Article

RAPID*prep*: A simple, fast protocol for RNA metagenomic sequencing of clinical samples

Rachel L. Tulloch [1,2,#], Karan Kim [1,2,#], Chisha Sikazwe [3,4], Alice Michie [3,4], Rebecca Burrell [2,5], Edward C. Holmes [2], Dominic E. Dwyer [2,6], Philip N. Britton [2,5], Jen Kok [6], and John-Sebastian Eden [1,2,*]

1 Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia; rachel.tulloch@sydney.edu.au, karan.kim@sydney.edu.au, js.eden@sydney.edu.au
2 Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia; Rebecca.Burrell@health.nsw.gov.au, philip.britton@sydney.edu.au, Edward.holmes@sydney.edu.au
3 PathWest Laboratory Medicine WA, Department of Microbiology, Nedlands, WA 6009, Australia; Chisha.Sikazwe@health.wa.gov.au, Alice.Michie@health.nsw.gov.au
4 School of Biomedical Sciences, The University of Western Australia, Crawley, WA 6009, Australia;
5 Departments of Infectious Diseases and Microbiology, The Children's Hospital at Westmead, Westmead, NSW, 2145, Australia;
6 NSW Health Pathology - Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia; dominic.dwyer@sydney.edu.au, jen.kok@health.nsw.gov.au
# Authors contributed equally
* Correspondence: js.eden@sydney.edu.au; Tel.: +61 2 8627 1817

Abstract: Emerging infectious disease threats require rapid response tools to inform diagnostics, treatment, and outbreak control. RNA-based metagenomics offers this; however, most approaches are time-consuming and laborious. Here, we present a simple and fast protocol – the RAPID*prep* assay – with the aim to provide cause agnostic laboratory diagnosis of infection within 24 hours of sample collection by sequencing ribosomal RNA-depleted total RNA. The method is based on the synthesis and amplification of double-stranded cDNA followed by short-read sequencing with minimal handling and clean-up steps to improve processing time. The approach was optimized and applied to a range of clinical respiratory samples to demonstrate diagnostic and quantitative performance. Our results showed robust depletion of both human and microbial rRNA, and library amplification across different sample types, qualities and extraction kits using a single protocol without input nucleic acid quantification or quality assessment. Furthermore, we demonstrate the genomic yield of both known and undiagnosed pathogens with complete genomes recovered in most cases to inform molecular epidemiological investigations and vaccine design. The RAPID*prep* assay is a simple and effective tool, and representative of an important shift towards integration of modern genomic techniques to infectious disease investigations.

Keywords: RNA sequencing; metagenomics; infectious diseases; diagnostics

1. Introduction

Despite major advancements in infectious disease diagnostics and treatment, infections remain a leading cause of death globally. Novel infectious agents and rapid pathogen evolution has led to considerable challenges for traditional diagnostics. At present, accepted methods for disease diagnostics rely on microbial isolation, targeted polymerase chain reaction (PCR), microarray-based assays and serology. As these traditional diagnostic methods are targeted, they are necessarily limited in in their capacity to identify novel pathogens and co-infections. For example, although the reverse transcription polymerase chain reaction (RT-PCR) is both fast and relatively inexpensive, it often fails to detect novel organisms or where genetic variation occurs in the binding region of known pathogens targeted by primers or probe [1]. Furthermore, many disease-causing agents are difficult

 doi:10.20944/preprints202212.0417.v1

to grow using culture-based methods or unculturable *in vitro*; such that these approaches are inherently slow and limited for uncovering novel pathogen diversity. Indeed, such limitations with identifying and characterizing novel pathogens through routine pathology laboratories, as seen with severe acute respiratory coronavirus 2 (SARS-CoV-2) [2], remains one of the greatest global public health challenges. However, the impact is also significant at the level of individual care where delays in diagnosis and treatment can dramatically affect clinical outcomes [3].

Advances in the cost and scale of genomic sequencing have provided important solutions to the challenges of emerging infectious diseases. Unbiased methods such as RNA-based metagenomic next-generation sequencing (RNA-mNGS) offers the capacity to recover and quantify sequences from pathogens with both DNA and RNA genomes [4], describe the microbiome and resistomes [5], and identify coinfections that may be associated with increased morbidity and mortality [6]. RNA-mNGS sequencing offers the unbiased detection of emerging pathogens with the greatest diagnostic potential as it does not require any prior knowledge as to the identity of the causative agent or its genomic sequence (i.e. cause agnostic). The diagnostic capacity of RNA-mNGS was clearly demonstrated during the COVID-19 pandemic and the rapid identification of SARS-CoV-2 in less than a week after realization the infections were likely caused by a novel agent [2]. The emergence of novel SARS-CoV-2 variants throughout the course of the pandemic and associated failures in RT-PCR primers for diagnostic [7] and whole genome sequencing [8] highlight the speed of pathogen evolution and the need for rapid and accurate unbiased sequencing. Whilst RNA-mNGS is indeed powerful there are some limitations when compared to traditional approaches. For example, the diagnostic sensitivity is lower compared to PCR or targeted enrichment due to the relatively low abundance of the viral sequences with respect to the high background from host or microbial nucleic acids. Deeper sequencing may circumvent some of the limitations in sensitivity, although this approach is more costly and often time consuming due to the turnaround times of higher output sequencing platforms and thorough library QC requirements. Ultimately, this highlights the fact that the advancement of mNGS and targeted sequencing into clinical diagnostics will require the development of multiple tools to address multiple needs.

In response to emerging disease threats, there is a need for simple and fast RNA-mNGS approaches to provide rapid and reliable identification of pathogens in a timely manner to inform better treatment and control. Here, we developed and validated a streamlined RNA-mNGS method capable of detecting pathogen RNA from sample collection in less than 24 hours. Furthermore, we developed the approach to utilize readily available reagents to ensure ease of access and reproducibility, particularly following the widespread adoption of amplicon-based whole genome sequencing (WGS) during the COVID-19 pandemic. The RAPID*prep* assay is designed to be simple with minimal handling and QC requirements. However, it is still robust and includes all important steps including genomic DNA (gDNA) removal and ribosomal RNA (rRNA) depletion to boost sensitivity. We developed, optimized and evaluated the utility of this approach on a range of clinical respiratory samples containing both known and unknown pathogens and compared the quantitative performance to quantitative RT-PCR. By providing real-time, high-resolution metagenomic data, the RAPID*prep* assay can inform the diagnosis of common and novel infections to control and monitor outbreaks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

2. Materials and Methods

2.1. Specimens

This study utilized common respiratory samples including nasopharyngeal swabs and aspirates, along with cultured material of A/pdmH1N1 2009 influenza viruses and ZymoBIOMICS Microbial Community Standard (Zymo #D6300). The samples were specifically representative of a range of known viruses (SARS-CoV-2 and respiratory syncytial virus (RSV)), sample qualities (storage in standard viral transport medium (VTM) or Zymo DNA/RNA shield reagent) and extraction platforms (Roche MagNA Pure 96 Viral NA small volume, Zymo Quick-RNA Viral or ZymoBIOMICS DNA/RNA Miniprep Kits). The sample processing followed the manufacturers' recommended protocols. The SARS-CoV-2 and RSV samples were quantified by RT-qPCR targeting the nucleocapsid [9] and nucleoproteins [10], respectively. Briefly, 5 µL of viral extract was converted to cDNA using the Invitrogen SuperScript IV VILO master mix before qPCR using IDT PrimeTime Gene Expression Master Mix with 500nM and 250nM of primers and probe, respectively. Finally, the study also included samples of unknown aetiology collected with parental consent from children with acute respiratory illnesses (mild and severe). This study was approved by the Sydney Children's Hospitals Network (SCHN) human research ethics committee (HREC; approval numbers HREC/18/SCHN/263 & 2020/ETH00837) and the Western Sydney Local Health District HREC (approval numbers LNR/17/WMEAD/128 and SSA/17/WMEAD/129).

