RESEARCH ARTICLE

Enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples [version 1; peer review: 2 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

v1 **First published:** 07 May 2021, **6**:99 https://doi.org/10.12688/wellcomeopenres.16756.1
Latest published: 07 May 2021, **6**:99 https://doi.org/10.12688/wellcomeopenres.16756.1

Abstract

Background: Nasopharyngeal samples contain higher quantities of bacterial and host nucleic acids relative to viruses; presenting challenges during virus metagenomics sequencing, which underpins agnostic sequencing protocols. We aimed to develop a viral enrichment protocol for unbiased whole-genome sequencing of respiratory syncytial virus (RSV) from nasopharyngeal samples using the Oxford Nanopore Technology (ONT) MinION platform.
Methods: We assessed two protocols using RSV positive samples. Protocol 1 involved physical pre-treatment of samples by centrifugal processing before RNA extraction, while Protocol 2 entailed direct RNA extraction without prior enrichment. Concentrates from Protocol 1 and RNA extracts from Protocol 2 were each divided into two fractions; one was DNase treated while the other was not. RNA was then extracted from both concentrate fractions per sample and RNA from both protocols converted to cDNA, which was then amplified using the tagged Endoh primers through Sequence-Independent Single-Primer Amplification (SISPA) approach, a library prepared, and sequencing done. Statistical significance during analysis was tested using the Wilcoxon signed-rank test.
Results: 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) 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.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 07 May 2021	✓ view	✓ view

1. **Martin M. Nyaga**, University of the Free State, Bloemfontein, South Africa
   Milton Mogotsi, University of the Free State, Bloemfontein, South Africa
2. **Gerald Mboowa**, Makerere University, Kampala, Uganda

Any reports and responses or comments on the article can be found at the end of the article.

Conclusions: DNase treatment was most effective in reducing host and bacterial contamination, but its effectiveness improved if done after RNA extraction than before. We attribute the incomplete genome segments to amplification biases resulting from the use of short length random sequence (6 bases) in tagged Endoh primers. Increasing the length of the random nucleotides from six hexamers to nine or 12 in future studies may reduce the coverage biases.

Keywords

metagenomics, sequencing, SISPA, RSV, centrifugal processing, Endoh primers, DNase, RNA

KEMRI wellcome

This article is included in the KEMRI | Wellcome Trust gateway.

Corresponding author: Jacqueline Wahura Waweru (jackiemaingih@gmail.com)

Author roles: Waweru JW: Conceptualization, Data Curation, Formal Analysis, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **de Laurent Z**: Conceptualization, Methodology, Validation; **Kamau E**: Conceptualization, Methodology, Validation, Writing – Review & Editing; **Mohammed KS**: Methodology, Validation; **Gicheru E**: Methodology, Validation; **Mutunga M**: Methodology, Validation; **Kibet C**: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Review & Editing; **Kinyua J**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Nokes DJ**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Visualization, Writing – Review & Editing; **Sande C**: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – Review & Editing; **Githinji G**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust, [102975], the National Institute for Health Research (NIHR) [project references 17/63/82] using UK aid from the UK Government to support global health research, and the DELTAS Africa Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107769/Z/10/Z] and the UK government. The views expressed in this publication are those of the author (s) and do not necessarily reflect those of Wellcome Trust, the UK government's official policies AAS or the NEPAD Agency. In addition, the funders had no role in the design of the study; collection, analyses, or interpretation of data; manuscript writing, or in the decision to publish the results.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Waweru JW, de Laurent Z, Kamau E *et al.* **Enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples [version 1; peer review: 2 approved]** Wellcome Open Research 2021, **6**:99 https://doi.org/10.12688/wellcomeopenres.16756.1

First published: 07 May 2021, **6**:99 https://doi.org/10.12688/wellcomeopenres.16756.1

Introduction

Unbiased sequencing of bacterial, fungal and viral communities has been used to characterize the microbial diversity in nasopharyngeal samples and aid in explaining diseases of unknown aetiologies (Camelo-Castillo *et al.*, 2019; Geliebter *et al.*, 2020; Lu *et al.*, 2020). Unlike targeted sequencing, unbiased sequencing strategies do not require prior knowledge of pathogens present in a sample thus eliminating relative abundance biases inherent to targeted sequencing (Camelo-Castillo *et al.*, 2019; Graf *et al.*, 2016). While bacterial and fungal metagenomics studies make use of the 16S and ITS (internal transcriber spacer) conserved markers for bacterial and fungal community amplification, respectively, viral communities lack conserved markers within viral families (Camelo-Castillo *et al.*, 2019; Conceição-Neto *et al.*, 2015; Geliebter *et al.*, 2020), making random priming also termed as Sequence Independent Single Primer Amplification (SISPA), a promising metagenomics strategy (Djikeng *et al.*, 2008).

SISPA was first developed by Reyes & Kim (1991), and 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). Though SISPA has previously proved effective in metagenomics studies, it results in preferential sequencing of the most abundant nucleic acid material in a nasopharyngeal sample; mainly host and bacteria (Djikeng *et al.*, 2008; Goya *et al.*, 2018). To counter this, methods 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 studies (Chrzastek *et al.*, 2017; Goya *et al.*, 2018; Lewandowski *et al.*, 2020) and deemed effective in enhancing viral read representation and in reducing bacterial and host contamination.

We endeavored to develop a metagenomics protocol for respiratory syncytial virus (RSV); a leading cause of lower respiratory tract infections among children under the age of five. 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). Roughly 48,000-74,500 in-hospital child deaths annually are attributed to RSV infections (Shi *et al.*, 2017). 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). Here, we utilized centrifugal filtration (Thurber *et al.*, 2009), DNase-treatment (Peret *et al.*, 1998) and SISPA (Nguyen *et al.*, 2016), as virus enrichment methods for RSV sequencing using the Oxford Nanopore Technology (ONT) MinION device: an affordable, long read and portable real-time single molecule sequencing device with potential for virus metagenomics studies (Lewandowski *et al.*, 2020; Miani *et al.*, 2020).

Methods

Study samples

Thirty-two nasopharyngeal swabs (NPS) collected between January 2012 and December 2015 from children under the age of five years presenting to the Kilifi County Hospital with clinical symptoms of severe pneumonia were selected for this study using the purposive sampling approach. All NPS samples used in this study were collected upon hospital admission by the clinicians on duty, stored in a universal transport media, kept at 8°C in an ice packed cool box, and transported to KEMRI-Wellcome Trust Research Programme laboratories four hours after collection where they were stored at -80°C. For samples to be included in this study, they had to have been confirmed positive for RSV using the indirect immunoflourescent antibody test (IFAT) and reverse transcription polymerase chain reaction (RT-PCR) method and recorded high viral load as identified by low cycle threshold scores (Ct < 24). In addition, samples included here had to have been sequenced using MiSeq (Illumina) by targeted amplification and full genomes obtained (Agoti *et al.*, 2015; Otieno *et al.*, 2018). We excluded samples with low cycle thresholds (Ct > 24) whose full genomes had not been unravelled before.

