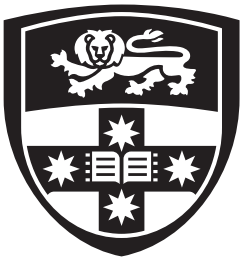
The genomics of viral respiratory disease in Australia: Understanding the spread evolution and pathogenesis of non-influenza respiratory viruses



THE UNIVERSITY OF SYDNEY

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Sydney Medical School
Faculty of Medicine and Health
The University of Sydney

Rachel Laurene Tulloch

BAnVetBioSci (Hons)

December 2022

Statement of Originality

The contents of this thesis are the product of my own original work, and to the best of my knowledge and belief, is original except as acknowledged in the text. I hereby declare that appropriate ethical review and approval was sought for this work and that the intellectual content of this thesis has not been submitted for the award of any other degree at The University of Sydney or any other academic institution unless otherwise stated.

Rachel Tulloch
December 2022

Abstract

Respiratory disease caused by emerging and endemic viruses present a major threat to public health both locally and globally. Human metapneumovirus (HMPV) is a significant cause of respiratory illness in children and adults, however, little is known about the sources and patterns of HMPV genetic diversity within Australia, which is important for their control. To address this, I developed an amplicon-based, whole genome sequencing method for HMPV. This assay was then applied to a collection of clinical samples spanning 10 years, that gave the first historical perspective of HMPV genomic diversity in Australia, and assessed the impact of the COVID-19 pandemic, highlighting a major collapse in circulating diversity. Next, I explored the role of metagenomic sequencing (mNGS) in clinical diagnostics of respiratory infections. A proof-of-concept study was performed to evaluate the utility of mNGS in identifying unknown pathogens in a cohort of children with undiagnosed severe acute respiratory illnesses. Several challenges were revealed regarding the use of residual clinical specimens, however, novel finding including the identification of a severe Influenza C virus infection. To facilitate the identification and characterisation of novel and emerging respiratory viruses, a simple and fast mNGS was developed with an aim to provide tangible results in less than 24 hours from sample collection. The RAPID*prep* assay was then applied across range of respiratory samples and sources, highlighting the effectiveness of the approach. Ultimately, this research has helped to further our understanding of the epidemiology, transmission and spread of HMPV in Australia, and contributed to the identification of unknown or novel respiratory pathogens using mNGS.

Dedication

To *my mother and father,*
you got me here.

To *my late Nunna,*
for your infinite love and strength.

To *my brother and sister,*
for pathing the way before and after me.

To *my partner,*
for unwavering love and support.

Acknowledgements

There are so many people I would like to thank for their encouragement, and support throughout the duration of my PhD candidature. I am sure I would not have been able to complete this task without you.

To my primary supervisor Dr John-Sebastian Eden, thank you for consistent guidance and support and for your patience and understanding throughout this journey. I will forever be thankful for you taking me on as your first ever PhD student. This PhD thesis is a testament to your ability and commitment to revolutionising viral genomics. I have watched our lab grow under your guidance and am proud to have been a part of such a driven research team. My eternal thanks to my auxiliary supervisor Dr Phillip Britton, who has been instrumental in providing a breadth of knowledge as a highly accomplished clinician and researcher. Thank you for your constant guidance from the inception of CSI Study, and my biggest gratitude for your invaluable guidance through the late state of my thesis submission. I am thankful to Professor Dominic Dwyer and Professor Tania Sorrell, without you my candidature would not exist, thank you for believing in both J-S and myself. To my entire supervisory team, your guidance, support, encouragement, and wisdom have been behind the development of my project every step of the way and have helped to shape the scientist I am today.

I would like to extend my thanks to the many researchers, academics, and collaborators I have been lucky enough to meet and work with throughout my candidature. To the wonderful people who call or have called WIMR their home away from home (ie. Work) thank-you for making the institute a welcoming place to all. To Katie Fisher, thank you times a million, from walking me through my first ever library preparation, to constantly bailing me out when I forgot my WIMR login, through to endless chats at our desks, thank you, for your friendship, kindness, and support. My express gratitude to Professor Sarah Palmer, for always including myself and our group in every Christmas Party and Lab Party, your kindness is second to none, thank you for making WIMR so special. To the entire level 6, thank you for creating a welcoming atmosphere, and good luck to you all in your respective endeavours.

To my friends and fellow postgraduate students, thank you. Your understanding, companionship and support has been invaluable through the highs and lows of candidature. Dr. Emma Wilkie, you so

deserve a special mention, talk about worlds collide, my very own veterinarian friend at Westmead, thank you for letting me accompany you to the vivarium, for our midday coffee dates and above all showing me the strength to follow your heart, infarctions and all, while I missed you terribly during the late stage of my candidature the time I did get to spend with you got me through some of the longest days. Thank you. To my valued colleague and friend Rebecca Burrell thank you for your friendship at Westmead, our corridor chats over dry ice are some of my favourite memories, thank you for always lending an ear and good luck with the remainder of your own candidature. To my high school friend, Andrew Fyffe, your friendship saw me through my HSC, and now through the submission of my PhD, our lunchtime "meetings" at Bear Brassier were the highlight of many a long day. I will think of you every time I eat spaghetti meatballs and garlic bread. Thank you for your unwavering friendship and understanding during some not so fun times, and the biggest congratulations on the close of your own PhD candidature.

To Bailey, Vinnie, Kitana, Edge, Charlotte, Axel, Derreck, Lucy, Marshall and about a million chickens, I could not have done this without you. My fur, feathered and hooved children, you have at times distracted me from my PhD, but throughout it all you have been there, with unwavering love and affection. Every day you inspire me to keep going. To my late Lucy and Marshall, while you weren't there to see me finish my candidature, you were both there at the start, you are missed every single day.

To my partner Rhys McGregor, where do I even start. From the moment I even considered taking on this endeavour you were there encouraging and supporting me to reach for the stars. You saw me through this every step of the way, every up and down, twist and turn, you were the lighthouse guiding me home again safely. We moved house, twice, bought our first home, brought a new car, welcomed new family members, and said goodbye to some, what a jam packed 3.5 years. Thank you will never be enough to express how much your love and encouragement has meant to me throughout this entire experience, but I can only continue to try. As this chapter of my life closes, I am so excited to see what the future holds for us.

Above all, thank you to my Mum and Dad, thank you. I am here because of you both. You have always put my needs above your own and supported me in every academic pursuit. Mum, you have shown me the importance of an education and made sacrifices at every turn to ensure I was afforded the best, you have encouraged me to do things I never thought possible and helped me to take every opportunity presented to me. Dad, you have shown me the importance of hard work and not stopping

until the job is done. Thank you for the sacrifices you have made for myself and our family, so many of the wonderful opportunities that have led me here would not have been possible. Mum and Dad, without your love and encouragement, I would not have been able to complete my PhD, you will forever be the reason this was possible. To my brother and sister, James and Kristen, you inspire me to be a better person, James with your generosity and can to attitude, and Kristen with your kind heart and unselfishness, thank you for being pillars in my life before, during and I'm sure after my PhD candidature. My beautiful family, my grandparents, aunts, and uncles, thank you for making our family so uniquely ours, you have all helped to shape me be the person I am today.

Preface

This thesis is comprised of seven chapters and is submitted for examination with both published and unpublished works. I, Rachel Laurene Tulloch, am the primary, or primary co-author on all included publications presented in this thesis. Chapter 1 is a general introduction to Human Metapneumovirus and associated literature. Chapter 2 is a body of work which has been peer-reviewed and published in a peer-reviewed scientific journal, Chapter 3 and 5, have been submitted to scientific journals and at the time of this thesis submission are currently under review. Chapter 4 is a general introduction to the use of metagenomic next generation sequencing. Chapter 6 is presented as a stand-alone body of work, constructed for the purpose of submission to a scientific journal. Chapter 7 includes a general discussion and conclusion of the thesis. An author attribution statement declaring each of the authors individual contributions are included at the beginning of each research chapter and the inclusion of published manuscripts adheres to the University of Sydney's Thesis and Examination Higher Degree by Research Policy, 2015, guidelines for thesis with publications.

Chapter 1 is a literature review which summarizes current understanding of Human Metapneumovirus, disease pathogenesis, and disease intervention strategies. Here scientific problems are identified and the aims of the next two chapters are discussed.

Chapter 2 is a research paper published in the peer-reviewed, open access virology journal, *Viruses*, which describes an overlapping amplicon-based method for the whole-genome sequencing of Human Metapneumovirus from clinical extracts.

Chapter 3 is a research paper currently under review with peer-reviewed, open access journal *Virus Evolution*, and provides the first comprehensive genome-scale molecular epidemiological analysis of Human metapneumovirus in Australia.

Chapter 4 is an introduction to metagenomics and metatranscriptomics with a review of the current diagnostic landscape in human health with a focus on acute respiratory infection. This chapter identifies the purpose and aims of the following two chapters.

Chapter 5 is an unpublished body of work, presented in a research paper format, which describes the use of metagenomic next-generation sequencing in the identification of pathogens in a cohort of children with unknown respiratory illnesses.

Chapter 6 is a research paper currently under review with the peer-reviewed, open access journal, *Viruses*. This paper details a simple, fast protocol for RNA metagenomic sequencing of clinical samples. Here the effectiveness of the RAPID*prep* method is evaluated using a range of clinical respiratory samples containing both known and unknown pathogens.

Chapter 7 is a general discussion of the work detailed within this thesis and includes the successes and shortcomings of the research and set expectations and suggestions for future research in these fields.

Supplementary material for each data chapter have been provided at the end of each body of work. Co-authored manuscripts completed during candidature are presented within the appendices section of this thesis **(Appendix A)**.

Publications

Publications Arising from Candidature

Tulloch, RL, Kok, J, Carter, I, Dwyer, DE, & Eden, J-S, 2021, 'An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus', *Viruses*, vol. 13, no. 3, p. 499–, doi: 10.3390/v13030499.

Tulloch, RL*, Kim, K*, Sikazwe, C, Michie, A, Burrell, R Holmes, EC, Dwyer, D, Britton, P, Kok, J & Eden, J-S, 2022, 'RAPIDprep: A simple, fast protocol for RNA metagenomic sequencing of clinical samples', *Viruses*, under review. * Denotes equal authorship contribution

Publications Under Review

Tulloch, RL, Horsburgh, BA, Cutmore, E, Tran, T, Jordan, P, Donovan, L, Ngo, C, Carter, I, Hammond, JM, Deveson, IW, Dwyer, D, Kok, J & Eden, J-S, 2022, 'The Genomic Epidemiology of Human Metapneumovirus in Australia and the impact of COVID-19' *Virus Evolution*, under review.

Publications Arising from Collaborative Work During Candidature

Eden, J-S, Sikazwe, C, Xie, R, Deng, Y-M, Sullivan, SG, Michie, A, Levy, A, Cutmore, E, Blyth, CC, Britton, PN, Crawford, N, Dong, X, Dwyer, DE, Edwards, KM, Horsburgh, BA, Foley, D, Kennedy, K, Minney-Smith, C, Speers, D, **Tulloch, RL,** Holmes, EC, Dhanasekaran, V, Smith, DW, Kok, J, & Barr, IG, 2022, 'Off-season RSV epidemics in Australia after easing of COVID-19 restrictions', *Nature Communications*, vol. 13, no. 1, pp. 2884–2884, doi: 10.1038/s41467-022-30485-3.

Robertson, M, Eden, J-S, Levy, A, Carter, I, **Tulloch, RL,** Cutmore, EJ, Horsburgh, BA, Sikazwe, CT, Dwyer, DE, Smith, DW, & Kok, J, 2021, 'The spatial-temporal dynamics of respiratory syncytial virus infections across the east–west coasts of Australia during 2016–17', *Virus Evolution*, vol. 7, no. 2, pp. veab068–veab068, doi: 10.1093/ve/veab068.

Hasan, MM*, **Tulloch, RL*,** Thomson, PC, Raadsma, HW, & Khatkar, MS, 2020, 'Meta-analysis of genetic parameters of production traits in cultured shrimp species', *Fish and Fisheries (Oxford, England)*, vol. 21, no. 6, pp. 1150–1174, doi: 10.1111/faf.12495. * Denotes equal authorship contribution

Eden, J-S, Rockett, R, Carter, I, Rahman, H, de Ligt, J, Hadfield, J, Storey, M, Ren, X, **Tulloch, RL**, Basile, K, Wells, J, Byun, R, Gilroy, N, O'Sullivan, MV, Sintchenko, V, Chen, SC, Maddocks, S, Sorrell, TC, Holmes, EC, Dwyer, DE, & Kok, J, 2020,, 'An emergent clade of SARS-CoV-2 linked to returned travellers from Iran', *Virus Evolution*, vol. 6, no. 1, pp. veaa027–veaa027, doi: 10.1093/ve/veaa027.

List of Figures

Chapter 3: The Genomic Epidemiology of Human Metapneumovirus in Australia and the Impact of COVID-19

Chapter 5: Diagnosis and Analysis of Unexplained Severe Respiratory Illness in Australian Children using Metagenomic Next-Generation Sequencing: A Pilot Study

Chapter 6: RAPIDprep: A simple, fast protocol for RNA metagenomic sequencing of clinical samples

List of Tables

Units of measurement

°C	Degree Celsius
µg	Microgram
µL	Microlitre
hr or h	Hour
mg	Milligram
mL	Millilitre
mM	Millimolar
ng	Nanogram
s	Second
mmol	Millimole

Table of contents

CHAPTER 1

Overview of Human Metapneumovirus and Clinical Disease

Chapter 1: Overview of Human Metapneumovirus and Clinical Disease

Abstract

Acute respiratory tract infections (ARI) are the leading cause of morbidity and mortality in paediatric patients globally (Wang et al., 2021). From a public health perspective, influenza virus has long been considered the most important human respiratory virus due to its propensity for causing large seasonal epidemics and pandemics. Epidemiological data show there are several other highly prevalent respiratory viruses that commonly cause ARI including rhinoviruses, parainfluenza virus and respiratory syncytial virus (RSV). In 2001, a previously unidentified respiratory virus, Human Metapneumovirus (HMPV), was identified as a global cause of ARI existing in the human population for up to 50 years before its discovery. HMPV is a single-stranded RNA virus belonging to the *Metapneumovirus* genus that has been recognised to be of increasing importance in the study of influenza-like illness (ILI) since its discovery and classification in 2001 (van Den Hoogen et al., 2002). HMPV is an endemic cause of ARI, with pooled analysis estimating it is responsible for 6.1-6.4% of ARI related hospital admissions in patients younger than 20 years of age, globally (Hamelin et al., 2004, Panda et al., 2014). Disease is especially burdensome in at risk populations including children and the elderly, with infants under the age of 1 having disproportionately high risk of severe human metapneumovirus infection (Cattoir et al., 2019, Wang et al., 2021). Genetically similar to RSV, HPMV has been identified as a common coinfecting agent, with one study reporting a simultaneous HPMV infection in up to 70% of infant RSV bronchiolitis cases requiring ventilation (Greensill et al., 2003). In this section, I aim to review recent literature pertaining to our understanding of HMPV as a neglected respiratory pathogen of global significance.

HMPV Overview

Discovery

HMPV was discovered in 2001 by scientists at the Departments of Virology and Paediatrics at the Erasmus Medical Centre in Rotterdam, the Netherlands (Bernadette et al., 2001). Respiratory samples from 28 cases of respiratory tract infection (RTI) where the causative could not be identified were inoculated on tertiary monkey kidney (tMK) cells. Supernatants of the infected tMK cells were analysed using electron microscopy which revealed the presence of a paramyxovirus-like pleiomorphic particles, later named HMPV. Retrospective serological studies have determined most children under the age of 5 in the Netherlands had been infected or exposed to HMPV. Such studies have also detected a high prevalence of HMPV seropositivity in adults over the age of 50, suggesting HMPV is a newly discovered virus rather than newly emerging (Gioula and Kyriazopoulou-Dalaina, 2009). Since its discovery HMPV has been detected on every continent, making HMPV a virus of global relevance (Shi et al., 2017, Kaslow et al., 2014, Broor et al., 2008, Kim et al., 2012, Ru-Ning et al., 2016).

Human Health Relevance

Human health and welfare is threatened by the occurrence of influenza like illness and other ARI. High rates of associated morbidity and mortality alone prompts comprehensive investigation into the causative viruses, associated epidemiology and approaches to mitigating disease burden. Non- influenza respiratory viruses are estimated to cost the global economy upwards of US$167 Billion per year, affecting 5-15% of the global population annually and accounting for up to 500,000 deaths per year (Xu et al., 2018, Zhang et al., 2014, Edwards et al., 2013). In cases of severe illness, there is some evidence for an increased disposition to mental illness (Ru-Ning et al., 2016). In Australia the cost of treating illnesses associated with the influenza virus in children is lower than the AUD$24-50 million per annum required to treat RSV related illnesses (Di Giallonardo et al., 2018a). The reduction of economic and health burden associated with the influenza virus has been associated with the development of effective vaccines, improved disease awareness and prevention and the development of targeted treatment plans, all of which relied heavily on genotypic and phenotypic monitoring of the responsible causative viruses (Liang et al., 2019, Dubois et al., 2019, Wang et al., 2019a). There are currently no targeted treatments or vaccines available for HMPV.

In recent studies HMPV has been identified in up to 40% of children hospitalised with acute respiratory illness and in as many as 13% all respiratory illness associated hospitalisations.

Serological studies estimate that almost all individuals will be exposed to HMPV by the age of 5 years old with infections in otherwise healthy individuals between the ages of 2-60 years old generally not requiring hospitalisation unless confounding factors, including chronic respiratory disorders, are present (Kaslow et al., 2014). Global data estimates HMPV detection rate of 3.9-7% in children hospitalised with lower respiratory tract infection, with outpatient detection rates ranging from 6.2-12% (Bernadette et al., 2001, McAdam et al., 2004, Williams et al., 2004). HMPV is a known cause of ARI in hospitalized adults and has increased morbidity and mortality in high risk populations, including immunosuppressed individuals, young children and the elderly (Falsey et al., 2003). A 2012 prospective study concluded that the hospitalisation rates of adults >50 years of age were statistically similar to those with influenza infections in the same region (Widmer et al., 2012). Like RSV, HMPV is a known causative pathogen of outbreaks in long-term care facilities (Falsey et al., 2003). In a 2002 study, it was reported that 96 of 364 residents of a long-term care facility developed respiratory disease over a 6-week period, of these, there were 6 confirmed HMPV infections, 3 of which resulted in patient death due to secondary pneumonia (Boivin et al., 2002b). A second study, published in 2007 also described an outbreak in a long- term care facility, where 26 of 171 residents were infected with HMPV, 8 patients developed pneumonia, of these patients 2 were hospitalized and 1 succumbed to the virus (Louie et al., 2007). The mean age of these patients was 70 years (Louie et al., 2007, Boivin et al., 2002b).

Respiratory infections remain a major global health burden and are considered the fourth most common cause of death in the elderly (Cunha et al., 2016). Clinical presentation of HMPV can range from mild respiratory illness to severe respiratory illness requiring hospitalisation and ventilation, unlike many other respiratory pathogens, HMPV has been reported as both focal and segmental infiltrates (Cesario, 2012). Documented symptoms of HMPV are broad, with characteristic clinical signs such as fever, cough, rhinitis, wheezing, bronchiolitis, asthma exacerbation, pneumonitis, community-acquired pneumonia and Chronic obstructive pulmonary disease (COPD) exacerbation (Karron et al., 2018). Since its discovery, several studies have shown that HMPV is clinically indistinguishable from closely related human-pathogenic paramyxovirus – RSV (Ditt et al., 2011, Bernadette et al., 2001, Schildgen et al., 2005b, van Den Hoogen, 2007, Sutherland et al., 2001, Linstow et al., 2008).

Biology of HMPV

Virion description

HMPV is highly pleomorphic in shape with both spherical and filamentous morphologies reported (Pangesti et al., 2018). The pleomorphic particles have been observed via electron microscopy, in infected cell supernatants measuring approximately 150 to 600nm, with short envelope projections of 13 to 17nm **(Figure 1)** (Chan et al., 2003). Viral infected LLC-MK2 supernatants have not indicated hemagglutination, similar to the closely related avian metapneumovirus (AMPV), viral replication relies heavily on the presence of trypsin (Xu et al., 2018).

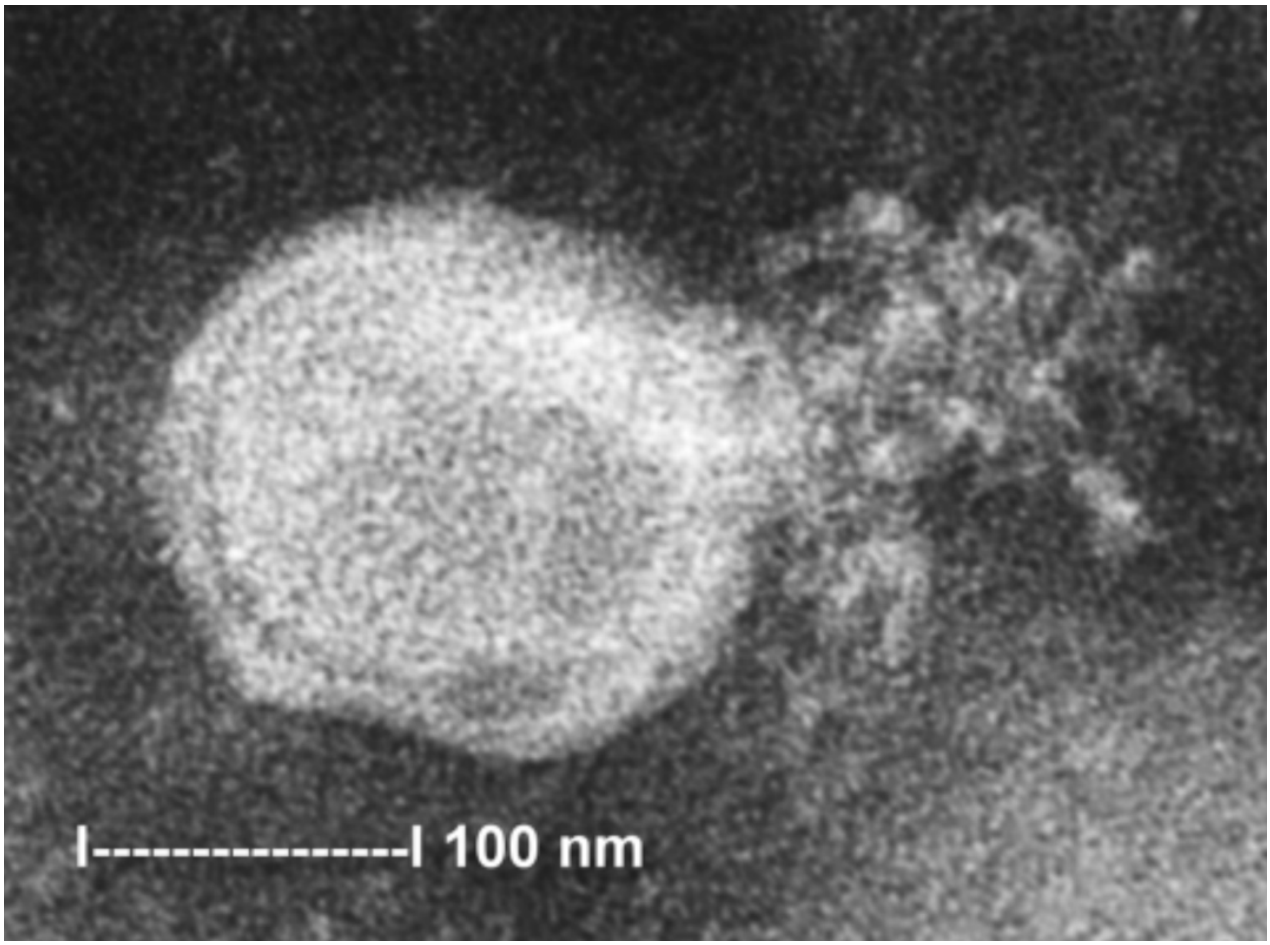


Figure 1: Electron micrograph of Human Metapneumovirus collected from the supernatant of rhesus monkey kidney (LLC-MK2) cell culture. A virion releasing a nucleocapsid is shown (Chan et al. 2003)

The virus genome is approximately 13.3kb in size and encodes eight genes, the genes occur in the following order: 3'-N-P-M-F-M2-SH-G-L-5' (Sutherland et al., 2001, El Najjar et al., 2016). The three transmembrane surface glycoproteins: the fusion protein (F), the attachment protein (G), and the small hydrophobic protein (SH), are densely packed on the viral envelope **(Figure 2, Table 1).**

Table 1: **Human Metapneumovirus protein size, and function.**

Gene	Protein	Approximate length (ammino acid)	Function
N	Nucleoprotein	390	RNA genome encapsulation
P	Phosphoprotein	290	Polymerase co-factor
M	Matrix protein	250	Aids in viral assembly and budding
F	Fusion protein	540	Virus-cell binding and membrane fusion
M2	M2-1 protein	190	RNA transcription processivity factor
	M2-2 protein	70	RNA transcription/replication regulation
SH	Small hydrophobic protein	170-180	Viroporin or innate immune inhibition
G	Attachment glycoprotein	230-240	Binds to cellular glycosaminoglycans
L	Large polymerase protein	2,000	Catalytic activity for viral replication

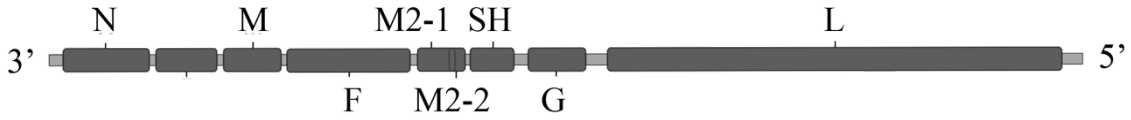


Figure 2: Schematic structure of the Human Metapneumovirus ssRNA genome. The 8 HMPV genomes are arranged in order from 3'-5'. Each gene is flanked by conserved sequences which are represented by the light grey shaded bar (Battles et al., 2017).

The lipid membrane envelope encapsulates the helical ribonucleoprotein (RNP) complex consisting of the nucleocapsid protein (N), phosphoprotein (P), large polymerase protein (L), and the M2-1 and M2-2 proteins. The N, P, a matrix protein (M2-1) and RNA dependent polymerase associated with the RNA genome together, form the RNA polymerase complex that carries out viral replication and transcription. The virion surface is comprised of three glycoproteins, the small-hydrophobic (SH) protein, the glycosylated (G) protein and the fusion (F) protein (Kamble et al., 2007). The role of the G and F protein has been readily defined as the catalyst of fusion and entry of the virus into a host cell through the heparan sulphate proteoglycan receptors of the host cells (Chang et al., 2012). The role of the SH protein is not completely understood, however some studies suggest it has a role in slowing and evading the host antiviral immune response by down regulating the type I IFN signalling pathway (Herfst et al., 2008, Herfst et al., 2004, Herd et al., 2010).

Studies over the past decade have increased our understanding of the mechanisms of entry and fusion of HMPV; however, the late stages in the replication cycle during which viral components

assemble and exit the cell are not well understood (Chang et al., 2012). The HMPV M protein, similar to other paramyxovirus M proteins, plays an essential role in production of virus particles, and budding of HMPV has been shown to occur in an endosomal sorting complex required for transport (ESCRT) independent manner (Sabo et al., 2011). Formation of HMPV virus like particles (VLPs) occurs following co-expression of the F and M proteins, with the G protein enhancing this process, thus indicating an important role for these proteins in the HMPV assembly process (Schowalter et al., 2006). In addition, morphogenesis of HMPV occurs at lipid-raft microdomains that are rich in actin, suggesting a role for lipid rafts and the actin cytoskeleton in virus assembly (Jumat et al., 2014). Initial sequencing of the viral genome revealed a close resemblance to fellow member of the *metapneumovirus* genus, avian metapneumovirus (AMPV). AMPV causes acute respiratory tract infections in avian species and shares up to 80% sequence identity with HMPV (Li et al., 2014). Despite HMPV and RSV being clinically indistinguishable, they differ greatly in genetic structure. HMPV lacks non-structural genes (NS1 And NS2) with different gene ordering. Recent advances in the imaging of a soluble form of the post-fusion hMPV F gene was captured allowing its structure to be deduced, interestingly extensive similarity was noted with the post-fusion RSV F gene despite having only ~38% sequence identity (Mas et al., 2016).

Taxonomy

Human metapneumovirus belongs to the order *Mononegavirales* and resides within the *Paramyxoviridae* (Embarek Mohamed et al., 2014, Wei et al., 2014). HMPV is further classified into the *Pneumovirinae* subfamily and further still into the *Metapneumovirus* genus, taking place as the first non-avian member of the Metapneumovirus genus (Karron et al., 2018, Bernadette et al., 2001).

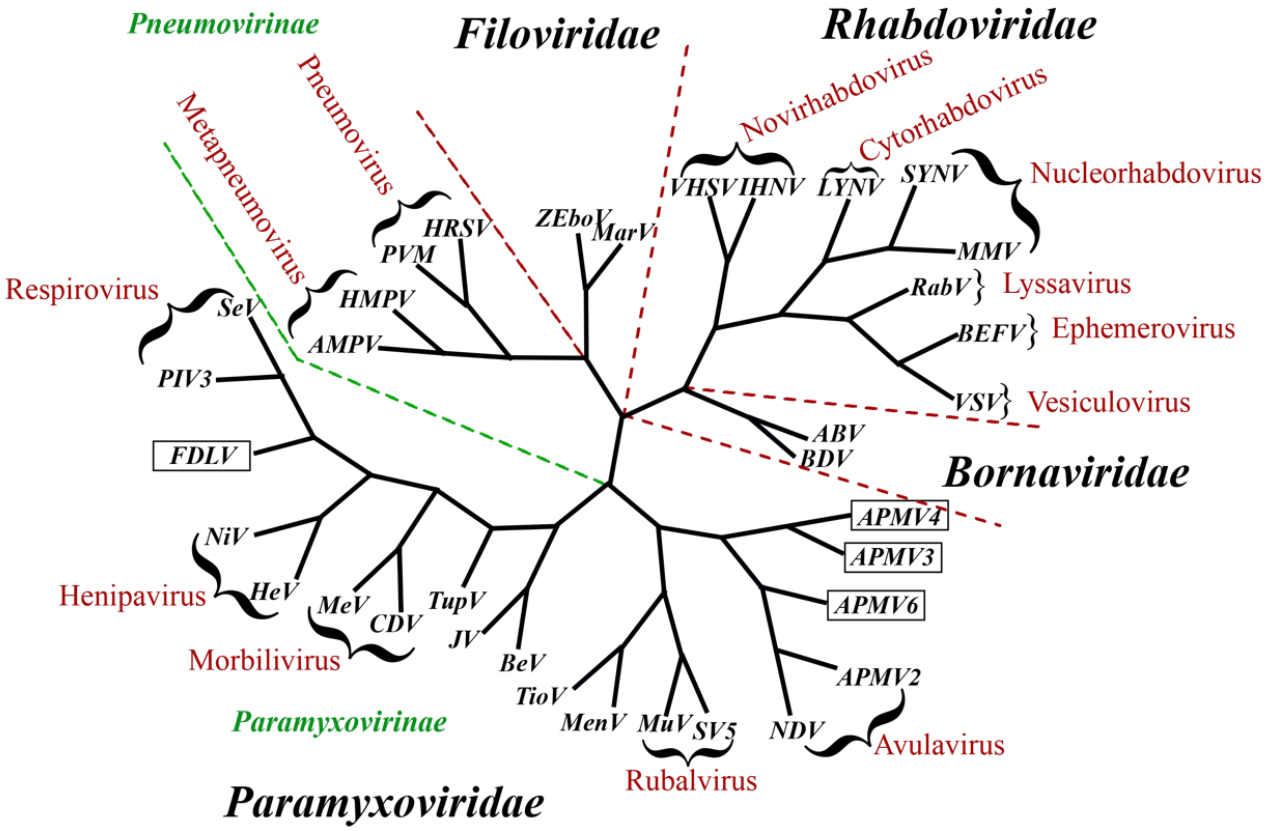


Figure 3: Un-rooted phylogenetic tree containing taxonomy of the Mononegavirales order (Adamson et al., 2013). Unrooted phylogenetic tree of members of the order Mononegavirales. The tree was constructed using the CLUSTALX and PHYLIP programs with the sequences of the conserved domain III of the polymerase proteins (Poch et al., 1990, Poch et al., 1989). Nine paramyxoviruses, not yet assigned to genera, are included (boxed): avian parainfluenza virus types 3 (APMV3), avian parainfluenza virus types 4 (APMV4), avian parainfluenza virus types 6 (APMV6), Beilong virus (BeV), Fer de Lance virus (FDLV), J virus (JV), Menangle virus (MenV), Tioman virus (TioV) and Tupaia paramyxovirus (TupV). Abbreviations: ABV, avian bornavirus; AMPV, avian metapneumovirus; APMV2, avian parainfluenza virus types 2; BDV, Borna disease virus; BEFV, bovine ephemeral fever virus; CDV, canine distemper virus; HeV, hendra virus; HMPV, human metapneumovirus; HRSV, human respiratory syncytial virus; IHNV, infectious hemorraghic necrosis virus; LNYV, lettuce necrotic yellows virus; MarV, Marburg virus; MeV, measles virus; MMV, maize mosaic virus; MuV - mumps virus; NDV, Newcastle disease virus; NiV, Nipah virus; PIV3, parainfluenza virus type 3; PVM, pneumonia virus of mice; RabV, rabies virus; SeV, Sendai virus; SV5, simian virus 5; SYNV, Sonchus yellow net virus; VSV, vesicular stomatitis Indiana virus; VHSV, viral hemorrhangic septicemia virus; ZeboV, Zaire ebolavirus.

Host infection and viral replication

Transmission of HMPV occurs via aerosolised viral particles and contaminated secretions, this includes saliva, droplets, or large particles (Sumino et al., 2005). HMPV infection occurs post contact with contaminated secretions and transmission is facilitated by the adhesion of viral particles to airway epithelial cells (Sumino et al., 2005). Whilst debate still surrounds the precise pathway of HMPV entry into host cells it is evident that both the G and F surface glycoproteins have important roles in the entry pathway of HMPV. The current accepted hypothesis is that the G gene mediates

initial HMPV virion interaction with host cell-surface proteoglycans (Battles et al., 2017). The G protein, the most variable across all HMPV isolates, is responsible for the attachment of the virus to the host cell. The unique amino acid sequence of the G attachment protein contains a single hydrophobic region located proximal to the N terminus. This hydrophobic region acts as an uncleaved signal peptide suggesting that the membrane-bound form of the virus acts as a membrane anchor. Adjacent cellular infection occurs due to the release of the virus post cell death, following intracellular replication (Kan-o et al., 2017, Hastings et al., 2015, Battles et al., 2017). The F protein promotes fusion of the viral and cell membrane post attachment and facilitates the entry of the viral ribonucleoprotein into the cell. The third transmembrane surface glycoprotein SH is a hydrophobic glycoprotein like RSV and AMPV. It contains 10 cysteine residues and has a high threonine/serine content of ~22% (van Den Hoogen et al., 2002). The role of the SH protein is to downregulate type I interferon (IFN) signalling by targeting signal transducer and activator of transcription-1 (STAT-1) (Hastings et al., 2015).

Clinical Presentation of HMPV Infections

Overview

The clinical presentation of HMPV is not dissimilar to other respiratory viruses, the disease often manifests as both upper and lower respiratory tract infections. Typical symptoms of HMPV include headache, sore throat, cough, rhinorrhoea, wheezing, difficulty breathing, hypoxia and other prodromal symptoms. Studies have shown that, typically, infected individuals present to health care practitioners within one week of becoming symptomatic (Shi et al., 2017). It has been hypothesised that HMPV shedding by infected individuals ranges from 6-28 days with a median PCR detection period of at 13 days. Typically an individual is non-infectious, determined by a negative culture at 13 days (Don and Korppi, 2011). HMPV is usually isolated to the respiratory tract, therefore making it unlikely to detect HMPV in any sample not derived from the respiratory tract. There has been one study which reported the identification of HMPV in serum using reverse-transcription-PCR (RT-PCR), however there have been no further studies to corroborate this (Maggi et al., 2003). A number of studies in nonhuman primates and other mammals have not been able to detect HMPV outside of the respiratory tract, further strengthening the hypothesis that HMPV exists solely in the respiratory tract (Cox et al., 2014, Cheemarla and Guerrero-Plata, 2015, Hastings et al., 2015, Boivin et al., 2002a, Li et al., 2014). HMPV can be associated with acute otitis media, where HMPV occasions otitis media, viral RNA can be detected in middle ear fluid (Schildgen et al., 2005a, Suzuki et al., 2005, Chang et al., 2012, Cox et al., 2012, Gioula et al., 2016). Reinfection with HMPV is not uncommon, this can be in the form of a different or the same strain, usually where reinfection has

occurred symptoms are usually limited to milder upper respiratory tract infections in otherwise healthy children or adults (Zhang et al., 2014). As the dominant HMPV strain may vary significantly by season, leading to antigenic escape and therefore become more likely to cause reinfection (Gioula and Kyriazopoulou-Dalaina, 2009, Ljubin-Sternak et al., 2008). HMPV infection has been identified as a trigger for respiratory disease exacerbation (Rudd et al., 2017). Several studies have been performed which suggest there may be a significant relationship between HMPV and the exacerbation of asthma (Edwards et al., 2013, Torres et al., 2012a, Williams et al., 2004). Several Maurine studies have also suggested airway hyperresponsiveness, a constituent of asthma, can be triggered by HMPV and persist *in vivo* (Rudd et al., 2017, Cheemarla and Guerrero-Plata, 2015).

Histopathological changes were observed in the epithelial cells of paediatric patients infected with HMPV (Wen et al., 2014, Aerts et al., 2015, Lai et al., 2016) that include the presence of hemosiderin-laden macrophages, red cytoplasmic inclusions, detached ciliary tufts and frequent neutrophils (Lai et al., 2016, Aerts et al., 2015). Studies of HMPV infections of non-human primates have demonstrated that the histopathologic changes caused by the virus remain restricted to the respiratory tract and remnants of the changes are not observed in any other organs (Nazly and Williams, 2018, Kaslow et al., 2014, Cheemarla and Guerrero-Plata, 2015, El Najjar et al., 2016). Thus, HMPV elicits primary tropism, which is confined to respiratory epithelia, as modelled in humans, cynomolgus macaques, mice, and rats. Viral expression of HMPV is commonly identified in the epithelial cells of nasal tissue through to the most distal part of the lungs, the bronchioles (Wen et al., 2014, McMichael et al., 2018). Reported radiological findings of hMPV lower respiratory tract infections show diffuse interstitial and alveolar infiltrates, bibasilar nodularity, opacities, small scattered nodules, basal shadowing, extensive air space consolidation, and pleural effusion (Boivin et al., 2002b, Boivin et al., 2002a, Maggi et al., 2003, Vicente et al., 2006, van Den Hoogen, 2007, Schildgen et al., 2011, Cesario, 2012, Dunn et al., 2013, Panda et al., 2014, Wen and Williams, 2015, Karron et al., 2018, Kamau et al., 2020). **Figure 4** shows chest computed tomography (CT) scans, alongside chest radiographs of paediatric patients with clinical abnormalities as a direct result of HMPV disease complications (Chu et al., 2014).

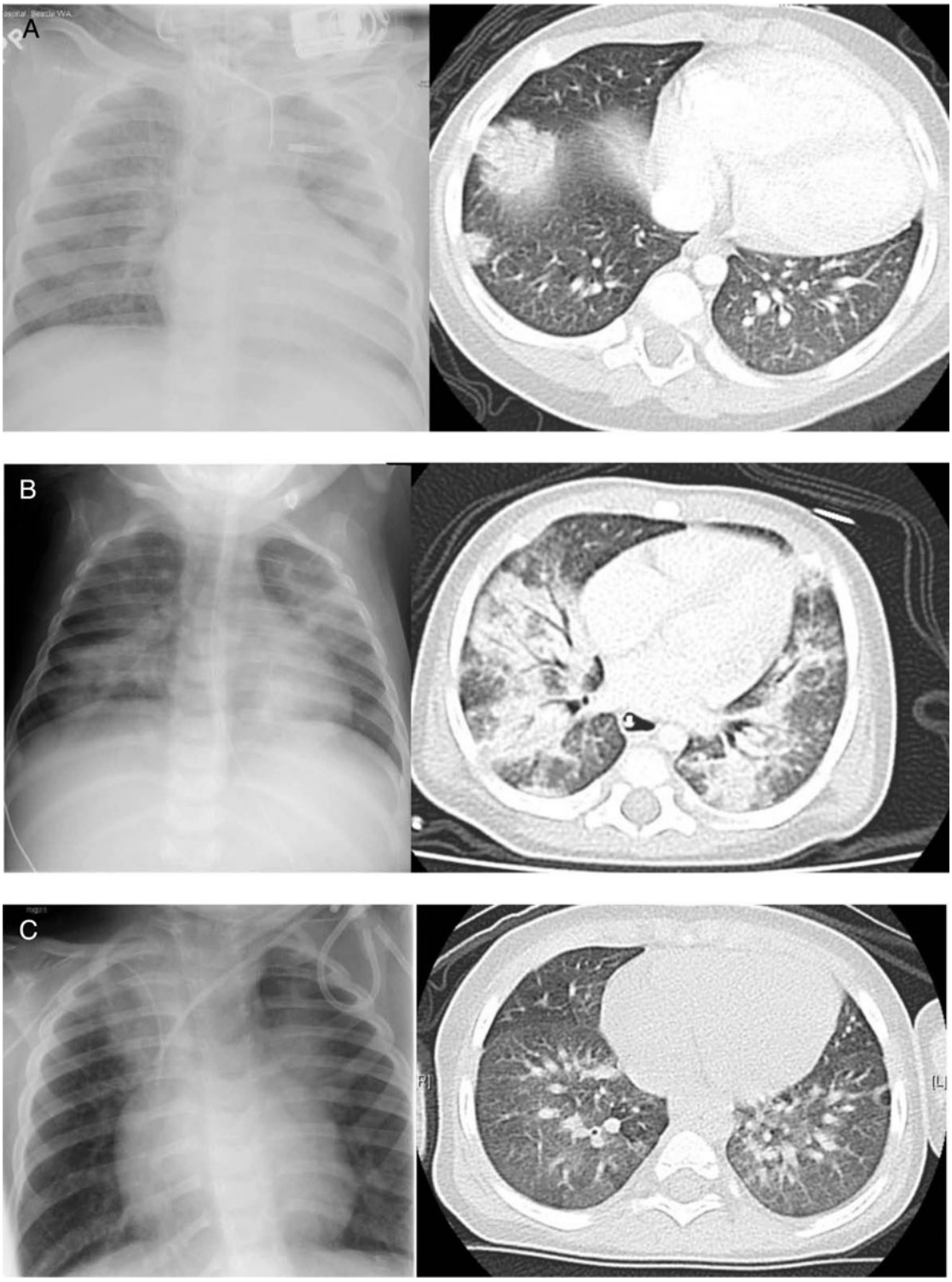


Figure 4 - Chest imaging findings of 3 immunocompromised patients infected with HMPV. (A) Patient 1 is an 11-month-old boy with severe combined immunodeficiency (SCID) who survived and cleared human metapneumovirus (hMPV) 511 days after diagnosis. His chest computed tomography (CT) scan shows right lower lobe consolidation. (B) Patient 2 is a 5-month-old girl with SCID who died of respiratory failure secondary to hMPV pneumonia. She has bilateral lower lobe consolidation on chest CT. (C) Patient 3 is a 2-year-old boy with aplastic anaemia who died of respiratory failure secondary to hMPV pneumonia. Chest CT shows airspace consolidation with air bronchograms (Jain et al., 2014).

At Risk Populations

Children

Children are often identified as key individuals at risk of severe HMPV infection (Edwards et al., 2013). HMPV causes significant disease throughout early childhood (Yan et al., 2017). Typically, the highest rate of HMPV associated clinic and emergency department visits occurred in paediatric patients between 6 and 11 months of age. Children are highly susceptible to coinfection which may have an impact on disease severity and symptom exacerbation although the full extent of this is yet to be quantified (Gioula and Kyriazopoulou-Dalaina, 2009, Lai et al., 2016). HMPV has been identified as a coinfection agent for children suffering severe upper and lower respiratory tract infections requiring hospitalisation. Many studies have reported co-infection of hMPV with other respiratory pathogens, including RSV, rhinovirus, enterovirus, parainfluenza virus, coronavirus, influenza A, and influenza B (Battles et al., 2017, Webster et al., 2014, Wei et al., 2013, Yang et al., 2009a). Co-infections were reported in most papers reviewed, in a study involving HMPV co-infection in paediatric cancer patients was estimated to account for 2.0-7.4% of the total virus distribution (Martinez-Rodriguez and Banos-Lara, 2020). HMPV co-infection has also been reported during an outbreak of severe acute respiratory syndrome (SARS). Studies have also found hMPV co-infection with bacterial pathogens like *Streptococcus pneumoniae*, *Mycoplasma pneumoniae*, and *Chlamydia pneumoniae* (Karron et al., 2018, Pangesti et al., 2018, Shi et al., 2017). Co-infections were described in detail by Torres et al., with HMPV reported as the sixth most frequent pathogen in sole infections, the fourth most frequent in mixed viral infections and the sixth most frequent in virus-bacteria mixed infections (Torres et al., 2012b).

Immunosuppressed Individuals

HMPV results in more severe disease in individuals who are immunocompromised, similarly to many respiratory viral infections including influenza virus (El Chaer et al., 2017, Adamson et al., 2013, Cheemarla and Guerrero-Plata, 2015, Nazly and Williams, 2018, Panda et al., 2014). Underlying diseases resulting in immunosuppression in HMPV studies have included solid tumours, underlying hematologic malignancy, organ transplant, rheumatologic disease, type one diabetes or are the recipients of immunosuppressant drugs (Adamson et al., 2013, Cheemarla and Guerrero-Plata, 2015, Panda et al., 2014). In some cohorts of paediatric patients with compromised immune systems HMPV has high morbidity and mortality. A recent metanalysis completed by Martinez-Rodriguez et al., report 1.3% of paediatric patients with cancer have been found to be positive for an HMPV infection and had a fatal outcome, whilst 2.3% of HMPV positive, transplant recipient, paediatric

cancer patients were reported to have infection related death (Martinez-Rodriguez and Banos-Lara, 2020). Using intravenous immunoglobulin and ribavirin to treat HMPV infected patients prior to the transplant demonstrated a drop from a 22% mortality rate in patients without treatment compared to 2% with treatment, and will be discussed further later in this review (Chu et al., 2014).

The Elderly

The presence of HMPV in long term aged care facilities, and it's increased propensity for morbidity and mortality in this environment is well documented (Cunha et al., 2016, Schuster et al., 2015, Zhang et al., 2014). Elderly patients for the purpose of this review are individuals over the age of 65 years old. A range of severe symptoms are often exhibited by elderly patients living in long term aged care facilities, these symptoms include respiratory failure, wheezing, bronchial wall thickening, fever and abnormal haematology (Ditt et al., 2011). A mouse model study performed by Darniot et al showed older mice developed more severe HMPV infections when compared to younger mice supported by increased mortality. A recent study identified an independent association between hypertension and anti-HMPV IgG titre concluding that there is a link between hypertension and HMPV in the elderly (Li et al., 2011). HMPV occasioning pneumonia in elderly patients has been reported as high as 50% of all positive cases and are usually accompanied by admission diagnoses such as COPD exacerbation, acute bronchitis, congestive heart failure and pneumonia (Reiche et al., 2014). Elderly patients were also reported to have a greater average length of hospitalization compared to any other age group with a median hospitalisation stay of 9 days for otherwise healthy elderly patients (Lim et al., 2020b, Cattoir et al., 2019, Panda et al., 2014). There has not been a clear cause identified for higher morbidity in elderly patients although it is hypothesised that quantitative and qualitative alteration of T cell response may contribute to delayed virus clearance.

Immune Response in Humans and Animals

Innate and Adaptive Immunity

Acute viral infection typically prompts an innate immune response the hallmark of which is induction of an antiviral type I interferon (IFN) response via JAK-STAT signalling pathways. (Kano et al., 2017, Nazly and Williams, 2018, Panda et al., 2014). Activation of the innate immune response is reliant upon the recognition of HMPV specific pathogen associated molecular patterns (PAMPs) by pattern recognition receptors (PRRs) present in the host (Velayutham et al., 2013). The HMPV specific PPRs include Toll-like receptors (TLRs) that are present in the endosomal compartments and cell surfaces. Previous studies have shown they are activated by viral infections

resulting in the regulation of the expression of proinflammatory and immune mediators (Leyrat et al., 2014). Toll-like receptor 4 (TLR4) plays an integral role in HMPV-induced secretion of proinflammatory cytokine and chemokines, alongside type I IFN and monocyte-derived dendritic cells (MoDCs) (Nazly and Williams, 2018). Single virus particle tracking and fusion assays have suggested that HMPV uses a bifurcated cellular entry strategy and therefore may be susceptible to inhibition by interferon-induced transmembrane protein (IFITM3)(Feeley et al., 2011).

HMPV antibodies are usually detected in individuals between the ages of 6 months and 1 year of age. The humoral response is a key component of adaptive immunity and is mediated by antibodies that target specific viral antigens. Higher levels of IgG and HMPV-neutralising antibodies are associated with protection against HMPV infection during cooler months (Falsey et al., 2009). In contrast, in animal models HMPV infection has been identified as eliciting an aberrant and somewhat weak adaptive immune response. Notably one study identified that the passive transfer of hyperimmune mouse derived HMPV-IgG sera to BALB/c mice, significantly increased passive immunity and decreased HMPV titres in the affected mouse lungs (Alvarez and Tripp, 2005).

Treatment and Prevention

Vaccine Development

At present, there are no commercially available or licenced vaccinations for the prevention of HMPV. Furthermore, treatment options are generic and not targeted for HMPV. Historically, approaches to the development of HMPV vaccines have been ineffective, more recently, the use of recombinant viral proteins and viral-vectored constructs has grown in popularity. A number of vaccine strategies have been evaluated and investigated, including protein-based recombinant vaccines, live-attenuated vaccines (LAV) and formalin-inactivated vaccines in animal models (Márquez-Escobar, 2017). To date, one LAV candidate has been evaluated in a phase I clinical study (registered at ClinicalTrials.gov, #NCT01255410). It is important to note that this study reported the vaccine candidate was not sufficient in providing protection in HMPV-seronegative children and all further studies pertained to its use was discontinued (Karron et al., 2018). Currently, consensus appears that LAV show greatest potential in eliciting both humoral and mucosal immunity and somewhat mimic natural viral replication routes. LAV are therefore regarded as highly suitable for both paediatric and geriatric immunization strategies (Cox et al., 2014, Dubois et al., 2019). Many potential targets have been identified for the rational design of epitope based HMPV vaccines, with preference given to those utilising vectored platforms. At present the most effective approaches to HMPV vaccine development appear to be live antigen delivery systems (viral vectors or live

attenuated viruses) which appear to cause effective T-cell memory formation. The consensus on HMPV vaccine development is that more research is required in the field of immunodominant CD8 and CD4 epitopes in HMPV proteins (Cheemarla and Guerrero-Plata, 2015, Zhang et al., 2014)

Antiviral Therapy

Ribavirin is a purine analog that is active against RNA and DNA viruses. At present aerosolised ribavirin has been approved with RSV related pneumonia. Since HMPV is phylogenetically and phenotypically related to RSV, the use of ribavirin was evaluated for the use against HMPV infection *in vitro* (Wyde et al., 2003). Whilst there is readily available anecdotal evidence to suggest ribavirin is effective in preventing HMPV related deaths in immunocompromised patients, studies utilising control groups and other statistically sound methods are not currently available (Collinson et al., 2021, El Chaer et al., 2017, Herd et al., 2010, Leyrat et al., 2014, McMichael et al., 2018). Studies have shown the use of ribavirin, with and without intravenous polyclonal immunoglobulins can be successful in treating HMPV infections (Pasikhova et al., 2018, Shachor-Meyouhas et al., 2011) it should be noted these studies are limited to small case series or case reports and require further replicated to be considered reliable. It is hypothesised ribavirin has a dual mode of action against HMPV infections, alongside the antiviral effect it also has been shown to awaken the hypo-responsive immune system after prolonged viremia (Safdar, 2008). With many unknowns and small sample sizes lacking control groups, the efficacy of ribavirin in preventing LRTI and mortality from these infections needs to be further investigated.

Another purine nucleotide analogue, favipiravir (T-705), has been evaluated for its use in the treatment of HMPV. Favipiravir works by inhibiting the RNA-dependent RNA polymerase, its use was evaluated in two clinical trials in the USA (registered at ClinicalTrials.gov #NCT02008344 and #NCT02026349). Both *in vitro* and *in vivo* results demonstrate antiviral potency in the inhibition of HMPV replication during early infection (Furuta et al., 2013). Anti-HMPV activity of favipiravir, has been hypothesised to be linked to HMPV polymerase inhibition, however it is accepted that additional studies are required before the progression of the drug to clinical trials (Jochmans et al., 2016).

Epidemiology

Virus seasonality

HMPV has a seasonal distribution which is similar to that of other viruses of the *Paramyxoviridae*, including RSV **(Figure 5)**. A study of patients admitted to the ICU in San

Sebastián (North of Spain) concluded two thirds of HMPV cases were detected in late winter-early spring which supports the hypothesis it occurs after the peak distribution of Influenza and RSV which occurs in early winter. Disease surveillance studies suggest multiple subgroups of HMPV circulate annually, with no disease severity variation associated with the genetically different strains. At present, the accepted method of HMPV detection is utilising real-time PCR, however serological diagnosis is possible using enzyme-linked immunoassays (ELISA), although this is not routinely used in clinical diagnostics. It is hypothesised due to the inability to distinguish HMPV infections from other acute respiratory illness pathogens, the true disease frequency is unknown in populations where HMPV is not tested for; this may include cases where the associated illness does not require hospitalisation or medical intervention.

HMPV infection is characteristically mild in otherwise healthy adults with increased disease severity and morbidity in the very young, elderly, and immunocompromised patients. Whilst HMPV infections have similar epidemiological features to RSV, infants typically between 6-23 months of age are more likely to be hospitalised with HMPV compared to RSV where the virus is most virulent in children under 6 months. International studies have reported the presence of HMPV in up to 40% of children hospitalised for acute respiratory illness. In infants less than 2 years of age the virus is rarely detected in asymptomatic individuals whilst almost all populations will experience a primary HMPV infection by age 5 (Panda et al., 2014). In a disease surveillance study conducted in Rochester, New York, HMPV was detected in up to 10% of hospitalised patients between 1999 to 2003 (Hildreth, 2009).

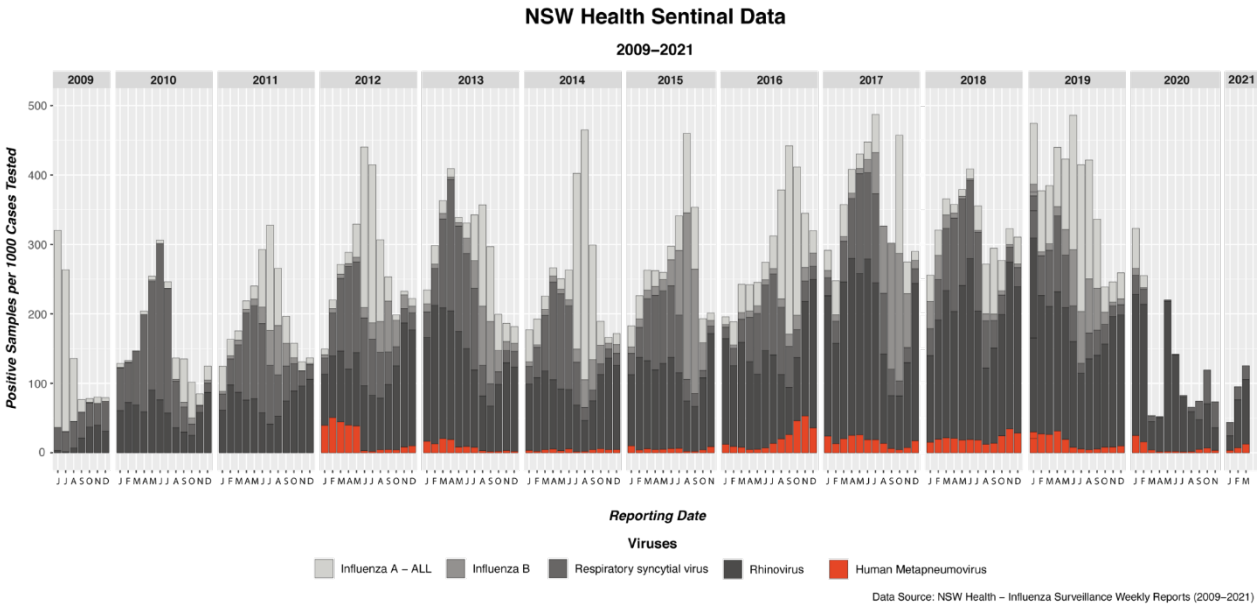


Figure 5 – **Seasonal distribution of common respiratory viruses found in NSW**, Influenza A, Influenza B, Respiratory syncytial virus, Rhinovirus and Human Metapneumovirus, Australia, and their seasonal distributions over the last 12 years. Data was sourced from the NSW influenza weekly surveillance reports. Data is further faceted into year, with each bar representing a month in that year. Data has been normalised to number of positive cases per 1000 samples tested.

Virus evolution and molecular epidemiology

The earliest description of HMPV hypothesised the presence of two HMPV subgroups based on genetic clustering observed in early phylogenetic analysis. In 2012, a submission was approved through the International Committee on Taxonomy of Viruses (ICTV, ICTV taxonomic proposal 2012.012V) to adopt official terminology for the HMPV nomenclature. There are two main types, referred to as "type A" and "type B" that are each comprised of 2 sublineages, A1, A2, B1, and B2. The distribution of published HMPV genomes can be seen in **Figure 6**. The nomenclature relies on G protein residues of the extracellular domain (Mackay et al., 2004). A 2006 study suggests a bipartition of subgroup A2 into two genetic clusters, A2a and A2b based on phylogenetic analysis of the F, N and G coding regions (Ljubin-Sternak et al., 2008, Huck et al., 2006). Phylogenetic analysis suggests the F protein between the A and B HMPV lineages is highly conserved with a maximum percent amino acid identity of 95-97%. In comparison, there was a maximum of 30-35% identity between G proteins. Long-term retrospective genome analysis of circulating HMPV subtypes in Uberlandia, Germany, South Korea and Croatia identified the co-circulation of multiple subtypes at any one given time (Carneiro et al., 2009, Chung et al., 2008). Previous studies have corroborated the cocirculation of HMPV strains suggesting multiple linages are capable of existing in the same population simultaneously.

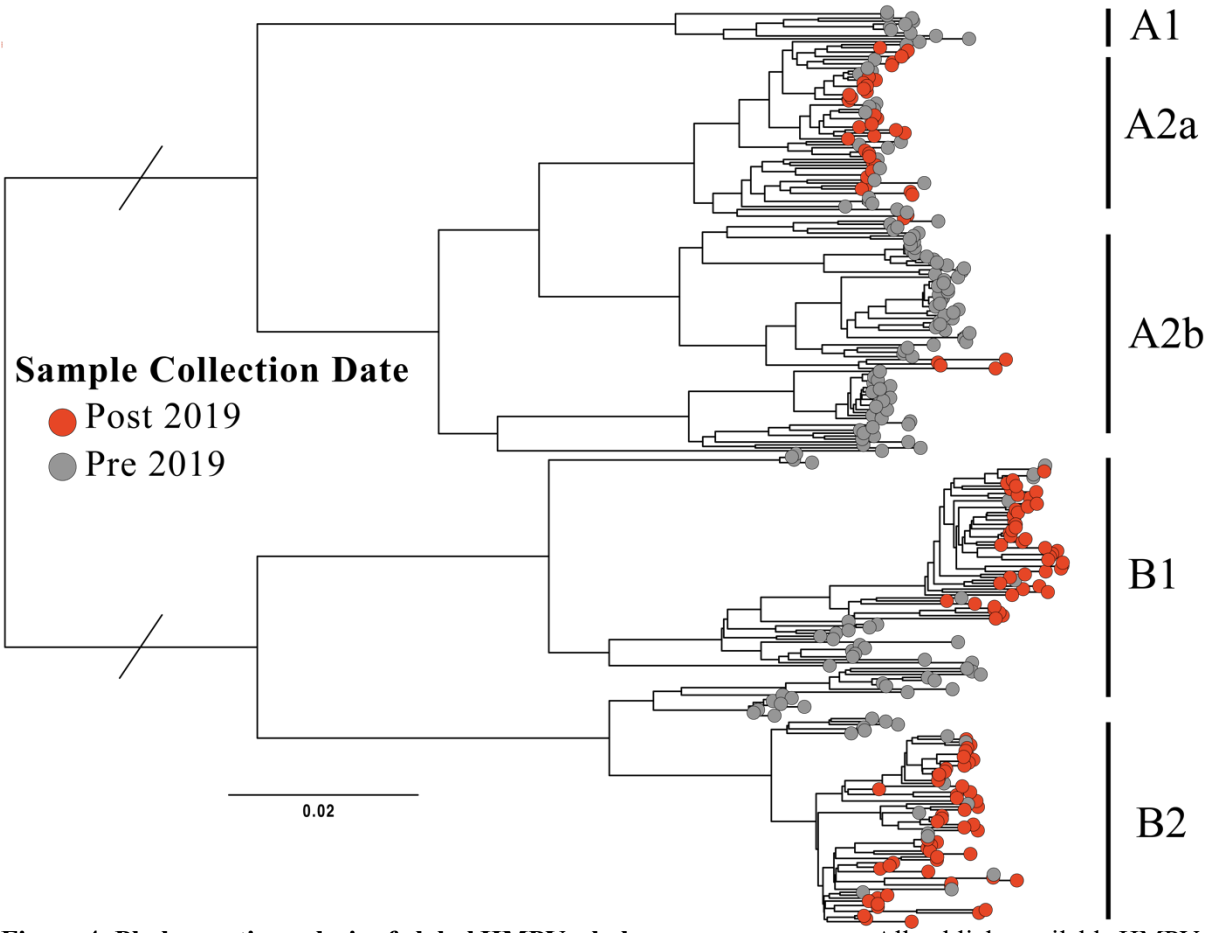


Figure 4: Phylogenetic analysis of global HMPV whole genome sequences. All publicly available HMPV whole genome sequences listed on NCBI GenBank, were downloaded. Collection date was used to classify samples into two categories, post 2019, capturing years 2019-2021 and pre-2019, anything collected before 2019. The 100 bootstrap, Phylogenetic maximum likelihood (PhyML) tree was created using Geneious Prime v2019.2.3).

Molecular epidemiological studies have revealed an epidemic peak of the A1 lineage epidemic occurred in 2002 with the A1 lineage rarely seen again, HMPV B1 and B2 lineages peaked in frequency in 2004. In contrast the A2 lineage has been observed in a long-long lasting global epidemic (Li et al., 2012). The A2 lineage adopts an atypical pattern which is likely due to the different prevalence of the A2a and A2b subtypes between the years of 2002 and 2006 (Kamau et al., 2020, Nazly and Williams, 2018). A more recent study between HMPV annual variations revealed prominent global epidemics of HMPV A2a in 2007, 2010, 2014, 2015, and 2016; B1 in 2011 and 2012; and B2 in 2008 and 2009. It appears a predominant HMPV genotype has fluctuated on a 3–4-year basis, oscillating between the three aforementioned strains with concurrent cocirculating strains.

Clinical Epidemiology

Clinical manifestations of HMPV range considerably in severity, from serious disease occasioning death to infrequently documented asymptomatic carriage. The observation of a diverse disease severity coupled with increasing genetic diversity has led to the hypothesis that a meaningful relationship likely exists between HMPV subtype and clinical severity (Chung et al., 2008), although this remains controversial. Vicente et al reported clinical severity of hMPV A infections was greater than that of hMPV B infections whilst Matsuzaki et al conveyed a relationship between specific symptoms and genotypes, for example wheezing was associated with HMPV B1 and B2 infections whilst laryngitis was associated with HMPV B1 infections (Matsuzaki et al., 2008, Vicente et al., 2006). Conversely there are several studies which report there is no correlation between genotype and disease severity (Mackay et al., 2004, Agapov et al., 2006, Wei et al., 2013). HMPV infection typically results in conspicuous disease, the illness is rarely reported in asymptomatic children or adults despite all populations experiencing primary HMPV infection by the age of 5. Studies have suggested that there is a significant underrepresentation of HMPV transmission and carriage in the global population due to asymptomatic carriers or mild cases going undetected (Don and Korppi, 2011). HMPV has a high presence in paediatric patients with initial HMPV infection occurring at around six months of age, likely due to the waning of maternal immunity (Paget et al., 2011).

Diagnostics

Cell Culture

A number of different cell lines have been used in the growth and isolation of HMPV, these include Vero cells, Hep-2 cells, Hep G2 cells, 293 cells and LLC-MK2 cells (Deffrasnes et al., 2005, Falsey et al., 2009, Sato et al., 2017). Recent studies suggest that the most suitable cell lines for the growth of HMPC are the human Chang conjunctiva cell line (clone 1-5Ca) and the feline kidney CRFK line (Nao et al., 2019). HMPV has a reported slow growth rate in cell culture, with delayed cytopathic effects, notably the rounding of cells and the detachment of cells from the culture matrix to small syncytium formations. For these reasons, direct fluorescence or ELISA-based assays are widely used for the detection of HMPV antigens in conjunction with cell culture. The sensitivity and specificity of cell culture methods were reportedly 68% and 99% respectively when compared to RT-PCR.

Fluorescent Immunoassays

Direct fluorescent antibody (DFA) testing is capable of directly identifying the presence of specific antigens with the fluorescently labelled antibodies, because of this, it has been widely used

in the detection of viruses in a clinical setting, where DFA is coupled with viral culture, sensitivity and specificity can be enhanced (Submaroney et al., 2022). Immunofluorescence microscopy requires extensive data interpretation, by determining the correct absorption wavelength required to excite the fluorophore tag bound to the antibody and detection of the released fluorescence reveals positive cells and indicate the presence of the target virus (Sato et al., 2017). DFA's have been used extensively in clinical diagnostics due to their simplicity and relatively rapid turn-around. In recent years with the increased availability of molecular diagnostic tools fluorescent immunoassays have lost popularity as a commercial diagnostic tool

Molecular diagnostics

The increasing availability and use of molecular detection of viruses has been critical to improved understand of epidemiology and informs efforts to improve the prevention of sudden acute respiratory infection causing viruses (Nao et al., 2019). HMPV diagnosis is usually dependent on the availability of RT-PCR due to culture yields being poor due to concurrent viral infections masking HMPV and its delayed cytopathic effect. RT-PCR has been found to have greater sensitivity, shown to be able to detect the virus in asymptomatic children with significantly lower viral loads, compared to symptomatic children. With the widely availability of molecular diagnostic tool in laboratories both locally and globally, multiplexing RT-PCR (mRT-PCR) has gained considerable popularity to detect HMPV in a rapid and more sensitive way (Paget et al., 2011). mRT-PCR is invaluable in time sensitive cases, where a suite of viruses can be tested for simultaneously. In principle a shortfall of RT-PCR include inability to detect variants where target amplicons undergo significant mutations leading to lower specificity in some circumstances (Subramoney et al., 2022). Most HMPV RT-PCR does not differentiate between A and B subtypes and therefore require additional investigation to determine subtype.

Introduction to the Scientific Problem

HMPV is now recognised as a leading cause of both lower and upper respiratory disease, especially in the immunocompromised, elderly, and young. Despite its propensity to cause severe disease in at-risk groups, we know very little with respect to the patterns of its genetic diversity and molecular epidemiology. Global surveillance of this disease is poor, despite it being identified on every continent. Whilst generally incidences of HMPV in the general population is mild, ongoing disease surveillance is necessary to increase our understanding of the disease, its emergence, pathogenicity and spread. Whole genome data for HMPV at present is lacking with less than 250 full length genome

sequences available globally (Tulloch et al., 2021). The lack of whole genome data potentially constrains further development of diagnostics and vaccines. As advancements are made in the affordability and feasibility of whole genome sequencing, we identified the need for a robust, whole genome sequencing method capable of capturing greater than 99% of the genome of all HMPV viruses to better define variant subtypes. We also identified large gaps in global and local HMPV strain identification and distribution, with no surveillance sequencing performed in NSW since the identification of the virus in the early 2000's (Tulloch et al., 2021a).

Aims of the Thesis and Chapter Outline

Ultimately the aim of the first half of this work is to enhance our current knowledge pertaining to the evolution, epidemiology and spread of HMPV within Australia. It is a hope that with the adoption of routine surveillance WGS a greater number of HMPV genomes may be made available for the purpose of phylogenetic analysis of historical and current circulating variants.

Chapter 2: An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus (*Viruses,* 2020).

The purpose of Chapter 2 was to develop a simple and robust approach for HMPV WGS sequencing which can be used to increase the breadth of HMPV whole genomes available for epidemiological analysis. A key component of this research is the development of a method which can utilise residual clinical extracts, meaning additional sampling is not required for ongoing disease surveillance.

Chapter 3: The Genomic Epidemiology of Human Metapneumovirus in Australia and the impact of COVID-19 (*Virus Evolution,* 2022, under review).

This research chapter has been constructed to provide greater genomic resolution of HMPV epidemiology in Australia over the period of 2013-2020 and highlight the importance of ongoing surveillance and sequencing of HMPV in the post-pandemic era, as previous patterns of disease may be subject to abrupt and unexpected change. In Chapter 3, we utilise the WGS method developed and evaluated in Chapter 2 to generate over 185 HMPV sequences generated from samples collected between 2013 and 2020. We utilise phylogenetic approaches to investigate the effect of non-pharmaceutical interventions in the suppression of co-circulating subtypes.

References

ADAMSON, P., THAMMAWAT, S., MUCHONDO, G., SADLON, T., WILLIAMS, J. & GORDON, D. 2013. hMPV Lineage Nomenclature and Heparin Binding. *Viruses,* 5**,** 2546-2547.

AERTS, L., RHÉAUME, C., CARBONNEAU, J., LAVIGNE, S., COUTURE, C., HAMELIN, M.-È. & BOIVIN, G. 2015. Adjuvant effect of the human metapneumovirus (HMPV) matrix protein in HMPV subunit vaccines. *The Journal of general virology,* 96**,** 767.

AGAPOV, E., SUMINO, K., GAUDREAULT-KEENER, M., STORCH, G. & HOLTZMAN, M. 2006. Genetic Variability of Human Metapneumovirus Infection: Evidence of a Shift in Viral Genotype without a Change in Illness. *Journal of Infectious Diseases,* 193**,** 396-403.

ALVAREZ, R. & TRIPP, R. A. 2005. The Immune Response to Human Metapneumovirus Is Associated with Aberrant Immunity and Impaired Virus Clearance in BALB/c Mice. *Journal of Virology,* 79**,** 5971-5978.

BATTLES, M., MÁS, V., OLMEDILLAS, E., CANO, O., VÁZQUEZ, M., RODRÍGUEZ, L., MELERO, J. & MCLELLAN, J. 2017. Structure and immunogenicity of pre-fusion-stabilized human metapneumovirus F glycoprotein. *Nature Communications,* 8**,** 1-11.

BERNADETTE, G. V. D. H., JAN, C. D. J., JAN, G., THIJS, K., RONALD DE, G., RON, A. M. F. & ALBERT, D. M. E. O. 2001. A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nature Medicine,* 7**,** 719.

BIACCHESI, S., SKIADOPOULOS, M. H., BOIVIN, G., HANSON, C. T., MURPHY, B. R., COLLINS, P. L. & BUCHHOLZ, U. J. 2003. Genetic diversity between human metapneumovirus subgroups. *Virology,* 315**,** 1-9.

BOIVIN, G., ABED, Y., PELLETIER, G., RUEL, L., MOISAN, D., COTE, S., PERET, T., ERDMAN, D. & ANDERSON, L. 2002a. Virological features and clinical manifestations associated with human metapneumovirus: A new paramyxovirus responsible for acute respiratory-tract infections in all age groups. *Journal of Infectious Diseases,* 186**,** 1330-1334.

BOIVIN, G., ABED, Y., PELLETIER, G., RUEL, L., MOISAN, D., COTE, S., PERET, T. C. T., ERDMAN, D. D. & ANDERSON, L. J. 2002b. Virological features and clinical manifestations associated with human metapneumovirus: A new Paramyxovirus responsible for acute respiratory-tract infections in all age groups. (Concise Communication). *Journal of Infectious Diseases,* 186**,** 1330.

BROOR, S., BHARAJ, P. & CHAHAR, H. 2008. Human metapneumovirus: a new respiratory pathogen. *Journal of Biosciences,* 33**,** 483-493.

CARNEIRO, B. M., YOKOSAWA, J., ARBIZA, J., COSTA, L. F., MIRAZO, S., NEPOMUCENO, L. L., OLIVEIRA, T. F. M., GOULART, L. R., VIEIRA, C. U., FREITAS, G. R. O., PAULA, N. T. & QUEIRÓZ, D. A. O. 2009. Detection of all four human metapneumovirus subtypes in nasopharyngeal specimens from children with respiratory disease in Uberlândia, Brazil. *Journal of Medical Virology,* 81**,** 1814-1818.

CESARIO, T. C. 2012. Viruses associated with pneumonia in adults. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America,* 55**,** 107-113.

CHAN, P. K. S., TAM, J. S., LAM, C.-W., CHAN, E., WU, A., LI, C.-K., BUCKLEY, T. A., NG, K.-C., JOYNT, G. M., CHENG, F. W. T., TO, K.-F., LEE, N., HUI, D. S. C., CHEUNG, J. L. K., CHU, I., LIU, E., CHUNG, S. S. C. & SUNG, J. J. Y. 2003. Human metapneumovirus detection in patients with severe acute respiratory syndrome.(Research). *Emerging Infectious Diseases,* 9**,** 1058.

CHANG, A., MASANTE, C., BUCHHOLZ, U. J. & DUTCH, R. E. 2012. Human Metapneumovirus (HMPV) Binding and Infection Are Mediated by Interactions between the HMPV Fusion Protein and Heparan Sulfate. *Journal of Virology,* 86**,** 3230.

CHEEMARLA, N. & GUERRERO-PLATA, A. 2015. Immune Response to Human Metapneumovirus Infection: What We Have Learned from the Mouse Model. *Pathogens,* 4**,** 682-696.

CHU, H. Y., RENAUD, C., FICKEN, E., THOMSON, B., KUYPERS, J. & ENGLUND, J. A. 2014. Respiratory Tract Infections Due to Human Metapneumovirus in Immunocompromised Children. *Journal of the Pediatric Infectious Diseases Society,* 3**,** 286-293.

CHUNG, J. Y., HAN, T. H., KIM, S. W. & HWANG, E. S. 2008. Genotype variability of human metapneumovirus, South Korea. *Journal of Medical Virology,* 80**,** 902-905.

COLLINSON, N., SNAPE, N., BEAGLEY, K., FANTINO, E. & SPANN, K. 2021. COPD Is Associated with Elevated IFN-β Production by Bronchial Epithelial Cells Infected with RSV or hMPV. *Viruses,* 13**,** 911.

COX, R. G., ERICKSON, J. J., HASTINGS, A. K., BECKER, J. C., JOHNSON, M., CRAVEN, R. E., TOLLEFSON, S. J., BOYD, K. L. & WILLIAMS, J. V. 2014. Human metapneumovirus virus-like particles induce protective B and T cell responses in a mouse model. *Journal of virology,* 88**,** 6368-6379.

COX, R. G., LIVESAY, S. B., JOHNSON, M., OHI, M. D. & WILLIAMS, J. V. 2012. The Human Metapneumovirus Fusion Protein Mediates Entry via an Interaction with RGD-Binding Integrins. *Journal of Virology,* 86**,** 12148.

CUNHA, B. A., IRSHAD, N. & CONNOLLY, J. J. 2016. Adult human metapneumonovirus (hMPV) pneumonia mimicking Legionnaire's disease. *Heart & Lung - The Journal of Acute and Critical Care,* 45.

DEFFRASNES, C., CÔTÉ, S. & BOIVIN, G. 2005. Analysis of Replication Kinetics of the Human Metapneumovirus in Different Cell Lines by Real-Time PCR. *Journal of Clinical Microbiology,* 43**,** 488-490.

DHANASEKARAN, V., SULLIVAN, S., EDWARDS, K. M., XIE, R., KHVOROV, A., VALKENBURG, S. A., COWLING, B. J. & BARR, I. G. 2022. Human seasonal influenza under COVID-19 and the potential consequences of influenza lineage elimination. *Nature communications,* 13**,** 1721-1721.

DI GIALLONARDO, F., KOK, J., FERNANDEZ, M., CARTER, I., GEOGHEGAN, J. L., DWYER, D. E., HOLMES, E. C. & EDEN, J.-S. 2018. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses,* 10.

DITT, V., LUSEBRINK, J., TILLMANN, R. L., SCHILDGEN, V. & SCHILDGEN, O. 2011a. Respiratory Infections by HMPV and RSV Are Clinically Indistinguishable but Induce Different Host Response in Aged Individuals.(Research Article). *PLoS ONE,* 6**,** e16314.

DITT, V., LÜSEBRINK, J., TILLMANN, R. L., SCHILDGEN, V. & SCHILDGEN, O. 2011b. Respiratory infections by HMPV and RSV are clinically indistinguishable but induce different host response in aged individuals. *PloS one,* 6**,** e16314-e16314.

DON, M. & KORPPI, M. 2011. Serologically defined human metapneumovirus in pediatric community-acquired pneumonia. *The Journal of Pediatrics,* 158**,** 173-173.

DUBOIS, J., PIZZORNO, A., CAVANAGH, M.-H., PADEY, B., NICOLAS DE LAMBALLERIE, C., UYAR, O., VENABLE, M.-C., CARBONNEAU, J., TRAVERSIER, A., JULIEN, T., LAVIGNE, S., COUTURE, C., LINA, B., HAMELIN, M.-È., TERRIER, O., ROSA-CALATRAVA, M. & BOIVIN, G. 2019. Strain-Dependent Impact of G and SH Deletions Provide New Insights for Live-Attenuated HMPV Vaccine Development. *Vaccines (Basel),* 7**,** 164.

DUNN, S. R., RYDER, A. B., TOLLEFSON, S. J., XU, M., SAVILLE, B. R. & WILLIAMS, J. V. 2013. Seroepidemiologies of Human Metapneumovirus and Respiratory Syncytial Virus in Young Children, Determined with a New Recombinant Fusion Protein Enzyme-Linked Immunosorbent Assay. *Clinical and Vaccine Immunology,* 20**,** 1654.

EDWARDS, K. M., ZHU, Y., GRIFFIN, M. R., WEINBERG, G. A., HALL, C. B., SZILAGYI, P. G., STAAT, M. A., IWANE, M., PRILL, M. M. & WILLIAMS, J. V. 2013. Burden of Human Metapneumovirus Infection in Young Children. *The New England Journal of Medicine,* 368**,** 633-643.

EL CHAER, F., SHAH, D. P., KMEID, J., ARIZA-HEREDIA, E. J., HOSING, C. M., MULANOVICH, V. E. & CHEMALY, R. F. 2017. Burden of human metapneumovirus infections in patients with cancer: Risk factors and outcomes. *Cancer,* 123**,** 2329-2337.

EL NAJJAR, F., CIFUENTES-MUNOZ, N., CHEN, J., ZHU, H., BUCHHOLZ, U. J., MONCMAN, C. L. & DUTCH, R. E. 2016. Human metapneumovirus Induces Reorganization of the Actin Cytoskeleton for Direct Cell-to-Cell Spread.(Research Article). *PLoS Pathogens,* 12**,** e1005922.

EMBAREK MOHAMED, M., REICHE, J., JACOBSEN, S., THABIT, A., BADARY, M., BRUNE, W., SCHWEIGER, B. & OSMANN, A. 2014. Molecular Analysis of Human Metapneumovirus Detected in

Patients with Lower Respiratory Tract Infection in Upper Egypt. *International Journal of Microbiology,* 2014.

FALSEY, A. R., ERDMAN, D., ANDERSON, L. J. & WALSH, E. E. 2003. Human metapneumovirus infections in young and elderly adults. (Major Article). *Journal of Infectious Diseases,* 187**,** 785.

FALSEY, A. R., HENNESSEY, P. A., FORMICA, M. A., CRIDDLE, M. M., BIEAR, J. M. & WALSH, E. E. 2009. Humoral immunity to human metapneumovirus infection in adults. *Vaccine,* 28**,** 1477-1480.

FEELEY, E. M., SIMS, J. S., JOHN, S. P., CHIN, C. R., PERTEL, T., CHEN, L.-M., GAIHA, G. D., RYAN, B. J., DONIS, R. O. & ELLEDGE, S. J. 2011. IFITM3 inhibits influenza A virus infection by preventing cytosolic entry. *PLoS pathogens,* 7.

GIOULA, G. & KYRIAZOPOULOU-DALAINA, V. 2009. Human metapneumovirus and pediatric patients: A mini review. 04**,** 351-356.

GIOULA, G., MELIDOU, A., EXINDARI, M., PAPALEXIS, P., XANTHIS, D. & MALISIOVAS, N. 2016. The burden of influenza, respiratory syncytial virus and human metapneumovirus in infants and young children in N. Greece, 2004-2013.(Original Article)(Report). *Acta Medica International,* 3**,** 154.

GREENSILL, J., MCNAMARA, P. S., DOVE, W., FLANAGAN, B., SMYTH, R. L. & HART, C. A. 2003. Human metapneumovirus in severe respiratory syncytial virus bronchiolitis. (Dispatches). *Emerging Infectious Diseases,* 9**,** 372.

HASTINGS, A. K., ERICKSON, J. J., SCHUSTER, J. E., BOYD, K. L., TOLLEFSON, S. J., JOHNSON, M., GILCHUK, P., JOYCE, S. & WILLIAMS, J. V. 2015. Role of type I interferon signaling in human metapneumovirus pathogenesis and control of viral replication. *Journal of virology,* 89**,** 4405-4420.

HERD, K. A., NELSON, M., MAHALINGAM, S. & TINDLE, R. W. 2010. Pulmonary infection of mice with human metapneumovirus induces local cytotoxic T-cell and immunoregulatory cytokine responses similar to those seen with human respiratory syncytial virus. *Journal of general virology,* 91**,** 1302-1310.

HERFST, S., DE GRAAF, M., SCHICKLI, J. H., TANG, R. S., KAUR, J., YANG, C. F., SPAETE, R., HALLER, A. A., VAN DEN HOOGEN, B. G., OSTERHAUS, A. & FOUCHIER, R. 2004. Recovery of Human Metapneumovirus Genetic Lineages A and B from Cloned cDNA. *Journal of Virology,* 78**,** 8264-8270.

HERFST, S., SCHRAUWEN, E. J. A., DE GRAAF, M., VAN AMERONGEN, G., VAN DEN HOOGEN, B. G., DE SWART, R. L., OSTERHAUS, A. D. M. E. & FOUCHIER, R. A. M. 2008. Immunogenicity and efficacy of two candidate human metapneumovirus vaccines in cynomolgus macaques. *Vaccine,* 26**,** 4224-4230.

HILDRETH, C. 2009. Human Metapneumovirus Infections in Adults: Another Piece of the Puzzle. *JAMA,* 301**,** 812.

HUCK, B., SCHARF, G., NEUMANN-HAEFELIN, D., PUPPE, W., WEIGL, J. & FALCONE, V. 2006. Novel human metapneumovirus sublineage.(DISPATCHES). *Emerging Infectious Diseases,* 12**,** 147.

JAIN, B., SINGH, A. K., DANGI, T., AGARWAL, A., VERMA, A. K., DWIVEDI, M., SINGH, K. P. & JAIN, A. 2014. High prevalence of human metapneumovirus subtype B in cases presenting as severe acute respiratory illness: an experience at tertiary care hospital. *The Clinical Respiratory Journal,* 8**,** 225-33.

JAIN, S., SELF, W. H., WUNDERINK, R. G., FAKHRAN, S., BALK, R., BRAMLEY, A. M., REED, C., GRIJALVA, C. G., ANDERSON, E. J., COURTNEY, D. M., CHAPPELL, J. D., QI, C., HART, E. M., CARROLL, F., TRABUE, C., DONNELLY, H. K., WILLIAMS, D. J., ZHU, Y., ARNOLD, S. R., AMPOFO, K., WATERER, G. W., LEVINE, M., LINDSTROM, S., WINCHELL, J. M., KATZ, J. M., ERDMAN, D., SCHNEIDER, E., HICKS, L. A., MCCULLERS, J. A., PAVIA, A. T., EDWARDS, K. M. & FINELLI, L. 2015. Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *The New England journal of medicine,* 373**,** 415-427.

JUMAT, M., NGUYEN HUONG, T., WONG, P., LOO, L., TAN, B., FENWICK, F., TOMS, G. & SUGRUE, R. 2014. Imaging analysis of human metapneumovirus-infected cells provides evidence for the involvement of F-actin and the raft-lipid microdomains in virus morphogenesis. *Virology Journal,* 11.

KAMAU, E., OKETCH, J., DE LAURENT, Z., PHAN, M., AGOTI, C., NOKES, D. & COTTEN, M. 2020. Whole genome sequencing and phylogenetic analysis of human metapneumovirus strains from Kenya and Zambia. *BMC Genomics,* 21**,** 1-10.

KAMBLE, R. T., BOLLARD, C., DEMMLER, G., LASALA, P. R. & CARRUM, G. 2007. Human metapneumovirus infection in a hematopoietic transplant recipient. *Bone Marrow Transplantation,* 40**,** 699.

KAN-O, K., RAMIREZ, R., MACDONALD, M. I., ROLPH, M., RUDD, P. A., SPANN, K. M., MAHALINGAM, S., BARDIN, P. G. & THOMAS, B. J. 2017. Human Metapneumovirus Infection in Chronic Obstructive Pulmonary Disease: Impact of Glucocorticosteroids and Interferon. *The Journal of Infectious Diseases,* 215**,** 1536-1545.

KARRON, R. A., SAN MATEO, J., WANIONEK, K., COLLINS, P. L. & BUCHHOLZ, U. J. 2018. Evaluation of a Live Attenuated Human Metapneumovirus Vaccine in Adults and Children. *Journal of the Pediatric Infectious Diseases Society,* 7**,** 86-89.

KASLOW, R. A., STANBERRY, L. R. & LE DUC, J. W. 2014. *Viral Infections of Humans Epidemiology and Control,* Boston, MA, Springer US.

KIM, H. R., CHO, A. R., LEE, M.-K., YUN, S. W. & KIM, T.-H. 2012. Genotype Variability and Clinical Features of Human Metapneumovirus Isolated from Korean Children, 2007 to 2010. *The Journal of Molecular Diagnostics,* 14**,** 61-64.

LAI, S.-H., LIAO, S.-L., WONG, K.-S. & LIN, T.-Y. 2016. Preceding human metapneumovirus infection increases adherence of Streptococcus pneumoniae and severity of murine pneumococcal pneumonia. *Journal of Microbiology, Immunology and Infection,* 49**,** 216-224.

LEYRAT, C., RENNER, M., HARLOS, K., HUISKONEN, J. T., GRIMES, J. M. & DÖTSCH, V. 2014. Drastic changes in conformational dynamics of the antiterminator M2-1 regulate transcription efficiency in Pneumovirinae. *eLife,* 3.

LI, J., REN, L., GUO, L., PARANHOS-BACCALÀ, G., VERNET, G. & WANG, J. 2012. Evolutionary Dynamics Analysis of Human Metapneumovirus Subtype A2: Genetic Evidence for Its Dominant Epidemic. *PLoS One,* 7**,** e34544.

LI, W., ZHU, S., SHE, R., HU, F., WANG, J., XU, Y., ZHANG, C., LIU, S., QUAN, R., LI, Z., DU, F., TING, W. & LIU, J. 2014. Viral Replication and Lung Lesions in BALB/c Mice Experimentally Inoculated with Avian Metapneumovirus Subgroup C Isolated from Chickens. *PLoS One,* 9**,** e92136.

LI, Z., RUI, C., KEIICHIRO, I., MIKIKO, A., YUMI, K., KAZUTO, S., OSAMU, I., MASASHI, O. & SHIGETO, M. 2011. Association between human metapneumovirus seroprevalence and hypertension in elderly subjects in a long-term care facility. *Hypertension Research,* 34**,** 474.

LINSTOW, M.-L., HØGH, M., NORDBØ, S., EUGEN-OLSEN, J., KOCH, A. & HØGH, B. 2008. A community study of clinical traits and risk factors for human metapneumovirus and respiratory syncytial virus infection during the first year of life. *European Journal of Pediatrics,* 167**,** 1125-1133.

LJUBIN-STERNAK, S., ŠANTAK, M., ČEPIN-BOGOVIĆ, J., BAĆE, A., VOJNOVIĆ, G., MLINARIĆ-GALINOVIĆ, G., FORČIĆ, D., DRAŽENOVIĆ, V. & FALSEY, A. R. 2008. Detection of genetic lineages of human metapneumovirus in Croatia during the winter season 2005/2006. *Journal of Medical Virology,* 80**,** 1282-1287.

LOUIE, J. K., SCHNURR, D. P., PAN, C.-Y., KIANG, D., CARTER, C., TOUGAW, S., VENTURA, J., NORMAN, A., BELMUSTO, V., ROSENBERG, J. & TROCHET, G. 2007. A summer outbreak of human metapneumovirus infection in a long-term-care facility. *Journal of Infectious Diseases,* 196**,** 705.

MACKAY, I. M., BIALASIEWICZ, S., WALIUZZAMAN, Z., CHIDLOW, G. R., FEGREDO, D. C., LAINGAM, S., ADAMSON, P., HARNETT, G. B., RAWLINSON, W., NISSEN, M. D. & SLOOTS, T. P. 2004. Use of the P gene to genotype human metapneumovirus identifies 4 viral subtypes. *The Journal of infectious diseases,* 190**,** 1913-1918.

MAGGI, F., PIFFERI, M., VATTERONI, M., FORNAI, C., TEMPESTINI, E., ANZILOTTI, S., LANINI, L., ANDREOLI, E., RAGAZZO, V., PISTELLO, M., SPECTER, S. & BENDINELLI, M. 2003. Human

Metapneumovirus Associated with Respiratory Tract Infections in a 3-Year Study of Nasal Swabs from Infants in Italy. *Journal of Clinical Microbiology,* 41**,** 2987.

MÁRQUEZ-ESCOBAR, V. A. 2017. Current developments and prospects on human metapneumovirus vaccines. *Expert review of vaccines,* 16**,** 419-431.

MAS, V., RODRIGUEZ, L., OLMEDILLAS, E., CANO, O., PALOMO, C., TERRON, M. C., LUQUE, D., MELERO, J. A. & MCLELLAN, J. S. 2016. Engineering, Structure and Immunogenicity of the Human Metapneumovirus F Protein in the Postfusion Conformation.(Research Article). *PLoS Pathogens,* 12**,** e1005859.

