ENRICHMENT PROTOCOL FOR METAGENOMICS SEQUENCING OF RESPIRATORY SYNCYTIAL VIRUS (RSV) USING THE OXFORD NANOPORE TECHNOLOGY (ONT) MINION DEVICE

JACQUELINE WAHURA WAWERU

MASTER OF SCIENCE

(Molecular Biology and Bioinformatics)

JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY

2022

Enrichment Protocol for Metagenomics Sequencing of Respiratory Syncytial Virus (RSV) Using the Oxford Nanopore Technology (ONT) Minion Device

Jacqueline Wahura Waweru

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Molecular Biology and Bioinformatics of the Jomo Kenyatta University of Agriculture and Technology

2022

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university

Signature……………………………………………..Date………………………

Jacqueline Wahura Waweru

This thesis has been submitted for examination with my approval as the university supervisor

Signature……………………………………………..Date………………………

Prof. Johnson Kinyua, PhD
JKUAT, Kenya

Signature……………………………………………..Date………………………

Dr. George Githinji, PhD
KEMRI Wellcome Trust Research Programme, Kenya

Signature……………………………………………..Date………………………

Prof. James Nokes, PhD
KEMRI Wellcome Trust Research Programme, Kenya

DEDICATION

I dedicate this work to my parents Mr. Samuel Timothy Maingi and Mrs. Alice Waweru, and my brother Allan Peter Maingi. Thank you very much for being such amazing support systems to me. For your love, tender care, financial and moral support, I am very grateful and am continually praying that God blesses you all immensely. I love you all so much!

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisors: Dr. George Githinji, Prof James Nokes and Prof. Johnson Kinyua for their support in conceptualizing this work. Special thanks to Dr. Charles Sande and Dr. Caleb Kibet who together with my supervisors helped in contributing ideologies that helped in shaping this project right from designing the lab experiments, trouble shooting, data analysis and thesis and manuscript writing. I am also very grateful for the mentorship you all accorded me; through this work, I have been well mentored and trained in the prime subjects in molecular biology and bioinformatics. I would like to acknowledge the Initiative to Develop African Research Leaders (IDeAL) for funding this work and KEMRI Wellcome Trust Research Programme for hosting me while I undertook my research. I thank Evelyn Kamau, Zaydah deLaurent, Elijah Gicheru and Martin Mutunga for supporting me during the lab experiments. I am greatly indebted. Special thanks to my current supervisor Dr. Juan Camilo Paredes for the support and even pushing me to prioritize on finalizing this work amidst all my other allocated duties, I am very grateful! Lastly and most importantly, my sincere gratitude to God for giving me good health, the strength, grace, perseverance, and a relentless spirit to complete this project.

"This research was funded in whole, or in part, by the Wellcome Trust [DEL-15-003] and the UK Foreign, Commonwealth & Development Office, with support from the Developing Excellence in Leadership, Training and Science in Africa (DELTAS Africa) programme. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission"

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

ATP	Adenosine Triphosphate
CDD	Coupled charged device
cDNA	complementary deoxyribonucleic acid
Ct	cycle threshold
DNA	Deoxyribonucleic acid
HMM	Hidden Markov Models
HS	high sensitivity
NEB	New England Biolabs
NGS	Next Generation Sequencing
nt	not treated
ONT	Oxford Nanopore Technology
PCR	Polymerase Chain Reaction
QC	Quality check
RNA	Ribonucleic acid
RNN	Recurrent Neural Networks
RSV	Respiratory Syncytial Virus
rRNA	ribosomal Ribonucleic acid
SISPA	Sequence independent single primer amplification

SMRT	single molecule real time
t	treated
TGS	Third generation sequencing
WGS	whole genome sequencing
ZMW	Zero mode waveguide

ABSTRACT

Respiratory syncytial virus (RSV) is a leading cause of lower respiratory tract infections among children under the age of five. Nonetheless, no effective vaccine against the virus exists, but there have been efforts to guide vaccine development by seeking to understand the transmission and evolutionary patterns of the virus using targeted partial and whole genome sequencing studies. This project aimed to develop a viral metagenomics enrichment protocol for RSV whole genome sequencing (WGS) using the ONT MinION device as a step towards unbiased sequencing of respiratory viruses. However, nasopharyngeal samples contain higher quantities of host and bacterial nucleic material relative to viral genetic material. This presents challenges during virus metagenomics sequencing which underpins agnostic sequencing protocols. Two unbiased viral enrichment protocols were therefore assessed using a similar set of samples. Protocol 1 involved physical pre-treatment of samples by centrifugal processing before RNA extraction, while protocol 2 entailed direct RNA extraction from samples without a pre-treatment step. From the centrifugal processing protocol, a pellet and supernatant were obtained after centrifugation at 8000rpm for five minutes, while concentrates and filtrates were obtained after centrifugal filtration of the supernatants at 14000 rpm for one hour using 3kD centrifugal filters, with the main fraction of interest being the concentrate. Concentrates from protocol 1 were divided into two fractions; one was DNase treated while the other was not, followed by RNA extraction. Extracted RNA from protocol 2 on the other hand, was divided into two fractions; one was DNase treated whilst the second was not. RNA from both protocols was converted to cDNA, amplified using the sequence independent single primer amplification (SISPA) approach, libraries prepared, and sequencing done. DNase-treated fractions from both protocols recorded significantly reduced host and bacterial contamination unlike the untreated fractions (in each protocol $p<0.01$). Additionally, DNase treatment after RNA extraction (Protocol 2) ($p<0.01$) enhanced host and bacterial read reduction compared to when done before (Protocol 1). However, neither protocol yielded whole RSV genomes. Sequenced reads mapped to parts of the nucleoprotein (N gene) and polymerase complex (L gene) from Protocol 1 and 2, respectively. The incomplete genome segments from both protocols were attributed to amplification biases introduced when part of the tag, in tagged Endoh primers anneals to the genome, due to the shortness (6 bases) of the random sequence. This study recommends that the random sequence in the tagged Endoh primers be extended in length to around 9-12 bases instead of six since the length of the random sequence in tagged random primers is an important factor for the success of SISPA.

CHAPTER ONE

INTRODUCTION

1.1: Background information

Respiratory syncytial virus (RSV) accounts for approximately 33.1 million cases and an estimated 3.2 million hospitalizations globally per year, among children under the age of five years (Shi *et al*., 2017). Out of all the global incidences, roughly 48,000-74,500 in-hospital child deaths annually are attributed to RSV infections (Shi *et al*., 2017). In 72 low- and middle-income countries, RSV is estimated to account for 20.8 million annual incidences, while in Kenya the number of annual reported cases among children are roughly 85000 (Li *et al*., 2020; Nokes *et al*., 2008). The virus also causes high morbidity and mortality among immunocompromised individuals and the elderly (Englund *et al*., 1991; Lee *et al*., 2013). The genome of the virus is a 15.2 kb non-segmented, negative-sense, single-stranded ribonucleic acid (RNA) virus (Mufson *et al*., 1985) belonging to the order *mononegavirales, pneumoviridae* family and the *Orthopneumovirus* genus (Rima *et al*., 2017).

This study endeavored to develop a viral enrichment protocol for the unbiased whole genome sequencing (WGS) of RSV using the Oxford Nanopore Technology (ONT) MinION device. The ONT MinION device is a pocket sized real-time single molecule sequencing device, advantages which were exploited during the Ebola and Zika virus epidemics in West Africa and Brazil respectively (Quick *et al.,* 2016, 2017). In addition, the ONT MinION sequencing device is a cheaper sequencing platform as compared to all the other available sequencing platforms. The lower sequencing cost have seen its vast usage in resource constrained countries during the SARS-CoV2 viral pandemic thus enhancing a representation of viral genomes in low- and middle-income countries(Bugembe *et al*., 2021; Githinji *et al*., 2021). Further, the ONT MinION sequencing device is capable of sequencing long reads which have demonstrated capacity to improve genome assembly especially in repetitive genomic regions (Kchouk *et al*., 2017).

The ONT MinION sequencing device has previously been used successfully in unravelling the genomic epidemiology, transmission and evolutionary patterns of a number of viruses including Zika virus (Quick *et al*., 2017), dengue virus (Mongan *et al*., 2019), influenza virus (Eckert *et al*., 2016), Ebola virus (Quick *et al*., 2016), SARS – CoV2 virus (Li *et al*., 2020) as well as plant viruses such as the cassava mosaic virus (Boykin *et al*., 2018), further making it possible to inform on policies. However, targeted virus enrichment approaches using either polymerase chain reaction (PCR) amplicon-based approaches or hybridization captures were used in all the studies. Though the targeted enrichment techniques sensitively result in the detection of the targeted viruses, prior knowledge of the pathogen present in a sample is often required to guide in designing the primers or hybridization probes (Hall *et al*., 2014), which is challenging in the case of studying novel viruses. In addition, targeted virus enrichment usually biases the relative abundance of the targeted virus relative to the others likely to be present in the samples, making the genotyping process of the rest difficult and especially in the case of co-infections within a sample (Hall *et al*., 2014). The challenges with targeted sequencing underscore the need for unbiased sequencing protocols.

Random priming also known as Sequence Independent Single Primer Amplification (SISPA) was used in a couple of previous viral metagenomics studies by Greninger *et al*. (2015) where the samples contained Ebola, Chikungunya and Hepatitis C viruses, detection of arboviruses in mosquitoes (Batovska *et al*., 2017), detection of Chikungunya and Dengue viruses (Kafetzopoulou *et al*., 2018) among others as a viral enrichment strategy. SISPA, first developed by Reyes & Kim (1991), entails the use of oligonucleotides consisting of random nucleotides on the 3’ end and a 5’ defined tag sequence that is mainly used for subsequent amplification (Chrzastek *et al*., 2017). Random priming has been a promising strategy in viral metagenomics because unlike bacterial and fungal communities which have 16S and ITS (internal transcriber spacer) conserved markers for bacterial and fungal community amplification, respectively, viral communities lack conserved markers across or even within viral families (Conceição-Neto *et al*., 2015).

While SISPA has been successfully used in enriching viral metagenomic datasets, other studies have shown that the approach can be challenging given the highly abundant host and bacterial material as compared to viral material. In perspective, SISPA has been shown to result in the preferential amplification of the overly abundant host and bacterial nucleic materials eventually resulting in their preferential sequencing and an under representation of viral material in the metagenomic dataset (Graf *et al*., 2016), creating the need to explore methods for reducing them. Alternatives to reduce host and bacterial reads often incorporate physical and enzymatic virus enrichment steps including centrifugal filtration and DNase treatment (Conceição-Neto *et al*., 2015; Goya *et al*., 2018; Thurber *et al*., 2009). SISPA, centrifugal filtration and DNase treatment were employed in several previous studies (Chrzastek *et al*., 2017; Goya *et al*., 2018; Yifei *et al*., 2018) and were deemed effective in enhancing viral read representation and reducing bacterial and host contamination. Here, the effectiveness of centrifugal filtration (Thurber *et al*., 2009), DNase-treatment (Peret *et al.*, 1998) and SISPA (Nguyen *et al*., 2016) were tested as virus enrichment methods for the unbiased RSV sequencing using the ONT MinION device. Confirmed RSV positive samples (using multiplex PCR) which had also been sequenced on the Illumina MiSeq platform were used in this study.

1.2: Statement of the problem

The burden of respiratory viruses is still high and has been grouped among the five most common causes of morbidity and mortality globally(Umuhoza *et al*., 2021). In addition, these infectious viruses have been implicated with high costs of management and treatment (Umuhoza *et al*., 2021). The quest to understand the evolutionary and transmission patterns of these viruses to better inform on better management policies for example takes a lot of time and has an implicated extra cost in the case where targeted sequencing is performed on samples with co-infections, creating the need for the unbiased whole genome viral sequencing protocols. However, a typical nasopharyngeal sample contains low quantities of viral nucleic material relative to bacterial and host nucleic material (Graf *et al*., 2016). Host and bacterial contaminants are a challenge for agnostic sequencing of nasopharyngeal samples because they are preferentially sequenced due to their higher relative abundance. To enrich for the viruses, which lack conserved and universal markers across or even within families, SISPA has been employed. Nonetheless, during the random priming process, lots of bacterial and host reads are also amplified and after sequencing, their reads are over-represented as compared to those of viruses in the metagenomics dataset (Graf *et al*., 2016), creating the need for better viral enrichment protocols. While metagenomics sequencing can be achieved on short read sequencing platforms such as Illumina, a challenge arises during the assembly process due to the presence of sequencing gaps and especially in the case novel viruses (Kchouk *et al*., 2017).

1.3: Justification of the study

This study was underscored as a step towards field epidemiological studies of respiratory viruses. An unbiased RSV WGS protocol was therefore an initial step towards metagenomics sequencing of respiratory viruses which would enable the detection of other viral communities present in nasopharyngeal samples during co-infections, enhancing their genotyping after only one sequencing experiment. The development and optimization of working unbiased respiratory sequencing protocols underpinned the enrichment of the lowly abundant virus nucleic material and enhanced reduction of host and bacterial reads in the metagenomics dataset (Conceição-Neto *et al*., 2015; Goya *et al*., 2018; Thurber *et al*., 2009). In addition, adoption of the ONT sequencing technology in viral metagenomics would enhance viral read assembly due to the long reads generated during sequencing process thus reducing the gaps characteristic to short read sequencing especially in the case of novel viruses. (Kchouk *et al*., 2017) Further, ONT sequencing technology would ensure rapid and improved genomic epidemiology of novel and endemic viruses at affordable sequencing costs.

1.4: Research questions

1. Does physical pre-treatment of RSV nasopharyngeal samples by centrifugal processing prior to RNA extraction and amplification using SISPA enhance viral read representation in the final metagenomic dataset?

2. How can respiratory viruses from clinical samples be sequenced using an agnostic approach?
3. Which available bioinformatics tools are suitable for the analysis of viral metagenomic datasets?

1.5: Objectives

1.5.1: General objective

To develop a viral metagenomics enrichment protocol for RSV WGS using the ONT MinION device.

1.5.2: Specific objectives

1. To optimize the RNA extraction method by including centrifugal processing and amplification using SISPA.
2. To develop an agnostic approach for sequencing respiratory viruses from clinical samples.
3. To analyze the generated data using the available bioinformatics tools and software.

CHAPTER TWO

LITERATURE REVIEW

2.1: Respiratory syncytial virus (RSV)

2.1.1: The discovery and pathology of RSV

The virus was first isolated in the upper airway secretions of Chimpanzee monkeys with bronchiolitis symptoms in 1956 and was named Chimpanzee Coryza agent (Chanock & Finberg, 1957). Another infection by the same infectious agent was later recorded among children with lower respiratory tract infections. The virus was then renamed to RSV based on the massive syncytial cells that resulted during an infection (Chanock & Finberg, 1957).