2.2. RAPIDprep assay

The assay is divided into the following steps: gDNA removal; rRNA depletion; first strand cDNA synthesis; second strand cDNA synthesis and cleanup; tagmentation; library amplification and cleanup, and sequencing. A simple step-by-step protocol has been made available from: https://www.protocols.io/view/rapidprep-a-simple-fast-protocol-for-rna-metagenom-rm7vzbjkxvx1. The specific reagents and their source have been listed in **Table 1**. For gDNA removal, 8 µL of sample extract (viral RNA, total DNA/RNA or purified RNA) was combined with 1 µL each of Invitrogen 10X ezDNase Buffer and enzyme before 10 min incubation at 37°C, then transferred to ice. For rRNA depletion, 1 µL of Qiagen FastSelect Mix (Equally combined QIAseq FastSelect Human, Mouse, Rat (HMR); Bacterial 5S/16S/23S; and water) was added to the previous reaction before a step-wise incubation from 75°C, 70°C, 65°C, 60°C, 55°C, 37°C and 25°C, holding 2 min at each step, then transferred to ice. For first strand cDNA synthesis, 4 µL of SuperScript IV VILO Master Mix (5X) and 5 µL of water were added to the previous reaction before incubation at 25°C for 10 min, 50°C for 20 min and 85°C for 5 min, then transferred to ice. For second strand cDNA synthesis, 8 µL of Sequenase reaction buffer (5X), 1 µL diluted Sequenase enzyme (Sequenase Dilution Buffer and Sequenase v2.0 DNA Polymerase at a ratio of 2:1), and 11 µL of water are added to the previous reaction before incubation starting at 4°C with a slow ramp (0.1°C/sec) to 37°C for 10 min, then 95°C for 2 min, then transferred to ice. The reaction was then topped up with a further 1 µL of diluted Sequenase enzyme before incubation at 37°C for 30 min. The double stranded cDNA (ds-cDNA) was then purified using Omega Bio-tek Mag-Bind Total Pure NGS cleanup beads with a 0.8X bead to sample ratio and a final elution with 22 µL of Qiagen EB. The purified ds-cDNA (5 µL) was then prepared for sequencing using the Nextera XT DNA Library Preparation Kit with the IDT for Illumina–Nextera DNA unique dual indexing kit as per manufacturer's instructions except for the following modifications: 16X cycles was used for library amplification followed by purification with Omega Bio-tek Mag-Bind Total Pure NGS cleanup beads using a 0.8X bead to sample ratio, and a final elution with 32 µL of Qiagen EB. Library QC was then performed using a High Sensitivity D1000 ScreenTape on the Agilent 2200 TapeStation system with gating of the fragments between 200 bp and 700 bp, before final dilution to 0.1 nM for loading and sequencing on an Illumina iSeq (paired-end 150 bp sequencing). As the minimal sequencing yield for each library should be 1 million paired reads, 1-4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

libraries can be multiplexed per iSeq run. For our large, batched run, we prepared and indexed 39 samples and one no template control (NTC). These were pooled evenly and sequenced on an Illumina NovaSeq SP 300 cycle lane generating at least 4 million paired reads per library (NCBI SRA SRR22726217 - SRR22726256).

Table 1 – Reagents used for RAPID*prep* assay

Reagent	Supplier	Catalogue
Invitrogen ezDNase Enzyme	Thermo Fisher	11766051
QIAseq FastSelect-rRNA HMR	Qiagen	334385
QIAseq FastSelect–5S/16S/23S	Qiagen	335921
Invitrogen SuperScript IV VILO Master	Thermo Fisher	11756050
Sequenase Version 2.0 DNA Polymerase	Thermo Fisher	70775Y200UN
Nextera XT DNA Library Preparation Kit	Illumina	FC-131-1096
IDT® for Illumina DNA/RNA UD Indexes	Illumina	20027213
Mag-Bind® TotalPure NGS	Omega Biotek	M1378-01
iSeq 100 i1 Reagent v2 (300-cycle)	Illumina	20031371

2.3. Optimization

We explored three aspects of optimizing the RAPID*prep* assay that were focused on simplifying the protocol to improve turnaround time and determining the optimal yield of the final libraries. These included testing: 1) rRNA depletion performance; 2) ds-cDNA yield; and 3) number of cycles for library amplification. For the rRNA experiments, the standard pre-cDNA hybridization step (as above) was compared against a simplified approach spiking 1 µL of depletion oligos (FastSelect mix) directly into the first strand cDNA reaction with the relative amount of rRNA following sequencing measured as output. For the ds-cDNA yield experiments, the standard Sequenase two-step reaction was compared to a single-step reaction combining the total amount of Sequenase enzyme (2 µL) and reaction time (40 min extension at 37°C). The output was measured by Agilent TapeStation to compare library yield of each approach. For the library amplification experiments, we titrated the number of indexing PCR cycles between 14X to 20X in two cycle steps. The output was also measured by Agilent TapeStation to compare library yield of the different cycles; however, the libraries were also sequenced to determine the sequence read duplication rate. For all the experiments, the same three respiratory sample extracts (clinical nasopharyngeal swabs collected in Zymo DNA/RNA shield and extracted with both the Zymo Viral RNA and ZymoBIOMICS DNA/RNA Miniprep kits used along with an NTC. Samples were run in duplicate with the mean and standard deviation values reported.

2.4. Severe acute respiratory infections in children cohort

A subset of the samples – the severe acute respiratory infections (SARI) in hospitalized children – had been previously sequenced using a commercial RNA sequencing assay (NCBI SRA SRR22838411 – SRR22838442). These data were used to compare against libraries made using the RAPID*prep* assay (**Supp Table 1**). Briefly, these RNA samples were prepared for sequencing using the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian with unique dual indexes (Takara, Japan) as per the manufacturer's instructions and sequenced on an Illumina NovaSeq with at least 40 million paired reads for library.

2.4. Bioinformatic analysis of RNA-mNGS data

Raw sequence reads were first quality trimmed and filtered using FastP v0.19.6 [11] with default parameters except the read length filter was 50 bp. The trimmed reads were then mapped to the human genome using STAR-aligner v2.6.1b [12], followed by Burrows-Wheeler Aligner (BWA) v0.7.17 [13] to ensure complete human sequence removal.

 doi:10.20944/preprints202212.0417.v1

The trimmed, human and non-human reads were then filtered into rRNA and non-rRNA and quantified using SortMeRNA v2.1b [14] before, the trimmed, non-human, non-rRNA reads were then *de novo* assembled using Megahit v1.1.3 [15] before annotation using blast+ v2.11 [16] and diamond v2.0.11 [17] against the NCBI GenBank database. A read-based analysis was also performed of the trimmed, non-human, non-rRNA datasets by mapping against the microbial taxonomic database in MetaPhlAn v3.0.13 [18]. Comparative analysis of microbial abundance was performance using calculated z-scores in R v3.4.3. Final viral read counts were also determined by alignment of trimmed, non-human, non-rRNA reads to the *de novo* assembled contigs and/or known viral reference genomes for the SARS-CoV-2, influenza virus and RSV samples using BBMap v 37.98 [19]. Maximum likelihood trees for individual viruses were estimated using PhyML v2.2.4 [20] with the GTR + Gamma substitution model and 1000 bootstrap replicates.

3. Results & Discussion

The aim of this study was to develop a simple yet robust workflow for RNA-mNGS of clinical samples that can provide a cause agnostic laboratory diagnosis in less than 24 hours. The RAPID*prep* assay is comparable to other meta-transcriptomic assays in that it aims to unbiasedly sequence the non-host, non-rRNA RNA for pathogen detection and quantification. However, it is unique in its simplicity, with reduced handling and a uniform protocol for sequencing across a range of sample types, qualities, and quantities. The first steps aim to remove gDNA and rRNA to improve target sensitivity before random double-stranded cDNA synthesis and amplification. Minimising the processing and handling was important to ensure the entire protocol could easily be completed in less than 6 hours. This was primarily achieved by the basic assay design with most steps being additive and performed in a single tube without the need for reaction clean-ups (bead-based purifications). However, we explored this further by attempting to simplify the rRNA-depletion and ds-cDNA synthesis steps and optimizing the library amplification yield through a range of experiments using three representative respiratory samples (RESP01-RESP03).