MATSUZAKI, Y., ITAGAKI, T., ABIKO, C., AOKI, Y., SUTO, A. & MIZUTA, K. 2008. Clinical impact of human metapneumovirus genotypes and genotype-specific seroprevalence in Yamagata, Japan. *Journal of Medical Virology,* 80**,** 1084-1089.

MCADAM, A. J., HASENBEIN, M. E., FELDMAN, H. A., COLE, S. E., OFFERMANN, J. T., RILEY, A. M. & LIEU, T. A. 2004. Human metapneumovirus in children tested at a tertiary-care hospital.(Major Article). *Journal of Infectious Diseases,* 190**,** 20.

MCMICHAEL, T. M., ZHANG, Y., KENNEY, A. D., ZHANG, L., ZANI, A., LU, M., CHEMUDUPATI, M., LI, J. & YOUNT, J. S. 2018. IFITM3 Restricts Human Metapneumovirus Infection. *The Journal of Infectious Diseases,* 218**,** 1582-1591.

NAO, N., SATO, K., YAMAGISHI, J., TAHARA, M., NAKATSU, Y., SEKI, F., KATOH, H., OHNUMA, A., SHIROGANE, Y., HAYASHI, M., SUZUKI, T., KIKUTA, H., NISHIMURA, H. & TAKEDA, M. 2019. Consensus and variations in cell line specificity among human metapneumovirus strains. *PloS one,* 14**,** e0215822-e0215822.

NAZLY, S. & WILLIAMS, J. 2018. Human metapneumovirus - what we know now. *F1000Research,* 7.

PAGET, S. P., ANDRESEN, D. N., KESSON, A. M. & EGAN, J. R. 2011. Comparison of human metapneumovirus and respiratory syncytial virus in children admitted to a paediatric intensive care unit. *Journal of Paediatrics and Child Health,* 47**,** 737-741.

PANDA, S., MOHAKUD, N. K., PENA, L. & KUMAR, S. 2014. Human metapneumovirus: review of an important respiratory pathogen. *International Journal of Infectious Diseases,* 25**,** 45-52.

PANGESTI, K. N. A., ABD EL GHANY, M., WALSH, M. G., KESSON, A. M. & HILL-CAWTHORNE, G. A. 2018. Molecular epidemiology of respiratory syncytial virus. *Reviews in Medical Virology,* 28**,** n/a-n/a.

PASIKHOVA, Y., HAYNE, J. & BALUCH, A. 2018. 555 - Oral Ribavirin for the Treatment of Respiratory Syncytial Virus (RSV) and Human Metapneumovirus (hMPV) Infections in Hematology Patients and Stem Cell Transplant (SCT) Recipients at a Nci-Designated Cancer Center. *Biology of Blood and Marrow Transplantation,* 24**,** S383-S383.

POCH, O., BLUMBERG, B. M., BOUGUELERET, L. & TORDO, N. 1990. Sequence Comparison of Five Polymerases (L proteins) of Unsegmented Negative-strand RNA Viruses: Theoretical Assignment of Functional Domains. *Journal of general virology,* 71**,** 1153-1162.

POCH, O., SAUVAGET, I., DELARUE, M. & TORDO, N. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *The EMBO journal,* 8**,** 3867-3874.

RU-NING, G., HUI-ZHEN, Z., CHUN-QUAN, O., LI-QUN, H., YONG, Z., XIN, Z., CAN-KUN, L., JIN-YAN, L., HAO-JIE, Z., TIE, S. & HUI-MING, L. 2016. Impact of Influenza on Outpatient Visits, Hospitalizations, and Deaths by Using a Time Series Poisson Generalized Additive Model. *PLoS ONE,* 11**,** e0149468.

RUDD, P. A., THOMAS, B. J., ZAID, A., MACDONALD, M., KAN-O, K., ROLPH, M. S., SOORNEEDI, A. R., BARDIN, P. G. & MAHALINGAM, S. 2017. Role of human metapneumovirus and respiratory syncytial virus in asthma exacerbations: where are we now? *Clinical science (London, England : 1979),* 131**,** 1713-1721.

SABO, Y., EHRLICH, M. & BACHARACH, E. 2011. The Conserved YAGL Motif in Human Metapneumovirus Is Required for Higher-Order Cellular Assemblies of the Matrix Protein and for Virion Production. *The Journal of Virology,* 85**,** 6594.

SAFDAR, A. 2008. Immune modulatory activity of ribavirin for serious human metapneumovirus disease: early i.v. therapy may improve outcomes in immunosuppressed SCT recipients. *Bone Marrow Transplantation,* 41**,** 707.

SATO, K., WATANABE, O., OHMIYA, S., CHIBA, F., SUZUKI, A., OKAMOTO, M., YOUNGHUANG, J., HATA, A., NONAKA, H., KITAOKA, S., NAGAI, Y., KAWAMURA, K., HAYASHI, M., KUMAKI, S., SUZUKI, T., KAWAKAMI, K. & NISHIMURA, H. 2017. Efficient isolation of human metapneumovirus using MNT-1, a human malignant melanoma cell line with early and distinct cytopathic effects. *Microbiology and immunology,* 61**,** 497-506.

SCHILDGEN, O., GEIKOWSKI, T., GLATZEL, T., SCHUSTER, J. & SIMON, A. 2005a. Frequency of human metapneumovirus in the upper respiratory tract of children with symptoms of an acute otitis media. *European Journal of Pediatrics,* 164**,** 400-401.

SCHILDGEN, O., GLATZEL, T., GEIKOWSKI, T., SCHEIBNER, B., MATZ, B., BINDL, L., BORN, M., VIAZOV, S., WILKESMANN, A., KNOPFLE, G., ROGGENDORF, M. & SIMON, A. 2005b. Human metapneumovirus RNA in encephalitis patient.(Dispatches). *Emerging Infectious Diseases,* 11**,** 467.

SCHILDGEN, V., DEN HOOGEN, B., FOUCHIER, R., TRIPP, R., ALVAREZ, R., MANOHA, C., WILLIAMS, J. & SCHILDGEN, O. 2011. Human Metapneumovirus: Lessons Learned over the First Decade. *Clinical Microbiology Reviews,* 24**,** 734.

SCHOWALTER, R. M., SMITH, S. E. & DUTCH, R. E. 2006. Characterization of Human Metapneumovirus F Protein-Promoted Membrane Fusion: Critical Roles for Proteolytic Processing and Low pH. *The Journal of Virology,* 80**,** 10931.

SCHUSTER, J. E., COX, R. G., HASTINGS, A. K., BOYD, K. L., WADIA, J., CHEN, Z., BURTON, D. R., WILLIAMSON, R. A. & WILLIAMS, J. V. 2015. A broadly neutralizing human monoclonal antibody exhibits in vivo efficacy against both human metapneumovirus and respiratory syncytial virus.(Report). *Journal of Infectious Diseases,* 211**,** 216.

SHACHOR-MEYOUHAS, Y., BEN-BARAK, A. & KASSIS, I. 2011. Treatment with oral ribavirin and IVIG of severe human metapneumovirus pneumonia (HMPV) in immune compromised child. 57**,** 350-351.

SHI, T., MCALLISTER, D. A., O'BRIEN, K. L., SIMOES, E. A. F., MADHI, S. A., GESSNER, B. D., POLACK, F. P., BALSELLS, E., ACACIO, S., AGUAYO, C., ALASSANI, I., ALI, A., ANTONIO, M., AWASTHI, S., AWORI, J. O., AZZIZ-BAUMGARTNER, E., BAGGETT, H. C., BAILLIE, V. L. & BALMASEDA, A. 2017. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *The Lancet,* 390**,** 946-958.

SUMINO, K. C., AGAPOV, E., PIERCE, R. A., TRULOCK, E. P., PFEIFER, J. D., RITTER, J. H., GAUDREAULT-KEENER, M., STORCH, G. A. & HOLTZMAN, M. J. 2005. Detection of severe human metapneumovirus infection by real-time polymerase chain reaction and histopathological assessment.(MAJOR ARTICLE). *Journal of Infectious Diseases,* 192**,** 1052.

SUTHERLAND, K. A., COLLINS, P. L. & PEEPLES, M. E. 2001. Synergistic Effects of Gene-End Signal Mutations and the M2-1 Protein on Transcription Termination by Respiratory Syncytial Virus. *Virology,* 288**,** 295-307.

SUZUKI, A., WATANABE, O., OKAMOTO, M., ENDO, H., YANO, H., SUETAKE, M. & NISHIMURA, H. 2005. Detection of Human Metapneumovirus From Children With Acute Otitis Media. *The Pediatric Infectious Disease Journal,* 24**,** 655-657.

TORRES, F. S., IOLSTER, J. T., SCHNITZLER, A. E., FARIAS, C. J., BORDOGNA, J. A., RUFACH, J. D., MONTES, G. M., SIABA, E. A., RODRÍGUEZ, E. M., JABORNISKY, E. R., COLMAN, E. C., FERNÁNDEZ, E. A., CAPROTTA, E. G., DIAZ, E. S., POTERALA, E. R., DE MEYER, E. M., PENAZZI, E. M., GONZÁLEZ, E. G., SAENZ, E. S., RECUPERO, E. O., ZAPICO, E. L., ALARCON, E. B., ARIEL, E. E., MINCES, E. P., MARI, E. E., CARNIE, E. A., GAREA, E. M. & JAEN, E. R. 2012. High mortality in patients with influenza A pH1N1 2009 admitted to a pediatric intensive care unit: A predictive model of mortality*. *Pediatric Critical Care Medicine,* 13**,** e78-e83.

VAN DEN HOOGEN, B. G. 2007. Respiratory tract infection due to human metapneumovirus among elderly patients. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America,* 44**,** 1159-1160.

VAN DEN HOOGEN, B. G., BESTEBROER, T. M., OSTERHAUS, A. D. M. E. & FOUCHIER, R. A. M. 2002. Analysis of the Genomic Sequence of a Human Metapneumovirus. *Virology,* 295**,** 119-132.

VELAYUTHAM, T. S., KOLLI, D., IVANCIUC, T., GAROFALO, R. P. & CASOLA, A. 2013. Critical Role of TLR4 in Human Metapneumovirus Mediated Innate Immune Responses and Disease Pathogenesis.(Research Article). *PLoS ONE,* 8**,** e78849.

VICENTE, D., MONTES, M., CILLA, G., PEREZ-YARZA, E. G. & PEREZ-TRALLERO, E. 2006. Differences in Clinical Severity between Genotype A and Genotype B Human Metapneumovirus Infection in Children. *Clinical Infectious Diseases,* 42**,** e111-e113.

WEBSTER, D. L., GARDNER, A. H., DYE, T. J. & CHIMA, R. S. 2014. Status epilepticus: a possible association with human metapneumovirus infection.(Case study). *Pediatrics,* 133**,** e747.

WEI, H.-Y., TSAO, K.-C., HUANG, C.-G., HUANG, Y.-C. & LIN, T.-Y. 2013. Clinical features of different genotypes/genogroups of human metapneumovirus in hospitalized children. *Journal of Microbiology, Immunology and Infection,* 46**,** 352-357.

WEI, Y., ZHANG, Y., CAI, H., MIRZA, A., IORIO, R., PEEPLES, M., NIEWIESK, S. & LI, J. 2014. Roles of the Putative Integrin-Binding Motif of the Human Metapneumovirus Fusion (F) Protein in Cell-Cell Fusion, Viral Infectivity, and Pathogenesis. *Journal of Virology,* 88**,** 4338-4352.

WEN, S., TOLLEFSON, S., JOHNSON, M., GILCHUK, P., BOYD, K., SHEPHERD, B., JOYCE, S. & WILLIAMS, J. 2014. Acute Clearance of Human Metapneumovirus Occurs Independently of Natural Killer Cells. *Journal of Virology,* 88**,** 10963-10969.

WEN, S. & WILLIAMS, J. 2015. New Approaches for Immunization and Therapy against Human Metapneumovirus. *Clinical and Vaccine Immunology,* 22**,** 858-866.

WIDMER, K., ZHU, Y., WILLIAMS, J. V., GRIFFIN, M. R., EDWARDS, K. M. & TALBOT, H. K. 2012. Rates of hospitalizations for respiratory syncytial virus, human metapneumovirus, and influenza virus in older adults.(VIRUSES)(Report). *Journal of Infectious Diseases,* 206**,** 56.

WILLIAMS, J. V., HARRIS, P. A., TOLLEFSON, S. J., HALBURNT-RUSH, L. L., PINGSTERHAUS, J. M., EDWARDS, K. M., WRIGHT, P. F. & CROWE, J. E. 2004. Human Metapneumovirus and Lower Respiratory Tract Disease in Otherwise Healthy Infants and Children. *The New England Journal of Medicine,* 350**,** 443-450.

XU, J., ZHANG, Y. & WILLIAMS, J. V. 2018. Development and optimization of a direct plaque assay for trypsin-dependent human metapneumovirus strains. *Journal of Virological Methods,* 259**,** 1-9.

YAN, X. L., LI, Y. N., TANG, Y. J., XIE, Z. P., GAO, H. C., YANG, X. M., LI, Y. M., LIU, L. J. & DUAN, Z. J. 2017. Clinical characteristics and viral load of respiratory syncytial virus and human metapneumovirus in children hospitaled for acute lower respiratory tract infection. *Journal of Medical Virology,* 89**,** 589-597.

YANG, C.-F., WANG, C. K., TOLLEFSON, S. J., PIYARATNA, R., LINTAO, L. D., CHU, M., LIEM, A., MARK, M., SPAETE, R. R., CROWE, J. E., JR. & WILLIAMS, J. V. 2009. Genetic diversity and evolution of human metapneumovirus fusion protein over twenty years.(Research)(Report). *Virology Journal,* 6**,** 138.

ZHANG, Y., WEI, Y., ZHANG, X., CAI, H., NIEWIESK, S. & LI, J. 2014. Rational design of human metapneumovirus live attenuated vaccine candidates by inhibiting viral mRNA cap methyltransferase. *Journal of virology,* 88**,** 11411-11429.

CHAPTER 2

An Amplicon-Based Approach for the Whole Genome Sequencing of Human Metapneumovirus

Chapter 2: An Amplicon-Based Approach for the Whole Genome Sequencing of Human Metapneumovirus

Author Attribution Statement for Thesis with Publications

I am writing this letter to specify the role of Rachel Tulloch in the preparation and submission of the following manuscript:

Tulloch, RL, Kok, J, Carter, I, Dwyer, DE, & Eden, J-S, 2021, 'An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus', *Viruses*, vol. 13, no. 3, p. 499–, doi: 10.3390/v13030499.

Contribution to the manuscript of all authors is as follows:

Task	**Role of Author**
Conceptualisation	R.L.T, J-S.E
Investigation	R.L.T, J-S.E
Methodology	R.L.T, J-S.E
Data curation	R.L.T, J-S.E
Software	R.L.T, J-S.E
Formal analysis	R.L.T, J-S.E
Visualisation	R.L.T, J-S.E
Resources	J.K, I.C, D.E.D
Writing – Original draft	R.L.T
Writing – Review and editing	R.L.T, J.K, I.C, D.E.D, J-S.E

I declare that the authorship attribution statements made above are true. In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author, my supervisor.

Rachel Tulloch
December 2022

As the primary supervisor for the candidature upon which the thesis is based, I can confirm that the above authorship attributions statements are correct.

John-Sebastian Eden
December 2022

An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus

Rachel L. Tulloch [1,2], Jen Kok [3], Ian Carter [3], Dominic E. Dwyer [2,3] 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
2. Marie Bashir Institute for Infectious Diseases and Biosecurity, Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia; dominic.dwyer@sydney.edu.au
3. NSW Health Pathology-Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia; jen.kok@health.nsw.gov.au (J.K.); ian.carter@health.nsw.gov.au (I.C.)

* Correspondence: js.eden@sydney.edu.au; Tel.: +61-2-8627-1817

Abstract - Human metapneumovirus (HMPV) is an important cause of upper and lower respiratory tract disease in individuals of all ages. It is estimated that most individuals will be infected by HMPV by the age of five years old. Despite this burden of disease, there remain caveats in our knowledge of global genetic diversity due to a lack of HMPV sequencing, particularly at the whole-genome scale. The purpose of this study was to create a simple and robust approach for HMPV whole-genome sequencing to be used for genomic epidemiological studies. To design our assay, all available HMPV full-length genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database and used to design four primer sets to amplify long, overlapping amplicons spanning the viral genome and, importantly, specific to all known HMPV subtypes. These amplicons were then pooled and sequenced on an Illumina iSeq 100 (Illumina, San Diego, CA, USA); however, the approach is suitable to other common sequencing platforms. We demonstrate the utility of this method using a representative subset of clinical samples and examine these sequences using a phylogenetic approach. Here we present an amplicon-based method for the whole-genome sequencing of HMPV from clinical extracts that can be used to better inform genomic studies of HMPV epidemiology and evolution.

Keywords: human metapneumovirus; whole-genome sequencing; genomic epidemiology

Introduction

Since its discovery in 2001, human metapneumovirus (HMPV) has become increasingly recognised as a major cause of acute respiratory illness (ARI), globally (van Den Hoogen et al., 2002). Serological studies estimate that almost all individuals will be exposed to HMPV by the age of five (Kaslow et al., 2014). Clinically, HMPV is indistinguishable from ARI caused by other respiratory

pathogens, including respiratory syncytial virus (RSV) (Biacchesi et al., 2003). While most infections are mild and self-limiting, HMPV has increased morbidity and mortality in high-risk populations, including immunosuppressed individuals, young children and the elderly (Falsey et al., 2003). Globally, HMPV is associated with 3.9–7% of children hospitalised with lower respiratory tract infections (LRTI), with outpatient detection rates ranging from 6.2 to 12%, highlighting its clinical significance as a cause of ARI in this age group (Biacchesi et al., 2003, Bernadette et al., 2001, McAdam et al., 2004, Williams et al., 2004). HMPV is also a known cause of ARI in hospitalised adults, with one study showing that the hospitalisation rates of adults >50 years of age were statistically similar to those with influenza infections in the same region (Widmer et al., 2012).

HMPV is a member of the *Pneumoviridae* family, with a negative-sense, single-stranded RNA genome of approximately 13.3 kb in length (Sutherland et al., 2001, El Najjar et al., 2016). HMPV is genetically similar to RSV; however, it lacks two nonstructural genes—NS1 and NS2. Phylogenetic analysis of the HMPV fusion (F) and glycoprotein (G) genes has led to the identification and classification of viruses into two major subtypes, HMPV A and HMPV B, which can further be subdivided into A1, A2a, A2b, B1 and B2 sublineages (Lim et al., 2020a). Epidemiological studies have revealed the cocirculation of HMPV subtypes globally, with the predominant subtype fluctuating throughout the year (Jain et al., 2014). Historically, HMPV molecular epidemiological studies have relied on the subgenomic sequencing of partial F or G protein genes to perform genomic and evolutionary studies (Reiche et al., 2014). Indeed, only 2.3% (n = 226/9795) of available sequences on the National Center for Biotechnology Information (NCBI) GenBank database are near-complete or complete genomes (as of November 2020). Therefore, our understanding of the genomic epidemiology, genetic diversity and evolution of HPMV remains limited.

Whole-genome sequencing (WGS) is a powerful tool for public health infectious disease surveillance; it can also inform the treatment and control of viruses in the population (Gardy et al., 2015, Houlihan et al., 2018). WGS offers increased resolution at multiple epidemiological scales, from investigating global virus traffic networks to elucidating individual transmission events within outbreaks (Houlihan et al., 2018, Gymoese et al., 2017). The recent SARS-CoV-2 epidemic has highlighted the utility of amplicon-based WGS methods as a cost-effective, rapid method to sequence the whole-genome approach (Eden et al., 2020a, Freed et al., 2020, Tyson et al., 2020). The purpose of this study was to develop a simple and robust amplicon-based method for sequencing the HMPV full-length genome with the aim to inform a better understanding of its molecular epidemiology.

Materials and Methods

Primer Design

Our approach was based on an existing WGS workflow designed to amplify and sequence the RSV genome using four amplicons between 3,528 and 4,375 nt in length, each with an overlapping region of at least 100nt (Di Giallonardo et al., 2018a). Given the similar genome lengths between HPMV and RSV, we focused on designing HMPV-specific primer sets that would also generate four overlapping amplicons of ~3.5 kb each that span the viral genome. To include historical and current circulating HMPV subtypes, all available full- or near-full-length (>13,000nt) HMPV genome sequences were obtained from the NCBI GenBank database (downloaded on 10 December 2019), and all sequence analysis was conducted using Geneious Prime 2019.2.3. Sequences were excluded from the initial analysis if they contained >20 continuous ambiguous bases. This result was a final set of 153 full- or near-full-length HMPV sequences, which were then aligned using MAFFT version 7.45 (Katoh and Standley, 2013). Phylogenetic analysis using a neighbour-joining approach was performed to show that all known HMPV subtypes were represented in the subset of sequences (Figure 1).

A sliding window approach was then used to plot sequence identity along the viral genome alignment and identify conserved regions for targeted primer design. These primers were designed to be ~25 bp in length with degeneracies where necessary to capture position variation between HMPV subtypes, as well as to have melting temperatures within 5 °C of each other, and avoiding potential dimer formations (Chuang et al., 2013). We also tested an existing published primer set designed to amplify the HPMV genome using a similar overlapping amplicon scheme (~3.5 kb each) from a study examining the local virus traffic and genetic diversity of HMPV in Peru (Pollett et al., 2018). The final set of primer pairs was designed to be run as four separate polymerase chain reactions (PCRs) performed in parallel (**Table 1**) and was tested on a selection of known viral extracts from HMPV-positive clinical specimens.

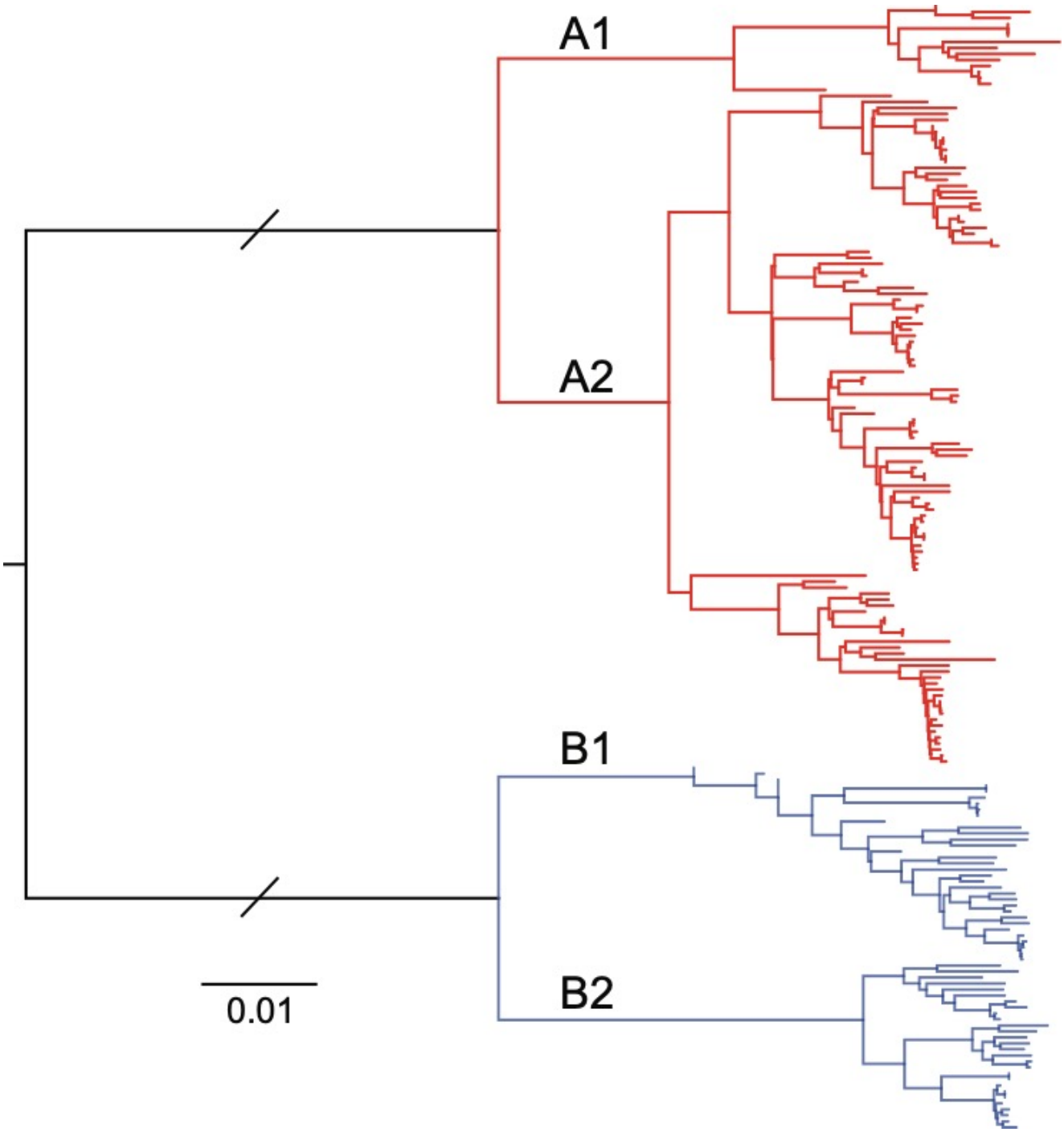


Figure 1: Neighbour-joining phylogenetic tree of 153 human metapneumovirus (HMPV) full-length or near-full-length genome sequences sourced from the National Center for Biotechnology Information (NCBI) GenBank database. The two major phylogenetic subtypes of HMPV are shown with different coloured branches, with known sublineages further differentiated by labelled branches (A1, A2, B1 and B2). The scale bar is proportional to the number of substitutions per site. The branch lengths between the A and B subtypes have been shortened for clarity.

Table 1: Primer pairs used in this study to complete the near-whole-genome sequencing of human metapneumovirus.

Assay	Primer Name	Sequence (5'-3')	Position (nt) *	PCR Amplicon Size (bp)
PCR1	HMPV1_F	GGGACAAATAAAAATGTCTCTTCA	41	4,125
	HMPV1_R	CTTCCTGTGCTRACYTTRCA	4,165	
PCR2	HMPV2_F	ACAGCAGCRGGRATYAATGT	4,074	4,010
	HMPV2_R	TAGTACTGAAYTGAGCATGYTCAG	8,083	
PCR3	HMPV3_F	AACTGTTAACATGGAAAGATGTGATG	7,823	3,229
	HMPV3_R	TAAGCTGGAACWGAWGCTG	11,051	
PCR4	HMPV4_F	TCAATAGGGAGTCTRTGTCARGAA	9,730	3,675
	HMPV4_R	GRCAAAAAAACCGTATACATYC	13,404	

* Nucleotide positions are relative to strain A1 (NCBI GenBank Accession KU821121).

Clinical Specimens and HMPV Isolation

Residual total nucleic acid from respiratory specimens submitted to NSW Health Pathology, Institute of Clinical Pathology and Microbiology Research (ICPMR) for diagnostic testing were utilised in this study as per protocols approved by the Westmead Hospital and Institute ethics and governance committees (LNR/17/WMEAD/128 and SSA/17/WMEAD/129, 16th November 2017). Total nucleic acid was extracted from each submitted respiratory specimen using a high-throughput, magnetic-bead-based extraction platform and screened against a panel of known respiratory viruses using an accredited multiplex quantitative RT-PCR (qRT-PCR). Viruses included on the respiratory panel are influenza A, influenza A subtype H3N2, influenza A H1N1 2009 pandemic, influenza B, adenovirus, parainfluenza 1,2 and 3, RSV, rhinovirus, enterovirus and HPMV. Archived clinical extracts that were reported as positive for HMPV were deidentified before inclusion in this study. No specific subtyping information was available on the archived samples; therefore, we selected ten random HMPV-positive nucleic acid specimens collected over seven years between 2013 and 2020 to attempt to capture historic HMPV genetic diversity in NSW, Australia. To estimate the levels of HMPV, an in-house singleplex qRT-PCR specific for HMPV was performed, with CT values ranging between 19.9 and 34.2 in our study samples.

Reverse Transcription

Complementary DNA (cDNA) synthesis was performed in a 20 µL reaction containing 4 µL of 5× SuperScript IV VILO MasterMix (Invitrogen, Carlsbad , CA, USA), 12 µL of nuclease-free water and 4 µL of the viral RNA template, as per the manufacturer's instructions. The thermal cycling

protocol used was as follows: random priming was performed at 25 °C for 10 min, followed by extension at 50 °C for 20 min and then enzyme denaturation at 85 °C for 5 min before holding at 4 °C. All incubation steps and reaction components were performed to the manufacturer's specifications on a SimpliAmp thermocycler (Applied Biosystems, Foster City , CA, USA). Viral cDNA was used immediately or stored at −80 °C until required.

HMPV Genome Amplification

The viral cDNA was then split across four separate PCR reactions, each amplifying one part of the HMPV genome (**Table 1**). Each PCR was performed in a 25 µL reaction containing 12.5 µL of 2× Platinum SuperFi MasterMix (Invitrogen), 1.25 µL of 10 µM forward primer, 1.25 µL of 10 µM reverse primer, 7 µL of nuclease-free water and 3 µL of cDNA template. The reactions were then incubated at 98 °C for 30 s, followed by 44 cycles of denaturation at 98 °C for 10 s, annealing at 60 °C for 20 s and extension at 72 °C for 2:10 min, with a final extension for 5 min at 72 °C, followed by holding at 4 °C. Amplicon size and yield were assessed by gel electrophoresis of 5 µL of PCR reactions using a 1% E-Gel-48 Agarose Gel (Invitrogen) with 500 ng of 1 Kb Plus DNA Ladder. The gels were pre-stained with ethidium bromide for amplicon visualisation under UV light for approximate quantification and sizing. Amplicon approximate quantity was estimated using the target PCR product band intensity. To ensure even coverage across the HPMV genome, the 4 amplicons of each clinical sample were pooled equally based on target amplicon intensity into a final pooled volume of 40 µL, adjusted with Qiagen EB buffer (Qiagen, Düsseldorf, Germany) if necessary. When nonspecific amplification was present, the band intensity of only the target amplicon was taken into consideration, and pooling was adjusted accordingly. The HMPV genome amplicon pools were purified using AMPure XP (Beckman Coulter, Pasadena, CA, USA) at a bead-to-sample ratio of 1× according to the manufacturer's protocol. The purified DNA was then quantified using the 1× double-stranded DNA high sensitivity (1× dsDNA HS) Qubit assay (Invitrogen) and measured on the Qubit 4 fluorometer. The pooled amplicons were then volumetrically diluted to 0.2 ng/µL, the required input concentration for library preparation.

Library Preparation and Sequencing

Amplicons were prepared for sequencing using the Nextera XT DNA Library Preparation Kit with the v2 Set B indexing kit (Illumina, Massachusetts, MA, USA), although any compatible indexing set of choice could be used in the reproduction of this approach. The manufacturer's protocol was followed for genomic DNA tagmentation, library amplification and clean-up, except

that all volumes were halved for reagent conservation. The purified DNA libraries were quantified using the 1× dsDNA HS Qubit assay and Qubit 4 fluorometer before normalisation using Qubit DNA concentrations. The final library pool molarity and fragment length distribution were determined using the 4200 TapeStation System with a high sensitivity D5000 tape (Agilent, Santa Clara, CA, USA) before dilution to 0.1 nM with Qiagen EB buffer (Qiagen, Düsseldorf, Germany) for loading and sequencing on an Illumina iSeq 100 platform (Illumina, San Diego, CA, USA) with a v1 300 cycle kit.

Viral Assembly

Raw sequences were initially quality scored using fastqc (Brown et al., 2017) following this, the reads were quality trimmed to a QC threshold of phred score 20 using bbduk (Bushnell, 2016). The trimmed reads were then *de novo* assembled using Megahit with default parameters (Li et al., 2015). The trimmed reads were then remapped onto the draft genome using bbmap (Bushnell, 2016) before the overall mapping alignment quality was assessed using the Geneious Prime 2019.2.3 and majority consensus genome extracted. The final genome was trimmed of terminal primer sequences and annotated using NCBI GenBank reference sequences.

Phylogenetic Analysis

Phylogenetic analysis was performed by aligning all sequences generated in this study against a reference set obtained from NCBI GenBank using MAFFT and PhyML (Silva et al., 2017), with node support estimated by 100 bootstrap replicates. Sequences obtained in this study were published to NCBI GenBank with the following accession IDs: MW221986-MW221995.

Results and Discussion

Designing Primers to Amplify the HMPV Genome

The aim of this study was to develop a simple and robust amplicon-based approach for amplifying and sequencing the HMPV genome. To do this, we adapted a previous approach used for RSV (Di Giallonardo et al., 2018a) to design four primer sets generating ~3.5 kb amplicons that overlap and span the viral genome. Our primers were based on all available HMPV genomes from the NCBI GenBank database and targeted conserved regions at suitably spaced positions in the genome (**Table 1 and Figure 2**).

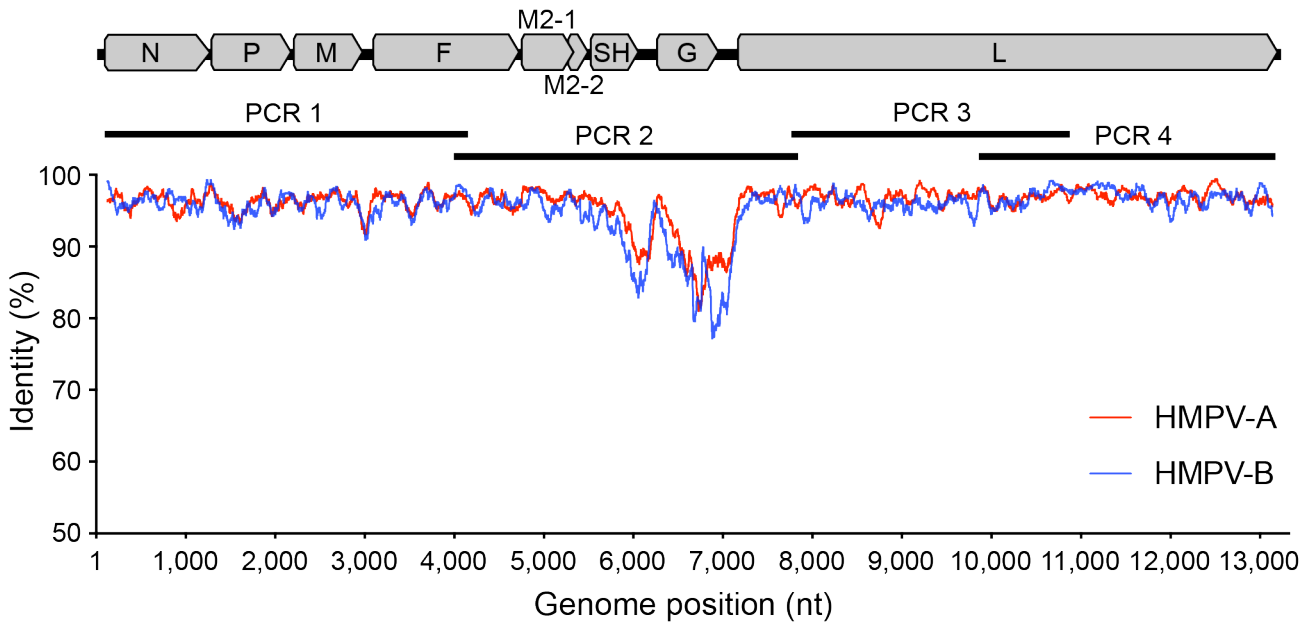


Figure 2: Genetic variability across the HPMV genome. A graphic representation showing the percentage identity at each nucleotide position along the genome with each HMPV subtype plotted separately—A (red) and B (blue). Included also is an HMPV genome shown to scale with annotated genes to highlight primer and amplicon positions.

The final primer sets were located in the terminal regions of the genome, as well as in the fusion and large protein genes, and avoided divergent regions of the genome such as in the viral glycoprotein (G protein) (**Figure 2**). Other amplicon-based methods for viral genome sequencing often employ shorter amplicon lengths (1,000 bp or less) to improve performance for low-viral-load or low-quality samples, such as with the ARTIC protocol for SARS-CoV-2 genome sequencing (Tyson et al., 2020), or even for enteric virus including human norovirus (Cotten et al., 2014). Here we chose targeted amplicons to be between 3,000 and 4,000 bp based on previous performance against RSV (Di Giallonardo et al., 2018a), where genomes would reliably amplify from 80 to 90% of clinical samples. Furthermore, there is greater diversity present in HMPV compared to SARS-CoV-2 such that there are less suitable target positions across the genome to readily amplify all subtypes. Indeed, to capture this diversity several degenerate nucleotides were included in our primers, and based on our current understanding of HMPV diversity, it would be expected our primers cover the vast majority of variants present in circulation (**Figure 3**).

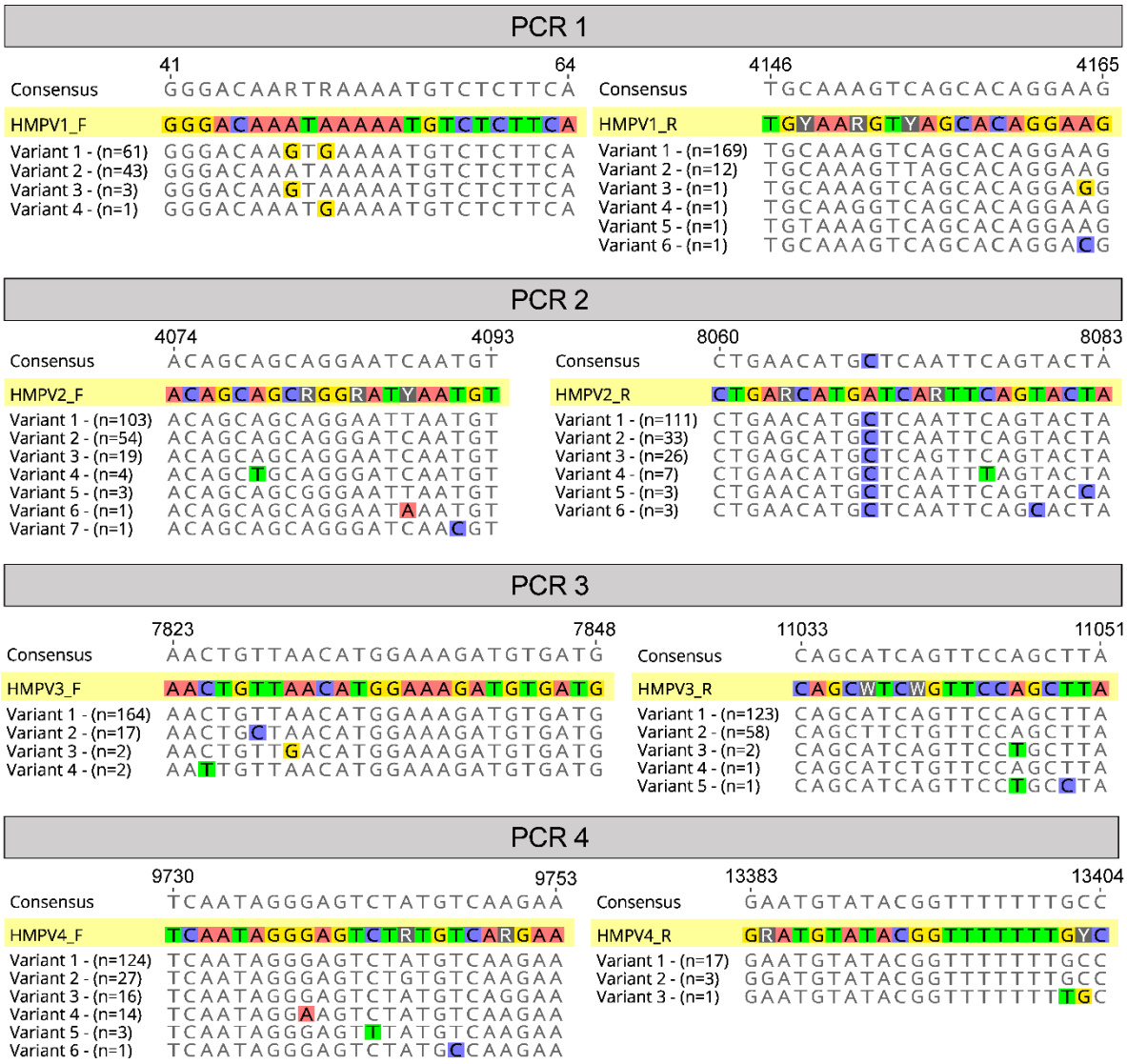


Figure 3: Primer alignments. A graphical representation of the four chosen primer pairs aligned against known variants (sequences in the alignment) at each binding site relative to strain A1 (NCBI GenBank Accession: KU821121). Variants are different possible sequences in the alignment, and "n" refers to the number of these sequences that were observed at the primer binding sites.

HMPV Genome RT-PCR Performance

To examine the performance of our newly designed primers, we tested them against a set of HMPV-positive extracts from clinical respiratory specimens. Since the subtype and sublineage classifications from our samples were unknown, we instead obtained samples across a wide time period (2013 to 2020) to attempt to capture a breadth of diversity. Initial end-point PCRs showed good levels of amplification across the four targets (data not shown); however, we then attempted further optimisation of the assay using a temperature gradient (59–61.5 °C) to establish the optimal annealing temperature was 60 °C to ensure efficient target amplification and minimise nonspecific amplification (**Figure 4A**). We also compared our primers to those previously published from the Peru WGS study (**Figure 4B**) and showed improved performance, particularly for the specific

amplification of the targeted HMPV region; however, this may be partly due to our initial optimisation of annealing and cycling conditions favouring our newly designed set. An additional RT-PCR was performed to ensure the assay was specific for HMPV and did not amplify the closely related virus, RSV (**Figure Supplementary 1**).

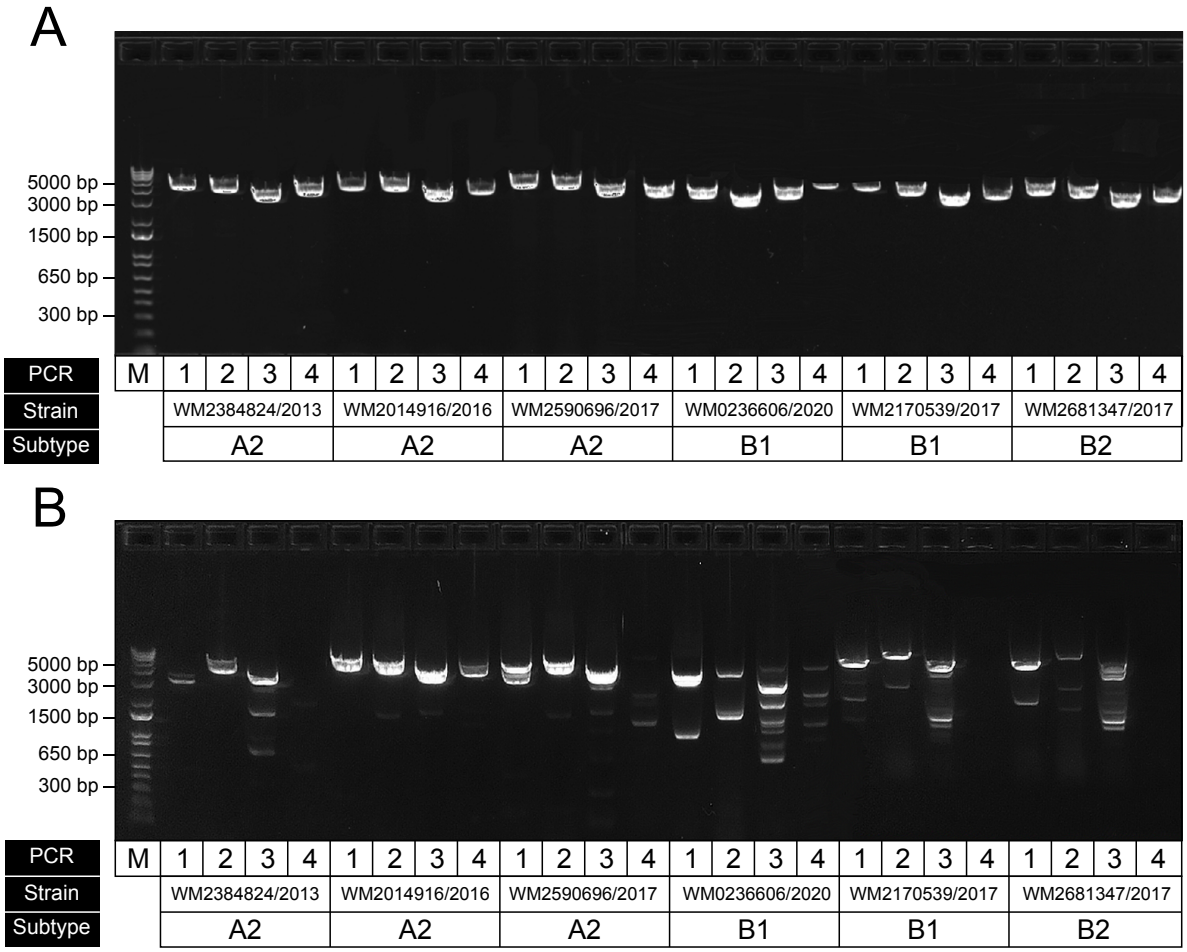


Figure 4: Representative gel electrophoresis result showing RT-PCR performance of amplicons for HMPV genome sequencing. (**A**) Results for our optimised assay using our newly designed primers; (**B**) same samples amplified using previously published primers using the same conditions (Pollett et al., 2018). For each panel, the four amplicons for each sample have been run in sequential order, with strain ID corresponding to the original sample accession and collected year. The samples shown here were of the subtypes A2, B1 and B2, as indicated.

Genome Sequencing, Assembly and Analysis

Following the successful amplification of all ten HMPV samples, the four amplicons from each were pooled, purified, and sequenced using the Nextera XT library prep kit and an Illumina iSeq 100. In this study, we sequenced the 10 clinical HMPV samples along with libraries from other projects. However, we targeted 100,000 paired reads per HMPV library to achieve an expected genome coverage depth between 800 and 1,000×, which is sufficient for calling a consensus genome. Given

an Illumina iSeq 100 run would yield a total of 5,000,000 paired reads, it would be possible to reliably multiplex up to 48 HMPV genomes per run. In this study, the samples included had a mean read depth of 1,125× across all samples (**Supplementary Figure 2**), and the coverage was found to be even across the genome, except where amplicon pooling was not equal. We then used a *de novo* assembly approach to generate the final consensus genomes for the ten HMPV samples. Reference mapping would also be an appropriate method for genome assembly; however, similar to RSV, there are notable structural variants (insertions) in the HMPV G protein (Di Giallonardo et al., 2018a) that may be misassembled when using an inappropriate reference strain for mapping such as a historical prototype. Therefore, a *de novo* approach would be recommended for both RSV and HMPV WGS, and users should examine coverage profiles for depth variability as an indication of structurally misassembled genomes. It is also important to note that the final genome sequences generated using our approach will be in-complete in the terminal regions and missing an expected 43 and 28 base pairs in the 5′ and 3′ ends, respectively. Ideally, our assay would have had primers designed from the very terminal regions; however, the limited sequence availability meant that such primers using existing GenBank sequences may not capture all circulating diversity. Importantly, the slightly inward placement of primers was not found to have any impact on the phylogeny of the virus, and the topology of phylogenetic trees using complete and "near-complete" genomes (i.e., just using the region our assay amplifies) were found to be identical (data not shown). Furthermore, it is also common practice in phylogenetic analysis to ensure sequences are trimmed to coding regions only, which our assay captures. This genome sequencing approach is also useful in investigating minor variants present in individual patient samples. Previous studies of other respiratory viruses have shown the merit of using amplicon-based methods to identify mixed infections containing multiple viral subtypes (Di Giallonardo et al., 2018a, Eden et al., 2020a). However, here in our study samples, no variants were observed at a frequency above 1%.

We then analysed our ten clinical HMPV genome sequences alongside four reference sequences and a selected subset of published hMPV sequences using a phylogenetic approach to determine their subtype. Of the samples presented in this study, four were identified as A2b strains, two as B1 and four B2 (**Figure 5**). There are limited data on the molecular epidemiology of HMPV in Australia and nothing published previously for the state of NSW, where these samples were collected. The finding of no A1 or A2a strains may be due to the under sampling in this current study; however, one study from Queensland, Australia, showed declining levels of A1 over the period 2001–2004 (Sloots et al., 2006), and since 2006, these subtypes have been rarely identified with A2b and B strains most commonly identified (Yi et al., 2019). Despite this, based on our alignments, we would expect the

primers and amplification to capture all subtypes, including A1 and A2, as these viruses were represented in our genome alignments (**Figures 1 and 3**), and this approach would be useful for ongoing genomic studies here in Australia and globally.

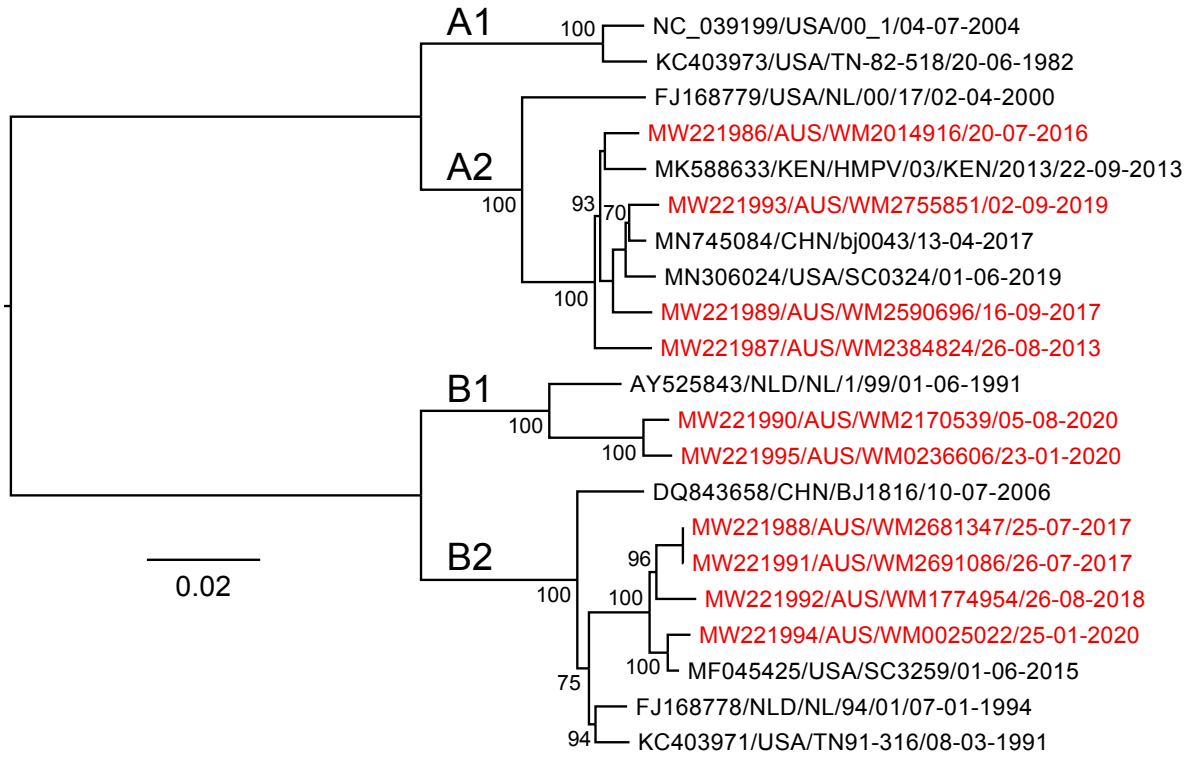


Figure 5: A maximum-likelihood tree constructed using near-full-length human metapneumovirus sequences generated in this study (those colored red). Node supports are indicated, and branch lengths are scaled according to sequence divergence (substitutions per site).

Conclusions

Using publicly available genome sequences representing the full known diversity of HMPV, we designed a simple and reliable assay for amplifying and sequencing HMPV genomes from clinical samples. Ten HMPV genomes were generated from residual-diagnostic specimens using this approach to demonstrate multiple subtypes circulating in NSW, Australia, since 2013. This work highlights the utility of amplicon-based sequencing for genomic epidemiological studies of respiratory viruses to inform public health investigations and understand the patterns of evolution and spread.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: Conceptualization, R.L.T. and J.-S.E.; methodology, R.L.T. and J.-S.E.; formal analysis, R.L.T. and J.-S.E.; resources, J.K., I.C., D.E.D. and J.-S.E.; data curation, R.L.T. and J.-S.E; writing—original draft preparation, R.L.T. and J.-S.E.; writing—review and editing, R.L.T., J.K., I.C., D.E.D. and J.-S.E.; supervision, J.-S.E.; project administration, J.-S.E; funding acquisition, J.K., D.E.D. and J.-S.E. All authors have read and agreed to the published version of the manuscript.

Funding: Funding was provided through the ICPMR Private Practice Trust fund, the National Health and Medical Research Council Center of Research Excellence in Emerging Infectious Diseases (1102962) and the Marie Bashir Institute for Infectious Diseases and Biosecurity at the University of Sydney.

Institutional Review Board Statement: The study was approved by the Westmead Hospital Ethics Committee of (LNR/17/WMEAD/128 and SSA/17/WMEAD/129, 16th November 2017).

Informed Consent Statement: Patient consent was waived for this study as it used de-identified, residual diagnostic specimens.

Data Availability Statement: The sequences generated in this study have been deposited to the NCBI GenBank database under the accession numbers: MW221986-MW221995.

Acknowledgments: We thank the members of the Virology team at NSW Health Pathology-ICPMR for the curation of the diagnostics specimens used in this study.

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

References

BERNADETTE, G. V. D. H., JAN, C. D. J., JAN, G., THIJS, K., RONALD DE, G., RON, A. M. F. & ALBERT, D. M. E. O. 2001. A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nature Medicine,* 7**,** 719.

BIACCHESI, S., SKIADOPOULOS, M. H., BOIVIN, G., HANSON, C. T., MURPHY, B. R., COLLINS, P. L. & BUCHHOLZ, U. J. 2003. Genetic diversity between human metapneumovirus subgroups. *Virology,* 315**,** 1-9.

BROWN, J., PIRRUNG, M. & MCCUE, L. A. 2017. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics,* 33**,** 3137-3139.

BUSHNELL 2016. BBMap short-read aligner, and other bioinformatics tools.

CHUANG, L.-Y., CHENG, Y.-H. & YANG, C.-H. 2013. Specific primer design for the polymerase chain reaction. *Biotechnology Letters,* 35**,** 1541-1549.

COTTEN, M., PETROVA, V., PHAN, M. V. T., RABAA, M. A., WATSON, S. J., ONG, S. H., KELLAM, P. & BAKER, S. 2014. Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. *Journal of virology,* 88**,** 11056-11069.

DI GIALLONARDO, F., KOK, J., FERNANDEZ, M., CARTER, I., GEOGHEGAN, J. L., DWYER, D. E., HOLMES, E. C. & EDEN, J.-S. 2018. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses,* 10.

EDEN, J.-S., ROCKETT, R., CARTER, I., RAHMAN, H., DE LIGT, J., HADFIELD, J., STOREY, M., REN, X., TULLOCH, R., BASILE, K., WELLS, J., BYUN, R., GILROY, N., O'SULLIVAN, M. V., SINTCHENKO, V., CHEN, S. C., MADDOCKS, S., SORRELL, T. C., HOLMES, E. C., DWYER, D. E. & KOK, J. 2020. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran.

EL NAJJAR, F., CIFUENTES-MUNOZ, N., CHEN, J., ZHU, H., BUCHHOLZ, U. J., MONCMAN, C. L. & DUTCH, R. E. 2016. Human metapneumovirus Induces Reorganization of the Actin Cytoskeleton for Direct Cell-to-Cell Spread.(Research Article). *PLoS Pathogens,* 12**,** e1005922.

FALSEY, A. R., ERDMAN, D., ANDERSON, L. J. & WALSH, E. E. 2003. Human metapneumovirus infections in young and elderly adults. (Major Article). *Journal of Infectious Diseases,* 187**,** 785.

FREED, N. E., VLKOVÁ, M., FAISAL, M. B. & SILANDER, O. K. 2020. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biology Methods & Protocols,* 5.

GARDY, J. L., NAUS, M., AMLANI, A., CHUNG, W., KIM, H., TAN, M., SEVERINI, A., KRAJDEN, M., PUDDICOMBE, D., SAHNI, V., HAYDEN, A. S., GUSTAFSON, R., HENRY, B. & TANG, P. 2015. Whole-Genome Sequencing of Measles Virus Genotypes H1 and D8 During Outbreaks of Infection Following the 2010 Olympic Winter Games Reveals Viral Transmission Routes. *The Journal of Infectious Diseases,* 212**,** 1574-1578.

GYMOESE, P., SORENSEN, G., LITRUP, E., OLSEN, J. E., NIELSEN, E. M. & TORPDAHL, M. 2017. Investigation of Outbreaks of Salmonella enterica Serovar Typhimurium and Its Monophasic Variants Using Whole-Genome Sequencing, Denmark.(RESEARCH). *Emerging Infectious Diseases,* 23**,** 1631.

HOULIHAN, C. F., FRAMPTON, D., FERNS, R. B., RAFFLE, J., GRANT, P., REIDY, M., HAIL, L., THOMSON, K., MATTES, F., KOZLAKIDIS, Z., PILLAY, D., HAYWARD, A. & NASTOULI, E. 2018. Use of Whole-Genome Sequencing in the Investigation of a Nosocomial Influenza Virus Outbreak. *The Journal of Infectious Diseases,* 218**,** 1485-1489.

JAIN, B., SINGH, A. K., DANGI, T., AGARWAL, A., VERMA, A. K., DWIVEDI, M., SINGH, K. P. & JAIN, A. 2014. High prevalence of human metapneumovirus subtype B in cases presenting as severe acute respiratory illness: an experience at tertiary care hospital. *The Clinical Respiratory Journal,* 8**,** 225-33.

KASLOW, R. A., STANBERRY, L. R. & LE DUC, J. W. 2014. *Viral Infections of Humans Epidemiology and Control,* Boston, MA, Springer US.

KATOH, K. & STANDLEY, D. M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution,* 30**,** 772-780.

LI, D., LIU, C.-M., LUO, R., SADAKANE, K. & LAM, T.-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

LIM, Y. K., KWEON, O. J., KIM, H. R., KIM, T.-H. & LEE, M.-K. 2020. Clinical Features, Epidemiology, and Climatic Impact of Genotype-specific Human Metapneumovirus Infections: Long-term Surveillance of Hospitalized Patients in South Korea. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America,* 70**,** 2683-2694.

MCADAM, A. J., HASENBEIN, M. E., FELDMAN, H. A., COLE, S. E., OFFERMANN, J. T., RILEY, A. M. & LIEU, T. A. 2004. Human metapneumovirus in children tested at a tertiary-care hospital.(Major Article). *Journal of Infectious Diseases,* 190**,** 20.

POLLETT, S., TROVAO, N. S., TAN, Y., EDEN, J.-S., HALPIN, R. A., BERA, J., DAS, S. R., WENTWORTH, D., OCANA, V., MENDOCILLA, S. M., ALVAREZ, C., CALISTO, M. E., GARCIA, J., HALSEY, E., AMPUERO, J. S., NELSON, M. I. & LEGUIA, M. 2018. The transmission dynamics and diversity of human metapneumovirus in Peru.(Report). *Influenza and Other Respiratory Viruses,* 12**,** 508.

REICHE, J., JACOBSEN, S., NEUBAUER, K., HAFEMANN, S., NITSCHE, A., MILDE, J., WOLFF, T. & SCHWEIGER, B. 2014. Human Metapneumovirus: Insights from a Ten-Year Molecular and Epidemiological Analysis in Germany. *PLoS One,* 9**,** e88342.

SILVA, J. O., ORELLANA, E. T. V. & TORRES, M. 2017. Development of a Parallel Version of PhyML 3.0 Using Shared Memory. *Revista IEEE América Latina,* 15**,** 959-967.

SLOOTS, T. P., MACKAY, I. M., BIALASIEWICZ, S., JACOB, K. C., MCQUEEN, E., HARNETT, G. B., SIEBERT, D. J., MASTERS, I. B., YOUNG, P. R. & NISSEN, M. D. 2006. Human Metapneumovirus, Australia, 2001–2004. *Emerging infectious diseases,* 12**,** 1263-1266.

SUTHERLAND, K. A., COLLINS, P. L. & PEEPLES, M. E. 2001. Synergistic Effects of Gene-End Signal Mutations and the M2-1 Protein on Transcription Termination by Respiratory Syncytial Virus. *Virology,* 288**,** 295-307.

TYSON, J. R., JAMES, P., STODDART, D., SPARKS, N., WICKENHAGEN, A., HALL, G., CHOI, J. H., LAPOINTE, H., KAMELIAN, K., SMITH, A. D., PRYSTAJECKY, N., GOODFELLOW, I., WILSON, S. J., HARRIGAN, R., SNUTCH, T. P., LOMAN, N. J. & QUICK, J. 2020. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*.

VAN DEN HOOGEN, B. G., BESTEBROER, T. M., OSTERHAUS, A. D. M. E. & FOUCHIER, R. A. M. 2002. Analysis of the Genomic Sequence of a Human Metapneumovirus. *Virology,* 295**,** 119-132.

WIDMER, K., ZHU, Y., WILLIAMS, J. V., GRIFFIN, M. R., EDWARDS, K. M. & TALBOT, H. K. 2012. Rates of hospitalizations for respiratory syncytial virus, human metapneumovirus, and influenza virus in older adults.(VIRUSES)(Report). *Journal of Infectious Diseases,* 206**,** 56.

WILLIAMS, J. V., HARRIS, P. A., TOLLEFSON, S. J., HALBURNT-RUSH, L. L., PINGSTERHAUS, J. M., EDWARDS, K. M., WRIGHT, P. F. & CROWE, J. E. 2004. Human Metapneumovirus and Lower Respiratory Tract Disease in Otherwise Healthy Infants and Children. *The New England Journal of Medicine,* 350**,** 443-450.

YI, L., ZOU, L., PENG, J., YU, J., SONG, Y., LIANG, L., GUO, Q., KANG, M., KE, C., SONG, T., LU, J. & WU, J. 2019. Epidemiology, evolution and transmission of human metapneumovirus in Guangzhou China, 2013-2017. *Scientific reports,* 9**,** 14022-9.

Supplementary Figures

The Genomic Epidemiology of Human Metapneumovirus in Australia and the impact of COVID-19

Tulloch, RL, Kok, J, Carter, I, Dwyer, DE, & Eden, J-S, 2021, 'An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus', *Viruses*, vol. 13, no. 3, p. 499–, doi: 10.3390/v13030499.

[1]Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia
[2]Sydney Institute for Infectious Diseases, Faculty of Health and Medicine, The University of Sydney, Westmead, NSW 2145, Australia
[3]Kirby Institute, The University of New South Wales, Kensington, NSW 2052, Australia
[4]NSW Health Pathology - Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia
[5]The Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
[6]The Kinghorn Centre for Clinical Genomics, Sydney, NSW 2010, Australia
[7]St Vincent's Clinical School, Department of Medicine, UNSW Sydney, Kensington, NSW 2052, Australia

Correspondence

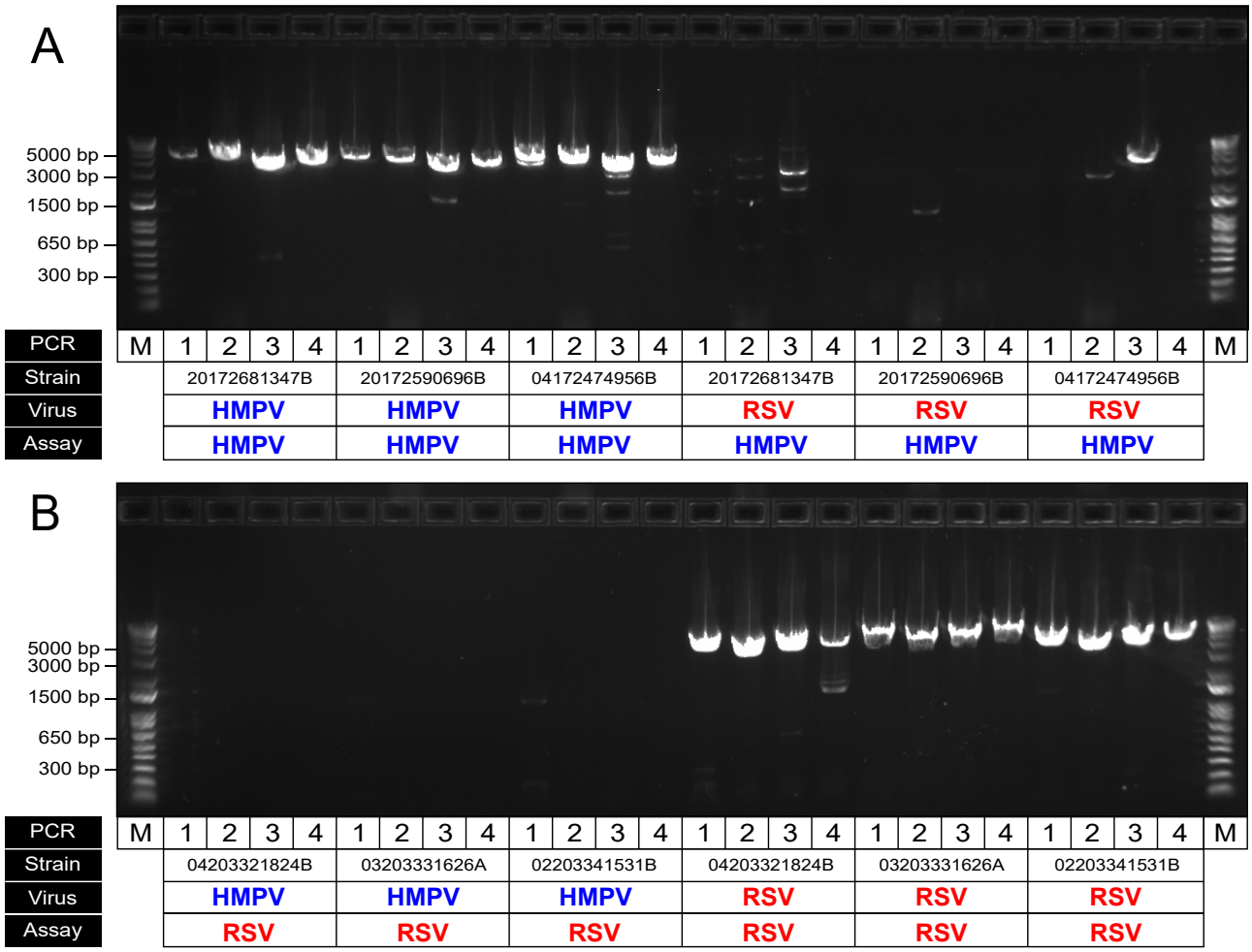
Dr John-Sebastian Eden

Email: js.eden@sydney.edu.au

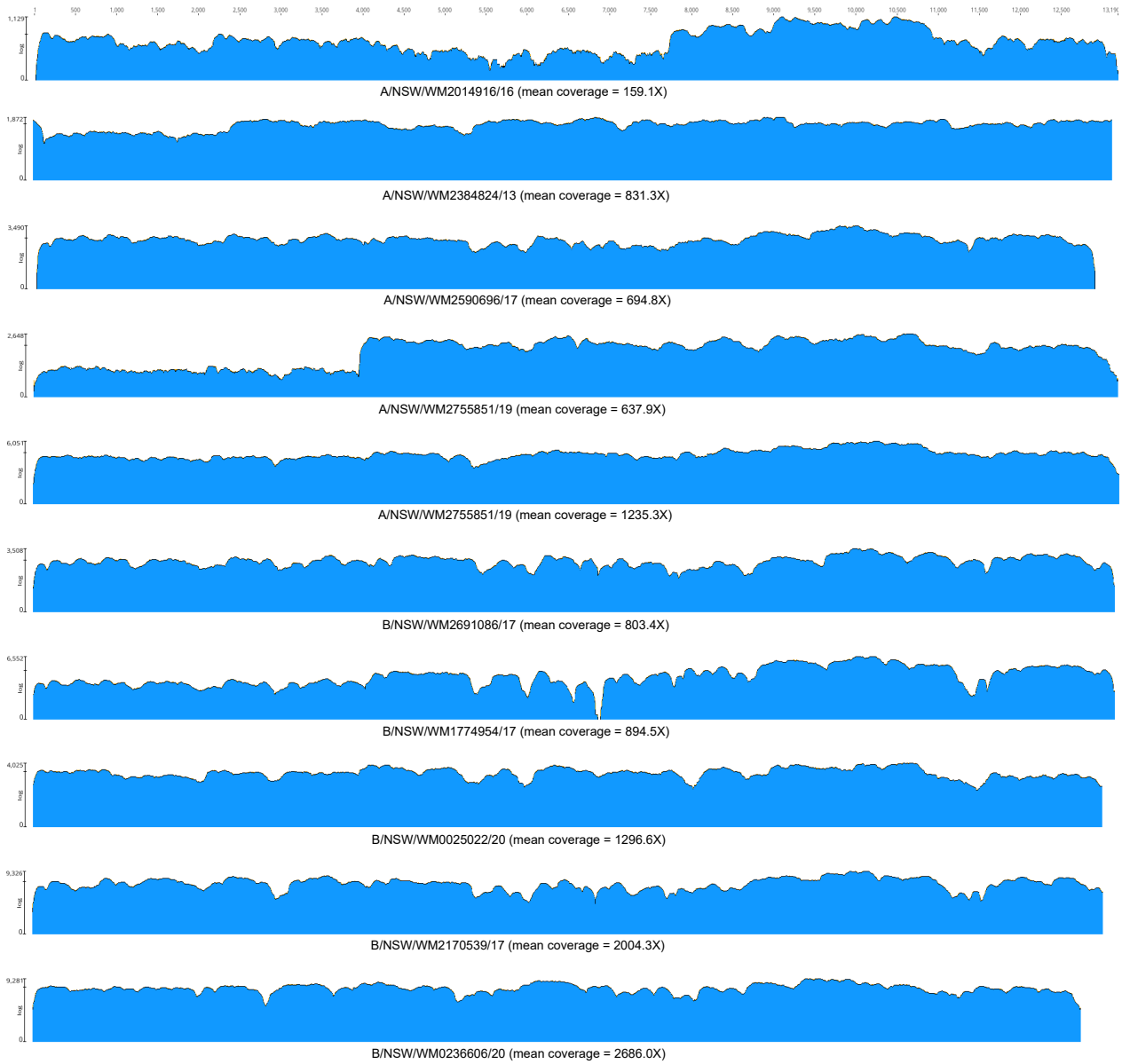
Phone: +61 2 8627 1817

Keywords

Human metapneumovirus; Whole-genome sequencing; Genomic epidemiology; COVID-19; non-pharmaceutical interventions.



Supplementary Figure 1. Gel electrophoresis result showing RT-PCR results for cross-reactivity between HMPV and RSV genome amplifcaiton primers. Panel A shows the results for our optimized HMPV assay using a set of HMPV and RSV positive samples (3 each) while Panel B shows the same samples amplified using previously published RSV primers (Di Giallonardo et al., 2018). For each panel, the labels show the PCR segment, Strain name, Virus sample is positive for and the WGS assay used for each RT-PCR.



Supplementary Figure 2. Coverage plots for the ten HMPV whole genome sequences generated in this study. The trimmed sequence reads were mapped onto draft reference genomes to examine coverage along the virus genome. The y-axis shows the log-scaled sequence reads depth (coverage) and the x-axis shows the relative genome position according to the scale at the top. The mean coverage is provided for each sample.

CHAPTER 3

The Genomic Epidemiology of Human Metapneumovirus in Australia and the Impact of COVID-19

Chapter 3: The Genomic Epidemiology of Human Metapneumovirus in Australia and the Impact of COVID-19

Author Attribution Statement for Thesis with Publications

I am writing this letter to specify the role of Rachel Tulloch in the preparation and submission of the following manuscript:

Tulloch, RL, Horsburgh, BA, Cutmore, E, Tran, T, Jordan, P, Donovan, L, Ngo, C, Carter, I, Hammond, JM, Deveson, IW, Dwyer, D, Kok, J & Eden, J-S, 2022, 'The Genomic Epidemiology of Human Metapneumovirus in Australia and the impact of COVID-19' *Virus Evolution*, under review.

Contribution to the manuscript of all authors is as follows:

Task	Role of Author
Conceptualisation	R.L.T, D.E.D, J.K, J-S.E
Investigation	R.L.T, J-S.E
Methodology	R.L.T, J-S.E
Data curation	R.L.T
Laboratory Work	R.L.T, B.A.H, E.C, T.T, P.J, L.D, C.N, I.C, J.M.H
Formal analysis	R.L.T, I.W.D J-S.E
Visualisation	R.L.T, J-S.E
Writing – Original draft	R.L.T
Writing – Review and editing	R.L.T, B.A.H, I.W.D, D.E.D, J.K, E.C, T.T, P.J, L.D, C.N, I.C, J.M.H, J-S.E

I declare that the authorship attribution statements made above are true. In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author, my supervisor.

Rachel Tulloch
December 2022

As the primary supervisor for the candidature upon which the thesis is based, I can confirm that the above authorship attributions statements are correct.

John-Sebastian Eden
December 202

The Genomic Epidemiology of Human Metapneumovirus in Australia and the impact of COVID-19

Rachel L. Tulloch[1,2], Bethany A. Horsburgh[1,2,3], Elena Cutmore[1,2], Tyna Tran[4], Penelope Jordan[4], Linda Donovan[4], Christine Ngo[4], Ian Carter[4], Jillian M. Hammond[5], Ira W. Deveson[5,6,7], Dominic E. Dwyer[2,4], Jen Kok[4] and John-Sebastian Eden[1,2]

[1]Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia
[2]Sydney Institute for Infectious Diseases, Faculty of Health and Medicine, The University of Sydney, Westmead, NSW 2145, Australia
[3]Kirby Institute, The University of New South Wales, Kensington, NSW 2052, Australia
[4]NSW Health Pathology - Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia
[5]The Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
[6]The Kinghorn Centre for Clinical Genomics, Sydney, NSW 2010, Australia
[7]St Vincent's Clinical School, Department of Medicine, UNSW Sydney, Kensington, NSW 2052, Australia

Correspondence

Dr John-Sebastian Eden

Email: js.eden@sydney.edu.au

Phone: +61 2 8627 1817

Keywords

Human metapneumovirus; Whole-genome sequencing; Genomic epidemiology; COVID-19; non-pharmaceutical interventions.

Abstract

Human metapneumovirus (HMPV) is an important cause of both upper and lower respiratory tract disease, particularly in the very young and the elderly. Despite this, we know little about the sources and patterns of HMPV genetic diversity as limited genomic surveillance is conducted. In this study, we used a whole genome sequencing and phylogenetic approach to examine circulating HMPV diversity and assess the impact of COVID-19 on the breadth and number of lineages in New South Wales (NSW), Australia between 2013-2021. The pre-COVID-19 pandemic period was defined by co-circulation of multiple HMPV subtypes including A2, B1 and B2, and typical winter seasonality. Interestingly, we identified two novel HMPV-A2 variants containing insertions in the viral glycoprotein that emerged globally but had not previously been reported locally. Furthermore, following the emergence of SARS-CoV-2 and the introduction of non-pharmaceutical interventions we observed a marked reduction in HMPV incidence, with few cases detected throughout 2020. However, during the late winter period of 2021, coinciding with the large outbreak of SARS-CoV-2 Delta variant in NSW, HMPV cases surged, associated with a single HMPV-A2 variant. This collapse in circulating HMPV diversity following COVID-19 restrictions is consistent with other respiratory viruses such as influenza virus and respiratory syncytial virus. We provide greater genomic resolution of HMPV epidemiology in Australia over the last nine years and highlight the importance of ongoing surveillance and sequencing of HMPV in the post-pandemic era, as previous patterns of disease may be subject to abrupt and unexpected change.

Introduction

Human metapneumovirus (HMPV) is a leading cause of acute respiratory infection (ARI) in vulnerable populations, including young children, the elderly and immunocompromised (Bernadette et al., 2001, Kaslow et al., 2014). While most infections are mild and localized to the upper respiratory tract, lower respiratory tract infections can lead to pneumonia and bronchiolitis with patient outcome largely reliant on comorbidities and coinfections. Indeed, HMPV is responsible for up to 10% of all lower respiratory tract infection hospitalizations in children (Heikkinen et al., 2008) and like influenza and SARS-CoV-2, may cause residential care facility outbreaks (Jones et al, 2022). HMPV is endemic globally (Jin et al., 2015, Liu et al., 2015, Pollett et al., 2018) with cases typically peaking around late winter and spring in temperate climates, whilst in tropical climates peaks are observed between late spring and early summer (Broor et al., 2008). Notably, globally, HMPV activity often follows the peak in respiratory syncytial virus (RSV) activity (Adams et al., 2015, Ditt et al., 2011, Paget et al., 2011, Cattoir et al., 2019).

HMPV is a member of *Pneumoviridae* family, with a negative-sense, single-stranded RNA genome of approximately 13.3kb in length. The viral genome contains eight genes (N, P, M, F, M2, SH, G, and L) encoding nine proteins (Sutherland et al., 2001, El Najjar et al., 2016). HMPV is genetically related to RSV, however it lacks two non-structural genes found in RSV NS1 and NS2 (Pangesti et al., 2018). Despite the increasing availability of whole genome sequencing (WGS), the vast majority of HMPV epidemiological studies have been performed using partial genome sequencing of either the glycoprotein (G), fusion (F), or nucleoprotein (N) genes (Carneiro et al., 2009, Embarek Mohamed et al., 2014, Li et al., 2012, Ljubin-Sternak et al., 2008), limiting phylogenetic resolution (Embarek Mohamed et al., 2014). HMPV can be broadly classified into two genetic lineages, genotypes A and B (van Den Hoogen, 2007). Recent studies have proposed the further subdivision of HMPV into four subtypes: A1, A2, B1 and B2, using phylogenetic analysis of the G protein extracellular domain (Peret et al., 2004).

Whilst HMPV genotypes and subtypes are clinically indistinguishable, phylogenetic analysis has allowed for the identification and classification of multiple variants co-circulating with global distributions (Agapov et al., 2006, Jain et al., 2014, Naganori et al., 2020), with frequent switching of variant dominance (Gaunt et al., 2011). For example, recent studies have identified the emergence of two novel HMPV-A2 variants carrying G gene duplications, one with a 111-bp duplication and another with a 180-bp duplication (Saikusa et al., 2017a, Nidaira et al., 2012). These variants have only been reported in Japan and their existence in Australia had not been documented (Saikusa et al.,

2017a, Saikusa et al., 2017b). Interestingly, recent RSV subtypes, RSV-A ON1 and RSV-B BA10, contain nearly identical duplication events in the same attachment G glycoprotein, suggesting some convergence in genotype (Di Giallonardo et al., 2018a, Linstow et al., 2008, Paget et al., 2011, Pangesti et al., 2018, Shi et al., 2017). In both cases, these variants replaced existing genotypes to become the globally dominant strains, suggesting some increased epidemiological fitness over the previous circulating genotypes (Liang et al., 2019, Eden et al., 2021, Di Giallonardo et al., 2018a). Some studies have proposed that the global dominance of the RSV-A ON1 genotype is due to the lengthening of the G glycoprotein, which possibly increases glycosylation patterns and facilitates improved cellular attachments, increasing transmission in human airways (Liang et al., 2019, Ma et al., 2021).

During the COVID-19 pandemic, seasonal epidemics of common respiratory viruses including influenza, RSV and HMPV were disrupted both locally and globally (Yeoh et al., 2020, Binns et al., 2022). For example, in Australia, RSV and influenza virus activity were reduced by 98% and 99.4%, respectively in Western Australia during 2020 when compared to the previous winter seasons of 2012-2019 (Yeoh et al., 2020). Similarly, RSV cases reported in New South Wales (NSW) between April and June 2020 were also reduced by 94.3% compared to 2015-2019 (Britton et al., 2020). This trend was observed globally with United States, Chile, Taiwan, and South Africa reporting declines in influenza virus circulation coinciding with the emergence of SARS-CoV-2 and the implementation of COVID-19-related non-pharmaceutical interventions (NPIs) such as staged lockdowns, mask mandates, isolation protocols and border closures to restrict the spread of SARS-CoV-2 (Olsen et al., 2020, Chen et al., 2021). While the control of SARS-CoV-2 varied across different countries and locally by state, the marked reduction in seasonal respiratory virus activity was dramatic, particularly for influenza, RSV, and also HMPV (Eden et al., 2021). Importantly, COVID-related NPIs have not only had great impact on the incidence and seasonality of these viruses, but also lead to major reductions in circulating viral diversity (Chen et al., 2021, Eden et al., 2021, Eden et al., 2022, Fricke et al., 2021, Tazinya et al., 2018). It remains unclear how the collapse of diverse sublineages and genetic bottlenecking observed with RSV and influenza will impact activity and molecular epidemiology of other respiratory viruses going forward, however this is a clear challenge for vaccine development and highlights the importance of routine genomic surveillance for all respiratory viruses beyond SARS-CoV-2 (Eden et al., 2021).

In NSW, COVID-19-related NPIs have varied in duration and intensity along with multiple waves of SARS-CoV-2 activity. During the COVID-19 pandemic, HMPV activity has been largely

reduced and no winter peak was observed during 2020. However, in the late winter period of 2021, HMPV activity rose with a major outbreak across NSW. As there are less than 300 publicly available whole genome sequences published globally, we lack the genomic information required to inform our understanding of the emergence, evolution and spread of HMPV, and how the molecular epidemiology has changed due to the COVID-19 pandemic. This study aims to describe the genetic diversity of HMPV in NSW and investigate the impact that circulating strains have on outbreaks in Australia before and after-the COVID-19 pandemic using whole genome sequencing and phylogenetic approaches. By generating a large dataset of new HMPV genomes from clinical cases in NSW during peak and off-peak seasons over a 9-year period, we demonstrate the evolutionary dynamics and impact of COVID-19 on circulating HMPV diversity.

Materials and Methods

Epidemiological data

Weekly published respiratory virus surveillance reports were sourced from NSW Health published between 2009 to 2021 (available at: https://www.health.nsw.gov.au/Infectious/covid-19/Pages/weekly-reports.aspx). These reports contain summaries of the number of detections and tests performed from sentinel labs across NSW. Respiratory pathogens tested in this panel include Influenza A, Influenza B, Adenovirus, Parainfluenza, RSV, Rhinovirus, HMPV, Enterovirus and post 2020, SARS-CoV-2; serological diagnoses are not included in the count data. Raw data was collated for the years 2013 and 2016-2021 with total positive cases standardized per 1000 tests.

Sample Collection

Residual respiratory specimens submitted to NSW Health Pathology-Institute of Clinical Pathology and Medical Research (NSWHP-ICPMR) at Westmead Hospital for diagnostic testing were utilized in this study as per protocols approved by Western Sydney Local Health District Human Research Ethics Committee (Approval numbers LNR/17/WMEAD/128 and SSA/17/WMEAD/129). RNA was extracted using the MagNA Pure 96 (Roche Diagnostics GmbH, Mannheim, Germany) with the Viral NA Small Volume Kit per the manufacturer's instructions (Roche Molecular Diagnostics, Mannhiem, Germany), and then screened against a panel of known respiratory viruses using a laboratory developed quantitative PCR targeting influenza viruses A & B, influenza A subtypes H3N2 & pandemic H1N1, adenovirus, parainfluenza viruses 1, 2 & 3, rhinovirus, enterovirus, RSV and HPMV (Ratnamohan et al., 2014). Clinical samples were de-identified, with basic metadata recorded, including date of collection, patient age, sex and sampling location. Here,

we sequenced 185 samples collected between the years 2013, 2016-2021; collection dates were staggered across different seasons throughout the year to capture seasonal diversity.

Whole Genome Sequencing

An overlapping reverse transcription PCR (RT-PCR) method was used to amplify the entire length of the HMPV genome (Tulloch et al., 2021a). Briefly, viral extracts were first reversed transcribed using the Invitrogen SuperScript IV VILO Master Mix (Invitrogen, Carlsbad, CA, USA) before genome amplification across four parallel PCR reactions as described (Tulloch et al., 2021a). Resultant amplicons were pooled equally and purified using AMPure XP (Beckman Coulter, Pasadena, CA, USA). The pooled and purified amplicons were then prepared for sequencing using the Nextera XT DNA Library Preparation Kit and indexed using the v2 Set B unique dual indexing kit (Illumina, Massachusetts, MA, USA). Sequencing was performed on an Illumina iSeq (150 paired end reads) with approximately 40 individually indexed libraries per run resulting in a coverage depth of at least 1000X reads per viral genome. To confirm structural variation in the viral glycoprotein, select samples were subjected to long-read sequencing by Oxford Nanopore Technologies (ONT). Briefly, libraries were prepared from the pooled and purified amplicons using the rapid barcoding kit and sequenced on a MinION using a R9.4.1 flow cell to a depth of at least 100X coverage across the glycoprotein region.

Raw sequences generated using the Illumina sequencing platform were first quality trimmed using bbduk v38.86 (Bushnell, 2016) before *de novo* assembly using megahit v1.0.2 (Li et al., 2015). The trimmed reads were then re-mapped onto the draft viral contig (genome) using bbmap v38.86 (Bushnell, 2016). Overall mapping and de novo alignment quality was assessed by visual inspection in Geneious Prime 2020.1.2 (available from: https://www.geneious.com). Final majority consensus sequences were then extracted from alignments with terminal primer sequences removed before gene annotation using National Center for Biotechnology Information (NCBI) GenBank reference sequences (HMPV-A NC_004148 and HMPV-B HM197719). All sequences generated in this study were submitted to NCBI GenBank with the following accessions: OP904002 - OP904181.

Sequence alignments

The purpose of this study was to analyse newly generated sequences in both a local and global context, we therefore downloaded all HMPV complete or near-complete (>13kb) genomes from NCBI GenBank (as of 16th Feb 2022). Reference sequences were first imported into Geneious Prime

2020.1.2 and then sorted into their respective subtypes HMPV A1, A2, B1 and B2 using MAFFT (Rozewicki et al., 2019). Sequences that lacked spatial or temporal collection data were excluded. Sequences that were generated using highly passaged viruses, were synthetic, or containing long stretches of ambiguous bases (>20bp long) were also excluded from the study. The curated reference sequences were then combined with our 185 newly generated genome sequences using MAFFT. The final dataset used in this study consisted of 367 full length HMPV sequences, 185 A and 182 B sequences. All four HMPV sub-types were analysed independently.

Phylogenetic and temporal analysis

Full-length genome maximum likelihood trees for all subtypes were estimated using PhyML (Guindon et al., 2010) with the GTR + Gamma substitution model and 1000 bootstraps. Date of collection information was utilised in conjunction with available full length genome ML phylogenies to examine temporal relationships within the data. This analysis was performed with TempEst v1.5.3 using regression of root-to-tip genetic distance against time (Rambaut et al., 2016) and plotted in RStudio 2022.07.1. Evolutionary rates and time-scale phylogenies were then estimated with a Bayesian approach using BEAST v1.10.4 (Suchard et al., 2018). To determine the best parameters for our data we compared different clock (strict versus relaxed clock) and tree (constant versus Bayesian skyride) models using path-sampling. For each analysis, the codon-based SRD06 model was used with the chain length set to 50 million, sampled every 10,000 steps with at least two independent runs. The BEAST output was imported into Tracer v1.7.1 and examined to ensure sufficient sampling and to facilitate the removal of between 10-20% of the posterior distribution to ensure sampling during the stationary phase.

Results and Discussion

The impact of COVID-19 on the seasonality of HMPV in New South Wales, Australia

The epidemiology of HMPV in Australia is poorly described and genomic studies are lacking. Furthermore, the COVID-19 pandemic has had major impacts on the seasonal patterns and circulating diversity of endemic respiratory viruses such as influenza virus and RSV (Baker et al., 2020, Britton et al., 2020, Dhanasekaran et al., 2022), and for HMPV, the impact remains unknown. To bridge the gap in our knowledge of HMPV molecular epidemiology, we collected incidence data from reports publish by NSW Health sentinel laboratories and conducted a genomic survey using whole genome sequencing from strains between 2013 and 2021. To examine the seasonal distribution of HMPV in NSW, the number of positive cases was plotted by month during the pre- and pandemic periods following the emergence of COVID-19 (**Figure 1**). In NSW, HMPV activity typically peaks in late winter to early spring, with historical peaks occurring between the months of September and October.

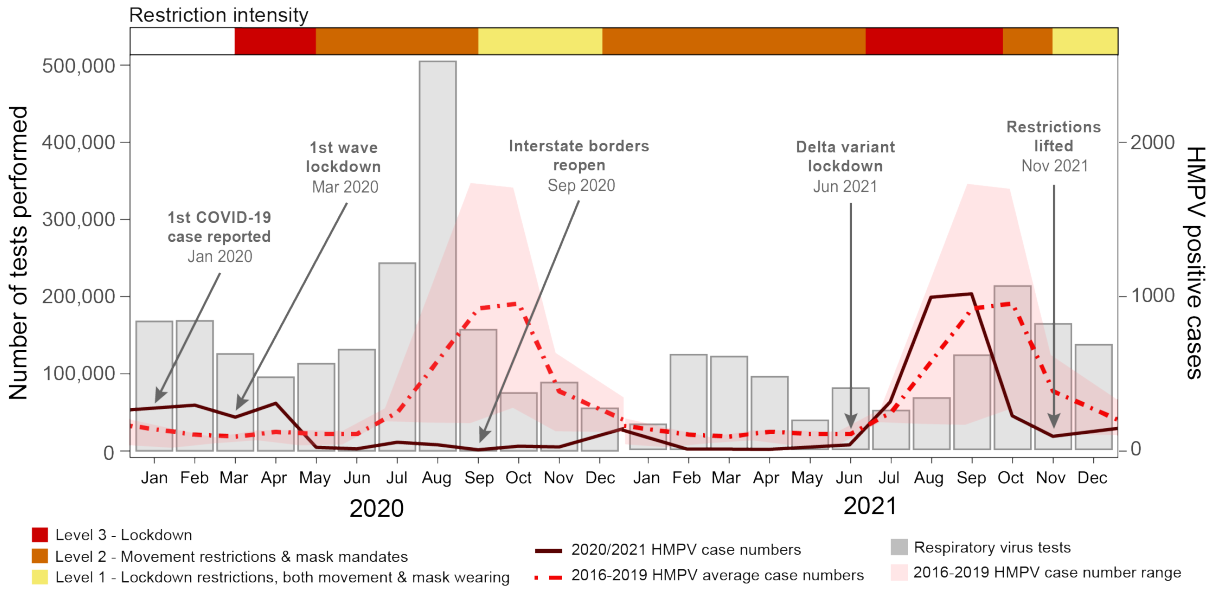


Figure 1: The epidemiology of HMPV in New South Wales between 2016 and 2021 – The grey shaded bar-plot with the left axis is the number of respiratory panel tests performed by NSW Health in 2020-21. The dot-dash red line represents the average number of HMPV positive cases by month between 2016 and 2019, while the solid blue and maroon lines are representative of the total number of HMPV cases per month in 2020 and 2021 respectively, the pink shaded region represents minimum and maximum monthly positive HMPV cases per month between 2016-2019, with those chart features corresponding to the right axis. The top panel shows the severity of public health restrictions and coloured according to the key provided.

