Development of Highly Efficient Methods for Comprehensive Pathogen Detection Using Next Generation Sequencing

(次世代シーケンサーを用いた
高効率網羅的病原体検出法の開発)

Patrick Reteng

Table of contents

Publications related to this dissertation

The contents of Chapter One have been published in Scientific Reports.

Reteng P, Nguyen Thuy L, Tran Thi Minh T, Mares-Guia MAMM, Torres MC, de Filippis AMB, Orba Y, Kobayashi S, Hayashida K, Sawa H, Hall WW, Nguyen Thi LA, Yamagishi J. A targeted approach with nanopore sequencing for the universal detection and identification of flaviviruses. Sci Rep. 2021;11(1):19031.

The contents of Chapter Two and Chapter Three have been published in mSphere.

Reteng P, Nguyen Thuy L, Rahman M, de Filippis AMB, Hayashida K, Sugi T, Gonzalez G, Hall WW, Nguyen Thi LA, Yamagishi J. Circular whole-transcriptome amplification (cWTA) and mNGS screening enhanced by a group testing algorithm (mEGA) enable high-throughput and comprehensive virus identification. mSphere. 2022;7(5):e0033222.

List of Abbreviations

BLAST	:	Basic local alignment search tool
bp	:	Base-pair
CDC	:	Center for Disease Control and Prevention
cDNA	:	Complementary DNA
cfDNA/cfRNA	:	Cell-free DNA/RNA
CHIKV	:	Chikungunya virus
COVID-19	:	Coronavirus disease 2019
Ct	:	Cycle threshold
cWTA	:	Circle whole transcriptome amplification
DAMIAN	:	Detection and analysis of viral and microbial infectious agents by next generation sequencing
DENV1-4	:	Dengue virus 1-4
ELISA	:	Enzyme-linked immunosorbent assay
ER	:	Endoplasmic reticulum
FBS	:	Fetal bovine serum
FREEBarcodes	:	Filled/truncated right end edit barcodes
FUO	:	Fever of unknown origin
HBV	:	Hepatitis B virus
HFF	:	Human foreskin fibroblast
HIV-1	:	Human immunodeficiency virus 1
JEV	:	Japanese encephalitis virus
kb	:	Kilo-base-pair
LAST	:	Local alignment search tool
MALBAC	:	Multiple annealing and looping based amplification cycles
MDA	:	Multiple displacement amplification
mEGA	:	Metagenomic next generation sequencing enhanced by a group testing algorithm
mNGS	:	Metagenomic next generation sequencing
NAT	:	Nucleic acid test
NGS	:	Next generation sequencing
NIHE	:	National Institute of Hygiene and Epidemiology
NS1-5	:	Non-structural protein 1-5
nt	:	Nucleotide
ORF	:	Open reading frame

PCR	:	Polymerase chain reaction
PFU	:	Plaque forming unit
RCA	:	Rolling circle amplification
RdRp	:	RNA-dependent RNA polymerase
RDT	:	Rapid diagnostic test
RT-PCR	:	Reverse transcription polymerase chain reaction
RT-qPCR	:	Reverse transcription quantitative polymerase chain reaction
SARS-CoV-2	:	Severe acute respiratory syndrome coronavirus 2
SHERLOCK	:	Specific high-sensitivity enzymatic reporter unlocking
SISPA	:	Sequence independent single primer amplification
SLEV	:	Saint Louis encephalitis virus
SNP	:	Single nucleotide polymorphism
ssDNA	:	single stranded DNA
TAM	:	Tyro3, Axl, and MerTK
TBEV	:	Tick-borne encephalitis virus
TIM	:	T-cell immunoglobulin domain and mucin domain
TTV	:	Torque teno virus
VIRTUS	:	Viral transcript usage sensor
WGA	:	Whole genome amplification
WNV	:	West Nile virus
YFV	:	Yellow fever virus
ZIKV	:	Zika virus

General Introduction

Challenges in diagnosing febrile illness

In the low resources area of the tropics, one of the most common symptoms reported by a person seeking medical care is fever (1). Fever can be accompanied by other signs and symptoms which can highlight the localization of the inflammation/infection, or it can occur in isolation. Though these signs and symptoms can aid in diagnosis, the causal pathogen remains difficult to be differentiated without further testing. On the other hand, a fever without any localization features (referred to as undifferentiated fever) is challenging for a health care worker to diagnose, mainly due to combination of non-specific symptoms during acute phase and the wide range of the aetiologies, that span across both infectious and non-infectious aetiology (1).

In the places where malaria is endemic, occurrence of fever is highly associated with malaria. However, development and deployment of rapid diagnostic tests (RDTs) to detect antigen originating from *Plasmodium* spp., the cause of malaria, has improved the diagnosis and therapy for malaria (2). Decreased incidence of malarial fever has given more space in the spotlight for other infectious aetiologies. In the past years, less common aetiologies of fever, including *Leptospira* spp., *Rickettsia* spp., hepatitis B virus, enterovirus, and cytomegalovirus have been reported (3,4). In the context of therapeutic management, deployment of malaria RDT has led to a more rational usage of antimalarial drugs in sub-Saharan Africa, but unsurprisingly it has also led to excessive prescription of antibacterial drugs (2). Unfortunately, self-limited viral infections (particularly respiratory viruses) were reported to have a high prevalence in primary care, suggesting that the antibacterial drugs might be inappropriately prescribed (2). Tests such as RDT (antigen or serology) are only currently available for several pathogens (*Plasmodium* spp. and dengue virus, among others) and sophisticated tests, such as culture and polymerase chain reaction (PCR), have limited availability. In addition, serology based RDTs have several limitations such as risk of cross-reactivity in closely-related pathogens (i.e. cross-reactivity between flaviviruses (5,6)), requiring extensive effort in development, and low levels of immunoglobulin during the early course of an infectious disease.

As a result, a considerable portion of febrile illness remains undiagnosed. Several studies concluded that up to 40.6% and 74.5% of febrile illness in South Asian and Southeast Asian countries were undiagnosed, respectively (7,8). Another review study showed that in resource-limited settings, the aetiology of acute febrile illness cases could not be identified in 3% to 63% of cases, depending on the testing capacity in the capital or on the aid from foreign countries (9). It is possible that those cases without identifiable pathogens are caused by non-infectious aetiology, such as cancer and autoimmune disease, but a review on the aetiology of fever of unknown origin (FUO) showed otherwise. Fever of unknown origin itself is defined as fever (>38.3°C) in immunocompetent patients lasting for more than 3 weeks with no diagnosis after a standard minimal diagnostic protocol, which include laboratory, microbiology, and imaging tests (10). Studies on FUO showed that by employing extensive testing, infectious aetiology can be detected in 16% to 55% of FUO cases depending on the decade, geographic region, age of the patients, and type of medical practice (11). The infectious aetiology found during the extensive workup suggests that the infection was missed during the initial workup. Collectively, those undiagnosed febrile illnesses might arise from undetected or even unknown infectious aetiology.

Importance of diagnosis of unknown infections

There are compelling reasons to pursue the diagnosis of those febrile illnesses that arise from unknown infectious origins. Firstly, the coronavirus disease 2019 (COVID-19) pandemic in 2019 has shown that early detection of an anomaly in emerging infectious disease is critical. Sequencing of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome through metagenomic sequencing has greatly facilitated the development of diagnosis and the vaccine research (12). Secondly, due to limited therapeutic means and the self-limited nature of many viral infections, viral infections are often overlooked. However, with many antivirals now under development, diagnosis of previously overlooked viral disease might improve the therapeutic outcome. Though the currently commercially available antivirals are limited for several viruses, other antivirals are being developed at an unprecedented rate (13). Finally, sophisticated molecular detection methods have become more affordable, boosting their availability and the usage of molecular approaches in developing countries (14).

Nucleic acid test has become the common approach for diagnosis of febrile illness

Diagnosis using nucleic acid test (NAT) is one of the most popular molecular approaches in diagnosing infectious diseases, and the PCR, both conventional and real-time, are now widely applied as diagnostic tools. Commercialized PCR assays are now available for a wide variety of pathogens, including *Mycobacterium tuberculosis* and its drug susceptibility testing, *Plasmodium* spp., dengue viruses, and SARS-CoV-2 (15–18). Additionally, as countries braced for SARS-CoV-2, they responded by increasing the testing capacity. This provides a huge boost to the availability of NAT, even in developing countries. In Indonesia for example, the government increased the number of laboratories capable of performing PCR and purchased additional thermal cycler machines to increase the testing capacity during the early phase of the pandemic (19). Even before the pandemic, testing for several pathogens have been shifted towards NAT as well. For example, the Pan-American Health Organization published a recommendation for the utilization of reverse transcription polymerase chain reaction (RT-PCR) in diagnosis and surveillance of yellow fever virus (YFV) and Zika virus (ZIKV) (20,21). Other non-PCR NAT approaches have also been developed using loop-mediated isothermal amplification, recombinase polymerase amplification, and multiple displacement amplification (MDA) (22–24). These techniques have also been combined with other nonconventional post-amplification steps, including nucleic-acid chromatography and specific high-sensitivity enzymatic reporter unlocking (SHERLOCK), to increase both the specificity and the portability (25–27).

Comprehensive or semi-comprehensive diagnostic approach using NGS

An ideal pathogen detection test should be free from the limitation of target pathogen. Semi-comprehensive detection with next generation sequencing (NGS) can be achieved by focusing on a genetic region conserved in a certain group of pathogens. These regions are known as “genetic barcodes”. For example, bacterial communities can be profiled by comparing sequences in their 16S rRNA gene. Several variable regions exist in this gene which can be used to identify the bacteria in the sample. The more variable the targeted region is, the more specific the taxonomic rank that can be assigned to a sequence. For example, analyzing only certain regions of the 16s rRNA gene might only result in

species- or family-level resolution, but full-length 16S rRNA sequencing allows identification of a bacterium up to the strain level (28). Other genetic barcodes exist in different organisms, for example 18S and 28s rRNA gene has been used to detect a wide variety of parasitic taxa from samples (including protozoa and helminths), while ITS region have been used to explore the fungal community (or mycobiome) (29,30). Even in RNA viruses, several conserved regions can be found which are mainly located in gene encoding RNA-dependent RNA polymerase (RdRp). This is true at least for family *Flavivirus*, *Coronavirus*, and *Paramyxovirus*, in which primer sets targeting this region can amplify target sequences originated from members of this family (31–33).

On the contrary, fully comprehensive NATs should be free from target limitations and applicable to any pathogens. This can be achieved through shotgun metagenomic sequencing, where all the nucleic acids present in the sample will be sequenced and analyzed. This method has greatly improved in the past years, supposedly due to improvements and developments in the sequencing technology. The development of nanopore sequencing technology, for example, has facilitated rapid turnaround time for metagenomic NGS (mNGS) (6 to 8 hours) (34–36). On the other hand, short-read platform mNGS is often favoured because it has less errors and has larger output in comparison to its long-read counterpart. Indeed, utilizations of mNGS in clinical settings have improved the diagnosis of infectious diseases (37–39). Metagenomic NGS is particularly successful in diagnosing infections of the central nervous system. In a case series, mNGS was able to detect a case of neuroleptospirosis and a case of neuroinvasive astrovirus, which were not detected using the standard practices (40).

The spread of SARS-CoV-2 has a silver lining, in which it boosted the availability of NAT and sequencing technology across the world. Thus, it is not impossible that sequencing would be adapted as the standard practice in the future. In order to establish mNGS as a standard practice, several studies have tried to compare mNGS and the conventional practice for pathogen detection. Though in general mNGS showed good performances, conflicting results were observed with mNGS having better performance or vice versa, with some of these studies showed that mNGS could detect more pathogens (41). At present, a commercial mNGS assay for pathogen detection in cell-free DNA is available (Karius test), further showcasing the potential of mNGS application as a diagnostic tool (42). For mNGS to be a standard practice, there are some pitfalls that need

to be addressed including running cost, turn-around time, library preparation, computational burden, and user-friendly analysis, among others. Fortunately, the recent development of portable sequencer provides several solutions, including reduced instrument and sequencing cost, easier library preparation, increased portability, and most importantly the ability to provide real-time sequencing data which is ideal for diagnostic purposes. These advances can greatly assist mNGS in peripheral laboratories. Given the potential for mNGS as a pathogen detection platform and the opportunity that arose from the development of portable sequencing, this study was designed to develop application of mNGS as a pathogen detection system with emphasis on utilization of portable sequencer and in optimizing the library preparation.

In Chapter One, the application of NGS to provide a semi-comprehensive approach of viral pathogens at a genus (or a family) level was explored using a targeted-sequencing approach. Viral pathogen shows a high diversity, making targeted sequencing approach (i.e. DNA barcode) ineffective. However, when the target resolution is narrowed down to a genus or a family level, a conserved region can exist among its members, providing a foundation for the semi-comprehensive approach. The amplified target regions then can be sequenced with NGS to identify the pathogens. Because the genus *Flavivirus* contains some important human pathogens and its members co-exist in highly populated places in the tropics, it was used as a model for a semi-comprehensive detection system. In this system, a broad-range RT-PCR for flaviviruses was combined with nanopore sequencing. Additionally, as library preparation remains a bottleneck, a multiplexing system using indexed primer was also developed and validated. This system was validated using a series of experiments using both spiked samples and clinical samples.

Having successfully developed a semi-comprehensive detection system for flavivirus, the stake was raised to then develop a comprehensive detection system for viral pathogens. In Chapter Two, an alternative unbiased amplification was developed and validated. By exploiting the limitations of phi29 enzyme being biased towards the circular template, the complementary DNA (cDNA) resulted from reverse transcription was circularized prior to non-specific amplification; hence it was designated as circular whole transcriptome amplification (cWTA). It was shown that the circularization improved the amplification and that the amplicons can be analyzed using NGS (nanopore sequencing).

However, an amplification bias was observed which indicate the need for future study to assess the bias systematically.

In Chapter Three, a method to minimize the library size that need to be constructed for large scale mNGS screening was applied, in combination with cWTA. Using a group testing algorithm, instead of constructed individual libraries for every sample, the samples were pooled according to a combinatorial group testing algorithm, which significantly reduced the number of libraries that needed to be constructed. This method was able to detect sequences homologous to several viruses with different genetic characteristics without prior knowledge. Thus, this approach provides an effort-saving, unbiased, hypothesis free approach for viral detection.

Chapter One

Universal High-Throughput Detection and Identification of Flaviviruses using Nanopore Sequencing

1.1 Summary

Identification of flaviviruses is usually carried out using NAT, especially quantitative PCR. Several semi-comprehensive NATs offer a genus-level identification of flaviviruses, but further species classification requires a post amplification analysis. To overcome this issue, a pan-flavivirus RT-PCR was combined with the portable nanopore sequencing, to provide a platform for broad-spectrum flavivirus detection. In addition, a sample multiplexing system was developed by modifying the primers to include unique nucleotide sequences at the 5'-end. This multiplexing system increased the number of samples that can be sequenced in one flowcell, further reducing the cost per sample. A streamlined bioinformatic pipeline was also developed and it enabled a plausible cut-off value for observed read counts to be defined. Through a series of validation, the system was shown to be able to amplify and detect a broad range of flavivirus. In addition, the rate of error in index assignment was found to be low (0.02%). The system was deployed to detect flavivirus from two different sets of clinical samples. Among the first set of clinical samples which were collected in Vietnam, DENV1, DENV2, and DENV4 were detected. The observed positive and negative agreement in comparison to a commercial NAT were 66.7% and 95.4%, respectively. The combination of pan-flavivirus and nanopore sequencing system was able to obtain more positive samples than a commercial NAT, suggesting a comparable performance. In a second clinical sample set, diverse flaviviruses were able to be detected and thus, supporting the broad-range aspect of the system. Collectively, a semi-comprehensive sequencing-based diagnostic system for the detection of flavivirus was developed, with a reasonable cost, considerable sensitivity, and a relatively easy procedure.

1.2 Introduction

Flaviviruses (family *Flaviviridae,* genus *Flavivirus*) are principally vectored by arthropods and this family consists of some important human pathogens such as YFV, ZIKV, Japanese encephalitis virus (JEV), St. Louis encephalitis virus (SLEV), dengue virus (DENV), West Nile virus (WNV), and tick-borne encephalitis virus (TBEV). Multiple outbreaks caused by these viruses have been reported across the globe, showing their significance to the global health. These include the 2015 yellow fever outbreak in Angola and its neighbouring countries, the recent 2022 JEV outbreak in Australia, and the ZIKV epidemic from 2015 to 2016 in the Americas (43–45). Meanwhile, in the Western Hemisphere, introduction of WNV was followed by a rapid geographical spread, with numerous cases of infection in humans with considerable mortality (46). Other members of the groups, including DENV, JEV, and TBEV put thousands of individuals at risk of infection. For example, up to a quarter of the world's population live in places where DENV is endemic and it is estimated to infect approximately 400 million people each year (46).

Flaviviruses are small (~50 nm), enveloped, spherical, positive-sense single-strand RNA viruses (46,47). The genome size varies between 10.5 kB to 11 kB, which contains a single open reading frame (ORF) (47). The endoplasmic reticulum (ER) translates this ORF into a polyprotein, which is then cleaved by proteases, resulting in 10 functional proteins, divided into structural (C, prM, E) and non-structural (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) proteins (47). Viral entry to the cell is mediated through interaction of structural proteins with multiple C-type lectins, binding of E protein to the glycosaminoglycans, and interactions between the viral lipid envelope to several surface proteins, including TIM (T-cell immunoglobulin domain and mucin domain) and TAM (Tyro3, Axl and, MerTK) family of phosphatidylserine receptors (46). Replication of the virus happens in the cytoplasm, initiated by translation of the viral RNA and then synthesis of complementary negative-sense RNA which serves as template for positive-sense RNA replication (47). Virus assembly happens in the ER and are released through exocytosis or cell lysis (47).

From the clinical aspects, diagnosing a flavivirus infection during the acute phase is challenging. Acute flavivirus infection manifests as an undifferentiated fever accompanied by other non-specific manifestations, making clinical diagnosis trivial (48).

Commonly used serological assays have potentials for cross-reactivity which is emphasized by the overlapping distribution of flaviviruses. Serological assay also has a limitation during the acute phase due to a low titer of antibody (5,49,50). Commercially available antigen test targeting specific circulating non-structural protein 1 (NS1) is limited to DENV. This test for DENV, however, could not differentiate the serotype, which is important in dengue infection cases due to potential of antibody-dependent enhancement phenomenon (51). Additionally, the sensitivity and specificity of the NS1 antigen tests vary, depending on which kit is used. In a previous report, the sensitivity and specificity of one NS1 antigen RDT were 81.5% and 66.7%, respectively, which is lower than those of enzyme-linked immunosorbent assay (ELISA) (89.9% sensitivity and 100% specificity) (52). Interestingly, one study reported that at high concentration, ZIKV, YFV, and Kunjin virus can give a false positive result when using DENV NS1 antigen RDT (6). Other than DENV, a similar antigen test targeting NS1 of YFV has been described, but it is not currently available for commercial use (53).

Due to limitations of the other tests, NAT has been drawing attention. The Pan American Health Organization has recommended RT-PCR (either conventional or quantitative) for yellow fever diagnosis and for ZIKV laboratory surveillance (20,21). In South American countries, reverse transcription quantitative polymerase chain reaction (RT-qPCR) is the benchmark method for diagnosis of yellow fever (54). While the United States has developed a multiplex DENV1-4 RT-qPCR assay and it has been adopted by the Center for Disease Control and Prevention (CDC) as one of the diagnostic approaches for diagnosing Dengue. This multiplex DENV test has been validated and shown to be reliable (17). Nevertheless, there are some limitations of NAT. Firstly, they require advanced and sophisticated equipment that hinders its application outside reference diagnostic centers and laboratories. Secondly, as these tests target a specific pathogen, a hypothesis of target pathogen should be made prior to testing. Abundance of causal pathogens that manifest as undifferentiated fever can muddle the hypothesis, leading to oversight of less common pathogens.

While the high sensitivity of NAT can be an upside, it can also be a limitation, i.e., multiple tests would be required to determine the causal pathogen. On the other hand, a broad-spectrum NAT increases the coverage of pathogens that can be detected using a test and can reduce the number of tests that need to be done. In flavivirus, a conserved

region at the region that encodes non-structural 5 protein (NS5) is ideal for a broad-spectrum flavivirus NAT. This region has been targeted for broad diagnosis of flaviviruses (55–59). Several RT-qPCR based pan-flavivirus approaches have already been published; some include post-amplification analysis to identify the viral species (31,60). Though conventional RT-PCR can be an affordable NAT approach for pan-flavivirus diagnosis, similar amplicon size limits the usage of conventional post-amplification analysis (i.e. gel electrophoresis) in differentiating the viruses. Multiplex RT-qPCR can address this problem by utilizing a probe; however, this approach requires sophisticated equipment and specialised training, but also designing a probe can be complicated.

A straightforward way to discriminate similar-sized amplicons would be sequencing. Although the conventional Sanger sequencing is low-cost, it is less suitable for targeted sequencing approaches, because presence of several amplicons in a sample will result in double peaks. Therefore, the result will only represent the dominant amplicons or when the amplicons are equally dominant, then the result can be uninterpretable. In addition, Sanger sequencing requires advanced machinery and high maintenance; a limitation presented by other NGSs as well. One NGS, the nanopore sequencer, has leveraged the sequencing process by providing portability through minimization of the laboratory-oriented processes. In addition, the nanopore sequencing can provide real time data which would be beneficial for diagnostic purposes. However, the platform has issues related to accuracy and cost. As a trade-off for its size and portability, the nanopore sequencing platform (prior to the release of new flowcell and chemistry version) lacks the accuracy when compared to other NGS platforms (5% to 15% error rate) (61). However, the producer (Oxford Nanopore Technology) promised an improved sequencing accuracy with the brand-new chemistry (the so called Q20) and flowcell (R10.4). It is currently on an early access program and only a handful of peer reviewed papers are published, with many still in the preprint server. Among those published, a study looking at a mock microbial community found a striking 99% of sequencing accuracy using the Q20 chemistry and R10.4 flowcell (62). In terms of cost, it depends on the type of the flowcell used, with running costs ranging from ~USD150 to ~USD600 per experiment. Running costs can be reduced by multiplexing samples to maximize sample throughput per run.

The aim of this study is to establish a broad-spectrum NAT for flavivirus identification. To achieve this, a combination of pan-flavivirus PCR assay and nanopore sequencing was employed. In addition, a multiplexing system using unique oligonucleotide indexes was included to maximize the sample throughput per sequencing run and to reduce the running costs.

1.3 Material and Methods

1.3.1 Primer modifications and index design

A degenerate primer set targeting the NS5 gene of flaviviruses described in another study was used to amplify the target region (31). The primer set consists of three primers; including an additional sense primer to improve the amplification for DENV4. The primer set produces approximately 260 base-pair (bp) long amplicons **(Table 1.1)**. The primer set was modified by concatenating a 26-mer unique oligonucleotide at the 5' end. These oligonucleotides are not to be confused with barcodes provided by Oxford Nanopore technologies which also will be mentioned later. The indexes were generated using Filled/truncated right end edit Barcodes (FREEBarcodes) (63). There are several criteria used by this package when generating barcodes, including balanced GC content (40% to 60%), no homopolymers, no triplet self-complementarity, and no GGC. At first, a set of 12-mer and 14-mer index sequences were generated. These sequences were then concatenated, to create the 26-mer index sequences. These indexes were then aligned to a database containing the newly designed index sequences using Local Alignment Search Tool (LAST) to ensure a maximum dissimilarity (64). When an index was found to be aligned to other indexes or to itself, then that index was excluded. Twenty-four index sequences were obtained at the end. These indexes were divided into two groups, the forward indexes were concatenated to the sense primers and the reverse indexes were concatenated to the anti-sense primer. Theoretically, by using different forward and reverse index combinations, a maximum of 144 combinations can be obtained. In addition, a different set of indexed primers (contains 14-mer index sequences) was also used as a prototype for the system.

1.3.2 Pan-flavivirus RT-PCR

One-step RT-PCR was carried on in a total volume of 15 µl using PrimeScript One Step RT-PCR Kit Ver.2 (TaKaRa). Each reaction consists of 7.5 µl 2× one step buffer, sense primers (Flavi_all_S and DEN4_F) at a final concentration of 250 nM each, primer Flavi_all_AS_2 at a final concentration of 500 nM, 1.2 µl PrimeScript II enzyme mix (TaKaRa) and nuclease-free water up to 30 µl. Conditions for the RT-PCR were as follows: 50°C for 30 minutes (cDNA synthesis), 94°C for 30 seconds, followed by 43 cycles of 94°C, 53°C, 72°C, 30 seconds each, and lastly 72°C for 5 minutes. The RT-PCR

products were then visualized on 1.5% agarose gels. The enzyme mix contains both Primescript RTase (TaKaRa) and TaKaRa Ex Taq Hot Start (TaKaRa),

1.3.3 Nanopore sequencing

Two different nanopore flowcells were used in this study: Flongle and MinION flowcell. The main difference between the two flowcells is the number of sequencing pores; a MinION flowcell has 2,048 nanopores while a Flongle flowcell has 126 nanopores. The difference in pore numbers subsequently affects the sequencing yield and depth. Flowcell quality control was performed prior to the sequencing to ensure adequate number of sequencing pores (at least 50% active pores).

Prior to sequencing, amplicons were purified using an equal volume of AMPure XP beads (Beckman Coulter) and eluted in 46 µl of nuclease-free water. Concentrations of each sample were measured using a Qubit fluorometer (Invitrogen). When using a MinION flowcell, 45 µl of the purified amplicons were subjected to end repair and dA-tailing in a reaction containing 7 µl Ultra II End-prep reaction buffer and 3 µl Ultra II End-prep enzyme mix (New England Biolabs), and nuclease-free water to a final volume of 60 µl. The reaction was incubated at 20°C for 5 minutes and at 65°C for 5 minutes. The end-prepped amplicons were then purified with an equal volume of AMPure XP beads (Beckman Coulter), and finally eluted in 25 µl of nuclease-free water. Reactions were carried out in half of the volume when using a Flongle flowcell.

When necessary, each sample was barcoded using Oxford Nanopore native barcode. Barcoding kit EXP-NBD103 and EXP-NBD114 (Oxford Nanopore) were used in this study. In experiments using MinION flowcell, ligation and tethering of the barcode were carried out using 22.5 µl of purified amplicons, 25 µl Blunt/TA ligase master mix and 2.5 µl native barcode. The reaction was incubated at 23°C for 10 minutes, then purified with equal volume of AMPure XP beads (Beckman Coulter), and then eluted in 10 µl of nuclease-free water. Reactions were carried out in half of the volume when using a Flongle flowcell.

The concentrations of each sample were measured using a Qubit fluorometer. For nanopore library construction, library kit SQK-LSK109 was used. For experiments using a MinION flowcell, equimolar amounts of barcoded samples were then pooled and diluted in nuclease-free water to a final volume of 50 µl. Purified amplicons or purified

barcoded amplicons were then subjected to adapter ligation and tethering. The reaction was carried out with 20 µl Barcode Adapter Mix, 20 µl NEB Next Quick Ligation Reaction Buffer (5×), and 10 µl Quick T4 DNA Ligase. The adapter-ligated sample was then incubated at 23°C for 10 minutes, then purified with 0.4× AMPure XP beads (Beckman Coulter), washed with Short Fragment Buffer, and eluted in 15 µl elution buffer. The MinION flowcell (version 9.4) was primed with 1 ml of a mix of 30 µl Flush Tether and 1,170 µl of Flush Buffer. Twelve µl of the adapter-ligated sample, 37.5 µl sequencing buffer and 22.5 µl loading beads were mixed and loaded into the flowcell. Sequencing was performed for up to 48 hours.

For experiments using a Flongle flowcell, purified amplicons or purified barcoded amplicons were pooled in a final volume of 32.5 µl. Adapter ligation was carried out using 2.5 µl Adapter Mix II, 10 µl NEB Next Quick Ligation Reaction Buffer (5×), and 5 µl Quick T4 DNA Ligase. The reaction was incubated at 23°C for 10 minutes. The adapter-ligated sample was then purified with 0.5× AMPure XP beads (Beckman Coulter), washed with short fragment buffer and eluted in 7 µl of elution buffer. The Flongle flowcell (version 9.4) was primed with 100 µl of a mix of 3 µl Flush Tether and 117 µl of Flush Buffer mix. Five µl of the adapter-ligated sample, 15 µl sequencing buffer and 10 µl loading beads were mixed and loaded into the flowcell. Sequencing was performed for up to 18 hours.

1.3.4 Sequence demultiplexing

Raw FAST5 files obtained from nanopore sequencer were basecalled using Guppy v 3.0 (Oxford Nanopore Technologies) to generate fastq files. Threshold for quality score was set to seven and only reads in the "pass" folder were used for downstream analysis. In case where a sample is barcoded using Nanopore native barcode, Guppy v 3.0 was used to demultiplex the reads according to Nanopore native barcode. A stricter parameter (barcode score of 90) was used, to limit the possible crosstalk which was reported previously (65).

Reads passing the filter above were then subjected to index demultiplexing. Three different tools/packages were evaluated for demultiplexing reads; FREEBarcodes (63), LAST (64), and Minibar (66). The simplified depiction of the pipeline can be seen in **Fig. 1.1.** The reads were filtered by length; only reads with length of 250-500 bp long were

included in the downstream analysis. Different parameters were applied for each pipeline, to ensure the best result. Then the read recovery rate and error rates were calculated for each sample. Reads obtained after the deindexing process will be referred as deindexed reads.

For the LAST-based pipeline, a database containing the primer sequences (including the index) was generated using the option -uNEAR -R01. The debarcoded reads were then aligned with the database using following options: -Q1 -q2 -r2 -a1 -b1 -e 20. Based on the highest alignment score, a forward and a reverse index will be assigned for each read. When demultiplexing index using Minibar, different edit distances were applied. The developer suggested to infer the ideal edit distance from the sequencing error rate of Nanopore sequencer, which is between 12% to 22% (66). This translates to an edit distance between 3 to 5, for a 26-mer index sequence. When deindexing the reads using FREEBarcodes, the adjusted parameter was filled/truncated right-end (FRE) *m*-edit, which essentially corresponds to the number of correctable errors.

The index-demultiplexed reads were then converted into fasta files for alignment searches using BLAST (basic local alignment search tool) (67). A local virus database which was constructed for READSCAN was used for the BLAST search (68). Reads were then filtered based on percentage identity (above 80%) and alignment length (250-290).

1.3.5 Validation of pan-flavivirus RT-PCR using the modified primer

The validation was carried out in two steps. Firstly, to imitate a serum sample, fetal bovine serum (FBS) was spiked with DENV1, DENV2, or YFV to achieve a final titer ranging from 10 to 10^5 PFU/mL. RNA was extracted from samples using QIAamp viral RNA mini kit (QIAGEN), according to the manufacturer's instructions. Then those spiked samples were subjected to pan-flavivirus RT-PCR as described in section 1.3.2.

Secondly, to demonstrate and evaluate the spectrum of flavivirus that can be detected, the indexed primer was used to amplify target sequences from nine different flaviviruses in comparison to the original, index-free primer. For a more accurate calculation of the template concentration and due to limited availability of template RNA, target sequence was cloned into a vector and transfected to *Escherichia coli*. In brief, viral RNA (DENV1, DENV2, DENV3, DENV4, JEV, YFV, WNV, ZIKV, and TBEV) was

subjected to RT-PCR as described in section 1.3.2, but using index-free primer. Up to 30 µl of amplicon were used for gel electrophoresis, using low-melt agarose. The gel was then cut, melted at 65°C, then directly ligated to pGEM-T vectors (Promega) and transformed into *E. coli* DH5α. Overnight grown colonies containing plasmid with insertions were then allowed to proceed to expansion for 12 hours. Plasmids were then purified using Wizard Plus SV Minipreps (Promega) according to the manufacturer's instructions. Colony PCR followed by Sanger sequencing was performed to ensure the correct target is inserted and to get the exact size of the inserts. The concentration was then measured using a Qubit fluorometer (Invitrogen). The estimated copy number was calculated using the following formula:

$$\text{copy number} = \frac{\text{DNA concentration (ng/µl)} \times 6.022 \times 10^{23}}{(\text{vector length} + \text{insert length}) \times 10^{9} \times 650}$$

The purified plasmids were then diluted to achieve plasmid concentration of 10^8 copies/ µl, then serially diluted up to 10^0 copies/µl. This serially diluted plasmid was then subjected to PCR using TaKaRa Ex Taq Hot Start (Takara), which is an identical polymerase used in the one-step RT-PCR. The PCR was carried in a total reaction volume of 15 µl containing 1 µl template, 250 nM each of indexed sense primers, 500 nM of indexed anti-sense primer. Cycling program was as follows: 94°C for 30 seconds, followed by 43 cycles of 94°C, 53°C, 72°C, 30 seconds each, and lastly 72°C for 5 minutes. Amplicons (1 µl) were then visualized in 1.5% agarose gel.