RSV infections are restricted to the ciliated epithelial cells in the respiratory system (Zhang *et al.*, 2002). Infections by the virus are preceded by inoculation of RSV aerosol particles to the nose or eyes resulting in viral replication in the nasopharynx (Hall, 1982). The incubation period for the virus is roughly 2-8 days followed by the spread of the virus to the lower respiratory tract 1-3 days later, which causes bronchitis and pneumonia (Piedimonte & Perez, 2014). Bronchitis is characterized by airway resistance, air trapping and wheezing, while hypoxia is most common in the case of pneumonia (Turner *et al*., 2014).

2.1.2: The RSV genome

Under an electron microscope, RSV particles appear as both spherical and filamentous. The RSV particle is medium sized roughly 120-300nm. Within the virus particle is a nucleocapsid that consists of the encapsulated RSV genome that has a helical symmetry (Bächi & Howe, 1973). The RSV genome encodes 10 genes that are translated into 11 proteins attributed to the two open reading frame M2-gene (Collins *et al*., 1990). The encoded proteins include the non-structural protein 1 (NS1), non-structural protein 2 (NS2), nucleoprotein (N), phosphoprotein (P), matrix

(M), short hydrophobic (SH), glycoprotein (G), fusion (F), M2-1, M2-2, and the large polymerase complex (L) (Cane *et al*., 1994).

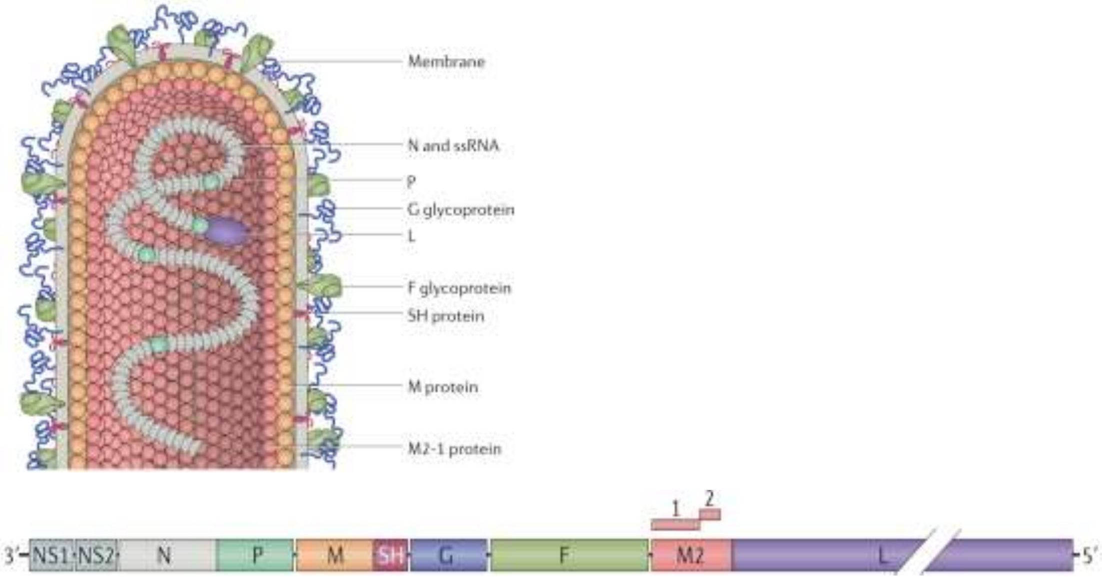


Figure 2. 1: **Diagrammatic representation of RSV genome organization of its ten genes.** Source (Battles & McLellan, 2019)

The N protein encapsulates the RSV viral genome. The non-structural 1 and 2 proteins are important in altering the host immune responses through an antagonistic action on interferon production and signaling. Further, they function in determining the host range and virulence of the virus, and in inhibiting apoptosis of the virus thus facilitating viral growth and replication regulation (Fearns & Collins, 1999) . The L protein is the main polymerase sub-unit that contains the main catalytic domains while the P protein is the main cofactor in RNA synthesis (Collins *et al*., 1984). The M2-1 and M2-2 are involved in transcription and modulation of balance between transcription and replication. The M protein is fundamental in viral assembly (Fearns & Collins, 1999).

The G, F and SH proteins are the main surface proteins. G is highly glycosylated and plays a significant role in the cell attachment process while the F protein aids in mediating fusion of the cell to viral membranes leading to the formation of syncytia (Simoes & Groothuis, 2002). The G and F proteins invoke the main immune responses as they are the main neutralization antigens in the virus (Simoes & Groothuis, 2002). Lastly, the SH protein plays a role in replication and immunity (Collins *et al*., 1990).

2.1.3: Previous RSV sequencing studies

Sequencing as an integral part of viral genomics aids in genotyping, elucidation of the evolutionary patterns and characterization of the transmission patterns of viruses at high resolutions. Previously, targeted partial and whole genome sequencing studies were conducted on the capillary and Illumina sequencing platforms respectively. Partial RSV sequencing mainly targeted the G gene of the virus (Johnson *et al*., 1987; Peret *et al*., 1998)because of its high variability and its role in eliciting immunological responses as was identified from immunological assays. From partial genome sequencing, it has been possible to characterize the evolutionary and transmission patterns of the virus (Johnson *et al*., 1987; Peret *et al*., 1998). However, partial genomes lack in resolution to comprehensively characterize the pathobiology of the virus, the evolutionary and transmission patterns of RSV, and in determining the conserved and variable regions of the genome (Rebuffo-Scheer *et al*., 2011).

To add resolution while characterizing the evolutionary and transmission patters of RSV, WGS using the Illumina platform was done (Agoti *et al.,* 2015, 2017; Otieno *et al.,* 2018). Since Illumina is a short read platform and most clinical samples contain low viral titers, sequencing the entire RSV genome was hard prompting the development of six and fourteen amplicon-based strategies (Agoti *et al*., 2015; Otieno *et al*., 2018) to amplify the entire genome. Overlapping primers that target six or fourteen different regions spanning the entire genome were used during whole genome amplification in six and fourteen different reactions. After amplification, small aliquots were drawn from each reaction into one tube that were then used in library preparation and sequencing (Agoti *et al*., 2015; Otieno *et al*., 2018). However, amplicon-based RSV sequencing resulted in biased relative abundance of RSV reads when compared to those of other viral and bacterial communities causing lower respiratory tract infections. The insufficient nucleotide information from other infectious microbial communities in the obtained reads made genotyping of the other co-infecting viruses difficult ((Hall *et al.*, 2014). Agnostic sequencing the endeavor of this project was an alternative to avoiding biases that come with targeted amplicon-based sequencing. Previous respiratory viruses' agnostic sequencing was attained through various strategies such as random priming protocols (Goya *et al.*, 2018). Physical methods, nuclease treatment and random amplification were previously used in enriching for viruses while altering the ratio of viruses versus bacteria and host genetic material in favor of viruses (Conceição-Neto *et al.*, 2015; Hall *et al.*, 2014; Rosseel *et al.*, 2013, 2015).

2.2: Available sequencing technologies

2.2.1: First generation sequencing technology

Watson and Crick solved the 3D structure of DNA in 1953 using the crystallographic analysis produced by Rosalind Franklin and Maurice Wilkins. Two decades later, nucleic acid sequencing was invented by Fred Sanger, and in 1977, this sequencing method became the first commercially successful sequencing technology (Sanger, 1988).

2.2.1.1: Sanger sequencing technology

Sanger sequencing is based on the di-deoxy nucleotide chain termination principle (Sanger, 1988) where the template is divided into four aliquots and all the substrates required for sequence synthesis added. Radiolabeled chain terminating di-deoxy-nucleotides, initially used during this sequencing process (ddATP, ddGTP, ddCTP, ddTTP), are added to each reaction to terminate the DNA strand synthesis. The terminated lengths of the template obtained from the platform are then separated using gels (Sanger, 1988). Later, Sanger sequencing technology was improved to the currently used capillary sequencing where fluorescently labelled chain terminating di-deoxy-nucleotides are used as chain terminators. The sequenced templates obtained from the platform are separated using capillary electrophoresis (Liu *et al.*, 2012; Swerdlow & Gesteland, 1990). Capillary sequencing results in highly accurate reads but it is labor intensive, low throughput, expensive and time consuming when used in whole genome sequencing (WGS) (Ari *et al*., 2016)**.**

2.2.2: Second generation sequencing technologies

The search for fast, low-cost, high throughput and accurate sequencing technologies resulted in the development of the so called “second generation sequencing” technologies. Second generation sequencing technologies are divided into two; 1) those based on the principle of sequencing by ligation such as the Applied Biosystems Sequencing by oligonucleotide ligation and detection (SOLID-ABI) and 2) sequencing by synthesis such as 454-pyrosequencing, Illumina sequencing and Ion torrent sequencing (Tipu & Shabbir, 2015).

2.2.2.1: SOLID-ABI sequencing technology

Briefly, sequencing using SOLID-ABI introduced in 2006, involves the use of a DNA ligase in ligating four fluorescently labelled bases competing to bind to an oligonucleotide complementing the template being sequenced. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length; however, it is not able to produce good read lengths and depth making assembly challenging (Liu *et al*., 2012).

2.2.2.3: 454-pyrosequencing technology

The 454-pyrosequencing introduced in 2005 uses the luminescent method in measuring pyrophosphate synthesis. During the sequencing by synthesis process, a pyrophosphate is released which is converted to ATP by the enzyme ATP sulfurase. The generated ATP then acts as a substrate of the enzyme luciferase resulting in the production of light proportional to the amount of pyrophosphate released (Tedersoo *et al*., 2010). Using this approach however, made it hard finding the number of nucleotides present in a row at a given position. Although the intensity of light released corresponded to the length of the homopolymer, noisy signals were generated if there were four or more identical nucleotides (Froehlich & Heindl, 2010).

2.2.2.4: Ion Torrent sequencing technology

Ion torrent sequencing introduced in 2010 on the other hand entails the detection of the differences in pH caused by the release of hydrogen ions when the sequencing process is on-going. If no complementary base is available to add to the growing template being sequenced, no ionic changes occur, demonstrating the high specificity of the process. The disadvantage of this sequencing process is that it is hard interpreting homopolymer sequences due to loss of signal as multiple matching dNTPs incorporate (Rothberg *et al*., 2011).

2.2.2.5: Illumina sequencing technology

Illumina sequencing developed in 2006 has a more stable chemistry making it the most widely used second generation sequencing platform. When sequencing, adapter ligated DNA sequences undergo bridge amplification to synthesize several identical copies of each sequence (clusters) (Bentley & Balasubramaniam, 2008). The clusters are then denatured and sequenced by synthesis using a DNA polymerase and fluorescently labelled bases. The inactive 3’ hydroxyl group in the nucleotides ensures that only one nucleotide is incorporated and after incorporation, specific light is emitted from laser excitation, which is then detected using a coupled charged device (CDD) camera (Tipu & Shabbir, 2015). Computer programs then help in translating the signals into nucleotide sequence. This sequencing platform however results in short read sequences which result in challenges when assembling genomes especially in regions that have repetitive sequences (Hengyun, & Francesca, 2016).

2.2.3: Third generation sequencing technologies

Third generation sequencing (TGS) platforms such as PacBio and Oxford Nanopore Technology (ONT) are capable of sequencing long reads and have demonstrated capacity to improve genome assembly especially in repetitive genomic regions (Kchouk *et al*., 2017). These technologies produce reads in real time implying shorter sequencing time compared to Illumina sequencing which takes at least 48 hours, with the ability to access the sequenced data only after the sequencing process is over. Real time sequencing improves the turnaround time especially when applied during diagnosis (Tyler *et al.*, 2018). The downside of TGS is that they have relatively higher error rates as compared to the other sequencing technologies. Error correction tools have nevertheless been developed. Additionally, since the advantages and disadvantages of SGS and TGS complement each other, hybrid sequence analysis strategies have been devised (Magi *et al*., 2017).

2.2.3.1: PacBio sequencing technology

PacBio sequencing details the use of single molecule real time (SMRT) cells which contain zero mode waveguides (ZMWs) described as wells of ten nanometers in diameter macro fabricated in a metal film (Travers *et al*., 2010). ZMWs facilitate light passage through openings of diameters less than the wavelength of the light preventing propagation of the light. The smaller diameters of the ZMW aid in decreasing the intensity of light along the wells hence illuminating the bottom of the wells (Travers *et al*., 2010). The ZMWs contain a DNA polymerase and nucleotide template attached to their bottom. During sequencing, the DNA polymerase incorporates fluorescently labelled nucleotides to the growing strand being synthesized (Eid *et al*., 2009; Rhoads & Au, 2015; Travers *et al*., 2010). When a nucleotide is incorporated, a luminous signal is released that is recorded by the sensors. Detection of the labelled nucleotides makes it possible to determine the DNA sequence (Rhoads & Au, 2015). Though this platform results in long reads and sequences in real time, it is quite expensive requiring considerable capital investment, limiting the accessibility of the technology in the general laboratory set ups. In addition, the technology is quite cumbersome in that the sequencing machine is large. This makes its applications during field epidemiology studies hard.

2.2.3.2: ONT MinION device sequencing technology

The ONT MinION device, the main platform in this study, remains as the current exciting sequencing device due to its small size and portability (Laver *et al*., 2015), providing potential applications during real time field epidemiology. ONT sequencing is based on the principle of nucleic materials passing through a nanopore which is a small hole. There are two types of nanopores namely biological nanopores and solid state nanopores (Haque *et al*., 2013; Liu *et al*., 2012). Biological nanopores are made when proteins are embedded in biological membranes. The proteins are then modified to change the internal structure and attach molecular motors for improved sequencing. Examples of proteins that have been used in creating these nanopores include α-haemolysin secreted by *Staphylococcus aureus* (Haque *et al*., 2013) and recently Curlin sigma S dependent growth subunit G secreted by *E. coli* (Goya *et al*., 2018). The solid state nanopores on the other hand are synthetically manufactured holes in solid materials such as graphene (Haque *et al*., 2013; Magi *et al*., 2017). Solid state nanopores are still under study and have not been used in sequencing.

Figure 2. 2: A pictorial representation of the ONT MinION device.

Source: Researcher 2022

During sequencing, the MinION device is plugged directly into a laptop with 1 terabyte storage, 16GB RAM and 4 core CPU without the need for additional infrastructure (Loman *et al*., 2015). A standard ONT MinION device flow cell has an array consisting of 512 channels each consisting of four nanopores, but only one nanopore per sensor is active at a time (Stoddart *et al*., 2009). To control the sequencing run process, the MinKNOW software is used. The software helps in assigning the experiment run parameters, data acquisition and in attaining feedback on how the experiment is progressing (Magi *et al*., 2017; Rang *et al*., 2018). Sequencing using the device entails unzipping of double stranded DNA and passing it through a chemically or biologically engineered pore by the action of a motor protein attached to the pore. This results in the generation of an ionic current caused by differences in the moving nucleotides occupying the pore (Stoddart *et al*., 2009).

