Application of rapid sequencing for the detection and epidemiology of respiratory pathogens

Alp Aydin

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

University of East Anglia

Quadram Institute Bioscience

August 2022

Abstract

Lower respiratory tract infections (LRTI) are a leading cause of morbidity and mortality globally, and the rise of Antimicrobial Resistance (AMR) complicates their treatment. To achieve the best patient outcomes and avoid contributing to the rise of AMR, timely and appropriate antimicrobial treatment needs to be prescribed. However, the current gold standard for aetiological investigation of LRTIs (microbiological culture) is too slow to guide initial therapy.

Clinical metagenomics (CMg) has emerged as a potential solution to this problem; however, existing methods are too laborious. In this study, we optimise our previously published CMg pipeline to achieve a sensitive workflow with a 3.5 hour turnaround time. Evaluating the workflow, we show efficient depletion (>99.8%) of host DNA with our new 15 minute host depletion method. Sensitivity and specificity are 90.5% and 62.5%, respectively, rising to 96.6% and 100% when qPCR is used to investigate discordance. We also show that 30 minutes of sequencing is sufficient to make an accurate pathogen call.

For pathogen surveillance, targeted sequencing approaches are more appropriate. Sequencing of SARS-CoV-2 for genomic epidemiology became a valuable tool during the ongoing COVID-19 pandemic. However, early on, methods were low-throughput and inflexible. We responded to this by developing a high-throughput library preparation method, CoronaHiT, which can be used for sequencing SARS-CoV-2 on Illumina or Oxford Nanopore Technologies platforms. The method was shown to be cheap and accurate, while also being more robust for samples with lower viral loads. CoronaHiT has subsequently been used to sequence hundreds of thousands of SARS-CoV-2 genomes in the UK.

In conclusion, we have developed and optimised two different approaches for investigating respiratory infections (CMg and targeted) for two different applications, demonstrating the potential of rapid sequencing. Methods like these will continue to reshape diagnostics and public health in the future.

Table of Contents

Acknowledgements

Firstly, I want to thank my supervisor and mentor Prof. Justin O'Grady. The immeasurable gratitude I have for you cannot be overstated. I appreciate all the support you gave me, every teaching moment, every opportunity. Thank you for the doors you opened for me, for believing in me and for pushing me. Over the past 4 years, you have made me a better scientist and person, and I hope I make you proud in the future.

Thank you also to my secondary supervisor, Prof. John Wain. I could always count on your support throughout my PhD and beyond. I appreciate every bit of invaluable advice and feedback you gave me.

I also want to give special thanks to Dr. Gemma Kay. You have been my rock for the past 4 years. You were there for me every step of the way, right from the very first email you sent me offering your help before I even stepped foot in Norwich. It is truly difficult to put into words just how much I appreciate your constant support and friendship.

Thank you also to the other members of the former O'Grady group including my good friends Themis, Alex, Lluis, Mike, Riccardo and honorary members Claire and Charlotte. PhDs are tough, but you all made it easier with your friendships. I could not have asked for a better group of people to be on this journey with.

Thank you to Dr. Vicky Enne, for nurturing my early career and enabling me to find this PhD in the first place. I will never take for granted your enduring support.

Thank you to the wider Quadram members and local COG-UK team. Working together with you during unprecedented times to deliver a huge public health service was a privilege. Also thank you to all the unsung staff who make PhD research possible, the laboratory managers, the technicians, the graduate school staff.

Thank you to my family and friends for always being there for me and putting up with me. To my dear parents, Sadiye and Hasan, you instilled in me the value of education right from the beginning and gave me everything in your means to ensure my success. My successes are your successes. To my only sibling, Yasemin, I do not think you realise just how important you are to me and how much I appreciate you.

And finally, my Alex. The amount of patience you have with me is nothing short of angelic. You and Lady kept me going during even the most stressful times. Thank you.

List of Figures

List of Tables

1. Introduction

1.1 Lower respiratory tract infections

Lower respiratory tract infections (LRTI) are infections that affect the airways. Before the recent Coronavirus Disease 2019 (COVID-19) pandemic, LRTIs ranked the 4th leading cause of death globally, killing 2.6 million people in 2019, making it the most lethal communicable disease globally [1,2]. It is an even larger problem in low-income countries, being the second ranking cause of death in 2019. COVID-19, which is a respiratory disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) can also progress to the lower respiratory tract and cause pneumonia, and has resulted in over 6.4 million confirmed deaths since the end of 2019 [3].

LRTIs affect the passages of the lower respiratory system including the trachea, primary bronchi and the lungs [4]. LRTI is often used interchangeably with pneumonia, which is an infection that causes inflammation of the air sacs themselves [5] and is a large constituent of LRTIs, however, LRTI also includes bronchitis which is inflammation of the lining of the bronchial tubes [6]. In contrast, upper respiratory tract infections (URTIs) affect the upper respiratory passages, including the nasal cavity, pharynx and larynx. URTIs are very common, especially during the winter, but compared to LRTIs, are generally benign and self-limiting with mild symptoms [7]. Causes include viruses such as the common cold (rhinovirus), influenza, adenovirus and respiratory syncytial virus (RSV) [7]. While URTIs have an economic burden due to time off sick and outpatient appointments, complications are rare, unless the infection spreads to other parts of the body (e.g., lower respiratory system, blood, brain). On the other hand, LRTIs are a much bigger problem with high hospitalisation and fatality rates, especially for infants and the elderly [8]. Risk factors of LRTIs include malnutrition, chronic health conditions, lack of immunisation and crowded living conditions.

Pneumonia is one of the most important LRTIs and can be further divided into categories depending on how it was acquired. These different forms of pneumonia have different

aetiologies and are generally treated differently. Community-acquired pneumonia (CAP) is pneumonia that develops in the community (i.e. outside of a hospital) [9]. Hospital-acquired pneumonia (HAP) is pneumonia that develops 48h after hospitalisation of a patient. Ventilator-associated pneumonia (VAP) is a form of HAP that develops while a patient is being mechanically ventilated. Healthcare-associated pneumonia (HCAP) was a previously used category for pneumonia for patients who interacted with the healthcare system without being hospitalised, however, this term has fallen out of favour as studies have shown that HCAP is a poor predictor of multi-drug resistance (MDR) and potentially led to the use of increased antibiotics without resulting in better patient outcomes [10]. VAP can be considered a subset of HAP, however, for treatment purposes, can be considered its own entity. Even though SARS-CoV-2 can cause pneumonia, COVID-19 has often been listed and managed separately from CAP/HAP/VAP [11].

1.1.1 Community Acquired Pneumonia

CAP has a high mortality rate, with 5-15% of patients hospitalised for CAP in the UK dying within 30 days, rising to 30% for those that need intensive care [12]. Additionally, CAP can have long-term effects, diminishing the quality of life and aggravating other underlying health issues such as cardiovascular disease and renal disease. Like all pneumonias, it affects older people disproportionately. CAP also has an indirect economic impact – in addition to the hospital costs, half of CAP patients in the UK require more than two weeks off work [12]. In Europe, the economic cost of work days lost to CAP is estimated to be 3.6 billion euros per annum [12].

Streptococcus pneumoniae is the leading bacterial cause of CAP in adults, followed by *Haemophilus influenzae* [13]. Atypical causes of CAP include *Mycoplasma pneumoniae*, *Legionella pneumophila* and *Chlamydia pneumoniae* (Figure 1.1). The proportion of patients where no aetiologic agent can be identified is high, with studies reporting a range between 50.2 - 67.1% [13]. MDR organisms account for less than 20% of CAP, with MDR

S. aureus and *P. aeruginosa* being most common [14]. The prevalence of respiratory viruses in CAP was previously underestimated, recent studies that include Polymerase Chain Reaction (PCR) have shown that 30-40% of CAP patients can have a respiratory virus [13]. The most common virus in adults with CAP is influenzae virus, with it being detected in 9% of patients on average [15]. Viral and bacterial coinfection can also be common, which means that identification of a virus alone cannot be used to rule out a bacterial infection and discontinue antibiotic treatment [13]. The detection of polymicrobial infections have increased due to the use of newer methods which have also led to the reduction of cases with no known aetiology; this can improve the accuracy of treatment [14].

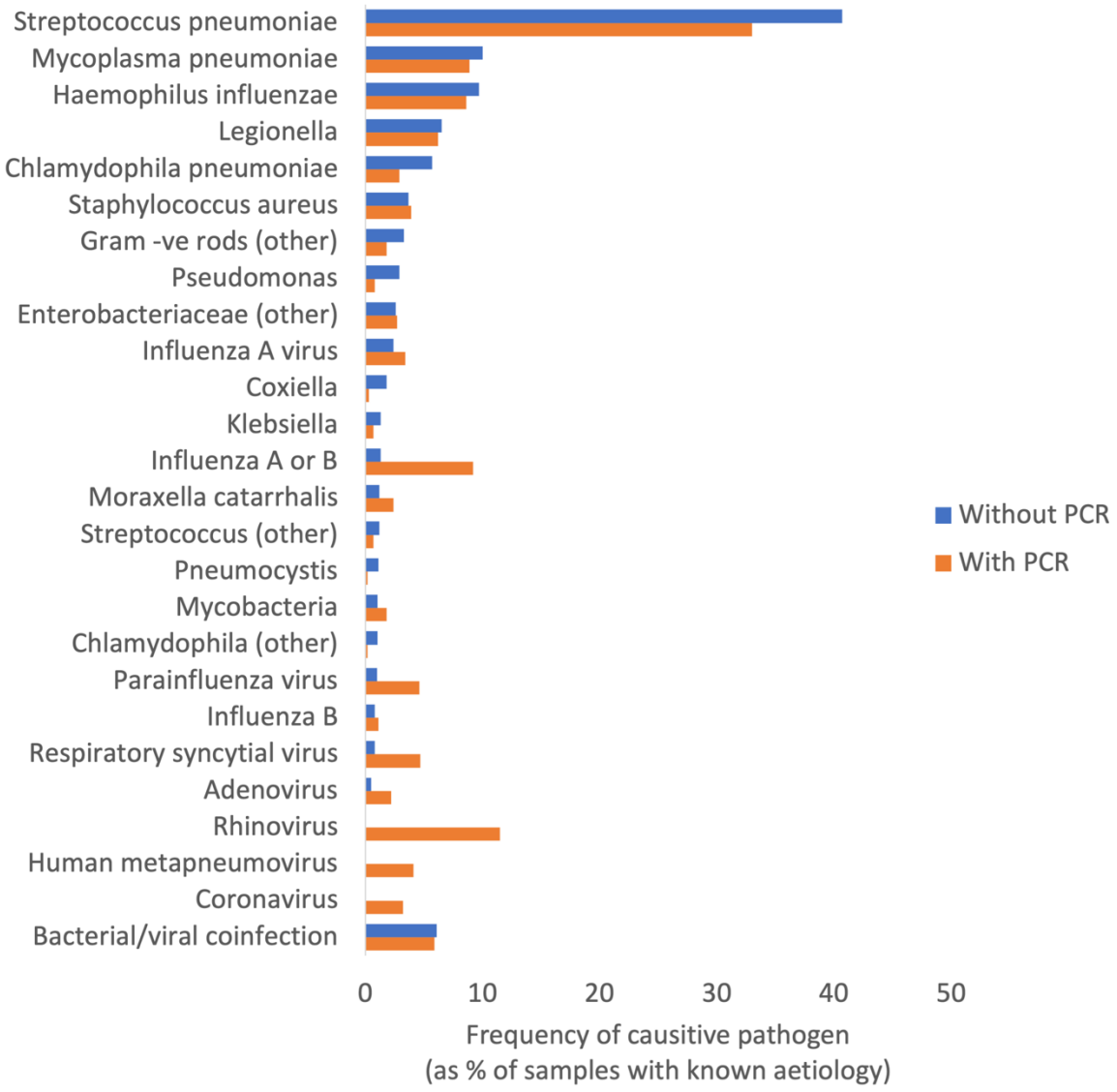


Figure 1.1 – Aetiology of CAP in adults, stratified by laboratory techniques used (Without PCR = culture and serological methods, With PCR = culture and serological methods + PCR for atypicals and viruses). Adapted from Shoar and Mushner, 2020 [13]

For patients that require intensive care, mortality can rise to as high as 50%, therefore, early and effective treatment is crucial [16]. Currently, treatment of CAP depends on the risk group and the most likely causative pathogen [16]. Treatment of non-severe cases is largely based around the assumption of likely *S. pneumoniae* infection. In the UK, amoxicillin is recommended as the first-choice antibiotic for low severity and moderate cases of CAP, plus a macrolide if an atypical is suspected (local AMR and surveillance data can guide this). For high severity cases, amoxicillin with clavulanic acid (co-amoxiclav) and a macrolide combination treatment is recommended, which provides

broad-spectrum cover for Gram-negatives and atypicals too [17]. This lines up with recommendations from the European Respiratory Society (ERS) which states that combination therapy should not be used for non-severe cases of CAP [14]. If a causative organism is detected or suspected, empiric treatment can be changed to a targeted therapy. Treatment failures can occur due to the emergence of MDR, which is on the rise [16]. Antibiotic resistance in *S. pneumoniae* can also be found, with high rates of penicillin resistance seen in some countries like Spain, however, amoxicillin is sufficient for treatment of *S. pneumoniae* in these cases [14].

1.1.2 Hospital Acquired Pneumonia and Ventilator Associated Pneumonia

HAP is defined as a pneumonia occurring 48 hours or more after hospital admission [18]. Risk factors include old age, being male and structural lung disease. Non-ventilator HAP has a prevalence of 1% in hospital inpatients, lengthens hospital stay by 8 days on average and has a high mortality rate [11]. One study showed that 15.5% of non-ventilator HAP patients die compared to 1.6% of other hospitalised patients and have a 19.0% chance of requiring ventilators compared to 3.9% for other hospitalised patients [19]. HAP also has a considerable cost to hospitals, with one study showing average costs of £43,100 per patient; antibiotic costs were only a small fraction of this cost at £321 [20]. Unfortunately, HAP is difficult to prevent and data on prevention strategies for HAP are either absent or of poor quality [18].

Gram-negative bacteria are the most common cause of HAP, namely *Pseudomonas aeruginosa, Klebsiella spp*, *Escherichia coli* and *Acinetobacter baumannii* [18]. *Staphylococcus aureus* is the most common Gram-positive cause. Many of these pathogens are considered antibiotic resistance threats globally [21] due to the difficulty in treating these infections, with particular concern for carbapenem-resistant Gram-negative bacteria. The World Health Organisation (WHO) released a priority list of pathogens that require urgent action, and many of the constituents are HAP pathogens (Table 1.1).

Viruses are not typically seen as important pathogens for severe HAP, however, viruses such as RSV and parainfluenza virus, as well as influenza and rhinovirus can be identified from bronchoalveolar lavages (BAL) [22]. Viruses are more commonly identified in immunocompromised patients compared to immunocompetent patients. In some case studies, it has been shown that no pathogen is identified in the vast majority of non-ventilated HAP cases (56.3% in one case study [19]). This likely depends on the diagnostic methods used and how it is used (e.g. culture with or without dilution, different PCR assays etc).

Table 1.1 – List of WHO priority pathogens for new antibiotics [23]

Priority 1: Critical	Priority 2: High	Priority 3: Medium
Acinetobacter baumannii carbapenem-resistance	*Enterococcus faecium*, vancomycin-resistant	*Streptococcus pneumoniae*, penicillin-non-susceptible
Pseudonomas aeruginosa carbapenem-resistant	*Staphylococcus aureus*, methicillin-resistant, vancomycin-intermediate and resistant	*Haemophilus influenzae*, ampicillin-resistant
Enterobacteriaceae carbapenem-resistant, ESBL-producing	*Helicobacter pylori*, clarithromycin-resistant	*Shigella spp.*, fluoroquinolone-resistant
	Campylobacter spp., fluoroquinolone-resistant	
	Salmonellae, fluoroquinolone-resistant	
	Neisseria gonorrhoeae, cephalosporin-resistant, fluoroquinolone-resistant	

VAP is pneumonia that develops 48-72 hours after a patient has been intubated [18]. It is estimated to develop in 10-20% of patients after 48 hours of mechanical ventilation, and VAP patients are twice as likely to die compared to non-VAP patients [24]. Additionally, VAP causes longer stays in intensive care units (ICU) meaning additional hospital costs.

The same organisms are typically involved in the cause of VAP as non-ventilator HAP, including *P. aeruginosa, Enterobacteriaceae, S. aureus* and *A. baumannii* [25] [26] [27] [28]. Additionally, some studies suggest Late-onset VAP (VAP that develops later during

intubation, typically after 4 or 5 days) vs. Early-onset VAP may correlate with different aetiological agents [27], however, other studies suggest no difference [26]. A high prevalence of MDR pathogens is common. Resistant *P. aeruginosa, A. baumannii* and methicillin resistant *S. aureus* (MRSA) are commonly found in ICUs around the world [25]. Viruses typically do not cause VAP but some studies suggest they can be associated with VAP; viruses such as Herpes Simplex Virus (HSV) and cytomegalovirus have been shown to reactivate in ventilated patients and cause bronchopneumonitis or VAP [1]. Additionally, patients with high viral loads generally have poorer outcomes compared to those with low or no virus.

National Institute for Health and Care Excellence (NICE) guidelines suggest that antibiotics should be given to patients suspected of having HAP as soon as possible (within 4 hours) and a sample sent for microbiological testing at the same time. The first-line oral antibiotic is typically co-amoxiclav; alternatives based on the patient and local resistance data include doxycycline, cefalexin, trimethoprim with sulfamethoxazole (co-trimoxazole), and levofloxacin (or clarithromycin for children). If symptoms are severe or if there is a high chance of resistance, other options include piperacillin/tazobactam, second/third generation cephalosporins (ceftazidime, ceftriaxone, cefuroxime), and the carbapenem meropenem. If MRSA is suspected, vancomycin, teicoplanin or linezolid should be added on top of the first-choice antibiotic [30].

1.1.3 Antimicrobial resistance

Antimicrobial resistance (AMR) is a major public health concern [31]., The Organisation for Economic Co-operation and Development (OECD) states that, if no effective action is taken, resistance to important second-line antibiotics will be 72% higher in 2030 compared to 2005 [31]. AMR is already responsible for 33,000 deaths per annum in European Union/European Economic Area (EU/EEA) countries, costs 1.1 billion euros annum and contributes to 700,000 deaths globally [32]. The health burden of AMR is comparable to

that of influenza, tuberculosis (TB) and human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) combined in the EU/EEA. 40% of the burden is due to bacteria that are resistant to last-line antibiotics (used as a last resort treatment when bacteria are resistant to other antibiotics). When bacteria are resistant to last-line antibiotics such as carbapenems and colistin, treatment can be difficult or impossible. By 2050, 10 million deaths may be attributed to AMR every year, if no action is taken [33]. This would be even higher than cancer which is projected to kill 8.2 million per year by 2050. The cost to the global economy could be up to 100 trillion dollars [33]. These numbers do not fully capture the problem with AMR, as the inability to treat infections due to resistance will have significant knock-on effects in healthcare. Treatments that suppress immune systems such as chemotherapy will become much riskier as any resulting infection may be difficult or impossible to treat. Routine surgeries such as joint replacement and organ transplantation will also become riskier. The effect across the globe will be unbalanced, with Asia and Africa seeing higher mortalities.

1.1.4 Antimicrobial stewardship

AMR is associated with antibiotic use, whereas antibiotic stewardship practises help reduce the development of AMR by preventing use of antibiotics when they are not necessary [31]. Antibiotic stewardship is the responsible use of antibiotics – this principle goes further than just individual patient health but also has the goal of preserving antimicrobial effectiveness in the future and therefore safeguarding public health [16]. The aims are to: achieve optimal outcomes for the patient by using the correct antibiotics; reduce toxicity by, for example, deescalating if a toxic antibiotic is not needed; reduce the costs associated with administering antibiotics and importantly; reduce the selection pressures which enable the proliferation of resistant strains.

ICUs, where HAP is often treated and/or where VAP can develop, are a focus for antibiotic stewardship programmes as this is where a large proportion of antibiotics are

administered. Additionally, due to better access to diagnostics and direct treatment of patients, there is a greater opportunity to de-escalate when microbiology results are negative. Rapid identification of the cause of pneumonia can lead to correct treatment given at the first instance or escalation/de-escalation earlier than normal. However, it has been shown that even when microbiological results identify susceptible pathogens, antibiotic therapy is only narrowed from broad-spectrum antibiotics in 30-40% of cases [34]. The tests must be reliable, validated and must be comprehensive enough to ensure pathogens are not missed, otherwise treatment is unlikely to be de-escalated.

The O'Neill review recommends several interventions to reduce the burden of AMR. In the final report, this is broadly divided into reducing the demand for antimicrobials and increasing the supply of new antimicrobials [32]. To increase the supply, the report suggests a global innovation fund for early-stage research and development (R&D) and better incentives to promote investment in new and existing drugs. In the reduction of demand category, the report suggests improving global public awareness, improving sanitation, reducing unnecessary use of antimicrobials in agriculture, improving global surveillance, promoting new vaccines, promoting people working in infectious diseases, and promoting new and rapid diagnostics. Surveillance involves monitoring infectious diseases globally, their resistances and the use of antimicrobials. This is something that requires laboratory capacity as resistance is often confirmed in the laboratory. Reducing antibiotic use will require new and rapid diagnostics. This is because antibiotics are often not prescribed depending on a diagnosis but empirically (decision based on observation and experience rather than microbiological data). Acutely ill patients need to be treated as soon as possible if they have a potential infection, so prescribers cannot wait for a definitive diagnosis, this means that prescribers are required to treat empirically, and this leads to unnecessary and/or over/under-prescribing. The OECD is another body that recommends rapid diagnostic tests as an important intervention to tackle the issue [31].

In the US, it was found that out of 40 million patients who are given antibiotics for respiratory reasons, only 13 million needed them, while 27 million received them

unnecessarily [35]. Additionally, respiratory diseases were the most common cause for the prescription of antibiotics (41% of antibiotics), followed by skin/mucosal conditions and urinary tract infections (UTI). Broad-spectrum antibiotic use is much higher than narrow-spectrum antibiotics in the majority of EU/EEA countries [31] and in the US [35]. This is a problem as use of broad-range antibiotics can lead to the increased risk of *Clostridium difficile* infection [36] and poor antibiotic stewardship that can drive the increase of antibiotic resistance. Rapid diagnosis of the infectious agent and any resistance can reduce over-prescribing. For example, if the aetiology is known to be exclusively viral, then antibiotics should not be prescribed as they will be ineffective.

According to the O'Neill report on diagnostics, the perfect rapid diagnostic test would determine:

1) whether the infection is bacterial or viral
2) what type of bacteria, if it is bacterial
3) whether the causative bacteria are resistant to available antibiotics
4) whether the causative bacteria are known to be susceptible to existing drugs [37].

Determining susceptibility rather than resistance is more advantageous, as this gives the treating clinician more confidence about the treatment they can use. Additionally, diagnostic tests would need to be widely deployable, in terms of healthcare settings (e.g., hospitals and primary care), and also globally (both the developed and developing world). Rapid diagnostics need to be as cheap as possible (as antibiotics are generally cheaper than running such tests), need to answer the right questions, and need to be accurate and provide trustworthy results.

1.2 Current methods for diagnosis of LRTIs

An effective LRTI diagnostic assay should guide appropriate antimicrobial therapy decisions soon after clinical onset/presentation. As described, pneumonia can be caused by a variety of bacteria and viruses (and in rare cases fungi). In order to determine the

causative agent, clinical symptoms alone are insufficient, and a microbiological diagnosis is needed alongside antibiotic resistance/susceptibility testing for bacterial infections.

1.2.1 Routine culture

1.2.1.1 Pathogen identification

The current gold-standard for aetiological investigation of LRTIs is microbiological culture [38]. The UK Health Security Agency (UKHSA) provides guidelines for the investigation of respiratory specimens such as BALs, sputum and associated specimens using culture and microscopy in the UK, however, there is variation in how these are implemented across the country [39]. Specimens are ideally collected before antimicrobial therapy is administered and processed the same day (or stored in the fridge until processing). Mucoid samples such as sputum are treated with dithiothreitol (DTT) or N-acetyl cysteine for liquefaction (a commercial product called Sputasol is typically used which contains DTT). BALs on the other hand are not treated unless viscous, they are centrifuged to concentrate the cells in a pellet. Sputum and similar specimens (e.g., endotracheal tube aspirates) are diluted 1 in 1000 and 1 μL is used to inoculate an agar plate, giving a growth threshold of 1 x 10^6 colony forming units (CFU) per mL, whereas BALs are serially diluted and inoculated onto agar plates for a semi-quantitative culture. The recommended diagnostic threshold for BALs in standard samples is 1 x 10^4 CFU/mL. Where *Legionella* is suspected, or the clinical context requires other organisms such as *Burkholderia* to be considered (for example if the patient has cystic fibrosis), then supplemental cultures may be set up that are not diluted. With *Legionella*, it is recommended that plates are inoculated with 0.1 mL of undiluted homogenised sputum directly. The standard bacterial cultures are grown for 40-48 hours (5 days for fungi), with cultures being read every day, however, this can be as long as 6 weeks for some fungi (e.g., if dimorphic fungi are suspected) (Table 1.2).

Table 1.2 – Recommended growth media and times for target organisms in sputum/BAL samples [39]

Organism	Standard growth media	Incubation time	Culture read
Haemophilus influenzae	Chocolate agar + Bacitracin disc	40-48 hours	Daily
Moraxella catarrhalis	Chocolate agar + Bacitracin disc	40-48 hours	Daily
Streptococcus pneumoniae	Chocolate agar + Bacitracin disc	40-48 hours	Daily
Staphylococcus aureus	Chocolate agar + Bacitracin disc (and Mannitol Salt / Chromogenic Agar in some clinical conditions)	40-48 hours	Daily
Enterobacteriaceae	CLED/MacConkey agar	40-48 hours	Daily
Pseudomonads	CLED/MacConkey agar	40-48 hours	Daily
Fungi	Sabouraud agar	5 days (or up to 6 weeks if dimorphic fungi suspected)	≥40 hours
Burkholderia cepacia complex	*Burkholderia cepacia* selective agar	5 days	Daily
Legionella species	Legionella selective agar	10 days	At 3, 7 and 10 days

If *Mycobacteria* or parasites are suspected, these standard methods will not be sufficient and other culture or molecular methods are required for investigation [40,41]. For viruses and atypical bacteria such as *M. pneumoniae* and *Chlamydophila* species, PCR screening is the gold standard [42], but, not all samples undergo these tests.

As samples can be taken after antibiotic treatment and since not all organisms are tested by default, culture has poor sensitivity (a pathogen is not reported in up to 75% of

pneumonia cases) [43]. This poor sensitivity, as well as the high turnaround time of culture (Table 1.2) means that patients are treated empirically with broad-spectrum antibiotics, leading to inappropriate treatment of pneumonia as discussed above [44].

1.2.1.2 Antibiotic susceptibility testing

Antibiotic susceptibility testing (AST) is performed to determine whether pathogens isolated from clinical samples are resistant or susceptible to selected antimicrobial therapies [45]. Resistance can be based on different mechanisms including mutations, acquired resistance genes, efflux and permeability. Some organisms are intrinsically resistant to particular antibiotics (e.g. *P. aeruginosa* to many β-lactam antibiotics) [46] and therefore only relevant antibiotics need to be tested.

Phenotypic AST can be performed in different ways. Agar and broth dilution methods can be used to determine the minimum inhibitory concentration (MIC) [45] by culturing bacteria in a select concentration of antibiotics or at single concentrations (e.g. at the breakpoint) and detecting presence/absence of growth. Agar dilution has the benefit of making it easier to detect contamination over microbroth dilutions whereas microbroth dilutions are able to test multiple antimicrobial drugs and concentrations using 96-well plates. These methods can be time-consuming, typically taking 24-48 hours [45].

The agar disk diffusion method is the standard approach used routinely by microbiology laboratories. This method is less resource-intensive than the other methods and is highly standardised [45]. In this approach, an agar plate is inoculated with a bacterial isolate, and filter paper disks that contain specific concentrations of antimicrobials are placed on the surface. The antimicrobials diffuse into the agar surrounding the disk with the concentration decreasing relative to the distance from the disk. Bacteria that are inoculated onto the plate grow in areas where it is not inhibited and do not grow in areas where it is inhibited by the antimicrobial agent (around the disks) creating a visible lawn of bacteria with areas of no growth. The areas of no growth are the zones of inhibition (ZOI)

and the diameter can be measured to determine whether an organism is susceptible or resistant.

The European Committee on Antimicrobial Susceptibility Testing (EUCAST) recommends the disk diffusion method and provides a standardised set of guidelines and clinical breakpoints to determine resistance or susceptibility [47]. The UK follows the EUCAST standard [48], which defines everything from the media used (typically Mueller Hinton), to the storage of plates, preparation of the inoculum, inoculation of the plates, application of the antimicrobial disks, incubation of plates (time and conditions) and importantly the measurement and interpretation of the results. Globally, 3 organisations set the breakpoints and interpretations, the Center for Drug Evaluation and Research (CDER), the Clinical and Laboratory Standards Institute (CLSI) and EUCAST [49]. These clinical breakpoints are reviewed and updated annually based on new information [50].

Advantages of the disk diffusion approach and why it has become the gold standard is that it is easy to perform with reproducible results, it is inexpensive, both in terms of equipment and reagents, and it provides simple categorical results [45]. Drawbacks are that it has poor performance with slow-growing and fastidious organisms, and despite standardisation, can still be prone to variability from operator handling and interpretation [51]. However, there are also now instruments that aim to reduce some of the interpretation error by taking pictures and analysing the ZOI in an automated fashion (e.g. Accuzone, Biomic, and Sirscan) [51]. One of the biggest drawbacks of the disk diffusion method is that it is culture-based and therefore takes too long. After initial culture of the primary clinical sample, the isolate has to be identified and cultured again using the disk diffusion method, meaning that results for ASTs can take ≥2 days after the sample is taken [52].

1.2.2 Rapid pathogen and antimicrobial resistance detection

More rapid methods are available for the detection of pathogens and AMR. Nucleic acid amplification tests (NAAT) such as PCR and loop-mediated isothermal amplification (LAMP) are fast methods for detecting the presence/absence of pathogens and resistance genes [51] and have evolved significantly from their early days of manual single target assays [53]. PCR is the gold standard for detecting respiratory viruses, as recommended by UKHSA [42], but typically target a relatively small panel of viruses such as Influenza A and B, RSV, Adenovirus and Parainfluenza.

Commercial sample-to-result NAATs have been developed for the point-of-care identification (ID) of influenza that take only about 20 minutes, e.g. the Alere I and cobas Liat Influenza A and B tests, but they are limited to a single analyte [54]. Broader multiplex sample-to-result PCR devices have also been developed, such as the Unyvero (Curetis) and Biofire Filmarray (BioMerieux) [43]. These have cartridges that are tailored for specific diseases and contain multiple targets for relevant organisms and resistance genes. The FilmArray has panels for upper respiratory, bloodstream, gastrointestinal, and joint infections, as well as meningitis/encephalitis and pneumonia. The pneumonia panel has 33 targets, with 15 bacteria, 8 viruses, 3 atypical bacteria and 7 AMR genes (Table 1.3) [55]. These tests have rapid turnaround times (~1 hour) and minimal manual handling as they perform nucleic acid extraction, purification and multiplex PCR and detection on board followed by automated analysis.

Table 1.3 - Full list of targets of the Biofire Filmarray Pneumonia panel

Bacteria	Viruses
Acinetobacteria calcoaceticus-baumannii complex	Adenovirus
Enterobacter cloacae complex	Coronavirus
Escherichia coli	Human metapneumovirus
Haemophilus influenzae	Human Rhinovirus
Klebsiella aerogenes	Influenza A
Klebsiella oxytoca	Influenza B
Klebsiella pneumoniae group	Parainfluenza virus
Moraxella catarrhalis	Respiratory Syncytial virus
Proteus spp.	**Antimicrobial resistance genes**
Pseudomonas aeruginosa	mecA/C and MREJ
Serratia marcescens	KPC
Staphylococcus aureus	NDM
Streptococcus agalactiae	Oxa-48-like
Streptococcus pneumoniae	VIM
Streptococcus pyogenes	IMP
Atypical bacteria	CTX-M
Chlamydia pneumoniae	
Legionella pneumophila	
Mycoplasma pneumoniae	

These devices have been well-studied with many evaluations of both the Unyvero pneumonia panel [56–58], and FilmArray Pneumonia panel [59–61]. Results from these studies vary significantly due to differences in how performance is reported (concordance vs specificity and sensitivity) as well as differences in sample type and geographic location, leading to differences in the frequency of pathogens detected. This makes comparison of such tests from different studies difficult. One multicentre study [43] directly compared the two devices and demonstrated that both PCR tests identified significantly more pathogens from samples compared to routine microbiology. Individual organism sensitivities ranged from 91.7-100% for FilmArray and 50-100% for Unyvero, and specificity was 87.5-99.5% for FilmArray and 89.4-99% for Unyvero. Sample-to-result times were significantly lower than culture for both devices, FilmArray taking 1 hour and 15 minutes and Unyvero taking 5 hours. Overall, the study found that the FilmArray was more sensitive than the Unyvero and was chosen to be tested in a randomised control trial. These devices are useful, as a study has shown that it was possible to change initial

empirical treatment within 5-6 hours for 33 patients using Unyvero compared to 96 hours for culture [56]. While these panels are significantly faster and easier to use, a drawback is that they are limited by the target pathogens on the panel, and only a few key resistance genes are detected [51]. Additionally, even if a wider range of resistance targets were included, multiplex PCR-based tests directly from clinical samples cannot determine the source of resistance genes. This was exemplified in the multicentre study where the Unyvero detected *mecA*/*mecC* in 70 samples that did not contain *S. aureus* (likely from other *Staphylococci* in the sample) [43].

Other rapid ID and AST methods have been developed that are phenotypic based including the Accelerate Pheno. This is an automated system that uses fluorescent in-situ hybridisation probes to detect organisms commonly associated with the disease, then immobilises the cells to grow them in the presence of different antibiotics. The cells are monitored with automated microscopy and their growth rates analysed to predict MIC values [62]. This method does not require full overnight culturing as is the case with the disk diffusion method; pathogen detection results can be available in less than 2 hours with AST results in less than 7 hours. This test is currently only available for bloodstream infections. Other proof-of-concept ID and AST methods include using digital LAMP to measure phenotypic responses from bacteria in clinical samples exposed to antibiotics (with results in 30 minutes) [63], and using probes on electrochemical biosensors that are complementary to bacterial 16S ribosomal RNA (rRNA) to quantify rRNA levels after a brief culture period in the presence of antibiotics [64]. However, these approaches have yet to be applied to pneumonia. We wrote a review paper summarising emerging technologies for rapid diagnosis of infection, and AMR in Current Opinion in Microbiology [52] (Appendix 1).

1.3 Sequencing

Advances in sequencing technologies have paved the way for sequencing to characterise infections, using whole genome, targeted and metagenomic approaches.

Whole genome sequencing (WGS) is a widely used tool for AMR detection, epidemiology and outbreak control. It is regularly used to identify and study outbreaks [65,66], and to study the origin and evolution of important clinical strains [67]. It has also been used to identify new AMR mechanisms and transmission routes [68,69]. Whilst WGS cannot be used to rapidly diagnose infection due to the need for culturing, it does provide comprehensive genomic information and has become an important tool for public health. Studies have been performed that use bait captures for pathogens to enable sequencing of whole genomes of one organism directly from primary clinical samples, without the need for culture, to determine resistance – an approach that is particularly beneficial for slow-growing pathogens such as *Mycobacterium tuberculosis* [70].

Targeted sequencing methods have been used to identify infection-causing agents by amplifying specific regions such as 16S or internal transcribed spacer (ITS) prior to sequencing [71]. These methods have the potential to be fast, as they do not require culture. 16S rRNA gene sequencing can be particularly useful in identifying fastidious bacteria, and as with multiplex PCR panels, detect and identify more pathogens than culture [72]. Targeted panels such as the BacCapSeq scheme use probes to target virulence and resistance genes directly from samples before sequencing to identify bacterial pathogens [73]. This method has also been applied to viruses (VirCapSeq) [74] and tick-borne pathogens (TBCCapSeq) [75]. It has also been applied specifically to study the resistome, targeting 78,600 non-redundant genes with ResCap [76]. In the case of TB, where resistance is mostly due to chromosomal mutations in conserved genes, it is possible to design PCR-sequencing panels to detect the presence of *Mycobacteria* and predict TB resistance [77].

Targeted sequencing methods can be fast, highly sensitive, and robust, however, by nature of being targeted, cannot provide information on all potential pathogens in the sample. Even 16S rRNA gene sequencing, which is pan-bacterial, does not cover fungi, and even with the inclusion of 18S/ITS, viruses are still missed. Additionally, these targeted methods do not provide the full breadth of information that sequencing more of the genome can provide.

A more recent application of sequencing has been clinical metagenomics (CMg) which is the sequencing of all the genetic material (DNA and/or RNA) in clinical samples to characterise the microorganisms present [78]. This is a rapidly expanding field that shows promise due to its target-free approach [78] and is covered in more detail in Section 1.4.

1.3.1 Sequencing technologies

Sanger sequencing was the dominant technology and gold standard for DNA sequencing for decades, however, the early 2000s brought new sequencing technologies (commonly referred to as Next Generation Sequencing - NGS) that allowed significantly higher throughput sequencing [79] and kickstarted the field of CMg [78]. 454 Life Sciences/Roche pyrosequencing, Solexa/Illumina sequencing, ABI SOLiD sequencing and Helicos single-molecule sequencing were the early NGS methods on the market that unlocked new capabilities for targeted, metagenomic and transcriptomic sequencing [79]. NGS techniques such as 454 sequencing quickly started being used in CMg applications to identify new pathogens (while traditional methods were failing to identify the cause of disease) [80]. Since then, the sequencing field has evolved rapidly. The turn of the 2010s brought Ion Torrent semiconductor sequencing (Thermofisher), and later long-read sequencing technologies by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) [81]. Short-read sequencers such as Illumina and Ion Torrent sequencers produce reads up to 600 bases, while long-read technologies can generate reads above 10 kb [82], up to megabases in the case of ONT. Short and long-read NGS

technologies are sometimes referred to as second and third generation sequencing respectively [83], however, as newer short-read technologies (e.g. by MGI Tech) are released [84], these naming conventions become less meaningful.

Illumina is currently the market leader in sequencing [84], offering devices that are cost-effective, high-throughput and accurate and also well-supported by pipelines and analysis tools [82]. However, Illumina sequencing has been limited by short reads that complicate reconstructing genomes which make *de novo* assembly, haplotype phasing and structural variant identification difficult, as well as stripping native DNA molecules of epigenetic modifications due to amplification of the template [82]. Additionally, short-read sequencers like Illumina and Ion Torrent have generally been limited by longer running times and the need for batching large numbers of samples to make sequencing cost effective [85].

PacBio and ONT are currently the only two producers of commercial long-read sequencing devices. PacBio sequencing uses a method called 'Single Molecule, Real-Time' (SMRT) sequencing, where a polymerase replicates circular DNA using fluorescently labelled nucleotides which releases unique emissions when incorporated [82]. The circular nature of the template means that each DNA strand can be sequenced multiple times by the polymerase (each sequencing of a strand is called a 'pass'), and provides highly accurate reads, reportedly up to 99.8%, (also known as HiFi sequencing) [86]. Nanopore sequencing uses protein pores that DNA or RNA molecules pass through, disrupting the applied ion current, giving a characteristic signal depending on the nucleotide. This produces a squiggle plot which can be decoded (basecalled) to give the genetic sequence. Nanopore sequencing currently has lower accuracies than other technologies, however recent chemistries ('Kit 14') are capable of raw read accuracies of 99.3% [87]. Both SMRT and nanopore sequencing have the capability of detecting epigenetic modifications directly, as they can both sequence molecules without the need for PCR amplification prior to sequencing [82,88].

When it comes to diagnostic purposes, the choice of sequencing technology is important, especially in terms of cost, time, footprint, usability, and analysis. Illumina technology has

been used in a wide range of viral [89–91] and bacterial [92,93] CMg applications. However, Illumina sequencing is not ideal to implement for rapid sequencing due to the long sequencing turnaround times, (over 24 hours for the NextSeq devices and between 5-24 hours for the MiniSeq) and the need for batching large numbers of samples [85]. Similarly, PacBio Sequel II has a runtime of 10 to 30 hours per SMRT-cell [94]. In contrast, due to the real-time data output of Nanopore sequencers, runtimes can be as short as minutes depending on the application [81]. Additionally, nanopore sequencing has rapid library preparation kits that take 10 minutes, compared to Illumina and PacBio which can take hours [81], meaning that nanopore sequencing is currently the sequencing technology of choice for speed. A study demonstrated that a CMg workflow could detect Ebola, chikungunya and hepatitis C viruses from blood samples within 10 minutes of sequencing using nanopore, and 6 hours from sample-to-result [95], whereas the Illumina version of the workflow took over 24 hours. The fast nature of Nanopore sequencing has also been demonstrated in other studies too [96][97][98], all achieving results in less than a day. The relatively low capital cost [99], and portability of ONT's MinION device [85] also make it a good candidate for clinical implementation [100].

One drawback in the past that has hindered clinical implementation has been the rapid evolution of the technology (meaning frequent updates), which can cause issues for clinical validation that requires standardisation and locked-down protocols [78]. However, the recent release of locked-down 'Q-Line' products that are produced and supported in the long-term will help with clinical adoption [101].

1.4 Clinical metagenomics

Unlike WGS, which sequences the whole genome of one organism, or targeted sequencing, which sequences specific pre-determined regions/organisms, metagenomics is a non-targeted approach that sequences all the DNA and/or RNA in a sample. Metagenomics has been used for environmental [102] and microbiome [103] studies, and

the potential to detect all microbes directly from primary clinical samples makes it a compelling tool for clinical use. The advent of next generation sequencing, which led to a significant increase in throughput and reductions in cost, kickstarted the field of CMg [78].