Ethical considerations

The study was ethically approved by the Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU #3103). Written informed consent had been collected from all the patient caregivers before using the samples for this study.

Sample processing

Each of the processes for the two protocols is set out in the flow diagram depicted in Figure 1.

Protocol 1: Centrifugal processing approach

Optimization. A set of 12 RSV positive samples were used at first to optimize the centrifugal pre-processing protocol. The 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 extracted from each of the three sample fractions (concentrate, filtrate and pellet from centrifugal processing) obtained from the 12 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 and to enhance the isolation of intact viral RNA, buffering conditions adjusted 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 IFAT. Quantity and quality of the RNA extracts were determined using Qubit RNA HS assay. RT-PCR assays for RSV (Hammitt *et al.*, 2011; Venter *et al.*, 2011) were used to

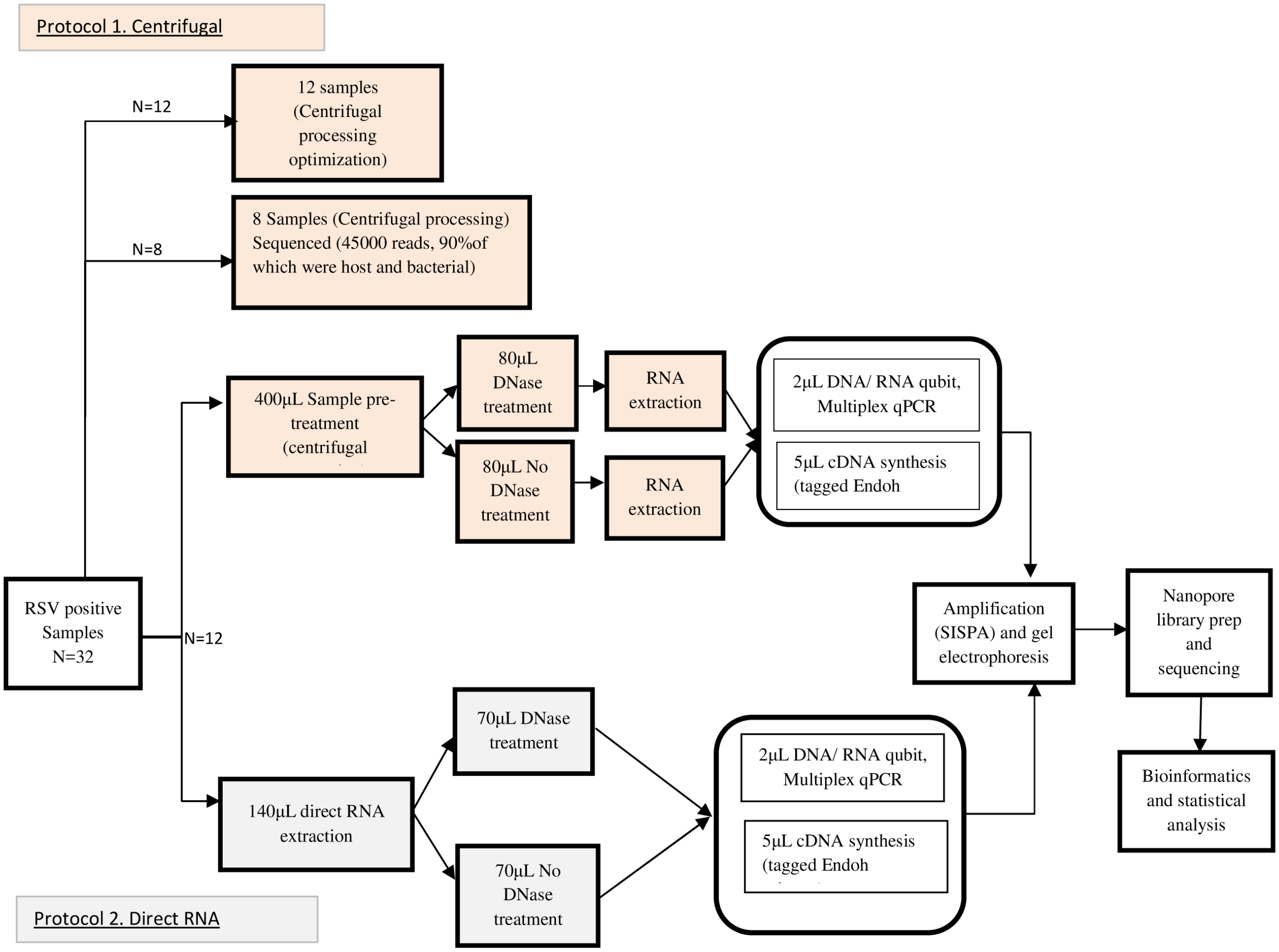
Protocol 1. Centrifugal
N=12
12 samples (Centrifugal processing optimization)
N=8
8 Samples (Centrifugal processing) Sequenced (45000 reads, 90%of which were host and bacterial)
400μL Sample pre-treatment (centrifugal
80μL DNase treatment
80μL No DNase treatment
RNA extraction
RNA extraction
2μL DNA/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

Figure 1. A flow chart representing the experimental setups tested in this study. In total, 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.

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 determined using conventional PCR, with primers that target the V3 and V4 region of the 16S ribosomal RNA (rRNA). Amplified PCR products were visualized in a 2% agarose gel.

Sequencing. All the sample volumes used during the centrifugal processing optimizations were depleted prompting us to select 8 additional RSV positive samples to assess the effectiveness of the approach during sequencing. We took the 8 additional samples through centrifugal processing approach, RNA extraction, cDNA synthesis, SISPA, library preparation and sequencing. However, only 45,000 reads were obtained from the sequencing run, 90% of which were host and bacterial, hindering further analysis. This prompted us to adopt a DNase treatment step after the centrifugal processing. Since the sample volumes for the eight samples also had depleted, we selected 12 additional samples. We used 400μL of each of the samples and took them through centrifugal processing and the resulting concentrate was divided into two equal fractions: the first was DNase treated to remove the genomic DNA concentration from our RNA using TURBO DNase (Thermo Fischer) while the second was not, followed by RNA extraction.

Protocol 2: Direct RNA extraction approach

From the remaining volume of the 12 samples, we used 140μL from each with the direct RNA extraction protocol. This 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. The resulting RNA was divided into two equal fractions, the first was DNase treated to also remove genomic DNA from our RNA of interest using TURBO DNase (Thermo Fischer), while the second was not.