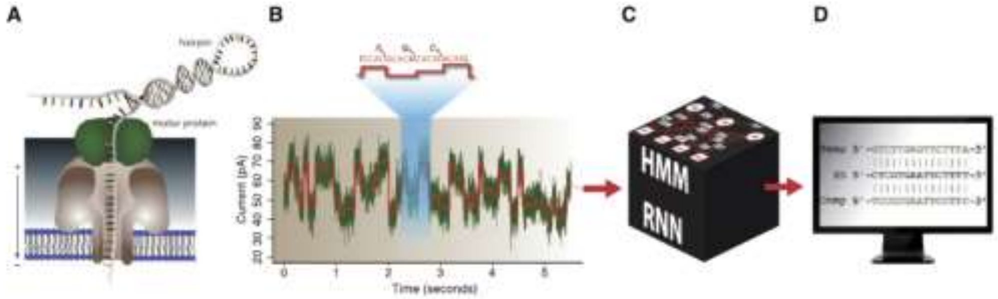


Figure 2. 3: A schematic representation of the nanopore sequencing approach

Source: Magi *et al*. (2017). (A) indicates the transition of a nucleotide through the pore, (B) indicates a squiggle chart of nanopore reads, (C) indicates the various ways of base calling and (D) is a pictorial representation of base called reads on a computer.

2.2.3.3: ONT reads bioinformatics analysis

Majority of the reads from the ONT MinION platform are long and error prone when compared to those from the short read and accurate Illumina platform (Loman & Quinlan, 2014). The disparity in the characteristics of the reads generated from the two platforms means that the tools that are used in analyzing Illumina reads cannot be adopted for analyzing MinION data. Specialized tools for analyzing long and error prone MinION data have been developed and others are still under development given that the technology is dynamically developing and therefore bioinformatics analysis algorithms need to keep up with the development pace (Plesivkova *et al*., 2019).

The first step in the analysis of MinION data entails base calling the FAST5 files to FASTQ. The accuracy of base calling is dependent on the flow cell chemistry used which has an influence on the signal to noise ratio, and how well the signal can be interpreted by a software used for base calling. Various base calling tools have been developed including those based on the Hidden Markov Models (HMM) such as Metrichor and the most recently used are based on recurrent neural networks (RNN) such as Guppy (Leggett & Clark, 2017; Magi *et al*., 2017; Plesivkova *et al*., 2019). MinKNOW software also has a real time base caller that can be used during the live base calling process when the run is still in progress. Alternatively, the option can be switched off and base calling done later either using MinKNOW or an alternative base caller. Base called reads are in FASTQ format which must be demultiplexed to separate the multiplexed reads and remove the adapters and barcodes (Oxford Nanopore Community).

To reduce the read per error rate, reads are polished using error correction tools such as nanopolish and genome assembly done based on either reference-based mapping or de novo assembly. Reference based assembly is used when one has a reference genome while De novo assembly is done when there are no reference genomes. Algorithmic approaches to assemble genomes fall into two classes, the hybrid, and the non-hybrid approaches (Magi *et al*., 2017). Hybrid algorithms correct the nanopore reads using the Illumina short read sequences because Illumina is more accurate (1% error rate) (Koren *et al*., 2012). The non-hybrid approach, on the other hand, involves self-correcting long reads by exploiting overlaps in the high coverage data (Chin *et al*., 2013). Moreover, the drastic improvements in nanopore chemistry by the production of flow cells with better chemistries and better base calling algorithms will lower the overall error rates in nanopore sequencing.

2.3: Sequence independent single primer amplification (SISPA)

SISPA technique is different from the original PCR in that it allows non-selective amplification of nucleic acids using a single primer (Reyes & Kim, 1991). The technique is particularly important when the nucleic components present in a sample are unknown and present in limited amounts (Reyes & Kim, 1991). This technique has found broad application in viral metagenomics as it allows untargeted genome amplification of viral pathogens allowing sensitive detection of diverse pathogens present in a sample on NGS platforms. Besides, the technique is also applicable in characterizing and genotyping the pathogens present in samples (Chrzastek *et al*., 2017).

The SISPA technique briefly entails the use of a common sequence (provided to all nucleic acid molecules present in a population) ligated to specially designed random primers through blunt end ligation. The asymmetric ends of the primer ensure directional ligation onto the target population. This permits one strand of the primer to be used as a primer in the subsequent enzymatic amplification of heterogeneous target DNA population (Reyes & Kim, 1991). In 1992, Froussard employed this technique in the amplification of MS2 phage RNA. During this study, Froussard (1992) used random hexamers tagged with a universal primer during the process of cDNA synthesis. After second strand synthesis Froussard (1992) then used a complement to the universal primer in amplifying the formed cDNA. Froussard (1992) then visualized the PCR products on agarose gel and reported the presence of a smear inferring that different fragment sizes from diverse populations had been amplified.

The use of random hexamers tagged with a universal primer had their challenges especially when applied in human viral metagenomics. The random hexamers on the 3’ end of the primers were unspecific resulting in their binding to the most abundant nucleic acid population in a sample which mainly is human nucleic acids. To reduce the host population, an alternative entailed the selection of poly-A RNA by oligo-dT column followed by oligo-dT priming to eliminate the influence of ribosomal RNAs on cDNA synthesis. Another alternative entailed the use of specific primers to the virus of your interest (Endoh *et al*., 2005). However, prior knowledge of the virus genome was required which would not have functioned well in characterizing diverse circulating virus strains whose complete genomes were yet to be sequenced and additionally, they required constant updating because viruses mutate at a high rate (Nguyen *et al*., 2016).

To alleviate the challenges above, Endoh *et al*., (2005) developed a whole new set of 96 primers inefficient for priming ribosomal RNA but effective at priming most of the genomes of an RNA virus. The primers were termed as the Endoh primers, and they were developed based on representational difference analysis. The primers existed as hexamers and were documented to be important in the identification of new viral agents as they did not require prior knowledge of the agent's class. The Endoh hexamers were thus used in generating the non-ribosomal cDNA (Endoh *et al.*, 2005).

Since the Endoh hexamers could not be used during the amplification process, Nguyen *et al*., (2016) modified them further by replacing the random motif of the 3' end of the FR26RV-N with those of 96 hexanucleotides designed by Endoh. A separate set of 96 separate primers consisting of an FR20 RV sequence at the 5' end was created. The universal FR20 primer allowed the amplification of cDNA generated using Endoh primers.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Ethical approval

The samples used in this study were obtained from Kilifi County Hospital from children under the age of five following informed written consent from each child's guardian or parent. Ethical approval to support the study had already been granted by the KEMRI Scientific and Ethics Review Unit (SERU), and the protocol number was SERU-3178.

3.2: Study population

This study was part of a larger study which aims to understand the transmission patterns of respiratory viruses in Kilifi through the continuous long-term surveillance of respiratory virus pathogens among the pediatric admissions to Kilifi County Hospital. Cumulatively, thirty-two nasopharyngeal swabs collected between January 2012 and December 2015 from children presenting with clinical symptoms of severe pneumonia at Kilifi County hospital were selected for this study. Given previous collections, all samples were stored in a biobank at -80°C.

The convenience sampling approach was used in this study when selecting the samples, with the inclusion criteria being samples with low RSV cycle threshold scores (Ct < 24) and which had been previously sequenced using Illumina MiSeq by targeted amplification yielding full genomes (Agoti *et al.*, 2015; Otieno *et al.*, 2018). Samples with low cycle threshold were most suited for this study because their viral load was high and thus had increased chances of yielding more viral reads spanning the entire genome. In addition, samples whose genomes had been sequenced on the Illumina Miseq platform were selected for sequencing with ONT MinION device for ease in comparison of results from the two platforms. The sample size for the study was determined based on the resources available for the project, and the available number of samples that met the inclusion criteria.

3.3: Sample processing

Each of the processes for the two protocols was set out in the flow diagram depicted in **Figure 3.1**.

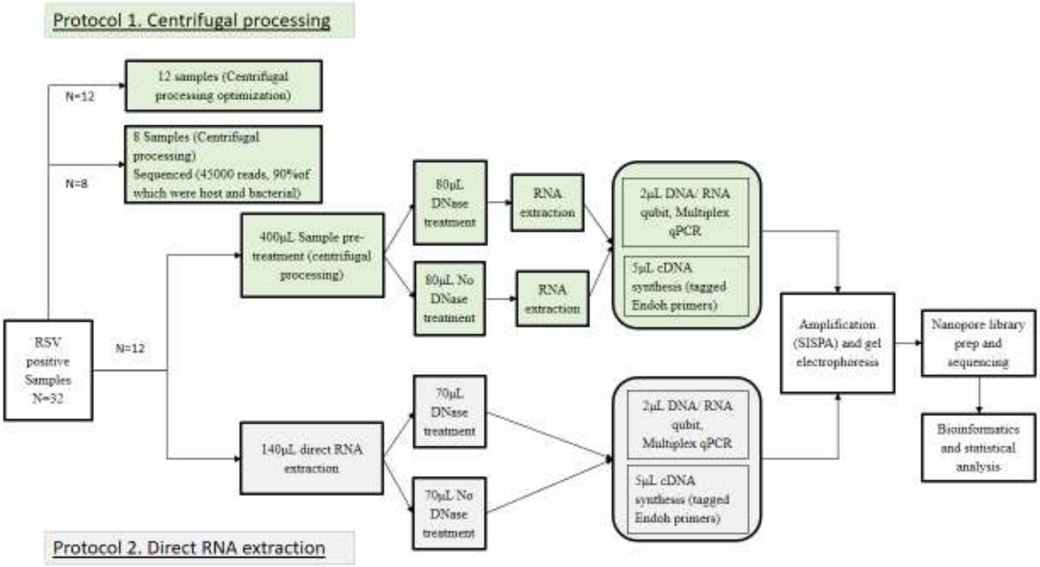

Protocol 1. Centrifugal processing
N=12
12 samples (Centrifugal processing optimization)
N=8
8 Samples (Centrifugal processing) Sequenced (45000 reads, 90%of which were host and bacterial)
RSV positive Samples N=32
N=12
400µL Sample pre-treatment (centrifugal processing)
80µL DNase treatment
80µL No DNase treatment
RNA extraction
RNA extraction
2µL DNA/ RNA qubit, Multiplex qPCR
5µL cDNA synthesis (tagged Endoh primers)
Amplification (SISPA) and gel electrophoresis
Nanopore library prep and sequencing
Bioinformatics and statistical analysis
140µL direct RNA extraction
70µL DNase treatment
70µL No DNase treatment
2µL DNA/ RNA qubit, Multiplex qPCR
5µL cDNA synthesis (tagged Endoh primers)
Protocol 2. Direct RNA extraction

Figure 3. 1: A flow chart representing the experimental setups tested in this study. Twelve (12) samples were selected and divided into two fractions: the first underwent centrifugal processing (Protocol 1) and the entire workflow is represented by the upper part of the flow chart while the second underwent direct RNA extraction (Protocol 2), and the entire workflow of the fractions treated using the approach is represented on the lower part of the flow chart. The arrows indicate the process from one step to the next.

3.3.1: Protocol 1: Centrifugal Processing Approach

3.3.1.1: Optimization

A set of 12 out of the 32 cumulative RSV positive samples were used at first to optimize the centrifugal pre-processing protocol. From the twelve samples, ten were taken through centrifugal processing, while two underwent normal RNA extraction without performing centrifugation prior to the extraction and used as controls. The centrifugal processing protocol involved centrifugation of 400µl of sample at 8000 rpm for 5 minutes, which resulted in a pellet constituted mainly of the dense host and bacterial content. A volume of 350µL supernatant was collected and transferred to the 3kD Scientific Centrifugal Filter (Thermo Fischer), for centrifugal filtration for one hour at 14,000rpm to recover, separately, concentrates and filtrates. RNA was then extracted from each of the three sample fractions (concentrate, filtrate, and pellet from centrifugal processing) obtained from the 10 samples using the QIAmp viral RNA kit (QIAGEN) according to the manufacturer's instructions. Briefly, samples were lysed under high denaturing conditions to inactivate RNases while isolation of intact viral RNA was enhanced by adjusted buffering conditions to provide optimum binding of the RNA to QIAMP membrane, contaminants washed away, and high-quality RNA precipitated and eluted in RNase free buffer ready for subsequent steps. The effectiveness of the pre-processing steps was assessed by performing RNA HS (high sensitivity) Qubit, multiplex RT-PCR and immunofluorescence antibody test (IFAT). Quantity and quality of the RNA extracts were determined using Qubit RNA HS assay. qRT-PCR assays for RSV and adenovirus(Hammitt *et al.*, 2011) were used to quantify the viral load in the three sample fractions. The differences in the viral Ct scores between the concentrate and the pellet were used to infer the extent of host contamination. IFAT using RSV DFA kit Light Diagnostics™ was further used to inform the extent of host contamination between the pellet and the concentrate by observing the intensity of red and green fluorescence (red fluorescence represents host cells while green represents viruses) in the two fractions. Bacterial contamination in the concentrate was measured using conventional PCR using the 314F-CCTACGGGNGGCWGCAG and 805R-GACTACHVGGGTATCTAATCC primers which amplify the V3 and V4 region of the 16S ribosomal RNA (rRNA). Amplified PCR products were visualized in a 2% gel.

3.3.1.2: Sequence independent single primer amplification (SISPA)

First-strand cDNA was synthesized in a 20µl reaction from 5µl viral RNA extracts using the Superscript III reverse transcriptase kit (Thermo Fischer Scientific) and the FR26-Endoh primers (Nguyen *et al.*, 2016). Briefly, the FR26-Endoh primers; created by replacing the 3' end of the FR26RV-N with those of 96 non ribosomal hexanucleotides designed by Endoh **(Appendix I)** (Endoh et al., 2005), were added to the template along with nuclease free water and deoxynucleoside triphosphate (dNTPs), and the mix heated at 65°C for 5 minutes. After heating, the mix was chilled on ice for one minute and the first strand synthesis mix constituted of first strand buffer, DTT, superscript III and RNaseOUT added, followed by incubation at 55°C for 40 minutes and inactivation of the reaction at 70°C for 15 minutes. Klenow fragment 3'-5' exo (NEB) was used to convert the first strand to second-strand cDNA. 20µl of the first-strand cDNA mixture was incubated at 37°C for 90 minutes in the presence of dNTPs, nuclease-free water, and 10X buffer. The RSV and Adenovirus qRT-PCR assay confirmed cDNA formation by excluding the RT step.