3.1. Optimization of the RAPIDprep assay

3.1.1. rRNA depletion

rRNA is the most abundant component of total RNA isolated from eukaryotic and microbial cells [21]. While the importance of rRNA-depleted libraries for improved coverage of mRNA for transcriptome sequencing is recognized [22-24], it is particularly important for the identification and genome recovery of viral pathogens with RNA genomes such as coronaviruses, influenza viruses and paramyxoviruses that are emerging disease threats. The FastSelect reagent blocks transcription with proprietary probes that bind to mammalian and microbial rRNA. As such, it does not necessarily deplete rRNA but rather prevents its synthesis during cDNA steps. To increase the speed of the protocol, we sought to determine if the FastSelect probes could be added directly to the first strand cDNA synthesis step without the need for pre-cDNA hybridization step that added approximately 30 mins of reaction and handling time. The relative abundance of rRNA in the final sequenced library between the two approaches was compared **(Figure 1A)**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

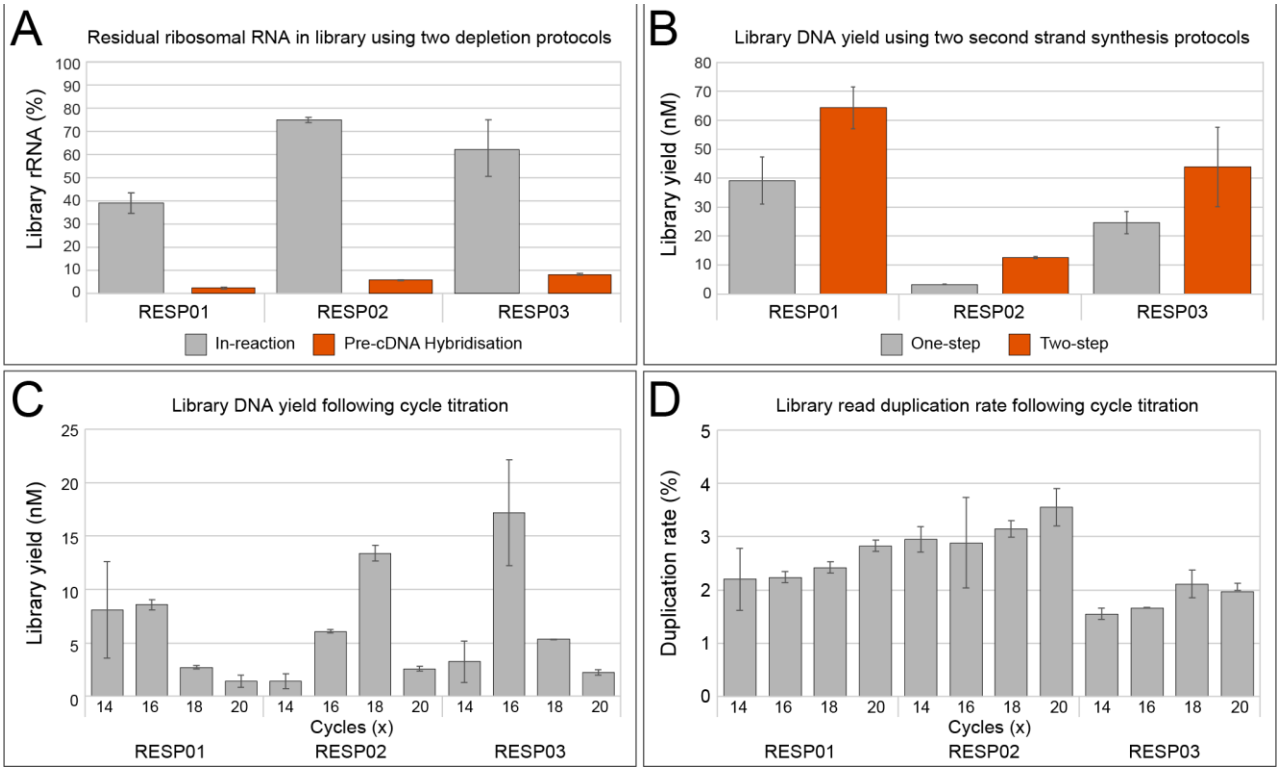

Figure 1 – RAPID*prep* optimization experiments. All results here are derived from the same sample extracts (RESP01-RESP03) run in duplicate and presented as mean values and error as standard deviation (SD); (**a).** The shaded bars are representative of the percentage of residual rRNA reads in the library following rRNA depletion with either an in-reaction cDNA synthesis method (grey) or a pre-cDNA hybridization approach (orange). The bars are clustered with respect to the sample they are derived from, labelled on the X axis. (**b**) A comparison in total library yield, in nanomolar generated using Tapestation values, following a parallel experiment with a one-step and two-step second strand synthesis step using the Sequenase enzyme. The grey and orange shaded bars are representative of the one-step and two-step protocols respectively. (**c**) Grey shaded bars represent total library yield of each sample, under different library amplification cycling conditions. The X-axis is marked with the number of amplification cycles and is sub-grouped by source sample. (**d**) The duplication rate of reads generated in the final libraries following cycle titration, the number of cycles for each sample is indicated on the X axis, and is sub-grouped by source sample.

A clear trend was observed across the three samples, where the final libraries made using a dedicated pre-cDNA hybridization step had a dramatically smaller proportion of rRNA in the final library yield (**Figure 1A**). To further investigate the effect of rRNA depletion method on the final library composition, we examined the residual rRNA by kingdom, and their relative abundance was once more compared across different samples and methods (**Table 2**). Similarly, all classes of rRNA were better depleted utilizing the pre-cDNA synthesis hybridization protocol with the residual rRNA not exceeding 4.0% (RESP03 bacterial 23S rRNA), and in most cases less than 1.0%, while the in-reaction approach had up to 22.0% residual rRNA (RESP02 eukaryotic 18S rRNA). FastSelect is a simple and effective solution for the removal of rRNA, although the probes clearly require dedicated steps to hybridize efficiently and, in this case, must occur prior to first strand cDNA synthesis. Whilst performing the rRNA step prior to cDNA hybridization increases the total protocol time slightly, the greatly improved rRNA depletion outweighs this and improves the sensitivity of the overall assay for better pathogen detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

Table 2 - Relative abundance of archaeal, bacterial, and eukaryotic rRNA using two different approaches.

	In-reaction						Pre-cDNA hybridisation					
rRNA	**RESP01**		**RESP02**		**RESP03**		**RESP01**		**RESP02**		**RESP03**	
Archaeal:16S	3.5%	2.9%	7.5%	7.3%	6.1%	4.0%	0.0%	0.0%	0.1%	0.1%	0.2%	0.1%
Archaeal:23S	10.9%	9.5%	19.2%	19.9%	22.0%	16.2%	0.1%	0.1%	0.8%	0.7%	1.6%	1.4%
Bacterial:5S	0.7%	0.8%	0.2%	0.2%	0.4%	0.6%	1.0%	1.3%	0.9%	0.8%	1.4%	1.3%
Bacterial:16S	0.7%	0.6%	2.1%	2.1%	3.0%	2.2%	0.1%	0.0%	0.2%	0.2%	0.2%	0.1%
Bacterial:23S	3.5%	3.1%	10.1%	10.7%	20.7%	18.1%	0.2%	0.3%	2.2%	2.1%	4.0%	3.9%
Eukaryotic:5.8S	0.5%	0.5%	1.1%	1.1%	0.9%	0.8%	0.0%	0.0%	0.1%	0.1%	0.1%	0.1%
Eukaryotic:18S	14.0%	11.9%	22.0%	22.1%	11.2%	7.7%	0.3%	0.4%	0.9%	0.8%	0.6%	0.5%
Eukaryotic:28S	8.5%	6.7%	12.6%	13.0%	6.6%	4.4%	0.2%	0.2%	0.7%	0.7%	0.4%	0.4%

Library rRNA	0.0%	5.0%	10.0%	15.0%	20.0%	25.0%

3.1.2. Double stranded cDNA synthesis

As the purpose of this method was to be robust yet rapid as possible, we next explored the feasibility of reducing the second strand synthesis of cDNA from a two-step process to one-step. Sequenase enzyme is a modified bacteriophage T7 DNA polymerase that lacks 3'→5' exonuclease activity with improved processivity and speed [25]. The standard reaction will occur in two steps, where initial double stranded cDNA from the first strand reaction will be produced from randomly primed single-stranded DNA template [26]. This will be followed by addition of further enzyme for final extension of the ds-cDNA products. We sought to compare this two-step approach to a simplified one-step protocol where the extension time and enzyme concentration during the first part was increased to match the overall two-step approach. Not only would this shorten the workflow by up to 15 mins, but it would also help minimising the handling and potential opportunities for contamination. However, across the three test samples we saw lower total library yields using the simplified one-step method **(Figure 1B).** While the yields of the one-step protocol were sufficient for sequencing, the desire for a single uniform protocol across varying sample types and qualities favoured here the approach with the greatest yield. Therefore, like the rRNA depletion optimization and despite a small trade-off in time and handling, our final assay utilized the two-step protocol that gave greatest performance.