1.3.6 System validation using spiked sample

Three sequencing experiments were carried out to evaluate the demultiplexing tools described in section 1.3.4. In each experiment, a total of 12 FBS samples were spiked with either DENV1 (10^5 plaque forming unit (PFU)/mL), DENV2 (10^4 PFU/mL), or YFV (10^5 PFU/mL); thus, there were four samples spiked with each virus in each experiment. These samples were subjected to RNA extraction and RT-PCR as above. The samples were barcoded using Oxford Nanopore native barcoding kit, then sequenced. The barcodes add a layer of information when decoding the sample origin of a read. The combination of index, barcode, and virus can be seen in **Table 1.3**. The sequencing data was then used to evaluate the demultiplexing tools and to calculate the demultiplexing error rate.

To demonstrate capability of nanopore sequencer and the downstream analysis to identify the flavivirus, additional experiments using plasmid-containing target sequences were carried out. Firstly, to provide information about the broad range of the system, the samples with the lowest template (plasmid) concentration on which an unambiguously visible band can be observed on the gel electrophoresis were identified (section 1.3.5). These samples were then subjected to pan-flavivirus RT-PCR then sequenced on a Flongle flowcell. Secondly, to explore the potential of deep sequencing with MinION in detecting the target sequence when an unambiguously visible band is not present, samples with template concentrations of 10^{-1} lower than the concentration mentioned above were subjected to pan-flavivirus RT-PCR and then sequenced with MinION.

1.3.7 Reads classification and threshold calculation

In order to evaluate each tool's performance in demultiplexing the reads, it is important to have a clear definition of the read classification. Essentially, a read will contain three different information: barcode, index, and viral sequences. When one information is incorrect, the other two information can be used to infer the sample origin and the incorrect information is assumed to be caused by an error related to that step. At first, the reads were classified as either true or false. True reads have all the decoded information to match the information in the experiment's plan, or simply the reads were classified into the correct pair of barcode, matched index, and virus (**Fig. 1.1**). When one or more information does not match, the read is classified as false. The false reads can be classified further based on which information was incorrectly decoded. Reads with matched index and virus sequence but have mismatched barcode were classified as false results in barcode. When index is the mismatched information, the reads are classified as false results in index and when the virus sequence is the mismatched information, the reads are classified as false results in sequencing.

Results with very limited read numbers (one or two reads) will be automatically omitted from the result. The false results in index represents the error rate (of index demultiplexing) and assuming that the false results in index rate would be constant, the reliability of the detection, i.e. threshold for positive or negative results, can be calculated based on the false results in index rate. Another assumption is that error in classifying

index sequence would happen only in one of the two indexes (either forward or reverse); thus, crosstalk sequences are expected to come from neighbouring samples with identical forward or reverse index. Reads resulting from index crosstalk were estimated through a calculation based on the error rate with an arbitrary safety margin of 10 times to minimize false positive results. The expected number of crosstalk reads was plotted to Poisson distribution. In the event (μ) where the upper cumulative Poisson probability ($P(x \geq \mu)$) was less than 0.01, that event will be determined as the threshold. Result calculation was based only on reads that have BLAST hit and counted at the result. If the number of reads from a certain virus passed the crosstalk threshold, the sample will be considered positive.

1.3.8 System validation using clinical samples

Clinical samples were provided by the National Institute of Hygiene and Epidemiology (NIHE) Vietnam. A total of 114 serum samples were collected from patients with undifferentiated fever in Namh Dinh province between May 2017 and May 2019. The samples were stored at -80^{o}C prior to the experiment. Among these samples, 71 were tested positive for an ELISA NS1 antigen test (Inbios). The NS1 antigen test positive samples were also subjected to CDC DENV-1-4 RT-qPCR Multiplex Assay as described previously (17). Six samples were found to be NS1 antigen test positive and RT-qPCR positive **(Fig. 1.2)**. These 114 samples were subjected to pan-flavivirus RT-PCR using the indexed primers. Combinations of forward and reverse index can be seen in **Fig. 1.2**. The amplicons were then pooled and sequenced using a Flongle flowcell. Samples with negative results from Flongle sequencing were pooled and then sequenced with a MinION flowcell.

Second set of clinical samples were obtained from flavivirus Laboratory, the Oswaldo Cruz Foundation, Brazil. A total of 24 serum samples from undifferentiated fever patients stored in the laboratory were used in this study. These samples were subjected to a flavivirus RT-qPCR as described previously (56). Positive samples were then subjected to virus-specific protocols as previously described (17,69–72). These serum samples were subjected to pan-flavivirus RT-PCR using the 14-mer indexed primer (**Table 1.2**). Amplicons were pooled and sequenced using a MinION flowcell.

Ethical permission was obtained both from NIHE (351/QD-VSDTTU) and Hokkaido University (Jinjyu1-3). For samples from Brazil, ethical permissions were

obtained from Oswaldo Cruz Foundation (no. 2.998.362) and Hokkaido University (Jinjyu30-4).

1.4 Results

1.4.1 Pan-flavivirus RT-PCR can amplify the genome of designated flavivirus

A set of index sequences were generated using FREEBarcodes (**Table 1.2**). The RT-PCR were performed using indexed primers using RNA extracted from the serially diluted FBS spiked with virus. The modified primers were shown to successfully amplify the target sequence from DENV1, DENV2, and YFV from the spiked samples up to 10 PFU/mL, 1 PFU/mL, and 10^4 PFU/mL, respectively (**Fig. 1.3)**. Then, the broad spectrum of the modified primer was validated using plasmid containing the target sequence. On gel electrophoresis, amplicons were unambiguously observed from reaction containing 10^3 copies (DENV1, DENV2, DENV3, ZIKV, WNV, TBEV, and YFV), and 10^2 copies (DENV4 and JEV) of template per reaction **(Fig. 1.4)**. However, a decrease in PCR efficacy is observed when the result of indexed primer is compared to index-free primers. In PCR reactions using original index-free primers, amplicons were visible in reactions with 10 to 10^2 copies of template per reaction lower compared to when using indexed primers (**Fig. 1.4**).

1.4.2 Optimization of bioinformatic parameters and threshold calculation

Three sets of spiked samples were processed with the system. Each set consists of 12 samples, that contains either 10^5 PFU/mL DENV1, 10^5 PFU/mL YFV, or 10^4 PFU/mL DENV2. One set of spiked samples was sequenced using MinION flowcell, while the remaining two sets were sequenced using Flongle flowcells. Three different tools were used and compared for the deindexing; FREEBarcodes, LAST version 9.16 (64), and Minibar version 0.2.1 (66). Based on these data, the overall recovery rate (percentage of index-demultiplexed reads to barcode-demultiplexed reads) was less than 50% in any of the three methods (**Table 1.4)**. The reads were then subjected to index demultiplexing using the three demultiplexing tools, while employing different parameters. Among the three tools evaluated, FREEBarcodes yielded low false results in index rate, but it was also the tool with the lowest recovery rate (**Fig. 1.5**). On the other hand, Minibar has a higher recovery rate, but it suffers from high false results in index rate (ranging from 1.11% to 3.11%). Therefore, the LAST based pipeline was adopted because it has low false results in index rate with a decent recovery rate (**Table 1.4** and **Fig. 1.5)**. Alignment score threshold of 70 was used to balance the recovery rate and error rate. By employing

LAST and the threshold, the system can correctly identify the viruses that were spiked in the sample in all three experiments (**Table 1.5, 1.6, and 1.7)**. In addition, the other parameters (such as recovery rate and false results in index rate) showed consistent results across the experiments.

Despite the adjustment of alignment score, the false results in index rate when demultiplexing using LAST was unable to be zero. It will make it difficult to discriminate true reads from and artifact derived from the false reads. Therefore, a method to determine the threshold for positive or negative call was developed. Firstly, it is assumed that the false results in index reads are most likely to be a crosstalk originating from samples sharing identical forward or reverse index. Secondly, it is assumed that the false results in index rate is constant and follows a Poisson distribution. Therefore, the read threshold was determined by assuming Poisson distribution given by the observed false result in index rate and the total number of reads from the neighbouring samples (**Fig. 1.6A**). Based on the initial sequencing experiments, the adopted sequencing system, and the bioinformatic pipeline, the error rate was found to be 0.02%. This rate was multiplied by 10 as a safety margin. As an example, in reads assigned to barcode 02 there were 1,055 and one DENV2 reads for the sample with i02-i14 and i07-i14 index pair, respectively (**Fig. 1.6A**). Given that the i07-i14 index pair was not used in this experiment and that index i14 was shared with the correct barcode and virus combination (DENV2 and i02-i14) (**Table 1.3**), while index i07 was shared with DENV1 and i07-i19, then it is highly likely that the read with error is originated from the sample with DENV2/i02-i14 combination. It is assumed that in that one read, i02 was incorrectly decoded as i07, resulting in the crosstalk. If such errors are assumed to occur in stochastic manner following Poisson distribution with the rate (λ) derived from number of reads (1,056) and error ratio (0.02%), then in this case when seven or more reads were observed, the probability that those reads are a result from crosstalk event is low ($P(x \geq 7) = 0.006$) (**Fig. 1.6A**).

The amplicons obtained from PCR with indexed primers were sequenced with Flongle, resulted in 244,851 raw reads, 53,894 debarcoded reads, and 16,366 deindexed reads. Alignment search results showed that the system can differentiate all nine flavivirus in the sample (**Table 1.8**). Reactions containing lower concentrations of template (10^{-1}) from the concentrations which resulted in visible band (**Table 1.8**) were then sequenced

using a MinION flowcell. The sequencing resulted in 3,377 raw reads, where substantial numbers of reads homologous to DENV3 and DENV4 can be detected (**Table 1.9**). Collectively, it was observed that reads could be detected at least from 10^3 copies (DENV1, DENV2, JEV, TBEV, WNV, YFV, and ZIKV), 10^2 copies (DENV3), and 10^1 copies (DENV4) of template per reaction.

1.4.3 Detection of flavivirus from the clinical samples

Result of pan-flavivirus RT-PCR for 114 samples from Vietnam can be seen in **Fig. 1.7.** From the electrophoresis, bands corresponding to the expected size were observed from three samples: sample 8 (i08-i20), sample 67 (i06-i19), and sample 71 (i06-i23). Sequencing of 114 clinical sample from Vietnam using Flongle yielded 3,126 deindexed reads, meanwhile deep sequencing of invisible amplicons using MinION yielded 80,579 deindexed reads. Distribution of deindexed reads for each sample can be seen in **Table 1.10**. Collectively, three dengue virus serotypes were detected in this test: DENV1, DENV2, and DENV4. Three samples (i01-i20, i06-i19, and i06-i23) showed bands on electrophoresis gel and all three were successfully sequenced by Flongle (**Table 1.11**). When conducted deep sequencing using MinION, reads homologous to dengue were obtained from additional five samples (i04-i15, i06-i15, i06-i20, i06-i22, and i08-i20) which do not show band on electrophoresis gel (**Table 1.12**). Interestingly, one of them (i08-i20) was obtained from a NS1 antigen test negative sample collected one day after fever onset (**Table 1.13**). In the result, 20 and 12 samples with limited number of DENV reads by Flongle and MinION, respectively, were regarded as negative because the numbers were below their threshold. For example, when the samples were sequenced by Flongle, three DENV1 reads were acquired from the i01-i19 indexed sample, but this samples were considered negative when subjected to our calculation (λ=0.48, $P(x\geq3)$ = 0.013, **Fig. 1.6B**); therefore, the sample were regarded as negative. Meanwhile in the result of MinION sequencing, a sample with i10-i13 index pair showed a very limited number of reads yet it is still above the threshold. This sample will be regarded as false positive as well, due to limited read number (< 3 reads). In comparison to the NS1 test, the positive agreement and negative agreement of the sequencing test using Flongle were 4.2% and 100%, respectively and those with the MinION-integrated test were 9.9% and 97.7%. In contrast, in comparison to the RT-qPCR test, the positive agreement and

negative agreement of the sequencing test using Flongle were 33.3% and 98.5%, respectively. Those of the MinION-integrated test were 66.7% and 95.4%. On the other hand, when compared to sequencing test using Flongle, the positive agreement and negative agreement of RT-qPCR were 66.7% and 94.1% respectively and when it is compared to MinION-integrated test, the sensitivity and specificity were 57.1% and 96.9%, respectively (**Table 1.14**).

For the 24 clinical samples from Brazil, the pan-flavivirus RT-PCR was also conducted using a prototype primer with shorter index sequence **(Table 1.3)**. Unambiguous bands were observed from 12 samples. Additionally, there were five samples with ambiguous results and seven samples with negative results. Sequencing of these PCR products, including the ambiguous and negative samples, with MinION yielded 1,698,573 reads and deindexing with LAST resulted in 1,316,420 deindexed reads. Then sequences homologous to SLEV, YFV, WNV, DENV, and ZIKV were identified (**Table 1.15**). Detection of flavivirus by RT-qPCR and the pan-flavivirus/sequencing system was mostly comparable for the positive amplicons; however, several discrepancies were observed for the ambiguous and negative samples. From this data, it was demonstrated that the system can detect broad-spectrum flavivirus from clinical samples.

1.5 Discussion

In this study, a targeted sequencing approach for detection and identification of flaviviruses was developed through a combination of pan-flavivirus RT-PCR and nanopore sequencing. The primers were modified by adding index sequences, allowing samples to be multiplexed. The RT-PCR system with modified primer was tested using FBS spiked with virus. When using target sequence-containing plasmid, detection of up to 10^2 or 10^3 copies of plasmid per reaction was achieved. Assuming 100% conversion of RNA to cDNA, 2:1 extraction volume (140 µl sample, eluted to 60 µl elution), and reaction was carried using 2 µl of extracted RNA, then the initial sample concentration can be inferred to be approximately 2×10^4 to 2×10^5 copies/mL. The viral load during some flavivirus infection were reported in a range from 10^2 to 10^8 RNA copies/mL and thus, cases with low viral load might not be detected by pan-flavivirus RT-PCR using the modified primers (73–77). The pan-flavivirus primer (without index modification) was able to detect 10-100 copies lower when compared to the modified primers (31). It strongly suggests that the modification is affecting the PCR efficacy. The long index sequence might affect the primer's annealing to the template and/or inhibit reaction through primer dimer formation. Nevertheless, the new R10.4 flowcell from nanopore with improved raw sequencing accuracy should allow a shorter index to be used, and subsequently improve the detection sensitivity.

Several demultiplexing pipelines were evaluated. One of them, the FREEBarcodes showed a low recovery rate, which is supposedly due to issue in decoding the index at the 3'-end. Because FREEBarcodes cannot decode dual index at the same time, indexes at the 5'-end and 3'-end of a read were decoded separately, in which a striking difference between the number of reads with decoded 5'-end index and 3'-end index was observed. Closer inspection of the sequencing at the 3'-end of the read, showed a tendency to have some additional bases (just before the index sequences) which affected the decoding process of FREEBarcodes. This observation suggests that an approach using a single index would be ideal when employing FREEBarcodes.

When the samples were multiplexed, a limited but substantial number of reads were detected from unassigned index sets. These reads are highly likely to originate from migration of reads from samples sharing identical forward or reverse index. Reads with errors at both sides can be ignored because the error ratio is squared. A model based on

Poisson distribution was used to differentiate migrating reads from true positive reads. By applying error rate (false results in index rate) as the rate and the total number of reads potentially spilled to a particular index, the total number of reads with low probability of being a crosstalk can be calculated. To further avoid any false positive result, an arbitrary 10 times higher error rate was applied. This "safety margin" can be adjusted along the way as more and more data are gained from the sequencing. Ultimately, this approach can exert more control in calling positive or negative results by minimizing the false positive result, as demonstrated in the result from clinical samples.

With RT-qPCR as a comparison, the positive agreement and negative agreement of the pan-flavivirus/sequencing system were calculated to be 66.7% and 95.4%, respectively. While the positive agreement is decent, the result is in concordance with the cycle threshold (Ct) values; samples with lower Ct values were also positive and in contrast, those with higher Ct values were negative in the pan-flavivirus/sequencing test (**Table 1.13**). Three additional RT-qPCR negative samples turned out to be positive when subjected to sequencing. In total, the pan-flavivirus/sequencing test showed more positive results then the RT-qPCR. When the RT-qPCR is compared to the pan-flavivirus/sequencing test, the consequent positive agreement and negative agreement of the RT-qPCR to our system was 57.14% and 96.88%, respectively. Given that the RT-qPCR is not the gold-standard, the newly developed system does have comparable performance to the commercial RT-qPCR assay used in this study (17). On the other hand, a low positive agreement was observed when the sequencing was compared to NS1 antigen test. Yet, DENV sequences can be detected in one NS1 antigen test negative sample. There are several explanations to this observation. Firstly, the difference in basic principle of the test (protein and nucleic acid detection). Secondly, there is a difference in the detectable window between antigen and viral RNA. The NS1 antigen can be detected from first to the ninth day (sixth day in secondary infection) after fever onset and can stay in blood two or three days longer after viremia, providing a longer window period for detection (78,79). Nevertheless, despite this outstanding performance of NS1 antigen test, the test did not provide serological information and a report regarding declining in sensitivity has also been reported, especially in secondary infection and in DENV4 infection (80).

The pan-flavivirus/sequencing system provides a broad detection platform of flavivirus. The comprehensiveness of the flavivirus identification was demonstrated by detection of the target sequence of nine different viruses. From Vietnam, DENV1, DENV2, and DENV4 were detected in clinical samples albeit the other flaviviruses were not detected because of distribution. In addition, SLEV, YFV, WNV, DENV, and ZIKV were detected from clinical samples in Brazil. Though, for this data, shorter index sequences (14-mer) were employed. Therefore, the observation still includes some ambiguities such as discrepancy to RT-qPCR test and a lack of differentiation between true positive and false positive. However, in cases where a substantial number of reads were assigned the existence of a virus would be likely enough.

It is observed that in the absence of an unambiguous band in gel electrophoresis, viral sequence can still be obtained from the sample. This has been observed from validation carried out using plasmid sequence (**Table 1.9**) and also from clinical samples. Thus, a workflow was proposed to accommodate this. In cases where post amplification analysis (gel electrophoresis) showed an unambiguous band, it can be directly sequenced with Flongle. As shown from the clinical testing, when signal from electrophoresis can be observed, the sequence can also be obtained using Flongle. Negative samples can be subjected to deep sequencing with MinION to recover some viral sequences which exist in a minute amount **(Fig. 1.8)**. Since electrophoresis result is rather a qualitative approach, it is possible to use a RT-qPCR instead of RT-PCR/gel electrophoresis in determining which sample should be sequenced with Flongle or MinION (i.e. RT-qPCR positive samples should be sequenced using a Flongle flowcell, while negative results can be sequenced using a MinION flowcell).

Infectious disease outbreaks in recent years have highlighted the importance of molecular detection and sequencing as an inseparable part in controlling the outbreaks. In this chapter, a combination of nanopore sequencing and pan-flavivirus RT-PCR has provided a broad detection for flaviviruses. This approach overcomes several limitations of the current molecular approaches for flavivirus detection, including lack of broad specificity for NAT and serology test, as well as cross reactivity between flavivirus species when serology test is employed. This approach includes a dual index system for multiplexing up to 144 samples in one sequencing run and a tailored bioinformatics pipeline to minimize crosstalk between samples. Comparison to a commercially available

NAT kit showed that pan-flavivirus RT-PCR/sequencing test could detect more positive samples. One downside of the multiplexing approach is that the index sequence appeared to interfere with the PCR reaction. Shorter index sequence in combination with the improved accuracy of the new flowcell and chemistry would be beneficial in resolving this issue. Taken altogether, the pan-flavivirus RT-PCR/sequencing detection system can be expected to provide a considerable contribution to the detection and surveillance of flaviviruses.

Table 1.1 List of pan-flavivirus primer used in this study (31).

Primer Name	Sequence (5' to 3')
Flavi_all_AS_2	GTGTCCCAGCCNGCKGTGTCATCWGC
DEN4_F	TACAACATGATGGGRAAACGTGAGAA
Flavi_all_S	TACAACATGATGGGGAARAGAGARAA

Table 1.2 List of index sequences used for multiplexing.

Name	26-mer Index Sequence (5' to 3')	14-mer Index Sequence (5' to 3')	Remarks
i01	ATTCTCTGGATCTCAAGCGGTCATTA	CTATACAGCATGAG	Forward
i02	AATGGTTCACACTCAAGGATACTCTC	AGAGTCTAGCTAGC	
i03	AACCGAGAATTAGGAAGTCACACGCC	TGCGACACATGTGA	
i04	GTTCAGACAGTATGACAAGAGATTCC	GACTATGCAGTGCA	
i05	CACTTATATCACGGACCTCCTGCGAA	ACGCGTGCATCTAC	
i06	ACGGTCTATTGTTGAGTATCTGGTGA	TCGAGTAGTCTCAG	
i07	AACAACAACAACCGATGCACCTCTGT	GTATCATGTCAGCA	
i08	TGCCATCTTGCGGAATTCATACCAGC	AGCTAGTAGCTACT	
i09	ACCATTGACCATAGCACTTCCGAGCT	CGAGACGATACTCT	
i10	TGGTGGTCGCTGAACCGGTGAGTTAA	TAGATGCTCGCGAG	
i11	CCTAACAATTCTGCCGTACCGGACAA	GCTACGCTGAGTAG	
i12	CCGAAGCTCCACTACTAACCATGAAG	TCTCAGCGCAGTGA	
i13	CGTCTTCCTCCATACTCTTAGCAGTA	TAGCTCGACTGCGA	Reverse
i14	CGTTGTGAGGTGGAGATTATGGTACG	AGCTGCTGATCACT	
i15	TTATGCCAATCGGAGCCTGACGCTTA	TGACAGTCAGTCGT	
i16	AACACCGCTCTTCTGCTGCAATATAC	GCATGTGTATACAC	
i17	CTCCGCATTAACTGGGTGGTGACAAG	CTCGCATCGATGCA	
i18	ACGCCACCTAAGATGTATTCCAGATG	TATGAGATCTGCTC	
i19	CGTAGGAACCAAGAGTCGCCACAACT	GACGTCATAGTGCA	
i20	AGTAAGCCTGTGCCTAGTGACGATGG	TCGATCGCGCATAG	
i21	CAGGAGTGTCTTAATATCATTCGCTC	TAGTACTGTGACAC	
i22	AAGGTAAGAGAAGGTCGAAGCGCGTT	CATGCACTGATCGT	
i23	AAGTTGCGGTTATGTGAGAGCCTATT	ACTCTGTCACGTGA	
i24	AAGAATGTATCGCCTTATGTGCTGCG	CTACGTGCGCTAGA	

Table 1.3 Information about the virus, barcode, and the 26-mer index combinations for each sample used for validations of the system.

Sample No	Spiked virus	MinION		Flongle 1		Flongle 2	
		Barcode	Index	Barcode	Index	Barcode	Index
1	DENV1	1	i01 i13	1	i01 i24	1	i01 i14
2	DENV2	2	i02 i14	5	i02 i23	2	i02 i15
3	YFV	3	i03 i15	9	i03 i22	3	i03 i16
4	DENV1	4	i04 i16	2	i04 i19	4	i04 i17
5	DENV2	5	i05 i17	6	i05 i20	5	i05 i18
6	YFV	6	i06 i18	10	i06 i14	6	i06 i19
7	DENV1	7	i07 i19	3	i07 i18	7	i07 i20
8	DENV2	8	i08 i20	7	i08 i17	8	i08 i21
9	YFV	9	i09 i21	11	i09 i13	9	i09 i22
10	DENV1	10	i10 i22	4	i10 i16	10	i10 i23
11	DENV2	11	i11 i23	8	i10 i21	11	i10 i24
12	YFV	12	i12 i24	12	i12 i15	12	i12 i13

DENV, dengue virus; YFV, yellow fever virus.

Table 1.4 The mean recovery reads, true result, and false results from the bioinformatic analysis using three different tools and various parameters.

Tools	Parameter	Recovery Rate[a] (%)				TRUE[b] (%)				False results in barcode[c] (%)				False results in index[d] (%)				False results in sequencing[e] (%)			
		MinION	Flongle 1	Flongle 2	Average	MinION	Flongle 1	Flongle 2	Average	MinION	Flongle 1	Flongle 2	Average	MinION	Flongle 1	Flongle 2	Average	MinION	Flongle 1	Flongle 2	Average
LAST	Score - 60	41.78	49.05	53.29	48.04	98.89	99.27	99.19	99.12	0.63	0.17	0.45	0.42	0.44	0.50	0.22	0.39	0.04	0.06	0.14	0.08
	Score - 65	31.97	38.89	41.63	37.50	99.17	99.60	99.37	99.38	0.67	0.25	0.44	0.45	0.12	0.08	0.05	0.08	0.05	0.07	0.14	0.09
	Score - 70	22.50	28.62	29.46	26.86	99.32	99.70	99.33	99.45	0.60	0.20	0.49	0.43	0.01	0.02	0.03	0.02	0.07	0.07	0.14	0.09
	Score - 75	14.05	18.92	18.56	17.18	99.29	99.70	99.28	99.42	0.63	0.24	0.55	0.47	0.00	0.00	0.01	0.00	0.08	0.07	0.16	0.10
	Score - 80	7.55	10.88	9.56	9.33	99.22	99.59	99.28	99.36	0.69	0.33	0.55	0.52	0.00	0.00	0.01	0.00	0.09	0.08	0.16	0.11
	Score - 85	2.86	4.52	3.48	3.62	98.54	99.65	99.18	99.12	1.19	0.26	0.62	0.69	0.00	0.00	0.00	0.00	0.28	0.09	0.20	0.19
	Score - 90	0.37	0.76	0.54	0.56	97.22	99.40	99.67	98.76	2.78	0.00	0.33	1.04	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.20
MINIBAR	Edit distance 6	28.53	37.76	49.47	38.59	95.77	95.06	98.15	96.33	0.58	0.17	0.47	0.41	3.57	4.50	1.24	3.10	0.08	0.27	0.14	0.16
	Edit distance 5	24.36	33.03	41.45	32.95	98.69	97.15	99.28	98.37	0.55	0.17	0.45	0.39	0.72	2.61	0.15	1.16	0.04	0.07	0.13	0.08
	Edit distance 4	19.53	27.15	32.79	26.49	98.67	97.14	99.27	98.36	0.57	0.18	0.46	0.40	0.72	2.64	0.16	1.17	0.04	0.04	0.11	0.06
	Edit distance 3	14.09	20.90	23.51	19.50	98.51	97.01	99.18	98.23	0.55	0.13	0.47	0.38	0.88	2.82	0.21	1.30	0.06	0.05	0.14	0.08
	Edit distance 2	8.81	13.45	14.27	12.18	98.80	97.37	99.10	98.42	0.51	0.13	0.55	0.40	0.61	2.46	0.27	1.11	0.08	0.04	0.08	0.07
FREEBARCODES	Edit distance 7	10.78	1.09	1.23	4.37	98.96	91.67	99.07	96.57	0.58	0.00	0.00	0.19	0.37	0.31	0.62	0.43	0.08	0.00	0.31	0.13
	Edit distance 5	9.92	1.19	1.18	4.10	99.43	91.67	99.74	96.95	0.57	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.09
	Edit distance 3	5.15	0.64	0.59	2.13	99.23	91.20	99.31	96.58	0.77	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.46	0.69	0.38

[a]Percentage of input reads to reads with decoded index (number of deindexed reads/numbers of input (debarcoded) reads).
[b]True results are reads with the correct barcode, index, and virus combination. Percentage was calculated as the number of true reads divided by the total number of true and false reads.
[c]False results in barcode are deindexed reads with matched index and virus combination but were binned into the wrong barcode. Percentage was calculated as the number of false in barcode reads divided by the total number of true and false reads.
[d]False results in index are deindexed reads with matched barcode and virus combination but were binned into the index pair not included in the sample. Percentage was calculated as the number of false in index reads divided by the total number of true and false reads.
[e]False results in sequencing are deindexed reads with matched index and barcode but hit other sequences beside the designated virus (Table 1.3). Percentage was calculated as the number of false in sequencing reads divided by the total number of true and false reads.

Table 1.5 Distribution of reads binned to index and barcode from sequencing of spiked samples with MinION flowcell (experiment 1/3) using the optimized pipeline (LAST, score 70).

Sample no	Spiked virus	MinION											
		Barcode	Index	Debarcoded reads[a]	Deindexed Reads[b]			TRUE[c]	False results in barcode[d]	False results in index[e]	False results in sequencing[f]		
					Correct Index	Other Index	No Index				Other Flaviviruses	Other viruses	Other
					(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample 01	DENV1	1	i01 i13	2,960	684 (23.11)	0 (0)	2,276 (76.89)	607 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 02	DENV2	2	i02 i14	5,062	1,454 (28.72)	2 (0.04)	3,606 (71.28)	1,055 (99.53)	1 (0.09)	1 (0.09)	2 (0.19)	0 (0)	1 (0.09)
Sample 03	YFV	3	i03 i15	3,066	397 (12.95)	9 (0.29)	2,660 (87.05)	350 97.77	8 (2.23)	0 0	0 (0)	0 (0)	0 (0)
Sample 04	DENV1	4	i04 i16	5,312	2068 (38.93)	4 (0.08)	3,240 (61.07)	1,784 (99.83)	3 (0.17)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 05	DENV2	5	i05 i17	2,541	271 (10.67)	0 (0)	2,270 (89.93)	203 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 06	YFV	6	i06 i18	5,942	1988 (33.46)	5 (0.08)	3,949 (66.54)	1,754 (99.72)	5 (0.28)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 07	DENV1	7	i07 i19	1,583	306 (19.33)	2 (0.13)	1,275 (80.67)	266 (99.63)	1 (0.37)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 08	DENV2	8	i08 i20	1796	498 (27.73)	14 (0.78)	1,284 (72.27)	353 (96.71)	12 (3.29)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 09	YFV	9	i09 i21	1978	506 (25.58)	0 (0)	1,472 (74.42)	457 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Sample 10	DENV1	10	i10 i22	2382	541 (22.71)	4 (0.17)	1,837 (77.29)	492 (99.19)	4 (0.81)	0 0	0 (0)	0 (0)	0 (0)
Sample 11	DENV2	11	i11 i23	4183	533 (12.74)	0 (0)	3,650 (87.26)	410 (99.51)	0 (0)	0 (0)	2 (0.49)	0 (0)	0 (0)
Sample 12	YFV	12	i12 i24	3253	456 (18.42)	0 (0)	2,797 (85.98)	403 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Average					808.5 (22.49)	3.33 (0.13)	2,526.33 (77.50)	677.83 (99.32)	2.83 (0.60)	0.08 (0.01)	0.33 (0.06)	0 (0)	0.08 (0.01)

[a]Number of reads passing the filter for the debarcoding step.
[b]Number of reads passing the filter for the deindexing step.
[c]True results are reads with correct barcode, index, and virus combination. Percentage was calculated as number of true reads divided by total of true and false reads.
[d]False results in barcode are deindexed reads with matched index and virus combination but were binned into the wrong barcode. Percentage was calculated as the number of false in barcode reads divided by the total number of true and false reads.
[e]False results in index are deindexed reads with matched barcode and virus combination but were binned into the index pair not included in the sample. Percentage was calculated as the number of false in index reads divided by the total number of true and false reads.
[f]False results in sequencing are deindexed reads with matched index and barcode but hit other sequences beside the designated virus (Table 1.3). Percentage was calculated as the number of false in sequencing reads divided by the total number of true and false reads.