Before the COVID-19 pandemic, HMPV seasonality was largely consistent with relatively minor fluctuations in the magnitude and timing of peak case numbers annually. However, in early to mid 2020, large disruptions to the expected pattern of HMPV incidence occurred, coinciding with the start of the COVID-19 pandemic. Here, unseasonably low cases of HMPV were observed between

April 2020 and December 2020, with the dramatic drop in case numbers largely due to the implementation of public health NPIs (**Figure 1**). The implementation of the first statewide lockdown in March 2020, and further introduction of NPIs to limit the spread of SARS-CoV-2 also greatly disrupted the typical seasonality of other common respiratory viruses such as the influenza viruses both locally and globally (Baker et al., 2020, Olsen et al., 2020, Rockett et al., 2020). A rise in HMPV cases was then observed during mid-2021, coinciding with the implementation of further COVID-19 restrictions to limit the increase of local SARS-CoV-2 transmission (Delta variant wave). Therefore, while the effectiveness of NPIs was apparent during 2020, the HMPV cases were seemingly sufficiently established that the 2021 surge suppressed by the introduction of COVID-19-related public health restrictions that included stay-at-home orders. A similar initial suppression of cases was seen in RSV during early 2020, however, this was maintained through-out the lockdown period and activity resumed only shortly after the easing of restrictions in late 2020. Of note, this resurgence was associated with a collapse in circulating diversity, with most pre-pandemic lineages becoming extinct, and the emergence of two geographically linked RSV-A variants that seeded further outbreaks interstate (Eden et al., 2022).

The impact of COVID-19 on the molecular epidemiology of HMPV in New South Wales, Australia

To examine the molecular epidemiology of HMPV in NSW before, during and after the COVID-19 restrictions were implemented, we sequenced the HMPV genome from 185 residual HMPV-positive clinical samples collected between 2013 and 2021 from Sydney and regional NSW. Noting that no samples were available to sequence between 2014-2016. Following genome sequencing, we used comparative analysis to identify the subtypes of samples sequenced in this study and describe the molecular epidemiology of HMPV in NSW during the study period **(Figure 2)**.

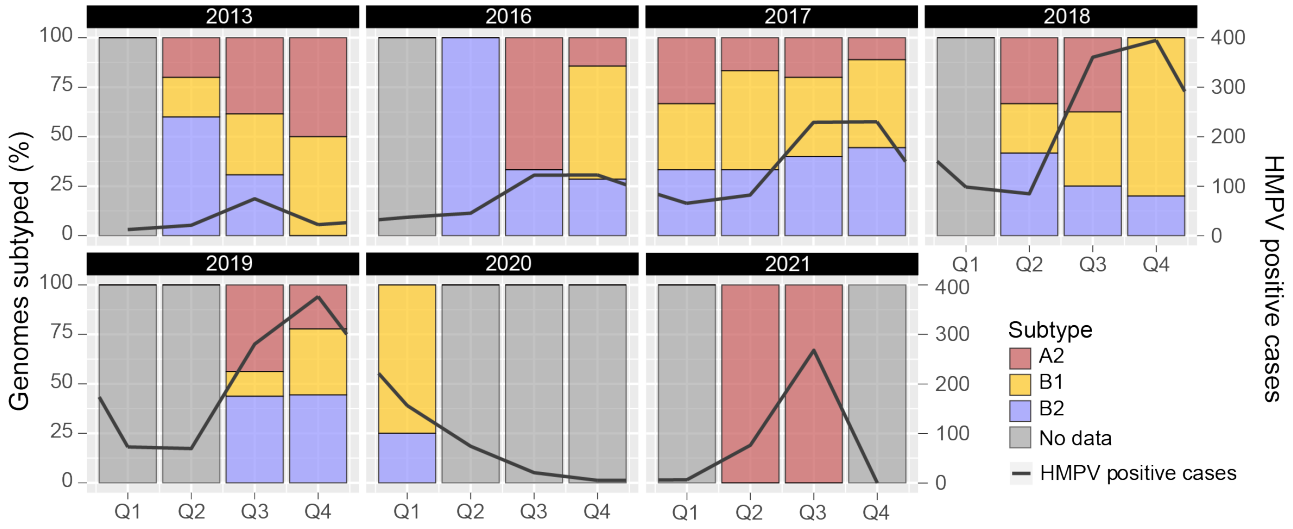


Figure 2: Seasonal distribution and subtypes detected by quarter - The total reported positive cases in NSW is portrayed on the right Y axis and shown by the bold line graph, whilst the number of genomes subtyped in this study on the right Y axis corresponds to the bar graph, and coloured according to the key provided. The X axis is labelled with quarters of each year.

In the pre-COVID era (2013-2019), the A2, B1 and B2 subtypes were found to be co-circulating with minor fluctuations in the predominant strain each season. Typically, each subtype was detected throughout the year except for Q2-Q3 2016, Q4 2018 and Q1 2020 however, there was on average less than 10 genomes per quarter for these periods. Therefore, this is likely due to under-sampling during these quarters, and does not necessarily reflect the temporal distribution of these subtypes in these periods. The A1 subtype was not identified in this study however, it is important to note that publicly available literature has not reported the identification of the A1 subtype since 2007, suggesting that this lineage may now extinct and no longer in circulation (Torres et al., 2012a). Interestingly, together the HMPV-B subtypes (B1 & B2) were largely dominant over HMPV-A (A2) in every year and quarter except Q3 2016 and throughout the entirety of 2021 in the COVID-19 period. Shifts in predominant subtypes between years and the apparent complex pattern of HMPV circulation in NSW has been mirrored in global long-term studies, including those conducted in the United States, Germany, China, Japan and South Korea (Lim et al., 2020b, Yang et al., 2009b, Reiche et al., 2014, Zhang et al., 2012). Following the decline in HMPV cases in 2020 and early 2021, we observed a collapse in diversity with only the A2 subtype detected in 2021, including during the period of peak activity in Q3. Interestingly, we did not identify any A2 subtype even during Q1 2020, immediately prior to COVID-19 restrictions and border closures. Therefore, our data suggests that following the relaxation of COVID-19 restrictions in NSW, a collapse in genetic diversity was observed, likely caused by COVID-19 related NPIs which is similar to influenza and RSV (Koutsakos et al., 2021).

Phylodynamics of HMPV in NSW, Australia - Evolution and spread pre-COVID-19

A phylogenetic analysis was performed by utilizing the 185 whole genome sequences generated in this study along with 192 reference sequences available on NCBI GenBank. The final alignments including sequences from 15 different countries across a 39-year period (1982-2021). This combined dataset of 377 samples comprised 190 and 187 HMPV-A and -B genomes, respectively. There is a clear lack of HMPV sampling and surveillance globally, consequently, approximately 49% of the total dataset was generated in this study from samples collected in NSW, between the years of 2013 and 2021 (**Supplementary Figure 1**). These significant sampling biases limited the scale of our phylogeographic analysis; therefore, sequences were grouped by continent of sampling (**Figure 3**). Our phylogenetic analysis of HMPV A2, B1 and B2 genomes showed that Australian circulating HMPV viruses were largely similar to those circulating globally. Like many respiratory pathogens, global diversity was linked to predominant subtypes and variants circulating in similar periods. This trend was also observed within our samples, at least prior to 2020 and the start of the COVID-19 pandemic (**Figure 3**).

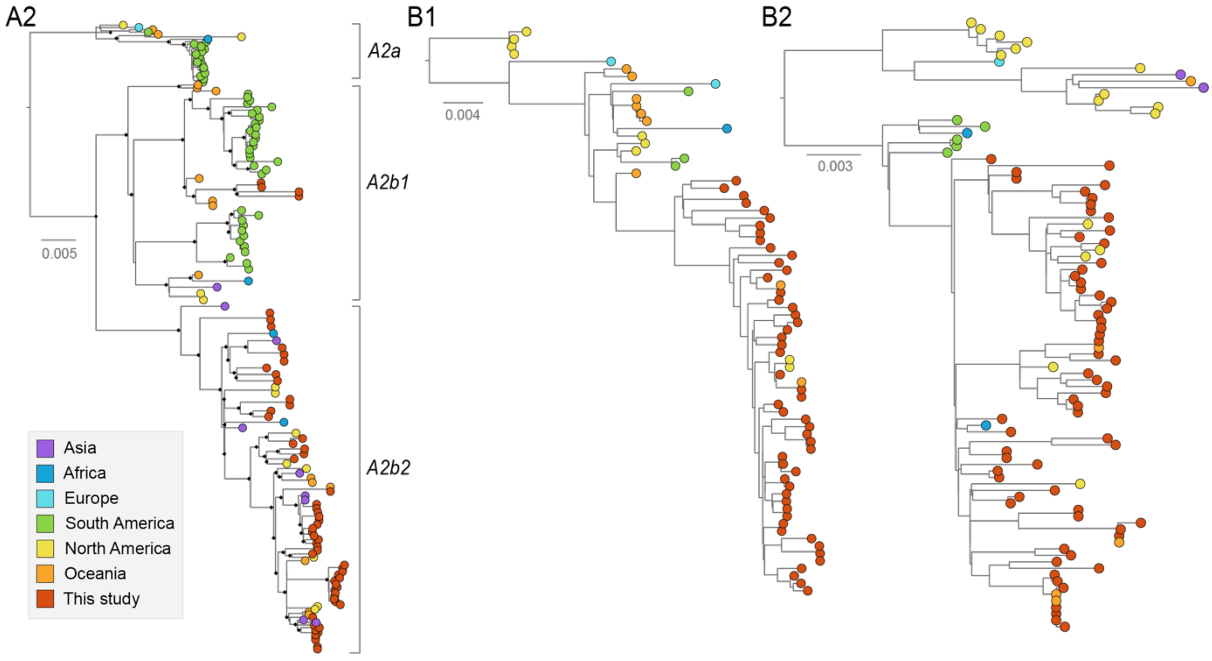


Figure 3: Phylogenetic analysis of global and local HMPV whole genome sequences. HMPV sequences in this study were aligned with all available HMPV sequences from NCBI GenBank and separated by subtype. Phylogenetic trees were then estimated for each subtype alignment using a minimum likelihood approach. Country of collection has been colored by broader region according to the key provided. Branches are proportional to the number of nucleotide substitutions per site. Diamonds at nodes indicate bootstrap support values >70%.

Currently, the A2 subtype at present can be further broken down into 3 distinct sub-lineages, A2a, A2b1 and A2b2 (Saikusa et al., 2017a), and here most HMPV-A2 strains were identified with

other global sequences as A2b2 sub-lineage. Within this A2b2 sub-lineage we identified two distinct duplication insertions (nt-dup) in the glycoprotein, one a 111bp insertion and the other a 180bp insertion (HMPV $A2b_{180nt\text{-}dup}$ or $A2b_{111nt\text{-}dup}$). These insertion subtypes have been identified globally, however this study is the first to report these variants locally in Australia (Naganori et al., 2020, Piñana et al., 2016, Saikusa et al., 2017a). Further analysis of the glycoprotein sequences generated in this revealed all A2 samples collected after 2017 were either the $A2b_{180nt\text{-}dup}$ or $A2b_{111nt\text{-}dup}$ variant, suggesting a replacement of previous HMPV A2 strains, and possible increased epidemiological fitness with these novel glycoprotein variants. Whilst we did not examine the phenotypes of these duplication events in this study, others have proposed these changes may be advantageous due to the lengthening of the major attachment protein of the virus and possible increased glycosylation patterns (Naganori et al., 2020, Saikusa et al., 2017a). Interestingly, equivalent duplication events have been described for RSV including the ON1 RSV-A genotype. Notably it contained a 72-nt duplication which quickly spread and seemingly replaced other RSV-A subtypes globally (Aneley Getahun et al., 2019, Shi et al., 2017, Tabatabai et al., 2014). No structural variants were observed for the HMPV-B viruses; however they remained circulating throughout the study period until the COVID-19-related bottleneck. Interestingly, the B1 and B2 viruses show distinct phylodynamic patterns, with the B1 viruses typified by shorter branch lengths and higher turnover of circulation lineages. While the B2 viruses showed deeper internal branches with higher degree of cocirculating clades compared to B1, with evidence of multi-year persistence of some B2 sub-lineages. Across both the HMPV-A and -B trees, some samples were found to cluster by time and place, suggesting possible local outbreaks.

The emergence of a novel HMPV-A2 variant in the post-COVID-19 period

As HMPV cases increased during 2021 in the post-COVID-19 period, we specifically sequenced 14 samples collected from seven geographically dispersed areas across NSW between June and July of 2021 (**Figure 3**) to compare to pre-COVID-19 circulating diversity and characterize the 2021 winter outbreak. The 14 sequences were found to be highly clonal, and all belonged to a novel A2b2 clade carrying the 111nt G gene duplication, here termed the NSW 2021 $A2b2_{111nt\text{-}dup}$ variant. The clade containing the post-COVID-19 sequences appeared to have no immediate ancestor based on whole-genome sequences (**Figure 3**). A significant limitation of this analysis was the sampling bias towards recent cases originating from NSW and the limited whole-genome data that is available for comparison. Therefore, to determine if there was any recent ancestor of the post-COVID-19 clade, we expanded our analysis to examine all available partial HMPV-A2 gene sequences available on

the NCBI database (**Supplementary Figure 2**). HMPV sequences in this study were aligned with all available HMPV-A2 sequences from NCBI GenBank. 3330 A2 Fusion gene sequences download from NCBI and combined with 190 study samples (total alignment sequences, n= 3520). Aligned with MAFFT and ML tree estimated using FastTree 2.1.11 with GTR subs model (Ward et al., 2020). The fusion glycoprotein coding region was extracted, and sequences less than 300 nucleotide were removed. No sequences sourced globally were found to be immediately related to 2021 outbreak clade suggesting the source remain unknown. Samples were grouped by country of origin and branch lengths constructed with relation to the number of nucleotide substitutions per site to provide a measure of sequence variation. Despite this, a recent ancestor could not be identified using the extended dataset. This again highlights major sampling biases towards limited regions including Europe, Asia and North America, and greater focus on other respiratory pathogens such as COVID-19, influenza virus and RSV.

Next, we sought to employ a Bayesian approach to estimate the time of origin of the novel NSW 2021 A2b2$_{111nt\text{-}dup}$ variant. While we saw clear temporal structure by root-to-tip distance plotted against collection date across the genome for each subtype alignment (**Supplementary Figure 3**), we limited our dating analysis to the coding region of RNA-directed RNA polymerase (L protein), due to its increased evolutionary stability and reduced impact of time-dependent rate inflation biases (Eden et al., 2022). Initially, we tested combinations of clock and coalescent models to obtain the best fitting statistical model according to path sampling and stepping-stone analysis (**Supplementary Table 1**), which for the A2 subtype dataset was a relaxed clock (uncorrelated lognormal distribution) with an exponential growth tree prior. This analysis showed that the common ancestor of the NSW 2021 outbreak clade was in circulation during 2020 (0.6 – 1.72 years prior to July 2021) and may have originated locally in Australia at the start of pandemic or after the first waves and initial lockdowns ended (**Figure 4**). Whilst a significant limitation of this study was the lack of global and regional sampling, it is important to recognize that the broader lineage defining the NSW 2021 outbreak viruses is poorly sampled and represents at least 6 years of unsampled diversity. Regardless of timing and origins, we demonstrate a remarkable collapse in circulating diversity for HMPV in Australia, similar to influenza virus and RSV (Eden et al., 2022), again highlighting how the molecular epidemiology of these endemic respiratory viruses have changed dramatically following the COVID-19 pandemic.

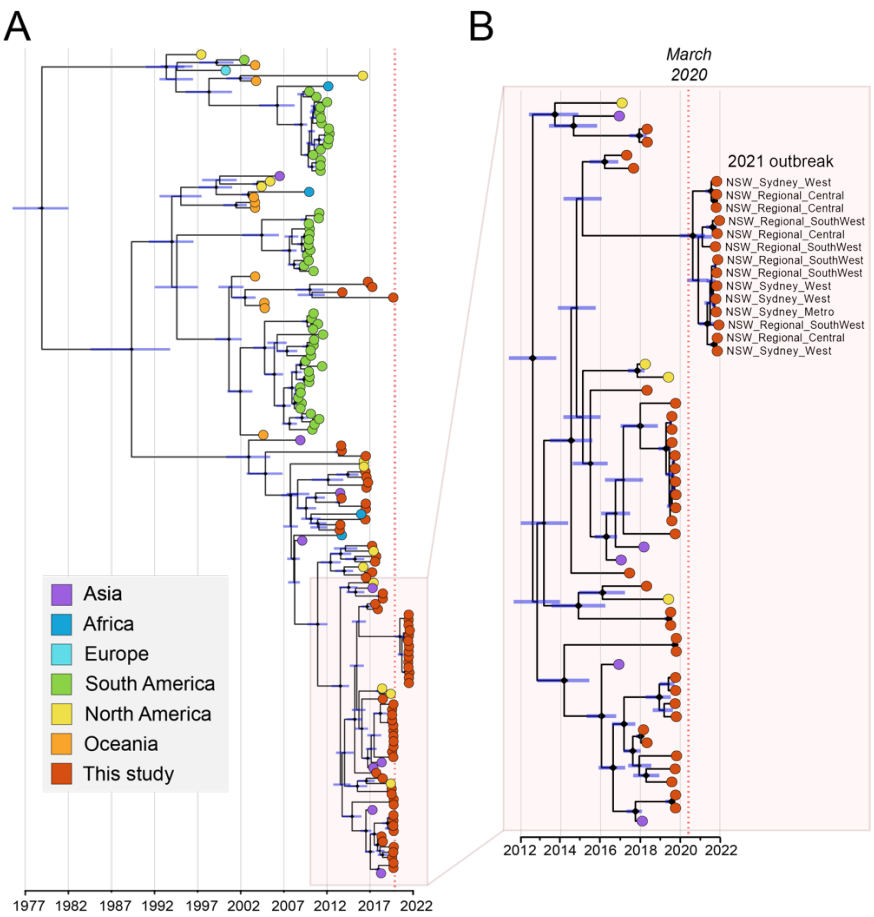


Figure 4: A time-scaled phylogenetic analysis of HMPV-A2 subtype. A. Complete HMPV-A2 L gene sequences generated in this study were aligned with NCBI GenBank sequences. These alignments were analysed using a Bayesian molecular clock approach estimated with BEAST (v1.10.a). **B.** An annotated subset of main phylogeny, with sampling location listed for samples collected post lockdown in the 2021 outbreak period. The dotted line represents the first wave of COVID-19 lockdowns in metropolitan NSW (16th March 2020). Country of collection has been coloured by broader region according to the key provided. Branches are proportional to the time with coloured bars at nodes indicating the 95% highest probability density for node ages. Diamonds at nodes indicate posterior probability support values >0.9.

Conclusions

HMPV remains a virus of significant importance to public health, especially in the young, elderly, and immune compromised. Here we use whole genome sequencing and a phylodynamic approach to provide a high-resolution genomic study of HMPV molecular epidemiology in Australia. Before the COVID-19 pandemic, we identified cocirculating subtypes both globally and locally, with predominant strains varying from year to year and season to season. We also report the identification of 111nt-dup and 180nt-dup A2 subtypes with structural duplications in the viral glycoprotein circulating within Australia that appear to have emerged globally in 2013 to replace wild-type A2 viruses. Post-COVID-19, we saw an initial dramatic decline in HMPV cases, likely due to the NPIs put in place to combat the initial pandemic, followed by a surge in cases in winter 2021 after relaxation

of public health measures before then peaking with a major wave of SARS-CoV-2 Delta variant. Of note, a major genetic bottleneck occurred during the COVID-19 restrictions of 2020 with the 2021 outbreak dominated by a single A2 clade, with the place of origin yet to be determined but possibly early in the pandemic and most therefore likely locally in Australia. This study also highlights the lack of HMPV whole genomes both globally and locally, and biases in sampling that overwhelmingly limits the comprehensive analysis of the global distribution and evolution of HMPV.

Author Contributions

Conceived the project: R.L.T, D.E.D., J.K & J.S.E.; processed samples: T.T., P.J., L.D., C.N. & I.C.; performed laboratory work: R.L.T. B.A.H., E.C., T.T., P.J., L.D., C.N., I.C & J.M.H.; analysed the data: R.L.T., I.W.D. & J.S.E.; wrote the paper: R.L.T. & J.S.E. All authors agree to be accountable for all aspects of the work and read and approved the final manuscript.

Acknowledgements

We acknowledge the University of Sydney's high-performance computing cluster Artemis for providing the computing resources used for this study

Funding

Funding was provided through the ICPMR Private Practice Trust fund to J.S.E., and J.K., the National Health and Medical Research Council Centre of Research Excellence in Emerging Infectious Diseases (#1102962) to J.S.E., and the Sydney Institute for Infectious Diseases at the University of Sydney to J.S.E

Competing interests

The authors declare that they have no competing interests.

References

ADAMS, O., WEIS, J., JASINSKA, K., VOGEL, M. & TENENBAUM, T. 2015. Comparison of human metapneumovirus, respiratory syncytial virus and Rhinovirus respiratory tract infections in young children admitted to hospital. *Journal of Medical Virology,* 87**,** 275-280.

AGAPOV, E., SUMINO, K., GAUDREAULT-KEENER, M., STORCH, G. & HOLTZMAN, M. 2006. Genetic Variability of Human Metapneumovirus Infection: Evidence of a Shift in Viral Genotype without a Change in Illness. *Journal of Infectious Diseases,* 193**,** 396-403.

ANELEY GETAHUN, S., PARRY, C. M., CRUMP, J. A., ROSA, V., JENNEY, A., NAIDU, R., MULHOLLAND, K. & STRUGNELL, R. A. 2019. A retrospective study of patients with blood culture-confirmed typhoid fever in Fiji during 2014-2015: epidemiology, clinical features, treatment and outcome. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 113**,** 764-770.

BAKER, R. E., PARK, S. W., YANG, W., VECCHI, G. A., METCALF, C. J. E. & GRENFELL, B. T. 2020. The impact of COVID-19 nonpharmaceutical interventions on the future dynamics of endemic infections. *Proceedings of the National Academy of Sciences - PNAS,* 117**,** 30547.

BERNADETTE, G. V. D. H., JAN, C. D. J., JAN, G., THIJS, K., RONALD DE, G., RON, A. M. F. & ALBERT, D. M. E. O. 2001. A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nature Medicine,* 7**,** 719.

BINNS, E., KOENRAADS, M., HRISTEVA, L., FLAMANT, A., BAIER-GRABNER, S., LOI, M., LEMPAINEN, J., OSTERHELD, E., RAMLY, B., CHAKAKALA-CHAZIYA, J., ENAGANTHI, N., SIMÓ NEBOT, S. & BUONSENSO, D. 2022. Influenza and respiratory syncytial virus during the COVID-19 pandemic: Time for a new paradigm? *Pediatric pulmonology,* 57**,** 38-42.

BRITTON, P. N., HU, N., SARAVANOS, G., SHRAPNEL, J., DAVIS, J., SNELLING, T., DALBY-PAYNE, J., KESSON, A. M., WOOD, N., MACARTNEY, K., MCCULLAGH, C. & LINGAM, R. 2020. COVID-19 public health measures and respiratory syncytial virus.

BROOR, S., BHARAJ, P. & CHAHAR, H. 2008. Human metapneumovirus: a new respiratory pathogen. *Journal of Biosciences,* 33**,** 483-493.

BUSHNELL 2016. BBMap short-read aligner, and other bioinformatics tools.

CARNEIRO, B. M., YOKOSAWA, J., ARBIZA, J., COSTA, L. F., MIRAZO, S., NEPOMUCENO, L. L., OLIVEIRA, T. F. M., GOULART, L. R., VIEIRA, C. U., FREITAS, G. R. O., PAULA, N. T. & QUEIRÓZ, D. A. O. 2009. Detection of all four human metapneumovirus subtypes in nasopharyngeal specimens from children with respiratory disease in Uberlândia, Brazil. *Journal of Medical Virology,* 81**,** 1814-1818.

CATTOIR, L., VANKEERBERGHEN, A., BOEL, A., VAN VAERENBERGH, K. & DE BEENHOUWER, H. 2019. Epidemiology of RSV and hMPV in Belgium: a 10-year follow-up. *Acta clinica belgica (English ed. Online),* 74**,** 229-235.

CHEN, A. P. L., CHU, I. Y. H., YEH, M. L., CHEN, Y. Y., LEE, C. L., LIN, H. H., CHAN, Y. J. & CHEN, H. P. 2021. Differentiating impacts of non-pharmaceutical interventions on non-coronavirus disease-2019 respiratory viral infections: Hospital-based retrospective observational study in Taiwan. *Influenza and other respiratory viruses,* 15**,** 478-487.

DHANASEKARAN, V., SULLIVAN, S., EDWARDS, K. M., XIE, R., KHVOROV, A., VALKENBURG, S. A., COWLING, B. J. & BARR, I. G. 2022. Human seasonal influenza under COVID-19 and the potential consequences of influenza lineage elimination. *Nature communications,* 13**,** 1721-1721.

DI GIALLONARDO, F., KOK, J., FERNANDEZ, M., CARTER, I., GEOGHEGAN, J. L., DWYER, D. E., HOLMES, E. C. & EDEN, J.-S. 2018. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses,* 10.

DITT, V., LÜSEBRINK, J., TILLMANN, R. L., SCHILDGEN, V. & SCHILDGEN, O. 2011. Respiratory infections by HMPV and RSV are clinically indistinguishable but induce different host response in aged individuals. *PloS one,* 6**,** e16314-e16314.

EDEN, J.-S., SIKAZWE, C., XIE, R., DENG, Y.-M., SULLIVAN, S. G., MICHIE, A., LEVY, A., CUTMORE, E., BLYTH, C. C., BRITTON, P. N., CRAWFORD, N., DONG, X., DWYER, D. E., EDWARDS, K. M., HORSBURGH, B. A., FOLEY, D., KENNEDY, K., MINNEY-SMITH, C., SPEERS, D., TULLOCH, R. L., HOLMES, E. C., DHANASEKARAN, V., SMITH, D. W., KOK, J. & BARR, I. G. 2022. Off-season RSV epidemics in Australia after easing of COVID-19 restrictions. *Nature communications,* 13**,** 2884-2884.

EL NAJJAR, F., CIFUENTES-MUNOZ, N., CHEN, J., ZHU, H., BUCHHOLZ, U. J., MONCMAN, C. L. & DUTCH, R. E. 2016. Human metapneumovirus Induces Reorganization of the Actin Cytoskeleton for Direct Cell-to-Cell Spread.(Research Article). *PLoS Pathogens,* 12**,** e1005922.

EMBAREK MOHAMED, M., REICHE, J., JACOBSEN, S., THABIT, A., BADARY, M., BRUNE, W., SCHWEIGER, B. & OSMANN, A. 2014. Molecular Analysis of Human Metapneumovirus Detected in Patients with Lower Respiratory Tract Infection in Upper Egypt. *International Journal of Microbiology,* 2014.

FRICKE, L. M., GLÖCKNER, S., DREIER, M. & LANGE, B. 2021. Impact of non-pharmaceutical interventions targeted at COVID-19 pandemic on influenza burden – a systematic review. *The Journal of infection,* 82**,** 1-35.

GAUNT, E. R., JANSEN, R. R., POOVORAWAN, Y., TEMPLETON, K. E., TOMS, G. L. & SIMMONDS, P. 2011. Molecular epidemiology and evolution of human respiratory syncytial virus and human metapneumovirus. *PloS one,* 6**,** e17427.

GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic biology,* 59**,** 307-321.

HEIKKINEN, T., OSTERBACK, R., PELTOLA, V., JARTTI, T. & VAINIONPÄÄ, R. 2008. Human metapneumovirus infections in children. *Emerging infectious diseases,* 14**,** 101-106.

JAIN, B., SINGH, A. K., DANGI, T., AGARWAL, A., VERMA, A. K., DWIVEDI, M., SINGH, K. P. & JAIN, A. 2014. High prevalence of human metapneumovirus subtype B in cases presenting as severe acute respiratory illness: an experience at tertiary care hospital. *The Clinical Respiratory Journal,* 8**,** 225-33.

JIN, L., ÖRVELL, C., MYERS, R., ROTA, P. A., NAKAYAMA, T., FORCIC, D., HIEBERT, J. & BROWN, K. E. 2015. Genomic diversity of mumps virus and global distribution of the 12 genotypes. *Reviews in medical virology,* 25**,** 85-101.

KASLOW, R. A., STANBERRY, L. R. & LE DUC, J. W. 2014. *Viral Infections of Humans Epidemiology and Control,* Boston, MA, Springer US.

KOUTSAKOS, M., WHEATLEY, A. K., LAURIE, K., KENT, S. J. & ROCKMAN, S. 2021. Influenza lineage extinction during the COVID-19 pandemic? *Nature reviews. Microbiology,* 19**,** 741-742.

LI, D., LIU, C.-M., LUO, R., SADAKANE, K. & LAM, T.-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

LI, J., REN, L., GUO, L., PARANHOS-BACCALÀ, G., VERNET, G. & WANG, J. 2012. Evolutionary Dynamics Analysis of Human Metapneumovirus Subtype A2: Genetic Evidence for Its Dominant Epidemic. *PLoS One,* 7**,** e34544.

LIANG, X., LIU, D.-H., CHEN, D., GUO, L., YANG, H., SHI, Y.-S., WANG, Y.-J., WANG, W.-K., XIE, Z.-P., GAO, H.-C., DUAN, Z.-J. & ZHANG, R.-F. 2019. Gradual replacement of all previously circulating respiratory syncytial virus A strain with the novel ON1 genotype in Lanzhou from 2010 to 2017. *Medicine (Baltimore),* 98**,** e15542-e15542.

LIM, Y. K., KWEON, O. J., KIM, H. R., KIM, T.-H. & LEE, M.-K. 2020a. Clinical Features, Epidemiology, and Climatic Impact of Genotype-specific Human Metapneumovirus Infections: Long-term Surveillance of Hospitalized Patients in South Korea. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America,* 70**,** 2683-2694.

LIM, Y. K., KWEON, O. J., KIM, H. R., KIM, T.-H. & LEE, M.-K. 2020b. Clinical Features, Epidemiology, and Climatic Impact of Genotype-specific Human Metapneumovirus Infections: Long-term Surveillance of Hospitalized Patients in South Korea. *Clinical infectious diseases,* 70**,** 2683-2694.

LINSTOW, M.-L., HØGH, M., NORDBØ, S., EUGEN-OLSEN, J., KOCH, A. & HØGH, B. 2008. A community study of clinical traits and risk factors for human metapneumovirus and respiratory syncytial virus infection during the first year of life. *European Journal of Pediatrics,* 167**,** 1125-1133.

LIU, L. P., OZA, S. M., HOGAN, D. P., PERIN, J. P., RUDAN, I. P., LAWN, J. E. P., COUSENS, S. P., MATHERS, C. P. & BLACK, R. E. P. 2015. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet (British edition),* 385**,** 430-440.

LJUBIN-STERNAK, S., ŠANTAK, M., ČEPIN-BOGOVIĆ, J., BAĆE, A., VOJNOVIĆ, G., MLINARIĆ-GALINOVIĆ, G., FORČIĆ, D., DRAŽENOVIĆ, V. & FALSEY, A. R. 2008. Detection of genetic lineages of human metapneumovirus in Croatia during the winter season 2005/2006. *Journal of Medical Virology,* 80**,** 1282-1287.

MA, Y., JIANG, H., WAN, Z., LI, S., LI, Y., WANG, W., JIN, X., LI, Y. & ZHANG, C. 2021. Evolutionary dynamics of group A and B respiratory syncytial virus in China, 2009-2018. *Archives of virology,* 166**,** 2407-2418.

NAGANORI, N., MIWAKO, S., KO, S., TSUYOSHI, S., SHUZO, U., NOBUKO, T., HIDEKAZU, N. & MAKOTO, T. 2020. Recent Molecular Evolution of Human Metapneumovirus (HMPV): Subdivision of HMPV A2b Strains. *Microorganisms (Basel),* 8**,** 1280.

NIDAIRA, M., TAIRA, K., HAMABATA, H., KAWAKI, T., GUSHI, K., MAHOE, Y., MAESHIRO, N., AZAMA, Y., OKANO, S., KYAN, H., KUDAKA, J., TSUKAGOSHI, H., NODA, M. & KIMURA, H. 2012. Molecular epidemiology of human metapneumovirus from 2009 to 2011 in Okinawa, Japan. *Japanese journal of infectious diseases,* 65**,** 337-340.

OLSEN, S. J., AZZIZ-BAUMGARTNER, E., BUDD, A. P., BRAMMER, L., SULLIVAN, S., PINEDA, R. F., COHEN, C. & FRY, A. M. 2020. Decreased Influenza Activity During the COVID-19 Pandemic - United States, Australia, Chile, and South Africa, 2020. *MMWR. Morbidity and mortality weekly report,* 69**,** 1305-1309.

PAGET, S. P., ANDRESEN, D. N., KESSON, A. M. & EGAN, J. R. 2011. Comparison of human metapneumovirus and respiratory syncytial virus in children admitted to a paediatric intensive care unit. *Journal of Paediatrics and Child Health,* 47**,** 737-741.

PANGESTI, K. N. A., ABD EL GHANY, M., WALSH, M. G., KESSON, A. M. & HILL-CAWTHORNE, G. A. 2018. Molecular epidemiology of respiratory syncytial virus. *Reviews in Medical Virology,* 28**,** n/a-n/a.

PERET, T. C. T., ABED, Y., ANDERSON, L. J., ERDMAN, D. D. & BOIVIN, G. 2004. Sequence polymorphism of the predicted human metapneumovirus G glycoprotein. *Journal of general virology,* 85**,** 679-686.

PIÑANA, M., VILA, J., GIMFERRER, L., VALLS, M., ANDRÉS, C., RAMÓN, J., CODINA, M. G., MARTÍN, M. D. C., FUENTES, F., SAIZ, R., ALCUBILLA, P. & RODRIGO, C. 2016. Genetic variability of human metapneumovirus A strain circulating in Catalonia during the 2014–2015 and 2015–2016 seasons: A 180-nucleotide G gene duplication reported. *Journal of Clinical Virology,* 82**,** S125-S125.

POLLETT, S., TROVAO, N. S., TAN, Y., EDEN, J.-S., HALPIN, R. A., BERA, J., DAS, S. R., WENTWORTH, D., OCANA, V., MENDOCILLA, S. M., ALVAREZ, C., CALISTO, M. E., GARCIA, J., HALSEY, E., AMPUERO, J. S., NELSON, M. I. & LEGUIA, M. 2018. The transmission dynamics and diversity of human metapneumovirus in Peru.(Report). *Influenza and Other Respiratory Viruses,* 12**,** 508.

RAMBAUT, A., LAM, T. T., MAX CARVALHO, L. & PYBUS, O. G. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution,* 2**,** vew007-vew007.

REICHE, J., JACOBSEN, S., NEUBAUER, K., HAFEMANN, S., NITSCHE, A., MILDE, J., WOLFF, T. & SCHWEIGER, B. 2014. Human Metapneumovirus: Insights from a Ten-Year Molecular and Epidemiological Analysis in Germany. *PLoS One,* 9**,** e88342.

ROCKETT, R. J., ARNOTT, A., LAM, C., SADSAD, R., TIMMS, V., GRAY, K.-A., EDEN, J.-S., CHANG, S., GALL, M., DRAPER, J., SIM, E. M., BACHMANN, N. L., CARTER, I., BASILE, K., BYUN, R.,

O'SULLIVAN, M. V., CHEN, S. C. A., MADDOCKS, S., SORRELL, T. C., DWYER, D. E., HOLMES, E. C., KOK, J., PROKOPENKO, M. & SINTCHENKO, V. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine,* 26**,** 1398-1404.

ROZEWICKI, J., LI, S., AMADA, K. M., STANDLEY, D. M. & KATOH, K. 2019. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic acids research,* 47**,** W5-W10.

SAIKUSA, M., KAWAKAMI, C., NAO, N., TAKEDA, M., USUKU, S., SASAO, T., NISHIMOTO, K. & TOYOZAWA, T. 2017a. 180-Nucleotide Duplication in the G Gene of Human metapneumovirus A2b Subgroup Strains Circulating in Yokohama City, Japan, since 2014. *Frontiers in microbiology,* 8**,** 402-402.

SAIKUSA, M., NAO, N., KAWAKAMI, C., USUKU, S., SASAO, T., TOYOZAWA, T., TAKEDA, M. & OKUBO, I. 2017b. A novel 111-nucleotide duplication in the G gene of human metapneumovirus. *Microbiology and immunology,* 61**,** 507-512.

SHI, T., MCALLISTER, D. A., O'BRIEN, K. L., SIMOES, E. A. F., MADHI, S. A., GESSNER, B. D., POLACK, F. P., BALSELLS, E., ACACIO, S., AGUAYO, C., ALASSANI, I., ALI, A., ANTONIO, M., AWASTHI, S., AWORI, J. O., AZZIZ-BAUMGARTNER, E., BAGGETT, H. C., BAILLIE, V. L. & BALMASEDA, A. 2017. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *The Lancet,* 390**,** 946-958.

SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. & RAMBAUT, A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus evolution,* 4**,** vey016-vey016.

SUTHERLAND, K. A., COLLINS, P. L. & PEEPLES, M. E. 2001. Synergistic Effects of Gene-End Signal Mutations and the M2-1 Protein on Transcription Termination by Respiratory Syncytial Virus. *Virology,* 288**,** 295-307.

TABATABAI, J., PRIFERT, C., PFEIL, J., GRULICH-HENN, J. & SCHNITZLER, P. 2014. Novel respiratory syncytial virus (RSV) genotype ON1 predominates in Germany during winter season 2012-13. *PloS one,* 9**,** e109191-e109191.

TAZINYA, A. A., HALLE-EKANE, G. E., MBUAGBAW, L. T., ABANDA, M., ATASHILI, J. & OBAMA, M. T. 2018. Risk factors for acute respiratory infections in children under five years attending the Bamenda Regional Hospital in Cameroon. *BMC pulmonary medicine,* 18**,** 7-7.

TORRES, F. S., IOLSTER, J. T., SCHNITZLER, A. E., FARIAS, C. J., BORDOGNA, J. A., RUFACH, J. D., MONTES, G. M., SIABA, E. A., RODRÍGUEZ, E. M., JABORNISKY, E. R., COLMAN, E. C., FERNÁNDEZ, E. A., CAPROTTA, E. G., DIAZ, E. S., POTERALA, E. R., DE MEYER, E. M., PENAZZI, E. M., GONZÁLEZ, E. G., SAENZ, E. S., RECUPERO, E. O., ZAPICO, E. L., ALARCON, E. B., ARIEL, E. E., MINCES, E. P., MARI, E. E., CARNIE, E. A., GAREA, E. M. & JAEN, E. R. 2012. High mortality in patients with influenza A pH1N1 2009 admitted to a pediatric intensive care unit: A predictive model of mortality*. *Pediatric Critical Care Medicine,* 13**,** e78-e83.

TULLOCH, R. L., KOK, J., CARTER, I., DWYER, D. E. & EDEN, J.-S. 2021. An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus. *Viruses,* 13**,** 499.

VAN DEN HOOGEN, B. G. 2007. Respiratory tract infection due to human metapneumovirus among elderly patients. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America,* 44**,** 1159-1160.

YANG, C.-F., WANG, C. K., TOLLEFSON, S. J., PIYARATNA, R., LINTAO, L. D., CHU, M., LIEM, A., MARK, M., SPAETE, R. R., CROWE, J. J. E. & WILLIAMS, J. V. 2009. Genetic diversity and evolution of human metapneumovirus fusion protein over twenty years. *Virology journal,* 6**,** 138-138.

YEOH, D. K., FOLEY, D. A., MINNEY-SMITH, C. A., MARTIN, A. C., MACE, A. O., SIKAZWE, C. T., LE, H., LEVY, A., BLYTH, C. C. & MOORE, H. C. 2020. Impact of Coronavirus Disease 2019 Public Health Measures on Detections of Influenza and Respiratory Syncytial Virus in Children During the 2020 Australian Winter. *Clinical Infectious Diseases,* 72**,** 2199-2202.

ZHANG, C., DU, L.-N., ZHANG, Z.-Y., QIN, X., YANG, X., LIU, P., CHEN, X., ZHAO, Y., LIU, E.-M. & ZHAO, X.-D. 2012. Detection and Genetic Diversity of Human Metapneumovirus in Hospitalized Children with Acute Respiratory Infections in Southwest China. *Journal of Clinical Microbiology,* 50**,** 2714-2719.

ZHANG, Y., WEI, Y., ZHANG, X., CAI, H., NIEWIESK, S. & LI, J. 2014. Rational design of human metapneumovirus live attenuated vaccine candidates by inhibiting viral mRNA cap methyltransferase. *Journal of virology,* 88**,** 11411-11429.

Supplementary Figures

The Genomic Epidemiology of Human Metapneumovirus in Australia and the impact of COVID-19

Rachel L. Tulloch[1,2], Bethany A. Horsburgh[1,2,3], Elena Cutmore[1,2], Tyna Tran[4], Penelope Jordan[4], Linda Donovan[4], Christine Ngo[4], Ian Carter[4], Jillian M. Hammond[5], Ira W. Deveson[5,6,7], Dominic E. Dwyer[2,4], Jen Kok[4] and John-Sebastian Eden[1,2]

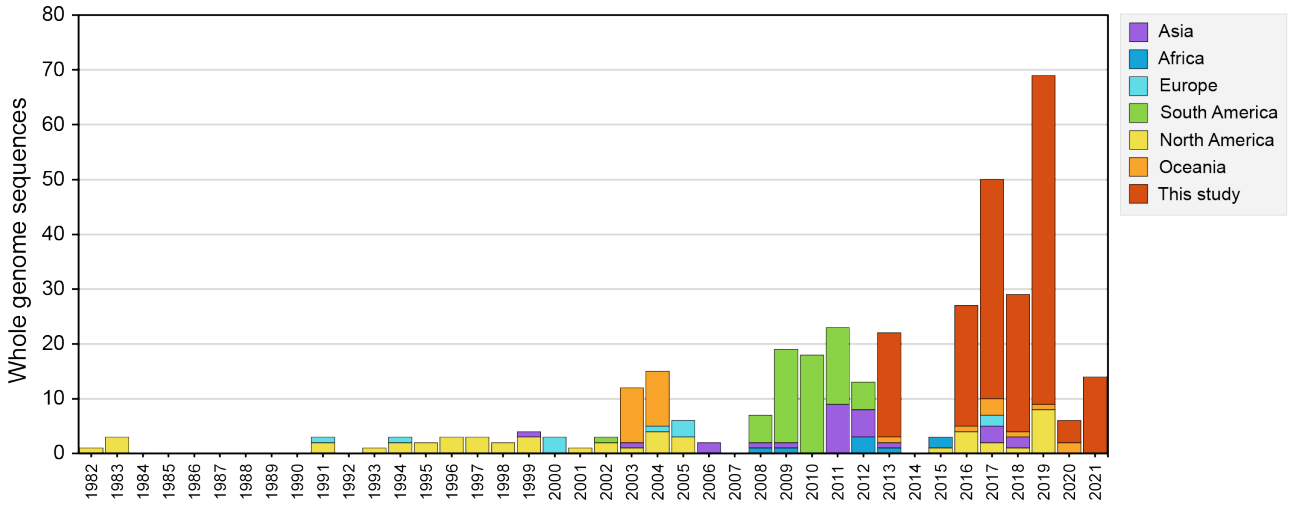
[1]Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia
[2]Sydney Institute for Infectious Diseases, Faculty of Health and Medicine, The University of Sydney, Westmead, NSW 2145, Australia
[3]Kirby Institute, The University of New South Wales, Kensington, NSW 2052, Australia
[4]NSW Health Pathology - Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia
[5]The Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
[6]The Kinghorn Centre for Clinical Genomics, Sydney, NSW 2010, Australia
[7]St Vincent's Clinical School, Department of Medicine, UNSW Sydney, Kensington, NSW 2052, Australia

Correspondence

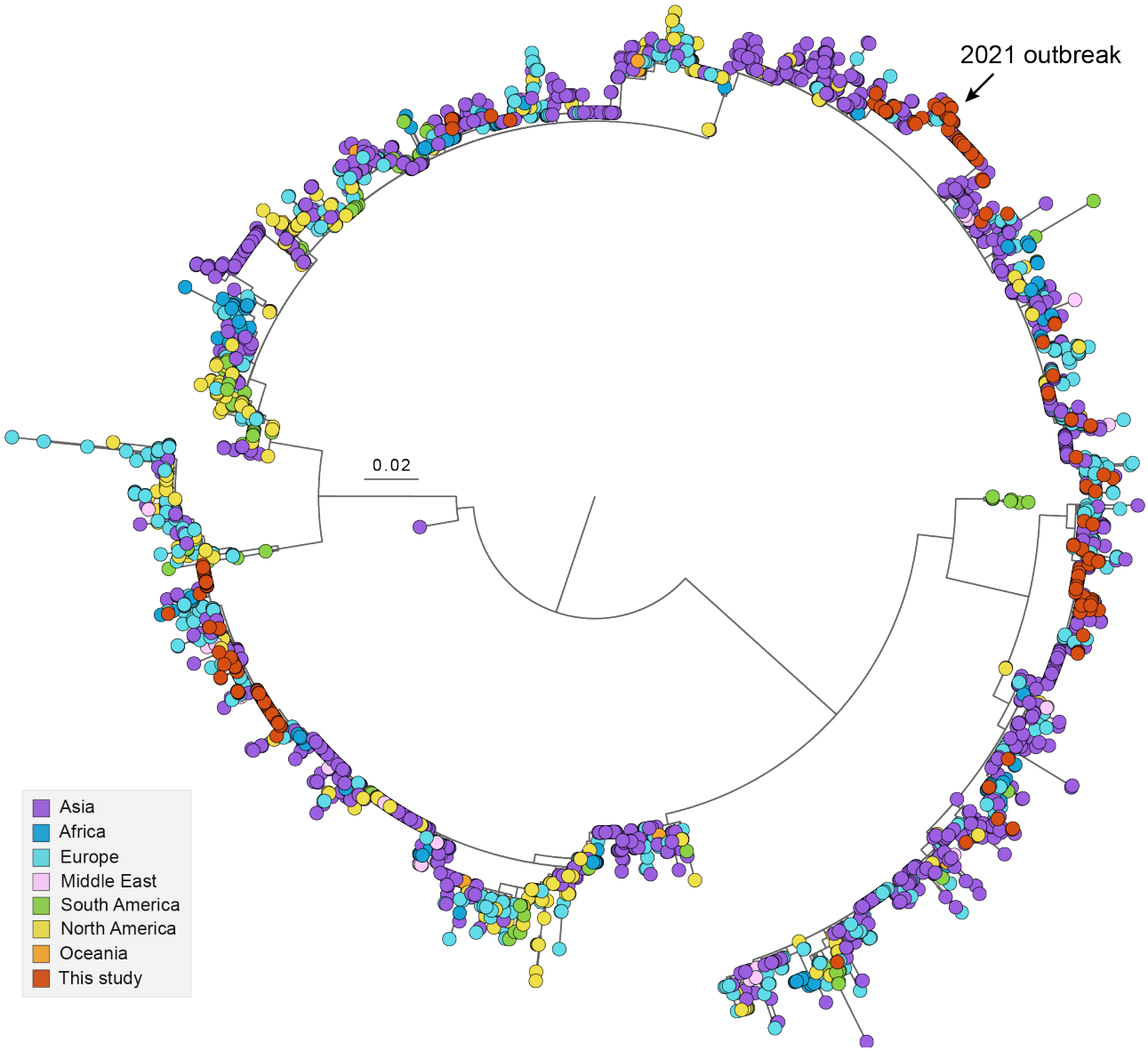
Dr John-Sebastian Eden

Email: js.eden@sydney.edu.au

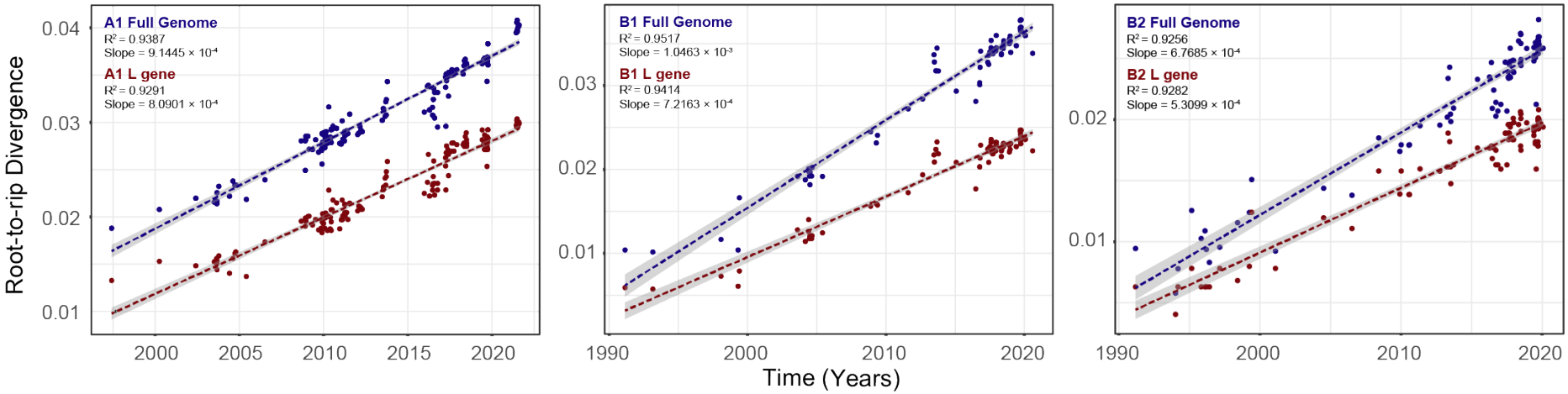
Phone: +61 2 8627 1817



Supplementary Figure 1: Geographic and temporal sampling distributions for genomic analysis. These plots show the number of sequences (y-axis) by collection year (x-axis) for the genome analysis presented in Figures 3 & 4, respectively. Sampling location has been coloured according to the key provided.



Supplementary Figure 2: Phylogenetic analysis of global and local HMPV fusion gene sequences. HMPV sequences in this study were aligned with all available HMPV-A2 sequences from NCBI GenBank. The fusion glycoprotein coding region was extracted, and sequences less than 300 nt were removed. No sequences sourced globally were found to be immediately related to 2021 outbreak clade suggesting the source remain unknown. Country of collection has been coloured by broader region according to the key provided. Branches are proportional to the number of nucleotide substitutions per site.



Supplementary Figure 3: Root-to-tip regression analysis of temporal structure. Plots of the root-to-tip genetic distance against sampling time (year), shown for phylogenies estimated from HMPV subtype alignments, A2, B1, B2, containing sequences generated in this study and publicly available genomes. Individual data points are coloured blue or red and are representative of individual sequences, the differing colours corresponds with either a full or partial genome analysis. The regression lines are similarly coloured with outer shading representing the 95% confidence intervals. The R^2 value and slope is displayed alongside the graph for each.

Supplementary Table S1: HMPV A2, B1 & B2 coalescent estimates

Subtype	Model	Rate of Evolution (subs/site/year)		Age (years)		Bayes Factor		
		Mean	***95%HPD***	***Mean***	***95%HPD***	***PS-BF***	***SS-BF***	***Rank***
A2	Root-to-tip Regression	8.09E-04	0.9639*	1985.3377	24.1973*	NA	NA	NA
	Strict Constant	6.54E-04	[5.904E-4, 7.1812E-4]	1978.7536	[1974.857, 1982.5175]	-29.27	-29.27	5
	Strict Skyride	6.63E-04	[5.9506E-4, 7.2749E-4]	1982.4869	[1979.3858, 1985.5221]	-43.06	-43.06	6
	Relaxed Constant	6.84E-04	[5.9489E-4, 7.7268E-4]	1977.7210	[1967.9026, 1986.1606]	-23.89	-23.89	4
	Relaxed Skyride	7.35E-04	[6.3718E-4, 8.3331E-4]	1990.1583	[1986.4465, 1993.6095]	-22.97	-22.97	3
	Relaxed Exponential	6.84E-04	[5.9423E-4, 7.7235E-4]	1977.6976	[1967.759, 1986.3028]	-19.63	-19.63	2
	Relaxed Skyline	6.68E-04	[5.8894E-4, 7.4945E-4]	1978.9201	[1970.5456, 1986.1631]	0	0	1
B1	Root-to-tip Regression	7.22E-04	0.9702*	1986.7627	29.4395*	NA	NA	NA
	Strict Constant	5.09E-04	[3.6796E-4, 6.7029E-4]	1979.9614	[1975.9692, 1983.894]	-22.29	-22.29	5
	Strict Skyride	5.06E-04	[3.5283E-4, 6.5833E-4]	1981.8723	[1978.3796, 1985.2835]	-23.59	-23.59	6
	Relaxed Constant	6.01E-04	[3.7487E-4, 9.0299E-4]	1979.5001	[1967.53, 1989.1406]	-2.68	-2.68	3
	Relaxed Skyride	6.34E-04	[3.8278E-4, 9.2772E-4]	1989.7916	[1985.747, 1991.1534]	-1.89	-1.89	2
	Relaxed Exponential	6.18E-04	[3.9442E-4, 9.1978E-4]	1982.8620	[1974.7855, 1991.1502]	-3.42	-3.42	4
	Relaxed Skyline	6.46E-04	[4.0614E-4, 9.4579E-4]	1985.4667	[1976.288, 1991.1534]	0	0	1
B2	Root-to-tip Regression	5.31E-04	0.9634*	1982.8486	28.8848*	NA	NA	NA
	Strict Constant	3.12E-04	[2.53E-4, 3.6946E-4]	1970.4735	[1963.4729, 1976.7943]	-38.22	-38.22	5
	Strict Skyride	3.03E-04	[2.509E-4, 3.5658E-4]	1973.6145	[1968.0398, 1978.8311]	-40.04	-40.04	6
	Relaxed Constant	3.71E-04	[2.6825E-4, 4.7583E-4]	1961.4532	[1935.1523, 1982.0427]	-4.34	-4.34	3
	Relaxed Skyride	3.57E-04	[2.7015E-4, 4.4991E-4]	1983.8530	[1978.2964, 1988.5495]	-8.64	-8.64	4
	Relaxed Exponential	3.97E-04	[3.0321E-4, 4.9749E-4]	1975.9732	[1963.9202, 1986.2543]	-0.31	-0.31	2
	Relaxed Skyline	3.98E-04	[2.7214E-4, 5.2087E-4]	1971.9243	[1954.0335, 1985.691]	0	0	1

*Coefficient of Variation (CV) generated during in Tempest (v1.5.3) using heuristic residual mean squared model fitted to data from the full-length genome HMPV PHYML trees.

CHAPTER 4

A Brief Introduction to Metagenomics

Chapter 4: A Brief Introduction to Metagenomics

For over 300 years the study of microorganisms has been based on growth in specific cultures, colony or cell morphology, and selection of biochemical profiles. While these methods enabled a detailed insight to the microbial world, they have now been augmented by methods superior in specificity and resolution (Oliver et al., 1991, Roszak et al., 1984). Attempts to define microbial communities have been reported in scientific literature as early 1986 where Begon et al. defined microbial communities as a set of microorganisms that coexist in the same time and space (Begon et al., 1986). The progressive and rapid development of culture-independent microbial analysis methods over the last 20 years facilitated by advances in technology now allows researchers to sequence unique microbial communities directly from environmental samples, a process known as metagenomics or sometimes referred to as community genomics (Chen et al., 2017). The term metagenomics was first introduced in 1998, as a way to describe the analysis of all collective genomes in a sample (Handelsman et al., 1998). Here we define metagenomics as shotgun, unbiased sequencing of the total genomic DNA of a sample. Metatranscriptomics involves the sequencing of the complete meta-transcriptome of the microbial community; it has the capacity to inform us of the genes that are expressed by the community as a whole, with the advantage of capturing double stranded organisms during their replication phase and RNA based pathogens (RNA viruses) and so is a valuable tool in capturing the entire microbiology community (Bashiardes et al., 2016).

This technology has a wide array of applications in both environmental and biomedical fields of science. The main research areas currently employing metagenomics as a research tool include studies of microbial diversity, population structure and composition, genetic and evolutionary relationships, functional activity, cooperative relationship, and environmental relationships. Whilst it exists as a useful tool in capturing the so-called microbiome of a sample unbiasedly, it also plays a part in pathogen identification and discovery where infectious agents are unknown (Brenner et al., 2018). Metagenomic sequencing has the capacity to sequence and identify nucleic acids from multiple different taxa when performed in a conservative unbiased manner. Metagenomics for the purpose of clinical applications has roots in the use of microarrays in the early 2000's where early success using this technology included the identification of SARS coronavirus, gene profiling of cancer mutations and the in-depth microbiome analysis of complex microbial populations within the human body (Sotiriou and Pusztai, 2009, Rota et al., 2003, Palmer et al., 2006). Gene expression microarrays have been commercially available for more than 10 years and have provided researchers with biologically meaningful snapshots during this time. Microarrays utilise probes to simultaneously

analyze the gene expression of thousands of genes. Whilst microarrays remain effective in identifying expression in known genes, they cannot detect previously unidentified genes or transcripts. The invention of next-generation sequencing (NGS) technology in 2005, has circumvented some of the limitations of microarrays and accelerated interest in the metagenomics field (Valencia et al., 2013, Bashiardes et al., 2016, Greninger, 2018). High throughput sequencing enabled millions to billions of reads to be generated in a single run, and so the analysis of the entire available genetic material of an environmental or clinical sample (Brenner et al., 2018, Fernandez-Marmiesse et al., 2018). With the increasing affordability of this technology and the advent of benchtop machines and all in one sequencing kits, NGS technology has been rapidly adopted for a variety of applications (Nimwegen et al., 2016). Specifically, metagenomics is a rapidly growing in popularity as a tool with varied use in clinical medicine where it has proved useful in the diagnosis of infectious diseases, outbreak tracking and tracing, infection control surveillance and disease mutation and pathogen discovery (Eden et al., 2021, Di Giallonardo et al., 2018a, Rockett et al., 2020, Baker et al., 2021, Yang et al., 2019).

Human Health Current Diagnostic Landscape

The term clinical microbiology encompasses both diagnostic microbiology and public health microbiology, that is it concerned with pathogen identification from clinical samples to guide management and treatment strategies for individual patients with infections, whilst also being concerned with the monitoring and surveillance of infectious disease outbreaks in the community. Direct pathogen detection methods, such as culture based, non-culture based, and indirect pathogen detection are methods commonly found in diagnostic laboratories. These methods include the selection, growth, and isolation of disease-causing microorganisms in culture, the detection of pathogen-specific antibodies also known as serology testing, antigen testing, molecular diagnostic and identification of microbial nucleic acids, DNA and RNA, most commonly using PCR. Typically, infectious diseases in patients are diagnosed by a physician selecting a suite of aforementioned tests in an attempt to identify the causative agent. Over the last two decades we have seen the transition of diagnostic pathology and laboratory medicine from traditional culture-based to molecular diagnostic methods. This phenomenon was largely attributed to the fact that much of molecular biology was first studied and defined in microorganisms. The number of commercially available microbiology test platforms, assays and test menus have increased exponentially over the last 10 years. The transition from culture-based to molecular-based diagnostic testing is potentially the biggest change in human disease diagnosis in many years.

Novel virus identification, investigation and limitations

Undoubtedly the discovery and identification of novel viruses is of great importance to human health and aid in containing infectious disease outbreaks. In recent years there has been a bloom in the number of virus discovery methods which are efficient in viral discovery and identification of its clinical association and relevance. The identification of an outbreak or patient with distinctive clinical disease features in the absence of a diagnosis using molecular diagnostic techniques is often the catalyst for novel disease investigation. Whilst possible techniques and technologies used to identify novel viruses are evaluated later in this chapter, here the limitations of causative agent discovery are here acknowledged. Virus discovery remains the first of many steps in establishing a clinical association between the presence of the virus and its likelihood as a causative agent. Where applicable, causation can be definitively established by measuring the fulfilment of traditional Koch postulates in animal models, this is not always readily available for viral agents. Robert Koch, in consultation with Friedrich Loeffler, devised strict criteria, to evaluate the correlation between human diseases and specific disease-causing microorganisms (Tammen et al., 2016, Linstow et al., 2008). These criteria helped to define the fundamental principles which established association between infectious disease- and disease-causing microorganism. This criteria requires proof of causality to be demonstrated, by the putative disease causing organism to be present in every case of the suspected disease and absent in other diseases, it also requires the organism to be isolated and cultured, after which it must be sufficient in reproducing disease by inoculation of a susceptible host, such as in a mouse (Linstow et al., 2008). Since the creation of the postulates at the end of the 19th century, modifications to them have been proposed repeatedly as our knowledge of infectious agents of disease has evolved, for example, viruses had not yet been discovered as agents of disease at the time of the development of the postulates. Further, limitations of Koch's postulates are widely acknowledged, including many agents of disease, such as viruses, cannot be grown in culture and that disease is not always caused by a single agent (Pendleton et al., 2017, Falkow, 2004).

Metagenomic sequencing of environmental and host-based samples has allowed for the discovery of several novel and potentially disease causing viruses (Shaffer et al., 2022). Using sequences generated using metagenomics is insufficient in assessing the potential pathogenicity of the virus, it may not provide sufficient evidence to even identify a possible host (Yang et al., 2019). The use of epigenetics and phylogenetic analysis allow scientist to draw meaningful connections between known and often characterised viruses which may be related, enabling a possibly method of host identification. An important consideration when utilising metagenomics as a tool to identify disease causing viruses include the identification of a virus in a sample taken at an acute stage of disease does not necessarily correlate with the disease-causing agent responsible for the pathology. This phenomenon is commonly observed where disease is caused by a virus which is shed from a host for a prolonged period, enteroviruses are good examples of long shedding viruses which may be detected long after symptoms of the disease have been resolved (Chen et al., 2021). Important considerations moving forward suggest the association between virus and disease in both humans and animals alike, will continue to depend on the results and findings of epidemiological, clinical and laboratory investigations. The results of these investigations should be considered with respect to Koch's postulates and the adapted versions to ensure the presence of viral sequences are not being incorrectly attributed to disease.

Point of care testing

Point of care testing (POCT) is rapidly emerging as an alternative to traditional laboratory-based diagnostics. Also known as near-patient testing, POCT have grown in recognition recently due to the roll-out of self-testing during the COVID-19 pandemic. POCT can be defined as diagnostic tests that are "conducted outside the laboratory, at or near the site of patient care, including the patient's bedside, the doctor's office and the patients home" (Khurshid et al., 2016). Several commercially available diagnostic tests and platforms are available, these platforms substantially reduce the turnaround time for the diagnosis of a host of microbial infections. Commonly used diagnostic tests currently used in diagnostic testing in Australia include the Cephied Xpert MRSA/SA SSTI[62] for MSSA and MRSA and the Biofire Film array Respiratory panel, capable of testing for 20 respiratory targets including *Bordertella pertussis, Mycoplasma pneumoniae* and *Chlamydophila pneumonia,* both commercially available tests have a turnaround time of less than an hour (Berry et al., 2022, Popowitch et al., 2013).Typically, POCT is conducted using automated or semi-automated systems or preprepared kits capable of integrating sample preparation, pathogen detection and identification. The purpose of POCT is to generate meaningful results in a timely manner with little to no user effort, including that involved in sample preparation or result interpretation. Depending on the type of test,

single or multiple pathogens (multiplexed) may be detected; recent advancements in tests have also enabled the detection of antimicrobial resistance genes at the patient's bedside. Whilst POC testing has its place inside the hospital, it also exists as an ideal way to test for infectious disease at home or in remote locations, such as isolated communities. POCT may have multiple other uses in the medical sphere including blood count testing, haemoglobin estimation, white blood cell counts, renal function, electrolyte function etc. For the purpose of this review, we will focus on POCT with respect to infectious diseases.

There are several considerations necessary to ensure a POCT device or test system is effective and fit for purpose, some of these requirements set out in a clinical trial by Bouzid et al. include the following suggestions (Bouzid et al., 2021):

- Be lightweight <10kg, portable and operate independent of a power source.
- Require only a few microliters of sample (blood or fluid) to be able to perform the test.
- A quick turnaround time, ideally less than 10 minutes.
- Use reagents that do not rely on cold store, reagents that remain stable at room temperature for extended periods of time are ideal.
- Deliver reproduceable results of adequate reliability and accuracy, preferably with comparable results to accredited pathology laboratories.

The COVID-19 pandemic has brought about changes in the way POCT were used globally. In Australia we saw a large upturn in at home rapid antigen testing using upper respiratory samples. The use of COVID-19 antigen tests grew exponentially in demand throughout the course of the pandemic. Immunochromatographic assays, like those used in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) rapid antigen test, were the most frequently used SARS-CoV2 POC tests globally and relied on saliva or nasal secretions as a testing sample. Rapid antigen testing involves the specific targeting of a specific region of the virus, usually the highly conserved N protein (Jacobs et al., 2020). These methods typically rely on a mobile phase, where a buffer mixed with the biological test specimen migrates through a stationary phase, usually nitrocellulose, tagged with antibodies against the viral antigen. In the case where the viral antigens are present, the antibodies attach to the virus and the accumulation can be detected by a visible chemical reaction which appears on the nitrocellulose membrane.

Studies that evaluated the sensitivity of typical SARS-CoV-2 assays using saliva as a test material had a concordance rate of up to 95% with conventional viral culture results (Uwamino et al., 2021). In a similar study, a satisfactory sensitivity result was reported for high viral load samples

using the Espline SARS-CoV-2 assay, where Ct values were below 25 for nasopharyngeal swabs (NPS) RT-qPCR detection was possible using the POCT cassette(Basso et al., 2021). Other notable immunochromatography-based POC COVID-19 tests include the COVID-VIRO assay, the Panbio assay and the hfx SARS-CoV-2 test kit. Whilst RT-PCR testing remains the most well-documented and widely accepted diagnostic test method of SARS-CoV-2, rapid antigen testing provided an at home convenience, requiring little or no training for the layman to use. Overall, the sampling method is like what is required for PCR-based methods, limitations include sensitivity dependent on viral load, however, is advantageous when compared to PCR as it requires no specific equipment or trained personnel. The turnaround of rapid antigen testing ranges from 5-30 minutes, cost per test is relatively low when compared to alternative methods including RT-PCR and viral culture, however the overall lower sensitivity of rapid antigen tests suggest risks a higher proportion of false negative tests.

Culture-based microbiology

Culture-based diagnostic methods have proven to be useful in providing phenotypic and antimicrobial resistance information for a multitude of infectious diseases (Wilson, 2015). Whilst long considered the gold standard method for pathogen detection there exists inherent limitations of the method as an adequate diagnostic tool for respiratory infections. Since the advent of molecular methods such as PCR and whole genome sequencing, culture-based methods are considered much less sensitive. Where mixed infections occur, it is possible for fast growing organisms to inhibit or outcompete slower growing organisms leading to incorrect identification of disease-causing pathogens (Hasan et al., 2020). The use of selective culture-based techniques can also exclude the growth and identification of unknown organisms. Where bacterial pathogens are causative, culture-based techniques remain a mainstay of pathogen diagnostics. Culture-based techniques are advantageous where pathogens are bacterial in origin due to the need of live bacteria for antimicrobial susceptibility testing (AST) (Wilson, 2015). Culture can also be useful in the detection of specific antimicrobial resistance genes (most commonly mecA and VanA), however this use of culture-based diagnostics is losing its primacy even in bacterial diseases with the advent of NGS.

Antigen Testing

There are two widely accepted antigen based diagnostic techniques, immunofluorescence (IF) and immunoassays. Immunofluorescence assays usually yields actionable results take between 2-4 hours and the technique is based on the microscopic detection of viral proteins on host cells using labelled antibodies. Typically, immunofluorescence is labour-intensive and does require specialised

microscope skill sets. Direct immunofluorescence is a method where the observed signal is on the antibody binding directly to the viral proteins, indirect immunofluorescence involves the targeted antibody that binds to a viral protein that produces an immunofluorescent signal. Immunoassays are most often commercially available, they are alike traditional immunofluorescent techniques in that a viral antigen is also targeted, but significantly differs in that it involves the macroscopic detection of signal-labelled antibodies bound to target proteins.

Rapid antigen-detection tests (RADTs) are a well-known form of commercially available immunoassays and are commonly used in the diagnosis of several molecular pathogens (Van der Palen et al., 2009). RADTs are available for an array of viruses, and from a variety of manufacturers in ready to use commercial kits. Typically, kits require an easy to obtain sample from the nasopharyngeal cavity and a qualitative positive or negative result is obtained within 10-30 minutes. Most RADTs have a “control” component to ensure test's reagents are functional and the test was performed properly. These kits can be operated with very little instruction and no specific virology expertise is required. RADTs are available in three distinct formats: immunochromatographic (ICR) tests, enzyme immunoassays (EIA) and optical immunoassays (OIA). Despite the variations the principle of these tests remains unchanged, if the pathogen is present a macroscopic colour change will occur. The colour change is the result of the binding of viral proteins from the sample to the target antibodies. Between and within the RADT types the apparatus, binding reaction sequence, and macroscopic signalling methods can differ.

Examples of an ICR test include a pharmacy-bought pregnancy test, these tests consist of a strip of nitrocellulose paper with a sample port at one end and a test result and control line at the other. The patient sample is added to the port, here antibodies targeting the virus fusion or in the case of a pregnancy test, the human chorionic gonadotrophin (hCG) tracer antibody, bind the antigen in the sample. Utilising capillary action the sample is migrated along the strip where it binds to a second set of specific antibodies, the second binding region sees a coalescence of antigen-antibody complexes, produces the appearance of a coloured line, and indicates the test results. EIAs differ from ICRs as the enzyme produces the colour change over the entire test membrane. EIAs can also be further broken down into two categories based on the separation criteria of immunocomplexes, homogeneous immunoassays and heterogenous immunoassays. Homogeneous assays do not require the separation of the immune complex from the reaction mixture, the change in enzyme activity is related to the concentration of the analyte. These assays are most used in the measurement of small analytes like drugs, due to the absence of a separation step this method is considered simple and faster

than its heterogenous counterpart. In heterogenous EIAs the separation of immune complexes is a pre-requisite for analyte estimation, this method is known as an enzyme-linked immunosorbent assay (ELISA). Here an antibodies or antigens are bound either noncovalently or covalently to a solid matrix, such as a microtiter plate, polystyrene tubes or beads, magnetic beads, or nitrocellulose membranes. Following binding the unreacted antigen or antibody is removed and the bound count is obtained. OIA's are the final type of test, they contain elements of both the ICR tests and EIAs however rely on silicon wafer as the test media. Antibodies are fixed to the surface and the enzyme acting on the substrate gives a colour change which denotes the test result. Whilst the performance profiles of the different types of RADTs are relatively similar, individual products are often chosen based on economical and logistical considerations including cost, ease of use and availability.

The emergence of SARS-CoV-2 in 2019 brought about significant challenges in disease detection for several reasons, including skilled labour availability, movement restrictions and cost of testing. At the height of the pandemic there was a large disparity between the number of tests needed in primary-care settings and the testing capacity of laboratories at the time. Antigen-based testing was heavily relied upon throughout the duration of the COVID-19 pandemic globally, RADT testing remains widely used today as a rapid diagnostic tool (Bigouette et al., 2021). Advantages of rapid antigen testing during the pandemic include but are not limited to:

- Tests are simple to perform at the point of care or at home.
- Tests require no medical knowledge or training to perform or interpret the results.
- Tests do not require specific equipment to perform.
- Tests are less expensive than conventional methods (RT-PCR or culture).
- Rapid turn-around of results (10-30 minutes).

Reported performance results of early RADTs used in COVID-19 diagnostics were extremely poor and acted as a deterrent against their use in the early days of the pandemic (Lambert-Niclot et al., 2020, Blairon et al., 2020). As with all rapid antigen methods sensitivity is a limitation of the assay. In two separate evaluations of the leading commercially available COVID-19 rapid antigen-detection tests, the Panbio and the Quidel Sofia 2, overall sensitivity of the assays were 79.3% and 76.3% respectively with a study size of 412 patients (54 positive RT-PCR) (Albert et al., 2021) and 255 patients (60 positive RT-PCR)(Young et al., 2020). International World Health Organisation (WHO) guidelines require that any commercially available SARS-CoV-2 RADT must demonstrate $\geq 80\%$ sensitivity and $\geq 97\%$ specificity compared to the reference RT-PCR assay. In a retrospective study using the Panbio test, clinical performance was analysed on frozen swabs from 1406 individuals

with correlating RT-PCR results available, of these samples 951 (67.6%) were positive and 487 (32.4%) were negative. Interestingly the sensitivity of the test was much higher in samples collected in the case of contact tracing (94.2%) when compared to asymptomatic screening (79.5%) (Alemany et al., 2021). This confounds the advice from WHO, which advises against the use of RADTs in screening asymptomatic populations due to the increased propensity of false positives (Basso et al., 2021, Uwamino et al., 2021). Overall, the performance of RADTs is dependent on the epidemiological situation, and the populations being tested. In a recent study 541 SARS-CoV-2 positive upper respiratory samples were tested using the TaqPath COVID-19 test (Lin et al., 2021)RT-PCR and then analysed with the "Panbio COVID-19 AG" (Albert et al., 2021) RADT test. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the Panbio RADT were calculated using RT-PCR as a reference. The sensitivity of the RADT was reported as 83.3% (65.3-94.4%) and the specificity was 99.8% (98.9-100%). The PPV was reported as 96.2% (80.4-99.9%) and the NPV was 99% (97.7-99.7%) (Aranaz-Andrés et al., 2022). The above values suggest in this scenario the RADT was relatively well performed within this context. Whilst the RADT shows good diagnostic performance in patients in this study, it is important to note the use of RADT tests are not recommended for asymptomatic patients (Aranaz-Andrés et al., 2022).

Polymerase Chain Reaction

Molecular diagnostic assays undoubtedly provide rapid and potentially cost effective way of diagnosing common infections. Nucleic acid testing using PCR can be employed for single target testing or can be multiplexed usefully into syndromic panels (Zumla et al., 2014, Cho et al., 2018). Syndromic panels generally include the most commonly found pathogens associated with the clinical syndrome, this can vary from region to region. Typical test panels include those for meningitis, encephalitis, acute respiratory infection, sepsis or diarrhoeal disease (Cho et al., 2018, Sharma et al., 2022, Liu et al., 2015). Limitations of conventional PCR diagnostic testing are centred around the specificity of the targets, that PCR cannot detect the presence of aetiological agents not included in the panel, that mutation at the sequence targets for PCR primers can cause improper probe binding and lead to false negative results. Much like all diagnostic tests, an accepted risk of PCR is some false positive and false negative results. False positive results in a diagnostic laboratory setting can be caused by non-specific amplification by environmental contaminants, contamination with previously amplified material and cross contamination of true positive samples. In a diagnostic laboratory setting an array of safeguards are put in place to avoid false positive results, including the

use of certified standards, external validation of commercialised tests and routine auditing (Li and Yan, 2021, Yuan et al., 2022).

Targeted enrichment

There have been significant advancements in the characterization of genetic sequencing data made available due to NGS technologies. A key challenge surrounding NGS technology lies within the inability to detect low-copy-number pathogens in complex biological mixtures without enrichment. A combination of small or very large genome size, low relative copy number, and heterogeneous genomic content can pose significant challenges of NGS diagnostic capabilities with respect to pathogen detection from complex samples. To circumvent this issue, virus isolation or culture methods are sometimes employed for viral, bacterial, or fungal agents of disease prior to NGS to increase the proportion of on target reads generated from clinical samples. To increase the proportion of NGS reads in a library dedicated to low abundance targets, an enrichment method, called targeted genome capture (TGC) can be used. TGC involves the use of oligonucleotide probes which enrich for specific nucleic acids in a heterogeneous extract such as a nasopharyngeal sample containing mixed upper respiratory taxa. Genomic regions of interest are captured using hybridization, that target-specific biotinylated probes, which are later isolated utilising magnetic separation techniques (Carpenter et al., 2013).Targeted sequencing affords the ability to target a higher amount of total gene content, allowing for the characterisation of both known and novel variants, enabling its use in variant discovery applications. High-throughput targeted capture was seen as a feasible alternative to whole genome sequencing (WGS) and multiplex PCR to obtain large datasets of orthologous loci in an attempt to address evolutionary questions in humans (Jones and Good, 2016). Target capture approaches are limited as they rely on prior sequence knowledge, although recent advancements have emerged where a reference genome for focal species is not necessary. To circumvent the use of a traditional reference genome, three common methods used include PCR capture, de novo assembly, and divergent reference (Dugat et al., 2014, Jones and Good, 2016).

Metagenomics in Medicine

Just as there are many varied uses for metagenomics there are a variety of methods used in practice. At present, the most widespread metagenomics method utilised is large scale sequencing of un-enriched samples. Through recent advancements in ultra-high throughput sequencing technologies that do not require prior cloning or PCR amplification, affordability of sequencing reads

has allowed the use and scope of metagenomics to increase exponentially. The analysis of the libraries themselves can also be undertaken in a variety of ways, with some analyses aimed at determining the biological diversity of samples and comparing this diversity against other samples of similar type, and others aimed at identification of novel microbes.

Clinical applications

Clinical metagenomic NGS (mNGS), involves the comprehensive analysis of microbial and host material (DNA and/or RNA) in patient samples. This technology is rapidly moving from the research sphere into clinical laboratories. Several studies have explored the usefulness of mNGS in clinical and public health settings, the most notable are those where mNGS findings contribute meaningfully by providing clinically actional information, such as a successful diagnosis which informs treatment plans. One of the earliest of such cases was reported by Wilson et al. where mNGS was used for the clinical diagnosis of neuroleptospirosis in a 14-year-old patient with meningitis. The correct diagnosis of the antilogical agent was attributed to mNGS, the treatment plan was tailored to the infectious agent and a positive outcome was achieved (Wilson et al., 2014). Until recently, the uptake of mNGS in clinical laboratories as a routine diagnostic tool had been slow despite the potential clinical utility of these methods.

The application of mNGS is varied, it exists as a useful tool for public health microbiology teams when used in conjunction with transmission network analysis. Notable examples of the use of this technique in public health settings include but are not limited to: epidemiological investigation of the *Neisseria meningitidis* serogroup C outbreak outside the meningitis belt in 2017, vancomycin-resistant enterococci (VRE) transmission in hospital settings and SARS Coronavirus-2 contact tracing in the 2019 pandemic (Neumann et al., 2020, Rockett et al., 2020, Bozio et al., 2018). Steadily, the ‘big data’ provided by NGS is being used, often retrospectively, for an array of purposes, including characterisation of antibiotic resistance and analysis of human host response data (transcriptomics) to predict causes of infection (Li and Yan, 2021, Moss et al., 2021).

Metagenomics and Metatranscriptomic Sequencing Strategies

Metagenomic sequencing can be broadly classified into two sequencing strategies: amplicon sequencing and shotgun sequencing. Amplicon, or polymerase chain reaction (PCR) is one of the fundamental methods currently used for taxonomic identification. PCR-based metagenomics is routine in gene prospecting by the amplification of specific genes while colony PCR is often used to

screen metagenomic libraries. PCR amplification of certain genes can be used to evaluate the species diversity utilising sequence composition. It is common practice to utilise 16S ribosomal RNA (rRNA) genes as a phylogenetic marker, this method allows for the imputation of functional content based on taxon abundance (Brumfield et al., 2020). The 16S rRNA gene sequencing method utilises universal PCR primers to amplify hypervariable regions of the 16S rRNA gene which enables taxonomic identification when libraries are aligned using bioinformatics against various rRNA sequence databases (Shaffer et al., 2022, Brumfield et al., 2020, Ahn et al., 2011). Common databases used as references for bioinformatic alignment include the Greengenes, the ribosomal database project (RDP) and SILVA Ribosomal RNA Gene Database Project (Bartram et al., 2011, McDonald et al., 2012, Cole et al., 2009). Database composition is variable between the three main databases, some prokaryotic taxonomies exist only in one of the three databases; furthermore, the SILVA and Greengenes databases do not have resolution to the species level (Cole et al., 2009, McDonald et al., 2012). Typically, 16S rRNA gene amplicon sequencing is restricted to taxonomic classification at genus level, limiting the amount of functional information. Well-recognized limitations of 16S rRNA gene sequencing include concerns about the reproducibility and accuracy of amplicon-based microbial community analysis (Barb et al., 2016).

Shotgun metagenomic sequencing offers a unique opportunity to classify all organism present in uncultured microbial communities of complex samples (Baker et al., 2021). Total DNA and RNA is extracted and sheared into fragments, which are then independently sequenced and aligned, taxonomic identification is facilitated by aligning reads to genomic databases including GenBank(Sayers et al., 2019), Reference Sequence (RefSeq) (O'Leary et al., 2016) or Pathosystems Resource Integration Center (PATRIC) (Wattam et al., 2014) databases. Shotgun metagenomics is also known as unbiased metagenomics and allows classification of bacteria to strain and species-level (Barb et al., 2016). Potential restrictions to shotgun sequencing uptake as a tool for large scale microbiome analysis include cost and demanding bioinformatics resources to facilitate manipulation of large datasets (Neumann et al., 2020).

Sequencing platforms

Two leading sequencing platforms exist for metagenome sequencing, short read and long read sequencing platforms. Short read sequencing platforms produce higher-throughput reads with approximate read length less than 500bp. Sequencing providers for short read sequencing platforms include Illumina and MGI Tech Co., Ltd. (MGI), within these two companies a number of platforms

exist, Illumina platforms include but are not limited to, the HiSeq, NovaSeq and iSeq platforms and variations, whilst MGI has two commonly used platforms, the BGISEQ-500 and DNBSEQ-T7 (Poulsen et al., 2022, Tulloch et al., 2021a, Kim et al., 2021). Significant differences between the platforms include read length, with most platforms providing reads between 90 -151bp, quality has also increased as sequencing platforms have evolved. The Illumina® platforms utilise single-stranded DNA-binding proteins, fluorescent-labelled deoxynucleoside triphosphates which are used to bridge amplified DNA template. Other short read sequencing technologies utilise clonal amplification, this is achieved using emulsion PCR (Kim et al., 2021). Emulsion or droplet PCR allows for the separation of DNA in physically separated hydrophobic water in oil droplets, which avoids unwanted interactions between DNA sequences. Clonal amplification platforms include Roche 454 pyrosequencing, Ion Torrent and SOLiD sequencing platforms (Singh et al., 2013, Frey et al., 2014, Luo et al., 2012). Short read sequencing platforms typically are unable to cover an entire gene of interest, as a result assembly is required before further analysis, this is an important consideration for the method as it requires complex bioinformatic analysis. A significant limitation of short read sequencing is the inability to sequence large stretches of DNA continuously given strands must be fragmented and amplified. Where sequencing depth is insufficient, short-read sequencing can fail to generate adequate read depth, causing difficulty in reconstructing highly complex or repetitive genomes.

As the name suggests, long read sequencing technologies can sequence longer lengths of DNA or RNA. This description is applied to platforms which can read between 5,000 and 30,000 base pairs. Long read sequencing platforms are commonly represented by two long read sequencing technologies, the single molecule Real-Time (SMRT) sequencer by Pacific Biosciences® (PacBio) and Oxford Nanopore Technologies (Nanopore). These technologies have grown in popularity over recent years and the length of sequencing reads has dramatically increased (Sanders et al., 2019). The PacBio system SMRT system (PacBio, Menlo Park, CA, USA) can produce an average read length of over 10kb in length (Das et al., 2019). The PacBio read protocol is utilises millions of microscopic wells with DNA polymerase at the base, known as waveguides, during sequencing two hairpin adaptors are ligated to the ends of the molecule as it passes through the wave guide, forming circular single-stranded DNA (Xiao et al., 2020, Schoonvaere et al., 2016). DNA polymerase is then used to sequence a complementary strand, and finally fluorescent nucleotides are added to the strand, with the signal is captured in real time (Das et al., 2019). The Nanopore platform can produce reads up to 1 million base pairs in length (Freed et al., 2020). Nanopore technology involves threading the DNA molecule through a bioengineered channel in a biological membrane, electrical current is pulsed

across the channel with signal interpreted to determine the base sequence (Jaworski et al., 2021, Tyson et al., 2020, Yang et al., 2019).

Long read sequencing circumvents amplification bias, as molecules are sequenced one at a time without shearing, a great advantage over short-read sequencing (Das et al., 2019). However long read technologies are notorious for high error rates, errors in nanopore sequencing are attributed to the inability of speed regulation of DNA translocation through the pore and poor sensitivity of read identification. SMRT base calling errors have been described as "random" and can be reduced using circular consensus sequencing (CSS). CSS is a method which allows the DNA to pass through the zero-mode waveguides multiple times. Error rates for SMRT are estimated between 1% and 12% whiles nanopore sequencing is approximately 5% (Oyola et al., 2021). In contrast, short read sequencing has a purported error rate of 0.23±0.06% (Pfeiffer et al., 2018). Some recent studies have explored the feasibility of combining long read sequencing platforms with short read sequencing platforms to generate more reliable, full length long read sequencing data (Sanders et al., 2019, Berbers et al., 2020).

Bioinformatics and analysis approaches

As metagenomics has grown in popularity, so too has the availability of bioinformatic tools, pipelines, and databases necessary for the interpretation of metagenomic datasets (Walsh et al., 2022). Tools can broadly be classified into 5 categories, short read taxonomic classifiers, short read functional classifiers, metagenomic assembly, metagenomic binning and genome analysis. The use of these tools is driven by several factors including the type of scientific question, sample type and reference database availability (Hasan et al., 2020). Typically, bioinformatic tools available for the interpretation of metagenomic data are tailored to short-read sequencing which is likely due to the widespread use of short-read sequencing platforms within the research community (Kim et al., 2021, Oyola et al., 2021, Berbers et al., 2020). Whole metagenome shot-gun sequencing can be very broadly assigned to two approaches, "read-based" and "assembly-based" approaches. These two approaches can be used separately or parallel. Read-based metagenomic and metatranscriptomic analysis differs to assembly-based analysis as read-based analysis relies solely on unassembled reads. Read-based analysis is useful in quantitative community profiling and relies on relevant reference sequences being available for mapping. This method is advantageous in allowing for rapid assignment of reads to relevant references and is less computationally demanding when compared to assembly-based workflows. Typically, bioinformatic assembly tools can be applied to metatranscriptomic data but must be done so with caution as many metagenomic tools have not been tested on metatranscriptome

datasets. Metatranscriptomic transcripts typically have nuances including introns and exons, not encountered in metagenomic sequencing (Shakya et al., 2019). Many similarities exist between metagenomic and metatranscriptomic bioinformatics and analysis.

Assembly Based

Quality control

Pre-processing is a necessary step for both metagenomic and metatranscriptomic analysis, erroneous reads are removed as part of a baseline (Yang et al., 2019) quality control (QC), there are dozens of publicly available QC tools including FastQC (Ward et al., 2020), fastp (Chen et al., 2018), FaQCs (Lo and Chain, 2014) and Trimmomatic (Bolger et al., 2014). Common pre-processing also includes removing target organisms from analysis, for example when looking for respiratory pathogens, host reads may be removed from sample libraries. Target host or human reads can be removed using traditional read mapping methods which tag and remove reads which map to the human genome, or alignment free methods which search for human specific k-mers in reads (Das et al., 2019). A significant difference between metagenomic and metatranscriptomic sequencing occurs during the pre-processing or quality control stage, this is the removal of highly abundant ribosomal RNA (rRNA) transcripts from the samples. Usually rRNA is depleted prior to library preparation and sequencing, despite this it is not uncommon for rRNA reads to constitute >90% of all sequence data (Walsh et al., 2022). In metatranscriptomic libraries, rRNAs are identified for removal in downstream analysis using tools like SortMeRNA, a rapid and efficient rRNA filter capable of sorting a large set of metatranscriptomic reads (Kopylova et al., 2012).

De Novo Assembly

High quality, pre-processed reads are assembled into putative transcripts using *de novo* assembly tools. *De novo* assembly is a method for constructing genomes from many DNA fragments, with no knowledge of the correct sequence or order of fragments. *De novo* assemblers provide a reference scaffold or set of joined-oriented contigs, that can provide a reference for a set of genes. This is advantageous where a microbiome is not adequately characterised using reference genomes. A key application of *de novo* assembly is to provide as a reference for mapping against during expression analysis but it can also be used to identify homologs and establish taxonomic origin. Common metagenomic assemblers include MEGAHIT (Li et al., 2016), IDBA-UD (Yu et al., 2012) and metaSPAdes (Nurk et al., 2017). Most metagenomic assemblers have been designed to assemble complex metagenomes which may contain population or strain-level variation, as well as highly

conserved regions with sequence similarity. Metatranscriptomic specific assemblers include Transcript Assembly Graph (TAG) (Ye and Tang, 2016), IDBA-MT (Leung et al., 2013) and IDBA-MTP (Leung et al., 2015). These assemblers consider the unique features of microbial communities and transcripts. The TAG and IDBA-MT assemblers both rely on multiple *k* values with respect to a de Bruijn graph (Ye and Tang, 2016, Leung et al., 2015, Leung et al., 2013). The IDBA-MT method accounts for features associated with mRNAs like uneven sequencing depth and repeat patterns (Leung et al., 2013). The TAG assembler assembles a metagenome using the de Bruijn graph, the metagenome is then used as a reference sequence to map the transcriptome reads and facilitate mRNA reconstruction (Ye and Tang, 2016). The IDBA-MTP assembler is based on the IDBA-MT assembler which utilises known protein sequences to guide assembly.

Taxonomic Classification

Taxonomic profiling is performed by a suite of tools which work by mapping short read sequences to subsections of metagenomic and metatranscriptomic data. For most metagenomic and metatranscriptomic samples, the species present in the sample are most likely unknown at the time of sequencing. This is especially true where the goal of sequencing is to determine the microbial composition of a sample. Many species have some detectable similarity to known species so sensitive alignment algorithms can detect similarity providing a basis for read classification. Where an organism is unlike anything else ever characterised, it is labelled "novel". The most widely accepted and well-known taxonomic labelling algorithm is the basic local alignment tool (BLAST). This tool classifies sequences by finding the best alignment to a large database of metagenomic sequences, this platform also provides a quality score for each alignment. A common short-read taxonomic tool is metagenomic phylogenetic analysis 3 (MetaPhlAn3) (Nousias and Montesanto, 2021), this taxonomic tool aligns read to an internal database of clade-specific and quasi marker genes through the ever updated nucleotide (BLASTn) (Altschul et al., 1990). The use of quasi markers is useful in expanding the analysis of viral and eukaryotic reads, markers act to validate positive and negative reads, circumventing the need to modify the initial metagenomic or transcriptomic dataset.