3.1.3. Library amplification

It is widely accepted that library preparation can introduce systematic bias to the characterization and representation of microbial communities in a sample [27-30]. Bias is most readily introduced during the library amplification stage. Some studies argue that the simplest means to mitigate this bias during PCR is to avoid library amplification all together. For the RAPID*prep* assay, a PCR-free library protocol would likely be unattainable due to the low concentrations of input total nucleic acid particularly from swabs and cell-free viral samples. Illumina Nextera XT is a commercially available library preparation kit that uses a transposase-based *tagmenention* reaction to fragment and add adapters onto template dsDNA [29]. Following this, limited cycle PCR is used to barcode and complete index adapters before sequencing [31]. While other library preparation kits would be compatible here, Nextera XT is simple and fast, and therefore an ideal partner for the RAPID*prep* assay. Furthermore, Nextera XT is widely used for amplicon-based WGS of viral pathogens [32-34], and offers potentially greater adoption compared to other library preparation kits. As low input total nucleic acid necessitates PCR amplification, we sought to identify an optimal cycle number which afforded greatest library yields whilst limiting potential amplification bias. A titration experiment was therefore performed with final library yield measured using DNA molarity as determined by Agilent Tapestation (**Figure**

 doi:10.20944/preprints202212.0417.v1

1C), and sample bias measured by calculating the read duplication rate of the sequenced libraries (**Figure 1D**). Percentage duplication rate is an ideal proxy for sequence bias as the redundant reads are typically introduced during library amplification PCR. Furthermore, duplicate reads will limit the entropy of the final dataset and potentially introduce bias for both sample identification and particularly quantification [35].

Here, we explored the optimal cycle conditions from 14X to 20X PCR cycles (**Figures 1C & 1D**). For two samples (RESP01 & RESP03), the DNA yield was greatest at 16X cycles, while for one (RESP02) it was 18X cycles. (**Figure 1C**) The apparent trend showing reduced yield at cycles 18X and above was due to over-amplification induced artefacts with fragments exceeding the upper range (1000 bp) of the Tapestation analyser (data not shown). While such large fragments are likely to be sequenced when denatured, they present a challenge for quantifying and final library QC. Similarly, the percentage duplication data demonstrated a clear increase in overamplification of libraries at cycles 18X and above for all three samples (**Figure 1D**)**.** A high PCR duplication rate cannot simply be overcome using deeper sequencing methods, this is a fundamental issue that can only be mitigated at the time of library preparation. Indeed, read duplicates can only be identified post sequencing, for this reason it is advantageous to choose PCR cycles that both maximises yield while minimising duplication rates. Here, this optimum seemed to be 16X PCR cycles, which was used for the final RAPID*prep* assay.

3.2. Application of the RAPIDprep assay to a panel of respiratory samples

To provide a broad assessment of RAPID*prep* performance, we selected a range of respiratory samples and control material (n=40) for a combined, proof-of-concept run, using the optimized protocol (**Table 3**). These samples varied in microbial composition, sample collection, quality, and extraction, and were designed to reflect a broad snapshot of real-world sampling performance. Libraries *RAPID01-12* were derived from SARS-CoV-2 positive respiratory swabs collected and processed within one week following collection from a household transmission study. *RAPID13* & *-14* were viral stocks collected from A/pdmH1N1 2009 influenza virus infected cells. Further known positive samples were prepared as libraries *RAPID25-32* that were RSV-positive and extracted through a diagnostic pathology service using a high-throughput bead-based platform (Roche MagNA Pure). *RAPID15* & *-16* were high-quality cultured material containing standard amounts of known bacteria and fungi – the ZymoBIOMICS Microbial Community Standards – and were process controls for a study investigating unknown SARI in hospitalized children. A subset of these SARI samples (libraries *RAPID17-24*) was included here as they represented residual, and often highly degraded, specimens collected through routine diagnostic services, and had existing deep sequencing data for comparison. Finally, high-quality respiratory samples of unknown aetiology collected in Zymo DNA/RNA shield were used (*RAPID33-39*) along with a NTC reaction (*RAPID40*). As per the protocol, no specific sample QC was performed and 8 µL of neat extract (total NA, Viral RNA or RNA) was used as input using the protocol as per section 2.2. All forty samples produced libraries with a mean yield of 5.0 nM (range: 0.7 nM to 8.7 nM) and were pooled and sequenced on a single Illumina NovaSeq SP lane (**Table 3**). The mean sequence yield per library was 15,577,978 reads (range: 9,453,054 to 28,187,178 reads). Each library was then analyzed for low-quality, human and rRNA content before taxonomic assignment and quantification using a standard mNGS pipeline.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

Table 3 - RAPID*prep* sample summary overview table

Library	Group	Virus	Type	Extraction method	Library Yield (nM)	Data output (reads)
RAPID01	COVID-19	SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	5.10	16,810,302
RAPID02		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	6.40	11,620,222
RAPID03		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	1.30	18,322,864
RAPID04		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	2.20	12,707,642
RAPID05		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	1.70	15,327,662
RAPID06		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	6.50	15,271,010
RAPID07		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	2.80	11,147,058
RAPID08		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	5.30	9,453,054
RAPID09		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	4.90	17,326,098
RAPID10		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	6.10	15,531,486
RAPID11		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	1.90	10,903,012
RAPID12		SARS-CoV-2	Nasopharyngeal swab	Zymo Quick-RNA Viral	6.10	12,670,186
RAPID13	Influenza A	pdmH1N1	Viral culture	Zymo Quick-RNA Viral	7.00	15,200,408
RAPID14		pdmH1N1	Viral culture	Zymo Quick-RNA Viral	3.00	12,405,816
RAPID15	Mock community	N/A	Mixed culture	ZymoBIOMICS DNA/RNA Miniprep	6.50	13,541,676
RAPID16		N/A	Mixed culture	ZymoBIOMICS DNA/RNA Miniprep	6.80	12,427,300
RAPID17	Kids SARI	Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	5.60	16,321,598
RAPID18		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	0.70	17,340,092
RAPID19		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	0.70	15,464,422
RAPID20		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	5.90	16,563,150
RAPID21		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	0.80	24,661,800
RAPID22		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	5.90	14,490,708
RAPID23		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	1.20	28,187,178
RAPID24		Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	1.80	19,042,138
RAPID25	RSV	RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	8.70	15,287,586
RAPID26		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	8.30	19,473,790
RAPID27		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	8.10	16,302,456
RAPID28		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	7.70	14,524,914
RAPID29		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	5.60	17,923,728
RAPID30		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	6.00	13,466,538
RAPID31		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	5.70	12,192,164
RAPID32		RSV	Nasopharyngeal swab	Roche MagNA Pure 96 Viral NA	6.80	17,199,224
RAPID33	Kids unknown	Unknown	Nasopharyngeal aspirate	ZymoBIOMICS DNA/RNA Miniprep	6.20	14,839,480
RAPID34		Unknown	Nasopharyngeal swab	Zymo Quick-RNA Viral	6.10	14,806,470
RAPID35		Unknown	Nasopharyngeal swab	Zymo Quick-RNA Viral	6.90	12,576,818
RAPID36		Unknown	Nasopharyngeal aspirate	Zymo Quick-RNA Viral	5.40	13,418,996
RAPID37		Unknown	Vomitus	Zymo Quick-RNA Viral	6.00	9,862,060
RAPID38		Unknown	Nasopharyngeal swab	Zymo Quick-RNA Viral	7.50	11,384,408
RAPID39		Unknown	Nasopharyngeal swab	Zymo Quick-RNA Viral	7.10	20,986,506
RAPID40	NTC	NTC	Water	N/A	1.50	26,137,100

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

The sequence reads for each library were filtered into five categories: low quality, human rRNA, human non-rRNA, non-human rRNA and non-human non-rRNA, and the relative proportions compared across the sample set and groups (**Figure 2**)**.** Low quality reads were of highest abundance in the samples from the SARI cohort (*RAPID17-24*) where the mean low-quality reads were 23.4% of total libraries. These were 'rescued' diagnostic specimens that had gone through multiple freeze-thaws in the pathology laboratory, with many likely degraded. Suboptimal sample quality also likely explains some of the variation in low-quality reads in the SARS-CoV-2 cohort (*RAPID01-12*), with delayed transport to the laboratory following collection at home. Low quality input samples generally result in an increased amount of homopolymers, and short fragment reads, and are hallmarks of endpoint sample degradation that can be used as a measure of sample quality [36]. While sample quality is an issue, low-biomass samples would be expected to have higher levels of low-complexity reads that will be removed during the initial QC steps such as the NTC library (*RAPID40*).