Table 1.6 Distribution of reads binned to index and barcode from sequencing of spiked samples using a Flongle flowcell (experiment 2/3) using the optimized pipeline (LAST, score 70).

Flongle_1													
Sample no	**Spiked virus**	**Barcode**	**Index**	**Debarcoded reads[a]**	**Deindexed Reads[b]**			**TRUE[c]**	**False results in barcode[d]**	**False results in index[e]**	**False results in sequencing[f]**		
					Correct Index	**Other Index**	**No Index**				**Other Flaviviruses**	**Other viruses**	**Other**
					(%)	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**
Sample 01	DENV1	1	i01 i24	1,669	447 (26.78)	0 (0)	1,222 (73.22)	408 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 02	DENV2	5	i02 i23	1,254	182 (26.92)	1 (0.15)	1,072 (73.08)	151 (99.34)	1 (0.66)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 03	YFV	9	i03 i22	964	362 (25.30)	1 (0.07)	602 (74.70)	329 (99.70)	1 (0.30)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 04	DENV1	2	i04 i19	1,237	312 (24.88)	2 (0.16)	925 (75.12)	276 (99.28)	2 (0.72)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 05	DENV2	6	i05 i20	676	483 (40.79)	3 (0.25)	193 (59.21)	400 (99.26)	1 (0.21)	1 (0.21)	1 (0.21)	0 (0)	0 (0)
Sample 06	YFV	10	i06 i14	1,184	272 (27.39)	0 (0)	912 (72.61)	254 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 07	DENV1	3	i07 i18	607	210 (21.78)	1 (0.10)	397 (78.22)	190 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 08	DENV2	7	i08 i17	2,304	91 (14.99)	0 (0)	2,213 (64.89)	70 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 09	YFV	11	i09 i13	1,431	371 (35.30)	0 (0)	1,060 (74.70)	338 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Sample 10	DENV1	4	i10 i16	993	459 (37.11)	0 (0)	534 (62.89)	401 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 11	DENV2	8	i11 i21	1,051	809 (35.11)	3 (0.13)	242 (64.70)	618 (98.88)	3 (0.48)	0 (0)	4 (0.64)	0 (0)	0 (0)
Sample 12	YFV	12	i12 i15	1,010	273 (27.03)	0 (0)	737 (72.97)	243 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Average				1,198.83	355.92 (28.62)	0.92 (0.07)	841.5 (71.38)	306.5 (99.70)	0.75 (0.20)	0.08 (0.02)	0.42 (0.07)	0 (0)	0 (0)

[a]Number of reads passing the filter for the debarcoding step.
[b]Number of reads passing the filter for the deindexing step.
[c]True results are reads with correct barcode, index, and virus combination. Percentage was calculated as number of true reads divided by total of true and false reads.
[d]False results in barcode are deindexed reads with matched index and virus combination but were binned into the wrong barcode. Percentage was calculated as the number of false in barcode reads divided by the total number of true and false reads.
[e]False results in index are deindexed reads with matched barcode and virus combination but were binned into the index pair not included in the sample. Percentage was calculated as the number of false in index reads divided by the total number of true and false reads.
[f]False results in sequencing are deindexed reads with matched index and barcode but hit other sequences beside the designated virus (Table 1.3). Percentage was calculated as the number of false in sequencing reads divided by the total number of true and false reads.

Table 1.7 Distribution of reads binned to index and barcode from sequencing of spiked samples using a Flongle flowcell (experiment 3/3) using the optimized pipeline (LAST, score 70).

Sample no	Spiked virus	Flongle_2											
		Barcode	Index	Debarcoded reads[a]	Deindexed Reads[b]			TRUE[c]	False results in barcode[d]	False results in index[e]	False results in sequencing[f]		
					Correct Index	Other Index	No Index				Other Flaviviruses	Other viruses	Other
					(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample 01	DENV1	1	i01 i14	6,632	2,291 (34.54)	2 (0.03)	4,339 (65.46)	1,872 (99.89)	2 (0.11)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 02	DENV2	2	i02 i15	5,243	1,711 (32.63)	6 (0.11)	3,526 (67.37)	1,187 (99.58)	4 (0.34)	1 (0.08)	0 (0)	0 (0)	0 (0)
Sample 03	YFV	3	i03 i16	3,970	1,085 (27.33)	1 (0.03)	2,884 (72.67)	943 (99.16)	1 (0.11)	0 (0)	7 (0.74)	0 (0)	0 (0)
Sample 04	DENV1	4	i04 i17	5,563	709 (12.74)	8 (0.14)	4,846 (87.26)	573 (98.62)	7 (1.20)	1 (0.17)	0 (0)	0 (0)	0 (0)
Sample 05	DENV2	5	i05 i18	2,384	737 (30.91)	11 (0.46)	1,636 (69.09)	515 (97.91)	10 (1.90)	0 (0)	1 (0.19)	0 (0)	0 (0)
Sample 06	YFV	6	i06 i19	687	155 (22.56)	0 (0)	532 (77.44)	134 (100)	0 (0.00)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 07	DENV1	7	i07 i20	4,163	1,666 (40.02)	1 (0.02)	2,496 (59.98)	1,388 (99.93)	1 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 08	DENV2	8	i08 i21	8,422	3,254 (38.64)	10 (0.12)	5,158 (61.36)	2,251 (99.51)	8 (0.35)	0 (0)	3 (0.13)	0 (0)	0 (0)
Sample 09	YFV	9	i09 i22	3,630	845 (23.28)	3 (0.08)	2,782 (76.72)	751 (99.73)	2 (0.27)	0 (0)	0 (0)	0 (0)	0 (0)

Sample 10	DENV1	10	i10 i23	5,323	1,790 (33.63)	8 (0.15)	3,525 (66.37)	1,555 (99.68)	5 (0.32)	0 (0)	0 (0)	0 (0)	0 (0)
Sample 11	DENV2	11	i11 i24	7288	2,045 (28.06)	23 (0.32)	5,220 (71.94)	1,400 (98.45)	18 (1.27)	1 (0.21)	3 (0.21)	0 (0)	0 (0)
Sample 12	YFV	12	i12 i13	2613	791 (29.12)	0 (0)	1,822 (70.88)	657 (99.55)	0 (0.00)	0 (0)	3 (0.45)	0 (0)	0 (0)
Average				4659.83	1,420.75 (29.46)	6.08 (0.12)	3,233 (70.54)	1,102.17 (99.33)	4.83 (0.49)	0.25 (0.03)	1.42 (0.14)	0 (0)	0 (0)

[a]Number of reads passing the filter for the debarcoding step.
[b]Number of reads passing the filter for the deindexing step.
[c]True results are reads with correct barcode, index, and virus combination. Percentage was calculated as number of true reads divided by total of true and false reads.
[d]False results in barcode are deindexed reads with matched index and virus combination but were binned into the wrong barcode. Percentage was calculated as the number of false in barcode reads divided by the total number of true and false reads.
[e]False results in index are deindexed reads with matched barcode and virus combination but were binned into the index pair not included in the sample. Percentage was calculated as the number of false in index reads divided by the total number of true and false reads.
[f]False results in sequencing are deindexed reads with matched index and barcode but hit other sequences beside the designated virus (Table 1.3). Percentage was calculated as the number of false in sequencing reads divided by the total number of true and false reads.

Table 1.8. Results of sequencing amplicons from pan-flavivirus PCR using plasmid template (10^2 or 10^3 copies/reaction) that contains sequences from nine different flaviviruses. Target sequence obtained from nine flaviviruses were cloned into plasmids, which then subjected to PCR using the 26-mer indexed primers. The resulted amplicons were sequenced using a Flongle flowcell.

Target sequence in the plasmid	Copies/ Reaction	DENV1	DENV2	DENV3	DENV4	JEV	ZIKV	YFV	WNV	TBEV	Other flaviviruses	Other viruses	Others
DENV1	10^3	**2,259**	0	0	0	0	0	0	0	0	0	0	0
DENV2	10^3	0	**1,143**	0	0	0	0	0	0	0	0	0	0
DENV3	10^3	0	0	**760**	0	0	0	0	0	0	0	0	0
DENV4	10^2	0	0	0	**869**	0	0	0	0	0	0	0	0
JEV	10^3	0	0	0	0	**1,190**	0	0	0	0	0	0	0
ZIKV	10^3	0	0	0	0	0	**947**	0	0	0	0	0	0
YFV	10^3	0	0	0	0	0	0	**2,110**	0	0	0	0	0
WNV	10^3	0	0	0	0	0	0	0	**1,270**	0	0	0	0
TBEV	10^3	0	0	0	0	0	0	0	0	**1,303**	0	0	0

Each row represents which viral sequence were inserted into the plasmid, which were used as the PCR template. Number in the cells represents the read homologous to virus in the column name.
DENV, Dengue Virus; JEV, Japanese Encephalitis Virus; ZIKV, Zika Virus; YFV, Yellow Fever Virus; WNV, West Nile Virus; TBEV, Tick-borne Encephalitis Virus.

Table 1.9. The MinION deep sequencing was able to retrieve sequence in invisible amplicons with lower concentration (10 to 10^2 copies/reaction)

ID	Target sequence in the plasmid	Copies/ reaction	DENV1	DENV2	DENV3	DENV4	JEV	TBEV	WNV	YFV	ZIKV	Other flaviviruses	Other viruses	Others
1	**DENV1**	10^2	0	0	0	0	0	0	0	0	0	0	0	0
2	**DENV2**	10^2	0	0	0	0	0	0	0	0	0	0	0	0
3	**DENV3**	10^2	0	0	**61**	0	0	0	0	0	0	0	0	0
4	**DENV4**	10	0	0	0	**26**	0	0	0	0	0	0	0	0
5	**JEV**	10	0	0	0	0	0	0	0	0	0	0	0	0
6	**TBEV**	10^2	0	0	0	0	0	0	0	0	0	0	0	0
7	**WNV**	10^2	0	0	0	0	0	0	0	0	0	0	0	0
8	**YFV**	10	0	0	0	0	0	0	0	0	0	0	0	0
9	**ZIKV**	10^2	0	0	0	0	0	0	0	0	0	0	0	0

Each row represents which viral sequence were inserted into the plasmid, which were used as the PCR template. Number in the cells represents the read homologous to virus in the column name. Concentration of template in each reaction is exactly 10^{-1} from the template concentration used in Table 1.8. DENV, Dengue Virus; JEV, Japanese Encephalitis Virus; ZIKV, Zika Virus; YFV, Yellow Fever Virus; WNV, West Nile Virus; TBEV, Tick-borne Encephalitis Virus.

Table 1.10 Distribution of reads binned to each index combination, based on Vietnam samples, sequenced with Flongle (top) and MinION (bottom).

Flongle												
	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	5	6	8	3	20	2	7	82	52	3	8	0
i02	13	2	5	2	5	2	2	4	27	5	0	3
i03	2	32	8	0	4	3	21	4	31	4	6	2
i04	3	2	2	8	5	0	24	5	18	1	0	1
i05	2	9	4	5	15	2	2	8	23	2	2	0
i06	0	1	1	2	4	0	211	0	26	2	47	4
i07	10	24	9	8	6	21	117	157	33	28	52	28
i08	7	8	0	5	4	1	17	4	10	7	3	3
i09	4	5	3	2	5	1	3	4	19	5	1	1
i10	4	3	1	8	2	2	6	2	48	5	2	0
i11	480	11	1	355	20	8	5	8	14	1	2	1
i12	2	0	1	3	4	1	74	434	11	134	34	0

MinION												
	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	349	515	481	251	2,111	133	312	184	3,498	402	771	446
i02	358	117	442	87	398	69	79	154	1,979	513	115	185
i03	164	2,846	765	129	161	174	961	252	1,206	236	356	180
i04	107	109	82	533	138	22	961	124	935	213	71	86
i05	72	302	57	180	276	110	66	521	760	104	66	101
i06	55	163	168	95	180	29	38	354	1,115	261	21	136
i07	474	3,287	501	762	245	1,092	5,797	19,501	2,361	2,138	3,005	3,477
i08	320	603	45	474	355	72	584	217	593	318	128	162
i09	101	195	120	164	562	36	77	195	841	112	75	109
i10	107	238	119	341	174	142	76	44	143	37	23	45
i11	0	0	0	0	1	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	1	0	0	0	0

Positive for NS1 **but negative** for DENV1-4 qPCR.

Positive for **both** NS1 and DENV1-4 qPCR.

Negative for NS1 and were **not** subjected to DENV1-4 qPCR.

Sample set from experiment not included in this study.

Unused index combination.

Table 1.11 Distribution of viral reads obtained from Vietnam sample, sequenced with Flongle.

DENV1

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	**3**	**3**	0	0	0	0
i02	0	0	0	0	0	0	0	**2**	0	0	0	0
i03	0	0	0	0	0	0	0	**1**	0	0	0	0
i04	0	0	0	0	0	0	0	**1**	0	0	0	0
i05	0	0	0	0	0	0	**1**	**1**	0	0	0	0
i06	0	0	0	**1**	**1**	0	**179**	0	**3**	0	**43**	**1**
i07	0	0	0	0	0	0	0	**1**	0	0	0	0
i08	0	0	0	0	0	0	0	0	0	0	**1**	0
i09	0	0	0	0	0	0	**1**	**2**	0	0	0	0
i10	0	0	0	0	0	0	**1**	0	0	0	0	0
i11	**1**	0	0	**272**	0	**6**	**1**	**4**	**2**	0	0	0
i12	0	0	**1**	**1**	0	0	**50**	**370**	**2**	**1**	0	0

DENV2

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	**2**	0	0	0	0	0	0	**44**	0	0	0	0
i02	0	0	0	0	0	0	0	0	0	0	0	0
i03	**1**	0	0	0	0	0	0	0	0	0	0	0
i04	0	0	0	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	0	0	0
i06	0	0	0	0	0	0	0	0	0	0	0	0
i07	0	0	0	0	0	0	0	**1**	0	0	0	0
i08	**1**	0	0	0	0	0	0	0	0	0	0	0
i09	**1**	0	0	0	0	0	0	0	0	0	0	0
i10	**1**	0	0	0	0	0	0	0	0	0	0	0
i11	**196**	0	0	0	0	0	**1**	**1**	**3**	**1**	0	0
i12	0	0	0	0	0	0	**9**	0	0	0	0	0

DENV3

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	0	0	0	0	0	0
i02	0	0	0	0	0	0	0	0	0	0	0	0
i03	0	0	0	0	0	0	0	0	0	0	0	0
i04	0	0	0	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	0	0	0
i06	0	0	0	0	0	0	0	0	0	0	0	0
i07	0	0	0	0	0	0	0	0	0	0	0	0
i08	0	0	0	0	0	0	0	0	0	0	0	0
i09	0	0	0	0	0	0	0	0	0	0	0	0
i10	0	0	0	0	0	0	0	0	0	0	0	0
i11	0	0	0	0	0	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	0	0	0	0	0

DENV4

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	0	0	0	0	0	0
i02	0	0	0	0	0	0	0	0	0	0	0	0
i03	0	0	0	0	0	0	0	0	0	0	0	0
i04	0	0	0	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	0	0	0
i06	0	0	0	0	0	0	0	0	0	0	0	0
i07	0	0	0	0	0	0	0	0	0	0	0	0
i08	0	0	0	0	0	0	0	0	0	0	0	0
i09	0	0	0	0	0	0	0	0	0	0	0	0
i10	0	0	0	0	0	0	0	0	0	0	0	0
i11	0	0	0	0	0	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	0	0	0	0	0

Positive for NS1 **but negative** for DENV1-4 qPCR.

Positive for **both** NS1 and DENV1-4 qPCR.

Negative for NS1 and were **not** subjected to DENV1-4 qPCR.

Sample set from experiment not included in this study.

Unused index combination.

1 Positive*

1 Negative*

*Positive or negative call was made based on the read number. When the detected viral read is < 3 reads, then the result is automatically omitted. When read number > 2, then the threshold should be calculated according to explanation in **Fig. 1.6**.

Table 1.12 Distribution of viral reads obtained from Vietnam sample, sequenced with MinION (bottom).

DENV1

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	0	0	0	0	0	0
i02	0	0	0	0	0	0	0	**1**	0	0	0	0
i03	0	0	0	0	0	0	0	**1**	0	0	0	0
i04	0	0	**12**	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	2	0	0
i06	**1**	0	0	0	**2**	0	0	**187**	0	**107**	0	0
i07	0	0	0	0	0	0	0	**1**	0	0	0	0
i08	0	0	0	0	0	0	0	**60**	0	0	0	0
i09	0	0	0	0	0	0	0	0	0	**1**	0	0
i10	**2**	0	0	**1**	0	0	0	0	0	**1**	0	**1**
i11	0	0	0	0	0	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	0	0	0	0	0

DENV2

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	0	0	0	0	0	0
i02	0	0	0	0	0	0	0	0	0	0	0	0
i03	0	0	0	0	0	0	0	0	0	0	0	0
i04	0	0	0	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	0	0	0
i06	0	0	0	0	0	0	0	0	0	0	0	0
i07	0	0	0	0	0	0	0	0	0	0	0	0
i08	0	0	0	0	0	0	0	0	0	0	0	0
i09	0	0	0	0	0	0	0	0	0	0	0	0
i10	0	0	0	0	0	0	0	0	0	0	0	0
i11	0	0	0	0	0	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	0	0	0	0	0

DENV3

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	0	0	0	0	0	0
i02	0	0	0	0	0	0	0	0	0	0	0	0
i03	0	0	0	0	0	0	0	0	0	0	0	0
i04	0	0	0	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	0	0	0
i06	0	0	0	0	0	0	0	0	0	0	0	0
i07	0	0	0	0	0	0	0	0	0	0	0	0
i08	0	0	0	0	0	0	0	0	0	0	0	0
i09	0	0	0	0	0	0	0	0	0	0	0	0
i10	0	0	0	0	0	0	0	0	0	0	0	0
i11	0	0	0	0	0	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	0	0	0	0	0

DENV4

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	0	0	0	0	0	0	0	0	0	0	0	0
i02	0	0	0	0	0	0	0	0	0	0	0	0
i03	0	0	0	0	0	0	0	0	0	0	0	0
i04	0	0	0	0	0	0	0	0	0	0	0	0
i05	0	0	0	0	0	0	0	0	0	0	0	0
i06	0	0	**16**	0	0	0	0	**1**	0	0	0	0
i07	0	0	0	0	0	0	0	0	0	0	0	0
i08	0	0	0	0	0	0	0	0	0	0	0	0
i09	0	0	0	0	0	0	0	0	0	0	0	0
i10	0	0	0	0	0	0	0	0	0	0	0	0
i11	0	0	0	0	0	0	0	0	0	0	0	0
i12	0	0	0	0	0	0	0	0	0	0	0	0

Positive for NS1 **<u>but</u> negative** for DENV1-4 qPCR.

Positive for **<u>both</u>** NS1 and DENV1-4 qPCR.

Negative for NS1 and were **<u>not</u>** subjected to DENV1-4 qPCR.

Unused index combination.

1 Positive

1 Negative

*Positive or negative call was made based on the read number. When the detected viral read is < 3 reads, then the result is automatically omitted. When read number >2, then the threshold should be calculated according to explanation in **Fig. 1.6**.

Table 1.13 Summary of clinical information, NS1 antigen-test, RT-qPCR, pan-flavivirus RT-PCR, and sequencing results.

Index	Day of fever	NS1-antigen test	Ct Value (qPCR)	RT-PCR[a]	Flongle	MinION	Number of reads					
							DENV1	DENV2	DENV3	DENV4	Other flaviviruses	Other viruses
i01-i20	4	+	-	+/-	DENV2	not tested	**3**	**44**	0	0	0	0
i03-i15	6	+	-	-	-	DENV1	**12**	0	0	0	0	0
i06-i15	2	+	-	-	-	DENV4	0	0	0	**16**	0	0
i06-i18	3	+	39	-	-	-						
i06-i19	1	+	28.2	++	DENV1	not tested	**179**	0	0	0	0	0
i06-i20	4	+	41.5	-	-	DENV1	**187**	0	0	0	0	0
i06-i21	2	+	35.8	-	-	-						
i06-i22	3	+	34	-	-	DENV1	**107**	0	0	0	0	0
i06-i23	2	+	32.5	+	DENV1	not tested	**43**	0	0	0	0	0
i08-i20	1	-	-	-		DENV1	**60**	0	0	0	0	0

[a]Intensity of gel electrophoresis of PCR products; ++, marked strong; +, strong; +/-, ambiguous; -, negative

NS1, non-structural protein 1; Ct, cycle threshold; RT, reverse transcription; PCR, polymerase chain reaction; qPCR, quantitative PCR; DENV, dengue virus.

Table 1.14 Positive and negative agreements between the tests.

Comparative assay	Test	Positive agreement	Negative agreement
NS1 antigen test	gel	4.2%	100.0%
NS1 antigen test	RT-qPCR	8.5%	100.0%
NS1 antigen test	Flongle	4.2%	100.0%
NS1 antigen test	Nanopore	9.9%	97.7%
RT-qPCR	Flongle	33.3%	98.5%
RT-qPCR	Nanopore	66.7%	95.4%
Flongle	RT-qPCR	66.7%	94.1%
Nanopore	RT-qPCR	57.1%	96.9%

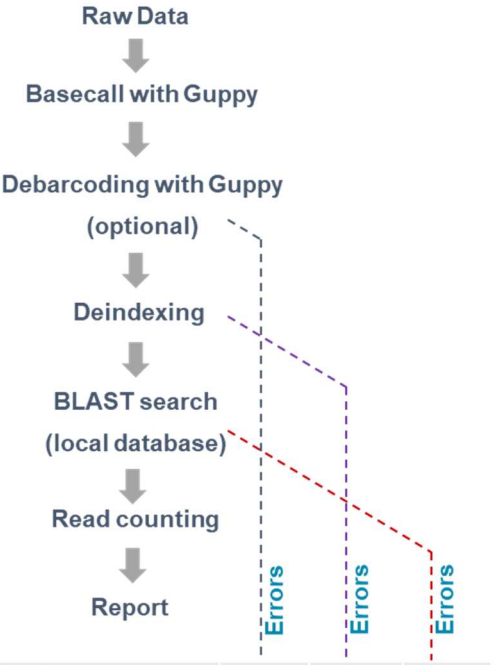
Table 1.15. Result of pan-flavivirus system in clinical sample from Brazil.

Sample No	Barcode	Index F	Index R	Result of RT-qPCR[a]	pan-flavi-PCR[b]	Virus reads							
						SLEV	YFV	WNV	DENV1	DENV2	DENV3	DENV4	ZIKV
1	3	i01	i13	SLEV	++	**118,350**	**2**	**1**	0	**1**	0	**47**	0
2	3	i02	i14	SLEV	++	**62,830**	**3**	**15**	**1**	**1**	0	**28**	0
3	3	i03	i15	WNV	+	**4**	**4**	**5,591**	0	0	0	**5**	0
4	3	i04	i16	WNV	++	**2**	0	**29,035**	0	**1**	**1**	**3**	0
5	3	i05	i17	YFV	++	**5**	**23,420**	0	0	**1**	0	**11**	0
6	3	i07	i19	DENV	++	**6**	**6**	**1**	**99**	**5,006**	**1,334**	**27,848**	0
7	3	i08	i20	DENV	+	**29**	**23**	**6**	**495**	**325**	**104**	**3,336**	0
8	1	i09	i21	ZIKA	+/-	**2**	0	**3**	**749**	0	0	0	0
9	1	i10	i22	ZIKA	-	**2**	0	0	0	0	0	0	0
10	1	i11	i23	DENV4 and ZIKV	++	**30**	**1**	0	0	0	0	**2560**	0
11	1	i12	i24	YFV	+	**3**	**4,018**	0	0	**1**	0	0	0
12	2	i01	i13	ZIKA	++	**105**	0	0	0	0	0	**27**	**1217**
13	2	i02	i14	YFV	+	**63**	**32,141**	0	0	0	0	**12**	0
14	2	i06	i18	DENV4	++	0	0	0	0	0	0	**21,974**	0
15	2	i08	i20	WNV	+/-	**1**	**2**	0	0	**2**	**1**	**24**	0
16	2	i09	i21	unknown	+/-	0	**150**	0	**2**	0	0	**4**	0
17	2	i05	i17	unknown	+/-	0	**38**	0	0	0	0	0	0
18	2	i07	i19	unknown	+/-	0	0	0	0	**2**	**2**	**23**	0
19	2	i04	i16	unknown	-	0	0	**30**	0	0	0	0	0
20	2	i10	i22	unknown	-	0	**1**	0	0	0	0	**5**	0
21	2	i11	i23	unknown	-	0	**2**	0	0	0	0	**32**	0
22	2	i12	i24	unknown	-	0	**13**	0	0	0	0	0	0
23	2	i03	i15	YFV	-	0	**1**	**8**	0	0	0	**2**	0
24	1	i06	i18	YFV	-	0	0	0	**1**	0	0	**7**	0

[a] Diagnostic decision by qRT-PCR. The question mark (?) represent uncertainty owing to their high Ct value.

[b] Intensity of gel electrophoresis of PCR products; ++, marked strong; +, strong; +/-, ambiguous; -, negative

SLEV, Saint Louis Encephalitis Virus; YFV, Yellow Fever Virus; WNV, West Nile Virus; DENV, Dengue Virus; ZIKV, Zika Virus.



TRUE	✓	✓	✓
FALSE result in barcode	×	✓	✓
FALSE result in index	✓	×	✓
FALSE result in sequence	✓	✓	×

Figure 1.1 Bioinformatic workflow to analyze the sequences. Errors were classified based on the properties assigned to each read (barcode, index, and virus sequence).

	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
i01	1	2	3	4	5	6	7	8	9	10	11	12
i02	13	14	15	16	17	18	19	20	21	22	23	24
i03	25	26	27	28	29	30	31	32	33	34	35	36
i04	37	38	39	40	41	42	43	44	45	46	47	48
i05	49	50	51	52	53	54	55	56	57	58	59	60
i06	61	62	63	64	65	66	67	68	69	70	71	72
i07	73	74	75	76	77	78	79	80	81	82	83	84
i08	85	86	87	88	89	90	91	92	93	94	95	96
i09	97	98	99	100	101	102	103	104	105	106	107	108
i10	109	110	111	112	113	114						
i11												
i12												

Positive for NS1 **but negative** for DENV1-4 qPCR.
Positive for **both** NS1 and DENV1-4 qPCR.
Negative for NS1 and were **not** subjected to DENV1-4 qPCR.
Sample set from experiment not included in this study.
Unused index combination.

Figure 1.2 Information about the forward and index combination applied to 114 samples from Vietnam. Sample numbers are written in each cell.

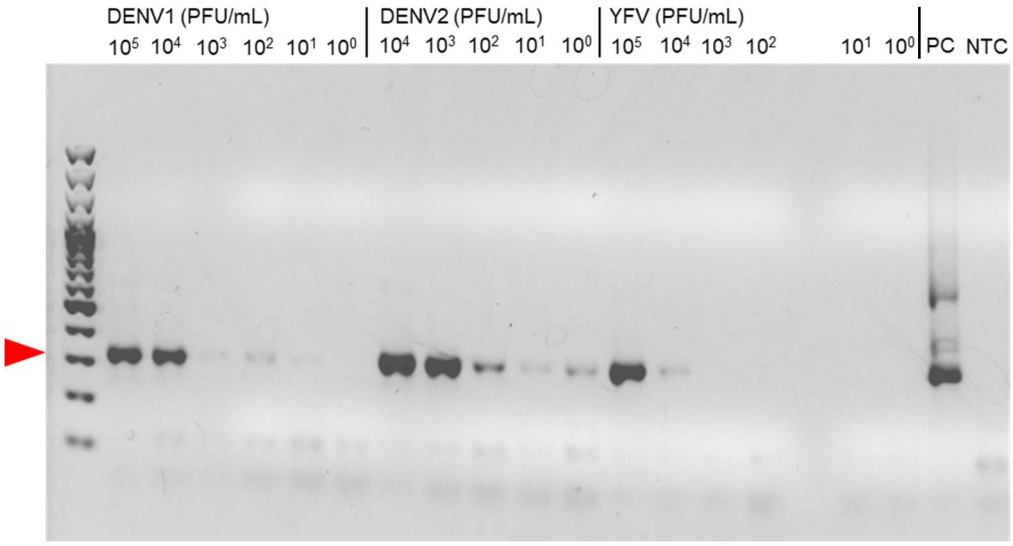


Figure 1.3 Gel electrophoresis image obtained from serial dilution experiment. Fetal bovine serum was spiked with viruses, then diluted accordingly to 1 PFU/mL. Viral RNA was extracted, then subjected to pan-flavivirus RT-PCR using the 26-mer index-added primer. DENV, dengue virus; YFV, yellow fever; PC, positive control; NTC, no template control. Red arrowhead denotes the expected amplicon size (~312 bp).

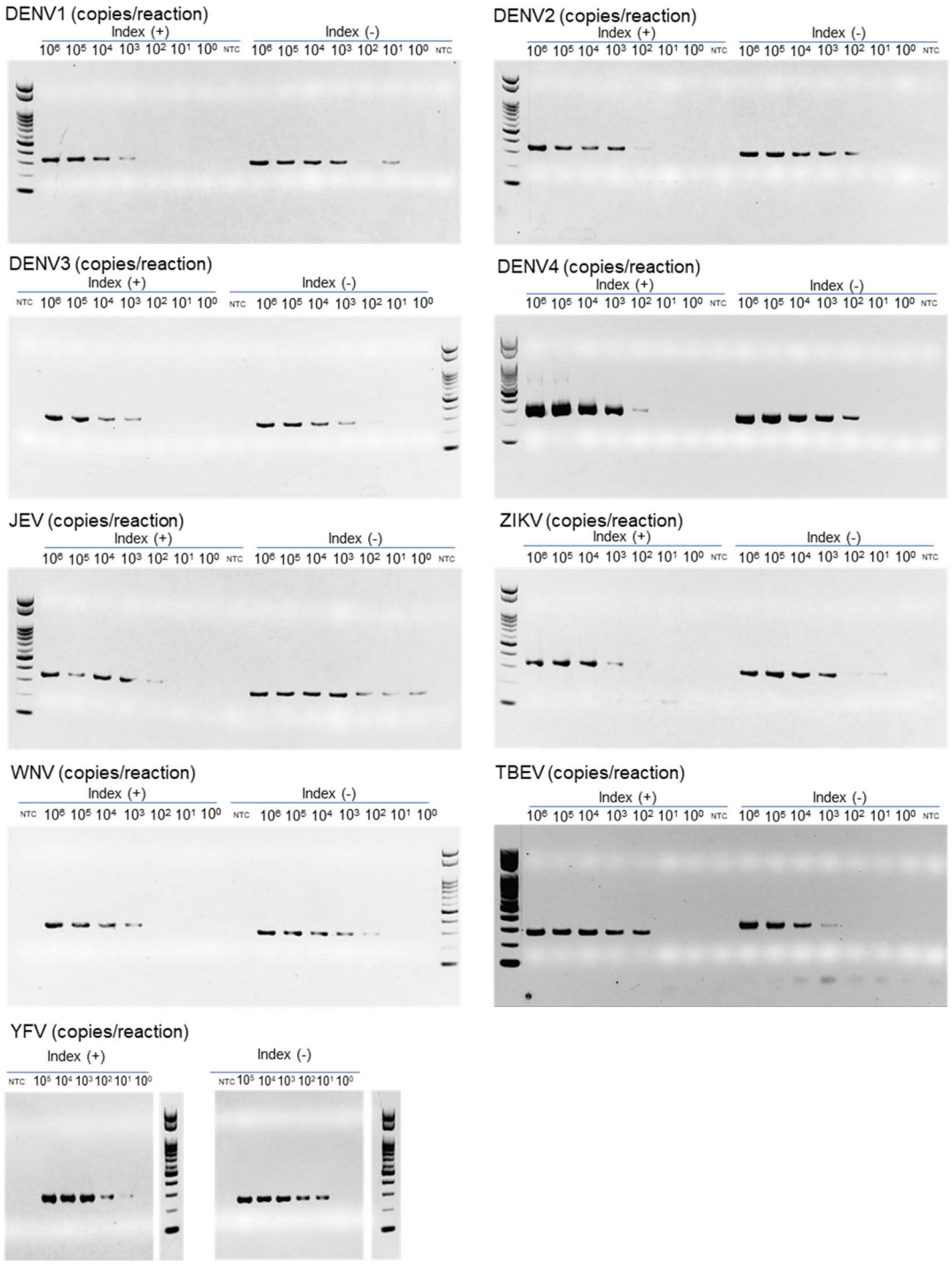


Figure 1.4 Gel electrophoresis image obtained from PCR using serially diluted plasmid template containing viral sequences as template. Index (+) indicate the result from 26-mer index-added primer sets while index (-) indicate index-free primer sets. DENV, Dengue Virus; JEV, Japanese Encephalitis Virus; ZIKV, Zika Virus; YFV, Yellow Fever Virus; WNV, West Nile Virus; TBEV, Tick Borne Encephalitis Virus.

