. baltic: baltic - backronymed adaptable lightweight tree import code for molecular phylogeny manipulation, analysis and visualisation. Development is back on the evogytis/baltic branch (i.e. here) [Internet]. Github; [cited 2023 Jul 28]. Available from: <https://github.com/evogytis/baltic>

Publisher's Note

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