Sequence independent single primer amplification (SISPA)

First-strand cDNA was synthesized in a 20µl reaction from 5µl viral RNA extracts from both protocols using the Superscript III reverse transcriptase kit (Thermo Fischer Scientific), according to the manufacturer's instructions and using 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 (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 RT-PCR assay was used to confirm cDNA formation by excluding the RT step during the PCR cycle because the reverse strand had been generated during the cDNA synthesis step.

The FR20RV primer 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).

Nanopore library preparation and sequencing

We prepared our library 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 15ul elution buffer, 2ul of which was used during quantification using qubit. A library mix containing 12ul of the DNA, 25.5ul of the loading beads and 37.5ul 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.

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). An alternative open source base calling tool is Scrappie (https://github.com/nanoporetech/scrappie), a technology illustrator also developed by ONT community. 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) (https://anaconda.org/bioconda/pycoqc/files?version=2.5.0.23) (Leger & Leonardi, 2019) after which taxonomic classification using Kraken2 (version 2.0.9beta) (https://anaconda.org/bioconda/kraken2/files?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) (https://anaconda.org/bioconda/minimap2/files?sort=ndownloads&sort_order=desc&version=2.17) (Li, 2018) and the resulting SAM files converted to a BAM file, sorted and indexed using SAMtools (version 1.7) (https://anaconda.org/bioconda/samtools/files?version=1.7) (Li *et al.*, 2009). Sorted bam files were visualized using Integrated Genomics Viewer (IGV) (version 11.0.1) (https://software.broadinstitute.org/software/igv/) (Thorvaldsdóttir *et al.*, 2013) to determine the regions they mapped to in the genome. We then searched the Endoh primers against a centroid genome generated from the consensus Illumina reads using Vsearch cluster (version 2.15.0) (http://phoenix.yiz-img.com/wulj2/vsearch) (Shen *et al.*, 2016), and the regions to which the primers mapped located using Seqkit locate (version 0.13.2) (https://anaconda.org/bioconda/seqkit/files?version=0.13.2&type=) (Rognes *et al.*, 2016). All statistical analyses to generate the bar graphs and boxplots presented in the results section were done in R version 3.6 (R Core Team, 2019).

Results

Protocol 1: Centrifugal processing approach optimization

3.1.1: Optimization. After comparing the RNA Qubit scores, cycle threshold (Ct) scores and IFAT images from the concentrate, filtrate and pellet (Waweru *et al.*, 2021), we observed that nucleic acid content in the concentrate and filtrate was undetectable compared to the pellet (Figure 2A). 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 2B). Samples taken through direct RNA extraction as described in Protocol 2 but not treated with DNAse termed as typical RSV positive samples here, had comparable Ct scores to the concentrates (Figure 2B). The concentrate's low RNA qubit scores and reduced viral load implied reduced host contaminants as compared to the pellet, as also confirmed by IFAT, where, 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 3). The differences in the red fluorescence is indicative of differences in the degree of host contamination in the two sample fractions (pellet > concentrate).

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 4A). Alternatives to reduce the contamination entailed adoption of DNase treatment using Turbo DNase or passing the extracted RNA through DNA columns. Of the two alternatives, DNase treatment appeared most effective in reducing the extent of bacterial contamination as compared to the use of DNA columns (Figure 4B). 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 5). This observation prompted us to treat the concentrates with DNase just before RNA extraction, a strategy that was 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).

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

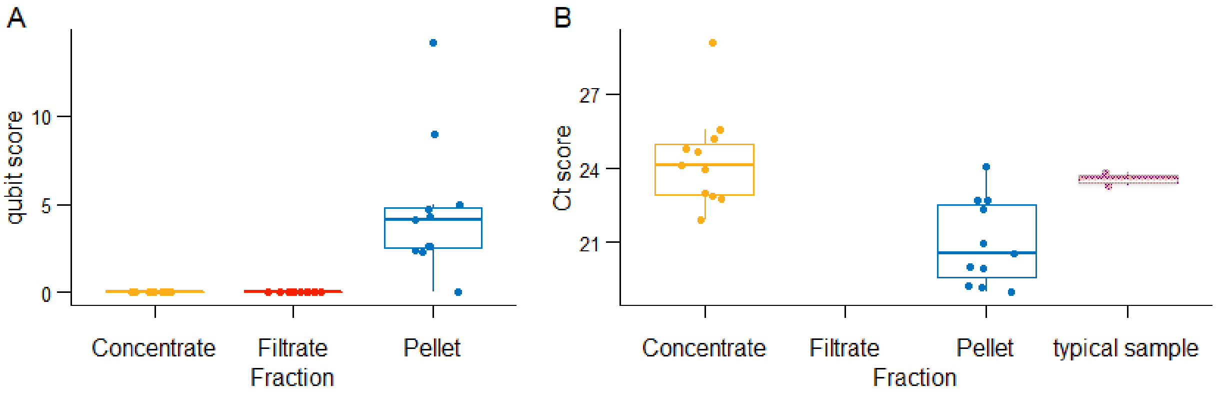

Figure 2. A. Boxplot of the qubit scores from eight centrifugal processed samples against sample fraction. **B.** A boxplot of RSV RT-PCR cycle threshold scores of twelve samples against the sample fractions (concentrate, filtrate and pellet). The colours represent the sample fractions. Filtrate in panel B is undefined, indicating a Ct value >=40.

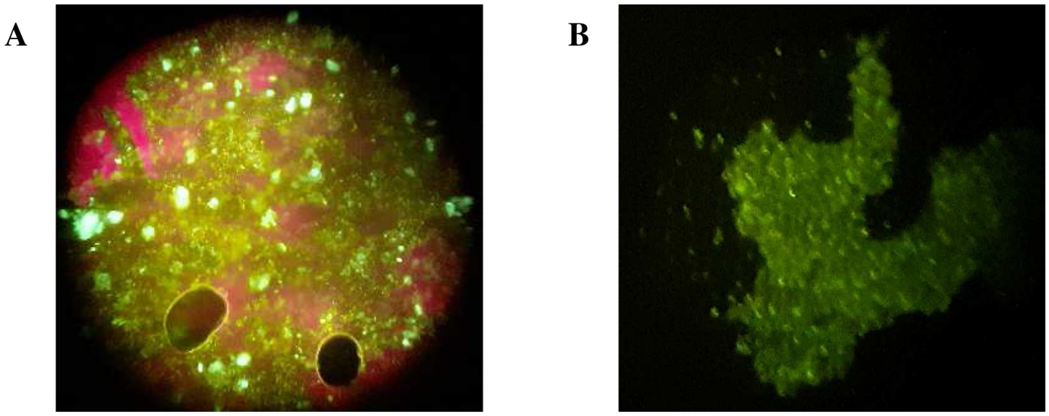

Figure 3. IFAT images of the pellet (**A**) and the concentrate (**B**). Red fluorescence in the pellet represents host cells while green fluorescence in both the pellet and the concentrate represents RSV particles.