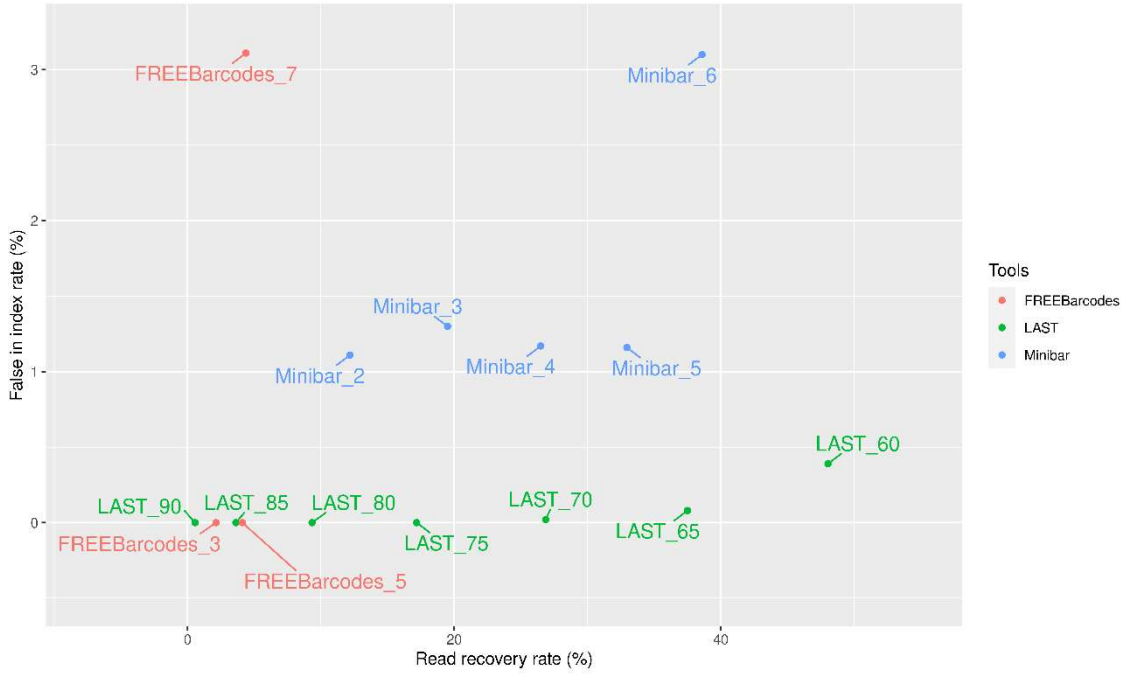


Figure 1.5 A plot showing the mean recovery rate (*x* axis) and mean error in index rate (*y* axis) obtained from three separate experiments using spiked sample. LAST-based demultiplexing system (green) showed a consistent performance, with modest recovery rate but lower error rate. The label on each dot represents the tools and the used parameters.

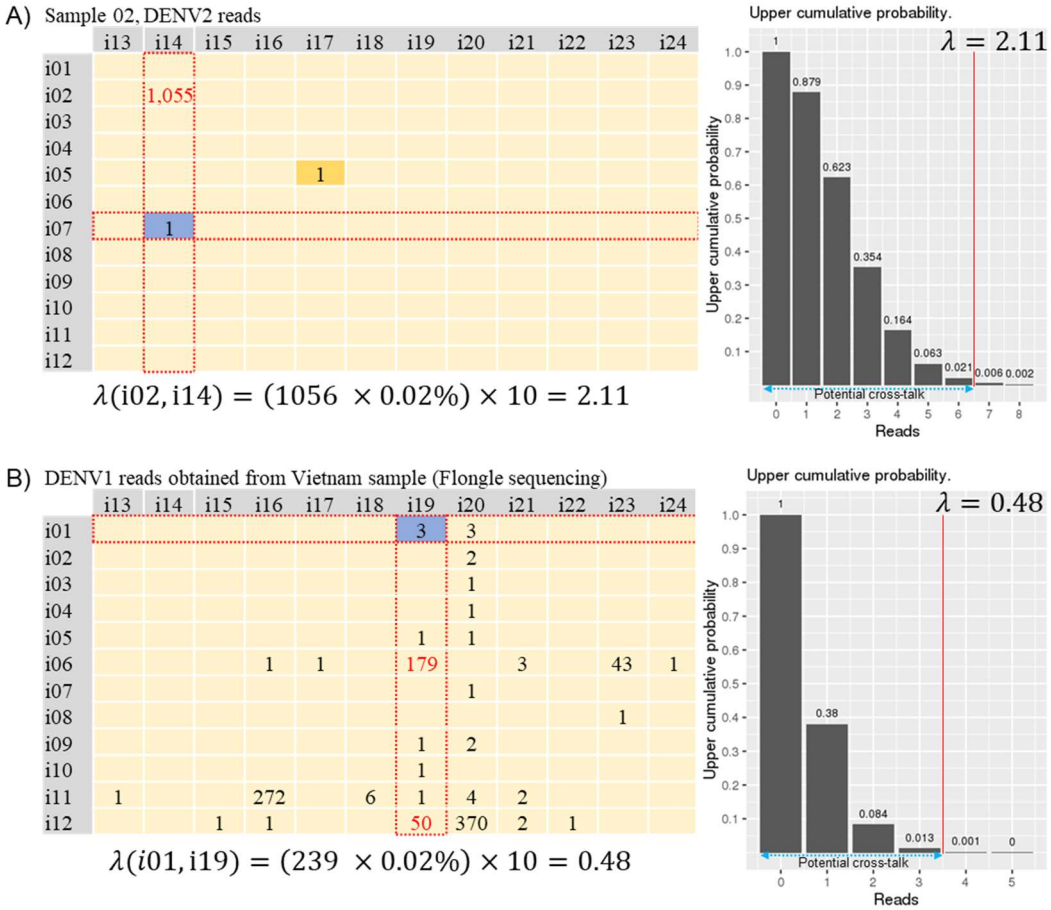


Figure 1.6 Calculation of threshold for potential crosstalk. Threshold was calculated based on assumption that spill-over reads originate from neighbouring samples with identical forward or reverse index. Error rate with a safety margin of ten times was used to estimate the λ value. The value was then plotted to Poisson cumulative distribution. In the event (μ) where the upper cumulative Poisson probability ($P(x \geq \mu)$) was less than 0.01 was determined as the threshold. A) Calculation of crosstalk threshold for i07-i14, showed that seven reads are needed for the result to be classified as positive. B) Similar calculation was performed on a clinical sample. The obtained 3 reads with i01-i19 index pair were classified as negative. The numbers colored in red is the possible source of crosstalk.

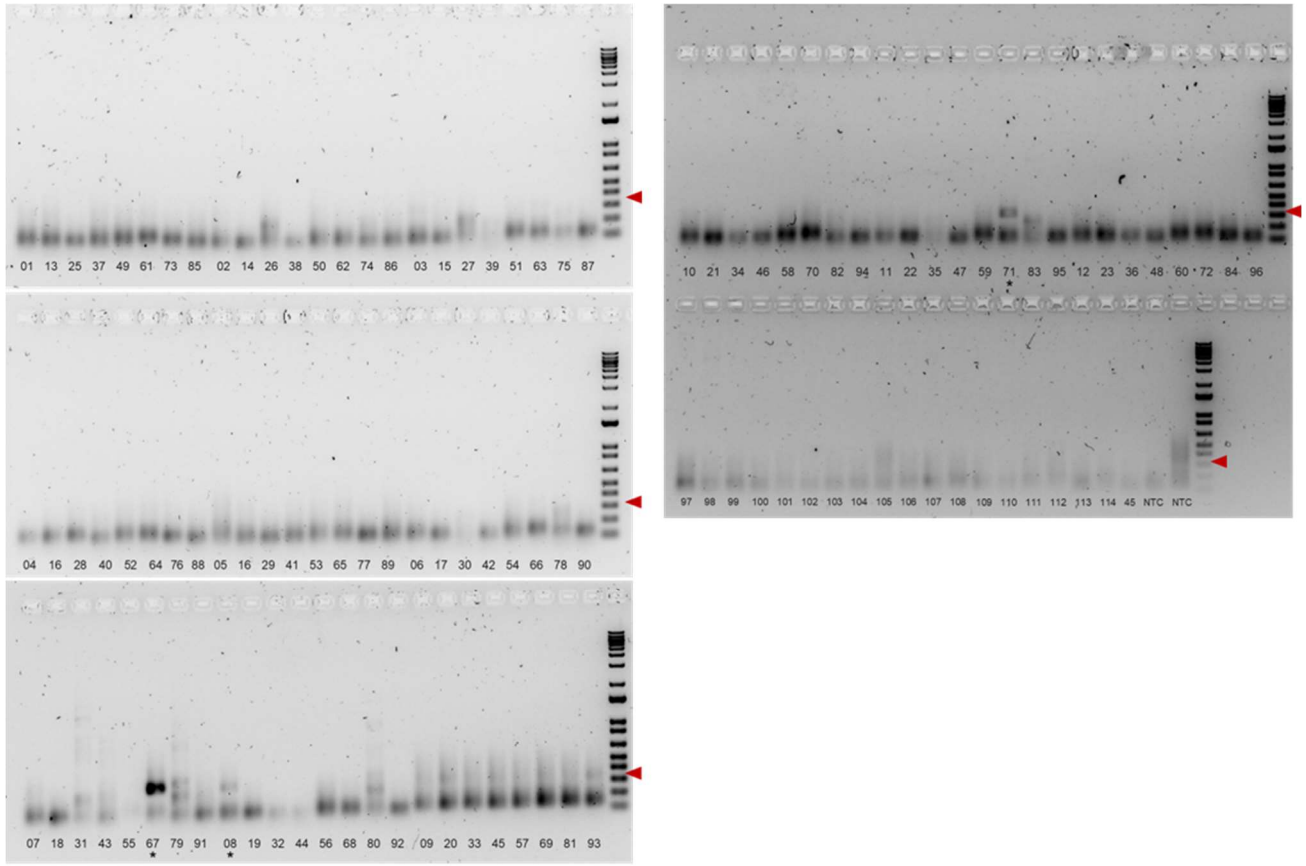


Figure 1.7 Visualization of the pan-flavivirus RT-PCR result of the 114 samples collected in Vietnam. Sample ID number can be seen at the bottom of the picture and the corresponding forward index (row) and reverse index (column) can be seen in the matrix (Figure 1.2). Asterisks represent positive sample and red arrowheads denote the expected amplicon size (~312 bp).

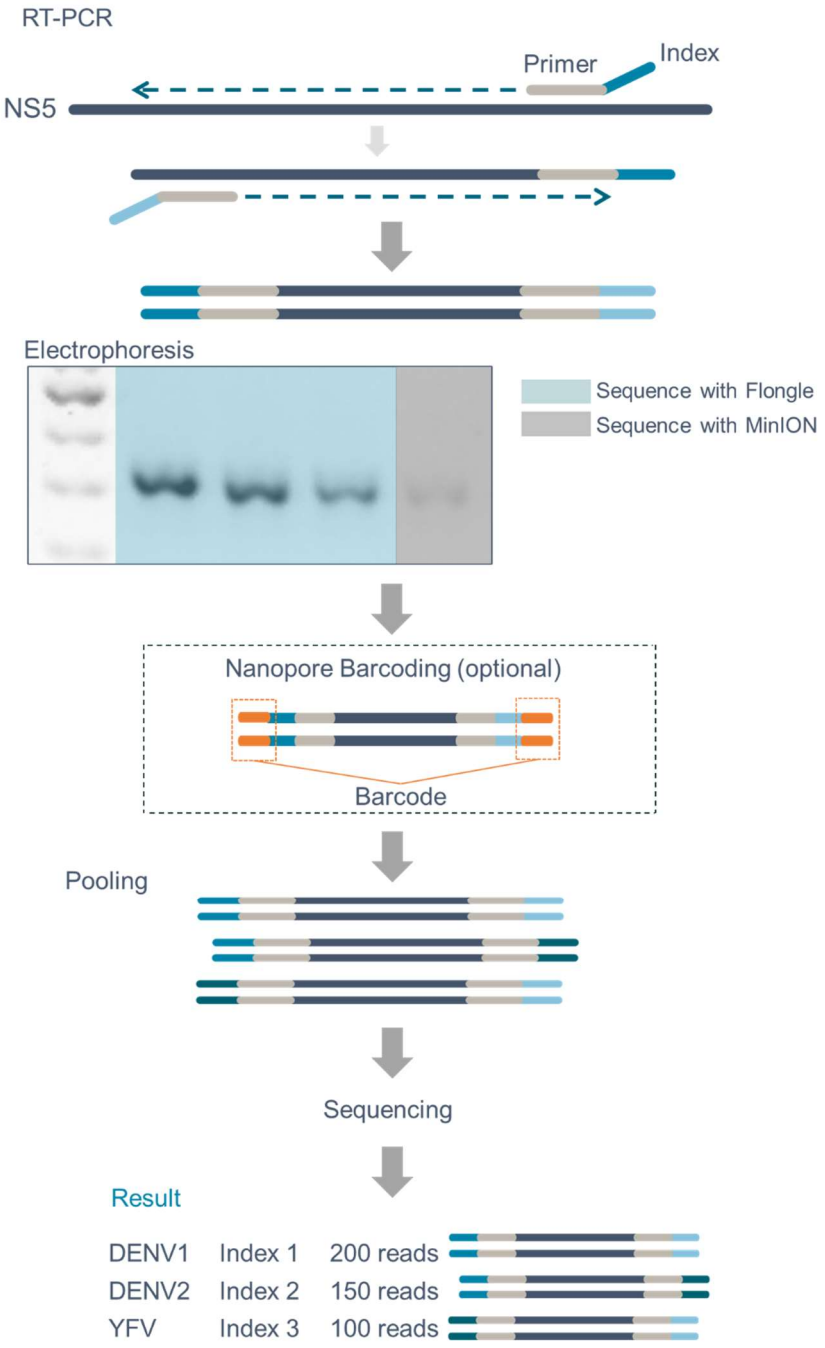


Figure 1.8 Proposed workflow for a sequencing-based detection of Flavivirus. When clear band is visible, the sample can be sequenced using Flongle. However, in the absence of clear band on gel electrophoresis, some viral sequences can be detected when MinION sequencing is employed.

Chapter Two

Circular Whole Transcriptome Amplification: An Alternative for Comprehensive RNAome Amplification

2.1 Summary

Limited amount of genetic material present in the sample challenges the application of NGS, due to the minimum requirement of DNA/RNA amount for library preparation. Comprehensive amplification is commonly used to increase the amount of genetic material. Among the available methods, PCR-based methods (such as sequence independent single primer amplification and its modification, SISPA) is one of the most common approaches. However, PCR-based amplification has several limitations including requirement of a precise temperature control and risk of introducing bias. Thus, in this chapter, an alternative unbiased amplification was developed and validated. This amplification relies on MDA, which uses phi29 polymerase, a high-fidelity polymerase that requires isothermal conditions. However, MDA is known to be less efficient in amplifying short template and it is also known to have an amplification bias towards circular single stranded DNA (ssDNA) templates. To overcome this issue, the synthesized cDNA was circularized using an enzyme that has a high affinity towards ssDNA. This method was termed circular whole transcriptome amplification, or cWTA. A series of experiments were carried out to validate cWTA. Firstly, comparison of the amplification result from reaction using circular or linear cDNA was able to show that circularization of the cDNA template improves the amplification. Secondly, the resulting amplicons were subjected to nanopore sequencing and viral sequences were able to be detected from sequencing, suggesting that the cWTA amplicon is suitable for NGS analysis. Lastly, the system was tested on clinical samples obtained from Bangladesh and Brazil. Sequences that are homologous to DENV were detected from samples obtained in Bangladesh, while sequences that are homologous to chikungunya virus (CHIKV) were detected from samples obtained in Brazil. The minimum requirement of cWTA in combination with nanopore allows for such system to be applied in laboratories or clinics with limited resource to leverage the genomic surveillance of infectious disease in such places.

2.2 Introduction

At present, one of the most popularly employed methods to detect pathogens is NAT. Among the available NATs, PCR is widely applied as a means of diagnosis. Not only limited to PCR, other non-PCR NATs have also been developed in the past years, including loop-mediated isothermal amplification, recombinase polymerase amplification, and MDA (22–24). Combination of these techniques with some non-conventional post amplification analysis, including nucleic acid chromatography (such as printed strip arrays) and specific high-sensitivity enzymatic reporter unlocking or SHERLOCK, have been described to increase the specificity (25–27).

Detection of nucleic acid is often highly specific to a certain target pathogen or a pathogen group, which is not suitable when there are numerous possible causes behind an undifferentiated febrile illness. Ideally a comprehensive NAT should be employed in such cases. One of the widely employed approaches for a semi-comprehensive NAT is by targeting conserved regions known as "genetic barcodes" (i.e., 16S rRNA gene, 18S rRNA gene, and ITS region) (40). In viruses, such sequences can be found only if targets are narrowed down into a genus or a family, as discussed in Chapter One. For further identification, additional analyses, such as microarrays, Sanger sequencing, or NGS, are required (60,81). In contrast, fully comprehensive NATs should be free from target limitations and applicable to any pathogens.

In principle, conventional RNA sequencing provides one possible approach, given that every pathogen has RNA as a genome or as transcripts. Indeed, transcriptomic data have been used to identify novel viruses, to characterize the virome present in the sample, or to detect viral contaminants in the laboratory reagents (82–84). Because these RNAs are often present in only a minute amount in samples, an amplification step is required for NGS. Among the current methods, random PCR is the most straightforward (85,86). Comprehensive amplification can also be achieved by modification of sequence independent single primer amplification (SISPA). Using random primers with a universal primer sequence attached on its 5'-end, the resulting cDNA can then be amplified using the attached universal primer sequence (86,87). Another amplification method, the multiple annealing and looping based amplification cycles (MALBAC) utilizes Bst polymerase to generate amplicons with universal sequence at both ends which complement each other (88). Because of that, the amplicons will form a "loop" preventing

the amplicon to be used as a template. A derived commercial product, TransPlex Whole Transcriptome Amplification (Sigma), is also available (89). However, bias, and limited reproductivity in the amplification step cannot be excluded, given that these methods depend on PCR. Other commercial products, such as a Direct RNA Sequencing Kit and a PCR-cDNA Sequencing Kit (Oxford Nanopore Technologies), are available for single-molecule sequencing, but their performance still requires optimization (90). A study utilized nanopore direct RNA sequencing for direct virus sequencing has been reported (91). The sequencing adapter contains a poly-dT sequence, and in the study *E. coli* poly(A) polymerase was used to add poly-A tail in viruses that lack it. Several unique amplification methods are designed to enrich the viral sequences. Combination of random primer and spiked primer designed to target certain groups of virus has successfully captured the target virus while preserving the metagenomic value of mNGS (92). Other method relies in "not-so-random primer", which were designed not to complement the host's ribosomal sequence, a major source of host sequence in transcriptomic data (93).

Since most methods rely on PCR, which tends to introduce bias in the results, other PCR-free amplification strategies are needed and should be considered for ease of handling and price. Thus, in this chapter, the focus will be on the application of MDA as an alternative for PCR-based amplification. It is known that MDA is unsuitable for the amplification of RNA or small cDNA fragments (< 2,000 bases) and is biased toward circular ssDNA templates (94–96). A novel approach was developed by combining MDA with cDNA circularized using a ligase with high affinity towards ssDNA. It is expected that the circular template will be amplified more efficiently by MDA and eliminate the template-size restriction.

2.3 Material and Methods

2.3.1 Samples and RNA extraction

Several types of samples were used in this study. For the feasibility experiments, approximately 10^6 of human foreskin fibroblast (HFF) cells and as a serum mimic, FBS samples that were spiked with DENV1 or DENV2 to achieve a concentration range from 10^2 to 10^5 PFU/mL, were used. Total RNA from the HFF cells was extracted using TRIzol (Thermo Fisher Scientific) and Direct-zol RNA Kits (Zymo Research) according to the manufacturer's instructions (including the DNase treatment). The amount of RNA was measured using a Qubit Fluorometer (Invitrogen). While RNA in the spiked FBS samples were extracted using QIAamp Viral RNA Mini kit (Qiagen) according to the manufacturer's instructions.

Two sets of serum samples obtained from Bangladesh and Brazil were used in this study. Two samples from Bangladesh were obtained from Evercare Hospital, Dhaka. From these 2 samples, nucleic acids of DENV2 were detected using a commercial RT-qPCR kit. The other set was a set of serum samples of patients with fever collected in 2018 and stored by the Flavivirus Laboratory, Oswaldo Cruz Institute/Oswaldo Cruz Foundation (Fiocruz) in Rio de Janeiro, Brazil. These two sets of clinical samples were processed independently. For both sets, RNA extractions were carried out using QIAamp Viral RNA Mini kit (Qiagen) according to the manufacturer's instructions. Ethical approvals were obtained from Apollo Hospital (present Evercare Hospital), Dhaka (ERC 16/2018-2), Fiocruz (90249218.6.1001.5248: 2.998.362), and Hokkaido University (Jinjyu1-3 and Jinjyu30-4).

2.3.2 Circular Whole Transcriptome Amplification

The extracted RNA was subjected to first strand cDNA synthesis using SuperScript IV Reverse Transcriptase (Invitrogen). The random primers were replaced with random 9-mer primers with phosphate modification at the 5'-end (N9P primer) at final concentration of 1.25 μM. For total RNA extracted from the cell culture, the RNA amount was adjusted so that each reaction contained 1 pg to 10 ng of RNA. For samples other than the culture, 11 μL of extracted RNA were used in each reaction. Other than the primer, the RT reaction was performed according to the manufacturer's instructions. The

cDNA was then purified with 1.8× volume of AmPure XP beads and finally eluted in 15 µL of nuclease-free water.

The purified cDNA was then circularized using CircLigase II ssDNA ligase (Lucigen). In brief, 12.5 µL of purified cDNA, 2 µL reaction buffer, 1 µL $MnCl_2$, 0.5 µL CircLigase II (Lucigen), and nuclease-free water were mixed and incubated at 60°C for 1 hour, and 80°C for 10 mins. The resulting circularized products were purified using 1.8× volume of AmPure XP beads.

Multiple displacement amplification was carried out using 1 µL of circularized cDNA. GenomiPhi V2 (Cytiva), a phi29 polymerase, was used in this reaction. The circularized cDNA was mixed with 9 µL sample buffer which contained 6-mer random primer. The mixture was incubated at 95°C for 5 minutes then placed on ice directly, to prime the random primer. Following the priming, 9 µL of reaction buffer and 1 µL of the enzyme were added. The MDA reaction was incubated at 30°C for 90 minutes, followed by inactivating the enzyme at 65°C for 5 minutes. Schematic process of the cWTA can be seen in **Fig. 2.1**.

Three different experiments were carried out to evaluate the feasibility of cWTA. Firstly, 1 pg to 10 ng of total RNA from HFF cells were subjected to cWTA. Secondly, a serially diluted spiked FBS sample containing 10^3, 10^4, and 10^5 PFU/mL of DENV1 virions and 10^2, 10^3, and 10^4 PFU/mL of DENV2 virions were subjected to cWTA. Thirdly, to demonstrate the importance of circularization a comparison of reaction using linear and circular template was carried out using RNA extracted from a spiked sample containing 10^5 pfu/mL of DENV1. In untreated samples, the enzymes (the ssDNA ligase or the phi29 polymerase) were replaced with nuclease-free water. The amplified product was then quantified with qPCR, using a pan-flavivirus primers set (see Chapter One). The reaction was carried out using the Kapa Taq Extra PCR Kit (Kapa Biosystems) in a total volume of 15 µL containing 1.5 µL of 50× diluted cWTA amplicons, 250 nM of each primer DEN4_F and flavi_all_S, 500 nM primer flavi_all_AS2, and 0.75 µL EvaGreen (Biotium). The qPCR was performed in a Bio-Rad CFX96 apparatus with temperature conditions for the PCR as follows: 94°C for 30 seconds, followed by 43 cycles of 94°C, 53°C, and 72°C for 30 seconds each, and finally at 72°C for 5 minutes.

2.3.3 Sequencing library preparation

The amplicons obtained from the serially diluted virus-spiked FBS samples were purified using 1.8× volume of AmPure XP beads, then prepared using sequencing kit SQK-RLB001 and SQK-LSK109 (Oxford Nanopore Technologies), per the manufacturer's instructions. Samples from Bangladesh and Brazil were prepared using library kit SQK-LSK108 (Oxford Nanopore Technologies) according to the manufacturer's instruction. When necessary, the libraries were barcoded with barcoding kit NBD-LSK103 (Oxford Nanopore Technologies). Sequencing was carried out on a MinION Flowcell (version 9.4) per the manufacturer's instruction.

2.3.4 Bioinformatic pipeline

Raw Fast5 data were basecalled with Albacore version 1.1.2 or Guppy version 3.2.4 to generate the fastq files. The qscore threshold was set to seven and only sequences in the pass folder were used on the subsequent analysis. In cases where barcodes were used, the debarcoding was performed using the default parameters. To eliminate the host sequence, the reads were aligned to the host sequence [*Homo sapiens* (GCF_000001405.39) or *Bos taurus* (GCF_002263795.1)] using Minimap2 version 2.17 (97). Unmapped reads were then extracted using SAMtools version 1.10 (98). For variant calling, the reads assigned to a virus were first extracted and then mapped to its reference genome using Minimap2. Then, variants were identified using LoFreq version 2.1.5 (99).

Host-decontaminated reads were subjected to an alignment search using BLAST version 2.9.0 with the word size parameter set to 11 (67). To reduce the time needed for alignment search, a small viral database containing important pathogenic viruses were used (83). This database contains 5,139 viruses. In this database, regions highly resembling human or bacterial genomes were masked and viral genomes with up to 90% of similarity were concatenated. Based on the results of the first BLAST search, sequences aligned to the database with e-value less than 1.0e-10 were extracted and subjected to the second BLAST search. The database for second BLAST consisted of the viral sequences, human genome, and representative prokaryotic genomes. The purpose of the second blast was to further remove any reads which possibly belong to bacteria or humans. The sequences were assigned to the top-hit feature based on the bit-score. When a read had more than one hit to a different virus, with identical bit-score, the read was not

counted at the final count. The schematic diagram of the bioinformatic pipeline can be seen in **Fig. 2.2**.

2.4 Results

2.4.5 Amplification of RNAome with cWTA

As an initial proof of the concept, the cWTA was shown to successfully amplify RNA extracted from the cell culture (**Fig, 2.3)**. The amplicons were visible up to reaction containing 10 pg of RNA. Interestingly, in a reaction using linear template, no bands were observed even when the reaction contained 10 ng of RNA. The observed high intensity of the band in the gel electrophoresis is a strong indication that template circularization improved the MDA.

In the next experiment, a spiked FBS sample containing 10^5 PFU/mL of DENV1 was subjected to cWTA **(Fig. 2.4)**. To further test the effect of linear and circular template, the RNA was subjected to different reactions where the ssDNA ligase or the phi29 enzyme was substituted with nuclease-free water. When the amplicons were visualized in gel electrophoresis, bands were observed only in reaction containing both CircLigase II and phi29, suggested that circularization of ssDNA template improved the amplification. Nevertheless, gel electrophoresis is a rather qualitative measurement, and to get a quantitative data, the amplicons were subjected to a qPCR experiment. When the template for qPCR originated from amplicons generated from circularized cDNA, the resulted Ct value was lower than when the amplicons generated from linear cDNA was used as the template (**Table 2.1)**. Taken on all accounts, it is obvious that circularization of cDNA templates improved the amplification considerably.

2.4.6 Identification of the viral genome sequences

In the next step, the feasibility of coupling cWTA and NGS was evaluated. Serially diluted spiked FBS samples were used for this purpose. On gel electrophoresis, amplicons were visible up to the lowest concentrations **(Fig 2.5A).** The amplicons were sequenced using nanopore sequencing, on a MinION flowcell. Basecalling resulted in 96,502, 62,893, 91,183, and 101,591 reads for samples containing 10^5, 10^4, 10^3, and 10^2 PFU/mL of DENV1, respectively. While for DENV2, the basecalling resulted in 75,906, 47,000, and 80,900 from samples containing 10^4, 10^3, and 10^2 PFU/mL DENV2, respectively. Among these reads, reads aligned to the Dengue virus genome were unambiguously detected from samples containing 10^5 and 10^4 PFU/mL DENV1 and 10^4 PFU/mL DENV2 (**Table 2.2)**. However, on gel electrophoresis, the amplicons were observed up to the

lowest concentration. Additionally, reads with tandem-repeat patterns presumably derived from the continuous amplification of the circular template were also observed (**Fig. 2.5B**). In one case, the size of each tandem-repeat unit was approximately 300 bp.

Further experiment was then conducted to see if the system can be applied to clinical samples. For this, two clinical sample sets were subjected to cWTA and then sequenced with nanopore sequencing (on a MinION flowcell). The first clinical sample set are two serum samples obtained in Bangladesh. The samples were tested positive for DENV2 at the hospital using a commercial assay. Sequencing of cWTA amplicons from these samples yielded 18,539 and 6,789 reads, of which three reads from each sample were found to be homologous to DENV2. The second clinical samples were obtained in Brazil. Sequencing of the cWTA amplicons from these samples yielded 7,266 to 77,877 reads. Reads homologous to CHIKV were obtained from six samples. These CHIKV reads from two samples were then extracted and aligned to the CHIKV reference genome, resulting in 90% coverage with depth range from 2× to 363× (**Fig. 2.6)**. This result showed that the system can detect the presence of viral genome in the sample and that it is also possible to use the data for subsequent analysis. Indeed, alignment of CHIKV reads originating from Fiocruz01 and Fiocruz02 to its reference genome showed a significant number of reads aligned at certain positions in the genome, particularly at the 5'-end of the genome, indicating an amplification bias. The similar pattern was observed from both the samples. Alignment results showed that the two viruses in the sample shares considerable number of single nucleotide polymorphisms (SNPs), however each sample has their own unique SNPs (**Fig 2.7**), suggesting that the sequences are likely originated from two different viruses and that the bias is reproducible in the two samples. Interestingly, the bias was also observed when DENV1 sequences obtained from sequencing of the spike sample using nanopore sequencing were aligned to its reference genome (**Fig. 2.6)**.