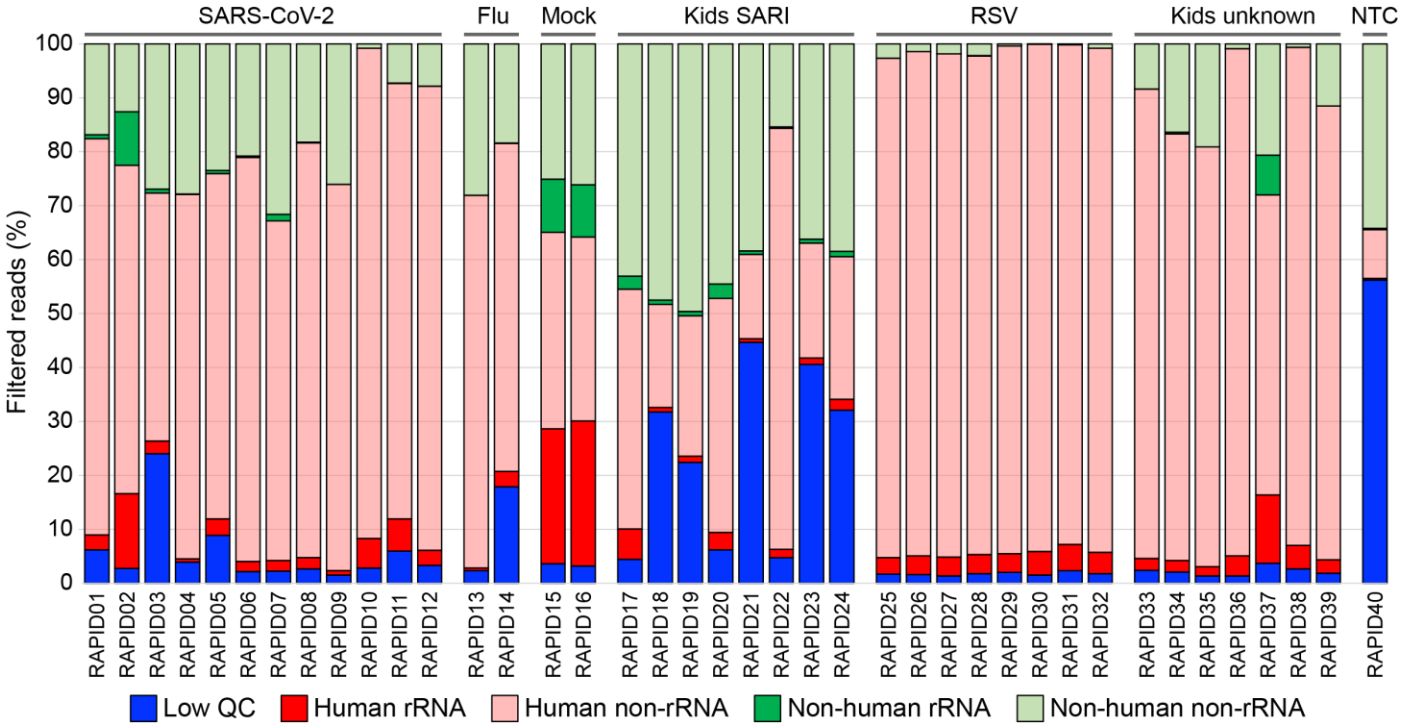

Figure 2 – Filtered read distribution and classification across forty RAPID*prep* libraries. The sequence reads were classified into five categories: low quality reads (blue), human rRNA reads (red), human non-rRNA (pink), non-human rRNA reads (green) and non-human non-rRNA reads (light green). Low quality, human rRNA, human non-rRNA and non-human rRNA are excluded from downstream analysis, and the non-human non-rRNA reads the sole target reads for pathogen detection. Relative distribution is calculated using the total number of reads for the individual library, divided by the number of reads mapping to the relative category and converted into a percentage by multiplying the value by 100. The results are ordered by library number and grouped by sample type.

The proportion of reads that mapped to the human genome had a mean value of 70.8% across all the libraries and sample type; however, this varied widely (range: 20.0% to 98.4%). Sequencing data from host-associated microbes may contain host cells, usually acquired at the time of sampling [22]. Several factors can influence the abundance of host material at the point of collection, including collection route, sampling device (such as flocked vs non-flocked swabs), technique and collector experience [24]. Furthermore, as the human genome is significantly larger than microbial and viral genomes, host-derived nucleic can easily be over-represented, even if in relatively small amounts. On average, human reads were the most common read assignment across all samples, except for the SARI cohort (*RAPID17-24*) and NTC (*RAPID40*). The collection method of samples in the SARI cohort varied, and many of these samples were acquired from sources other than nasopharyngeal swabs including aspirates due to age and hospitalization, which likely

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

contributed to the variable yields. Contamination of sequencing data by human nucleic acid can readily occur, with putative sources including adjacent samples or from the collector [37]. The relative abundance of human reads can also be affected by sample processing steps including the extraction kit used. For example, the highest levels of human RNA were found in the RSV-positive respiratory swabs (**Figure 2**) that were all processed using the Roche MagNA Pure 96 Viral NA small volume kit. In contrast to the other Zymo extraction kits, this platform includes an initial Proteinase K digestion that likely increases the relative yield of human DNA and RNA [38]. High levels of human sequences not only limit the sensitivity of target non-human, non-rRNA but also increases the risk of residual human DNA being deposited in public archives, which presents an ethical concern and potentially indefinable information. Care must be taken at the point of sampling or in processing to reduce the amount of unnecessary human tissue acquired.

Overall, residual rRNA was limited across the sequence libraries highlighting the performance of the RAPID*prep* assay where the mean non-human rRNA was only 1.3% (**Figure 2**). However, despite our extensive optimization experiments (**Figure 1A**), rRNA depletion remained incomplete in some samples such as *RAPID02*, *RAPID15* and *RAPID16* that contained between 9.7% and 9.9% non-human rRNA reads of each library. One reason could be the limited microbial diversity captured by rRNA depletion probes such as Qiagen FastSelect and equivalent products. In both *RAPID15* and *RAPID16*, taxonomic assignment of the residual rRNA showed a predominance of *Bacillus spp* (63.0% to 73.0% rRNA reads), while the same organisms were at much reduced abundance in non-rRNA data (18.0% to 22.0% non-rRNA reads). Such an imbalance was not noted for other taxa present in the mock community suggesting some failure of targeting of the *Bacillus spp* by the FastSelect probes. However, in these same libraries residual human rRNA was also present suggesting more likely that the level of FastSelect probes and for higher input RNA such as these cultures or even whole tissue, the concentration might need to be increased (**Figure 2**). While the overall rRNA-depletion performance was good, incomplete rRNA depletion will limit the detection of target species [39]. The overarching goal of this and other meta-transcriptomic approaches is to produce sufficient non-human non-rRNA reads to identify pathogens of interest. Where non-human, non-rRNA reads are unexpectedly low, the depth of sequencing becomes an important consideration. Across the samples in this study, the mean number of non-human non-rRNA reads was 3,125,035, which would be considered an acceptable read number for pathogen identification [39]. However, the lowest yielding non-human non-rRNA library was *RAPID30* with only 5,898 reads. At this sequencing depth potentially no pathogen sequences will be identified, and it is also difficult to rule out infections (often a goal of clinical mNGS), therefore target yields of >1M non-human, non-rRNA would be ideal.

3.3. Viral sequence identification, genome recovery and quantitative performace

For each library, the non-human, non-rRNA sequences were taxonomically assigned and quantified with a focus on viral reads expressed as log transformed-read per million (RPM) values (**Table 4**). In samples where known pathogens were detected, e.g. SARS-CoV-2, RSV and A/pdmH1N1 influenza virus, the $_{log}$RPM values ranged between 2.64 and 6.00. The mean $_{log}$RPM values for the sample groups were 5.17, 3.74 and 5.88, respectively, and in all samples the expected respiratory pathogen was identified. In addition to detecting known pathogens, we sought to evaluate the utility of the RAPID*prep* assay in identifying unknown pathogens in two sample groups. The first, a SARI cohort (*RAPID17-24*) and the other, children with mild respiratory infections (*RAPID33-39*) (**Table 3**). For the SARI cohort, two samples returned a positive result using the RAPID*prep* assay, and the identification of a possible causative pathogens which had not been identified using conventional diagnostic methods (**Table 4**). Human cytomegalovirus (CMV) was identified at low levels in *RAPID17* with a $_{log}$RPM value of 0.62, that mapped to multiple viral genes, and were likely true hits and not host-derived. Interestingly, abundant Influenza C virus was identified in *RAPID22* with a *log*RPM of 5.77 (**Table 4**). As this sample group was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

comprised of samples stored for up to six weeks at 4°C before transfer to -80°C , and also thawed and refrozen multiple times, the subsequent identification of possible pathogens emphasises the clear diagnostic potential of the RAPID*prep* assay and RNA-mNGS approaches. Across the mild unknown cases, human rhinovirus was detected in 5 of 7 samples with $_{log}$RPM ranging between 4.79 and 6.00 (**Table 4**). Incidentally, human betaherpesvirus 7 (HHV-7) was also detected in *RAPID36*, although not likely responsible for the acute respiratory illness symptoms, it remains an important detection.