Studies have demonstrated the potential of CMg in bone and joint infections [93], CAP [104], HAP and VAP [92,105], meningitis and encephalitis [106], bloodstream infections [107], UTIs [108,109], and in faecal samples [110]. CMg has also been applied specifically for viruses, for the detection of influenza, ebola, chikungunya and other viruses directly from sample [89,91,111]. Due to the difficulty in culturing viruses, molecular methods are critical and identification of unusual or new viruses often must be performed by sequencing. There have been examples of CMg identifying organisms when the aetiology of disease was unknown, as in the case of an astrovirus infection [90], which is an unusual encephalitis pathogen that would not have been detected without metagenomics as it is not routinely tested for in this context. It has also been used to confirm infections that have led to the altered treatment and cure of patients, such as a rare *Leishmania* co-infection in a patient with HIV [112].

Metagenomics has also been used to study AMR; investigating resistance determinants, mechanisms and mobility [113,114]. Recently, it has also been used in a clinical context for detecting resistance genes and mutations to predict AMR. One study used CMg to detect beta-lactam resistance in samples from HAP/VAP patients [115], showcasing examples where both escalation and de-escalation would be recommended based on CMg results. Another study attempted to detect mutational resistance determinants in *S. pneumoniae* from nasopharyngeal samples, for example, showing expected mutations in *folA* and *folP* genes in 87/98 co-trimoxazole resistant isolates [116]. Inferring phenotypes from CMg is difficult and is limited by a wide range of factors. One limitation is resistance gene databases; none of the existing databases are exhaustive and there is no consensus on which should be used for CMg [117]. There is also the challenge of not knowing enough about the genetic basis of resistance in some organisms or not being able to easily decipher phenotype from DNA data alone due to expression-mediated

resistance, for example in *P. aeruginosa*, which has multiple efflux pumps that can be overexpressed to confer resistance [117]. Transcriptomics in conjunction with machine learning has been used to improve phenotypic prediction in this area [118]. One of the biggest challenges in using CMg for AMR is matching mobile genetic elements with the host organism, as resistance determinants may be sequenced from background organisms that are not the cause of infection. Hi-C is a method that ligates DNA sequences based on proximity and can therefore link plasmids to the host chromosome [119], however, is currently complex and adds a lengthy step that would significantly increase time-to-result and cost [117].

Given the complexity of detecting resistance, one potential method is predicting resistance by lineage association [120]. Using a reference database of known isolates with resistances, it is possible to use metagenomic data to infer the resistance or susceptibility of a sequenced pathogen by determining their lineage/closest lineage. This was demonstrated using the tool Resistance-Associated Sequence Elements (RASE) for *S. pneumoniae* from CMg data and was very rapid (less than 10 minutes of analysis) [120]. This method works best when databases are comprehensive and may require supplementing with local data. However, there needs to be enough within-species diversity to accurately call the nearest lineage and there needs to be a strong association between resistance and lineage to make a reliable call on phenotype.

A review in 2019 found that sensitivity and specificity for agnostic CMg methods are, on average, 88% and 86% respectively for pathogen ID [78]. Genotypic AMR prediction agreement with phenotypic results are typically 83% on average, with false susceptible prediction rates being 9% and false resistant prediction rates being 1% on average [78]. The median time from sample-to-result of methods reviewed was 23.5 hours, with the shortest time being 7 hours; sequencing time on average accounted for 10 hours of the total time [78]. The cost per sample ranged from $128 to $685. The review also found that only 61% of studies made use of negative controls, which are necessary to determine if pathogen detections are real of if they are contamination.

A drawback of current CMg methods is their complexity and requirement of highly trained individuals to carry out the complex manual steps which can be laborious and prone to error. Additionally, since CMg is untargeted, any nucleic acid that is introduced during the process has the potential to be sequenced and cause false-positive results, therefore, extra care has to be taken to avoid contamination with these methods, and/or must be accounted for in the analysis [121]. Cost is also an important consideration, as sequencing methods can be expensive. Batching samples can reduce the cost, however, this limits the flexibility of running samples when needed and can increase time-to-result. In addition, the skilled staff required also adds to the cost [78].

Despite some of the current limitations, there are CMg methods that have been validated for testing patient samples such as cerebrospinal fluid [106], blood plasma [107] and respiratory specimens (for viruses) [122]. None of these methods can provide results in the same working day, therefore don't have a significant turnaround time advantage over culture and have limited clinical utility (typically reserved as last resort tests when none of the standard investigations have provided any useful results). This is due to the method or choice of sequencing technology, but even if the technology is not the issue, long sequencing times may be required due to the high amount of background human DNA. One of the main challenges in clinical metagenomics is that human DNA dominates clinical specimens [78]. This can make detection of microorganisms using CMg challenging, especially if the pathogen is not typically present at high concentrations, for example, if the patient is in the early stages of infection. In sputum samples, white blood cells can be present at around 4.1×10^6 cells per gram of sputum [123], and this is usually elevated during an infection. BALs, which are more dilute, have been shown to contain 2.8×10^6 cells per mL during bacterial pneumonia and 3×10^5 during viral pneumonia [124]. A human cell contains approximately 6.5 pg of DNA [125], whereas a 4.6 Mb genome *E. coli* cell contains approximately 5 fg of DNA, meaning a factor of >1000X difference. For viruses this difference can be even larger due to their small genome sizes (Figure 1.2). Influenza B for example, which often causes CAP, has a single-stranded RNA (ssRNA)

genome of 14.6 Kb [126], a factor of >100,000X difference. Viral loads can range anywhere from 1 x 10^1 to 1 x 10^9 copies/mL, (average of 1 x 10^4 to 1 x 10^6) [127,128], which makes it extremely challenging to sequence viruses using CMg. These ratios explain why less than 1% of reads from clinical samples are typically from the infectious agent [78]. Tissues which have a higher proportion of cells to body fluids are even more challenging and result in even lower sensitivity for CMg. Therefore, depletion of human nucleic acid or alternatively, enrichment of microbial DNA/RNA is an important step in CMg [129]. Methods that apply human depletion in sputum samples have shown a significant increase in the ratio of microbial : human reads, which have led to reductions in sequencing times required to reliable detect pathogens and resistance genes/SNPs [97].

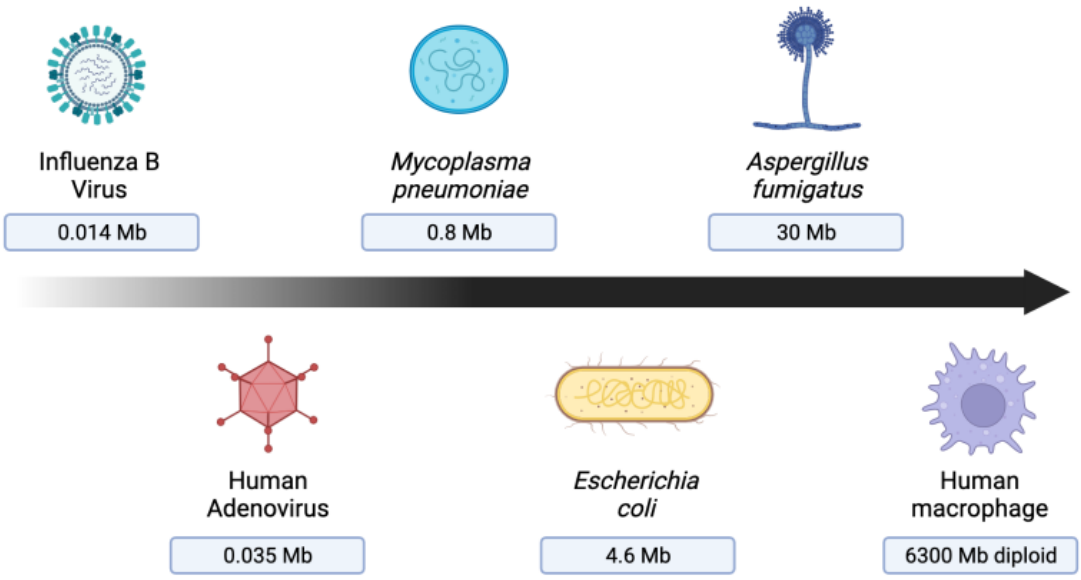


Figure 1.2 – Comparison of genome sizes (in megabases) of some respiratory pathogens compared to a human macrophage

1.4.1 Host depletion

A method of improving the sensitivity of CMg is to take a partially targeted approach, where specific organisms or genes are enriched over the background. Metagenomic sequencing with spiked primer enrichment (MSSPE) introduces primers for select viruses in addition to the standard random primers, which gives approximately 10-fold enrichment

of the viruses, while still being able to sequence non-enriched viruses [130]. This was demonstrated for a panel of 14 important viruses, including Zika, Ebola, Chikungunya, among others. However, this increases bias towards selected targets. To try and keep metagenomics untargeted, a method that depletes host DNA or universally enriches all pathogens is desirable.

There are broadly two approaches to improve the ratio of microbial DNA relative to the host; methods that differentiate:

- on a cellular level pre-extraction
- using the nucleic acid post-extraction [131].

Pre-extraction approaches act by taking advantage of the different cellular properties of host and microbial cells. One approach is to differentiate based on cell size, as host cells are larger than microbial cells, however, attempts to filter human cells while allowing microbial cells to pass through have been unsuccessful, at least for bacteria [131]. Viruses on the other hand, which are even smaller have successfully been enriched by filtering out host and bacterial cells using 0.45 μM filters [132]. Alternatively, centrifugation is also an option, and is commonly used to deplete host cells by differential centrifugation and collecting viruses and/or bacteria in the supernatant [109,111].

For bacteria, pre-extraction methods typically focus on differentially lysing host cells while leaving microbial cells intact, an approach that uses differences in properties of the human cell membrane compared to bacteria and fungi, which have cell walls protecting their membranes, and the protein capsid of some viruses [133]. Differential lysis uses chemicals or enzymes that breakdown phospholipid bilayers, typically followed by degrading the exposed nucleic acid, for example by nuclease treatment. Different detergents have been used for differential host lysis: Triton-X, Tween 20, Chaps cell extract buffer [134]. One of the most widely used host depletion reagents is saponin [97,134,135], which shows efficient lysis of human cells with minimal effect on bacterial cells. Saponin is a non-ionic surfactant that has a hydrophobic steroid core with a high

affinity for cholesterol which is abundant in plasma membranes [136]. It works by forming complexes with the cholesterol which leads to the formation of pores in the plasma membrane and leads to a loss of membrane integrity. Other examples of chemicals that permeabilise plasma membranes using the same mechanism are digitonin (a specific saponin) and filipin [136]. Due to the lower cholesterol content of intracellular organelles such as the endoplasmic reticulum and mitochondria, these chemicals can leave intracellular organelles largely intact [136], however, saponin is better at permeabilising intracellular organelles than digitonin [137]. Alternatively, cholesterol-dependent cytolysins can also be used for differential cell lysis. These are proteins produced by bacteria such as *S. pyogenes, Listeria monocyotgenes* and *Clostridium perfringens* that also form pores in the plasma membrane leading to cell lysis and/or programmed cell death [138].

Following host cell lysis, the most common approach is to then digest the exposed nucleic acid with a nuclease. DNase I is an option, however, increasingly Benzonase has been used due to its activity in a wider range of conditions and ability to degrade DNA into very short fragments [133]. HL-SAN is another option that has been used, which is a highly salt resistant nuclease [97]. Alternatively, instead of degrading the DNA, it is also possible to treat the DNA with an intercalator such as propidium monoazide (PMA), which is cell membrane impermeable and therefore only intercalates exposed DNA [131]. When treated and exposed to visible light, DNA covalently bonds to the PMA, rendering it unusable for downstream applications such as PCR and sequencing. An important prerequisite of differential lysis methods is that the microbial cells have to be intact, as any extracellular DNA will be lost. A potential downside of this is if a patient has been treated with antibiotics prior to the sample being taken or if the sample is not fresh (i.e. old or frozen), the microbial cells may not be intact and may be lost. Additionally, even without these factors, there is evidence that bacteria with different cell wall rigidities may be affected differently by cell lysis methods [139], particularly those with no cell walls like *M. pneumoniae*.

In contrast to cellular methods, post-extraction methods take advantage of differences in the properties of the DNA or the DNA sequence itself. Methylation density is different between human genomic DNA and microbial DNA and can be used to capture and remove human DNA, leaving behind microbial DNA. Methyl CpG binding domain proteins fused to a human antibody can be incubated with the DNA along with paramagnetic beads that bind to the antibody and then removed using a magnet [140]. The reverse can also be applied, where proteins that have affinity to non-CpG methylated DNA are used to capture and enrich the microbial DNA instead, which has been shown to reduce host DNA [141]. Use of restriction endonucleases to differentially bind methylated or non-methylated DNA have also shown promise but are not widely used currently [142].

Commercial kits are available that take advantage of both principles; differential lysis and post-extraction methylation (Table 1.4), however, it has been shown that differential lysis kits perform better at depleting host DNA than methylation-based depletion [143].

Table 1.4 – Commonly used commercial host depletion/microbial enrichment kits [133]

Principle	Kit	Manufacturer	Method Time (min)	Cost per sample ($)
Differential cell lysis	QIAmp DNA Microbiome	Qiagen	160	13
	MolYsis Complete/Ultra-Deep Microbiome Prep	Molzym	120	11
	HostZERO Mucrobial DNA kit	Zymo	30	10
Methylated host DNA capture	NEBNext Microbiome DNA Enrichment	New England Biolabs	30	39

An alternative post-extraction approach is to use the DNA sequence to deplete host DNA. For example, Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Cas9 technology has been utilised to create libraries of guide RNAs to target and degrade host DNA based on the sequence, leaving non-host DNA intact [144]. Even further

downstream, it is possible to selectively choose strands of DNA to sequence, which can be used to 'enrich' for microbial reads during sequencing. This is possible with nanopore sequencing owing to the real-time analysis of DNA strands as they translocate through the pores (adaptive sampling). This means that DNA strands can be sequenced or rejected out of pores if they are unwanted [145]. This has been shown to increase microbial depth of coverage by 1.7-fold. Decisions on rejecting reads are made around 400 bp [145], meaning that reads have to be sufficiently long enough for this method to be viable. While the amount of enrichment may not be sufficient on its own, it can be used in addition to other methods to further improve enrichment.

Even though there are some disadvantages to the current methods available, host depletion has been successfully used in a wide-range of studies with difficult samples such as cerebrospinal fluid [146], synovial fluid [147], blood [148], and tissue [149,150]. For respiratory samples such as sputum, BALs and ETAs, the saponin-based differential lysis with heat-labile Salt Active Nuclease (HL-SAN) has been demonstrated to be successful in multiple studies [97,115,151].

1.4.2 Sample preparation and automation

An important consideration when using CMg, is how to extract and purify nucleic acid from clinical samples. An inefficient extraction method for example, can not only lead to bias within the sample but can result in false-negative calls. Respiratory samples such as sputum can be particularly difficult, as they are typically comprised of thick and complex matrices [152]. The first step in extracting from sputum is usually to pre-treat samples with a homogenising agent such as DTT or N-acetyl-L-cysteine which has shown to improve the detection of pathogens from sputum samples [153].

Another important consideration in the extraction process is the method used to lyse the microbes. There are many different ways in which microbial cells can be lysed: thermal, alkaline, detergent, enzymatic, mechanical or electrochemical [154]. However, not all of

these methods are applicable to all cell types. For example, Gram-positive bacteria such as *S. aureus* have thick peptidoglycan cell walls which make them harder to lyse than Gram-negative bacteria, and therefore thermal or detergent lysis may not be sufficient. Specific measures have to be taken to efficiently lyse these cells, such as pre-treatment with enzymes like lysostaphin and lysozyme [155]. An alternative approach for hard-to-lyse microbes is mechanical disruption such as bead-beating. A study has shown that bead-beating may not be as efficient at lysing bacterial cells in pure cultures compared to enzymatic lysis (even for Gram-positive bacteria), but in complex clinical samples types, performed equivalent to chemical lysis [156]. For some microbes, such as *Bacillus* spores and *Mycobacteria*, mechanical disruption is unavoidable, as these organisms have thick multilayer structures or waxy cell walls which can be highly resistant to chemical treatments [157]. Therefore, more vigorous methods such as sonication and beadbeating are common. This typically involves the use of laboratory benchtop beadbeaters, however, miniaturised and disposable beadbeating devices have also been used [157,158]. Since beadbeating is an efficient method of lysis for a wide range of cells, including those difficult-to-lyse, this makes it a good choice for CMg. However, the choice of bead diameter, density, and speed of beadbeating is also important. Downsides of beadbeating include fragmentation of nucleic acids, leading to shorter reads and degradation of less stable products such as RNA [154].

The sensitive and untargeted nature of metagenomics makes it very susceptible to contamination, especially in low biomass samples, so the choice of extraction kit/method is important for reducing contamination [159]. Sources of contamination can include the skin, the air, laboratory/clinical equipment, cross-contamination between samples, or the laboratory reagents and kits themselves [160]. Contamination from extraction kits have been well documented [121,161,162]. Use of automated extraction robots have the potential to reduce contamination caused by manual errors but may still introduce contaminants through the kit reagents. Automated extractions can also reduce hands-on time and decrease variability. It has been shown that automated methods are just as

efficient as manual extractions [163] and there are numerous automated DNA extraction machines available commercially that can perform extractions in as little as 6 minutes or as cheap as $2 per sample [164].

A study comparing multiple extractions methods found that a manual phenol:chloroform extraction method was the only approach that introduced contamination in the negative control, whereas an automated method was clean [165]. To minimise contamination, in addition to kit choices, additional steps can be taken, such as treating kits to reduce contaminating DNA and processing negative controls, which can be used to identify contamination during the analysis [121].

In addition to extraction, there are further steps that can be automated such as PCR and library preparation. Liquid-handling robots are capable of streamlining these processes and freeing up staff time [166]. These devices are divided into different categories, ranging from Tier 4 to Tier 1, depending on sophistication [167]. Some robots such as the OT-2 (Opentron) can be relatively cheap ($5000) but are basic automated pipetting devices, whereas more sophisticated devices such as the Microlab STAR (Hamilton Robotics) and Biomek i5 (Beckman Coulter) have sensors that can detect liquid levels and blockages, but cost >$120,000 [167]. Higher end robots also have washing and ultraviolent (UV) light modules which can help to reduce contamination. If PCR amplification is used in the CMg pipeline, at least two robots may be preferable, one for pre-PCR and one for post-PCR, to reduce the risk of amplicon cross-contamination [78], which can mean that end-to-end automation can be costly and have a high laboratory footprint. This is an option for high-throughput diagnostic laboratories and sequencing centres, however, is not feasible in many other settings. The development of smaller microfluidic robots such as the VolTRAX (ONT) which uses electrowetting to move small quantities of liquids around could reduce cost and footprint and potentially bring CMg closer to point-of-care settings [166].

1.4.3 Bioinformatic analysis

1.4.3.1 Metagenomic analysis

One of the big questions in metagenomics is how best to infer the composition/relative abundance of the microbial community and how to decipher contamination from microbes present in the sample [168]. Historically, the early method for metagenomic classification was using Basic Local Alignment Search Tool (BLAST) to compare reads to sequences in GenBank databases (either comparing nucleotides or translated amino acid sequences). Tools such as MEGAN were designed for this task [169], which takes BLAST results and computes and orders taxonomic content. However, due to the size of current microbial databases and sequencing outputs, this is computationally intensive and slow [170]. This led to the development of specific tools capable of dealing with large metagenomic datasets and databases, some of which are summarise in Table 1.5)=.

Table 1.5 – Commonly used metagenomic classification tools

Classifier	Approach	Reference
Kaiju	K-mer	[168]
CLARK	K-mer	[171]
DUDes	Alignment to custom reference type (genomes, genes, proteins)	[172]
DIAMOND	Alignment to protein reference database	[173]
GOTTCHA	Alignment to unique genome signatures	[174]
metaCache	K-mer	[175]
metaPhlAn2	Alignment to marker genes	[176]
MetaPhyler	Alignment to marker genes	[177]
MetaMaps	Alignment to genomes in a RefSeq database	[178]
mOTUs2	Alignment to market genes	[179]
Centrifuge	K-mer	[180]
Kraken 2	K-mer	[181] [182]
LMAT	K-mer	[183]

DIAMOND is an alternative to BLASTx that is up to 20,000 times faster due to using an index-based structure [173] and along with MEGAN can be used for taxonomic binning by aligning to a protein reference database [184]. Another approach to speeding up analysis is by focusing on specific regions of the genome. Classifiers such as metaPhlAn2 [176]

and mOTUs [179] align to a pre-determined selection of marker genes. DUDes allows the user to select a custom type of database [172] and GOTTCHA maps to the non-redundant unique fraction of reference genomes [174]. Classifiers that map to specific regions require standard alignment tools such as BWA, Bowtie or Minimap2 as part of the process.

'K-mer' based approaches have also become popular. This is where reference genomes are pre-processed by breaking them into all possible substrings of a fixed length, k, and storing them in an index for fast lookup [168]. K-mers in the metagenomic dataset can then be searched in the index and classified based on matches. K-mer based classifiers can use different approaches giving different results, for example Clark [171] only uses discriminative k-mers between reference genomes at a particular rank, and it can only give results at one taxonomic level whereas Kraken 2 can assign reads at different taxonomic levels [182]. Kaiju uses a k-mer approach on a protein-level, which can reduce the number of unclassified reads compared to DNA-level classifiers [168]. A number of benchmarking studies have been performed on the various tools and Kraken 2 is often found to be a fast and accurate option [185,186].

Another approach is to perform assembly-based analysis. It has been shown that assembly methods perform well (precision and recall rates) compared to non-assembly based approaches, particularly when starting with short reads [187]. However, assembly-based approaches can be computationally demanding and can require a large amount of memory. [188] Assembly-free classification is generally faster.

When a metagenomic classifier provides an estimation of relative taxonomic abundance, it is called a taxonomic profiler [186]. However, depending on how the tool works, particularly depending on the reference database used, it may be reporting sequence abundance as opposed to taxonomic abundance. Profilers which compare reads to comprehensive metagenome or protein databases such as Kraken and Kaiju can underestimate the proportion of microbes with smaller genomes/fewer proteins [189]. Whereas profilers which use single-copy markers, such as MetaPhlAn2, are significantly

better at estimating taxonomic abundance. It is possible to convert sequence abundance to taxonomic abundance using genome size and ploidy calculations, however, this is challenging and often does not work very well, therefore care has to be taken when interpreting microbial abundance from metagenomic data [189].

Most databases used for metagenomic analysis use genomes and draft genomes from GenBank, a repository for global genome sequence data [170]. GenBank is part of an international collaboration of nucleotide sequence databases which also includes the European Nucleotide Archive (ENA) and DNA Data Bank of Japan (DDBJ) [190]. A challenge of analysing metagenomic data is the sampling bias in the distribution of reference genomes. Some organisms that are commonly studied or isolated are overrepresented while other organisms, for example those that are not possible to culture easily, are underrepresented [168]. Another issue with public genomes in GenBank, particularly draft genomes is that they may be incorrectly labelled or contaminated with reads from other organisms (e.g. lab contaminants) as they are user submitted and not curated [170]. RefSeq is a curated database of non-redundant GenBank sequences that aims to fix some of these issues. The latest release (02 May 2022), contained entries from 119,373 organisms (genome, transcript, protein sequence) [191]. There are selected reference/representative genomes for 4,076 prokaryotes in RefSeq (3,837 bacteria and 239 Archaea), but only 22 for fungi (Table 1.6). Classifiers such as Kraken 2 use RefSeq bacterial, archaeal and viral sequences by default [192], however, there are a series of indexed RefSeq databases of varying sizes that also include protozoa and fungi [193].

Table 1.6 – RefSeq statistics as of June 2022. (Reference and representative genomes are single genomes representing a species) [194]

	Number of species with RefSeq entries (genome, transcript, protein)	**Number of RefSeq complete genomes**	**Number of representative/ reference genomes**
Bacteria	68,260	27,351	3,837
Archaea	1,410	477	239
Fungi	16,581	22	22
Viruses	11,620	11,303	47

Use of bioinformatic tools for metagenomics generally requires knowledge of the command-line, however, there are packaged workflows with graphical user interfaces (GUIs) available. Illumina provides the DRAGEN pipeline [195] on its data analysis hub BaseSpace, which uses Kraken 2 to analyse Illumina data and provides an organism detection report. ONT also provides a cloud-based solution for nanopore data called EPI2ME with workflows such as What's In My Pot (WIMP). WIMP uses Centrifuge for classification and provides an online result report (Figure 1.3) [196]. Additionally, ONT offers a more customisable local workflow on the EPI2ME-Labs platform called wf-metagenomics that uses Kraken 2 with the option of a custom database [197].

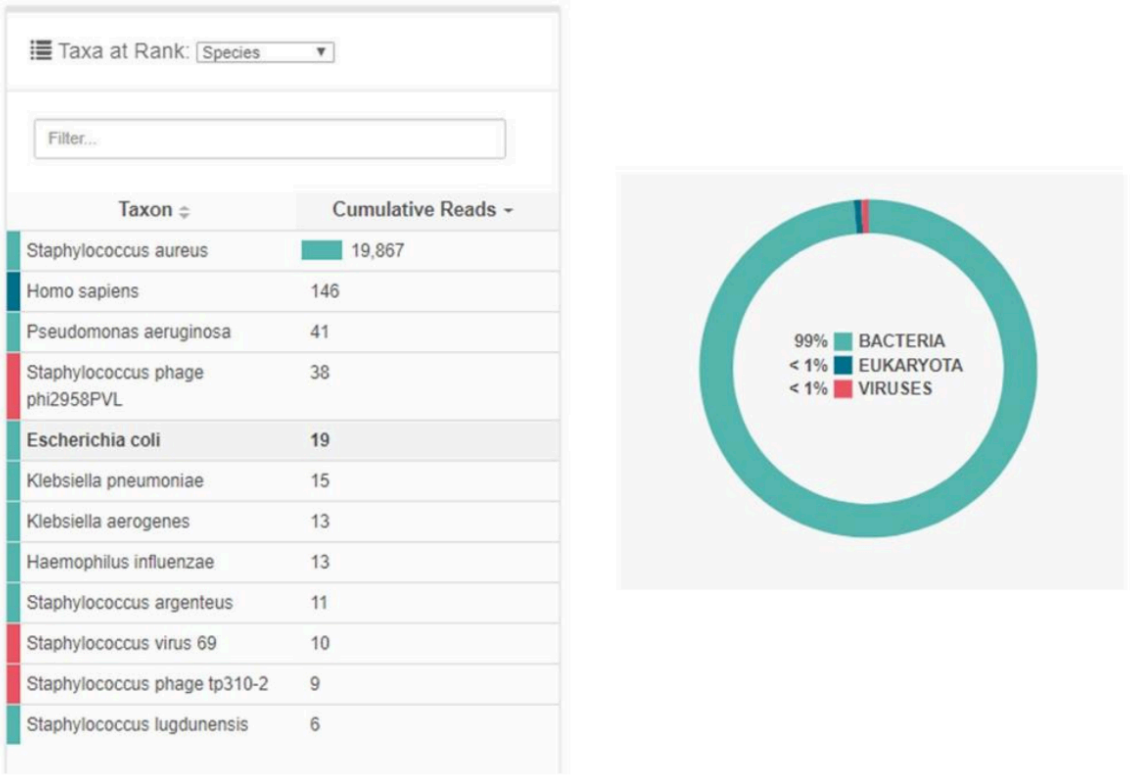


Figure 1.3 – Graphical user interface of a WIMP classification result

1.4.3.2 Antimicrobial resistance analysis

Beyond just simple taxonomic classification, there are many tools that can be used analysing AMR data too [198]:

- ABRIcate [199] and sraX [200] work by aligning contigs against a database
- starAMR [201] and AMR Finder Plus [202] both use contigs for BLAST-based analysis
- ResFinder 4.0 also aligns reads to a database but does not need an assembly [203]
- PointFinder is a subtool of ResFinder for the detection of chromosomal point mutations, limited to a few organisms [204]
- shortBRED uses a marker gene database to search the data for relevant protein families [205]
- RGI (Resistance gene identifier) uses protein homology and SNP models from contigs, plasmids, low quality assemblies and merged metagenomic reads [206].

ResFinder and RGI have web-based inputs for detection and therefore do not require knowledge of the command-line. AMR Finder Plus, RGI and ResFinder's subtool PointFinder [204] can detect chromosomal point mutations in addition to acquired resistance.

There are a number of AMR databases that can be utilised. The Comprehensive Antibiotic Resistance Database (CARD) is a widely used curated collection of resistance determinants derived from peer-reviewed publications [206]. Other databases include ARG-ANNOT [207], Resfinder's database [208], MEGARes 2.0 [209], NCBI AMRFinderPlus's Reference Gene Database [202], Antibiotic Resistance Genes Database (ARDB) [210] and UniProt [211]. Tools such as ABRIcate come bundled with many of these databases by default, such as ResFinder, NCBI and CARD, as well as virulence and plasmid databases such as PlasmidFinder and Ecoli_VF [199]. However, the onus is on the user to know which database to select for their application

For CMg data, the most basic AMR analysis is to simply align reads to a resistance gene database and report the results. ONT's EPI2ME Antimicrobial resistance workflow uses minimap2 to map uncorrected individual reads to the CARD database. However, this is inadequate. The standard approach for AMR analysis with CMg data is taxonomic binning followed by assembly, however, these can suffer from lack of sensitivity [212]. An alternative approach is to use a combination of taxonomic binning, AMR gene mapping, and assembly [212]. This has been shown to improve AMR gene detection compared to traditional methods.

1.5 Genomic Epidemiology

1.5.1 Epidemiology using metagenomic data

Metagenomic data can be used for more than pathogen ID and resistance gene detection. Once a pathogen has been identified, the genome can be used for subtyping and

monitoring outbreaks (for example in hospitals) [78]. While there have been studies showing the potential of using metagenomic data to generate whole genomes [213] and use them for high resolution bacterial typing [214], studies taking advantage of this for surveillance in a clinical setting have been sparse. A recent study did use CMg to investigate an unexpectedly high rate of *Klebsiella* infections in a hospital for the potential of patient-to-patient transmission [115]. Metagenomic sequencing data (~24hr sequencing) was analysed to call the sequence types of the organisms and identified a likely ST307 *K. pneumonia* circulating within the hospital [115]. It is possible to compare sequences at a higher resolution too by analysing Single Nucleotide Polymorphisms (SNP). The study generated consensus genomes and identified SNP differences between all samples to determine the distance, identifying a likely transmission of *Cornyebacterium striatum* between patients – epidemiological data later showed that most patients with *C. striatum* infection had overlapping stays, supporting the idea of a *C. striatum* outbreak in the hospital.

CMg has also been used for tracking viral outbreaks. Metagenomic sequencing of norovirus in a large paediatric hospital revealed discrete transmission clusters and pointed to chronic shedding of the virus from a specific immunocompromised patient [215]. CMg has also been used outside of the clinical setting to confirm outbreaks. A study in the US investigating two severe foodbourne outbreaks using metagenomic sequencing of stool samples was able to identify two distinct strains that were the cause [216]. In addition to identifying the strains, by performing CMg, the study was also able to detect *S. aureus* coinfections, as well as changes in the gut microbiome, identifying potential diagnostic signatures.

CMg has also been applied to detecting outbreaks caused by new organisms, with the original SARS-CoV-2 genome sequence elucidated by metagenomic sequencing of a clinical sample, before any targeted scheme existed for it [217]. While this is possible when the pathogen load is high, genomic epidemiology by CMg can be limited by low abundance of organisms. In many examples, samples have to be removed from analyses

due to low coverage [115]. Therefore, there is still an important role for targeted sequencing of organisms for genomic epidemiology.

1.5.2 Targeted sequencing for genomic epidemiology

Surveillance of pathogens can be significantly enhanced by utilising genomic epidemiology to help control outbreaks and the spread of disease. For example, sequencing of portions of the HIV genome has been used for genotyping to aid in reducing transmission [218]. The Centre for Excellence in HIV/AIDS (CFE) Laboratory Program in British Columbia routinely genotypes HIV cases (covering over half estimated cases in 2015), mainly using the 1497 bp HIV *pol* gene. New genotypes are added to the system every day, which triggers automated analysis of the data pool. The system is able to identify and report clusters of new HIV cases, and in 2014 was able to detect a growing cluster of 11 new HIV cases in a short period of time, which led to a public health intervention. The data could also be used to track the transmission of drug resistance. Such interventions have led to increased patients taking up treatment and reduction of onward transmission [218].

Viral pathogens causing disease often evolve very quickly, with genomes from closely related cases showing nucleotide differences in a short timescale (over weeks and months) [219]. Sequencing of a new virus in an outbreak provides important information about the virus, features of its genome and relatedness to known viruses, as was the case with the metagenomic sequencing of the SARS-CoV-2 genome [217]. This allows the development of diagnostic NAATs and targeted sequencing panels. As more viruses are sequenced, the genetic diversity can be used to estimate the speed of disease spread and create models. This was demonstrated in the 2009 H1N1 influenza A pandemic when a sequence-based estimate of transmission was shown to be similar to traditional epidemiological estimates [219]. Sequencing was also used to investigate nosocomial transmission of flu [220], with a study by Blackburn et al. (2019) retrospectively showing

that nosocomial transmission plays a significant role in the hospital burden of influenza [221].

The Ebola virus epidemic in West Africa (2014-2016) was the first example of the application of wide-scale real-time genomic epidemiology using MinION sequencing [222]. The epidemic started in Guinea and quickly spread to other countries in 2014, causing over 28,000 cases and 11,325 deaths by 2016 [223]. The viral genome substitution rate was shown to be around 1.19×10^{-3} mutations per site per year [222] which is equivalent to approximately 22 mutations in a genome per year – meaning that sequences diverged enough to detect sub-lineages over the course of the outbreak. Quick et al. demonstrated that the necessary lab equipment could be transported to Guinea in airline luggage and then used to sequence Ebola viruses on site. This was a tiling PCR approach with 38 primer pairs that amplified the whole EBV genome, which was sequenced on a MinION locally and on the Illumina MiSeq in UK. MinION results were obtained in less than 24 hours from collection, with the sequencing time taking as little as 15-60 minutes. Over 6 months, 142 samples were sequenced to provide good coverage of the cases over that period. The data was used to identify two separate lineages and make inferences about transmission between Sierra Leone and Guinea. The study was an important proof-of-concept which provided important tools and know-how for future virus outbreaks.

In the past decade, there have been many arbovirus threats globally, including Zika virus, chikungunya virus and dengue virus. Other regional mosquito-borne viruses such as yellow fever and West Nile virus also continue to persist and cause significant morbidity [224]. During the 2016 Zika virus epidemic genomic surveillance was performed using targeted sequencing to track transmission in the Americas. Multiple different methods were, all requiring either amplification or enrichment of the Zika virus genome due to the high C_T values in these samples (low genome copies) [225]. One approach taken by Thézé et al. (2018) used spiked primer enrichment to preferentially amplify the viral genome, yielding Zika virus reads in 71 out of 81 samples, with an average coverage of 64% [226]. This study demonstrated that there were multiple independent introductions

into Central America and Mexico, likely from Brazil. Another approach developed by Quick et al. (2017) amplified the Zika virus genome using a 35 amplicon primer scheme (ZikaAsian) [225] prior to sequencing, similar to the method used in the West Africa Ebola epidemic. This method could be used for sequencing on any device and was demonstrated with Illumina sequencing to show that there were multiple introductions of the virus into the US (4-40) linked to the Caribbean [227]. This study also demonstrated the use of another enrichment method using probes to capture Zika Virus RNA instead of using targeted PCR [227]. The ZikaAsian PCR tiling approach was used with MinION sequencing in a mobile genomics laboratory in Brazil to sequence 54 Zika Virus genomes. Results from the study showed that Zika virus was disseminated from northeast Brazil both nationally and internationally and that there was cryptic transmission prior to first detection [228]. This tiling approach was generated using a publicly available tool, PrimalScheme (by Josh Quick and Andy Smith) and has also been used to generate a PCR scheme for chikungunya virus [225].

Genomic epidemiology has become an extremely valuable tool in the study of viral outbreaks. Genome sequence data can now be used for a number of purposes, such as confirming the geographic origin of cases, identifying lineages with increased virulence and/or transmissibility and identifying lineages that can evade natural or vaccine induced immunity [224]. The tools developed during the evolution of this field over the past decade proved invaluable when applied during the ongoing SARS-CoV-2 pandemic.

1.6 COVID-19

SARS-CoV-2, which is the cause of COVID-19, emerged in late 2019 in Wuhan, China, and quickly spread across the globe over the course of 2020, leading to the current pandemic [229]. SARS-CoV-2 belongs to a broad range of viruses called betacoronaviruses and is related to other known coronaviruses that are highly pathogenic in humans i.e. Middle East respiratory syndrome coronavirus (MERS-CoV) and severe

acute respiratory syndrome coronavirus (SARS-CoV). The full genome of the SARS-CoV-2 was first published on the 10th of January 2020, sequenced by metagenomic cDNA sequencing of a BAL sample taken from a patient with severe pneumonia [229]. It shares 79% of its genome sequence with SARS-CoV but is much more closely related to horseshoe bat and pangolin coronaviruses, though is still genetically distinct. Currently the intermediate host between the bat/pangolin coronaviruses and human SARS-CoV-2 virus has not been found. The virus has 4 structural proteins encoded by 4 genes; S (spike), E (envelope), M (membrane) and N (nucleocapsid) [229]. The spike protein has been of particular interest, as this contains the receptor binding domain that binds to the human ACE2 receptor, allowing the virus to enter host cells. Mutations in the spike gene have been linked to important characteristic changes in the virus, such as transmissibility [230].

SARS-CoV-2 can infect anyone at any age; however, the severity of COVID-19 is correlated with increased age and co-morbidities. Young people are often asymptomatic or have mild disease, whereas older people and those with co-morbidities are more likely to develop severe disease, leading to hospitalisation and ultimately death in a minority of cases. The most common symptoms of disease caused by the first circulating lineages included fever, fatigue, loss of taste and smell, and a continuous cough. Less common symptoms included headache, sputum production, chest pain, chills, nausea, and sore throat. However, over the course of the pandemic, the emergence of new lineages/variants has led to different symptom profiles, with the later Omicron variant being less associated with loss of smell and more associated with a sore throat [231]. The disease can ultimately lead to respiratory failure, septic shock, and other organ failures. According to the WHO Coronavirus Dashboard, SARS-CoV-2 has reached every country in the world and at present (June 2022) there are over 542 million confirmed cases globally (potentially a significant underestimation of real cases) and 6.3 million deaths [3].

The massive impact the pandemic had on society led to a global drive for vaccines and treatments. As of June 2022, there were 853 unique activate compounds, with 243 vaccines, 261 antivirals (drugs that interact with the virus to disrupt replication) and 349

treatments (drugs that treat the various illnesses resulting from the virus) [232]. Some of the vaccines, such as BNT162b2 (Pfizer-BioNTech) and mRNA-1273 (ModernaTX) use new messenger RNA (mRNA) vaccine technology, where modified mRNA for the SARS-CoV-2 spike protein are introduced in lipid nanoparticles to be expressed in the body. Others such as the ChAdOx1 vaccine (Oxford-AstraZeneca) use a chimpanzee adenovirus as a vector to carry SARS-CoV-2 spike proteins. Clinical trials have shown that these vaccines have high efficacy, with the BNT162b2 vaccine originally being reported to be as high as 95% effective in preventing COVID-19 after the second dose [233]. Many of these vaccines received emergency use approval with vaccination programmes rapidly deployed globally. As of June 2022, over 5.23 billion people have received at least one dose of a COVID-19 vaccine globally (68.1% of the world), with the ChAdOx1 (Oxford-AstraZeneca) vaccine having the highest global reach, being used in 185 countries, followed by the BNT162b2 (Pfizer-BioNTech), used in 164 countries [234].

Effective treatments for COVID-19 have been more difficult to develop, however, some progress has been made. The Randomised Evaluation of COVID-19 Therapy (RECOVERY) trial was a global trial that tested multiple treatment options for COVID-19. It recruited 47,761 cumulative participants with 199 active sites (as of June 2022) [235]. The trial made an early breakthrough discovery, showing that using dexamethasone, a corticosteroid, for the treatment of COVID-19 cut the rate or mortality by 1/3rd for patients on ventilators and by 1/5th for patients receiving oxygen [236]. Other findings from the trial indicated that monoclonal antibody therapy may also reduce deaths in COVID-19 patients [237]. Since then, other treatments such as Baricitinib, Tocilizumab, [238] and Paxlovid [239] were also found to reduce COVID-19 deaths in hospitalised patients.

Diagnosis of COVID-19 is crucial for preventing the spread of infection and to identify which samples need to be sequenced for genomic epidemiology purposes. Molecular tests, particularly reverse transcription qPCR (RT-qPCR) have been used for SARS-CoV-2 detection from the beginning of the pandemic [240][241]. Most nucleic acid tests target either ORF1b, or the nucleocapsid, envelope or spike protein genes [229]. Time-to-result

can take minutes to hours depending on the test. SARS-CoV-2 can be detected from many respiratory sources, including throat swabs, nasal swabs, sputum, bronchial fluid, and saliva, but can also be detected in faeces and blood [229]. In fact RT-qPCR of wastewater has become an important analysis tool in the pandemic, giving early warnings as infection levels rise to allow timely intervention [242]. In addition to RT-qPCR, which detects the RNA of SARS-CoV-2, antigen Lateral Flow Tests (LFT) have become widely used, which utilise monocloncal antibodies to target the N protein produced by SARS-CoV-2 [243]. LFTs enable mass community testing without requiring central laboratories as is typically the case with RT-qPCR. It has been shown that LFTs have a very high specificity (>98%) but relatively low sensitivity (65-89%) [243]. Sensitivity is high (>90%) for samples with lower SARS-CoV-2 qPCR C_T (i.e. higher viral loads). RT-qPCR remains the gold standard, and has provided more information than just detection of the virus – important deletions in the spike gene causing 'drop-outs' in commonly used commercial kits such as the ThermoFisher TaqPath COVID-19 PCR assay could be used to make inferences about the variant in the sample [244]. This is useful because as the pandemic progresses, evolution of the virus leads to variants that can pose an increased risk (e.g. higher transmissibility, higher virulence, immune evasion). Therefore, it is useful to track so called Variants of Concern (VOC). This is why multiplex panels were later designed specifically to be able to identify VOCs by targeting specific deletions. For example, a simple 3-target multiplex was able to differentiate between B.1.1.7 (Alpha), B.1.351 (Beta) and P.1 (Gamma) variants in one PCR [244]. However, these PCR tests could only be designed due to the initial characterisation of these variants by sequencing. Additionally, as SARS-CoV-2 evolves, old variants are rapidly replaced by new ones, such as Delta and Omicron which followed Alpha, Beta and Gamma, meaning that these schemes become outdated quite quickly.