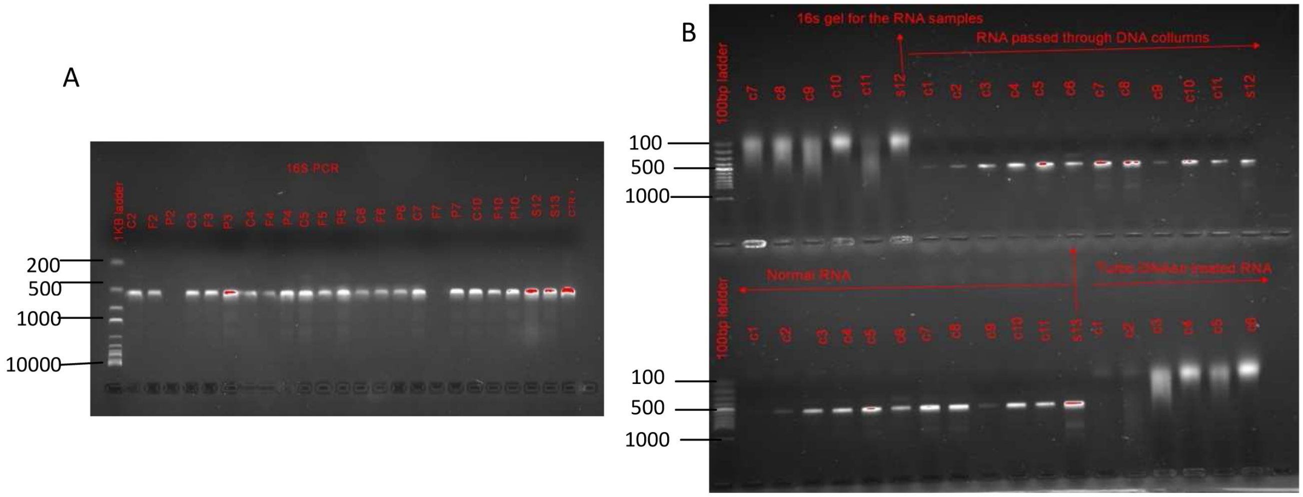

Figure 4. 16s rRNA gel images. Gel image **A** demonstrates bacterial contamination in the various sample fractions. Gel image **B** is an illustration of the impact of DNase treatment and DNA columns in reducing bacterial contamination.

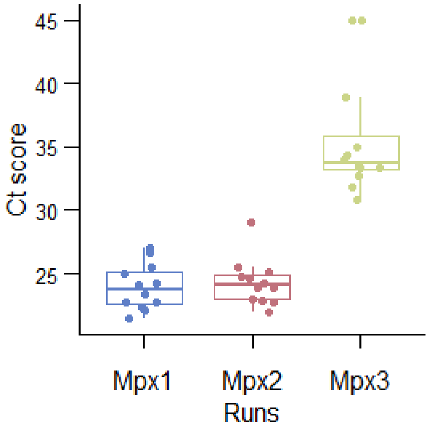

Figure 5. A boxplot of Ct values against runs which demonstrates the effect of DNase treatment in reducing viral load content in the concentrate. Mpx1 represents the Ct values when selecting the samples, Mpx2 the Ct scores from the concentrates after centrifugal processing and Mpx3 the Ct values after treating the concentrates with DNase.

not treated with DNase (Figure 6). The varying lengths in the band sizes demonstrated that the SISPA approach was successful in untargeted amplification of nucleic material present in each sample.

Protocol 1: Centrifugal processing results

We recovered 8.2 million reads from this protocol, 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 7A). 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 8A). No full RSV genome was recovered from this protocol and the sequenced reads mainly mapped to part of the N gene (Figure 9A), 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.

Protocol 2: Direct RNA extraction results

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 7B). 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 8B). 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 (Figure 9B), 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

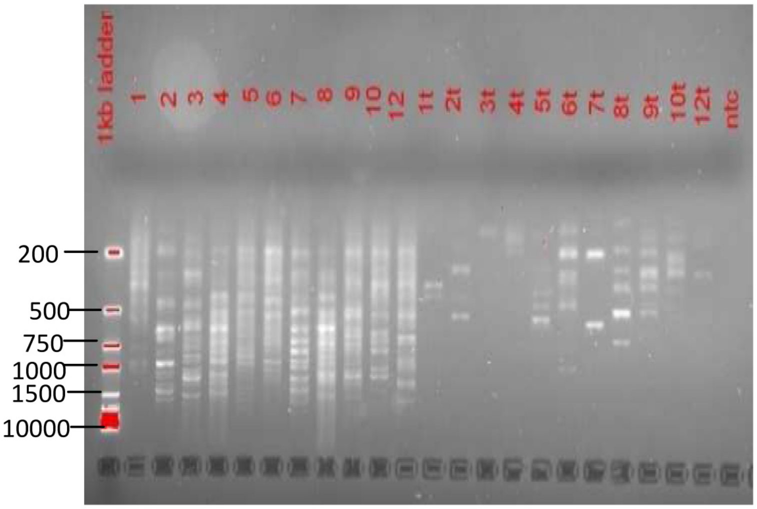

Figure 6. A gel image after performing SISPA. DNase treated sample fractions are denoted with a't' after the sample ID while traces with the sample ID alone denotes the untreated fractions.

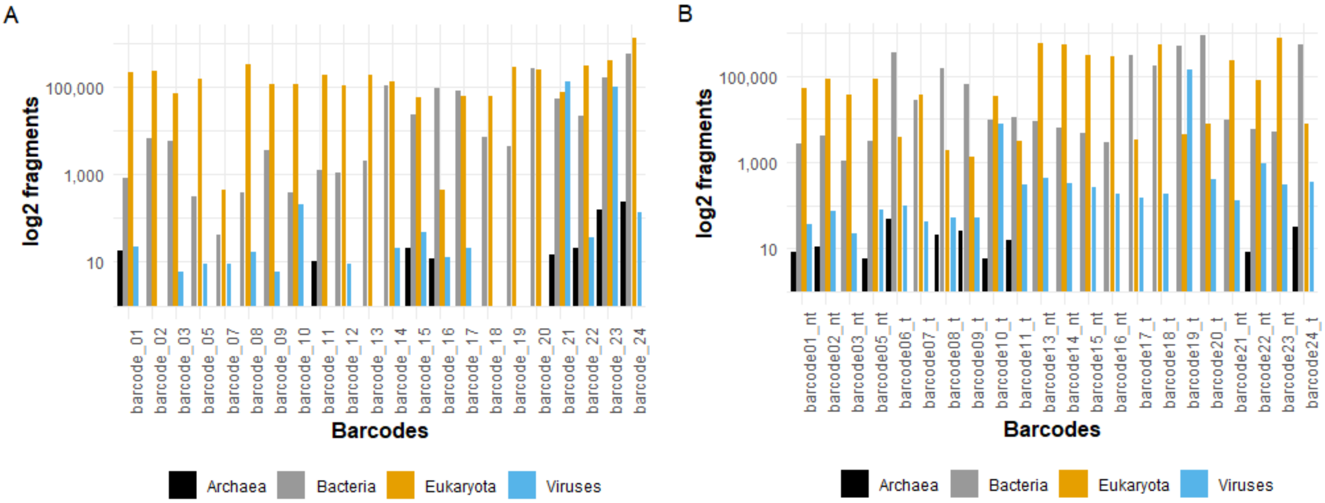

Figure 7. A graphical representation of the domains present in the obtained reads per barcode. Panel **A** represents the domains present in the sample fractions that underwent centrifugal processing (Protocol 1), while panel **B** represents the domains present in the sample fractions that underwent direct RNA extraction (Protocol 2).