To assess the genome recovery of the RAPID*prep* assay, we examined the sequence coverage of the SARS-CoV-2 (*RAPID01-12*) and RSV libraries (*RAPID25-32*) by mapping against the viral genome. For the SARS-CoV-2 data, all libraries (n=12) produced genome coverage >99.9% at a mean depth of 7,270X (range: 22X to 23,085X). For the RSV data, only half the libraries (n=4) produced genome coverage >90%, while the remaining ranged between 41.9% and 77.8%. The reduced genomic recovery was due to lower coverage depth (mean: 7X, range: 1X to 23X) (**Supp Table 2**). As mentioned previously, the reduced genomic yield in the RSV samples was due to an over-abundance of human sequences (**Figure 2**). Despite this, the genomic recovery was more than sufficient to subtype both the SARS-CoV-2 and RSV cases, as well as the previously undiagnosed rhinovirus sequences from the unknown mild infections, using a phylogenetic approach (**Supp Figure 1A-1C**). This not only demonstrates the diagnostic performance of the RAPID*prep* assay but also utility for allowing further epidemiological investigation of potential pathogens. Furthermore, this genomic data could also be used to design new diagnostic assays and inform vaccine development as seen during the COVID-19 pandemic with the sequencing and release of the first SARS-CoV-2 genome [2].

To assess the quantitative performance of the RAPID*prep* assay, we utilized the cycle threshold values generated using RT-qPCR and compared these to $_{log}$RPM values from the RAPID*prep* SARS-CoV-2 and RSV positive libraries (**Figure 3**). The SARS-CoV-2 sample group comprised 12 PCR positive samples (*RAPID01-12*). Here, we identified SARS-CoV-2 in all 12 samples using the RAPID*prep* method and revealed a strong linear relationship (R^2=0.86) between $_{log}$RPM and CT values (**Figure 3A**). As expected, $_{log}$RPM increases as CT values decrease, indicating that the RAPID*prep* method was sensitive to relative viral load. A similar result was observed for the RSV data (*RAPID25-32*) where the read $_{log}$RPM and RT-qPCR CT values were well-correlated (R^2=0.85) (**Figure 3B**). Together, these results highlight the quantitative performance of the RAPID*prep* assay.

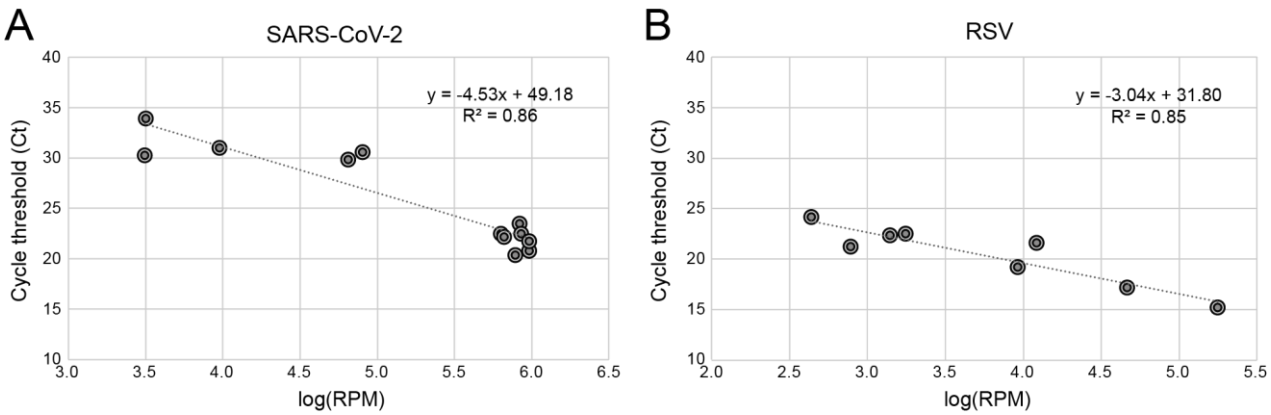

Figure 3 – Quantitative detection of SARS-COV-2 and RSV sequences. A simple linear regression model was applied to both SARS-CoV-2 (**A**) and RSV (**B**) data sets with a line of best fit estimating the relationship between log transformed reads per million (RPM) and cycle threshold (CT) values. The linear-regression slope coefficient, and intercept parameter are printed on the top right of each plot with R^2 calculated to measure the goodness of fit.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

Table 4 - Log transformed read-per-million (RPM) virus distribution across RAPID*prep* samples.

Library	SARS-CoV2	RSV-A	RSV-B	Flu-A pdmH1N1	Flu-C	Rhinovi-rus	GB virus C	CMV	HHV7
RAPID01	5.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID02	3.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID03	3.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID04	5.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID05	3.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID06	5.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID07	5.82	0.00	0.00	0.00	0.00	0.00	1.15	0.00	0.00
RAPID08	5.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID09	5.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID10	4.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID11	4.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID12	5.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID13	0.00	0.00	0.00	5.94	0.00	0.00	0.00	0.00	0.00
RAPID14	0.00	0.00	0.00	5.83	0.00	0.00	0.00	0.00	0.00
RAPID15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.00
RAPID18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID22	0.00	0.00	0.00	0.00	5.77	0.00	0.00	0.00	0.00
RAPID23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID25	0.00	0.00	2.64	0.00	0.00	0.00	0.00	0.00	0.00
RAPID26	0.00	3.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID27	0.00	0.00	3.25	0.00	0.00	0.00	0.00	0.00	0.00
RAPID28	0.00	2.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID29	0.00	0.00	4.09	0.00	0.00	0.00	0.00	0.00	0.00
RAPID30	0.00	5.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID31	0.00	4.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID32	0.00	0.00	3.14	0.00	0.00	0.00	0.00	0.00	0.00
RAPID33	0.00	0.00	0.00	0.00	0.00	5.61	0.00	0.00	0.00
RAPID34	0.00	0.00	0.00	0.00	0.00	5.94	0.00	0.00	0.00
RAPID35	0.00	0.00	0.00	0.00	0.00	6.00	0.00	0.00	0.00
RAPID36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.35
RAPID37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPID38	0.00	0.00	0.00	0.00	0.00	4.79	0.00	0.00	0.00
RAPID39	0.00	0.00	0.00	0.00	0.00	5.98	0.00	0.00	0.00
RAPID40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Log-RPM	6.00	5.00	4.00	3.00	2.00	1.00	0.00

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0417.v1

3.4. Comparison of RAPIDprep to commercial assay

Finally, we sought to compare the performance of the RAPID*prep* assay against a commercial assay – Takara SMARTer Stranded Total RNA-Seq Kit v2. The SMARTer-Seq libraries were prepared previously from eight residual diagnostic samples and two Zymo mock community controls as part of a study into the possible infectious causes of SARI in children (**Table 2**). These libraries labelled as *ICU15-24* had the same source RNA extracts for the RAPID*prep* libraries (*RAPID15-24*), with each RNA labelled with same sample number (i.e. ICU15 & RAPID15 share the same RNA source – see **Supp Table 1**). For the analysis, we processed each library by removal of low quality, human and non-human rRNA sequences before extracting 1M non-human rRNA for alignment and taxonomic assignment using MetaPhlAn3. An un-clustered heatmap of microbial abundance (Z-score for the top 24 taxa) was used to compare the sensitivity and specificity of the mNGS protocols (**Figure 4**). Overall, there was good concordance between the SMARTer-Seq and RAPID*prep* assays with conservation of nasal-oral taxa across protocols, particularly for the predominant species (*ICU/RAPID17*). Furthermore, the Zymo mock community control samples, displayed good repeatability across methods presenting similar row z-scores. As anticipated, there was some variation between methods likely due to the depth of sequencing and batch effects. For example, the increased abundance of *Escherichia coli* sequences across the RAPID*prep* libraries indicates a common source, and most likely from using different reagents. This highlights the need for positive and non-template controls, as well as reagent batching when performing mNGS studies, particularly with low-biomass samples [40]. Finally, the influenza C virus detected in *RAPID22* was also identified in *ICU22* (**Table 4**), again confirming the diagnostic value were largely comparable. The SMARTer-Seq protocol is slower (2-day protocol) and more costly (~2X) but is designed for very low-inputs (RNA amounts <1 ng), and therefore, more suitable for mNGS of low-biomass sample types such as cerebrospinal fluid (CSF), where it has been used for pathogen discovery [41]. This aspect of the RAPID*prep* assay is yet to be explored.