Kraken (Wood and Salzberg, 2014) is an ultrafast sequence classifier tool which utilises *k*-mer exact-match database queries, it differs from other taxonomic tools as it does not rely on partial sequence alignment. Sequence alignment and machine learning techniques are aimed at improving the accuracy of BLAST. The Naïve Bayes Classifiers (NBC) approach applies a Bayesian rule to *k*-mer distribution within a genome, this facilitates a higher accuracy than BLAST alone. It is apparent

that classifier tools can be loosely categorised into two groups, NBC based classifiers aim to classify all sequences as accurately as possible, whilst other classifiers like Kraken, BLAST and MetaPhlAn3 leave sequences unclassified if insufficient information is available to accurately classify them. Nonselective classifiers like NBC based tools tend to achieve higher sensitivity than selective classifiers, but this is a trade-off between precision, with non-selective methods significantly lower in precision. It is therefore important to consider which taxonomic profiling method or tool is used and take into consideration the type of algorithm used when interpreting sequence identification.

Nucleotide and non-redundant protein annotation

A fundamental task for metatranscriptomic and metagenomics is the annotation of nucleotide sequences (metagenomics) and non-redundant proteins (metatranscriptomics). Genome annotation is broadly the annotation of features on assembled DNA/RNA sequences (Panda and Slotkin, 2020). These features include genes and their location, encoding proteins or open reading frames (ORFs) and encoding ribosomal or transfer RNA (tRNA) molecules (Rota et al., 2003). At present there are few annotation tools available, typically these tools are either online or standalone platforms (Kim et al., 2021, O'Leary et al., 2016). The annotation of genomes can be challenging. Typically, metagenome and metatranscriptome libraries contain sequence data from organisms originating from all domains of life so they may contain understudied lineages with encoding genes with low similarity to reference genes.

Role of Metagenomics as a Diagnostic Tool

Historically, diagnostic molecular assays have been the mainstay of pathogen diagnosis in clinical settings. An accepted limitation of molecular diagnostic assays is that a limited number of pathogens can be targeted in any one assay using a combination of primers and probes. The advantage here of metagenomic/metatranscriptome sequencing is the ability to characterize all DNA or RNA present in any given sample. Resultant sequence data also allows for the analysis of the entire microbiome as well as human host genome, presenting a unique opportunity to characterise antimicrobial resistance genes or host integrating viruses. Metagenomic approaches have an extensive history in the identification of novel viral pathogens and the characterization of the human virome in both healthy and diseased states. Recent studies have investigated the diagnostic value of metagenomic sequencing have primarily been retrospective studies focusing on proof-of-concept experiments, measuring diagnostic specificity and sensitivity in comparison to clinically endorsed diagnostic methods including culture and quantitative PCR (qPCR).

A 2016 study by Cummings et al., analysed 15 sputum samples collected from cystic fibrosis patients and compared Illumina 16S rRNA NGS results to traditional culture methods (Cummings et al., 2016). This study demonstrated the ability of NGS to detect all pathogens grown using culture techniques as well as 3 additional disease-causing bacteria that culture did not detect (Cummings et al., 2016). A retrospective study of 41 respiratory specimens was performed in 2019, samples were collected from immunocompromised children between the years of 2014 and 2016 (Zinter et al., 2019). The study utilised Illumina® paired-end sequencing and the national centre for biotechnology information (NCBI) nucleotide reference database and identified pathogens in 11 of 24 culture negative specimens. Similar studies have also been performed using long-read sequencing platforms and comparing NGS to alternative diagnostic techniques including qPCR (Charalampous et al., 2019). Diagnostic applications of NGS have been explored in a variety of sample types including cerebrospinal fluid (CSF) from patients with suspected central nervous system (CNS) infections (Wilson et al., 2019).

The promise of metagenomic and metatranscriptomic sequencing in capturing DNA or RNA of novel or unexpected pathogens is widely acknowledged. Despite this, uptake of the method in clinical settings have been slow (Hogan et al., 2021). Recently a small number of reference laboratories, including the University of California San Francisco (UCSF) and the Centre for Clinical Infection and Diagnostics Research (CIDR), Kings' College London, have begun offering NGS diagnostics to infectious disease clinicians, as such the initiative is being adopted globally (Serpa et al., 2022). However, the use of NGS is not considered an accredited diagnostic method in Australia or the United States of America. To facilitate the uptake of metagenomic and metatranscriptomic assays in a clinical setting an attempt to standardise the sequencing approach is necessary. The key steps in developing a diagnostic assay include optimising the total nucleic acid extraction for a variety of sample types, increasing repeatability for sequencing quality, utilising negative controls to monitor sample contamination, standardising the bioinformatics pipeline, and establishing a biologically relevant threshold for reporting pathogenic organisms. Previous studies have reported a correlation between read counts and PCR cycles when comparing single target qPCR and NGS; however, NGS specificity has been shown to be considerably lower than qPCR, with one study reporting the specificity of NGS is up to 11% lower than conventional qPCR (van Boheemen et al., 2020). High-throughput genome sequencing without viral enrichment has demonstrated capacity in identifying RNA and DNA viruses, it also provides rich bacterial and fungal microbiome data which has varied research and diagnostic applications (Collinson et al., 2021, Zinter et al., 2019, Kan-o et al., 2017).

Metagenomic and metatranscriptomic NGS (mNGS) has accelerated in filling a niche in diagnostic medicine where other microbiological, serological, and molecular based diagnostic methods fail to yield a clinically relevant diagnosis. mNGS is most advantageous in the testing of small volume clinical samples and those from difficult to sample regions, for example CNS. There have been real world applications of mNGS in commercial laboratories with the usefulness of metagenomics evident in the SARS-CoV-2 pandemic. COVID-19 changed the landscape for clinical metagenomics/metatranscriptomics and highlighted the capacity of the tool to characterise emerging pathogens in a rapid manner. Identification of SARS-CoV-2 using mNGS facilitated the early phylodynamic analysis which were used to investigate viral origins and zoonotic reservoirs. Early knowledge of the viral genome allowed for the acceleration of vaccine development and targeted therapy (Quer et al., 2022). During the beginning of the pandemic mMGS allowed for the easy identification of coinfections during a time where the clinical symptoms of COVID-19 were not completely understood, this allowed for the identification of a suite of COVID-19 specific symptoms which were instrumental in disease identification and control of spread. Rapid viral mutation of the SAR-CoV-2 virus presented challenges for traditional molecular tests and therapeutics; unbiased whole-genome sequencing became a well-used tool for the identification of important mutations in real time.

Barriers to the uptake of routine clinical mNGS have also become evident during the pandemic, despite repeated demonstration of its feasibility and impact. A key obstacle to the prevention of global uptake of mNGS includes access to skilled personal. Bioinformaticians are a limited resource and necessary for the purpose of mNGS with most analysis pipelines involving a suite of analysis tools which require bioinformatics expertise with a sound understanding of command line skills. Movement has been made towards the automation of analysis pipelines with the introduction of graphical user interface (GUI) user friendly analysis pipelines which negate the need for coding proficiency. While improving, pipeline analysis is not yet at a point that can circumvent the need for bioinformatician input. Incorrect interpretation of data can mean the difference between a correct or incorrect diagnosis. Access to sequencing platforms also poses a considerable hurdle in the uptake of mNGS. Alongside access issues to the sequencing instrument itself, flow cells and reagents kits, as well as routine and emergency service technicians to maintain the instruments can be difficult to source at this time.

Start-up and ancillary expenses of purchasing and maintaining a sequencing platform can cost in excess of $70,000AUD, costs that are prohibitive for many diagnostic laboratories. Cost per sample is also considered high in relation to commonly used pathogen detection tests, where one mNGS run averages around $800AUD per sample "in house" and up to $4000AUD for a single sample sequenced in a commercial laboratory, a RADT can cost as little as $2AUD per sample (Simner et al., 2018). It is anticipated further refinements in sequencing technology will result in reduced associated costs; however, affordability will likely continue to be a prohibitive factor in the uptake of mNGS in healthcare systems (Govender et al., 2021). Finally, the time taken to perform NGS can be considered a prohibitive factor in the uptake of mNGS in situations where rapidly evolving disease outbreaks require prompt, tangible results (Gu et al., 2021, Charalampous et al., 2019). The time taken from the start of sample extraction to useable results varies greatly due to multiple factors, including sequencing platform, extraction method, library preparation and sequencing platform access (Gu et al., 2021). In cases where a sequencing platform is not available at the point of sample collection, an external provider may need to be used or samples may need to be relocated to another location for the purpose of sequencing. Sequencing platforms can have very variable run times, with even more variation within platforms such as the Illumina platform with smaller runs taking as little as 4 hours to up to 10 days; estimated run times can vary on cluster density and computer resources (Poulsen et al., 2022). Rapid pathogen detection and their usefulness has been explored in recent studies; one study demonstrated a rapid mNGS testing method which gave useable results in 6 hours. This method was validated using 182 body fluid samples which revealed test specificity of 75% for bacteria and fungi using Illumina® and nanopore sequencing (Gu et al., 2021). Whilst mNGS remains a very useful technology with the potential to be broadly applied to the detection, control and prevention of disease-causing pathogens, its uptake may be slower than expected due to the obstacles challenging the method.

Introduction to the scientific problem

Infectious diseases remain a leading cause of morbidity and mortality worldwide. Pathogen identification of acute respiratory infection can be difficult and is acknowledged as a significant issue for infectious disease clinicians in hospital settings (Serpa et al., 2022). Conventional culture methods have comparatively low detection rates when compared to culture independent methods such as nucleic acid amplification tests (PCR). Precision diagnosis using culture techniques can be challenging where pathogens are fastidious or take a prolonged time to culture. Culture independent techniques such as PCR and serological assays have proven extremely useful for broadening the scope of detectible pathogens and remain the pillars for current clinical diagnostic techniques. Current

diagnostic techniques require prior knowledge, or suspicion of infectious pathogen origin, this is sometimes impractical especially if an infectious disease manifests in an atypical manner. Reviews of current testing practices have reported that up to 60% of all ARI cases treated have no potential pathogens detected despite access to comprehensive testing methods (Gu et al., 2021, Duan et al., 2021, Miao et al., 2018). Delay in the timely diagnosis of a causative pathogen may result in a delay of precision antimicrobial treatment, leading to unnecessary broad-spectrum antibiotic usage, increasing healthcare costs, and introducing anti-microbial resistance. Metagenomic and metatranscriptomic sequencing are unbiased approaches which have the potential to detect most pathogens in a clinical sample, and is most suitable for rare, novel, or atypical aetiologies of infectious diseases. The purpose of the next two data chapters were to address the usefulness of mNGS as a clinical diagnostic tool and as a tool for use in disease outbreak and novel virus identification.

Aims of the Thesis and Chapter Outline

Chapter 5: Diagnosis and Analysis of Unexplained Severe Respiratory Illness in Australian Children using Metagenomic Next-Generation Sequencing: A Pilot Study (Unpublished)

The overarching aim of this chapter was to conduct a proof-of-concept study to evaluate the utility of parallel DNA/RNA mNGS in the identification of unknown pathogens in a cohort of children with unknown infections presenting with ARI like symptoms. Here the use of mNGS was calculated with the purpose of eventual integration of the method in a clinical diagnostic capacity. Care was taken to ensure relevant quality control measures were taken with a goal of ensuring method repeatability at the forefront of the experimental design. Reagent batching, and extraction of DNA/RNA in parallel as well as the inclusion of microbial mock communities sought to ensure meaningful correlations and conclusions could be made about the methods performance.

Chapter 6: RAPIDprep: A simple, fast protocol for RNA metagenomic sequencing of clinical samples (*Viruses,* 2022, under review).

In this Chapter I aim to develop and validate a streamlined RNA-mNGS method for the express purpose of reducing the time taken to obtain tangible results. This method has been proven capable of detecting potential pathogen RNA in less than 24 hours after sample collection. Another aim of this method was to develop an approach which utilizes readily available reagents to ensure ease of access and reproducibility. The purpose of this research body of work was also to evaluate the capacity of the method to detect possible pathogens in a variety of sample types and quality.

References

AHN, J., YANG, L., PASTER, B. J., GANLY, I., MORRIS, L., PEI, Z. & HAYES, R. B. 2011. Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. *PloS one,* 6**,** e22788.

ALBERT, E., TORRES, I., BUENO, F., HUNTLEY, D., MOLLA, E., FERNÁNDEZ-FUENTES, M. Á., MARTÍNEZ, M., POUJOIS, S., FORQUÉ, L., VALDIVIA, A., SOLANO DE LA ASUNCIÓN, C., FERRER, J., COLOMINA, J. & NAVARRO, D. 2021. 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,* 27**,** 472.e7-472.e10.

ALEMANY, A., BARÓ, B., OUCHI, D., RODÓ, P., UBALS, M., CORBACHO-MONNÉ, M., VERGARA-ALERT, J., RODON, J., SEGALÉS, J., ESTEBAN, C., FERNÁNDEZ, G., RUIZ, L., BASSAT, Q., CLOTET, B., ARA, J., VALL-MAYANS, M., G-BEIRAS, C., BLANCO, I. & MITJÀ, O. 2021. Analytical and clinical performance of the panbio COVID-19 antigen-detecting rapid diagnostic test. *The Journal of infection,* 82**,** 186-230.

ALTSCHUL, S. F., GISH, W., MILLER, W., MEYERS, E. W. & LIPMAN, D. J. 1990. Basic Local Alignment Search Tool. *Journal of molecular biology,* 215**,** 403-410.

ARANAZ-ANDRÉS, J. M., CHÁVEZ, A. C. F., LASO, A. M., ABREU, M., NÚÑEZ, P. M., GALÁN, J. C. & MORENO, R. C. 2022. Analysis of the diagnostic accuracy of rapid antigenic tests for detection of SARS-CoV-2 in hospital outbreak situation. *European journal of clinical microbiology & infectious diseases,* 41**,** 305-312.

BAKER, J. L., MORTON, J. T., DINIS, M., ALVAREZ, R., TRAN, N. C., KNIGHT, R. & EDLUND, A. 2021. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome research,* 31**,** 64-74.

BARB, J. J., OLER, A. J., KIM, H.-S., CHALMERS, N., WALLEN, G. R., CASHION, A., MUNSON, P. J. & AMES, N. J. 2016. Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. *PloS one,* 11**,** e0148047-e0148047.

BARTRAM, A. K., LYNCH, M. D. J., STEARNS, J. C., MORENO-HAGELSIEB, G. & NEUFELD, J. D. 2011. Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. *Applied and environmental microbiology,* 77**,** 5569-5569.

BASHIARDES, S. S., ZILBERMAN-SCHAPIRA, G. G. & ELINAV, E. E. 2016. Use of Metatranscriptomics in Microbiome Research. *Bioinformatics and Biology Insights,* 2016**,** 19-25.

BASSO, D., AITA, A., PADOAN, A., COSMA, C., NAVAGLIA, F., MOZ, S., CONTRAN, N., ZAMBON, C.-F., MARIA CATTELAN, A. & PLEBANI, M. 2021. Salivary SARS-CoV-2 antigen rapid detection: A prospective cohort study. *Clinica chimica acta,* 517**,** 54-59.

BEGON, M., TOWNSEND, C. R. & HARPER, J. L. 1986. *Ecology : individuals, populations and communities,* Oxford, Blackwell Scientific.

BERBERS, B., SALTYKOVA, A., GARCIA-GRAELLS, C., PHILIPP, P., ARELLA, F., MARCHAL, K., WINAND, R., VANNESTE, K., ROOSENS, N. H. C. & DE KEERSMAECKER, S. C. J. 2020. Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. *Scientific reports,* 10**,** 4310-4310.

BERRY, G. J., ZHEN, W., SMITH, E., MANJI, R., SILBERT, S., LIMA, A., HARINGTON, A., MCKINLEY, K., KENSINGER, B., NEFF, C. & LU, D. 2022. Multicenter Evaluation of the BioFire Respiratory Panel 2.1 (RP2.1) for Detection of SARS-CoV-2 in Nasopharyngeal Swab Samples. *Journal of clinical microbiology,* 60**,** e0006622-e0006622.

BIGOUETTE, J. P., FORD, L., PRAY, I., LANGOLF, K., KAHRS, J., ZOCHERT, T., TATE, J. E., GIERYN, D., KIRKING, H. L., WESTERGAARD, R. & KILLERBY, M. E. 2021. Application of a Serial Antigen-Based Testing Strategy for Severe Acute Respiratory Syndrome Coronavirus 2 and Student Adherence in a University Setting: Wisconsin, October-November 2020. *Open forum infectious diseases,* 8**,** ofab472-ofab472.

BLAIRON, L., WILMET, A., BEUKINGA, I. & TRÉ-HARDY, M. 2020. Implementation of rapid SARS-CoV-2 antigenic testing in a laboratory without access to molecular methods: Experiences of a general hospital. *Journal of clinical virology,* 129**,** 104472-104472.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BOUZID, D., CASALINO, E., MULLAERT, J., LAURENT, O., DUVAL, X., LESCURE, F. X., PEIFFER SMADJA, N., TUBIANA, S., ARMAND LEFÈVRE, L., DESCAMPS, D., FIDOUH, N., CHOQUET, C., LUCET, J.-C. & VISSEAUX, B. 2021. Added value of rapid respiratory syndromic testing at point of care versus central laboratory testing: a controlled clinical trial. *Journal of antimicrobial chemotherapy,* 76**,** iii20-iii27.

BOZIO, C. H., VUONG, J., DOKUBO, E. K., FALLAH, M. P., MCNAMARA, L. A., POTTS, C. C., DOEDEH, J., GBANYA, M., RETCHLESS, A. C., PATEL, J. C., CLARK, T. A., KOHAR, H., NAGBE, T., CLEMENT, P., KATAWERA, V., MAHMOUD, N., DJINGAREY, H. M., PERROCHEAU, A., NAIDOO, D., STONE, M., GEORGE, R. N., WILLIAMS, D., GASASIRA, A., NYENSWAH, T., WANG, X., FOX, L. M., KONWAY, Y., WIAH, S. Q., DOEDEH, V., BAO, U., SENNEH, G., GORWOR, L., GONOTEE, P., PAASEWE, T., TAMATAI, G., YARKEH, J., SMITH, S., BRIMA-DAVIS, A., DAUDA, G., MONGER, T., GORNOR-PEWU, L. W., LOMBEH, S., WILSON, H. W., KORVAYAN, M., DOVILLIE, N., JETOH, R., TAWEH, F., VERA WALKER, Y., HARDY, P., FREEMAN, M., GEORGE, G., KERWILLAIN, G., TOE, S., GHARTEY, E., LARWAY, L., GWEH, D., ALLEN, D., FRIESEN, S., GWESA, G., KINKADE, C., REED, M., CHANG, A., GEORGE, J., SCHIER, J., THOMAS, J., DIAZ, M. H., JENKINS, L. T., MAHON, B. E., SCHMINK, S. E., JOSEPH, S. J., WALLER, J. L., WHALEY, M. J., WINCHELL, J. M., ARTHUR, R. R., FULLER, S., CHRISTIAN, K., REDD, J. T., YEALUE, K. D. M., NAIENE, J., ASAMOAH FRIMPONG, J., AMO-ADDAE, M., STEPHEN, O., VON GOTTBERG, A. & TAHA, M. 2018. Outbreak of Neisseria meningitidis serogroup C outside the meningitis belt—Liberia, 2017: an epidemiological and laboratory investigation. *The Lancet infectious diseases,* 18**,** 1360-1367.

BRENNER, T., DECKER, S. O., GRUMAZ, S., STEVENS, P., BRUCKNER, T., SCHMOCH, T., PLETZ, M. W., BRACHT, H., HOFER, S., MARX, G., WEIGAND, M. A. & SOHN, K. 2018. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine (Baltimore),* 97**,** e9868-e9868.

BRUMFIELD, K. D., HUQ, A., COLWELL, R. R., OLDS, J. L. & LEDDY, M. B. 2020. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PloS one,* 15**,** e0228899.

CARPENTER, MEREDITH L., BUENROSTRO, JASON D., VALDIOSERA, C., SCHROEDER, H., ALLENTOFT, MORTEN E., SIKORA, M., RASMUSSEN, M., GRAVEL, S., GUILLÉN, S., NEKHRIZOV, G., LESHTAKOV, K., DIMITROVA, D., THEODOSSIEV, N., PETTENER, D., LUISELLI, D., SANDOVAL, K., MORENO-ESTRADA, A., LI, Y., WANG, J., GILBERT, M. THOMAS P., WILLERSLEV, E., GREENLEAF, WILLIAM J. & BUSTAMANTE, CARLOS D. 2013. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *American journal of human genetics,* 93**,** 852-864.

CHARALAMPOUS, T., KAY, G. L., RICHARDSON, H., AYDIN, A., BALDAN, R., JEANES, C., RAE, D., GRUNDY, S., TURNER, D. J., WAIN, J., LEGGETT, R. M., LIVERMORE, D. M. & O'GRADY, J. 2019. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature biotechnology,* 37**,** 783-792.

CHEN, S., ZHOU, Y., CHEN, Y. & GU, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *bioRxiv*.

CHEN, Z., COLLIN, E., PEDDIREDDI, L., CLEMENT, T., GAUGER, P. & HAUSE, B. M. 2017. Genetic diversity in envelope genes of contemporary U.S. porcine reproductive and respiratory syndrome virus strains influences viral antigenicity. *Research in veterinary science,* 115**,** 432-441.

CHO, J., MATUSHEK, S. & CHARNOT-KATSIKAS, A. 2018. Optimizing the Utilization of Multiplex PCR Respiratory Panels at Two Emergency Departments Over Three Consecutive Influenza Seasons. *American journal of clinical pathology,* 150**,** S147-S148.

COLE, J. R., WANG, Q., CARDENAS, E., FISH, J., CHAI, B., FARRIS, R. J., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., MARSH, T., GARRITY, G. M. & TIEDJE, J. M. 2009. Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research,* 37**,** D141-D145.

COLLINSON, N., SNAPE, N., BEAGLEY, K., FANTINO, E. & SPANN, K. 2021. COPD Is Associated with Elevated IFN-β Production by Bronchial Epithelial Cells Infected with RSV or hMPV. *Viruses,* 13**,** 911.

CUMMINGS, L. A., KUROSAWA, K., HOOGESTRAAT, D. R., SENGUPTA, D. J., CANDRA, F., DOYLE, M., THIELGES, S., LAND, T. A., ROSENTHAL, C. A., HOFFMAN, N. G., SALIPANTE, S. J. & COOKSON, B. T. 2016. Clinical Next Generation Sequencing Outperforms Standard Microbiological Culture for Characterizing Polymicrobial Samples. *Clinical chemistry (Baltimore, Md.),* 62**,** 1465-1473.

DAS, A. K., GOSWAMI, S., LEE, K. & PARK, S.-J. 2019. A hybrid and scalable error correction algorithm for indel and substitution errors of long reads. *BMC genomics,* 20**,** 948-948.

DI GIALLONARDO, F., KOK, J., FERNANDEZ, M., CARTER, I., GEOGHEGAN, J. L., DWYER, D. E., HOLMES, E. C. & EDEN, J.-S. 2018. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses,* 10.

DUAN, H., LI, X., MEI, A., LI, P., LIU, Y., LI, X., LI, W., WANG, C. & XIE, S. 2021. The diagnostic value of metagenomic next**-**generation sequencing in infectious diseases. *BMC infectious diseases,* 21**,** 62-62.

DUGAT, T., LOUX, V., MARTHEY, S., MOROLDO, M., LAGRÉE, A.-C., BOULOUIS, H.-J., HADDAD, N. & MAILLARD, R. 2014. Comparative genomics of first available bovine Anaplasma phagocytophilum genome obtained with targeted sequence capture. *BMC genomics,* 15**,** 973-973.

EDEN, J.-S., SIKAZWE, C., XIE, R., DENG, Y.-M., SULLIVAN, S. G., MICHIE, A., LEVY, A., CUTMORE, E., BLYTH, C. C., BRITTON, P. N., CRAWFORD, N., DONG, X., DWYER, D. E., EDWARDS, K. M., HORSBURGH, B. A., FOLEY, D., KENNEDY, K., MINNEY-SMITH, C., SPEERS, D., TULLOCH, R. L., HOLMES, E. C., DHANASEKARAN, V., SMITH, D. W., KOK, J., BARR, I. G. & GROUP, A. R. S. 2021. Off-season RSV epidemics in Australia after easing of COVID-19 restrictions. *medRxiv***,** 2021.07.21.21260810.

FERNANDEZ-MARMIESSE, A., GOUVEIA, S. & COUCE, M. L. 2018. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Current medicinal chemistry,* 25**,** 404-432.

FREED, N. E., VLKOVÁ, M., FAISAL, M. B. & SILANDER, O. K. 2020. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biology Methods & Protocols,* 5.

FREY, K. G., HERRERA-GALEANO, J. E., REDDEN, C. L., LUU, T. V., SERVETAS, S. L., MATECZUN, A. J., MOKASHI, V. P. & BISHOP-LILLY, K. A. 2014. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC genomics,* 15**,** 96-96.

GOVENDER, K. N., STREET, T. L., SANDERSON, N. D. & EYRE, D. W. 2021. Metagenomic Sequencing as a Pathogen-Agnostic Clinical Diagnostic Tool for Infectious Diseases: a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *Journal of clinical microbiology,* 59**,** e0291620-e0291620.

GRENINGER, A. L. 2018. The challenge of diagnostic metagenomics. *Expert review of molecular diagnostics,* 18**,** 605-615.

GU, W., DENG, X., LEE, M., SUCU, Y. D., AREVALO, S., STRYKE, D., FEDERMAN, S., GOPEZ, A., REYES, K., ZORN, K., SAMPLE, H., YU, G., ISHPUNIANI, G., BRIGGS, B., CHOW, E. D., BERGER, A., WILSON, M. R., WANG, C., HSU, E., MILLER, S., DERISI, J. L. & CHIU, C. Y. 2021. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nature medicine,* 27**,** 115-124.

HANDELSMAN, J., RONDON, M. R., BRADY, S. F., CLARDY, J. & GOODMAN, R. M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology,* 5**,** R245-R249.

HASAN, M. R., SUNDARARAJU, S., TANG, P., TSUI, K.-M., LOPEZ, A. P., JANAHI, M., TAN, R. & TILLEY, P. 2020. A metagenomics-based diagnostic approach for central nervous system infections in hospital acute care setting. *Scientific reports,* 10**,** 11194-11194.

HOGAN, C. A., YANG, S., GARNER, O. B., GREEN, D. A., GOMEZ, C. A., DIEN BARD, J., PINSKY, B. A. & BANAEI, N. 2021. Clinical Impact of Metagenomic Next-Generation Sequencing of Plasma Cell-Free DNA for the Diagnosis of Infectious Diseases: A Multicenter Retrospective Cohort Study. *Clinical infectious diseases,* 72**,** 239-245.

JACOBS, J., KÜHNE, V., LUNGUYA, O., AFFOLABI, D., HARDY, L. & VANDENBERG, O. 2020. Implementing COVID-19 (SARS-CoV-2) Rapid Diagnostic Tests in Sub-Saharan Africa: A Review. *Frontiers in medicine,* 7**,** 557797-557797.

JAWORSKI, E., LANGSJOEN, R. M., MITCHELL, B., JUDY, B., NEWMAN, P., PLANTE, J. A., PLANTE, K. S., MILLER, A. L., ZHOU, Y., SWETNAM, D., SOTCHEFF, S., MORRIS, V., SAADA, N., MACHADO, R. R., MCCONNELL, A., WIDEN, S. G., THOMPSON, J., DONG, J., REN, P., PYLES, R. B., KSIAZEK, T. G., MENACHERY, V. D., WEAVER, S. C. & ROUTH, A. L. 2021. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *eLife,* 10.

JONES, M. R. & GOOD, J. M. 2016. Targeted capture in evolutionary and ecological genomics. *Molecular ecology,* 25**,** 185-202.

KAN-O, K., RAMIREZ, R., MACDONALD, M. I., ROLPH, M., RUDD, P. A., SPANN, K. M., MAHALINGAM, S., BARDIN, P. G. & THOMAS, B. J. 2017. Human Metapneumovirus Infection in Chronic Obstructive Pulmonary Disease: Impact of Glucocorticosteroids and Interferon. *The Journal of Infectious Diseases,* 215**,** 1536-1545.

KHURSHID, Z., ZOHAIB, S., NAJEEB, S., ZAFAR, M. S., SLOWEY, P. D. & ALMAS, K. 2016. Human Saliva Collection Devices for Proteomics: An Update. *International journal of molecular sciences,* 17**,** 846-846.

KIM, H.-M., JEON, S., CHUNG, O., JUN, J. H., KIM, H.-S., BLAZYTE, A., LEE, H.-Y., YU, Y., CHO, Y. S., BOLSER, D. M. & BHAK, J. 2021. Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *Gigascience,* 10.

KOPYLOVA, E., NOE, L. & TOUZET, H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics,* 28**,** 3211-3217.

LAMBERT-NICLOT, S., CUFFEL, A., LE PAPE, S., VAULOUP-FELLOUS, C., MORAND-JOUBERT, L., ROQUE-AFONSO, A.-M., LE GOFF, J. & DELAUGERRE, C. 2020. Evaluation of a Rapid Diagnostic Assay for Detection of SARS-CoV-2 Antigen in Nasopharyngeal Swabs. *Journal of clinical microbiology,* 58.

LEUNG, H. C. M., YIU, S.-M. & CHIN, F. Y. L. 2015. IDBA-MTP: A Hybrid Metatranscriptomic Assembler Based on Protein Information. *Journal of computational biology,* 22**,** 367-376.

LEUNG, H. C. M., YIU, S.-M., PARKINSON, J. & CHIN, F. Y. L. 2013. IDBA-MT: De Novo Assembler for Metatranscriptomic Data Generated from Next-Generation Sequencing Technology. *Journal of computational biology,* 20**,** 54-550.

LI, B. & YAN, T. 2021. Next generation sequencing reveals limitation of qPCR methods in quantifying emerging antibiotic resistance genes (ARGs) in the environment. *Applied microbiology and biotechnology,* 105**,** 2925-2936.

LI, D., LUO, R., LIU, C.-M., LEUNG, C.-M., TING, H.-F., SADAKANE, K., YAMASHITA, H. & LAM, T.-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods (San Diego, Calif.),* 102**,** 3-11.

LIN, L., CARLQUIST, J., SINCLAIR, W., HALL, T., LOPANSRI, B. K. & BENNETT, S. T. 2021. Experience With False-Positive Test Results on the TaqPath Real-Time Reverse Transcription-Polymerase Chain Reaction Coronavirus Disease 2019 (COVID-19) Testing Platform. *Archives of pathology & laboratory medicine (1976),* 145**,** 259-261.

LIU, L. P., OZA, S. M., HOGAN, D. P., PERIN, J. P., RUDAN, I. P., LAWN, J. E. P., COUSENS, S. P., MATHERS, C. P. & BLACK, R. E. P. 2015. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet (British edition),* 385**,** 430-440.

LO, C.-C. & CHAIN, P. S. G. 2014. Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC bioinformatics,* 15**,** 366-366.

LUO, C., TSEMENTZI, D., KYRPIDES, N., READ, T. & KONSTANTINIDIS, K. T. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one,* 7**,** e30087-e30087.

MCDONALD, D., PRICE, M. N., GOODRICH, J., NAWROCKI, E. P., DESANTIS, T. Z., PROBST, A., ANDERSEN, G. L., KNIGHT, R. & HUGENHOLTZ, P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal,* 6**,** 610-618.

MIAO, Q., MA, Y., WANG, Q., PAN, J., ZHANG, Y., JIN, W., YAO, Y., SU, Y., HUANG, Y., WANG, M., LI, B., LI, H., ZHOU, C., LI, C., YE, M., XU, X., LI, Y. & HU, B. 2018. Microbiological Diagnostic Performance of Metagenomic Next-generation Sequencing When Applied to Clinical Practice. *Clinical infectious diseases,* 67**,** S231-S240.

MOSS, S., ATSAWARUNGRUANGKIT, A., DANG, L., CHUA, D., ZHOU, Y., CHONG, Z. Z., ZHANG, H. & GRAHAM, D. Y. 2021. S1381 Rapid Prediction of H. pylori Antibiotic Resistance Using Next Generation Sequencing of Stool Samples Compared to Gastric Biopsies. *The American journal of gastroenterology,* 116**,** S634-S634.

NEUMANN, B., BENDER, J. K., MAIER, B. F., WITTIG, A., FUCHS, S., BROCKMANN, D., SEMMLER, T., EINSELE, H., KRAUS, S., WIELER, L. H., VOGEL, U. & WERNER, G. 2020. Comprehensive integrated NGS-based surveillance and contact-network modeling unravels transmission dynamics of vancomycin-resistant enterococci in a high-risk population within a tertiary care hospital. *PloS one,* 15**,** e0235160-e0235160.

NIMWEGEN, K. J. M. V., SOEST, R. A., VELTMAN, J. A., NELEN, M. R., WILT, G. J. V. D., PEART-VISSERS, L. E. L. M. & GRUTTERS, J. P. C. 2016. Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. *Clinical chemistry (Baltimore, Md.),* 62**,** 1458-1464.

NOUSIAS, O. & MONTESANTO, F. 2021. Metagenomic profiling of host-associated bacteria from 8 datasets of the red alga Porphyra purpurea with MetaPhlAn3. *Marine genomics,* 59**,** 100866-100866.

NURK, S., MELESHKO, D., KOROBEYNIKOV, A. & PEVZNER, P. A. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome research,* 27**,** 824-834.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C. M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V. S., KODALI, V. K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K. M., MURPHY, M. R., O'NEILL, K., PUJAR, S., RANGWALA, S. H., RAUSCH, D., RIDDICK, L. D., SCHOCH, C., SHKEDA, A., STORZ, S. S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R. E., VATSAN, A. R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M. J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T. D. & PRUITT, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research,* 44**,** D733-D745.

OLIVER, J. D., NILSSON, L. & KJELLEBERG, S. 1991. Formation of nonculturable Vibrio vulnificus cells and its relationship to the starvation state. *Applied and Environmental Microbiology,* 57**,** 2640-2644.

OYOLA, S. O., HENSON, S. P., NZAU, B., KIBWANA, E. & NENE, V. 2021. Access to ultra-long IgG CDRH3 bovine antibody sequences using short read sequencing technology. *Molecular immunology,* 139**,** 97-105.

PALMER, C., BIK, E. M., EISEN, M. B., ECKBURG, P. B., SANA, T. R., WOLBER, P. K., RELMAN, D. A. & BROWN, P. O. 2006. Rapid quantitative profiling of complex microbial populations. *Nucleic acids research,* 34**,** e5-e5.

PANDA, K. & SLOTKIN, R. K. 2020. Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of Transposable Elements. *The Plant cell,* 32**,** 2687-2698.

PFEIFFER, F., GRÖBER, C., BLANK, M., HÄNDLER, K., BEYER, M., SCHULTZE, J. L. & MAYER, G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific reports,* 8**,** 10950-14.

POPOWITCH, E. B., O'NEILL, S. S. & MILLER, M. B. 2013. Comparison of the Biofire FilmArray RP, Genmark eSensor RVP, Luminex xTAG RVPv1, and Luminex xTAG RVP Fast Multiplex Assays for Detection of Respiratory Viruses. *Journal of Clinical Microbiology,* 51**,** 1528-1533.

POULSEN, C. S., EKSTRØM, C. T., AARESTRUP, F. M. & PAMP, S. J. 2022. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology spectrum,* 10**,** e0009022-e0009022.

QUER, J., COLOMER-CASTELL, S., CAMPOS, C., ANDRÉS, C., PIÑANA, M., CORTESE, M. F., GONZÁLEZ-SÁNCHEZ, A., GARCIA-CEHIC, D., IBÁÑEZ, M., PUMAROLA, T., RODRÍGUEZ-FRÍAS, F., ANTÓN, A. & TABERNERO, D. 2022. Next-Generation Sequencing for Confronting Virus Pandemics. *Viruses,* 14**,** 600.

ROCKETT, R. J., ARNOTT, A., LAM, C., SADSAD, R., TIMMS, V., GRAY, K.-A., EDEN, J.-S., CHANG, S., GALL, M., DRAPER, J., SIM, E. M., BACHMANN, N. L., CARTER, I., BASILE, K., BYUN, R., O'SULLIVAN, M. V., CHEN, S. C. A., MADDOCKS, S., SORRELL, T. C., DWYER, D. E., HOLMES, E. C., KOK, J., PROKOPENKO, M. & SINTCHENKO, V. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine,* 26**,** 1398-1404.

ROSZAK, D. B., GRIMES, D. J. & COLWELL, R. R. 1984. Viable but nonrecoverable stage of Salmonella enteritidis in aquatic systems. *Canadian journal of microbiology,* 30**,** 334-338.

ROTA, P. A., OBERSTE, M. S., MONROE, S. S., NIX, W. A., CAMPAGNOLI, R., ICENOGLE, J. P., PEÑARANDA, S., BANKAMP, B., MAHER, K., CHEN, M., TONG, S., TAMIN, A., LOWE, L., FRACE, M., DERISI, J. L., CHEN, Q., WANG, D., ERDMAN, D. D., PERET, T. C. T., BURNS, C., KSIAZEK, T. G., ROLLIN, P. E., SANCHEZ, A., LIFFICK, S., HOLLOWAY, B., LIMOR, J., MCCAUSTLAND, K., OLSEN-RASMUSSEN, M., FOUCHIER, R., GÜNTHER, S., OSTERHAUS, A., STEN, C., PALLANSCH, M. A., ANDERSON, L. J. & BELLINI, W. J. 2003. Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *Science (American Association for the Advancement of Science),* 300**,** 1394-1399.

SANDERS, J. G., NURK, S., SALIDO, R. A., MINICH, J., XU, Z. Z., ZHU, Q., MARTINO, C., FEDARKO, M., ARTHUR, T. D., CHEN, F., BOLAND, B. S., HUMPHREY, G. C., BRENNAN, C., SANDERS, K., GAFFNEY, J., JEPSEN, K., KHOSROHEIDARI, M., GREEN, C., LIYANAGE, M., DANG, J. W., PHELAN, V. V., QUINN, R. A., BANKEVICH, A., CHANG, J. T., RANA, T. M., CONRAD, D. J., SANDBORN, W. J., SMARR, L., DORRESTEIN, P. C., PEVZNER, P. A. & KNIGHT, R. 2019. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology,* 20**,** 226-226.

SAYERS, E. W., CAVANAUGH, M., CLARK, K., OSTELL, J., PRUITT, K. D. & KARSCH-MIZRACHI, I. 2019. GenBank. *Nucleic acids research,* 47**,** D94-D99.

SCHOONVAERE, K., DE SMET, L., SMAGGHE, G., VIERSTRAETE, A., BRAECKMAN, B. P. & DE GRAAF, D. C. 2016. Unbiased RNA Shotgun Metagenomics in Social and Solitary Wild Bees Detects Associations with Eukaryote Parasites and New Viruses. *PloS one,* 11**,** e0168456-e0168456.

SERPA, P. H., DENG, X., ABDELGHANY, M., CRAWFORD, E., MALCOLM, K., CALDERA, S., FUNG, M., MCGEEVER, A., KALANTAR, K. L., LYDEN, A., GHALE, R., DEISS, T., NEFF, N., MILLER, S. A., DOERNBERG, S. B., CHIU, C. Y., DERISI, J. L., CALFEE, C. S. & LANGELIER, C. R. 2022. Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. *Genome medicine,* 14**,** 1-74.

SHAFFER, J. P., CARPENTER, C. S., MARTINO, C., SALIDO, R. A., MINICH, J. J., BRYANT, M., SANDERS, K., SCHWARTZ, T., HUMPHREY, G., SWAFFORD, A. D. & KNIGHT, R. 2022. A comparison of six DNA extraction protocols for 16S, ITS and shotgun metagenomic sequencing of microbial communities. *BioTechniques,* 73**,** 34-46.

SHAKYA, M., LO, C.-C. & CHAIN, P. S. G. 2019. Advances and Challenges in Metatranscriptomic Analysis. *Frontiers in genetics,* 10**,** 904-904.

SHARMA, N., GAUTAM, H., TYAGI, S., RAZA, S., MOHAPATRA, S., SOOD, S., DHAWAN, B., KAPIL, A. & DAS, B. 2022. Clinical use of multiplex-PCR for the diagnosis of acute bacterial meningitis. *Journal of family medicine and primary care,* 11**,** 593-598.

SIMNER, P. J., MILLER, S. & CARROLL, K. C. 2018. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clinical infectious diseases,* 66**,** 778-788.

SINGH, N., CHIKARA, S. & SUNDAR, S. 2013. SOLiD™ sequencing of genomes of clinical isolates of Leishmania donovani from India confirm leptomonas co-infection and raise some key questions. *PloS one,* 8**,** e55738-e55738.

SOTIRIOU, C. & PUSZTAI, L. 2009. Gene-Expression Signatures in Breast Cancer. *The New England journal of medicine,* 360**,** 790-800.

TULLOCH, R. L., KOK, J., CARTER, I., DWYER, D. E. & EDEN, J.-S. 2021. An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus. *Viruses,* 13**,** 499.

TYSON, J. R., JAMES, P., STODDART, D., SPARKS, N., WICKENHAGEN, A., HALL, G., CHOI, J. H., LAPOINTE, H., KAMELIAN, K., SMITH, A. D., PRYSTAJECKY, N., GOODFELLOW, I., WILSON, S. J., HARRIGAN, R., SNUTCH, T. P., LOMAN, N. J. & QUICK, J. 2020. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*.

UWAMINO, Y., NAGATA, M., AOKI, W., NAKAGAWA, T., INOSE, R., YOKOTA, H., FURUSAWA, Y., SAKAI-TAGAWA, Y., IWATSUKI-HORIMOTO, K., KAWAOKA, Y., HASEGAWA, N. & MURATA, M. 2021. Accuracy of rapid antigen detection test for nasopharyngeal swab specimens and saliva samples in comparison with RT-PCR and viral culture for SARS-CoV-2 detection. *Journal of infection and chemotherapy : official journal of the Japan Society of Chemotherapy,* 27**,** 1058-1062.

VALENCIA, C. A., PERVAIZ, M. A., HUSAMI, A., QIAN, Y. & ZHANG, K. 2013. *Next Generation Sequencing Technologies in Medical Genetics,* Dordrecht, Springer.

VAN BOHEEMEN, S., VAN RIJN, A. L., PAPPAS, N., CARBO, E. C., VORDERMAN, R. H. P., SIDOROV, I., VAN `T HOF, P. J., MEI, H., CLAAS, E. C. J., KROES, A. C. M. & DE VRIES, J. J. C. 2020. Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients. *The Journal of molecular diagnostics : JMD,* 22**,** 196-207.

VAN DER PALEN, M., GILLET, P., BOTTIEAU, E., CNOPS, L., VAN ESBROECK, M. & JACOBS, J. 2009. Test characteristics of two rapid antigen detection tests (SD FK50 and SD FK60) for the diagnosis of malaria in returned travellers. *Malaria journal,* 8**,** 90-90.

WALSH, L. H., COAKLEY, M., WALSH, A. M., O'TOOLE, P. W. & COTTER, P. D. 2022. Bioinformatic approaches for studying the microbiome of fermented food. *Critical reviews in microbiology,* ahead-of-print**,** 1-33.

WARD, C. M., TO, T.-H. & PEDERSON, S. M. 2020. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. *Bioinformatics,* 36**,** 2587-2588.

WATTAM, A. R., ABRAHAM, D., DALAY, O., DISZ, T. L., DRISCOLL, T., GABBARD, J. L., GILLESPIE, J. J., GOUGH, R., HIX, D., KENYON, R., MACHI, D., MAO, C., NORDBERG, E. K., OLSON, R., OVERBEEK, R., PUSCH, G. D., SHUKLA, M., SCHULMAN, J., STEVENS, R. L., SULLIVAN, D. E., VONSTEIN, V., WARREN, A., WILL, R., WILSON, M. J. C., YOO, H. S., ZHANG, C., ZHANG, Y. & SOBRAL, B. W. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research,* 42**,** D581-D591.

WILSON, M. L. 2015. Diagnostic microbiology: the accelerating transition from culture-based to molecular-based methods. *American journal of clinical pathology,* 143**,** 766-767.

WILSON, M. R., NACCACHE, S. N., SAMAYOA, E., BIAGTAN, M., BASHIR, H., YU, G., SALAMAT, S. M., SOMASEKAR, S., FEDERMAN, S., MILLER, S., SOKOLIC, R., GARABEDIAN, E., CANDOTTI, F., BUCKLEY, R. H., REED, K. D., MEYER, T. L., SEROOGY, C. M., GALLOWAY, R., HENDERSON,

S. L., GERN, J. E., DERISI, J. L. & CHIU, C. Y. 2014. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *The New England journal of medicine,* 370**,** 2408-2417.

WILSON, M. R., SAMPLE, H. A., ZORN, K. C., AREVALO, S., YU, G., NEUHAUS, J., FEDERMAN, S., STRYKE, D., BRIGGS, B., LANGELIER, C., BERGER, A., DOUGLAS, V., JOSEPHSON, S. A., CHOW, F. C., FULTON, B. D., DERISI, J. L., GELFAND, J. M., NACCACHE, S. N., BENDER, J., DIEN BARD, J., MURKEY, J., CARLSON, M., VESPA, P. M., VIJAYAN, T., ALLYN, P. R., CAMPEAU, S., HUMPHRIES, R. M., KLAUSNER, J. D., GANZON, C. D., MEMAR, F., OCAMPO, N. A., ZIMMERMANN, L. L., COHEN, S. H., POLAGE, C. R., DEBIASI, R. L., HALLER, B., DALLAS, R., MARON, G., HAYDEN, R., MESSACAR, K., DOMINGUEZ, S. R., MILLER, S. & CHIU, C. Y. 2019. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *The New England journal of medicine,* 380**,** 2327-2340.

WOOD, D. E. & SALZBERG, S. L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology,* 15**,** R46-R46.

XIAO, M., LIU, X., JI, J., LI, M., LI, J., YANG, L., SUN, W., REN, P., YANG, G., ZHAO, J., LIANG, T., REN, H., CHEN, T., ZHONG, H., SONG, W., WANG, Y., DENG, Z., ZHAO, Y., OU, Z., WANG, D., CAI, J., CHENG, X., FENG, T., WU, H., GONG, Y., YANG, H., WANG, J., XU, X., ZHU, S., CHEN, F., ZHANG, Y., CHEN, W., LI, Y. & LI, J. 2020. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome medicine,* 12**,** 57-57.

YANG, L., HAIDAR, G., ZIA, H., NETTLES, R., QIN, S., WANG, X., SHAH, F., RAPPORT, S. F., CHARALAMPOUS, T., METHÉ, B., FITCH, A., MORRIS, A., MCVERRY, B. J., O'GRADY, J. & KITSIOS, G. D. 2019. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respiratory research,* 20**,** 265-12.

YE, Y. & TANG, H. 2016. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics,* 32**,** 1001-1008.

YOUNG, S., TAYLOR, S. N., CAMMARATA, C. L., VARNADO, K. G., ROGER-DALBERT, C., MONTANO, A., GRIEGO-FULLBRIGHT, C., BURGARD, C., FERNANDEZ, C., ECKERT, K., ANDREWS, J. C., REN, H., ALLEN, J., ACKERMAN, R. & COOPER, C. K. 2020. Clinical Evaluation of BD Veritor SARS-CoV-2 Point-of-Care Test Performance Compared to PCR-Based Testing and versus the Sofia 2 SARS Antigen Point-of-Care Test. *Journal of clinical microbiology,* 59.

YU, P., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-1428.

YUAN, G., WANG, H., ZHAO, Y., MAO, E., LI, M., WANG, R., ZHOU, F., JIN, S., ZHANG, Z., XU, K., XU, J., LIANG, S., LI, X., JIANG, L., ZHANG, L., SONG, J., YANG, T., GUO, J., ZHANG, H., ZHOU, Y., WANG, S., QIU, C., JIANG, N., AI, J., WU, J. & ZHANG, W. 2022. Early identification and severity prediction of acute respiratory infection : a study protocol for a randomized controlled trial. *BMC infectious diseases,* 22**,** 1-632.

ZINTER, M. S., DVORAK, C. C., MAYDAY, M. Y., IWANAGA, K., LY, N. P., MCGARRY, M. E., CHURCH, G. D., FARICY, L. E., ROWAN, C. M., HUME, J. R., STEINER, M. E., CRAWFORD, E. D., LANGELIER, C., KALANTAR, K., CHOW, E. D., MILLER, S., SHIMANO, K., MELTON, A., YANIK, G. A., SAPRU, A. & DERISI, J. L. 2019. Pulmonary Metagenomic Sequencing Suggests Missed Infections in Immunocompromised Children. *Clinical infectious diseases,* 68**,** 1847-1855.

ZUMLA, A. P., AL-TAWFIQ, J. A. M. D., ENNE, V. I. P., KIDD, M. P., DROSTEN, C. P., BREUER, J. P., MULLER, M. A. P., HUI, D. P., MAEURER, M. P., BATES, M. P., MWABA, P. P., AL-HAKEEM, R. M. D., GRAY, G. P., GAUTRET, P. P., AL-RABEEAH, A. A. F., MEMISH, Z. A. P. & GANT, V. F. 2014. Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections—needs, advances, and future prospects. *The Lancet infectious diseases,* 14**,** 1123-1135.

CHAPTER 5

Diagnosis and Analysis of Unexplained Severe Respiratory Illness in Australian Children using Metagenomic Next-Generation Sequencing: A Pilot Study

Chapter 5: Diagnosis and Analysis of Unexplained Severe Respiratory Illness in Australian Children using Metagenomic Next-Generation Sequencing: A Pilot Study

Rachel L. Tulloch[1,2], Rebecca Burrell[2,3], John-Sebastian Eden[1,2*] & Philip N. Britton[2,3*]

Unpublished

[1]Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia

[2]Sydney Institute for Infectious Diseases, Faculty of Health and Medicine, The University of Sydney, Westmead, NSW 2145, Australia

[3]Departments of Infectious Diseases and Microbiology, The Children's Hospital at Westmead, Westmead, NSW, 2145, Australia

*Contributed equally as co-senior authors.

Correspondence

Dr John-Sebastian Eden

Email: js.eden@sydney.edu.au

Phone: +61 2 8627 1817

A/Prof Philip Britton

Email: philip.britton@health.nsw.gov.au

Phone: +61 2 9845 3139

Keywords

Metagenomics; Metatranscriptomics; Respiratory infections; Diagnostics; Paediatrics

Abstract

Traditional diagnostic techniques such as Polymerase chain reaction (PCR) for respiratory pathogens remain widely used globally as affordable and reliable methods of pathogen detection. These techniques are pathogen-specific and require some knowledge of the infectious agent prior to testing. On the other hand, metagenomic next-generation sequencing (mNGS) exists as an unbiased method of detection with no prior knowledge of potential pathogens required. Despite this, the translation of mNGS into clinical pathology has been slow and a number of technical challenges remains. We sought to conduct a proof-of-concept study to evaluate the utility of parallel DNA/RNA mNGS in the identification of unknown pathogens in a cohort of children with acute respiratory infection (ARI) like symptoms. In total, 40 children were consented for inclusion in this study with 23 residual clinical diagnostic samples collected between throughout a single respiratory season. Using uniform protocols and reagent batching, the total DNA and RNA were extracted from all samples and deep sequenced in parallel. Microbial mock communities acted as a positive control throughout the study to ensure study validity. Sequencing reads were classified using a suite of metagenomic analysis tools to determine taxonomic identities. The diagnostic results and clinical notes were utilised to make an inference about the reliability of metagenomic data. We identified multiple abundant, cross-domain respiratory microbiomes containing bacteria, fungi, and viral sequences across all patient samples. Potentially pathogenic bacteria were ubiquitous across many clinical samples; however, overabundance of these species could not be reliably classified as possible causes of disease as many were commensal in origin. Novel diagnoses were detected utilising mNGS in two clinical samples, whilst all but one culture positive result was replicated by mNGS method in this study. Time of sampling and sample quality were identified to be the biggest factors in affecting mNGS data usefulness. A purpose developed parallel DNA/RNA metagenomic pipeline demonstrates capacity to identify unknown pathogens and elucidate information of the pulmonary microbiome of children admitted with unknown ARI. Potential pathogens were best identified in samples of high quality, where relative abundance can act as a tool for causative pathogen identification. Ongoing investigation is required before mNGS can be offered as a routine diagnostic service, with many constraints centralised around reliably differentiating pathogenic microbes and normal respiratory commensals.

Introduction

Acute respiratory illness (ARI) is the leading cause of morbidity and mortality in paediatric patients globally (Liu et al., 2015). It is estimated that approximately 1.3 million deaths in children under the age of 5 can be attributed to ARI globally (Tazinya et al., 2018). The presentation of ARI is ambiguous and differential diagnosis of etiological agents can be difficult using clinical features only. It is accepted that a key element to successful ARI management is the early and accurate targeting of causative pathogens (Vaughn et al., 2019). Thus, the prompt and accurate identification of causal pathogens is critical to optimisation of an effective treatment regime and necessary in determining infection control measures (Brenner et al., 2018).

Traditionally, laboratory-based testing has been considered optimal for patients admitted to hospital with ARI. However, laboratory testing using culture-based methods for viral or bacterial isolation have low sensitivity and lengthy turnaround times (48-72 hours from sampling to actionable results) (Jain et al., 2015). Laboratory advancement in involving rapid polymerase-chain reaction (PCR) tests have more recently gained prominence as a diagnostic tool. Whilst more sensitive than culture-based approaches, the fundamental limitation of PCR diagnostics is that they are targeted and requiring the pre-selection of pathogens prior to sample analysis (Zumla et al., 2014). As novel pathogens emerge, it is likely that most will fail to be identified using PCR and assays may require constant updating in the face of pathogen evolution (Jaworski et al., 2021). The latest advance has been cause agnostic, culture-independent diagnostic methods in the form of clinical metagenomic next-generation sequencing (mNGS) that offers a way to resolve patient microbial communities in a variety of clinical syndromes including ARI (Zumla et al., 2014, Fernandez-Marmiesse et al., 2018). The simplest application of mNGS involves the unbiased sequencing of all the DNA (meta-genomics) or RNA (meta-transcriptomics) in a sample (Binns et al., 2022). Typically, mNGS tools demonstrate good sensitivity for low abundance pathogens, including viruses, however the greatest appeal of these methods is the ability to identify pathogens without the need for prior knowledge or speculation around the pathogen origin (Annand et al., 2022, Sikazwe et al., 2022a). Despite many proof-of-concept studies highlighting the usefulness of mNGS in novel and unknown pathogen identification, there exists a plethora of barriers preventing the uptake of these methods in accredited laboratories as a reliable pathogen identification method. Many of these barriers are centralised around the difficulty in standardising mNGS methods across multiple platforms and sample types in a way that is both reproducible and reliable.

The development of short read metagenomic sequencing, such as Illumina and MGI Tech Co., Ltd. (MGI) platforms, present unprecedented capacity for the detailed profiling of microbial communities (Pendleton et al., 2017, Kim et al., 2021). Unlike assays that target the 16S and/or 28S/ITS ribosomal RNA amplicons, unbiased metagenomic next-generation sequencing (mNGS) allows for the characterisation and detection of viral, bacterial, and fungal pathogens at a species level (Caporaso et al., 2012, Pendleton et al., 2017). The utility of mNGS as a diagnostic tool has been demonstrated in limited studies of adults with ARI, and has demonstrated usefulness in the etiologic diagnosis of severe pneumonia in mechanically-ventilated ICU patients (Yang et al., 2019, Wilson et al., 2019, Bal et al., 2018, Zhang et al., 2018). During the SARS-CoV-2 pandemic mNGS was utilised as a tool for the identification, diagnosis and epidemiological analysis of the virus (Yuan et al., 2022, Quer et al., 2022, Fricke et al., 2021, Pfefferle et al., 2021). Throughout the pandemic the clinical and public health utility of the use of mNGS to characterise emerging novel variants of the virus became apparent. The actional information generated through the use of mNGS methods assisted in the prevention or mitigation of emerging variants, provided insights in to viral transmission patterns and was critical in elucidating information on co-circulating respiratory pathogens and variants during the pandemic (Pfefferle et al., 2021, Jaworski et al., 2021, Yang et al., 2021, John et al., 2021).

In this proof-of-concept study, we sought to evaluate the technical feasibility and clinical utility of Illumina metagenomic sequencing as a diagnostic tool for children with severe ARI of unknown origin in the ICU. We explore the usefulness of mNGS in a diagnostic capacity and evaluate the applications of the technology where the causative pathogen remains undetected using traditional, culture-based or PCR, diagnostic tools. The clinical characteristics, treatments, and traditional diagnostic methods available for each patient and sample were evaluated alongside mNGS findings in a cohort of paediatric patients with severe ARI. We hypothesized that an optimized mNGS assay may improve the characterization of the pulmonary microbiome and may assist in the identification of novel pulmonary pathogens in the study cohort. This study also aims to measure the effectiveness of mNGS in identify putative causative pathogens in cases where viral, bacterial or fungal pathogen origin is unknown or presents in an atypical manner. Finally, we hope to demonstrate the usefulness of data generated in clinical mNGS studies in characterising microbial diversity in complex hospitalised patients at risk of opportunistic infection and for the detection of otherwise unknown/novel pathogens.

Methods

Cohort selection

All respiratory samples that had tested negative on the Seegene Allplex Respiratory Full Panel Assay were screened between August and December at a major tertiary children's hospital in New South Wales (NSW), Australia. A total of 581 negative respiratory samples were linked to 469 separate admissions. Eligible cases required an associated ICU admission and availability of a contemporaneous chest X-ray (CXR) and a residual sample for salvage. Admissions were checked against key components of a subsequent clinical categorisation that included: [1] associated ICU admission (n=108), and [2] availability of a contemporaneous CXR (n=73), and [3] immunocompromise (n=12). The families of a total of 40 eligible children were contacted for consent, from which we obtained 23 samples from 21 children. Most samples were nasopharyngeal aspirates (n=21) with 1 endotracheal tube aspirate and 1 throat swab included in the study cohort. CXR images and reports were assessed by an infectious disease's paediatrician to determine if they showed changes suggestive of inflammatory/infectious aetiology or changes that could be attributed to an alternate diagnosis (i.e. pulmonary oedema). Cases were categorised using the above criteria with expected correlation with the likelihood of an infectious aetiology of their clinical disease:

- Category 1, ICU admitted AND CXR suggested inflammation/infection.
- Category 2, ICU admitted AND immunocompromised.
- Category 3, ICU admitted only.

The study was approved by the Sydney Children's Hospitals Network (SCHN) human research ethics committee (Local Ethics Committee of the Sydney Children's Hospitals Network; approval number HREC18/SCHN/263).

Nucleic Acid Extraction, Quantification and Analysis

DNA and RNA was extracted from each sample in parallel using the ZymoBIOMICS DNA/RNA kit, in accordance with the manufacturer's recommendations. The kit uses bead homogenisation and column-based purification to isolate purified DNA and RNA from biological fluids (with on-column DNase treatment for residual DNA removal of purified RNA). Zymo mock microbial communities and water controls were extracted alongside the clinical samples. The samples were extracted in three separate batches using identical batches of reagents to minimise reagent variation. Nucleic acid extracts were initially quantified using Qubit 4 fluorometer (Invitrogen Life technologies, Carlsbad, CA, USA) using both the Qubit dsDNA HS Assay Kit and Qubit RNA HS Assay Kit (Life Technologies, Eugene, OR, USA)

following the manufacturer's instructions. DNA extracts were screened with the Zymo Femto DNA Quantification Kit, a quantitative PCR (qPCR) based kit, to detect and quantify the amount of human, bacterial and fungal DNA present in each of the samples. Input DNA was calculated by multiplying the input volume by the sample concentration. These values were tabulated and relationships between input DNA and bacterial, fungal, and human DNA were measured. Correlations were calculated using the Pearson correlation coefficient. Statistical analysis and correlation plots were drawn in RStudio (Team, 2020).

Preparation of DNA and RNA Libraries

DNA and RNA libraries were prepared separately from the extracted samples and controls. DNA samples were prepared for sequencing using the Nextera DNA Flex Library Prep kit (Illumina, San Diego, CA, USA) in accordance with the manufacturer's instructions. Following quantification, the extracted DNA was grouped based on input concentration: low 1-9 ng, medium 10-24 ng and high 100-500 ng to determine the minimum number of PCR cycles for library amplification, which was here, 13X, 8X and 5X cycles, respectively. The RNA extracted were prepared for sequencing using the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Takara Bio, Kusatsu, Shiga, JPN) in accordance with manufacturers recommendations. For both the DNA and RNA libraries, unique dual indexes were used limit indexing cross-over during sequencing. Library quantity was determined for both RNA and DNA prepared libraries using the Qubit™ dsDNA HS Assay Kit, whilst quality and size was estimated using a High Sensitivity D1000 ScreenTape® on the 2200 TapeStation® system (Agilent, Waldbronn, Germany). DNA and RNA libraries were each pooled separately with libraries represented equimolarly. The separate DNA and RNA pooled libraries were sequenced on an Illumina NovaSeq 6000 (Illumina, CA, USA) to generate paired-end sequencing reads of approximately 150bp in length with an approximate coverage of 40-50 million reads per library.

Metagenomic Analysis

The raw DNA and RNA mNGS libraries were processed initially by removing low-quality and adaptor sequences using bbduk v37.98 with a phred score threshold of 20 (Bushnell, 2016, Brown et al., 2017). The trimmed and filtered reads were then mapped to the human genome using STAR-aligner v2.6.1b (Dobin et al., 2013) to extract non-human sequences for down-stream analysis. The high-quality, non-human reads from the DNA libraries were then *de novo* assembled using Megahit v1.1.3 (Li et al., 2015) before annotation using blast+ v2.11 (Camacho et al., 2009) and diamond v2.0.11 (Buchfink et al., 2015) against the NCBI GenBank

database. The RNA libraries were analysed using a similar pipeline expect that the high-quality, non-human reads were first rRNA filtered using SortMeRNA v2.1b (Kopylova et al., 2012) with the non-rRNA the focus of the analysis. A final summary table was constructed summarising all blast, diamond and *de-novo* assembly data and relative abundance determined by k-mer counts during assembly. Metagenomic sequence libraries generated in this study have been submitted to the NCBI short read archive (SRA), accession numbers SRR22838411 – SRR22838466. Where sequences were identified as a putative cause of disease, quality filtered reads were mapped to respective references sequences using Geneious Prime 2020.1.2 (https://www.geneious.com). Maximum likelihood trees for possible causative pathogens were estimated using PhyML (Guindon et al., 2010) with the GTR + Gamma substitution model and 1000 bootstraps.

Statistical Analysis

We utilised MetaPhlAn 4.0 to profile the composition of the microbial communities and calculate relative abundance (Blanco-Miguez et al., 2022) . Raw abundance was normalised by taking the total number of reads (quality filtered, non-human DNA or non-rRNA) and calculating the number of positive reads per 1000 total reads. Relative abundance was calculated for all identified unambiguous taxonomic assignments to allow for ease of interpretation where library depth was not uniform. A clustered heatmap was constructed in *R* (v4.2.2) utilising the ggplots package (version 3.1.3) (R Development Core Team, 2010, Gómez-Rubio, 2017). Relative abundance data was imported into *R* in a as a two dimensional array, the data-frame is then converted to a matrix, using the as.matrix function (*Base R*) (R Development Core Team, 2010). Hierarchical clustering is performed on both rows and columns of the data matrix using the h, from this row and column dendrograms are constructed. The row and cluster dendrograms were used to order both samples and rows, first the rows are re-ordered based on row means, the columns or samples are the reordered based on the column means, the addition of dendrograms helps to cluster both microbe species and samples with those that have similar microbial communities. The heatmap.2 function was used to build an enhanced heatmap from the (Gómez-Rubio, 2017). The row Z-Score scaling method was applied to improve visualisation of clusters in a heatmap, the z-score describes the abundance of a microbe relative to other microbes within the same sample measured in standard deviation units.

Results and Discussion

Cohorts and clinical presentation

Twenty families consented for their child's samples to be used in the study. The twenty cases involved 21 admissions, with participating children aged from 1 week to 17 years (mean 4.1 years, SD=5.23). There was one sample per admission for all cases except for one case, in which three negative respiratory samples were recovered from one admission. It was also this child also had a subsequent admission later in the study period from which one eligible sample was available. There was a slight male predominance in consented cases (62% male), which was representative of the respiratory negative ICU admitted population during the screening period (63%). Mean length of stay was 34.1 days (SD=20.24; min=4; max=84), with a mean ICU length of stay of 21.4 days (SD=15.02). The majority of samples were nasopharyngeal aspirates (NPA) (91%), with only one throat swab and one endotracheal aspirate being collected for analysis. This predominance of upper respiratory tract samples is representative of the overall negative respiratory sample types collected from all ICU admitted patients (98% upper respiratory tract; 87% nasopharyngeal aspirates, 6% endotracheal aspirates, 5% respiratory swabs) (**Table 1**)

Table 1: Clinical Summary of Test Cohort.

Case Number	Clinical Category	Sex	Specimen Number	Date of Admission	Sample Date	Sample Source	Length of Stay (days)	Length of ICU Stay (days)	Sample Taken on Day of Admission	Age At Sample (years)	Medical Comorbidity	Immune Compromised?[#]
ICU-001	2	M	1	19/07/2018	3/08/2018	NPA	54	41	15	0.14	Yes	Yes, Steroid
ICU-002	2	F	1	21/07/2018	3/08/2018	NPA	24	9	13	0.98	Yes	Yes, Transplant
ICU-007	1	M	1	25/07/2018	7/08/2018	NPA	44	43	13	10.36	Yes	No
ICU-007	1	M	2	25/07/2018	21/08/2018	NPA	44	43	27	10.40	*As above*	*As above*
ICU-007	1	M	3	25/07/2018	31/08/2018	NPA	44	43	37	10.43	*As above*	*As above*
ICU-019	3	M	1	3/08/2018	23/08/2018	NPA	41	25	20	0.10	Yes	No
ICU-027	3	M	1	22/08/2018	30/08/2018	NPA	11	5	8	0.12	Yes	No
ICU-037	3	M	1	6/09/2018	13/09/2018	NPA	14	4	7	0.24	Yes	No
ICU-053	3	F	1	20/09/2018	28/09/2018	NPA	15	40	8	5.61	Yes	Yes, Steroid
ICU-059	2	F	1	17/10/2018	18/10/2018	NPA	5	4	1	9.83	Yes	Yes, Transplant
ICU-062	1	M	1	16/10/2018	22/10/2018	NPA	51	27	6	0.02	Yes	No
ICU-063	1	M	1	23/10/2018	23/10/2018	NPA	8	3	0	0.43	Yes	No
ICU-068	1	F	1	15/10/2018	3/11/2018	NPA	44	7	19	5.70	Yes	Yes, Chemotherapy
ICU-069	1	F	1	10/10/2018	3/11/2018	NPA	49	28	24	0.36	Yes	No
ICU-072	2	M	1	6/11/2018	7/11/2018	NPA	4	1	1	3.58	No	No
ICU-075*	1	M	1	7/11/2018	9/11/2018	NPA	40	33	2	10.62	Yes	No
ICU-077	1	F	1	5/11/2018	13/11/2018	NPA	25	25	8	0.05	No	No
ICU-080	2	M	1	1/11/2018	15/11/2018	NPA	84	10	14	7.28	Yes	No
ICU-081	3	M	1	4/11/2018	15/11/2018	NPA	17	17	11	0.11	Yes	No
ICU-086	3	F	1	12/11/2018	23/11/2018	Throat	15	1	11	17.24	Yes	No
ICU-102	2	M	1	15/12/2018	16/12/2018	NPA	60	19	1	0.05	Yes	No
ICU-104	1	M	1	18/12/2018	19/12/2018	ETA	50	35	1	13.73	Yes	Yes, Steroid
ICU-107	3	F	1	7/12/2018	29/12/2018	NPA	41	30	22	0.21	Yes	No

*ICU-007 repeat admission; [#]Steroid treatment included extended use/high dose Hydrocortisone, Prednisolone, Dexamethasone, and Deflazacort. Post-transplant immunosuppressants included Tacrolimus and Myfortic. Chemotherapy agents used, not mentioned prior, included Methotrexate, and Methylprednisolone.

Detection of putative pathogenic infectious microbes in clinical samples of respiratory samples

Unbiased shotgun sequencing exists as an untapped diagnostic tool to assist in identifying and quantifying known and novel pathogens (Bal et al., 2018, Hasan et al., 2020). Common reasons metagenomic sequencing has not been widely accepted as a routine diagnostic tool include the cost and infrastructure required to perform the sequencing, as well as the time and resources necessary to interpret the results (Greninger, 2018). Whilst there are these potential prohibitive reasons to using next generation sequencing in place of conventional diagnostic tools such as quantitative PCR, the promise of this tool where conventional methods do not yield a result is apparent. Here, we examined respiratory specimens collected from a cohort of children admitted to hospital with severe, acute respiratory illness (SARI). All samples and controls were processed by the parallel mNGS of extracted DNA and RNA, and the non-human sequences were taxonomically assigned and quantified (**Figure 1**).

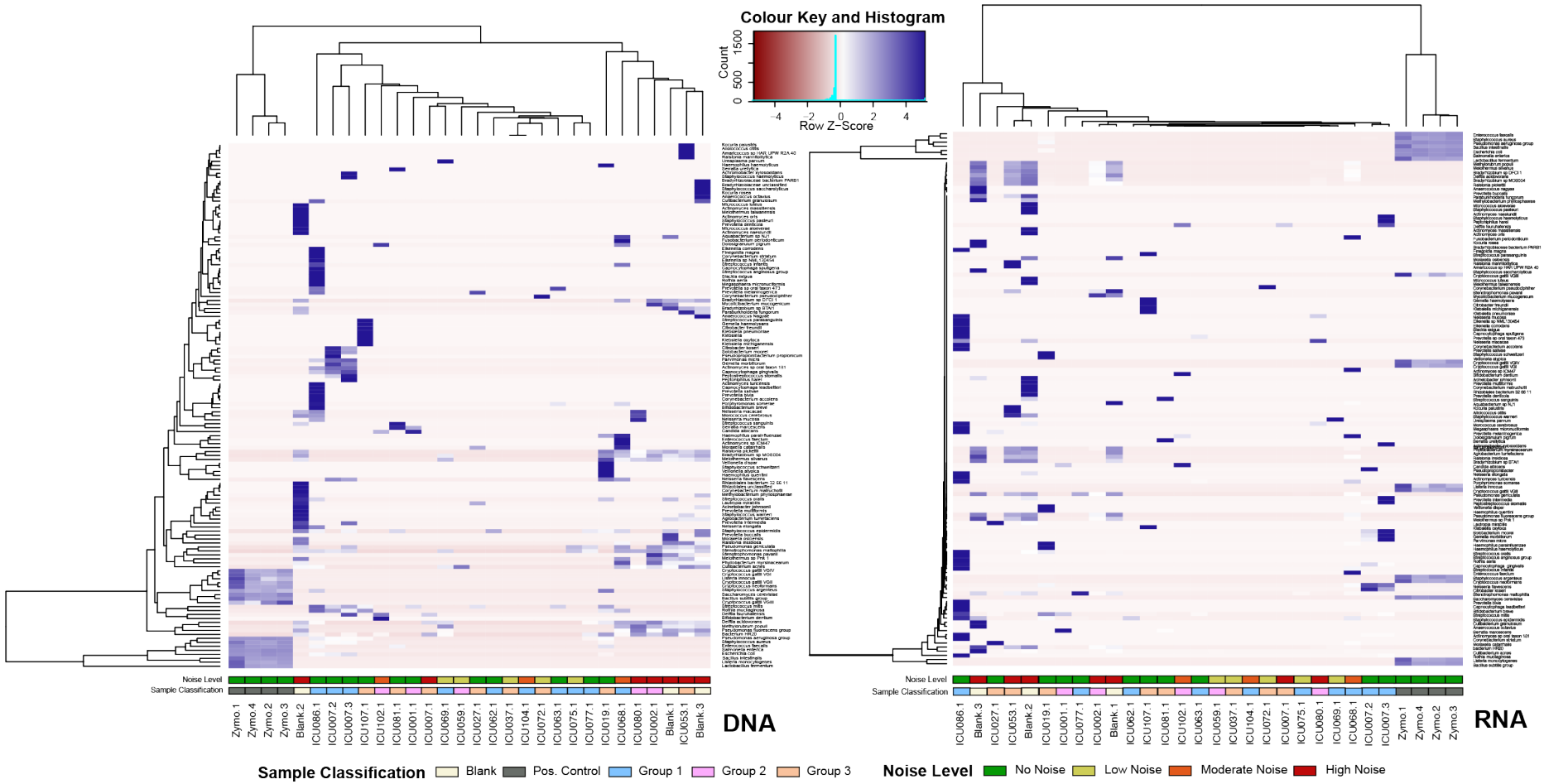


Figure 1: Heatmap of DNA and RNA metagenomic (mNGS) libraries with adjusted z-score ranks and clustering. Here, we compared the predominant microbial taxa for both the DNA (left panel) and RNA (right panel) mNGS libraries using clustered heatmaps of microbial abundance (Z-score). Differences between 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. Study samples are shown along the X-axis with colour coding depending on the sample type or clinical, as well as the level of microbial noise (based on taxa present in blank controls).

Bacteria

Potentially pathogenic bacteria genera including *Staphylococcus, Streptococcus, Stenotrophomonas, Pseudomonas, Klebsiella, Enterococcus* and *Escherichia* were identified in samples at levels above those in the negative controls (Figure 1). *Pseudomonas aeruginosa* was identified in 6 of the 23 patient samples using parallel RNA/DNA NGS (ICU007.2, ICU007.3, ICU019.1, ICU077.1, ICU086.1 and ICU107.1), with average relative abundance of 0.44% (± 0.78%) and 7.73% (± 14.09%) respectively. *P. aeruginosa* had been detected using culture for the patient of sample ICU107.1 on three separate occasions, the first detection was on an eye swab on, followed by sputum on and urine on a couple of weeks later. The upper respiratory sample we sequenced here of ICU107.1 was collected on immediately in the days before the sputum and urine detections and was positive for *P. aeruginosa* by both DNA and RNA sequencing. In contrast, for case ICU080.1, a urine sample returned a positive *P. aeruginosa* culture on the 18/11/2018, however here, the nasopharyngeal sample collected 3 days before had no *P. aeruginosa* detected utilising the DNA/RNA NGS method. Clinical sample ICU080.1 is noted as having high noise and is of low sample quality, this library only had 130391 reads mapped to known clades, suggesting the sequencing depth may not have been sufficient for the identification of all present bacterial species. In other case, potentially linking longitudinally collected specimens, an eye swab was taken from cases ICU081.1 was positive by culture for *Serratia marcescens* and *Enterococcus faecalis*, with our mNGS of the previously negative NPA returning a positive for both organisms, with relative abundances of 59.68% and 4.74% in the DNA library and 2.41% and 0.56% in the RNA library, for *Serratia marcescens* and *Enterococcus faecalis,* respectively. Low quality samples were revealed by microbial communities similar to that of the blank samples. For example, ICU080.1, ICU053.1, ICU002.2 and ICU007.1 were considered very low-quality samples and returned abundance profiles that were nearly identical to the negative controls with predominant taxa including *Bradyrhizobium sp DFCI 1*, *Delftia acidovorans, Methylorubrum populi* and *Pseudomonas fluorescens group.* It is evident that sample quality impacts the ability for DNA/RNA NGS to detect potentially pathogenic bacteria from NPA samples.

Fungi

Relative to the portion of bacterial reads, fungal reads were significantly less prevalent in the study cohort. Using DNA and RNA sequencing potentially pathogenic fungi, *Saccharomyces* and *Cryptococcus*, were identified in the four Zymo mock community samples (Fig. 1). Relative abundance averages across the mock community samples varied between RNA and DNA sequencing methods and between identified fungi; however, was expected at

relative abundance of 2% based on the spiked proportions. DNA metagenomic sequencing detected *Saccharomyces* at 0.44% (± 0.02%) relative abundance and *Cryptococcus* at 0.16% (± 0.03%), whilst RNA metatranscriptomic sequencing detected *Saccharomyces* at 12.44% (± 4.20%) relative abundance and *Cryptococcus* at 0.18% (± 0.13%). This discrepancy highlights differences between quantifying DNA genomes and RNA transcripts in eukaryotic organisms. The Zymo mock community was intended to quantify genomic DNA and makes no guarantees of RNA levels in the cells, therefore the result is mostly in-line with what was expected (0.16-0.44% versus 2%), and our lower yield is mostly likely due to incomplete extraction of the difficult to lyse fungi.

DNA metagenomic shot-gun sequencing identified *Candida albicans* in 2 out of the 23 patient samples, ICU001.1 and ICU102.1, relative abundance was extremely varied with approximate non-human sequence abundance of 0.33% and 55.97%, respectively by metaphlan4 (Fig. 1). RNA metatranscriptomic sequencing identified *C. albicans* in patient sample ICU102.1 with a relative abundance of 0.01%. Patient sample ICU001.1 returned a *C. albicans* positive culture from a swab taken of the patient's central venous line 6 days after the collection date of the nasal phalangeal sample included in this study. The finding of *C. albicans* using culture of a sample from a different location and at a later timepoint demonstrates an agreement in findings between metagenomic sequencing and diagnostic culture techniques. The identification of *C. albicans* in patient sample ICU102.1 was not reflected in the clinical notes, however it is not known if culture was requested for this sample. *C. albicans* is the most common causative agent of mucosal and systemic infections, responsible for approximately 70% of all fungal infections globally and although infections are usually mild, life-threatening invasive infections possess a mortality rate of approximately 40%, especially in hospital settings (Basmaciyan et al., 2019, da Silva Dantas et al., 2016). It is important to note, culture may not have been requested for all patient samples and therefore culture results may not be available for each sample. Where fungicidal or fungistatic medication had been administered to patients prior to sampling, detection of fungal species may be impacted, this data was not available for patients included in this study.

Viruses

Within our study cohort, all respiratory samples included for analysis had tested negative by commercial RT-qPCR using the Allplex Respiratory Full Panel Assay (Vandendriessche et al., 2019). Here, mNGS extended the results with the identification of possible viral pathogens in two separate cases. In one sample - ICU068.1, we detected a moderate number of respiratory

syncytial virus (RSV) reads, and further inspection of the clinical history revealed a positive RSV result by RT-qPCR on a patient's sample three weeks prior. Interestingly, Influenza C virus was identified in the NPA sample ICU075.1, with relatively high abundance of 3% of the total metagenomic reads in the RNA library. Whilst not typically viewed to cause serious disease, influenza C is a neglected pathogen from a human health perspective with fewer than 100 published genomes available on public databases. Furthermore, it is rarely included in commercial respiratory multiplex PCR assays (Cunha et al., 2016). A maximum likelihood tree was constructed for the haemagglutinin gene of the Influenza C virus (**Figure 2**), however the lack of sampling locally and globally, does not allow for meaningful phylodynamic inference. Nonetheless, the usefulness of DNA/RNA NGS is highlighted here, where a clear pathogen is not suspected or where symptoms present in an atypical way, the detection of viral pathogens using conventional methods may be insufficient. Correct and early diagnosis of influenza C virus infection could be treated with appropriate antiviral medication, highlighting the potential yield of metagenomics in case management. Our data demonstrates the meta-transcriptomic exploration of potential infectious microbes and identify the ways it can add value in selected cases despite its caveats. Although there were no suspected herpesvirus pneumonitis cases in this cohort, *Cytomegalovirus, HHV-4* and *HHV7*, was identified in 6 clinical samples. Further incidental findings of viral genera of unknown or unlikely pathogenicity include *Papillomavirus, Torquetenoviruses, and WU Polyomavirus* in ICU086.1, ICU027.1 and ICU068.1 respectively, as well as several bacteriophages likely derived from oral microbes.

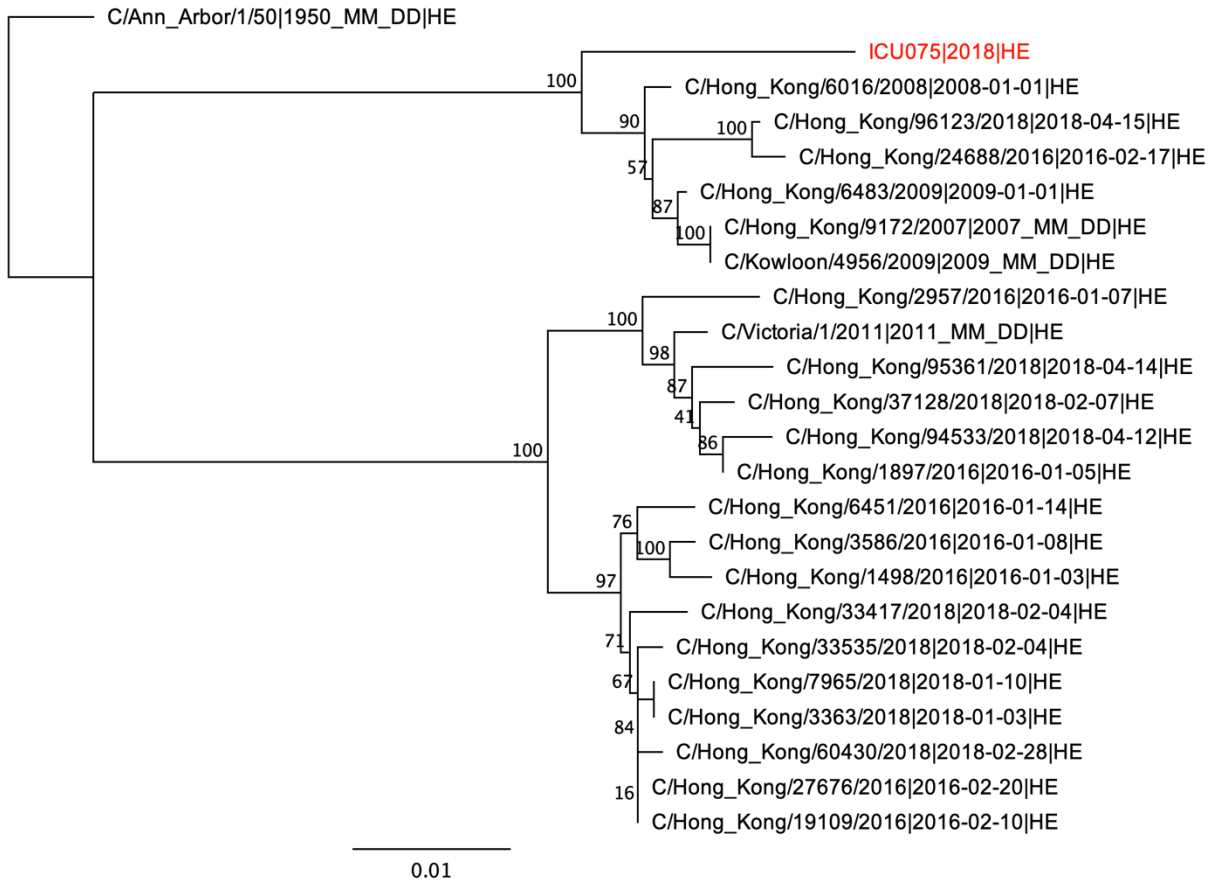


Figure 2: Influenza C virus phylogeny from case ICU075 using haemagglutinin gene sequences. The assembled sequence from case ICU075 corresponding to the viral haemagglutinin (HE) gene was extracted (colored red) and aligned against known reference strains (colored black) sourced from the GISAID database using MAFFT (Rozewicki et al., 2019). A maximum likelihood phylogeny was then estimated using PhyML with bootstrapping support values indicated at nodes. The tree was rooted using the Ann Arbor strain collected in 1950. The branch lengths are proportional the number of substitutions per site.

Microbial DNA as the best measure of metagenome quality of respiratory samples

We sought to define the limitations of our metagenomic pipeline with respect to samples of sub-optimal quality, by introducing quality checks prior to library preparation and correlated these values with sequencing results. Sample quality is a determining factor in the metagenomic performance of a sample. Samples in this study were considered rescued and not collected for the purpose of metagenomic sequencing (Chen et al., 2017, Bal et al., 2018). Typically, residual diagnostic samples are of low quality, a finding reflected in our study. Of the 23 samples sequenced in this study, 7 had high background noise determined by the resemblance of the sample microbiome to the blank or kit microbiome, a further five samples showed some level of noise in the generated microbiome. Samples with high background noise were associated with a distinct lack of infectious agent detection, with much of the metagenomic library being uninterpretable. High noise was a marker for poor sample quality

and was a critical factor in the capturing of an unbiased sample infectome. We determined metagenomics noise to be an objective measurement whereby the genomic data appeared to be skewed towards the kit microbiome, indicative of insufficient sample biomass to capture an unbiased snapshot of the true infectome (Luhung et al., 2021). To ensure samples yield unbiased high quality metagenomic libraries, samples to be taken forward into a metagenomics testing pathway need to be collected and stored appropriately as soon as possible after sampling. (Hasan et al., 2020).

Parametric correlation analysis DNA was performed to determine if there was a meaningful relationship between input or total DNA, measured using fluorometry, and human, bacterial and fungal DNA by qPCR (Sedgwick, 2012). Quantification data obtained for study samples were analysed using the Pearson's correlation coefficient (**Table 2**). The relationship between input DNA and Human DNA was found to be statistically significant (r (23) = .72, p < .001), and positively correlated (**Figure 3**). Human DNA is highly correlated with overall sample quantification as human DNA appears to provide the bulk of the sample biomass. The proportion of human DNA in a sample can vary, and likely for a number of reasons including sample type and technique, with nasal swabs often resulting in a large amount of incidental human DNA being collected, due to epithelial cell contamination. A positive correlation was also identified between input DNA and bacterial DNA, this relationship was found to be statistically significant at the 95% confidence interval (r (23) = .39 p = 0.05) (**Figure 3**). Bacterial DNA quantification using qPCR appeared to be the best early indicator of mNGS performance for respiratory samples. Where the CT value of the bacterial PCR was above cycle 33/34, libraries were clearly lower quality and had a greater amount of noise from reagent-associated taxa. An inverse correlation was identified between the input DNA and the total fungal DNA, (Fig. 3) this relationship was found to not be statistically significant (r (23) = -.08 p = 0.69), very few libraries in this study contained fungal DNA, as such we can infer fungal DNA had very little influence in total sample biomass. Where samples did contain fungal DNA, those with a lower CT typically corresponded well with the mNGS data. For example, the libraries with fungi identified – ICU001.1 and ICU102.1 – had CT values of 32 and 31, respectively. Metagenome data is richly diverse in non-human content, whilst we have identified human DNA is a good indicator of overall sample biomass, our data suggest fluorometry may provide insight into library preparation but is not sufficient in providing an indication of mNGS performance (signal > noise), and as such targeted quantification of bacterial or fungal DNA may be a more reliable indicator of sample performance, and may be considered for batched mNGS runs in a clinical setting.

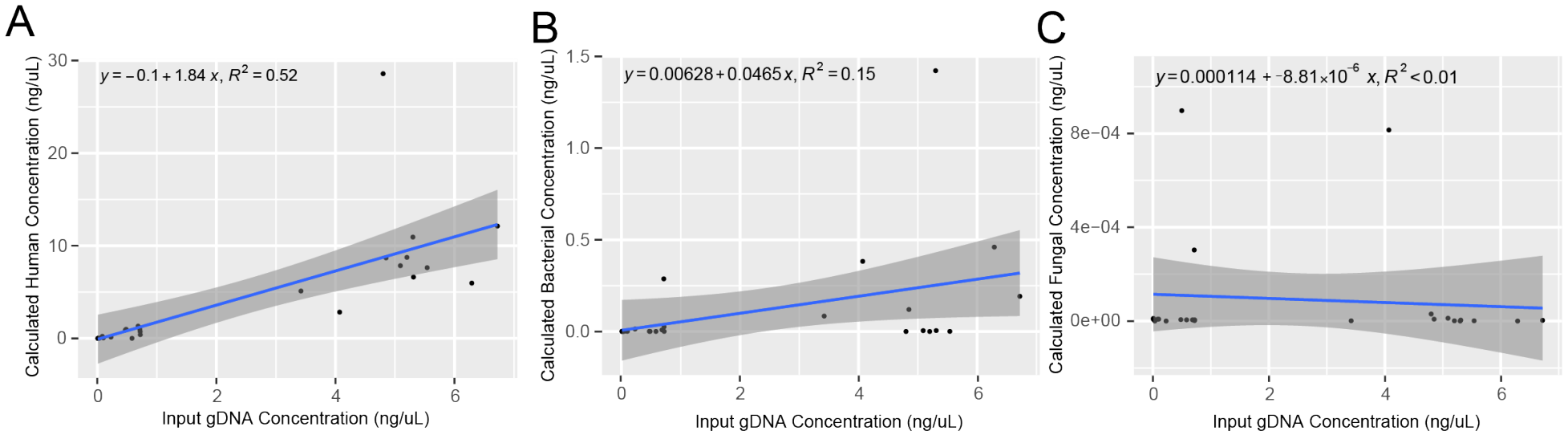


Figure 3: Correlation plots between total gDNA and source DNA. The total gDNA of each sample was measured using the Qubit HS ds-DNA assay, and then compared to quantitative measurements of human (A), bacterial (B) and fungal (C) DNA assayed using qPCR. The correlations between measurements were calculated using the Pearson correlation coefficient with equations and R^2 value shown.

Table 2: Pearson's correlation test of total gDNA and source DNA

Source DNA	T-test statistic	df	p-value	Conf.int (correlation coefficient at 95%)	Sample estimate (correlation coefficient)
Human	5.0311	23	0.000043	[0.4603623, 0.8701027]	0.723827
Bacterial	2.0245	23	0.04469	[-0.007364723, 0.679597970]	0.388898
Fungal	-0.40177	23	0.6916	[-0.4633290, 0.3222801]	-0.083482

Samples clustered by metagenome quality rather than clinical groups

In this study, the cases were sorted into three different categories to reflect clinical likelihood of an infectious agent as a cause for disease. We hypothesised that case category may have an impact on the distribution of organisms within the individual metagenomic library. We sort to identify any meaningful correlations between sample classification and infectome, with the sample set available in this study. A hierarchical clustering heatmap of all patient and control samples was constructed, metatranscriptomic and metagenomic data was plotted separately for comparison (**Figure 1**). Organisms identified in the samples were grouped along the y-axis and samples were grouped on the x-axis, the row Z-Score scaling method was utilised for visualisation, differences between samples could be identified by a deeper blue shading, whilst organisms conserved across samples were lighter blue through to pink. For comparison, we also plotted the raw relative abundances (reads per 1000 reads) as unclustered heatmaps (**Supp. Figure 2 & 3**).

The clustering dendrogram allowed immediate identification of conserved organisms across mock community controls and blank samples in both DNA and RNA sequencing approaches. Clustering was also observed around low biomass and blank samples for the DNA libraries, which seemed to have the best-preserved signal for microbial taxa versus the RNA libraries. This finding suggests that in instances where biomass is insufficient to create a library, the kit microbiome is instead captured and tends to skew the sample metagenome towards the kit noise. Low noise, higher biomass samples appeared to have unique microbiomes with unique organism clusters in varying abundances and distributions. Some organisms, such as *Streptococcus mitis* were conserved across multiple upper respiratory tract samples, which may be expected given the predominance of upper respiratory samples. Other typical oral taxa were well represented across several samples, whilst samples with an overabundance of human DNA often shared the presence of the four major taxa normally present on the human skin *Cutibacterium*, *Staphylococcus*, *Streptococcus*, and *Corynebacterium* (Baker et al., 2021). It is important to consider sample size and sample quality when analysing these results, of the available samples in this study 15 out of 23 were impacted by less-than-optimal quality. There was also an unequal number of samples in each category further influencing the clustering of samples within the heatmap dendrogram.