The FR20RV primer (GCCGGAGCTCTGCAGATATC) and Q5 PCR kit (NEB) were then used to amplify 13µl of the double-stranded cDNA as follows: 98°C for 30s, 38 cycles of 98°C for 10s, 55°C for 30s and 72°C for 1 min. This PCR was run twice to complete any partial amplicons resulting from used up dNTPs and primers in the first amplification. PCR products were visualized in a 1% gel and purified using Agencourt AMPure XP beads (Beckman Coulter).

3.3.1.3: Nanopore library preparation and sequencing

The library was prepared by multiplexing up to 24 end repaired samples using the Oxford Nanopore 1D ligation sequencing kit (SQL-LSK 109). In brief, all the samples were barcoded using the native barcoding kits (EXP-NBD 104 and EXP-NBD 114), and the enzyme T4 ligase. After barcoding, the samples were washed using the AMPure XP beads (Beckman Coulter) and eluted using an elution buffer. 1ul of barcoded samples were used in quantification using the Invitrogen Qubit double stranded DNA HS kit (Thermo Fisher) and the obtained concentrations used during the normalization process. Normalization was done to ensure that equimolar amounts of the barcoded samples were picked when pooling the samples together. To the pooled barcoded samples, adapter ligation was done using Adapter mix II (AMII), Nebnext Ultra II ligation master mix and Nebnext ligation enhancer. After a 10min incubation to enhance the adapter ligation process, a clean-up using the AMPure XP beads and short fragment buffer (SFB) in place of ethanol was done.

The adapter ligated samples were eluted using 15µl elution buffer, 2µl of which was used during quantification using qubit. A library mix containing 12µl of the DNA, 25.5 of the loading beads and 37.5µl of the sequencing buffer was prepared and loaded on a QC-ed R9.4.1 flow cell (FLO-MIN106) and sequencing performed using MinKNOW software (version 19) for 12 hours.

3.3.1.4: Sequencing

All the sample volumes used during the centrifugal processing optimizations were depleted and to assess the effectiveness of the approach for sequencing, a set of eight additional RSV positive samples were selected, and taken through the centrifugal processing approach, RNA extraction, cDNA synthesis, SISPA, library preparation and sequencing as described in the previous steps. From sequencing the eight samples, insufficient reads were obtained (45000 reads), 90% of which were host and bacterial, which prompted the modification of the protocol to include a DNase treatment step on the concentrates after the centrifugal processing step. Since the sample volumes for the eight samples also got depleted, 12 additional samples were further selected. Out of the 12 samples, 11 were taken through this protocol since two samples had insufficient volumes to allow them to undergo the two protocols in test. 400µL of each of the 11 samples was picked and taken through centrifugal processing and the resulting concentrate divided into two equal fractions: the first was DNase treated while the second was not, followed by RNA extraction.

3.4: Protocol 2: Direct RNA extraction approach

140µL of the remaining volumes from each of the remaining 11 samples out of the 12 were picked and taken through direct RNA extraction protocol. Direct RNA extraction protocol involved extracting RNA from the samples without a prior physical or enzymatic enrichment step using QIAmp viral RNA kit (QIAGEN) according to the manufacturer’s instructions described previously. The resulting RNA was divided into two equal fractions, the first was DNase treated while the second was not. Next, screening for RSV positivity was done as described in the previous sections using qRT-PCR assays for RSV and adenovirus (Hammitt *et al.*, 2011; Venter *et al.*, 2011). Next, untargeted amplification using SISPA was performed, followed by nanopore library preparation and sequencing as previously described in the previous sections.

3.5: Bioinformatics analysis

The reads generated from both protocols were taken through bioinformatics analysis using open-source tools other than for the Guppy base-calling software (version 3.1.5, ONT technologies). The output FAST5 files were base called and de-multiplexed using Guppy version 3.1.5 and then quality checked (QC) using PycoQC (version 2.5.0.23) (Leger & Leonardi, 2019) after which taxonomic classification using Kraken2 (version 2.0.9beta) (Wood *et al.*, 2019) was done. All the reads that passed QC (Phred score >7) test were then mapped to the corresponding 12 RSV references generated from Illumina using Minimap2 (version 2.17) (Li, 2018) and the resulting SAM files converted to a BAM file, sorted, and indexed using SAMtools (version 1.7) (Li *et al.*, 2009). Sorted BAM files were visualized using Integrated Genomics Viewer (IGV) (version 11.0.1) (Thorvaldsdóttir *et al.*, 2013) to determine the regions they mapped to in the genome. Next, the regions to which the Endoh primers matched were located using Seqkit locate (version 0.13.2) (Shen *et al.*, 2016), against a centroid genome generated from the consensus Illumina reads using Vsearch cluster (version 2.15.0) (Rognes *et al.*, 2016). All statistical analyses were conducted in R version 3.6 (R Core Team, 2019).

CHAPTER FOUR

RESULTS

4.1: Protocol 1: Centrifugal processing approach optimization

4.1.1: Optimization

The nucleic acid content in the concentrate and filtrate were undetectable compared to that of the pellet, after comparing the RNA Qubit scores from the concentrate, filtrate and pellet, **(Figure 4.1A)**. The filtrate was RSV negative suggesting little or no virus loss during centrifugal filtration while the pellet had a lower Ct score than the concentrate suggesting more viral content in the pellet relative to the concentrate **(Figure 4.1B)**. Typical RSV positive samples – those that underwent direct RNA extraction without pre-treatment (through centrifugal processing) had lower Ct scores compared to the concentrates **(Figure 4.1B)**. The concentrate's low RNA qubit scores and reduced viral load inferred reduced host contaminants as compared to the pellet and the typical sample.

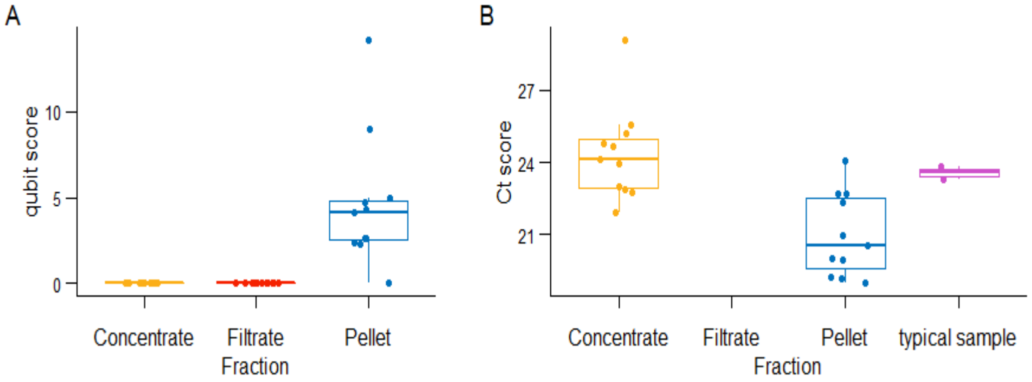


Figure 4.1: **Qubit and RT-PCR results from the different centrifugal processing fractions.** (A) Boxplot of the qubit scores from ten samples that underwent centrifugal processing against sample fraction. (B) A boxplot of RT-PCR cycle threshold scores of twelve samples against the sample fractions (concentrate, filtrate, and pellet) and typical samples (those that underwent direct RNA extraction without prior physical enrichment). The colors represent the sample fractions.

A comparison of the IFAT images from the concentrate and the pellet indicated that in addition to the green fluorescence signifying virus particles, the pellet had more red fluorescence indicative of host cells as compared to the concentrate, as shown in the images in **(Figure 4.2)**. The differences in the extents of fluorescence were indicative of the extents of host contamination in the two sample fractions.

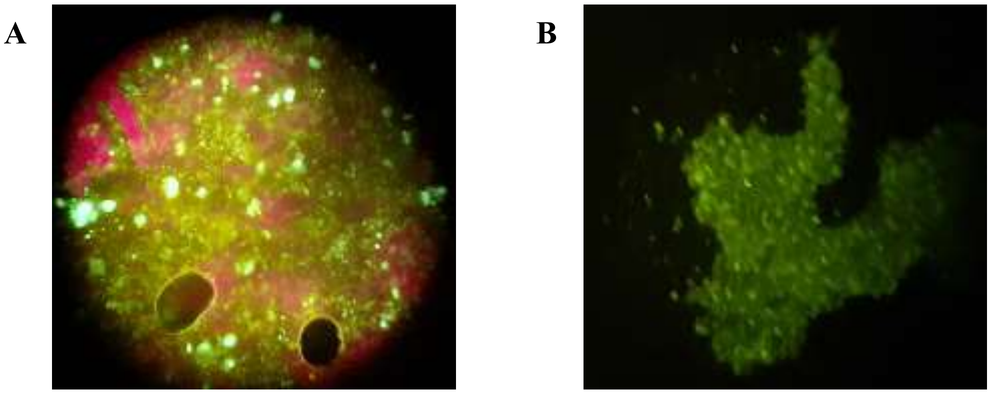
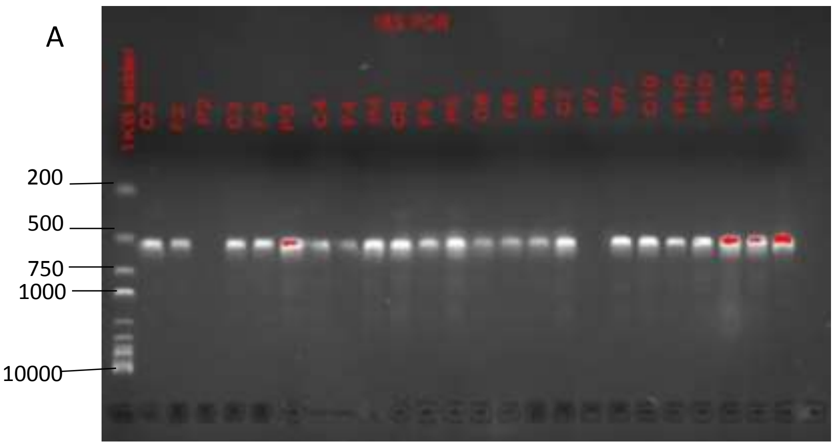


Figure 4.2: IFAT images from different centrifugal processing fractions. (A) from the pellet and (B) from the concentrate. Red fluorescence in the pellet represents host cells while green fluorescence in both the pellet and the concentrate represents RSV particles.

An analysis of the 16S rRNA PCR results indicated that the concentrate, which was the main sample fraction of focus in this study, still contained a lot of bacterial contamination **(Figure 4.3A)**. Alternatives to reduce the contamination entailed adoption of a DNase treatment step or passing the extracted RNA through DNA columns. Of the two alternatives, the DNase treatment step turned out to be the most effective in reducing the extent of bacterial contamination as compared to the use of DNA columns **(Figure 4.3B)**. However, treating the concentrates with DNase reduced the viral load initially present in the concentrates, as confirmed by a rise in Ct scores in the concentrates treated with DNase **(Figure 4.4)**. This observation led to the concentrates being treated with DNase just before RNA extraction, a strategy that deemed effective at reducing host contaminants while protecting the viral genomes from digestion and enhancing viral reads representation in the final metagenomics dataset in a study by (Lewandowska *et al*., 2017).



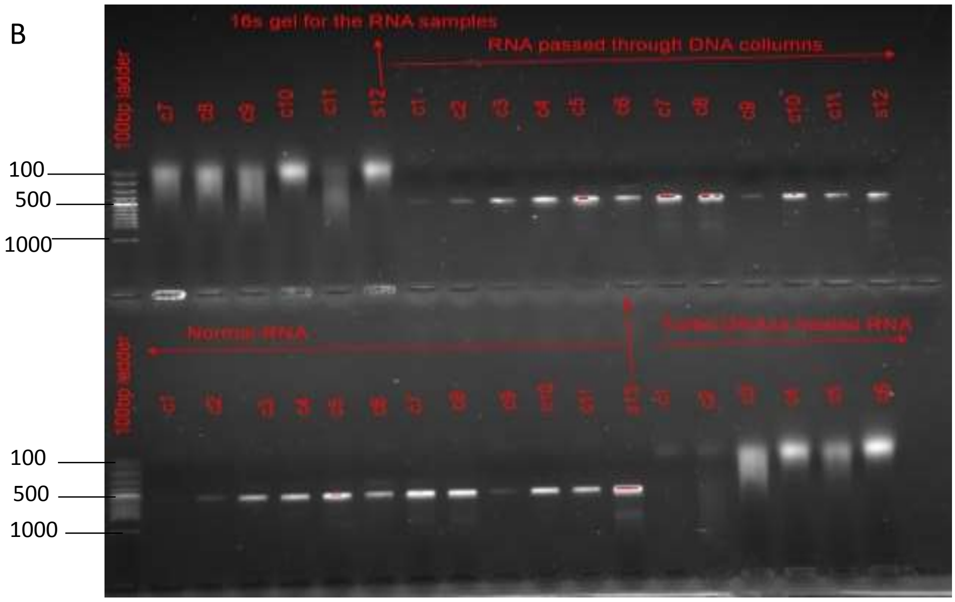


Figure 4.3: 16s rRNA gel images from the concentrates. Gel image (A) demonstrates bacterial contamination in the various sample fractions (1kb ladder was

used). Gel image (B) is an illustration of the impact of DNase treatment and DNA columns in reducing bacterial contamination (100bp ladder was used). A similar set of concentrates were used to inform on the extent of bacterial contamination and the best protocol for their depletion.

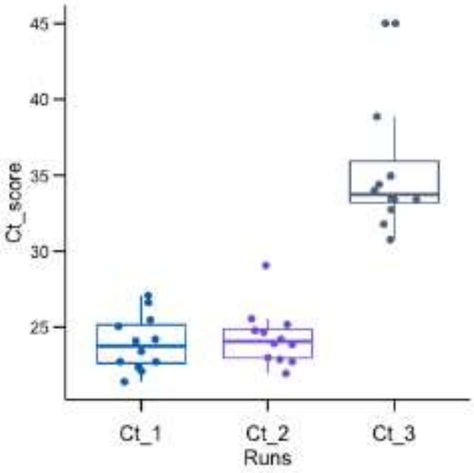


Figure 4.4: A boxplot of Ct values against runs. This boxplot demonstrates the effect of DNase treatment in reducing viral load content in the concentrate. Ct_1 represents the Ct values when selecting the samples (typical samples), Ct_2 the Ct scores from the concentrates after centrifugal processing and Ct_3 the Ct values after treating the concentrates with DNase

4.1.2: Sequence independent single primer amplification (SISPA)

Random amplification using SISPA resulted in PCR products of varying lengths ranging between 250 bases to 1500 bases. The varying PCR products were more prominent in the samples not treated with DNase **(Figure 4.5)**. The varying lengths in the band sizes demonstrated that the SISPA approach was successful in untargeted amplification of nucleic material present in each sample, in that, no specific band size was targeted and sequenced.