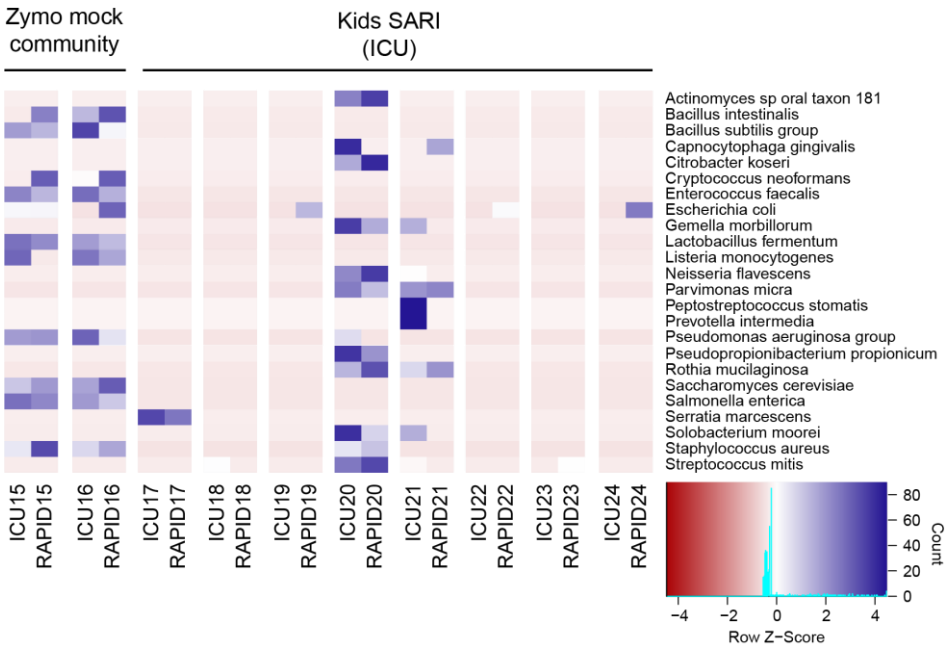

Figure 4 – Comparison of RAPID*prep* to commercial RNA library preparation kit. Using previously generated data for the kids SARI cohort, we compared the twenty-four most abundant species identified across both protocols for the same set of samples. An unclustered heatmap of microbial abundance (Z-score) is shown with differences between samples identified by a deeper blue shading, whilst organisms conserved across samples were lighter blue through to red. A frequency histogram is overlayed on the colour key and signifies the count of each Z score at any given point. Tick labels on the X-axis in the ICUXX format represent deep RNA sequencing generated previously, while tick labels in the RAPIDXX format represent sequencing data generated in this study using the RAPID*prep* assay for the corresponding samples.

 doi:10.20944/preprints202212.0417.v1

4. Conclusions

We present a simple yet robust workflow for the mNGS of RNA from clinical respiratory samples. Our RAPID*prep* assay has been designed specifically as a rapid response tool, and has proven to be effective in novel disease investigations, including identifying the first cases of emergent Japanese encephalitis virus during the 2021-22 outbreak in south-eastern Australia [42,43], and characterising the first cases of COVID-19 in NSW [33]. The assay has also been used to investigate non-human diseases, and was critical to the genome recovery of a novel Hendra virus variant detected initially from a fatal equine infection by pan-paramyxovirus RT-PCR [1]. In the future, pathogen-agnostic mNGS testing will likely assume a greater role in identifying and quantifying novel, emerging, and re-emerging pathogens to guide individual patient management and public health responses as part of communicable disease control.

Supplementary Materials: The following supporting information can be downloaded online.

Author Contributions: Conceptualization, J.S.E.; methodology, R.L.T, K.K., C.S., A.M. and J.S.E.; formal analysis, R.L.T, K.K. and J.S.E.; resources, R.B., E.C.H., D.E.D., P.N.B. and J.K..; data curation, R.L.T, K.K. and J.S.E.; writing—original draft preparation, R.L.T, K.K. and J.S.E.; writing—review and editing, R.L.T, K.K., C.S., E.C.H., D.E.D., P.N.B., J.K., and J.S.E.; funding acquisition, J.S.E. All authors have read and agreed to the published version of the manuscript.

Funding: Funding was provided through the Snow Medical Foundation BEAT COVID-19 research program, the National Health and Medical Research Council Centre of Research Excellence in Emerging Infectious Diseases (#1102962), and the Medical Research Future Fund (#FSPGN000045).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Local Ethics Committee of the Sydney Children's Hospitals Network (approval numbers HREC/18/SCHN/263 & 2020/ETH00837), and the Western Sydney Local Health District (approval numbers LNR/17/WMEAD/128 and SSA/17/WMEAD/129).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Metagenomic sequence libraries used in this study have been submitted to the NCBI short read archive (SRA) with accession numbers: SRR22726217 - SRR22726256.

Acknowledgments: We acknowledge the University of Sydney's high-performance computing cluster Artemis for providing the computing resources used for this study. A/pdmH1N1 2009 influenza viruses were kindly provided by Dr Maryam Shojaei.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Annand, E.J.; Horsburgh, B.A.; Xu, K.; Reid, P.A.; Poole, B.; de Kantzow, M.C.; Brown, N.; Tweedie, A.; Michie, M.; Grewar, J.D.; et al. Novel Hendra Virus Variant Detected by Sentinel Surveillance of Horses in Australia. *Emerging infectious diseases* **2022**, *28*, 693-704, doi:10.3201/eid2803.211245.
2. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A new coronavirus associated with human respiratory disease in China. **2020**.
3. Zhang, D.; Lou, X.; Yan, H.; Pan, J.; Mao, H.; Tang, H.; Shu, Y.; Zhao, Y.; Liu, L.; Li, J.; et al. Metagenomic analysis of viral nucleic acid extraction methods in respiratory clinical samples. *BMC genomics* **2018**, *19*, 773-773, doi:10.1186/s12864-018-5152-5.
4. Shi, M.; Zhao, S.; Yu, B.; Wu, W.-C.; Hu, Y.; Tian, J.-H.; Yin, W.; Ni, F.; Hu, H.-L.; Geng, S.; et al. Total infectome characterization of respiratory infections in pre-COVID-19 Wuhan, China. **2022**.
5. Serpa, P.H.; Deng, X.; Abdelghany, M.; Crawford, E.; Malcolm, K.; Caldera, S.; Fung, M.; McGeever, A.; Kalantar, K.L.; Lyden, A.; et al. Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. *Genome medicine* **2022**, *14*, 1-74, doi:10.1186/s13073-022-01072-4.
6. Li, Y.; Deng, X.; Hu, F.; Wang, J.; Liu, Y.; Huang, H.; Ma, J.; Zhang, J.; Zhang, F.; Zhang, C. Metagenomic analysis identified co-infection with human rhinovirus C and bocavirus 1 in an adult suffering from severe pneumonia. *The Journal of infection* **2018**, *76*, 311-313, doi:10.1016/j.jinf.2017.10.012.
7. Subramoney, K.; Mtileni, N.; Bharuthram, A.; Davis, A.; Kalenga, B.; Rikhotso, M.; Maphahlele, M.; Giandhari, J.; Naidoo, Y.; Pillay, S.; et al. Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021. *Journal of medical virology* **2022**, *94*, 3676-3684, doi:10.1002/jmv.27797.
8. Itokawa, K.; Sekizuka, T.; Hashino, M.; Tanaka, R.; Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLoS One* **2020**, *15*, e0239403, doi:10.1371/journal.pone.0239403.
9. Lu, X.; Wang, L.; Sakthivel, S.K.; Whitaker, B.; Murray, J.; Kamili, S.; Lynch, B.; Malapati, L.; Burke, S.A.; Harcourt, J.; et al. US CDC Real-Time Reverse Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging infectious diseases* **2020**, *26*, 1654-1665, doi:10.3201/eid2608.201246.
10. Wang, L.; Piedra, P.A.; Avadhanula, V.; Durigon, E.L.; Machablishvili, A.; López, M.-R.; Thornburg, N.J.; Peret, T.C.T. Duplex real-time RT-PCR assay for detection and subgroup-specific identification of human respiratory syncytial virus. *Journal of virological methods* **2019**, *271*, 113676-113676, doi:10.1016/j.jviromet.2019.113676.
11. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *bioRxiv* **2018**, doi:10.1101/274100.
12. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15-21, doi:10.1093/bioinformatics/bts635.
13. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754-1760, doi:10.1093/bioinformatics/btp324.
14. Kopylova, E.; Noe, L.; Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **2012**, *28*, 3211-3217, doi:10.1093/bioinformatics/bts611.
15. Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **2015**, *31*, 1674-1676, doi:10.1093/bioinformatics/btv033.
16. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: architecture and applications. *BMC bioinformatics* **2009**, *10*, 421-421, doi:10.1186/1471-2105-10-421.
17. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **2015**, *12*, 59-60, doi:10.1038/nmeth.3176.
18. Segata, N.; Waldron, L.; Ballarini, A.; Narasimhan, V.; Jousson, O.; Huttenhower, C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **2012**, *9*, 811-814, doi:10.1038/nmeth.2066.