mapped to part of the L gene (Figure 9C) with the total number of sequenced bases being roughly 258 spanning from around 12890 bases to 13160 bases.

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 10A. When we compared RSV

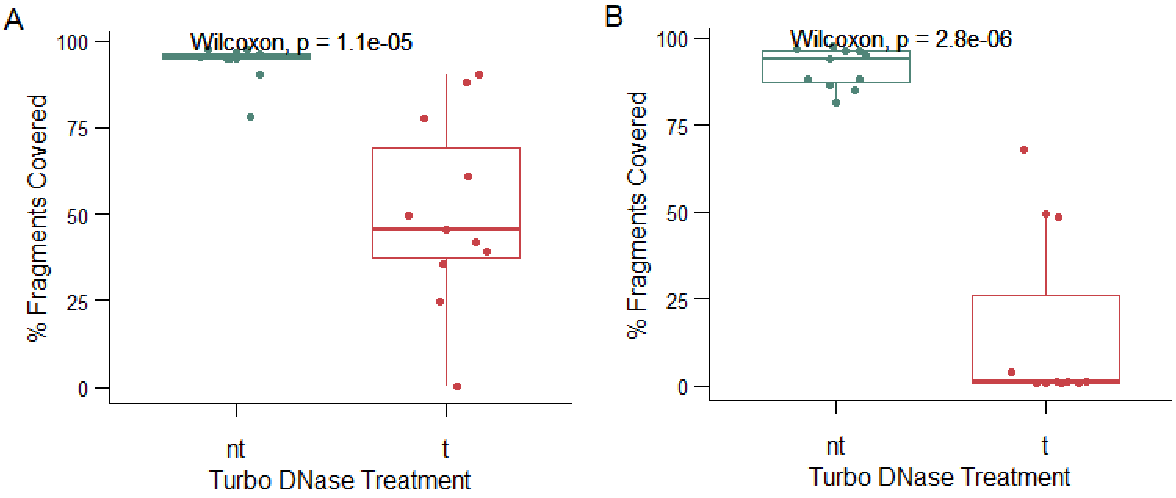

Figure 8. A boxplot of the distribution of host reads between the DNase treated (t) and the non-treated (nt) sample fractions in sample fractions that were processed using Protocol 1 in panel **A** and those that were processed with Protocol 2 in panel **B**.

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 10B).

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 3A). Centrifugal processing recovered freely floating viruses in the concentrate consisting of reduced host cells, although its viral load was reduced. However, centrifugal processing showed little impact in reducing bacterial contamination as confirmed by 16s rRNA PCR (Figure 4A), but DNase treatment was deemed most effective at reducing the extent of bacterial contamination but at the expense of reduced viral content (Figure 5). Despite these processes, we were unable to recover full RSV genome from either protocol.

A comparison of our findings in Figure 2 and Figure 3 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). 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 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 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. 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