2.5 Discussion

In this chapter, an alternative amplification method targeting RNAome termed as circle whole transcriptome amplification was developed. The amplification relies on phi29 polymerase, a polymerase isolated from bacteriophage phi29 (100). This enzyme amplifies DNA strand in isothermal conditions with high fidelity (due to proofreading ability) and it is also capable of producing long amplicons (up to 70 kilo-base-pair (kb)) (100,101). One of the well-known applications of this enzyme is MDA, which is utilized in whole genome amplification. However, amplification of MDA is known to be less efficient in amplifying short templates because short templates have less priming site and thus, resulting in fewer hyperbranching cycles (94). Circular template, which serves as an infinite length template, is known to be preferrable for phi29 (96,102,103). Thus, several other methods generate circular templates to be amplified with phi29, including rolling circle amplification (RCA) and its variations (25,104–106). Though in general RCA uses random primer for the amplification, a prior step to create the circular template is usually target-specific and therefore it lacks the comprehensiveness. For example, one of the modifications of RCA relies on a splint oligonucleotide whose sequences are complementary to the 5'- and 3'-end of the target sequence (105,107). On the contrary, cDNA template for cWTA is generated from random primer and circularization is carried out using a ligase with known affinity towards ssDNA, allowing cWTA to provide a comprehensive RNAome amplification.

Due to the bias towards circular ssDNA template, the cDNA circularization significantly improved the amplification. The striking difference in the band intensity was clearly observed from the circular and the linear template. Amplification of phi29 relies on the generated hyperbranching cycle, and this number is in linear correlation with the length of the template. Using circular templates, phi29 was able to generate a long-concatemeric DNA which was confirmed from sequencing using the long-read platform. This long-concatemeric read is likely the key to the improvement, because it can serve as a template for amplification and generating more hyperbranching cycles.

An amplification step is often necessary, given that the amount of genomic RNA or transcripts of pathogens in the sample are limited. These constraints come from the minimum amount of genetic material that is required for NGS (i.e., 200 ng for Illumina TruSeq library preparation and 100 to 200 ng for nanopore library prep). Several studies

using synovial fluids required unbiased amplification (WGA) to obtain sufficient amounts of DNA for library preparation (108,109). With unbiased amplifications, it is expected that the amount of genetic material can be increased without affecting the detection. In a previous study focusing on cerebrospinal fluid, it was reported that reads that come from pathogens were enriched 1 to 9 times when the sample was amplified with WGA (110). In this study, such systematic comparison using NGS data was not carried out. However, when the amplicons were subjected to a qPCR assay, comparison of amplicons produced from MDA reaction using circular and linear templates showed that amplification using circularized template resulted in nearly 32 times more amplicons compared to the result from linear template. This observation is true at least for the target region of the qPCR assay (the 260 bp region in the NS5 gene) and this observation is less likely to be the result of amplification bias since the bias was observed to be more prominent at the 5'-end of the viral genome. Taken altogether, it can be concluded that circularization of template cDNA improved MDA.

There are some limitations of this study. Firstly, random amplification can potentially introduce bias, which was not systematically assessed in this study. Several studies have reported amplification bias of MDA on the viral composition, with the most notable report being the bias towards circular ssDNA virus (96). Additionally, this study did not compare the bias with other random amplification methods (such as random primer amplification, modified SISPA, or MALBAC). Yet, a bias towards the 5'-end of the viral genome was observed in CHIKV and DENV1 samples. It is hypothesized that given the unique structure of CHIKV and DENV1 genome which has a cap at the 5'-end, the region closer to the cap might be more resistant to degradation, resulting in enrichment of region in the virus genome close to the 5'-end. A second hypothesis is that phi29 might have bias towards different template length. Evaluation on the effect of template lengths (range: 68 to 95 nucleotides (nt)) on RCA, showed a template dependent bias with a sinusoidal pattern (111). When an RT primer anneals to a location close to the 5'-end of RNA, the synthesized cDNA will have a shorter length. Yet, the correlation of template length (above 100 nt) and efficiency of phi29 amplification is currently unknown; thus, it is possible that the template length can affect the efficiency of cWTA. Given the current findings, a future study that includes a systematic evaluation (using NGS) of the amplicon

resulted from linear and circular template, while employing different types of viruses will help elucidate potential biases.

Another limitation is that most of the reads were dropped during host sequence decontamination. This is a common finding in metagenomic sequencing, and a host decontamination step will benefit the workflow. Though serum samples are known to have less genetic material, cell free DNA and RNA (cfDNA and cfRNA) that originate from the host (or pathogens) can be found. From the experiment using serially diluted spiked samples, amplification was observed even at lower concentrations of spiked virus. This amplicons are likely to originate from the host's cfDNA or cfRNA. Several pathogen detection methods targeting these cell-free nucleic acid have been described before (112–115). These cfDNA and cfRNA are usually highly fragmented but still be amplified because phi29 is shown to amplify circular templates with only 43 bp long (25). Yet, despite this limitation, analysis of cWTA amplicons with NGS was able to detect the virus in the clinical sample.

This chapter established cWTA as a non-specific amplification which can be analyzed with NGS. Indeed, pathogen sequences were able to be detected from clinical samples collected in Bangladesh and Brazil. Additionally, the reads obtained from at least two samples from Brazil was enough to cover almost the entire genome of CHIKV (depth range from 2× to 363x for Fiocruz01). Additional downstream analysis to uncover additional information about the pathogen can be carried out when enough reads can be obtained. For example, the similarity in coverage of CHIKV alignments from Fiocruz01 and Fiocruz02 suggests that the reads in Fiocruz02 might come from contamination during sample preparation or a result from barcode crosstalk from Fiocruz01. However, analysis of the variants between the two samples suggests that the reads come from two different viruses. Given that the nanopore sequencer is portable, stand-alone, and has low implementation costs, the application of cWTA and nanopore sequencer is promising for detection of pathogens in field settings or local clinics with limited resources. At last, implementation of cWTA-nanopore sequencing system can assist rural clinics in detection and also obtaining genetic information from pathogens with high potential threat.

Table 2.1 Comparison of Ct value obtained from subjecting cWTA amplicon to a RT-qPCR assay targeting NS5 region of flavivirus. RNA extracted from FBS spiked with 10^5 PFU/mL of DENV1 was subjected to reverse transcription then amplification with phi29. The cWTA reactions were carried out using either circularized or linear cDNA and either with or without amplification with phi29.

CircLigase II	phi29	Ct value		Mean
		rep1	rep2	
+	+	32.22	31.8	32.01
+	-	ND	ND	ND
-	+	37.01	36.95	36.98
-	-	ND	ND	ND

Ct, cycle threshold; ND, not detected; rep, technical replicates.

Table 2.2 Sequencing statistics and viral reads detected from the spiked and clinical samples using nanopore sequencing.

Sample		Number of raw reads	Reads mapped to host (%)		Unmapped reads (%)		DENV1	DENV2	CHIKV	Other viruses
Spiked sample	DENV1 10^5	96,502	45,309	(46.95)	51,193	(53.05)	**35**	0	0	0
	DENV1 10^4	62,893	46,859	(74.51)	16,034	(25.49)	**8**	0	0	0
	DENV1 10^3	91,183	68,702	(75.35)	22,481	(24.65)	**1**	0	0	0
	DENV1 10^2	101,591	78,776	(77.54)	22,815	(22.46)	0	0	0	0
	DENV2 10^4	75,906	34,125	(44.96)	41,781	(55.04)	0	**36**	0	0
	DENV2 10^3	47,000	36,622	(77.92)	10,378	(22.08)	0	**1**	0	0
	DENV2 10^2	80,900	63,060	(77.95)	17,840	(22.05)	0	0	0	0
Clinical sample	Bangladesh01	18,539	16,596	(89.52)	1,943	(10.48)	0	**3**	0	0
	Bangladesh02	6,789	6,048	(89.09)	741	(10.91)	0	**3**	0	0
	Fiocruz01	11,068	8,854	(80.00)	2,214	(20.00)	0	0	**1,576**	0
	Fiocruz02	9,868	8,930	(90.49)	938	(9.51)	0	0	**417**	0
	Fiocruz03	10,148	9,372	(92.35)	776	(7.65)	0	0	**1**	0
	Fiocruz04	11,897	10,983	(92.32)	914	(7.68)	0	0	**4**	0
	Fiocruz05	7,266	6,955	(95.72)	311	(4.28)	0	0	**15**	0
	Fiocruz06	11,190	10,715	(95.76)	475	(4.24)	0	0	**43**	0

DENV, dengue virus; CHIKV, chikungunya virus.

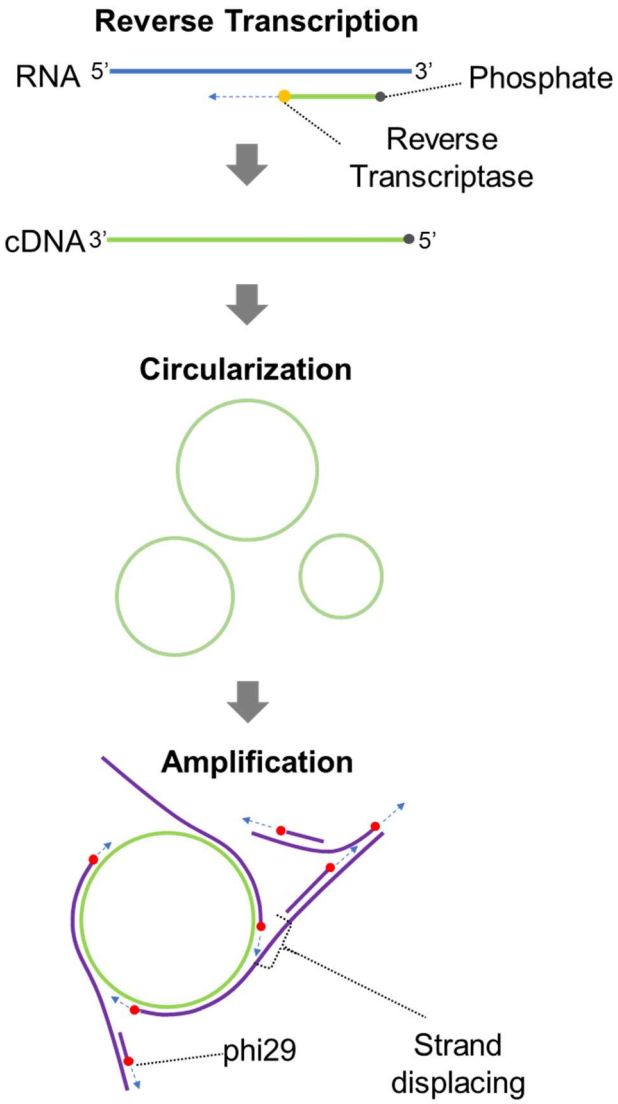


Figure 2.1 Diagram showing the concept of circular whole transcriptome amplification. Complementary DNA was synthesized using random 9-mer primer with 5'-phosphate modification. The resulted cDNA then circularized, before being amplified with random 6-mer primer and ph29 polymerase.

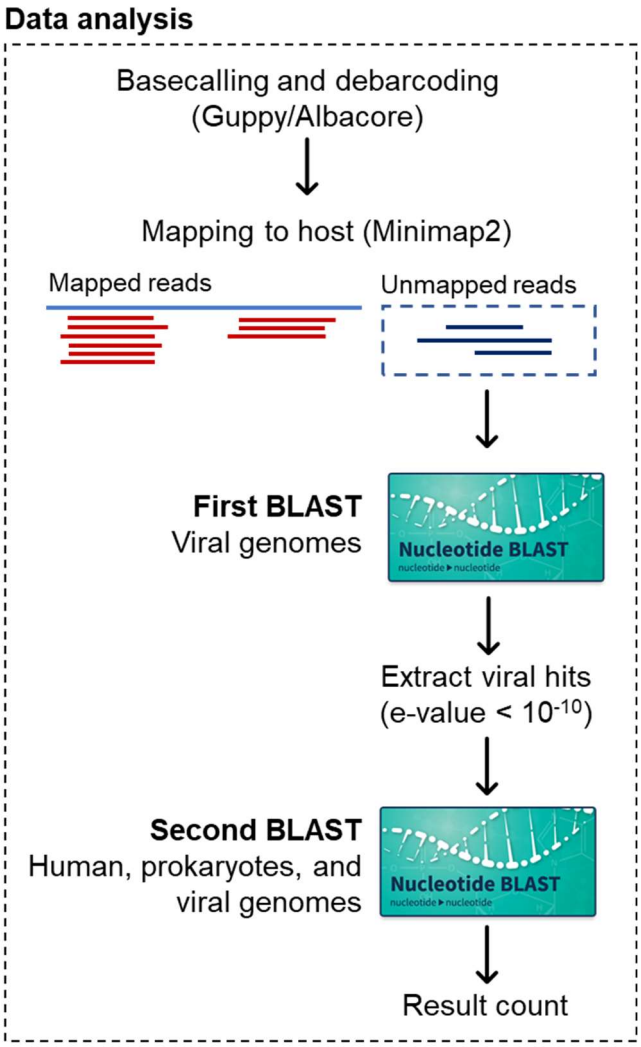


Figure 2.2 Schematic diagram of bioinformatic analysis. Raw sequence data were basecalled using Guppy or Albacore (depends on the kit version). The data were then decontaminated from host sequence by aligning it to the host genome. The identification of virus is based on two steps BLAST search as described previously (82).

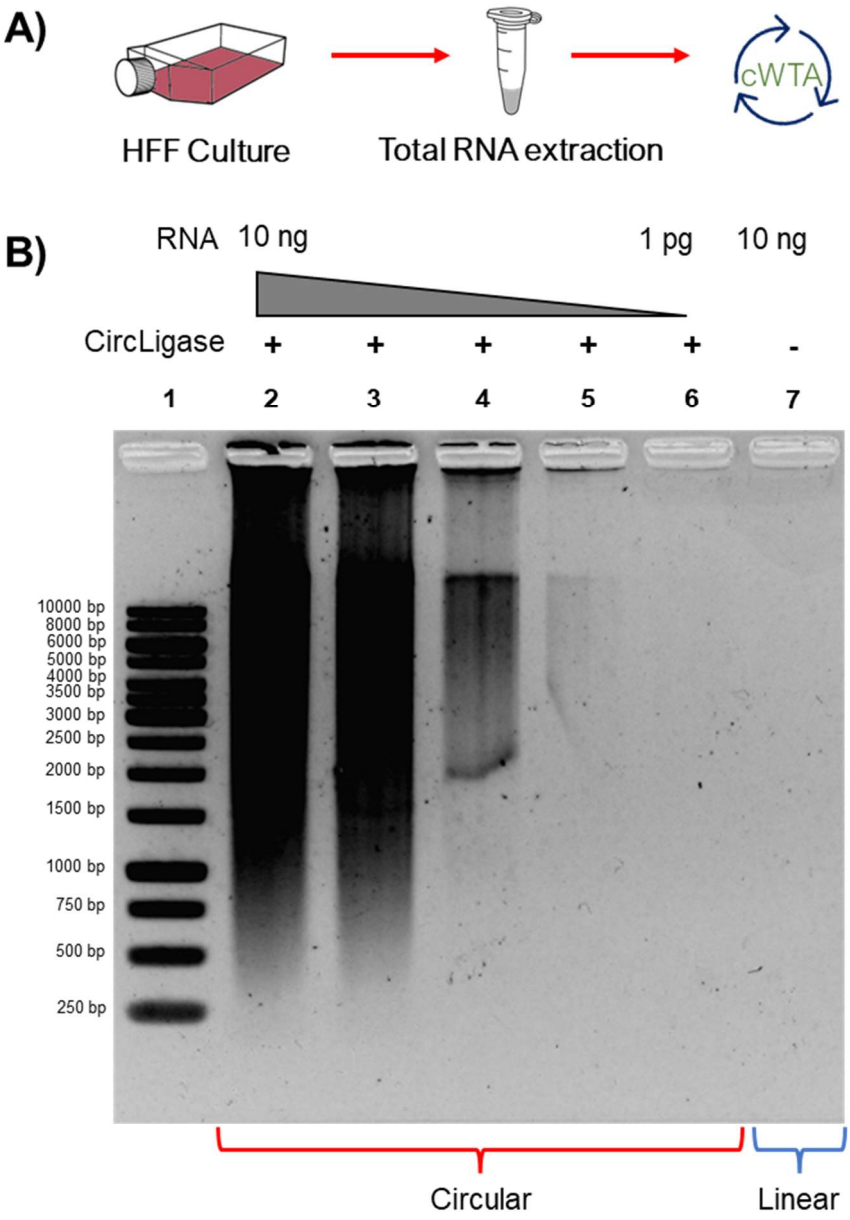


Figure 2.3 Amplification of total RNA with cWTA. A) The scheme of the experiment. Total RNA extracted from approximately 10^6 HFF cells were subjected to cWTA. B) Gel electrophoresis result showing amplicons can be observed up to 10 pg (lane 5) when the template is circularized, but no observable bands when the template is linear (lane 7). Lane 1, 1 kb marker; lanes 2 to 6, amplicons from reactions containing 10 ng, 1 ng, 100 pg, 10 pg, and 1 pg total RNA; lane 7, reaction containing 10 ng without circularization.

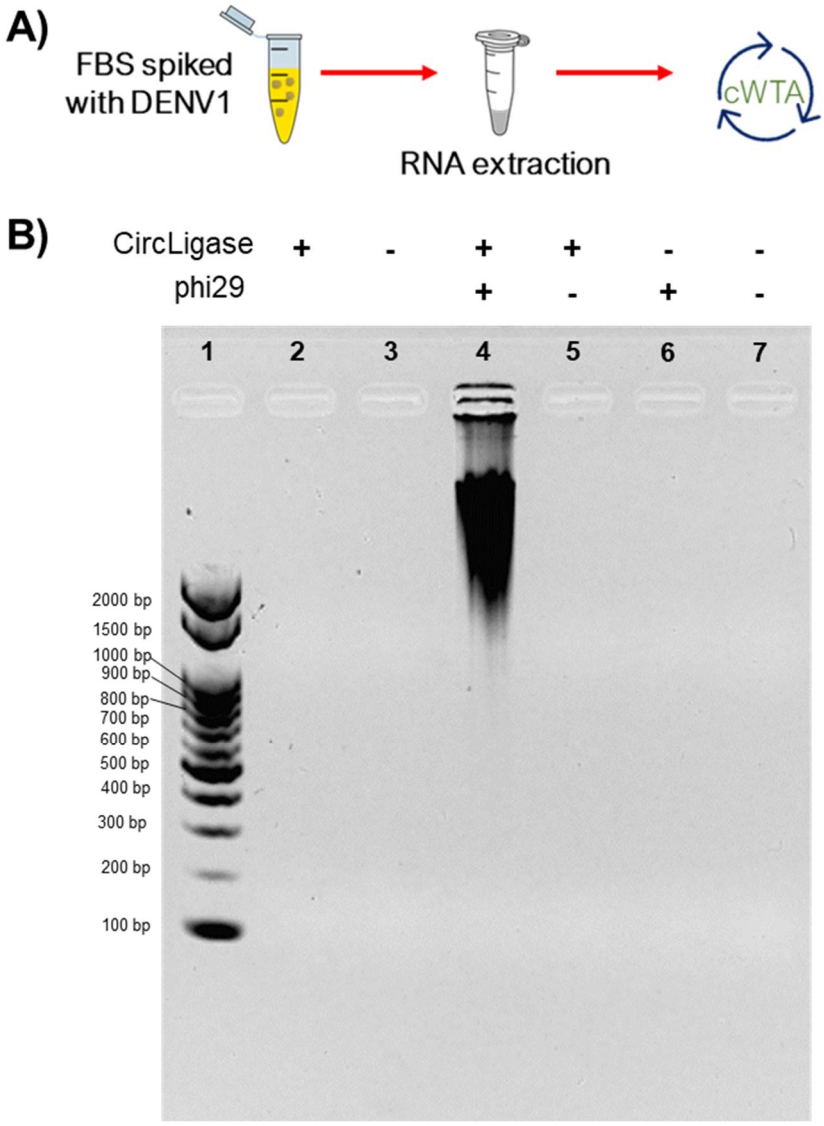


Figure 2.4 Amplification of total RNA with cWTA. A) The scheme of the experiment. Viral RNA were extracted from a spiked FBS sample containing 10^5 PFU/mL of DENV1. B) Gel electrophoresis result showing amplicons when the template is circularized, but no observable bands when the template is linear. Lane 1, 100 bp marker; lane 2, stock circularized cDNA; lane 3, stock linear cDNA; lane 4, circularized template amplified with phi29; lane 5, circularized template without amplification; lane 6, linear template amplified with phi29; line 7, linear template without amplification.

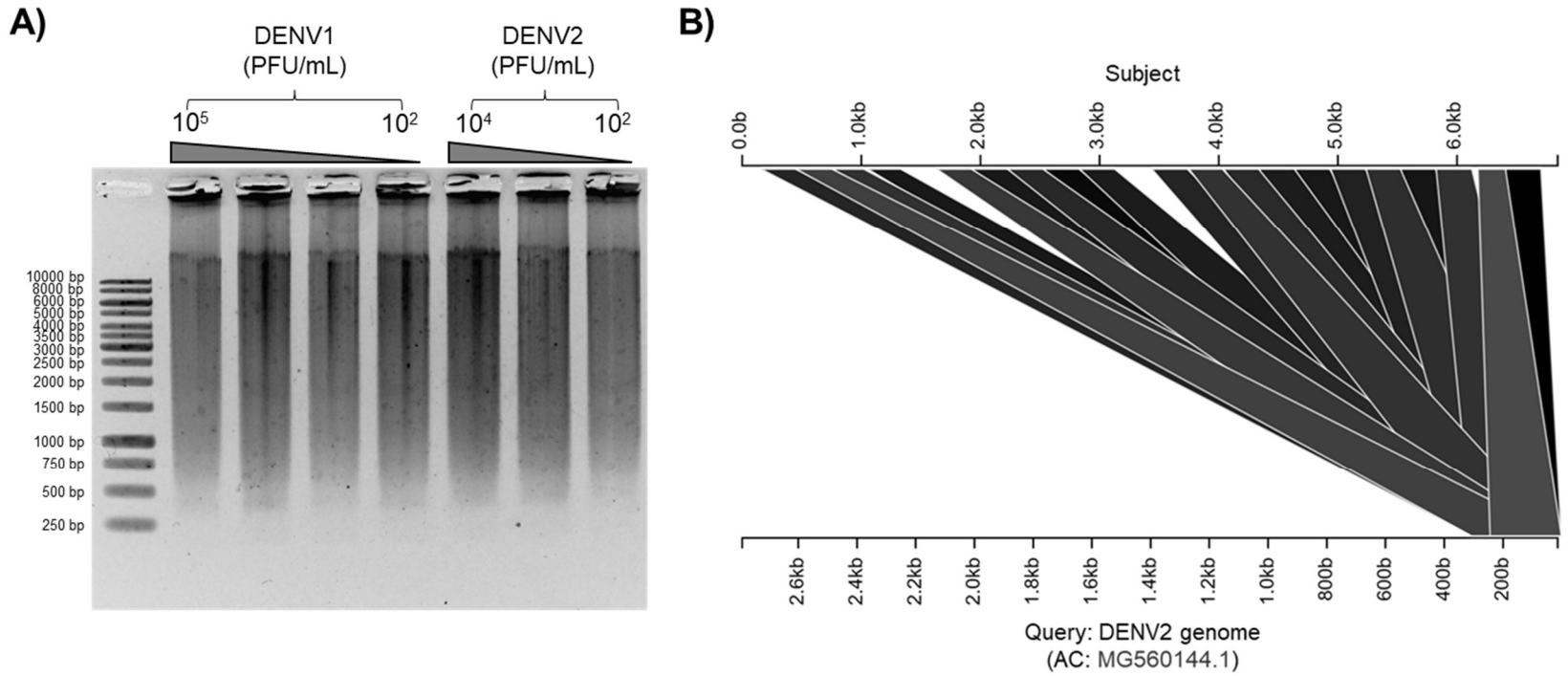


Figure 2.5 Visualizations of cWTA amplicon. A) Amplification of RNA in the spiked samples. Amplification was observed in FBS spiked with DENV1 and DENV2 virions. Lane 1, 100 bp marker; lanes 2 to 4, 10^5 to 10^3 PFU/mL DENV1; lanes 5 to 7, 10^4 to 10^2 pfu/mL DENV2. B) Visualization of the alignment result of one read obtained from nanopore sequencing showing that the 6 kbp read consists of an approximately 300 bp tandem repeat. The trapezoids represent projection of the subject onto the query. The color represents the quality of alignment, with darker shade meaning stronger alignment (based on e-value and bit-score).

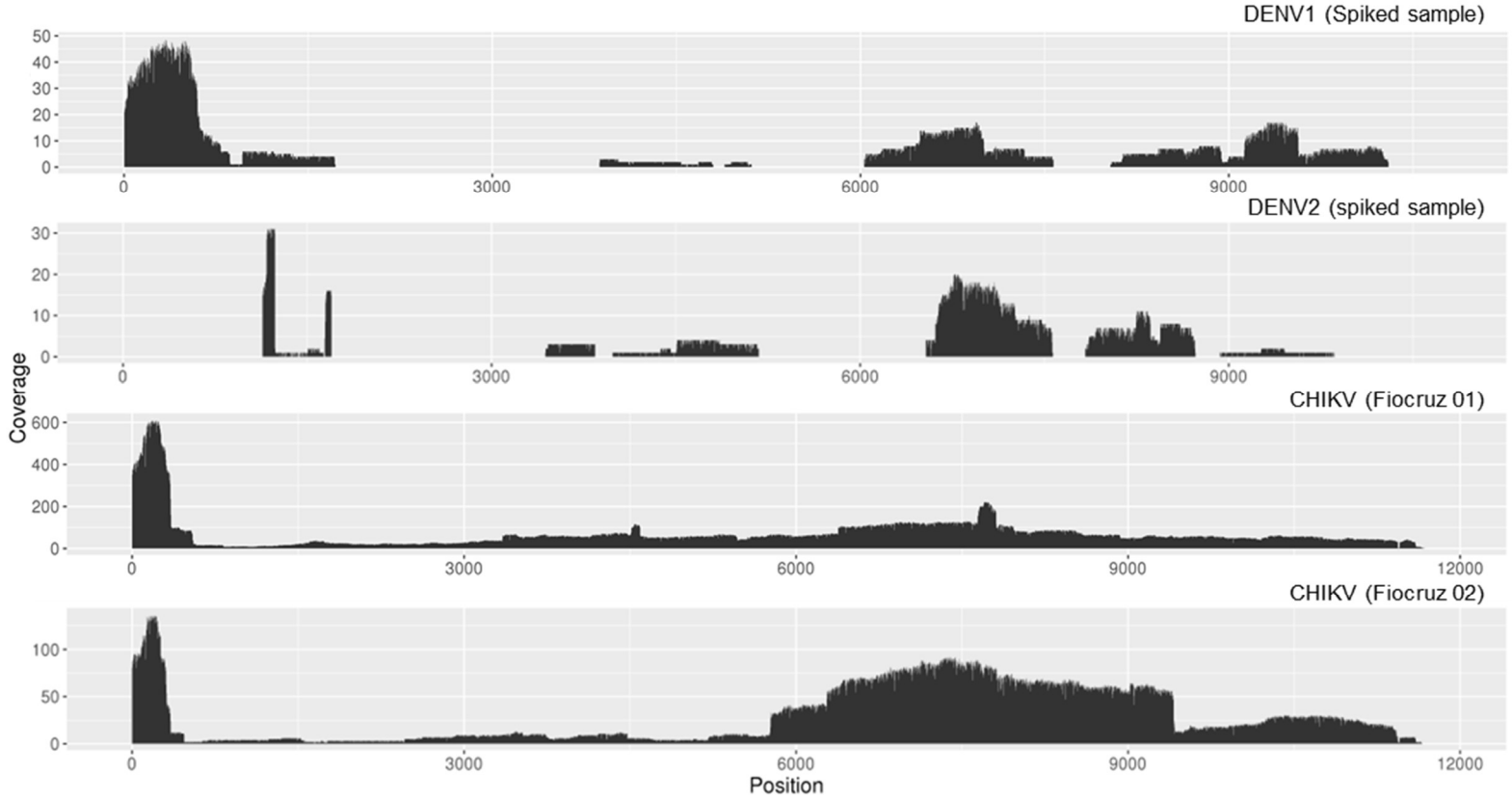


Figure 2.6 Depth and coverage obtained from mapping viral reads obtained from nanopore sequencing to its respective viral reference genome. The *y* axis denotes the coverage depth, and the *x* axis denotes the genome position. The accession numbers for the reference genomes were NC_001477 (DENV1), and NC_004162 (CHIKV).

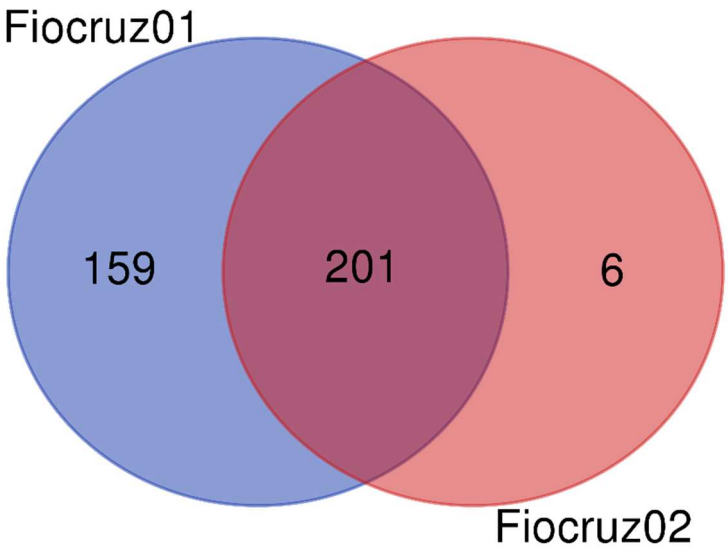


Figure 2.7 Number of shared and unique SNPs between sequences mapped to CHIKV detected in Fiocruz01 and Fiocruz02. CHIKV reads were extracted and aligned to the reference genome using Minimap2. The variant calling was performed using LoFreq.

Chapter Three

High Throughput and Comprehensive Virus Identification with Metagenomic Next Generation Sequencing Enhanced by A Group Testing Algorithm (mEGA)

3.1 Summary

A single test that can detect multiple pathogens is an ultimate approach for diagnosing infectious diseases. A hypothesis free mNGS, can serve as a comprehensive detection system. However, it is constrained by a complicated library construction as it is often costly and complicated. Furthermore, to increase the sample throughput and to further reduce the cost per sample, a combination of cWTA, Illumina sequencer, and group testing algorithm were employed to detect viral pathogens in 44 serum samples obtained in Vietnam, which included one sample known to be positive for DENV1 as a positive control. Combination of mNGS and cWTA were termed mNGS enhanced by a group testing algorithm (mEGA). Due to the large number of samples being pooled together, Illumina sequencer which can generate more reads was used. Utilizing a group testing algorithm, the number of sequencing libraries needed to be constructed were reduced to 11 libraries, while retaining the sample information. The origin of viral reads detected in the pools were able to be traced back to each sample, which were validated using pathogen-specific test. From these 43 samples (one sample is positive control), DENV2, hepatitis B virus (HBV), and human parvovirus B19, were able to be detected without prior knowledge.

3.2 Introduction

The application of mNGS has been shown to be useful for identifying pathogens in clinical settings. Many studies have been conducted with various findings. Several studies have reported that mNGS was able to detect more positive samples when compared to the traditional method under certain situations (116,117). Studies comparing mNGS to conventional testing reported various sensitivity and specificity, ranging from 55% to 98% and 63% to 98%, respectively (41). Interestingly, one study explores utilization of nanopore sequencer to provide rapid mNGS diagnosis for upper respiratory infection with turnaround time of only six hours, much faster than the culture method (34). These reports showed that mNGS has potential as a complementary of the conventional method, or even replacing the conventional method. However, one of the major bottlenecks in application of mNGS is the library preparation. The library preparation usually consisted of multiple steps that require significant cost and labour.