1.6.1 COVID-19 Genomics (COG) Consortium and SARS-CoV-2 surveillance

The COVID-19 Genomics UK Consortium (COG-UK) was a consortium of academics set up by Sharon Peacock to sequence SARS-CoV-2 genomes which launched in March, 2020. The main aim was to sequence SARS-CoV-2 to track transmission, identify mutations, and assess how the viral genome interacts with other factors relating to COVID-19 (e.g. consequences) [245]. This information could potentially identify vaccine and therapeutic targets, but the primary aim was genomic epidemiology. One of the main priorities of COG-UK was to ensure that sequence data was matched with metadata, including patient clinical information and non-genomic epidemiology to enhance the value of the genetic data [245]. It is the first time such a large scale national effort for genomic epidemiology has been made, and results were used to guide decision-makers such as the Scientific Advisory Group for Emergencies (SAGE).

COG-UK was comprised of an integrated network of sequencing sites across the UK, co-ordinated from the University of Cambridge, led by Professor Sharon Peacock. Samples were collected and sequenced locally across the country with 16 sequencing hubs and over 70 partners involved, including academic partners, public health agencies, hospitals and other sequencing sites [246]. The advantages of this decentralisation meant that sequencing could be performed rapidly at or close to the site of sample collection, providing results quickly [245]. Sequencing in some regional sequencing sites was performed in almost real time, with 1-2 days turnaround time. There have also been sites that provided higher volume capacity for national sequencing, such as the Wellcome Sanger Institute, providing sequencing for sites without the local capabilities. Quadram Institute Bioscience (QIB) was one such site that in addition to performing local sequencing, also performed sequencing of samples from Lighthouse Laboratories and from studies such as the REal-time Assessment of Community Transmission (REACT) study [247]. QIB was also involved in sequencing of samples from international sites such as Zimbabwe [248].

Weekly analysis reports were provided to SAGE, giving information on whether cases were predicted to be locally transmitted versus imported, the rate of epidemic growth, spatial movement, transmission chains, genetic changes (e.g. mutations), and identification of genetic changes that could potentially affect diagnostic tests or therapies [245]. With the huge number of sequences deposited and a lack of universal classification system, a new nomenclature was created for the SARS-CoV-2 lineages. This was a dynamic classification system that focused on active lineages and was adopted globally which assisted in the tracking of COVID-19 globally [249].

COG-UK developed an end-to-end computer infrastructure to handle the amount of data from all parts of the country. The Cloud Infrastructure for Microbial Bioinformatics (CLIMB) compute facility was used as the hub that the system was built around [250]. Sites that were part of COG-UK performed alignments against the SARS-CoV-2 genome and uploaded the alignments (as BAM files) as well as consensus FASTA files. This allowed sites to have control over their data and prevented human reads from being uploaded to the cloud. It also meant that it removed unnecessary delays from uploading raw sequence data so that actionable data could be provided as fast as possible. In addition to sequence data, mandatory metadata about the sample was also uploaded, which included a central sample identifier, date of sample collection, the country code, and the sampling strategy. It was also recommended that the county the sample was collected from was also detailed. The data was subjected to a quality control (QC) check by the database to limit which sequences would be made available to downstream pipelines. And different QCs could be applied based on whether the sequencing was performed on Illumina or ONT [250]. The data was also disseminated internationally, by uploading consensus sequences to the Global Initiative on Sharing Influenza Data (GISAID) which became the de facto international repository for SARS-CoV-2 sequences. Data from COG-UK was routinely analysed and along with some international data from GISAID was visualised publicly on Microreact, showing the rise and fall of variants, with data from March 2020 up to February 2022 [251].

COG-UK was hugely successful, generating over 550,000 public sequences in just over a year from its inception [250]. As of February 2022, the UK had sequenced and uploaded over 2 million SARS-CoV-2 genomes to GISAID, meaning that a quarter of all sequences globally were from the UK. This accounted for approximately 10% of all estimated COVID-19 cases in the UK, allowing detailed surveillance throughout the pandemic [252]. COG-UK contributed over 20 reports to the government, over 50 publications in journals and aided in the control of outbreaks across the country [250].

One of first examples of COG-UK data being used to show the evolution of SARS-CoV-2 was with the 2020 variant containing the spike gene D614G mutation. It was shown that increases in the D614G variant were likely due to a selective advantage of the mutation. It was associated with higher viral loads and there was evidence to support higher transmissibility (but not higher mortality or clinical severity) [253]. Following on from this, continued investigation of developing lineages led to the identification and characterisation of the B.1.1.7 (Alpha) variant [254] which was designated as the first VOC by Public Health England (now UKHSA). Investigation of spike gene target failures from RT-qPCR, as well as COG-UK sequence data (31,390 Alpha variant sequences and 52,795 non-Alpha SARS-CoV-2 sequences) showed that the new variant had a 50-100% higher reproduction number (R_t) than normal [254].

COG-UK data was also used to study the dynamics of transmission from abroad and within the UK. A study on the early pandemic showed that the first wave of the pandemic in the UK started with hundreds of independent introductions into the UK [255]. The study estimated that the vast majority of transmission lineages in the first wave were due to arrivals from Spain, France, and Italy. As non-pharmaceutical interventions (NPIs) such as lockdowns and travel restrictions were introduced, lineages quickly became extinct and diversity of lineages was lost. Genomic epidemiology has aided in driving public policy. For example, the Alpha variant identified in December 2020 was one of the reasons the UK government imposed a Christmas lockdown [256]. The discovery of a cluster of Beta variants in parts of South London led to deployment of surge testing in the region in an

attempt to suppress the variant [257]. High prevalence of VOCs in other countries led to the imposition of travel bans to certain countries to prevent importation of VOCs into the UK [258]. Rising cases of the Delta variant was the reason the UK government delayed relaxation of lockdown rules in 2021 [259].

Sequencing of SARS-CoV-2 has continued to be used for surveillance of emerging and circulating lineages around the world. Currently the WHO designates one variant as a VOC, Omicron, due to the widespread transmission globally. Since February 2022, over 98% of sequences uploaded to GISAID were Omicron [260]. Omicron continues to evolve, with descendent lineages arising. Currently, Omicron is split into 5 further sub-lineages BA.1, BA.2, BA.3, BA.4 and BA.5. The BA.4 and BA.5 variants were driving a summer surge of COVID-19 infections in the UK at the time of writing (July 2022) [261].

1.6.2 SARS-CoV-2 genome sequencing methods

In early 2020, Josh Quick from the ARTIC network [262] released a method for sequencing SARS-CoV-2 using a tiling PCR method similar to the those used for Ebola and Zika viruses. This method was designed for nanopore sequencing, enabling rapid turnaround for actionable results in a short timeframe. Version 1 of the protocol used a ~400 bp 98-amplicon scheme (divided into two pools) to amplify the whole SARS-CoV-2 genome, designed using PrimalScheme [263]. The method was intended for sequencing using ONT's Ligation kit, which had 24 available barcodes at the time, meaning the maximum number of samples per flowcell was 24 (or fewer, if controls are included). Illumina and multiple academic groups developed methods for SARS-CoV-2 sequencing based on the ARTIC primer scheme which increased throughput [264–266]. Changes to the primer scheme to optimise and fix various amplicon drop-outs caused by new variants led to multiple iterations of the ARTIC primers, with the latest being V4 [267]. Subsequently ONT increased the number of native barcodes to 96, increasing the number of samples that could be sequenced per flowcell [268,269].

Other amplicon schemes were later developed, such as a 1.2 kb scheme (the ‘Midnight’ scheme) [270], a 2 kb amplicon scheme [271], and a 2.5 kb scheme which only required 14 amplicons [272]. It was later shown in a comparison that the original ARTIC 400 bp scheme produces more data, however, due to the unevenness of coverage, it produces fewer finished sequences than using the longer amplicons. The use of longer amplicons led to more even coverage across the genome [273]. The original ARTIC method, designed for nanopore sequencing, was adapted for Illumina sequencing [265]. The 2.5 kb amplicon protocol was also intended for Illumina sequencing [272], while the 1.2 kb amplicon scheme was adapted by ONT into a kit that used its rapid transposase chemistry instead of the traditional ligation library preparation [274]. It is possible to avoid PCR tiling altogether, as a study showed that using bait capture enrichment was a viable alternative to PCR, however, required very long hybridisation times [275].

1.7 Aims of PhD

The overall aim of my PhD is to demonstrate the application of rapid sequencing for the detection and epidemiology of respiratory pathogens. Key to this aim is the optimisation of a CMg workflow for rapid and accurate detection of pathogens directly from clinical samples. Aside from metagenomics applications however, the COVID-19 pandemic provided an opportunity to develop and test a targeted sequencing approach for genomic epidemiology of the respiratory pathogen SARS-CoV-2.

The objectives are:

- To optimise a respiratory CMg workflow for the detection of bacteria and fungi by reducing test turnaround time and complexity, thereby bringing it closer to clinical implementation
- To develop a viral metagenomics workflow capable of detecting DNA and RNA viruses from respiratory samples that could be used in parallel to the bacterial/fungal test
- To develop a method for high-throughput sequencing of SARS-CoV-2 to aid in the expansion of genomic epidemiology

2. Methods

2.1 Clinical sample processing

2.1.1 Sample collection ethics

Excess respiratory samples were collected under University College London (UCL) infection DNA bank ethics used by the INHALE study (REC 12/LO/1089) which allowed for the collection of residual diagnostic samples for research. Samples were not study specific and were collected after being processed by routine microbiology. No patient data was collected and culture results for the samples were pseudoanonymised by the microbiology department. Informed consent was therefore not required. Similarly, for SARS-CoV-2, COVID-19 Genomics UK Consortium excess diagnostic samples were used with no patient identifiable information collected or used.

2.1.2 Respiratory sample collection and storage

For development and testing, surplus sputa and endotracheal aspirates (ETA) were requested from the NNUH microbiology department. Samples included routine diagnostic positives (where a pathogen is detected) and negatives, i.e., normal respiratory flora (NRF), where growth is deemed to be the normal commensal lung community, or no growth. Sputum samples were already treated by the microbiology department with a 1:1 volume of Sputasol and incubated at 35-37°C for 15 minutes. Samples that had already been processed by the microbiology department were selected by staff at the Innovation Centre, Norfolk and Norwich University Hospital (NNUH) and aliquoted into 20 mL sterile universal tubes for collection. The aliquots were collected from the Innovation Centre and transported to the Bob Champion Research and Education (BCRE) building or QIB at ambient temperatures in under 10 minutes and then stored at 4°C until depletion and extraction. Most samples were processed within 24 hours after collection and a maximum of 72 hours. After 72 hours, any unused samples were discarded by clinical waste routes.

2.1.3 Culturing and storage of bacteria

E. coli was grown for RNA extraction to be used in Sequence-Independent, Single-Primer Amplification (SISPA) experiments. Starter cultures were created by transferring 20 μL of an *E. coli* glycerol stock to 4 mL of Luria Bertani Broth (NaCl, 10 g/L, Tryptone, 10 g/L, Yeast Extract, 5 g/L) in a 20 mL universal tube and grown overnight in a shaking incubator (180 rpm) for 18 hours at 37 °C. From the starter culture, 20 μL was transferred to a new a 20 mL universal tube with 4 mL of LB broth and grown in a shaking incubator (180 rpm) at 37 °C for 2.5 hours. Cells were harvested by splitting the 4 mL culture into two in 2 mL Eppendorfs and centrifuged at 1000xg for 5 minutes. The supernatant was discarded in both tubes, leaving bacterial pellets.

2.1.4 Pre-host depletion sample processing

Samples were treated with sputasol depending on viscosity (even if they had been treated previously). A working stock of sputasol was made by adding 92.5 mL of water to 7.5 mL of sputasol (Oxoid) (or an equivalent ratio) and mixed by vortexing. An equal volume of freshly made sputasol was added to samples that were too viscous to be readily pipetted and vortexed for 15 seconds. The sample was then incubated at 37 °C until fully homogenised. Following homogenisation (checked by the ability to pipette), the full volume was centrifuged at 12,000 xg for 3 minutes. Half of the supernatant was removed and the rest was resuspended in the remaining volume by pipetting to achieve the same starting volume before the addition of sputasol. If internal control spikes were used in the experiment, they were added to the sample prior to sputasol treatment.

2.1.5 Host depletion and controls

Host depletion was performed on surplus sputum samples from NNUH to test and compare the original method host depletion method with the new one-pot method (Section 2.1.5).

A working stock of 5% saponin solution was made by adding 250 mg of saponin (Tokyo Chemical Industry UK) to 5 mL of PBS. HL-SAN buffer was made by dissolving 11.69 g of NaCl (5M) and 0.38 g of $MgCl_2$ (100 mM) in 30 mL of water and making up to 40 mL with water then filter sterilised using a SCFA Syringe Filter 0.45 μM (Corning). HL-SAN buffer was stored at room temperature and always vortexed prior to use in case of salt precipitation.

To 400 μL of sample (or max available volume topped-up with PBS to 400 μL) was centrifuged at 8000 xg for 5 minutes. The supernatant was discarded and the pellet was re-suspended in 250 μl of PBS and mixed by pipetting and vortexing. If the pellet was difficult to resuspend, the pipette tip was used to mechanically disrupt the pellet. To the resuspended pellet, 200μl of 5% saponin solution was added and mixed by vortexing. The tubes were incubated at room temperature for 10 minutes. After the saponin treatment, 350 μL of H_2O was added and incubated at room temperature for 30 seconds. Following this, 12 μL of 5M NaCl was added and vortexed. The sample was centrifuged at 6000 xg for 5 minutes and the supernatant discarded. The pellet was resuspended in 100 μL of PBS and 100 μL of HL-SAN buffer was added and mixed by vortexing. 10 μL of HL-SAN (ArticZymes) was added and mixed by pipetting up and down. The reaction was incubated at 37°C for 15 minutes at 800 rpm in a thermoshaker. After incubation, the pellet was washed with the addition of 800 μL of PBS and centrifuged at 6000 xg for 3 minutes. The supernatant was discarded and the pellet was resuspended again in 1 mL of PBS. The mix was centrifuged at 6000 xg for 3 minutes and the supernatant discarded again. The pellet was then used for DNA extraction.

For all depletion experiments, a non-depletion control was processed with DNA extraction only (this was an equivalent volume to the sample). DNA extraction was performed on the sample along with the depleted samples. Additionally, a negative process control was processed with every batch of depletions; 400 μL of PBS was processed just as all the depletion samples.

2.1.5 One-pot host depletion

Working saponin solution (1%) was made by adding 50 mg of saponin to 5 mL of PBS. Saponin solution was stored at room temperature in a dark place and used within 1 week.

To 200 μL of sample, 40 μL of 1% saponin solution, 200 μL of HL-SAN buffer and 10 μL of HL-SAN was added. This was incubated in a thermoshaker for 10 minutes at 37 °C at 1000 rpm. After incubation, 1 mL of PBS was added and centrifuged at 12,000 xg for 3 minutes. The supernatant was slowly aspirated and discarded, leaving approximately 50 μL of liquid behind to not disturb the pellet. The pellet was then used for DNA extraction.

2.1.6 Bacterial/fungal DNA extraction using the MagNAPure Compact

The host-depleted pellet was resuspended in 700 μL of Bacterial Lysis Buffer (Roche) and mixed by pipetting up and down at least 15 times. The resuspended pellet was transferred to a Lysing Matrix E tube (MP Biomedicals). Samples were homogenised in a TissueLyser LT (Qiagen) for 3 minutes at 50 Hz. The tubes were centrifuged at 20,000 xg for 1 minute and 400 μL of the clear supernatant was transferred to a DNA LoBind Tube (Eppendorf) with 20 μL of Proteinase K (Qiagen) per sample. This was mixed and incubated on a thermomixer at 65 °C for 5 minutes at 1000 rpm. A MagNA Pure Compact cartridge was set up by shaking the cartridge to resuspend the beads and loaded into the compact. After 5 minutes, the full volume of the Proteinase K treated sample was transferred to MagNA Pure Compact cartridge (Roche) from the Nucleic Acid Isolation Kit I. Samples were eluted in 50 μL of elution buffer. DNA extracts were stored at -20 °C.

2.1.7 Post-extraction clean-up

DNA extracts were SPRI cleaned with AMPure XP beads (Agencourt) at 1.2X (beads to sample volume). 20 μL of extract was made up to 50 μL with water and 60 μL of resuspended beads was added. This was incubated for 5 minutes at room temperature with periodic flick-mixing. The tube was then added to a magnetic rack (DynaMag 2,

Thermofisher) for 3 minutes to pellet. The liquid was discarded, and 200 μL of 70% ethanol was added and removed after 30 seconds – this step was repeated for a total of two washes. The tubes were pulse centrifuged and any residual ethanol was removed. The beads were dried with the lids of the tubes open for less than one minute, then 15 μL of Nuclease Free Water (Ambion) was added to elute the DNA. This was left to incubate at room temperature off the magnetic rack for 2 minutes, then added back to the magnetic rack to collect the eluate without the beads.

2.2 Library preparation and sequencing

2.2.1 Published library preparation method

Extracted and SPRI cleaned DNA was tagmented by adding 2.5 μL of FRM (SQK-RPB004, ONT) to 10 ng of sample (or maximum available) in 7.5 μL for a total reaction size of 10 μL. This was gently flick mixed and incubated in a thermocycler (conditions Table 2.1).

Table 2.1 – Tagmentation conditions using FRM

Temperature	Time
30 °C	1 min
80 °C	1 min
4 °C	Hold

Following tagmentation, a mastermix using LongAmp Taq (NEB) was made with the components:

- 50 μL LongAmp Taq Mastermix (NEB)
- 2 μL RPB004 barcode (ONT)
- 38 μL Nuclease Free water (Ambion)
- 10 μL tagmented DNA

Total volume: 100 μL

The reaction was mixed by pulse vortexing for 2-3 seconds, centrifuged and transferred to a thermocycler for PCR (conditions Table 2.2) using a Veriti 96 Well Thermal Cycler (Applied Biosystems).

Table 2.2 – PCR cycling conditions for the library preparation method prior to optimisation

Cycles	Temperature	Time
1	95 °C	3 min
25	95 °C	15 sec
	56 °C	15 sec
	65 °C	6 min
1	65 °C	6 min
1	4 °C	Hold

PCR products were quantified with Qubit dsDNA HS (as detailed in Section 2.4.1).

2.2.2 Rapid library preparation

Optimisation of this method is detailed in the results. Tagmentation was performed the same way as the original method (detailed in Table 2.1).

A mastermix was made per sample with the Takara GXL PCR components:

- 20 μL 5X GLX buffer (Takara)
- 8 μL dNTPs (Takara)
- 56 μL Nuclease Free Water (Ambion)
- 4 μL GLX polymerase (Takara)
- 2 μL RPB004 barcode (ONT)
- 10 μL tagmented DNA

The reaction was mixed by pulse vortexing for 2-3 seconds, centrifuged and added to a thermocycler (conditions Table 2.3).

Table 2.3 – Conditions for the rapid PCR after optimisation

Cycles	Temperature	Time
1	98 °C	2 min
25	98 °C	15 sec
	56 °C	15 sec
	68 °C	45 sec
1	68 °C	4 min
1	4 °C	Hold

PCR products were quantified with Qubit dsDNA HS (detailed in Section 2.4.1).

2.2.3 Post-PCR pooling and SPRI cleaning

For each run, samples were pooled post-barcode PCR and quantified (for Flongle, 1-3 samples, or MinION, 6 samples). 500 ng of each sample was added to 1.5 mL Eppendorf and mixed. If only one sample was being sequenced, then 1000 ng was taken instead. The maximum volume of the process negative control was added to the pool in each batch. If 500 ng was not reached for a particular sample, then the maximum volume of sample was added instead. The pooled library was Solid Phase Reversible Immobilization (SPRI) cleaned with 0.6X (volume) AMPure XP beads and incubated at room temperature on a hulamixer for 5 minutes. After incubation, the tube was added to a magnetic rack for 3 minutes to separate the beads from solution. The beads were washed by adding 500 μL of 70% ethanol and removed, this was repeated for a total of two washes. The pellet was air dried for less than 1 minute and the DNA was eluted in 12 μL of MinION buffer (10 mM Tris-HCl and 50 mM NaCl). The post-washed DNA was quantified using Qubit High Sensitivity and fragment size analysis using the Tapestation (detailed in Section 2.4.1).

2.2.4 Sequencing using MinION and Flongle

Samples were sequenced individually or in batches of 2-3, plus a negative control per Flongle flowcell. The final loading library was prepared with a total of 50 fmol of the pooled

library. 0.5 μL of RAP (ONT) was added to 5 μL of library and incubated at room temperature for 5 minutes. The Flongle flowcell was flushed with 100 μL of flush buffer mix (117 μL flush buffer and 3 μL flush tether), and the final loading library was made up with 13.5 μL sequencing buffer, 11 μL loading beads and 5.5 μL of adapted library. 30 μL was loaded on the Flongle flowcell. Sequencing was performed for a minimum of 2 hours up to a maximum of 24 hours.

For developmental work, sequencing was performed on MinION using manufacturer's instructions. However, the final loading library was prepared with a total of 100 fmol of the pooled library, rather than 50 fmol as recommended. 1 μL of RAP (ONT) was added to 10 μL of library and incubated at room temperature for 5 minutes. The MinION flowcell was primed with 800 μL of priming mix (30 μL of FLT added to a whole tube of FB, ONT) through the priming port. After 5 minutes, another 200 μL of priming mix was added with the SpotON port open, and the final loading library was loaded (34 μL of SQB, 25.5 μL of LB, 4.5 μL of nuclease free water and 11 μL of adapted library). Sequencing time varied depending on the experiment.

2.2.5 Flowcell washing

Flowcells were washed using a community nuclease wash method [276] prior to the release of official nuclease wash kits by ONT.

The community buffer composition:

- 300 mM KCl,
- 2 mM $CaCl_2$
- 10 mM $MgCl_2$
- 15 mM HEPES

  pH 8.0

480 μL of this buffer was mixed with 20 μL of DNase I (Qiagen) and 500 μL of this mix was loaded through the priming port. Flowcells were incubated in the MinION device for 30 minutes and then flushed with 500 μL of Storage Buffer from the Flowcell Wash Kit (ONT). The priming port was then closed, and the waste chamber was emptied by pipetting. Following the release of nuclease-based wash kits by ONT, (Flowcell wash Kits V3 and V4, ONT), the reagents in these kits were used instead.

2.3. SISPA and viral metagenomics

2.3.1 Viral RNA extraction

For tests with viruses, HIV-1 was used as a surrogate for a single-stranded RNA (ssRNA) virus and excess sputum or PBS were spiked with 100 μL of AcroMetrix™ HIV-1 High Control and centrifuged at 12,000 xg for 3 minutes. The supernatant was transferred to a new Eppendorf and used as the virus-containing part of the sample.

Samples were extracted using the Promega Maxwell RSC. To 300 μL of sample, 300 μL of Lysis Buffer and 30 μL of Proteinase K from the Maxwell® RSC Viral Total Nucleic Acid Purification Kit was added and incubated at 56 °C on a shaking incubator for 10 minutes, at 1000 RPM. The Maxwell was loaded as per manufacturer's instructions and 50 μL of elution buffer was used as recommended. The full lysed sample volume was added to the same well and the extraction started. Extracts were stored at 4 °C if used the same day or at -80 °C in the long-term.

2.3.2 Bacterial RNA extraction for SISPA experiments

E. coli RNA was used to test the SISPA method and extracted with the Qiagen AllPrep DNA/RNA/Protein Mini Kit method (with modifications). 10 uL of B-mercaptoethanol was added to 1 mL of RLT buffer. 600ul of RLT buffer was then added to one tube with the centrifuged bacterial cells to resuspend the pellet and then transferred to the other tube to resuspend the remaining pellet. The full volume was transferred to a Lysing Matrix E tube

and homogenized on a FastPrep 24 for 20 seconds at 6.0 m/s. 350 μL of the supernatant was transferred to the AllPrep DNA spin column. This tube was spun at 8000 xg for 30 s. To the flow-through RNA, 250 μL of 100% ethanol was added and mixed and transferred to an RNeasy spin column. This column was centrifuged at 8000 xg for 15 seconds. The wash was performed with 3 subsequent spins: 700 μL of RW1 was added to the column and centrifuged at 8000 xg for 15 s, then 500ul of RPE was added and centrifuged at 8000 xg for 15 s and finally, 500 μL of RPE was added and centrifuged at 8000 xg for 2 mins. Elution was performed with 50 μL of RNase free water and centrifuged at 8000 xg for 1 min. RNA was stored at -80 °C.

2.3.3 DNase treatment of RNA (Turbo DNA free)

To 50 μL of extracted RNA, 0.1x volume (5 μL) of TURBO DNase buffer (Invitrogen) was added followed by 2 μL of TURBO DNase enzyme. This was pipette-mixed and incubated at 37 °C in a thermoshaker at 300 rpm for 30 min. After DNase treatment, 0.2x volume (10 μL) of DNase inactivation buffer was added and incubated at room temperature for 5 minutes with resuspension by flick-mixing every 2 minutes. The solution was centrifuged at 10,000 xg for 2 mins and the supernatant containing the RNA was aspirated, avoiding the DNase inactivation beads. DNase treated RNA was stored at -80 °C.

2.3.4 RNA concentration

RNA extracts for SISPA were SPRI cleaned and concentrated following the Agencourt RNAClean XP (Agencourt) protocol. 30 μL of RNA extract was made up to 50 μL with nuclease free water. 90 μL of beads was added and mixed by pipetting and incubated at room temperature for 5 minutes. The tube was transferred to a magnetic rack for 5 minutes and the solution was aspirated and discarded. 500 μL of 70% ethanol was added and removed, and this was repeated for a total of 3 washes. Samples were dried on the rack until no visible ethanol remained and then 7 μL of nuclease free water was added

and mixed to elute. This was incubated at room temperature for 2 minutes and then added back to the magnetic rack

2.3.5 SISPA library preparation for *E. coli* RNA / viruses

The SISPA library preparation was adapted from Greninger et al. 2015 [95] 2 μL of RNA extract was mixed with 2 μL of nuclease free water and 1 μL of 40 μM 9N-primer (TTTTTCGTGCGCCGCTTCAACNNNNNNNNN). This reaction was incubated at 65 °C for 5 min and then 23 °C for 5 min.

The following First Strand synthesis reaction was prepared:

- 2 μL 5x Superscript IV buffer (Thermofisher)
- 1 μL 10 mM dNTPs
- 0.5 μL DTT
- 1 μL nuclease free water
- 0.5 μL Superscript IV RT (Thermofisher)
- 5 μL RNA/primer mix

This mix was incubated at 42 °C for 10 min. Following this, the second strand synthesis mix was made using Sequenase Version 2.0 DNA Polymerase reagents (Thermofisher):

- 1 μL Sequenase buffer (Thermofisher)
- 4.85 μL nuclease free water
- 0.15 μL Sequenase (Thermofisher)
- 5 μL of First Strand synthesis mix

This was incubated at 37 °C for 8 min.

An additional 0.45 μL sequenase dilution buffer (Thermofisher) and 0.15 μL Sequenase was added to the reaction after 8 min and incubated for another 8 min resulting in a total volume of 15.6 μL at the end of 2nd strand synthesis.

The library preparation PCR was prepared as follows

- 10 μL 5x GXL buffer (Takara)
- 4 μL dNTP
- 2 μL GXL Polymerase
- 1 μL ONT adapter primer (100 μM) (TTTTTCGTGCGCCGCTTCA)
- 28 μL nuclease free water
- 5 μL from 2nd Strand Synthesis as template

The reaction was mixed and PCR was performed with the rapid barcoding conditions (Table 2.3).

Following this, the barcoding PCR was set up.

Reaction composition:

- 10 μL 5X GXL buffer
- 4 μL dNTP
- 2 μL GXL Polymerase
- 1 μL RPB004 barcode (ONT)
- 28 μL nuclease free water
- 5 μL from the first PCR

This was mixed and added to a thermocycler with a modified version of the rapid barcoding PCR (Table 2.4).

Table 2.4 – Modified version of the rapid barcoding PCR with fewer cycles

Cycles	Temperature	Time
1	98 °C	2 min
14	98 °C	15 sec
	56 °C	15 sec
	68 °C	45 sec
1	68 °C	4 min
1	4 °C	Hold

Final PCR products were sequenced as described in the Post-PCR pooling and SPRI cleaning section (2.2.3).

2.4 DNA QC and qPCR

2.4.1 Quantification and fragment size analysis

DNA was quantified using the Qubit dsDNA High Sensitivity assay kit. For every sample, 1 ul of dye was diluted in 199 uL of buffer, including standards. 190 ul of the working solution was aliquoted per standard and 199 uL per sample. 10 ul of standard 1 and 2 were added to their respective tubes and 1 ul of DNA/library was added to the Qubit tubes. Each tube was vortexed for 4 seconds and incubated at room temperature for 2 minutes. The readings were then recorded on a Qubit Fluorimeter 3.0. New standards were made every time the buffer was remade.

For library sizing, Tapestation 2200 (Agilent) was used with Genomic Screentape. The Tapestation buffer was left to equilibrate to room temperature and then 10 uL of buffer was aliquoted into Tapestation optical tubes (Agilent) for every sample, plus the ladder. 1 ul of DNA sample/library was then added per tube, including the ladder and vortexed briefly. The tubes were centrifuged and loaded in the Tapestation with the Genomic Screentape.

2.4.2 Quantitative PCR

qPCR was used to detect presence/absence of pathogens and to estimate bacterial/human DNA loss after host depletion.

Host depletion was estimated by calculating the delta-Cq between depleted samples and the non-depleted controls from a probe-based qPCR assay targeting human RNA polymerase II subunit A. For loss of bacteria, a SYBR green qPCR assay was used targeting bacterial 16S DNA.

The human probe assay PCR composition:

- 5 μL of Lightcycler 480 probe master (Roche)
- 1 μL of primer-probe mix (5 μM forward primer, 5 μM reverse primer, 2 μM probe)
- 3 μL of nuclease free water
- 1 μL of template,

  To a final volume of 10 μL

The bacterial 16S SYBR green PCR composition:

- 5 μL of Lightcycler 480 SYBR Green I Mastermix (Roche)
- 1 μL of primer mix (5 μM forward primer, 5 μM reverse primer)
- 3 μL of nuclease free water
- 1 μL of template

  To a final volume of 10 μL

qPCR was performed using a Lightcycler 480 system (Roche) (conditions Table 2.5)

Table 2.5 – qPCR conditions for the human/bacterial host depletion assay

Cycles	Temperature	Time
1	95 °C	5 min
40	95 °C	30 sec
	55 °C	15 sec
	72 °C	30 sec

For confirmation of presence/absence of organisms, qPCR was performed on extracts using species-specific probe assays (using conditions in Table 2.6).

The reaction composition:

- 10 μL of Lightcycler 480 probe master (Roche)
- 2 μL of species-specific primer-probe mix (5 μM forward primer, 5 μM reverse primer, 2 μM probe)
- 3 μL of nuclease free water

- 5 μL of template

  To a final volume of 20 μL

Table 2.6 – qPCR conditions for the bacterial detection probe assay

Cycles	Temperature	Time
1	95 °C	5 min
45	95 °C	15 sec
	72 °C	15 sec

All primers and probes are listed in Table 2.7.

Table 2.7 – Primers and probes used for qPCR assays

Target	Primers and probes (5'-3')	Reference
Human	Forward: TGAAGCCGTGCGGAAGG	[277]
	Reverse: ACAAGAGAGCCAAGTGTCG	
	Probe: [6FAM]-TACCACGTCATCTCCTTTGATGGCTCCTAT-[BHQ1]	
Bacteria 16S universal	Forward: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG	[278]
	Reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVG GGTATCTAATC	
HIV-1	Forward: CATGTTTTCAGCATTATCAGAAGGA	[279]
	Reverse: TGCTTGATGTCCCCCCACT	
	Probe: [6FAM]-CCACCCCACAAGATTTAAACACCATGCTAA-[BHQ1]	
Escherichia coli	Forward: CGATAATCGCCAGATGGC	[97]
	Reverse: CCTAAGTTGCAGGAGATGG	
	Probe: [6FAM]-TAGAGCGCCTTCGGTGTCGGT-[BHQ1]	
Staphylococcus aureus	Forward: ACTGTAACTTTGGCACTGG	[280]
	Reverse: GCAGATACCTCATTACCTGC	
	Probe: [6FAM]-ATCGCAACGACTGGCGCTA-[BHQ1]	
Pseudomonas aeruginosa	Forward: AGCCTTCCTGGTCCCCTTAC	[280]
	Reverse: CCTAATGAACCCCAGTGTATAAGTTTG	
	Probe: [6FAM]-TGAACTGACGGTCGCCAACGGTT-[BHQ1]	
Moraxella catarhalis	Forward: GGTGAGTGCCGCTTTTACAAC	[280]
	Reverse: TGTATCGCCTGCCAAGACAA	
	Probe: [6FAM]-TGCTTTTGCAGCTGTTAGCCAGCCTAAG-[BHQ1]	
Klebsiella pneumoniae	Forward: CGGGCGTAGCGCGTAA	[280]
	Reverse: GATACCCGCATTCACATTAAACAG	
	Probe: [6FAM]-CCCGGCATGGATCGTTCCGA-[BHQ1]	
Haemophillus influenzae	Forward: AGCGGCTTGTAGTTCCTCTAACA	[280]
	Reverse: CAACAGAGTATCCGCCAAAAGTT	
	Probe: [6FAM]-CGATGCTGCAGGCAATGGTGCT-[BHQ1]	
Streptococcus pneumoniae	Forward: GCTTATGGGCGCCAAGTCTA	[280]
	Reverse: CAAAGCTTCAAAAGCAGCCTCTA	
	Probe: [6FAM]-CTCAAGTTGGAAACCACGAGTAAGAGTGATGAA-[BHQ1]	

SYBR Green qPCR assays were designed for qPCR of *Imtechella halotolerans* (NZ_AJJU01000002.1) and *Allobacillus halotolerans* (NZ_JAHLZF000000000.1) by using Primer BLAST to generate primers for <150 bp targets.

2.4.3 RT-qPCR

For RNA extracted from viruses, and for RNA extracted from *E. coli,* RT-qPCR was used instead. This was performed using a One Step PCR using qPCRBIO Probe 1-Step Go (PCR Biosystems).

The RT-qPCR reaction composition was:

- 10 μL of 1-Step Go Mastermix
- 1 μL of RTase
- 2 μL Primer-probe mix (5 μM forward primer, 5 μM reverse primer, 2 μM probe)
- 5 μL nuclease free water
- 2 μL of template to a final volume of 20 μL

qPCR was performed on a Roche Lightcycler 480 (conditions Table 2.8).

Table 2.8 – RT-qPCR cycling conditions

Cycles	Temperature	Time
1	55 °C	10 min
1	95 °C	2 min
40	95 °C	5 sec
	60 °C	30 sec

2.5 Analysis of respiratory data

2.5.1 Raw read processing

Raw nanopore data was basecalled live on a MinIT (ONT) using high accuracy mode (Guppy version 3.0.6+9999d81]. Basecalled FASTQ files were demultiplexed using qcat

v1.1.0 with default parameters. Human reads were removed from demultiplexed FASTQ files using minimap2 aligning to the human hg38 genome (GRCh38.p13) with default parameters for long-read data (-a -x map-ont). Unassigned reads were exported to a BAM file and converted back to FASTQ format using SAMtools. These FASTQ files were then analysed to identify pathogens and antibiotic resistance genes.

2.5.2 Presence/absence analysis

For evaluation of the new rapid workflow, data was analysed using the new bespoke CLInical Metagenomics and AnTimicrobial rEsistance (CLIMATE) pipeline (described in the results section). Presence or absence of HIV-1 was tested by using What's In My Pot (v2.3.10) EPI2ME (ONT). These results were confirmed by mapping reads to the HIV-1 reference genome (AF033819.3) using minimap2 with default parameters and using the map-ont flag.

2.6. SARS-CoV-2 genomics

2.6.1 ARTIC LoCost SARS-CoV-2 cDNA synthesis and multiplex tiling PCR

cDNA and PCR reactions were broadly prepared using Version 3 (LoCost) of the ARTIC nCoV-2019 sequencing protocol [269] . 8 μL of RNA was added to 2 μL of LunaScript RT Supermix kit and mixed by vortexing. The reaction was incubated in a thermocycler at 25 °C for 2 minutes, 55 °C for 10 minutes and 95 °C for 1 minute, followed by a hold at 4 °C. Synthesised cDNA was stored at 4 °C in the short term, prior to use or -80 °C for long-term storage. If necessary, RNA samples were diluted prior to the cDNA synthesis reaction, C_T <15 was diluted 100-fold and C_T 15-18 was diluted 10 fold.

After cDNA synthesis, the V3 CoV-2 primers (Integrated DNA Technologies) were used to perform the multiplex PCR for SARS-CoV-2 according to the ARTIC protocol. For each sample, Pool 1 and Pool 2 reactions were prepared with the following composition:

- 5 μL 5X Q5 Reaction Buffer
- 0.5 μL 10 mM dNTPs
- 0.25 μL Q5 Hot Start DNA Polymerase
- 0.4 μL of the Pool 1 or 2 primers
- 16.35 μL of nuclease free water.
- 2.5 μL of cDNA from the previous step

This was mixed by vortexing and pulse centrifuged briefly and transferred to a thermocycler for PCR (conditions Table 2.9). PCR products were stored at 4 °C overnight (if necessary) and -20 °C for any longer period.

Table 2.9 – PCR conditions for the ARTIC PCR used in all SARS-CoV-2 sequencing experiments

Cycles	Temperature	Time
1	98 °C	30 seconds
35	98 °C	15 sec
	65 °C	5 min
1	4 °C	Hold

2.6.2 Preparation of CoronaHiT Barcodes

Corona High Throughput (CoronaHiT) barcodes were purchased as standard oligos in 100 uM concentrated stocks (Sigma). Diluted working stocks were made. In each well, 80 µL of nuclease free water was added and then for each barcode, 10 µl of the forward primer and 10 µl of the reverse primer were added (total 100 µL) and mixed well. Aliquots of the diluted barcodes were stored at -20 °C.

2.6.3 CoronaHiT library preparation for MinION sequencing (final method)

Optimisation of this final method is detailed in the results section. PCR products were diluted 1 in 5, by directly adding 2.5 µL of Pool 1 and 2.5 µL of Pool 2 to 20 µl NFW. Tagmentation was performed with the composition:

- 0.5 µL TB1 Tagmentation Buffer 1

- 0.5 µL BLT Bead-Linked Transposase (DNA Prep, (M), Illumina)
- 4 µL NFW
- 2 µL of 1/5 diluted PCR product

With multiple samples, a mastermix was made of the tagmented mix and aliquoted for each sample. The samples were pulse centrifuged briefly to avoid pelleting of the beads. The reaction conditions were 55 °C for 15 min in a thermal cycler (heated lid 65 °C) and cooled to 10 °C.

PCR barcoding was performed using Kapa 2G Robust PCR kit (Sigma) with the following composition:

- 4 μL Reaction buffer GC
- 0.4 μL dNTPs
- 0.08 μL Kapa 2G Robust Polymerase
- 7.52 μL PCR grade water
- 1 μL of working barcode stock (10 μM each)
- 12 μL tagmented sample

This was briefly mixed and pulse centrifuged and added to a thermocycler (conditions Table 2.10).

Table 2.10 – Barcoding PCR conditions CoronaHiT library preparation

Cycles	Temperature	Time
1	95 °C	1 min
14	95 °C	10 sec
	55 °C	20 sec
	72 °C	1 min
1	72 °C	3 min
1	4 °C	Hold

Following PCR, 2 µl of each sample was pooled in a 1.5 mL Eppendorf and 40 µL of this pool SPRI cleaned using 36 µL (0.8X) AMPure XP beads. The beads were incubated for 5 minutes to bind DNA, and this was followed by 2 washes using 200 µL 70% ethanol

without resuspension. For the experiment with a lower proportion of beads (0.6X instead of 0.8X) (described in results Section 3.4.4.1), 100 µL of the pool was washed with 60 µL AMPure XP instead (washes the same). Pools were eluted in 20 µl of EB (Qiagen). The barcoded pool was quantified using Qubit High Sensitivity (detailed in Section 2.4.1).

The SQK-LSK109 protocol for amplicon sequencing was then followed for nanopore sequencing. The end-prep reaction composition was:

- 7 µL Ultra II end prep buffer (NEB)
- 3 µL Ultra II end prep enzyme mix (NEB)
- 40 µL nuclease free water
- 10 µL of washed barcoded pool from the previous step

The reaction was incubated at room temperature for 15 min and then 65 °C in a thermocycler for 10 min (with heated lid at 75 °C), followed by a hold at 4 °C for at least 1 min. This was SPRI cleaned using 60 µL of AMPure Beads (1X) and two 200 µL 70% ethanol washes as before and eluted in 61 µL nuclease free water. Adapter ligation was performed with the reaction composition:

- 25 µL LNB (ONT)
- 10 µL NEBNext Quick T4 Ligase (NEB)
- 5 µL AMX (ONT) combined and mixed.
- 30 μL nuclease free water
- 30 μL end-prepped DNA

The reaction was incubated at room temperature for 20 min. After incubation, 40 µL AMPure XP beads were added, incubated for 10 minutes, followed by 2 washes using 250 µL SFB (ONT) with resuspension of beads both times followed by 3 minutes on the magnetic rack for pelleting. After the second wash, DNA was eluted in 15 µl of EB and incubated for 5 minutes before adding back to the magnetic rack and removing the supernatant (ONT). The final library was quantified using Qubit High Sensitivity and fragment size analysis using the Tapestation D5000 tape (detailed in Section 2.4.1).

37.5 µl SQB and 25.5 µl LB was added to 12 µL of the library and loaded on the MinION flowcell (R9.4.1) for sequencing.