19. Bushnell. BBMap short-read aligner, and other bioinformatics tools. **2016**.
20. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic biology* **2010**, *59*, 307-321, doi:10.1093/sysbio/syq010.
21. Zhao, S.; Zhang, Y.; Gamini, R.; Zhang, B.; von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific reports* **2018**, *8*, 4781-4712, doi:10.1038/s41598-018-23226-4.
22. Albert, E.; Torres, I.; Bueno, F.; Huntley, D.; Molla, E.; Fernández-Fuentes, M.Á.; Martínez, M.; Poujois, S.; Forqué, L.; Valdivia, A.; et al. Field evaluation of a rapid antigen test (Panbio™ COVID-19 Ag Rapid Test Device) for COVID-19 diagnosis in primary healthcare centres. *Clinical microbiology and infection* **2021**, *27*, 472.e477-472.e410, doi:10.1016/j.cmi.2020.11.004.
23. Bal, A.; Pichon, M.; Picard, C.; Casalegno, J.S.; Valette, M.; Schuffenecker, I.; Billard, L.; Vallet, S.; Vilchez, G.; Cheynet, V.; et al. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. *BMC infectious diseases* **2018**, *18*, 537-537, doi:10.1186/s12879-018-3446-5.
24. Dudas, G.; Bedford, T. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC evolutionary biology* **2019**, *19*, 1-232, doi:10.1186/s12862-019-1567-0.
25. Tabor, S.; Richardson, C.C. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J Biol Chem* **1989**, *264*, 6447-6458.
26. Zhu, B. Bacteriophage T7 DNA polymerase - sequenase. *Frontiers in microbiology* **2014**, *5*, 181-181, doi:10.3389/fmicb.2014.00181.
27. Brenner, T.; Decker, S.O.; Grumaz, S.; Stevens, P.; Bruckner, T.; Schmoch, T.; Pletz, M.W.; Bracht, H.; Hofer, S.; Marx, G.; et al. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine (Baltimore)* **2018**, *97*, e9868-e9868, doi:10.1097/MD.0000000000009868.
28. Peddu, V.; Shean, R.C.; Xie, H.; Shrestha, L.; Perchetti, G.A.; Minot, S.S.; Roychoudhury, P.; Huang, M.-L.; Nalla, A.; Reddy, S.B.; et al. Metagenomic Analysis Reveals Clinical SARS-CoV-2 Infection and Bacterial or Viral Superinfection and Colonization. *Clinical chemistry (Baltimore, Md.)* **2020**, *66*, 966-972, doi:10.1093/clinchem/hvaa106.
29. Poulsen, C.S.; Ekstrøm, C.T.; Aarestrup, F.M.; Pamp, S.J. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology spectrum* **2022**, *10*, e0009022-e0009022, doi:10.1128/spectrum.00090-22.
30. Wilson, M.R.; Naccache, S.N.; Samayoa, E.; Biagtan, M.; Bashir, H.; Yu, G.; Salamat, S.M.; Somasekar, S.; Federman, S.; Miller, S.; et al. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *The New England journal of medicine* **2014**, *370*, 2408-2417, doi:10.1056/NEJMoa1401268.
31. Peck, M.A.; Sturk-Andreaggi, K.; Thomas, J.T.; Oliver, R.S.; Barritt-Ross, S.; Marshall, C. Developmental validation of a Nextera XT mitogenome Illumina MiSeq sequencing method for high-quality samples. *Forensic science international : genetics* **2018**, *34*, 25-36, doi:10.1016/j.fsigen.2018.01.004.
32. Di Giallonardo, F.; Kok, J.; Fernandez, M.; Carter, I.; Geoghegan, J.L.; Dwyer, D.E.; Holmes, E.C.; Eden, J.S. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses* **2018**, *10*, doi:10.3390/v10090476.
33. Eden, J.S.; Rockett, R.; Carter, I.; Rahman, H.; de Ligt, J.; Hadfield, J.; Storey, M.; Ren, X.; Tulloch, R.; Basile, K.; et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol* **2020**, *6*, veaa027, doi:10.1093/ve/veaa027.

34. Tulloch, R.L.; Kok, J.; Carter, I.; Dwyer, D.E.; Eden, J.S. An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus. *Viruses* **2021**, *13*, doi:10.3390/v13030499.
35. Bansal, V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC bioinformatics* **2017**, *18*, 43-43, doi:10.1186/s12859-017-1471-9.
36. Schoonvaere, K.; De Smet, L.; Smagghe, G.; Vierstraete, A.; Braeckman, B.P.; de Graaf, D.C. Unbiased RNA Shotgun Metagenomics in Social and Solitary Wild Bees Detects Associations with Eukaryote Parasites and New Viruses. *PloS one* **2016**, *11*, e0168456-e0168456, doi:10.1371/journal.pone.0168456.
37. Meadow, J.F.; Altrichter, A.E.; Bateman, A.C.; Stenson, J.; Brown, G.Z.; Green, J.L.; Bohannan, B.J.M. Humans differ in their personal microbial cloud. *PeerJ (San Francisco, CA)* **2015**, *3*, e1258-e1258, doi:10.7717/peerj.1258.
38. Qamar, W.; Khan, M.R.; Arafah, A. Optimization of conditions to extract high quality DNA for PCR analysis from whole blood using SDS-proteinase K method. *Saudi journal of biological sciences* **2017**, *24*, 1465-1469, doi:10.1016/j.sjbs.2016.09.016.
39. Yang, L.; Haidar, G.; Zia, H.; Nettles, R.; Qin, S.; Wang, X.; Shah, F.; Rapport, S.F.; Charalampous, T.; Methé, B.; et al. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respiratory research* **2019**, *20*, 265-212, doi:10.1186/s12931-019-1218-4.
40. Porter, A.F.; Cobbin, J.; Li, C.X.; Eden, J.S.; Holmes, E.C. Metagenomic Identification of Viral Sequences in Laboratory Reagents. *Viruses* **2021**, *13*, doi:10.3390/v13112122.
41. Tuddenham, R.; Eden, J.S.; Gilbey, T.; Dwyer, D.E.; Jennings, Z.; Holmes, E.C.; Branley, J.M. Human pegivirus in brain tissue of a patient with encephalitis. *Diagn Microbiol Infect Dis* **2020**, *96*, 114898, doi:10.1016/j.diagmicrobio.2019.114898.
42. Sikazwe, C.; Neave, M.J.; Michie, A.; Mileto, P.; Wang, J.; Cooper, N.; Levy, A.; Imrie, A.; Baird, R.W.; Currie, B.J.; et al. Molecular detection and characterisation of the first Japanese encephalitis virus belonging to genotype IV acquired in Australia. *PLoS Negl Trop Dis* **2022**, *16*, e0010754, doi:10.1371/journal.pntd.0010754.
43. Waller, C.; Tiemensma, M.; Currie, B.J.; Williams, D.T.; Baird, R.W.; Krause, V.L. Japanese Encephalitis in Australia - A Sentinel Case. *N Engl J Med* **2022**, *387*, 661-662, doi:10.1056/NEJMc2207004.