Group testing algorithms can be a solution to overcome this bottleneck. At the core, group testing is how samples can be pooled to minimize the number of tests required to screen a population. In general, there are two main approaches for group testing: hierarchical and combinatorial (non-hierarchical) (**Fig. 3.1**). The first mention of group testing was a hierarchical approach proposed by Dorfman in 1943 to conduct a large screening of a sexually transmitted disease (118). He proposed that instead of testing each person, the sample can be pooled and only people in the positive pool will be tested individually. This approach would later be known as hierarchical group testing algorithm. On the other hand, in non-hierarchical or combinatorial group testing, a sample is tested several times in pools consisting of different samples (119). Despite the mixing, combinatorial group testing preserves the sample identification and makes it possible for the positive result to be traced back to its individual sample.

In the dawn of COVID-19 pandemic, several approaches have been made to make the most out of the limited testing resource using a group testing algorithm. One of the approaches is the hypercube algorithm, a form of combinatorial group testing in which 3^n samples can be pooled so that only 3n tests are needed (120). Application of this algorithm was successfully applied in COVID-19 screening in 81 samples at once using 12 tests (121). The current application of group testing is limited to certain tests targeting a

specific pathogen. It would be interesting to combine group testing with a broad range mNGS platform.

Thus, in this chapter the performance of a comprehensive NAT in combination with a combinatorial group testing algorithm has been validated, which was termed mEGA. This method enabled the number of mNGS libraries of 2^n samples to be significantly compressed to just 2n libraries, while preserving the broad pathogen coverage.

3.3 Material and Methods

3.3.1 Clinical samples, pooling, and sequencing

A total of 44 serum samples collected from patients in Nam Dinh Province, Vietnam, from May 2017 to May 2019, were used to validate mEGA. These serum samples were stored at −80°C prior to RNA extraction and they were also screened for dengue virus with NS1 ELISA antigen test (Inbios). Extraction of RNA was carried out using QIAamp Viral RNA Mini kit (Qiagen) according to the manufacturer's instructions. Samples were subjected to cWTA and later sequenced on an Illumina sequencer. One sample positive for DENV1 confirmed by RT-qPCR (CDC DENV-1-4 Real-Time RT-PCR Multiplex Assay) was included as a positive control. Serum samples were collected and stored at NIHE, Hanoi, Vietnam. Ethical approvals were obtained from NIHE (351/QD-VSDTTU) and Hokkaido University (Jinjyu30-1).

The extracted RNA from all the 44 samples were subjected to cWTA as described in section 2.4.1. The amplicons from each sample were then pooled into 11 pools, encoded by number (1 to 5) and letter (a, b, or c). The details of the pools can be found in **Table 3.1**. In brief, amplicons from a sample was pooled in five different pools. The combination of these 5 pools was unique to each sample and thus, allowed for the sample ID to be traced back. Prior to pooling, a 500 ng of cWTA product from each sample was diluted with nuclease-free water to a final volume of 10 µL. Then, 1.5 µL from each diluted sample were added to five different pools as assigned from the table. From each pool, 200 ng of DNA were used to construct the sequencing library. Sequencing libraries were prepared with TruSeq Nano DNA Library prep (Illumina) according to the manufacturer's instructions. Libraries were then quantified with a fluorometer, and the quality was assessed with Bioanalyzer (Agilent). The concentration of the libraries was

adjusted to a final concentration of 4 nM. Sequencing was performed using the Illumina MiSeq to generate 2 × 150 bp, paired end reads.

3.3.2 Bioinformatic pipeline

Essentially, the pipeline is similar to the pipeline explained in section 2.3.4 but tailored for a short-read platform. Adapter sequences were r-moved, and the quality filter of bases was carried out with Cutadapt version 2.10 (122). Reads passing filters were then processed with Trimmomatic version 0.3.9 (123) for second adapter removal and quality filtering. The parameters set for Trimmomatic were as follows: leading 3, trailing 3, sliding window [4, 15], and minimal length 36. Surviving sequences were then filtered for low-complexity sequencing with Komplexity (124) version 0.3.6 (the complexity score threshold set to 0.55). Filtered reads were then mapped against the human sequence (GCF_000001405.39) using Bowtie2 (125) version 2.2.4. Unmapped reads were extracted using SAMtools (98) using the following flags: -f 12 -F 256. Decontaminated reads were then subjected to three pipelines for detection of viral reads. First tool is a two-step alignment search with BLAST (as described in section 2.3.4). In short, the unmapped reads were subjected to BLAST using a custom viral database (83). The word size parameter was set to 11. Based on the result of the first BLAST, the sequences aligned to the database with e-value less than 1.0e-10 were extracted and subjected to the second BLAST search which used a database comprised of the viral sequences, human genome, and representative prokaryotic genomes. The sequences were assigned to the top-hit feature based on the bit-score. If a read has multiple hits, with identical bit-score value, the read will not be counted. Additionally, because the sequencer generated a paired end read, if the hit with the highest bit-score for R1 is not the same as the hit with the highest score for R2, then the read will not be counted. The schematic diagram of the pipeline and result counting can be seen in **Fig. 3.2.** As a comparison, the reads were also subjected to two other pipelines. First comparison tool is DAMIAN (Detection & Analysis of viral and Microbial Infectious Agents by NGS) (126). This tool performed de novo assembly (using IDBA-UD) from the decontaminated reads, to obtain a longer contig, which will give higher specificity in the alignment search with MEGABLAST. The third tool is VIRTUS (VIRal Transcript Usage Sensor), which is made specifically for analysing transcriptomic data (127). While the two other pipelines used BLAST for similarity

search, VIRTUS aligns the host-decontaminated reads to a database of virus using STAR. The mapped reads were then counted.

3.3.3 Cross validation

To increase the confidence of the findings (with mEGA being an unbiased approach), subsequent pathogen specific PCR must be carried out to validate the findings. The list of primers used for cross validation can be found in **Table 3.2** (128–133). Each 10 µL reaction consisted of 1 µL template, 2 µL Taq Buffer, 0.8 µL deoxynucleotide triphosphates (5 mM each), 0.05 µL ExTaq Hot Start polymerase (TaKaRa), and 200 to 500 nM of each primer (forward and reverse). The cDNA synthesized from extracted RNA (linear) was used as the template, except for probable human immunodefiency virus 1 (HIV-1) samples, in which the cWTA product was used due to limited amount of the remaining cDNA. Cycling conditions for each PCR can be found in **Table 3.2.** For semi nested or nested PCR, the template for the second PCR was 100 times diluted product of the first PCR. As much as 1 µL PCR products were visualized with gel electrophoresis on 1.5% agarose gels. Furthermore, all PCR amplicons were cloned into the pGEM-T vector (Promega) and transformed into *Escherichia coli* D5Hα. The amplicons were cloned to the plasmids due to some amplicons being too short to be sequenced using Sanger sequencing. The cloned plasmids were then subjected to PCR with the SP6 and T7 primers. The amplicons were then processed for Sanger sequencing.

3.4 Results

3.4.1 Sequencing statistics

Combination of cWTA and mEGA was performed on a set of serum samples collected in Vietnam. The group testing algorithm significantly reduced the number of sequencing libraries (from 44 to 11), while preserving the information related to sample number. Total raw reads obtained ranged from 2,852,907 to 5,054,500 (median, 4,260,704). On average, 16.41% ± 0.5% (mean ± standard deviation) of the raw reads were filtered out during the quality control. Up to half of the raw reads (57.97% ± 0.99%) were aligned to the human genome and then subsequently filtered out. The remaining reads (25.61% ± 1.02%) were then subjected to the alignment search with BLAST. The number of reads with hits to a certain pathogen can be seen in **Table 3.3**.

3.4.2 Detection of viral reads

In general, the BLAST-based pipeline and the two other pipelines detected the same viruses **(Table 3.4)**. These viruses are DENV1, DENV2, HBV, human parvovirus B19, and HIV-1. However, the comparison pipelines failed to detect certain viruses in certain pools. Viruses with limited read were particularly difficult to be detected in DAMIAN due to the contig-building step. Indeed, the contig-based approach increases the specificity as demonstrated by not detecting human erythrovirus V9 (a closely related virus to human parvovirus B19). Meanwhile, VIRTUS and BLAST-based pipeline showed an almost identical result, except for HIV-1. The inflated read count in VIRTUS is due to the paired reads being counted individually as two reads. While the difference in read number is likely due to difference in database used and also difference in the algorithm between BLAST search and STAR alignment. Taken altogether, the BLAST-based pipeline was able to detect more viruses compared to other pipelines.

Additionally, reads homologous to human erythrovirus V9 and torque teno virus (TTV) were detected **(Table 3.3)**. Human erythrovirus V9 and human parvovirus B19 are closely related viruses, and it is highly likely that these reads are misassigned reads. To investigate it, the read assigned as human erythrovirus V9 (n = 258 reads) was extracted and aligned to the reference genome. The result showed that the reads were mapped only to a certain part of the genome **(Fig. 3.3)**, suggesting that this result was most probably an artifact. Alignment of human erythrovirus V9 genome and genome of several human

parvovirus B19 strains shows that these regions are highly similar among the virus and thus explained the misassignment. In addition, some reads have BLAST with hits for both human parvovirus B19 and human erythrovirus V9 with identical bit-score and e-value (n = 1,198 reads, including the 258 reads mentioned above). As it is impossible to tell them apart, these reads (total 1198 reads, 0.4% of all detected parvovirus B9 and human erythrovirus V9 reads) were filtered out. On the other hand, reads homologous to TTV were only detected in less than five pools, not sufficient to trace the sample origin.

Interestingly, among these three tools, only the BLAST-based pipeline was able detect HIV-1 in at least five different pools (enough to track the sample number). Alignment to the reference genome showing that the reads were aligned to only a particular region, raising the possibility that these reads are an artifact, or a possibility that it came from unclassified human endogenous retrovirus **(Fig. 3.4)**. On the contrary, at least in pool 2a, HIV-1 can be detected by all pipelines, suggesting that the HIV-1 sequence does exist among the samples. However, further testing using pathogen-specific PCR would be necessary to confirm these findings.

It is also known that index hopping is a common issue when performing multiplex sequencing. Using the same strategy applied in Chapter One and by assuming 0.1% as the index hopping rate, it is possible to estimate the possible number of reads caused by the index hopping (134,135). Number of reads (μ) that resulted in upper cumulative Poisson probability ($P(x \geq \mu)$) less than 0.01 was determined as the cut-off value. Based on this approach, the cut-off value for B19 was 341 reads; then reads in pools 1a, 3a, and 4a were a possible migration from the other pools (**Table 3.5**).

3.4.3 Identification of samples from which the viral read originated from.

By comparing the pool in which a specific virus was detected to the pool where a sample was pooled, the origin of the sample can be traced back. For example, a DENV1 positive sample was included as a positive control as sample no. 67. This sample was pooled into pool 1a, 2a, 3a, 4b, and 5b (**Table 3.1, Fig. 3.5**). The BLAST result showed that DENV1 reads were detected in those five pools, and therefore it can be inferred that the DENV1 reads must have originated from sample no. 67. Using the similar logic, reads homologous to DENV2 in pools 1a, 2a, 3a, 4a, and 5b originated from sample no. 8. While reads homologous to human parvovirus B19 and hepatitis B virus (HBV) were presumed to

originate from samples no. 98 and no. 99, respectively. In case of HIV-1, there are two pools with HIV-1 reads in the pool group 5, and thus there are two possible samples (no. 96 and no. 97) from which the HIV-1 read might have originated.

Presence of DENV1, DENV2, HBV, and human parvovirus B19 RNA or DNA in the cDNA sample were able to be validated using pathogen specific PCR (**Fig. 3.6**). The amplicons were then cloned to a vector, multiplied, then subjected to Sanger sequencing. Similarity search of the amplicons sequence returned a hit with more than 95% to each respective virus target (**Table 3.6**). For HIV-1, several primer sets were not able to yield a positive result. Thus, a specific nested primer set was generated to target the location in which the HIV-1 reads were aligned (the primer locations can be seen in **Fig. 3.7**). Additionally, due to the limited amount of the samples, the cWTA amplicon was diluted and used as the template. Amplicons were obtained from sample no. 97 but not sample no. 96 and when subjected to a BLAST search, the sequence was found to be homologous to an HIV-1 group M subtype B isolate, with 92.49% identity.

3.5 Discussion

In this chapter, a combinatorial group testing algorithm was successfully combined with mNGS for a broad range of pathogen detection. Unlike the hierarchical group testing algorithm, all samples were included in multiple pools (in this case five pools), resulting in each sample having distinct combinations of five pools they are in. This information was later used to trace the origin of the samples, which is the most promising point of the group testing algorithm. Using this combinatorial strategy, 2^n samples can be pooled into 2n libraries. Sequencing of such pooled samples would naturally require a greater depth, and thus making nanopore sequencing less useful for this purpose. Instead, for a high throughput screening, short-read sequencing (Illumina) was employed. Indeed, 44 amplicons were able to be pooled into eleven libraries and viral reads were successfully obtained.

Sequencing of these 11 pooled samples was able to detect the positive control (DENV1) and additional three viruses (DENV2, human parvovirus B19, and HBV) without prior knowledge. Interestingly, the four viruses pose different genomic characteristics (single-stranded RNA, ssDNA, and partially double-stranded DNA) suggesting the broad spectrum of this approach. Mapping of other virus reads to the reference genome of the corresponding virus resulted in 30.45% to 100% coverage, denying presumably artifacts. Reads aligned to human parvovirus B19 were sufficient to cover the entire genome. There could be two explanations to this observation. Firstly, it is presumed that this observation was due to the high viral load of human parvovirus B19 in the serum sample, which was supported by a previous report (136). Secondly, it is possible that the ssDNA genome of the virus was extracted from intact virion found in serum and undergone preferential amplification by phi29. The alignment result showed that the reads are mapped across the viral genome, supporting the notion that the reads come from genomic DNA. Nevertheless, the amplification and sequencing method showed that both RNA and DNA virus can be detected.

In contrast, when the reads assigned to HIV-1 were extracted and aligned to the reference genome, the reads are aligned around the *gag-pol* region. It is possible that these results could be an artifact. However, a PCR assay targeting this particular region resulted in amplicons with 92.5% similarity with an isolate of HIV-1 subtype B. Given the high diversity in *gag* sequence among HIV-1 subtypes (up to 35%) and the diversity within

the genome of HIV-1 subtypes (5% to 10%), it is possible that these reads are truly coming from HIV-1 (137,138). It can be speculated that the low number of reads might be a result from low viral load because of anti-retroviral therapy. Unfortunately, clinical history of HIV diagnosis and treatment for this patient is unavailable.

In a group testing algorithm, the number of samples that can be pooled will have a direct consequence on the sensitivity. Simply speaking, pooling a positive sample among 20 samples is equal to 20× dilution, at least when using a pathogen-specific test. This study did not assess the effect of the pooling on the sensitivity. However, when NGS is employed, the problem can potentially be resolved by getting more reads (137). Application of short-read NGS sequencer might offer additional benefit by providing larger sequencing depth, and thus increase the chance for the viral reads to be detected.

Group testing algorithm was originally developed for the effective diagnosis for syphilitic antigen in 1974 (118). Currently, a derivative of the algorithm was also applied for the screening of SARS-CoV-2 (120). A hypercube algorithm that creates 3n pools for 3^n samples instead of making 2n pools for 2^n samples was used in this study; however, essentially in both approaches the number of libraries grew in a logarithmic fashion (**Fig. 3.8**). In contrast, an algorithm that comprehensively identifies multiple pathogens was expanded, in which pathogens could be identified without prior knowledge of their origin. Therefore, the identification of potential pathogens from patients with FUO in hospital settings, retrospective screening of biobank samples for identification of potentially zoonotic pathogens, and routine testing of blood transfusions can be optimised. In contrast, similar to other group testing algorithms, identification/tracing of the sample becomes impossible if the same pathogen reads are detected from more than two additional pools, because the number of potentially positive samples will exponentially increase according to the number of additional pools. Ideally, when the samples are grouped into two pools (i.e., 1a and 1b), the pathogen would be detected only from one of the pools. However, when the pathogen reads can be detected from both pools, the number positive samples will become two. In short, when pathogen reads can be detected from both pools in a column (**Table 3.1**), the number of possibly positive sample will be two and it will increase in an exponential manner for every column containing positive pools. Nevertheless, this disadvantage is not critical because the major targets of the method are rare, undiagnosed, neglected pathogens. In addition, if enough reads were obtained for

pathogens, genomic polymorphism could be employed to discriminate the same pathogens originated from different samples.

This chapter describes mEGA and its validation as an application of the group testing algorithm for comprehensive pathogen detection, and to reduce the number of sequencing libraries that need to be constructed to successfully identify unknown pathogens without prior knowledge. In the previous chapter, it has been discussed that given the affordability and portability of the nanopore sequencer and the isothermal nature of cWTA, the workflow is feasible for more peripheral laboratories with limited settings. In contrast, application of mEGA using the Illumina platform will improve the scalability, making it more suitable for large-scale screening. Lastly, although this study has only focused on viruses, it is theoretically possible for this pipeline to provide a broad range pathogen detection, including protozoa, fungi, and bacteria. Taken together, this approach and workflow may provide one of the better practices in NAT for a comprehensive approach for pathogen identification.

Table 3.1 List of samples and their respective pools. Each sample is added into five different pools denote by the pool in each row.

Sample no	Pools				
5	1a	2a	3a	4a	5a
8	1a	2a	3a	4a	5b
16	1a	2a	3a	4a	5c
36	1a	2a	3a	4b	5a
67	1a	2a	3a	4b	5b
72	1a	2a	3a	4b	5c
73	1a	2a	3b	4a	5a
74	1a	2a	3b	4a	5b
75	1a	2a	3b	4a	5c
76	1a	2a	3b	4b	5a
77	1a	2a	3b	4b	5b
78	1a	2b	3a	4a	5a
79	1a	2b	3a	4a	5b
80	1a	2b	3a	4a	5c
81	1a	2b	3a	4b	5a
82	1a	2b	3a	4b	5b
83	1a	2b	3a	4b	5c
84	1a	2b	3b	4a	5a
85	1a	2b	3b	4a	5b
86	1a	2b	3b	4a	5c
87	1a	2b	3b	4b	5a
88	1a	2b	3b	4b	5b

Sample no	Pools				
89	1b	2a	3a	4a	5a
90	1b	2a	3a	4a	5b
91	1b	2a	3a	4a	5c
92	1b	2a	3a	4b	5a
93	1b	2a	3a	4b	5b
94	1b	2a	3a	4b	5c
95	1b	2a	3b	4a	5a
96	1b	2a	3b	4a	5b
97	1b	2a	3b	4a	5c
98	1b	2a	3b	4b	5a
99	1b	2a	3b	4b	5b
100	1b	2b	3a	4a	5a
101	1b	2b	3a	4a	5b
102	1b	2b	3a	4a	5c
103	1b	2b	3a	4b	5a
104	1b	2b	3a	4b	5b
105	1b	2b	3a	4b	5c
106	1b	2b	3b	4a	5a
107	1b	2b	3b	4a	5b
108	1b	2b	3b	4a	5c
109	1b	2b	3b	4b	5a
110	1b	2b	3b	4b	5b

Table 3.2 List of primers used for cross validation and their PCR condition.

Target pathogen	Name	Sequences (5' - 3')	Cycling condition	Reference
B19	e1905f	TGCAGATGCCCTCCACCCA	45 cycles of 95°C for 30 seconds, 60°C for 1 minute.	(128)
	e1987r	GCTGCTTTCACTGAGTTCTTC		
HBV	SP1s	GCTCCGACTATTGCCTCTCTCACA	45 cycles of 95°C for 30 seconds, 60°C for 1 minute.	(129)
	SP1a	TGTAACACGAGAAGGGGTCCTAGGA		
DENV	D1	TCAATATGCTGAAACGCGCGAGAAACCG	1st PCR (primer D1 and D2): 40 cycles of 94°C for 30 seconds, 55°C for 1 minute, 72°C for 2 minutes. 2nd PCR (primer D1 and TS1/TS2): 20-25 cycles f 94°C for 30 seconds, 55°C for 1 minute, 72°C for 2 minutes.	(130)
	D2	TTGCACCAACAGTCAATGTCTTCAGGTTC		
	TS1	CGTCTCAGTGATCCGGGGG		
	TS2	CGCCACAAGGGCCATGAACAG		
HIV-1	6F-HIV	CATGTTTTCAGCATTATCAGAAGGA	45 cycles of 95°C for 30 seconds, 60°C for 1 minute.	(131)
	84R-HIV	TGCTTGATGTCCCCCCACT		
	HIV-intF	CCCTACAATCCCCAAAGTCA	35 cycles of 95°C for 30 seconds, 55°C for 30, 72°C for 30 seconds.	(132)
	HIV-intR	CTTGCCACACAATCATCACC		
	SK39	TTTGGTCCTTGTCTTATGTCCAGAATGC	1st PCR (SK39 and SK145): 45 cycles of 95°C for 30 seconds, 60°C for 1 minute. 2nd PCR (SK39 and SK101): 25 cycles of 95°C for 30 seconds, 60°C for 1 minute.	(133)
	SK101	GCTATGTCAGTTCCCCTTGGTTCTC		
	SK145	AGTGGGGGGACATCAAGCAGCCATGCAAAT		
	HIV1-outer-F	AGGGCTGTTGGAAATGTGGA	1st PCR (outer primer): 45 cycles of 95°C for 30 seconds, 60°C for 1 minute. 2nd PCR (inner primer): 25 cycles of 95°C for 30 seconds, 60°C for 1 minute.	This study
	HIV1-outer-R	ACGTTGACAGGTGTAGGTCC		
	HIV1-inner-F	TGGAAATGTGGAAAGGAAGG		
	HIV1-inner-R	GCCAAAGAGTGATTTGAGGGC		

B19, human parvovirus 19; HBV, hepatitis B virus; DENV, dengue virus; HIV, human immunodeficiency virus.

Table 3.3 Total raw reads obtained from sequencing, reads dropped during downstream analysis, viral reads, and confirmation PCR results of Vietnam samples using cWTA and mEGA.

Pool	Raw reads	Reads dropped from QC (%)		Reads mapped to host (%)		Unmapped reads (%)		Reads							
								DENV1	DENV2	B19	HIV-1	HBV	TTV8	TTV24	V9
1a	3,548,149	594,485	(16.75)	2,025,030	(57.07)	928,634	(26.17)	**9**	**84**	**4**	0	0	0	0	0
1b	4,776,701	780,689	(16.34)	2,815,856	(58.95)	1,180,156	(24.71)	0	0	**47,820**	**2**	**3**	0	**1**	0
2a	2,852,907	471,288	(16.52)	1,632,297	(57.22)	749,322	(26.27)	**5**	**101**	**34,861**	**8**	**5**	0	**1**	0
2b	3,960,684	639,447	(16.14)	2,348,298	(59.29)	972,939	(24.56)	0	0	0	0	0	**1**	0	0
3a	3,616,010	600,703	(16.61)	2,055,922	(56.86)	959,385	(26.53)	**10**	**37**	**5**	0	0	0	0	0
3b	5,054,500	859,593	(17.01)	2,919,607	(57.76)	1,275,300	(25.23)	0	0	**73,521**	**12**	**3**	0	0	**1**
4a	4,861,811	839,593	(17.27)	2,785,501	(57.29)	1,236,717	(25.44)	0	**137**	**31**	**4**	0	0	0	0
4b	4,664,334	737,011	(15.80)	2,731,045	(58.55)	1,196,278	(25.65)	**22**	0	**46,792**	0	**1**	0	0	0
5a	4,782,575	768,445	(16.07)	2,682,222	(56.08)	1,331,908	(27.85)	0	0	**96,198**	0	0	0	0	**1**
5b	3,964,608	618,340	(15.60)	2,367,500	(59.72)	978,768	(24.69)	**9**	**134**	0	**1**	**2**	0	0	0
5c	4,260,704	697,628	(16.37)	2,510,115	(58.91)	1,052,961	(24.71)	0	0	0	**9**	0	0	0	0
Sample no								67	8	98	96/97	99	UD	UD	UD
Confirmation PCR								+	+	+	-	+	NT	NT	NT

Serum Samples from fever patients (n = 43) and one sample positive for DENV, were subjected to cWTA and then sequenced on Illumina platform. The bottom row shows the sample number and result of the confirmation PCR. DENV, dengue virus; B19, human parvovirus 19; HIV, human immunodeficiency virus; HBV, hepatitis B virus; TTV, torque teno virus; V9, human erythrovirus V9; UD, undetermined; NT, not tested.

Table 3.4 Comparison of result using three different identification pipelines.

Pool	BLAST						VIRTUS (Reads)						DAMIAN					
	DENV1	DENV2	V9	B19	HIV-1	HBV	DENV1	DENV2	V9	B19	HIV-1	HBV	DENV1	DENV2	V9	B19	HIV-1	HBV
1a	**9**	**84**	0	**4**	0	0	**1**	**14**	0	**6**	0	0	-	+	-	-	-	-
1b	0	0	0	**47,820**	**2**	**3**	0	0	**34**	**85,749**	**0**	**4**	-	-	-	+	-	-
2a	**5**	**101**	0	**34,861**	**8**	**5**	**2**	**28**	**34**	**62,342**	**2**	**2**	-	+	-	+	+	+
2b	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-
3a	**10**	**37**	0	**5**	0	0	**2**	**10**	0	**9**	0	0	-	+	-	-	-	-
3b	0	0	**1**	**73,521**	**12**	**3**	0	0	**66**	**131,363**	**6**	**4**	-	-	-	+	-	-
4a	0	**137**	0	**31**	**4**	0	0	**30**		**50**	0	0	-	+	-	+	-	-
4b	**22**	0	0	**46,792**	0	**1**	**10**	0	**32**	**83,860**	0	**2**	-	-	-	+	-	-
5a	0	0	**1**	**96,198**	0	0	0	0	**98**	**175,590**	0	0	-	-	-	+	-	-
5b	**9**	**134**	0	0	**1**	**2**	**4**	**32**	0	0	0	**4**	-	+	-	-	-	-
5c	0	0	0	0	**9**	0	0	0	0	0	**4**	0	-	-	-	-	-	-

For BLAST-based method and VIRTUS, number denotes number of reads assigned to a particular virus. For DAMIAN, +/- denotes the assembled contig. Cell highlighted in green are in agreements with the BLAST-based method, while cells highlighted in red are in disagreements. DENV; Dengue virus V9; human erythrovirus V9, B19; human parvovirus B19, HIV; Human immunodeficiency Virus, HBV; Hepatitis B virus; DAMIAN, Detection & Analysis of viral and Microbial Infectious Agents by NGS; VIRTUS, VIRal Transcript Usage Sensor.

Table 3.5 Sequencing reads from the pooled experiment aligned to parvovirus B19 and the subsequent threshold calculation.

Pool	B19 reads
1a	**4**
1b	47,820
2a	34,861
2b	0
3a	**5**
3b	73,521
4a	**31**
4b	46,792
5a	96,198
5b	0
5c	0
Total viral read	299,232
Index hopping rate (%)	0.1
λ*	299.232
positive threshold (≥)	341

Table 3.6 Result of BLAST search of the amplified fragment from each virus.

Query	Accession	Subject	Perc. Ident.	Length	E-value	Bit-score
B19	MK097259.1	**Primate erythroparvovirus 1 isolate B19V**_CbaAR_2017.16 nonstructural protein 1 gene, partial cds	100	103	5.00e-45	191
DENV2	AY079174.1	**Dengue virus type 2** Sullana-Peru 6682-01 capsid protein gene, partial cds	98.32	119	2.00e-50	209
DENV1	KX595191.1	**Dengue virus 1** strain Hue265/2013, complete genome	98.93	468	0	837
HBV	MH925939.1	**Hepatitis B virus** isolate TG_62 large S protein (S) gene, partial cds	97.27	110	7.00e-44	187
HIV-1	JF683794.1	**HIV-1** isolate CY251 from Cyprus, partial genome	92.49	213	4e-79	305

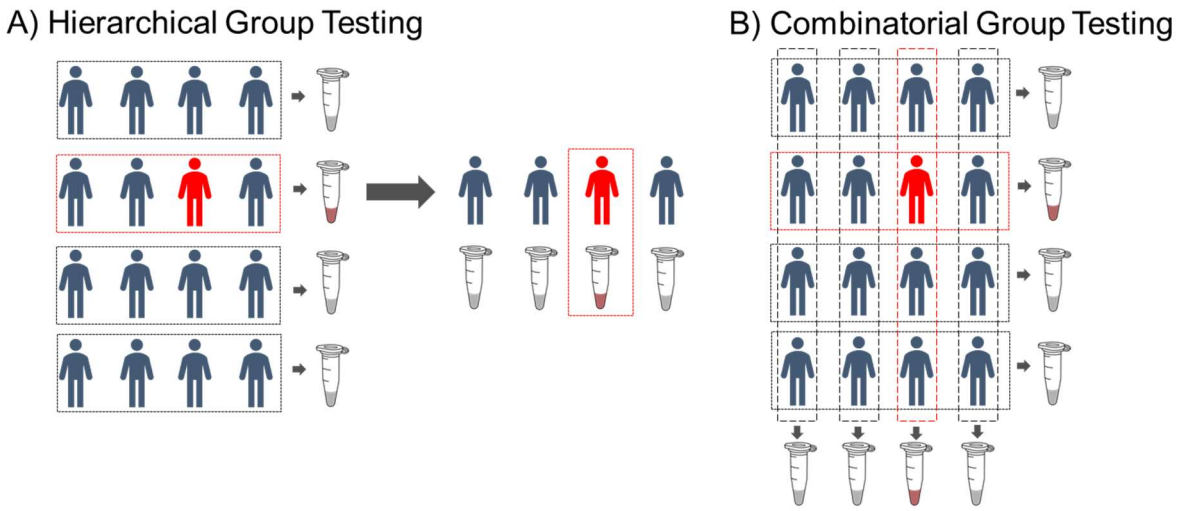
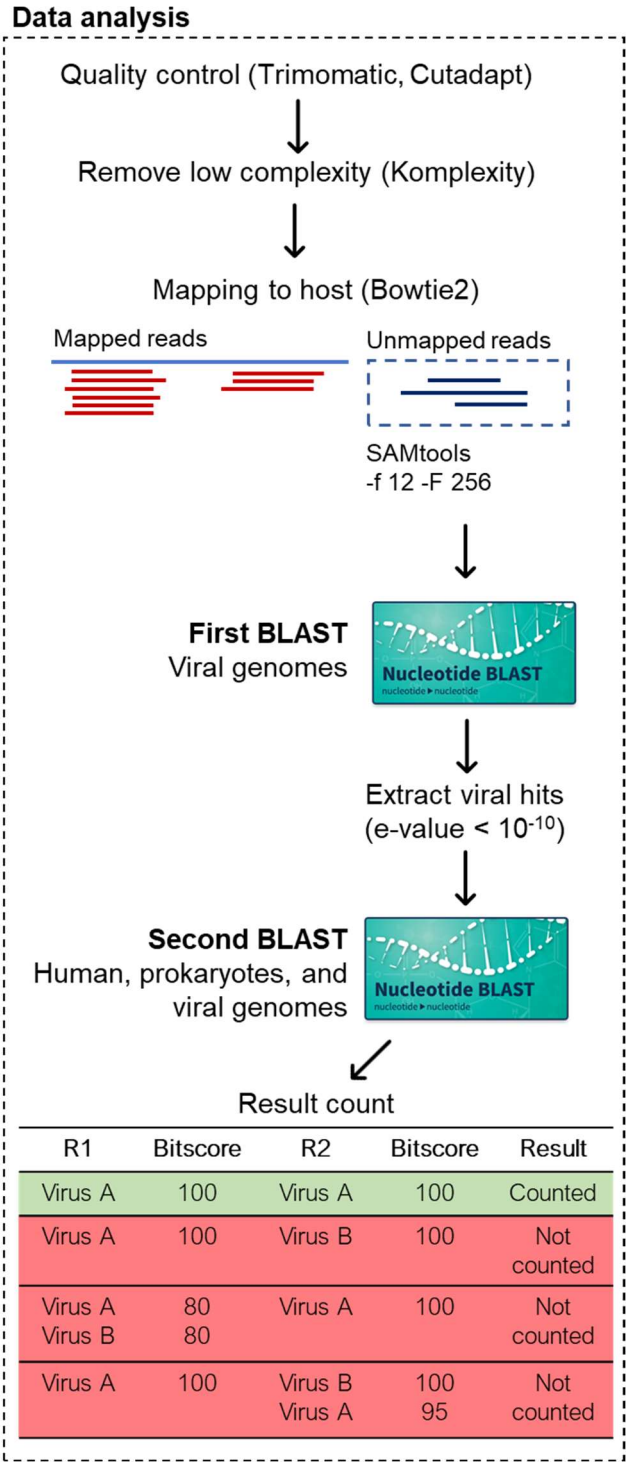


Figure 3.1 Illustration of two different group testing algorithm. A) In hierarchical group testing, samples from a group were pooled then tested. When a pathogen is detected from a pool, then the samples that were pooled in that pool will be tested individually to find the positive sample. B) In combinatorial group testing, a sample will be pooled in several pools and then tested in parallel. The positive sample then can be inferred based on the positive pools.