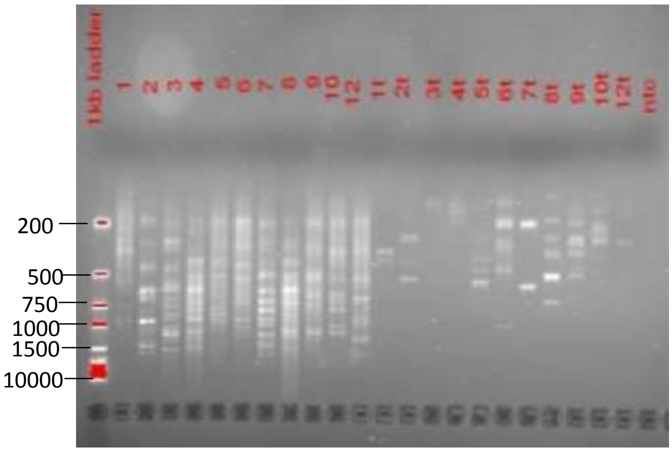

1kb ladder
1
2
3
4
5
6
7
8
9
10
12
1I
2I
3I
4I
5I
6I
7I
8I
9I
10I
12I
ntc
200
500
750
1000
1500
10000

Figure 4.5: A representation of the appearance of the gel images after performing SISPA. DNase treated sample fractions are denoted with a ‘t’ after the sample ID while the sample ID only denotes the untreated fractions.

4.2: Sequencing results from Protocol 1

This protocol yielded 8.2 million reads, 7.2 million of which passed quality check (QC) with their median read quality being 11.11. Taxonomic classification of all the reads that passed QC from this protocol using Kraken2 indicated that the most abundant domains were Eukaryota, and Bacteria as compared to those from viruses **(Figure 4.6A)**. A comparison of the extent of host and bacterial contamination between the DNase treated and untreated sample fractions indicated that DNase treated sample fractions had significantly lower contamination extents as compared to the untreated (*p*= 0.000011), **(Figure 4.7A)**. No full RSV genome was recovered from this protocol and the sequenced reads mainly mapped to part of the N gene **(Appendix IIA)**, with the total number of sequenced bases being roughly 470, spanning from around 1350 bases to around 1800 bases. Additional reads in samples labelled with barcodes 10 and 21 from the same protocol mapped to part of G and L genes respectively with the total number of sequenced bases being 271 and 266 spanning the regions between 4970 to 5245 and 12900 to 13166 respectively.

4.3: Sequencing results from Protocol 2

This protocol yielded 8.2 million reads, 6.8 million of which passed quality check (QC). The median read quality for all the reads that passed QC was 10.33. Taxonomic classification of the reads that passed QC using Kraken2 indicated that the most abundant domains from this protocol were also Eukaryota and Bacteria as compared to those from viruses **(Figure 4.6B)**. A comparison of bacterial and host contamination extents between the DNase treated and untreated sample fractions from this protocol also showed significantly lower contamination extents in the DNase treated fractions as compared to the untreated (*p*= 0.0000028) **(Figure 4.7B)**. Nonetheless, no full RSV genome was recovered from this protocol either with reads from barcodes 01 and 06 mapping to part of the G gene **(Appendix IIB)**, with the total number of sequenced reads being roughly 305 spanning the regions between

4900 to roughly 5200. Reads from barcodes 13-24 on the other hand mainly mapped to part of the L gene **(Appendix IIC)** with the total number of sequenced bases being roughly 258 spanning from around 12890 bases to 13160 bases.

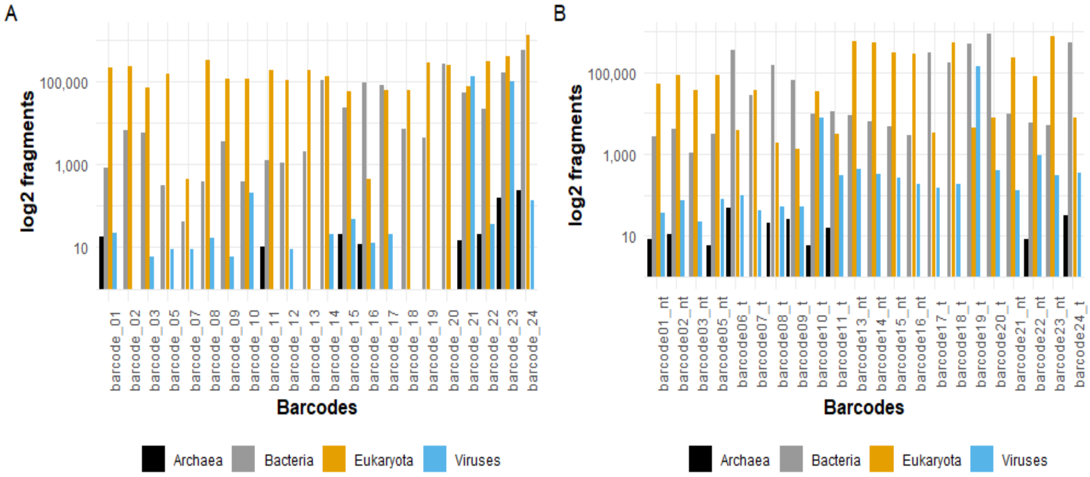


Figure 4.6: A graphical representation of the domains present per barcode. (A) represents the domains present in the sample fractions that underwent centrifugal processing, while (B) represents the domains present in the sample fractions that underwent direct RNA extraction.

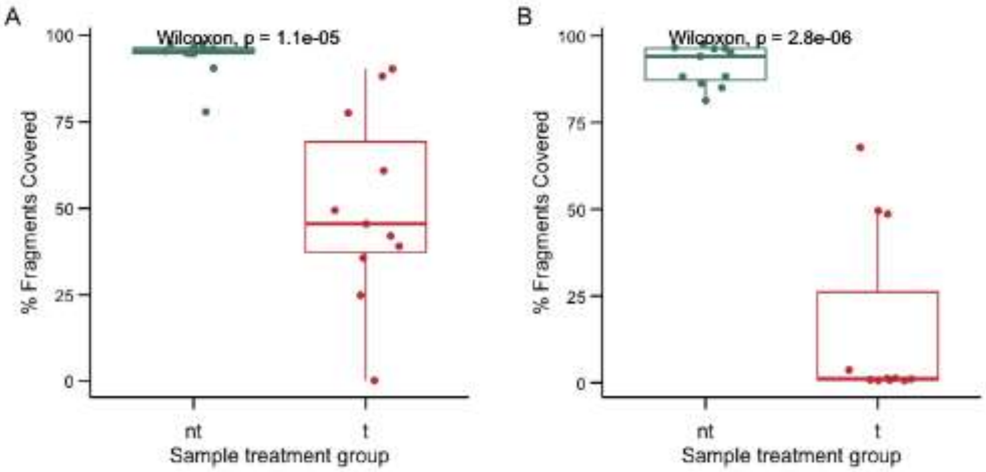


Figure 4.7: Boxplots of the distribution of host reads between sample treatment groups. (A) represents the distribution of host reads in DNase treated (t) and the non-treated (nt) sample groups in samples that underwent centrifugal processing while those that underwent direct RNA extraction are labelled (B).

sequence	**Primer name**	**Pattern**	**Strand**	**Start**	**End**	**Matched**
113388	Primer 59	CATATTG	-	12879	12885	CATATTG
113388	Primer 87	GATATCATGTTA	+	1355	1366	GATATCATGTTA
113388	Primer 92	CCATACT	+	4974	4980	CCATACT

4.4:5 Amplification tag bases annealed to the centroid genome

When the tagged Endoh primers were matched against the centroid genome generated using Vsearch cluster, primers 59, 89 and 92 were found to have some bases constituting the tag annealing to the centroid genome. These bases were speculated to assist the random hexamers at the 3’ end in annealing during the first strand synthesis.

Table 4.1: A tabulation of the Endoh primers that could have played a role in preferential amplification of the genomic regions of RSV in this study

4.5: Comparison of centrifugal processing and direct RNA extraction protocols

Given that the same 12 samples were sequenced in both protocols; we observed that the regions that the reads span varied per run with the average percentage genome coverage in reads that underwent centrifugal processing being 3% and 1% for those that underwent direct RNA extraction. In addition, when we compared the proportions of host reads between the DNase treated and untreated fractions from the two protocols, we observed that there was a significant difference in the treated fractions (p = 0.04), with greater reductions in those extracted using Protocol 2, while there was no significant difference in the untreated fractions (p = 0.44) between the two protocols **Figure 4.8A**. When we compared RSV reads yield from the two protocols, we observed a significant difference in the proportion of RSV reads between the DNase treated (p = 0.013) and untreated fractions (p = 0.0085) from both experimental setups with the more RSV reads in the DNase treated and directly extracted samples compared to those that underwent centrifugal processing **(Figure 4.8B)**.

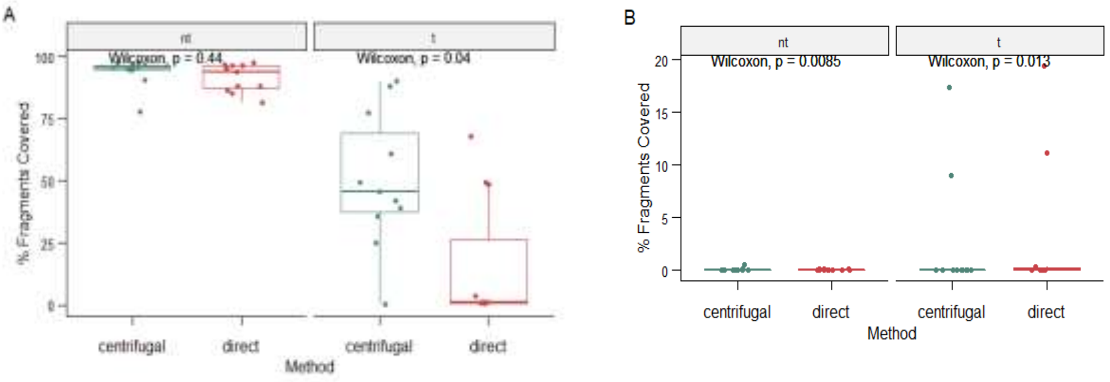


Figure 4.8: Boxplots comparing the proportion of host and RSV reads across methods. (A) shows the comparison of proportion of host reads and (B) the proportion of RSV reads between the treated (t) and untreated (nt) sample fractions in samples that underwent direct RNA extraction labelled direct and those that underwent centrifugal processing prior to extraction labelled centrifugal.

4.5: Bioinformatics analysis

From the bioinformatics analysis conducted in this study, it was evident that Kraken2 and Minimap2 gave inconsistent results when the number of RSV reads were quantified. Specifically, RSV reads quantification from Kraken2 showed that barcode 23 from protocol 1 had the greatest number of reads **(Figure 4.9A)** while reference mapping using the corresponding Illumina consensus sequence, showed that barcode21 had the highest number of reads **(Table 4.2)**. The same disparity was observed even with reads from protocol 2 where Kraken2 reports demonstrated that all the barcodes used in the run had some RSV reads while Minimap2 demonstrated that samples labelled with barcodes 13-24 and barcode 01 and 06 had reads that mapped, while those labelled with the rest of the barcodes had no reads that mapped (**Table 4.3**). In addition, a quantification of the RSV reads based on Kraken2 indicated that barcode19 had the greatest number of reads **(Figure 4.9B)** while Minimap2 showed that barcode22 had the highest number of reads **(Table 4.3).**

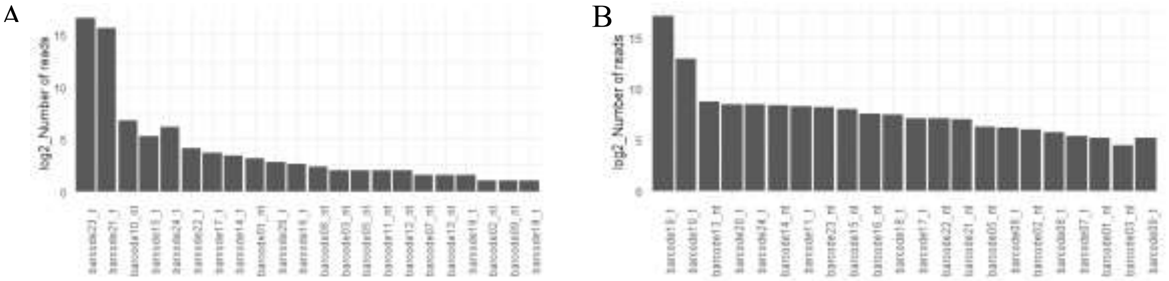


Figure 4.9: Graphical representation of the RSV reads distribution. RSV read distribution from (A) centrifugal processing and (B) direct RNA extraction based on Kraken2 taxonomic classification

Table 4.2: **A tabulation of Ct scores, barcodes, contig length, depth, and genome coverage from protocol 1.**

Sample Id	Fraction	RNA Ct values	cDNA Ct values	Barcode	RSV Contig length	RSV read depth	RSV % Genome coverage	Ref. length
128635	1_nt	22.78	23.35	barcode01	474	10	3.14	15064
	1_t	25.02	25.41	barcode14	473	9	3.14	15064
128247	2_nt	21.55	22.19	barcode02	474	12	3.14	15046
	2_t	23.72	25.13	barcode15	475	9	3.16	15046
129722	3_nt	22.59	24.26	barcode03	473	5	3.14	15060
	3_t	24.53	31.53	barcode16	473	6	3.14	15060
128367	4_nt	34.02	26.49	barcode05	0	0	0	15065
	4_t	27.22	27.0	barcode17	462	3	3.07	15065
129384	5_nt	25.72	25.02	barcode07	473	5	3.14	15062
	5_t	25.72	26.68	barcode18	500	7	3.32	15062
113388	6_nt	22.24	27.67	barcode08	473	19	3.10	15224
	6_t	22.89	26.68	barcode19	480	22	3.15	15224
113732	7_nt	24.54	27.67	barcode09	473	9	3.14	15184
	7_t	26.74	29.17	barcode20	473	37	3.14	15184
116032	8_nt	25.67	23.78	barcode10	745	87	4.91	15179
	8_t	21.75	24.01	barcode21	772	4771	5.09	15179
116026	9_nt	22.89	25.5	barcode11	473	6	3.12	15180
	9_t	23.61	27.18	barcode22	473	20	3.14	15180
116235	10_nt	23.49	23.71	barcode12	484	7	3.19	15184
	10_t	23.49	29.93	barcode23	478	66	3.15	15184
116410	12_nt	23.49	26.88	barcode13	474	15	3.11	15236
	12_t	23.46	29.72	barcode24	480	55	3.15	15236

11 samples were included in this protocol because two (116410 and 116469) were insufficient to be taken through both protocols

Table 4.3: A tabulation of Ct scores, barcodes, contig length, depth, and genome coverage from protocol 2.