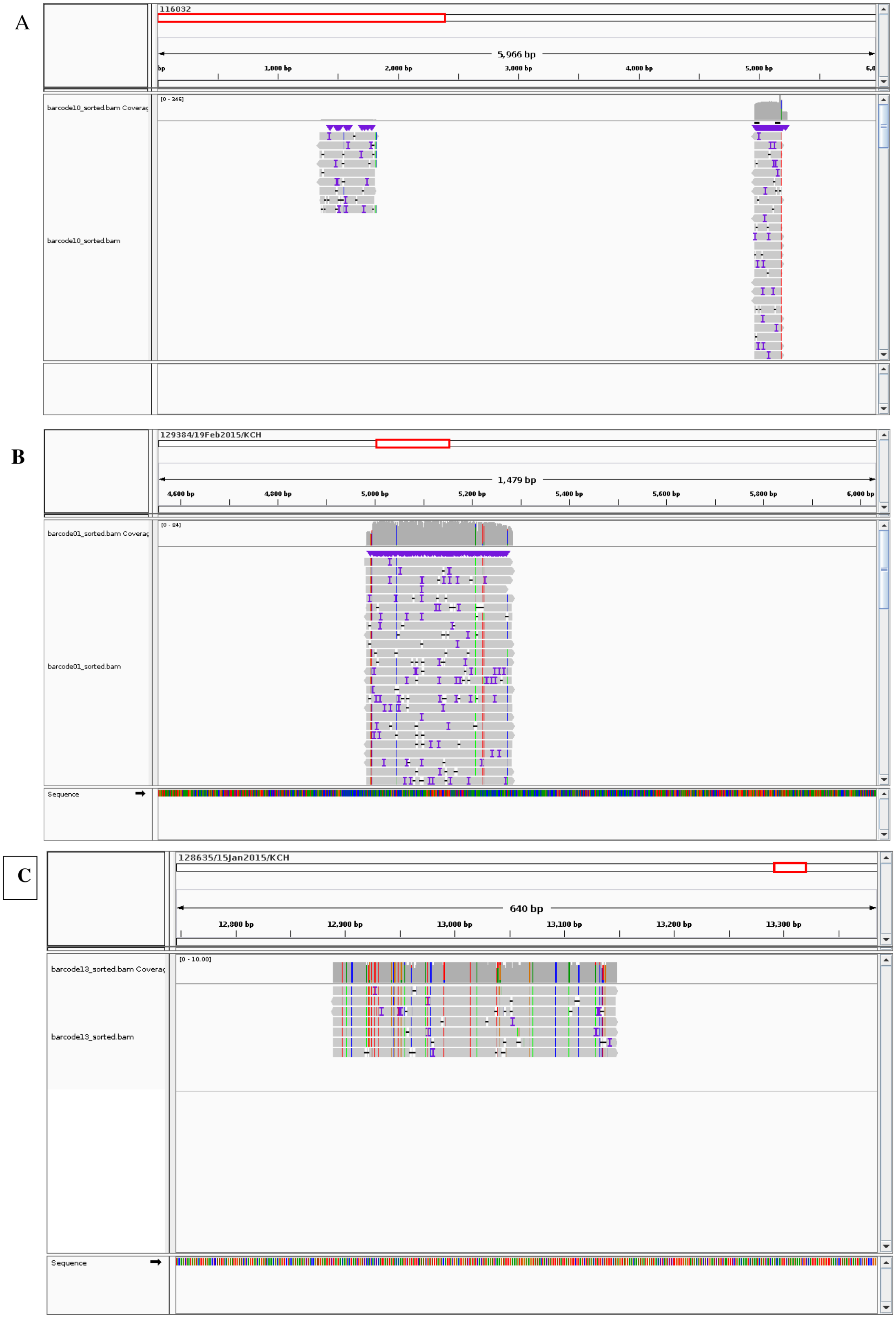

Figure 9. Screen shots of the regions to which RSV reads mapped using Illumina consensus references. (**A**) illustrates the region to which the reads from Protocol 1 mapped: part of the N gene, with some additional reads mapping to part of the G gene while (**B**) and (**C**) illustrates the regions to which the reads from the Protocol 2 mapped: part of the RSV G and L genes respectively

for PCR amplification) assist the random hexamers at the 3' end in annealing during first strand synthesis. In our study, we also speculate that the uneven distribution of the reads across the RSV genome and the variation in the regions that the reads span per run was as a result of part of the tag annealing to the genomic sequence. Of interest in this study was the random primers named 59, 87 and 92 which had some bases on the tag annealing to the centroid sequence and resulting to the over-amplification of the main regions that our reads span (Table 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. A high number of host 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.

In summary, this study demonstrates that although physical virus enrichment approaches such as centrifugal processing help in enriching for the viruses in a viral metagenomics dataset, they cannot be used independently in metagenomics studies. Large amounts of host and bacterial reads are still recovered even after physical enrichment thus making it paramount to include an enzymatic depletion step using DNase, although at the expense of decreasing the virus component. DNase activity should be done after RNA extraction to achieve the best DNase activity in depleting host and bacterial contaminants. During random priming, it is important to consider the length of the random primers being used to avoid preferential amplification biases introduced by using short hexamers in this study. Increasing the length of the random nucleotides

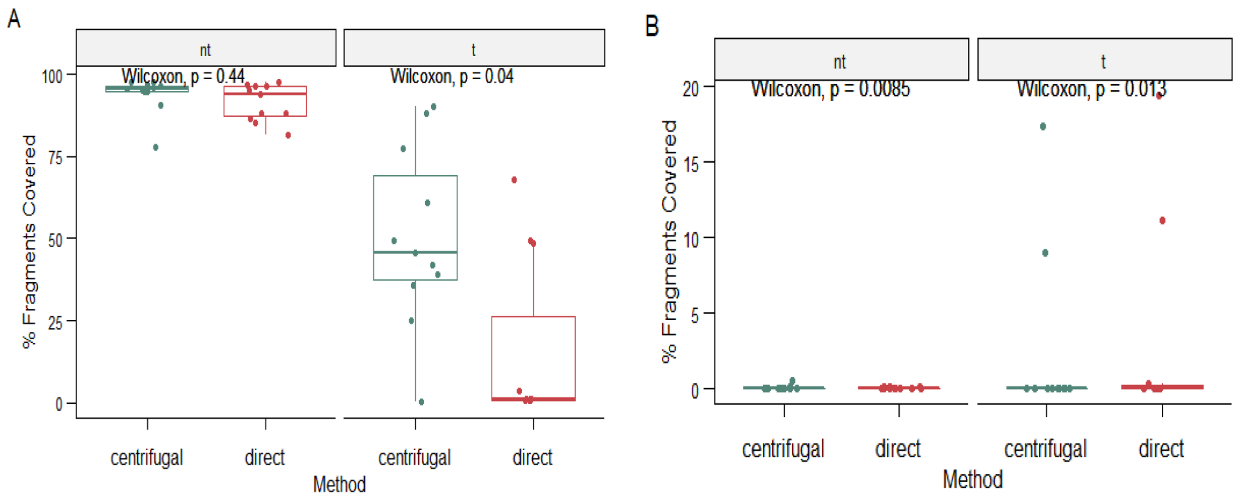

Figure 10. Box plot in panel **A** shows the comparison of proportion of host reads between the two protocols while that in **B** shows the proportion of RSV reads between the treated and untreated sample fractions with those treated with Protocol 1 labelled centrifugal and those processed with Protocol 2 labelled direct.

Table 1. A tabulation of the primers that could have played a role in preferential amplification of some genomic regions of the RSV region in our study.

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

from six hexamers to 9 or 12 in future studies is merited as FR20RV-9mer or FR20RV-12mer have been shown to be more stable and enhanced the chance of their equal distribution across the genome (Rosseel *et al.*, 2013).

Data availability

Harvard Dataverse: Replication Data for: Enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples. https://doi.org/10.7910/DVN/28LOAI (Waweru *et al.*, 2021).

This project contains the following underlying data:

- Data.zip (Raw datasets)
- Boxplots_script.R; taxonomic_analysis.R (The analysis scripts)
- Bacterial_contamination_gel.jpg; bacterial_contamination_reduction_gel.jpg; sispa_gel.jpg (Gel images)

Acknowledgements

We would like to thank the DELTAS Africa initiative along with the other funding partners; Wellcome Trust and the UK government. Special thanks to the KEMRI Wellcome Trust Research Programme, Viral Epidemiology and Control Research Group, field, clinical and laboratory staff, for their contributions to the work reported here. We would also like to thank Nickson Murunga, the data manager for his assistance in curating the data.

References

Agoti CN, Otieno JR, Munywoki PK, *et al.*: **Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing.** *J Virol.* 2015; **89**(7): 3444–3454.
PubMed Abstract | Publisher Full Text | Free Full Text

Allander T, Emerson SU, Engle RE, *et al.*: **A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.** *Proc Natl Acad Sci U S A.* 2001; **98**(20): 11609–11614.
PubMed Abstract | Publisher Full Text | Free Full Text

Camelo-Castillo A, Henares D, Brotons P, *et al.*: **Nasopharyngeal microbiota in children with invasive pneumococcal disease: Identification of bacteria with potential disease-promoting and protective effects.** *Front Microbiol.* 2019; **10**: 11.
PubMed Abstract | Publisher Full Text | Free Full Text

Chrzastek K, Lee DH, Smith D, *et al.*: **Use of Sequence-Independent, Single-Primer-Amplification (SISPA) for rapid detection, identification, and characterization of avian RNA viruses.** *Virology.* 2017; **509**: 159–166.
PubMed Abstract | Publisher Full Text | Free Full Text