R1	Bitscore	R2	Bitscore	Result
Virus A	100	Virus A	100	Counted
Virus A	100	Virus B	100	Not counted
Virus A Virus B	80 80	Virus A	100	Not counted
Virus A	100	Virus B Virus A	100 95	Not counted

Figure 3.2 Schematic diagram showing the bioinformatic pipeline and how the reads were counted.

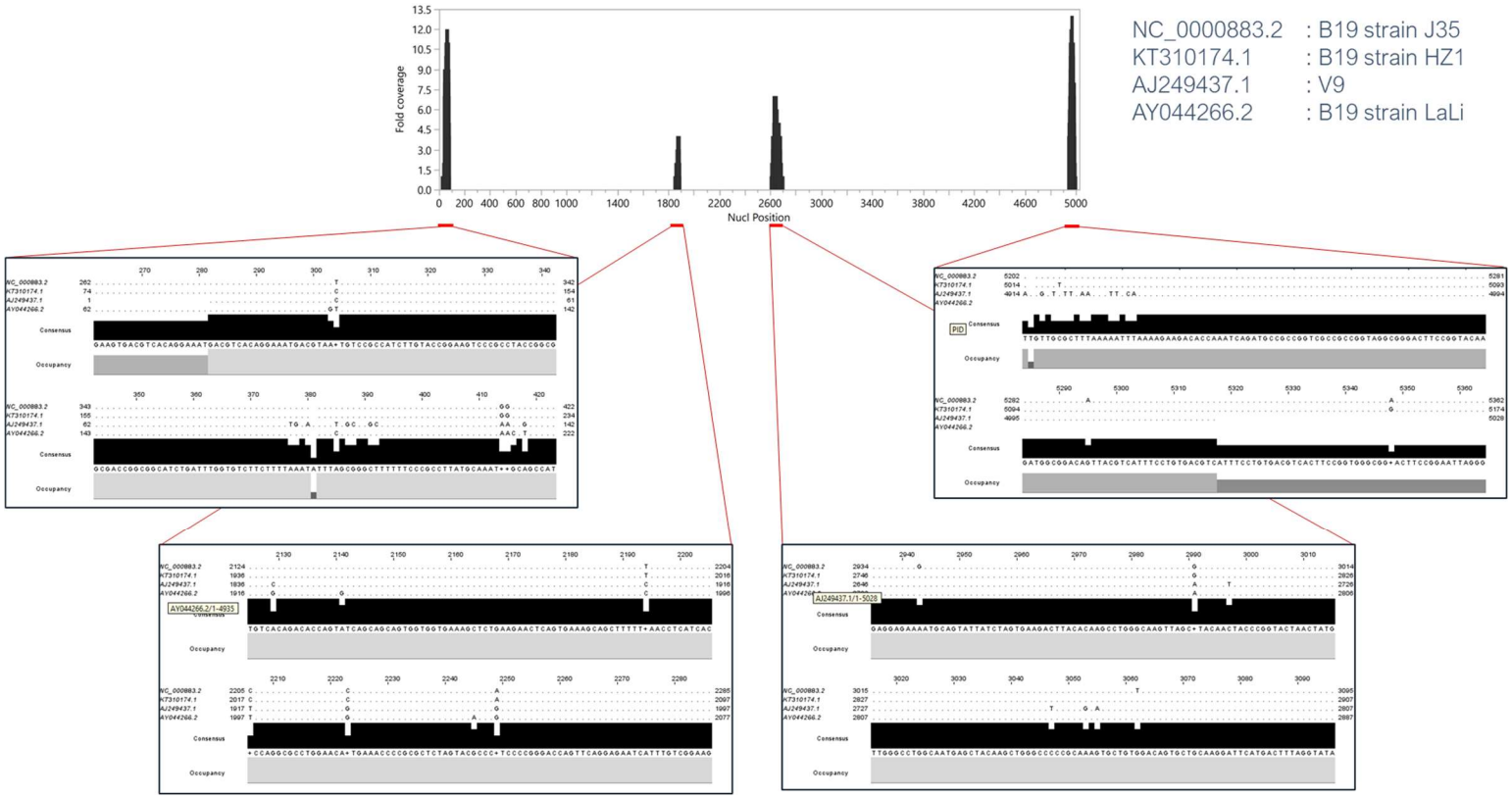


Figure 3.3 Alignment of reads assigned to human erythrovirus V9 showed that those reads are mapped in certain genomic region, which is highly similar to other parvoviruses.

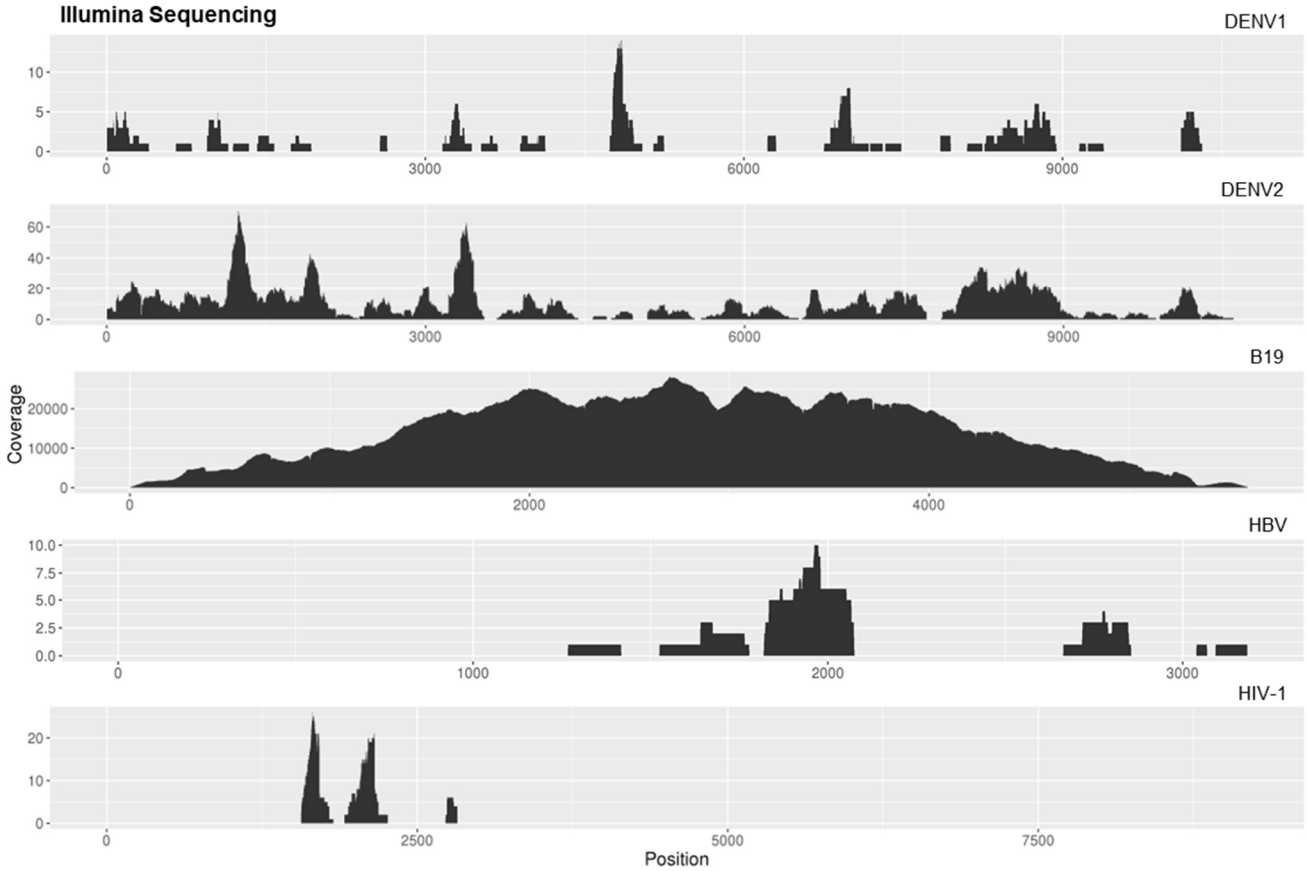


Figure 3.4 Mapping of viral reads obtained from Illumina sequencing to the respective viral reference genome. The accession numbers for the reference genomes were NC_001477 (DENV1), NC_001474 (DENV2), NC_000883 (human parvovirus B19), NC_003977 (HBV), and NC_001802 (HIV-1). DENV, dengue virus; B19, human parvovirus B19; HBV, hepatitis B virus; HIV, human immunodeficiency virus.

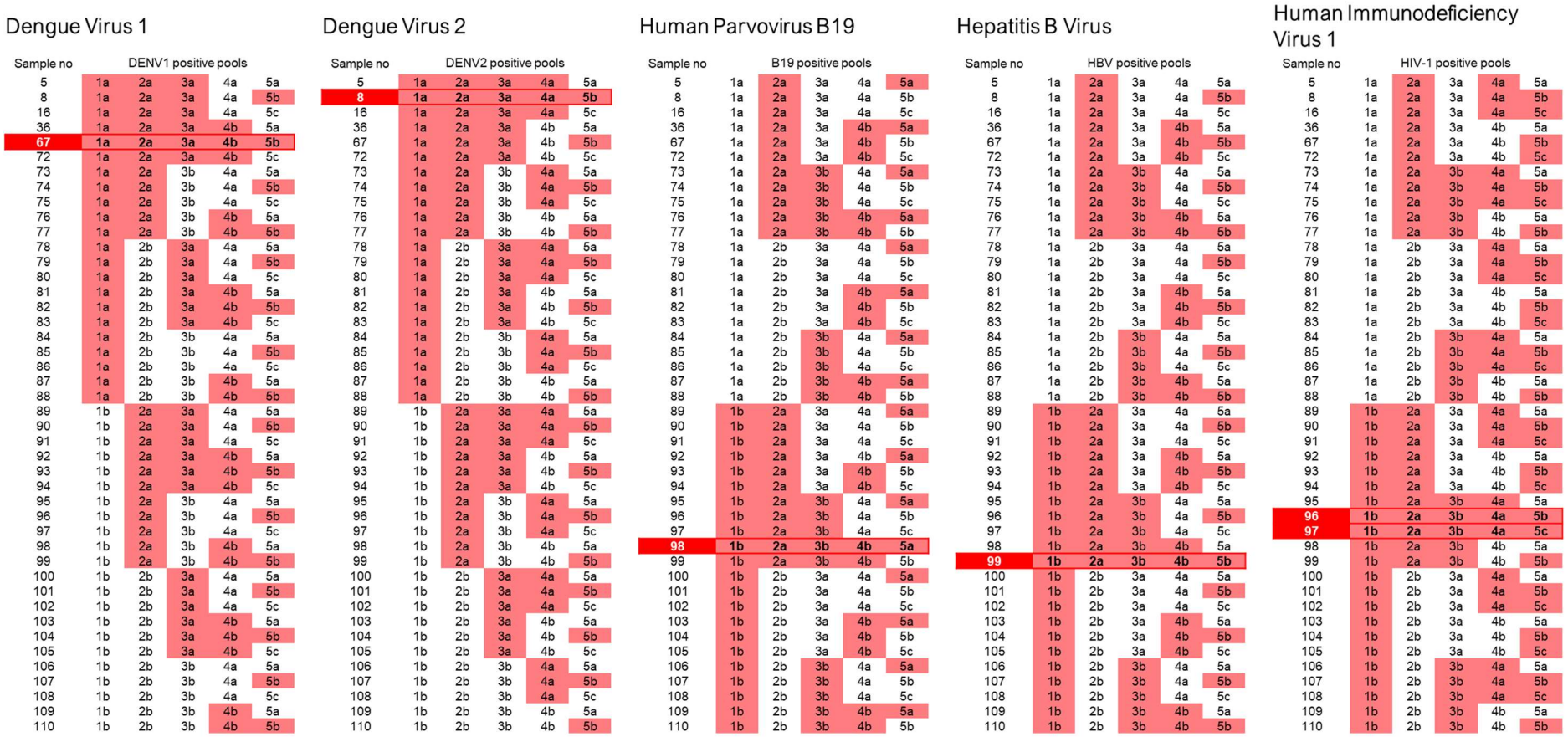


Figure 3.5 Summary of mEGA result and how the sample identity was recovered. When viral reads can be detected from five different pools, into which a sample was pooled (see Table 3.1), then it can be assumed that the viral reads must have originated from that sample. Pools where the viral reads were detected are shaded in red, while sample that were pooled into those pools are outlined in red.

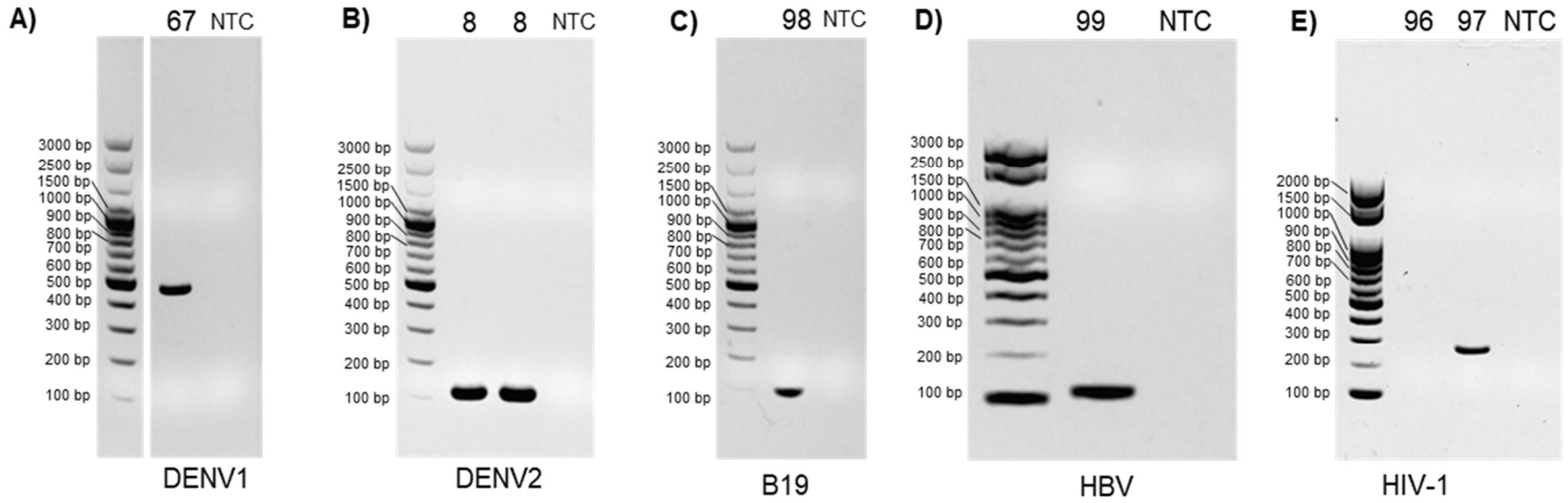


Figure 3.6 Visualization of the amplicons obtained from pathogen specific PCR. For DENV, human parvovirus B19, and HBV, the cDNA synthesized from the extracted RNA were used as the template for the pathogen specific PCR. While for HIV-1, due to limitation of sample amount, diluted cWTA amplicons were used as template for PCR (using a newly designed primer).

DENV, dengue virus; B19, human parvovirus B19; HBV, hepatitis B virus; HIV, human immunodeficiency virus.

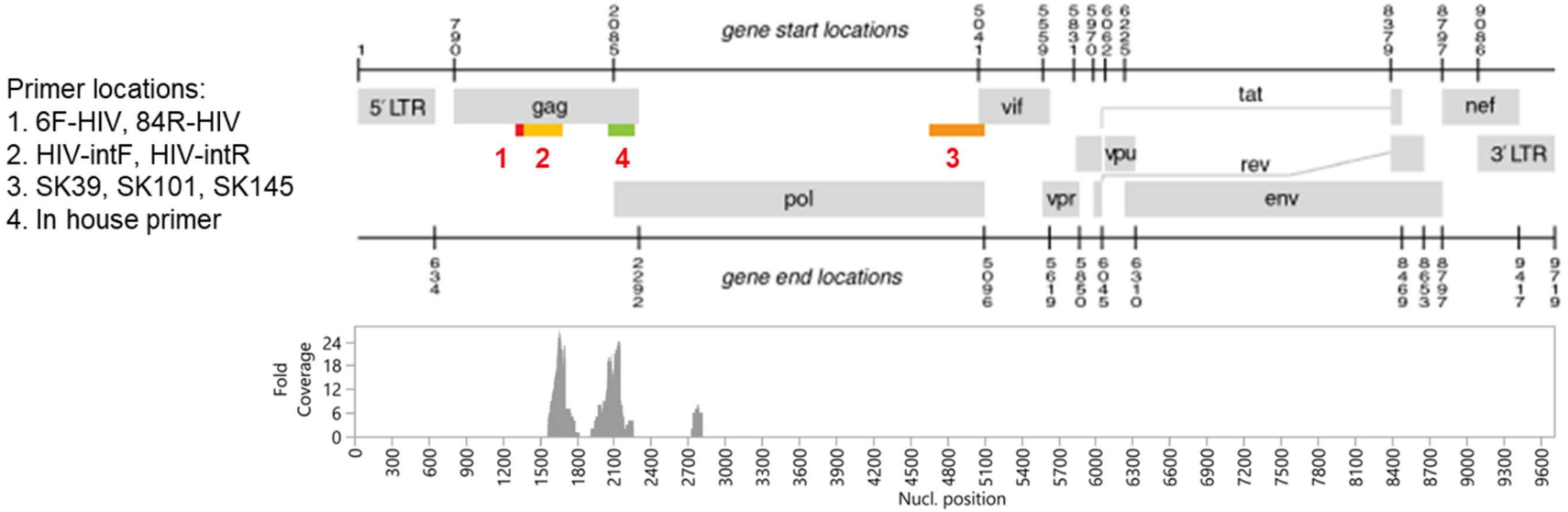


Figure 3.7 Alignment result and location of each primer targeting HIV-1 used in this study. The top part shows HIV-1 genome and its ORFs, the bottom part shows where the reads homologous to HIV-1 (obtained from mEGA) were aligned in the reference genome. The locations of the sequence targeting by each primer are denoted using colored bar and number in the top image. Amplicons were obtained when using the newly designed primer (primer number 4; the target location is marked with green bar) (**Fig. 3.6E**). This amplicon was then subjected to sequencing. Primer sequences are listed in **Table 3.2**.

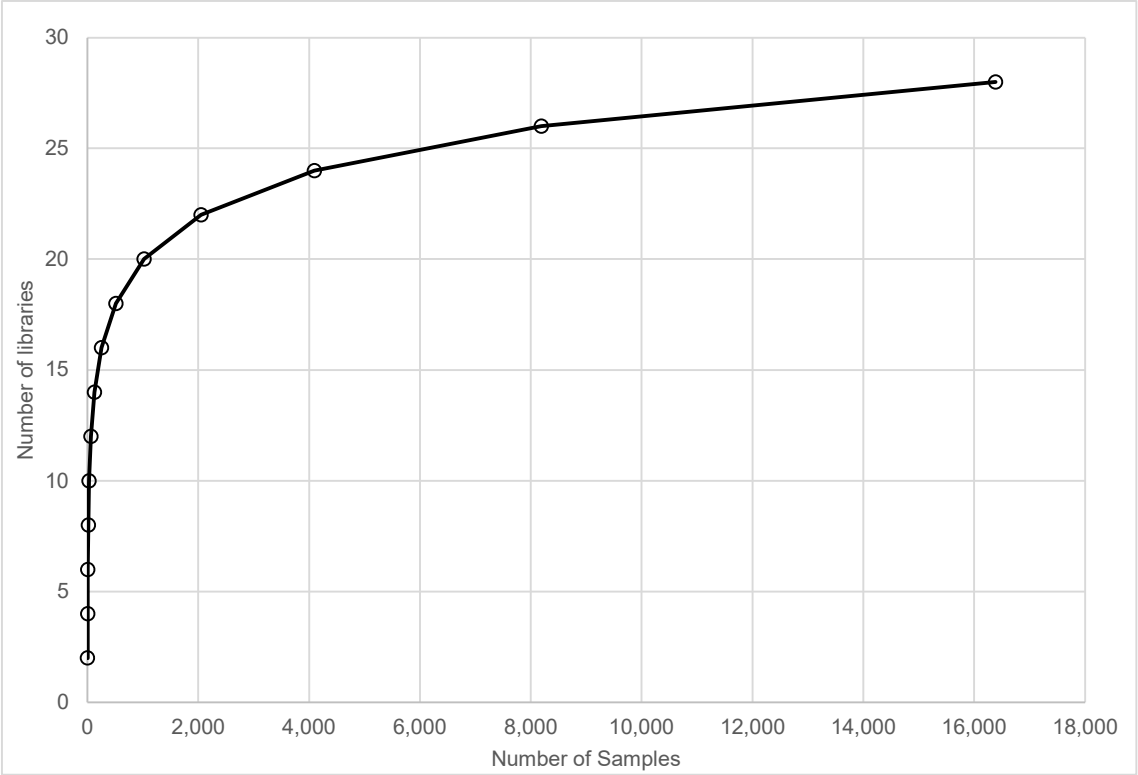


Figure 3.8 Figure depicting the number of libraries (*y* axis) that need to be reconstructed for *n* number of samples (*x* axis) when using mEGA (2n libraries for 2^n samples).

General Conclusion

Determining the etiological pathogen behind a febrile illness can be a challenging task. The numerous possible etiologies in combination with non-specific clinical presentation are the main reasons. The current hypothesis-based approach relies on medical practitioner's expertise to determine the most possible pathogens that need further testing. Unfortunately, though this approach is beneficial for common pathogens, this approach risks missing less common pathogens. As a result, a considerable portion of febrile illness remains to be diagnosed. While it is possible that those cases arise from non-infectious aetiology, it is also possible that the infection is missed from the initial workup. Thus, an unbiased approach would be beneficial for febrile illness. Metagenomic NGS has gained popularity for clinical usage in recent years, though the usage is currently limited as a complementary to the conventional practice or only in certain cases of infections where the diagnosis is difficult. Additionally, previous works shows that mNGS were able to detect more pathogens compared to those of conventional approaches further cementing mNGS as a promising candidate for pathogen diagnostics in the coming years.

The technological development in the NGS field has also boosted the application of mNGS. While the conventional, high-throughput, high-accuracy short-read sequencing (Ilumina) remains to be the standard practice, the portable, long-read sequencer (nanopore) has gained more grounds in resource limited settings. The accuracy boost in sequencing accuracy received by nanopore sequencer recently will make the platform even more reliable for usage in remote laboratories. Though there are limitations on the application of NGS on standard practice, this study explores two main limitations for large scale application of NGS; one being cost and the other being complicated and time-consuming library preparation. Through this study, several alternative ways in creating sequencing libraries were explored, including multiplexing samples to reduce the number of libraries that need to be constructed, while preserving the comprehensiveness of mNGS.

In Chapter One a foundation for comprehensive NGS detection was laid out by establishing a semi-comprehensive platform for universal detection of a virus family. This chapter focuses on flaviviruses, a notoriously known cause of febrile illness. The universal identification of the flavivirus was achieved by amplifying a conserved region in the NS5 gene of flaviviruses. In a sense, this approach is similar to targeted mNGS

which usage lacks in viruses due to their genetic diversity. Indeed, through a targeted sequencing approach, various flavivirus were detected either from spiked samples or from clinical samples. Additionally, sample throughput per sequencing run was improved through a dual indexing system. The portability of nanopore sequencing in combination with sample multiplexing to reduce cost per sequencing run, it is expected that the system can be used for middle to large scale screening or diagnosis in resource-limited settings.

Having established a semi-comprehensive detection system using targeted sequencing approaches, the target was then shifted towards establishing an unbiased viral detection workflow. In Chapter Two, a comprehensive RNAome amplification was developed and validated. To achieve an unbiased detection, the focus was set to RNAome assuming that it can capture the genome of RNA virus and the viral transcripts, which include DNA viruses. In short, the method relies on circularization of the cDNA template prior to amplification with phi29. It is shown that the circularization significantly improves the amplification and that viral sequences were able to be detected using NGS. From two sets of clinical samples, DENV2 and CHIKV were successfully detected from the clinical sample. The portability of nanopore sequence and relatively easy cWTA will be a useful application in peripheral laboratories to gain genetic insight from an infectious agent.

In Chapter Three, mNGS was used to provide an unbiased screening platform for viral pathogens. As library preparation remains costly, time consuming, and complicated, sequencing large sample numbers can be burdensome. To overcome this issue, a pooling method known as group testing algorithm was employed together with cWTA. To achieve high sample throughput screening, the short-read platform was used to generate a larger sequencing depth. In theory, the combinatorial group testing approach can reduce the number of libraries for 2^n samples to 2n libraries with the sample information preserved. This approach was validated on a set of 44 clinical serum samples originated from patients with fever, which were pooled into 11 libraries. Indeed, virus sequences (DENV2, HBV, human parvovirus B19) were detected from those samples. Compared to the approach in Chapter Two, this approach will be more applicable in central laboratories for it requires a deep sequencing.

Overall, these studies have shown that mNGS can be utilized for large-scale and broad-range detection of viral pathogens. These studies have also shown that mNGS can

serve as a powerful tool for a hypothesis-free approach and that both short-read or long-read NGS can be used for high-throughput screening, depending on the capacity of the laboratory.

References

1. Crump JA, Newton PN, Baird SJ, Lubell Y. Febrile illness in adolescents and adults. In: Holmes KK, Bertozzi S, Bloom BR, Jha P, editors. Major infectious diseases. 3rd ed. Washington D.C.: The International Bank for Reconstruction and Development / The World Bank; 2017. p. 365–85.
2. Maze MJ, Bassat Q, Feasey NA, Mandomando I, Musicha P, Crump JA. The epidemiology of febrile illness in sub-Saharan Africa: implications for diagnosis and management. Clin Microbiol Infect. 2018;24(8):808–14.
3. Shrestha P, Roberts T, Homsana A, Myat TO, Crump JA, Lubell Y, Newton PN. Febrile illness in Asia: gaps in epidemiology, diagnosis and management for informing health policy. Clin Microbiol Infect. 2018;24(8):815–26.
4. Elven J, Dahal P, Ashley EA, Thomas NV, Shrestha P, Stepniewska K, Crump JA, Newton PN, Bell D, Reyburn H, Hopkins H, Guérin PJ. Non-malarial febrile illness: A systematic review of published aetiological studies and case reports from Africa, 1980-2015. BMC Med. 2020;18(1):279.
5. Maeki T, Tajima S, Ikeda M, Kato F, Taniguchi S, Nakayama E, Takasaki T, Lim CK, Saijo M. Analysis of cross-reactivity between flaviviruses with sera of patients with Japanese encephalitis showed the importance of neutralization tests for the diagnosis of Japanese encephalitis. J Infect Chemother Off J Japan Soc Chemother. 2019;25(10):786–90.
6. Tan LK, Wong WY, Yang HT, Huber RG, Bond PJ, Ng LC, Maurer-Stroh S, Hapuarachchi HC. Flavivirus cross-reactivity to dengue nonstructural protein 1 antigen detection assays. Diagnostics. 2020;10(1):11.
7. Wangdi K, Kasturiaratchi K, Nery SV, Lau CL, Gray DJ, Clements ACA. Diversity of infectious aetiologies of acute undifferentiated febrile illnesses in south and Southeast Asia: A systematic review. BMC Infect Dis. 2019;19(1):577.
8. Shrestha P, Dahal P, Ogbonnaa-Njoku C, Das D, Stepniewska K, Thomas NV, Hopkins H, Crump JA, Bell D, Newton PN, Ashley EA, Guérin PJ. Non-malarial febrile illness: A systematic review of published aetiological studies and case reports from Southern Asia and South-eastern Asia, 1980–2015. BMC Med. 2020;18(1):299.

9. Robinson ML, Manabe YC. Review article: Reducing uncertainty for acute febrile illness in resource-limited settings: The current diagnostic landscape. Am J Trop Med Hyg. 2017;96(6):1285–95.
10. Haidar G, Singh N. Fever of unknown origin. N Engl J Med. 2022;386(5):463–77.
11. Wright WF, Auwaerter PG. Fever and fever of unknown origin: Review, recent advances, and lingering dogma. Open Forum Infect Dis. 2020;7(5):1–12.
12. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.
13. De Clercq E, Li G. Approved antiviral drugs over the past 50 years. Clin Microbiol Rev. 2016;29(3):695–747.
14. Okeke IN, Ihekweazu C. The importance of molecular diagnostics for infectious diseases in low-resource settings. Nat Rev Microbiol. 2021;19(9):547–8.
15. Deng S, Sun Y, Xia H, Liu Z, Gao L, Yang J, Zhao Y, Huang F, Feng J, Wang L, Huan S, Zhan S. Accuracy of commercial molecular diagnostics for the detection of pulmonary tuberculosis in China: a systematic review. Sci Rep. 2019;9(1):4553.
16. Nuin NA, Tan AF, Lew YL, Piera KA, William T, Rajahram GS, Jelip J, Dony JF, Mohammad R, Cooper DJ, Barber BE, Anstey NM, Chua TH, Grigg MJ. Comparative evaluation of two commercial real-time PCR kits (QuantiFast™ and abTES™) for the detection of *Plasmodium knowlesi* and other *Plasmodium* species in Sabah, Malaysia. Malar J. 2020;19(1):1–11.
17. Santiago GA, Vergne E, Quiles Y, Cosme J, Vazquez J, Medina JF, Medina F, Colón C, Margolis H, Muñoz-Jordán JL. 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(7):e2311.
18. Hur KH, Park K, Lim Y, Jeong YS, Sung H, Kim MN. Evaluation of four commercial kits for SARS-CoV-2 real-time reverse-transcription polymerase chain reaction approved by emergency-use-authorization in Korea. Front Med. 2020;7:521.
19. Sucahya PK. Barriers to covid-19 RT-PCR testing in Indonesia: a health policy perspective. J Indones Heal Policy Adm. 2020;5(2):36–42.

20. Pan American Health Organization. Guidelines for surveillance of Zika virus disease and its complications [Internet]. Washington D.C.: Pan American Health Organization; 2016. Available from: http://iris.paho.org/xmlui/bitstream/handle/123456789/28405/9789275118948_eng.pdf?sequence=1&isAllowed=y
21. Pan American Health Organization. Laboratory diagnosis of yellow fever virus infection [Internet]. 2018. Available from: https://www.paho.org/hq/index.php?option=com_docman&view=download&category_slug=guidelines-5053&alias=46877-laboratory-diagnosis-of-yellow-fever-virus-infection&Itemid=270&lang=en
22. Salim B, Hayashida K, Mossaad E, Nakao R, Yamagishi J, Sugimoto C. Development and validation of direct dry loop mediated isothermal amplification for diagnosis of *Trypanosoma evansi*. Vet Parasitol. 2018;260(August):53–7.
23. Lobato IM, O'Sullivan CK. Recombinase polymerase amplification: Basics, applications and recent advances. Trends Anal Chem. 2018;98:19–35.
24. Huang M, Yang F, Fu J, Xiao P, Tu J, Lu Z. Reaction parameter comparison and optimization of multiple displacement amplification. Anal Methods. 2019;12(1):46–53.
25. Ning L, Wang X, Xu K, Song S, Li Q, Yang X. A novel isothermal method using rolling circle reverse transcription for accurate amplification of small RNA sequences. Biochimie. 2019;163:137–41.
26. Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, Freije CA, Myhrvold C, Bhattacharyya RP, Livny J, Regev A, Koonin E V., Hung DT, Sabeti PC, Collins JJ, Zhang F. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science. 2017;356(6336):438–42.
27. Moonga LC, Hayashida K, Kawai N, Nakao R, Sugimoto C, Namangala B, Yamagishi J. Development of a multiplex loop-mediated isothermal amplification (LAMP) method for simultaneous detection of spotted fever group rickettsiae and malaria parasites by dipstick DNA chromatography. Diagnostics. 2020;10(11):897.
28. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome

analysis. Nat Commun. 2019;10(1):5029.