Sample Id	Fraction	RNA Ct values	cDNA Ct values	Barcode	RSV Contig length	RSV read depth	RSV % Genome coverage	Ref. length
128635	1_nt	26.11	27.62	barcode13	258	7	1.71	15064
	1_t	28.6	30.23	barcode17	258	2	1.71	15064
128247	2_nt	24.1	25.79	barcode14	258	5	1.71	15046
	2_t	27.06	25.67	barcode18	258	3	1.71	15046
129722	3_nt	24.75	26.88	barcode15	262	4	1.74	15060
	3_t	27.13	26.98	barcode19	258	5	1.71	15060
128367	4_nt	23.35	26.48	barcode16	258	2	1.71	15065
	4_t	26.81	27.15	barcode20	258	6	1.71	15065
129384	5_nt	26.23	29.13	barcode01	306	5	2.03	15062
	5_t	29.46	27.42	barcode06	304	7	2.01	15062
113388	6_nt	21.8	24.62	barcode02	0	0	0	15224

	6_t	24.29	25.49	barcode07	0	0	0	15224
113732	7_nt	22.35	25.44	barcode03	0	0	0	15184
	7_t	24.45	25.3	barcode08	0	0	0	15184
116032	8_nt	23.03	25.37	barcode05	0	0	0	15179
	8_t	25.04	24.96	barcode09	0	0	0	15179
116026	9_nt	21.77	23.66	barcode21	274	7	1.81	15180
	9_t	23.49	24.3	barcode24	274	9	1.81	15180
116235	10_nt	21.17	24.5	barcode22	282	841	1.86	15184
	10_t	23.47	23.67	barcode10	0	0	0	15184
116469	11_nt	22.05	24.69	barcode23	274	10	1.81	15236
	11_t	23.56	25.81	barcode11	0	0	0	15236

11 samples were included in this protocol because two (116410 and 116469) were insufficient to be taken through both protocols

CHAPTER FIVE

DISCUSSION

5.1: Discussion

In this study, centrifugal processing, nuclease treatment using DNase and random amplification using SISPA were tested for metagenomics sequencing of clinical respiratory viruses in RSV positive specimens. The results from the sample extraction optimization step demonstrated that most of the viruses were embedded in the pellet, which was highly abundant in host cells **(Figure 4.2A)**. Centrifugal processing recovered freely floating viruses in the concentrate consisting of reduced host cells, although its viral load was also reduced. However, centrifugation processing showed little impact in reducing bacterial contamination as confirmed by 16s rRNA PCR**,** but DNase treatment method was deemed the most effective at reducing the extent of bacterial contamination but at the expense of reduced viral content. Despite these processes, no full RSV genome was recovered from either protocol.

A comparison of our findings from centrifugal processing optimization (**Figure 4.1 and 4.2**) showed congruence with what has been done previously since Hall et al. (2014), Goya *et al.* (2018) and Thurber *et al.* (2009) showed that the adoption of centrifugal filtration prior to RNA extraction at moderate speeds helped in reducing host contaminants and increased the recovery of viruses. Thurber *et al.* (2009) demonstrated that centrifugal processing was a suitable sample pre-treatment process because viruses are encapsulated enabling them to withstand concentration without resulting in the degradation of the nucleic material. Nevertheless, Hall *et al.* (2014) cautioned on the speed and time set while running centrifugal processing since the process results in reduced viral load and the loss was more significant with increased centrifugation speeds and time due to the continuous precipitation of the particles including viruses present in a sample. Low centrifugation speeds, on the other hand, had no impact in reducing host contaminants (Hall *et al.*, 2014).

This study further demonstrated that the use of centrifugal processing did not reduce the amount of bacterial contamination in the samples (**Figure 4.3A**). Hall *et al.* (2014) indicated that though the centrifugal filters reduced bacterial contamination in a clinical sample, their efficiency in facilitating bacterial loads reduction in a clinical specimen was reduced. DNase treatment as recommended by metagenomics studies by Goya *et al.* (2018), Allander *et al.* (2001) and Rosseel *et al.* (2015) was deemed most effective at improving the identification of viruses and reducing the extent of bacterial and host contaminants. The highly abundant host and bacterial reads compared to viruses in our dataset even after DNase treatment (**Figure 4.6**) confirmed how challenging it is to deplete the two major contaminants.

Reference mapping analysis from this study indicated that no complete RSV genome was recovered from either of the two protocols, with the identified genomic segments spanning varying regions of the genome from both protocols (**Appendix II**). These observations suggest an incidence of preferential amplification of the most abundant regions of the genome when SISPA was done. Rosseel *et al.* (2013) and Victoria *et al*. (2009) made closely similar observations and reported that the SISPA technique introduced coverage depth distribution bias. In their studies, Rosseel *et al*. (2013) and Victoria *et al.* (2009) observed gaps in areas of low complexity and exaggerated sequence depths in the preferentially amplified regions. Rosseel *et al*. (2013) attributed the SISPA coverage depth bias to annealing biases introduced by the primer used, where the annealing of the random hexamers is enhanced when some nucleotides termed as annealing sites specific to the 5’ amplification tag (designed for PCR amplification) assist the random hexamers at the 3’ end in annealing during first strand synthesis. Uneven distribution of the reads across the RSV genome and the variation in the regions that the reads span per run in this study are speculated to be because of part of the tag annealing to the genomic sequence and resulting to the over-amplification of the main regions that our reads span (**Table 4.1**). The primer labelled 87 specifically which presumably amplified part of the N gene recovered in this study, had six bases constituting the tag annealing to our centroid genome.

Additionally, the results from this study demonstrated that significant depletion of host and bacteria reads from viral reads was dependent on whether DNase was done prior to RNA extraction or after RNA extraction. Significant reduction in contamination levels was more evident in samples that were extracted using the direct RNA protocol and treated with DNase after RNA extraction as compared to those that underwent centrifugal processing and their concentrate treated with DNase prior to RNA extraction (**Figure 4.8A**). A high number of hosts reads after centrifugal processing and DNase treatment, as seen in this study, could be attributed to ribosomes held within the concentrate (Rosseel *et al*., 2015). Rosseel *et al*. (2015) indicated that pre-treating the concentrate with DNase prior to RNA extraction had no impact on ribosomal RNA as they stayed protected from the nucleases and were released during the RNA extraction process, resulting in high host reads relative abundances after extraction.

Another distinct observation from this study was the incongruence in the number of RSV reads from Kraken2 and from reference mapping using Minimap2. The distinctiveness could be indicative of the differences in the principles of the two tools with Kraken2 being kmer based (Wood *et al*., 2019) and minimap2 being alignment based (Li, 2018). Kraken2 being an ultrafast classifier (Wood *et al*., 2019) and assigning reads to a given taxa based on best matches even if part of the kmers match to a different taxon but at low similarity percentages compared to the one assigned, makes it highly error prone. Given that each barcode was mapped to the respective single Illumina consensus sequence, it is likely that some reads unlikely to map if several reference sequences were used, mapped to the reference further contributing to the incongruence.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1: Conclusion

The findings in this study show that viral enrichment using centrifugal processing is effective in reducing host contamination but not bacterial contamination. In addition, amplification using SISPA ensures unbiased amplification of nucleic material in the sample given that no specific fragments are targeted during the process. However, the approaches cannot be used independently since large amounts of host and bacterial reads are still recovered even after the two enrichment approaches are used independently, making it paramount to include enzymatic treatment using DNase. Most effective DNAse activity against constituting contaminants is evident if done after RNA extraction although with centrifugal processing it is accompanied by a significant loss in viruses. Further, for success when using SISPA approach during unbiased viral enrichment, the length of the random primers used is very important to avoid instances of preferential amplification biases introduced by using short hexamers in this study. Lastly, bioinformatics analysis results congruence is highly affected by the distinctiveness in principles underlying the tools used.

6.2: Recommendations

- The three enrichment approaches used in this study should be used together since when used concurrently they enhance the representation of viral reads better in the final dataset.
- When employing random priming as an enrichment approach, it is fundamental to lengthen the random part of the random primers to reduce the chances of preferential sequencing in some genomic regions
- Alternative approaches to enhance viral reads representation such as real time selective sequencing by reversing sequencing of reads that do not meet certain selected sequencing criteria such as length can be used as alternatives.

REFERENCES

Agoti, C. N., Munywoki, P. K., Phan, M. V. T., Otieno, J. R., Kamau, E., Bett, A., Kombe, I., Githinji, G., Medley, G. F., Cane, P. A., Kellam, P., Cotten, M., & Nokes, D. J. (2017). Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis. In *Virus Evolution* (Vol. 3, Issue 1). https://doi.org/10.1093/ve/vex006

Agoti, C. N., Otieno, J. R., Munywoki, P. K., Mwihuri, A. G., Cane, P. A., Nokes, D. J., Kellam, P., & Cotten, M. (2015). Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. *Journal of Virology*, *89*(7), 3444–3454. https://doi.org/10.1128/JVI.03391-14.

Allander, T., Emerson, S. U., Engle, R. E., Purcell, R. H., & Bukh, J. (2001). A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11609–11614. https://doi.org/10.1073/pnas.211424698

Ari, Sule, Arikan, M. (2016). Next-Generation Sequencing: Advantages, Disadvantages, and the Future. *Research Gate*, *October*. https://doi.org/10.1007/ 978-3-319-31703-8

Bächi, T., & Howe, C. (1973). Morphogenesis and ultrastructure of respiratory syncytial virus. *Journal of Virology*, *12*(5), 1173–1180.

Batovska, J., Lynch, S. E., Rodoni, B. C., Sawbridge, T. I., & Cogan, N. O. (2017). Metagenomic arbovirus detection using MinION nanopore sequencing. *Journal of Virological Methods*, *249*, 79–84. https://doi.org/10.1016/j.jviromet.2017.08 .019

Battles, M. B., & McLellan, J. S. (2019). Respiratory syncytial virus entry and how to block it. *Nature Reviews Microbiology*, *17*(4), 233–245. https://doi.org/10.1038/ s41579-019-0149-x

Bentley D.R, Balasubramaniam S, ..Smith A.J. (2008). Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517.Accurate

Boykin, L., Ghalab, A., Rossitto De Marchi, B., Savill, A., M. Wainaina, J., Kinene, T., Lamb, S., Rodrigues, M., Kehoe, M., Ndunguru, J., Tairo, F., Sseruwagi, P., Kayuki, C., Mark, D., Erasto, J., Bachwenkizi, H., Alicai, T., Okao-Okuja, G., Abridrabo, P., … Kiarie, S. (2018). Real time portable genome sequencing for global food security. *F1000Research*, *7*, 1101. https://doi.org/10.12688/f1000research.15507.1

Bugembe, D. L., Phan, M. V. T., Ssewanyana, I., Semanda, P., Nansumba, H., Dhaala, B., Nabadda, S., O'Toole, Á. N., Rambaut, A., Kaleebu, P., & Cotten, M. (2021). Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. *Nature Microbiology*, *6*(8), 1094–1101. https://doi.org/10.1038/s41564-021-00933-9

Cane, P. A., Matthews, D. A., & Pringle, C. R. (1994). Analysis of respiratory syncytial virus strain variation in successive epidemics in one city. *Journal of Clinical Microbiology*, *32*(1), 1–4.

Chanock, R., & Finberg, L. (1957). Recovery From Infants With Respiratory Illness of a Virus Related To Chimpanzee Coryza Agent (Cca) Epidemiologic. *American Journal of Epidemiology*, *M*(4), 291–300. https://doi.org/10.1093/oxfordjournals .aje.a119902

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., &

Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. https://doi.org/10.1038/ nmeth.2474

Chrzastek, K., Lee, D. hun, Smith, D., Sharma, P., Suarez, D. L., Pantin-Jackwood, M., & Kapczynski, D. R. (2017). Use of Sequence-Independent, Single-Primer-Amplification (SISPA) for rapid detection, identification, and characterization of avian RNA viruses. *Virology*, *509*, 159–166. https://doi.org/10.1016/j.virol.2017.06.019

Collins, P. L., Huang, Y. T., & Wertz, G. W. (1984). Identification of a Tenth mRNA of Respiratory Syncytial Virus and Assignment of Polypeptides to the 10 Viral Genes. *Journal of Virology*, *49*(2), 572–578.

Collins, P. L., Olmsted, R. A., & Johnson, P. R. (1990). The small hydrophobic protein of human respiratory syncytial virus: Comparison between antigenic subgroups A and B. *Journal of General Virology*, *71*(7), 1571–1576. https://doi.org/10.1099/0022-1317-71-7-1571

Collins, P. L., Olmstedt, R. A., & Johnson, P. R. (1990). The small hydrophobic protein of human respiratory syncytial virus : comparison between antigenic subgroups A and B. *Journal of General Virology*, *71*, 1571–1576.

Conceição-Neto, N., Zeller, M., Lefrère, H., de Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., Ranst, M. van, Heylen, E., & Matthijnssens, J. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. *Scientific Reports*, *5*(October), 1–14. https://doi.org/10.1038/srep16532

Eckert, S. E., Chan, J. Z.-M., Houniet, D., Breuer, J., & Speight, G. (2016). Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. *Microbial Genomics*, *2*(9). https://doi.org/10.1099/mgen.0.000087

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/ science.1162986

Endoh, D., Mizutani, T., Kirisawa, R., Maki, Y., Saito, H., Kon, Y., Morikawa, S., & Hayashi, M. (2005). Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. *Nucleic Acids Research, 33*(6). https://doi.org/10.1093/nar/gni064

Englund, J. A., Anderson, L. J., & Rhame, F. S. (1991). Nosocomial transmission of respiratory syncytial virus in immunocompromised adults. *Journal of Clinical Microbiology*, *29*(1), 115–119.

Fearns, R., & Collins, P. L. (1999). Role of the M2-1 transcription antitermination protein of respiratory syncytial virus in sequential transcription. *J Virol.*, *73*(7), 5852–5864.

Froussard, P. (1992). A random-POR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Cold Spring Harbor Laboratory Press*, *20*(11), 2900.

Githinji, G., de Laurent, Z. R., Mohammed, K. S., Omuoyo, D. O., Macharia, P. M., Morobe, J. M., Otieno, E., Kinyanjui, S. M., Agweyu, A., Maitha, E., Kitole, B., Suleiman, T., Mwakinangu, M., Nyambu, J., Otieno, J., Salim, B., Kasera, K., Kiiru, J., Aman, R., … Agoti, C. N. (2021). Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya.