2.6.4 CoronaHiT library preparation for Illumina sequencing

The libraries were performed like the CoronaHiT-ONT libraries up until the barcoding PCR, at which point CoronaHiT-ONT barcodes were switched for Nextera XT Index Kit barcodes (Sets A to D, Illumina). The PCR mastermix volumes were adjusted slightly accounting for barcode volume differences:

- 4 μL Reaction buffer GC
- 0.4 μL dNTPs
- 0.08 μL Kapa 2G Robust Polymerase
- 4.52 μL PCR grade water
- 2 μL P5 primer
- 2 μL P7 primer
- 12 μL tagmented sample

The PCR conditions were the same. Following barcode PCR, final libraries were prepared by Dave Baker to load on the NextSeq 500. 5 μL of each sample was pooled and 100 μL was SPRI cleaned with 0.8X AMPure XP beads (80 μL), and 2X 200 μL 80% ethanol washes. The final library was eluted in 50 μL of 10 mM Tris-HCL. The final library was quantified using the QuantiFluor ONE dsDNA system (Promega) according to manufacturer instructions and the size was determined using Tapestation D5000 (Agilent) to calculate the molarity. 1.5 pM was loaded on a 500 Mid Output v2 flowcell and sequenced using a Nextseq500.

2.6.5 Original ARTIC 1-24 sample library preparation (Nanopore)

For the initial comparison, the ONT version of the library preparation was followed post-PCR, “PCR tiling of COVID-19 virus” (revision E, released on 6th February 2020). Following the ARTIC cDNA synthesis and multiplex PCR (described previously in Section 2.6.1), the total volume of Pool 1 and Pool 2 PCR reactions were pooled (50 μL total) and SPRI cleaned using KAPA Pure Beads (Roche). 50 μL of KAPA Pure beads were added to each pool incubated for 5 minutes then added to a magnetic rack. The liquid was removed and the beads were washed twice with 200 μL of 80% ethanol, then eluted in 30 μL of 10 mM Tris-HCL and incubated for 2 minutes before adding to the magnetic rack and retaining the elution. The DNA was quantified using the QuantiFluor ONE dsDNA System. Each sample was normalised to 50 ng in 12.5 μL total volume. End-prep was then prepared with the following composition:

- 1.75 μl Ultra II end prep buffer (NEB)
- 0.75 μl Ultra II end prep enzyme mix (NEB)
- 12.5 μL of washed amplicon (50 ng total)

Samples were mixed by pipetting and incubated at 20°C for 5 min and 65°C for 5 min in a thermocycler. A tenth of this was used directly for native barcode ligation:

- 10 μL NEBNext Ultra II Ligation Master mix (NEB)
- 0.5 μL of NEBNext Ligation Enhancer (NEB)
- 2.5 μL of Native Barcode from EXP-NBD104 and EXP-NBD114 (ONT)
- 5.5 μL nuclease free water
- 1.5 μl end-prepped DNA,

  (Final volume 20 μl)

This was incubated in a thermocycler at 20°C for 20 min and 65°C for 10 min. All barcoded samples were pooled together (480 μL) and underwent a 0.4X AMPure SPRI clean. 192 μL of AMPure beads were added and incubated for 10 minutes, added to the magnetic rack and the solution removed, then the pellet was washed with two 700 μl SFB

washes and one 80% ethanol wash. DNA was then eluted in 35 µl of nuclease free water. Adapter ligation was performed using 30 µL of the washed pool. The composition was:

- 5 µl Adapter Mix II (ONT)
- 10 µL NEBNext Quick Ligation Reaction Buffer (5X)
- 5 µl Quick T4 DNA Ligase,
- 30 µL pooled washed barcoded amplicons

The ligation reaction was incubated at room temperature for 20 min and washed with 0.4X AMPure beads (20 µL beads to 50 µL). This was washed twice with 125 µL SFB (ONT) and the library was eluted in 15 µL EB (ONT). The library was quantified with Qubit High Sensitivity and 20 ng of the adapted library was used for final loading.

2.6.6 ARTIC LoCost library preparation

The final comparison was performed following the nCoV-2019 LoCost (V3) sequencing protocol [269]. 2.5 µL of Pool 1 and 2.5 µL of Pool 2 were added to 45 µL nuclease free water. End prep was performed on the dilution per sample with the composition:

- 1.2 µL Ultra II end prep buffer (NEB)
- 0.5 µL Ultra II end prep enzyme mix (NEB)
- 5 µL nuclease free water
- 3.3 µL dilution from previous step

The reaction was incubated at room temperature for 15 min and 65 °C in a thermocycler for 15 minutes. Then native barcode ligation was prepared:

- 5 µL Blunt/TA Ligase Master Mix (NEB)
- 1.25 µL native barcode (NBD-196, ONT)
- 3 µL nuclease free water
- 0.75 µL end-prepped DNA from previous step

The reaction was incubated at room temperature for 20 minutes and then 65 °C in a thermocycler for 10 min. Samples were pooled together and SPRI cleaned (0.4X) using AMPure XP beads. The beads were washed twice with 250 μL of SFB with resuspension and then one 70% ethanol wash without resuspending and eluted in 30 µl of EB (Qiagen). Adapter ligation was performed with the following composition:

- 5 μL Adapter Mix II (ONT)
- 10 μL NEBNext 5X Quick Ligation Reaction Buffer (NEB)
- 5 μL Quick T4 DNA Ligase (NEB)
- 30 μL barcoded amplicon pool

This was incubated at room temperature for 20 minutes. The adapter ligated pool was SPRI cleaned with 1X AMPure XP Beads and washed twice with 250 μL of SFB (ONT) with resuspension of beads. The final library was eluted in 15 μL of EB (ONT) and quantified by Qubit high sensitivity. 15 ng of the adapted library was loaded for MinION sequencing.

2.6.7 SARS-CoV-2 genome analysis

Basecalling of all nanopore data was performed using Guppy v.4.2.2 (ONT) with high accuracy mode (model dna_r9.4.1_450bps_hac) and demultiplexed using guppy_barcoder v.4.2.2 (ONT) with 'require_barcode_both_ends'. For CoronaHIT-ONT data, a custom arrangement of the barcodes was used due to the different barcodes [281].

Basecalled data was run through the various pipelines internally by Thanh Le Viet (QIB). The ARTIC pipeline (v1.1.3) [282] was used to generate consensus sequences for ARTIC comparison data while an internally modified version [283] of the pipeline was used to generate a consensus sequence for CoronaHiT-ONT data. For Illumina data, raw reads were demultiplexed using bcl2fastq v.2.20 (Illumina). Illumina data was trimmed and a consensus built using a modified version of iVAR v.1.2.3 [284][285].

Phylogeny trees were generated from consensus genomes by Leonardo de Oliveira Martins (QIB). A multiple FASTA alignment was created by aligning all samples to the reference genome MN908947.3 with MAFFT v7.470 and a maximum likelihood tree as created using IQTREE2 v2.0.4 under the HKY model, and branches smaller than 10−7 were collapsed into a polytomy. SNPs were identified using SNP-sites v2.5.1 and the tree was visualised with FigTree v1.4.4.

3. Results

3.1. Optimisation of the respiratory metagenomics method

We previously developed a metagenomic method for bacterial and fungal detection from respiratory samples, which I contributed to with analysis and experimental planning [97] (Appendix 2). This method involves performing a saponin-based depletion to remove human DNA, followed by microbial DNA extraction. The DNA is prepared for sequencing using a modified version of ONT's Rapid PCR Barcoding library preparation kit (RPB004) which barcodes and adapts the DNA using transposase, followed by single primer PCR from the adapter sequence to amplify the library (detailed in section 2.2.1). The library is sequenced on a MinION for 2 hours before the data is analysed and pathogens and AMR genes are identified. Sequencing can be performed for longer to acquire more data for genomic epidemiology applications.

The respiratory metagenomics workflow takes 7 hours from sample to result if a DNA clean-up step is included after extraction or ~6.5 hours if the clean-up is omitted (however, this step typically is included to concentrate low biomass samples). However, for widescale adoption of CMg in practice, these methods need to be faster and simpler. The three longest steps of our CMg method are the PCR, which takes 2 hours and 30 minutes (depending on PCR machine), the sequencing, which we perform for 2 hours, and the host depletion, which takes 45 minutes, and which is the most variable step and a common source of human error. The aim was to simplify these steps by reducing complexity and/or turnaround time. The PCR step is where the most gains can be made in terms of time, so this was optimised first.

3.1.1. PrimeSTAR Max polymerase testing

LongAmp Taq (New England Biolabs), hereafter referred to as LAT, is a long-range polymerase recommended by ONT in the RPB004 library preparation kit as described in section 2.21. This kit was chosen as it is suitable for low biomass samples (recommended for ≥5ng input DNA, but capable of detecting 5-50pg in our hands) and can multiplex up to

12 samples (we typically multiplexed 6 samples per library). The use of a transposase to add adapters produces fragments of varying lengths ranging from hundreds to thousands of bases. Therefore, a long range polymerase is required for PCR amplification and the ONT recommended LAT enzyme requires a long extension time (50 seconds per kb) to amplify the longer fragments efficiently. ONT recommends a 6 minute extension time, however, this was previously reduced to 4 minutes in our group, leading to a reduction in the size of reads [97]. This 25 cycle PCR takes 2.5 hours even with the 2 minute reduction of the elongation step. Therefore I searched for long range PCR polymerases with faster processing speed than LAT that had similar sensitivity.

PrimeSTAR Max DNA Polymerase (Takara Bio Inc.) was selected to be tested as an alternative due to having an elongation factor that makes it faster. This enzyme was tested to determine if it was capable of amplifying DNA from clinical extracts using a short extension time and if the products could be sequenced using the RPB004 kit. A previously host depleted DNA extract from sputum, Test Sample 1 (T1), with known DNA concentration and microbial profile was amplified using PrimeSTAR Max Polymerase. This sample was tested using the standard approach and contained *S. pneumoniae* and *H. influenzae*.

The sample was divided and amplified using PrimeSTAR with 3 different extension times at 72°C :

A) 2 minute extension

B) 1 minute extension

C) 30 second extension

All other temperatures and times were kept the same as the default LAT method (detailed in Section 2.1). Quantification of PCR products indicated that all 3 tests produced PCR products. Post-PCR concentrations were 22.5 ng/μL, 20.1 ng/μL and 19.3 ng/μL for the 2 minutes, 1 minute and 30 second extensions respectively. Analysis by Tapestation (capillary electrophoresis) confirmed that the PCR had worked and that the peak for the 2 minute extension was 5,841 bp (Figure 3.1).

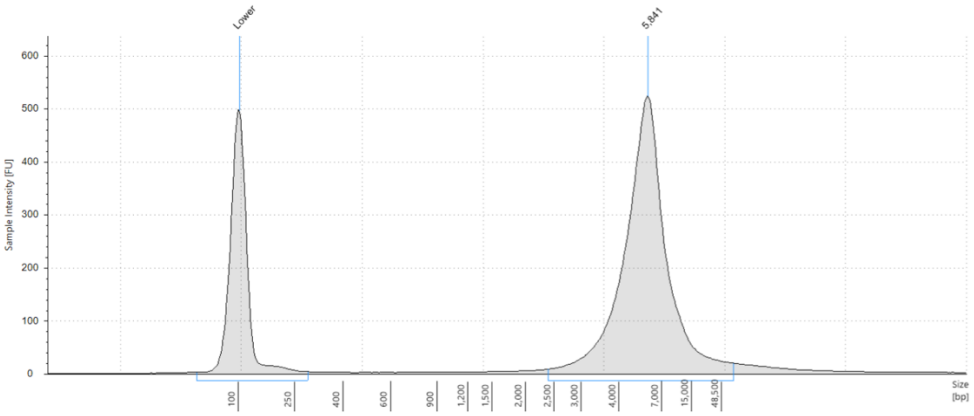


Figure 3.1 – Fragment size analysis result for PCR products using PrimeSTAR Max with a 2 minute extension.

Samples were multiplexed and sequenced using a MinION to check that the rapid chemistry of the RPB004 kit worked with the amplified products and that the expected organisms were detected. The sequencing run produced 1.56 million reads and 5.84 Gb of data in 22 hours. A 2-hour subset of this data was analysed using the WIMP workflow on EPI2ME (Table 3.1). The number of reads on average per sample was 21,000, which is approximately 50% lower than what was observed for the same sample using the standard approach [97], where on average, each sample produced 41,000 reads after 2 hours (however, this was highly variable with a range of 2300-108,600 reads per sample).

Table 3.1 – Metrics for the sequencing output of the 3 extension times tested using PrimeSTAR Max polymerase

	Number of reads after 2 hours	**Mean read size**	**Number of pathogen reads (proportion of reads)**
A) 2 minute extension	20,563	5,087	*H. influenzae:* 5,473 (26.6%) *S. pneumoniae:* 796 (3.9%)
B) 1 minute extension	19,389	4,940	*H. influenzae:* 5,593 (28.8%) *S. pneumoniae:* 724 (3.7%)
C) 30 seconds extension	23,214	3,996	*H. influenzae:* 7,158 (30.8%) *S. pneumoniae:* 832 (3.6%)

The correct pathogens (*H. influenzae* and *S. pneumoniae*) were identified by sequencing with similar read numbers and proportions using the different extension times, however,

there was a reduction in the mean read size below 1 minute, dropping from 4,940 kb to 3,996 at 30 second extension.

These conditions were further tested against LAT, using *E. coli* DNA. *E. coli* DNA was serially diluted, tagmented and split in two aliquots for testing with a final input of 250, 50, 25, 5 and 2.5 pg in the PCRs. The 1 minute extension was tested for PrimeSTAR Max whereas LAT was tested with default conditions (Section 2.1).

PrimeSTAR Max resulted in a lower yield across all input concentrations (Figure 3.2). At 250 pg (approximately 5 x 10^4 *E. coli* cell equivalents), LAT produced 27.5 ng/μL of DNA compared to 3.15 ng/μL for PrimeSTAR Max. At 25 pg (5 x 10^3 *E. coli* cell equivalents), LAT produced 5.11 ng/μL while PrimeSTAR Max produced 0.53 ng/μL which was similar to the negative control at 0.31 ng/μL. This suggested that the limit of detection (LOD) for the faster enzyme was likely to be lower than LAT.

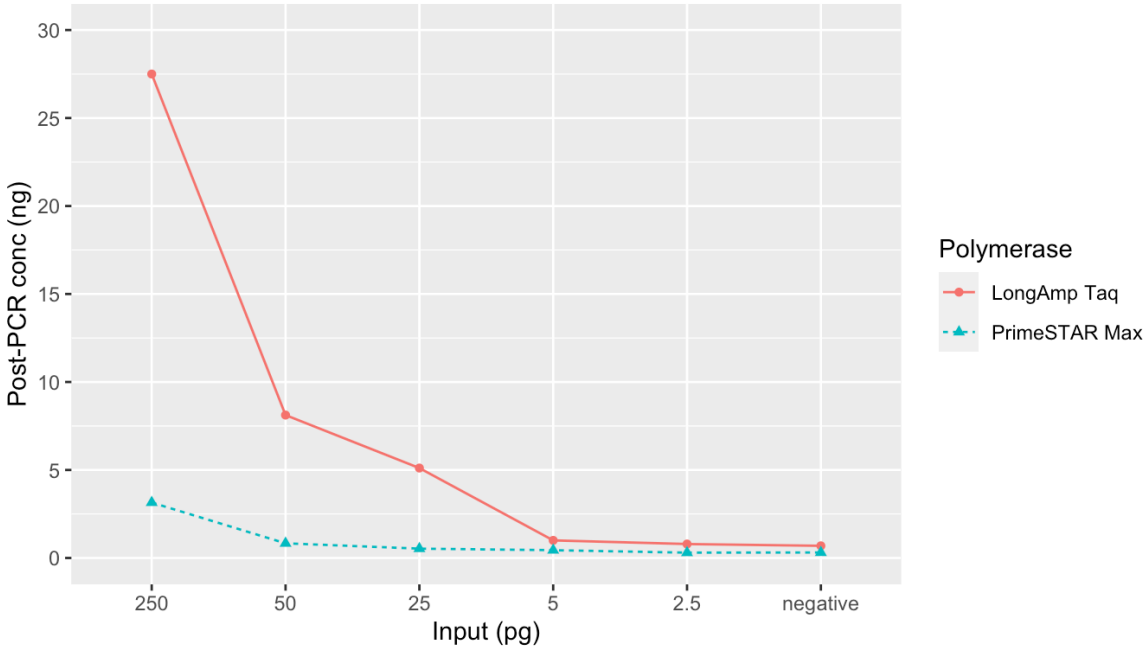


Figure 3.2 – DNA concentrations of the PCR products using LAT polymerase and PrimeSTAR Max with different DNA inputs

3.1.2. PrimeSTAR Max and LAT polymerases combined

To improve the LOD, an experiment was performed to test if the two polymerases could work together, providing both sensitivity and a faster PCR. Enzyme blends are commonly

used to combine different features of different polymerases [286]. The LAT polymerase was tested with PrimeSTAR Max to provide sensitive amplification in the early cycles, followed by fast PCR in the latter cycles.

LAT was used at manufacturer recommended concentrations (10 units per 100 μL) in the PrimeSTAR Max mastermix and also at a reduced concentration (2.5 units per 100 μL). This was tested on a clinical sample extract, T2, which contained *Moraxella catarrhalis.* DNA was diluted to 1 ng/μL to test if the PCR would work. The PCR was divided into two stages, with the first 5 cycles favouring LAT conditions and remaining 20 cycles favouring the faster PrimeSTAR Max conditions, taking approximately 1 hour 30 minutes (Table 3.2).

Table 3.2 – Cycling conditions for the 2 enzyme PCR

Number of cycles	PCR step	Temperature	Time
1 cycle	Initial Denaturation	95 °C	3 min
Stage 1 (5 cycles)	Denaturation	95 °C	15 sec
	Annealing	56 °C	15 sec
	Extension	65 °C	4 min
Stage 2 (20 cycles)	Denaturation	95 °C	15 sec
	Annealing	56 °C	15 sec
	Extension	72 °C	1 min
1 cycle	Final Extension	65 °C	4 min
1 cycle	Storage	4 °C	Hold

LAT and PrimeSTAR Max-only PCR reactions were performed as controls. Post-PCR quantification results showed that LAT and PrimeSTAR Max performed similarly, producing 34.6 ng/μL and 33.1 ng/μL respectively. The 2 enzyme PCR performed similarly when using 10 units of LAT, yielding 30 ng/μL, but did not perform as well when using the lower enzyme concentration, giving 17.6 ng/μL. Tapestation results showed that the average read size was higher for the 2 enzyme PCRs (5385-5598 bp) compared to LAT (2455 bp) and PrimeSTAR Max (4108 bp) and the range of sizes was narrower (Figure 3.3). Sequencing confirmed the presence of *M. catarrhalis* in all samples, however, the 2 enzyme PCR produced fewer reads overall. LAT produced 13,623 *M. catarrhalis* reads out of 14,252 classified reads (95.5%), PrimeSTAR Max produced

10,623 out of 11,150 classified (95.2%), while the hybrid tests produced 3,685 out of 3,898 (94.5%) and 2,391 out of 2,513 (95.1%) *M. catarrhalis* reads for the high and low LAT concentrations respectively, suggesting that most of the PCR-products could not be sequenced or that the PCR product quantification wasn't accurate by Qubit. In fact, the Tapestation quantification was higher for the enzymes on their own, which might explain the difference in sequencing yield.

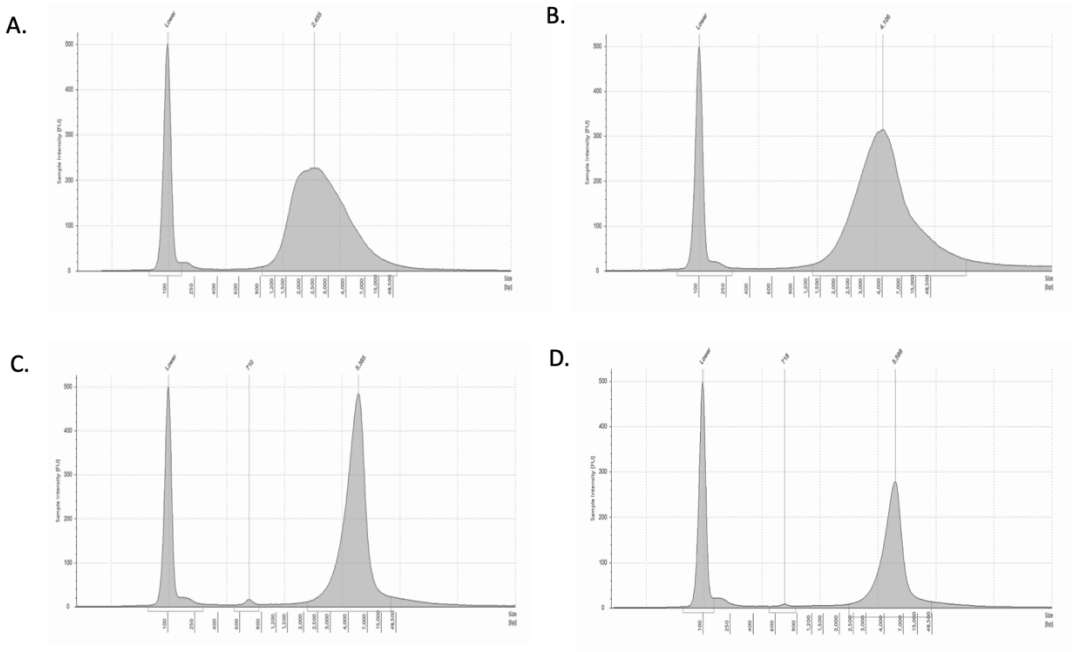


Figure 3.3 – Size distribution of reads using the enzymes on their own versus when used in combination. A) LAT on its own B) PrimeSTAR Max on its own C) Hybrid PCR with 10 units of LAT D) Hybrid PCR with 2.5 units of LAT.

We then investigated whether the faster 2 enzyme PCR improved the LOD compared to PrimeSTAR alone. The 2 enzyme PCR (with 10 units of LAT) and LAT alone were tested at lower concentrations using diluted T2 clinical sample (50 pg and 10 pg inputs). Data from 2 hours of sequencing showed that *M. catarrhalis* reads were the predominant classified reads in all samples, with >20,000 reads of *M. catarrhalis* at 50 pg for both LAT and the hybrid test, however, at 10 pg, LAT produced 48,940 reads whereas the hybrid PCR only produced 3 reads, similar to the negative control which produced 1 read (Table 3.3).

Table 3.3 – Number of reads of *M. catarrhalis* using EPI2ME WIMP for LAT versus the hybrid PCR conditions

	Number of *M. catarrhalis* reads using LAT	**Number of *M. catarrhalis* reads using the hybrid PCR**
50 pg	31,073 (94%)	25,767 (93.4%)
10 pg	48,940 (94%)	3 (100%)
Negative	1 (100%)	1 (100%)

A further test was performed increasing the number of cycles in the first stage of the hybrid PCR by 3 cycles. This increased the hybrid PCR time to 1 hour and 45 minutes (compared to 2 hours and 30 minutes for the default LAT). After 2 hours of sequencing, LAT produced 8,180 reads of *M. catarrhalis* while the hybrid PCR only produced 7 reads of *M. catarrhalis*. This suggested that the 2 enzyme PCR didn't improve the LoD compared to PrimeSTAR alone, so this approach was abandoned.

3.1.3. GXL polymerase

An alternative polymerase was selected for testing. A long-range polymerase, GXL, (Takara Bio Inc.), as it was described as having higher processivity than LAT and was provided as individual PCR components rather than as a master mix, allowing more optimisation. The manufacturer recommends that for faster PCR, double the polymerase concentration can be used, with potential extension rates of 10 seconds/kb.

GXL was tested on sample T2 at 10 pg concentration to test if there was an improvement in LOD and to confirm if it was compatible with the RPB004 library preparation kit. Double polymerase concentration was used (2.5 Units per μL instead of 1.25 Units) with a short extension time of 45 seconds, (at 10 sec/kb, fragments up to 4.5 kb could be amplified). A longer 1 minute and 15 seconds extension was also tested (Table 3.4).

Table 3.4 – The two cycling conditions tested for the new enzyme with different extension times

Number of cycles	PCR step	Temperature	Time
1 cycle	Initial Denaturation	98 °C	2 min
25 cycles	Denaturation	98 °C	15 sec
	Annealing	56 °C	15 sec
	Extension	68 °C	45 seconds or 1 min 15 seconds
1 cycle	Final Extension	68 °C	4 min
1 cycle	Storage	4 °C	Hold

Post-PCR concentrations were low as expected near the limit of detection; 0.220 ng/μL for double GXL with the 45 second extension and 0.226 ng/μL for the 1 minute 15 second extension. The samples were pooled in equimolar amounts and sequenced for 2 hours to determine the number of reads. Both samples had *M. catarrhalis* reads in higher quantities than the 2 enzyme approach. The 1 minute 15 second extension PCR had 835 reads of *M. catarrhalis* (92.3% of classified reads), and there were *M. catarrhalis* 493 reads (94.8%) for the 45 second extension (Table 3.5).

Table 3.5 – Sequencing metrics with the new GXL polymerase

Test	PCR time	*M. catarrhalis* reads	Average read size (Kb)	Yield (Mb)
GXL 1 minute 15 seconds	1 hour 15 minutes	835 (92.3%)	3.2	2.7
GXL 45 seconds	1 hour	493 (94.8%)	3.1	1.5
Negative control	1 hour	0 (0%)	N/A	N/A

Showing more promise than previous enzyme, the new polymerase and conditions were tested against LAT with a range of DNA inputs to determine an analytical LOD. The number of cycles was increased from 25 to 35 in an additional test to see if this would further improve the LOD. A 25 cycle PCR with GXL and a 45 second extension takes 1 hour, which is a 1 hour and 30-minute time saving compared to LAT. A 35 cycle PCR with GXL is 1 hour and 15 minutes in total, which is still a significant improvement compared to LAT. Culture extracts of both *E. coli* and *S. aureus* were used for the test, both were

combined at approximately equal DNA concentrations and diluted to test inputs of 100, 50, 10, and 5 pg for the LOD experiments.

Post-PCR quantification of the products showed that GXL produced higher concentrations than LAT, particularly at higher inputs. 35 cycles with GXL produced higher concentrations than 25 cycles (Figure 3.4). This is in contrast to the originally tested PrimeSTAR Max, which gave lower concentrations than LAT.

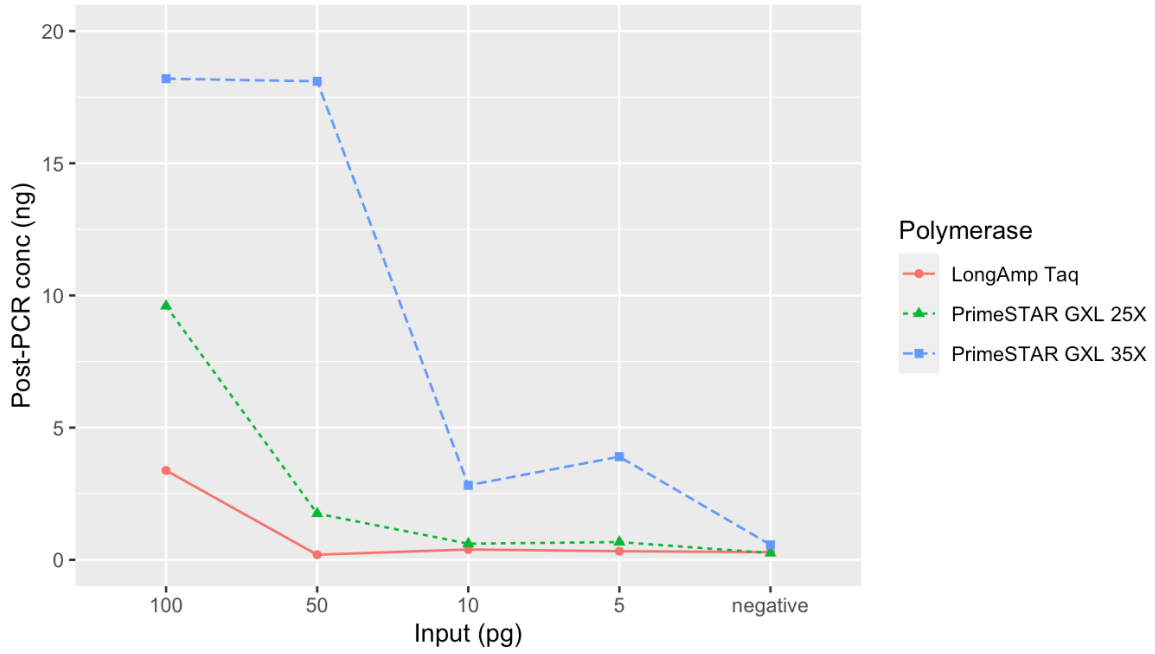


Figure 3.4 – DNA concentrations PCR products produced using LAT and GXL (25 and 35 cycles) with different DNA inputs

All samples were sequenced, however, due to limited RPB004 barcodes (n=12), had to be sequenced in two runs. GXL with 25 cycles (GXL-25) and 35 cycles (GXL-35) were sequenced on two separate runs, with LAT on both runs as a control. Sequencing results verified post-PCR concentrations, with GXL giving the most reads across the range of inputs (Table 3.6A and 3.6B). For GXL-25, *E. coli* and *S. aureus* were both detectable down to 5 pg, similar to LAT, but produced more reads across all inputs. At 100 pg input, GXL-25 produced 111,779 reads in total compared to 38,127 for LAT, and at 5 pg, produced 2,315 reads compared to 746 for LAT. GXL-35 produced considerably more data, even at the low concentration end, with 32,905 reads in total from 5 pg of input, the

majority of which were *E. coli* and *S. aureus* as expected. GXL-25 and LAT produced very similar read sizes, particularly at high inputs (2.4 kb average at 100 pg for both), and GXL-35 produced slightly longer reads than LAT (2.9 kb average at 100 pg).

More cycles with GXL led to a higher levels of contamination. With GXL-25, there were 24 reads in total produced in the negative control and 0 reads were *E. coli* or *S. aureus.* With LAT, there were a total of 28-40 reads in the negative control in the two runs, with 2-7 *E. coli* reads and 0-2 *S. aureus* reads. In contrast, there were 1,692 reads in the negative control for GXL-35, with 30 *E. coli* reads, 1.8% of reads. We have previously used a 1% abundance threshold for calling pathogens, so this could be problematic.

The proportion of *E. coli* and *S. aureus* reads were consistent between GXL-25 and LAT, especially at higher inputs. At 100 pg input, *E. coli* was 52.9% of reads for GXL-25 and 51.5% for LAT, and *S. aureus* was 15% for GXL-25 and 15.9% for LAT. In contrast, GXL-35 proportions diverged more from LAT. At 100 pg input, *E. coli* was 42.9% of reads for GXL-35 and 50.6% of reads for LAT, and *S. aureus* was 26.5% of reads for GXL-35 and 16% of reads for LAT. More cycles led to *S. aureus* being a higher proportion of the total reads. As the *E. coli* genome is approx. double the size of *S. aureus*, the GXL-35 results reflect organism abundance most accurately with approx. double the number of *E. coli* reads compared to *S. aureus*.

These results show that GXL performs similarly, if not better than LAT when 25 cycles are used, with similar read sizes and similar proportions of reads for the organisms down to 5 pg, unlike previous enzymes tested. GXL with 35 cycles has a higher yield, however, is likely to be more prone to contamination and PCR bias. The 25-cycle PCR, with the conditions described in Table 3.4, and a 45 second elongation (rather than 1 minute 15 second) was taken forward for further evaluation. This PCR takes 1 hour.

Table 3.6A – Sequencing metrics for the first run comparing GXL polymerase with 25 cycles versus the default LAT polymerase

Sequencing Run 1 – 25 cycle test						
	GXL (25 cycles)			**LAT**		
Input (pg)	**Total number of reads**	**Mean read size (kb)**	***E. coli* and *S. aureus* reads (proportion)**	**Total number of reads**	**Mean read size (kb)**	***E. coli* and *S. aureus* reads (proportion)**
100	111,779	2.4	*E. coli*: 59,217 (52.9%) *S. aureus*: 16,818 (15.0%)	38,127	2.4	*E. coli*: 19,662 (51.5%) *S. aureus*: 6,096 (15.9%)
50	23,316	2.2	*E. coli*: 11,084 (47.5%) *S. aureus*: 4,240 (18.2%)	1,225	2.1	*E. coli*: 556 (45.4%) *S. aureus:* 205 (16.7%)
10	3,273	2.2	*E. coli*: 1,562 (47.8%) *S. aureus*: 538 (16.4%)	1,055	2.0	*E. coli*: 504 (47.8%) *S. aureus*: 139 (13.2%)
5	2,315	2.1	*E. coli*: 982 (42.4%) *S. aureus*: 430 (18.6%)	746	1.7	*E. coli*: 294 (39.4%) S. aureus: 98 (13.1%)
0	24	3.1	*E. coli*: 0 (0%) *S. aureus*: 0 (0%)	40	2.4	*E. coli*: 7 (17.5%) *S. aureus:* 0 (0%)

Table 3.6B - Sequencing metrics for the second run comparing GXL polymerase with 35 cycles versus the default LAT polymerase

Sequencing Run 2 – 35 cycle test						
	GXL (35 cycles)			**LAT**		
Input (pg)	**Total number of reads**	**Mean read size (kb)**	***E. coli* and *S. aureus* reads (proportion)**	**Total number of reads**	**Mean read size (kb)**	***E. coli* and *S. aureus* reads (proportion)**
100	66,041	2.9	*E. coli*: 28,344 (42.9%) *S. aureus*: 17,508 (26.5%)	29,642	2.3	*E. coli*: 14,989 (50.6%) *S. aureus*: 4,733 (16.0%)
50	89,225	2.6	*E. coli*: 37,915 (42.5%) *S. aureus*: 22,999 (25.8%)	924	2.1	*E. coli*: 431 (46.6%) *S. aureus*: 170 (18.4%)
10	23,411	2.7	*E. coli*: 8,621 (36.8%) *S. aureus*: 6,280 (26.8%)	794	2.0	*E. coli*: 381 (48.0%) *S. aureus*: 113 (14.2%)
5	32,905	2.6	*E. coli*: 10,521 (32.0%) *S. aureus*: 8,732 (26.5%)	637	1.7	*E. coli*: 245 (38.5%) *S. aureus*: 91 (14.3%)
0	1,692	3.2	*E. coli*: 30 (1.8%) *S. aureus*: 0 (0%)	28	2.6	*E. coli*: 2 (7%) *S. aureus*: 2 (7%)

The PCRs were performed on two separate PCR machines, so to check that there was no significant performance difference between machines, a sample was tested on the two machines under identical conditions. The new GXL PCR was tested in duplicate on clinical sample T3 at a low concentration. After the PCR, the samples were quantified, giving almost identical concentrations for both machines: 9.93 ng/μL on PCR Machine 1 and 9.92 ng/μL on PCR Machine 2.

3.1.4. Host depletion

The host depletion is another step that takes a significant amount of time and is a manual process with multiple steps. The method our group developed and published in Nature Biotechnology [97] was a differential lysis approach using saponin which was based on similar methods described in the literature. This method took approx. 45 minutes with

separate incubations for differential cell lysis using saponin and human DNA digestion with HL-SAN nuclease. We wanted to simplify this method so tested the potential to combine the cell lysis and nuclease treatment steps into one step and to remove some of the other steps in the procedure that added time and complexity but may not improve performance, including the osmotic shock step and a pellet wash. These changes combined reduced the host depletion time by approximately 30 minutes compared to the published method. Table 3.7 shows the differences between the two protocols with the steps that were removed and the remaining timed steps, i.e., incubations and centrifugation times. Non-timed steps in between these steps (i.e., pipetting, transferring) will take a variable amount of time depending on the samples and user, taking seconds for single samples but longer for more. The total timesaving of the one-pot method is therefore an underestimate, as the reduction in the number of steps also reduces manipulation time, which is useful when multiple samples are being processed.

Table 3.7 – The differences between the published method and the one pot method. Steps that were removed for the One pot method are indicated. Times are for timed steps such as incubations and centrifugations.

Published method		**One pot method**	
Step	**Time (min)**	**Step**	**Time (min)**
Centrifuge sputasol treated sample at 8000 xg	5	[Step Removed]	
Discard supernatant, resuspend in 250 μL PBS		[Step Removed]	
Add 200 μL of 5% saponin to resuspended pellet		Add 40 μL of 1% saponin directly to 200 μL sample	
Incubate at room temperature	10	[Step Removed]	
Add 350 μL water, wait 30 seconds	0.5	[Step Removed]	
Add 12 μL 5M salt		[Step Removed]	
Centrifuge 6000 xg	5	[Step Removed]	
Discard supernatant, resuspend in 100 μL PBS		[Step Removed]	
Add 100 μL HL-SAN buffer		Add 200 μL HL-SAN buffer	
Add 10 μL HL-SAN		Add 10 μL HL-SAN	
Incubate at 37 °C 800 rpm	15	Incubate at 37 °C, 1000 rpm	10
Add 800 μL PBS		Add 1 mL PBS	
Centrifuge 6000 xg	3	Centrifuge 12000 xg	3
Discard supernatant, resuspend in 1 mL PBS		[Step Removed]	
Centrifuge 6000 xg	3	[Step Removed]	
Discard supernatant and start lysis/extraction		Discard supernatant and start lysis/extraction	
Total timed steps:	**41.5**	**Total timed steps:**	**13**

This new ‘one pot’ method showed promise on one sample (tested by Gemma Kay, QIB, results not shown) and to confirm that the method performed well and was reproducible, it was tested on more sputum samples, as well as with a higher saponin concentration and compared to the published method. Three excess sputum samples (T4, T5 and T6) positive for pathogens using culture methods at the NNUH were depleted using 1) the one pot method (0.1% final saponin concentration), 2) the one pot method using 5 times the concentration of saponin (0.5% final) 3) the previously published method.

Host depletion with all 3 methods resulted in a qPCR C_T >35 for all samples (Table 3.8). The highest depletion was 10.76 ΔC_T (the difference in human qPCR C_T between the depleted sample and the undepleted control), which is approximately a 1000-fold

depletion. The lowest was 5.33 ΔC_T (~40-depletion), however, this is because the amount of human DNA in this sample was high to begin with. In terms of loss of bacteria, ΔC_T was <1, (less than 50% loss of bacteria), with the exception of the published method in sample 2 which led to a 6.52 C_T loss of bacteria. There was no difference in performance between the different concentrations of saponin tested.

Sample T6 was mucoid despite prior sputasol treatment by routine microbiology. For the depletion experiments, T6 was re-treated with sputasol to homogenise the sample and reduce viscosity. As a test, prior to sputasol treatment, an aliquot of the sample was taken and the one pot depletion method was performed on the non-treated viscous sample. qPCR results showed that the sputasol-treated sample had a human C_T of <35 whereas the non-treated depleted sample had a C_T of 30.21, giving a host depletion ΔC_T of only 0.54. The viscosity had a negative impact on the host depletion, supporting previous experience that reducing viscosity is important for effective host depletion.

Table 3.8 – Host depletion results on 3 sputum samples using ΔC_T of qPCR assays for human and bacterial 16S pre- and post-depletion

Sample	Test	Non-depleted human C_T	Depleted human C_T	ΔC_T human	Non-depleted bacteria C_T	Depleted bacteria C_T	ΔC_T bacteria
T4	One pot 0.1%	24.24	>35	10.76	17.88	17.64	-0.24
	One pot 0.5%		>35	10.76		17.91	0.03
T5	One pot 0.1%	25.36	>35	9.64	17.43	18.04	0.61
	One pot 0.5%		>35	9.64		17.97	0.54
	Previous method		>35	9.64		23.95	6.52
T6	One pot 0.1%	29.67	>35	5.33	24.48	24.75	0.27
	One pot 0.5%		>35	5.33		24.07	-0.41
	Previous method		>35	5.33		24.55	0.07

The new ‘one pot’ method performed as well as the published method but was much faster and easier to perform. Hence this method was used for the respiratory

metagenomics method moving forward. A patent application was written and submitted covering the novel 'one pot' method (PCT/GB2020/052986) which I am a co-inventor of (Appendix 3).

3.1.5. Rapid CMg workflow on Flongle

Flongle flowcells were released by ONT in March 2019 offering lower capacity sequencing at a lower cost ($90USD). Flongle flowcells can be used as an alternative to MinION flowcells to run fewer samples at once, which reduces/removes the necessity for batching, allowing more rapid diagnostic use in a clinical setting. Initial testing of early access Flongle flowcells was hampered by poor quality, with the average available pore count being ~40 out of a possible 126. Additionally, the RPB004 library prep kit was not initially supported on the Flongle so had to be tested. A clinical sample extract, T7, was quantified and tested using the new faster PCR protocol (detailed in Section 3.1.3). A flowcell with a higher number of available pores (n=70) was selected for testing. 100 fmol of DNA was loaded on the flowcell and sequencing was performed for 2 hours. The duty plot indicated that the sequencing worked successfully (Figure 3.5), with a yield of 200 Mb in 2 hours (Flongles are advertised as capable of generating 1+ Gb in 24 hours).

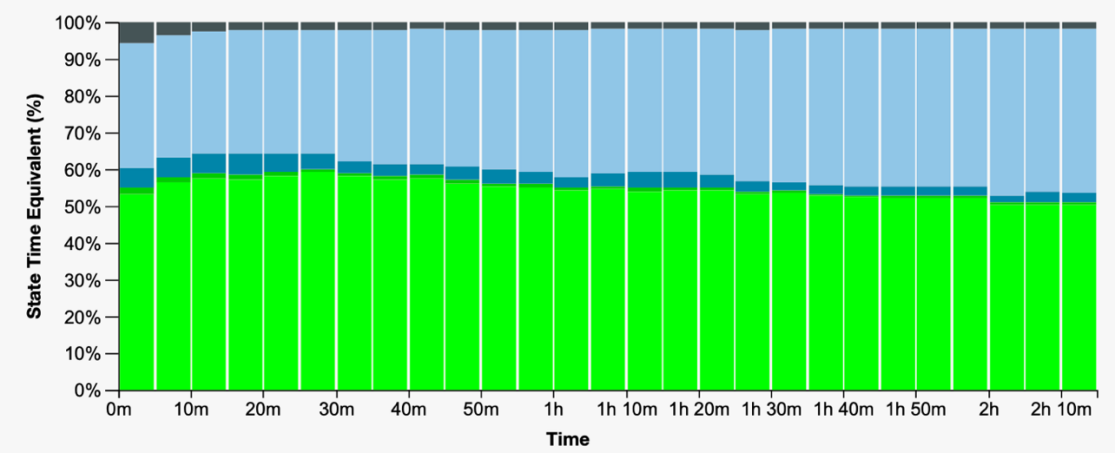


Figure 3.5 – Sequencing duty plot indicating consistent pore occupancy over 2 hours, with a slight decline over time.