29. Kounosu A, Murase K, Yoshida A, Maruyama H, Kikuchi T. Improved 18S and 28S rDNA primer sets for NGS-based parasite detection. Sci Rep. 2019;9:15789.
30. Tiew PY, Mac Aogain M, Ali NABM, Thng KX, Goh K, Lau KJX, Chotirmall SH. The mycobiome in health and disease: Emerging concepts, methodologies and challenges. Mycopathologia. 2020;185(2):207–31.
31. Patel P, Landt O, Kaiser M, Faye O, Koppe T, Lass U, Sall AA, Niedrig M. Development of one-step quantitative reverse transcription PCR for the rapid detection of flaviviruses. Virol J. 2013;10:58.
32. Holbrook MG, Anthony SJ, Navarrete-Macias I, Bestebroer T, Munster VJ, van Doremalen N. Pan-coronavirus PCR assay. Viruses. 2021;13:599.
33. Tong S, Chern SWW, Li Y, Pallansch MA, Anderson LJ. Sensitive and broadly reactive reverse transcription-PCR assays to detect novel paramyxoviruses. J Clin Microbiol. 2008;46(8):2652–8.
34. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J, Leggett RM, Livermore DM, O'Grady J. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37(7):783–92.
35. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM, Dodd R, Mulembakani P, Schneider BS, Muyembe-Tamfum JJ, Stramer SL, Chiu CY. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 2015;7(1):99.
36. Chen L, Gao X, Xue W, Yuan S, Liu M, Sun Z. Rapid metagenomic identification of two major swine pathogens with real-time nanopore sequencing. J Virol Methods. 2022;306:114545.
37. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, Federman S, Stryke D, Briggs B, Langelier C, Berger A, Douglas V, Josephson SA, Chow FC, Fulton BD, DeRisi JL, Gelfand JM, Naccache SN, Bender J, Dien Bard J, Murkey J, Carlson M, Vespa PM, Vijayan T, Allyn PR, Campeau S, Humphries RM, Klausner JD, Ganzon CD, Memar F, Ocampo NA, Zimmermann LL, Cohen SH, Polage CR, DeBiasi RL, Haller B, Dallas R, Maron G, Hayden R, Messacar K,

Dominguez SR, Miller S, Chiu CY. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. N Engl J Med. 2019;380(24):2327–40.

38. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, Stryke D, Pham E, Fung B, Bolosky WJ, Ingebrigtsen D, Lorizio W, Paff SM, Leake JA, Pesano R, DeBiasi R, Dominguez S, Chiu CY. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 2019;29(5):831–42.
39. Kufner V, Plate A, Schmutz S, Braun DL, Günthard HF, Capaul R, Zbinden A, Mueller NJ, Trkola A, Huber M. Two years of viral metagenomics in a tertiary diagnostics unit: Evaluation of the first 105 cases. Genes (Basel). 2019;10(9):661.
40. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341–55.
41. Govender KN, Street TL, Sanderson ND, Eyre DW. Metagenomic sequencing as a pathogen-agnostic clinical diagnostic tool for infectious diseases: A systematic review and meta-analysis of diagnostic test accuracy studies. J Clin Microbiol. 2021;59(9):e0291620.
42. Shishido AA, Noe M, Saharia K, Luethy P. Clinical impact of a metagenomic microbial plasma cell-free DNA next-generation sequencing assay on treatment decisions: a single-center retrospective study. BMC Infect Dis. 2022;22(1):1–7.
43. Bagcchi S. Looking back at yellow fever in Angola. Lancet Infect Dis. 2017;17(3):269–70.
44. Waller C, Tiemensma M, Currie BJ, Williams DT, Baird RW, Krause VL. Japanese encephalitis in Australia - a sentinel case. N Engl J Med. 2022;387(7):661–2.
45. Ferguson NM, Cucunubá ZM, Dorigatti I, Nedjati-Gilani GL, Donnelly CA, Basáñez MG, Nouvellet P, Lessler J. Countering the Zika epidemic in Latin America. Science. 2016;353(6297):353–4.
46. Pierson TC, Diamond MS. The continued threat of emerging flaviviruses. Nat Microbiol. 2020;5(6):796–812.
47. Burrell CJ, Howard CR, Murphy FA. Chapter 36 - Flaviviruses. In: Fenner and White's Medical Virology. 5th ed. London: Academic Press; 2017. p. 493–518.
48. Endy TP. Viral febrile illnesses and emerging pathogens. In: Ryan ET, Hill DR, Solomon T, Aronson NE, Endy TP, editors. Hunter's Tropical Medicine and

Emerging Infectious Diseases. 10th ed. Canda: Elsevier; 2020. p. 325–50.

49. Souza NCSE, Felix AC, de Paula AV, Levi JE, Pannuti CS, Romano CM. Evaluation of serological cross-reactivity between yellow fever and other flaviviruses. Int J Infect Dis. 2019;81:4–5.

50. Lustig Y, Sofer D, Bucris ED, Mendelson E. Surveillance and diagnosis of West Nile virus in the face of flavivirus cross-reactivity. Front Microbiol. 2018;9:2421.

51. Vaughn DW, Green S, Kalayanarooj S, Innis BL, Nimmannitya S, Suntayakorn S, Endy TP, Raengsakulrach B, Rothman AL, Ennis FA, Nisalak A. Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. J Infect Dis. 2000;181(1):2–9.

52. Gaikwad S, Sawant SS, Shastri JS. Comparison of nonstructural protein-1 antigen detection by rapid and enzyme-linked immunosorbent assay test and its correlation with polymerase chain reaction for early diagnosis of dengue. J Lab Physicians. 2017;9(3):177–81.

53. Kim YH, Kim TY, Park JS, Park JS, Lee J, Moon J, Chong CK, Junior IN, Ferry FR, Ahn HJ, Bhatt L, Nam HW. Development and clinical evaluation of a rapid diagnostic test for yellow fever non-structural protein 1. Korean J Parasitol. 2019;57(3):283–90.

54. Domingo C, Charrel RN, Schmidt-Chanasit J, Zeller H, Reusken C. Yellow fever in the diagnostics laboratory. Emerg Microbes Infect. 2018;7(1):129.

55. Sánchez-Seco MP, Rosario D, Domingo C, Hernández L, Valdés K, Guzmán MG, Tenorio A. Generic RT-nested-PCR for detection of flaviviruses using degenerated primers and internal control followed by sequencing for specific identification. J Virol Methods. 2005;126(1–2):101–9.

56. Maher-Sturgess SL, Forrester NL, Wayper PJ, Gould EA, Hall RA, Barnard RT, Gibbs MJ. Universal primers that amplify RNA from all three flavivirus subgroups. Virol J. 2008;5:1–10.

57. Ayers M, Adachi D, Johnson G, Andonova M, Drebot M, Tellier R. A single tube RT-PCR assay for the detection of mosquito-borne flaviviruses. J Virol Methods. 2006;135(2):235–9.

58. Johnson N, Wakeley PR, Mansfield KL, McCracken F, Haxton B, Phipps LP, Fooks AR. Assessment of a novel real-time pan-flavivirus RT-polymerase chain

reaction. Vector-Borne Zoonotic Dis. 2010;10(7):665–71.

59. Chao DY, Davis BS, Chang GJJ. Development of multiplex real-time reverse transcriptase PCR assays for detecting eight medically important flaviviruses in mosquitoes. J Clin Microbiol. 2007;45(2):584–9.

60. Vina-Rodriguez A, Sachse K, Ziegler U, Chaintoutis SC, Keller M, Groschup MH, Eiden M. A Novel Pan-Flavivirus Detection and Identification Assay Based on RT-qPCR and Microarray. Biomed Res Int. 2017;2017:4248756.

61. Mongan AE, Tuda JSB, Runtuwene LR. Portable sequencer in the fight against infectious disease. J Hum Genet. 2020;65(1):35–40.

62. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19(7):823–6.

63. Hawkins JA, Jones SKJ, Finkelstein IJ, Press WH. Indel-correcting DNA barcodes for high-throughput sequencing. Proc Natl Acad Sci U S A. 2018;115(27):E6217–26.

64. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–93.

65. Xu Y, Lewandowski K, Lumley S, Pullan S, Vipond R, Carroll M, Foster D, Matthews PC, Peto T, Crook D. Detection of viral pathogens with multiplex nanopore MinION sequencing: Be careful with cross-Talk. Front Microbiol. 2018;9:2225.

66. Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, Shoobridge JD, Graham N, Patel NH, Gillespie RG, Prost S. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. Gigascience. 2019;8(5):giz006.

67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.

68. Naeem R, Rashid M, Pain A. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics.

2013;29(3):391–2.

69. Lanciotti RS, Kerst AJ. Nucleic acid sequence-based amplification assays for rapid detection of West Nile and St. Louis encephalitis viruses. J Clin Microbiol. 2001;39(12):4506–13.

70. Domingo C, Patel P, Yillah J, Weidmann M, Méndez JA, Nakouné ER, Niedrig M. Advanced yellow fever virus genome detection in point-of-care facilities and reference laboratories. J Clin Microbiol. 2012;50(12):4054–60.

71. Lanciotti RS, Kerst AJ, Nasci RS, Godsey MS, Mitchell CJ, Savage HM, Komar N, Panella NA, Allen BC, Volpe KE, Davis BS, Roehrig JT. Rapid detection of west nile virus from human clinical specimens, field-collected mosquitoes, and avian samples by a TaqMan reverse transcriptase-PCR assay. J Clin Microbiol. 2000;38(11):4066–71.

72. Lanciotti RS, Kosoy OL, Laven JJ, Velez JO, Lambert AJ, Johnson AJ, Stanfield SM, Duffy MR. Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. Emerg Infect Dis. 2008;14(8):1232–9.

73. Corman VM, Rasche A, Baronti C, Aldabbagh S, Cadar D, Reusken CBEM, Pas SD, Goorhuis A, Schinkel J, Molenkamp R, Kümmerer BM, Bleicker T, Brünink S, Eschbach-Bludau M, Eis-Hübinger AM, Koopmans MP, Schmidt-Chanasit J, Grobusch MP, De Lamballerie X, Drosten C, Drexler JF. Assay optimization for molecular detection of Zika virus. Bull World Health Organ. 2016;94(12):880–92.

74. Saksida A, Jakopin N, Jelovšek M, Knap N, Fajs L, Lusa L, Lotrič-Furlan S, Bogovič P, Arnež M, Strle F, Avšič-Županc T. Virus RNA load in patients with tick-borne encephalitis, Slovenia. Emerg Infect Dis. 2018;24(7):1315–23.

75. Busch MP, Kleinman SH, Tobler LH, Kamel HT, Norris PJ, Walsh I, Matud JL, Prince HE, Lanciotti RS, Wright DJ, Linnen JM, Caglioti S. Virus and antibody dynamics in acute West Nile virus infection. J Infect Dis. 2008;198(7):984–93.

76. Belsher JL, Gay P, Brinton M, DellaValla J, Ridenour R, Lanciotti R, Perelygin A, Zaki S, Paddock C, Querec T, Zhu T, Pulendran B, Eidex RB, Hayes E. Fatal multiorgan failure due to yellow fever vaccine-associated viscerotropic disease. Vaccine. 2007;25(50):8480–5.

77. Hernandez S, Cardozo F, Myers DR, Rojas A, Waggoner JJ. Simple and Economical Extraction of Viral RNA and Storage at Ambient Temperature.

Microbiol Spectr. 2022;10(3):e00859-22.

78. Pang J, Chia PY, Lye DC, Leo YS. Progress and challenges towards point-of-care diagnostic development for dengue. J Clin Microbiol. 2017;55(12):3339–49.

79. Muller DA, Depelsenaire ACI, Young PR. Clinical and laboratory diagnosis of dengue virus infection. J Infect Dis. 2017;215(Suppl 2):S89–95.

80. Ahmed NH, Broor S. Comparison of NS1 antigen detection ELISA, real time RT-PCR and virus isolation for rapid diagnosis of dengue infection in acute phase. J Vector Borne Dis. 2014;51(3):194–9.

81. Hu H, Jung K, Wang Q, Saif LJ, Vlasova AN. Development of a one-step RT-PCR assay for detection of pancoronaviruses (α-, β-, γ-, and δ-coronaviruses) using newly designed degenerate primers for porcine and avian ‘fecal samples. J Virol Methods. 2018;256:116–22.

82. Annand EJ, Horsburgh BA, Xu K, Reid PA, Poole B, de Kantzow MC, Brown N, Tweedie A, Michie M, Grewar JD, Jackson AE, Singanallur NB, Plain KM, Kim K, Tachedjian M, van der Heide B, Crameri S, Williams DT, Secombe C, Laing ED, Sterling S, Yan L, Jackson L, Jones C, Plowright RK, Peel AJ, Breed AC, Diallo I, Dhand NK, Britton PN, Broder CC, Smith I, Eden JS. Novel Hendra virus variant detected by sentinel surveillance of horses in Australia. Emerg Infect Dis. 2022;28(3):693–704.

83. Kumata R, Ito J, Takahashi K, Suzuki T, Sato K. A tissue level atlas of the healthy human virome. BMC Biol. 2020;18(1):55.

84. Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, Thompson EJ, Meric-Bernstam F, Medeiros LJ, Weinstein JN, Su X. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 sases using RNA-Seq. J Virol. 2013;87(16):8916–26.

85. Froussard P. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. Nucleic Acids Res. 1992;20(11):2900.

86. Djikeng A, Halpin R, Kuzmickas R, DePasse J, Feldblyum J, Sengamalay N, Afonso C, Zhang X, Anderson NG, Ghedin E, Spiro DJ. Viral genome sequencing by random priming methods. BMC Genomics. 2008;9:1–9.

87. Reyes GR, Kim JP. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. Mol Cell Probes. 1991;5(6):473–81.

88. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338(6114):1622–6.

89. Parker JK, Chang TY, Meschke JS. Amplification of viral RNA from drinking water using TransPlex^{TM} whole-transcriptome amplification. J Appl Microbiol. 2011;111:216–23.

90. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun. 2019;10(1):1–14.

91. Wongsurawat T, Jenjaroenpun P, Taylor MK, Lee J, Lavado Tolardo A, Parvathareddy J, Kandel S, Wadley TD, Kaewnapan B, Athipanyasilp N, Skidmore A, Chung D, Chaimayo C, Whitt M, Kantakamalakul W, Sutthent R, Horthongkham N, Ussery DW, Jonsson CB, Nookaew I. Rapid sequencing of multiple RNA viruses in their native form. Front Microbiol. 2019;10:260.

92. Deng X, Achari A, Federman S, Yu G, Somasekar S, Bártolo I, Yagi S, Mbala-Kingebeni P, Kapetshi J, Ahuka-Mundeke S, Muyembe-Tamfum JJ, Ahmed AA, Ganesh V, Tamhankar M, Patterson JL, Ndembi N, Mbanya D, Kaptue L, McArthur C, Muñoz-Medina JE, Gonzalez-Bonilla CR, López S, Arias CF, Arevalo S, Miller S, Stone M, Busch M, Hsieh K, Messenger S, Wadford DA, Rodgers M, Cloherty G, Faria NR, Thézé J, Pybus OG, Neto Z, Morais J, Taveira N, R. Hackett J, Chiu CY. Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. Nat Microbiol. 2020;5(3):443–54.

93. Bhargava V, Ko P, Willems E, Mercola M, Subramaniam S. Quantitative transcriptomics using designed primer-based amplification. Sci Rep. 2013;3(1):1740.

94. Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, Dillon D, Segraves R, Vossbrinck B, González A, Pinkel D, Albertson DG, Costa J, Lizardi PM. Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. Genome Res. 2003;13(2):294–307.

95. Berthet N, Reinhardt AK, Leclercq I, Van Ooyen S, Batéjat C, Dickinson P,

Stamboliyska R, Old IG, Kong KA, Dacheux L, Bourhy H, Kennedy GC, Korfhage C, Cole ST, Manuguerra JC. Phi29 polymerase based random amplification of viral RNA as an alternative to random RT-PCR. BMC Mol Biol. 2008;9:1–7.

96. Parras-Moltó M, Rodríguez-Galet A, Suárez-Rodríguez P, López-Bueno A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. Microbiome. 2018;6(1):119.
97. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
98. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
99. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189–201.
100. Blanco L, Bernad A, Lázaro JM, Martín G, Garmendia C, Salas M. Highly Efficient DNA Synthesis by the Phage ϕ 29 DNA Polymerase. J Biol Chem. 1989;264(15):8935–40.
101. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, Li C, Meyerson M, Sellers WR. Genome coverage and sequence fidelity of ϕ29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res. 2004;32(9):e71–e71.
102. Kim KH, Bae JW. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol. 2011;77(21):7663–8.
103. Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Coleman ML, Breitbart M, Sullivan MB. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. Woyke T, editor. PeerJ. 2016;4:e2777.
104. Mohsen MG, Kool ET. The discovery of rolling circle amplification and rolling circle transcription. Acc Chem Res. 2016;49(11):2540–50.
105. Li XY, Du YC, Zhang YP, Kong DM. Dual functional Phi29 DNA polymerase-

triggered exponential rolling circle amplification for sequence-specific detection of target DNA embedded in long-stranded genomic DNA. Sci Rep. 2017;7(1):6373.

106. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001;11(6):1095–9.
107. Takahashi H, Ohkawachi M, Horio K, Kobori T, Aki T, Matsumura Y, Nakashimada Y, Okamura Y. RNase H-assisted RNA-primed rolling circle amplification for targeted RNA sequence detection. Sci Rep. 2018;8(1):7770.
108. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, Abdel MP, Patel R. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods. 2016;127:141–5.
109. Ruppé E, Lazarevic V, Girard M, Mouton W, Ferry T, Laurent F, Schrenzel J. Clinical metagenomics of bone and joint infections: a proof of concept study. Sci Rep. 2017;7(1):7718.
110. Simner PJ, Miller HB, Breitwieser FP, Monsalve GP, Pardo CA, Salzberg SL, Sears CL, Thomas DL, Eberhart CG, Carrolla KC. Development and optimization of metagenomic next-generation sequencing methods for cerebrospinal fluid diagnostics. J Clin Microbiol. 2018;56(9).
111. Joffroy B, Uca YO, Prešern D, Doye JPK, Schmidt TL. Rolling circle amplification shows a sinusoidal template length-dependent amplification bias. Nucleic Acids Res. 2018;46(2):538–45.
112. Cheng AP, Burnham P, Lee JR, Cheng MP, Suthanthiran M, Dadhania D, De Vlaminck I. A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection. Proc Natl Acad Sci. 2019;116(37):18738–44.
113. Chen P, Li S, Li W, Ren J, Sun F, Liu R, Zhou XJ. Rapid diagnosis and comprehensive bacteria profiling of sepsis based on cell-free DNA. J Transl Med. 2020;18(1):1–10.
114. Wanda L, Ruffin F, Hill-Rorie J, Hollemon D, Seng H, Hong D, Blauwkamp T, Kertesz M, Fowler Jr. V. Direct detection and quantification of bacterial cell-free DNA in patients with bloodstream infection (BSI) using the Karius plasma next generation sequencing (NGS) test. Open Forum Infect Dis. 2017;4:S613–S613.
115. Mannerström B, Paananen RO, Abu-Shahba AG, Moilanen J, Seppänen-

Kaijansinkko R, Kaur S. Extracellular small non-coding RNA contaminants in fetal bovine serum and serum-free media. Sci Rep. 2019;9(1):5538.

116. Duan H, Li X, Mei A, Li P, Liu Y, Li X, Li W, Wang C, Xie S. The diagnostic value of metagenomic next-generation sequencing in infectious diseases. BMC Infect Dis. 2021;21(1):1–13.
117. Gu W, Deng X, Lee M, Sucu YD, Arevalo S, Stryke D, Federman S, Gopez A, Reyes K, Zorn K, Sample H, Yu G, Ishpuniani G, Briggs B, Chow ED, Berger A, Wilson MR, Wang C, Hsu E, Miller S, DeRisi JL, Chiu CY. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. Nat Med. 2021;27(1):115–24.
118. Dorfman R. The Detection of Defective Members of Large Populations. Ann Math Stat. 1943;14(4):436–40.
119. McDermott JH, Stoddard D, Woolf PJ, Ellingford JM, Gokhale D, Taylor A, Demain LAM, Newman WG, Black G. A nonadaptive combinatorial group testing strategy to facilitate health care worker screening during the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) outbreak. J Mol Diagnostics. 2021;23(5):532–40.
120. Mutesa L, Ndishimye P, Butera Y, Souopgui J, Uwineza A, Rutayisire R, Ndoricimpaye EL, Musoni E, Rujeni N, Nyatanyi T, Ntagwabira E, Semakula M, Musanabaganwa C, Nyamwasa D, Ndashimye M, Ujeneza E, Mwikarago IE, Muvunyi CM, Mazarati JB, Nsanzimana S, Turok N, Ndifon W. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. Nature. 2020;589:276–80.
121. Singh L, Anyaneji UJ, Ndifon W, Turok N, Mattison SA, Lessells R, Sinayskiy I, San EJ, Tegally H, Barnett S, Lorimer T, Petruccione F, de Oliveira T. Implementation of an efficient SARS-CoV-2 specimen pooling strategy for high throughput diagnostic testing. Sci Rep. 2021;11(1):17793.
122. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10–2.
123. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
124. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee JJ, Fett B, Bushman FD, Bittinger

K. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome. 2019;7(1):46.

125. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

126. Alawi M, Burkhardt L, Indenbirken D, Reumann K, Christopeit M, Kröger N, Lütgehetmann M, Aepfelbacher M, Fischer N, Grundhoff A. DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. Sci Rep. 2019;9(1):16841.

127. Yasumizu Y, Hara A, Sakaguchi S, Ohkura N. VIRTUS: a pipeline for comprehensive virus analysis from conventional RNA-seq data. Bioinformatics. 2021;37(10):1465–7.

128. Servant A, Laperche S, Lallemand F, Marinho V, De Saint Maur G, Meritet JF, Garbarg-Chenon A. Genetic diversity within human erythroviruses: identification of three genotypes. J Virol. 2002;76(18):9124–34.

129. Sosa-Jurado F, Meléndez-Mena D, Rosas-Murrieta NH, Guzmán-Flores B, Mendoza-Torres MA, Barcenas-Villalobos R, Márquez-Domínguez L, Cortés-Hernández P, Reyes-Leyva J, Vallejo-Ruiz V, Santos-López G. Effectiveness of PCR primers for the detection of occult hepatitis B virus infection in Mexican patients. PLoS One. 2018;13(10):e0205356.

130. Lanciotti RS, Calisher CH, Gubler DJ, Chang GJ, Vorndam A V. Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction. J Clin Microbiol. 1992;30(3):545–51.

131. Wiegand A, Maldarelli F. Single-copy quantification of HIV-1 in clinical samples. Methods Mol Biol. 2014;1087:251–60.

132. Wang JH, Cheng L, Wang CH, Ling WS, Wang SW, Lee GB. An integrated chip capable of performing sample pretreatment and nucleic acid amplification for HIV-1 detection. Biosens Bioelectron. 2013;41:484–91.

133. Whetsell AJ, Drew JB, Milman G, Hoff R, Dragon EA, Adler K, Hui J, Otto P, Gupta P, Farzadegan H, Wolinsky SM. Comparison of three nonradioisotopic polymerase chain reaction-based methods for detection of human immunodeficiency virus type 1. J Clin Microbiol. 1992;30(4):845–53.

134. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, Light M,

Lai K, Jarosz M, McNeill MS, Ducar MD, Meyerson M, Thorner AR. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics. 2018;19(1):1–10.

135. Reteng P, Nguyen Thuy L, Tran Thi Minh T, Mares-Guia MAM de M, Torres MC, de Filippis AMB, Orba Y, Kobayashi S, Hayashida K, Sawa H, Hall WW, Nguyen Thi LA, Yamagishi J. A targeted approach with nanopore sequencing for the universal detection and identification of flaviviruses. Sci Rep. 2021;11(1):19031.
136. Slavov SN, Rodrigues ES, Sauvage V, Caro V, Diefenbach CF, Zimmermann AM, Covas DT, Laperche S, Kashima S. Parvovirus B19 seroprevalence, viral load, and genotype characterization in volunteer blood donors from southern Brazil. J Med Virol. 2019;91(7):1224–31.
137. Burrell CJ, Howard CR, Murphy FA. Chapter 23 - Retroviruses. In: Fenner and White's Medical Virology. 5th ed. London: Academic Press; 2017. p. 317–44.
138. Désiré N, Cerutti L, Le Hingrat Q, Perrier M, Emler S, Calvez V, Descamps D, Marcelin AG, Hué S, Visseaux B. Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. Retrovirology. 2018;15:80.

Acknowledgement

I would like to express my deepest appreciation for Dr. Junya Yamagishi for giving me an opportunity to pursue a study under his supervision. I also would like to thank him for his guidance during my doctoral course, the knowledge he has shared with me, the encouragement, and the constructive criticisms that had helped me to become a better researcher. I would also like to extend my deepest gratitude to Dr. Kyoko Hayashida and Dr. Tatsuki Sugi, whose relentless support and insightful suggestions have assisted me numerous times during my course and during the writing of this dissertation. The inspiration and motivation that I have received from them are invaluable, and for that I am thankful. I would also extend my appreciation to Dr. Yuki Eshita, whose kindness and helpful advice have supported me through my course. I would also extend my appreciation to Ms. Naoko Kawai for her valuable contribution in this dissertation and for her assistance in laboratory experiments. I would also like to mention all the collaborators who were contributed to the success of this study, especially to Prof. William W. Hall for his helpful advice, and to Dr. Lanh Anh Nguyen Thi and Dr. Anna Maria Bispo de Filippis for their continuous support.

I would also like to express my deepest appreciation to Dr. Yutaka Suzuki, for his unparalleled support and for allowing me to take part in a research project in his laboratory. The knowledge and experience I have obtained from the research are invaluable. In the same laboratory, I would also like to extend the gratitude to Dr. Yukie Kashima for her guidance and support while I was conducting research there and to all members of Suzuki Laboratory. I would also like to send special thanks to Dr. John-Sebastian Eden, for precious experiences that I obtained during my visit and for awakening my interest in the study of viral evolution.

I would also like to thank the committee that evaluated this dissertation, Dr. Hirofumi Sawa and Dr. Nariaki Nonaka. I am thankful for the thorough perusal of this manuscript and for the insightful comments and suggestions that have improved the content of this dissertation.

I also would like to thank the lab members of Collaboration Education and the supportive environments that they created. I could not imagine going through this course

without you. Many thanks to family and friends, for their encouragement and support, especially to those who I pestered for grammar-related questions.

Finally, I am thankful to the Japanese government, especially the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for given me a chance to study at Hokkaido University through a scholarship. I would also like to extend my thanks to WISE Program (Doctoral Program for World-leading Innovative & Smart Education) for supporting my research activity and overseas internship.

Japanese Abstract (和文要旨)

感染による発熱性疾患は、臨床症状が非特異的である場合が多く、また、病因となりうる病原体も多岐にわたることから、その原因特定は容易ではない。従って多くの場合、医療従事者の専門知識や経験に依存して、数種類の最も蓋然性の高い病原体に限定した検査が行われているのが実情である。この方法は感染者数の多い主要な病原体に対しては有効だが、そうではない病原体を見逃してしまう危険性がある。それらは、個々の患者数が少ないとしても、それを引き起こす病原体が多数存在することから、全体として、かなりの部分が診断されないままになっていると考えられている。それらを特定するアプローチとして、近年、次世代シーケンサー（NGS）を応用した metagenomic NGS（mNGS）の臨床応用が進みつつある。これまでの研究では、mNGS による病原体ゲノムの網羅的検出により、従来の手法に比べてより多くの病原体を検出できることが示されている。しかしながら、現状では、コストや手技の煩雑さから、従来の診療を補完するもの、あるいは診断が困難な特定の感染症に限定して用いられており、標準的な診断方法として利用されるには至っていない。一方、NGS 分野における技術開発が近年加速しており、mNGS 実用化へ向けた追い風となっている。Ilumina 社が提供するプラットフォームは、高精度を特徴とするショートリードをハイスループットに出力するのに対し、Oxford nanopore technologies 社のプラットフォームは、精度には劣るものの、デバイスが安価かつ持ち運び可能で、ロングリードを出力できることもあり、開発途上国等のリソースの限られた環境下でより多くの支持を得ている。さらに、弱点とされる精度の向上も報告されている。mNGS による病原体診断の社会実装には、解析に係る高額な費用と複雑で時間を要するライブラリ調製が妨げになっている。そこで本研究では、これらの問題の解決に資する代替方法の開発を行った。

第 1 章では、熱性疾患の原因としてよく知られているフラビウイルスをモデルに、同一の属や科に属するウイルスを網羅的に検出するための方法を開発した。具体的にはフラビウイルス属の NS5 遺伝子の保存領域を PCR により増幅し、nanopore シーケンサーで解析することにより、同属ウイルスの網羅的な同定を実現した。さらに、PCR プライマーにインデックス配列を付与することで、マルチプレックス解析によるシーケンスランあたりのコスト削減が可能なこと

を実証した。本技術は、リソースが限られた環境での中・大規模なスクリーニングや診断に利用することが期待される。

第 2 章では、RNAome 増幅法の開発と検証を行った。RNAome は、RNA ウイルスのゲノムと、DNA ウイルスを含むウイルスの転写産物を含んでおり、理論上、mNGS を用いることで全てのウイルスを検出可能である。しかしながら、微量な RNA を検出するためには、NGS 解析の前に網羅的な増幅を行う必要があるため、簡便かつ偏りの少ない RNA 増幅法の開発が求められていた。そこで本研究では、cDNA を一本鎖 DNA 特異的な ligase により環状化した後、phi29 で増幅する circular whole-transcriptome amplification (cWTA)の開発に成功した。本法により実験的に増幅したウイルスの検出が可能なことも示され、臨床サンプルから DENV2 と CHIKV を検出することにも成功した。実施が比較的容易な cWTA と可搬的な nanopore 型 NGS を組み合わせることで、ゲノム配列に基づく網羅的な感染症診断が、辺境地域でも実施可能になることが期待される。

第 3 章では、mNGS による網羅的な病原体検出をより効率的に行うためのライブラリー調整法の開発を行った。本研究では、グループテストアルゴリズムと呼ばれるプーリング方法を採用し、mNGS と組み合わせた。本法を用いることで、2^n 個のサンプルに対するライブラリの数を、サンプル情報を保持したまま 2n 個のライブラリに削減し、そこに含まれる病原体を網羅的に検出することができる。実証試験として、発熱患者から採取した 44 の臨床血清試料を 11 のライブラリにプールし検証を行った。その結果，これらのサンプルから、事前情報を用いることなくウイルス配列（DENV2，HBV，ヒトパルボウイルス B19）を検出することに成功した。第 2 章でのアプローチと比較すると、本アプローチは大規模検体の一括解析に適し、より高出力の NGS プラットフォームを必要とするため、中央研究所でより適用しやすいと思われる。

以上、中央研究所だけでなく、小規模研究室や辺境地域等、状況に応じた mNGS による網羅的病原体解析を実施可能とする種々のライブラリー調整方法の開発に成功し、mNGS による網羅的病原体解析の社会実装を妨げるコストに関する課題を解決する道筋が示された。