Nature Communications, *12*(1). https://doi.org/10.1038/s41467-021-25137-x

Goya, S., Valinotto, L. E., Tittarelli, E., Rojo, G. L., Nabaes Jodar, M. S., Greninger, A. L., Zaiat, J. J., Marti, M. A., Mistchenko, A. S., & Viegas, M. (2018). An optimized methodology for whole genome sequencing of RNA respiratory viruses from nasopharyngeal aspirates. *PLoS ONE*, *13*(6), 1–15. https://doi.org/10.1371/journal.pone.0199714

Graf, Erin H.Schlaberg, R., Tardif, K. D., Flygare, S., Simmon, K. E., Hymas, W., Eilbeck, K., & Yandell, M. (2016). Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel. *Journal of Clinical Microbiology*, *54*(4), 1000–1007. https://doi.org/10.1128/jcm.03060-15

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J. M., Dodd, R., Mulembakani, P., Schneider, B. S., Stramer, S. L., & Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, 1–13. https://doi.org/10.1186/s13073-015-0220-9

Hall, C. B. (1982). Respiratory syncytial virus: Its transmission in the hospital environment. *Yale Journal of Biology and Medicine*, *55*(3–4), 219–223.

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., Moore, N. E., Ren, X., Huang, Q. S., Carter, P. E., & Peacey, M. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *Journal of Virological Methods*, *195*, 194–204. https://doi.org/10.1016/j.jviromet.2013.08.035

Hammitt, L. L., Kazungu, S., Welch, S., Bett, A., Onyango, C. O., Gunson, R. N., Scott, J. A. G., & Nokes, D. J. (2011). Added Value of an Oropharyngeal Swab in Detection of Viruses in Children Hospitalized with Lower Respiratory Tract Infection. *Journal of Clinical Microbiology*, *49*(6), 2318–2320. https://doi.org/10.1128/JCM.02605-10

Haque, F., Li, J., Wu, H.-C., Liang, X.-J., & Guo, P. (2013). Solid-State and Biological Nanopore for Real Time Sensing of Single Chemical and Sequencing DNA. *NIH Public Acccess*, *8*(1), 1–7. https://doi.org/10.1038/jid.2014.371

Johnson, P. R., Spriggs, M. K., Olmsted, R. A., & Collins, P. L. (1987). The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins. *Proceedings of the National Academy of Sciences*, *84*(16), 5625–5629. https://doi.org/10.1073/pnas.84.16.5625

Kafetzopoulou, L. E., Efthymiadis, K., Lewandowski, K., Crook, A., Carter, D., Osborne, J., Aarons, E., Hewson, R., Hiscox, J. A., Carroll, M. W., Vipond, R., & Pullan, S. T. (2018). Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Eurosurveillance*, *23*(50), 1–13. https://doi.org/10.2807/1560-7917.ES.2018.23.50.1800228

Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine*, *09*(03). https://doi.org/10.4172/0974-8369.1000395

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. https://doi.org/10.1038/nbt.2280

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. https://doi.org/10.1016/j.bdq.2015.02.001

Lee, N., Lui, G. C. Y., Wong, K. T., Li, T. C. M., Tse, E. C. M., Chan, J. Y. C., Yu, J., Wong, S. S. M., Choi, K. W., Wong, R. Y. K., Ngai, K. L. K., Hui, D. S. C., & Chan, P. K. S. (2013). High morbidity and mortality in adults hospitalized for respiratory syncytial virus infections. *Clinical Infectious Diseases*, *57*(8), 1069–1077. https://doi.org/10.1093/cid/cit471

Leger, A., & Leonardi, T. (2019). PycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open-Source Software*, *4*(34), 1236. https://doi.org/10.21105/joss.01236

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany*, *68*(20), 5419–5429. https://doi.org/10.1093/jxb/erx289

Lewandowska, D. W., Zagordi, O., Geissberger, F. D., Kufner, V., Schmutz, S., Böni, J., Metzner, K. J., Trkola, A., & Huber, M. (2017). Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. *Microbiome*, *5*(1), 94. https://doi.org/10.1186 /s40168-017-0317-z

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*,*34*(18),3094–3100.https://doi.org/10.1093/bioinformatics /bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.10 93/bioinformatics/btp352

Li, J., Wang, H., Mao, L., Yu, H., Yu, X., Sun, Z., Qian, X., Cheng, S., Chen, S., Chen, J., Pan, J., Shi, J., & Wang, X. (2020). Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. *Scientific Reports*, *10*(1), 1–10. https://doi.org/10.1038/s41598-020-74656-y

Li, X., Willem, L., Antillon, M., Bilcke, J., Jit, M., & Beutels, P. (2020). Health and economic burden of respiratory syncytial virus (RSV) disease and the cost-effectiveness of potential interventions against RSV among children under 5 years in 72 Gavi-eligible countries. *BMC Medicine, 18*(1). https://doi.org/10.1186/s12916-020-01537-6

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*.https://doi.org/10.1155/2012/251364

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *BioRxiv*, 1–21.

Loman, N. J., & Quinlan, A. R. (2014). Poretools: A toolkit for analyzing nanopore sequence data. *Bioinformatics*, *30*(23), 3399–3401. https://doi.org/10.1093/bioinformatics/btu555

Lu Hengyun, Giordano Francesca, N. Z. (2016). Oxford Nanopore MinION Sequencing and GenomeAssembly.*GenomicsProteomics Bioinformatics,* *14*(2016), 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., D'Aurizio, R. (2017). Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in Bioinformatics*, *1*(1), 1–17.

Marietjie Venter, Ria Lassaunie` re, Tina Louise Kresfelder, Yvette Westerberg, A., & Visser, A. (2011). Contribution of Common and Recently Described Respiratory Viruses to Annual Hospitalizations in

Children in South Africa. *Journal of Medical Virology*, *83*(November 2005), 1458–1468. https://doi.org/10.1002/jmv

Mongan, A. E., Tuda, J. S. B., & Runtuwene, L. R. (2019). Portable sequencer in the fight against infectious disease. *Journal of Human Genetics*, 35–40. https://doi.org/10.1038/s10038-019-0675-4

Mufson A. Maurice, Orvell Claes, Rafnar Bjorg, N. E. (1985). Two Distinct Subtypes of Human Respiratory Syncytial Virus. *Journal of General Virology*, *66*(1), 2111–2124.

Nguyen, A. T., Tran, T. T., Hoang, V. M. T., Nghiem, N. M., Nguyen, N., & Le, T. (2016). Development and evaluation of a non- ribosomal random PCR and next- generation sequencing based assay for detection and sequencing of hand, foot, and mouth disease pathogens. *Virology Journal*, 1–10. https://doi.org/10.1186/s12985-016-0580-9

Nokes, D. J., Okiro, E. A., Ngama, M., Ochola, R., White, L. J., Scott, P. D., English, M., Cane, P. A., & Medley, G. F. (2008). Respiratory syncytial virus infection and disease in infants and young children observed from birth in Kilifi District, Kenya. *Clinical Infectious Diseases*, *46*(1), 50–57. https://doi.org/10.1086/524019

Otieno, J. R., Kamau, E. M., Oketch, J. W., Ngoi, J. M., Gichuki, A. M., Binter, Š., Otieno, G. P., Ngama, M., Agoti, C. N., Cane, P. A., Kellam, P., Cotten, M., Lemey, P., & Nokes, D. J. (2018). Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. *Virus Evolution*, *4*(2), 1–13. https://doi.org/10.1093/ve/vey027

Peret, T. C. T., Hall, C. B., Schnabel, K. C., Golub, J. A., & Anderson, L. J. (1998). Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *Journal of*

General Virology, *79*(9), 2221–2229. https://doi.org/10.1099/0022-1317-79-9-2221

Piedimonte, G., & Perez, M. K. (2014). Respiratory Syncytial Virus Infection and Bronchiolitis. *Pediatrics in Review*, *35*(12), 519–530. https://doi.org/10.1542/pir.35-12-519

Plesivkova, D., Richards, R., & Harbison, S. (2019). A review of the potential of the MinION^{TM} single-molecule sequencing system for forensic applications. *Wiley Interdisciplinary Reviews: Forensic Science*, *1*(1), e1323. https://doi.org/10.1002/wfs2.1323

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S. C., Black, A., Bedford, T., Carroll, M. W., Nunes, M., … Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, *12*(6), 1261–1266. https://doi.org/10.1038/nprot.2017.066

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H. J., Becker-Ziaja, B., Boettcher, J. P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L. L., … Carroll, M. W. (2016). Real-time,portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. https://doi.org/10.1038/ nature 16996

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to base pair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, *19*(1), 1–11. https://doi.org/10.1186/s 13059-018-1462-9

Rebuffo-Scheer, C., Bose, M., He, J., Khaja, S., Ulatowski, M., Beck, E. T., Fan, J., Kumar, S., Nelson, M. I., & Henrickson, K. J. (2011). Whole genome sequencing and evolutionary analysis of human respiratory syncytial virus A and B from Milwaukee, WI 1998-2010. *PLoS ONE*, *6*(10). https://doi.org/10.1371/ journal. pone.0025468

Reyes, G. R., & Kim, J. P. (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molecular and Cellular Probes*, *5*(6), 473–481. https://doi.org/10.1016/S0890-8508(05)80020-9

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*,*13*(5),278–289. https://doi.org /10.1016/ j.gpb.2015.08.002

Rima, B., Collins, P., Easton, A., Fouchier, R., Kurath, G., Lamb, R. A., Lee, B., Maisner, A., Rota, P., & Wang, L. (2017). ICTV virus taxonomy profile: Pneumoviridae. *Journal of General Virology*, *98*(12), 2912–2913. https://doi.org/10.1099/jgv.0.000959

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). *VSEARCH :* a versatile open source tool for metagenomics. PEERJ, 1–22. https://doi.org/10.7717/peerj.2584

Rosseel, T., Ozhelvaci, O., Freimanis, G., & van Borm, S. (2015). Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *Journal of Virological Methods*, *222*, 72–80. https://doi.org/10.1016/j.jviromet.2015.05.010

Rosseel, T., van Borm, S., Vandenbussche, F., Hoffmann, B., van den Berg, T., Beer, M., & Höper, D. (2013). The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. *PLoS ONE*, *8*(9), 1–9. https://doi.org/10.1371/ journal. pone.0076144

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. https://doi.org/10.1038/nature10242

Sanger, F. (1988). SEQUENCES, SEQUENCES , AND SEQUENCES. *Annual Review of Bichemistry*, *57*(1), 1–28.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit : A Cross-Platform and Ultrafast Toolkit for FASTA / Q File Manipulation. *PLoS ONE,* 1–10. https://doi.org/10.1371/journal.pone.0163962

Shi, T., McAllister, D. A., O'Brien, K. L., Simoes, E. A. F., Madhi, S. A., Gessner, B. D., Polack, F. P., Balsells, E., Acacio, S., Aguayo, C., Alassani, I., Ali, A., Antonio, M., Awasthi, S., Awori, J. O., Azziz-Baumgartner, E., Baggett, H. C., Baillie, V. L., Balmaseda, A., … Nair, H. (2017). Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *The Lancet*, *390*(10098), 946–958. https://doi.org/10.1016/S0140-6736(17)30938-8

Simoes, E. A. F., & Groothuis, J. R. (2002). Respiratory syncytial virus prophylaxis - The story so far. *Respiratory Medicine, 96*(SUPPL. 2), 15–24. https://doi.org/10.1053/rmed.2002.1296

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(19), 7702–7707. https://doi.org/10.1073/pnas.0901054106

Swerdlow, H., & Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, *18*(6), 1415–1419. https://doi.org/10.1093/nar/18.6.1415

Tedersoo, L., Nilsson, R. H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., Bahram, M., Bechem, E., Chuyong, G., & Kõljalg, U. (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, *188*(1), 291–301. https://doi.org/10.1111/j.1469-8137.2010.03373.x

Thomas Froehlich, Dieter Heindl, A. R. (2010). Miniaturized, High-Throughput Nucleic Acid Analysis. *Patent Application Publication*, *1*(19).

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics, 14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Thurber, R. v., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature Protocols*, *4*(4), 470–483. https://doi.org/10.1038/nprot.2009.10

Tipu, H. N., & Shabbir, A. (2015). Evolution of DNA sequencing. *Journal of College of Physicians and Surgeons Pakistan*, *25*(4), 210–215.

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research, 38*(15). https://doi.org/10.1093/nar/gkq543

Turner, T. L., Kopp, B. T., Paul, G., Landgrave, L. C., Hayes, D., & Thompson, R. (2014). Respiratory syncytial virus: Current and emerging treatment options. *ClinicoEconomics and Outcomes Research*, *6*(1), 217–225. https://doi.org/10.2147/ceor.s60710

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., & Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-29334-5

Umuhoza, T., Bulimo, W. D., Oyugi, J., Musabyimana, J. P., Kinengyere, A. A., & Mancuso, J. D. (2021). Prevalence of human respiratory syncytial virus, parainfluenza and adenoviruses in East Africa Community partner states of Kenya, Tanzania, and Uganda: A systematic review and meta-analysis (2007–2020). In *PLoS ONE* (Vol. 16, Issue 4 April 2021). Public Library of Science. https://doi.org/10.1371/journal.pone.0249992

Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., & Delwart, E. (2009). Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis. *Journal of Virology*, *83*(9), 4642–4651. https://doi.org/10.1128/JVI.02301-08

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *BioRxiv*, 762302. https://doi.org/10.1101/762302

Xu Yifei, Lewandowski Kuiama, Lumley Sheila, Sanderson Nicholas, Vaughan Ali, Vipond Richard, Carroll Miles, Jeffery Katie, Foster Dona, Walker A Sarah, Peto Timothy, Crook Derrick, Pullan Steven, M. P. (2018). Nanopore metagenomic sequencing of full length human metapneumovirus (HMPV) within a unique sub-lineage. *BioRxiv*, 1–37. https://doi.org/10.1093/annonc/mdy039/4835470