The optimised host depletion and PCR steps were then combined and tested on a fresh respiratory sample acquired from NNUH microbiology lab (T8) and sequenced on a Flongle. The one pot host depletion was performed using 0.1% concentration saponin, and the PCR was performed using the 45 second extension with 25 cycles. The sample was SPRI cleaned and concentrated (as was performed for the published CMg method) before performing the rapid PCR.

The post-PCR concentration was 30.2 ng/μL. The sample was sequenced for 2 hours on a Flongle and produced 51,000 reads with an average length of 2,810 bp. 47,703 reads were classified by EPI2ME, with 29,612 (62%) reads for *P. aeruginosa* and 289 reads (0.6%) of reads for human., suggesting the infection was *P. aeruginosa* (clinical sample result was unknown) and that the host depletion step and the entire process worked very well.

3.1.6. Optimisation of the bioinformatics pipeline

The EPI2ME (ONT) analysis pipeline was used for the published CMg method. EPI2ME is cloud-based, which means that it requires uploading potentially sensitive clinical sequence data to the cloud, which can limit clinical use. Additionally, it can be unreliable and slow during times of high traffic and is largely a black box tool that can be changed or discontinued by ONT, so isn't suitable for clinical use. As an alternative, a basic customised pipeline was developed with Riccardo Scotti (Post-Doc in O'Grady Lab). The CLIMATE pipeline (CLInical Metagenomics and AnTimicrobial rEsistance) was developed using Galaxy, and includes a new AMR resistance filtering tool, Scagaire [287], developed by Andrew Page (Head of Informatics at QIB).

The pipeline takes demultiplexed reads with and first performs a quality control step using fastp to remove reads with a Phred score <7. For the taxonomic assignment step, Kraken 2 is used as it provides a fast classification with low memory requirement. The Loman Lab 'maxikraken2_1903_140GB' database is used for classification. Kraken 2 classifies reads at multiple levels on the tree meaning that the total number of reads listed for an organism

is not accurate as many reads will be classified at a higher taxonomic level (e.g. Genus) and significantly underestimate the abundance of the species. Therefore, the Kraken 2 report is used as an input for Bracken [288] which performs sequence abundance estimation. The following thresholds are then applied: pathogens must have a minimum of 100 reads for a pathogen ID call, with no reads of the pathogen in the negative control. In the final output, all organisms with relative abundance ≥1% are reported, except for pathobionts *S. pneumoniae* and *H. influenzae* (which can be present in NRF samples at low abundances), for which the threshold is set at ≥5%.

For resistance detection, the filtered fastq files are analysed with ABRicate to screen for AMR genes against the ResFinder database. However, ABRicate reports all resistance genes in the metagenomic sample, including those associated with commensal bacteria. Therefore, the ABRicate results are fed into a new tool – Scagaire - which takes the output and filters the list of AMR genes to include only those which have been observed in the given pathogen/s identified in the sample (using the above thresholds). Scagaire, contains an in-built list of common pathogens for respiratory and gastrointestinal infections. This list is non-exhaustive, (however can be expanded with a custom database):

- *Clostridioides difficile*
- *Campylobacter jejuni*
- *Escherichia coli*
- Enterobacter
- *Enterococcus faecalis*
- *Enterococcus faecium*
- *Enterobacter aerogenes*
- *Enterobacter cloacae*
- *Escherichia coli*
- *Haemophilus influenzae*
- *Klebsiella oxytoca*
- *Klebsiella pneumoniae*
- *Listeria monocytogenes*
- *Moraxella catarrhalis*
- *Mycobacterium leprae*
- *Proteus mirabilis*
- *Pseudomonas aeruginosa*
- *Salmonella enterica*
- *Serratia marcescens*
- *Staphylococcus aureus*
- *Streptococcus pneumoniae*
- *Staphylococcus aureus*

- *Streptococcus agalactiae*
- *Streptococcus pyogenes*
- *Vibrio cholerae*

This resistance gene detection and filtering was used to analyse sample data from our previous publication, Charalampous et al. 2019 [97] with has well characterised antimicrobial susceptibility data. Analysis focused on three pathogens in which acquired resistance is important, *S. aureus, E. coli, and K. pneumoniae*. Results from Scagaire were compared to phenotypic resistance data reported in the paper, with the EPI2ME resistance data to check in the case of discrepancies (Table 3.9).

Samples were classified as concordant, discordant and partially concordant. Concordance is when detected phenotypic resistance can be explained by reported genes, or phenotypic susceptibility is matched with a lack of resistance genes. Where a phenotypic resistance is not explained by a gene or the presence of a gene falsely suggests resistance when the organism is susceptible, the result is discordant. If an organism contains both a mix of results, it is partially concordant.

Of the 13 organisms, 6 had concordant results, 4 had partially concordant results, and 3 were discordant. While most non-relevant genes were correctly filtered for the organisms, 3 major filtering errors by Scagaire were detected, one of which may be responsible for a discordant result. In sample T38, the bla_{TEM} and tetC genes should have been filtered out, as they are likely Gram-Negative genes – this could explain the discrepancy with the susceptible *S. aureus* result. Additionally, there are two examples of genes being incorrectly filtered out, bla_{OXY} for *K. oxytoca* in T5 and bla_{TEM} for *H. influenzae* T39. Many of the discordance was due to Co-amoxiclav, as the presence of bla_{TEM} genes is not often enough to call co-amoxiclav resistance and gene copy number and mutation information may be required to give expression to identify co-amoxiclav resistance accurately. These results suggest that while Scagaire has potential, there are significant improvements that need to be made. Results were reported back so that they could be further investigated.

Table 3.9 – Comparison of CLIMATE and Scagaire filtering to EPI2ME results

Sample	Pathogen (detected by both)	Routine Antibiogram (S = Susceptible, I = Intermediate, R = Resistant)	Scagaire species-specific genes	Climate pre-filtered	EPI2ME comparison
P1	*Escherichia coli*	Amoxicillin R, Gentamicin S, Co-amoxiclav R, Co-trimoxazole R, Tazocin I, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime S, Ceftriaxone S, Cefuroxime S, Amikacin S, Ertapenem S, Tigecycline S, Tobramycin S, Cefepime S	*aadA5* *bla*$_{TEM}$ *dfrA17* *dfrA7* *mdfA* *mphA* *sul1*	*aadA5* *bla*$_{TEM}$ *dfrA17* *dfrA7* *mdfA* *mphA* *sul1* *tet34*	*aadA5* *dfrA17* *bla*$_{TEM}$ *mphA* *sul1* *bla*$_{ACT}$
P2	*Klebsiella pneumoniae*	Amoxicillin R, Gentamicin S, Co-amoxiclav R, Co-trimoxazole S, Tazocin I, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime S, Ceftriaxone S, Cefuroxime S, Amikacin S, Ertapenem S, Tigecycline S, Tobramycin S, Cefepime S	*bla*$_{SHV}$ *fosA* *oqxA* *oqxB*	*bla*$_{SHV}$ *fosA* *lnuA* *oqxA* *oqxB*	*lnuA* *oqxA* *oqxB* *tetM*
P3	*Klebsiella oxytoca*	Gentamicin S, Co-trimoxazole S, Tazocin I, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime S, Amikacin S, Tigecycline S, Tobramycin S, Levofloxacin S, Colistin S, Cefepime S, Imipenem S, Minocycline S, Ticarcillin R	*aph(3')-Ia* *oqxB*	*aph(3')-Ia* *bla*$_{OXY}$ *oqxB*	*aph(3')-Ia* *oqxB* *bla*$_{OXY}$ *vgaC*
P4	*Staphylococcus aureus*	Flucloxacillin S, Erythromycin/clarithromycin S, Clindamicin S, Fuscidic acid S, Tetracycline/doxycycline S, Mupirocin S	*bla*$_{Z}$	*bla*$_{Z}$	*bla*$_{Z}$

P5	*Escherichia coli*	Co-amoxiclav R, Co-trimoxazole R, Tazocin S, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime S, Ceftriaxone S, Cefuroxime S, Amikacin S, Ertapenem S, Tigecycline S, Tobramycin R, Cefepime S	*aac(3)-IIa* *aac(3)-IId* *aadA2* *bla*$_{TEM}$ *dfrA12* *mdfA* *mphA* *sul1* *tetA*	*aac(3)'-IIa* *aac(3)-IId* *aadA2* *bla*$_{TEM}$ *dfrA12* *ermC* *mdfA* *mphA* *sul1* *tetA*	*aac(3')-IIa* *mphA* *aac(3')-IIc* *aadA2* *dfrA12* *ermC* *sul1* *bla*$_{TEM}$ *tetC*
P6	*Staphylococcus aureus*	Flucloxacillin S, Erythromycin/clarithromycin S, Clindamicin S, Fuscidic acid S, Tetracycline/doxycycline S, Mupirocin S	*bla*$_{Z}$	*bla*$_{BRO}$ *blaZ* *mefA* *msrD* *tetM*	*mel* *tet38* *tetM* *tetQ*
P7	*Staphylococcus aureus*	Flucloxacillin S, Erythromycin/clarithromycin S, Clindamicin S, Fuscidic acid S, Tetracycline/doxycycline S, Mupirocin S	*tetM*	*bla*$_{TEM}$ *fosA7* *mefA* *msrD* *penA* *tetM* *tetW* *tetA* *tetB*	*mefA* *mel* *bla*$_{TEM}$ *tet38* *tetM* *tetW*
P8	*Staphylococcus aureus*	Penicillin R, Flucloxacillin R, Oxacillin R, Erythromycin S, Clindamycin S, Trimethoprim R, Gentamicin R, Ciprofloxacin R, Fusidic acid R, Mupirocin S, Rifampicin S, Vancomycin S, Teicoplanin S, Tigecycline S, Linezolid S	*bla*$_{Z}$ *ermB* *fosD* *fusC* *mecA* *tetM*	*blaZ* *ermB* *fosD* *fusC* *mecA* *tetC* *tetM*	*ermB* *mecA* *bla*$_{TEM}$ *tet38* *tetC* *tetM*

P9	*Staphylococcus aureus*	Flucloxacillin S, Erythromycin R, Clindamycin R, Fuscidic acid S, Tetracycline S, Mupirocin S	*ermT*	*ermT* *tetM*	*ermT* *mefA* *tet38* *tetM*
	Haemophilus influenzae	Amoxicillin S, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Cotrimoxazole S			
P10	*Escherichia coli*	Amoxicillin R, Gentamicin S, Co-amoxiclav R, Co-trimoxazole S, Tazocin R, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime I, Ceftriaxone S, Cefuroxime S, Amikacin S, Ertapenem S, Tobramycin S	bla_{TEM} *mdfA*	*Caz-lo* *Caz-3''* bla_{CTXM} *type* bla_{IRT} bla_{TEM} bla_{YOU} *mdfA*	bla_{ACT} bla_{TEM} *tetC*
P11	*Staphylococcus aureus*	Flucloxacillin S, Erythromycin S, Clindamycin S, Fuscidic acid S, Tetracycline S, Mupirocin S	bla_{TEM} bla_{Z} *tetC*	bla_{TEM} *blaZ* *crpP* *tetC*	bla_{TEM} *tetC* *tet38*
P12	*Staphylococcus aureus*	Penicillin R, Flucloxacillin R, Oxacillin R, Erythromycin S, Doxycycline S, Clindamycin S, Trimethoprim S, Gentamicin S, Ciprofloxacin R, Fuscidic acid S, Rifampicin S, Vancomycin S, Teicoplanin S, Tigecycline S, Linezolid S, Daptomycin S, Chloramphenicol S	*mecA*	*mecA* *tetC*	*mecA* *tet38* *tetC* bla_{TEM}
P13	*Staphylococcus aureus*	Flucloxacillin S, Erythromycin S, Clindamycin S, Fuscidic acid S, Tetracycline S, Mupirocin S		bla_{TEM} bla_{Z} *tetC*	bla_{TEM} *tetC*

3.2. Evaluation of the optimised CMg workflow

Thirty-seven surplus sputum samples were tested with the optimised rapid CMg pipeline across 21 different sequencing runs. Samples were sequenced on Flongle flowcells in batches of 1-3 for 2 hours. A mix of positive and negative (normal respiratory flora or no growth) samples were requested from NNUH microbiology and processed in small batches. Samples were all host depleted fresh (without freezing) using the new one pot depletion method (0.1% saponin concentration) including additional treatment with sputasol if samples were viscous. DNA was then extracted, washed and the fast PCR was performed using GXL polymerase.

3.2.1. One-pot host depletion

The one pot host depletion method reduced human DNA by 99.8% on average (average ΔC_T 8.7 between depleted and undepleted samples). The range of host depletions was ΔC_T 1.44 to 13.35 with lower and upper quartiles of ΔC_T 7.16 and 10.79 (Table 3.10). As the LightCycler analysis software doesn't provide an accurate C_T above 35 (when running 40 cycles), this is likely an underestimate of the level of depletion. The highest ΔC_T was 13.35 which equates to a >10,000-fold depletion of human DNA.

Bacterial ΔC_T can sometimes be negative, because the C_T after the depletion is slightly lower than the C_T of the non-depleted sample. When this is <1 C_T, this may be due to random variation in the qPCR. Since it is not possible to gain bacteria, the ΔC_T for these samples are reverted to 0 (no loss) for the purposes of the average calculation. There was no significant loss of bacteria, with the average less than 2-fold (ΔC_T <1) depletion of bacterial DNA. The highest loss of bacterial DNA was ΔC_T 2.98 in sample S36 (approximately 8-fold). This compares to ΔC_T 12.23 depletion of depletion of human DNA (~4800 fold) in the same sample, demonstrating that even though there was some bacterial loss, the host depletion step was still worthwhile.

Table 3.10 – Cycle thresholds for the human and 16S universal bacteria assays, comparing depleted samples to non-depleted controls

	Human			16S		
Sample	**Depleted C_T**	**Non-depleted C_T**	**ΔC_T**	**Depleted C_T**	**Non-depleted C_T**	**ΔC_T**
S1	>35	23.22	11.78	22.4	20.77	1.63
S2	>35	25.98	9.02	24.08	21.58	2.5
S3	32	23.51	8.49	16.71	16.26	0.45
S4	32.99	25.43	7.56	16.77	15.55	1.22
S5	32.88	28.24	4.64	18.68	17.99	0.69
S6	32.26	24.24	8.02	22.16	20.73	1.43
S7	32.23	22.61	9.62	17.23	18.94	0
S8	29.43	22.27	7.16	19.87	21.42	0
S9	31.41	24.48	6.93	15.46	15.51	0
S10	29.68	22.15	7.53	22.8	21.43	1.37
S11	28.33	22.61	5.72	19.22	23.2	0
S12	>35	21.76	13.24	27.64	25.83	1.81
S13	32.31	24.76	7.55	19.92	18.29	1.63
S14	31.87	30.43	1.44	21.86	19.82	2.04
S15	33.99	26.28	7.71	19.69	18.56	1.13
S16	>35	25.76	9.24	21.95	20.43	1.52
S17	34	22.6	11.4	22.56	21.56	1
S18	29.49	23.06	6.43	16.3	16.79	0
S19	>35	28.06	6.94	23.61	21.29	2.32
S20	32.3	23.19	9.11	24.17	23.21	0.96
S21	31.67	27.16	4.51	19.03	16.82	2.21
S22	>35	24.21	10.79	17.64	17.61	0.03
S23	32.59	22.15	10.44	30.7	28.04	2.66
S24	30.12	21.69	8.43	25.32	24.54	0.78
S25	31.26	21.24	10.02	24.05	23.72	0.33
S26	32.09	23.14	8.95	16.54	16.43	0.11
S27	>35	24.11	10.89	16.28	16.03	0.25
S28	33.62	27.11	6.51	22.51	24.23	0

S29	31.85	23.66	8.19	19.44	19.2	0.24
S30	32.09	22.45	9.64	20.49	19.57	0.92
S31	>35	23.79	11.21	17.27	16.8	0.47
S32	34.54	23.63	10.91	17.67	18.17	0
S33	30.8	22.18	8.62	21.3	22.57	0
S34	>35	22.35	12.65	15.91	15.79	0.12
S35	>35	21.65	13.35	17.82	18.74	0
S36	>35	22.77	12.23	22.73	19.75	2.98
S37	>35	32.71	2.29	21.1	20	1.1

3.2.2. Pathogen detection performance

Of the 37 samples tested, 21 samples were reported as positive for pathogens by routine microbiology (25 pathogens reported (Table 3.11)). The remaining 16 samples were reported as having normal respiratory flora (n=13) or no bacterial growth (n=3). Of the 21 positive samples, the rapid CMg pipeline detected the culture reported pathogens in 19, resulting in a sensitivity of 90.5% compared to routine microbiology. The two missed pathogens were *S. aureus* in sample S1 and *S. pneumoniae* and sample S4. Of the 16 culture negative samples, 10 samples were negative by CMg, with additional respiratory pathogen detections in the remaining 6 samples, resulting in a specificity of 62.5%. In the 6 NRF/no growth samples with additional detections, CMg detected 4 *S. pneumoniae* and one each of *P. aeruginosa, K. pneumoniae, H. influenzae* and *M. catarrhalis* (Table 3.11). Species-specific qPCR was used to investigate discordant results. PCR was performed on undepleted extracts of discordant samples (to ensure the depletion step didn't affect the result), along with matched samples negative for the pathogen and positive controls (to ensure qPCR assay was working appropriately). Using culture combined with qPCR as the gold standard, the sensitivity of the CMg pipeline was 96.6% and specificity was 100%. The *S. aureus* in sample S1 was not detected by qPCR, agreeing with metagenomics, whereas the *S. pneumoniae* in S4 was detected by qPCR at C_T 20.61, agreeing with culture. This sample was therefore classified as a false negative by CMg. Further analysis of S4 confirmed that a significant proportion of *S. pneumoniae* DNA was

lost during the depletion with a ΔC_T 6.48 between the depleted sample and the non-depleted control. *S. pneumoniae* reads were detected in this sample, but below the 5% threshold set for pathobionts. There was evidence of an average 7-fold loss of *S. pneumoniae* during the depletion process (ΔC_T 2.9), however, this was variable (ΔC_T 1.32 to 6.48 C_T) between samples. All of the additional pathogens detected in the 6 NRF/no growth samples using CMg were also detected by qPCR, making them true positives.

All sequencing runs contained water process controls, taken through the full process from depletion to sequencing and analysis. None of the controls had more than 10 classified reads after 2 hours of sequencing, with one exception. The water control for S6 had 1,647 *Terribacillus saccharophilus* reads, which was also detected in sample S6 from the same run. However, as it is not a respiratory pathogen and was in low abundance in the sample (3.6%), hence this contamination event was not considered to affect the integrity of the result.

The samples in this subset were largely susceptible apart from the *P. aeruginosas* meaning that the resistance prediction was not particularly relevant. In the case of *P. aeruginosa* and *M. catarrhalis*, the pipeline detected intrinsic bla_{OXA} and bla_{BRO} genes in every instance. *S. aureus* penicillin resistance could correctly be called in both samples where *S. aureus* was detected, with the bla_Z gene being correctly identified within 2 hours of sequencing, however, this would be assumed from *S. aureus* anyway.

Table 3.11 – Detection of pathogens and AMR genes from the optimised pipeline compared to routine microbiology. Green indicates concordance, Red is a missed pathogen by the CMg pipeline and Yellow is additional pathogens detected by CMg pipeline.

Pathogens				Antimicrobial resistance		
Sample	**Pathogen cultured by routine microbiology**	**Pathogen identified from metagenomic pipeline**	**Relative abundance (%)**	**Antibiotic susceptibility testing by routine microbiology**	**Gene reported by pipeline**	**Number of reads**
S1	*Staphylococcus aureus*					
	Pseudomonas aeruginosa	*Pseudomonas aeruginosa*	39.8	Gentamicin (S), Pip/Tazobactam (S), Ciprofloxacin (S), Ceftazidime (S), Meropenem (S)	*bla*$_{PAO}$ *catB7*	2 3
S2	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	95.3	Gentamicin (S), Pip/Tazobactam (S), Ciprofloxacin (S), Ceftazidime (S), Meropenem (S)	*bla*$_{OXA}$ *bla*$_{PAO}$ *catB7* *crpP* *fosA*	2 2 2 6 11
S3	*Haemophilus influenzae* *Pseudomonas sp.*	*Haemophilus influenzae* †*Pseudomonas sp.*	74.4 1.0			
S4	*Streptococcus pneumoniae*					
	Haemophilus influenzae	*Haemophilus influenzae*	96			
S5	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	51.2	Gentamicin (S), Pip/Tazobactam (S), Ciprofloxacin (S), Ceftazidime (S), Meropenem (S)	*bla*$_{OXA}$ *bla*$_{PAO}$ *catB7* *crpP* *fosA*	6 6 3 3 9 5
		Serratia marcescens	46.2		*aac(6')-Ic* *bla*$_{SRT}$	15 2 3
S6	*Streptococcus pneumoniae*	*Streptococcus pneumoniae*	6.2	Erythromycin (S), Clindamycin (S), Tetracycline (S), Penicillin (S),	*mefA* *msrD*	4 5

				Ampicillin (S), Trimethoprim/Sulfamethoxazole (S), Vancomycin (S), Levofloxacin (S), Moxifloxacin (S), Rifampicin (S), Linozalid (S), Oxacillin (S)		
S7	*Haemophilus influenzae*	*Haemophilus influenzae*	87.4			
S8	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	100	Gentamicin (S), Pip/Tazobactam (S), Ciprofloxacin (S), Ceftazidime (S), Meropenem (S)	*catB7*	2
S9	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	97.6	Gentamicin (S), Pip/Tazobactam (R), Ciprofloxacin (R), Ceftazidime (R), Meropenem (I), Ceftriaxone (R), Levofloxacin (S), Aztreonam (R), Amikacin (S), Tigecycline (R), Tobramycin (S), Imipenem (R)	*aph(3')-IIb* *bla*$_{OXA}$ *bla*$_{PAO}$ *catB7* *crpP* *fosA*	9 18 6 13 31 34 14
S10	*Streptococcus pneumoniae*	*Streptococcus pneumoniae*	66.3			
		Haemophilus influenzae	9.0			
		Escherichia coli	7.1			
S11	NRF	*Moraxella catarrhalis*	41.6			
		Streptococcus pneumoniae	5.0		*msrD*	3
S12	*Enterobacter* complex	*Enterobacter cloacae* complex	100			
S13	NRF	None				
S14	NRF	None				
S15	NRF	None				
S16	*Streptococcus pneumoniae*	*Streptococcus pneumoniae*	15.2	Erthromycin (S), Clindamycin (S), Tetracycline (S), Penicillin (S),	*msrD*	3

				Ampicillin (S), Trimethoprim/Sulfamethoxazole (S), Vancomycin (S), Levofloxacin (S), Moxifloxacin (S), Rifampicin (S), Linozalid (S)		
S17	*Streptococcus pneumoniae*	*Streptococcus pneumoniae*	62.3			
S18	*Moraxella catarrhalis*	*Moraxella catarrhalis*	59	Tetracycline (S), Ciprofloxacin (S), Ampicillin (R), Ceftriaxone (S), Augmentin (S), Trimethoprim/Sulfamethoxazole (S)	*bla*$_{BRO}$	11
	Streptococcus pneumoniae	*Streptococcus pneumoniae*	40.6			
S19	NRF	None				
S20	NRF	None				
S21	NRF	None				
S22	NRF	None				
S23	NBG	None				
S24	NBG	None				
S25	NBG	*Streptococcus pneumoniae*	5.8		*mefA*	2
S26	*Haemophilus influenzae*	*Haemophilus influenzae*	89.5			
S27	*Staphylococcus aureus*	*Staphylococcus aureus*	51.6	Flucloxacillin (S), Erthromycin (R), Clindamycin (R), Fuscidic Acid (S), Tetracycline (S), Penicillin (R)	*bla*$_{Z}$ *ermT*	8 2
S28	*Moraxella catarrhalis*	*Moraxella catarrhalis*	96.6	Tetracycline (S), Ciprofloxacin (S), Ceftriaxone (S), Augmentin (S), Trimethoprim/Sulfamethoxazole (S)	*bla*$_{BRO}$	29
S29	*Staphylococcus aureus*	*Staphylococcus aureus*	59.6	Flucloxacillin (S), Erthromycin (S), Clindamycin (S), Fuscidic Acid (S), Tetracycline (S), Penicillin (R)	*bla*$_{Z}$ *fosD*	7 3

		Haemophilus influenzae	16.3		
S30	*Haemophilus influenzae*	*Haemophilus influenzae*	94.4		
S31	NRF	*Klebsiella pneumoniae*	87.8	*bla*$_{SHV}$ *fosA* *fosA6* *oqxA* *oqxB*	7 3 12 5 10
		Pseudomonas aeruginosa	5.7		
S32	*Haemophilus influenzae*	*Haemophilus influenzae*	84.5		
S33	NRF	*Streptococcus pneumoniae*	8.5	*mefA* *msrD* *tetM*	10 9 5
S34	*Haemophilus influenzae*	*Haemophilus influenzae*	87.8		
S35	NRF	*Haemophilus influenzae*	20.1	*bla*$_{TEM}$	2
S36	NRF	None			
S37	NRF	*Streptococcus pneumoniae*	5.6	*mefA* *msrD* *tetO*	8 9 4

NRF, Normal respiratory flora; NBG, No bacterial growth, † Routine microbiology did not report a species.

3.2.4. Sequencing data and timepoint analysis

The Flongle sequencing runs produced an average yield of 28.8 Mb of passed data (q-score >7) after 30 minutes of sequencing with a mean read length of 3,239 bp and an average q-score of 10.2. The number of reads per sample was dependant on the number of samples sequenced per Flongle flowcell (runs listed in Table 3.12). When sequencing 3 samples per flowcell, the average number of reads per sample was 9,700, for 2 samples it was 16,400 and for 1 sample it was 27,600 reads after 2 hours of sequencing.

Some runs were sequenced for longer to observe the course of full runs. Sequencing output steadily declines over the 24 hours, and there is variability in the performance of runs. In some cases, sequencing output rapidly deteriorates after the first few hours, for example, in Run 2 (Figure 3.6) while other runs continue producing data until the end of the run. This was not related to the pore count of the flowcell and was therefore difficult to predict – it could be due to performance issues with the Flongle, however library quality cannot be ruled out. In most cases, the vast majority of data is acquired in the first half of the run with diminishing returns as the run goes on.

Table 3.12 – List of sequencing runs with number of reads for each sample. S12 can be seen as the outlier with a very low number of reads.

Run number	Number of Flongle Pores	Sample	Number of reads after 30 min	Number of reads after 2 hours
1	64	S1	2273	8834
		S2	2193	8257
		S3	3082	12091
2	59	S4	7969	30665
3	55	S5	6393	26902
4	67	S6	6234	23266
5	46	S7	3036	10900
		S8	3148	10938
6	76	S9	10325	41292
7	77	S11	10935	38752
		S12	207	726
8	72	S10	2397	9174
		S13	5290	19584
9	63	S14	4277	17496
		S15	4493	17517
10	58	S16	4110	14439
11	62	S17	3854	14816
		S19	2739	10696
12	54	S18	3239	11430
		S20	2465	9131
13	74	S21	5315	19902
		S22	5409	20164
14	76	S23	4409	14901
		S24	7362	24061
15	68	S25	8562	29025
16	71	S27	5314	19753
		S28	4513	15341
17	74	S26	4715	18353
		S29	5190	20004
18	54	S30	3208	13035
		S31	4494	17734
19	69	S32	3305	13449
		S33	7694	30493
20	72	S34	6551	25787
		S35	3364	13313
21	58	S36	2107	8564
		S37	3552	13552

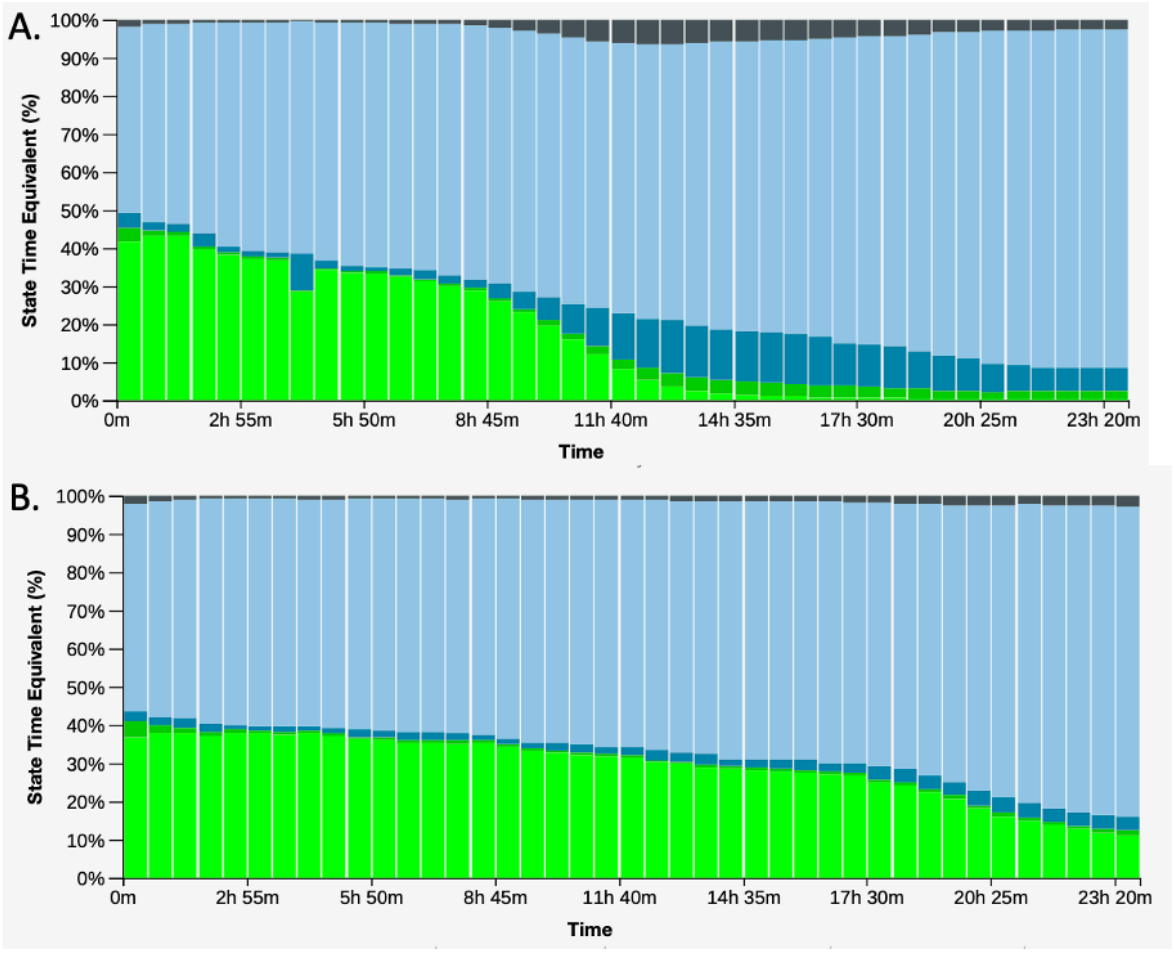


Figure 3.6 – Sequencing duty plots of two full 24 hour Flongle runs. A) Run 2 B) Run 11

Timepoint analysis on the 37 samples showed that the pathogens detected after 30 minutes of sequencing were the same as after 1 or 2 hours. For most samples, 15 minutes of sequencing was sufficient to correctly identify the pathogen with the exception of sample S12, which required 30 minutes of sequencing for the pathogen to be detected above thresholds. This was due to a lower sequencing yield compared to other samples (Table 3.12). Sample S12 had only 95 reads in 15 minutes and 726 reads in 2 hours whereas all other samples had >1,000 reads in 15 minutes of sequencing, and >5,000 reads after 2 hours. Sample 12 had low bacterial biomass (the highest bacterial C_T from a positive sample – 27.64 post depletion) and almost no human DNA remaining after host depletion which explains the low read numbers. The abundance of pathogenic reads remained very consistent over time (Figure 3.7), with no change of results between 30

minutes and 2 hours. It is therefore possible to call pathogen ID after 30 minutes of sequencing instead of waiting 2 hours.

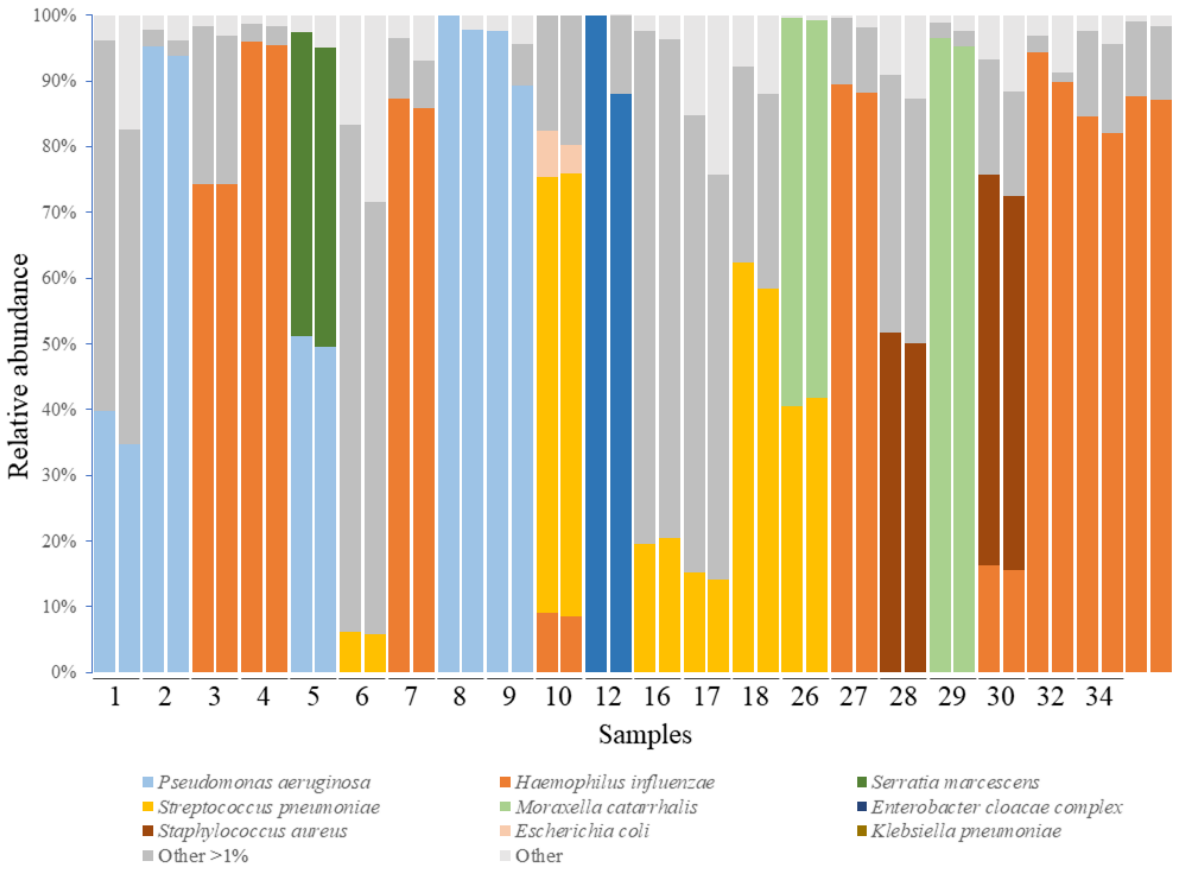


Figure 3.7 – Relative abundance of pathogens detected in positive samples. Each sample has two bars, the first shows abundance of pathogen after 30 minutes of sequencing and the second shows abundance after 2 hours.

The results from the evaluation of the rapid CMg workflow are currently being prepared in a manuscript for publication.

3.3. Further optimisation of the CMg workflow

Following the successful evaluation of the rapid CMg workflow, further experiments were performed to improve the method. This included removing steps to reduce time and complexity, testing an internal process control to control for method failure and

development of a separate viral metagenomic arm of the procedure that can be performed in parallel.

3.3.1. Necessity of the post-extraction 1.2X SPRI clean

The rapid CMg workflow contains a SPRI clean-up step for individual samples after DNA extraction. This is to concentrate the sample when the microbial biomass is low, improving PCR success rate; it is also an additional purification of the DNA which may help improve tagmentation and PCR during the library preparation by removing inhibitors. However, this adds approx. 20 minutes to the procedure, and is performed on an individual sample basis, adding to the manual labour and increases the risk of error/contamination. The necessity for the bead wash was tested by comparing PCR results of SPRI cleaned vs uncleaned sputum sample DNA extracts. Five clinical samples with a mix of concentrations were split and tested in 3 ways:

A. 7.5 μL uncleaned DNA extract used for tagmentation followed by PCR
B. 7.5 μL DNA SPRI cleaned, eluted in 7.5 μL and 7.5 μL used for tagmentation followed by PCR
C. 37.5 μL DNA SPRI cleaned, eluted in 7.5 μL and 7.5 μL used for tagmentation followed by PCR

B tests the effect of the SPRI clean without attempting to concentrate the sample, whereas C concentrates the sample in addition to the clean-up.
In samples that were quantifiable by Qubit after the concentrating SPRI clean (C), recovery rates of the DNA ranged from 75% - 100%, with an average of 93.2% (Table 3.13), suggesting that very little DNA was being lost in the SPRI cleaning process.

Where extracts were cleaned without concentrating (B), the post-PCR concentrations were always lower. In 3 of the samples (T11, T12, T13), the PCR on the cleaned extracts failed the PCR. In the samples that were concentrated in the wash (C) post-PCR concentrations were better than for B, however, were still lower than A (not doing a clean-

up). The PCR for sample T13, which was not quantifiable by Qubit pre-PCR; failed using the SPRI concentration (0.42 ng/μL) but produced 16.8 ng/μL without SPRI clean. This suggested that the 1.2X SPRI wash was not required and may even have a negative impact on test performance. However, the samples that were tested were previously frozen and therefore may have been degraded and have behaved differently to fresh samples – confirmation of these results with fresh samples needs to be performed. If these results hold in fresh samples, this would allow the removal of the post extraction SPRI clean, reducing test turnaround time by another 20 minutes and simplifying the workflow.

Table 3.13 – SPRI clean vs uncleaned results for 5 clinical samples and a negative control

Sample	Pre-PCR conc. (ng/μL)	DNA recovery post SPRI clean	No SPRI PCR product conc. (ng/μL)	SPRI clean PCR product conc. (ng/μL)	Concentrated SPRI clean PCR product conc. (ng/μL)
T9	9.76	100%	37.0	9.4	26.0
T10	4.48	97.8%	14.4	10.1	7.2
T11	0.416	75%	40.6	0.27	23.6
T12	0.190	100%	3.92	0.32	3.5
T13	Undetectable	N/A	16.8	0.31	0.42
Negative	Undetectable	N/A	0.316	0.29	N/A

3.3.2. Internal process control

Where a sample fails to sequence successfully this may be because of a failure of the process or it may be due to the sample having a low/no microbial biomass (i.e. no growth by routine microbiology). To distinguish between a process failure and a genuine negative clinical sample, an internal process control is necessary. This is a control (typically a microbe never found in the type of sample being tested) that is added to each sample prior to DNA extraction and carried through the whole process – reads from the internal control organism need to be present for the test to be valid, if the sample produces few or no reads (samples that produce microbial/human reads in abundance clearly didn't fail the

process and are valid even if no control reads are detected). Otherwise, the sample has failed due to e.g. depletion failure, extraction failure, PCR failure, sequencing failure. ZymoBIOMICS Spike-in Control I (Zymo) was identified as a potentially suitable process control for the CMg test. This consists of two organisms not typically found in human samples, *Imtechella halotolerans* (Gram-negative) and *Allobacillus halotolerans* (Gram-positive). The control is provided as frozen tubes containing mixed organisms (50:50) at 2 x 10^7 per organism per prep (20 μL). This was diluted and spiked into two samples (PBS and a no growth clinical sample S24) at 1 x 10^5 and 1 x 10^4 CFU. Samples were depleted using the new one-pot host depletion method and DNA extracted. qPCR results (Table 3.14) show that there was a significant loss of the control during the depletion – an average ΔC_T 5.11 loss for *I. halotolerans* and ΔC_T 4.6 for *A. halotolerans*. This is likely to be an underestimate due to the inaccurate C_T values above 35 on the Lightcycler.

Table 3.14 – Host depletion results on ZymoBIOMICS spike in control

Spike organism	Sample	Spike quantity	Non-depleted C_T	Depleted C_T	ΔC_T
I. halotolerans	PBS	10^4	31.09	>35	3.91
		10^5	28.15	>35	6.85
	S24	10^4	31.81	>35	3.19
		10^5	28.51	>35	6.49
A. halotolerans	PBS	10^4	31.39	>35	3.61
		10^5	28.21	31.43	3.22
	S24	10^4	29.90	>35	5.1
		10^5	26.28	32.76	6.48

The 10^5 spike samples were amplified and prepared and sequenced using the new protocol. The sequencing did not produce any reads from the spiked organisms when analysed by EPI2ME. To confirm that this is not a database issue, all reads were mapped to the genomes of *I. halotolerans* and *A. halotolerans*, which produced 0 mapped reads, confirming the lack of the spiked organism reads. Further investigation into the ZymoBIOMICS product revealed that the organisms are stored in DNA/RNA Shield which is also a lysis agent, meaning that the organisms are likely disrupted in the storage

solution and the DNA degraded during the depletion process. There was evidence that there was less loss of the Gram-positive organism DNA, with C_T of 31.43 and 32.76 for the 10^5 spikes post depletion, however, this was not sufficient for detection using our pipeline. An organism that can be grown in the lab, frozen in aliquots of non-lysing buffer, be spiked into clinical samples and show no loss post depletion is required for future use as a positive control.