Timing of sampling is critical for acute diagnosis and sample quality

The use of antibiotics and antibiotic exposure has been shown to affect distribution of bacterial communities within airways, collecting samples prior to the administration of broad-spectrum antibiotics is critical in ensuring an unbiased microbiome is captured (Serpa et al., 2022). Microbial diversity can be decreased, and certain types of antibiotics may influence the presence of dominant pathogenic bacteria, for example, *P. aeruginosa* are resistant to many different classes of antibiotics including aminoglycosides, quinolones, and β-lactams. In this study only 4 of the 6 positive *P. aeruginosa* samples were collected from patients who had been treated by one or more different antibiotics, common antibiotics used between patients include cephazolin, gentamicin, ceftazidime and amoxycillin-clavulanate, whilst ICU019.1 and ICU077.1 had not been treated with antibiotics at the time of sample collection. The effect of antimicrobial administration can cause a skew in the representation of species of bacteria resistant to antibiotic use, where no prior pathogen has been identified care must be taken to not identify an over-abundance of antibiotic-resistant bacteria as a causative pathogen.

Longitudinal samples ICU007.1-ICU007.3 were collected over a 23-day period, the first sample ICU007.1 was collected 2 weeks after admission, and *P. aeruginosa* was not detected (although of low-sample quality), ICU007.2 and ICU007.3, collected 4 and 19 days respectively after the collection of ICU007.1 returned positive results for *P. aeruginosa* with relative library abundance of <2% for both samples. Positive culture results were returned for *Citrobacter koseri* and *Stenotrophomonas maltophilia* were reported for a sputum culture collected 4 days prior to the collection of ICU007.1, in ICU007.1 we identified these 2 pathogens as well as other predominant taxa such as *Neisseria flavescens; Gemella morbillorum & Parvimonas micra*. These findings suggest there may be merit to the use of mNGS to extend the breadth of results to other potentially pathogenetic species when compared to culture-based methods, however the identification of a causative pathogen remains a complicated task. During the patients first admission they were administered gentamicin, ceftazidime and amoxycillin-clavulanate. In a groin swab taken 6 days before the collection of ICU007.2, methicillin-resistant *Staphylococcus aureus* (MRSA) was identified using culture, MRSA was identified using mNGS in ICU007.2 at a relative abundance of 11.5%. Whilst not explicitly explored in this study, there exists a potential for microbial mNGS data to be utilised in the identification of antimicrobial resistant microbes both in real time and retrospective studies.

The time of sample collection can have a considerable impact of the usefulness of resultant mNGS data. Where samples are collected at the acute onset of symptoms, or during the early admission of a patient, we can infer information about the source of the infection, for example, ICU075.1 was positive for influenza C, the NPA was taken 2 days after admission (**Table 1**). The time of sample collection with respect to hospital admission date and symptom onset suggests that the infection was likely community acquired. In situations where sample collection occurs early during symptom onset and hospital admission the nature of the patient microbiome remains relevant. In this study, on average samples were collected from patients on day 12 of hospital admission, with some taken as late as 37 days of admission or as early as the day of admission. A plethora of studies have suggested early intervention with metagenomic sequencing may be useful in informing treatment decisions (Gosiewski et al., 2017, Schulz et al., 2022, Hogan et al., 2021). Where mNGS is likely to be employed in a diagnostic capacity, it is a recommendation of this study that early and deliberate sample collection and storage is necessary. Whilst we have demonstrated the capacity of this method on rescued residual extracts, mNGS is best performed on samples collected and stored for the purpose of metagenomics, with sample quality dependent on appropriate collection, storage and preparation of total nucleic acid samples (Yuan et al., 2022, Poulsen et al., 2022, Uwamino et al., 2021). Therefore, the greatest diagnostic value would be to perform mNGS upfront or reserve samples stored appropriately for later mNGS screening, rather than relying on degraded, rescued samples, only considered for mNGS when all other avenues are exhausted. There is also potential added value of mNGS upfront for precision medicine and the potential detection of treatment resistant organisms that are harboured by individual patients.

Conclusions

In summary, here we present a comprehensive mNGS study utilising the parallel sequence of DNA and RNA across a range of clinical respiratory samples collected from severe, acute respiratory infections. This study demonstrated the utility of mNGS in identifying diverse bacterial, fungal and viral pulmonary species in 23 rescued clinical samples. The capacity for mNGS technology to identify microbial diversity in complex hospitalised patients as risk of opportunistic infection has been demonstrated by the characterisation of potentially pathogenic bacterial and fungal species in over abundance. We also identify a previously undiagnosed influenza C infection and provide proof of principle for the use of mNGS in the detection of otherwise unknown or novel pathogens. It is a recommendation of this study that the use of mNGS in pathogen detection could brought forward to inform diagnosis earlier in hospitalisation, and preferably prior to antibiotic exposure. We acknowledge that further

research must be conducted to facilitate the use of mNGS as a clinical diagnostic tool and envisage movement towards further translation in pathology labs.

Author Contributions

Conceived the project- J.S.E & P.N.B; collected samples- R.B & P.N.B.; performed laboratory work- R.L.T. & J.S.E.; analysed the data- R.L.T., R.B., J.S.E & P.N.B.; wrote the paper- R.L.T., R.B., J.S.E & P.N.B. All authors agree to be accountable for all aspects of the work and read and approved the final manuscript.

Acknowledgements

The authors acknowledge the generous support and work of the lab and pathologist staff at the Sydney Children's Hospital Microbiology at Westmead, and protocol support from A/Prof Jason Kwong at Austin Health

.

Funding

Funding was provided by the National Health and Medical Research Council Centre of Research Excellence in Emerging Infectious Diseases (#1102962), and philanthropic support through the University of Sydney - Sydney Infectious Diseases Institute.

Competing interests

The authors declare that they have no competing interests.

References

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., JACKSON, A. E., SINGANALLUR, N. B., PLAIN, K. M., KIM, K., TACHEDJIAN, M., VAN DER HEIDE, B., CRAMERI, S., WILLIAMS, D. T., SECOMBE, C., LAING, E. D., STERLING, S., YAN, L., JACKSON, L., JONES, C., PLOWRIGHT, R. K., PEEL, A. J., BREED, A. C., DIALLO, I., DHAND, N. K., BRITTON, P. N., BRODER, C. C., SMITH, I. & EDEN, J.-S. 2022. Novel Hendra Virus Variant Detected by Sentinel Surveillance of Horses in Australia. *Emerging infectious diseases,* 28**,** 693-704.

BAKER, J. L., MORTON, J. T., DINIS, M., ALVAREZ, R., TRAN, N. C., KNIGHT, R. & EDLUND, A. 2021. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome research,* 31**,** 64-74.

BAL, A., PICHON, M., PICARD, C., CASALEGNO, J. S., VALETTE, M., SCHUFFENECKER, I., BILLARD, L., VALLET, S., VILCHEZ, G., CHEYNET, V., ORIOL, G., TROUILLET-ASSANT, S., GILLET, Y., LINA, B., BRENGEL-PESCE, K., MORFIN, F. & JOSSET, L. 2018. 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,* 18**,** 537-537.

BASMACIYAN, L., BON, F., PARADIS, T., LAPAQUETTE, P. & DALLE, F. 2019. "Candida Albicans Interactions With The Host: Crossing The Intestinal Epithelial Barrier". *Tissue barriers,* 7**,** 1612661-1612661.

BINNS, E., KOENRAADS, M., HRISTEVA, L., FLAMANT, A., BAIER-GRABNER, S., LOI, M., LEMPAINEN, J., OSTERHELD, E., RAMLY, B., CHAKAKALA-CHAZIYA, J., ENAGANTHI, N., SIMÓ NEBOT, S. & BUONSENSO, D. 2022. Influenza and respiratory syncytial virus during the COVID-19 pandemic: Time for a new paradigm? *Pediatric pulmonology,* 57**,** 38-42.

BLANCO-MIGUEZ, A., BEGHINI, F., CUMBO, F., MCIVER, L. J., THOMPSON, K. N., MORENO, Z., MANGHI, P., DUBOIS, L., HUANG, K. D., THOMAS, A. M., PICCINNO, G., PIPERNI, E., PUNČOCHÁŘ, M., VALLES-COLOMER, M., TETT, A., GIORDANO, F., DAVIES, R., WOLF, J., BERRY, S. E., SPECTOR, T. D., FRANZOSA, E. A., PASOLLI, E., ASNICAR, F., HUTTENHOWER, C. & SEGATA, N. 2022. Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

BRENNER, T., DECKER, S. O., GRUMAZ, S., STEVENS, P., BRUCKNER, T., SCHMOCH, T., PLETZ, M. W., BRACHT, H., HOFER, S., MARX, G., WEIGAND, M. A. & SOHN, K. 2018. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine (Baltimore),* 97**,** e9868-e9868.

BROWN, J., PIRRUNG, M. & MCCUE, L. A. 2017. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics,* 33**,** 3137-3139.

BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods,* 12**,** 59-60.

BUSHNELL 2016. BBMap short-read aligner, and other bioinformatics tools.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC bioinformatics,* 10**,** 421-421.

CAPORASO, J. G., LAUBER, C. L., WALTERS, W. A., BERG-LYONS, D., HUNTLEY, J., FIERER, N., OWENS, S. M., BETLEY, J., FRASER, L., BAUER, M., GORMLEY, N., GILBERT, J. A., SMITH, G. & KNIGHT, R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal,* 6**,** 1621-1624.

CHEN, Z., COLLIN, E., PEDDIREDDI, L., CLEMENT, T., GAUGER, P. & HAUSE, B. M. 2017. Genetic diversity in envelope genes of contemporary U.S. porcine reproductive and respiratory syndrome virus strains influences viral antigenicity. *Research in veterinary science,* 115**,** 432-441.

CUNHA, B. A., IRSHAD, N. & CONNOLLY, J. J. 2016. Adult human metapneumonovirus (hMPV) pneumonia mimicking Legionnaire's disease. *Heart & Lung - The Journal of Acute and Critical Care,* 45.

DA SILVA DANTAS, A., LEE, K. K., RAZIUNAITE, I., SCHAEFER, K., WAGENER, J., YADAV, B. & GOW, N. A. R. 2016. Cell biology of Candida albicans –host interactions. *Current opinion in microbiology,* 34**,** 111-118.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

FERNANDEZ-MARMIESSE, A., GOUVEIA, S. & COUCE, M. L. 2018. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Current medicinal chemistry,* 25**,** 404-432.

FRICKE, L. M., GLÖCKNER, S., DREIER, M. & LANGE, B. 2021. Impact of non-pharmaceutical interventions targeted at COVID-19 pandemic on influenza burden – a systematic review. *The Journal of infection,* 82**,** 1-35.

GÓMEZ-RUBIO, V. 2017. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *Journal of statistical software,* 77**,** 1-3.

GOSIEWSKI, T., LUDWIG-GALEZOWSKA, A. H., HUMINSKA, K., SROKA-OLEKSIAK, A., RADKOWSKI, P., SALAMON, D., WOJCIECHOWICZ, J., KUS-SLOWINSKA, M., BULANDA, M. & WOLKOW, P. P. 2017. Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method - the observation of DNAemia. *European journal of clinical microbiology & infectious diseases,* 36**,** 329-336.

GRENINGER, A. L. 2018. The challenge of diagnostic metagenomics. *Expert review of molecular diagnostics,* 18**,** 605-615.

GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic biology,* 59**,** 307-321.

HASAN, M. R., SUNDARARAJU, S., TANG, P., TSUI, K.-M., LOPEZ, A. P., JANAHI, M., TAN, R. & TILLEY, P. 2020. A metagenomics-based diagnostic approach for central nervous system infections in hospital acute care setting. *Scientific reports,* 10**,** 11194-11194.

HOGAN, C. A., YANG, S., GARNER, O. B., GREEN, D. A., GOMEZ, C. A., DIEN BARD, J., PINSKY, B. A. & BANAEI, N. 2021. Clinical Impact of Metagenomic Next-Generation Sequencing of Plasma Cell-Free DNA for the Diagnosis of Infectious Diseases: A Multicenter Retrospective Cohort Study. *Clinical infectious diseases,* 72**,** 239-245.

JAIN, S., SELF, W. H., WUNDERINK, R. G., FAKHRAN, S., BALK, R., BRAMLEY, A. M., REED, C., GRIJALVA, C. G., ANDERSON, E. J., COURTNEY, D. M., CHAPPELL, J. D., QI, C., HART, E. M., CARROLL, F., TRABUE, C., DONNELLY, H. K., WILLIAMS, D. J., ZHU, Y., ARNOLD, S. R., AMPOFO, K., WATERER, G. W., LEVINE, M., LINDSTROM, S., WINCHELL, J. M., KATZ, J. M., ERDMAN, D., SCHNEIDER, E., HICKS, L. A., MCCULLERS, J. A., PAVIA, A. T., EDWARDS, K. M. & FINELLI, L. 2015. Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *The New England journal of medicine,* 373**,** 415-427.

JAWORSKI, E., LANGSJOEN, R. M., MITCHELL, B., JUDY, B., NEWMAN, P., PLANTE, J. A., PLANTE, K. S., MILLER, A. L., ZHOU, Y., SWETNAM, D., SOTCHEFF, S., MORRIS, V., SAADA, N., MACHADO, R. R., MCCONNELL, A., WIDEN, S. G., THOMPSON, J., DONG, J., REN, P., PYLES, R. B., KSIAZEK, T. G., MENACHERY, V. D., WEAVER, S. C. & ROUTH, A. L. 2021. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *eLife,* 10.

JOHN, G., SAHAJPAL, N. S., MONDAL, A. K., ANANTH, S., WILLIAMS, C., CHAUBEY, A., ROJIANI, A. M. & KOLHE, R. 2021. Next-Generation Sequencing (NGS) in COVID-19: A Tool for SARS-CoV-2 Diagnosis, Monitoring New Strains and Phylodynamic Modeling in Molecular Epidemiology. *Current issues in molecular biology,* 43**,** 845-867.

KIM, H.-M., JEON, S., CHUNG, O., JUN, J. H., KIM, H.-S., BLAZYTE, A., LEE, H.-Y., YU, Y., CHO, Y. S., BOLSER, D. M. & BHAK, J. 2021. Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *Gigascience,* 10.

KOPYLOVA, E., NOE, L. & TOUZET, H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics,* 28**,** 3211-3217.

LI, D., LIU, C.-M., LUO, R., SADAKANE, K. & LAM, T.-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

LIU, L. P., OZA, S. M., HOGAN, D. P., PERIN, J. P., RUDAN, I. P., LAWN, J. E. P., COUSENS, S. P., MATHERS, C. P. & BLACK, R. E. P. 2015. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet (British edition),* 385**,** 430-440.

LUHUNG, I., UCHIDA, A., LIM, S. B. Y., GAULTIER, N. E., KEE, C., LAU, K. J. X., GUSAREVA, E. S., HEINLE, C. E., WONG, A., PREMKRISHNAN, B. N. V., PURBOJATI, R. W., ACERBI, E., KIM, H. L., JUNQUEIRA, A. C. M., LONGFORD, S., LOHAR, S. R., YAP, Z. H., PANICKER, D., KOH, Y., KUSHWAHA, K. K., ANG, P. N., PUTRA, A., DRAUTZ-MOSES, D. I. & SCHUSTER, S. C. 2021. Experimental parameters defining ultra-low biomass bioaerosol analysis. *NPJ biofilms and microbiomes,* 7**,** 37-37.

PENDLETON, K. M., ERB-DOWNWARD, J. R., BAO, Y., BRANTON, W. R., FALKOWSKI, N. R., NEWTON, D. W., HUFFNAGLE, G. B. & DICKSON, R. P. 2017. Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. *American journal of respiratory and critical care medicine,* 196**,** 1610-1612.

PFEFFERLE, S., GÜNTHER, T., KOBBE, R., CZECH-SIOLI, M., NÖRZ, D., SANTER, R., OH, J., KLUGE, S., OESTEREICH, L., PELDSCHUS, K., INDENBIRKEN, D., HUANG, J., GRUNDHOFF, A., AEPFELBACHER, M., KNOBLOCH, J. K., LÜTGEHETMANN, M. & FISCHER, N. 2021. SARS Coronavirus-2 variant tracing within the first Coronavirus Disease 19 clusters in northern Germany. *Clinical microbiology and infection,* 27**,** 130.e5-130.e8.

POULSEN, C. S., EKSTRØM, C. T., AARESTRUP, F. M. & PAMP, S. J. 2022. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology spectrum,* 10**,** e0009022-e0009022.

QUER, J., COLOMER-CASTELL, S., CAMPOS, C., ANDRÉS, C., PIÑANA, M., CORTESE, M. F., GONZÁLEZ-SÁNCHEZ, A., GARCIA-CEHIC, D., IBÁÑEZ, M., PUMAROLA, T., RODRÍGUEZ-FRÍAS, F., ANTÓN, A. & TABERNERO, D. 2022. Next-Generation Sequencing for Confronting Virus Pandemics. *Viruses,* 14**,** 600.

R DEVELOPMENT CORE TEAM 2010. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statisitical Computing

ROZEWICKI, J., LI, S., AMADA, K. M., STANDLEY, D. M. & KATOH, K. 2019. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic acids research,* 47**,** W5-W10.

SCHULZ, E., GRUMAZ, S., HATZL, S., GORNICEC, M., VALENTIN, T., HUBER-KRASSNITZE, B., KRIEGL, L., UHL, B., DEUTSCH, A., GREINIX, H., KRAUSE, R. & NEUMEISTER, P. 2022. Pathogen Detection by Metagenomic Next-Generation Sequencing During Neutropenic Fever in Patients With Hematological Malignancies. *Open forum infectious diseases,* 9**,** ofac393-ofac393.

SEDGWICK, P. 2012. Pearson's correlation coefficient. *BMJ (Online),* 345**,** e4483-e4483.

SERPA, P. H., DENG, X., ABDELGHANY, M., CRAWFORD, E., MALCOLM, K., CALDERA, S., FUNG, M., MCGEEVER, A., KALANTAR, K. L., LYDEN, A., GHALE, R., DEISS, T., NEFF, N., MILLER, S. A., DOERNBERG, S. B., CHIU, C. Y., DERISI, J. L., CALFEE, C. S. & LANGELIER, C. R. 2022. Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. *Genome medicine,* 14**,** 1-74.

SIKAZWE, C., NEAVE, M. J., MICHIE, A., MILETO, P., WANG, J., COOPER, N., LEVY, A., IMRIE, A., BAIRD, R. W., CURRIE, B. J., SPEERS, D., MACKENZIE, J. S., SMITH, D. W. & WILLIAMS, D. T. 2022. Molecular detection and characterisation of the first Japanese encephalitis

virus belonging to genotype IV acquired in Australia. *PLoS neglected tropical diseases,* 16**,** e0010754-e0010754.

TAZINYA, A. A., HALLE-EKANE, G. E., MBUAGBAW, L. T., ABANDA, M., ATASHILI, J. & OBAMA, M. T. 2018. Risk factors for acute respiratory infections in children under five years attending the Bamenda Regional Hospital in Cameroon. *BMC pulmonary medicine,* 18**,** 7-7.

TEAM, R. 2020. RStudio: Integrated Development for R. RStudio. *PBC, Boston, MA,* URL http://www.rstudio.com/.

UWAMINO, Y., NAGATA, M., AOKI, W., NAKAGAWA, T., INOSE, R., YOKOTA, H., FURUSAWA, Y., SAKAI-TAGAWA, Y., IWATSUKI-HORIMOTO, K., KAWAOKA, Y., HASEGAWA, N. & MURATA, M. 2021. Accuracy of rapid antigen detection test for nasopharyngeal swab specimens and saliva samples in comparison with RT-PCR and viral culture for SARS-CoV-2 detection. *Journal of infection and chemotherapy : official journal of the Japan Society of Chemotherapy,* 27**,** 1058-1062.

VANDENDRIESSCHE, S., PADALKO, E., WOLLANTS, E., VERFAILLIE, C., VERHASSELT, B. & COOREVITS, L. 2019. Evaluation of the Seegene Allplex™ Respiratory Panel for diagnosis of acute respiratory tract infections. *Acta clinica belgica (English ed. Online),* 74**,** 379-385.

VAUGHN, V. M., FLANDERS, S. A., SNYDER, A., CONLON, A., ROGERS, M. A. M., MALANI, A. N., MCLAUGHLIN, E., BLOEMERS, S., SRINIVASAN, A., NAGEL, J., KAATZ, S., OSTERHOLZER, D., THYAGARAJAN, R., HSAIKY, L., CHOPRA, V. & GANDHI, T. N. 2019. Excess Antibiotic Treatment Duration and Adverse Events in Patients Hospitalized With Pneumonia: A Multihospital Cohort Study. *Annals of internal medicine,* 171**,** 153-163.

WILSON, M. R., SAMPLE, H. A., ZORN, K. C., AREVALO, S., YU, G., NEUHAUS, J., FEDERMAN, S., STRYKE, D., BRIGGS, B., LANGELIER, C., BERGER, A., DOUGLAS, V., JOSEPHSON, S. A., CHOW, F. C., FULTON, B. D., DERISI, J. L., GELFAND, J. M., NACCACHE, S. N., BENDER, J., DIEN BARD, J., MURKEY, J., CARLSON, M., VESPA, P. M., VIJAYAN, T., ALLYN, P. R., CAMPEAU, S., HUMPHRIES, R. M., KLAUSNER, J. D., GANZON, C. D., MEMAR, F., OCAMPO, N. A., ZIMMERMANN, L. L., COHEN, S. H., POLAGE, C. R., DEBIASI, R. L., HALLER, B., DALLAS, R., MARON, G., HAYDEN, R., MESSACAR, K., DOMINGUEZ, S. R., MILLER, S. & CHIU, C. Y. 2019. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *The New England journal of medicine,* 380**,** 2327-2340.

YANG, L., HAIDAR, G., ZIA, H., NETTLES, R., QIN, S., WANG, X., SHAH, F., RAPPORT, S. F., CHARALAMPOUS, T., METHÉ, B., FITCH, A., MORRIS, A., MCVERRY, B. J., O'GRADY, J. & KITSIOS, G. D. 2019. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respiratory research,* 20**,** 265-12.

YANG, Q., XIAO, X., GU, X., LIANG, D., CAO, T., MOU, J., HUANG, C., CHEN, L. & LIU, J. 2021. Surveillance of common respiratory infections during the COVID-19 pandemic demonstrates the preventive efficacy of non-pharmaceutical interventions. *International journal of infectious diseases,* 105**,** 442-447.

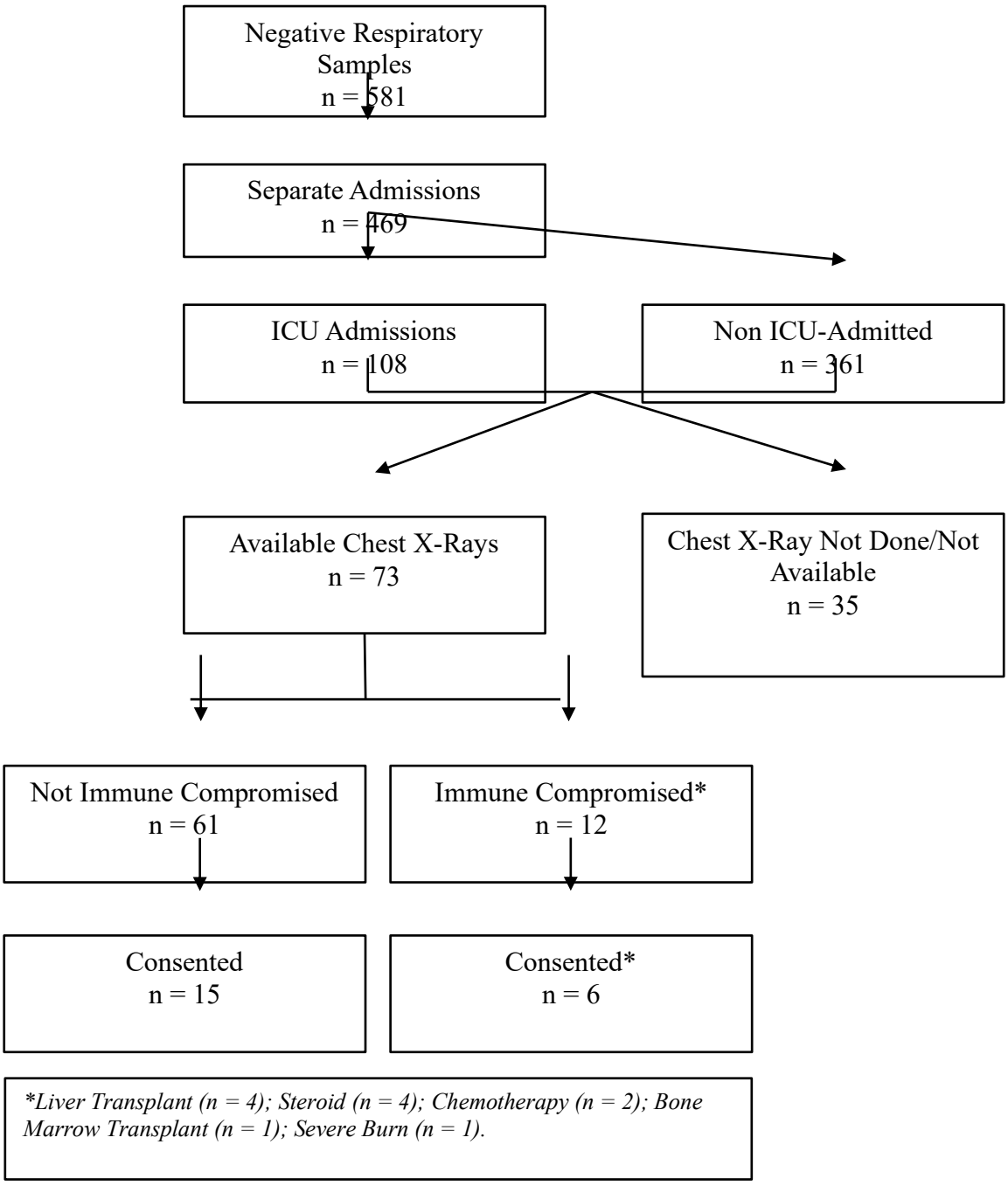
YUAN, G., WANG, H., ZHAO, Y., MAO, E., LI, M., WANG, R., ZHOU, F., JIN, S., ZHANG, Z., XU, K., XU, J., LIANG, S., LI, X., JIANG, L., ZHANG, L., SONG, J., YANG, T., GUO, J., ZHANG, H., ZHOU, Y., WANG, S., QIU, C., JIANG, N., AI, J., WU, J. & ZHANG, W. 2022. Early identification and severity prediction of acute respiratory infection : a study protocol for a randomized controlled trial. *BMC infectious diseases,* 22**,** 1-632.

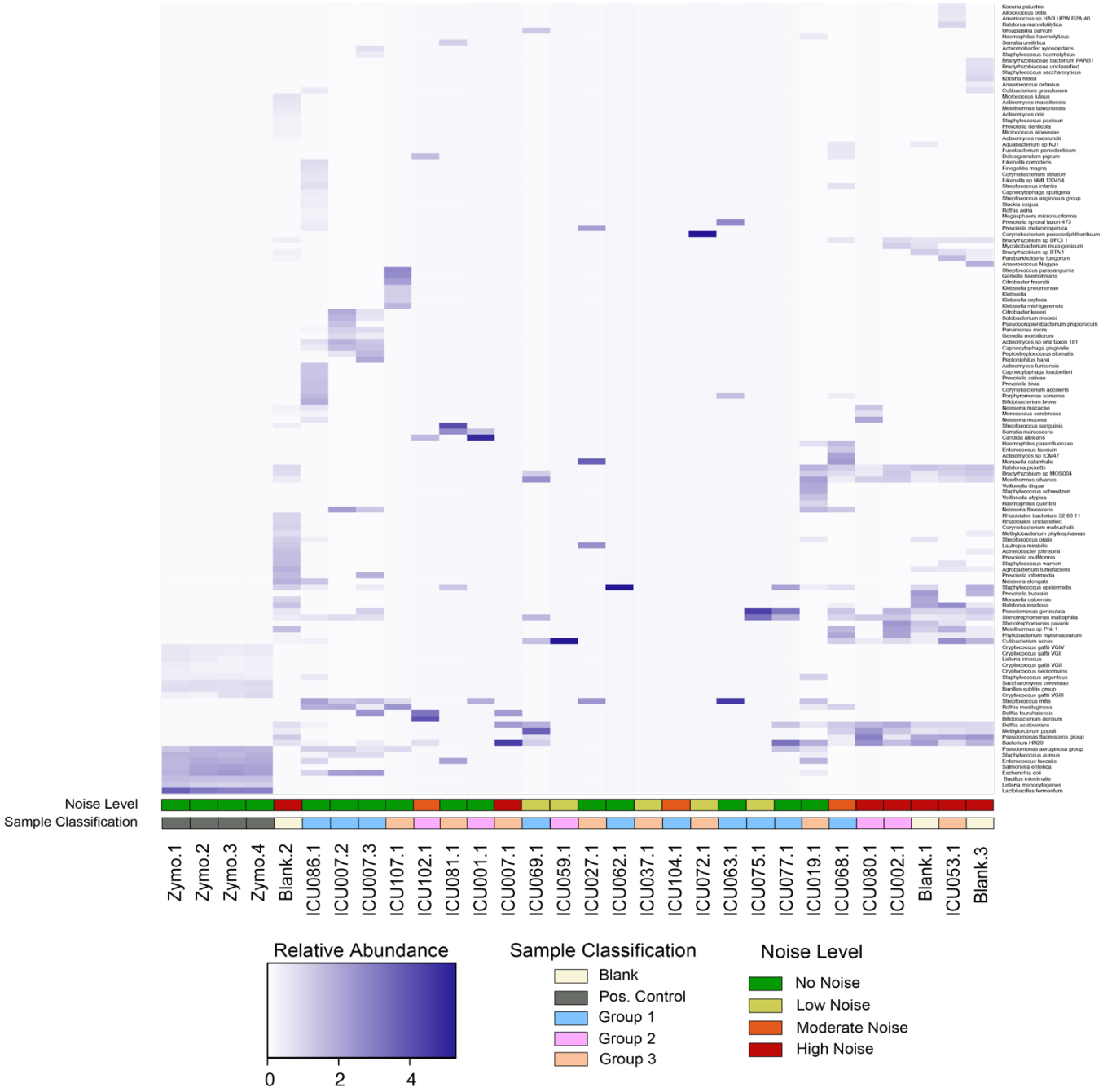
ZHANG, D., LOU, X., YAN, H., PAN, J., MAO, H., TANG, H., SHU, Y., ZHAO, Y., LIU, L., LI, J., CHEN, J., ZHANG, Y. & MA, X. 2018. Metagenomic analysis of viral nucleic acid extraction methods in respiratory clinical samples. *BMC genomics,* 19**,** 773-773.

ZUMLA, A. P., AL-TAWFIQ, J. A. M. D., ENNE, V. I. P., KIDD, M. P., DROSTEN, C. P., BREUER, J. P., MULLER, M. A. P., HUI, D. P., MAEURER, M. P., BATES, M. P., MWABA, P. P., AL-HAKEEM, R. M. D., GRAY, G. P., GAUTRET, P. P., AL-RABEEAH, A. A. F., MEMISH, Z. A. P. & GANT, V. F. 2014. Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections—needs, advances, and future prospects. *The Lancet infectious diseases,* 14**,** 1123-1135.

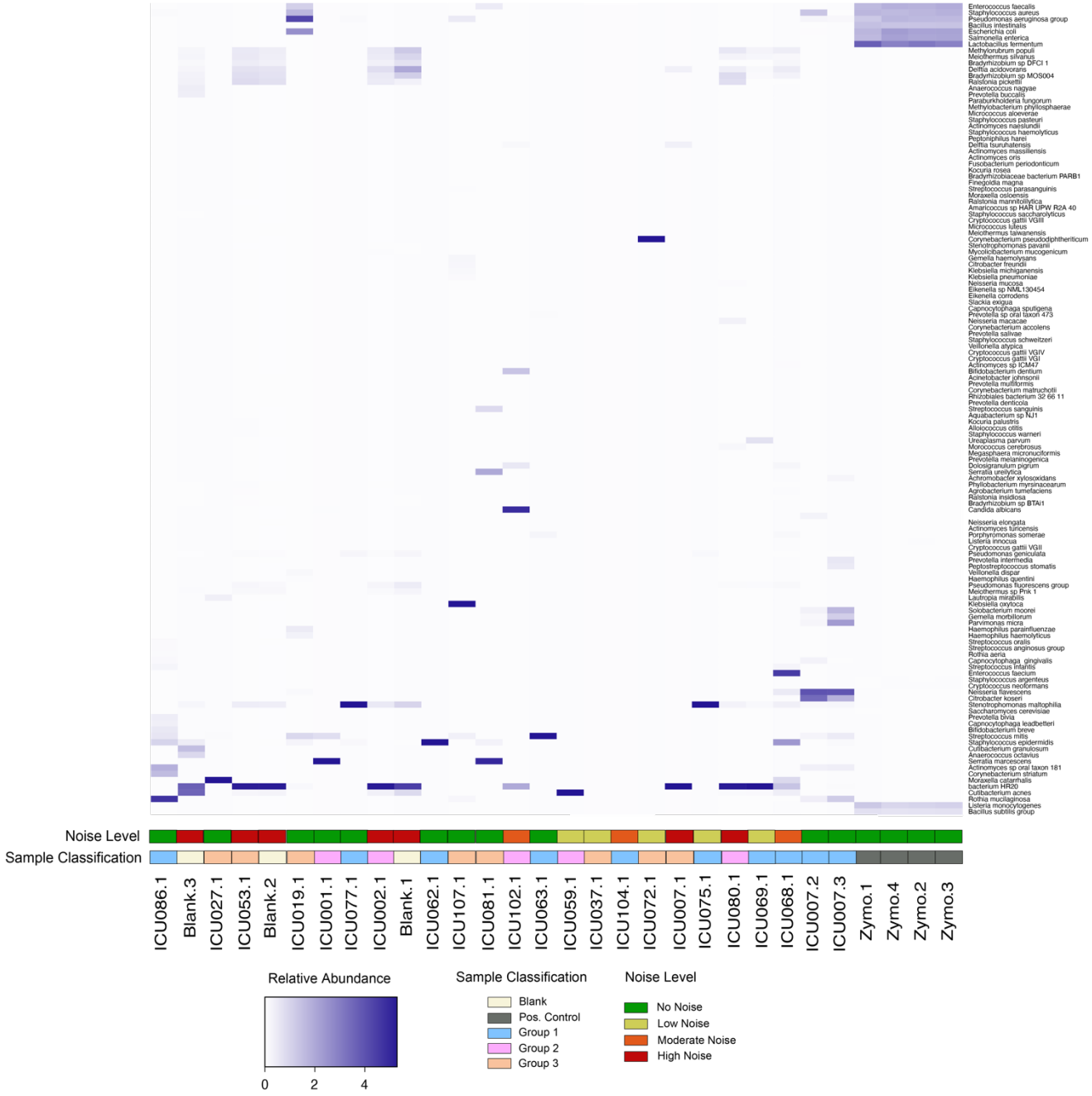
Supplementary Information

Supplementary Figure 1: Cohort case flow chart. A flow chart beginning with total number of samples screened, through to the number of ICU cases identified, with chest X-rays available, and immune compromised.





Supplementary Figure 2: DNA abundance heatmap. Here we present an unclustered heatmap which displays the relative abundance of each organism. A colour key shows a change from white to blue, and is measured in relative abundance %. Samples are further classified by a noise level bar and a sample classification bar to help differentiate between sample type and sample quality, a key for these classifications can be found on the lower middle of the figure.



Supplementary Figure 3: RNA abundance heatmap. Here we present an unclustered heatmap which displays the relative abundance of each organism. A colour key shows a change from white to blue and is measured in relative abundance %. Samples are further classified by a noise level bar and a sample classification bar to help differentiate between sample type and sample quality, a key for these classifications can be found on the lower middle of the figure.

CHAPTER 6

RAPIDprep: A simple, fast protocol for RNA metagenomic sequencing of clinical sample

Chapter 6: RAPIDprep: A simple, fast protocol for RNA metagenomic sequencing of clinical samples

Author Attribution Statement for Thesis with Publications

I am writing this letter to specify the role of Rachel Tulloch in the preparation and submission of the following manuscript:

Tulloch, RL*, Kim, K*, Sikazwe, C, Michie, A, Burrell, R Holmes, EC, Dwyer, D, Britton, P, Kok, J & Eden, J-S, 2022, 'RAPIDprep: A simple, fast protocol for RNA metagenomic sequencing of clinical samples', *Viruses*, under review.

* Denotes equal authorship contribution

Contribution to the manuscript of all authors is as follows:

Task	**Role of Author**
Conceptualisation	R.L.T, J-S.E
Investigation	R.L.T, J-S.E
Methodology	R.L.T, K.K, C.S, A.M, J-S.E
Data curation	R.L.T, K.K, J-S.E
Formal analysis	R.L.T, K.K, J-S.E
Visualisation	R.L.T, J-S.E
Writing – Original draft	R.L.T, 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, J-S.E

I declare that the authorship attribution statements made above are true. In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author, my supervisor.

Rachel Tulloch
December 2022

As the primary supervisor for the candidature upon which the thesis is based, I can confirm that the above authorship attributions statements are correct.

John-Sebastian Eden
December 2022

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

Introduction

Despite major advancements in infectious disease diagnostics and treatment, respiratory tract 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 (Annand et al., 2022). Furthermore, many disease-causing agents are difficult 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) (Wu et al., 2020), 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 (Zhang et al., 2018).

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 (Shi et al., 2022), describe the microbiome and resistomes (Serpa et al., 2022), and identify coinfections that may be associated with increased morbidity and mortality (Li et al., 2018). 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 (Wu et al., 2020). The emergence of novel SARS-CoV-2 variants throughout the course of the pandemic and associated failures in RT-PCR primers for diagnostic (Subramoney et al., 2022) and whole genome sequencing (Itokawa et al., 2020) 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, optimised, 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.

Materials and Methods

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 (Lu et al., 2020) and nucleoproteins (Wang et al., 2019a), 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).

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 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 Mix	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

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.

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.

Bioinformatic analysis of RNA-mNGS data

Raw sequence reads were first quality trimmed and filtered using FastP v0.19.6 (Chen et al., 2018) 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 (Dobin et al., 2013), followed by Burrows-Wheeler Aligner (BWA) v0.7.17 (Li and Durbin, 2009) to ensure complete human sequence removal. The trimmed, human and non-human reads were then filtered into rRNA and non-rRNA and quantified using SortMeRNA v2.1b (Kopylova et al., 2012) before, the trimmed, non-human, non-rRNA reads were then *de novo* assembled using Megahit v1.1.3 (Li et al., 2015) before annotation using blast+ v2.11 (Camacho et al., 2009) and diamond v2.0.11 (Buchfink et al., 2015) 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 (Segata et al., 2012). 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 (Bushnell, 2016). Maximum likelihood trees for individual viruses were estimated using PhyML v2.2.4 (Guindon et al., 2010) with the GTR + Gamma substitution model and 1000 bootstrap replicates.

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).

Optimization of the RAPIDprep assay

rRNA depletion

rRNA is the most abundant component of total RNA isolated from eukaryotic and microbial cells (Zhao et al., 2018). While the importance of rRNA-depleted libraries for improved coverage of mRNA for transcriptome sequencing is recognized (Albert et al., 2021, Bal et al., 2018, Dudas and Bedford, 2019), 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)**.

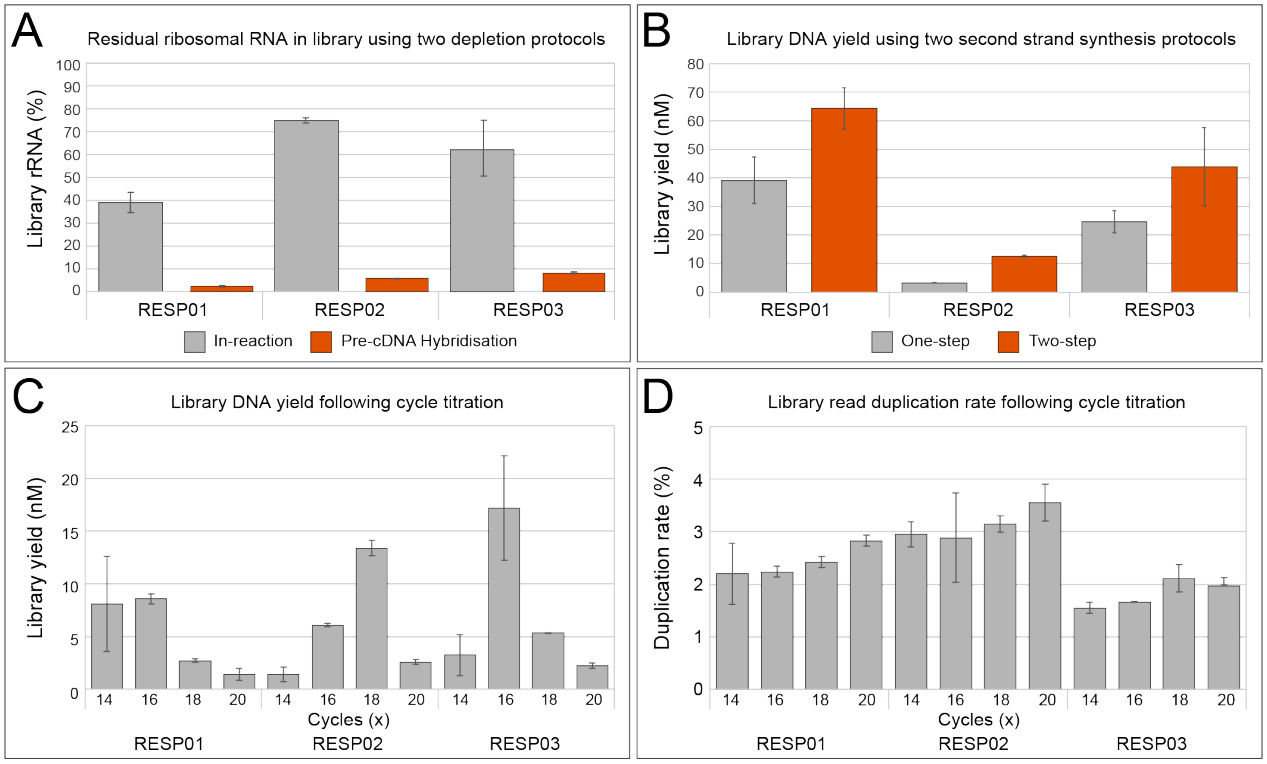


Figure 1: RAPID*prep* optimization experiments. All results here are derived from the same sample extracts (RESP01-RESP03) run in duplicate and presented as mean values and error as standard deviation (SD); (**a).** The shaded bars are representative of the percentage of residual rRNA reads in the library 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.

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%

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 (Tabor and Richardson, 1989). 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 (Zhu, 2014). 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.

Library amplification

It is widely accepted that library preparation can introduce systematic bias to the characterization and representation of microbial communities in a sample (Brenner et al., 2018, Peddu et al., 2020, Poulsen et al., 2022, Wilson et al., 2014). 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 (Poulsen et al., 2022). Following this, limited cycle PCR is used to barcode and complete index adapters before sequencing (Peck et al., 2018). 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 (Di Giallonardo et al., 2018b, Eden et al., 2020b, Tulloch et al., 2021b), 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 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 (Bansal, 2017).

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.

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.

Table 3: RAPID*prep* sample summary overview table.

Library	Group	Virus	Type	Extraction method	LibraryYield (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	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	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		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	Kids unknown	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

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 (Schoonvaere et al., 2016). 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*).

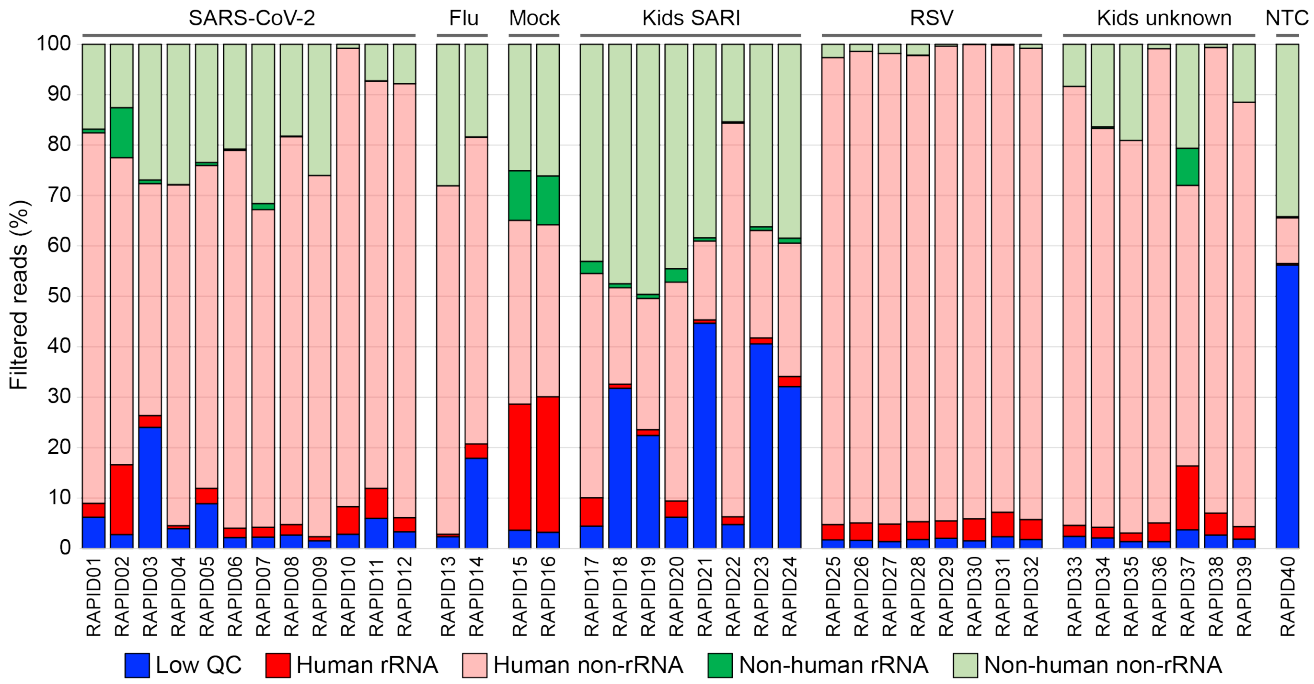


Figure 2: Filtered read distribution and classification across forty RAPID*prep* libraries. The sequence reads were classified into five categories: low quality reads (blue), human rRNA reads (red), human non-rRNA (pink), non-human rRNA reads (green) and non-human non-rRNA reads (light green). Low quality, 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 (Albert et al., 2021). 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 (Dudas and Bedford, 2019). 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 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 (Meadow et al., 2015). 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 (Qamar et al., 2017). 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 (Yang et al., 2019). 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 (Yang et al., 2019). 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.

Viral sequence identification, genome recovery and quantitative performance

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 logRPM 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 logRPM 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 logRPM of 5.77

(**Table 4**). As this sample group was 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 logRPM 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 (Wu et al., 2020).

To assess the quantitative performance of the RAPID*prep* assay, we utilized the cycle threshold values generated using RT-qPCR and compared these to logRPM 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 logRPM and CT values (**Figure 3A**). As expected, logRPM 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 logRPM 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.

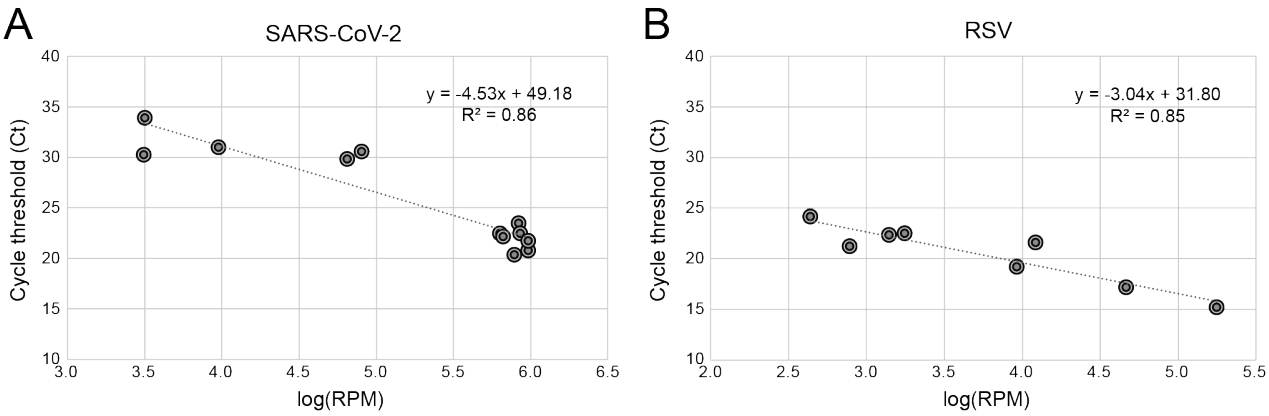


Figure 3: Quantitative detection of SARS-COV-2 and RSV sequences. A simple linear regression model was applied to both SARS-CoV-2 (**A**) and RSV (**B**) data sets with a line of best fit estimating the relationship between log transformed reads per million (RPM) and cycle threshold (CT) values. The 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.

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	Rhinovirus	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

Comparison of RAPIDprep to commercial assay

Finally, we sought to compare the performance of the RAPIDprep 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 RAPIDprep 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 RAPIDprep 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 RAPIDprep 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 (Porter et al., 2021). 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 (Tuddenham et al., 2020). This aspect of the RAPIDprep assay is yet to be explored.

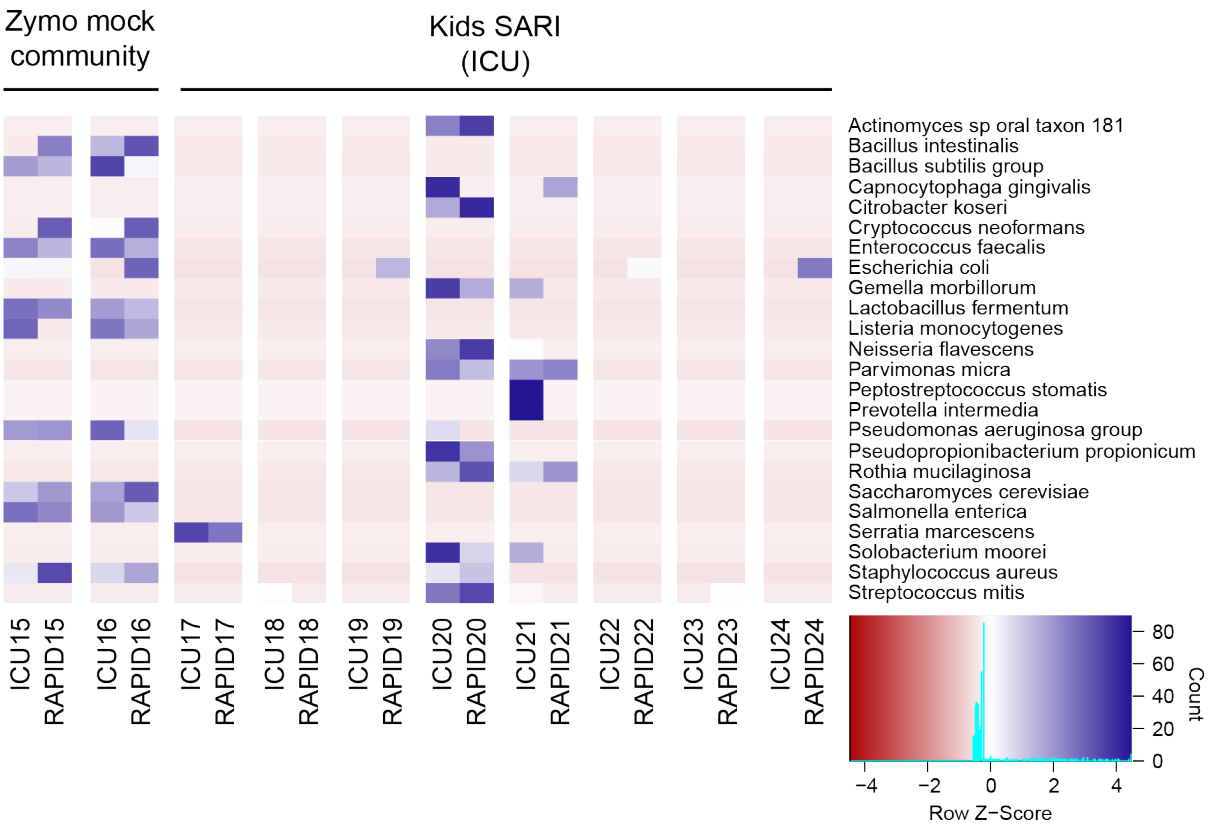


Figure 4: Comparison of RAPID*prep* to commercial RNA library preparation kit. Using previously generated data for the kids SARI cohort, we compared the twenty-four most abundant species identified across both protocols for the same set of samples. An unclustered heatmap of microbial abundance (Z-score) 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.

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 (Sikazwe et al., 2022b, Waller et al., 2022a), and characterising the first cases of COVID-19 in NSW (Eden et al., 2020b). 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 (Annand et al., 2022). 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 at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

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

ALBERT, E., TORRES, I., BUENO, F., HUNTLEY, D., MOLLA, E., FERNÁNDEZ-FUENTES, M. Á., MARTÍNEZ, M., POUJOIS, S., FORQUÉ, L., VALDIVIA, A., SOLANO DE LA ASUNCIÓN, C., FERRER, J., COLOMINA, J. & NAVARRO, D. 2021. 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,* 27**,** 472.e7-472.e10.

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., JACKSON, A. E., SINGANALLUR, N. B., PLAIN, K. M., KIM, K., TACHEDJIAN, M., VAN DER HEIDE, B., CRAMERI, S., WILLIAMS, D. T., SECOMBE, C., LAING, E. D., STERLING, S., YAN, L., JACKSON, L., JONES, C., PLOWRIGHT, R. K., PEEL, A. J., BREED, A. C., DIALLO, I., DHAND, N. K., BRITTON, P. N., BRODER, C. C., SMITH, I. & EDEN, J.-S. 2022. Novel Hendra Virus Variant Detected by Sentinel Surveillance of Horses in Australia. *Emerging infectious diseases,* 28**,** 693-704.

BAL, A., PICHON, M., PICARD, C., CASALEGNO, J. S., VALETTE, M., SCHUFFENECKER, I., BILLARD, L., VALLET, S., VILCHEZ, G., CHEYNET, V., ORIOL, G., TROUILLET-ASSANT, S., GILLET, Y., LINA, B., BRENGEL-PESCE, K., MORFIN, F. & JOSSET, L. 2018. 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,* 18**,** 537-537.

BANSAL, V. 2017. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC bioinformatics,* 18**,** 43-43.

BRENNER, T., DECKER, S. O., GRUMAZ, S., STEVENS, P., BRUCKNER, T., SCHMOCH, T., PLETZ, M. W., BRACHT, H., HOFER, S., MARX, G., WEIGAND, M. A. & SOHN, K. 2018. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine (Baltimore),* 97**,** e9868-e9868.

BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods,* 12**,** 59-60.

BUSHNELL 2016. BBMap short-read aligner, and other bioinformatics tools.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC bioinformatics,* 10**,** 421-421.

CHEN, S., ZHOU, Y., CHEN, Y. & GU, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *bioRxiv*.

DI GIALLONARDO, F., KOK, J., FERNANDEZ, M., CARTER, I., GEOGHEGAN, J. L., DWYER, D. E., HOLMES, E. C. & EDEN, J. S. 2018. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses,* 10.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DUDAS, G. & BEDFORD, T. 2019. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC evolutionary biology,* 19**,** 1-232.

EDEN, J. S., ROCKETT, R., CARTER, I., RAHMAN, H., DE LIGT, J., HADFIELD, J., STOREY, M., REN, X., TULLOCH, R., BASILE, K., WELLS, J., BYUN, R., GILROY, N., O'SULLIVAN, M. V., SINTCHENKO, V., CHEN, S. C., MADDOCKS, S., SORRELL, T. C., HOLMES, E. C., DWYER, D. E., KOK, J. & NCO, V. S. G. 2020. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol,* 6**,** veaa027.

GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic biology,* 59**,** 307-321.

ITOKAWA, K., SEKIZUKA, T., HASHINO, M., TANAKA, R. & KURODA, M. 2020. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLoS One,* 15**,** e0239403.

KOPYLOVA, E., NOE, L. & TOUZET, H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics,* 28**,** 3211-3217.

LI, D., LIU, C.-M., LUO, R., SADAKANE, K. & LAM, T.-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, Y., DENG, X., HU, F., WANG, J., LIU, Y., HUANG, H., MA, J., ZHANG, J., ZHANG, F. & ZHANG, C. 2018. Metagenomic analysis identified co-infection with human rhinovirus C and bocavirus 1 in an adult suffering from severe pneumonia. *The Journal of infection,* 76**,** 311-313.

LU, X., WANG, L., SAKTHIVEL, S. K., WHITAKER, B., MURRAY, J., KAMILI, S., LYNCH, B., MALAPATI, L., BURKE, S. A., HARCOURT, J., TAMIN, A., THORNBURG, N. J., VILLANUEVA, J. M. & LINDSTROM, S. 2020. US CDC Real-Time Reverse Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging infectious diseases,* 26**,** 1654-1665.

MEADOW, J. F., ALTRICHTER, A. E., BATEMAN, A. C., STENSON, J., BROWN, G. Z., GREEN, J. L. & BOHANNAN, B. J. M. 2015. Humans differ in their personal microbial cloud. *PeerJ (San Francisco, CA),* 3**,** e1258-e1258.

PECK, M. A., STURK-ANDREAGGI, K., THOMAS, J. T., OLIVER, R. S., BARRITT-ROSS, S. & MARSHALL, C. 2018. Developmental validation of a Nextera XT mitogenome Illumina MiSeq sequencing method for high-quality samples. *Forensic science international : genetics,* 34**,** 25-36.

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., PHUNG, Q., REINHARDT, A., JEROME, K. R. & GRENINGER, A. L. 2020. Metagenomic Analysis Reveals Clinical SARS-CoV-2 Infection and Bacterial or Viral Superinfection and Colonization. *Clinical chemistry (Baltimore, Md.),* 66**,** 966-972.

PORTER, A. F., COBBIN, J., LI, C. X., EDEN, J. S. & HOLMES, E. C. 2021. Metagenomic Identification of Viral Sequences in Laboratory Reagents. *Viruses,* 13.

POULSEN, C. S., EKSTRØM, C. T., AARESTRUP, F. M. & PAMP, S. J. 2022. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology spectrum,* 10**,** e0009022-e0009022.

QAMAR, W., KHAN, M. R. & ARAFAH, A. 2017. Optimization of conditions to extract high quality DNA for PCR analysis from whole blood using SDS-proteinase K method. *Saudi journal of biological sciences,* 24**,** 1465-1469.

SCHOONVAERE, K., DE SMET, L., SMAGGHE, G., VIERSTRAETE, A., BRAECKMAN, B. P. & DE GRAAF, D. C. 2016. Unbiased RNA Shotgun Metagenomics in Social and Solitary Wild Bees Detects Associations with Eukaryote Parasites and New Viruses. *PloS one,* 11**,** e0168456-e0168456.

SEGATA, N., WALDRON, L., BALLARINI, A., NARASIMHAN, V., JOUSSON, O. & HUTTENHOWER, C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods,* 9**,** 811-814.

SERPA, P. H., DENG, X., ABDELGHANY, M., CRAWFORD, E., MALCOLM, K., CALDERA, S., FUNG, M., MCGEEVER, A., KALANTAR, K. L., LYDEN, A., GHALE, R., DEISS, T., NEFF, N., MILLER, S. A., DOERNBERG, S. B., CHIU, C. Y., DERISI, J. L., CALFEE, C. S. & LANGELIER, C. R. 2022. Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. *Genome medicine,* 14**,** 1-74.

SHI, M., ZHAO, S., YU, B., WU, W.-C., HU, Y., TIAN, J.-H., YIN, W., NI, F., HU, H.-L., GENG, S., TAN, L., PENG, Y., SONG, Z.-G., WANG, W., CHEN, Y.-M., HOLMES, E. C. & ZHANG, Y.-Z. 2022. Total infectome characterization of respiratory infections in pre-COVID-19 Wuhan, China.

SIKAZWE, C., NEAVE, M. J., MICHIE, A., MILETO, P., WANG, J., COOPER, N., LEVY, A., IMRIE, A., BAIRD, R. W., CURRIE, B. J., SPEERS, D., MACKENZIE, J. S., SMITH, D. W. & WILLIAMS, D. T. 2022. Molecular detection and characterisation of the first Japanese encephalitis virus belonging to genotype IV acquired in Australia. *PLoS Negl Trop Dis,* 16**,** e0010754.

SUBRAMONEY, K., MTILENI, N., BHARUTHRAM, A., DAVIS, A., KALENGA, B., RIKHOTSO, M., MAPHAHLELE, M., GIANDHARI, J., NAIDOO, Y., PILLAY, S., RAMPHAL, U., RAMPHAL, Y., TEGALLY, H., WILKINSON, E., MOHALE, T., ISMAIL, A., MASHISHI, B., MBENENGE, N., OLIVEIRA, T., MAKATINI, Z., FIELDING, B. C. & TREURNICHT, F. K. 2022. Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021. *Journal of medical virology,* 94**,** 3676-3684.

TABOR, S. & RICHARDSON, C. C. 1989. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J Biol Chem,* 264**,** 6447-58.

TUDDENHAM, R., EDEN, J. S., GILBEY, T., DWYER, D. E., JENNINGS, Z., HOLMES, E. C. & BRANLEY, J. M. 2020. Human pegivirus in brain tissue of a patient with encephalitis. *Diagn Microbiol Infect Dis,* 96**,** 114898.

TULLOCH, R. L., KOK, J., CARTER, I., DWYER, D. E. & EDEN, J. S. 2021. An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus. *Viruses,* 13.

WALLER, C., TIEMENSMA, M., CURRIE, B. J., WILLIAMS, D. T., BAIRD, R. W. & KRAUSE, V. L. 2022. Japanese Encephalitis in Australia - A Sentinel Case. *N Engl J Med,* 387**,** 661-662.

WANG, L., PIEDRA, P. A., AVADHANULA, V., DURIGON, E. L., MACHABLISHVILI, A., LÓPEZ, M.-R., THORNBURG, N. J. & PERET, T. C. T. 2019. Duplex real-time RT-PCR assay for detection and subgroup-specific identification of human respiratory syncytial virus. *Journal of virological methods,* 271**,** 113676-113676.

WILSON, M. R., NACCACHE, S. N., SAMAYOA, E., BIAGTAN, M., BASHIR, H., YU, G., SALAMAT, S. M., SOMASEKAR, S., FEDERMAN, S., MILLER, S., SOKOLIC, R., GARABEDIAN, E., CANDOTTI, F., BUCKLEY, R. H., REED, K. D., MEYER, T. L., SEROOGY, C. M., GALLOWAY, R., HENDERSON, S. L., GERN, J. E., DERISI, J. L. & CHIU, C. Y. 2014. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *The New England journal of medicine,* 370**,** 2408-2417.

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., YUAN, M.-L., ZHANG, Y.-L., DAI, F.-H., LIU, Y., WANG, Q.-M., ZHENG, J.-J., XU, L., HOLMES, E. C. & ZHANG, Y.-Z. 2020. A new coronavirus associated with human respiratory disease in China.

YANG, L., HAIDAR, G., ZIA, H., NETTLES, R., QIN, S., WANG, X., SHAH, F., RAPPORT, S. F., CHARALAMPOUS, T., METHÉ, B., FITCH, A., MORRIS, A., MCVERRY, B. J., O'GRADY, J. & KITSIOS, G. D. 2019. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respiratory research,* 20**,** 265-12.

ZHANG, D., LOU, X., YAN, H., PAN, J., MAO, H., TANG, H., SHU, Y., ZHAO, Y., LIU, L., LI, J., CHEN, J., ZHANG, Y. & MA, X. 2018. Metagenomic analysis of viral nucleic acid extraction methods in respiratory clinical samples. *BMC genomics,* 19**,** 773-773.

ZHAO, S., ZHANG, Y., GAMINI, R., ZHANG, B. & VON SCHACK, D. 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific reports,* 8**,** 4781-12.

ZHU, B. 2014. Bacteriophage T7 DNA polymerase - sequenase. *Frontiers in microbiology,* 5**,** 181-181.

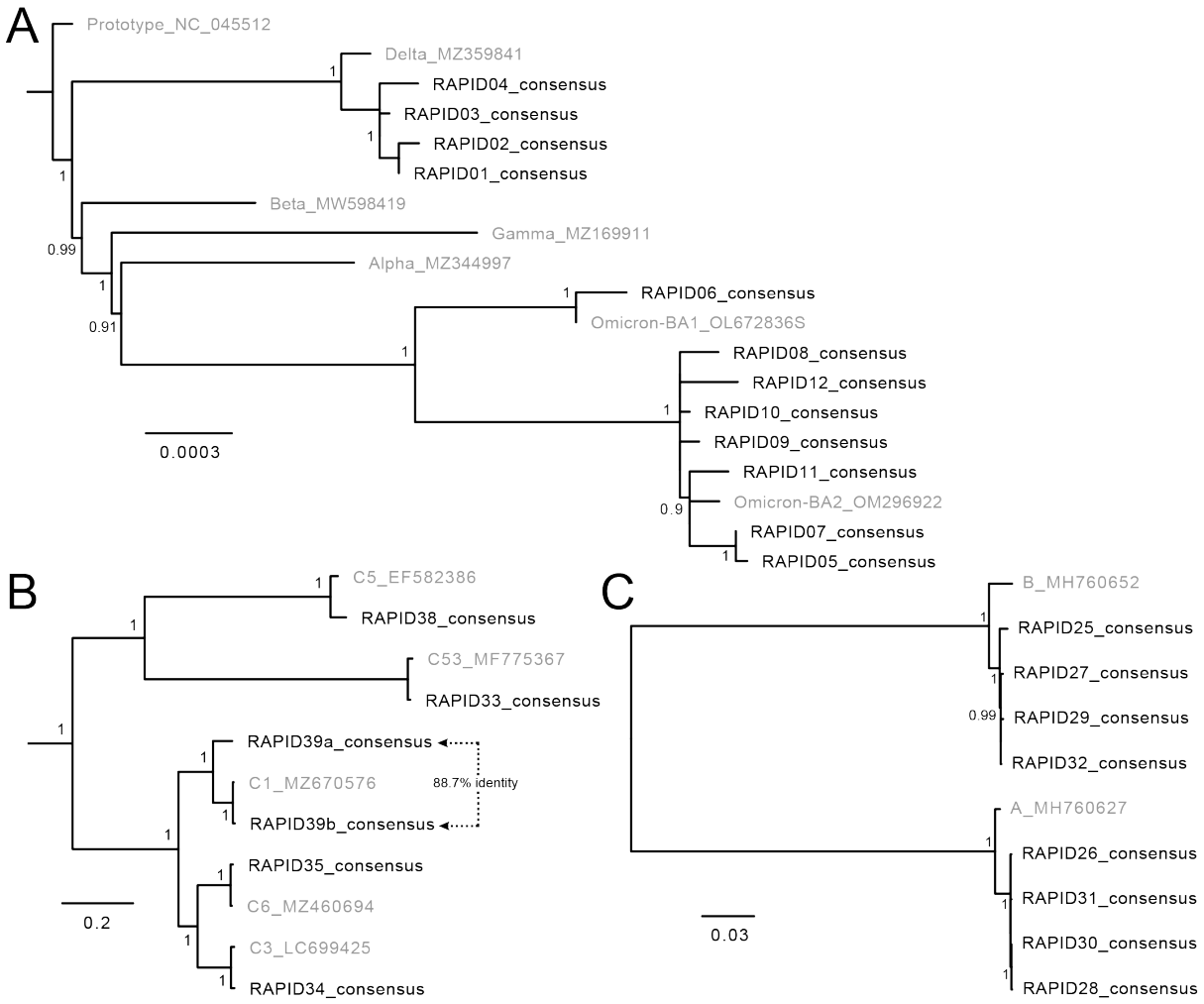
Supplementary Information

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



Supplementary Figure 1: Phylogenetic analysis of viruses identified by RAPID*prep* assay. Maximum likelihood phylogenies were estimated for (**a**) SARS-CoV-2, (**b**) rhinovirus and (**c**) RSV strain identified in this study (black text) against known reference strains (grey text). The GTR + Gamma DNA substitution model was used with 1000 bootstrap replicates to measure support (proportions indicates at each node). Each sequence is labelled according to the library name with multiple viruses indicated with an additional letter (i.e. RAPID39a and RAPID39b). The scales are proportional to the number of substitutions per site.

Supplementary Table 1 – Samples sequenced using RAPIDprep and a commercial assay

Sample	RAPID*prep*		SMARTer Stranded Total RNA-Seq Kit v2	
	Library	*NCBI SRA accession*	*Library*	*NCBI SRA accession*
Zymo_2	RAPID15	SRR22726250	ICU15	SRR22838413
Zymo_3a	RAPID16	SRR22726249	ICU16	SRR22838412
ICU-001_1	RAPID17	SRR22726248	ICU17	SRR22838439
ICU-002_1	RAPID18	SRR22726247	ICU18	SRR22838438
ICU-007_1	RAPID19	SRR22726246	ICU19	SRR22838437
ICU-007_2	RAPID20	SRR22726245	ICU20	SRR22838436
ICU-007_3	RAPID21	SRR22726243	ICU21	SRR22838435
ICU-075_1	RAPID22	SRR22726242	ICU22	SRR22838423
ICU-077_1	RAPID23	SRR22726241	ICU23	SRR22838422
ICU-102_1	RAPID24	SRR22726240	ICU24	SRR22838417

Supplementary Table 2: **Genomic yield, coverage, recovery and qPCR results**

Library	Virus	Non-host, non-rRNA	Viral reads			Coverage		qPCR
		reads	*reads*	*RPM*	*Log (RPM)*	*depth X*	*genome %*	*Ct*
RPD01	SARS-CoV-2	2,834,326	1,792,076	632,275.89	5.80	9,159.00	100.00	22.46
RPD02	SARS-CoV-2	1,461,808	4,552	3,113.95	3.49	21.50	99.50	30.23
RPD03	SARS-CoV-2	4,932,332	15,696	3,182.27	3.50	70.30	99.50	33.92
RPD04	SARS-CoV-2	3,530,136	2,768,186	784,158.46	5.89	15,815.00	100.00	20.37
RPD05	SARS-CoV-2	3,592,638	34,308	9,549.53	3.98	159.50	100.00	31.00
RPD06	SARS-CoV-2	3,170,060	2,651,566	836,440.32	5.92	13,626.00	100.00	23.47
RPD07	SARS-CoV-2	3,517,722	2,336,874	664,314.58	5.82	12,450.00	100.00	22.16
RPD08	SARS-CoV-2	1,723,666	1,470,634	853,201.26	5.93	7,675.00	100.00	22.46
RPD09	SARS-CoV-2	4,502,692	4,334,060	962,548.63	5.98	23,085.00	100.00	20.78
RPD10	SARS-CoV-2	119,716	9,580	80,022.72	4.90	46.30	99.90	30.57
RPD11	SARS-CoV-2	789,160	51,130	64,790.41	4.81	254.30	99.97	29.85
RPD12	SARS-CoV-2	988,654	949,790	960,689.99	5.98	4,877.00	100.00	21.75
RPD25	RSV	403,584	176	436.09	2.64	1.20	41.90	24.15
RPD26	RSV	269,136	2,466	9,162.65	3.96	22.90	99.97	19.20
RPD27	RSV	294,604	518	1,758.29	3.25	4.40	77.80	22.54
RPD28	RSV	316,712	248	783.05	2.89	2.30	72.00	21.23
RPD29	RSV	66,036	804	12,175.18	4.09	7.70	90.10	21.60
RPD30	RSV	5,898	1,042	176,670.06	5.25	10.10	98.50	15.21
RPD31	RSV	17,584	816	46,405.82	4.67	7.90	93.80	17.18
RPD32	RSV	134,974	188	1,392.86	3.14	1.80	54.80	22.38

CHAPTER 7

General Discussion

Chapter 7: General Discussion

Thesis Overview

Infectious diseases remain one of the most significant threats to human health and wellbeing globally, permeating constraints of age, gender, and locality (Yuan et al., 2022). Despite significant advancements in diagnostics, infections remain a leading cause of death globally (Man et al., 2017). The world health organisation (WHO) 2019 Global Health Estimates, showed that lower respiratory infections remain one of the deadliest communicable disease globally (2020). Ranking as the fourth leading cause of human death globally, acute respiratory infections resulted in 2.6 million deaths worldwide in 2019 (2020). Respiratory tract infections are of major concern to vulnerable populations, including young children, the elderly and immunocompromised (Bernadette et al., 2001, Cheemarla and Guerrero-Plata, 2015, Márquez-Escobar, 2017). Etiological agents of respiratory infections are extremely diverse; however, most are derived from bacterial or viral origins (Zumla et al., 2014). Timely detection of pathogens in individuals is paramount in achieving favourable clinical outcomes and public health surveillance.

Human metapneumovirus (HMPV) remains a virus of clinical importance, both locally and globally (Takashita et al., 2021). A review of HMPV epidemiological studies revealed large gaps in our knowledge regarding the evolution and spread of the virus and can thus be considered neglected from an epidemiological perspective. This issue is further amplified by a lack of genomic data – only a few hundred genomes are available globally. This lack of resolution around HMPV molecular epidemiology was identified as an area for improvement and inspired the investigation of HMPV in Australia using whole genome sequencing and phylodynamic approaches in this thesis (Chapters 2 & 3). Here, residual diagnostic extracts collected through routine testing at a major public pathology lab (ICPMR at Westmead Hospital), was used to provide a historical and recent context to the study. The amplicon-based method developed in Chapter 2 of this thesis, uses four overlapping primer pairs and was developed to capture the whole HMPV genome (Tulloch et al., 2021a). The utility of this approach has been evident through similar approaches adapted for viruses such as respiratory syncytial virus (RSV), influenza and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Foley et al., 2022, Binns et al., 2022, Jaworski et al., 2021). Full-length genome

sequencing data is useful in tracking disease outbreaks, characterising transmission pathways and is of critical importance in the analysis of virus genetic diversity and evolution (Shaffer et al., 2022, Subramoney et al., 2022, Tulloch et al., 2021a).

The COVID-19 pandemic has resulted in tremendous impacts to global human health; however the impacts of the pandemic have also dramatically altered the epidemiology and diversity of the endemic respiratory pathogen. Globally, SARS-CoV-2 associated non-pharmaceutical interventions (NPIs) have significantly altered the epidemiology of flu, RSV and HMPV (Dhanasekaran et al., 2022, Eden et al., 2021, Olsen et al., 2020, Binns et al., 2022, Mohebi et al., 2022). Indeed during the first waves of the pandemic, SARS-CoV-2 was largely controlled in Australia by NPIs, and interestingly this ultimately led to the additional suppression of the other endemic respiratory viruses with little to no activity during the winter of 2020 (Eden et al., 2021, Yeoh et al., 2020, Britton et al., 2020). This led to important questions around when and how normal respiratory virus activity might return, this formed one of the aims for Chapter 3 where through molecular investigation the circulation of HMPV in the post-COVID-19 era was compared to our collection of data from 2013-2019. Our WGS studies of HMPV molecular epidemiology provided a unique opportunity to describe circulating viruses pre- and post-COVID-19 pandemic, and, like RSV, identifying a major genetic bottleneck had occurred likely due to the NPIs and COVID-19 related restrictions (Chapter 3). This works highlights the need to continue the genomic surveillance of the respiratory pathogens such as HMPV to better how the disruption caused by COVID-19 might impact us into the future.

The value of culture independent diagnostic methods such as metagenomics (mNGS) has also been increasingly apparent over the course of the pandemic. mNGS has been proven as a powerful tool in genetic detection, classification, and characterisation of novel infectious agents like SARS-CoV-2 (Subramoney et al., 2022). In recent years, culture independent sequencing methods have enabled the rapid development of novel molecular assays, mRNA and DNA based vaccines, disease outbreak identification, characterisation of spill over events and the detection of new variants (Houlihan et al., 2018, Di Giallonardo et al., 2018a, Bal et al., 2018, Gardy et al., 2015). While there is some potential for mNGS to replace routine molecular diagnostic panels, emerging infectious disease threats require bespoke rapid response tools to

inform diagnoses, treatment, and outbreak control, a caveat of this technology has always been prohibitive measures including sensitivity, cost, and turnaround time (Pfefferle et al., 2021).RNA-based mNGS (RNA-mNGS) is particularly powerful as it provides near universal coverage of infecting pathogens by sequencing not only pathogens with RNA genomes but also mRNA expressed by those with DNA genomes. We designed and optimised a rapid, RNA-mNGS method, known as the RAPID*prep* assay, and demonstrated the utility of the approach in producing tangible results from respiratory samples and in less 24 hours from collection (Chapter 6).

Culture-independent diagnostic methods can be particularly advantageous where viruses have similar symptom profiles. Indeed, acute respiratory infections (ARI) can be clinically indistinguishable with ambiguous upper respiratory symptoms consistent across a variety of disease-causing pathogens, including viruses, bacteria, and fungi (Panda et al., 2014, Ditt et al., 2011). In Chapter 5, we assess the utility of parallel DNA and RNA sequencing against a cohort of 23 paediatric samples, obtained from patients with severe ARI and where traditional clinical diagnostics were insufficient. Here, the use of mNGS was evaluated as a solution where a single test could be used to diagnose any pathogen or pathogens present, irrespective if the pathogen is novel or carrying high levels of genetic diversity, factors which are both inherently illusive to target based methods such as PCR or culture. This work highlighted the value of early intervention of mNGS into the disease investigation to take advantage of potentially higher pathogen load and higher sample quality.

Whole genome sequencing (WGS) provides valuable insight into the epidemiology and evolution of a neglected respiratory pathogen - Human Metapneumovirus (HMPV)

Historically, very little interest has been shown in the epidemiology, evolution and spread of HMPV, when compared to pathogens such as influenza, and more recently SARS-CoV-2 and RSV. Whole genome sequencing (WGS) provides valuable resources for characterising viral epidemiology, identifying transmission events and examining outbreaks. WGS sequencing methods provide significant advantages over sub-genomic fragments, with viral full length sequencing data important for diagnosis, disease management, molecular epidemiology and infection tracing and control (Chapter 2) (Shaffer et al., 2022, Serpa et al., 2022, Tulloch et al., 2021a). Between the discovery of HMPV in 2001 and 2019, less than 200 full length genomes had been generated and published globally (Tulloch et al., 2021a). This is partly due to the lack

of simple and reliable methods for the WGS of HMPV, in Chapter 2, we leveraged an overlapping amplicon approach to amplify and then sequence the HMPV genome. The workflow is simple and designed for large-scale molecular epidemiological investigations (Tulloch et al., 2021a). Utilising this approach, 185 HMPV whole genome sequences were generated from residual diagnostic extracts collected between 2013 and 2021 (Chapter 3) allowing for the observation of epidemiological patterns of subtype circulation between 2013 and 2020, with predominant strains varying from year to year and season to season, A2, B2 and B2 subtypes were found to be co-circulating within Australia. In this thesis, we demonstrate the co-circulation of three major HMPV subtypes (A2, B1 & B2), and the local emergence of several globally distributed A2 subtypes carrying novel insertions in the viral glycoprotein, an observation not noted in Australia prior to the publication of this research (Chapter 2 & 3)(Tulloch et al., 2021a). Our studies also demonstrated the global seeding of HMPV variants in Australia, which is important for our understanding of the sources of new diversity and the maintenance of transmission of HMPV. Co-circulation of different variants has been identified in many common respiratory pathogens, including Influenza, RSV and SARS-CoV-2, the co-circulation of multiple viral lineages in the same region is hypothesized to be necessary to sustain transmission during and between seasons (Chow et al., 2016, Tabatabai et al., 2014, Dhanasekaran et al., 2022). The sequences generated in chapter 3 of this thesis, using the whole genome sequencing method developed in Chapter 2 (Tulloch et al., 2021a), has provided the first epidemiological analysis of the circulating diversity of HMPV in Australia at genome-scale resolution. Furthermore, in Chapter 3 of this thesis, we also documented for the first time the local circulation of two novel HMPV A2 variations carrying insertions (duplicates), of 111nt-dup and 180nt-dup length, in the viral attachment glycoprotein. The earliest evidence of this variant in Australia was an A2 strain collected in 2015 which was identified in this thesis, and the replacement of the wildtype A2 strain suggests these variants carry some epidemiological fitness, and interestingly, similar glycoprotein insertions have been observed in recent evolution of both RSV subtypes (Di Giallonardo et al., 2018a). Prior to this study, the duplication variants had only been identified in Japan, documented in a 2018 study and its existence in Australia had not yet been reported (Saikusa et al., 2017b), the identification of the variants within Australia is important for contextualizing the endemic strains of HMPV in Australia, and their global sources.

The changing landscape of endemic respiratory viruses post-COVID-19

Non-pharmaceutical interventions (NPIs) including handwashing, social distancing and wearing face masks these methods were all effective in slowing the spread of SARS-CoV-2 and were shown to be generally effective in preventing the spread of many other respiratory viruses during the pandemic (Baker et al., 2020, Olsen et al., 2020, Yang et al., 2021). During the COVID-19 pandemic restrictions, and in the months subsequent to the ease of restrictions, an clear decline in HMPV cases was noted, likely due to the NPIs put in place to combat the slow the spread of SARS-CoV-2 (Chapter 3). Rates of common respiratory infections including influenza and RSV also dramatically declined both globally and locally during the pandemic (Eden et al., 2022, Mohebi et al., 2022, Bhardwaj et al., 2022). In Australia and New Zealand a significant decrease in influenza, RSV and bronchiolitis was observed in late 2020 (Binns et al., 2022). Toddlers and young school aged children have been identified as main drivers of seasonal RSV, Influenza and HMPV outbreaks. It is reasonable to hypothesize that the restricted movements of key populations may be a factor in contributing to the changing landscape of HMPV in NSW as observed the cocirculation of multiple sublineages to the collapse of diversity during the lockdown period (Chapter 3). A paper by McNab et al. described the delayed re-emergence of RSV in Victoria when compared to NSW and WA, which occurred in the context of Victoria having a prolonged lockdown period (McNab et al., 2021). The findings of the McNab et al. study suggest that RSV seasonality is possibly dependent on behaviors that mitigate viral transmission such as the implementation of NPIs, strengthening the hypothesis that HMPV re-emergence and seasonality may not be solely reliant on colder temperatures (McNab et al., 2021). However, HMPV and RSV are possibly unique in that the key reservoir population (children) are likely to sustain transmission during periods of restricted movement (like during the COVID-19 lockdowns) in contrast to influenza virus, where older populations and adult transmission maintain cases through global networks.

Following the abnormally low case numbers of HMPV in Australia during 2020 and 2021, a marked collapse in diversity was observed. Sequencing data for samples collected between 2020 and 2021 was generated, analysis of this data revealed tightly clustered genomes, indicating genetic bottlenecking in NSW HMPV strains, the A2 subtype was the only variant detected during a rapid spike in cases observed in NSW during September of 2020 (Chapter 3 & Chapter 1, Fig 4). Further analysis is required in respect to the recent publication by Foley et al. (Foley et al., 2022), which identifies an outbreak in Western Australia of the HMPV A2

lineage in July of 2020, 8 weeks before the spike of cases in NSW (Chapter 1 & Chapter 3). It is likely that the outbreak in Western Australia seeded the re-emergence of the A2 lineage following prolonged suppression of HMPV cases in NSW (the sequence data for this study was not available for analysis). Epidemiological analysis of September 2020 NSW HMPV samples and July 2020 WA HMPV samples would be useful in determining the origin of the predominant A2 lineage observed in NSW after the lifting of COVID-19 restrictions. Whilst this analysis was not performed in this thesis, identifying the seed for the NSW outbreak will be useful in prediction of future outbreaks and for the identification of potential reservoirs of HMPV which have allowed for the survival of the viral throughout multiple phases of lockdown.

Valuable information about the epidemiology, evolution and spread of HMPV in the pre and post COVID-19 pandemic are presented here, however this research has also revealed a distinct lack of genome sampling both locally and globally, as well as historically. Epidemiological questions remain unanswered in the wake of the COVID-19 pandemic, including when and if the regular seasonality of HMPV will return, it is imperative ongoing HMPV surveillance is implemented to monitor the disruption of this virus due to the pandemic. Overwhelmingly, the lack of HMPV whole genome sequences available limits the comprehensive analysis of the global distribution and evolution of the virus. The usefulness of the amplicon-based sequencing protocol outlined in Chapter 2 of this thesis is evident, the method was to generate an additional 53 coding complete HMPV genomes in a Western Australia based study conducted by Foley et al., (Foley et al., 2022). It is anticipated that the method outlined in Chapter 2 may continue to be used to increase the number of whole genome sequences generated for HMPV and allow for better genomic resolution of HMPV evolution and epidemiology into the future.

The mNGS diagnosis revolution is close but still faces several challenges.

The progressive development of culture-independent microbial analysis methods has grown at pace in recent years (Wang et al., 2019b). Shotgun mNGS is an unbiased approach which theoretically has the capacity to detect all pathogens in a clinical sample and is suitable for the identification of rare, novel and atypical aetiologies of infectious diseases (Schulz et al.,

2022). Historically diagnostic molecular based assays and culture based diagnostic methods have remained the pillars on which clinical pathogen detection rests (Wang et al., 2019a). The diagnostic value of metagenomic and metatranscriptomic sequencing has been explored in retrospective proof of concept studies, such as the study detailed in Chapter 6. Indeed, tt appears mNGS has the capacity to fulfil a niche in diagnostic medicine where other microbiological, serological, and molecular based diagnostic methods fail to yield a clinically relevant diagnosis. The usefulness of mNGS in capturing the DNA or RNA of novel or unexpected pathogens is demonstrated in Chapter 6 of this thesis, where mNGS was used to identify a severe case of Influenza C in a child with an unknown respiratory illness, and a plethora of potentially pathogenic microbes in several other rescues respiratory samples.

Despite the ongoing demonstration of mNGS as a useful diagnostic tool, the uptake of the method in clinical settings have been relatively slow (Hogan et al., 2021). An aim of chapters 5 & 6 of this thesis sought to identify and pragmatically evaluate barriers to the integration of mNGS in clinical settings. Overwhelmingly, barriers to mNGS uptake which could not be circumvented utilising adjustments to operation methods included, set-up and ancillary costs of obtaining and maintaining a sequencing platform, the availability of skilled individuals including bioinformaticians or individuals familiar with command line programming, and technical personnel to support the use and function of the sequencing platforms. Whilst the cost and operational requirements were not objectively assessed in this study, this remains an area of concern which may need to be explored urgently in the future.

In Chapter 5 of this thesis, clinical characteristics, treatments, and traditional diagnostic methods were evaluated alongside novel mNGS findings. Sample quality was used to determine the effect of sample integrity on pathogen identification, overwhelmingly sample quality was deemed an important consideration for use in metagenomic and metatranscriptomic sequencing. Where sample quality was compromised, the ability to detect candidate pathogens was seriously impaired with 15 of 23 samples included for analysis in this study returning low quality libraries, and limited diagnostic yield. Findings within this thesis suggest the use of metagenomic sequencing in a diagnostic capacity needs to be considered earlier in the diagnostic sphere. Part of the reason is that mNGS has long been considered by clinicians as a “last ditch” attempt, when traditional diagnostic methods are exhausted. The delay in the employment of mNGS methods can mean that often samples submitted for sequencing are

rescued extracts, taken for the purpose of alternative diagnostic testing as evident in the samples utilised in Chapter 5 & 6.

Findings in Chapter 5 & 6 demonstrate the importance of collecting samples for the purpose of metagenomics, with samples collected for the purpose of metagenomics used in Chapter 6 demonstrating superior quality and resultant metagenomic libraries. Findings in Chapter 5 also suggest the timing of sample acquisition can also affect the capability of mNGS to provide useful diagnostic information. Timely and accurate diagnosis remains a highly challenging task for patients hospitalized with respiratory infections with novel, rare or aetiologically pathogens. Many acute respiratory infections present with indistinguishable clinical signs, usually panels of multiplexed tests are ordered based on clinical symptoms. Whilst multiplexed diagnostic qPCR casts a wide net for pathogen detection, there is room for error where potential pathogens are not tested for, testing is hypothesis driven and requires a priori suspicion of the causative pathogen. When samples are obtained early during patient admission, the acute infectome is perhaps most significant, where there has not been the introduction of antibiotics and the pathogen load is likely high. Several studies have provided congruent results which suggest the early introduction of mNGS using samples collected for the purpose of metagenomic sequencing is advantageous in the identification of disease-causing agents and necessary to assist in the differentiation of potential pathogens compared to non-disease causing commensal species (Yuan et al., 2022, Schulz et al., 2022) .

The lack of standardisation across metagenomic/metatranscriptomic sampling, extraction, sequencing method and analysis pipeline is prohibitive in the wider adoption of mNGS as a diagnostic tool (Simner et al., 2018). The lack of standardisation across metagenomic/metatranscriptomic sampling, extraction, sequencing method and analysis pipeline is prohibitive in the wider adoption of mNGS as a diagnostic tool (Simner et al., 2018). Criteria for a positive mNGS result remains ambiguous and varied across published proof of concept studies (Serpa et al., 2022, Nagy-Szakal et al., 2021). The definition of a standardised criteria for the characterisation of a positive mNGS detection has been suggested in numerous studies evaluating the use of mNGS as a clinical diagnostic tool (Bal et al., 2018, Schulz et al., 2022, Govender et al., 2021, Wang et al., 2019b). Over 70% of the samples sequenced in the children's cohort detailed in chapter 5, returned a negative mNGS result, whilst all samples had

yielded clinically negative results using conventional methods, it was anticipated a greater number of samples would yield a positive return for a potential pathogen. Where a sample returns a negative mNGS result, care must be taken to differentiate a true negative result from a potential pathogen which is below the detection limit of the method. Future studies should be aimed at developing structure and standardisation around sample collection and extraction, as well as the introduction of internal and external controls to set defined detection limits. This is particularly important when moving mNGS into an accredited framework. Metagenomics and metatranscriptomics will remain a niche product with these numerous barriers inhibiting its widespread uptake in a clinical space. While it is unlikely clinical mNGS will replace the likes of qPCR and culture in the immediate future, its use demonstrates great potential particularly for syndromes including gastroenteritis, respiratory infections, and CNS disease, where the diagnosis yield is low. With appropriate patient selection, sampling time, sample handling, and data interpretation, mNGS has the potential to emerge as a promising technology for precision diagnosis and tailored therapy for clinical infectious diseases despite its limitations.

Bridging the gap: Rapid RNA-based metatranscriptomic next-generation sequencing (RNA-mNGS) demonstrates potential for use novel disease outbreak situations.