Conceição-Neto N, Zeller M, Lefrère H, *et al.*: **Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis.** *Sci Rep.* 2015; **5**: 16532.
PubMed Abstract | Publisher Full Text | Free Full Text

Djikeng A, Halpin R, Kuzmickas R, *et al.*: **Viral genome sequencing by random priming methods.** *BMC Genomics.* 2008; **9**: 5.
PubMed Abstract | Publisher Full Text | Free Full Text

Endoh D, Mizutani T, Kirisawa R, *et al.*: **Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription.** *Nucleic Acids Res.* 2005; **33**(6): e65.
PubMed Abstract | Publisher Full Text | Free Full Text

Englund JA, Anderson LJ, Rhame FS: **Nosocomial transmission of respiratory syncytial virus in immunocompromised adults.** *J Clin Microbiol.* 1991; **29**(1): 115–119.
PubMed Abstract | Publisher Full Text | Free Full Text

Geliebter J, Reyes N, Montes O, *et al.*: ***Staphylococcus aureus* nasal carriage and microbiome composition among medical students from Colombia: a cross-sectional study [version 2; peer review: 2 approved].** *F1000Res.* 2020; **9**: 78.
PubMed Abstract | Publisher Full Text | Free Full Text

Goya S, Valinotto LE, Tittarelli E, *et al.*: **An optimized methodology for whole genome sequencing of RNA respiratory viruses from nasopharyngeal aspirates.** *PLoS One.* 2018; **13**(6): e0199714.
PubMed Abstract | Publisher Full Text | Free Full Text

Graf EH, Simmon KE, Tardif KD, *et al.*: **Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel.** *J Clin Microbiol.* 2016; **54**(4): 1000–1007.
PubMed Abstract | Publisher Full Text | Free Full Text

Hall RJ, Wang J, Todd AK, *et al.*: **Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery.** *J Virol Methods.* 2014; **195**: 194–204.
PubMed Abstract | Publisher Full Text | Free Full Text

Hammitt LL, Kazungu S, Welch S, *et al.*: **Added Value of an Oropharyngeal Swab in Detection of Viruses in Children Hospitalized with Lower Respiratory Tract Infection.** *J Clin Microbiol.* 2011; **49**(6): 2318–2320.
PubMed Abstract | Publisher Full Text | Free Full Text

Lee N, Lui GCY, Wong KT, *et al.*: **High morbidity and mortality in adults hospitalized for respiratory syncytial virus infections.** *Clin Infect Dis.* 2013; **57**(8): 1069–1077.
PubMed Abstract | Publisher Full Text

Leger A, Leonardi T: **pycoQC, interactive quality control for Oxford Nanopore Sequencing.** *J Open Source Softw.* 2019; **4**(34): 1236.
Publisher Full Text

Lewandowska DW, Zagordi O, Geissberger FD, *et al.*: **Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples.** *Microbiome.* 2017; **5**(1): 94.
PubMed Abstract | Publisher Full Text | Free Full Text

Lewandowski K, Xu Y, Pullan ST, *et al.*: **Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples.** *J Clin Microbiol.* 2020; **58**(1): e00963–19.
PubMed Abstract | Publisher Full Text | Free Full Text

Li H: **Minimap2: Pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text

Lu R, Zhao X, Li J, *et al.*: **Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.** *Lancet.* 2020; **395**(10224): 565–574.
PubMed Abstract | Publisher Full Text | Free Full Text

Miani MAT, Baumann C, Barbani MT, *et al.*: **Whole genome sequencing of human enteroviruses from clinical samples by nanopore direct RNA sequencing.** *Viruses.* 2020; **12**(8): 841.
PubMed Abstract | Publisher Full Text | Free Full Text

Mufson MA, Orvell C, Rafnar B, *et al.*: **Two distinct subtypes of human respiratory syncytial virus.** *J Gen Virol.* 1985; **66**(Pt 10): 2111–2124.
PubMed Abstract | Publisher Full Text

Nguyen AT, Tran TT, Hoang VMT, *et al.*: **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.** *Virol J.* 2016; **13**: 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Otieno JR, Kamau EM, Oketch JW, *et al.*: **Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains.** *Virus Evol.* 2018; **4**(2): vey027.
PubMed Abstract | Publisher Full Text | Free Full Text

Peret TC, Golub JA, Anderson LJ, *et al.*: **Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community.** *J Gen Virol.* 1998; **79**(9): 2221–2229.
PubMed Abstract | Publisher Full Text

R Core Team: **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing. 2019.
Reference Source

Reyes GR, Kim JP: **Sequence-independent, single-primer amplification (SISPA) of complex DNA populations.** *Mol Cell Probes.* 1991; **5**(6): 473–481.
PubMed Abstract | **Publisher Full Text**

Rima B, Collins P, Easton A, *et al.*: **ICTV virus taxonomy profile: Pneumoviridae.** *J Gen Virol.* 2017; **98**(12): 2912–2913.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rognes T, Flouri T, Nichols B, *et al.*: **VSEARCH: a versatile open source tool for metagenomics.** *PeerJ.* 2016; **4**: e2584.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rosseel T, Ozhelvaci O, Freimanis G, *et al.*: **Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples.** *J Virol Methods.* 2015; **222**: 72–80.
PubMed Abstract | **Publisher Full Text**

Rosseel T, Van Borm S, Vandenbussche F, *et al.*: **The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing.** *PLoS One.* 2013; **8**(9): e76144.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Shen W, Le S, Li Y, *et al.*: **SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA / Q File Manipulation.** *PLoS One.* 2016; **11**(10): e0163962.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Shi T, McAllister DA, O'Brien KL, *et al.*: **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.** *Lancet.* 2017; **390**(10098): 946–958.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration.** *Brief Bioinform.* 2013; **14**(2): 178–192.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Thurber RV, Haynes M, Breitbart M, *et al.*: **Laboratory procedures to generate viral metagenomes.** *Nat Protoc.* 2009; **4**(4): 470–483.
PubMed Abstract | **Publisher Full Text**

Venter M, Lassauniere R, Kresfelder TL, *et al.*: **Contribution of Common and Recently Described Respiratory Viruses to Annual Hospitalizations in Children in South Africa.** *J Med Virol.* 2011; **83**(8): 1458–1468.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Victoria JG, Kapoor A, Li L, *et al.*: **Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis.** *J Virol.* 2009; **83**(9): 4642–4651.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Waweru JW, de Laurent Z, Kamau E, *et al.*: **"Replication Data for: Enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples"**, Harvard Dataverse, V3. 2021.
http://www.doi.org/10.7910/DVN/28LOAI

Wood DE, Lu J, Langmead B: **Improved metagenomic analysis with Kraken 2.** *BioRxiv.* 2019; 762302.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 11 May 2024

https://doi.org/10.21956/wellcomeopenres.18478.r46402

✓ **Gerald Mboowa**
The African Center of Excellence in Bioinformatics and Data-Intensive Sciences, The Infectious Disease Institute, Makerere University, Kampala, Uganda

The article describes an enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples. Suggestions within sentences are given in square brackets.