3.3.3 Testing a parallel viral metagenomics arm to the CMg workflow

The CMg workflow was originally designed for characterising bacteria and fungi as part of a research programme on the rapid diagnosis of nosocomial pneumonia. Due to the nature of the depletion, and the focus on DNA, respiratory RNA viruses, such as Flu A and B, RSV, rhinovirus, and SARS-CoV-2 are missed. Double stranded DNA viruses (such as adenoviruses) could, in principle, be sequenced by the workflow, however, the centrifugation based wash step during the host depletion likely results in viruses being discarded. Additionally, the impact of saponin on viruses has not been tested, and viruses that have an envelope (outer layer that is derived from the host membrane) may be susceptible to saponin treatment and their nucleic acid lost in the nuclease digestion step. Therefore, a separate arm of the procedure needed to be designed to sequence DNA and RNA viruses. The viral arm required a depletion step for human, fungal and bacterial DNA/RNA prior to nucleic acid extraction and cDNA generation. This should ideally run alongside the existing CMg pipeline and the nucleic acid from both arms be combined for sequencing to provide comprehensive results in a rapid and cost effective manner.

3.3.3.1 Comparison of viral nucleic acid extraction methods

In order to identify a rapid and efficient nucleic acid extraction method for viruses, a number of commercial kits were tested and compared. A commercial HIV-1 Control (AcroMetrix) was chosen as an initial model virus for testing, as this had been used in the group previously and is an encapsulated virus provided fully intact (made safe by genetic

modification). HIV-1 has similar features to respiratory viral pathogens such as Flu which is ssRNA and is enveloped. An inactivated SARS-CoV-2 control (Qnostics) was also tested in the extraction, however, this control is inactivated by heat treatment and gamma irradiation, meaning its use was limited.

The RSC Viral Total Nucleic Acid Purification Kit (Promega), MagNA Pure LC Total Nucleic Acid Isolation Kit (Roche) and Quick-DNA/RNA Viral Kit (Zymo) were tested. The Promega and Roche kits are performed on automated extraction machines using magnetic beads and the Zymo kit is a column-based manual extraction kit. All kits were used according to manufacturer instructions to isolate HIV-1 and SARS-CoV-2 RNA. Extractions on both viruses were performed in duplicate. The Maxwell Viral kit and MagNA Pure Total Isolation kit performed similarly for HIV-1, whereas the Zymo kit gave a higher C_T (Table 3.15). For SARS-CoV-2, Promega Maxwell outperformed the MagNA Pure and Zymo kits, with Zymo again providing the lowest RNA yield. The Promega Maxwell extraction kit was therefore taken forward for viral extraction experiments.

Table 3.15 – Extraction comparison of the two RNA viruses

	Maxwell Extract C_T	**MagNA Pure Extract C_T**	**Zymo Quick-DNA/RNA Extract C_T**
HIV-1 1	30.16	30.06	31.08
HIV-1 2	29.65	29.75	30.74
SARS-CoV-2 1	29.90	31.97	32.59
SARS-CoV-2 2	30.65	32.68	33.13

3.3.3.2 Viral library preparation

A library preparation approach for viruses was designed using Sequence-Independent, Single-Primer Amplification (SISPA). This is a random priming method that converts RNA into cDNA followed by amplification in a sequence-independent manner. The SISPA method reported by Chrzastek et al. 2017 [289], was adapted for the CMg pipeline. The main steps of the method are:

1. Primers that contain random octamers with a universal primer tag are used for first strand cDNA synthesis from extracted RNA

2. The products from first strand cDNA synthesis are used in a second reaction with Klenow enzyme using the same random octamer primers. This leads to the synthesis of double stranded cDNA
3. The double stranded cDNA has tags for universal priming on both strands and is amplified in PCR using a single primer complementary to the tag sequence

The first two steps of this protocol were kept the same, however, a different random primer construct was used which contains the priming site for the ONT RPB004 library preparation kit primers (TTTTTCGTGCGCCGCTTCAACNNNNNNNNN). In the normal RPB004 library preparation, these tags are introduced in the enzymatic transposase tagmentation step, however, this step is unnecessary when introducing the priming site during first and second strand synthesis of cDNA. Another change to the Chrzastek method was the use of the rapid GXL polymerase for the single primer PCR (the full adapted method is described in Section 2.3.5). The first strand synthesis step takes 30 minutes and the second strand synthesis takes 1 hour.

This library preparation approach was performed on HIV-1. The same volume of HIV-1 control used for the extraction experiments (C_T ~30) was spiked into 100 μL of PBS and 100 μL of sputum. A negative control was also included. The sputum sample was centrifuged for 5 minutes at 12,000 x g to pellet human and bacterial/fungal cells, and the supernatant was taken for extraction. The PBS sample was not centrifuged, acting as a control. Following library preparation, PCR concentrations were 5 ng/μL for the sputum HIV-1 sample, 25 ng/μL for the PBS HIV-1 sample and 0.6 ng/μL for the negative control. Sequencing produced a very low yield with no HIV-1 reads present in any of the samples, suggesting the library preparation was not optimal or the number of HIV particles spiked was below the LOD of the approach. For the sputum sample, 88% of the 5,893 reads were human, and for the PBS sample, 99.8% of the 62,109 were human.

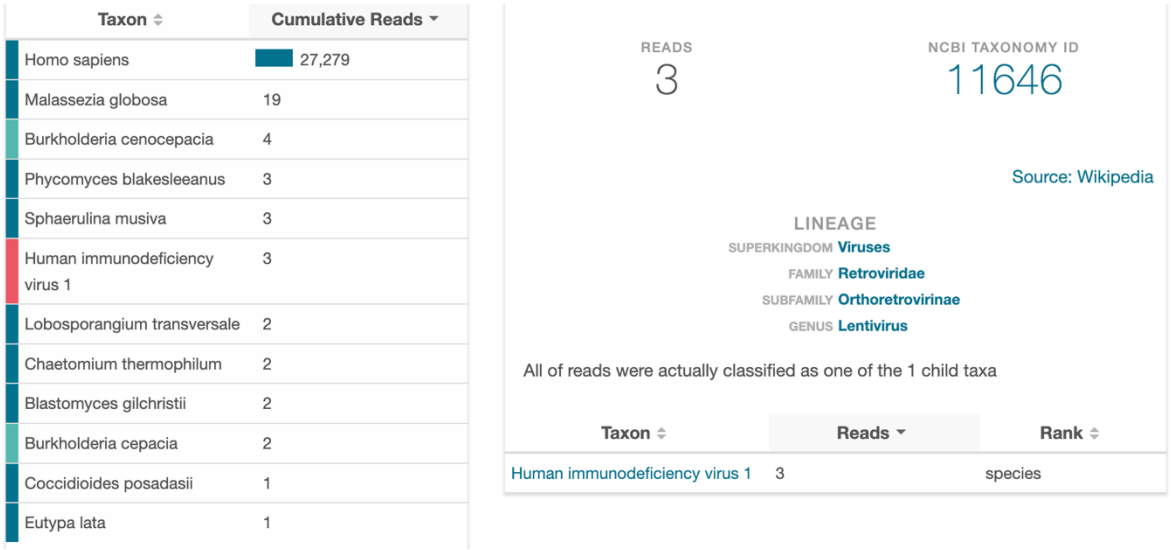
Several changes were then made to the protocol, with some elements taken from a method by Greninger et al. 2015 [95]. This SISPA method follows the same principles as

Chrzastek but uses Sequenase (Affymetrix) for the second strand synthesis rather than Klenow enzyme, which takes 16 minutes instead of 1 hour. Additionally, a shorter first strand synthesis step was introduced which uses SuperScript IV Reverse Transcriptase (Invitrogen). This reduced first and second strand synthesis to 36 minutes prior to the amplification. The PCR was also modified in an attempt to improve sensitivity. A shorter primer that does not contain the barcode and click chemistry overhangs was used for the first 25 cycles, followed by addition of RPB004 primers and a further 10 PCR cycles to add the barcodes.

This modified method was first tested on RNA from *E. coli* to confirm that it worked before testing on more challenging viral samples that contain low concentration of RNA. *E. coli* RNA was extracted and treated with Turbo DNase (RT-qPCR confirmed the absence of contaminating DNA). The treated RNA was tested with the new modified protocol including a control without the reverse transcriptase in the cDNA step to confirm absence of *E. coli* DNA reads. The concentration after the barcoding PCR for the RNA sample with reverse transcriptase (RT) was 41.6 ng/μL and for the RT sample without RT was 1.34 ng/μL. After 15 minutes of sequencing, there were 593 *E. coli* reads in the RT sample and no reads in the control, confirming that the method works for RNA. The average read length was 626 b – this was expected due to the use of random priming in the protocol.

The new protocol was then tested on HIV-1. HIV-1 was extracted without adding to a clinical sample. RT-qPCR using the HIV-1 assay showed that the extract had a C_T of 29.5, which approximates ~1000 viral copies per μL. 15 μL of this was used for the procedure. Post-barcode PCR concentration was 11.8 ng/μL. Sequencing of this product on a Flongle yielded 27,393 classified reads, of which 3 reads were HIV-1 (0.01%). The vast majority of reads, 27,279 (99.6%) were human (Figure 3.8). The average read size was 852 bases. The presence of HIV-1 reads was confirmed by aligning to the HIV-1 genome using minimap2. Since HIV-1 is an RNA virus, this shows that the process of first strand DNA synthesis and amplification worked and could be further optimised. However, human DNA

dominated the sequencing run. This human nucleic acid was coming from the human plasma matrix that the HIV-1 was stored in.



Taxon	Cumulative Reads
Homo sapiens	27,279
Malassezia globosa	19
Burkholderia cenocepacia	4
Phycomyces blakesleeanus	3
Sphaerulina musiva	3
Human immunodeficiency virus 1	3
Lobosporangium transversale	2
Chaetomium thermophilum	2
Blastomyces gilchristii	2
Burkholderia cepacia	2
Coccidioides posadasii	1
Eutypa lata	1

Taxon	Reads	Rank
Human immunodeficiency virus 1	3	species

Figure 3.8 - EPI2ME result showing the presence of HIV-1 in a high background of human DNA

3.3.3.3 Host depletion for viruses

The dominance of human reads indicated that host depletion would be required for successful viral CMg. It is possible to treat the extracted nucleic acid with a DNase so that only RNA is processed downstream. We showed this when a HIV-1 extract was treated with TURBO DNase (Invitrogen), which led to a human C_T of >40 (not detected) post-treatment. However, DNase treatment after extraction leads to loss of DNA viruses such as Adenovirus. Nuclease treatment prior to extraction while the virus is intact can reduce non-viral DNA and RNA, without losing DNA or RNA viruses. Two nucleases were tested on the HIV-1 sample prior to extraction, Micrococcal Nuclease (MNase) and HL-SAN nuclease. Results showed minimal depletion of human DNA when MNase was used, with a ΔC_T of 1.08, and only a small loss of HIV-1 with a ΔC_T of 0.63. HL-SAN with high salt buffer performed better for host depletion with ΔC_T 2.36, however, also led to more loss of HIV-1 with ΔC_T 2.41.

The nucleases were tested in combination with differential lysis to test if that would improve depletion of host nucleic acid. Saponin and Phospholipase C (PLC), a cytolysin

that also disrupts phospholipid bilayer membranes, were tested. These were tested together with either MNase or HL-SAN and PLC/Saponin in the same tube simultaneously. Unlike the standard host depletion, samples were not centrifuged after the incubation as this could lead to loss of viruses – samples were extracted directly after depletion. A non-depletion control was used as the baseline to determine HIV-1/human loss. The results show that both the use of PLC and saponin results in the complete removal of human nucleic acid (Table 3.16), however, significant loss of HIV-1 was also observed (more than the human nucleic acid removal), indicating that saponin and PLC can lyse enveloped viruses. In one sample, where HL-SAN was used in combination with saponin, there was little loss of both human and HIV-1, however, this was likely just a failure of the method.

Table 3.16 – Depletion of human DNA and loss of HIV-1. Starting human material in the PBS sample is C_T 35.5.

Test	**Membrane disruption**	**Nuclease**	**Human nucleic acid removal (ΔC_T)**	**HIV-1 RNA loss (ΔC_T)**
1	Saponin	HL-SAN	0.58	1.77
2	PLC	HL-SAN	>4	>10
3	Saponin	MNase	>4	6.5
4	PLC	MNase	>4	6.9

Loss of HIV-1 nucleic acid using saponin and PLC indicated that they cannot be used for viral host depletion, as envelope viruses are significantly affected and are important causes of respiratory diseases. Further work is required to find a suitable host depletion approach for viral CMg. One approach that removes the need for depletion is to target the whole genome of the virus of interest using tiling PCR.

3.4. Development of a high-throughput sequencing method for SARS-CoV-2

In April 2020, the COG-UK consortium was established for sequencing of SARS-CoV-2 genomes at 16 sites across the UK, including at QIB. Initially, the ARTIC method [285] was used at most sites, which involved a tiling PCR primer scheme followed by ONT's Native Barcoding Ligation library preparation. The library preparation method required cleaning of PCR products, A-tailing the ends of amplicons, and using ligation chemistry to attach barcodes per sample. The samples were then pooled, and the sequencing adaptor was added by ligation. The library preparation took 3 hours, on top of the RNA extraction, reverse transcription, and PCR steps. Additionally, native barcoding was limited to 24 barcodes, with a minimum of 7 samples recommended to have enough material for successful sequencing, meaning that the method was relatively low throughput and quite inflexible. A higher throughput, more flexible method, preferably with a shorter turnaround time was urgently required, particularly during peaking waves of the pandemic.

3.4.1 CoronaHiT method development

We developed the SARS-**Corona**virus-2 **Hi**gh **T**hroughput (CoronaHiT) method to improve sequencing throughput, flexibility, and turnaround time. CoronaHiT uses Illumina's Nextera transposase chemistry to introduce PCR adapters, followed by a short PCR to introduce barcodes (Table 2.9 in Methods section). The barcoded products can then be sequenced using nanopore (Figure 3.9). This method allowed for a larger number of barcodes to be used for Nanopore sequencing compared to ONT's 24 barcodes (ONT later released 96 native barcodes to allow higher throughput, but this wasn't available at the time).

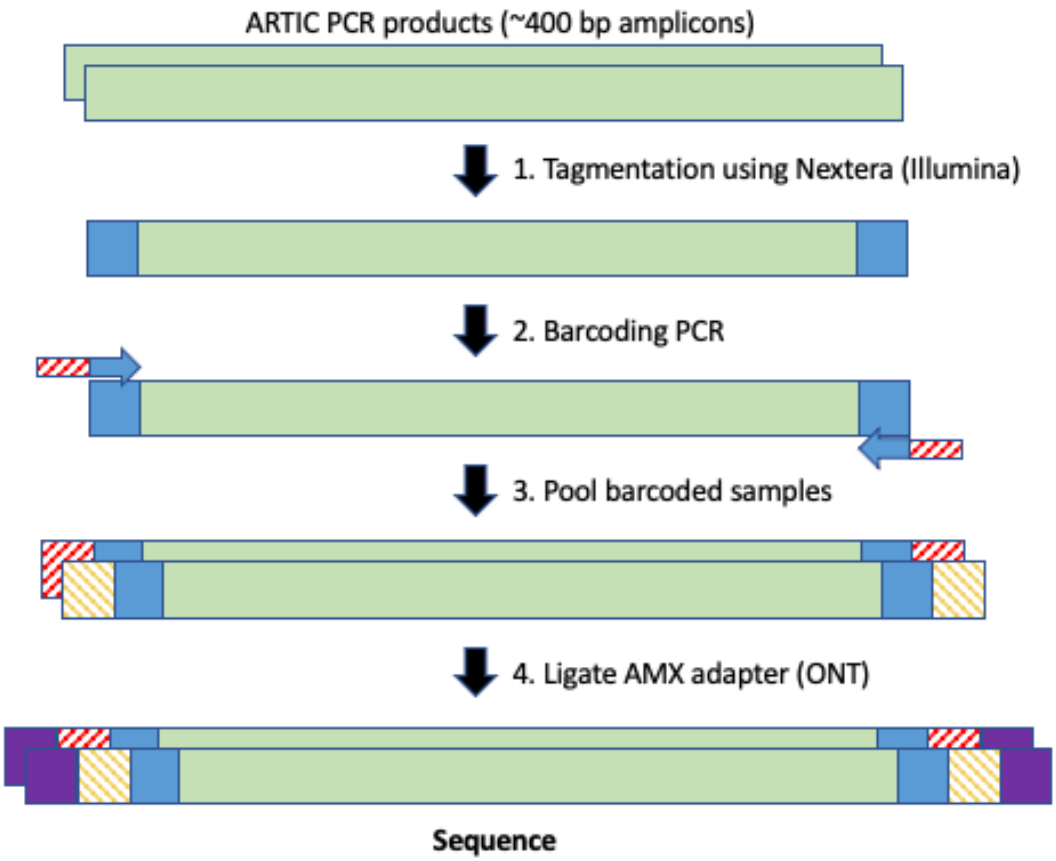


Figure 3.9 – CoronaHiT principle. The first step is tagmentation of ARTIC PCR products with Nextera transposase chemistry. The Nextera adapters can then be used to perform PCR barcoding using primers complementary to the sequences with barcode overhangs. After barcoding, the samples are pooled and prepared for nanopore sequence by the addition of the nanopore adapter using ligation.

The first barcode construct, (designed by Dave Baker, Head of Sequencing at QIB) used 16-bp barcodes sequences from Pacific Biosciences [290] combined with Illumina P5 and P7 sequences [291] complementary to the adapters introduced by the Nextera transposase. This allowed for the use of up to 384 unique barcodes.

The construct was designed as follows, where N = 16 bp barcode sequence:

Barcode with P5 complement: 5‘- NNNNNNNNNNNNNNNNGTCTCGTGGGCTCGG –‘3

Barcode with P7 complement: 5’- NNNNNNNNNNNNNNNNTCGTCGGCAGCGTC - ‘3

To test if the barcoding method worked as intended, samples were sequenced using the barcode constructs across 2 sequencing runs, one nanopore run with 95 samples and one nanopore run with the same 95 samples plus an additional 63 (therefore 158 samples in

total). The same samples were also sequenced on Illumina (with Nextera barcodes). As a control, 23 of the samples were also sequenced with the default ARTIC method (with ONT native barcodes). Initial comparison of the results from the 95 nanopore CoronaHiT samples vs 23 ARTIC samples showed that CoronaHiT had lower coverage compared to ARTIC. This was because the proportion of unclassified reads for CoronaHiT (50%) was almost double that for ARTIC (28%). Additionally, the coverage for the different barcodes was very uneven for CoronaHiT (M = 104x, SD = 71x) compared to ARTIC (M = 124x, SD = 25x). Illumina was similar to ARTIC in terms of barcode evenness (M = 131 Mb, SD = 26); the coefficient of variance for barcode yield was identical (0.20) for both ARTIC and Illumina, indicating similar variation, but higher (0.69) for CoronaHiT (Figure 3.10).

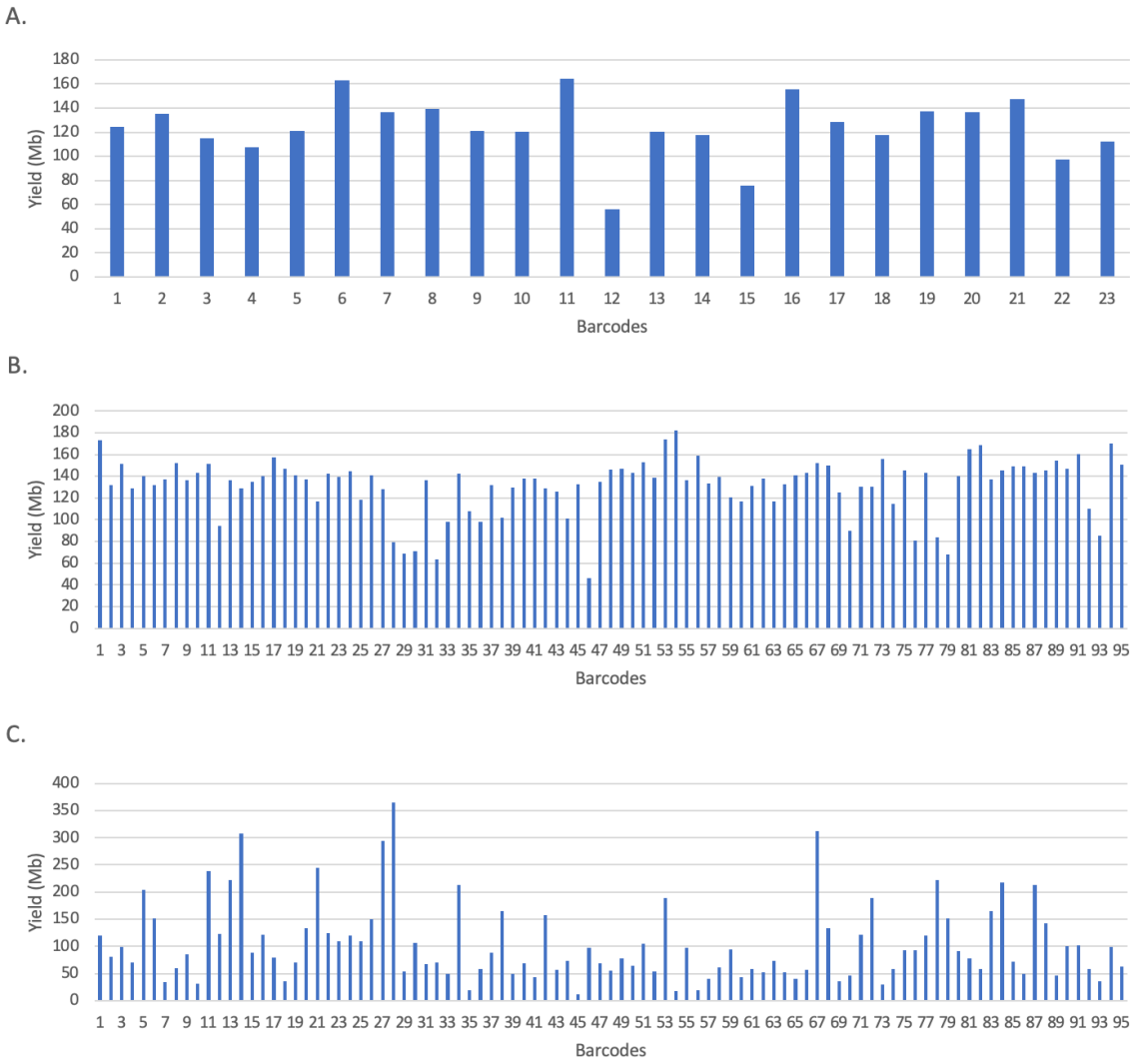


Figure 3.10 – Sequencing yield (Mb) per barcode with the 3 sequencing methods showing high variation with the CoronaHiT barcodes. A) ARTIC method, B) Illumina sequencing and C) CoronaHiT method.

Further investigation of the coverage unevenness for nanopore CoronaHiT showed that this not related the concentration of SARS-CoV-2 in the sample i.e. the diagnostic PCR C_T (Figure 3.11A). Also, there was also no correlation between Illumina and CoronaHiT for the same samples (Figure 3.11B). However, there was a very strong correlation in coverage between the two separate CoronaHiT runs for the 95 shared samples (Figure 3.11C). Knowing that this correlation was not sample related, and that the same barcodes are over/under-represented in two separate runs suggested that the barcodes constructs themselves were causing the problem.

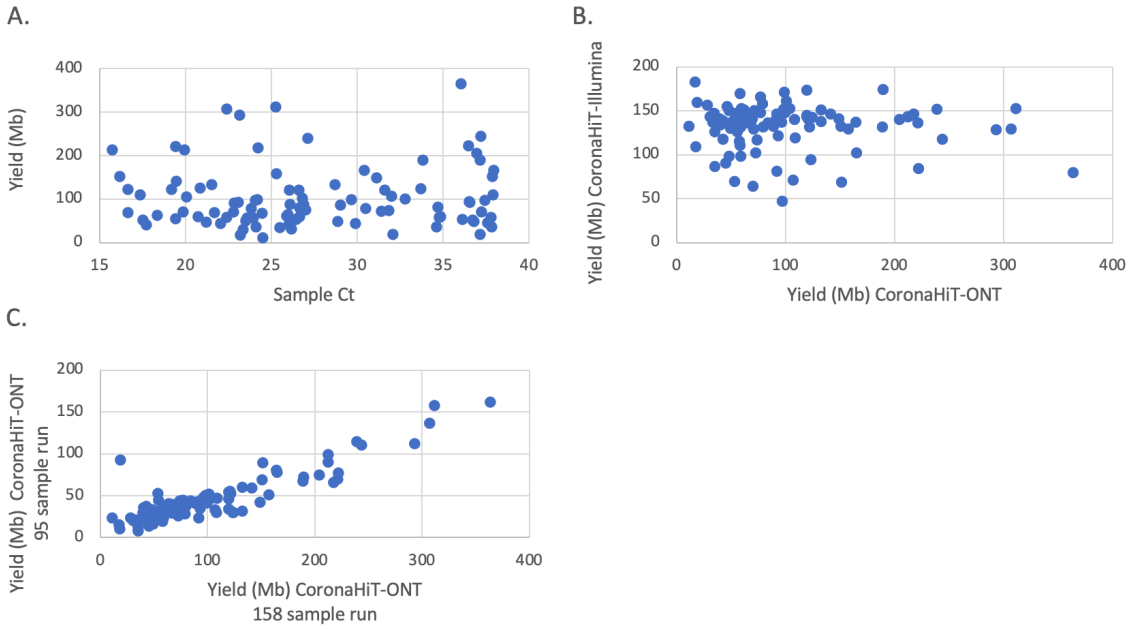


Figure 3.11 – Comparisons of yield. A) Sample yield of the CoronaHiT run versus SARS-CoV-2 qPCR C_T showing no correlation. B) Sample yield of Illumina versus sample yield for the same samples in CoronaHiT showing no correlation. C) Sample yields for the CoronaHiT run with 95 samples, versus the same 95 samples in a separate CoronaHiT run.

The unclassified CoronaHiT fastq reads were analysed in detail, which revealed that these reads all had 5' truncated barcode sequences. An example of the expected sequencing construct is provided below with the nanopore sequencing adapter in blue, the barcode green (barcode 1 sequence given here as an example) and the P5/P7 sequence (yellow):

5' AATGTACTTCGTTCAGTTACGTATTGCTCACATATCAGAGTGCGTCGTCGGCAGCGTC

However, the unclassified reads were missing some 5' barcode sequence e.g.

5' AATGTACTTCGTTCAGTTACGTATTGCT - - CATATCAGAGTGCGTCGTCGGCAGCGTC

The presence of the nanopore adapter sequence suggested that this is not a bioinformatic truncation of the read but a physical truncation of the barcoded primers. This was likely caused by truncated primer sequences being synthesised by the oligonucleotide provider [292]. It was postulated that the barcodes were too short for nanopore sequencing, given the sequencing error rate and exacerbated by the shorter primers. To overcome this issue, the barcodes were redesigned based on the design of ONT barcodes. The new

barcode constructs increased barcode sequence length from 16 to 24 bp (the same as used by ONT). A 7 bp buffer sequence was added at the 5' end to mitigate for any truncation of the primer sequence during manufacture. Additionally, a 7 bp spacer sequence was added between the barcode and the Nextera adapter complement as ONT have in their barcode constructs. The demultiplexing parameters in the analysis pipeline were also changed to include the new flanking regions of the new barcodes.
The new barcode primer was construdted as follows (the 24 Ns represent the unique barcode):

5' GGTGCTGNNNNNNNNNNNNNNNNNNNNNNNNTTAACCTGTCTCGTGGGCTCGG

A total of 96 unique barcodes were made with the new design and tested in the CoronaHiT method on the same samples. The unclassified rate dropped from 50% with the previous set of barcodes to 34.2%, closer to the rates seen in ARTIC library preparation (28.4%). Additionally, the evenness between barcodes also improved (Figure 3.12), with the coefficient of variation reducing from 0.69 with the previous set to 0.36 with the new set. There were two outliers in terms of low yield. One was definitively determined to be human error and therefore excluded from further analysis. The second had no obvious cause, so was not excluded.

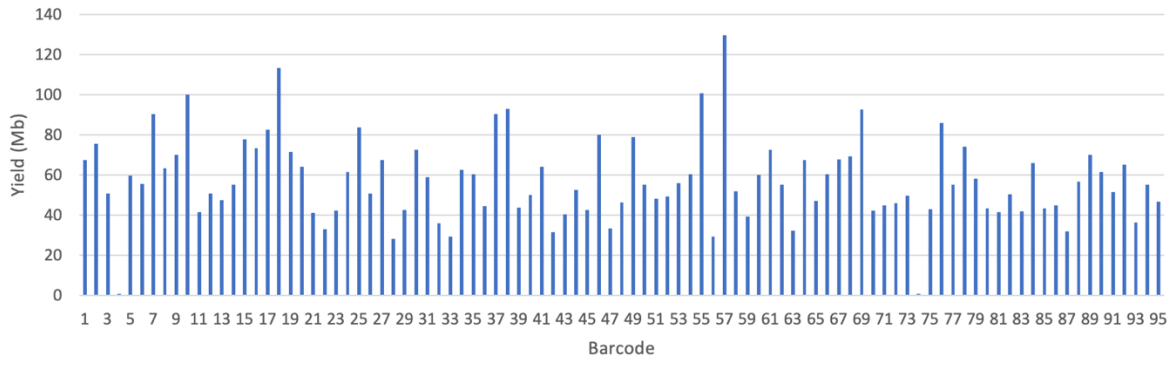


Figure 3.12 - Yield per barcode with the new barcode constructs, showing improved evenness compared to the previous experiment (Figure 3.9C).

3.4.2 CoronaHiT versus ARTIC ONT

A second CoronaHiT sequencing run was performed with the new CoronaHiT barcodes with approximately half the number of samples (48) to test the impact of sequencing yield on the quality of the genomes. A subset of 23 samples (due to limited barcodes) was tested with the ARTIC ONT protocol as a control. The two methods produced different amounts of data, affecting coverage. CoronaHiT with 94 samples yielded 9.2 Gb of data after 43 hours of sequencing resulting in 757X average coverage per sample. CoronaHiT with 48 samples yielded slightly more data in the same sequencing time, 11.1 Gb. Average coverage per sample increased to 2037X (due to lower sample number and higher sequencing yield). The ARTIC sequencing produced 9 Gb of data after 23 hours of sequencing resulting in 7509X average coverage for the 23 samples.

The coverage impacted the quality of the genomes. Two quality thresholds were used for determining whether a genome was of sufficient quality to be uploaded to COG-UK and GISAID databases. The COG-UK QC threshold was > 50% genome coverage (at ≥20x for nanopore and 10x for Illumina) and the presence of at least 1 contiguous assembled sequence of >10,000 bases (a third of the SARS-CoV-2 genome). The GISAID database upload QC was stricter, requiring >90% genome coverage. Both methods had a similar pass rate for the COG-UK QC with 81.3% (48 sample run) and 80.0% (94 sample run) of samples passing the COG-UK QC using CoronaHiT compared to 82.6% for ARTIC. However, for the higher GISAID QC, the gap was wider, with 73.9% passing for ARTIC and 66.7% passing for CoronaHiT (48 samples) and 65% for CoronaHiT (94 samples). For samples with lower C_T than 32 (higher viral load), 100% of the samples passed GISAID QC for ARTIC, whereas 94.1% passed for CoronaHiT (48 samples) and 92.2% for CoronaHiT (94 samples) (Table 3.17).

Table 3.17 – Comparison of QC pass rates between ARTIC ONT and CoronaHiT with different numbers of samples sequenced on one flowcell.

Results for all samples			
	CoronaHiT-ONT (48 samples)	**CoronaHiT-ONT (94 samples)**	**ARTIC ONT (23 samples)**
Passing COG-UK QC	81.3%	80.0%	82.6%
Passing GISAID QC	66.7%	65.0%	73.9%
Failing COG-UK QC	18.8%	20.0%	17.4%
Failing GISAID QC	43.2%	35.0%	26.1%
Median Ns of COG-UK passed	128	490	121
Average SNPs of COG-UK passed	8.3	8.0	8.4
Results for samples with $C_T \leq 32$			
Passing COG-UK QC	97.05%	98.4%	100%
Passing GISAID QC	94.11%	92.2%	100%
Failing COG-UK QC	2.95%	1.6%	0%
Failing GISAID QC	5.89%	7.8%	0%
Median Ns of COG-UK passed	128	369	121
Average SNPs of COG-UK passed	8.4	8.0	8.3

The higher number of samples was associated with an increase in Ns (positions where the minimum depth of coverage was not reached so a confident basecall could not be made). CoronaHiT with 94 samples had a median of 490 Ns in samples that passed the COG-UK QC in contrast to the CoronaHiT with 48 samples which had a median of 128 Ns, similar to the ARTIC ONT method (121 Ns).

Since coverage is directly related to the number of samples, it was difficult to directly high and lower throughput methods. To make comparison between methods easier, downsampling was used to provide an average of 900X per sample for all methods. The CoronaHiT method still contained more Ns in the genome assemblies than the ARTIC

ONT method, with samples below C_T 22 typically still containing Ns using CoronaHiT (Figure 3.13).

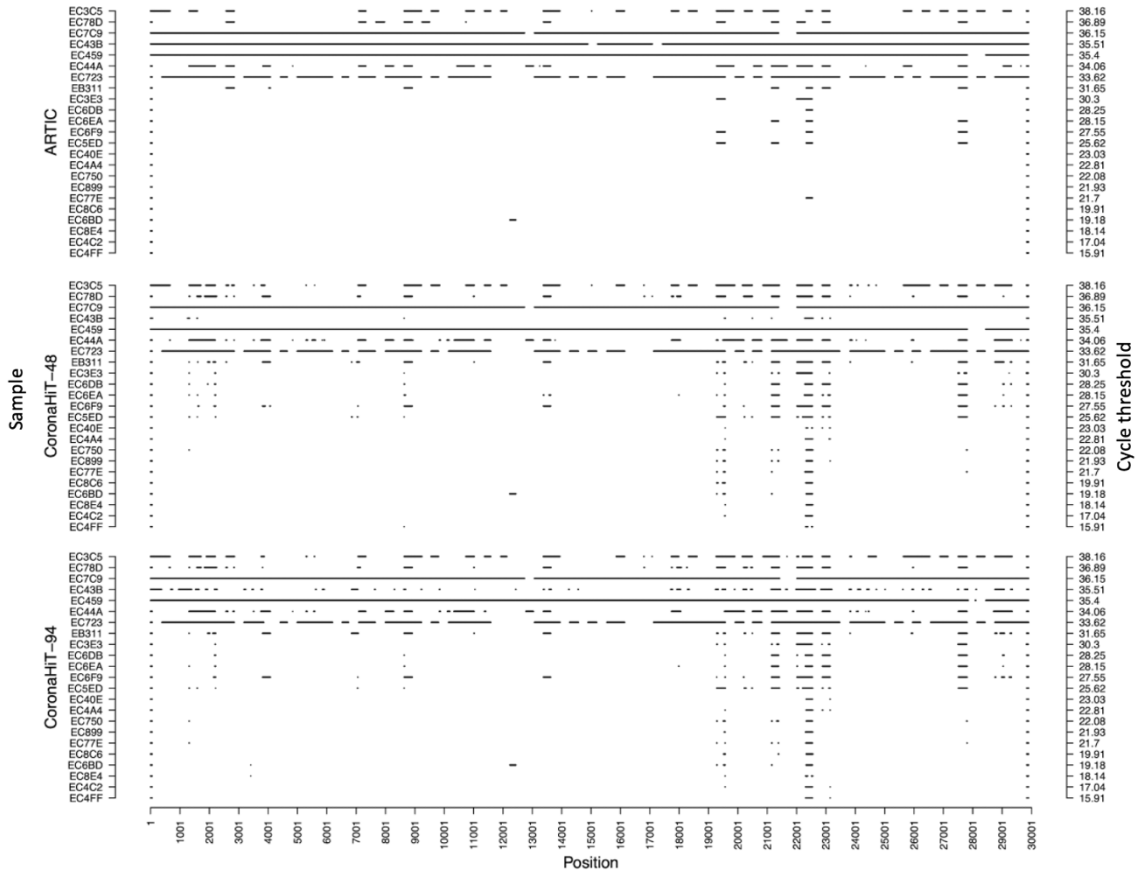


Figure 3.13 – Position of Ns (unknown bases) for the CoronaHiT method versus the ARTIC method for the downsampled data showing that number of Ns increases with C_T.

3.4.3 Optimisation of CoronaHiT

Further optimisation of the CoronaHiT method was required to improve the yield of the sequencing, reduce the amount of time and reduce the cost. Firstly, the transposase was switched from Nextera to Nextera Flex (Illumina). This is the same transposase, however, the enzymes are attached to beads at uniform density. This means that there is a regular distance between the transposases which dictates the frequency at which the amplicons are fragmented. The advantage of this approach is that DNA can be mixed with the Bead Linked Transposase (BLT) without normalisation, as the transposase can only make a limited number of cuts regardless of the sample concentration (whereas with Nextera, the

concentration ratio dictates the length of the fragmented products). Additionally, the individual sample washes were removed, and instead samples were diluted 1/5 in nuclease free water to reduce PCR carryover into the Tagmentation reaction. This removed the need for laborious individual sample washes, quantification, and normalisation, which can take 1-2 hours depending on the number of samples. To increase method efficiency in the absence of washing and normalising the samples, the incubation for the end-prep step (post-barcoding PCR) was increased from 5 to 15 minutes, and the nanopore adapter ligation step was increased from 10 to 20 minutes.

Changes were also made to the default Illumina protocols for SARS-CoV-2 sequencing by Dave Baker. Firstly, the Flex transposase step, which is used for CoronaHiT was also used for Illumina sequencing, and in both cases, the reactions were miniaturised. BLT is normally used at 11 μL per sample according to manufacturer's instructions, however, this was reduced to 0.5 μL per sample, a >20-fold reduction. Additionally, the post-tagmentation deactivation and clean-up step was eliminated, so that the tagged products could be added directly to the barcoding PCR. Finally, the Illumina indexes were replaced with custom barcodes based on a paper by Perez-Sepulveda et al. (2021) [293], consisting of 438 bespoke 9 bp barcode sequences designed for Illumina sequencing. The P5 and P7 sequences remain the same and the new barcode sequences were added at the 5' end. This expanded the multiplexing capacity of Illumina to 438 samples per flowcell if using the barcodes symmetrically or 191,844 if using the barcodes asymmetrically (different barcodes at both ends).

These changes meant that the library preparation procedure for sequencing ARTIC PCR products on Illumina and nanopore were the same, with the exception of the different barcodes and sequencing adapters used. The CoronaHiT method was therefore platform agnostic and is hereafter referred to as CoronaHiT-ONT and CoronaHiT-Illumina.

3.4.4 Optimised CoronaHiT versus ARTIC LoCost method

During optimisation of CoronaHiT, ONT released the Native Barcoding Expansion kit (EXP-NBD196) which increased the number of available barcodes from 24 to 96. This was used in an improved version of the ARTIC method, ARTIC LoCost, which reduced the cost and turnaround time of the method and expanded the capacity to 96 samples [269]. The CoronaHiT method (nanopore and Illumina) was compared to the updated ARTIC method on two sets of samples, 95 routinely processed COG-UK samples and a set of 59 rapidly processed samples associated with a local outbreak (hereafter referred to as the rapid response set). For the routine samples, 30 hours of sequencing data was used for CoronaHiT-ONT and ARTIC LoCost, and for the rapid response set, 18 hours of sequencing data was used. The full dataset was used for both CoronaHiT-Illumina runs as these runs included samples unrelated to this study. CoronaHiT-ONT and ARTIC LoCost were directly comparable (same sample number and sequencing time).

3.4.4.1 Sequencing metrics and depth

Post demultiplexing of the nanopore runs, 74.7% of the reads were successfully classified for the routine dataset, and 81.9% of reads were classified for ARTIC LoCost. For the rapid response set, 69.6% of reads were classified for CoronaHiT-ONT and 71.6% for ARTIC LoCost, suggesting there is variation depending on the samples, but CoronaHiT-ONT performed similarly to ARTIC-LoCost in the rapid set. The negative controls in the nanopore runs had no reads map to SARS-CoV-2. The CoronaHiT-Illumina routine dataset had mapped reads in the negative control, however, only 4 were >40 bp and the remainder were primer-dimers which did not impact the results.

For the routine samples, average depth of coverage per sample for CoronaHiT-ONT was 1145X, for ARTIC LoCost was 1719X and for CoronaHiT-Illumina was 4649X (Table 3.17). For the rapid response data, average coverage was 1104X for CoronaHiT-ONT, 1421X for ARTIC LoCost and 3010X for CoronaHiT-Illumina. CoronaHiT-ONT had less

variation in sample coverage compared to the ARTIC LoCost runs in both the routine and rapid response datasets, with lower standard deviation relative to the mean (Table 3.18). The lower coverage for CoronaHiT-ONT compared to ARTIC LoCost was largely due to shorter reads with a higher proportion of adapter sequence (due to the transposase fragmentation and longer barcodes). ARTIC LoCost produced average read sizes of 448 and 457 for the two runs (Table 3.18) with average of 386 and 384 bases bases mapping. CoronaHiT-ONT on the other hand produced an average read length of 374 bp for the routine dataset with only 205 bp mapping. For the rapid response samples, a minor adjustment was made to the CoronaHiT-ONT procedure whereby the 0.8X SPRI clean after the barcoding PCR was decreased to a 0.6X SPRI clean select for longer reads. This increased average read length to 413 bp for the rapid response dataset, which also increased the average mapped read length to 241 bp. This change improved the depth of coverage for CoronaHiT-ONT in comparison to ARTIC. CoronaHiT-Illumina produced the shortest mean mapped read length (135 and 131 bp for the two datasets); this was due to the sequencing chemistry used (paired end 151 bp chemistry).

Table 3.18 - Run metrics for the CoronaHiT-ONT, CoronaHiT-Illumina and ARTIC LoCost runs. † Illumina runs contained other samples in addition to the test samples.