The early detection of causative pathogens in patients with severe infection is critical in directing clinical intervention. The use of qPCR is flawed where variants which escape detection are causative, or where a pathogen causes uncharacterised or unexpected clinical signs (Yuan et al., 2022). Utilising culture-based methods can be a reasonable diagnostic technique where a potential pathogen can be grown in a timely fashion, however this does not account for difficult to culture or unculturable microorganisms (Rhodes et al., 2010, Wyde et al., 2003). Priori knowledge is not required for the use of mNGS, which as a tool, relies on the unbiased shotgun sequencing of the entire microbiome of a sample. Traditional mNGS sequencing time can typically vary between a few hours to upwards of 10 days (Simner et al., 2018), longer turnaround times precludes the use of the method in cases where a rapid diagnosis is required. Failure to diagnose infections or a delay in diagnosis has been reported to increase hospitalisation length and can ultimately lead to increased morbidity and mortality (Stoddard et al., 2009). During the construction of this thesis a need was identified to decrease the amount of time taken for the collection, preparation and sequencing of a diagnostic sample as time to results has historically been a prohibitive factor in the adoption of mNGS as a rapid diagnostic tool (Li and Yan, 2021, Gu et al., 2021). An aim of this thesis was to streamline the use of

RNA-mNGS in a rapid and readily available way to enable the identification of unknown, and novel pathogens in emerging disease outbreaks (Chapter 6). A robust approach was developed to detect pathogen RNA from sample collection in less than 24 hours, critical analysis of the study results revealed that whilst the speed to results can be lowered, library depth and usefulness may be compromised (Chapter 6). Genome recovery was shown to be lower than expected, with PCR positive RSV samples yielding between 44-99% of the whole viral genome, less than optimal genome recovery may be due to the variation in library sequencing depth and sample quality by the way of over-abundance of human sequences, likely due to variation in sample collection. To validate the method across a suite of different sample types and quality, quantitative RT-PCR was utilised in Chapter 6 and a positive correlation was identified between total input gDNA and the calculated human concentration (ng/uL). The study suggests higher gDNA concentrated samples may not preform optimally and may also indicate an overabundance of human genomic material. Key advances underlying this approach include minimal handling throughout the library preparation, the use of widely available reagents and lack of QC requirements for library generation.

Whilst the outcome of the investigation of a streamlined, fast approach to mNGS for the diagnosis of novel or unknown pathogens may not replace routine diagnostic tests, the RAPID*prep* assay was shown again to be fit for purpose in the characterisation of both human and non-human disease outbreaks (Annand et al., 2022, Sikazwe et al., 2022a, Waller et al., 2022b). The RAPID*prep* assay was utilised to identify the first cases of Japanese encephalitis virus during the 2021-2022 outbreak in South-Eastern Australia (Sikazwe et al., 2022a, Waller et al., 2022b) as well as a novel Hendra variant in a horse from Queensland (Annand et al., 2022, Sikazwe et al., 2022a). It is anticipated this method will continue to be adapted and utilised in novel disease outbreaks and that the proof-of-concept studies performed in this thesis involving both the rapid diagnostic method and the routine clinical mNGS method may ultimately assist in the eventual uptake of these methods in a more formal diagnostic capacity, and in the nearer future, assist in the movement of these methods into the clinical trials realm. It is anticipated that pathogen-agnostic mNGS testing will assume a greater role in the identification of novel, emerging and re-emerging pathogens and it is hoped that the findings presented in this thesis will facilitate this. The limitations of this study are acknowledged and

whilst it was not within the scope of this thesis, exploration of Oxford Nanopore® sequencing as an alternative to illumina® based sequencing approaches should be explored.

References

AHN, J., YANG, L., PASTER, B. J., GANLY, I., MORRIS, L., PEI, Z. & HAYES, R. B. 2011. Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. *PloS one,* 6**,** e22788.

ALBERT, E., TORRES, I., BUENO, F., HUNTLEY, D., MOLLA, E., FERNÁNDEZ-FUENTES, M. Á., MARTÍNEZ, M., POUJOIS, S., FORQUÉ, L., VALDIVIA, A., SOLANO DE LA ASUNCIÓN, C., FERRER, J., COLOMINA, J. & NAVARRO, D. 2021. 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,* 27**,** 472.e7-472.e10.

ALEMANY, A., BARÓ, B., OUCHI, D., RODÓ, P., UBALS, M., CORBACHO-MONNÉ, M., VERGARA-ALERT, J., RODON, J., SEGALÉS, J., ESTEBAN, C., FERNÁNDEZ, G., RUIZ, L., BASSAT, Q., CLOTET, B., ARA, J., VALL-MAYANS, M., G-BEIRAS, C., BLANCO, I. & MITJÀ, O. 2021. Analytical and clinical performance of the panbio COVID-19 antigen-detecting rapid diagnostic test. *The Journal of infection,* 82**,** 186-230.

ALTSCHUL, S. F., GISH, W., MILLER, W., MEYERS, E. W. & LIPMAN, D. J. 1990. Basic Local Alignment Search Tool. *Journal of molecular biology,* 215**,** 403-410.

ARANAZ-ANDRÉS, J. M., CHÁVEZ, A. C. F., LASO, A. M., ABREU, M., NÚÑEZ, P. M., GALÁN, J. C. & MORENO, R. C. 2022. Analysis of the diagnostic accuracy of rapid antigenic tests for detection of SARS-CoV-2 in hospital outbreak situation. *European journal of clinical microbiology & infectious diseases,* 41**,** 305-312.

BAKER, J. L., MORTON, J. T., DINIS, M., ALVAREZ, R., TRAN, N. C., KNIGHT, R. & EDLUND, A. 2021. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome research,* 31**,** 64-74.

BARB, J. J., OLER, A. J., KIM, H.-S., CHALMERS, N., WALLEN, G. R., CASHION, A., MUNSON, P. J. & AMES, N. J. 2016. Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. *PloS one,* 11**,** e0148047-e0148047.

BARTRAM, A. K., LYNCH, M. D. J., STEARNS, J. C., MORENO-HAGELSIEB, G. & NEUFELD, J. D. 2011. Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. *Applied and environmental microbiology,* 77**,** 5569-5569.

BASHIARDES, S. S., ZILBERMAN-SCHAPIRA, G. G. & ELINAV, E. E. 2016. Use of Metatranscriptomics in Microbiome Research. *Bioinformatics and Biology Insights,* 2016**,** 19-25.

BASSO, D., AITA, A., PADOAN, A., COSMA, C., NAVAGLIA, F., MOZ, S., CONTRAN, N., ZAMBON, C.-F., MARIA CATTELAN, A. & PLEBANI, M. 2021. Salivary SARS-CoV-2 antigen rapid detection: A prospective cohort study. *Clinica chimica acta,* 517**,** 54-59.

BEGON, M., TOWNSEND, C. R. & HARPER, J. L. 1986. *Ecology : individuals, populations and communities,* Oxford, Blackwell Scientific.

BERBERS, B., SALTYKOVA, A., GARCIA-GRAELLS, C., PHILIPP, P., ARELLA, F., MARCHAL, K., WINAND, R., VANNESTE, K., ROOSENS, N. H. C. & DE KEERSMAECKER, S. C. J. 2020. Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. *Scientific reports,* 10**,** 4310-4310.

BERRY, G. J., ZHEN, W., SMITH, E., MANJI, R., SILBERT, S., LIMA, A., HARINGTON, A., MCKINLEY, K., KENSINGER, B., NEFF, C. & LU, D. 2022. Multicenter Evaluation of the BioFire Respiratory Panel 2.1 (RP2.1) for Detection of SARS-CoV-2 in Nasopharyngeal Swab Samples. *Journal of clinical microbiology,* 60**,** e0006622-e0006622.

BIGOUETTE, J. P., FORD, L., PRAY, I., LANGOLF, K., KAHRS, J., ZOCHERT, T., TATE, J. E., GIERYN, D., KIRKING, H. L., WESTERGAARD, R. & KILLERBY, M. E. 2021. Application of a Serial Antigen-Based Testing Strategy for Severe Acute Respiratory Syndrome Coronavirus 2 and Student Adherence in a University Setting: Wisconsin, October-November 2020. *Open forum infectious diseases,* 8**,** ofab472-ofab472.

BLAIRON, L., WILMET, A., BEUKINGA, I. & TRÉ-HARDY, M. 2020. Implementation of rapid SARS-CoV-2 antigenic testing in a laboratory without access to molecular methods: Experiences of a general hospital. *Journal of clinical virology,* 129**,** 104472-104472.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BOUZID, D., CASALINO, E., MULLAERT, J., LAURENT, O., DUVAL, X., LESCURE, F. X., PEIFFER SMADJA, N., TUBIANA, S., ARMAND LEFÈVRE, L., DESCAMPS, D., FIDOUH, N., CHOQUET, C., LUCET, J.-C. & VISSEAUX, B. 2021. Added value of rapid respiratory syndromic testing at point of care versus central laboratory testing: a controlled clinical trial. *Journal of antimicrobial chemotherapy,* 76**,** iii20-iii27.

BOZIO, C. H., VUONG, J., DOKUBO, E. K., FALLAH, M. P., MCNAMARA, L. A., POTTS, C. C., DOEDEH, J., GBANYA, M., RETCHLESS, A. C., PATEL, J. C., CLARK, T. A., KOHAR, H., NAGBE, T., CLEMENT, P., KATAWERA, V., MAHMOUD, N., DJINGAREY, H. M., PERROCHEAU, A., NAIDOO, D., STONE, M., GEORGE, R. N., WILLIAMS, D., GASASIRA, A., NYENSWAH, T., WANG, X., FOX, L. M., KONWAY, Y., WIAH, S. Q., DOEDEH, V., BAO, U., SENNEH, G., GORWOR, L., GONOTEE, P., PAASEWE, T., TAMATAI, G., YARKEH, J., SMITH, S., BRIMA-DAVIS, A., DAUDA, G., MONGER, T., GORNOR-PEWU, L. W., LOMBEH, S., WILSON, H. W., KORVAYAN, M., DOVILLIE, N., JETOH, R., TAWEH, F., VERA WALKER, Y., HARDY, P., FREEMAN, M., GEORGE, G., KERWILLAIN, G., TOE, S., GHARTEY, E., LARWAY, L., GWEH, D., ALLEN, D., FRIESEN, S., GWESA, G., KINKADE, C., REED, M., CHANG, A., GEORGE, J., SCHIER, J., THOMAS, J., DIAZ, M. H., JENKINS, L. T., MAHON, B. E., SCHMINK, S. E., JOSEPH, S. J., WALLER, J. L., WHALEY, M. J., WINCHELL, J. M., ARTHUR, R. R., FULLER, S., CHRISTIAN, K., REDD, J. T., YEALUE, K. D. M., NAIENE, J., ASAMOAH FRIMPONG, J., AMO-ADDAE, M., STEPHEN, O., VON GOTTBERG, A. & TAHA, M. 2018. Outbreak of Neisseria meningitidis serogroup C outside the meningitis belt—Liberia, 2017: an epidemiological and laboratory investigation. *The Lancet infectious diseases,* 18**,** 1360-1367.

BRENNER, T., DECKER, S. O., GRUMAZ, S., STEVENS, P., BRUCKNER, T., SCHMOCH, T., PLETZ, M. W., BRACHT, H., HOFER, S., MARX, G., WEIGAND, M. A. & SOHN, K. 2018. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine (Baltimore),* 97**,** e9868-e9868.

BRUMFIELD, K. D., HUQ, A., COLWELL, R. R., OLDS, J. L. & LEDDY, M. B. 2020. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PloS one,* 15**,** e0228899.

CARPENTER, MEREDITH L., BUENROSTRO, JASON D., VALDIOSERA, C., SCHROEDER, H., ALLENTOFT, MORTEN E., SIKORA, M., RASMUSSEN, M., GRAVEL, S., GUILLÉN, S., NEKHRIZOV, G., LESHTAKOV, K., DIMITROVA, D., THEODOSSIEV, N., PETTENER, D., LUISELLI, D., SANDOVAL, K., MORENO-ESTRADA, A., LI, Y., WANG, J., GILBERT, M. THOMAS P., WILLERSLEV, E., GREENLEAF, WILLIAM J. & BUSTAMANTE, CARLOS D. 2013. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *American journal of human genetics,* 93**,** 852-864.

CHARALAMPOUS, T., KAY, G. L., RICHARDSON, H., AYDIN, A., BALDAN, R., JEANES, C., RAE, D., GRUNDY, S., TURNER, D. J., WAIN, J., LEGGETT, R. M., LIVERMORE, D. M. & O'GRADY, J. 2019. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature biotechnology,* 37**,** 783-792.

CHEN, S., ZHOU, Y., CHEN, Y. & GU, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *bioRxiv*.

CHEN, Z., COLLIN, E., PEDDIREDDI, L., CLEMENT, T., GAUGER, P. & HAUSE, B. M. 2017. Genetic diversity in envelope genes of contemporary U.S. porcine reproductive and respiratory syndrome virus strains influences viral antigenicity. *Research in veterinary science,* 115**,** 432-441.

CHO, J., MATUSHEK, S. & CHARNOT-KATSIKAS, A. 2018. Optimizing the Utilization of Multiplex PCR Respiratory Panels at Two Emergency Departments Over Three Consecutive Influenza Seasons. *American journal of clinical pathology,* 150**,** S147-S148.

COLE, J. R., WANG, Q., CARDENAS, E., FISH, J., CHAI, B., FARRIS, R. J., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., MARSH, T., GARRITY, G. M. & TIEDJE, J. M. 2009. Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research,* 37**,** D141-D145.

COLLINSON, N., SNAPE, N., BEAGLEY, K., FANTINO, E. & SPANN, K. 2021. COPD Is Associated with Elevated IFN-β Production by Bronchial Epithelial Cells Infected with RSV or hMPV. *Viruses,* 13**,** 911.

CUMMINGS, L. A., KUROSAWA, K., HOOGESTRAAT, D. R., SENGUPTA, D. J., CANDRA, F., DOYLE, M., THIELGES, S., LAND, T. A., ROSENTHAL, C. A., HOFFMAN, N. G., SALIPANTE, S. J. & COOKSON, B. T. 2016. Clinical Next Generation Sequencing Outperforms Standard Microbiological Culture for Characterizing Polymicrobial Samples. *Clinical chemistry (Baltimore, Md.),* 62**,** 1465-1473.

DAS, A. K., GOSWAMI, S., LEE, K. & PARK, S.-J. 2019. A hybrid and scalable error correction algorithm for indel and substitution errors of long reads. *BMC genomics,* 20**,** 948-948.

DI GIALLONARDO, F., KOK, J., FERNANDEZ, M., CARTER, I., GEOGHEGAN, J. L., DWYER, D. E., HOLMES, E. C. & EDEN, J.-S. 2018. Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses,* 10.

DUAN, H., LI, X., MEI, A., LI, P., LIU, Y., LI, X., LI, W., WANG, C. & XIE, S. 2021. The diagnostic value of metagenomic next**-**generation sequencing in infectious diseases. *BMC infectious diseases,* 21**,** 62-62.

DUGAT, T., LOUX, V., MARTHEY, S., MOROLDO, M., LAGRÉE, A.-C., BOULOUIS, H.-J., HADDAD, N. & MAILLARD, R. 2014. Comparative genomics of first available bovine Anaplasma phagocytophilum genome obtained with targeted sequence capture. *BMC genomics,* 15**,** 973-973.

EDEN, J.-S., SIKAZWE, C., XIE, R., DENG, Y.-M., SULLIVAN, S. G., MICHIE, A., LEVY, A., CUTMORE, E., BLYTH, C. C., BRITTON, P. N., CRAWFORD, N., DONG, X., DWYER, D. E., EDWARDS, K. M., HORSBURGH, B. A., FOLEY, D., KENNEDY, K., MINNEY-SMITH, C., SPEERS, D., TULLOCH, R. L., HOLMES, E. C., DHANASEKARAN, V., SMITH, D. W., KOK, J., BARR, I. G. & GROUP, A. R. S. 2021. Off-season RSV epidemics in Australia after easing of COVID-19 restrictions. *medRxiv***,** 2021.07.21.21260810.

FERNANDEZ-MARMIESSE, A., GOUVEIA, S. & COUCE, M. L. 2018. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Current medicinal chemistry,* 25**,** 404-432.

FREED, N. E., VLKOVÁ, M., FAISAL, M. B. & SILANDER, O. K. 2020. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biology Methods & Protocols,* 5.

FREY, K. G., HERRERA-GALEANO, J. E., REDDEN, C. L., LUU, T. V., SERVETAS, S. L., MATECZUN, A. J., MOKASHI, V. P. & BISHOP-LILLY, K. A. 2014. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC genomics,* 15**,** 96-96.

GOVENDER, K. N., STREET, T. L., SANDERSON, N. D. & EYRE, D. W. 2021. Metagenomic Sequencing as a Pathogen-Agnostic Clinical Diagnostic Tool for Infectious Diseases: a Systematic Review and

Meta-analysis of Diagnostic Test Accuracy Studies. *Journal of clinical microbiology,* 59**,** e0291620-e0291620.

GRENINGER, A. L. 2018. The challenge of diagnostic metagenomics. *Expert review of molecular diagnostics,* 18**,** 605-615.

GU, W., DENG, X., LEE, M., SUCU, Y. D., AREVALO, S., STRYKE, D., FEDERMAN, S., GOPEZ, A., REYES, K., ZORN, K., SAMPLE, H., YU, G., ISHPUNIANI, G., BRIGGS, B., CHOW, E. D., BERGER, A., WILSON, M. R., WANG, C., HSU, E., MILLER, S., DERISI, J. L. & CHIU, C. Y. 2021. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nature medicine,* 27**,** 115-124.

HANDELSMAN, J., RONDON, M. R., BRADY, S. F., CLARDY, J. & GOODMAN, R. M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology,* 5**,** R245-R249.

HASAN, M. R., SUNDARARAJU, S., TANG, P., TSUI, K.-M., LOPEZ, A. P., JANAHI, M., TAN, R. & TILLEY, P. 2020. A metagenomics-based diagnostic approach for central nervous system infections in hospital acute care setting. *Scientific reports,* 10**,** 11194-11194.

HOGAN, C. A., YANG, S., GARNER, O. B., GREEN, D. A., GOMEZ, C. A., DIEN BARD, J., PINSKY, B. A. & BANAEI, N. 2021. Clinical Impact of Metagenomic Next-Generation Sequencing of Plasma Cell-Free DNA for the Diagnosis of Infectious Diseases: A Multicenter Retrospective Cohort Study. *Clinical infectious diseases,* 72**,** 239-245.

JACOBS, J., KÜHNE, V., LUNGUYA, O., AFFOLABI, D., HARDY, L. & VANDENBERG, O. 2020. Implementing COVID-19 (SARS-CoV-2) Rapid Diagnostic Tests in Sub-Saharan Africa: A Review. *Frontiers in medicine,* 7**,** 557797-557797.

JAWORSKI, E., LANGSJOEN, R. M., MITCHELL, B., JUDY, B., NEWMAN, P., PLANTE, J. A., PLANTE, K. S., MILLER, A. L., ZHOU, Y., SWETNAM, D., SOTCHEFF, S., MORRIS, V., SAADA, N., MACHADO, R. R., MCCONNELL, A., WIDEN, S. G., THOMPSON, J., DONG, J., REN, P., PYLES, R. B., KSIAZEK, T. G., MENACHERY, V. D., WEAVER, S. C. & ROUTH, A. L. 2021. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *eLife,* 10.

JONES, M. R. & GOOD, J. M. 2016. Targeted capture in evolutionary and ecological genomics. *Molecular ecology,* 25**,** 185-202.

KAN-O, K., RAMIREZ, R., MACDONALD, M. I., ROLPH, M., RUDD, P. A., SPANN, K. M., MAHALINGAM, S., BARDIN, P. G. & THOMAS, B. J. 2017. Human Metapneumovirus Infection in Chronic Obstructive Pulmonary Disease: Impact of Glucocorticosteroids and Interferon. *The Journal of Infectious Diseases,* 215**,** 1536-1545.

KHURSHID, Z., ZOHAIB, S., NAJEEB, S., ZAFAR, M. S., SLOWEY, P. D. & ALMAS, K. 2016. Human Saliva Collection Devices for Proteomics: An Update. *International journal of molecular sciences,* 17**,** 846-846.

KIM, H.-M., JEON, S., CHUNG, O., JUN, J. H., KIM, H.-S., BLAZYTE, A., LEE, H.-Y., YU, Y., CHO, Y. S., BOLSER, D. M. & BHAK, J. 2021. Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *Gigascience,* 10.

KOPYLOVA, E., NOE, L. & TOUZET, H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics,* 28**,** 3211-3217.

LAMBERT-NICLOT, S., CUFFEL, A., LE PAPE, S., VAULOUP-FELLOUS, C., MORAND-JOUBERT, L., ROQUE-AFONSO, A.-M., LE GOFF, J. & DELAUGERRE, C. 2020. Evaluation of a Rapid Diagnostic Assay for Detection of SARS-CoV-2 Antigen in Nasopharyngeal Swabs. *Journal of clinical microbiology,* 58.

LEUNG, H. C. M., YIU, S.-M. & CHIN, F. Y. L. 2015. IDBA-MTP: A Hybrid Metatranscriptomic Assembler Based on Protein Information. *Journal of computational biology,* 22**,** 367-376.

LEUNG, H. C. M., YIU, S.-M., PARKINSON, J. & CHIN, F. Y. L. 2013. IDBA-MT: De Novo Assembler for Metatranscriptomic Data Generated from Next-Generation Sequencing Technology. *Journal of computational biology,* 20**,** 54-550.

LI, B. & YAN, T. 2021. Next generation sequencing reveals limitation of qPCR methods in quantifying emerging antibiotic resistance genes (ARGs) in the environment. *Applied microbiology and biotechnology,* 105**,** 2925-2936.

LI, D., LUO, R., LIU, C.-M., LEUNG, C.-M., TING, H.-F., SADAKANE, K., YAMASHITA, H. & LAM, T.-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods (San Diego, Calif.),* 102**,** 3-11.

LIN, L., CARLQUIST, J., SINCLAIR, W., HALL, T., LOPANSRI, B. K. & BENNETT, S. T. 2021. Experience With False-Positive Test Results on the TaqPath Real-Time Reverse Transcription-Polymerase Chain Reaction Coronavirus Disease 2019 (COVID-19) Testing Platform. *Archives of pathology & laboratory medicine (1976),* 145**,** 259-261.

LIU, L. P., OZA, S. M., HOGAN, D. P., PERIN, J. P., RUDAN, I. P., LAWN, J. E. P., COUSENS, S. P., MATHERS, C. P. & BLACK, R. E. P. 2015. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet (British edition),* 385**,** 430-440.

LO, C.-C. & CHAIN, P. S. G. 2014. Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC bioinformatics,* 15**,** 366-366.

LUO, C., TSEMENTZI, D., KYRPIDES, N., READ, T. & KONSTANTINIDIS, K. T. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one,* 7**,** e30087-e30087.

MCDONALD, D., PRICE, M. N., GOODRICH, J., NAWROCKI, E. P., DESANTIS, T. Z., PROBST, A., ANDERSEN, G. L., KNIGHT, R. & HUGENHOLTZ, P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal,* 6**,** 610-618.

MIAO, Q., MA, Y., WANG, Q., PAN, J., ZHANG, Y., JIN, W., YAO, Y., SU, Y., HUANG, Y., WANG, M., LI, B., LI, H., ZHOU, C., LI, C., YE, M., XU, X., LI, Y. & HU, B. 2018. Microbiological Diagnostic Performance of Metagenomic Next-generation Sequencing When Applied to Clinical Practice. *Clinical infectious diseases,* 67**,** S231-S240.

MOSS, S., ATSAWARUNGRUANGKIT, A., DANG, L., CHUA, D., ZHOU, Y., CHONG, Z. Z., ZHANG, H. & GRAHAM, D. Y. 2021. S1381 Rapid Prediction of H. pylori Antibiotic Resistance Using Next Generation Sequencing of Stool Samples Compared to Gastric Biopsies. *The American journal of gastroenterology,* 116**,** S634-S634.

NEUMANN, B., BENDER, J. K., MAIER, B. F., WITTIG, A., FUCHS, S., BROCKMANN, D., SEMMLER, T., EINSELE, H., KRAUS, S., WIELER, L. H., VOGEL, U. & WERNER, G. 2020. Comprehensive integrated NGS-based surveillance and contact-network modeling unravels transmission dynamics of vancomycin-resistant enterococci in a high-risk population within a tertiary care hospital. *PloS one,* 15**,** e0235160-e0235160.

NIMWEGEN, K. J. M. V., SOEST, R. A., VELTMAN, J. A., NELEN, M. R., WILT, G. J. V. D., PEART-VISSERS, L. E. L. M. & GRUTTERS, J. P. C. 2016. Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. *Clinical chemistry (Baltimore, Md.),* 62**,** 1458-1464.

NOUSIAS, O. & MONTESANTO, F. 2021. Metagenomic profiling of host-associated bacteria from 8 datasets of the red alga Porphyra purpurea with MetaPhlAn3. *Marine genomics,* 59**,** 100866-100866.

NURK, S., MELESHKO, D., KOROBEYNIKOV, A. & PEVZNER, P. A. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome research,* 27**,** 824-834.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C. M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V. S., KODALI, V. K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K. M., MURPHY, M. R., O'NEILL, K., PUJAR, S., RANGWALA, S. H., RAUSCH, D., RIDDICK, L. D., SCHOCH, C., SHKEDA, A., STORZ, S. S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R. E., VATSAN, A. R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M. J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T. D. & PRUITT, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research,* 44**,** D733-D745.

OLIVER, J. D., NILSSON, L. & KJELLEBERG, S. 1991. Formation of nonculturable Vibrio vulnificus cells and its relationship to the starvation state. *Applied and Environmental Microbiology,* 57**,** 2640-2644.

OYOLA, S. O., HENSON, S. P., NZAU, B., KIBWANA, E. & NENE, V. 2021. Access to ultra-long IgG CDRH3 bovine antibody sequences using short read sequencing technology. *Molecular immunology,* 139**,** 97-105.

PALMER, C., BIK, E. M., EISEN, M. B., ECKBURG, P. B., SANA, T. R., WOLBER, P. K., RELMAN, D. A. & BROWN, P. O. 2006. Rapid quantitative profiling of complex microbial populations. *Nucleic acids research,* 34**,** e5-e5.

PANDA, K. & SLOTKIN, R. K. 2020. Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of Transposable Elements. *The Plant cell,* 32**,** 2687-2698.

PFEIFFER, F., GRÖBER, C., BLANK, M., HÄNDLER, K., BEYER, M., SCHULTZE, J. L. & MAYER, G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific reports,* 8**,** 10950-14.

POPOWITCH, E. B., O'NEILL, S. S. & MILLER, M. B. 2013. Comparison of the Biofire FilmArray RP, Genmark eSensor RVP, Luminex xTAG RVPv1, and Luminex xTAG RVP Fast Multiplex Assays for Detection of Respiratory Viruses. *Journal of Clinical Microbiology,* 51**,** 1528-1533.

POULSEN, C. S., EKSTRØM, C. T., AARESTRUP, F. M. & PAMP, S. J. 2022. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology spectrum,* 10**,** e0009022-e0009022.

QUER, J., COLOMER-CASTELL, S., CAMPOS, C., ANDRÉS, C., PIÑANA, M., CORTESE, M. F., GONZÁLEZ-SÁNCHEZ, A., GARCIA-CEHIC, D., IBÁÑEZ, M., PUMAROLA, T., RODRÍGUEZ-FRÍAS, F., ANTÓN, A. & TABERNERO, D. 2022. Next-Generation Sequencing for Confronting Virus Pandemics. *Viruses,* 14**,** 600.

ROCKETT, R. J., ARNOTT, A., LAM, C., SADSAD, R., TIMMS, V., GRAY, K.-A., EDEN, J.-S., CHANG, S., GALL, M., DRAPER, J., SIM, E. M., BACHMANN, N. L., CARTER, I., BASILE, K., BYUN, R., O'SULLIVAN, M. V., CHEN, S. C. A., MADDOCKS, S., SORRELL, T. C., DWYER, D. E., HOLMES, E. C., KOK, J., PROKOPENKO, M. & SINTCHENKO, V. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine,* 26**,** 1398-1404.

ROSZAK, D. B., GRIMES, D. J. & COLWELL, R. R. 1984. Viable but nonrecoverable stage of Salmonella enteritidis in aquatic systems. *Canadian journal of microbiology,* 30**,** 334-338.

ROTA, P. A., OBERSTE, M. S., MONROE, S. S., NIX, W. A., CAMPAGNOLI, R., ICENOGLE, J. P., PEÑARANDA, S., BANKAMP, B., MAHER, K., CHEN, M., TONG, S., TAMIN, A., LOWE, L., FRACE, M., DERISI, J. L., CHEN, Q., WANG, D., ERDMAN, D. D., PERET, T. C. T., BURNS, C., KSIAZEK, T. G., ROLLIN, P. E., SANCHEZ, A., LIFFICK, S., HOLLOWAY, B., LIMOR, J., MCCAUSTLAND, K., OLSEN-RASMUSSEN, M., FOUCHIER, R., GÜNTHER, S., OSTERHAUS, A., STEN, C., PALLANSCH, M. A., ANDERSON, L. J. & BELLINI, W. J. 2003. Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *Science (American Association for the Advancement of Science),* 300**,** 1394-1399.

SANDERS, J. G., NURK, S., SALIDO, R. A., MINICH, J., XU, Z. Z., ZHU, Q., MARTINO, C., FEDARKO, M., ARTHUR, T. D., CHEN, F., BOLAND, B. S., HUMPHREY, G. C., BRENNAN, C., SANDERS, K., GAFFNEY, J., JEPSEN, K., KHOSROHEIDARI, M., GREEN, C., LIYANAGE, M., DANG, J. W., PHELAN, V. V., QUINN, R. A., BANKEVICH, A., CHANG, J. T., RANA, T. M., CONRAD, D. J., SANDBORN, W. J., SMARR, L., DORRESTEIN, P. C., PEVZNER, P. A. & KNIGHT, R. 2019. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology,* 20**,** 226-226.

SAYERS, E. W., CAVANAUGH, M., CLARK, K., OSTELL, J., PRUITT, K. D. & KARSCH-MIZRACHI, I. 2019. GenBank. *Nucleic acids research,* 47**,** D94-D99.

SCHOONVAERE, K., DE SMET, L., SMAGGHE, G., VIERSTRAETE, A., BRAECKMAN, B. P. & DE GRAAF, D. C. 2016. Unbiased RNA Shotgun Metagenomics in Social and Solitary Wild Bees Detects Associations with Eukaryote Parasites and New Viruses. *PloS one,* 11**,** e0168456-e0168456.

SERPA, P. H., DENG, X., ABDELGHANY, M., CRAWFORD, E., MALCOLM, K., CALDERA, S., FUNG, M., MCGEEVER, A., KALANTAR, K. L., LYDEN, A., GHALE, R., DEISS, T., NEFF, N., MILLER, S. A., DOERNBERG, S. B., CHIU, C. Y., DERISI, J. L., CALFEE, C. S. & LANGELIER, C. R. 2022. Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. *Genome medicine,* 14**,** 1-74.

SHAFFER, J. P., CARPENTER, C. S., MARTINO, C., SALIDO, R. A., MINICH, J. J., BRYANT, M., SANDERS, K., SCHWARTZ, T., HUMPHREY, G., SWAFFORD, A. D. & KNIGHT, R. 2022. A comparison of six DNA extraction protocols for 16S, ITS and shotgun metagenomic sequencing of microbial communities. *BioTechniques,* 73**,** 34-46.

SHAKYA, M., LO, C.-C. & CHAIN, P. S. G. 2019. Advances and Challenges in Metatranscriptomic Analysis. *Frontiers in genetics,* 10**,** 904-904.

SHARMA, N., GAUTAM, H., TYAGI, S., RAZA, S., MOHAPATRA, S., SOOD, S., DHAWAN, B., KAPIL, A. & DAS, B. 2022. Clinical use of multiplex-PCR for the diagnosis of acute bacterial meningitis. *Journal of family medicine and primary care,* 11**,** 593-598.

SIMNER, P. J., MILLER, S. & CARROLL, K. C. 2018. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clinical infectious diseases,* 66**,** 778-788.

SINGH, N., CHIKARA, S. & SUNDAR, S. 2013. SOLiD™ sequencing of genomes of clinical isolates of Leishmania donovani from India confirm leptomonas co-infection and raise some key questions. *PloS one,* 8**,** e55738-e55738.

SOTIRIOU, C. & PUSZTAI, L. 2009. Gene-Expression Signatures in Breast Cancer. *The New England journal of medicine,* 360**,** 790-800.

TULLOCH, R. L., KOK, J., CARTER, I., DWYER, D. E. & EDEN, J.-S. 2021. An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus. *Viruses,* 13**,** 499.

TYSON, J. R., JAMES, P., STODDART, D., SPARKS, N., WICKENHAGEN, A., HALL, G., CHOI, J. H., LAPOINTE, H., KAMELIAN, K., SMITH, A. D., PRYSTAJECKY, N., GOODFELLOW, I., WILSON, S. J., HARRIGAN, R., SNUTCH, T. P., LOMAN, N. J. & QUICK, J. 2020. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*.

UWAMINO, Y., NAGATA, M., AOKI, W., NAKAGAWA, T., INOSE, R., YOKOTA, H., FURUSAWA, Y., SAKAI-TAGAWA, Y., IWATSUKI-HORIMOTO, K., KAWAOKA, Y., HASEGAWA, N. & MURATA, M. 2021. Accuracy of rapid antigen detection test for nasopharyngeal swab specimens and saliva samples in comparison with RT-PCR and viral culture for SARS-CoV-2 detection. *Journal of infection and chemotherapy : official journal of the Japan Society of Chemotherapy,* 27**,** 1058-1062.

VALENCIA, C. A., PERVAIZ, M. A., HUSAMI, A., QIAN, Y. & ZHANG, K. 2013. *Next Generation Sequencing Technologies in Medical Genetics,* Dordrecht, Springer.

VAN BOHEEMEN, S., VAN RIJN, A. L., PAPPAS, N., CARBO, E. C., VORDERMAN, R. H. P., SIDOROV, I., VAN `T HOF, P. J., MEI, H., CLAAS, E. C. J., KROES, A. C. M. & DE VRIES, J. J. C. 2020. Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients. *The Journal of molecular diagnostics : JMD,* 22**,** 196-207.

VAN DER PALEN, M., GILLET, P., BOTTIEAU, E., CNOPS, L., VAN ESBROECK, M. & JACOBS, J. 2009. Test characteristics of two rapid antigen detection tests (SD FK50 and SD FK60) for the diagnosis of malaria in returned travellers. *Malaria journal,* 8**,** 90-90.

WALSH, L. H., COAKLEY, M., WALSH, A. M., O'TOOLE, P. W. & COTTER, P. D. 2022. Bioinformatic approaches for studying the microbiome of fermented food. *Critical reviews in microbiology,* ahead-of-print**,** 1-33.

WARD, C. M., TO, T.-H. & PEDERSON, S. M. 2020. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. *Bioinformatics,* 36**,** 2587-2588.

WATTAM, A. R., ABRAHAM, D., DALAY, O., DISZ, T. L., DRISCOLL, T., GABBARD, J. L., GILLESPIE, J. J., GOUGH, R., HIX, D., KENYON, R., MACHI, D., MAO, C., NORDBERG, E. K., OLSON, R., OVERBEEK, R., PUSCH, G. D., SHUKLA, M., SCHULMAN, J., STEVENS, R. L., SULLIVAN, D. E., VONSTEIN, V., WARREN, A., WILL, R., WILSON, M. J. C., YOO, H. S., ZHANG, C., ZHANG, Y. & SOBRAL, B. W. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research,* 42**,** D581-D591.

WILSON, M. L. 2015. Diagnostic microbiology: the accelerating transition from culture-based to molecular-based methods. *American journal of clinical pathology,* 143**,** 766-767.

WILSON, M. R., NACCACHE, S. N., SAMAYOA, E., BIAGTAN, M., BASHIR, H., YU, G., SALAMAT, S. M., SOMASEKAR, S., FEDERMAN, S., MILLER, S., SOKOLIC, R., GARABEDIAN, E., CANDOTTI, F., BUCKLEY, R. H., REED, K. D., MEYER, T. L., SEROOGY, C. M., GALLOWAY, R., HENDERSON, S. L., GERN, J. E., DERISI, J. L. & CHIU, C. Y. 2014. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *The New England journal of medicine,* 370**,** 2408-2417.

WILSON, M. R., SAMPLE, H. A., ZORN, K. C., AREVALO, S., YU, G., NEUHAUS, J., FEDERMAN, S., STRYKE, D., BRIGGS, B., LANGELIER, C., BERGER, A., DOUGLAS, V., JOSEPHSON, S. A., CHOW, F. C., FULTON, B. D., DERISI, J. L., GELFAND, J. M., NACCACHE, S. N., BENDER, J., DIEN BARD, J., MURKEY, J., CARLSON, M., VESPA, P. M., VIJAYAN, T., ALLYN, P. R., CAMPEAU, S., HUMPHRIES, R. M., KLAUSNER, J. D., GANZON, C. D., MEMAR, F., OCAMPO, N. A., ZIMMERMANN, L. L., COHEN, S. H., POLAGE, C. R., DEBIASI, R. L., HALLER, B., DALLAS, R., MARON, G., HAYDEN, R., MESSACAR, K., DOMINGUEZ, S. R., MILLER, S. & CHIU, C. Y. 2019. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *The New England journal of medicine,* 380**,** 2327-2340.

WOOD, D. E. & SALZBERG, S. L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology,* 15**,** R46-R46.

XIAO, M., LIU, X., JI, J., LI, M., LI, J., YANG, L., SUN, W., REN, P., YANG, G., ZHAO, J., LIANG, T., REN, H., CHEN, T., ZHONG, H., SONG, W., WANG, Y., DENG, Z., ZHAO, Y., OU, Z., WANG, D., CAI, J., CHENG, X., FENG, T., WU, H., GONG, Y., YANG, H., WANG, J., XU, X., ZHU, S., CHEN, F., ZHANG, Y., CHEN, W., LI, Y. & LI, J. 2020. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome medicine,* 12**,** 57-57.

YANG, L., HAIDAR, G., ZIA, H., NETTLES, R., QIN, S., WANG, X., SHAH, F., RAPPORT, S. F., CHARALAMPOUS, T., METHÉ, B., FITCH, A., MORRIS, A., MCVERRY, B. J., O'GRADY, J. & KITSIOS, G. D. 2019. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respiratory research,* 20**,** 265-12.

YE, Y. & TANG, H. 2016. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics,* 32**,** 1001-1008.

YOUNG, S., TAYLOR, S. N., CAMMARATA, C. L., VARNADO, K. G., ROGER-DALBERT, C., MONTANO, A., GRIEGO-FULLBRIGHT, C., BURGARD, C., FERNANDEZ, C., ECKERT, K., ANDREWS, J. C., REN, H., ALLEN, J., ACKERMAN, R. & COOPER, C. K. 2020. Clinical Evaluation of BD Veritor SARS-CoV-2 Point-of-Care Test Performance Compared to PCR-Based Testing and versus the Sofia 2 SARS Antigen Point-of-Care Test. *Journal of clinical microbiology,* 59.

YU, P., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-1428.

YUAN, G., WANG, H., ZHAO, Y., MAO, E., LI, M., WANG, R., ZHOU, F., JIN, S., ZHANG, Z., XU, K., XU, J., LIANG, S., LI, X., JIANG, L., ZHANG, L., SONG, J., YANG, T., GUO, J., ZHANG, H., ZHOU, Y., WANG, S., QIU, C., JIANG, N., AI, J., WU, J. & ZHANG, W. 2022. Early identification and severity prediction of acute respiratory infection : a study protocol for a randomized controlled trial. *BMC infectious diseases,* 22**,** 1-632.

ZINTER, M. S., DVORAK, C. C., MAYDAY, M. Y., IWANAGA, K., LY, N. P., MCGARRY, M. E., CHURCH, G. D., FARICY, L. E., ROWAN, C. M., HUME, J. R., STEINER, M. E., CRAWFORD, E. D., LANGELIER, C., KALANTAR, K., CHOW, E. D., MILLER, S., SHIMANO, K., MELTON, A., YANIK, G. A., SAPRU, A. & DERISI, J. L. 2019. Pulmonary Metagenomic Sequencing Suggests Missed Infections in Immunocompromised Children. *Clinical infectious diseases,* 68**,** 1847-1855.

ZUMLA, A. P., AL-TAWFIQ, J. A. M. D., ENNE, V. I. P., KIDD, M. P., DROSTEN, C. P., BREUER, J. P., MULLER, M. A. P., HUI, D. P., MAEURER, M. P., BATES, M. P., MWABA, P. P., AL-HAKEEM, R. M. D., GRAY, G. P., GAUTRET, P. P., AL-RABEEAH, A. A. F., MEMISH, Z. A. P. & GANT, V. F. 2014. Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections—needs, advances, and future prospects. *The Lancet infectious diseases,* 14**,** 1123-1135.

APPENDIX A

Publications Arising from Collaborative Research

Appendix A: Publications Arising from Collaborative Research

Appendix A1: Off-Season RSV Epidemics in Australia after Easing of COVID-19 Restrictions

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-30485-3 OPEN

Off-season RSV epidemics in Australia after easing of COVID-19 restrictions

John-Sebastian Eden[1,2,20], Chisha Sikazwe[3,4,20], Ruopeng Xie[5,6,20], Yi-Mo Deng[7,8], Sheena G. Sullivan[7,9], Alice Michie[4], Avram Levy[3], Elena Cutmore[1,2], Christopher C. Blyth[3,10,11,12], Philip N. Britton[2,13], Nigel Crawford[14,15,16], Xiaomin Dong[7,8], Dominic E. Dwyer[2,17], Kimberly M. Edwards[5,6], Bethany A. Horsburgh[1,2], David Foley[3], Karina Kennedy[18], Cara Minney-Smith[3], David Speers[3,10], Rachel L. Tulloch[1,2], Edward C. Holmes[2], Vijaykrishna Dhanasekaran[5,6,21✉], David W. Smith[3,10,21✉], Jen Kok[17,21✉], Ian G. Barr[7,8,21✉] & the Australian RSV study group*

Human respiratory syncytial virus (RSV) is an important cause of acute respiratory infection with the most severe disease in the young and elderly. Non-pharmaceutical interventions and travel restrictions for controlling COVID-19 have impacted the circulation of most respiratory viruses including RSV globally, particularly in Australia, where during 2020 the normal winter epidemics were notably absent. However, in late 2020, unprecedented widespread RSV outbreaks occurred, beginning in spring, and extending into summer across two widely separated regions of the Australian continent, New South Wales (NSW) and Australian Capital Territory (ACT) in the east, and Western Australia. Through genomic sequencing we reveal a major reduction in RSV genetic diversity following COVID-19 emergence with two genetically distinct RSV-A clades circulating cryptically, likely localised for several months prior to an epidemic surge in cases upon relaxation of COVID-19 control measures. The NSW/ACT clade subsequently spread to the neighbouring state of Victoria and to cause extensive outbreaks and hospitalisations in early 2021. These findings highlight the need for continued surveillance and sequencing of RSV and other respiratory viruses during and after the COVID-19 pandemic, as mitigation measures may disrupt seasonal patterns, causing larger or more severe outbreaks.

A full list of author affiliations appears at the end of the paper.

Each year respiratory syncytial virus (RSV) causes an estimated 3.2 million hospital admissions and 118,200 deaths in children under five years of age, predominantly in low- and middle-income countries[1]. While this burden is greatest in the young, RSV is clinically significant for all age groups, as re-infection can occur throughout life[2]. The elderly and immunocompromised are particularly at risk of severe infection with intensive care admission and mortality rates similar to that of influenza, posing a considerable threat to residents of long-term care facilities[3,4]. RSV causes seasonal epidemics in both tropical and temperate regions of the world[5]. In Australia, most temperate regions experience seasonal RSV outbreaks during the autumn and winter, often peaking in June–July[6] and usually preceding the influenza season[7]. In the more tropical northern parts of Australia, RSV activity correlates with the rainfall and humidity patterns of the rainy season from December to March[8].

Non-pharmaceutical interventions (NPIs) to limit the spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have disrupted the typical seasonality of other common respiratory pathogens in many countries[9,10]. Australia's initial SARS-CoV-2 epidemic was effectively controlled by NPIs[11], and those same restrictions also suppressed seasonal respiratory virus circulation, particularly for influenza virus and RSV, as the usual winter epidemics were notably absent during 2020[12–14]. While the control measures of each Australian state and territory varied in stringency and duration[15], all occurred throughout the usual peak of RSV seasonal activity. Interestingly, the impact of NPIs was not consistent across all the common respiratory viruses: rhinoviruses and, to a lesser extent, adenoviruses continued to circulate in Australia during this pandemic period after an initial disruption to usual circulation[14]. The suppression of influenza virus and RSV activity during the COVID-19 pandemic in the southern hemisphere was also seen in South Africa[16] and New Zealand[17], where similarly following an initial reduction in circulation, RSV activity rebounded in late 2020[18] and early 2021, respectively. Marked reductions in RSV activity have also been seen in the northern hemisphere since early 2020, although some European countries such as France[19], Iceland[20], as well as in Israel[21], and some US states[22] have recently reported out-of-season spikes in RSV activity in early-mid 2021. There has also been a dearth of RSV sequences submitted to public databases since early 2020, including these most recent outbreaks presumably a reflection of the lack of RSV circulation in many countries and a focus on SARS-CoV-2 sequencing.

Here, we show a major shift in the epidemiology of RSV in Australia following the emergence of SARS-CoV-2 with large scale outbreaks of disease occurring out-of-season during the summer of 2020–21. These outbreaks across three states coincided with the easing of COVID-19-related restrictions and the emergence of two novel yet related RSV-A ON1-like lineages. Using genome sequencing, we demonstrate a remarkable collapse in circulating RSV genetic diversity and uncertainty around the origins and timings of the outbreak strains due to limited global sampling. This work highlights the shifting landscape of respiratory virus molecular epidemiology in the post-COVID-19 world and the rapid pace at which activity can rebound once public health restrictions have eased. Furthermore, this work underscores the urgent need for on-going surveillance for RSV, influenza virus and other major respiratory pathogens to examine changes in their genetic diversity, particularly towards informing vaccine compositions.

Results and discussion

Summer outbreaks of RSV after COVID-19 restriction easing. In late 2020, severe out-of-season RSV outbreaks occurred in several Australian states and territories beginning in New South Wales and the Australian Capital Territory (NSW/ACT) and Western Australia (WA). This was followed by outbreaks in Victoria (VIC) throughout the summer in early 2021. To understand the change in seasonal prevalence of RSV in Australia we examined RSV testing data from January 2017 to March 2021, comparing the proportion positive and overall testing capacity in NSW, WA, and VIC (Fig. 1 and Supplementary Figs. 1 and 2). Before 2020, RSV activity consistently began during mid-autumn (April–May) and normally persisted for six months with an epidemic peak in the middle of the Australian winter (middle of July, weeks 27–29). In contrast, RSV activity in 2020 occurred between six and nine months later than historically observed, and at the peak of RSV activity across each state and territory, laboratory-confirmed RSV positivity rates were considerably higher than those of the previous three seasons (Fig. 1 and Supplementary Fig. 2).

In Australia, the suppression of RSV activity in early 2020 coincided with restrictions in response to increasing community cases of COVID-19 (Fig. 1). During March 2020, individual Australian state and territory governments implemented a range of NPIs, which included limits on international arrivals and strict quarantine requirements for persons with COVID-19 (minimum of 14 days), internal border closures, social distancing, school closures or encouraging parents to keep their children home, and hygiene protocols to minimise SARS-CoV-2 transmission. Importantly, and most relevant for RSV, childcare centres mostly remained open during these restriction periods. Indeed, during the 2020/21 RSV season in Europe, where RSV activity was overall very low, the only countries with major RSV outbreaks were those with policies to keep primary school and childcare centres open throughout lockdowns[20]. Prior to the implementation of COVID-19 control measures, a gradual increase in RSV activity was observed in early 2020 across all four Australian states and territories with laboratory test RSV positivity rates being comparable to monthly averages over the previous three seasons (Supplementary Figs. 1 and 2). Introduction of NPIs led to a rapid decline in RSV incidence in each state. Examination of sentinel hospital records for bronchiolitis by ICD-10 Australian modification (AM) codes (including both RSV-confirmed bronchiolitis and bronchiolitis of unknown cause) showed a marked decline that mirrored the decrease in laboratory-confirmed RSV (Supplementary Fig. 1). The subsequent RSV epidemics in late 2020 and early 2021 resulted in test positivity and ICD-10 AM admission levels equivalent to or exceeding the normal winter seasonal RSV activity seen in any of the previous three years. For NSW, the epidemic began in September 2020 with bimodal peaks in activity in mid-November (reaching 26% positive) and early January (reaching 24% positive) (Fig. 1). The dual peaks likely reflect inconsistent testing over the Christmas and New Year holiday period, as the peak in bronchiolitis hospitalisations in NSW occurred between late December and early January and coincided with this period (Fig. 1). Furthermore, a December 2020 peak was also observed in the ACT, which is a territory within the state of NSW, and where 46% of tests were positive in this period (Supplementary Fig 2). For WA, the RSV epidemic began in late September 2020 and peaked in December 2020 at 37% positivity, with a matching peak in bronchiolitis hospitalisations (Fig. 1 and Supplementary Fig. 1).

Over the course of the pandemic, the stringency of COVID-19 restrictions has varied across the different states and territories. In VIC, a second wave of COVID-19 from July to August 2020 necessitated a longer SARS-CoV-2 control period, which likely contributed to a three-month delay in the onset of RSV activity relative to epidemic outbreaks in NSW/ACT and WA. RSV activity in VIC began in early January 2021 and then peaked in

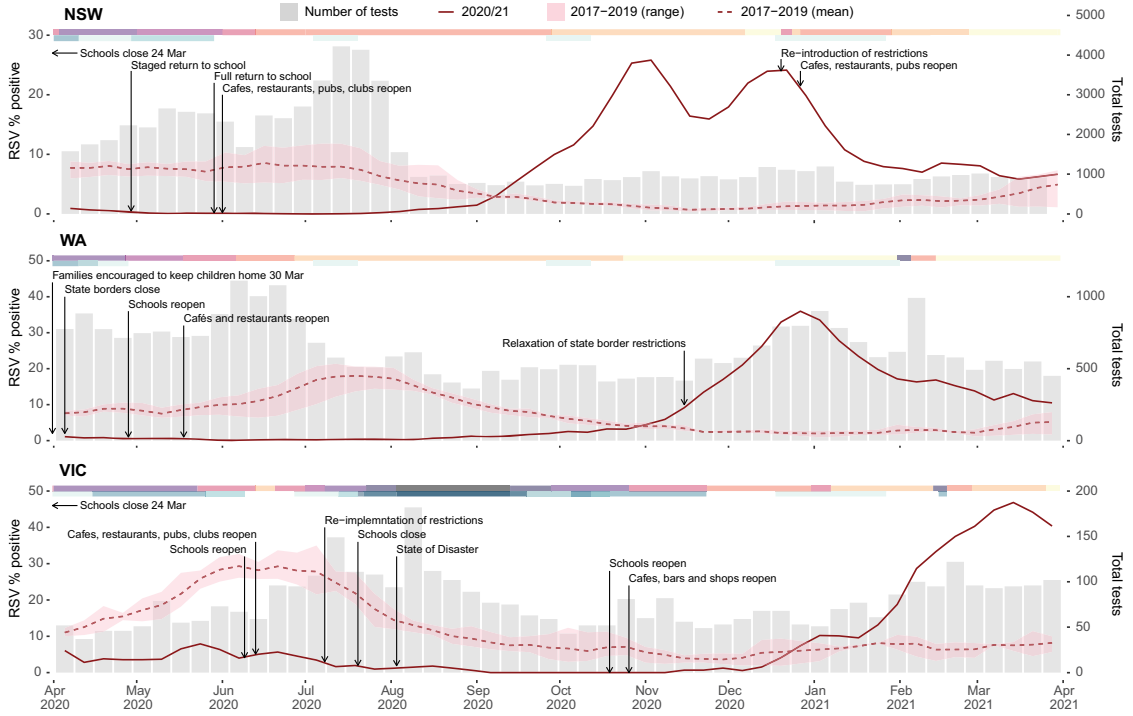


Fig. 1 The epidemiology of RSV detections in three Australian states—New South Wales (NSW), Western Australia (WA), and Victoria (VIC). Laboratory testing for RSV in 2020 as weekly percent positive (red line, left *y*-axis) and as total number of tests performed (grey bars, right *y*-axis). In each panel, the dashed red line represents mean monthly RSV percent positive over the three previous seasons, and corresponding red shading represents minimum and maximum weekly percent positive. Pink-shaded bars across the top of each plot indicate the severity of pandemic restrictions, with darker colours indicative of greater stringency. Blue bars across the top of each plot indicate the periods during which students did not attend school either due to pandemic restrictions or school holiday periods, with darker colours indicative of more stringent school restrictions.

early March that year with 48% of tests positive (Fig. 1). In each state, the peaks in RSV activity occurred a few months after the relaxation of COVID-19 restrictions (Fig. 1).

Reduced genomic diversity in the post-COVID-19 period. Whole genome sequencing (WGS) was performed on RSV positive specimens collected before (July 2017–March 2020), during and after (April 2020–March 2021) the implementation of COVID-19 restrictions in NSW ($n = 253$), ACT ($n = 47$), WA ($n = 216$) and VIC ($n = 178$). The samples were mostly collected from young children (median ages between 0.78 and 2.34 years for the different states), although all age groups including adults and the elderly were represented. Sampling of gender was even and included geographically diverse locations (Supplementary Figs. 3 and 4). Historically in Australia, both RSV-A and -B subtypes have co-circulated with variable but relatively even prevalence[23,24]. This trend continued in the pre-COVID-19 period, where RSV-A represented 45–79% of cases. However, from late 2020 to early 2021 there was an overwhelming predominance of the RSV-A subtype (>95% for all four states and territories). This suggested that RSV-A viruses were responsible for both the NSW/ACT and WA outbreaks in late 2020, as well as the surge in RSV activity in VIC seen in early 2021.

Phylogenetic analysis of all available RSV-A genomes revealed that the Australian RSV-A viruses belonged to the ON1-like genotype first reported in Canada in December 2010[25]. These viruses have since become globally dominant and have frequently been re-introduced into Australia[23,26]. Indeed, prior to the emergence of SARS-CoV-2 and the implementation of related control measures, multiple RSV-A ON1-like sub-lineages co-circulated (Fig. 2) with genetic diversity sustained from both endemic and imported sources[23]. While the viruses sampled before March 2020 were distributed amongst those circulating globally, RSV-A viruses circulating after the post-COVID-19 onset formed two geographically distinct monophyletic lineages (Fig. 2A). One lineage was associated with cases from NSW, ACT, and VIC, while the other was associated with cases from WA, hereafter referred to as the NSW/ACT/VIC 2020 and WA 2020 lineages, respectively (Fig. 2B). The NSW, ACT, and WA outbreaks during late 2020 occurred at a time of minimal RSV activity in other Australian states including VIC (Fig. 1 and Supplementary Fig. 1). Genomic analysis showed that cases in VIC in early 2021 were associated with multiple importations of the NSW/ACT/VIC 2020 lineage and a small number of importations of WA 2020 lineage (Fig. 2A, B).

Notably, both genetic lineages were defined by several non-synonymous changes in the genome, most of which were under significant selective pressure as measured by Contrast-FEL[27] (Fig. 2B, bold residues and Supplementary Table 2). In the WA 2020 lineage, five amino acid changes were observed under selection in the glycoprotein (T80A, T129I, and S174N), nucleocapsid (I104F) and small hydrophobic (H51Q) proteins, while 11 such lineage defining changes were identified in the NSW/ACT/VIC 2020 virus glycoprotein localised to the C-terminus region (V225A, E263Q, L265P, Y273H, S277P, S291P, Y297H, and L316P), large polymerase protein (D755G)

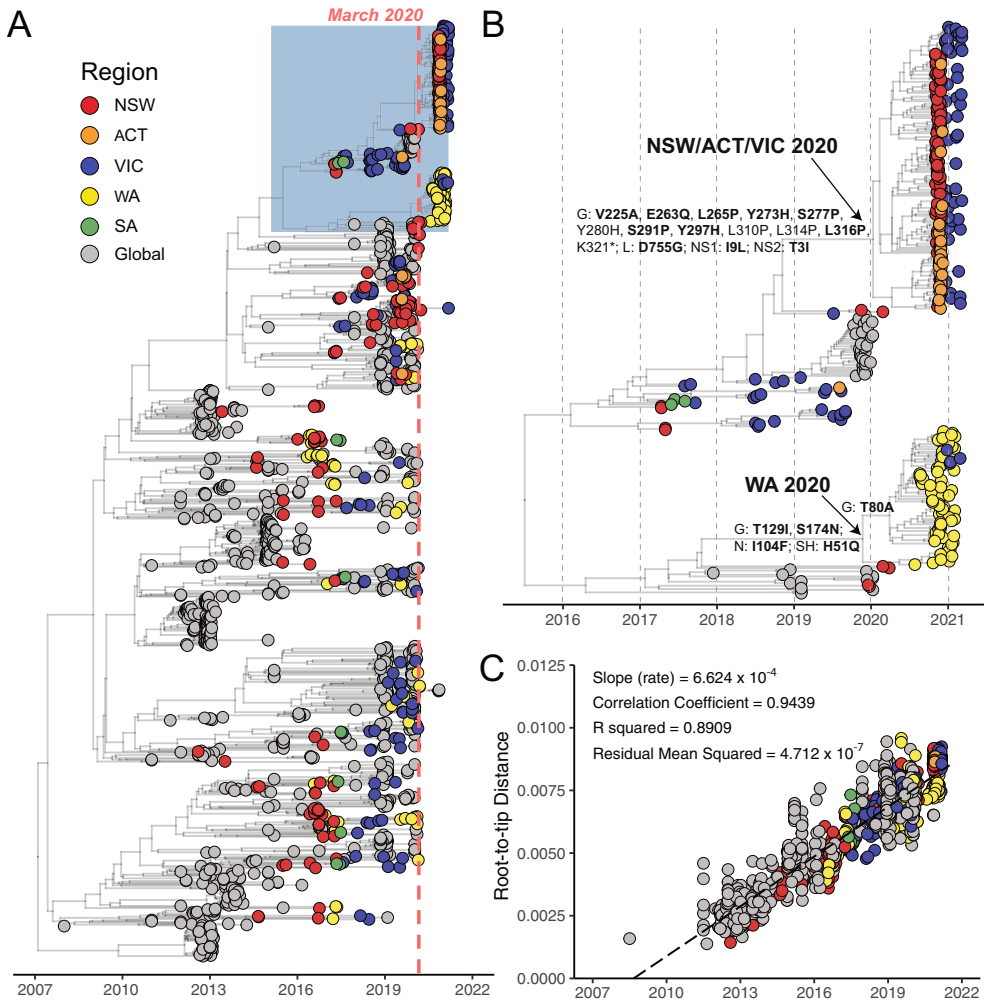


Fig. 2 Phylogenetic analysis of global and Australian RSV-A genome sequences. A RSV-A genome sequences were aligned with NCBI GenBank reference sequences and analysed using a Bayesian molecular clock approach estimated with BEAST (v1.10) focused on recent ON1-like viruses. Australian states—New South Wales (NSW), Australia Capital Territory (ACT), Western Australia (WA) and Victoria (VIC), and South Australia (SA)—and globally-derived sequences are coloured according to the key provided. The light red dotted line marks March 2020 and the beginning of extensive COVID-19-related restrictions. The blue shaded box is expanded in panel (**B**), which is a focused analysis of NSW/ACT/VIC and WA 2020 lineages. Amino acid mutations inferred by TreeTime are labelled on select branches, and those under significant selection pressure are shown in bold. **C** Temporal signal in RSV-A genomic dataset determined by linear-regression of root-to-tip distance (*y*-axis) against sample collection date (*x*-axis).

and non-structural proteins (NS1: I9L, NS2: T3I), most of which do not appear to have been previously reported. Despite this, an analysis using aBSREL[28], a model that tests phylogenetic branches for episodic diversifying selection, showed that branch-wise selection of the two outbreak lineages was not significant in comparison to a subset of global background sequences.

We then used RELAX[29] to further test the strength of natural selection along the outbreak branches, and found selection intensifying on NSW/ACT/VIC 2020 and WA 2020 branches (intensity parameter $k > 2$; Supplementary Table 3) in comparison to the global RSV background (i.e., sites under moderate purifying selection in the outbreak branches were subject to stronger purifying selection than non-outbreak branches). Approximately 0.25% of sites in the WA 2020 clade were under positive selection, whereas the NSW/ACT/VIC 2020 clade contained more than 10 times as many sites (2.60–7.88%) under positive selection (Supplementary Table 3).

Cryptic origins of the novel RSV-A lineages linked to outbreaks. To maximise spatiotemporal sampling of the RSV-A ON1-like viruses, we expanded our phylogenetic analysis to include all available RSV glycoprotein (G) sequences (Fig. 3A), which are more numerous than published whole RSV genomes. The genome-based phylogeny contained 1520 RSV-A ON1-like genomes (Fig. 2A), and an additional 3527 sequences were included in the G gene phylogenetic analysis (Fig. 3). Despite the additional sequences, the G gene phylogeny found that the viruses from the Australian 2020–21 epidemics did not cluster with any other RSV-A viruses sampled nationally or internationally to date. As such, the initial source of these two novel RSV-A lineages

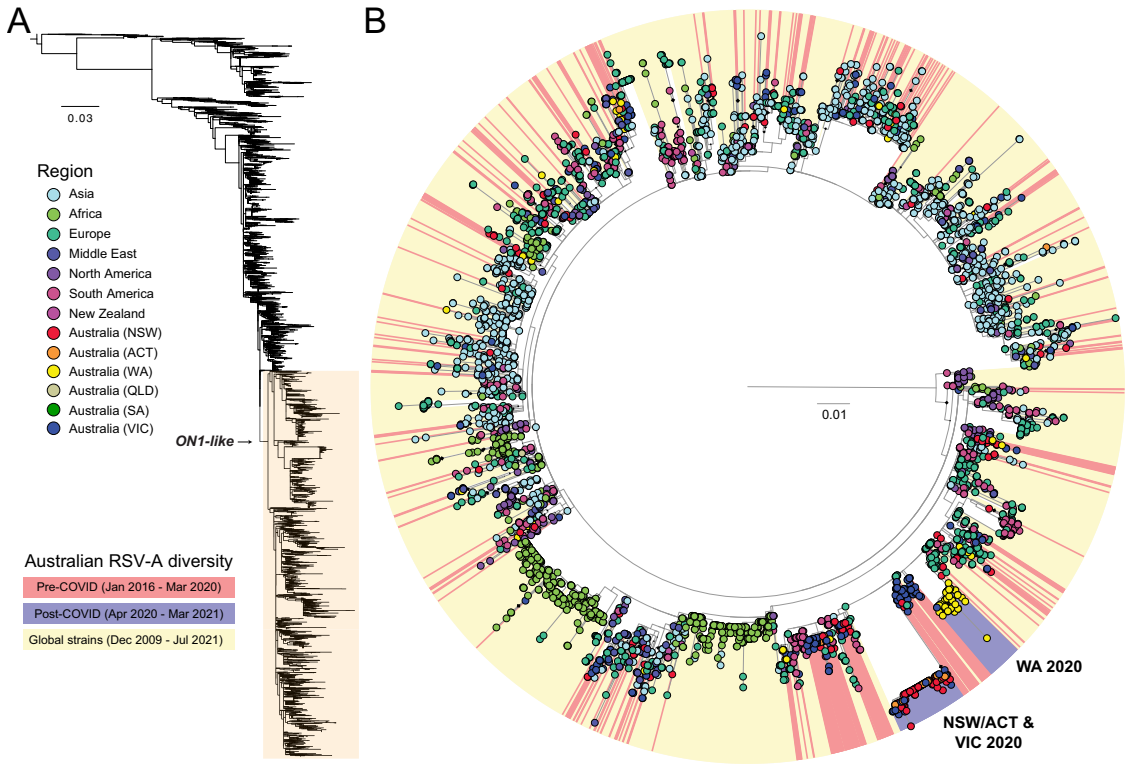


Fig. 3 Phylogenetic analysis of global and Australian RSV-A glycoprotein sequences. A RSV-A sequences in this study were aligned with all available RSV-A sequences from NCBI GenBank. The glycoprotein coding region was extracted, and sequences less than 300 nt were removed. **B** A detailed examination of recently circulating ON1-like viruses showed only two pre-COVID-19 lineages (coloured red) survived into the post-COVID-19 period (blue). These two lineages were associated with outbreaks in NSW/ACT and WA in late 2020, and VIC in early 2021. No sequences sourced globally (yellow) were found to be related to the these lineages, suggesting the sources remain unknown. Australian states—New South Wales (NSW), Australia Capital Territory (ACT), Western Australia (WA) and Victoria (VIC), South Australia (SA) and Queensland (QLD)—and globally-derived sequences are coloured according to the key provided. Diamonds at nodes indicate bootstrap support values >70%. Branches are proportional to the number of nucleotide substitutions per site.

remains undetermined (Fig. 3A, B). This most likely reflects a lack of RSV genomic surveillance during the COVID-19 pandemic, where clear gaps remain in sampling (Supplementary Fig. 5). However, like the genome-scale analysis, the G gene phylogeny revealed a major genetic bottleneck through which the genetic diversity of other ON1-like lineages had co-circulated from 2016 and 2020 were largely eliminated during these outbreak periods (Fig. 3B). We also examined RSV-B diversity, and while the number of detections were very low in the post-COVID-19 period, a similar pattern was observed to RSV-A diversity, whereby previously established lineages were mostly absent, and a single lineage was dominant during the 2020–21 outbreak (Supplementary Fig 6). Taken together, these results illustrate a remarkable collapse in the genetic diversity of RSV in Australia during the implementation of COVID-19 related restrictions (Fig. 3B).

Phylogenetic analysis showed there was sufficient genetic diversity within the outbreak samples to indicate the circulation of these viruses in NSW and WA prior to implementation of COVID-19 restrictions (Fig. 2B). The NSW/ACT/VIC 2020 and WA 2020 genomes contained extensive rate variation ($R^2 < 0.2$) and did not conform to temporality in a root-to-tip regression (Supplementary Fig. 7), which is essential for accurate estimation of divergence times[30]. Hence, we inferred the time to most recent common ancestor (tMRCA) of NSW/ACT/VIC and WA 2020 lineages from the large Bayesian molecular clock analysis presented in Fig. 2A, B, followed by a post hoc correction for excess deleterious mutations[31] and/or selection pressure (Supplementary Tables 3 and 4), as described previously[32]. The corrected tMRCA estimates suggest origins around March 2020 (Table 1), with confidence intervals spanning the period immediately prior to implementation of pandemic restrictions in Australia and globally (Fig. 1). Surveillance data support this inference for the WA 2020 lineage, which was first detected in central and southern non-metropolitan regional WA from late July to September, then in the Perth metropolitan area in October (Supplementary Fig. 4), indicating low-level regional circulation and rapid spread soon after introduction into areas of high population density, in this case, metropolitan Perth. The genomic data also showed two transcontinental transmissions from WA to VIC after November 2020, at a time when interstate travel bans had eased (Figs. 2 and 3). In contrast, the origin and early dissemination of the NSW/ACT/VIC 2020 lineage was less certain.

Despite the widespread occurrence and relatively high genetic diversity of outbreaks in NSW and ACT (Figs. 2 and 3B), no precursor virus(es) of the same lineage were detected prior to the outbreak onset, despite efforts to identify and sequence cases from the low-activity period in early-mid 2020. In addition to a

Table 1 Mean time of most recent common ancestors (tMRCA) for NSW/ACT/VIC 2020 and WA 2020 clades estimated by different methods.

Method	tMRCA (NSW/ACT/VIC)[a]	tMRCA (WA)[a]	Evolution rate (per site per year)
Beast	2020-01-10 [2019-11-05, 2020-03-16]	2019-11-23 [2019-08-08, 2020-02-20]	6.624×10^{-4}
Beast-SLAC[b]	2020-05-03 [2020-03-16, 2020-06-20]	2020-06-03 [2020-04-01, 2020-07-25]	NSW/ACT/VIC: 9.092×10^{-4} WA: 1.124×10^{-3}
Beast-SNAP[b]	2020-03-25 [2020-01-31, 2020-05-18]	2020-05-06 [2020-02-26, 2020-07-03]	NSW/ACT/VIC: 8.070×10^{-4} WA: 1.021×10^{-3}
Beast-Contrast-FEL[b]	2020-05-08 [2020-03-22, 2020-06-25]	2020-06-01 [2020-03-30, 2020-07-24]	NSW/ACT/VIC: 9.245×10^{-4} WA: 1.117×10^{-3}

[a]Ranges in brackets show confidence intervals (95% highest probability densities).
[b]tMRCA corrected according to *dN/dS* ratio.

lack of RSV positive samples from the middle of 2020, our analysis was hampered by an inflated evolutionary rate due to recent sampling[33], multiple novel non-synonymous substitutions in the G gene, and variable evolutionary rates which complicated reliable phylogenetic dating estimates. As per our estimates of the origins for NSW/ACT/VIC 2020 and WA 2020 lineages, the viruses were already likely to be present in the country during the early stages of the COVID-19 pandemic. However, the origins of these novel virus lineages into Australia remains unclear as the branches leading to the outbreaks represent an unsampled diversity of >1 year for NSW/ACT/VIC and >2 years for WA lineage prior to March 2020 (Table 1), compounded by major biases in recent RSV genomic surveillance (Supplementary Fig. 5).

An examination of RSV diversity in Australia before and after the implementation of COVID-related NPIs using WGS has shown a major collapse in lineages that circulated prior to 2020[23,26]. This coincided with the emergence of two distinct, but phylogenetically related RSV-A lineages associated with summer outbreaks in NSW/ACT and WA, respectively. Both lineages were subsequently imported into VIC, where the NSW/ACT lineage caused a major outbreak in early 2021. Our analysis suggests cryptic circulation of both of these lineages, while other RSV-A and -B lineages were largely eliminated through COVID-19 related NPIs. While the estimation of the source of the epidemic was hindered by a paucity of sequence data from other jurisdictions and globally, the genetic diversity observed during these outbreaks strongly indicates undetected local circulation prior to the widespread outbreak. Travel restrictions and other social distancing measures may have significantly reduced RSV transmission and slowed spread but were unable to eliminate RSV in metropolitan WA and in NSW, and after a considerable delay, a substantial out-of-season epidemic occurred.

The near absence and subsequent resurgence of RSV-A in Australia has provided a unique opportunity to increase our understanding about how RSV epidemics occur and to identify measures for better control of RSV and other respiratory viruses in the future. The outbreak branches were characterised by intensifying selection, and a distinct set of mutations with unsampled ancestral branches extending over 1 and 2 years, respectively. This indicates these were distinct genotypes that had evolved under different evolutionary constraints in comparison to the general RSV background; the reasons for which require further study. Furthermore, our study highlights how quickly respiratory pathogens can rebound, even leading to unseasonal epidemics. Delayed or forgone RSV seasons may increase the cohort of young children susceptible to RSV infection and increase the age of first infection leading to larger outbreaks of RSV when they do finally occur. Increasing the age of first infection may be expected to coincide with reduced hospitalisations given that RSV burden is most pronounced in infants less than 6 months old[34]; however, this has not been reflected in bronchiolitis admissions (Supplementary Fig 1), with peak admissions in WA and VIC higher than in prior seasons. By increasing the pool of susceptible children, including those with underlying risk factors such as congenital heart disease, extreme prematurity, or chronic lung disease, outbreaks may also be more severe with regard to hospitalisations and intensive care admissions. Recent modelling studies predict delayed and severe RSV outbreaks in the US during the 2021–2022 winter[9] but not early out-of-season outbreaks as was observed for some states[35]. Whether these large summer epidemics experienced in NSW/ACT, WA, and VIC have sufficiently reduced the susceptible population to result in a smaller than usual winter seasons going forward is uncertain. The early winter data of 2021 appears to support this premise in WA and VIC and possibly to a lesser extent in NSW/ACT.

It also remains unclear how long it will take for normal winter RSV seasonality to resume in Australia and globally. The H1N1 2009 influenza pandemic impacted respiratory virus circulation for a number of years[36]. The findings from this study exemplify the need to be prepared for the occurrence of large outbreaks of RSV outside of normal seasonal periods and for health systems to be prepared to combat future severe RSV outbreaks. It also raises important questions as to how the epidemiological and evolutionary dynamics of RSV outbreaks might inform the re-emergence of influenza virus, which is still expected, and given the smaller role children play in population-scale transmission of influenza compared to RSV, may require the re-opening of international borders to import the influenza variants required to effectively seed new local outbreaks. Nonetheless, our study highlights the propensity for COVID-19-related NPIs to cause immense disruption in seasonal patterns of respiratory virus circulation and evolution. Furthermore, this study provides a timely warning to countries emerging from pandemic restrictions: the burden of disease from other respiratory pathogens such as RSV, may have all but disappeared and will likely rebound in the near future, possibly at unusual times and with stronger impact.

Methods

RSV surveillance and epidemiology. Respiratory specimen testing with quantitative RT-PCR assays was performed at six sites including (i) NSW Health Pathology—Institute of Clinical Pathology and Microbiology Research (ICPMR), Westmead, NSW, (ii) The Children's Hospital at Westmead, NSW, (iii) PathWest Laboratory Medicine WA, Perth, WA[37], (iv) The Royal Children's Hospital, Melbourne, VIC, (v) Monash Pathology, Monash Medical Centre, Clayton, VIC and (vi) ACT Pathology, Canberra, ACT. NSW Health Pathology—ICPMR and PathWest laboratories are both major diagnostic hubs that provide state-wide testing for respiratory viruses in NSW and WA, respectively. Monash Pathology provides services for all ages for a region of Melbourne, while the Children's

Hospital at Westmead and Royal Children's Hospitals are major metropolitan hospitals in NSW and VIC, respectively. ACT Pathology provides diagnostic services for all adult and paediatric hospital emergency department presentations and a proportion of outpatient community requests for the ACT. Weekly counts for RSV testing were collated for the period January 2017–March 2021, and derived from three laboratories: PathWest Laboratory in Perth, WA, NSW Health Pathology—ICPMR in Sydney, NSW, ACT Pathology, Canberra, ACT, and the Bio21 Royal Children's Hospital in Melbourne, VIC. PathWest, ICPMR, and ACT Pathology are public health laboratories testing children and adults across their respective states, whereas Bio21 only provided testing data for children receiving care at the Royal Children's Hospital in Melbourne, and therefore only includes results for children under 18 years of age. The proportion of tests that were RSV positive was calculated and smoothed using a 3-week, centred moving average. Data were plotted in time series to compare observed RSV activity for April 2020–March 2021, versus the average for April 2017–March 2020. Monthly bronchiolitis admissions for the three children's hospitals in Perth, Sydney, and Melbourne were collated for the period January 2017–March 2021. Only admissions with a J21 ICD-10 AM code were considered. Prior work has shown that admissions for bronchiolitis are heavily represented by children aged <2 years. School and other restrictions for each state were collated from media releases and official public health directions issued in each of the three states. In addition, school holiday periods for all years January 2017–March 2021 were collated to visually assess the role of school holidays as a proxy for student mixing in RSV seasonality. The period between relaxation of restrictions and increased RSV activity was visually compared.

RSV subtyping and whole genome sequencing. Samples were sequenced from cases collected for routine diagnostic purposes as part of public health responses and from on-going research studies with approval from the local Human Research Ethics Committees of the Royal Children's Hospital and Western Sydney Local Health District with approval numbers 37185 and LNR/17/WMEAD/128, respectively. Total nucleic acid was extracted from RSV positive respiratory specimens archived at −80 °C using high-throughput bead-based protocols. RSV WGS was conducted using established protocols[24] for a subset of samples selected to provide temporal and geographical representation of (i) the pre-COVID-19 period, inclusive of July 2017–March 2020, and (ii) the post-COVID-19 period, inclusive of April 2020–March 2021. Briefly, viral cDNA was prepared from extracted nucleic acid using SuperScript IV VILO Master Mix or SuperScript IV (Invitrogen, Carlsbad, CA, USA), followed by RT-PCR amplification of four long overlapping fragments spanning the RSV genome using Platinum SuperFi Master Mix (Invitrogen). The four target amplicons were then combined equally before DNA purification with AMPure XP (Beckman Coulter, Indianapolis, IN, USA). The purified and pooled amplicons were diluted to 0.2 ng/µl and prepared for sequencing using the Nextera XT library preparation kit with v2 indexes (Illumina, San Diego, CA, USA). Multiplexed libraries were then sequenced either on an Illumina iSeq 100 or MiSeq producing at least 200,000 paired end reads (2x150nt) per library. For genome assembly, the sequence reads were QC trimmed using BBDuk v37.98[38] before de novo assembly with MEGAHIT v1.1.3[39] or reference-based assembly with IRMA[40]. To confirm assembly, the trimmed sequence reads were re-mapped onto the draft genome with BBMap v37.98 and visually assessed using the Geneious Prime v.2020.0.3 before the final majority consensus genome was extracted.

Phylogenetic analysis. RSV sequences generated in this study were analysed along with reference sequences sourced from NCBI GenBank and GISAID databases (see Supplementary Data 2) or from the NIAID Virus Pathogen Database and Analysis Resource (ViPR)[41] at http://www.viprbrc.org/. Specifically, all available full-length genomes and partial G gene sequences (greater than 300 nt) with collection dates were downloaded on 1st October 2021. Multiple sequence alignments were performed independently with MAFFT v.7[42] and examined using TempEst v.1.5[30] to identify and exclude excessively divergent sequences in a preliminary maximum likelihood (ML) tree generated in FastTree v.2.1[43]. G gene phylogenies were estimated using RAxML v.8[44] using the GTR-Γ nucleotide substitution model, with branch support estimated by 1000 bootstrap replicates.

Time-scaled phylogenetic trees of the full-length alignments were inferred using a Bayesian molecular clock phylogenetic analysis pipeline proposed by du Plessis et al.[45] for large datasets using BEAST (v1.10)[46]. In this pipeline, a maximum likelihood tree with branch lengths in nucleotide substitutions (genetic distances) and a time-calibrated tree estimated using TreeTime[47] were used as starting inputs for Bayesian dating. To maintain a map between branches of both trees, any clade disruptions were rejected, and only polytomy resolutions and branch durations could be optimised under a strict clock model with a log-normal prior mean rate of 6.624×10^{-4} substitutions/site/year, which we inferred using a root-to-tip regression in TempEst v.1.5[30] based on a ML tree generated by IQ-TREE v.2.0[48] with the best-fit nucleotide substitution model. The Skygrid population model with 14 grid points (13.18 years duration in this study) and a Laplace root-height prior with mean equal to the time-calibrated tree estimated by TreeTime was used with scale set to 20% of the mean. We performed two MCMC runs for 100 million steps, sampled every 10,000 steps, and discarded 10% as burn-in, ensuring ESS for each parameter is larger than 200 using Tracer v.1.7.1. We estimated the effect of excess non-synonymous/synonymous substitutions in outbreak clades on the tMRCA estimates using a protocol described previously[32], in which SLAC[49] (Single Likelihood Ancestor Counting), SNAP v.2.1.1[50] (Synonymous Non-synonymous Analysis Programme) and Contrast-FEL were used to estimate the difference between the mean ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site (d_N/d_S) of the NSW/ACT/VIC ($n = 166$) and WA ($n = 79$) datasets versus a subsampled global dataset ($n = 112$) comprising genomes from major clades.

aBSREL[28] was used to test whether a proportion of sites evolved under positive selection in any of the branches, and RELAX[29] was applied to test for strength of selection (relaxed/intensified) in a set of test branches. In this study, the clades of the subsampled global dataset were defined as background, and the NSW/ACT/VIC and WA clades were regarded as foreground separately. Furthermore, Contrast-FEL[27] was used to identify codon sites that evolved differently for the outbreak clades. For this, the subsampled global dataset, four sequences from the largest NSW/ACT/VIC 2020 clade and three sequences from the largest WA 2020 clade were used, with the ancestral branches of the outbreak clades assigned as foreground for each of the two clades respectively.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequence data generated in this study have been deposited in the NCBI GenBank database under accession codes OM857140 - OM857397, and the GISAID EpiRSV database with accession numbers EPI_ISL_1653938 to EPI_ISL_1653948, EPI_ISL_2543762 to EPI_ISL_2543853, and EPI_ISL_2839170 to EPI_ISL_2839457 (see Supplementary Table 1 and Supplementary Data 1). Reference sequences were downloaded from the NCBI GenBank and GISAID EpiRSV databases (see Supplementary Data 2).

Code availability

The code used for genome assembly can be found at https://github.com/jsede/virus_assembly/ (https://doi.org/10.5281/zenodo.6360724)[51], while code for individual analyses can be found at https://github.com/jsede/RSV_2020 (https://doi.org/10.5281/zenodo.6360678)[52].

Received: 22 July 2021; Accepted: 27 April 2022;
Published online: 24 May 2022

References

1. Shi, T. et al. Global disease burden estimates of respiratory syncytial virus-associated acute respiratory infection in older adults in 2015: A systematic review and meta-analysis. *J. Infect. Dis.* **222**, S577–S583 (2020).
2. Griffiths, C., Drews, S. J. & Marchant, D. J. Respiratory syncytial virus: Infection, detection, and new options for prevention and treatment. *Clin. Microbiol. Rev.* **30**, 277–319 (2017).
3. Fleming, D. M. et al. Modelling estimates of the burden of respiratory syncytial virus infection in adults and the elderly in the United Kingdom. *BMC Infect. Dis.* **15**, 443 (2015).
4. Lee, N. et al. High morbidity and mortality in adults hospitalized for respiratory syncytial virus infections. *Clin. Infect. Dis.* **57**, 1069–1077 (2013).
5. Li, Y. et al. Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: A systematic analysis. *Lancet Glob. Health* **7**, e1031–e1045 (2019).
6. Saravanos, G. L. et al. Respiratory syncytial virus-associated hospitalisations in Australia, 2006–2015. *Med. J. Aust.* **210**, 447–453 (2019).
7. Price, O. H., Sullivan, S. G., Sutterby, C., Druce, J. & Carville, K. S. Using routine testing data to understand circulation patterns of influenza A, respiratory syncytial virus and other respiratory viruses in Victoria, Australia. *Epidemiol. Infect.* **147**, e221 (2019).
8. Fagan, P., McLeod, C. & Baird, R. W. Seasonal variability of respiratory syncytial virus infection in the Top End of the Northern Territory (2012–2014): Seasonality of RSV in the NT. *J. Paediatr. Child Health* **53**, 43–46 (2017).
9. Baker, R. E. et al. The impact of COVID-19 nonpharmaceutical interventions on the future dynamics of endemic infections. *Proc. Natl Acad. Sci. USA* **117**, 30547–30553 (2020).
10. Olsen, S. J. et al. Decreased influenza activity during the COVID-19 pandemic-United States, Australia, Chile, and South Africa, 2020. *Am. J. Transpl.* **20**, 3681–3685 (2020).

11. Rockett, R. J. et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat. Med.* **26**, 1398–1404 (2020).
12. Yeoh, D. K. et al. The impact of COVID-19 public health measures on detections of influenza and respiratory syncytial virus in children during the 2020 Australian winter. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa1475 (2020).
13. Britton, P. N. et al. COVID-19 public health measures and respiratory syncytial virus. *Lancet Child Adolesc. Health* **4**, e42–e43 (2020).
14. Sullivan, S. G. et al. Where has all the influenza gone? The impact of COVID-19 on the circulation of influenza and other respiratory viruses, Australia, March to September 2020. *Euro Surveill.* **25**, 2001847 (2020).
15. corporateName=Commonwealth Parliament, Address=Parliament House, Canberra, ACT & contact=+. COVID-19: A chronology of state and territory government announcements (up until 30 June 2020) – Parliament of Australia. https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp2021/Chronologies/COVID-19StateTerritoryGovernmentAnnouncements (2021).
16. Tempia, S. et al. Decline of influenza and respiratory syncytial virus detection in facility-based surveillance during the COVID-19 pandemic, South Africa, January to October 2020. *Euro Surveill.* **26**, 2001600 (2021).
17. Huang, Q. S. et al. Impact of the COVID-19 nonpharmaceutical interventions on influenza and other respiratory viral infections in New Zealand. *Nat. Commun.* **12**, 1001 (2021).
18. SOUTH AFRICA WEEK. Weekly respiratory pathogens surveillance report. https://www.nicd.ac.za/wp-content/uploads/2020/12/Weekly-Respiratory-Pathogens-Report-Week-50-2020.pdf.
19. Casalegno, J.-S. et al. Characteristics of the delayed respiratory syncytial virus epidemic, 2020/2021, Rhône Loire, France. *Euro Surveill.* **26**, 2100630 (2021).
20. van Summeren, J. et al. Low levels of respiratory syncytial virus activity in Europe during the 2020/21 season: What can we expect in the coming summer and autumn/winter? *Euro Surveill.* **26**, 2100639 (2021).
21. Weinberger Opek, M. et al. Delayed respiratory syncytial virus epidemic in children after relaxation of COVID-19 physical distancing measures, Ashdod, Israel, 2021. *Euro Surveill.* **26**, 2100706 (2021).
22. National Center for Immunization and Respiratory Diseases, RSV National Trends. https://www.cdc.gov/surveillance/nrevss/rsv/natl-trend.html (2021).
23. Robertson, M. et al. The spatial-temporal dynamics of respiratory syncytial virus infections across the east-west coasts of Australia during 2016–17. *Virus Evol* **7**, veab068 (2021).
24. Di Giallonardo, F. et al. Evolution of human respiratory syncytial virus (RSV) over multiple seasons in New South Wales, Australia. *Viruses* **10**, 476 (2018).
25. Eshaghi, A. et al. Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. *PLoS One* **7**, e32807 (2012).
26. Otieno, J. R. 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.* **4**, vey027 (2018).
27. Kosakovsky Pond, S. L., Wisotsky, S. R., Escalante, A., Magalis, B. R. & Weaver, S. Contrast-FEL-A test for differences in selective pressures at individual sites among clades and sets of branches. *Mol. Biol. Evol.* **38**, 1184–1198 (2021).
28. Smith, M. D. et al. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).
29. Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L. & Scheffler, K. RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* **32**, 820–832 (2015).
30. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
31. Pybus, O. G. et al. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol. Biol. Evol.* **24**, 845–852 (2007).
32. Smith, G. J. D. et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
33. Duchêne, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. Biol. Sci.* **281**, 20140732 (2014).
34. Nair, H. et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *Lancet* **378**, 1917–1930 (2011).
35. Olsen, S. J. et al. Changes in influenza and other respiratory virus activity during the COVID-19 pandemic—United States, 2020–2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 1013–1019 (2021).
36. Mak, G. C., Wong, A. H., Ho, W. Y. Y. & Lim, W. The impact of pandemic influenza A (H1N1) 2009 on the circulation of respiratory viruses 2009–2011. *Influenza Other Respir. Viruses* **6**, e6–e10 (2012).
37. Chidlow, G. R., Harnett, G. B., Shellam, G. R. & Smith, D. W. An economical tandem multiplex real-time PCR technique for the detection of a comprehensive range of respiratory pathogens. *Viruses* **1**, 42–56 (2009).
38. Bushnell, B. BBMap. https://sourceforge.net/projects/bbmap/ (2021).
39. 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* **31**, 1674–1676 (2015).
40. Shepard, S. S. et al. Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. *BMC Genomics* **17**, 708 (2016).
41. Pickett, B. E. et al. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, D593–D598 (2012).
42. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
43. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
44. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
45. du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
46. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
47. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
48. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
49. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
50. Rodrigo, A. G. & Learn, G. H. *Computational and Evolutionary Analysis of HIV Molecular Sequences* (Springer, 2014).
51. laulambr & jsede. *jsede/virus_assembly: v.1.0.0.* (Zenodo). https://doi.org/10.5281/ZENODO.6360724 (2022).
52. jsede. *jsede/RSV_2020: v.0.1.0-inital.* (Zenodo). https://doi.org/10.5281/ZENODO.6360678 (2022).

Acknowledgements

The WHO Collaborating Centre for Reference and Research on Influenza is supported by the Australian Government Department of Health. Funding was provided through the ICPMR Private Practice Trust fund to J.S.E., and J.K., the National Health and Medical Research Council Centre of Research Excellence in Emerging Infectious Diseases (#1102962) to J.S.E., E.C.H., and D.S.W. and the Sydney Institute for Infectious Diseases at the University of Sydney to J.S.E. We would finally like to thank all the authors who have kindly shared genome data on GISAID to the EpiRSV database (Supplementary Data 3).

Author contributions

This study was designed by J.S.E., C.S., S.S., V.D., D.W.S., J.K., and I.G.B. Epidemiological data and sample provision were performed by S.S., A.L., C.C.B., P.N.B., N.C., D.E.D, K.M.E., D.F., K.K., D.S., D.W.S., and J.K. Genome sequencing was performed by J.S.E., C.S., Y.M.D., A.M., E.C., X.D., B.A.H., C.M.S., and R.L.T. Phylodynamic analysis was performed by J.S.E., C.S., R.X., X.D., E.C.H., and V.D. Data visualisation was performed by J.S.E., C.S., R.X., S.S., and V.D. Project supervision was performed by E.C.H., V.D., D.W.S., J.K., and I.G.B. The original draft of this manuscript was prepared by J.S.E., C.S., R.X., S.S., V.D., D.W.S., J.K., and I.G.B. and was reviewed and edited by all remaining authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30485-3.

Correspondence and requests for materials should be addressed to Vijaykrishna Dhanasekaran, David W. Smith, Jen Kok or Ian G. Barr.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

[1]Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia. [2]Sydney Institute for Infectious Diseases, Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia. [3]PathWest Laboratory Medicine WA, Department of Microbiology, Nedlands, WA 6009, Australia. [4]School of Biomedical Sciences, The University of Western Australia, Crawley, WA 6009, Australia. [5]School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [6]HKU-Pasteur Research Pole, School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [7]WHO Collaborating Centre for Reference and Research on Influenza, Royal Melbourne Hospital, at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia. [8]Department of Microbiology and Immunology, University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia. [9]Department of Infectious Diseases, University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia. [10]School of Medicine, The University of Western Australia, Crawley, WA 6009, Australia. [11]Department of Infectious Diseases, Perth Children's Hospital, Nedlands, WA 6009, Australia. [12]Wesfarmers Centre of Vaccines and Infectious Diseases, Telethon Kids Institute, Nedlands, WA 6009, Australia. [13]Departments of Infectious Diseases and Microbiology, The Children's Hospital at Westmead, Westmead, NSW 2145, Australia. [14]Murdoch Children's Research Institute, Melbourne, VIC 3000, Australia. [15]Department of Paediatrics, University of Melbourne & Royal Children's Hospital, Melbourne, VIC 3000, Australia. [16]Immunisation Service, Royal Children's Hospital, Melbourne, VIC 3000, Australia. [17]NSW Health Pathology - Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia. [18]Departments of Clinical Microbiology and Infectious Diseases, Canberra Hospital, Garran, ACT 2605, Australia. [20]These authors contributed equally: John-Sebastian Eden, Chisha Sikazwe, Ruopeng Xie. [21]These authors jointly supervised this work: Vijaykrishna Dhanasekaran, David W. Smith, Jen Kok, Ian G. Barr. *A list of authors and their affiliations appears at the end of the paper. ✉email: veej@hku.hk; David.Smith@health.wa.gov.au; jen.kok@health.nsw.gov.au; ian.barr@influenzacentre.org

the Australian RSV study group

Annette Alafaci[14], Ian Carter[17], Andrew Daley[15], Michelle Francis[19], Alison Kesson[2,13], Hannah Moore[12], Christine Ngo[17] & Tyna Tran[17]

[19]Microbiology Laboratory, Monash Pathology, Monash Health, Melbourne, VIC, Australia.