Abstract:

Methods

- Assessed two protocols using RSV positive samples. Clearly indicate whether these were fresh or preserved archived samples?Protocol 1 include physical pre-treatment enrichment centrifugal i.e the details such as speed, time, and temperature.

  Concentrates from Protocols 1 & 2 were each divided into two fractions; one was DNase treated while the other was not. How was normalization between DNase treated and untreated samples from the two protocols done?

Conclusions

- 'We attribute the incomplete genome segments to amplification biases resulting from the use of short length random sequence (6 bases) in tagged Endoh primers. Increasing the length of the random nucleotides from six hexamers to nine or 12 in future studies may reduce the coverage biases' Did the authors mean improve capture and coverage of the template?

Introduction:

- 'The virus also causes high morbidity and mortality among immunocompromised individuals and the elderly (Englund et al., 1991' This reference is too outdated.

Methods:

- Study samples 'transported to KEMRI-Wellcome Trust Research Programme laboratories [within] four hours after collection where they were stored at -80°C.'

Ethical considerations:

- 'The study was ethically approved by the Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU #3103)' Include the date of this approval.

- Written informed consent had been collected from all the patient caregivers before using the samples for this study. I think it is important that authors indicate that they sought waiver of consent to use these samples from the IRB because assent and consent would be possible to be obtained from the study participants.

Protocol 1: Centrifugal processing approach:

Optimization:

- 'A set of 12 RSV positive samples were used at first to optimize the centrifugal pre-processing protocol. The 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)...' This should be 'Thermo Fisher Scientific', check this spelling throughout the article. Also include vendor's the city and country, same for Beckman Coulter.

- 'The effectiveness of the pre-processing steps was assessed by performing RNA HS (high sensitivity) qubit, multiplex RT- PCR and IFAT.' Write IFAT in full.

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). An alternative open-source base calling tool is Scrappie (https://github.com/nanoporetech/scrappie), a technology illustrator also developed by ONT community.' Did the authors call the bases using guppy or scrappie? Otherwise, this statement is not necessary if they called bases using Guppy as seen in the subsequent sentence.

- 'All the reads that passed QC (Phred score >7) test were then mapped to the corresponding 12 RSV references generated from Illumina'. Authors should include the reference or source where these 12 genomes can be located.

Results:

Protocol 1: Centrifugal processing approach optimization

- 3.1.1: Optimization. After comparing the RNA Qubit scores, cycle threshold (Ct) scores and IFAT images from the concentrate, filtrate and pellet (Waweru et al., 2021), we observed that nucleic acid content in the concentrate and filtrate was undetectable compared to the pellet (Figure 2A). 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 [quantity] in the pellet relative to the concentrate (Figure 2B). Samples taken through direct RNA extraction as described in Protocol 2 but not treated with DNAse termed as typical RSV positive samples here, had comparable Ct scores to the concentrates (Figure 2B).

Protocol 1: Centrifugal processing results:

- We recovered 8.2 million reads from this protocol, 7.2 million of which passed quality check (QC) [abbreviate this once] with their median read quality being 11.11. Taxonomic classification of all the reads that passed QC from this protocol using Kraken2 indicated.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 31 January 2022

https://doi.org/10.21956/wellcomeopenres.18478.r47101

✓

Martin M. Nyaga
Next Generation Sequencing Unit, Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa
Milton Mogotsi

Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa

In the study “Enrichment approach for unbiased sequencing of respiratory syncytial virus directly from clinical samples” by Waweru *et al*., 2021, the authors attempted to develop a metagenomics protocol for enrichment of respiratory syncytial virus (RSV), which entailed centrifugal filtration, nuclease treatment and random amplification.

This is a very impressive study that was thoroughly performed, and the authors compared their results extensively with the relevant international literature on the subject. Below are several comments and suggestions from the reviewer.

The authors described that their enrichment procedure in “protocol 1” entailed centrifugation and filtration of the RSV positive samples. The resulting pellet following centrifugation and the filtrate and concentrate resulting from filtration were all subjected to RNA extraction.

The authors report that after comparing the RNA Qubit scores, Ct values and immunofluorescence images of the antibody tests, the nucleic acid content of the concentrate and filtrate was undetectable compared to the pellet. This is expected since a larger fraction of the cells (host and bacterial cells), and therefore nucleic material, will be retained in the pellet.

The authors should take into account that centrifugation conditions can also influence the reduction of viral particles and non-viral particles such as bacteria. Centrifugation is expected to pellet larger particles, while viruses remain in solution. Although procedures differ among studies, for a more thorough optimization, authors could also look into testing and comparing different centrifugation speeds and times to determine which conditions yield better results (improved viral recovery).

After subjecting the supernatant to filtration using a 3kD centrifugal filter, the authors stated that the filtrate was RSV negative suggesting no virus loss during filtration. In my understanding, it seems the aim here was to retain the viruses in the “concentrate”, i.e., to prevent virus particles from passing through the filter.

I think it would have been more efficient if the authors opted to collect virus particles in the “filtrate” rather than the “concentrate” being the main fraction of focus since viruses are generally smaller in size and therefore expected to easily pass through the filter. However, with that approach, it is advisable to filters of different particle sizes with indicated pore-sizes (0.22 µm, 0.45 µm, 0.8 µm), preferably 0.22 µm since the RSV virion ranges from 150-250nm in diameter. The majority of the larger eukaryotic and prokaryotic cells are not expected to pass through the filter, thereby enhancing viral recovery. Additionally, authors must also consider subjecting the sample to homogenisation prior to centrifugation. A non-homogenous suspension can result in clogging of the filter. This should, however, be done carefully as it may result in the destruction of virus particles (must be done at the appropriate speed, and without beads).

After sequencing, the authors report that 90% of the 45 000 reads generated belonged to bacteria and host. As a result, the experiment was repeated but this time around the concentrate was subjected to DNase treatment to remove genomic DNA. The authors reported that after analysis 16s rRNA, the bacterial contamination was still very high. The reason could be that the bacterial genome is intact within the bacterial cell, the DNase treatment can only digest the free-floating nucleic material.

As mentioned by the authors, it is crucial to include a DNase treatment to remove host and bacterial contaminants, however, due to the larger sizes of host and bacterial genomes compared to that of viruses, a bulk of host and bacterial reads will be recovered even after enrichment, dominating the sequence data.

Overall, this was a very interesting study, the authors have done excellent work. The results and discussion parts are well described and were able to sufficiently support their findings with literature/previous studies. There are no major issues to be addressed by the authors.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virology; Next Generation Sequencing; Molecular Biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.