	Routine samples			**Rapid Response samples**		
	CoronaHiT -ONT	**ARTIC LoCost**	**CoronaHit -Illumina**	**CoronaHiT -ONT**	**ARTIC LoCost**	**CoronaHiT -Illumina**
No. of samples	95	95	95	59	59	59
Run time (h)	30	30	25.4	18	18	24.4
Yielded bases (Gb)	10.3	8.5	43.9†	6.3	4.8	48.6†
Bases demultiplexed (Gb)	9.6	8.0	15.7	5.7	4.5	7.3
Number of 1000s of reads sequenced (>Q7)	24,765 k	15,733 k	113,756 k	13,045 k	8,824 k	53,678 k
Average PHRED score	13.47	13.11	33.15	13.2	12.98	33.48

Average coverage (X)	1145X	1719X	4649X	1104X	1421X	3010X
Standard deviation of coverage (X)	698X	1683X	4352X	439X	1145X	3496X
Average read length (bases)	374	448	135	413	457	135
Average (Mean) mapped length	205	386	135	241	384	131

3.4.4.2 Genome assembly pass rates and breadth of coverage

The COG-UK and GISAID QC criteria are directly dependant on the breadth of coverage (defined in section 3.4.2). The breadth of coverage was correlated with SARS-CoV-2 C_T values in the clinical samples. Samples with low genome coverage were associated with higher C_T values (Figure 3.14). Samples with a C_T above 32 (approximately 100 viral genome copies in the PCR reaction) generally failed COG-UK and GISAID QC thresholds with the number of Ns (missing bases in the consensus genome) increasing steeply for samples with C_T >32 (Figure 3.14). CoronaHiT-ONT and Illumina perform better than ARTIC LoCost in terms of the number of Ns. Comparing the routine samples with a C_T of 32 or below, the average number of Ns was 815 for ARTIC LoCost, whereas it was 682 and 111 for CoronaHiT-ONT and Illumina respectively. If the higher C_T samples were also included, the average number of Ns increases to 1635 for ARTIC LoCost, 1504 for CoronaHiT-ONT and 688 for CoronaHiT-Illumina. The difference in the number of Ns in the sequence in ARTIC compared to CoronaHiT is more pronounced at higher C_T (Figure 3.14).

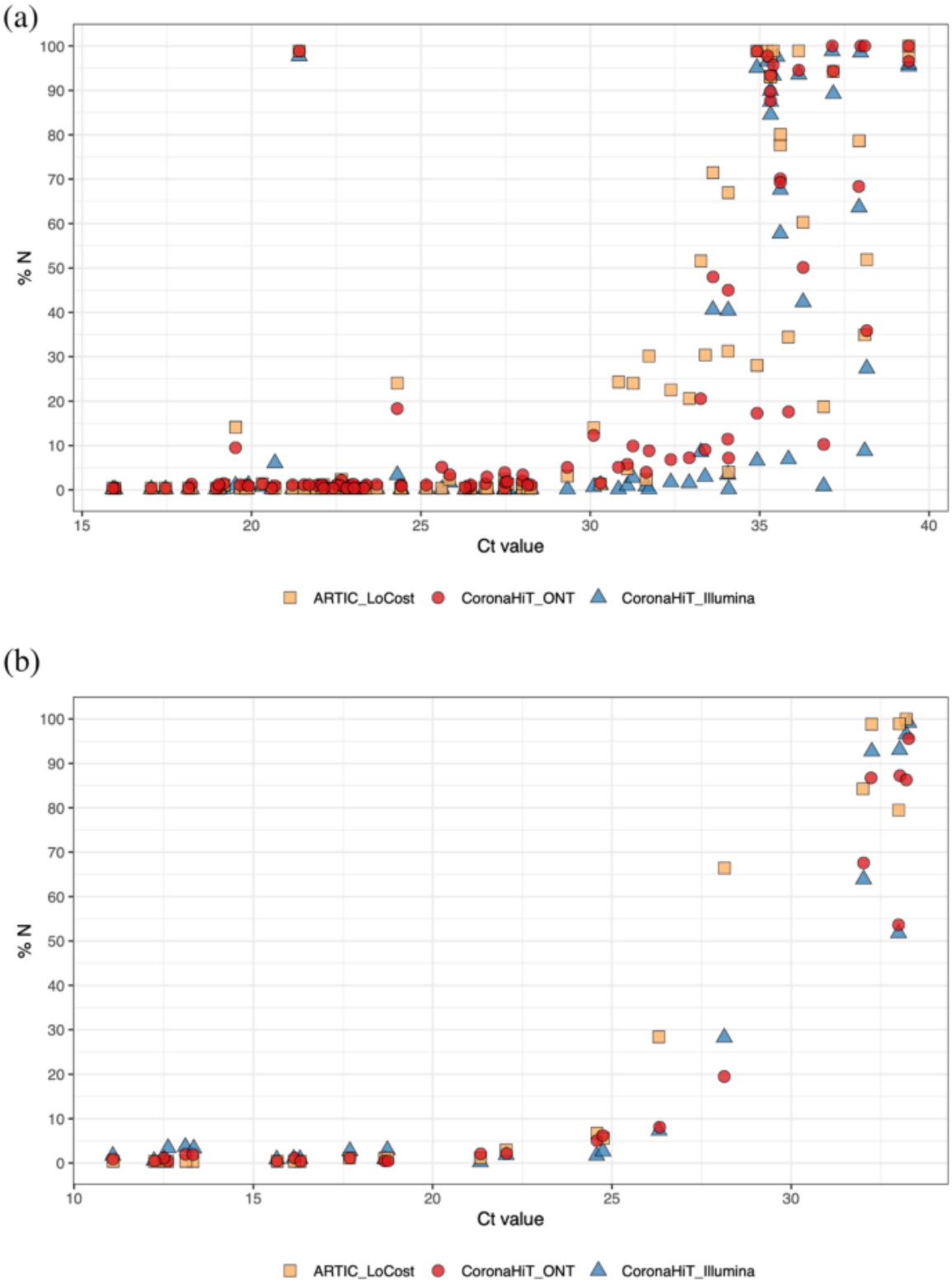


Figure 3.14 – Coverage of samples (represented by % of N positions) against the sample C_T for A) the routine dataset and B) the rapid response dataset. [266]

The difference in the percentage of Ns in the consensus genomes using the different methods impacts the proportion of samples passing COG-UK and GISAID QC thresholds.

When considering samples with C_T <32, a similar number of samples passed the COG-UK threshold for the methods; 81 samples pass for ARTIC LoCost, and 82 pass for both CoronaHiT-ONT and CoronaHiT-Illumina. However, taking all sequenced samples (including high C_T samples), the gap between the methods widens. The number of samples passing the COG-UK QC criteria for all samples is 117 for ARTIC LoCost, 124 for CoronaHiT-ONT and 126 for CoronaHiT-Illumina (full breakdown in Table 3.19). Despite having lower average depth of coverage compared to ARTIC LoCost, more CoronaHiT-ONT samples passed QC thresholds. The stricter GISAID QC threshold resulted in a lower genome pass rate for all methods, however, CoronaHiT still outperformed ARTIC LoCost. 99 samples passed the GISAID QC for ARTIC LoCost, 110 for CoronaHiT-ONT and 118 pass for CoronaHiT-Illumina (Table 3.18). This means the overall pass rate for the GISAID QC threshold was 64.3% for ARTIC LoCost, 71.4% for CoronaHiT-ONT and 76.6% for CoronaHiT-Illumina. With higher viral load samples (≤C_T 32), the GISAID QC pass rate was 89.2% for ARTIC LoCost, 95.2% for CoronaHiT-ONT, and 97.6% for CoronaHiT-Illumina (breakdown in Table 3.19).

Table 3.19 - Comparison of QC pass rates between the sequencing methods

	Routine samples			Rapid Response samples		
	CoronaHiT-ONT	**ARTIC LoCost**	**CoronaHiT-Illumina**	**CoronaHiT-ONT**	**ARTIC LoCost**	**CoronaHiT-Illumina**
No. of samples sequenced	**95**	**95**	**95**	**59**	**59**	**59**
Consensus genomes	98.95% (94)	96.84% (92)	100% (95)	96.61% (57)	91.53% (54)	100% (59)
Passing COG-UK QC	80.00% (76)	76.84% (73)	82.11% (78)	81.36% (48)	74.58% (44)	81.36% (48)
Passing GISAID QC	69.47% (66)	62.11% (59)	77.89% (74)	74.58% (44)	67.80% (40)	74.58% (44)
Failing COG-UK QC	20.00% (19)	23.16% (22)	17.89% (17)	18.64% (11)	25.42% (15)	18.64% (11)
Failing GISAID QC	30.53% (29)	37.89% (36)	22.11% (21)	25.42% (15)	32.20% (19)	25.42% (15)
Avg. (Mean) Ns of COG-UK passed	1504	1635	688	977	1101	911
Avg SNPs of COG-UK passed	7.99	7.99	11.0	18.3	18.2	20.4
Results for samples with C_T ≤32						
No. of samples	**65**	**65**	**65**	**18**	**18**	**18**
Consensus genomes	100% (65)	100% (65)	100% (65)	100% (18)	100% (18)	100% (18)
Passing COG-UK QC	98.46% (64)	98.46% (64)	98.46% (64)	100%(18)	94.44% (17)	100% (18)
Passing GISAID QC	95.38% (62)	89.23% (58)	98.46% (64)	94.44% (17)	88.89% (16)	94.44% (17)
Failing COG-UK QC	1.54% (1)	1.54% (1)	1.54% (1)	0% (0)	5.56% (1)	0% (0)
Failing GISAID QC	4.62% (3)	10.77% (7)	1.54% (1)	5.56% (1)	11.11% (2)	5.56% (1)
Average (mean) Ns of COG-UK passed	682	815	111	895	911	1064
Average SNPs of COG-UK passed	8.19	8.17	10.2	18.8	18.9	20

3.4.4.3 Relatedness of samples sequenced using the different methods

Maximum likelihood trees were constructed for all samples to investigate any differences in clustering between the same samples sequenced using the different methods. The 72 consensus genomes that passed the COG-UK QC from the routine dataset and the 44 passed genomes from the rapid response dataset were used. When the consensus genomes were placed on a phylogenetic tree (Figure 3.15A and B) for the routine set, all three methods had the same clustering pattern, except for three samples (EB1DB, EC741 and EC644). This was due to the presence of Ns in the ARTIC LoCost sequences and ambiguous bases in CoronaHiT-Illumina sequences. In all 3 of these samples, ARTIC LoCost had substantially lower breadth of coverage (Figure 3.16).

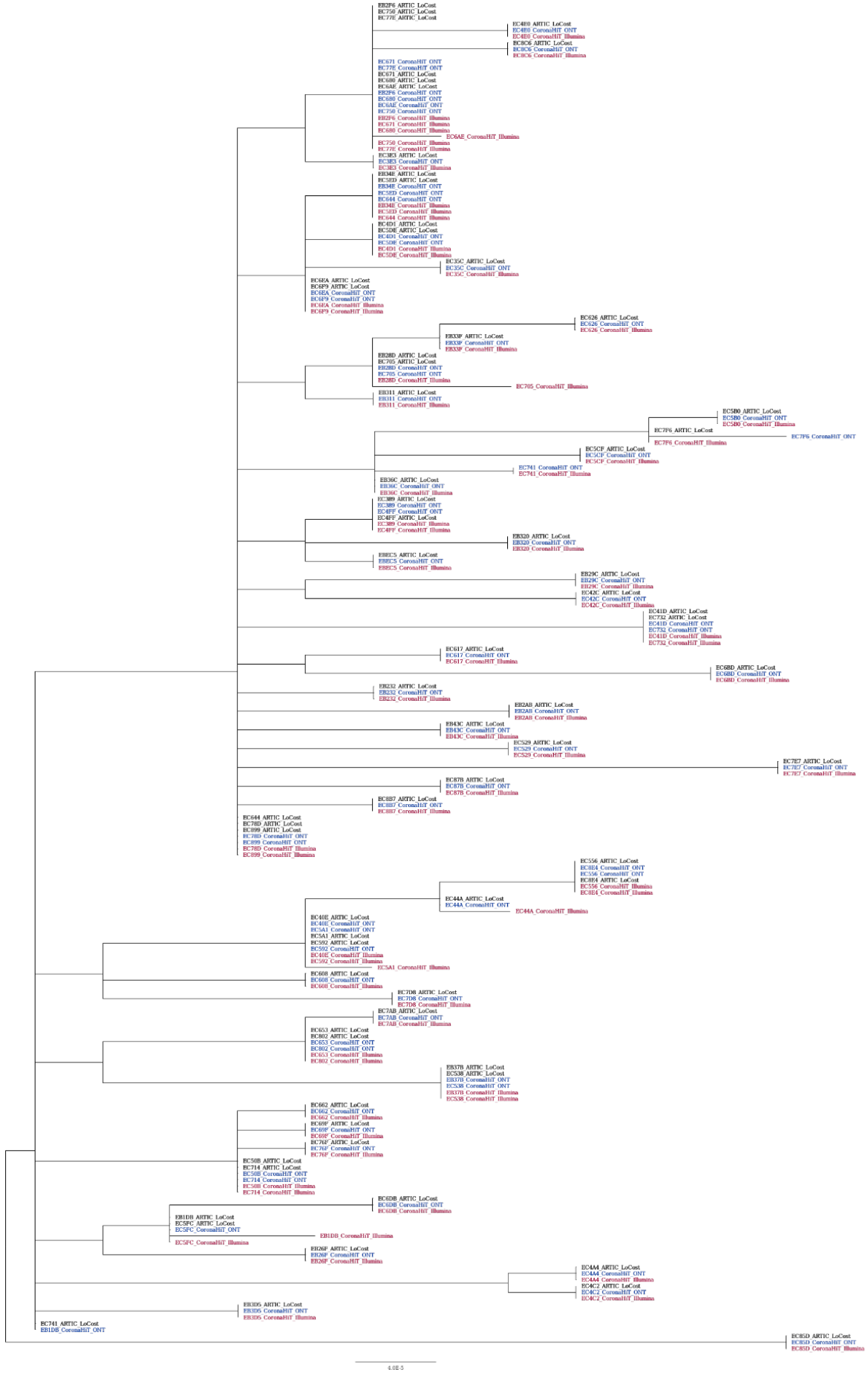


Figure 3.15A - Maximum likelihood tree of samples that generated a consensus genome for the routine set [266]

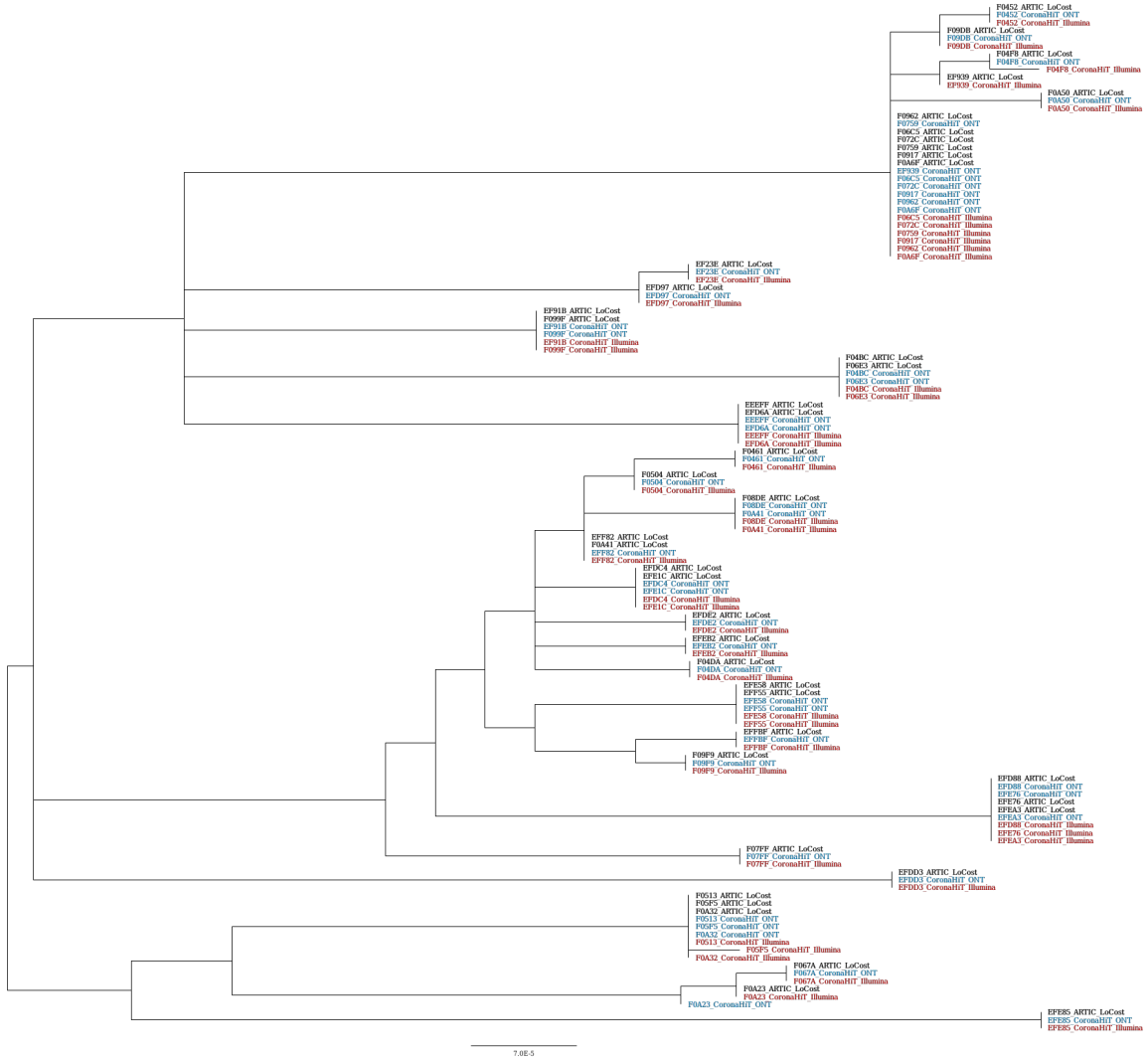


Figure 3.15B - Maximum likelihood tree of samples that generated a consensus genome for the rapid response set [266]

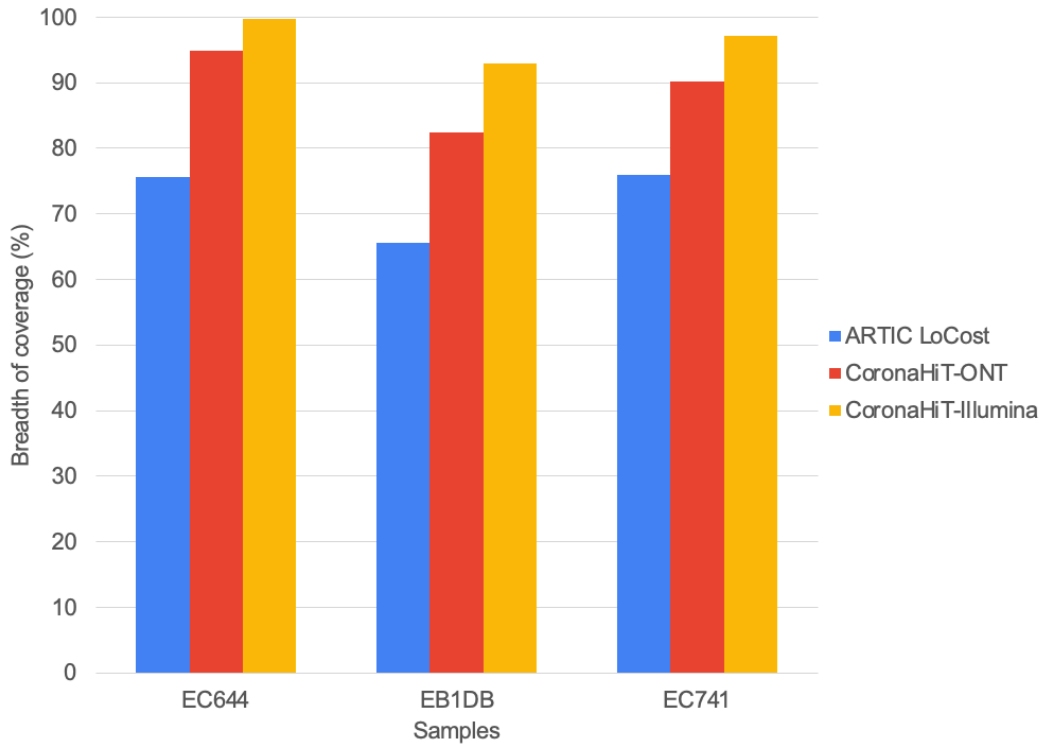


Figure 3.16 – Breadth of coverage for the 3 samples with clustering discrepancies

There were only two SNP discrepancies between consensus genomes generated using the different methods. One SNP difference in sample F04F8 between CoronaHiT-ONT and CoronaHiT-Illumina, with ARTIC LoCost showing an N and the other in sample F0A23, with CoronaHiT-ONT disagreeing with CoronaHiT-Illumina and ARTIC LoCost (Table 3.20). However, these SNP differences did not affect the classification of the samples (i.e. the closest sequence in the database was the same for all methods). The only other difference detect, which led to varying branch lengths, was related to ambiguous Illumina base calls i.e. a base that cannot reliably be called as an A, T, G or C is given a letter from The International Union of Pure and Applied Chemistry (IUPAC) code. The average number of SNPs between the SARS-CoV-2 reference genome (Genbank: MN908947.3) and the consensus genomes were similar for the two nanopore methods, 8.17 for ARTIC LoCost and 8.19 for CoronaHiT-ONT for the routine samples, and 18.9 and 18.8 respectively for the rapid response dataset. The average number of SNPs was higher for CoronaHiT-Illumina, 10.2 for the routine dataset and 20 for the rapid response dataset (Table 3.19). This was related to the ambiguous Illumina bases.

Table 3.20 – SNP discrepancies between the library preparation methods

Sample	Sequencing method	SNP difference	Breadth of coverage
F04F8	ARTIC LoCost	N	84.7%
	CoronaHiT ONT	A	97.0%
	CoronaHiT Illumina	G	99.8%
F0A23	ARTIC LoCost	T	99.6%
	CoronaHiT ONT	C	99.6%
	CoronaHiT Illumina	T	99.8%

Despite these small discrepancies which did not have an impact on lineage determination, we have shown that CoronaHiT is a powerful high throughput alternative to the ARTIC LoCost method, which produces higher quality genomes in low viral load samples.

Our CoronaHiT method has been published in Genome Medicine on which I am a joint first co-author (Appendix 4).

3.5. Epidemiology and local outbreak surveillance

Throughout 2020 and early 2021, I was involved in the COG-UK consortium, aiding in the set-up of SARS-CoV-2 sequencing at the QIB, developing CoronaHiT for higher throughput sequencing and using a combination of ARTIC and CoronaHiT methods to sequence thousands of local and national SARS-CoV-2 samples. Local Norfolk and surrounding region samples – including samples from the 'Rapid response' dataset sequenced during the development of CoronaHiT by me – were analysed by QIB staff (led by Andrew Page) to study the local epidemiology [294].

Analysis revealed the transient nature of lineages, with some lineages being identified and going extinct in the region within a month and being replaced by others. While the main lineage in the region at the time B.1.1 (44.83% of cases) was the same as the main lineage in the UK (39.7% of cases), some other lineages were over- and under-

represented in Norfolk relative to the rest of the UK. For example, B.1.11 was 10.7% of cases in Norfolk while it was only 2.0% in the rest of the UK (Figure 3.17).

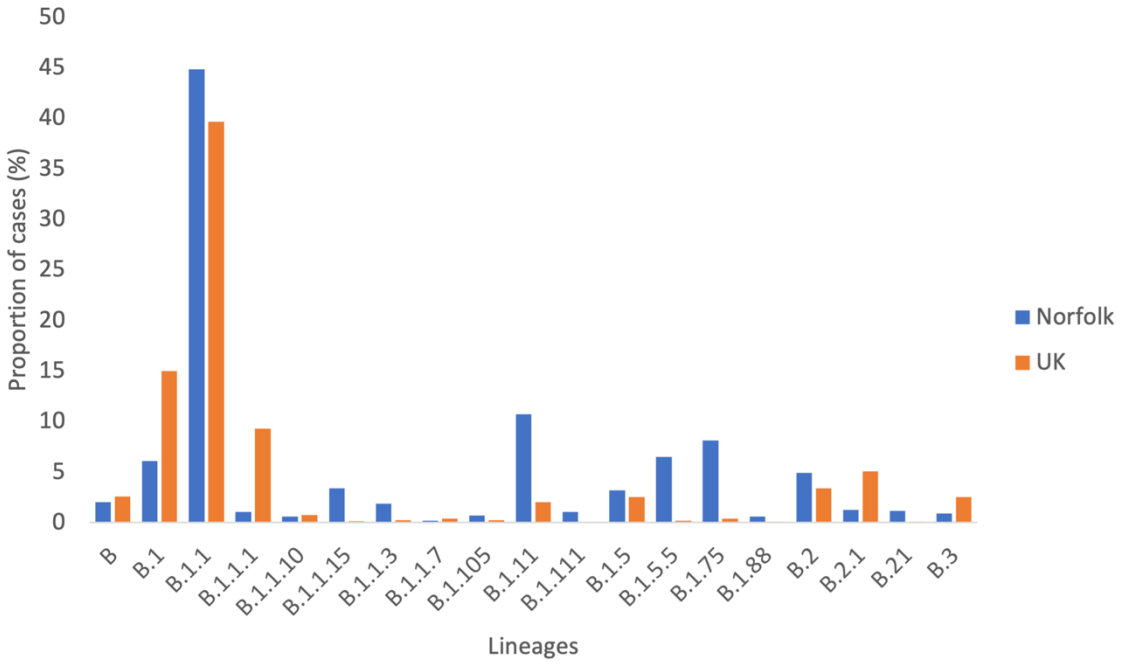


Figure 3.17 – Proportion of lineages in Norfolk versus the rest of the UK during the first wave of the pandemic

In August 2020, an outbreak was detected in a food processing facility in Norfolk. More than 120 staff members tested positive for COVID-19 forcing the shutdown of the facility [295]. 35 of the positive samples were sequenced in the rapid response samples described previously the same day they were received, with results the next day (ARTIC-ONT data was generated first, so this was used for the outbreak analysis). 27 of the samples generated lineages and were all shown to be B.1.1.15. Not only was B.1.1.15 rare in the UK at the time with only 7 other samples identified the previous month, the other UK B.1.1.15 samples lacked 3 signature SNPs that samples from the food facility outbreak had. The other UK samples had two SNPs of their own that were absent in the outbreak set. With an evolutionary rate of 2 SNPs per month, all these data suggest that there was a common ancestor 1-2 months prior [294]. Analysing global data from GISAID, the ancestral B.1.1.15 lineage was only observed in Portugal in that timeframe, strongly suggesting an importation event from abroad. In contrast to the food processing facility

outbreak, analysis of a different dataset from a suspected hospital outbreak revealed that there were multiple co-occurring lineages from the samples sequenced, ruling out a single large hospital outbreak. Our data show the power of genomic epidemiology and has been published

3.6. Additional collaborative studies

As part of my PhD, my expertise in nanopore sequencing and SARS-CoV-2 led to small collaborations with other groups at QIB to aid in metagenomics and SARS-CoV-2 detection projects. These studies were published and are very briefly detailed below.

3.6.1 Metagenomics of the human gut microbiome

Bacteria in the gut are capable of fulfilling functions that human guts themselves are not capable of, for example breaking down certain complex carbohydrates [296]. This symbiotic relationship is therefore important for human gut health. However, the impact of certain carbohydrates on the gut microbiome composition is not fully understood. Led by Fred Warren (Group Leader, QIB), research was undertaken to use metagenomics to profile model guts at different timepoints in response to 6 different types of carbohydrates. Both Illumina and nanopore PromethION sequencing were used (with PromethION sequencing performed by me) to enable hybrid metagenome assembled genomes (MAGs). This led to the assembly of 509 high-quality MAGs which showed enrichment of certain species with predicted carbohydrate degrading functions over time. This study also demonstrated the value of using long-read sequencing technology for MAGs. Even though the DNA was sheared (ranging from 1400-5200 bp) due to the extraction, use of nanopore to supplement Illumina data still led to higher N50s and fewer contigs. This work has been published as a preprint on BioRxiv [297] (Appendix 6) and has been accepted for publication in Communications Biology.

3.6.2 Detection of SARS-CoV-2 in stool

SARS-CoV-2 can be detected in stool samples of patients who are or were recently infected, and viable SARS-CoV-2 has been isolated from this sample type [298]. This is a risk for patients receiving faecal microbiota transplantations (FMT), so a highly sensitive procedure is required for screening of SARS-CoV-2 from stool. During the pandemic, I tested and characterised the Centers for Disease Control and Prevention (CDC) RT-qPCR assay at QIB using the Probe 1-Step Go No Rox (PCR Biosystems) and LightCycler 480 (Roche) for the detection of SARS-CoV-2. This was then used for detection in stool extracts in a study led by Arjan Narbad's (Group Leader, QIB) team. With the combination of their optimised RNA extraction from stool and the RT-qPCR assay, the method was shown to have an LOD as low as 1 viral particle per mg of stool. This work has been published in BMC Microbiology [299] (Appendix 7).

4. Discussion

4.1 Rapid clinical metagenomics

LRTIs are a leading cause of morbidity and mortality globally. For appropriate treatment to be given, the cause of infection needs to be determined as soon as possible. Currently, most antimicrobials are prescribed empirically, which can lead to inappropriate use. This is associated with poor patient outcomes, increased healthcare costs, and can lead to the emergence of AMR [300]. In the US alone, 27 million people annually are prescribed antibiotics unnecessarily for respiratory issues [37].

Microbiological culture is the current gold standard for identifying bacterial and fungal pathogens but can take 24-48 hours for results, with a further day required for antimicrobial susceptibility testing. Additionally, culture is insensitive, with no causative pathogen being reported in the majority of cases. This means that current culture methods are inadequate and need to be replaced or supplemented with better methods. Targeted molecular methods such as PCR panels are becoming more widely used as they can provide results faster, however, they are not comprehensive and miss unusual pathogens and can target only a few common resistance genes. CMg is a potential solution to these problems, providing rapid, comprehensive results.

4.1.1 Faster CMg

Our group previously developed a CMg method that could be applied to lower respiratory tract samples to determine the cause of LRTIs such as pneumonia. This method takes approximately 6-7 hours from sample to result, which is one of the fastest CMg methods published, with most methods taking over 7 to 20+ hours [78]. However, real-life processing factors and multiple samples can mean that the procedure can often take longer and therefore results may not be possible to obtain in the same working day. Additionally, the host depletion step is a manual 45-minute step which requires technical laboratory skills and training, which is a barrier to clinical implementation. To improve the

chances of implementation, the method needs to be simpler, with less room for human error, and faster, to be able to compete with other rapid molecular methods. We halved the sample-to-result time to 3.5 hours, as well as making it simpler, requiring less hands-on time (Figure 4.1).

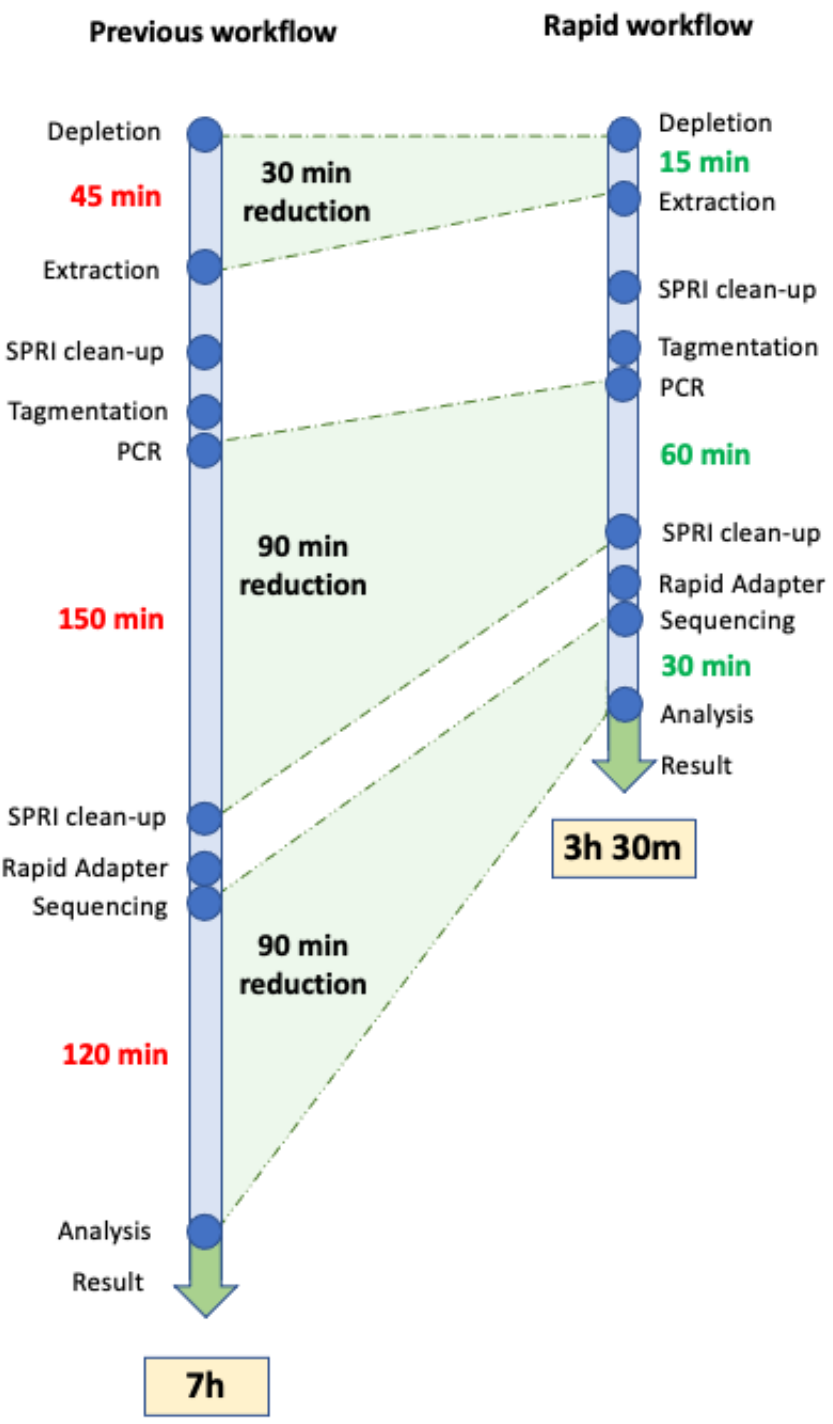


Figure 4.1 - The CMg workflow before and after optimisation, with the time reductions shown.

We reduced the PCR from 150 minutes to 60 minutes by switching to a faster polymerase and using a higher concentration of polymerase. Secondly, we significantly simplified the human DNA depletion, combining the cell lysis and nuclease treatments into one simultaneous incubation ('one pot') and removing 9 out of 16 steps (Table 3.7). This reduced the depletion from 45 to 15 minutes. Additionally, we showed that there was no difference in the pathogen identification result after 30 minutes of sequencing in 37 clinical samples, meaning that sequencing time could be reduced from our default of 2 hours to 30 minutes before a pathogen call is made (with further sequencing for more data being optional). We also demonstrated the potential of using lower throughput Flongle flowcells for CMg applications. Using Flongle allows samples to be rapidly tested as required, rather than being batched for cost effectiveness, which significantly impacts test turnaround time. These advances bring CMg a step closer to more routine clinical use. Identifying potential pathogens within hours will make diagnosis more reliable and guide early appropriate antimicrobial therapy.

While the difference between 7 hours and 3.5 hours may not seem significant, antibiotics should be administered as soon as possible when pneumonia is suspected. In the case of HAP, NICE guidelines state that antibiotic treatment should be initiated within 4 hours of being diagnosed, and 1 hour for patients who meet high risk criteria or are suspected of having sepsis [30]. This is also the case for CAP patients who have been admitted to hospital or are diagnosed in the community [301]. This means that 7 hours would not meet these guidelines and empirical treatment would likely have to be initiated, whereas a method providing results in half that time would mean that treatment could be administered based on rapid CMg results. These turnaround times do not take into account the steps before and after the laboratory processing of samples, such as the ordering of tests, transportation of samples, and reporting. These steps can take as much as 96% of the total turnaround time [302]. Therefore, improvements to the laboratory workflow are not enough. However, another benefit of shortening and simplifying the CMg method is that it decreases the necessity for it to be performed in specialised central labs.

The faster and less manual the method, the more likely it can be automated and processed near patient. Bedside testing of samples would reduce pre- and post-processing steps such as transport and reporting time. While this method is not automated yet, the simplification of the protocol and time-saving is a big step toward enabling this in the future.

The call for rapid diagnostics focuses heavily on how easily tests can be performed and how fast results can be delivered. The 2014 UK government AMR review calls for tests (such as those that can detect bacteria in sputum and susceptibility to antibiotics) that take 1 hour or less [37]. The Longitude Prize, which is a contest for solving global health challenges, is awarding £8 million for an accurate affordable and easy-to-use point-of-care diagnostic AMR test that takes 1 hour or less [303]. CMg has the potential to be highly impactful, even more so if it can meet these requirements and be used at rapid point-of-care.

4.1.2 Host depletion

Key to the new rapid turnaround is a simplified and fast one-pot host DNA depletion method which makes it possible to identify pathogens in respiratory samples with very little sequence data. Without a host depletion step, human DNA makes up the vast majority of sequencing reads (often >99%) from clinical samples [78]. To acquire enough microbial reads, sequencing time must be extended significantly, increasing test cost and turnaround time. Our group has previously shown that in 2 hours of sequencing, a depleted clinical respiratory sample produced over 40X coverage of the *E. coli* genome, whereas in the same time, its non-depleted control had less than 1X coverage [97]. This stark difference demonstrates why host depletion is essential to provide sufficient microbial reads in a short timeframe. Therefore, the question becomes which depletion method should be used. Many commercial host depletion/microbial enrichment kits exist on the market as described previously (Table 1.3). However, these can be expensive (~

$10-$39 per sample) and take too long (30-160 minutes). A comparison of the main methods on tissue samples showed that the 3 differential cell lysis methods performed significantly better than the methylated host DNA capture method, with the QIAmp DNA Microbiome and HostZERO Microbial DNA kits performing the best. These kits increased the percentage of bacterial reads 10-fold, from 6.7% to 71% (QIAmp) and 79.9% (HostZERO) [304]. The cheapest and fastest commercial method is the HostZERO Microbial DNA Kit which takes 30 minutes and costs approx. $10 per sample. This kit has also performed well on other sample types such as skin and saliva, showing that bacterial reads were the majority of reads after host depletion [305]. However, a recent study by Chen et al. 2022 which used the method on sputum samples to detect pulmonary infections showed that in the 50 samples processed, on average, only 11.5% of reads (median 4.8%) were microbial, showing that the method was potentially not as effective in sputum [306].

Our group previously developed an in-house host depletion method using saponin for human cell lysis, following by DNase treatment. While saponin has been well-described in the literature as being an efficient human cell lysis chemical [134], it is commonly used with nucleases such as DNase I. Key to our method's success in sputum is the combination of a high concentration salt buffer and a salt-tolerant nuclease HL-SAN. Previous work within the group (not shown) demonstrated that the depletion is not as efficient with other nucleases or lower salt concentration buffers, meaning that the high salt is an important element of efficient depletion. This is because high salt induces release of DNA from chromatin structures [307], making DNA more available for rapid nuclease digestion. The complexity of sputum is also likely to be a reason why high salt helps. Use of saline is an established method for reducing viscosity of mucus in cystic fibrosis patients - it works by reducing ionic bonds within the mucus and dissociating DNA from mucoproteins [308]. Therefore, the use of high salt may aid in the homogenisation of sputum and making DNA more available for nuclease digestion by freeing it from the mucus. Given the importance of salt in sputum, this necessitates the use of a nuclease

that is highly tolerant to salt, otherwise repeated salt and wash treatments may be required before nuclease digestions. Commonly used nucleases such as DNase I lose activity at high salt concentrations. HL-SAN is specifically activated by salt and is tolerant to high concentrations, so is ideal for this application. This method was shown to be highly effective, with 87.3% of reads being bacterial on average after depletion (median 97.8%) [97]. In addition, the reagents used makes the method cheaper than existing commercial methods as the method costs approximately $5 with the most expensive constituent being the HL-SAN enzyme.

The main downside was the turnaround time (45 minutes) and manual nature of the method. Additionally, it required know-how for robust performance, as incorrect pipetting at various stages can lead to loss of bacteria. We significantly simplified this method by combining the human cell lysis and nuclease treatment steps so that it could be performed simultaneously in the same incubation. The idea of a 'one-pot' method was novel and is the basis of a patent application titled "Method for digesting nucleic acid in a sample" (PCT/GB2020/052986 – on which I am an inventor). The method performs at least as well as the original method and takes 1/3 of the time (15 mins). In the 37 samples tested, on average 80.9% of reads were bacterial (median 98.8%). In some samples, only 10 of the >10,000 reads sequenced were human, showcasing the success of the depletion. Additionally, like the original method, 16S qPCR showed a <2 C_T loss of bacteria. While limited data exists comparing the methods directly, there is evidence that the shorter and simpler method leads to less loss of bacteria, which would be expected as there are fewer centrifugation, vortexing, and pipetting steps that can lead to loss of bacterial biomass. We did observe some loss of *S. pneumoniae*, and out of the 6 culture positive *S. pneumoniae* samples, this loss led to one false negative result. This issue was reported previously [97]. *S. pneumoniae* is a fastidious species that has an unusual property of undergoing autolysis, particularly when it reaches the stationary phase of growth, mediated by autolysins which degrade the cell wall [309]. This effect could potentially be induced by reagents/conditions used during the host depletion method, however, does not

always seem to have the same impact, as we observed losses ranging from 2-100-fold. It is possible that the length of time between sample collection and processing, the nature of the sample and even strain of *S. pneumoniae* may be factors that contribute to this. Further experiments could be performed to determine whether a particular element of our depletion method is the cause (i.e. testing each component individually), to investigate potential ways of reducing the loss.