Appendix A2: The Spatial-Temporal Dynamics of Respiratory Syncytial Virus Infections Across the East-West Coasts of Australia During 2016-17

VIRUS EVOLUTION

Virus Evolution, 2021, 7(2), 1–8
DOI: https://doi.org/10.1093/ve/veab068
Advance access publication date: 22 July 2021
Research Article

The spatial-temporal dynamics of respiratory syncytial virus infections across the east–west coasts of Australia during 2016–17

Mark Robertson,[1,†] John-Sebastian Eden,[2,3,†,‡,*] Avram Levy,[4,5] Ian Carter,[1] Rachel L. Tulloch,[2,3] Elena J. Cutmore,[2,3] Bethany A. Horsburgh,[2,3] Chisha T. Sikazwe,[4,5] Dominic E. Dwyer,[1,3] David W. Smith,[4,6] and Jen Kok[1,*]

[1]NSW Health Pathology-Institute for Clinical Pathology and Medical Research, NSW Health Pathology, Redbank Road, Westmead Hospital, Westmead, NSW 2145, Australia, [2]Centre for Virus Research, Westmead Institute for Medical Research, 176 Hawkesbury Road, Westmead, NSW 2145, Australia, [3]Marie Bashir Institute for Infectious Diseases and Biosecurity, Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia, [4]Department of Microbiology, PathWest Laboratory Medicine WA, Hospital Ave, Nedlands, WA 6009, Australia, [5]School of Biological Sciences, The University of Western Australia, Crawley, WA 6009, Australia and [6]School of Medicine, The University of Western Australia, Crawley, WA 6009, Australia
[†]https://doi.org/0000-0003-1374-3551
[‡]These authors contributed equally to this work.
*Corresponding authors: E-mail: js.eden@sydney.edu.au; jen.kok@health.nsw.gov.au

Abstract

Respiratory syncytial virus (RSV) is an important human respiratory pathogen. In temperate regions, a distinct seasonality is observed, where peaks of infections typically occur in early winter, often preceding the annual influenza season. Infections are associated with high rates of morbidity and mortality and in some populations exceed that of influenza. Two subtypes, RSV-A and RSV-B, have been described, and molecular epidemiological studies have shown that both viruses mostly co-circulate. This trend also appears to be the case for Australia; however, previous genomic studies have been limited to cases from one Eastern state—New South Wales. As such, the broader spatial patterns and viral traffic networks across the continent are not known. Here, we conducted a whole-genome study of RSV comparing strains across eastern and Western Australia during the period January 2016 to June 2017. In total, 96 new RSV genomes were sequenced, compiled with previously generated data, and examined using a phylodynamic approach. This analysis revealed that both RSV-A and RSV-B strains were circulating, and each subtype was dominated by a single genotype, RSV-A ON1-like and RSV-B BA10-like viruses. Some geographical clustering was evident in strains from both states with multiple distinct sub-lineages observed and relatively low mixing across jurisdictions, suggesting that endemic transmission was likely seeded from imported, unsampled locations. Overall, the RSV phylogenies reflected a complex pattern of interactions across multiple epidemiological scales from fluid virus traffic across global and regional networks to fine-scale local transmission events.

Key words: respiratory syncytial virus; molecular epidemiology; Australia; whole-genome sequencing; phylogenetics

1. Introduction

Respiratory syncytial virus (RSV) is a major cause of acute respiratory tract infections in patients of all ages, producing significant morbidity and mortality (Shi et al., 2017). The greatest burden of disease is in children under 1 year old, where it is the most common cause of acute respiratory tract infection and in this age group is second only to malaria as a cause of death globally (Hall et al., 2009; Griffiths, Drews, and Marchant 2017). Importantly, RSV infection in young children may also lead to long-term sequelae such as asthma, chronic bronchitis, and obstructive pulmonary disease (Beigelman and Bacharier 2016; Griffiths, Drews, and Marchant 2017). Other populations particularly impacted by RSV include those over 65 years and immunosuppressed patients, such as solid organ and bone marrow transplant recipients (Beigelman and Bacharier 2016; Griffiths, Drews, and Marchant 2017).

Despite being discovered in the 1950s, the burden of RSV disease has only recently been appreciated, which is in part due to more reliable methods of diagnosis and detection. The laboratory diagnosis of RSV was initially reliant on viral isolation and the visualisation of characteristic syncytial cytopathic effects, for which its name is derived (Henrickson and Hall 2007). These techniques were slow and required technical expertise. The most commonly used modalities now include rapid antigen (such as lateral flow immunochromatography and fluorescent immunoassays) and nucleic acid amplification tests (NATs). Molecular assays, including both commercial and in-house NATs, offer increased sensitivity and specificity, as well as the ability to be multiplexed to detect other respiratory pathogens such as influenza, parainfluenza, rhinovirus, and human metapneumovirus (Mahony et al., 2007). Rapid test assays can also offer the additional benefit of early diagnosis, which allows for appropriate infection control

interventions, rationalisation of unnecessary antibiotic therapy, and shorter hospitalisation periods (Henrickson and Hall 2007). More generally, improved diagnostics and reporting have begun to shed light on the true incidence, seasonal patterns, and peaks of activity of RSV. In temperate regions such as the southern major metropolitan areas of Australia, the seasonal peak typically occurs in the late autumn to early winter period (late-March to mid-August) in the months leading into the influenza season (Henrickson and Hall 2007; Di Giallonardo et al., 2018; Yeoh et al., 2020).

RSV can be divided into two antigenically and genetically distinct subtypes, RSV-A and RSV-B. These may be further divided into genetic groups based on the viral glycoprotein (G gene), termed genotypes, with at least 11 and 23 for RSV-A and RSV-B, respectively. However, recent work has proposed shifting to a genotype classification based on whole-genome sequencing (WGS) in order to increase phylogenetic resolution (Ramaekers et al., 2020). Serological studies have shown that the majority of people are infected by age 2, and whilst primary infections are typically more severe, they are not protective against repeated infection (Griffiths, Drews, and Marchant 2017). The spectrum of disease severity associated with RSV infection remains an important yet controversial topic. Studies have made associations between increased severity and particular subtypes/genotypes (Vandini, Biagi, and Lanari 2017); however, these are complicated by both host (Tal et al., 2004) and viral factors (DeVincenzo, El Saleeby, and Bush 2005), as well as their interactions. Molecular epidemiological studies have shown that both RSV-A and RSV-B co-circulate during a season and often at similar levels (Otieno et al., 2017; 2018; Pangesti et al., 2018; Park et al., 2017; Cattoir et al., 2019). Furthermore, for each of these subtypes, a single genotype will tend to predominate, such as with the recent RSV-A ON1-like and RSV-B BA10-like viruses (Eshaghi et al., 2012; Pretorius, Van Niekerk, and Tempia 2013; Di Giallonardo et al., 2018). Our understanding of the basic molecular epidemiology of RSV has been strengthened by WGS, which, similar to other pathogens, is becoming increasingly common in its application in infectious disease surveillance (Dapat et al., 2010; Agoti et al., 2014; Di Giallonardo et al., 2018). The added resolution from WGS has been particularly useful for elucidating transmission networks at both local (Agoti et al., 2015a) and epidemiological scales (Di Giallonardo et al., 2018), as well as for identifying and classifying the introduction and spread of new genotypes (Agoti et al., 2014; Ramaekers et al., 2020).

We recently performed the first genome-scale study of RSV molecular epidemiology in Australia (Di Giallonardo et al., 2018), demonstrating a wide diversity of co-circulating RSV lineages, with limited evidence of strong age and geographical clustering. However, this study was limited to cases obtained from eastern Australia in New South Wales (NSW) through the Western Sydney Local Health District. Here, we expand on these initial investigations to perform a transcontinental study of RSV genomic epidemiology in Australia. We compare RSV strains obtained from Western Australia (WA) to those from NSW on the east coast over an equivalent time period to examine the phylogenetic distribution of strains and to describe viral traffic between these regions.

2. Materials and methods

2.1 Sample collection and processing

WGS was conducted on residual RSV-positive specimens collected and tested through routine testing at diagnostic laboratories in two major Australian diagnostic laboratories. Specimens were de-identified with basic demographic information collected, including age, sex, and postcode, as per protocols approved by local ethics and governance committees (LNR/17/WMEAD/128). One hundred and three RSV-A- and RSV-B-positive cases from WA with a reverse transcription polymerase chain reaction (RT-PCR) cycle threshold <30 in the original screen were collected between January 2016 and May 2017. Sample aliquots (typically, nasopharyngeal swabs in viral transport medium) were transferred to the Institute of Clinical Pathology and Medical Research (ICPMR), NSW, and subsequently extracted using the Qiagen EZ1 Advance XL extractor with the EZ1 Virus Mini Kit v2.0 (Qiagen, Germany). For the eastern Australian RSV cases, 45 previously untyped RSV-positive viral extracts from cases collected during early 2017 at the ICPMR were obtained from storage at –80°C archive. Additional cases from 2016 in NSW were sequenced in a previous study (Di Giallonardo et al., 2018).

2.2 Whole-genome sequencing

A previously published approach was used to amplify both RSV-A and RSV-B genomes (Di Giallonardo et al., 2018). In short, RT-PCR was used to amplify four overlapping amplicons (each ~4 kb) that together span the RSV genome. The size and yield of each RT-PCR was determined by agarose gel electrophoresis, and the four targets pooled equally. The pooled RSV amplicons were then purified with Agencourt AMpure XP beads (Beckman Coulter, USA) and quantified using the Quant-iT PicoGreen dsDNA Assay (Invitrogen, USA). The purified DNA was then diluted to 0.25 ng/μl and prepared for sequencing with the Nextera XT DNA library prep kit (Illumina, USA). Libraries were sequenced on an Illumina MiSeq using a 300 cycle v2 kit (150-nt paired-end reads). Raw paired sequence reads were trimmed using Trim Galore (https://github.com/FelixKrueger/TrimGalore; last accessed 11 April 2021) and then *de novo* assembled using Trinity (Grabherr et al., 2011). RSV contigs were identified by a local Blastn (Altschul et al., 1990) using a database of RSV reference genomes from NCBI RefSeq. The trimmed reads were remapped to draft genome contigs using BBMap (https://sourceforge.net/projects/bbmap/; last accessed 11 April 2021) to check the assembly, and the final majority consensus was extracted for each sample.

2.3 Phylodynamic analysis

The aim of this study was to compare the spatial and temporal dynamics of RSV infection in Australia. To do this, we used a phylogenetic approach to compare the distribution of strains circulating in two major, geographically distinct regions—NSW and WA—representing eastern Australia and WA, respectively, during the period from January 2016 to June 2017. All WA data, as well as NSW data from the year 2017, were generated within this study. To ensure even sampling across sites, these data were combined with NSW data from 2016 that was obtained from a previous study (Di Giallonardo et al., 2018). To provide additional context, RSV genome data was also sourced from NCBI GenBank where location (country) and collection date (year) was known. The Australian and global RSV genomes were first aligned with MAFFT (Katoh and Standley 2013), using the FFT-NS-i algorithm followed by manual inspection of gapped regions particularly in the G gene and non-coding regions. The alignments were then trimmed to include only the coding regions and screened for potential recombinants using RDP4 (Martin et al., 2015) with default parameters that were then removed before analysis. The RSV-A alignment included 1,190 sequences with a length of 15,747 nt, whilst the RSV-B alignment included 1,121 sequences with a length of 15,646 nt. To increase sampling resolution, we also generated comparable data sets

Table 1. Demographic details of NSW and WA RSV cohorts.

	NSW	WA		NSW	WA
Age (years)	Number of samples (%)		Sex	Number of samples (%)	
≤5	65 (51.6)	64 (62.1)	Male	66 (52.4)	48 (46.6)

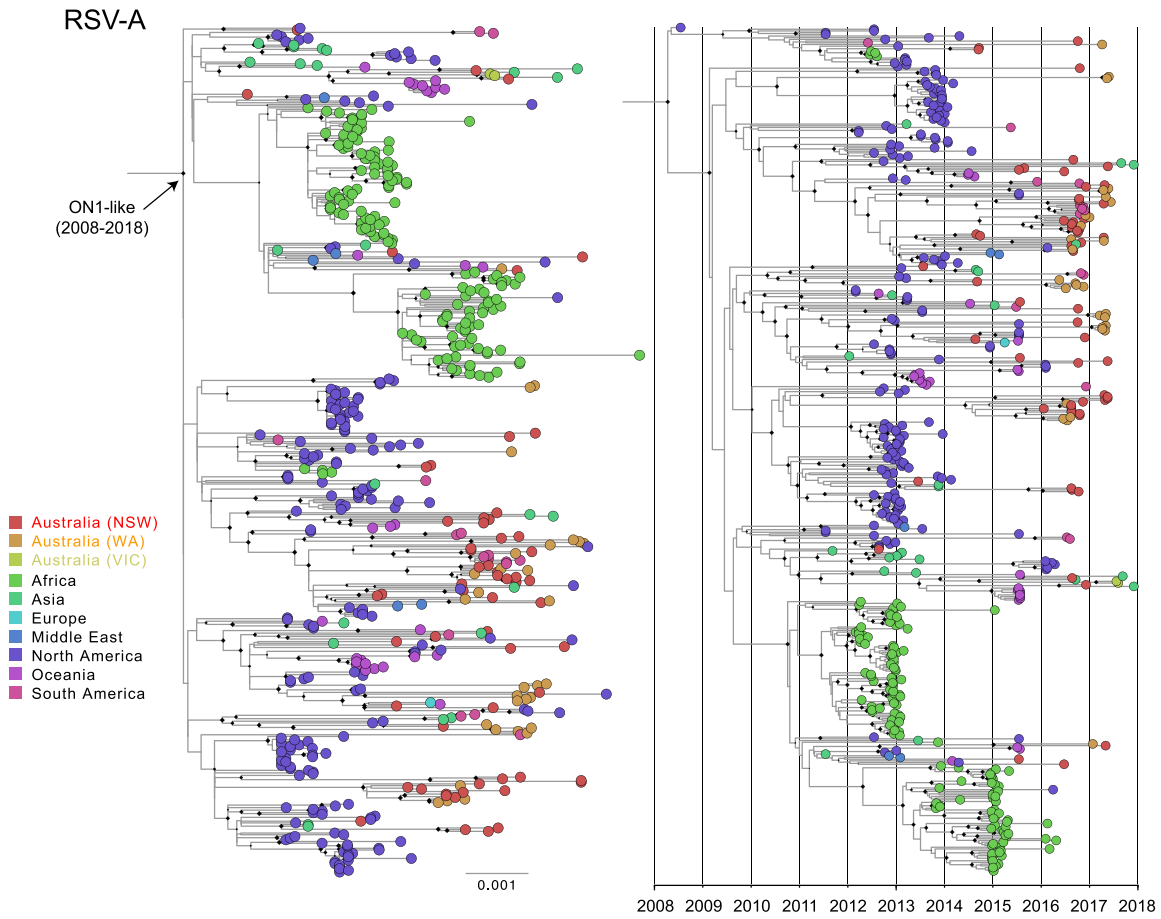


Figure 2. Phylogenetic analysis of RSV-A strains circulating in WA and NSW between January 2016 and June 2017. The global phylogenies were first estimated using alignments of complete genome sequences of RSV-A ($n = 1{,}190$) strains circulating since the late 1970s; however for clarity, only the recently circulating RSV-A ON1 lineage has been shown ($n = 499$), which is defined by the specific branch marked with an arrow. The panel on the left shows the tree estimated using a ML approach, whilst the tree on the right shows the same phylogeny estimated using the time-scaled Bayesian approach. The taxa, shown as small circles, have been coloured by sampling location as per the key provided. Diamonds at nodal positions indicate branching support with bootstrap replicate values >70 per cent and posterior probability values >0.9 for the ML and time-scaled Bayesian trees, respectively. The left-sided scale bar represents the number of substitutions per site, whilst the right is scaled to time (years).

$n = 18/103$ from patients ≥65 years), reflecting the distribution of infection at both extremes of age. The distribution between sexes was approximately equal for both study sites (Table 1).

Virus genomes were successfully extracted and sequenced from 63 of the 103 respiratory samples collected in WA during 2016 and 2017, and 32 of the 45 NSW cases collected in 2017 (total new genomes $n = 96$, where one case from NSW was an RSV-A/B mixed infection). These data were combined with existing genomes from NSW during 2016 ($n = 93$) generated previously (Di Giallonardo et al., 2018), bringing the total number of Australian genomes for analysis across both states in 2016 and 2017 in the current study to 189. The breakdown between sampling location and RSV subtypes determined by WGS shows that both RSV-A and RSV-B subtypes were present in both NSW and WA populations (Fig. 1). In NSW, there was a higher proportion of RSV-B strains (63 per cent, $n = 80/126$), which was consistent across the entire study period. Furthermore, in NSW, the peak in RSV activity typically occurs in May–June each year (NSW Health influenza surveillance data); however, most of our genome sequences in 2016 were from isolates in August–October that year, that is, they were from the later part of the RSV season. In WA, there was even representation of RSV-A and RSV-B in the WGS data across 2016, with an increase in the relative proportion of RSV-A during 2017. Whilst our sampling for WA was intentionally even with regard to time, subtype, age, and setting, the predominance of RSV-A during 2017 was observed by the initial diagnostic testing where the assay used resolves RSV subtypes. Specifically, with this lab testing data, a transition from RSV-B to RSV-A was observed between the 2016 and 2017 seasons in WA (data not shown). These results are consistent with other molecular epidemiological studies globally that show both RSV-A and RSV-B subtypes co-circulate with shifting predominance across seasons (Di Giallonardo et al., 2018; Luo et al., 2020; Razanajatovo Rahombanjanahary et al., 2020; Yun, Choi, and Lee 2020).

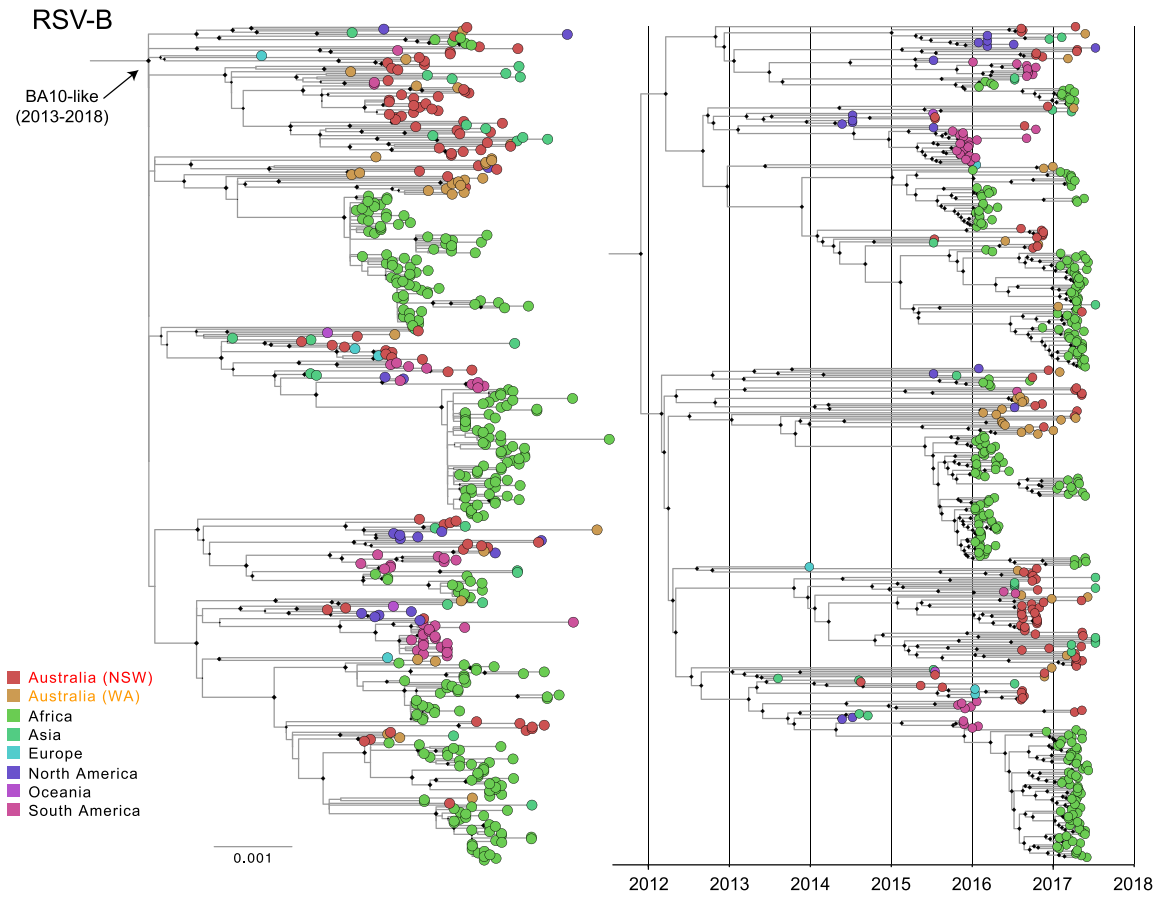


Figure 3. Phylogenetic analysis of RSV-B circulating in WA and NSW between January 2016 and June 2017. The global phylogenies were first estimated using alignments of complete genome sequences of RSV-B ($n = 1,121$) strains circulating since the 1960s; however, for clarity, only the recently circulating RSV-B BA10 lineage has been shown ($n = 470$), which is defined by the specific branch marked with an arrow. The panel on the left shows the tree estimated using a ML approach, whilst the tree on the right shows the same phylogeny estimated using the time-scaled Bayesian approach. The taxa, shown as small circles, have been coloured by sampling location as per the key provided. Diamonds at nodal positions indicate branching support with bootstrap replicate values >70 per cent and posterior probability values >0.9 for the ML and time-scaled Bayesian trees, respectively. The left-sided scale bar represents the number of substitutions per site, whilst the right is scaled to time (years).

3.2 Phylodynamics of RSV infections in NSW and WA in a global context

The WGS data from NSW and WA were aligned and compared to global genome references to provide genetic context for local strains. Phylogenetic analysis using a ML approach was then employed to examine the diversity of RSV-A and RSV-B strains (Figs 2 and 3). This analysis revealed that for each subtype, regardless of location, a single genotype was predominant. For RSV-A, recent viruses from both NSW and WA were derived from the ON1 lineage (Eshaghi et al., 2012), which has been the predominant RSV-A genotype globally since it first emerged in 2011 (Eshaghi et al., 2012; Agoti et al., 2014; Tabatabai et al., 2014; Di Giallonardo et al., 2018) (Fig. 2). Similarly, the RSV-B phylogeny showed that circulating viruses were mostly of the BA10 lineage (Fig. 3). From the limited available sequence data on GenBank, it appears that the BA10-like viruses are circulating globally; however, there are few published molecular epidemiology studies to support the suggestion they are predominant. It is clear, however, that in Australia, in both NSW and WA that this has been the major RSV-B genotype since at least 2014 (Di Giallonardo et al., 2018).

Across both RSV-A and RSV-B phylogenies, there does not appear to be strict spatial clustering when comparing viruses from NSW and WA (Figs 2 and 3), that is, the viruses from WA and NSW do not form monophyletic groups. Such monophyletic clustering would not be expected based on what is known for other respiratory pathogens—such as influenza—which are characterised by high levels of gene flow and viral traffic at global scales (Rambaut et al., 2008; Vijaykrishna et al., 2015). As such, we observed multiple co-circulating sub-lineages distributed across the entire diversity of RSV-A ON1-like and RSV-B BA10-like viruses (Figs 2 and 3). However, within these sub-lineages, the viruses often formed clusters based on sampling location; therefore, local spatial structure was apparent in our analysis. To explore this formally, we conducted a clustering analysis with BaTS and a posterior set of trees estimated using a time-scaled phylogenetic analysis in BEAST. We limited the analysis to the RSV-A ON1-like

Table 2. Phylogeny-trait association test for RSV in Australia and other regions globally.

	RSV-A		RSV-B	
Location	Seqs (*n*)	*P*-value	Seqs (*n*)	*P*-value
Overall clustering[a]	499	<0.001	470	<0.001
Asia	24	0.002	26	0.006
South America	16	0.001	42	0.001
North America	185	0.001	17	0.001
Middle East	5	0.008	0	~
Oceania[b]	28	0.001	2	1.000
Africa	147	0.001	263	0.001
Australia (NSW)	60	0.006	84	0.001
Australia (VIC)	2	0.002	0	~
Australia (WA)	31	0.001	32	0.001
Europe	1	1.000	4	1.000

[a]Determined by AI and PS.
[b]New Zealand only.

and RSV-B BA10-like viruses, and the maximum clade credibility tree for each RSV subtype was found to be congruent with the ML trees (Figs 2 and 3). Using the posterior set of trees, we then measured the degree of clustering based on sampling location (Table 2).

When sequences were grouped by broad geographical regions, specifically, continents and the two Australian states NSW and WA (and VIC for RSV-A data), phylogeny-trait association tests indicated a significant pattern of overall geographical clustering (*P*-values <0.001) as measured by AI and PS (Table 2). For individual locations, most were found to cluster with significant scores. This includes Asia, the Americas, the Middle East, Africa, Oceania, NSW, WA, and VIC for RSV-A viruses (all *P*-values <0.008) and then similarly all for RSV-B where the sampling was sufficient (n >4 sequences). This analysis demonstrates that, at larger epidemiological scales, viral lineages are often imported and established locally. We note, however, that this analysis is biased for regions that have been highly sampled, as highlighted by the American and Kenyan viruses, which are mostly from one specific city and/or hospital (Supplementary Figs S1 and S2) (Otieno et al., 2018). Our sampling of NSW and WA strains is similarly constrained, except that our sampling sites are both major diagnostics hubs that perform testing across wide geographic regions that are representative of large and diverse populations (Di Giallonardo et al., 2018), and in our comparison, the period of sampling was equivalent. Despite this and any apparent geographical clustering, the phylogenies both reflect a lack of data from many jurisdictions globally, and our insights into global and local viral traffic will remain hampered until these genome sequences become available. To address this, we also performed ML phylogenetic analysis on a more well-sampled data set derived from sequences of the G gene region only (Supplementary Figs S3–S5). Whilst this analysis improved the overall location sampling, particularly for European and Asian sequences, there was less phylogenetic resolution; therefore, a trade-off exists between using the less sampled, higher-resolution whole-genome data versus the more well-sampled partial genome (G gene) sequences available on NCBI GenBank (Supplementary Fig. S3). Regardless, as shown below, these data can shed light on the possible sources and global transmission networks for RSV.

3.3 Lineages and local clusters

A detailed examination of individual lineages (Supplementary Figs S1 and S2) confirmed that most only contained viruses from either state. There were some instances where individual viruses from NSW were nested within diversity solely comprised of WA strains, and vice versa, and which may represent viral migration events between the jurisdictions. The overall pattern suggests low levels of mixing between NSW and WA, perhaps due to the large geographic area of each state and the distance that separates them. Moreover, we had suggested previously that some evidence of viral persistence across seasons was present in NSW (Di Giallonardo et al., 2018). Here, we found no direct link between the 2016 and 2017 seasons in both jurisdictions and that viral diversity is likely maintained and seeded from imported yet unsampled locations. An examination of the better-sampled G gene phylogenies (Supplementary Figs S4 and S5) did not clarify potential additional sources, except for sporadic examples where NSW or WA were clustered genetically and temporally with Asian or European strains. Importantly, we also found examples where two Australian sequences clustered in the G gene phylogeny but were then sufficiently different at a genome scale to not be classified as clusters such as seen with the NSW and WA strains WM1079A/14 August 2016 (MH760625) and PW3375488K/9 August 2016 (MW160759) (Supplementary Figs S1 and S4). This is consistent with other studies that have shown that local persistence makes a minimal contribution to the diversity of RSV strains in any given area (Agoti et al., 2015b; Zou et al., 2016). Rather, viral importation and global mixing of strains seem to be the major driver of RSV diversity and their sources. Despite the data suggesting limited east–west viral traffic across the Australian continent, the Sydney–Perth air route remains one of the busiest in the country, nor can we discount other regions within Australia such as the tropical North, where persistence of viral lineages may occur as important sources of viral diversity. Indeed, from our understanding of the influenza virus in Australia, the synchronised dissemination of viral strains across the country is driven by both multiple introductions from the global population and strong domestic connectivity (Geoghegan et al., 2018). Increasing the breadth of WGS data for RSV cases nationally would assist in confirming what domestic viral traffic networks exist, particularly between tropical and temperate regions, as well as between major population centres on the east coast such as Sydney and Melbourne.

Next, we examined genetic clusters defined by high genetic identity (<10-bp difference across the genome) and similar sampling periods (collection dates within 2 weeks) to consider what features such as patient age and sampling location define them. For RSV-A, we identified 8 clusters (Supplementary Fig. S1 and Table S1), and for RSV-B, we identified 10 clusters (Supplementary Fig. S2 and Table S2). In this study, 75 per cent ($n = 6/8$) and 80 per cent ($n = 8/10$) of RSV-A and RSV-B clusters, respectively, could be linked to common localities and/or institutions, including individual hospital wards or emergency departments. Whilst two of the RSV-B clusters were found to be repeat samples of the same patient (Supplementary Table S2), overwhelmingly, these clusters of genetically and temporal related sequences most likely represent fine-scale transmission events.

4. Conclusions

In summary, our analysis has identified a number of features of RSV epidemiology and patterns of spread, including (1) the co-circulation of both major subtypes—RSV-A and RSV-B, (2) a single genotype for each subtype predominates each season, (3) multiple distinct sub-lineages of each genotype will co-circulate and which are associated with regional and local clustering/outbreaks,

(4) little viral mixing across the east–west coasts of Australia despite apparent overall geographic clustering, (5) that genetically and temporal-related sequences most likely represent fine-scale transmission events such as institutional outbreaks, and (6) that whole-genome sequence is required and encouraged over partial G gene sequencing for elucidating clusters and transmission pathways. Taken together, this presents a complex phylodynamic pattern with globally circulating diversity with viral mixing across different regions, yet, finer-scale patterns revealing multiple endemic sub-lineages and clusters consistent with local transmission events and outbreaks. This highlights the connections of genomic data across multiple epidemiological scales and further strengthens the need for a much greater sampling of RSV, not just here in Australia, but globally.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We thank all the members of the Virology and Microbiology teams at NSW Health Pathology—ICPMR in Westmead, NSW, and the PathWest QE2 Medical Centre laboratories in Perth, WA, for all their contributions towards processing the diagnostics specimens used in this study.

Funding

Funding was provided through the Institute for Clinical Pathology and Medical Research (ICPMR) Private Practice Trust fund, the National Health and Medical Research Council Centre of Research Excellence in Emerging Infectious Diseases (1102962) and the Marie Bashir Institute for Infectious Diseases and Biosecurity at the University of Sydney.

Conflict of interest: None declared.

References

Agoti, C. N. et al. (2014) 'Rapid Spread and Diversification of Respiratory Syncytial Virus Genotype ON1, Kenya', *Emerging Infectious Diseases*, 20: 950–9.

——— et al. (2015a) 'Local Evolutionary Patterns of Human Respiratory Syncytial Virus Derived from Whole-Genome Sequencing', *Journal of Virology*, 89: 3444–54.

——— et al. (2015b) 'Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence', *Journal of Virology*, 89: 11630–42.

Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215: 403–10.

Beigelman, A., and Bacharier, L. B. (2016) 'Early-Life Respiratory Infections and Asthma Development: Role in Disease Pathogenesis and Potential Targets for Disease Prevention', *Current Opinion in Allergy and Clinical Immunology*, 16: 172–8.

Cattoir, L. et al. (2019) 'Epidemiology of RSV and hMPV in Belgium: A 10-Year Follow-up', *Acta Clinica Belgica*, 74: 229–35.

Dapat, I. C. et al. (2010) 'New Genotypes within Respiratory Syncytial Virus Group B Genotype BA in Niigata, Japan', *Journal of Clinical Microbiology*, 48: 3423–7.

Darriba, D. et al. (2012) 'jModelTest 2: More Models, New Heuristics and Parallel Computing', *Nature Methods*, 9: 772.

DeVincenzo, J. P., El Saleeby, C. M., and Bush, A. J. (2005) 'Respiratory Syncytial Virus Load Predicts Disease Severity in Previously Healthy Infants', *The Journal of Infectious Diseases*, 191: 1861–8.

Di Giallonardo, F. et al. (2018) 'Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia', *Viruses*, 10: 476.

Drummond, A. J. et al. (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29: 1969–73.

Eshaghi, A. et al. (2012) 'Genetic Variability of Human Respiratory Syncytial Virus A Strains Circulating in Ontario: A Novel Genotype with A 72 Nucleotide G Gene Duplication', *PLoS One*, 7: e32807.

Geoghegan, J. L. et al. (2018) 'Continental Synchronicity of Human Influenza Virus Epidemics despite Climactic Variation', *PLoS Pathogens*, 14: e1006780.

Grabherr, M. G. et al. (2011) 'Full-length Transcriptome Assembly from RNA-Seq Data without a Reference Genome', *Nature Biotechnology*, 29: 644–52.

Griffiths, C., Drews, S. J., and Marchant, D. J. (2017) 'Respiratory Syncytial Virus: Infection, Detection, and New Options for Prevention and Treatment', *Clinical Microbiology Reviews*, 30: 277–319.

Hall, C. B. et al. (2009) 'The Burden of Respiratory Syncytial Virus Infection in Young Children', *The New England Journal of Medicine*, 360: 588–98.

Henrickson, K. J., and Hall, C. B. (2007) 'Diagnostic Assays for Respiratory Syncytial Virus Disease', *The Pediatric Infectious Disease Journal*, 26: S36–40.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Luo, H.-J. et al. (2020) 'Epidemiological Characteristics and Phylogenic Analysis of Human Respiratory Syncytial Virus in Patients with Respiratory Infections during 2011–2016 in Southern China', *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases*, 90: 5–17.

Mahony, J. et al. (2007) 'Development of a Respiratory Virus Panel Test for Detection of Twenty Human Respiratory Viruses by Use of Multiplex PCR and a Fluid Microbead-based Assay', *Journal of Clinical Microbiology*, 45: 2965–70.

Martin, D. P. et al. (2015) 'RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes', *Virus Evolution*, 1: vev003.

Otieno, J. R. et al. (2017) 'Spread and Evolution of Respiratory Syncytial Virus A Genotype ON1, Coastal Kenya, 2010–2015', *Emerging Infectious Diseases*, 23: 264–71.

——— et al. (2018) 'Whole Genome Analysis of Local Kenyan and Global Sequences Unravels the Epidemiological and Molecular Evolutionary Dynamics of RSV Genotype ON1 Strains', *Virus Evolution*, 4: vey027.

Pangesti, K. N. A. et al. (2018) 'Molecular Epidemiology of Respiratory Syncytial Virus', *Reviews in Medical Virology*, 28: e1968.

Park, E. et al. (2017) 'Molecular and Clinical Characterization of Human Respiratory Syncytial Virus in South Korea between 2009 and 2014', *Epidemiology and Infection*, 145: 3226–42.

Parker, J., Rambaut, A., and Pybus, O. G. (2008) 'Correlating Viral Phenotypes with Phylogeny: Accounting for Phylogenetic Uncertainty', *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 8: 239–46.

Pretorius, M. A., van Niekerk, S., and Tempia, S. (2013), *Replacement and Positive Evolution of Subtype A and B Respiratory Syncytial Virus G-protein Genotypes from 1997–2012 in South Africa. The Journal of Infectious Diseases.* <https://academic.oup.com/jid/article-abstract/208/suppl_3/S227/789685> Accessed 21 Jul 2021.

Ramaekers, K. et al. (2020) 'Towards a Unified Classification for Human Respiratory Syncytial Virus Genotypes', *Virus Evolution*, 6: veaa052.

Rambaut, A. et al. (2008) 'The Genomic and Epidemiological Dynamics of Human Influenza A Virus', *Nature*, 453: 615–9.

Razanajatovo Rahombanjanahary, N. H. et al. (2020) 'Genetic Diversity and Molecular Epidemiology of Respiratory Syncytial Virus Circulated in Antananarivo, Madagascar, from 2011 to 2017: Predominance of ON1 and BA9 Genotypes', *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology*, 129: 104506.

Shi, T. et al. (2017) 'Global, Regional, and National Disease Burden Estimates of Acute Lower Respiratory Infections Due to Respiratory Syncytial Virus in Young Children in 2015: A Systematic Review and Modelling Study', *The Lancet*, 390: 946–58.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Tabatabai, J. et al. (2014) 'Novel Respiratory Syncytial Virus (RSV) Genotype ON1 Predominates in Germany during Winter Season 2012–13', *PLoS One*, 9: e109191.

Tal, G. et al. (2004) 'Association between Common Toll-like Receptor 4 Mutations and Severe Respiratory Syncytial Virus Disease', *The Journal of Infectious Diseases*, 189: 2057–63.

Vandini, S., Biagi, C., and Lanari, M. (2017) 'Respiratory Syncytial Virus: The Influence of Serotype and Genotype Variability on Clinical Course of Infection', *International Journal of Molecular Sciences*, 18: 1717.

Vijaykrishna, D. et al. (2015) 'The Contrasting Phylodynamics of Human Influenza B Viruses', *eLife*, 4: e05055.

Yeoh, D. K. et al. (2020) 'The Impact of COVID-19 Public Health Measures on Detections of Influenza and Respiratory Syncytial Virus in Children during the 2020 Australian Winter', *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 72: 2199–202.

Yun, K. W., Choi, E. H., and Lee, H. J. (2020) 'Molecular Epidemiology of Respiratory Syncytial Virus for 28 Consecutive Seasons (1990–2018) and Genetic Variability of the Duplication Region in the G Gene of Genotypes ON1 and BA in South Korea', *Archives of Virology*, 165: 1069–77.

Zou, L. et al. (2016) 'Evolution and Transmission of Respiratory Syncytial Group A (RSV-A) Viruses in Guangdong, China 2008–2015', *Frontiers in Microbiology*, 7: 1263.

Appendix A3: Meta-Analysis of Genetic Parameters of Production Traits in Cultured Shrimp Species

Received: 10 September 2019 | Revised: 2 June 2020 | Accepted: 14 July 2020

DOI: 10.1111/faf.12495

ORIGINAL ARTICLE

FISH and FISHERIES WILEY

Meta-analysis of genetic parameters of production traits in cultured shrimp species

Md. Mehedi Hasan[1,2] | Rachel L. Tulloch[1,2] | Peter C. Thomson[2,3] | Herman W. Raadsma[1,2] | Mehar S. Khatkar[1,2]

[1]Faculty of Science, Sydney School of Veterinary Science, The University of Sydney, Camden, NSW, Australia

[2]ARC Research Hub for Advanced Prawn Breeding, Townsville, QLD, Australia

[3]Faculty of Science, School of Life and Environmental Sciences, The University of Sydney, Camden, NSW, Australia

Correspondence
Mehar S. Khatkar, Faculty of Science, Sydney School of Veterinary Science, The University of Sydney, 425 Werombi Road, Camden, NSW 2570, Australia.
Email: mehar.khatkar@sydney.edu.au

Funding information
Australian Research Council

Abstract

Genetic parameters of growth, reproductive efficiency and survival are important for designing successful genetic improvement programmes. In this study, a database was assembled and a meta-analysis was conducted using 87 peer-reviewed published studies on genetic parameters of 63 economically important traits reported across nine major farmed shrimp species. A total of 641 estimates of heritability, 978 estimates of genetic correlations and 658 estimates of phenotypic correlations were included in this study. Pacific white shrimp (*Penaeus vannamei*) was the most widely reported species followed by giant freshwater prawn (*Marcobrachium rosenbergii*) and black tiger prawn (*Penaeus monodon*). Published studies demonstrate large variability across heritability and correlation estimates. Consensus estimates were based on both unweighted and weighted averages. The traits were broadly classed into weight, morphology, growth rate, body composition, feeding efficiency, reproductive efficiency and survival. This meta-analysis revealed moderate-to-high heritability estimates for weight (0.30 ± 0.002), morphological (0.26 ± 0.003), growth rate (0.27 ± 0.01) and feeding efficiency (0.58 ± 0.07) traits, but low for body composition (0.06 ± 0.009), reproductive efficiency (0.11 ± 0.006) and survival traits (0.15 ± 0.003). Strong responses to selection were found among the shrimp species studied. We have developed a ShinyApp (https://khatkar.shinyapps.io/shrimp_genetics/) for visualization of the individual heritability estimates. In addition to the meta-analysis of genetic parameters, we have calculated the correlated responses among economically important traits and highlighted response to selection, realized heritability, social interaction and common environmental effects in shrimp breeding programmes in terms of industry implications. Consensus estimates suggest that there is sufficient additive genetic variation to carry out effective genetic improvement programmes in these species.

KEYWORDS

genetic and phenotypic correlation, genetic improvement, heritability, prawn, review

1 | INTRODUCTION

Aquaculture has emerged as a food source of increasing importance to the growing global human population. It is evident that an expanding population and change in socio-demographics have led to an increase in demand for animal-based protein (Henchion, Hayes, Mullen, Fenelon, & Tiwari, 2017). Although shrimp and shrimp products have historically been viewed as a high-value species and considered a luxury product, demand for shrimp has grown with the increase in demand for animal-based protein. In response, annual global aquaculture production of shrimp has grown from 1.2 million tonnes to 4.8 million tonnes over the last two decades to meet this demand (Anderson, Valderrama, & Jory, 2018). Three shrimp species, Pacific white shrimp (*Penaeus vannamei*, Penaeidae), black tiger prawn (*Penaeus monodon*, Penaeidae) and giant freshwater prawn (*Marcobrachium rosenbergii*, Palaemonidae), account for 65% of the total global annual production (FAO, 2018). Although significant work has been done for improving hatchery and seed production techniques in Pacific white shrimp, much more information is needed for improved farming of other shrimp species (Norman-López et al., 2016).

Animals that grow faster, reproduce well, have strong resistance to pathogens and resilient to environmental variations are the key to profitable aquaculture production systems (Gjedrem, Robinson, & Rye, 2012) and form the basis of genetic improvement programmes. Estimates of genetic parameters of these economically important traits, including heritability, and genetic and phenotypic correlations, are the foundation to the development of breeding programmes and genetic improvement of shrimp (Gjedrem, 2012). Over the last few decades, a number of studies on genetic parameter estimates for commercially important traits in shrimp have been published. Heritability estimates have been derived using a variety of methods and study designs and often have varying reliability. There is a large disparity among estimates across studies as well as between estimates within a species. Although genetic parameter estimates are considered population- and species-specific, it is likely that small sample sizes, statistical models and limited data used to estimate genetic parameters, at the individual study level, can cause variation between studies (Gitterle, Ødegård, Gjerde, Rye, & Salte, 2006; Visscher, Hill, & Wray, 2008). Furthermore, shrimp are typically raised in environments which experience extreme variation where management practices vary within species and between farms; consequently, genetic parameter estimates show a lack of consensus. Moreover, due to the low heritability estimates of some traits of economic importance, as well as a small resource population of phenotyped shrimp, the estimates of genetic parameters in individual studies may lack precision. As a result of variation among published estimates, consensus estimates of genetic parameters have not been available to shrimp breeders and researchers. To obtain a consensus summary of the available estimates, a statistical meta-analysis was conducted which combined data from multiple studies to derive pooled estimates of parameters that are closest to the unknown true parameters (Sutton et al., 2000).

The purpose of this study was to provide a quantitative, unbiased synthesis of published findings and provide summary estimates of important genetic parameters for economically important traits. The resultant outputs will enable the expedited development of breeding schemes for nine of the leading commercially farmed shrimp

species internationally. The cumulative use of data available in a meta-analysis summary can aid in improving the precision of the consensus estimates of the genetic parameters. We have developed a ShinyApp for visualization of the individual heritability estimates. In addition to the meta-analysis of genetic parameters, we have calculated the correlated responses among economically important traits and highlight response to selection, realized heritability and genomic selection in shrimp breeding programmes in terms of industry implications. Furthermore, we have reviewed studies on non-additive genetic variation and impact of genotype-by-environment interaction in shrimp breeding programmes. A review of this nature had not yet been performed on the genetic parameters of commercially farmed shrimp species.

2 | MATERIALS AND METHODS

2.1 | Data collection

A systematic approach to the inclusion of literature related to genetic parameter estimates of shrimp production traits was applied. The species were selected for their recognized presence in farmed shrimp production, these species are as follows: Pacific white shrimp, black tiger prawn, banana shrimp (*Fenneropenaeus merguiensis*, Penaeidae), Chinese white shrimp (*Fenneropenaeus chinensis*, Penaeidae), giant freshwater prawn, Indian white prawn (*Fenneropenaeus indicus*, Penaeidae), kuruma prawn (*Penaeus japonicus*, Penaeidae), green tiger prawn (*Penaeus semisulcatus*, Penaeidae) and western blue prawn (*Penaeus stylirostris*, Penaeidae). Studies were included if they met the following criteria: (a) heritability estimates were provided for commercially important traits to one of nine shrimp species; (b) the studies clearly identify the method of evaluation used to estimate heritabilities; (c) studies that contained genetic and phenotypic correlations were included, but lack of these estimates was not cause for exclusion; (d) presented estimates of variation (*SE* or *SD*) for the genetic parameter estimates, however, lack of these variation measures was also not cause for exclusion; and (e) they were published in English.

Studies targeting genetic parameter estimates for shrimp were extracted using the Google Scholar and PubMed search engines with the topic search terms 'shrimps', 'prawn' and 'genetic parameter estimates'. The latest date of literature search was December 2019. The preliminary search yielded 98 articles on the search topics. However, 11 studies were removed because they were either not available in English or provided inadequate information for the meta-analysis. The final data were sourced from 87 peer-reviewed studies (Table 1). For each study, the year of publication, species name, country of study, number of individuals and families used, trait means, genetic parameter estimates and software used to calculate genetic parameters were recorded. The studies included here report original research conducted in Mexico, China, Hawaii, Thailand, Vietnam, Indonesia, the United States of America, Australia, India and Columbia. In total, 641 estimates of heritability, 978 genetic and 658 phenotypic correlation estimates across 63 commercially relevant traits of shrimp production were included in this study. The total number of genetic parameter estimates for the nine selected species may be greater than what has been included in this analysis during the specified time period, this can be as a result of data that are unpublished or unavailable to the public. Data were recorded to four decimal places, and values were rounded where necessary.

Traits deemed commercially important are those which are relevant to shrimp production globally and are likely to be included as key determinants of breeding objectives. A total of 63 traits were compiled and grouped into eight overarching classes: weight (7), morphology (23), growth rate (2), body composition (6), feeding efficiency (3), reproductive efficiency (14), survival measurements (6) and other (2). A list of traits and the corresponding abbreviations are presented in Table 2a,b. In Table 2a, trait abbreviations are described, and in Table 2b, trait groups are defined along with their basis of classification.

Traits were measured at various time points post-stocking; consequently, traits were further grouped into three broad periods, pre-harvest (0–119 days), harvest (120–180 days) and mature (>180 days). These three classifications were constructed around the optimal harvest time for the shrimp species included in this study (Araneda, Hernández, & Gasca-Leyva, 2011; Coman, Arnold, Wood, & Kube, 2010). Estimates classified as "pre-harvest" are any measurements that have been taken prior to the optimal harvest period, whilst the "harvest" classification encompasses measurements taken during the optimal harvest period and estimates deemed "mature" are those based on measurements post-optimal harvest period. In cases where no age information was provided, the data were used only to compute overall estimates.

2.2 | Data processing

The data set comprised estimates of three genetic parameters, namely heritability, and genetic and phenotypic correlations as described above. There were a few heritability estimates outside of the zero to one range, and these were excluded from the meta-analysis. Phenotypic means were recorded for each trait where available, along with sample size and standard deviation. The recorded parameter estimates utilized in this meta-analysis were analysed under the assumption of normality. This assumption was tested by creating box plots for traits with more than 20 estimates and observing the distribution, in all cases the distributions were approximately normal, after relevant data transformation (see Appendix S1). Where more than one report of a specific genetic parameter estimate or trait mean was available in the literature, the values were averaged in an unweighted and weighted manner, and presented separately. Where only one value was available, the data were presented as given in the original study. All parameter estimates and trait means were separated into groups relative to the time (number of days)

TABLE 1 Number of studies, study ID, species, country and trait abbreviation[a] of the traits studied

Study No.	Species	Origin	Trait abbreviation[a]	References
1	Pacific white shrimp	Mexico	BW, NN	Caballero-Zamora, Montaldo, et al. (2015)
2	Pacific white shrimp	China	BW	Lu et al. (2016)
3	Pacific white shrimp	Mexico	BW	Castillo-Juárez et al. (2007)
4	Pacific white shrimp	China	BW, BL, CL, CAW, CAH	Sun et al. (2015)
5	Pacific white shrimp	Australia	BW	Coman et al. (2010)
6	Pacific white shrimp	Australia	EN, NN	Macbeth et al. (2007)
7	Pacific white shrimp	Mexico	BW, SUR	Campos-Montes et al. (2013)
8	Pacific white shrimp	China	BW	Luan et al. (2015)
9	Pacific white shrimp	Mexico	TL, AL, BW, FASW	Pérez-Rostro and Ibarra (2003b)
10	Pacific white shrimp	USA	SUR	Moss et al. (2013)
11	Pacific white shrimp	Colombia	BW, SUR	Gitterle, Salte, et al. (2005)
12	Pacific white shrimp	China	GR	Luan et al. (2012)
13	Pacific white shrimp	China	GR	Sui, et al. (2016)
14	Pacific white shrimp	China	BW, SUR	Li et al. (2015)
15	Pacific white shrimp	Hawaii	BW, BL, CL, AL, SUR	Tan, Luan, et al. (2017)
16	Black tiger prawn	Australia	GR, BW	Kenway et al. (2006)
17	Banana shrimp	Australia	BW, BL, HL, BW, TL, MY	Nguyen et al. (2014)
18	Giant freshwater prawn	Vietnam	BW, BL, HL, BWI, TW, MY	Hung et al. (2013)
19	Pacific white shrimp	China	BW, BL, TL, FASD, FASW, TASD, PCL	Andriantahina, Liu, Huang, and Xiang (2013)
20	Giant freshwater prawn	Vietnam	BW, BL, CL, CW, AL, AW	Hung et al. (2014)
21	Pacific white shrimp	China	SUR	Lu, Luan, Cao, Sui, et al. (2017)
22	Giant freshwater prawn	Thailand	BW, CLAL, TL, BL, CL	Kitcharoen, Rungsin, Koonawootrittriron, and Na-Nakorn (2012)
23	Pacific white shrimp	China	BW	Sui, Luan, Luo, Meng, Cao, et al. (2016)
24	Pacific white shrimp	Mexico	BW	Montaldo et al. (2013)
25	Pacific white shrimp	China	BW, BL, EN, NN, SF, K, SS	Tan, et al. (2017)
26	Pacific white shrimp	Colombia	BW, SUR	Gitterle, Rye, et al. (2005)
27	Pacific white shrimp	USA	GR, SUR, BW	Argue et al. (2002)
28	Pacific white shrimp	Mexico	BW	Ibarra and Famula (2008)
29	Pacific white shrimp	China	TL, BL, FASD, TASD, FASW	Andriantahina et al. (2012)
30	Pacific white shrimp	China	BL, SUR, SURT	Lu, Luan, Cao, Meng, et al. (2017)
31	Black tiger prawn	India	BW, SUR	Krishna et al. (2011)
32	Pacific white shrimp	Mexico	BW, SUR	Alejandra Caballero-Zamora, Cienfuegos-Rivas, et al. (2015)
33	Pacific white shrimp	China	CL, AL, BL, BW, SUR	Zhang et al. (2017)
34	Pacific white shrimp	China	BL, BW, SUR	Zhang et al. (2017)
35	Pacific white shrimp	China	ASL, BL, BW, SUR	Yuan, Hu, Liu, and Zhang (2018)
36	Pacific white shrimp	China	BW, SUR	Li, Lu, et al. (2016)
37	Chinese white shrimp	China	BW, BL, SUR, SUR_TIME, SUR_TEMP	Wang, Kong, et al. (2017)
38	Pacific white shrimp	Mexico	BW, TW, TP	Campos-Montes et al. (2017)
39	Pacific white shrimp	China	RFI, FER, ADG, DFI	Dai, Luan, Lu, Luo, Meng, et al. (2017)
40	Pacific white shrimp	Mexico	BW, FA	Nolasco-Alzaga et al. (2018)

(Continues)

TABLE 1 (Continued)

Study No.	Species	Origin	Trait abbreviation[a]	References
41	Banana shrimp	Australia	HPV, BW, BL, BWI, MY, CRS, CCS, FS, YH	Phuthaworn, Nguyen, Quinn, and Knibb (2016)
42	Western blue prawn	France	GR	Goyard et al. (2002)
43	Pacific white shrimp	Mexico	TL, CL, AL, BW, AWT, FASW, CWT	Pérez-Rostro and Ibarra (2003a)
44	Pacific white shrimp	Mexico	BW	Campos-Montes, Montaldo, Martínez-Ortega, and Castillo-Juárez (2009)
45	Green tiger prawn	Egypt	BW, SUR	Megahed (2019)
46	Giant freshwater prawn	Vietnam	BW, FEC, SUR	Vu and Nguyen (2019)
47	Kuruma prawn	China	BLAL, BW	Liu, Zheng, and Liu (2019)
48	Pacific white shrimp	Norway	SUR	Ødegård et al. (2011)
49	Giant freshwater prawn	Indonesia	BW, TL, SL	Imron and Suprapto (2013)
50	Pacific white shrimp	Mexico	BW, SUR	de los Ríos-Pérez, Campos-Montes, Martínez-Ortega, Castillo-Juárez, and Montaldo (2015)
51	Black tiger prawn	Thailand	BW, SUR	Thiengpimol, Koonawootrittriron, and Wongtripop (2014)
52	Pacific white shrimp	China	BW, TL, BL, FASD, TASD, FASW, PCL	Xu, Chinh, Faisal, and Liu (2016)
53	Pacific white shrimp	Mexico	BW, SUR	Caballero-Zamora et al. (2014)
54	Pacific white shrimp	Colombia	BW, SUR	Gitterle, Gjerde, et al. (2006)
55	Pacific white shrimp	Mexico	SUR	Cala, Montaldo, Campos-Montes, and Castillo-Juárez (2014)
56	Giant freshwater prawn	China	SUR	Luan et al. (2014)
57	Giant freshwater prawn	Vietnam	SUR, BL, BW, AW	Vu et al. (2017)
58	Black tiger prawn	Australia	BW, TL	Benzie, Kenway, and Trott (1997)
59	Giant freshwater prawn	Vietnam	BW, REP	Dinh and Nguyen (2014)
60	Chinese white shrimp	China	BW, SUR	Li, Luan, et al. (2016)
61	Black tiger prawn	India	SUR	Hayes et al. (2010)
62	Giant freshwater prawn	USA	GR	Malecha, Masuno, and Onizuka (1984)
63	Chinese white shrimp	China	GR	Zhang et al. (2011)
64	Kuruma prawn	Australia	GR	Hetzel et al. (2000)
65	Pacific white shrimp	Mexico	NS	Ibarra, Arcos, Famula, Palacios, and Racotta (2005)
66	Pacific white shrimp	Mexico	BW, DTFS, EN, ED, VIT, FA	Arcos, Racotta, and Ibarra (2004)
67	Giant freshwater prawn	India	BW, BL, CL	Pillai et al. (2017)
68	Pacific white shrimp	Venezuela	BW, GR, SUR	De Donato et al. (2008)
69	Pacific white shrimp	Mexico	VTG, NO, DO	Ibarra, Famula, and Arcos (2009)
70	Pacific white shrimp	Mexico	SUR	Ibarra, Pérez-Rostro, et al. (2007)
71	Pacific white shrimp	Mexico	NO, DO, OA, DTFS, FFS, FEC, NS	Arcos, Racotta, Palacios, and Ibarra (2005)
72	Pacific white shrimp	Thailand	BW, BL, AL, CL, BW/BL	Kanchanachai, Poompuang, Koonawootrittriron, and Uraiwan (2011)
73	Pacific white shrimp	Mexico	BL, FASL, SASL, CW, FASW, SASW, CWT, AL	Pérez-Rostro, Ramirez, and Ibarra (1999)
74	Giant freshwater prawn	China	SUR	Luan et al. (2015)
75	Indian white prawn	Egypt	SUR	Megahed, Elmesiry, Mohamed, and Dhar (2019)
76	Kuruma prawn	India	GR, SUR	(Gopikrishna et al., 2012)

(Continues)

TABLE 1 (Continued)

Study No.	Species	Origin	Trait abbreviation[a]	References
77	Pacific white shrimp	Mexico	BL	Campos, Montaldo, Arechavaleta, and Castillo-Juárez (2006)
78	Pacific white shrimp	Colombia	BW, SUR	Gitterle et al. (2007)
79	Pacific white shrimp	Colombia	SUR	Gitterle, Gjerde, et al. (2006)
80	Black tiger prawn	Australia	SUR, GAV_LOAD	Noble et al. (2019)
81	Pacific white shrimp	Vietnam	SUR, GR	Trinh et al. (2019)
82	Pacific white shrimp	China	GR, ADG, FER, DFI	Dai et al. (2019)
83	Pacific white shrimp	Vietnam	GR, SUR	Trinh et al. (2019)
84	Giant freshwater prawn	China	BW	Sui, Luan, Luo, Meng, Lu, et al. (2016)
85	Pacific white shrimp	Mexico	BW, SUR	Hernández-Ruíz, Montaldo, Bustos-Martínez, Campos-Montes, and Castillo-Juárez (2019)
86	Chinese white shrimp	China	ADG, FER	Dai, Luan, Lu, Luo, Cao, et al. (2017)
87	Pacific white shrimp	Vietnam	BW	Nguyen, Ninh, and Hung (2019)

[a]Trait abbreviations are described in Table 2a.

measurements were made post-stocking as described above. Where the number of days post-stocking information was not provided, parameter estimates and trait means were still included in the overall calculation.

2.3 | Handling of missing standard errors of parameter estimates

A portion of estimates did not include standard errors of parameter estimates: approximately 33% of genetic, 51% of phenotypic and 4% of heritability standard error (SE) values were missing. For these estimates, a conservative approximate SE value of the upper quartile of known SE estimates was used and attributed to the estimate. This imputation procedure was applied within the trait classes for SE of heritability, whereas all the available SE estimates were used for imputing the missing SE of the correlations. Finally, for the trait means, only values where standard deviation was reported were used in the meta-analysis.

2.4 | Unweighted average trait means and genetic parameters

For each trait mean and the genetic parameter estimates, an unweighted average value was calculated for each of the 63 traits across species. The data were partitioned and re-analysed with time of measurement, that is the three stages. For trait class-specific means, unweighted averages were calculated as above; however, data were subdivided into species before analysis within trait and time of measurement. This was due to the relevance of the physical measurement to both the species and the trait itself as well as the influence of time of measurement on the value itself.

2.5 | Weighted parameter estimates

Weighted genetic parameter and correlation estimates were calculated for all traits where more than one estimate was available, including estimates across species. As indicated above, weighted trait means were also calculated within species and within each individual trait, where more than one estimate was available. Where there was only one estimate or trait mean in any class, the value was recorded as it appeared in the original study.

Previous studies have identified the inverse of the sampling variance, that is inverse of the standard error squared, as the optimal weighting method for a weighted meta-analysis (Hedges & Olkin, 1985; Marín-Martínez & Sánchez-Meca, 2010), and this weighting method was applied here. The detailed procedures for computing weighted estimates for heritability and correlations are provided in Appendix S1. The SE of a weighted estimate provides precision of the consensus estimate. All resultant weighted outputs were tabulated and presented in corresponding tables, alongside unweighted values for comparison and ease of interpretation. Data analysis and visualizations were performed in R (R Core Team, 2020).

3 | RESULTS

The heritability estimates are presented in Table 3 and Table 4, and genetic and phenotypic correlations in Table 5 and Figure 1. Species-specific trait means and weighted heritabilities are presented in Tables S1–S9 (Appendix S2). The distribution of individual heritability estimates for different trait classes is shown in Figure 2. The heritability estimates based on i) whether or not common environmental effect was included in the statistical models, ii) rearing conditions and iii) study origins have been provided in Appendix S2, and

TABLE 2A Definition of traits used across studies and grouped in eight major trait classes[a] in this meta-analysis

Trait abbreviation	Trait	Trait class
AWT	Abdominal weight	WT
ADG	Average daily gain	GR
AGE_PL	Days to reach post-larvae	OTHER
AL	Abdominal length	MOR
ASL	Abdominal segment length	MOR
AW	Abdominal width	MOR
BL	Body length	MOR
BW	Body weight	WT
BW:BL	Body weight-to-body length ratio	MOR
BWI	Body width	MOR
CAH	Carapace height	MOR
CAW	Carapace width	MOR
CL	Carapace length	MOR
CLAL	Claw length	MOR
CW	Cephalothorax width	MOR
CWT	Cephalothorax weight	WT
DC	Dark colour	MOR
DFI	Daily feed intake	FEED
DO	Oocyte diameter	REP
DTFS	Days to first spawn	REP
ED	Egg diameter	REP
EN	Egg number	REP
ET	Egg triacylglycerides	COMP
ETL	Egg total lipid	COMP
ETP	Egg total protein	COMP
EV	Egg vitellin	COMP
FA	Fatty acid	COMP
FASL	First abdominal segment length	MOR
FER	Feed efficiency ratio	FEED
FS	Flesh streaks	MOR
FSAD	First abdominal segment depth	MOR
FSAW	First abdominal segment width	MOR
GR	Growth rate	GR
GAV_LOAD	Gill-associated virus load	SUR
HATCH	Hatching rate	REP
HL	Head length	MOR
HPV	Hepatopancreatic parvovirus titre	SUR
K	Condition factor	REP
MY	Meat yield	WT
NN	Number of nauplii	REP
NO	Number of oocytes	REP
NS	Number of spawns	REP
OA	Ovary area	REP
OM	Ovary maturity	REP
PCL	Partial carapace length	MOR
RC	Red colour	MOR
REP_S	Reproductive status	REP
RFI	Residual feed intake	FEED
SASL	Sixth abdominal segment length	MOR
SASW	Sixth abdominal segment width	MOR
SF	Spawning frequency	REP
SOW	Skeleton off weight	WT
SS	Spawning success	REP
SUR	Survival	SUR
SUR_TEMP	Survival temperature	SUR
SUR_TIME	Time survived after treatment	SUR
TASD	Third abdominal segment depth	MOR
TOW	Telson off weight	WT
TP	Tail percentage	MOR
TW	Tail weight	WT
VTG	Vitellogenin	COMP
VL	Virus load	SUR
YH	Yellow hepatopancreas	OTHER

[a]The trait class abbreviations are described in Table 2b.

Figures S1–S3, respectively. We only draw attention to individual species and trait-specific weighted heritabilities if five or more individual observations were reported.

We have developed a ShinyApp (https://khatkar.shinyapps.io/shrimp_genetics/) for visualization of the database of individual heritability estimates and trait means as reported in different studies. The species and trait group can be selected from the left-hand side panel. A screen grab of the App is shown in Figure S4 (Appendix S2). The specific findings of trait means and heritability estimates are described below.

3.1 | Heritability estimates

3.1.1 | Weight traits

The overall average weighted heritability of for all traits grouped in the weight class was moderate with a value of 0.30 ± 0.002 (Table 4). Age-specific heritability analysis revealed a moderate estimate of 0.26 ± 0.002 for weight traits at pre-harvest

stage, whilst the estimates were slightly higher at the harvest (0.30 ± 0.005) and mature stage (0.36 ± 0.01). Of the traits that comprise the weight class, BW was recorded most extensively with 185 observations for this trait. The heritability estimates for this class ranged from 0.001 to 0.89. The mean weighted heritability estimates for BW differed across time of measurement; the heritability estimate for body weight was lowest at the pre-harvest (0.26 ± 0.003), followed by the harvest (0.29 ± 0.005) and highest at the mature time period (0.35 ± 0.015) (Table 3). For Pacific white shrimp, average weighted heritability for BW was in the range of 0.30–0.32 depending on the age of measurement, whilst for black tiger prawn it was higher in the range of 0.39–0.48. For Chinese white shrimp, the average heritability estimate for body weight was 0.26 ± 0.01 and lowest average estimate was observed for kuruma prawn (0.18 ± 0.01), whilst for giant freshwater prawn this estimate ranged from 0.14 to 0.41 (Appendix S2, Tables S1–S9).

Among the traits included in the weight class, cephalothorax weight (CWT) is the most heritable (0.81 ± 0.12) at the pre-harvest period, and meat yield (MY) was least heritable (0.04 ± 0.02). Growth rate, another very important trait for aquaculture, showed a low estimate of heritability of 0.16 ± 0.01 at the pre-harvest time but a high estimate at the harvest period (0.45 ± 0.05) and mature stage (0.43 ± 0.03) (Table 3). For Pacific white shrimp, average weighted heritability for GR ranged from 0.47 to 0.53, whilst for giant freshwater prawn estimates of weighted average heritability for GR were substantially lower at 0.11 (Appendix S2, Tables S1–S9).

3.1.2 | Morphometric traits

There were 213 individual reported estimates of heritability across the 23 traits included in the morphometric class (Table 3). The overall weighted heritability estimates of the morphometric trait class are comparable across different life stages with the average value of 0.26 ± 0.005, 0.27 ± 0.004 and 0.23 ± 0.002 for pre-harvest, harvest and mature stages, respectively (Table 3). Body length (BL) was the most frequently reported trait ($n = 87$) in the morphological trait class. The mean heritability of BL trait was 0.29 ± 0.006 across all stages. For Pacific white shrimp, average weighted heritability for BL was of 0.35 ± 0.01 and 0.32 ± 0.01 at pre-harvest and harvest stages, respectively. However, for kuruma prawn heritability estimates were lower at 0.17 ± 0.02 for BL measured across all stages.

Among the morphological trait class, the highest heritability was observed for width of the sixth abdominal segment (SASW) with a mean value of 0.77 ± 0.12, and the lowest estimate was reported for claw length (CLAL) (0.01 ± 0.02; Table 3). Species-specific heritability estimates of note were seen for Pacific white shrimp where average weighted heritability for abdominal length (AL) was in the range of 0.25 to 0.34 depending on the age of measurement, with relatively

TABLE 2B Definition of trait classes, according to which individual traits were classified

Trait class	Abbreviation	Definition and corresponding trait abbreviations[a]
Weight	WT	Aspects of body weight of the shrimp: AWT, BW, CWT, MY, SOW, TW.
Morphology	MOR	All aspects of morphological features: AL, ASL, AW, BL, BW:BL, BWI, CAH, CAW, CL, CLAL, CW, DC, FASL, FS, FSAD, FSAW, HL, PCL, RC, SASL, SASW, TASD, TL, TP.
Growth rate	GR	Daily weight gain or difference in weight gain after specific time period: GR.
Body composition	COMP	Lipid or fatty acid contents in various body composition estimates: ET, ETL, ETP, EV, FA, VTG.
Feeding efficiency	FEED	Daily feed intake and feed efficiency: DFI, FER, RFI.
Reproductive efficiency	REP	Various aspects of reproduction: REP_STATUS, DO, DTFS, ED, EN, HATCH, K, NN, NO, NS, OA, OM, SF, SS, VTG.
Survival	SUR	Overall survival (%) after certain period/treatment, or in any challenge test, any measure of survival, including pathogenic and environmental challenge: SUR, SUR_TIME, SUR_TEMP.
Other	OTHER	Traits those do not fit within any of the above trait classes: AGE_PL and YH.

[a]Trait abbreviations are described in Table 2a.

higher estimates for third abdominal segment depth (TASD) at 0.30 to 0.41, 0.27 to 0.40 for first abdominal segment depth (FASD) and 0.37 for first abdominal segment width (FASW). In the same species, average weighted heritability estimates for carapace length (CL) were lower and ranged from 0.17 to 0.22 and higher for PCL with estimates ranging from 0.33 to 0.42. For kuruma prawn, weighted average heritability for AL was 0.17 ± 0.01, whilst for giant freshwater prawn it was 0.22 ± 0.01 with similar estimates for abdominal width (AW) 0.22 ± 0.01 and for carapace width (CW) 0.22 ± 0.01 in this species. All the individual estimates can be found in Appendix S2 (Tables S1–S9).

3.1.3 | Growth rate

For growth trait class, there was only a single trait named growth rate itself. There were altogether 28 reported estimates of heritability for this trait. The overall heritability estimate was 0.27 ± 0.01, and estimates ranged from 0.05 to 0.84. The heritability estimate was higher in harvest (0.45 ± 0.05) and mature (0.43 ± 0.03) stages as compared to pre-harvest stage (0.16 ± 0.01) (Table 4).

3.1.4 | Body composition and feed traits

There were 33 estimates of heritability reported for the traits included in the body composition class. Overall, the heritability for body composition was low with a mean estimate of 0.06 ± 0.009 ranging from 0.001 to 0.29 (Table 3). Heritability for traits included in this class was only recorded at harvest or mature stage (Table 4).

Traits related to feed trait class include daily feed intake (DFI), feeding efficiency ratio (FER) and residual feed intake (RFI). A total of seven heritability estimates have been reported for these traits, all of which were categorised in the pre-harvest stage class. The heritability estimate of the feed-related trait was high and ranged from 0.25 to 0.70 with a mean value of 0.58 ± 0.07 (Table 4).

3.1.5 | Reproduction traits

The overall heritability estimate for the 14 traits included in the reproduction class was 0.11 ± 0.006 (Table 4). The highest mean heritability estimate was observed for ovary maturity (OM) (0.71 ± 0.26), and the lowest estimate was recorded for egg diameter (ED) trait (0.001 ± 0.001). All the heritability estimates in this class were reported for the harvest and mature stages. This is likely due to traits of reproductive performance needing to be measured at the point of sexual maturity which for all species occurs in the mature stage. For individual species, very few studies on reproductive traits (e.g. fecundity and reproductive status) have sufficient replication to warrant specific mention, with the only exception for giant freshwater prawn where a weighted average heritability for reproductive status was 0.17 ± 0.01 based on eight studies (Appendix S2, Tables S1–S9).

3.1.6 | Survival traits

The survival trait class is comprised of traits related to survival including, survival (SUR), survival time (SUR_TIME), survival temperature (SUR_TEMP) and virus load (VL), which is measured as a percentage of animals which have survived after experimental treatments. However, we have highlighted the challenges and limitations of conducting meta-analysis of this group of traits in the discussion section. This meta-analysis included 119 measures of survival across a number of different species. Overall, mean estimates suggest heritability of this trait class is low, 0.15 ± 0.003. Survival time and survival temperature had five and one observation, respectively, with the mean heritability estimates of 0.46 ± 0.03 and 0.07 ± 0.05, respectively. Virus load had four heritability estimates with a mean value of 0.18 ± 0.01. For Pacific white shrimp, average weighted heritability for SUR was 0.12 ± 0.004, whilst for giant freshwater prawn it was higher in the range of 0.21 ± 0.01 and highest for green tiger prawn at 0.31 ± 0.01.

3.2 | Genetic and phenotypic correlations

Weighted and unweighted correlation estimates were calculated among each of the broad trait class, viz. weight, growth, morphometric, reproduction and survival. Weighted means were also calculated for all traits within the same trait class and across classes, for example morphometric traits were compared with other morphometric traits as well as all other trait classes, including weight, reproduction, survival and growth. The number of correlation estimates contributing to weighted mean genetic and phenotypic correlations for the trait class pairs ranged from 2 to 217 (Table 5). There were more estimates of genetic correlation than phenotypic correlation. The details on the estimates within and across trait classes are described in the following sections.

3.2.1 | Weight traits

There were 394 genetic and 300 phenotypic correlation estimates reported for weight-related traits. The weighted average estimate for genetic correlation among the traits within the weight trait class was 0.64 ± 0.01 and ranged from −0.39 to 0.99 (Table 5). There were 47 phenotypic correlations among weight traits within the range of −0.06 to 0.99 with a weighted average of 0.70 ± 0.006. High genetic and phenotypic correlations were found between weight and morphological traits, with values of 0.81 ± 0.002 and 0.68 ± 0.002, respectively (Table 5). For weight traits, there were negative genetic and phenotypic correlations with body composition traits (−0.13 ± 0.21 and −0.13 ± 0.03 for genetic and phenotypic

TABLE 3 Trait-wise heritability (h^2) estimates pooled across nine shrimp species

Trait Abbreviation[b]	Stage[a]	*n*	Weighted h^2	Min	Max	Unweighted h^2
AWT	Overall	8	0.31 ± 0.04	0.16	0.97	0.46 ± 0.11
AWT	Pre-harvest	7	0.36 ± 0.05	0.16	0.97	0.50 ± 0.12
AWT	Mature	1	0.18 ± 0.06	0.18	0.18	0.18
BW	Overall	185	0.30 ± 0.002	0.001	0.89	0.30 ± 0.01
BW	Pre-harvest	59	0.26 ± 0.003	0.01	0.89	0.32 ± 0.02
BW	Harvest	89	0.29 ± 0.005	0.001	0.8	0.27 ± 0.01
BW	Mature	36	0.35 ± 0.01	0.02	0.71	0.34 ± 0.02
CWT	Overall	3	0.33 ± 0.09	0.15	0.89	0.60 ± 0.23
CWT	Pre-harvest	2	0.81 ± 0.12	0.78	0.89	0.83 ± 0.05
CWT	Mature	1	0.15 ± 0.06	0.15	0.15	0.15
GR	Overall	28	0.27 ± 0.01	0.05	0.84	0.38 ± 0.04
GR	Pre-harvest	14	0.16 ± 0.01	0.05	0.84	0.36 ± 0.07
GR	Harvest	2	0.45 ± 0.05	0.25	0.50	0.37 ± 0.12
GR	Mature	11	0.43 ± 0.03	0.12	0.55	0.43 ± 0.04
MY	Overall	1	0.04 ± 0.02	0.04	0.04	0.04
MY	Harvest	1	0.04 ± 0.02	0.04	0.04	0.04
SOW	Overall	4	0.31 ± 0.05	0.20	0.41	0.27 ± 0.04
SOW	Pre-harvest	4	0.31 ± 0.05	0.20	0.41	0.27 ± 0.04
TOW	Overall	4	0.28 ± 0.04	0.17	0.39	0.24 ± 0.05
TOW	Pre-harvest	4	0.28 ± 0.04	0.17	0.39	0.24 ± 0.05
TW	Overall	2	0.38 ± 0.06	0.16	0.49	0.32 ± 0.16
TW	Harvest	2	0.38 ± 0.06	0.16	0.49	0.32 ± 0.16
SUR	Overall	109	0.14 ± 0.002	0.00	0.43	0.11 ± 0.01
SUR	Pre-harvest	55	0.17 ± 0.003	0.001	0.43	0.16 ± 0.01
SUR	Harvest	40	0.07 ± 0.004	0.001	0.40	0.05 ± 0.01
SUR	Mature	7	0.10 ± 0.009	0.008	0.14	0.08 ± 0.02
SUR_TEMP	Overall	1	0.07 ± 0.05	0.07	0.07	0.07
SUR_TEMP	Pre-harvest	1	0.07 ± 0.05	0.07	0.07	0.07
SUR_TIME	Overall	5	0.46 ± 0.03	0.15	0.78	0.45 ± 0.11
SUR_TIME	Pre-harvest	5	0.46 ± 0.03	0.15	0.78	0.45 ± 0.11
DO	Overall	2	0.31 ± 0.11	0.23	0.57	0.40 ± 0.17
DO	Mature	2	0.31 ± 0.11	0.23	0.57	0.40 ± 0.17
DTFS	Overall	2	0.44 ± 0.12	0.41	0.47	0.44 ± 0.03
DTFS	Mature	1	0.41 ± 0.20	0.41	0.41	0.41
ED	Overall	2	0.001 ± 0.001	0.001	0.01	0.006 ± 0.004
ED	Mature	2	0.001 ± 0.001	0.001	0.01	0.006 ± 0.004
EN	Overall	4	0.17 ± 0.04	0.09	0.41	0.18 ± 0.07
EN	Harvest	1	0.13 ± 0.04	0.13	0.13	0.13
EN	Mature	2	0.11 ± 0.07	0.09	0.12	0.10 ± 0.01
HATCH	Overall	1	0.18 ± 0.16	0.18	0.18	0.18
K	Overall	1	0.10 ± 0.06	0.10	0.10	0.10
K	Mature	1	0.10 ± 0.06	0.10	0.10	0.10
NN	Overall	3	0.02 ± 0.01	0.001	0.27	0.10 ± 0.08
NN	Harvest	1	0.03 ± 0.04	0.03	0.03	0.03

(Continues)

TABLE 3 (Continued)

Trait Abbreviation[b]	Stage[a]	*n*	Weighted h^2	Min	Max	Unweighted h^2
NN	Mature	1	0.001 ± 0.001	0.001	0.001	0.001
NO	Overall	2	0.01 ± 0.009	0.001	0.11	0.05 ± 0.05
NO	Mature	2	0.01 ± 0.009	0.001	0.11	0.05 ± 0.05
NS	Overall	1	0.20 ± 0.15	0.20	0.20	0.20
NS	Mature	1	0.20 ± 0.15	0.20	0.20	0.20
OA	Overall	1	0.001 ± 0.001	0.001	0.001	0.001
OA	Mature	1	0.001 ± 0.001	0.001	0.001	0.001
OM	Overall	1	0.71 ± 0.26	0.71	0.71	0.71
OM	Mature	1	0.71 ± 0.26	0.71	0.71	0.71
REP_STATUS	Overall	8	0.17 ± 0.01	0.001	0.56	0.14 ± 0.06
REP_STATUS	Mature	8	0.17 ± 0.01	0.001	0.56	0.14 ± 0.06
SF	Overall	1	0.06 ± 0.06	0.06	0.06	0.06
SF	Mature	1	0.06 ± 0.06	0.06	0.06	0.06
SS	Overall	1	0.37 ± 0.14	0.37	0.37	0.37
SS	Mature	1	0.37 ± 0.14	0.37	0.37	0.37
AGE_PL	Overall	1	0.30 ± 0.28	0.30	0.30	0.30
AGE_PL	Pre-harvest	1	0.30 ± 0.28	0.30	0.30	0.30
HPV	Overall	1	0.41 ± 0.08	0.41	0.41	0.41
HPV	Harvest	1	0.41 ± 0.08	0.41	0.41	0.41
YH	Overall	1	0.02 ± 0.01	0.02	0.02	0.02
YH	Harvest	1	0.02 ± 0.01	0.02	0.02	0.02
AL	Overall	23	0.20 ± 0.01	0.07	0.78	0.29 ± 0.04
AL	Pre-harvest	16	0.20 ± 0.01	0.07	0.78	0.31 ± 0.05
AL	Harvest	5	0.19 ± 0.03	0.11	0.35	0.22 ± 0.04
AL	Mature	2	0.25 ± 0.06	0.23	0.33	0.28 ± 0.05
ASL	Overall	4	0.35 ± 0.04	0.28	0.42	0.34 ± 0.03
ASL	Pre-harvest	3	0.37 ± 0.04	0.28	0.42	0.37 ± 0.04
ASL	Harvest	1	0.28 ± 0.11	0.28	0.28	0.28
AW	Overall	7	0.22 ± 0.01	0.10	0.35	0.20 ± 0.03
AW	Pre-harvest	6	0.22 ± 0.01	0.10	0.35	0.20 ± 0.03
AW	Harvest	1	0.22 ± 0.04	0.22	0.22	0.22
BL	Overall	87	0.29 ± 0.006	0.07	0.86	0.34 ± 0.02
BL	Pre-harvest	47	0.29 ± 0.009	0.09	0.86	0.41 ± 0.03
BL	Harvest	34	0.29 ± 0.007	0.07	0.46	0.27 ± 0.01
BL	Mature	6	0.26 ± 0.03	0.17	0.51	0.28 ± 0.04
BW_BL	Overall	2	0.42 ± 0.11	0.34	0.58	0.46 ± 0.12
BW_BL	Pre-harvest	1	0.58 ± 0.20	0.58	0.58	0.58
BW_BL	Harvest	1	0.34 ± 0.13	0.34	0.34	0.34
BWI	Overall	1	0.45 ± 0.03	0.45	0.45	0.45
BWI	Harvest	1	0.45 ± 0.03	0.45	0.45	0.45
CAH	Overall	1	0.13 ± 0.01	0.13	0.13	0.13
CAH	Harvest	1	0.13 ± 0.01	0.13	0.13	0.13
CAW	Overall	1	0.15 ± 0.01	0.15	0.15	0.15
CAW	Harvest	1	0.15 ± 0.01	0.15	0.15	0.15

(Continues)

TABLE 3 (Continued)

Trait Abbreviation[b]	Stage[a]	*n*	Weighted h^2	Min	Max	Unweighted h^2
CL	Overall	22	0.18 ± 0.007	0.05	0.59	0.21 ± 0.02
CL	Pre-harvest	12	0.18 ± 0.01	0.06	0.59	0.24 ± 0.04
CL	Harvest	6	0.17 ± 0.009	0.05	0.26	0.16 ± 0.03
CL	Mature	4	0.20 ± 0.03	0.17	0.25	0.20 ± 0.01
CLAL	Overall	1	0.01 ± 0.02	0.01	0.01	0.01
CLAL	Harvest	1	0.01 ± 0.02	0.01	0.01	0.01
CW	Overall	8	0.23 ± 0.01	0.10	0.94	0.35 ± 0.10
CW	Pre-harvest	7	0.23 ± 0.01	0.10	0.94	0.38 ± 0.11
CW	Harvest	1	0.19 ± 0.03	0.19	0.19	0.19
DC	Overall	1	0.18 ± 0.05	0.18	0.18	0.18
DC	Harvest	1	0.18 ± 0.05	0.18	0.18	0.18
FASD	Overall	9	0.30 ± 0.008	0.17	0.41	0.28 ± 0.03
FASD	Pre-harvest	1	0.40 ± 0.02	0.40	0.40	0.40
FASD	Harvest	8	0.27 ± 0.009	0.17	0.41	0.27 ± 0.03
FASL	Overall	3	0.67 ± 0.11	0.59	0.89	0.74 ± 0.08
FASL	Pre-harvest	2	0.65 ± 0.12	0.59	0.74	0.66 ± 0.07
FASL	Harvest	1	0.89 ± 0.21	0.89	0.89	0.89
FASW	Overall	15	0.33 ± 0.01	0.14	0.97	0.40 ± 0.06
FASW	Pre-harvest	4	0.43 ± 0.02	0.22	0.97	0.64 ± 0.18
FASW	Harvest	9	0.34 ± 0.01	0.25	0.43	0.33 ± 0.02
FASW	Mature	2	0.18 ± 0.05	0.14	0.35	0.24 ± 0.1
FS	Overall	1	0.03 ± 0.02	0.03	0.03	0.03
FS	Harvest	1	0.03 ± 0.02	0.03	0.03	0.03
HL	Overall	1	0.14 ± 0.03	0.14	0.14	0.14
HL	Harvest	1	0.14 ± 0.03	0.14	0.14	0.14
PCL	Overall	9	0.37 ± 0.01	0.23	0.42	0.32 ± 0.02
PCL	Pre-harvest	1	0.42 ± 0.03	0.42	0.42	0.42
PCL	Harvest	8	0.33 ± 0.01	0.23	0.42	0.31 ± 0.02
RC	Overall	1	0.08 ± 0.03	0.08	0.08	0.08
RC	Harvest	1	0.08 ± 0.03	0.08	0.08	0.08
SASL	Overall	2	0.70 ± 0.12	0.67	0.74	0.70 ± 0.03
SASL	Pre-harvest	2	0.70 ± 0.12	0.67	0.74	0.70 ± 0.03
SASW	Overall	2	0.77 ± 0.12	0.74	0.82	0.78 ± 0.04
SASW	Pre-harvest	2	0.77 ± 0.12	0.74	0.82	0.78 ± 0.04
TASD	Overall	9	0.32 ± 0.01	0.19	0.42	0.30 ± 0.03
TASD	Pre-harvest	1	0.41 ± 0.03	0.41	0.41	0.41
TASD	Harvest	8	0.30 ± 0.01	0.19	0.42	0.28 ± 0.03
TP	Overall	3	0.11 ± 0.03	0.05	0.24	0.13 ± 0.05
TP	Pre-harvest	1	0.24 ± 0.11	0.24	0.24	0.24
TP	Harvest	2	0.09 ± 0.02	0.05	0.12	0.08 ± 0.03
DFI	Overall	2	0.68 ± 0.17	0.66	0.69	0.68 ± 0.01
DFI	Pre-harvest	2	0.68 ± 0.17	0.66	0.69	0.68 ± 0.01
FER	Overall	2	0.62 ± 0.17	0.57	0.68	0.63 ± 0.05
FER	Pre-harvest	2	0.62 ± 0.17	0.57	0.68	0.63 ± 0.05

(Continues)

TABLE 3 (Continued)

Trait Abbreviation[b]	Stage[a]	*n*	Weighted h^2	Min	Max	Unweighted h^2
RFI	Overall	2	0.67 ± 0.17	0.64	0.70	0.67 ± 0.03
RFI	Pre-harvest	2	0.67 ± 0.17	0.64	0.70	0.67 ± 0.03
ET	Overall	1	0.20 ± 0.20	0.20	0.20	0.20
ET	Mature	1	0.20 ± 0.20	0.20	0.20	0.20
ETL	Overall	1	0.001 ± 0.001	0.001	0.001	0.001
ETL	Overall	1	0.001 ± 0.001	0.001	0.001	0.001
ETP	Overall	1	0.13 ± 0.23	0.13	0.13	0.13
ETP	Mature	1	0.13 ± 0.23	0.13	0.13	0.13
EV	Overall	1	0.28 ± 0.22	0.28	0.28	0.28
EV	Mature	1	0.28 ± 0.22	0.28	0.28	0.28
FA	Overall	28	0.06 ± 0.01	0.001	0.19	0.07 ± 0.01
FA	Harvest	28	0.06 ± 0.01	0.001	0.19	0.07 ± 0.01
VL	Overall	4	0.18 ± 0.01	0.08	0.42	0.22 ± 0.07
VL	Pre-harvest	2	0.18 ± 0.01	0.18	0.21	0.19 ± 0.01
VL	Harvest	2	0.17 ± 0.16	0.08	0.42	0.25 ± 0.17
VTG	Overall	1	0.29 ± 0.12	0.29	0.29	0.29
VTG	Mature	1	0.29 ± 0.12	0.29	0.29	0.29

[a]Stages (pre-harvest = <120 days, harvest = 120–180 days and mature = >180 days), number of observations (*n*), weighted and unweighted heritability estimates with minimum (Min) and maximum (Max) value of range.

[b]Trait abbreviations are described in Table 2a.

correlations, respectively), suggesting a potential trade-off among these traits.

3.2.2 | Morphometric traits

A total of 450 genetic and 459 phenotypic correlation estimates were recorded for morphological traits. A weighted average mean for the genetic correlations among morphological traits was 0.77 ± 0.003 with a range of −0.99 to 0.99 for individual estimates. The corresponding average phenotypic correlation was 0.62 ± 0.002 (range: −0.99 to 0.99). Morphological traits depicted very low phenotypic correlation with reproduction traits (e.g. 0.03 ± 0.06 and 0.01 ± 0.004 for genetic and phenotypic correlations, respectively; Table 5).

3.2.3 | Growth, survival and reproductive traits

There were 58 genetic and 32 phenotypic correlations for growth traits of shrimp species (Table 5). High genetic and phenotypic correlation estimates were found among growth traits at 0.45 ± 0.05 and 0.35 ± 0.01, respectively. No phenotypic correlations were reported between growth traits and any other trait classes. Genetic correlations between traits in the growth class and reproduction class were low (0.05 ± 0.11), similarly for growth and survival (0.17 ± 0.07).

For survival traits, a total of 102 genetic and 83 phenotypic correlations were recorded. The highest weighted genetic correlation was 0.62 ± 0.01 among survival traits (range: −0.99 to 0.99), whilst the weighted average phenotypic correlation was 0.49 ± 0.003 among survival traits. There were 87 reports for genetic and 56 observations for phenotypic correlations of reproductive traits. Overall, the weighted phenotypic correlation estimate was low (0.07 ± 0.01), and similarly, genetic correlation was close to zero among reproduction traits (0.09 ± 0.09; Table 5).

3.2.4 | Correlated response among various trait classes

In selective breeding programmes, it is important to understand how the improvement of one trait can cause simultaneous change in other traits. Here, we present estimates of predicted correlated responses using the heritability and genetic correlation among different trait classes (Table 6). As these estimates are based on an average across traits within a trait class, the results should be considered for illustration only. The response, in standard deviation units of a correlated trait, was calculated by multiplying the genetic correlation by the product of the square root of the heritability estimates of both traits and the genetic standard deviation of the correlated trait, where one standard deviation unit of selection differential was applied on the selected trait (Falconer, 1989). The correlated response was compared to direct response in the trait under selection using the same

TABLE 4 Trait class-wise heritability (h^2) estimates pooled across nine shrimp species

Trait class[b]	Stage[a]	*n*	Weighted h^2	Min	Max	Unweighted h^2
WT	Overall	236	0.30 ± 0.002	0.001	0.97	0.32 ± 0.01
WT	Pre-harvest	91	0.26 ± 0.002	0.01	0.97	0.35 ± 0.02
WT	Harvest	94	0.30 ± 0.005	0.001	0.80	0.28 ± 0.01
WT	Mature	49	0.36 ± 0.01	0.02	0.71	0.36 ± 0.02
SUR	Overall	119	0.15 ± 0.003	0.00	0.78	0.13 ± 0.01
SUR	Pre-harvest	63	0.19 ± 0.003	0.001	0.78	0.18 ± 0.02
SUR	Harvest	42	0.07 ± 0.004	0.001	0.42	0.06 ± 0.01
SUR	Mature	7	0.10 ± 0.009	0.008	0.14	0.08 ± 0.02
REP	Overall	30	0.11 ± 0.006	0.001	0.71	0.18 ± 0.03
REP	Harvest	2	0.11 ± 0.035	0.03	0.13	0.08 ± 0.05
REP	Mature	24	0.09 ± 0.006	0.001	0.71	0.17 ± 0.04
MOR	Overall	213	0.26 ± 0.003	0.01	0.97	0.33 ± 0.014
MOR	Pre-harvest	106	0.26 ± 0.005	0.06	0.97	0.39 ± 0.024
MOR	Harvest	93	0.27 ± 0.004	0.01	0.89	0.26 ± 0.013
MOR	Mature	14	0.23 ± 0.02	0.14	0.51	0.25 ± 0.025
FEED	Overall	7	0.58 ± 0.07	0.25	0.70	0.60 ± 0.06
FEED	Pre-harvest	7	0.58 ± 0.07	0.25	0.70	0.60 ± 0.06
COMP	Overall	33	0.06 ± 0.009	0.001	0.29	0.08 ± 0.01
COMP	Harvest	28	0.06 ± 0.009	0.001	0.19	0.07 ± 0.01
COMP	Mature	5	0.06 ± 0.025	0.001	0.29	0.18 ± 0.05
OTHER	Overall	3	0.16 ± 0.03	0.02	0.41	0.24 ± 0.11
OTHER	Pre-harvest	1	0.30 ± 0.28	0.30	0.30	0.30
OTHER	Harvest	2	0.17 ± 0.04	0.02	0.41	0.21 ± 0.19

[a]Stages (pre-harvest = <120 days, harvest = 120–180 days and mature = >180 days), number of observations (*N*), weighted and unweighted heritability estimates with minimum (Min) and maximum (Max) value of range.