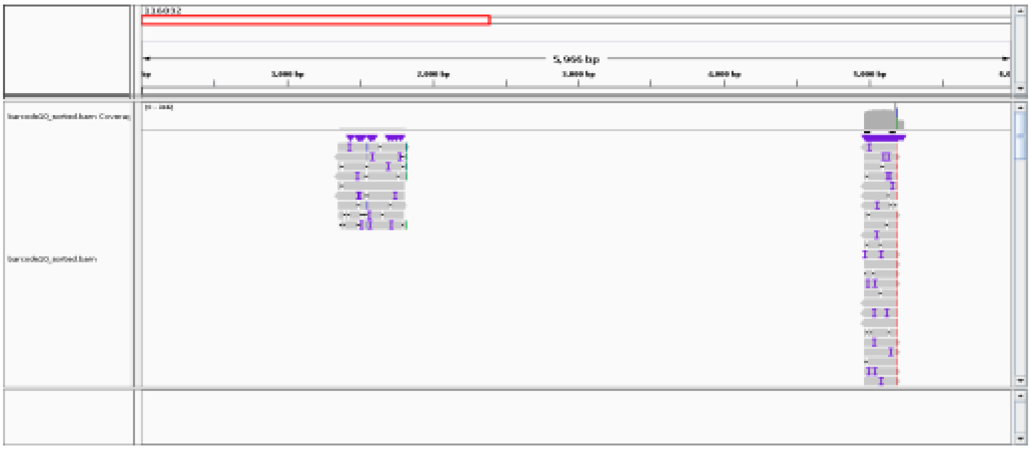
Zhang, L., Peeples, M.E., Boucher, R.C., Collins, P.L., Pickles, R. (2002). Respiratory Syncytial Virus Infection of Human Airway Epithelial Cells Is Polarized, Specific to Ciliated Cells, and without Obvious Cytopathology. *Journal of Virology*, *76*(11), 5654–5666. https://doi.org/10.1128/JVI.76.11.5654

APPENDICES

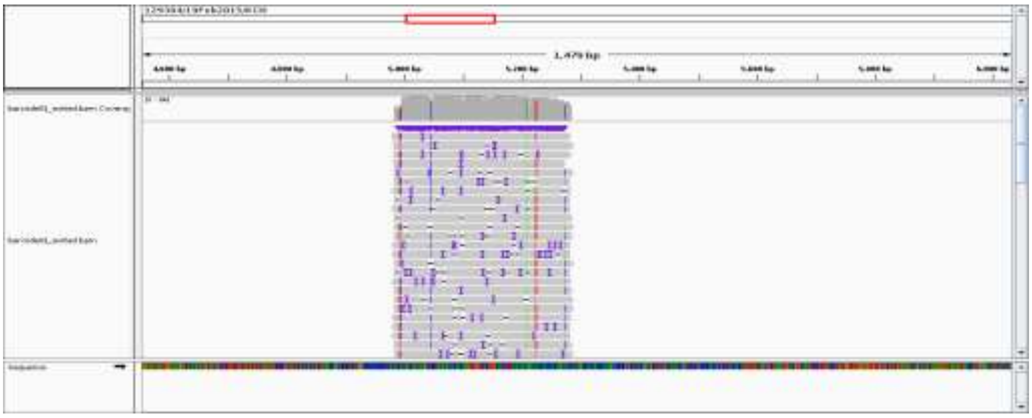
Appendix I: Set of Endoh non-ribosomal hexamers used in this study. The random hexamers tabulated here were tagged to the FR20RV tag.

GATATC	GATACT	CGATAT	ACTACT	ATAGTC	CTTAGT	ACTAAG	AACTTA
TAGTAT	CGTATA	GTATAC	TAACGA	CTAGTA	CTTACA	GCATAC	ATAACG
TATAGT	GTATAG	AATCCA	CGACTA	GTACTA	TTATGC	CAATAT	ATGTTA
TATATA	CGGTTA	TAGCAC	TACTAG	TAAGTT	ATACGC	ACCGTA	TGGTAT
ATACTA	AATAGT	ATATCG	AGTAGT	ATATCC	CGCTTA	GTGCTA	TGCGTA
ATATAT	CGCATA	AATATT	GTTAAC	TCGATA	TAACGC	ACGCTA	GGATAT
GTGCAC	ATTACG	TATAGC	GTCTAC	GTACCA	GGTCAT	ATGTCG	CATAGC
ACTATA	TTAACA	CTTGTA	TACAAG	GTATCA	CTCATA	AGCTTA	CATACT
CGTAAT	AGTATC	TAGTCG	TACCAG	ATACTC	AATTTG	CGACAT	CGGATA
CTATAC	TGTTAA	GTAGAC	TGGATT	ACATTA	CTGGTA	GCTATA	TTACTA
TATACG	ACTATT	CTATAG	TCGTTA	ATATTG	TTCATG	GCTATG	ACTCGT
TATGCG	TAACCG	TAGCTA	ATAGTA	CGTCTA	GCGATA	TGTAAG	TAAGGT

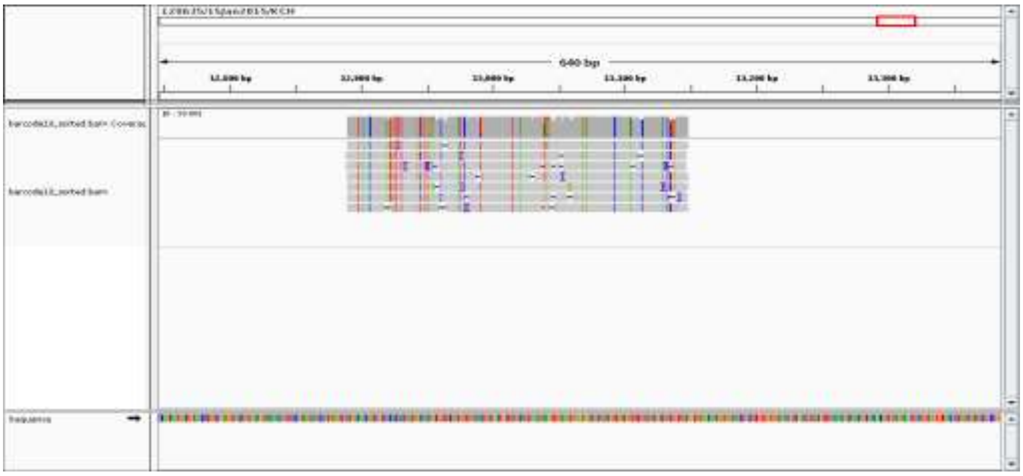
a


5,966 bp
barcode20_sorted.bam Coverag
barcode20_sorted.bam

b

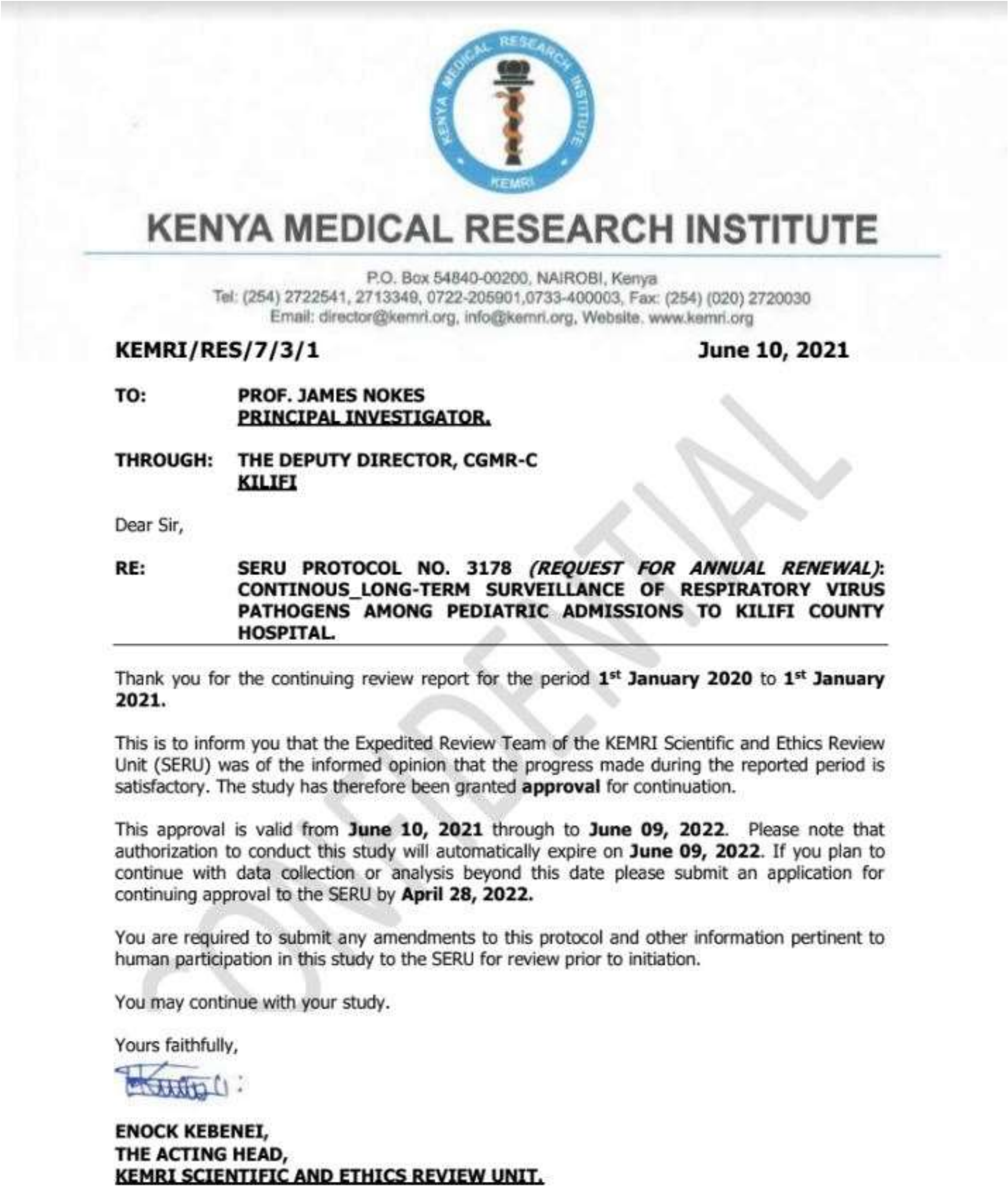

1,470 bp
barcode01_sorted.bam Coverag
barcode01_sorted.bam
Sequence

c


640 bp
barcode13_sorted.bam Coverag
barcode13_sorted.bam
Sequence

Appendix II: Screen shots of the regions to which RSV reads mapped using Illumina consensus references. (a) illustrates the region to which the reads from the centrifugal processing protocol mapped: part of the N gene, with some additional reads mapping to part of the G gene while (b) and (c) illustrates the regions to which the reads from the direct RNA extraction protocol mapped: part of the RSV G and L genes respectively

Appendix III: Ethical approval for this study



KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840-00200, NAIROBI, Kenya
Tel: (254) 2722541, 2713349, 0722-205901,0733-400003, Fax: (254) (020) 2720030
Email: director@kemri.org, info@kemri.org, Website. www.kemri.org

KEMRI/RES/7/3/1 **June 10, 2021**

TO: **PROF. JAMES NOKES**
PRINCIPAL INVESTIGATOR.

THROUGH: **THE DEPUTY DIRECTOR, CGMR-C**
KILIFI

Dear Sir,

RE: **SERU PROTOCOL NO. 3178 *(REQUEST FOR ANNUAL RENEWAL)*: CONTINOUS_LONG-TERM SURVEILLANCE OF RESPIRATORY VIRUS PATHOGENS AMONG PEDIATRIC ADMISSIONS TO KILIFI COUNTY HOSPITAL.**

Thank you for the continuing review report for the period **1st January 2020** to **1st January 2021.**

This is to inform you that the Expedited Review Team of the KEMRI Scientific and Ethics Review Unit (SERU) was of the informed opinion that the progress made during the reported period is satisfactory. The study has therefore been granted **approval** for continuation.

This approval is valid from **June 10, 2021** through to **June 09, 2022**. Please note that authorization to conduct this study will automatically expire on **June 09, 2022**. If you plan to continue with data collection or analysis beyond this date please submit an application for continuing approval to the SERU by **April 28, 2022.**

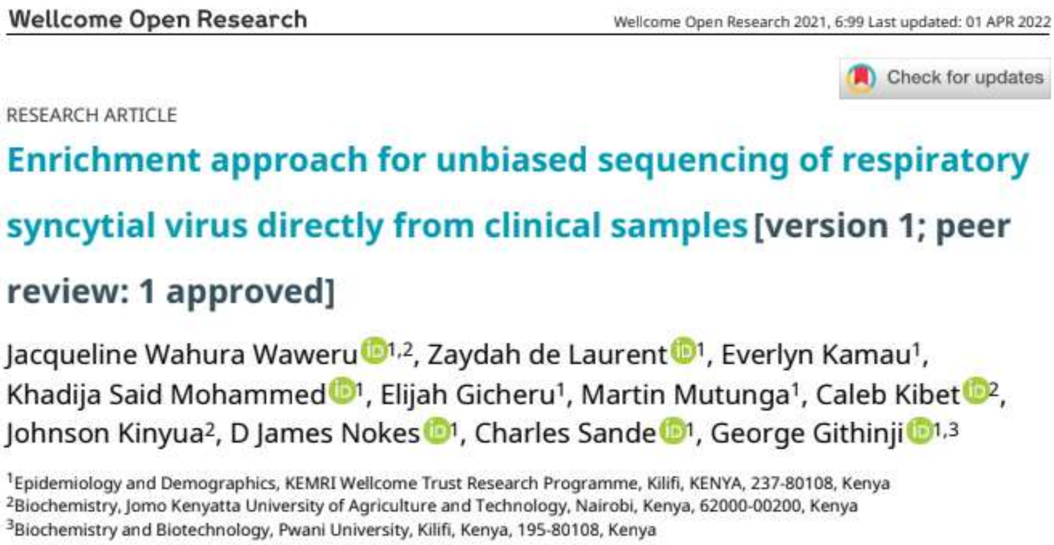
You are required to submit any amendments to this protocol and other information pertinent to human participation in this study to the SERU for review prior to initiation.

You may continue with your study.

Yours faithfully,

ENOCK KEBENEI,
THE ACTING HEAD,
KEMRI SCIENTIFIC AND ETHICS REVIEW UNIT.

Appendix IV: Publication from this study



Wellcome Open Research

Wellcome Open Research 2021, 6:99 Last updated: 01 APR 2022

Check for updates

RESEARCH ARTICLE

Enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples [version 1; peer review: 1 approved]

Jacqueline Wahura Waweru[1,2], Zaydah de Laurent[1], Everlyn Kamau[1], Khadija Said Mohammed[1], Elijah Gicheru[1], Martin Mutunga[1], Caleb Kibet[2], Johnson Kinyua[2], D James Nokes[1], Charles Sande[1], George Githinji[1,3]

[1]Epidemiology and Demographics, KEMRI Wellcome Trust Research Programme, Kilifi, KENYA, 237-80108, Kenya
[2]Biochemistry, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya, 62000-00200, Kenya
[3]Biochemistry and Biotechnology, Pwani University, Kilifi, Kenya, 195-80108, Kenya