[b]The trait class abbreviations are described in Table 2b.

selection intensity. The computed estimates suggested selection on weight traits can indirectly improve morphological traits; however, it can only result in small changes in survival and reproductive traits. Similarly, selection of growth traits can improve feed, reproduction and survival traits indirectly. However, selection on weight traits can negatively impact survival and body composition traits (Table 6).

3.3 | Trait means

There were 409 estimates for Pacific white shrimp, 90 estimates for giant freshwater prawn, 39 for Chinese white shrimp, 17 for kuruma prawn, 20 for black tiger prawn, 11 for banana shrimp, 10 for green tiger prawn and 4 for Indian white prawn across 57 traits. In general, body weight was the most frequently reported trait across the studies. The average body weight information across species suggests that black tiger prawn are, in general, the heaviest at mature stage as compared to the other shrimp species. Among the morphometric traits, body length (BL) is most frequently reported across the shrimp species. The analysis revealed that banana shrimp is the longest with an average body length of 13.9 cm at harvest time period, compared with more popular aquaculture species such as black tiger prawn (9.26 cm) and Pacific white shrimp (7.96 cm). Growth rate was measured in grams of gain per day over a specific time period. The highest growth rate was found for the giant freshwater prawn during the pre-harvest period with an average estimate of 0.11 g/day. Body composition was measured mainly as per cent of fatty acid in the haemolymph and eggs of shrimp. The average unweighted fatty acid content was 9.43%, with a range of 0.16 to 41.03% in Pacific white shrimp. Daily feed intake (DFI) and feed efficiency ratio (FER) were estimated as the measures of feed efficiency. The mean unweighted FER was higher (0.66) in Pacific white shrimp compared with Chinese white shrimp (0.24) in the pre-harvest stage (Dai, Luan, Lu, Luo, Meng, et al., 2017). A wide range of reproductive traits were recorded across studies at the mature stage. A condition factor (*K*) estimate of 1.41 was recorded at the mature stage for Pacific white shrimp. There was one recorded measurement for spawning frequency (this measurement is a count of how frequently an animal spawned over a 55-day period) of 1.70 (Tan, Kong, et al., 2017) for Pacific white shrimp. Spawning success (SS) (per cent of females spawned) also had a single value recorded for Pacific white shrimp at maturity of 76.11% (Tan, Kong, et al., 2017). Egg number

TABLE 5 Genetic and phenotypic correlations[a] among different traits classes[b]

Trait class 1	Trait class 2	*n*	Genetic correlation	Min	Max	*n*	Phenotypic correlation	Min	Max
COMP	COMP	159	0.07 ± 0.04	−0.99	0.99	6	0.01 ± 0.02	−0.11	0.88
COMP	REP	14	0.06 ± 0.11	−0.82	0.99	12	−0.04 ± 0.01	−0.37	0.24
COMP	WT	4	−0.13 ± 0.21	−0.72	0.14	4	−0.13 ± 0.03	−0.19	−0.1
FEED	FEED	11	0.49 ± 0.07	−0.76	0.86	9	0.10 ± 0.01	−0.61	0.70
FEED	GR	6	0.38 ± 0.07	−0.01	0.85	6	0.47 ± 0.01	−0.007	0.81
GR	GR	31	0.45 ± 0.05	−0.26	0.98	26	0.35 ± 0.01	−0.22	0.98
GR	REP	12	0.05 ± 0.11	−0.48	0.63	NA	NA	NA	NA
GR	SUR	9	0.17 ± 0.09	−0.02	0.46	NA	NA	NA	NA
MOR	MOR	217	0.77 ± 0.003	−0.99	0.99	219	0.62 ± 0.002	−0.99	0.99
MOR	REP	8	0.03 ± 0.06	−0.98	0.72	8	0.01 ± 0.004	−0.77	0.30
MOR	SUR	28	0.26 ± 0.01	−0.43	0.99	28	0.08 ± 0.004	−0.17	0.41
MOR	WT	197	0.81 ± 0.002	−0.6	0.99	204	0.68 ± 0.002	−0.11	0.99
REP	REP	24	0.09 ± 0.09	−0.77	0.99	17	0.07 ± 0.01	−0.30	0.85
REP	SUR	NA	NA	NA	NA	2	0.01 ± 0.04	−0.06	0.08
REP	WT	29	0.04 ± 0.05	−0.95	0.93	17	0.02 ± 0.004	−0.65	0.99
SUR	SUR	65	0.62 ± 0.01	−0.99	0.99	25	0.49 ± 0.003	−0.94	0.88
SUR	WT	61	−0.06 ± 0.02	−0.98	0.82	28	0.02 ± 0.005	−0.72	0.62
WT	WT	103	0.64 ± 0.01	−0.39	0.99	47	0.70 ± 0.006	−0.06	0.99

[a]Weighted correlation estimates.

[b]The trait class abbreviations are described inTable 2b

(EN) was reported for the Pacific white shrimp and black tiger prawn species as mean counts of 116.9 and 216, respectively. A total of 82 estimates of survival traits were included in this study, 52 of which were based on measurements taken at the pre-harvest stage, and the remaining 30 based on measurements made in the harvest and mature stage. The species-specific estimates of all trait means are provided in Tables S1–S9 and individual estimates in the ShinyApp (https://khatkar.shinyapps.io/shrimp_genetics/).

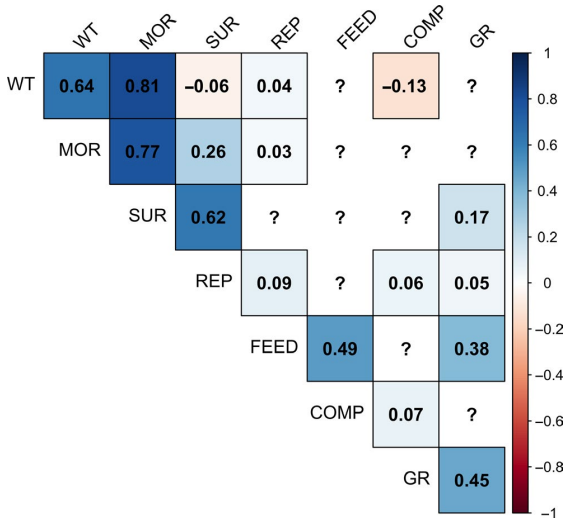


FIGURE 1 Mean genetic correlations among different trait classes pooled across nine shrimp species (missing values are indicated as "?"). The trait class abbreviations are described in Table 2b [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

This meta-analysis presents the first quantitative analysis of genetic parameters derived from selected studies in the nine main shrimp species farmed commercially. The 87 studies in our analysis suggest a substantial global research effort in the genetic analysis of shrimp species. This meta-analysis combined different published estimates of heritability and genetic correlation for economically important traits of farmed shrimp species and shows both the degree of consensus and variability among estimates in the literature. As such, this study provides the most complete consensus view of genetic parameters which are critically important for genetic improvement programmes in these farmed aquatic species.

Pacific white shrimp was the species most represented in this study, which reflects its successful domestication and importance in aquaculture among the commercially farmed shrimp species. A dramatic increase in the global production of this species has been seen in the last two decades, rising from approximately 0.1 million tonnes in 1995 to 3 million tonnes in 2017, especially in South-East Asia (Global Aquaculture Alliance, 2018). The need for ongoing aquaculture development in this species led to the development of

resource populations and increased research effort. Although black tiger prawn is considered another important shrimp species for aquaculture, with only 20 trait means reported, it was poorly represented when compared to Pacific white shrimp with 379 trait means reported. This may be due to the lack of successful domestication of this species and limited availability of resource populations for genetic analysis.

Across all studies, body weight was the most common trait reported on, followed by morphological traits, which may be due to the ease of recording and also being the core focus in shrimp breeding programmes for increased biomass production. Body weight is important as it provides industry with a direct assessment of an animal's performance and potential farm productivity (Boyd, Tucker, McNevin, Bostick, & Clay, 2007). At present, weight is a major determining factor of price paid for stock at the farm gate. Morphometric traits encompass the largest number of individual traits with a large proportion (33.2%) of all heritability estimates. By contrast, reproduction traits, which are of considerable importance to the industry, account for less than 4.6% of all heritability estimates. This can be attributed to the difficulty of measuring this trait accurately, combined with the challenges faced in closing the life cycle of many shrimp species in captivity. Further explanation for the substantially lower representation of reproductive traits in this analysis can be found in the structure of many breeding programmes and facilities. The housing and breeding protocols of many shrimp producers impose inherent difficulties associated in obtaining the exact number of eggs or nauplii spawned by any one dam. Genetic parameter estimates for other traits of economic importance, including meat yield, tail and carapace length, were less frequently reported despite being influential in the profitability of shrimp farming. The increased difficulty of obtaining measurements related to these more complex traits may be the reason that these traits are underrepresented in the meta-analysis. The majority of reproduction traits were measured at the mature stage, whilst the majority of weight, morphometric and growth traits were measured at the harvest stage, and survival was most frequently measured at the pre-harvest stage.

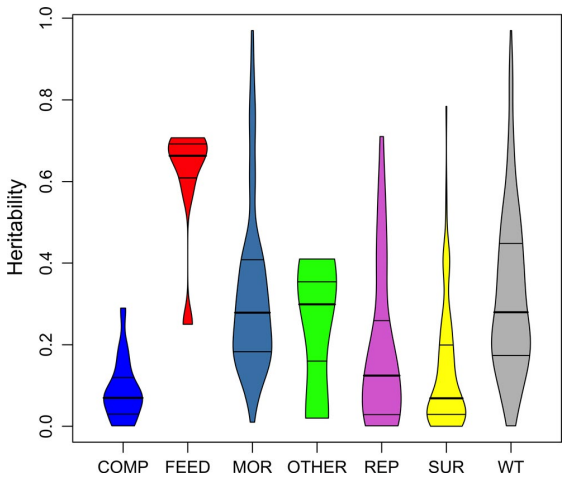


FIGURE 2 Distribution of trait class-specific heritability estimates shown with violin plots. The trait class abbreviations are described in Table 2b [Colour figure can be viewed at wileyonlinelibrary.com]

Across the studies, we observed a faster growth rate for giant freshwater prawn at the early life stage. This suggests that this species might be preferable to culture where multiple/ crops are targeted in a short period of time and/or in areas with restricted growing season, for example in the temperate regions. On the other hand, the highest average body weight, body length and growth rate till harvest stage were found in Pacific white shrimp, suggesting the potential for this species to provide better yields compared with the other species. This might be the reason why this species is considered an excellent aquaculture species. However, black tiger prawn are heaviest when they are harvested in the mature stage (Appendix S2, Table S2), suggesting an extended culture period is needed to gain higher weight from this species.

4.1 | Heritability estimates

This meta-analysis revealed that economically important traits, such as body weight, length and growth among selected shrimp species, are moderately to highly heritable. We would consider the estimates obtained in our meta-analysis to be the norm since estimates of heritability for these traits are consistently high across a wide range of species, including fish (Bolivar & Newkirk, 2002), beef cattle (Bourdon & Brinks, 1982) and human (Yang et al., 2010). Meta-analysis studies with cattle, sheep, pig and goat also revealed moderately higher heritability estimates of growth traits (Akanno, Schenkel, Quinton, Friendship, & Robinson, 2013; Jembere et al., 2017; Safari, Fogarty, & Gilmour, 2005). Most of the aquaculture species are raised in the diverse water environments characterized by highly variable aquatic parameters among different ponds or production units, for example variation in temperature, salinity, pH and level of dissolved oxygen, which might greatly affect overall animal production (Mengistu, Mulder, Benzie, & Komen, 2019). Moreover, due to ease in transportation of fertilized eggs and larvae, aquaculture species are disseminated across many locations and experience great variability in rearing conditions. These variable environmental factors can critically impact on heritability estimates of economically important traits (e.g. growth) of aquaculture species. However, despite the potential impact of environmental variation, the observed medium-to-high heritability of the growth traits suggests that there are sufficient additive gene actions regulating variation of these traits. This in turn offers great potential for selection and genetic improvement of these traits as is discussed further below.

Heritability estimates for growth-related traits also varied greatly among shrimp species. For banana shrimp and Pacific white shrimp, body weight and body length were highly heritable whereas growth rate was highly heritable in black tiger prawn, suggesting each species may have unique attributes that can be exploited for selective breeding (Appendix S2, Tables S1–S9).

TABLE 6 Expected correlated response among various trait classes[a]

Trait class 1	Trait class 2	Heritability trait 1 (h_1^2)	Heritability trait 2 (h_2^2)	Genetic correlation traits 1 and 2 (r_g)	Correlated response trait 2 (CR)[b]	Comparative performance of indirect selection[c]
MOR	WT	0.26	0.30	0.81	0.22	75.40%
WT	MOR	0.30	0.26	0.81	0.22	87.01%
MOR	SUR	0.26	0.16	0.26	0.05	33.14%
MOR	REP	0.26	0.11	0.03	0.005	4.61%
WT	SUR	0.30	0.16	−0.06	−0.01	−8.21%
WT	COMP	0.30	0.06	−0.13	−0.01	−29.07
COMP	REP	0.06	0.11	0.06	0.005	4.43%
FEED	GR	0.58	0.27	0.38	0.15	55.69%
GR	FEED	0.27	0.58	0.38	0.15	25.92%
GR	REP	0.27	0.11	0.05	0.009	7.83%
GR	SUR	0.27	0.16	0.17	0.03	22.08%
REP	WT	0.11	0.30	0.04	0.007	2.42%

[a]The trait class abbreviations are described in Table 2b.

[b]Correlated response in trait 2 with 1 *SD* selection differential in trait 1, that is $CR_2 = r_g \times h_1 \times h_2 \times SD_2$ where SD_2 is the *SD* of trait 2.

[c]Correlated response in trait 2 as a percentage of gain possible from direct selection for trait 2 $\frac{CR_2}{SD_2} \times 100 = (r_g \times h_1/h_2) \times 100$.

The heritability estimate of survival traits was low with the exception of repeated studies in green tiger prawn where the heritability for survival was moderate at 0.31. Although survival traits depicted a low heritability, they do express genetic variation which suggests genetic improvement of this trait is feasible through selective breeding (Argue, Arce, Lotz, & Moss, 2002; Campos-Montes, Montaldo, Martínez-Ortega, Jiménez, & Castillo-Juárez, 2013; Zhang, Cao, Liu, Yuan, & Hu, 2017). However, heritability of survival trait class has to be interpreted with caution and this should be considered as overall estimate. This is because survival traits are highly heterogeneous, and the meta-analysis of survival trait class consists of data of survival from a wide range of environmental conditions with various biotic (e.g. pathogenic strains) and abiotic (e.g. salinity, ammonia and temperature tolerance) conditions. This might be reflected through the wide range of heritability estimates of survival traits. Heritability estimates for specific environmental conditions were in the following ranges: commercial pond (0.00 to 0.43), tanks (0.36 to 0.72), population density (0.22 to 0.61), hypoxia tolerance (0.07 to 0.15), ammonia tolerance (0.13 to 0.17), salinity tolerance (0.21 to 0.78), cold tolerance (0.019 to 0.26), WSSV resistance (0.00 to 0.17), HPV (0.41), AHPND (0.24 to 0.43) and TSV (0.07 to 0.41). The details of individual estimates are provided in the ShinyApp (https://khatkar.shinyapps.io/shrimp_genetics/).

Survival, which is a fitness trait, is assumed to show lower heritability due to strong selection which ultimately can cause a reduction in genetic variation. Moreover, increased environmental variability can result in low estimates of heritability even in the presence of significant genetic variation (Merilä & Sheldon, 1999). In spite of the importance of survival and reproductive traits, the implementation of selection for these traits in shrimp breeding programmes is limited (Ibarra, Pérez-Rostro, Ramírez, & Ortega-Estrada, 2007; Ibarra, Racotta, Arcos, & Palacios, 2007). Despite the limitations and difficulty in measurement and low heritability, their high economic importance warrants further investigation on how they can best be incorporated in selective breeding programmes. In particular, the use of indirect or genomic selection technologies may provide an attractive avenue as discussed below.

Another interesting observation from our meta-analysis was that heritability estimates for all traits showed life stage-specific variation and appear to be strongly age-dependent, with higher additive genetic variance observed for harvest-size shrimp. Since response to selection could vary due to animals' life stages (Charmantier, Perrins, McCleery, & Sheldon, 2005), age of selection could be an important factor in the design of breeding programmes.

4.2 | Genetic and phenotypic correlation among traits

In order to optimize and direct-controlled breeding programmes, it is essential to understand the phenotypic, genetic and environmental correlations among important traits. A limited number of studies report genetic and phenotypic correlations for paired trait combinations in shrimp species and highlight the need for additional estimates to be reported on in shrimp species. Morphological traits have more reports on correlation estimates followed by correlations within weight, survival, growth and reproductive trait classes. In general, genetic correlations were high between weight and morphological traits. The genetic correlation between morphological and body weight traits has been reported to be very high or close to unity for several other aquatic species. For example, the genetic correlation between body weight and carapace length of black tiger prawn was 1 (Kristjánsson & Arnason, 2016; Sun et al., 2015). This suggests that these correlated traits can be improved by selecting for body

weight. Similar to body weight traits, strong genetic and phenotypic correlations among morphological traits were noted. This suggests that these traits are regulated by the same set of genes and have a high degree of autocorrelation because of the nature of physical body measurements. Therefore, selection of one morphological trait will lead to a change in other correlated morphological traits. In this perspective, body condition factor (e.g. length-to-weight ratio) can be considered to avoid selection of extreme body types, that is long skinny versus short fat shrimp (Li, Zhou, et al., 2016).

The survival traits had low genetic and phenotypic correlations with growth and morphological traits, whilst with body weight it was close to zero. However, genetic correlation between survival and growth traits has been found to be both negative and positive in shrimp (Argue et al., 2002; Gitterle, Salte, et al., 2005; Li et al., 2015; Moss, Moss, & Lotz, 2013; Zhang, Cao, Liu, & Yuan, 2017) and in a range of other aquaculture species (Charo-Karisa, Rezk, Bovenhuis, & Komen, 2005; Perry, Martyniuk, Ferguson, & Danzmann, 2005; Zhang et al., 2014). The correlation between survival and growth traits also varies greatly across growth phase, species, experimental conditions and pathogens (Zhang et al., 2017). This suggests correlation between survival and growth traits is species- and life stage-specific, and thus, it should be interpreted with care. Therefore, shrimp genetic improvement programmes should consider the correlation between growth and survival traits on an individual case basis to design optimal breeding schemes.

Overall, our analysis of genetic correlations among different traits suggests that the selection for higher body weight has mostly positive or only a marginal negative impact on other traits. A multi-trait selection index composed of the key economic traits will be the most efficient means to bring improvement in overall profitability.

4.3 | Non-additive genetic variation in shrimp breeding programmes

Selection is generally carried out utilizing the existence of additive genetic variance. However, non-additive genetic effects (e.g. dominance and epistasis) can have important roles in improving production efficiency in breeding programmes. Cross-breeding can bring complementary additive effects by crossing superior lines and can exploit non-additive effects (dominance and epistasis), thus generating heterosis (Maluwa & Gjerde, 2006). To date, limited studies have reported the benefits of cross-breeding to improve the yield from the shrimp breeding programmes. Lu, Luan, Cao, Sui, et al. (2017) documented a higher survival rate of line-crossed Pacific white shrimp compared with the control population. Lu et al. (2016) assessed the growth performance of hybrid Pacific white shrimp by crossing several strains and revealed that 75% of hybrids have positive heterosis for harvest body weight, suggesting the possibility of dominant gene action controlling this trait. It is expected that higher production efficiency could be obtained by combining cross-breeding and selective breeding whilst negating the negative effects of inbreeding.

4.4 | The social interaction and indirect genetic effects in shrimp

Genetic effects due to social interaction (e.g. mutualism, antagonism, aggression, dominance and population density) have been widely studied in animals and plants. The social interaction impact is also evident in shrimp species. In particular, aggression and cannibalism are very common in shrimp (Chak, Bauer, & Thiel, 2015). Furthermore, these behaviours become more evident when stocking density is high and feeding frequency is low. Cannibalism has been reported more frequently during moulting periods in shrimp (Karplus, 2005). Luan et al. (2015) reported that social interaction accounted a substantial amount of heritable variation in growth traits in Pacific white shrimp. Dinh and Nguyen (2014) found notable differences in additive genetic variance of body weight between morphotypes of giant freshwater shrimp. Hung, Nguyen, Hurwood, and Mather (2014) also reported morphotypes, and reproductive status of giant freshwater prawn had significant effects on estimates of genetic parameters of body weight. Ellen et al. (2014) reported a low-to-moderate heritability of between 0.1 and 0.4 for social interaction traits in farmed animals, suggesting possibilities of selecting these traits for better growth and survival. Studies on the assessment of social effects on economically important traits in shrimp species are limited. Future shrimp breeding programmes should consider the presence of indirect social genetic effects on the heritable variation of target traits.

4.5 | Common environmental effects

In animal breeding, phenotypic records may be similar for environmental reasons. This occurs when the related animals share the same environment, for example when separate families are raised in separate tanks. If the common environmental effects are omitted from the statistical models, then inflated heritability estimates can be obtained (Appendix S2, Figure S1) (Rutten, Komen, & Bovenhuis, 2005). Across the literature, the common environment effect ranged from 0.00 to 0.22, suggesting it may account for a substantial amount of the phenotypic variation in shrimp. However, we observed that only 29% of studies reported common environment effects. Of these, only nine studies reported that common environment effect was significant. Due to this low number reporting common environmental effects, we were unable to include this information in our meta-analysis. In most experiments, rearing families in separate tanks made it difficult to separate common environment and genetic effects (Lu, Luan, Cao, Sui, et al., 2017; Sui, Luan, Luo, Meng, Lu, et al., 2016). Overall, there is evidence that economically important traits in shrimp are affected by common environmental effects, and therefore, it is essential to design breeding experiments where these effects can be accounted in the statistical models to avoid obtaining inflated heritability estimates (Montaldo, Castillo-Juárez, Campos-Montes, & Pérez-Enciso, 2013). Moreover, common environmental effect can be significantly reduced by maintaining

standardized hatchery and management practices (Trinh, Nguyen, Nguyen, Wayne, & Nguyen, 2019). A number of studies have been conducted in the presence of genotype-by-environment interaction (G × E) in shrimp; however, these were not related to the impact of G × E on the genetic parameters. Therefore, future studies should be directed to quantify the impact of G × E on the estimates of genetic parameters.

4.6 | Industry implications

From the meta-analyses, it is clear that growth traits are moderately to highly heritable and can be selected for in a breeding improvement programme. This is also supported by a number of selection experiments. Hetzel, Crocos, Davis, Moore, and Preston (2000) reported the response to selection for growth in kuruma prawn. The average response to selection was 10.7% in a single generation. Andriantahina, Liu, Huang, and Xiang (2012) also reported a 10.7% response to selection for a growth trait in first generation in Pacific white shrimp. Within the same species, Preston, Crocos, Keys, Coman, and Koenig (2004) showed that, after four generations of selection, selected shrimp were 9 to 14% heavier than the non-selected population. Sui, Luan, Luo, Meng, Cao, et al. (2016) evaluated the response to selection of harvest body weight of the Chinese white shrimp. After five generations of multi-trait selection, the accumulated realized genetic gain was 18.6%. Hung, Nguyen, Ponzoni, Hurwood, and Mather (2013) reported an average selection response in body weight of 7.4% per generation in a selected giant freshwater prawn population. Sui et al. (2019) and Luan et al. (2012) also reported 6.6% and 14% realized response, respectively, per generation for body weight of giant freshwater prawn. Goyard et al. (2002) carried out a selection breeding experiment for higher growth rate of western blue prawn and observed that after the five generations of selection, shrimp growth rate was 21% higher than the non-selected population (Goyard et al., 2002). Overall, the results from these studies suggest that significant genetic improvement of growth traits can be achieved following only a few generations of selection. Although some of the estimates of response to selection were calculated from only one cycle of selection, this provides useful information to predict responses up to a middle term, for example up to five generations (Falconer, 1989). The response to selection for survival traits has also been reported. In a multi-trait selection breeding programme of Pacific white shrimp, Gitterle et al. (2007) reported an average selection response per generation of 5.7% for survival and 1.7% for traits related to WSSV resistance. After five generations of selection, an accumulated genetic change of 8.46% was obtained for survival trait of giant freshwater prawn. The average genetic gain per generation was low, with a value of 1.69% (Luan et al., 2012). Vu, Trong, and Nguyen (2017) reported an 18% higher survival rate of a selected line compared with the unselected wild stock of giant freshwater prawn species. A selective breeding experiment of Pacific white shrimp for growth and resistance to Taura syndrome virus revealed after one generation the selected line grew 21% faster and had an 18.4% higher survival than the control group (Argue et al., 2002). In summary, the results from these selection experiments are broadly consistent with the expectation from genetic parameters and suggested that significant improvement can be made in shrimp body size, survival and disease resistance traits through selective breeding.

Analysis of comparative performance of indirect selection suggested that selection for weight can effectively change morphological traits with an efficiency of 87% (Table 6), whereas selection for morphological traits can be 75% efficient in changing weight indirectly. Moreover, selection of morphological traits can improve survival traits indirectly with an efficiency of 33% and would be much easier to measure. However, selection for increased body weight has shown to have a marginal positive impact on reproductive traits. Direct selection on growth traits can increase feed traits (e.g. feed conversion ration; FCR) by 26% indirectly (Table 6). Overall, these examples suggest that morphological, weight, growth, feeding and survival traits can be simultaneously selected and index selection can be effective in shrimp selection breeding programmes.

4.7 | Future directions

Large-scale phenotyping is critical for obtaining reliable estimates of genetic parameters and implementing efficient selective breeding programmes. This is becoming more important for the application of genomic selection where the accuracy of models is established on the basis of phenotyping large training populations. The manual recording of phenotypes is error-prone and labour-intensive, and induces stress to the animals in culture. High-throughput phenotyping can enhance genetic gain within a breeding programme. Studies with wheat breeding have shown that high-throughput phenomics can be successfully applied to collect data from thousands of plots to uncover the genetic factors and increase the rate of gain for complex traits (Singh et al., 2019). Recently, complex phenotypes have been recorded and analysed using machine learning-based high-throughput platform for a number of aquaculture species, for example tilapia (Zhao et al., 2017) and salmon (Pinkiewicz, Purser, & Williams, 2011). However, to date high-throughput phenotyping has not been employed for genetic parameter estimation and selection breeding programmes of aquaculture species. Recent advances in phenomics, which combine sensor and imaging technologies with artificial intelligence, can help to develop systems to accurately record novel high-throughput phenotypes for incorporation into aquaculture breeding programmes (Zenger et al., 2018).

Access to accurate estimates of genetic parameters, in particular heritability and correlation estimates among groups of economically important production traits, is a prerequisite for building successful breeding programmes. Ideally, these parameter estimates are obtained under environments in which shrimp are selected and produced. The analysis presented in this study will

serve as the baseline reference for genetic parameter estimates for shrimp species which are commercially farmed. However, our analysis was greatly hindered by limited availability of comprehensive information on the background details of the genetic parameter studies. Therefore, we recommend that the future studies should report the following information, at a minimum (a) breeding design, (b) origin of the population, (c) number of samples and families used, (d) standardized definition of traits and phenotyping protocols, (e) statistical methods used for estimating genetic parameters, (f) standard error(s), and (g) estimates of additive and residual variance–covariance components.

Most of the genetic parameter estimates are reported from the resource population raised in hatcheries or in controlled environments. The estimates of genetic parameters and selection based on data from commercial ponds are required to ensure that the selective breeding programmes are evaluated in the target environment. This can be achieved by using genomic information. Selection in breeding programmes in shrimp species is mostly based on pedigree information, but see Khatkar et al. (2017); Wang, Yu, Li, Zhang, and Xiang (2017); Wang, Yu, Yuan, et al. (2017); Wang et al. (2019), for studies using genomic information. Availability of low-cost and high-throughput SNP panels and incorporating whole-genome selection will be of utmost importance for implementing efficient breeding programmes (Dekkers, 2012; Zenger et al., 2018).

5 | CONCLUSION

Overall, the moderate-to-high heritability and genetic correlation estimates for body size and growth rate recorded in this meta-analysis suggest that there is substantial genetic variability for economically important traits in all shrimp species studied to date, with realistic potential for genetic improvement of these traits. Although low heritability estimates were reported for reproduction and survival traits, there is sufficient underlying genetic variation to suggest that selecting for improved survival and reproduction is possible. The genetic correlations among morphological, weight, growth, feeding and survival traits suggest these traits can be simultaneously improved using index selection. This meta-analysis identified that there is a lack of studies of several economically important traits in shrimp species, namely disease resistance and feed efficiency. Currently, most of the data available are from Pacific white shrimp: in future, additional data from other shrimp species are required to gain a precise understanding of genetic parameter estimates. The derived estimates reported in this meta-analysis along with online ShinyApp will be an important key reference point for designing future genomic selection programmes in shrimp species.

ACKNOWLEDGEMENTS

The study was conducted under the research project of the Australian Research Council Industrial Transformation Research Program IH130200013, funded by the Australian Research Council (ARC) Industrial Transformation Research Hub scheme, awarded to James Cook University and in collaboration with the Commonwealth Scientific Industrial Research Organisation (CSIRO), the Australian Genome Research Facility (AGRF), the University of Sydney and Seafarms Group Pty Ltd. MMH is a recipient of the Research Training Program Stipend (RTP) International Scholarship from the University of Sydney and Francis Henry Loxton Supplementary Scholarship from the Sydney Institute of Agriculture, University of Sydney.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

MK, MH, RT, PT and HR conceptualized and designed the study. MH, RT and MK acquired the data. MK, MH, PT and HR conducted the data analysis and interpretation of results. MH, MK, HR and PT wrote the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The data analysed in this manuscript are from published research articles from the public domain. Table 1 contains information on the articles. Specific information of each study can be found in the ShinyApp (https://khatkar.shinyapps.io/shrimp_genetics/). The R-scripts used in the meta-analysis can be obtained from the authors on request.

ORCID

Md. Mehedi Hasan https://orcid.org/0000-0002-7838-2315
Rachel L. Tulloch https://orcid.org/0000-0002-8358-3066
Peter C. Thomson https://orcid.org/0000-0003-4428-444X
Herman W. Raadsma https://orcid.org/0000-0002-4753-1824
Mehar S. Khatkar https://orcid.org/0000-0002-4554-1404

REFERENCES

Akanno, E., Schenkel, F., Quinton, V., Friendship, R., & Robinson, J. (2013). Meta-analysis of genetic parameter estimates for reproduction, growth and carcass traits of pigs in the tropics. *Livestock Science*, *152*(2–3), 101–113. https://doi.org/10.1016/j.livsci.2012.07.021

Anderson, J. L., Valderrama, D., & Jory, D. (2018). Global shrimp production review and forecast: Steady growth ahead. *Global Aquaculture Alliance*, *43*. https://www.aquaculturealliance.org/advocate/global-shrimp-production-review-and-forecast-steady-growth-ahead/. Accessed May 9, 2020.

Andriantahina, F., Liu, X., Huang, H., & Xiang, J. (2012). ZResponse to selection, heritability and genetic correlations between body weight and body size in Pacific white shrimp, *Litopenaeus vannamei*. *Chinese Journal of Oceanology and Limnology*, *30*(2), 200–205. https://doi.org/10.1007/s00343-012-1066-2

Andriantahina, F., Liu, X., Huang, H., & Xiang, J. (2013). Selection for growth performance of tank-reared Pacific white shrimp, *Litopenaeus vannamei*. *Chinese Journal of Oceanology and Limnology*, *31*(3), 534–541. https://doi.org/10.1007/s00343-013-2220-1

Araneda, M. E., Hernández, J. M., & Gasca-Leyva, E. (2011). Optimal harvesting time of farmed aquatic populations with nonlinear size-heterogeneous growth. *Natural Resource Modeling*, *24*(4), 477–513. https://doi.org/10.1111/j.1939-7445.2011.00099.x

Arcos, F. G., Racotta, I. S., & Ibarra, A. M. (2004). Genetic parameter estimates for reproductive traits and egg composition in Pacific white shrimp Penaeus (Litopenaeus) vannamei. *Aquaculture*, *236*(1–4), 151–165. https://doi.org/10.1016/j.aquaculture.2004.03.003

Arcos, F. G., Racotta, I. S., Palacios, E., & Ibarra, A. M. (2005). Ovary development at the onset of gametogenesis is genetically determined and correlated with reproductive traits at maturity in shrimp Litopenaeus (Penaeus) vannamei. *Marine Biology*, *148*(2), 339–346. https://doi.org/10.1007/s00227-005-0087-1

Argue, B. J., Arce, S. M., Lotz, J. M., & Moss, S. M. (2002). Selective breeding of Pacific white shrimp (*Litopenaeus vannamei*) for growth and resistance to Taura Syndrome Virus. *Aquaculture*, *204*(3–4), 447–460. https://doi.org/10.1016/S0044-8486(01)00830-4

Benzie, J., Kenway, M., & Trott, L. (1997). Estimates for the heritability of size in juvenile Penaeus monodon prawns from half-sib matings. *Aquaculture*, *152*(1–4), 49–53. https://doi.org/10.1016/S0044-8486(96)01528-1

Bolivar, R. B., & Newkirk, G. F. (2002). Response to within family selection for body weight in Nile tilapia (Oreochromisniloticus) using a single-trait animal model. *Aquaculture*, *204*(3–4), 371–381. https://doi.org/10.1016/S0044-8486(01)00824-9

Bourdon, R., & Brinks, J. (1982). Genetic, environmental and phenotypic relationships among gestation length, birth weight, growth traits and age at first calving in beef cattle. *Journal of Animal Science*, *55*(3), 543–553. https://doi.org/10.2527/jas1982.553543x

Boyd, C. E., Tucker, C., McNevin, A., Bostick, K., & Clay, J. (2007). Indicators of resource use efficiency and environmental performance in fish and crustacean aquaculture. *Reviews in Fisheries Science*, *15*(4), 327–360. https://doi.org/10.1080/10641260701624177

Caballero-Zamora, A., Cienfuegos-Rivas, E. G., Montaldo, H. H., Campos-Montes, G. R., Martínez-Ortega, A., & Castillo-Juárez, H. (2015). Genetic parameters for spawning and growth traits in the Pacific white shrimp (Penaeus (Litopenaeus) vannamei). *Aquaculture Research*, *46*(4), 833–839. https://doi.org/10.1111/are.12235

Caballero-Zamora, A., Montaldo, H., Campos-Montes, G., Cienfuegos-Rivas, E., Martínez-Ortega, A., & Castillo-Juárez, H. (2014). *Estimation of (co) variance components for body weight and survival in the presence of a White Spot Syndrome Virus (WSSV) natural outbreak in the Pacific White Shrimp Penaeus (Litopenaeus) vannamei*. Paper presented at the 10th World Congress of Genetics Applied to Livestock Production.

Caballero-Zamora, A., Montaldo, H. H., Campos-Montes, G. R., Cienfuegos-Rivas, E. G., Martínez-Ortega, A., & Castillo-Juárez, H. (2015). Genetic parameters for body weight and survival in the Pacific White Shrimp Penaeus (Litopenaeus) vannamei affected by a White Spot Syndrome Virus (WSSV) natural outbreak. *Aquaculture*, *447*, 102–107. https://doi.org/10.1016/j.aquaculture.2014.08.028

Cala, N., Montaldo, H., Campos-Montes, G., & Castillo-Juárez, H. (2014). *Evolution of genetic variability for growth-out survival rate in a selected population of Pacific white shrimp Penaeus (Litopenaeus) vannamei*. Paper presented at the 10th World Congress of Genetics Applied to Livestock Production.

Campos, G., Montaldo, H., Arechavaleta, M., & Castillo-Juárez, H. (2006). *Heritability of body length at 25 days of age in the Pacific white shrimp (Litopenaeus vannamei)*. Paper presented at the Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 13–18 August, 2006.

Campos-Montes, G. R., Montaldo, H. H., Armenta-Córdova, M., Martínez-Ortega, A., Caballero-Zamora, A., & Castillo-Juárez, H. (2017). Incorporation of tail weight and tail percentage at harvest size in selection programs for the Pacific white shrimp Penaeus (Litopenaeus) vannamei. *Aquaculture*, *468*, 293–296. https://doi.org/10.1016/j.aquaculture.2016.10.034

Campos-Montes, G. R., Montaldo, H. H., Martínez-Ortega, A., & Castillo-Juárez, H. (2009). Genotype by environment interaction effects for body weight at 130 days of age in the Pacific white shrimp [Penaeus (Litopenaeus) vannamei]. *Veterinaria México*, *40*(3), 255–268.

Campos-Montes, G. R., Montaldo, H. H., Martínez-Ortega, A., Jiménez, A. M., & Castillo-Juárez, H. (2013). Genetic parameters for growth and survival traits in Pacific white shrimp Penaeus (Litopenaeus) vannamei from a nucleus population undergoing a two-stage selection program. *Aquaculture International*, *21*(2), 299–310. https://doi.org/10.1007/s10499-012-9553-1

Castillo-Juárez, H., Casares, J. C. Q., Campos-Montes, G., Villela, C. C., Ortega, A. M., & Montaldo, H. H. (2007). Heritability for body weight at harvest size in the Pacific white shrimp, Penaeus (Litopenaeus) vannamei, from a multi-environment experiment using univariate and multivariate animal models. *Aquaculture*, *273*(1), 42–49. https://doi.org/10.1016/j.aquaculture.2007.09.023

Chak, S. T., Bauer, R., & Thiel, M. (2015). *Social behaviour and recognition in decapod shrimps, with emphasis on the Caridea*. *In* Social Recognition in Invertebrates (pp. 57–84). Springer.

Charmantier, A., Perrins, C., McCleery, R. H., & Sheldon, B. C. (2005). Age-dependent genetic variance in a life-history trait in the mute swan. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1583), 225–232. https://doi.org/10.1098/rspb.2005.3294

Charo-Karisa, H., Rezk, M. A., Bovenhuis, H., & Komen, H. (2005). Heritability of cold tolerance in Nile tilapia, Oreochromis niloticus, juveniles. *Aquaculture*, *249*(1–4), 115–123. https://doi.org/10.1016/j.aquaculture.2005.04.029

Coman, G. J., Arnold, S. J., Wood, A. T., & Kube, P. D. (2010). Age: Age genetic correlations for weight of Penaeus monodon reared in broodstock tank systems. *Aquaculture*, *307*(1–2), 1–5. https://doi.org/10.1016/j.aquaculture.2010.06.027

Dai, P., Kong, J., Meng, X., Luo, K., Lu, X., Chen, B., … Luan, S. (2019). Genotype by environment interaction for feed efficiency trait of the juvenile Pacific white shrimp *Litopenaeus vannamei* held in individuals vs. in groups. *Aquaculture*, *500*, 506–513. https://doi.org/10.1016/j.aquaculture.2018.10.063

Dai, P., Luan, S., Lu, X., Luo, K., Cao, B., Meng, X., & Kong, J. (2017). Genetic evaluation of feed efficiency in the breeding population of Fenneropenaeus chinensis "Huanghai No. 2" using phenotypic, pedigree and genomic information. *Aquaculture International*, *25*(6), 2189–2200. https://doi.org/10.1007/s10499-017-0182-6

Dai, P., Luan, S., Lu, X., Luo, K., Meng, X., Cao, B., & Kong, J. (2017). Genetic assessment of residual feed intake as a feed efficiency trait in the Pacific white shrimp *Litopenaeus vannamei*. *Genetics Selection Evolution*, *49*(1), 61. https://doi.org/10.1186/s12711-017-0334-1

De Donato, M., Ramirez, R., Howell, C., Verginelli, R., Orta, T., Cabrera, S., … Manrique, R. (2008). Artificial family selection based on growth rate in cultivated lines of *Litopenaeus vannamei* (Decapoda, Penaeidae) from Venezuela. *Genetics and Molecular Biology*, *31*(4), 850–856. https://doi.org/10.1590/S1415-47572008005000006

de los Ríos-Pérez, L., Campos-Montes, G. R., Martínez-Ortega, A., Castillo-Juárez, H., & Montaldo, H. H. (2015). Inbreeding effects on body weight at harvest size and grow-out survival rate in a genetic selected population of pacific white shrimp penaeus (Litopenaeus) vannamei. *Journal of the World Aquaculture Society*, *46*(1), 53–60. https://doi.org/10.1111/jwas.12169

Dekkers, J. C. M. (2012). Application of genomics tools to animal breeding. *Current Genomics*, *13*(3), 207–212. https://doi.org/10.2174/138920212800543057

Dinh, H., & Nguyen, N. H. (2014). Genetic inheritance of female and male morphotypes in giant freshwater prawn Macrobrachium rosenbergii. *PLoS One*, *9*(2), e90142. https://doi.org/10.1371/journal.pone.0090142

Ellen, E. D., Rodenburg, T. B., Albers, G. A., Bolhuis, J. E., Camerlink, I., Duijvesteijn, N., … Reimert, I. (2014). The prospects of selection for social genetic effects to improve welfare and productivity

in livestock. *Frontiers in Genetics*, *5*, 377. https://doi.org/10.1111/jwas.12169

Falconer, D. S. (1989). *Introduction to quantitative genetics*. Edinburgh, London, UK: Oliver And Boyd.

FAO (2018). *FAO. The state of world fisheries and aquaculture (SOFIA)* (p. 23). Rome, Italy: Food and Agriculture Organization. Retrieved from http://www.fao.org/3/i9540en/i9540en.pdf

Gitterle, T., Gjerde, B., Cock, J., Salazar, M., Rye, M., Vidal, O., … Salte, R. (2006). Optimization of experimental infection protocols for the estimation of genetic parameters of resistance to White Spot Syndrome Virus (WSSV) in Penaeus (Litopenaeus) vannamei. *Aquaculture*, *261*(2), 501–509. https://doi.org/10.1016/j.aquaculture.2006.07.017

Gitterle, T., Johansen, H., Erazo, C., Lozano, C., Cock, J., Salazar, M., & Rye, M. (2007). Response to multi-trait selection for harvest weight, overall survival, and resistance to white spot syndrome virus (WSSV) in Penaeus (Litopenaeus) vannamei. *Aquaculture*, *272*, S262. https://doi.org/10.1016/j.aquaculture.2007.07.073

Gitterle, T., Ødegård, J., Gjerde, B., Rye, M., & Salte, R. (2006). Genetic parameters and accuracy of selection for resistance to White Spot Syndrome Virus (WSSV) in Penaeus (Litopenaeus) vannamei using different statistical models. *Aquaculture*, *251*(2–4), 210–218. https://doi.org/10.1016/j.aquaculture.2005.06.008

Gitterle, T., Rye, M., Salte, R., Cock, J., Johansen, H., Lozano, C., … Gjerde, B. (2005). Genetic (co) variation in harvest body weight and survival in Penaeus (Litopenaeus) vannamei under standard commercial conditions. *Aquaculture*, *243*(1–4), 83–92. https://doi.org/10.1016/j.aquaculture.2004.10.015

Gitterle, T., Salte, R., Gjerde, B., Cock, J., Johansen, H., Salazar, M., … Rye, M. (2005). Genetic (co) variation in resistance to White Spot Syndrome Virus (WSSV) and harvest weight in Penaeus (Litopenaeus) vannamei. *Aquaculture*, *246*(1–4), 139–149. https://doi.org/10.1016/j.aquaculture.2005.02.011

Gjedrem, T. (2012). Genetic improvement for the development of efficient global aquaculture: A personal opinion review. *Aquaculture*, *344*, 12–22. https://doi.org/10.1016/j.aquaculture.2012.03.003

Gjedrem, T., Robinson, N., & Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture*, *350*, 117–129. https://doi.org/10.1016/j.aquaculture.2012.04.008

Global Aquaculture Alliance (2018). *Shrimp production review – Global aquaculture alliance*. Retrieved from https://www.aquaculturealliance.org/wp-content/uploads/2017/06/Day1_JimAnderson.pdf

Gopikrishna, G., Gopal, C., Shekhar, M., Kumar, K. V., Kannappan, S., & Ponniah, A. (2012). Growth performance and white spot syndrome virus resistance in families of kuruma shrimp (*Marsupenaeus japonicus*). *Indian Journal of Geo-Marine Sciences*, *41*(1), 56–60.

Goyard, E., Patrois, J., Peignon, J., Vanaa, V., Dufour, R., Viallon, J., & Bedier, E. (2002). Selection for better growth of Penaeus stylirostris in Tahiti and New Caledonia. *Aquaculture*, *204*(3–4), 461–468. https://doi.org/10.1016/S0044-8486(01)00831-6

Hayes, B. J., Gitterle, T., Gopikrishna, G., Gopal, C., Krishna, G., Jahageerdar, S., … Rye, M. (2010). Limited evidence for genetic variation for resistance to the white spot syndrome virus in Indian populations of Penaeus monodon. *Aquaculture Research*, *41*(11), e872–e877. https://doi.org/10.1111/j.1365-2109.2010.02611.x

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Henchion, M., Hayes, M., Mullen, A., Fenelon, M., & Tiwari, B. (2017). Future protein supply and demand: Strategies and factors influencing a sustainable equilibrium. *Foods*, *6*(7), 53. https://doi.org/10.3390/foods6070053

Hernández-Ruíz, H., Montaldo, H. H., Bustos-Martínez, J., Campos-Montes, G. R., & Castillo-Juárez, H. (2019). Heritability and genetic correlations for infectious hypodermal and hematopoietic necrosis virus load, body weight at harvest, and survival rate in Pacific white shrimp (*Litopenaeus vannamei*). *Journal of the World Aquaculture Society*, *51*(1), 312–323. https://doi.org/10.1111/jwas.12664

Hetzel, D. J., Crocos, P. J., Davis, G. P., Moore, S. S., & Preston, N. C. (2000). Response to selection and heritability for growth in the Kuruma prawn, *Penaeus japonicus*. *Aquaculture*, *181*(3–4), 215–223. https://doi.org/10.1016/S0044-8486(99)00237-9

Hung, D., Nguyen, N. H., Hurwood, D. A., & Mather, P. B. (2014). Quantitative genetic parameters for body traits at different ages in a cultured stock of giant freshwater prawn (Macrobrachium rosenbergii) selected for fast growth. *Marine and Freshwater Research*, *65*(3), 198–205. https://doi.org/10.1071/MF13111

Hung, D., Nguyen, N. H., Ponzoni, R. W., Hurwood, D. A., & Mather, P. B. (2013). Quantitative genetic parameter estimates for body and carcass traits in a cultured stock of giant freshwater prawn (Macrobrachium rosenbergii) selected for harvest weight in Vietnam. *Aquaculture*, *404*, 122–129. https://doi.org/10.1016/j.aquaculture.2013.04.027

Ibarra, A. M., Arcos, F. G., Famula, T. R., Palacios, E., & Racotta, I. S. (2005). Heritability of the categorical trait 'number of spawns' in Pacific white female shrimp Penaeus (Litopenaeus) vannamei. *Aquaculture*, *250*(1–2), 95–101. https://doi.org/10.1016/j.aquaculture.2005.03.018

Ibarra, A. M., & Famula, T. R. (2008). Genotype by environment interaction for adult body weights of shrimp Penaeus vannamei when grown at low and high densitie. *Genetics Selection Evolution*, *40*(5), 541. https://doi.org/10.1186/1297-9686-40-5-541

Ibarra, A. M., Famula, T. R., & Arcos, F. G. (2009). Heritability of vitellogenin in hemolymph, a pre-spawning selectable trait in Penaeus (Litopenaeus) vannamei, has a large genetic correlation with ovary maturity measured as oocytes mean diameter. *Aquaculture*, *297*(1–4), 64–69. https://doi.org/10.1016/j.aquaculture.2009.09.015

Ibarra, A. M., Pérez-Rostro, C. I., Ramírez, J. L., & Ortega-Estrada, E. (2007). Genetics of the resistance to hypoxia in postlarvae and juveniles of the Pacific white shrimp Penaeus (Litopenaeus) vannamei (Boone 1931). *Aquaculture Research*, *38*(8), 838–846. https://doi.org/10.1111/j.1365-2109.2007.01734.x

Ibarra, A. M., Racotta, I. S., Arcos, F. G., & Palacios, E. (2007). Progress on the genetics of reproductive performance in penaeid shrimp. *Aquaculture*, *268*(1–4), 23–43. https://doi.org/10.1016/j.aquaculture.2007.04.028

Imron, I., & Suprapto, R. (2013). Heritability for growth related traits in giant freshwater prawn (*Macrobrachium rosenbergii*) at various developmental stages and culture conditions estimated by intraclass correlation. *Indonesian Aquaculture Journal*, *8*(1), 35–45. https://doi.org/10.15578/iaj.8.1.2013.35-45

Jembere, T., Dessie, T., Rischkowsky, B., Kebede, K., Okeyo, A., & Haile, A. (2017). Meta-analysis of average estimates of genetic parameters for growth, reproduction and milk production traits in goats. *Small Ruminant Research*, *153*, 71–80. https://doi.org/10.1016/j.smallrumres.2017.04.024

Kanchanachai, Y., Poompuang, S., Koonawootrittriron, S., & Uraiwan, S. (2011). Estimating genetic parameters for weight and body size of Pacific white shrimp (*Litopenaeus vannamei*) by restricted maximum likelihood method. *Kasetsart Journal (Natural Science)*, *45*, 1047–1057.

Karplus, I. (2005). Social control of growth in Macrobrachium rosenbergii (De Man): A review and prospects for future research. *Aquaculture Research*, *36*(3), 238–254. https://doi.org/10.1111/j.1365-2109.2005.01239.x

Kenway, M., Macbeth, M., Salmon, M., McPhee, C., Benzie, J., Wilson, K., & Knibb, W. (2006). Heritability and genetic correlations of growth and survival in black tiger prawn Penaeus monodon reared in tanks. *Aquaculture*, *259*(1–4), 138–145. https://doi.org/10.1016/j.aquaculture.2006.05.042

Khatkar, M., Zenger, K., Jones, D., Prochaska, J., van der Steen, H., Jerry, D., & Raadsma, H. (2017). Quantitative genomic analyses in the pacific whiteleg shrimp *Litopenaeus vannamei*. *Proceedings of the*

Conference of the Association for the Advancement of Animal Breeding and Genetics, *22*, 121–124.

Kitcharoen, N., Rungsin, W., Koonawootrittriron, S., & Na-Nakorn, U. (2012). Heritability for growth traits in giant freshwater prawn, Macrobrachium rosenbergii (de Mann 1879) based on best linear unbiased prediction methodology. *Aquaculture Research*, *43*(1), 19–25. https://doi.org/10.1111/j.1365-2109.2011.02796.x

Krishna, G., Gopikrishna, G., Gopal, C., Jahageerdar, S., Ravichandran, P., Kannappan, S., … Hayes, B. (2011). Genetic parameters for growth and survival in Penaeus monodon cultured in India. *Aquaculture*, *318*(1–2), 74–78. https://doi.org/10.1016/j.aquaculture.2011.04.028

Kristjánsson, T., & Arnason, T. (2016). Heritability of economically important traits in the Atlantic cod Gadus morhua L. *Aquaculture Research*, *47*(2), 349–356. https://doi.org/10.1111/are.12496

Li, W., Lu, X., Luan, S., Luo, K., Sui, J., & Kong, J. (2016). Heritability of body weight and resistance to ammonia in the Pacific white shrimp *Litopenaeus vannamei* juveniles. *Chinese Journal of Oceanology and Limnology*, *34*(5), 1025–1033. https://doi.org/10.1007/s00343-016-5034-0

Li, W., Luan, S., Luo, K., Sui, J., Lu, X., Wang, Q., & Kong, J. (2016). Genetic parameters for cold tolerance and body weight of Chinese fleshy prawn, *Fenneropenaeus chinensis*. *Acta Oceanologica Sinica*, *35*(6), 12–18. https://doi.org/10.1007/s13131-016-0845-6

Li, W., Luan, S., Luo, K., Sui, J., Xu, X., Tan, J., & Kong, J. (2015). Genetic parameters and genotype by environment interaction for cold tolerance, body weight and survival of the Pacific white shrimp Penaeus vannamei at different temperatures. *Aquaculture*, *441*, 8–15. https://doi.org/10.1016/j.aquaculture.2015.02.013

Li, Y., Zhou, F., Ma, Z., Huang, J., Jiang, S., Yang, Q., … Qin, J. G. (2016). Length–weight relationship and condition factor of giant tiger shrimp, Penaeus monodon (Fabricius, 1798) from four breeding families. *SpringerPlus*, *5*(1), 1279. https://doi.org/10.1186/s40064-016-2979-6

Liu, J., Zheng, J., & Liu, J. (2019). Genetic parameters for growth-related traits and survival with age in the Kuruma shrimp Marsupenaeus japonicus. *Aquaculture Research*, *50*(1), 42–48. https://doi.org/10.1111/are.13862

Lu, X., Luan, S., Cao, B., Meng, X., Sui, J., Dai, P., … Kong, J. (2017). Estimation of genetic parameters and genotype-by-environment interactions related to acute ammonia stress in Pacific white shrimp (*Litopenaeus vannamei*) juveniles at two different salinity levels. *PLoS One*, *12*(3), e0173835. https://doi.org/10.1371/journal.pone.0173835

Lu, X., Luan, S., Cao, B., Sui, J., Dai, P., Meng, X., … Kong, J. (2017). Heterosis and heritability estimates for the survival of the Pacific white shrimp (*Litopenaeus vannamei*) under the commercial scale ponds. *Acta Oceanologica Sinica*, *36*(2), 62–68. https://doi.org/10.1007/s13131-016-0942-6

Lu, X., Luan, S., Luo, K., Meng, X., Li, W., Sui, J., … Kong, J. (2016). Genetic analysis of the Pacific white shrimp (*Litopenaeus vannamei*): Heterosis and heritability for harvest body weight. *Aquaculture Research*, *47*(11), 3365–3375. https://doi.org/10.1111/are.12820

Luan, S., Luo, K., Chai, Z., Cao, B., Meng, X., Lu, X., … Kong, J. (2015). An analysis of indirect genetic effects on adult body weight of the Pacific white shrimp *Litopenaeus vannamei* at low rearing density. *Genetics Selection Evolution*, *47*(1), 95. https://doi.org/10.1186/s12711-015-0164-y

Luan, S., Yang, G., Wang, J., Luo, K., Chen, X., Gao, Q., … Kong, J. (2014). Selection responses in survival of Macrobrachium rosenbergii after performing five generations of multi-trait selection for growth and survival. *Aquaculture International*, *22*(3), 993–1007. https://doi.org/10.1007/s10499-013-9722-x

Luan, S., Yang, G., Wang, J., Luo, K., Zhang, Y., Gao, Q., … Kong, J. (2012). Genetic parameters and response to selection for harvest body weight of the giant freshwater prawn Macrobrachium rosenbergii. *Aquaculture*, *362*, 88–96. https://doi.org/10.1016/j.aquaculture.2012.05.011

Macbeth, M., Kenway, M., Salmon, M., Benzie, J., Knibb, W., & Wilson, K. (2007). Heritability of reproductive traits and genetic correlations with growth in the black tiger prawn Penaeus monodon reared in tanks. *Aquaculture*, *270*(1–4), 51–56. https://doi.org/10.1016/j.aquaculture.2007.03.018

Malecha, S. R., Masuno, S., & Onizuka, D. (1984). The feasibility of measuring the heritability of growth pattern variation in juvenile freshwater prawns, Macrobrachium rosenbergii (de Man). *Aquaculture*, *38*(4), 347–363. https://doi.org/10.1016/0044-8486(84)90339-9

Maluwa, A. O., & Gjerde, B. (2006). Genetic evaluation of four strains of *Oreochromis shiranus* for harvest body weight in a diallel cross. *Aquaculture*, *259*(1–4), 28–37. https://doi.org/10.1016/j.aquaculture.2006.06.003

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*(1), 56–73. https://doi.org/10.1177/0013164409344534

Megahed, M. E. (2019). Quantitative genetics of disease resistance in different groups of Penaeus semisulcatus from preliminary controlled challenge test with Vibrio parahaemolyticus the aetiological agent of acute hepatopancreatic necrosis disease (AHPND). *Journal of Applied Aquaculture*, *31*(1), 17–33. https://doi.org/10.1080/10454438.2018.1545723

Megahed, M. E., Elmesiry, G., Mohamed, K., & Dhar, A. K. (2019). Genetic parameters of survival in four generations of Indian white shrimp (Fenneropeneaus indicus). *Journal of Applied Aquaculture*, *31*(2), 121–139. https://doi.org/10.1080/10454438.2018.1551169

Mengistu, S. B., Mulder, H. A., Benzie, J. A., & Komen, H. (2019). A systematic literature review of the major factors causing yield gap by affecting growth, feed conversion ratio and survival in Nile tilapia (*Oreochromis niloticus*). *Reviews in Aquaculture*, *12*, 524–541. https://doi.org/10.1111/raq.12331

Merilä, J., & Sheldon, B. C. (1999). Genetic architecture of fitness and nonfitness traits: Empirical patterns and development of ideas. *Heredity*, *83*(2), 103–109. https://doi.org/10.1046/j.1365-2540.1999.00585.x

Montaldo, H. H., Castillo-Juárez, H., Campos-Montes, G., & Pérez-Enciso, M. (2013). Effect of the data family structure, tank replication and the statistical model, on the estimation of genetic parameters for body weight at 28 days of age in the Pacific white shrimp (Penaeus (Litopenaeus) vannamei B oone, 1931). *Aquaculture Research*, *44*(11), 1715–1723. https://doi.org/10.1111/j.1365-2109.2012.03176.x

Moss, D. R., Moss, S. M., & Lotz, J. M. (2013). Estimation of genetic parameters for survival to multiple isolates of Taura syndrome virus in a selected population of Pacific white shrimp Penaeus (Litopenaeus) vannamei. *Aquaculture*, *416*, 78–84. https://doi.org/10.1016/j.aquaculture.2013.07.037

Nguyen, N. H., Ninh, N. H., & Hung, N. H. (2019). Evaluation of two genetic lines of Pacific White leg shrimp Liptopenaeus vannamei selected in tank and pond environments. *Aquaculture*, *516*, 734522. https://doi.org/10.1016/j.aquaculture.2019.734522

Nguyen, N. H., Quinn, J., Powell, D., Elizur, A., Thoa, N. P., Nocillado, J., … Knibb, W. (2014). Heritability for body colour and its genetic association with morphometric traits in Banana shrimp (Fenneropenaeus merguiensis). *BMC Genetics*, *15*(1), 132. https://doi.org/10.1186/s12863-014-0132-5

Noble, T. H., Coman, G. J., Wade, N. M., Thomson, P. C., Raadsma, H. W., Khatkar, M. S., … Jerry, D. R. (2019). Genetic parameters for tolerance to gill-associated virus under challenge-test conditions in the black tiger shrimp (*Penaeus monodon*). *Aquaculture*, *516*, 734428. https://doi.org/10.1016/j.aquaculture.2019.734428

Nolasco-Alzaga, H. R., Perez-Enriquez, R., Enez, F., Bestin, A., Palacios-Mechetnov, E., & Haffray, P. (2018). Quantitative genetic parameters of growth and fatty acid content in the hemolymph of the Whiteleg shrimp *Litopenaeus vannamei*. *Aquaculture*, *482*, 17–23. https://doi.org/10.1016/j.aquaculture.2017.09.015

Norman-López, A., Sellars, M. J., Pascoe, S., Coman, G. J., Murphy, B., Moore, N., & Preston, N. (2016). Productivity benefits of selectively breeding Black Tiger shrimp (*Penaeus monodon*) in Australia. *Aquaculture Research*, *47*(10), 3287–3296. https://doi.org/10.1111/are.12782

Ødegård, J., Gitterle, T., Madsen, P., Meuwissen, T. H. E., Yazdi, M. H., Gjerde, B., … Rye, M. (2011). Quantitative genetics of taura syndrome resistance in pacific white shrimp (Penaeus vannamei): A cure model approach. *Genetics Selection Evolution*, *43*(1), 14. https://doi.org/10.1186/1297-9686-43-14

Pérez-Rostro, C. I., & Ibarra, A. M. (2003a). Heritabilities and genetic correlations of size traits at harvest size in sexually dimorphic Pacific white shrimp (*Litopenaeus vannamei*) grown in two environments. *Aquaculture Research*, *34*(12), 1079–1085. https://doi.org/10.1046/j.1365-2109.2003.00913.x

Pérez-Rostro, C. I., & Ibarra, A. M. (2003b). Quantitative genetic parameter estimates for size and growth rate traits in Pacific white shrimp, Penaeus vannamei (Boone 1931) when reared indoors. *Aquaculture Research*, *34*(7), 543–553. https://doi.org/10.1046/j.1365-2109.2003.00851.x

Pérez-Rostro, C., Ramirez, J., & Ibarra, A. (1999). Maternal and cage effects on genetic parameter estimation for Pacific white shrimp Penaeus vannamei Boone. *Aquaculture Research*, *30*(9), 681–693. https://doi.org/10.1046/j.1365-2109.1999.00377.x

Perry, G. M., Martyniuk, C. M., Ferguson, M. M., & Danzmann, R. G. (2005). Genetic parameters for upper thermal tolerance and growth-related traits in rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, *250*(1–2), 120–128. https://doi.org/10.1016/j.aquaculture.2005.04.042

Phuthaworn, C., Nguyen, N. H., Quinn, J., & Knibb, W. (2016). Moderate heritability of hepatopancreatic parvovirus titre suggests a new option for selection against viral diseases in banana shrimp (*Fenneropenaeus merguiensis*) and other aquaculture species. *Genetics Selection Evolution*, *48*(1), 64. https://doi.org/10.1186/s12711-016-0243-8

Pillai, B. R., Lalrinsanga, P. L., Ponzoni, R. W., Khaw, H. L., Mahapatra, K. D., Mohanty, S., … Jayasankar, P. (2017). Phenotypic and genetic parameters for body traits in the giant freshwater prawn (*Macrobrachium rosenbergii*) in India. *Aquaculture Research*, *48*(12), 5741–5750. https://doi.org/10.1111/are.13397

Pinkiewicz, T., Purser, G., & Williams, R. (2011). A computer vision system to analyse the swimming behaviour of farmed fish in commercial aquaculture facilities: A case study using cage-held Atlantic salmon. *Aquacultural Engineering*, *45*(1), 20–27. https://doi.org/10.1016/j.aquaeng.2011.05.002

Preston, N. P., Crocos, P. J., Keys, S. J., Coman, G. J., & Koenig, R. (2004). Comparative growth of selected and non-selected Kuruma shrimp Penaeus (Marsupenaeus) japonicus in commercial farm ponds; implications for broodstock production. *Aquaculture*, *231*(1–4), 73–82. https://doi.org/10.1016/j.aquaculture.2003.09.039

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rutten, M. J., Komen, H., & Bovenhuis, H. (2005). Longitudinal genetic analysis of Nile tilapia (Oreochromis niloticus L.) body weight using a random regression model. *Aquaculture*, *246*(1–4), 101–113. https://doi.org/10.1016/j.aquaculture.2004.12.020

Safari, E., Fogarty, N., & Gilmour, A. R. (2005). A review of genetic parameter estimates for wool, growth, meat and reproduction traits in sheep. *Livestock Production Science*, *92*(3), 271–289. https://doi.org/10.1016/j.livprodsci.2004.09.003

Singh, D., Wang, X. U., Kumar, U., Gao, L., Noor, M., Imtiaz, M., … Poland, J. (2019). High-throughput phenotyping enabled genetic dissection of crop lodging in wheat. *Frontiers in Plant Science*, *10*, 394. https://doi.org/10.3389/fpls.2019.00394

Sui, J., Luan, S., Luo, K., Meng, X., Cao, B., Liu, N., … Kong, J. (2016). Genetic parameters and response to selection of harvest body weight of the Chinese shrimp *Fenneropenaeus chinensis* after five generations of multi-trait selection. *Aquaculture*, *452*, 134–141. https://doi.org/10.1016/j.aquaculture.2015.08.011

Sui, J., Luan, S., Luo, K., Meng, X., Lu, X., Cao, B., … Xu, S. (2016). Genetic parameters and response to selection for harvest body weight of Pacific white shrimp, *Litopenaeus vannamei*. *Aquaculture Research*, *47*(9), 2795–2803. https://doi.org/10.1111/are.12729

Sui, J., Luan, S., Yang, G., Xia, Z., Luo, K., Tang, Q., … Kong, J. (2019). Genetic parameters and selection response for the harvest body weight of the giant freshwater prawn (*Macrobrachium rosenbergii*) in a breeding program in China. *PLoS One*, *14*(8), https://doi.org/10.1371/journal.pone.0218379

Sun, M. M., Huang, J. H., Jiang, S. G., Yang, Q. B., Zhou, F. L., Zhu, C. Y., … Su, T. F. (2015). Estimates of heritability and genetic correlations for growth-related traits in the tiger prawn P enaeus monodon. *Aquaculture Research*, *46*(6), 1363–1368. https://doi.org/10.1111/are.12290

Sutton, A. J., Abrams, K. R., Jones, D. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research* (Vol. *39*). Chichester, UK: John Wiley & Sons.

Tan, J., Kong, J., Cao, B., Luo, K., Liu, N., Meng, X., … Luan, S. (2017). Genetic parameter estimation of reproductive traits of *Litopenaeus vannamei*. *Journal of Ocean University of China*, *16*(1), 161–167. https://doi.org/10.1007/s11802-017-3127-9

Tan, J., Luan, S., Luo, K., Guan, J., Li, W., Sui, J., … Kong, J. (2017). Heritability and genotype by environment interactions for growth and survival in *Litopenaeus vannamei* at low and high densities. *Aquaculture Research*, *48*(4), 1430–1438. https://doi.org/10.1111/are.12978

Thiengpimol, P., Koonawootrittriron, S., & Wongtripop, S. (2014). Genetic parameters for body weight and survival rate of black tiger shrimp (*Penaeus monodon*). *Genomics and Genetics*, *7*(3), 158–163. https://doi.org/10.14456/tjg.2014.19

Trinh, T. T., Nguyen, N. H., Nguyen, H. H., Wayne, K., & Nguyen, N. H. (2019). Genetic variation in disease resistance against White Spot Syndrome Virus (WSSV) in *Liptopenaeus vannamei*. *Frontiers in Genetics*, *10*, 264. https://doi.org/10.3389/fgene.2019.00264

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), 255. https://doi.org/10.1038/nrg2322

Vu, N. T., & Nguyen, N. H. (2019). Quantitative genetic changes in reproductive performance of giant freshwater prawn after 10 years of selection for increased growth rate. *Reproduction in Domestic Animals*, *54*(2), 199–206. https://doi.org/10.1111/rda.13326

Vu, N. T., Trong, T. Q., & Nguyen, N. H. (2017). Effects of selection for fast growth on survival rate during grow-out phase in giant freshwater prawn (Macrobrachium rosenbergii). *BMC Genetics*, *18*(1), 56. https://doi.org/10.1186/s12863-017-0521-7

Wang, M., Kong, J., Meng, X., Luan, S., Luo, K., Sui, J., … Shi, X. (2017). Evaluation of genetic parameters for growth and cold tolerance traits in Fenneropenaeus chinensis juveniles. *PLoS One*, *12*(8), e0183801. https://doi.org/10.1371/journal.pone.0183801

Wang, Q., Yu, Y., Li, F., Zhang, X., & Xiang, J. (2017). Predictive ability of genomic selection models for breeding value estimation on growth traits of Pacific white shrimp *Litopenaeus vannamei*. *Chinese Journal of Oceanology and Limnology*, *35*(5), 1221–1229. https://doi.org/10.1007/s00343-017-6038-0

Wang, Q., Yu, Y., Yuan, J., Zhang, X., Huang, H., Li, F., & Xiang, J. (2017). Effects of marker density and population structure on

the genomic prediction accuracy for growth trait in Pacific white shrimp *Litopenaeus vannamei*. *BMC Genetics*, *18*(1), 45. https://doi.org/10.1186/s12863-017-0507-5

Wang, Q., Yu, Y., Zhang, Q., Zhang, X., Huang, H., Xiang, J., & Li, F. (2019). Evaluation on the genomic selection in *Litopenaeus vannamei* for the resistance against Vibrio parahaemolyticus. *Aquaculture*, *505*, 212–216. https://doi.org/10.1016/j.aquaculture.2019.02.055

Xu, R.-W., Chinh, P. X., Faisal, R. M., & Liu, X.-L. (2016). *Estimation of the Heritability for Growth Traits at 3 Months and 5 Months of Pacific White Shrimp, Litopenaeus Vannamei*. Paper presented at The International Conference on Biological Sciences and Technology.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565. https://doi.org/10.1038/ng.608

Yuan, R., Hu, Z., Liu, J., & Zhang, J. (2018). Genetic parameters for growth-related traits and survival in pacific white shrimp, *Litopenaeus vannamei* under conditions of high ammonia-N concentrations. *Turkish Journal of Fisheries and Aquatic Sciences*, *18*(1), 37–47. https://doi.org/10.4194/1303-2712-v18_1_05

Zenger, K. R., Khatkar, M. S., Jones, D. B., Khalilisamani, N., Jerry, D. R., & Raadsma, H. W. (2018). Genomic selection in aquaculture: Application, limitations and opportunities with special reference to marine shrimp and pearl oysters. *Frontiers in Genetics*, *9*, 693. https://doi.org/10.3389/fgene.2018.00693

Zhang, J., Cao, F., Liu, J., & Yuan, R. (2017). Genetic parameters for growth and survival traits in *Litopenaeus vannamei* at different ages. *Aquaculture International*, *25*(5), 1901–1911. https://doi.org/10.1007/s10499-017-0163-9

Zhang, J., Cao, F., Liu, J., Yuan, R., & Hu, Z. (2017). Genetic parameters for growth and hypoxic tolerance traits in pacific white shrimp *Litopenaeus vannamei* at different ages. *North American Journal of Aquaculture*, *79*(1), 75–83. https://doi.org/10.1080/15222055.2016.1194923

Zhang, T., Kong, J., Liu, B., Wang, Q., Cao, B., Luan, S., & Wang, W. (2014). Genetic parameter estimation for juvenile growth and upper thermal tolerance in turbot (*Scophthalmus maximus* Linnaeus). *Acta Oceanologica Sinica*, *33*(8), 106–110. https://doi.org/10.1007/s13131-014-0460-3

Zhang, T., Kong, J., Luan, S., Wang, Q., Luo, K., & Tian, Y. (2011). Estimation of genetic parameters and breeding values in shrimp *Fenneropenaeus chinensis* using the REML/BLUP procedure. *Acta Oceanologica Sinica*, *30*(1), 78–86. https://doi.org/10.1007/s13131-011-0093-8

Zhao, J., Bao, W. J., Zhang, F. D., Ye, Z. Y., Liu, Y., Shen, M. W., & Zhu, S. M. (2017). Assessing appetite of the swimming fish based on spontaneous collective behaviors in a recirculating aquaculture system. *Aquacultural Engineering*, *78*, 196–204. https://doi.org/10.1016/j.aquaeng.2017.07.008

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Hasan MM, Tulloch RL, Thomson PC, Raadsma HW, Khatkar MS. Meta-analysis of genetic parameters of production traits in cultured shrimp species. *Fish Fish*. 2020;21:1150–1174. https://doi.org/10.1111/faf.12495

Appendix A4: An emergent clade of SARS-CoV-2 linked to return travellers from Iran

VIRUS EVOLUTION

Virus Evolution, 2020, 6(1): veaa027

doi: 10.1093/ve/veaa027
Rapid Communication

An emergent clade of SARS-CoV-2 linked to returned travellers from Iran

John-Sebastian Eden,[1,2,*,†] Rebecca Rockett,[1,3,4] Ian Carter,[3] Hossinur Rahman,[3] Joep de Ligt,[5] James Hadfield,[6] Matthew Storey,[5] Xiaoyun Ren,[5] Rachel Tulloch,[1,2] Kerri Basile,[3*] Jessica Wells,[3] Roy Byun,[7] Nicky Gilroy,[3] Matthew V. O'Sullivan,[3,4] Vitali Sintchenko,[1,3,4] Sharon C. Chen,[1,3,4] Susan Maddocks,[3] Tania C. Sorrell,[1,2,3] Edward C. Holmes,[1,‡] Dominic E. Dwyer,[1,3,4] and Jen Kok[3,4]; for the 2019-nCoV Study Group[§]

[1]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences & School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia, [2]Centre for Virus Research & Centre for Infectious Diseases and Microbiology, Westmead Institute for Medical Research, PO Box 412, Westmead, NSW 2145, Australia, [3]Centre for Infectious Diseases and Microbiology Laboratory Services, NSW Health Pathology – Institute of Clinical Pathology and Medical Research, Westmead, NSW 2145, Australia, [4]Centre for Infectious Diseases and Microbiology – Public Health, Westmead Hospital, Westmead, NSW 2145, Australia, [5]Institute of Environmental Science and Research, Porirua 5240, New Zealand, [6]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA and and [7]NSW Ministry of Health, North Sydney, NSW 2059, Australia

*Corresponding authors: E-mails: js.eden@sydney.edu.au; kerri.basile@health.nsw.gov.au
†https://orcid.org/0000-0003-1374-3551
‡https://orcid.org/0000-0001-9596-3552
§The members of 2019-nCoV Study Group are listed in the Acknowledgments.

Abstract

The SARS-CoV-2 epidemic has rapidly spread outside China with major outbreaks occurring in Italy, South Korea, and Iran. Phylogenetic analyses of whole-genome sequencing data identified a distinct SARS-CoV-2 clade linked to travellers returning from Iran to Australia and New Zealand. This study highlights potential viral diversity driving the epidemic in Iran, and underscores the power of rapid genome sequencing and public data sharing to improve the detection and management of emerging infectious diseases.

Key words: COVID-19; SARS-CoV-2; genome sequencing; phylogenetics.

1. Introduction

From a public health perspective, the real-time whole-genome sequencing (WGS) of emerging viruses enables the informed development and design of molecular diagnostic assays, and tracing patterns of spread across multiple epidemiological scales (i.e. genomic epidemiology). However, WGS capacities and data sharing policies vary in different countries and jurisdictions, leading to potential sampling bias due to delayed or underrepresented sequencing data from some areas with substantial SARS-CoV-2 activity. Herein, we show that the genomic analyses of SARS-CoV-2 strains from Australian returned travellers with COVID-19 disease may provide important insights into viral diversity present in regions currently lacking genomic data.

2. SARS-CoV-2 emergence and dissemination

In late December 2019, a cluster of cases of pneumonia of unknown aetiology in Wuhan city, Hubei province, China was reported by health authorities (Wuhan Municipal Health Commission 2019). A novel betacoronavirus, designated SARS-CoV-2, was identified as the causative agent (Wu et al. 2020) of the disease now known as COVID-19, with substantial human-to-human transmission (Lu et al. 2020). To contain a growing epidemic, Chinese authorities implemented strict quarantine measures in Wuhan and surrounding areas in Hubei province. Significant delays in the global spread of the virus were achieved, but despite these measures, cases were exported to other countries. As of 9 March 2020, these numbered more than 100 countries, on all continents except Antarctica; the total number of confirmed infections exceeded 110,000 and there were nearly 4,000 deaths (Dong, Du, and Gardner 2020). Although the majority of cases have occurred in China, major outbreaks have also been reported in Italy, South Korea, and Iran (World Health Organisation 2020a). Importantly, there is widespread local transmission in multiple countries outside China following independent importations of infection from visitors and returned travellers.

3. WGS of SARS-CoV-2 cases in Australia and New Zealand

Viral extracts were prepared from respiratory tract samples where SARS-CoV-2 was detected by reverse-transcription polymerase chain reaction (RT-PCR) using World Health Organisation (2020b) recommended primers and probes targeting the E and RdRp genes. In New South Wales (NSW), Australia, WGS for SARS-CoV-2 was developed based on an existing amplicon-based Illumina sequencing approach (Di Giallonardo et al. 2018). Viral extracts were reverse transcribed with SSIV VILO cDNA master mix and then used as input for multiple overlapping PCR reactions (~2.5 kb each) spanning the viral genome using Platinum SuperFi master mix (primers provided in Supplementary Table S1). Amplicons were pooled equally, purified, and quantified. Nextera XT libraries were prepared and sequencing was performed with multiplexing on an Illumina iSeq (300 cycle flow cell). In New Zealand, the ARTIC network protocol was used for WGS (Quick 2020). In short, 400-bp tiling amplicons designed with Primal Scheme (Grubaugh et al. 2019a) were used to amplify viral cDNA prepared with SuperScript III. A sequence library was then constructed using the Oxford NanoPore ligation sequencing kit and sequenced on a R9.4.1 MinION flow cell. Near-complete viral genomes were then assembled *de novo* in Geneious Prime 2020.0.5 or through reference mapping with RAMPART V1.0.6 (Hadfield 2019) using the ARTIC network nCoV-2019 novel coronavirus bioinformatics protocol (Loman and Rambaut 2020). In total, 13 SARS-CoV-2 genomes were sequenced from cases in NSW diagnosed between 24 January and 3 March 2020, as well as a single genome from the first patient in Auckland, New Zealand sampled on 27 February 2020 (Table 1). Australian and New Zealand sequences were aligned to global reference strains sourced from GISAID with MAFFT (Katoh 2002) and then compared phylogenetically using a maximum-likelihood approach—PhyML v2.2.4 (Guindon and Gascuel 2003).

4. A distinct clade of SARS-CoV-2 identified in travellers returned from Iran

The Australian strains of SARS-CoV-2 were dispersed across the global SARS-CoV-2 phylogeny (Fig. 1A). The first four cases of COVID-19 disease in NSW occurred between 24 and 26 January 2020, and these were closely related with 1–2 single nucleotide polymorphisms (SNPs) difference to the dominant variant circulating in Wuhan at the time (prototype strain MN908947/SARS-CoV-2/Wuhan-Hu-1). As the four patients identified in this period had recently returned from China, this region was the likely source of infection. From 1 February 2020, travel to Australia from mainland China was restricted to returning Australian residents and their children, who were placed in home quarantine for 14 days. Despite the intensive testing of such returning travellers, no further cases of COVID-19 were detected in NSW until 28 February 2020, when SARS-COV-2 was detected in an individual returning from Iran (NSW05). A close contact of this individual also tested positive (NSW14) providing the first evidence of local transmission within NSW. This was followed by further Iran travel-linked cases in NSW (NSW06, NSW11, NSW12, and NSW13) and New Zealand (NZ01).

Of note, the genomes of all patients with a history of travel to Iran were part of a monophyletic group defined by three nucleotide substitutions (G1397A, T28688C, and G29742T) in the SARS-CoV-2 genome relative to the Wuhan prototype strain (Fig. 1B). G1397A and T28688C both occur in coding regions with G1397A producing a non-synonymous change (V378I) in the ORF1ab-encoded non-structural protein 2 region. G29742T occurs in the 3′-UTR. In addition to the Australian and New Zealand strains, this clade also included a traveller who had returned to Canada from Iran (BC_37_0-2), providing further evidence of its likely link to the Iranian epidemic. Indeed, a search of all currently available GISAID sequences and metadata revealed no other complete genome sequences from patients with documented history of travel to or residence in Iran (as of 9 March 2020). A search of partial sequences identified two SARS-CoV-2 sequences which originated in Iran (413553/IRN/Tehran15AW/2020-02-28 and 413554/IRN/Tehran9BE/2020-02-23) spanning a 363 nt region of the viral nucleoprotein (N). Although short in length, these two sequences covered one of the informative SNPs defining this clade—T28688C, and both Iranian strains matched the sequences from patients with travel histories to Iran and grouped by phylogenetic analysis (Supplementary Figs. S1 and S2).

5. Discussion

Technological advancements and the widespread adoption of WGS in pathogen genomics have transformed public health and infectious disease outbreak responses (Popovich and Snitkin 2017). Previously, disease investigations often relied on the targeted sequencing of a small locus to identify genotypes and

Table 1. SARS-CoV-2 genomes sequenced in this study

GISAID ID	Virus name	Location	Collection date	Travel history
EPI_ISL_408976	408976/Australia/Sydney-2/2020-01-22	Sydney, Australia	22 January 2020	China
EPI_ISL_407893	407893/Australia/NSW01/2020-01-24	Sydney, Australia	24 January 2020	China
EPI_ISL_408977	408977/Australia/Sydney-3/2020-01-25	Sydney, Australia	25 January 2020	China
EPI_ISL_413490	413490/New_Zealand/01/2020-02-27	Auckland, New Zealand	27 February 2020	Iran
EPI_ISL_412975	412975/Australia/NSW05/2020-02-28	Sydney, Australia	28February 2020	Iran
EPI_ISL_413594	413594/Australia/NSW08/2020-02-28	Sydney, Australia	28 February 2020	SE Asia
EPI_ISL_413595	413595/Australia/NSW09/2020-02-28	Sydney, Australia	28 February 2020	SE Asia
EPI_ISL_413213	413213/Australia/NSW06/2020-02-29	Sydney, Australia	29 February 2020	Iran
EPI_ISL_413214	413214/Australia/NSW07/2020-02-29	Sydney, Australia	29 February 2020	None
EPI_ISL_413596	413596/Australia/NSW10/2020-02-28	Sydney, Australia	1 March 2020	SE Asia
EPI_ISL_413597	413597/AUS/NSW11/2020-03-02	Sydney, Australia	2 March 2020	Iran
EPI_ISL_413600	413600/AUS/NSW14/2020-03-03	Sydney, Australia	3 March 2020	None
EPI_ISL_413598	413598/AUS/NSW12/2020-03-04	Sydney, Australia	4 March 2020	Iran
EPI_ISL_413599	413599/AUS/NSW13/2020-03-04	Sydney, Australia	4 March 2020	Iran

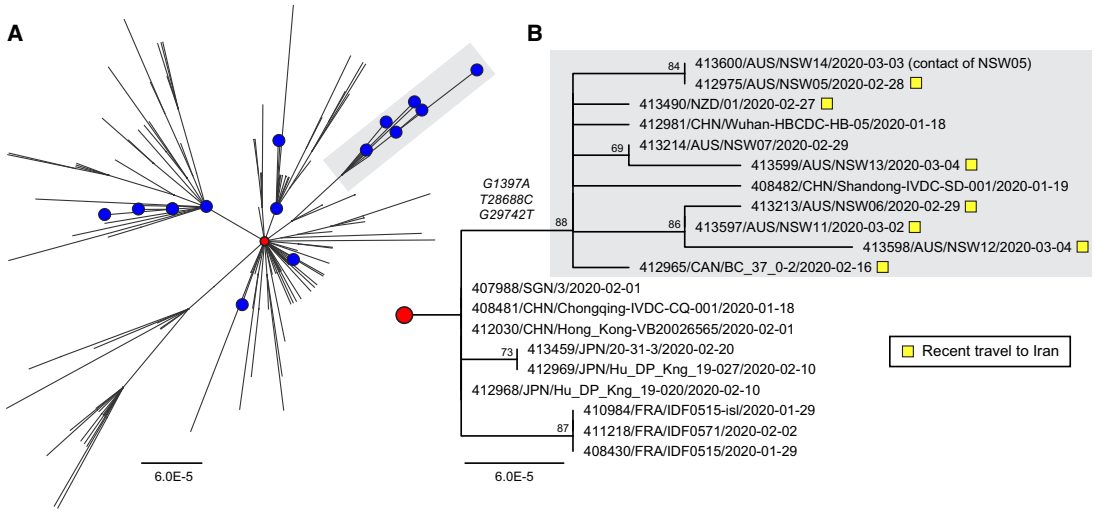


Figure 1. Phylogenetic analysis of SARS-CoV-2 genome sequences highlighting a clade of imported cases from Iran. (A) Global diversity of circulating SARS-CoV-2 strains including Australian sequences (blue circles, $n = 19$). The prototype strain Wuhan-Hu-1 is shown as a red circle. An emergent clade containing cases imported from Iran is highlighted with grey shading. (B) Sub-tree showing the informative branch containing imported Iranian cases (highlighted with yellow squares) and defined by substitutions at positions G1397A, T28688C, and G29742T. Node support is provided as bootstrap values of 100 replicates. For both (A) and (B), the scales are proportional to the number of substitutions per site.

infer patterns of spread along with epidemiological data (Dudas and Bedford 2019). As seen with the recent West African Ebola (Dudas et al. 2017) and Zika virus epidemics (Grubaugh et al. 2018), rapid WGS significantly increases resolution of diagnosis and surveillance thereby strengthening links between genomic, clinical, and epidemiological data (Grenfell 2004), and potentially uncovering outbreaks in unsampled locations (Grubaugh et al. 2019b). This advance improves our understanding of pathogen origins and spread that ultimately lead to stronger and more timely intervention and control measures (Grubaugh et al. 2019c). Following the first release of the SARS-CoV-2 genome (Wu et al. 2020), public health and research laboratories worldwide have rapidly shared sequences on public data repositories such as GISAID (Shu and McCauley 2017) ($n = 236$ genomes as of 9 March 2020) that have been used to provide near real-time snapshots of global diversity through public analytic and visualization tools (Hadfield et al. 2018).

Although all known cases linked to Iran are contained in this clade, it is important to note the presence of two Chinese strains sampled during mid-January 2020 from Hubei and Shandong provinces. It is expected that further Chinese strains will be identified that fall in this clade, and across the entire phylogenetic diversity of SARS-CoV-2 as this is where the outbreak started, including likely for the outbreak in Iran itself. However, while we cannot completely discount that the cases in Australia and New Zealand came from other sources including China, our phylogenetic analyses, as well as epidemiological (recent travel to Iran) and clinical data (date of symptom onset), provide evidence that this clade of SARS-CoV-2 is directly linked to the Iranian epidemic, from where genomic data are currently lacking. Importantly, the seemingly multiple importations of very closely related viruses from Iran into Australia suggest that this diversity reflects the early stages of SARS-CoV-2 transmission within Iran.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

The members of the nCoV-2019 Study Group also include Linda Donovan, Shanil Kumar, Tyna Tran, Danny Ko, Christine Ngo, Tharshini Sivaruban, Verlaine Timms, Connie Lam, Mailie Gall, Karen-Ann Gray, Rosemarie Sadsad, and Alicia Arnott. The authors acknowledge the Sydney Informatics Hub and the use of the University of Sydney's high-performance computing cluster, Artemis, and all the laboratories that have referred SARS-CoV-2 samples to the Centre for Infectious Diseases and Microbiology Laboratory Services, NSW Health Pathology – Institute of Clinical Pathology and Medical Research, Westmead Hospital. We would finally like to thank all the authors who have kindly shared genome data on GISAID, and we have included a table (Supplementary Table S2) outlining the authors and institutes involved. Data including the sequences in this study are available for download from https://www.gisaid.org/.

Author contributions

Study concept and design by J.S.E., E.C.H., and J.K. Sample processing and testing by I.C. and H.R. Sequencing and analysis by J.S.E., R.R., J.D.L., J.H., M.S., X.R., R.T., and E.C.H. Study coordination by K.B., J.W., R.B., N.G., M.V.O., V.S., S.C.C., S.M., T.C.S., D.E.D., and J.K. J.S.E. wrote the first manuscript draft with editing from E.C.H., J.D.L., R.R., T.C.S., V.S., and J.K. The final manuscript was approved by all authors.

Funding

This study was supported by the Prevention Research Support Programme funded by the New South Wales Ministry of Health and the NHMRC Centre of Research Excellence in Emerging Infectious Diseases (GNT1102962).

Conflict of interest: None declared.

References

Di Giallonardo, F. et al. (2018) 'Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia', *Viruses*, 10: 476.

Dong, E., Du, H., and Gardner, L. (2020) 'An Interactive Web-Based Dashboard to Track COVID-19 in Real Time', *The Lancet Infectious Diseases*, 20: 30120–1.

Dudas, G., and Bedford, T. (2019) 'The Ability of Single Genes vs Full Genomes to Resolve Time and Space in Outbreak Analysis', *BMC Evolutionary Biology*, 19: 1–17.

—— et al. (2017) 'Virus Genomes Reveal Factors That Spread and Sustained the Ebola Epidemic', *Nature*, 544: 309–15.

Grenfell, B. T. (2004) 'Unifying the Epidemiological and Evolutionary Dynamics of Pathogens', *Science*, 303: 327–32.

Grubaugh, N. D. et al. (2018) 'Genomic Insights into Zika Virus Emergence and Spread', *Cell*, 172: 1160–2.

—— et al. (2019a) 'An Amplicon-Based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and iVar', *Genome Biology*, 20: 8.

—— et al. (2019b) 'Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic', *Cell*, 178: 1057–71.

—— et al. (2019c) 'Tracking Virus Outbreaks in the Twenty-First Century', *Nature Microbiology*, 4: 10–9.

Guindon, S., and Gascuel, O. (2003) 'A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood', *Systematic Biology*, 52: 696–704.

Hadfield, J. (2019) *ARTIC Network RAMPART* <https://github.com/artic-network/rampart> accessed 10 March 2020.

—— et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Katoh, K. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.

Loman, N., and Rambaut, A. (2020) *ARTIC Network Bioinformatics SOP* <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html> accessed 10 March 2020.

Lu, R. et al. (2020) 'Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding', *The Lancet*, 395: 565–74.

Popovich, K. J., and Snitkin, E. S. (2017) 'Whole Genome Sequencing-Implications for Infection Prevention and Outbreak Investigations', *Current Infectious Disease Reports*, 19: 15.

Quick, J. (2020) *nCoV-2019 Sequencing Protocol* <https://www.protocols.io/view/ncov-2019-sequencing-protocol-single-sample-bdbfi2jn> accessed 10 March 2020.

Shu, Y., and McCauley, J. (2017) 'GISAID: Global Initiative on Sharing All Influenza Data—From Vision to Reality', *Eurosurveillance*, 22: 30494.

World Health Organisation (2020a) Coronavirus Situation Report—8th *March 2020* <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200308-sitrep-48-covid-19.pdf?sfvrsn=16f7ccef_4> accessed 9 March 2020.

—— (2020b) *Coronavirus Disease (COVID-19) Technical Guidance: Laboratory Testing for 2019-nCoV in Humans* <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance> accessed 10 March 2020.

Wu, F. et al. (2020) 'A New Coronavirus Associated with Human Respiratory Disease in China', *Nature*, 579: 265–9.

Wuhan Municipal Health Commission. (2019) *Briefing on the Current Pneumonia Epidemic Situation in Our City 2019* <http://wjw.wuhan.gov.cn/front/web/showDetail/2019123108989> accessed 9 March 2020.