4.1.3 PCR in CMg

Library amplification in CMg is not used in all protocols, with some methods forgoing a PCR step e.g. Pendleton et al. 2017 [310]. Downsides of universal amplification include the addition of a time consuming step to the protocol, the higher risk of contamination, and the introduction of bias leading to over/under-estimation of abundance [311]. However, avoiding library amplification PCR leads to lower sensitivity and diagnostic calls being made on very few reads. This was the case in Pendleton et al. 2017 study, where organisms were reported based on as few as 6 reads [312]. For good quality nanopore sequencing runs, the DNA inputs that are required for PCR-free library preparation are not typically achievable from clinical samples. Nanopore ligation kits require at least 1000 ng of DNA for the library preparation and PCR-free rapid kits require 400 ng of DNA, whereas host-depleted clinical samples typically have <1 ng/μl of DNA. We regularly process samples with DNA that is not quantifiable by Qubit post depletion and extraction. Use of amplification in CMg methods is therefore best practice in our opinion [313].

While PCR adds time to the protocol, the increase in the amount of sequencable DNA means that data is acquired faster during the sequencing, therefore, reducing turnaround time overall. Additionally, PCR bias may not be a large concern if the data is mainly used to diagnose (a qualitative result) rather than make inferences about the relative proportions of organisms in a community (e.g. for microbiome studies). However, there are corrective procedures that can also help ensure the results are accurate such as the

use of calibration controls, which can improve abundance prediction and even provide quantitative information (similar to cfu/mL estimates provided by semi-quantitative culture) [311]. Also keeping the number of cycles as low as possible can reduce bias introduced by PCR, as more cycles are associated with more bias [314]. We reduced the PCR time by switching to a faster polymerase and using a higher concentration of polymerase, and we were able to do this without having to increase the number of cycles. Using more cycles (35 instead of 25), added more time, more reads in the negative control and there was evidence that the proportions of *E. coli* and *S. aureus* diverged from those seen when using LongAmp Taq for 25 cycles. However, further experiments need to be performed with mock communities to characterise the effect of extra PCR cycles and if it is likely to impact the sensitivity for some organisms.

The choice of polymerase is important, due to the different properties that can affect the time and results. The fidelity of the enzyme affects how regularly errors are introduced. Fidelity is not the biggest priority for this application, as even the highest polymerase error rates are not enough to impact classification of species or detection of resistance genes, and is only likely to cause issues if attempting to detect minor frequency SNPs [315], where nanopore sequencing errors supersede polymerase errors. Nevertheless, the new GXL polymerase has higher fidelity (1 error per 21,000 bp) compared to LongAmp Taq (1 error per 7,000 bp). More importantly, GXL is significantly faster, capable of replicating 1 kb of DNA in 10 seconds, compared to 1kb in 50 seconds for LongAmp Taq. Other factors that are important are sensitivity and tolerance to inhibitors. Clinical samples, particularly those that contain blood, can be particularly inhibitory to PCR due to the presence of substances such as haemoglobin which is a known inhibitor and can persist even after extraction [316]. Therefore, not all polymerases are appropriate for clinical extracts and should be tested to see how well they perform. We found GXL polymerase worked well in amplifying from clinical extracts and was compatible with transposase-adapted DNA from sputum samples (RPB004 library preparation kit).

4.1.4 Sequencing technology and flowcell

The only sequencing technology on the market that can provide the rapid turnaround times required for infection diagnosis (within minutes/hours) is nanopore. This is largely what other groups also use for rapid CMg methods [95,313]. Sequencing time is the longest part of the procedure if other platforms such as Illumina are used, where runtimes exceed 16 hours. There are still benefits to CMg on these technologies as they can be used as a last resort method for aetiological investigation when other methods fail to determine a pathogen, however, it is not an adequate improvement on routine culture in terms of time.

As described previously (Section 1.4.3), ONT currently produces 3 sequencing technologies for different throughput needs. The PromethION is the highest throughput option, generating in excess of 120 Gb of data from a single flowcell. Given that the RPB004 library kit is limited to 12 barcodes currently, this significantly exceeds the required yields per sample, and is not cost effective. While wash and reuse of nanopore flowcells is an option, this is best performed when different barcodes are used in subsequent runs as repeated use of the same barcodes could lead to carryover and false positive results. MinION flowcells typically generate 10-20 Gb of data. In our published method, we show that with 2 hours of sequencing of 6 samples on a MinION, we generated on average ~115 Mb per sample, or ~40,000 reads [97]. In a full run, it is possible to generate in excess of 2 Gb per sample, or 700,000 reads, however, this would take more than a day of sequencing.

Flongle is the lowest throughput option provided by ONT with only 126 pores compared to a maximum of 2048 in MinION. The max yield is advertised as 1.8 Gb [317]. In our sequencing runs using Flongle, we averaged ~116 Mb in 2 hours, and for the full sequencing runs (letting the flowcell run for a full 24 hours), the maximum yield was 831 Mb. This lower yield is likely to do with teething problems with the product that meant they were not very stable. This was later reported to partly be a problem with the stability of the

reagents due to plastic contamination from the storage containers [318], and this was partly fixed with the change to glass containers for reagents [319].

Flongle flowcells can be used to sequence a smaller number of samples. In high throughput central laboratories, this may not be needed as there will be a high volume of samples meaning that batching is unlikely to be a bottleneck in the process. However, Flongle could allow for sequencing in urgent and low throughput settings where it is advantageous to run single samples when needed (e.g. small ICUs, if the method was safe and automated) and allow for near patient testing. We demonstrate that Flongle can be used to sequence 1-3 samples and produce enough data to make a pathogen ID call within 30 minutes of sequencing, and we show that the proportion of reads do not change substantially after 30 minutes, however, the amount of data produced in this time is limited and depends on the number of samples per run.

If two samples are sequenced on one flowcell for 30 minutes, the yield is 14 Mb per sample. Whereas if one sample is sequenced, the average yield is 24 Mb per sample after 30 minutes. If a single sample is sequenced and the pathogen significantly dominates the sequencing such as in S2 in our study where 95.3% of reads were *P. aeruginosa*, this means an average 4X coverage of the *P. aeruginosa* genome in 30 minutes and 15X in 2 hours. If two samples are sequenced, the data per sample are approximately halved. Due to uneven coverage of genomes, the higher the coverage, the more reliable the detection of present resistance genes. It has been estimated that approximately 7X depth may be sufficient for 99.9% coverage of a genome [109]. Additionally, if resistance detection is to include higher resolution analysis to determine alleles (e.g. for bla_{TEM} genes) or to detect single point resistance mutations, e.g. in *gyrA*, this requires higher sequencing depths (potentially approx. 30x) [320]. This may therefore require longer sequencing run times than 2 hours. Another thing to consider is that the detected pathogen does not always dominate the sequencing; for example, the *S. pneumoniae* in sample S8 only makes up 6.2% of reads, therefore significantly more data is required for good coverage of the genome. However, in our 37 samples, the average proportion of pathogen reads in the

data was 67.7% (median 74.4%), so most reported pathogens in our data dominate the microbial community present in the sample and hence dominate the sequencing data (once the human DNA is removed). Our data shows that Flongle can be used reliably for pathogen ID, even with multiple samples, but for reliable resistance detection, fewer samples may need to be run and for longer. Advances that improve the yield and improve the quality will make Flongle sequencing more viable for resistance detection. Assuming continued improvements, if the advertised yield of 1.8 Gb is met, it would be possible to sequence a pathogen such as *E. coli* (~4.5 Mb) with ~300X average coverage for the full run (24 hours), or ~50X in 4 hours (presuming the pathogen makes up ~70% of reads).

Flongle is also cost effective. The cost of running 1-2 samples on a Flongle is comparable to running 6-12 samples on a MinION (Table 4.1). Flongle can therefore be used to run single samples, which is not feasible on a MinION. However, MinION can also be used for running single samples to high depths followed by nuclease wash and re-use of the same flowcell with different barcodes. This is a way of using MinION flowcells in a cost-effective way while avoiding batching. However, there is the added step of washing flowcells in between whereas Flongle flowcells can be used and discarded. Additionally, sequencing a single sample with Flongle flowcells has the advantage of avoiding any potential barcode crosstalk, which is clearly important in clinical settings. However, a negative control should be added so a single sample per flowcell isn't practical when applying Flongle to diagnostic applications.

Table 4.1 – Cost of using Flongle versus using MinION for sequencing respiratory samples

	Flongle		MinION		
	Cost per 1 sample (£)	**Cost per 2 samples (£)**	**Cost per 1 sample (£)**	**Cost per 6 samples (£)**	**Cost per 12 samples (£)**
Host depletion	4	4	4	4	4
Extraction (MagNA Pure LC)	6.6	6.6	6.6	6.6	6.6
PCR (GXL)	4	4	4	4	4
Library prep	18.5	18.5	18.5	18.5	18
Flongle	75	37.5	-	-	
MinION	-	-	500	83	42
Total	**108**	**70**	**533**	**116**	**75**

4.1.5 Pathogen and resistance detection

We detected the correct pathogen in 19 out of 21 of microbiology positive samples when we applied the optimised CMg method to respiratory samples. Two pathogens reported by culture were missed, one of which was an *S. aureus* that was not detected by our species-specific qPCR. The negative qPCR suggests that the *S. aureus* wasn't present in the sample, as the qPCR assay used is highly sensitive (LoD <10 cell equivalents) [280]. We perform the qPCR on a non-host depleted extract of the sample in case bacteria were lost in the host depletion. However, the possibility that the *S. aureus* was not efficiently extracted cannot be ruled out. If the CMg result is correct as confirmed by the qPCR, this would mean that the reported *S. aureus* is a false positive by routine microbiology. Misidentification of this organism is possible and in fact is a common reported issue with *S. aureus* in particular as other *Staphylococci* and yeasts can resemble *S. aureus* morphologically and in some chemical tests [321]. However, we did not observe any yeasts or *Staphylococci* in the CMg results. Alternatively, it could be a case of contamination, as clinical microbiology laboratories are environments with pathogen contamination observed on many instruments, surfaces, and on the lab coat cuffs of the

technicians [322,323]. Ideally, to resolve cases like these while evaluating CMg in the future, samples would be analysed using multiple methods, including independent cultures and multiplex PCR tests.

The other organism not detected by CMg was a *S. pneumoniae* and that was detected by qPCR and therefore confirmed to be present. While *S. pneumoniae* reads were present in the CMg data, this was below the abundance threshold and was therefore classed as not detected. The reason for this false negative was related to the loss of *S. pneumoniae* during the host depletion step. This is a known issue that was discussed previously (in Section 4.1.3). However, this issue did not prevent the majority of *S. pneumoniae* positive samples being detected in the evaluation. From the 6 *S. pneumoniae* samples reported by routine microbiology, 5 were detected by our workflow.

An additional complication with *S. pneumoniae* is that, along with *H. influenzae*, it is a common pathobiont, meaning that it can be part of the commensal flora colonising the respiratory airways, but can also become a pathogen [324,325]. *S. pneumoniae* is a common cause of CAP [326] and is therefore an important organism to detect and report. In routine microbiology, reporting of *S. pneumoniae* as a pathogen relies heavily on the clinical context and quality of the lower respiratory tract sample, i.e. evidence of upper respiratory contamination is called as NRF [327]. In previous analysis of the CMg workflow in a clinical trial (INHALE) [328], we observed pathobionts such as *S. pneumoniae* and *H. influenzae* at low levels in a number of sputum samples (and not in the negative controls) – likely low level contamination from the upper respiratory tract. Therefore, CMg requires a system of distinguishing NRF from pathogen (discussed more in Section 4.1.6). Additionally, we observed significant metagenomic misclassification between *Streptococci* species. This is because members of the *Streptococcus* genus are genetically very similar due to horizontal gene transfer and homologous recombination [329]. *S. pneumoniae* for example is known to receive a significant amount of genetic material from commensals such as *S. mitis* in a unidirectional manner [330]. This means that metagenomic analyses can struggle to correctly classify these sequences,

complicating detection of *S. pneumoniae*. We mitigate this in our analysis by elevating the abundance threshold used to identify it in a sample (described in section 4.1.6).

There were 4 additional *S. pneumoniae* and 3 additional *H. influenzae* detections by CMg compared to culture – these were true positives that were not reported by culture, as confirmed by qPCR. There were 5 other organisms that routine microbiology did not detect, 1 of each of *E. coli, K. pneumoniae, P. aeruginosa, S. marcescens* and *M. catarrhalis*. These detections were all confirmed by species specific qPCR performed on the non-depleted extracts. Most of these organisms were present at low abundance in the CMg data and with high qPCR C_T, therefore their clinical significance is questionable, however, there was at least one example (S31) where evidence pointed to *K. pneumoniae* dominating the sample; 87.8% of the microbial reads were *K. pneumoniae* and the qPCR for *K. pneumoniae* returned a C_T of 22.2 (>10^5 genome equivalents in the PCR, therefore at least 10^6 cfu/mL). Cross-contamination or a processing mix-up on our behalf was ruled out as no other samples processed in the entire set were reported to have *K. pneumoniae*. It is likely that this organism was missed/not reported by routine microbiology. All these data point to CMg being more sensitive than routine microbiology.

We also showed that the microbial community profile largely stayed the same over sequencing time and the pathogens called by the pipeline did not change in any of the samples after 30 minutes. Between 15 minutes and 30 minutes of sequencing, all but one pathogen ID call stayed the same. The one sample that changed was due to the low number of overall reads which meant that more data was required. While the host depletion worked well for this sample (ΔC_T 13.24), the PCR did not work as this sample was a clear outlier in terms of sequencing reads (all samples had 2,000-11,000 reads after 30 minutes whereas this sample had 207 reads). The cause of this is likely the low biomass in the sample, with a post-depletion 16S C_T of 27.64, the highest bacterial C_T of any of the positive samples. This is close to the LOD of the method, therefore more time was required to get sufficient reads to identify the pathogen. This sample highlights the importance of an internal control to differentiate negative vs failed samples. A standard

quantity of internal control spiked at the beginning and known to produce reads in the absence of other bacteria would indicate whether a sample failed or performed sub-par. Internal controls such as DNA/RNA bacteriophages have been shown to be good process controls [106]. However, they were spiked into samples post-beadbeating, meaning that steps prior to this were not controlled for. An ideal process control would be introduced at the very start of the protocol (for example, in our case would be introduced prior to the host depletion).

In terms of resistance, we were limited by the largely susceptible nature of the organisms detected in the clinical samples. This is not surprising as the East of England has relatively lower incidence of resistant bacteria of concern such as MRSA [331] and Carbapenemase-producing Enterobacteriaceae (CPEs) [332]. We did however show that we were capable of detecting some genes that explain phenotypic resistance, however, some of these detections would be assumed from the pathogen ID itself (i.e. intrinsic resistance). In S27 and S29, we detect *bla*$_{Z}$ which explains penicillin resistance in both *S. aureus,* though the vast majority of *S. aureus* are assumed to be resistant to penicillin anyway [333]. In S27, we also detect *erm(T)* which can cause macrolide resistance (the *S. aureus* was phenotypically resistant to Erythromycin and Clindamycin) [334] but this gene has not been widely reported in *S. aureus*. We also detect *bla*$_{BRO}$ in S18 for *M. catarrhalis* which explains the ampicillin resistance in this organism, however, most *M. catarrhalis* strains produce β-lactamase, so this would also be assumed [335]. This shows that there is value to detecting pathogen ID itself, as inherent resistance and local epidemiology can guide treatment based on the organism.

The other two highly resistant organisms are the *P. aeruginosa* in S2 and *E. cloacae* complex in S12 which are both likely to be caused by mutations (e.g AmpC derepression in *E. cloacae complex*). Currently our analysis pipeline does not detect mutational resistance and it may be highly challenging to do so with metagenomic data. The future in complex resistance detection may be to use genomic and transcriptomic data with machine learning to predict resistance as Khaledi et al. 2020 demonstrated [118] or to

predict resistance based on rapid lineage calling as described by Brinda et al. (2020) [120]. Since sequence variations within genes is also an important factor, higher coverage of genomes may be needed to make more accurate predictions.

4.1.6 Analysis pipeline

The choice of taxonomic classification tool used in the CLIMATE pipeline was made based on the literature. In Charalampous et al.,(2019) [97], bioinformatic analysis was performed using the EPI2ME Antimicrobial Resistance pipeline (ONT, v.2.59.1896509), which uses the Centrifuge tool for taxonomic classification. However, this pipeline is unreliable and non-customisable. We made our own pipeline that takes unprocessed fastq reads and outputs the taxonomic profile and resistance genes. We chose Kraken 2 as it is one of the fastest classifiers with the minimum amount of memory required. Compared to Centrifuge, Kraken 2 has been shown to perform better for classification precision and recall, accuracy of abundance profile estimation, and false positive classifications [185]. While Centrifuge can give multiple assignments per read, Kraken 2 gives each read one taxonomic assignment. Bracken is then used for more accurate abundance estimation, using a probabilistic approach rather than just using proportions of reads. Specifically, it uses a Bayesian algorithm to integrate reads Kraken 2 classified at higher taxonomic levels into the desired taxonomic level (genus or species). Most classifier tools including Kraken 2 perform well when taxa are genetically distinct but are poor at distinguishing below the species level, so choice of tools are limited in this sense.

A major issue with most classifiers is that they report low abundance false positives. This is particularly an issue when there is host DNA reads present, as these human reads can be misclassified as microbial due to missing reference sequence [185]. In our pipeline, aside from depleting human DNA, human reads are filtered prior to being classified, as other groups also do [106,107]. This, combined with abundance thresholds (which Is also

common practice), meant that we did not observe any false positives (additional detections were all confirmed by qPCR).

We applied an elevated abundance threshold for pathobionts that are typically found in the upper respiratory tract, raising the threshold from 1% of classified reads to 5% of classified reads. This helps differentiate colonisation from infection, reporting *S. pneumoniae* only in samples where dominates the microbial community, for example in Sample S17, where *S. pneumoniae* was 62.3% of the reads, but removes low level bioinformatic and upper respiratory tract contaminants. However, choosing a set positivity threshold is not ideal, as relative abundance is likely to overlap between organisms that are commensal versus pathogenic (or about to become pathogenic). This challenge of distinguishing colonisation from disease has long been an issue in routine microbiology, with no clear answer on the best approach to solve the problem [327]. The rich data provided by CMg may allow machine learning techniques to be applied to detect potential patterns and associations [336]. Meta-transcriptomic data can in particular be helpful here, as it has been shown that *S. pneumoniae* infection and colonisation leads to different host and bacterial gene expression profiles [337].

Other bespoke metagenomic pipelines also exist such as SURPI+, which has graphical interface tools to make the results easier to interpret [106]. SURPI+ also has a pathogen detection threshold which is based on the ratio of sample reads per million (RPM) to no template control RPM. This ratio has to be equal to or above 10 for a pathogen to be considered as ‘detected’. While we included negative controls in all runs, we did not observe any pathogen reads in them and therefore did not need to include them in the analysis. If pathogen reads were detected in the negative control, detection of that pathogen in a sample would be invalid.

4.1.7 CMg versus multiplex PCR panels

Alternative diagnostic solutions already exist to fix the problem of slow turnaround time of culture, therefore CMg must offer benefits over these methods to be adopted. One of the main advances in rapid diagnostics has been sample-in answer-out multiplex PCR machines. These machines utilise single-use cartridges to process clinical samples directly by extracting DNA/RNA and performing multiplex PCR and detection all in one machine. Examples include the ePlex System (Roche), Unyvero (Curetis) and Biofire FilmArray (Biomérieux). ePlex offers a respiratory panel that is aimed at URTI (taking nasopharyngeal swabs and targeting mainly viruses), but do not have an LRTI/pneumonia panel, whereas Unyvero and FilmArray have panels for pneumonia. One of the main benefits of these machines is that they can be very fast. The Unyvero device takes 4-5 hours from sample to result [338] and the FilmArray takes 1 hour [55]. This puts the rapid CMg workflow, which takes 3.5 hours in a similar ballpark to these devices, faster than the Unyvero but slower than the FilmArray. All of these methods are a significant improvement on routine microbiology which takes at least 1 day, and more often longer (median time to result for culture in the INHALE multicentre HAP/VAP trial was 72 hours [43]). The market leading multiplex panel for pneumonia is the BioFire Filmarray Pneumonia Plus Panel which has some advantages and disadvantages compared to our rapid CMg method (Table 4.2).

Table 4.2 – Comparison of BioFire FilmArray to the rapid CMg pipeline. Sensitivity and specificity for BioFire are those reported by the manufacturer and for the rapid CMg from our internal evaluation of 37 samples with qPCR confirmation. The cost of FilmArray Pneumonia Panel is estimated from a different panel from the same manufacturer.

Biofire FilmArray Pneumonia Panel		**Rapid CMg pipeline**
1 hour	**Total time to result**	3.5 hours
2 minutes	**Hands-on-time**	70 minutes
£177	**Cost-per-test**	£108
96.3%	**Sensitivity**	96.6%
97.2%	**Specificity**	100%
33	**Number of targets**	Not limited
Low	**Complexity of procedure**	High
No	**Genomic epidemiology**	Yes
No	**Comprehensive AMR detection**	Yes
No	**Comprehensive pathogen detection**	Yes

Multiplex Panel PCRs are also very sensitive with good LODs depending on the assay [339]. UKHSA recommends a sputum processing method that would only lead to visible growth at 10^6 CFU/mL which means that Multiplex Panel PCR and CMg methods can both easily match this. We have previously shown that with CMg, we can achieve LODs of 10^3-10^4 CFU/mL depending on the level of background [97]. However, in some cases (depending on varying diagnostic laboratory procedures), lower CFU/mL may be reported. Additionally, different culture procedures are undertaken depending on clinical context, for example samples from immunocompromised patients are cultured undiluted, which can lead to a significantly lower numbers of organisms being detected by culture. Multiplex PCR methods can potentially achieve these LODs, however, BioFire has a deliberate lower cut-off at 5 x 10^3 CFU/mL and will not report detections below this level.

While the speed and sensitivity of multiplex PCR tests may be difficult to beat, a major advantage of CMg over these tests is the comprehensiveness and capacity to detect more. The BioFire panel has 33 targets, of which 7 are resistance genes. Whereas CMg can in theory detect any resistance gene from a chosen or custom database. Virulence

factors (which contribute to pathogenecity) can also be detected, as there is no limit to the number of genes included in the analysis [340]. Additionally, the sequencing data can be used to perform genomic surveillance and for example, identify outbreaks in hospitals in near real-time. This was demonstrated by Charalampous et al. 2021, where a *K. pneumoniae* outbreak involving 4 patients was quickly identified using MLST from the sequence data [115]. Ideally in the future, if CMg becomes routinely implemented, public health agencies will be able to monitor the evolution of resistance in real-time on a regional or national basis and identify outbreaks early. This could be used as an early warning system to change antibiotic prescribing guidelines and prevent further resistance development and transmission. CMg, unlike targeted PCR, also has the capacity to detect new and unusual pathogens. The SARS-CoV-2 genome was first sequenced using metagenomics [217]. If CMg was routinely used, new and unusual sequence signatures could be flagged and acted upon (with an appropriate analysis pipeline in place).

4.1.8 Viral metagenomics

Our method was designed for the detection of bacteria and is also capable of detecting fungi, however, this is insufficient for some respiratory diseases, such as CAP. Viruses are a common cause of CAP, especially in children [341]. While bacteria are the principal cause of HAP/VAP, viruses such as influenza, respiratory syncytial virus and human metapneumovirus can all cause hospital acquired respiratory infections [22]. More recently, SARS-CoV-2 is also shown to cause pneumonia [342]. Our current CMg method is not capable of detecting most viruses. This is mainly because many viral respiratory pathogens contain RNA (such as all the ones listed earlier) and our protocol does not have an RT step. Additionally, our host depletion is likely to lose a significant amount of virus, as it involves pelleting bacteria/fungi and discarding the supernatant. We have previously attempted this method without centrifuging and washing the bacterial pellet, however this led to significantly less efficient human depletion. This means that a viral arm

most likely needs to be introduced as a separate but parallel workflow, combining the cDNA and DNA from viral and bacterial arms together before sequencing.

Viral CMg methods tend to focus on RNA viruses, as these make up most human pathogens [95,111,343], but some split the processing in two arms so that RNA and DNA can be detected [106]. This is because the sample preparation process for RNA viruses usually includes the removal of DNA (human, and bacterial) with a DNase treatment post-extraction. If DNA viruses were included in the same reaction, they would be degraded, therefore, they are processed separately. This would not be ideal in our case, as it would mean the addition of a third arm to the protocol. And focusing on just RNA viruses would also not be optimal, as there are DNA viruses such as adenovirus that can cause human disease, (however, the process of RNA metagenomics may help capture mRNAs of actively replicating DNA viruses [344]).

The principle behind the method we devised is that a sample would be fractioned into a bacterial/fungal layer (the pellet) and the viral layer (the supernatant), which would capture both RNA and DNA viruses without having to treat them separately – with no DNase step after extraction. While the introduction of a reverse transcriptase and 2nd strand synthesis step was shown to successfully convert RNA into sequencable cDNA in principle, human reads dominated the sequencing. Attempts to treat the supernatant with a nuclease pre-extraction while viruses were still intact failed and did not give sufficient reduction of human DNA. This could be because the small extracellular vesicles containing DNA/RNA remain in the supernatant; presence of so called exosomes have been characterised In biological fluids such as blood and respiratory samples [345,346]. These structures could protect human DNA and RNA from digestion. However, we have demonstrated that using saponin should not be used to release this DNA because it can disrupt virus envelopes and lead to loss of virus – this is expected as viral envelopes taken directly from host membranes during maturation of the virus, having the same lipids as the host cell [347]. Other host DNA depletion methods need to be investigated that do not lead to loss of virus. One potential method may be to use beadbeating to lyse ‘soft’ host tissue, using

specific beads that leave viruses intact – this was used by Oechslin. et al. (2018), and they demonstrated viruses were not lost, however, the depletion not perform as well in clinical samples [146]. Truly agnostic viral metagenomics is challenging, so perhaps the best option may be a semi-targeted approach where some viruses are enriched with spiked primers, such as the method described in Deng et al. 2020 [130]. This increases the sensitivity for select viruses, while still allowing non-targeted viruses to still be detected (albeit less sensitively).

These challenges in viral metagenomic sequencing mean that targeted approaches are still required, particularly when reliable sequencing of whole genomes is required from clinical samples (for genomic epidemiology purposes).

4.2 Development of a high-throughput sequencing method for SARS-CoV-2

4.2.1 Genomic epidemiology of SARS-CoV-2

The COVID-19 pandemic has led to the rapid expansion of viral genomic sequencing globally and has changed how we perform epidemiology. While the value of viral genomic epidemiology was becoming increasingly clear with outbreaks such as the 2013-2016 Western African Ebola epidemic and 2015-2016 Zika virus outbreak [219], there was initially scepticism that widescale sequencing of SARS-CoV-2 would be useful, and the creation of COG-UK was met with doubt [348]. This is because SARS-CoV-2 has a relatively modest mutation rate (2 substitutions per month), and the usefulness of genomic epidemiology depends on the mutation and transmission rates of the pathogen, since if enough genetic variation is not generated, it is not possible to determine outbreaks [348]. However, throughout the pandemic, it became clear just how useful genomic epidemiology of SARS-CoV-2 was despite the relatively slow mutation rate. Genomic epidemiology of SARS-CoV-2 has been used to:

- Investigate transmission dynamics in hard hit regions [349], and in close communities such as universities [350] to determine how the virus spreads and how to limit its spread
- Identify importation events of lineages into countries [255]
- Assess the impact of quarantine measures on the importation and transmission of the virus [351]
- Characterise new lineages [254] or specific mutations [253] to determine their transmissibility
- Investigate the effectiveness of vaccines in preventing transmission [352]
- Investigate suspected outbreaks [294,353] in work and care settings

The relatively slower mutation rate does mean that there is a limit, as exact reconstruction of transmission chains between close contacts is not always possible [354], at least from consensus genomes (it may be possible using within-host diversity). But the overall impact of SARS-CoV-2 genomic epidemiology has been very clear.

At time of writing (July 2022), GISAID has over 12,500,000 SARS-CoV-2 genomes in the database submitted from over 200 countries [355]. 5,300,000 of these alone are of the Omicron variant which was only reported in November 2021 [356]. One of the reasons for the significant expansion of genomic epidemiology has been due to advances in sequencing technologies and the development of rapid and portable technologies such as the MinION and Flongle (ONT). We can now produce data in real-time, bringing sequencing to the outbreaks rather than sending samples to reference laboratories. Additionally, due to the low cost and small footprint of devices such as the MinION, as well as the lack of requirement for maintenance contracts, sequencing has become significantly more accessible and is deployed much more easily in developing countries. Technologies such as Illumina, however, are still important, as these devices can provide the very high throughput capacity required during outbreak peaks and in central laboratories. This is why we sought to develop a flexible library preparation method for

sequencing SARS-CoV-2 that was platform agnostic and could be used depending on throughput requirement.

4.2.2 SARS-CoV-2 sequencing and CoronaHiT

In early 2020, the ARTIC tiling PCR method was developed by John Quick (University of Birmingham) for sequencing of SARS-CoV-2 genomes on a MinION and was widely adopted across the world. At the beginning of the pandemic, this method was very low throughput and inflexible. It used ligation-based addition of barcodes and adapter which meant that it was limited by ONT's native barcode kits, allowing a maximum of 24 samples to be sequenced at once. Additionally, in our experience, sequencing anything less than ~10 samples was unreliable, as there was a risk of generating insufficient library for a good sequencing run. This meant that there was a small range of 10 to 24 samples that could be processed per flowcell.

To improve flexibility, in response to constantly changing requirements during the pandemic, we developed CoronaHiT, for high throughput and cost-effective sequencing of SARS-CoV-2. Over the course of the pandemic, optimisations of other methods meant that throughput increased for other methods too. However, our method still has some important benefits. Firstly, it provides more even coverage between samples, resulting in a more samples passing QC. The ARTIC method fails more often with lower viral load samples because coverage is less even, and some samples drop below the threshold QC. The reduced variability between samples for CoronaHiT could be related to the transposase step, which limits the quantity of DNA that is tagged (and therefore no sample dominates) or due to the 14 additional cycles of PCR during barcoding, which likely improves the LOD of the method.

Secondly, CoronaHiT is designed to be adaptable for nanopore or Illumina sequencing, whereby it is possible to change between nanopore and Illumina barcodes depending on available technology and throughput need. With the use of asymmetric barcode primers

for Illumina, it is possible to sequence sample at very high throughputs. At QIB, we have sequenced over a thousand SARS-CoV-2 genomes on a single Illumina NextSeq High Output run (data not shown). The CoronaHiT-Illumina library preparation method is also cheaper and more streamlined than standard Illumina library preparations. This is because there is no sample washing or quantification before pooling, the absence of stop solution, and no bead clean-up after the tagmentation and barcoding of PCR products.

Tiling PCR approaches are prone to significant genome coverage variation due to variable primer efficiency in multiplex reactions. Some amplicons can have much higher coverage than adjacent amplicons. In the case of the 400 bp amplicon tiling scheme (ARTIC), an average coverage of 1000X is required across the genome to obtain at least 20X coverage of regions where primers are less efficient. It is possible to achieve an average of >1000X SARS-CoV-2 genome coverage in approximately 20 minutes per sample with CoronaHiT on a MinION. With a full set of 95 samples, this takes 30 hours. While the CoronaHiT-ONT runs described here are very consistent, sequencing yield depends on flowcell quality. At least 100 Mb of sequencing data per sample is required for >1000X coverage/sample (average across flowcell) using CoronaHiT-ONT.

A limitation of CoronaHiT-ONT is that while it is possible to sequence up to 95 samples on a single MinION flowcell (plus a negative control), this is highly dependent on the quality of the library and the flowcell itself. The quality of the flowcell is out of the user's control and is due to the biological nature of the pores. MinION flowcells have a minimum warranty of 800 active pores, however, higher pore numbers are required to reliably sequence 95 samples. Pore number directly correlates with the yield, as each pore has a finite life. However, this limitation also applies to ARTIC sequencing which also uses nanopore.

We showed that all methods are unreliable when sequencing higher RT-qPCR C_Ts (above 32, approximately 100 viral genome copies); however, CoronaHiT produces fewer Ns in these samples compared to ARTIC LoCost, likely due to the more even coverage between samples as mentioned earlier. While more samples pass both QC measures with

CoronaHiT compared to ARTIC LoCost, primer dropout regions can be more pronounced. The cause of this is unknown, however, since the same ARTIC PCR products were used for all methods, it means that it occurs in the library preparation post ARTIC PCR, potentially relating to the shorter size of CoronaHiT reads (caused by tagmentation of PCR products). Reducing ARTIC PCR annealing temperature from 65 °C to 63 °C may help improve coverage in these difficult regions according to report [269]. Nevertheless, data produced from CoronaHiT is sufficient to provide accurate consensus genomes that result in the same lineages being called as ARTIC LoCost. Therefore, we have demonstrated that it is possible to multiplex 95 samples using CoronaHiT on a single flowcell and still achieve the correct result. If the ARTIC PCR step is optimised to balance the amplicons, less overall coverage may be required, and more samples can be multiplexed using all methods.

The cost for CoronaHiT-ONT was £8.46 per sample when sequencing 95 samples on a MinION flowcell, slightly cheaper but similar to ARTIC LoCost sequencing which costs £9.75 per sample [266]. It is possible to achieve even cheaper per samples costs with Illumina. For example, if 384 samples are sequenced on an Illumina NextSeq Mid output run with CoronaHiT, the per sample cost is £6.22. Significant cost savings are made the more samples are sequenced using CoronaHiT-Illumina.

Since the development and publication of the method, there have been other advances to SARS-CoV-2 genome sequencing. There are now alternative tiling PCR methods such as the Midnight method developed by Freed et al. [270]. These methods use longer 1200 bp amplicons which are more suitable for use with transposase library preparation methods like the Rapid Barcoding Kit (RBK110.96, ONT). The benefit of this is that with fewer amplicons, you get more even coverage and fewer primer pairs to optimise for even coverage. This method was originally limited by 12 rapid barcodes, however there are now 96 Rapid Barcodes. ONT has adopted the “Midnight” method of sequencing SARS-CoV-2. This method is faster than both the ARTIC method and CoronaHiT as there are no lengthy ligation and barcode PCR steps (barcode addition takes 5 minutes using rapid

chemistry). A downside of longer amplicons is that they are less efficient at amplifying low quantity and degraded RNA. Using shorter amplicons schemes like ARTIC may be more beneficial in high C_T samples and in applications like wastewater SARS-CoV-2 sequencing [357].

As far as we are aware, the CoronaHiT method, particularly when used with high output Illumina runs, is still one of the cheapest methods for sequencing SARS-CoV-2 genomes. It has been used to sequence over 90,000 samples at the QIB, which included local samples [294], national surveillance samples (for the REACT-1 study), and international samples from Lebanon, Pakistan and Zimbabwe [358]. It has also been adopted by other laboratories in the COG-UK consortium, having a big impact on SARS-CoV-2 epidemiological research.

4.3 Conclusions

The field of CMg is advancing rapidly. CMg methods have become very accurate for detecting and characterising pathogens with rapid same day results. In this work, we demonstrate a cost-effective CMg workflow that has a 3.5 hour turnaround time and a rapid simple to use Galaxy bioinformatic pipeline. Turnaround time is significantly faster than routine microbiology and in the same ballpark as targeted PCR-based approaches, which will allow clinicians to modify antimicrobial therapy before a second dose of empiric therapy is administered (within 8 hours). In fact, our collaborators at St Thomas' Hospital in London (Prof Jonathan Edgeworth) have implemented a version of our CMg method and find that clinicians are willing to wait for the results before prescribing antimicrobials, avoiding empiric therapy altogether. At ~£108 per sample, if sequencing one sample, or ~£70 if sequencing two samples, on a Flongle, the cost of metagenomic sequencing using this method is cheaper than commercial multiplex PCR panel tests, while providing more comprehensive results (with the ability to detect a wider range of organisms and resistance genes and use the data for genomic epidemiology and infection control).

However, for widespread adoption, processing will need to be automated and the results will need to be presented in easy-to-interpret clinician facing reports. Resistance detection is still a major challenge, with increases in yield and better analysis needed to improve AMR prediction. Use of transcriptomics and machine learning can help in this area. Where CMg isn't suitable, targeted and WGS methods can be used as alternatives. We demonstrated the significant value of a cheap, high-throughput library preparation method that we developed for sequencing SARS-CoV-2 genomes at the height of the pandemic which was used to sequence hundreds of thousands of genomes for surveillance at QIB and in other institutes. The rise of rapid, simple, and cheap sequencing for clinical infectious diseases applications will be used for diagnostic and public health applications in the future to improve patient outcomes and antimicrobial stewardship, and to control outbreaks.

4.4 Future work

- Further simplification and automation of the CMg workflow with the ultimate aim of a sample-in answer-out solution which can be used near patient with minimal hands on time
- Inclusion of an appropriate internal/calibration control to be able to detect process failures and to provide quantitative or semi-quantitative results
- Integration of viral metagenomics for wider application of the method
- Investigate and optimise the host depletion to reduce the loss of some bacteria, for more reliable results.
- Better analysis for resistance/susceptibility prediction, particularly the incorporation of chromosomal mutations for the ability to predict resistance in organisms like *P. aeruginosa*
- Develop appropriate methods to better associate resistance genes with their host and differentiate commensals from pathogens

Appendices

Appendix 1

Trotter AJ, Aydin A, Strinden MJ, O'Grady J. Recent and emerging technologies for the rapid diagnosis of infection and antimicrobial resistance. Curr Opin Microbiol. 2019 Oct 1;51:39–45

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Recent and emerging technologies for the rapid diagnosis of infection and antimicrobial resistance

Alexander J Trotter[1,2], Alp Aydin[1,2], Michael J Strinden[1,2] and Justin O'Grady[1,2]

The rise in antimicrobial resistance (AMR) is predicted to cause 10 million deaths per year by 2050 unless steps are taken to prevent this looming crisis. Microbiological culture is the gold standard for the diagnosis of bacterial/fungal pathogens and antimicrobial resistance and takes 48 hours or longer. Hence, antibiotic prescriptions are rarely based on a definitive diagnosis and patients often receive inappropriate treatment. Rapid diagnostic tools are urgently required to guide appropriate antimicrobial therapy, thereby improving patient outcomes and slowing AMR development. We discuss new technologies for rapid infection diagnosis including: sample-in-answer-out PCR-based tests, BioFire FilmArray and Curetis Unyvero; rapid susceptibility tests, Accelerate Pheno and microfluidic tests; and sequencing-based approaches, focusing on targeted and clinical metagenomic nanopore sequencing.

Addresses
[1] University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR4 7TJ, UK
[2] Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, NR4 7UQ, UK

Corresponding author: O'Grady, Justin (justin.ogrady@quadram.ac.uk)

Current Opinion in Microbiology 2019, **51**:39–45

This review comes from a themed issue on **Antimicrobials**

Edited by **Mattew I Hutchings, Andrew W Truman and Barrie Wilkinson**

For a complete overview see the Issue and the Editorial

Available online 9th May 2019

https://doi.org/10.1016/j.mib.2019.03.001

Introduction

More than 700 000 people die per year globally due to antimicrobial resistance (AMR) according to an estimate from the UK government-commissioned review on AMR (O'Neill report) [1]. At the current rate, by 2050, AMR is predicted to cause 10 million deaths annually and cost the world economy $100 trillion in total. It is widely recognised that rapid diagnostics are crucial in the fight against AMR, to improve the management of life threatening infections such as sepsis and pneumonia and to enable earlier and more precise targeting of pathogens with appropriate antibiotics (i.e. improved antibiotic stewardship) [2]. The final O'Neill report states that by 2020 all antibiotic prescriptions should be supported by a rapid diagnostic test where available [1].

Current standard methods for diagnosing bacterial infection are based on microbiological culture and have long turn-around times, offer poor clinical sensitivity and are not fit-for-purpose for acute serious infection such as sepsis, pneumonia and meningitis. Acute infections force clinicians into early broad-spectrum treatment, before culture results become available, highlighting the need for rapid diagnostics [3–6]. A paradigm shift in diagnostic microbiology is urgently required, with the ultimate goal of providing pathogen identification and resistance/susceptibility information to clinicians before antibiotics are administered.

In this review, we highlight recent and emerging tests for the rapid diagnosis of pathogens, antimicrobial resistance and antimicrobial susceptibility and their current/future clinical applications. We describe some of the current tests that utilise genotypic methods such as PCR for pathogen identification and antibiotic resistance testing. We also describe technologies and techniques that combine pathogen identification with rapid phenotypic antibiotic susceptibility testing (AST). Finally, we outline key advances in the application of DNA sequencing for the rapid diagnosis of infection and AMR that could be implemented clinically in the near future.

Rapid PCR-based pathogen and antimicrobial resistance detection

Discerning bacterial from viral infections is the simplest level of diagnosis that can be clinically useful to guide antimicrobial therapy, reducing unnecessary antibiotic prescriptions. FebriDx® is a dipstick test measuring c-reactive protein and Myxovirus resistance protein A levels in blood, differentiating bacterial from viral infections using an inflammation biomarker [7]. Polymerase chain reaction (PCR)-based systems such as ID Now and cobas® Liat® have specific tests for specific targets such as influenza A&B [8]. However, an ideal diagnostic test will identify the specific pathogen and provide guidance on appropriate antimicrobial therapy. This is particularly important in clinical syndromes such as urinary tract infections (UTIs), pneumonia and sepsis, which can be caused by many different pathogens (bacteria, fungi or

Appendix 2

Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019 Jul;37(7):783–92

nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-019-0156-5

Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection

Themoula Charalampous[1,8], Gemma L. Kay[1,2,8], Hollian Richardson[1,8], Alp Aydin[2], Rossella Baldan[1,3], Christopher Jeanes[4], Duncan Rae[4], Sara Grundy[4], Daniel J. Turner[5], John Wain[1,2], Richard M. Leggett[6], David M. Livermore[1,7] and Justin O'Grady[1,2*]

The gold standard for clinical diagnosis of bacterial lower respiratory infections (LRIs) is culture, which has poor sensitivity and is too slow to guide early, targeted antimicrobial therapy. Metagenomic sequencing could identify LRI pathogens much faster than culture, but methods are needed to remove the large amount of human DNA present in these samples for this approach to be feasible. We developed a metagenomics method for bacterial LRI diagnosis that features efficient saponin-based host DNA depletion and nanopore sequencing. Our pilot method was tested on 40 samples, then optimized and tested on a further 41 samples. Our optimized method (6 h from sample to result) was 96.6% sensitive and 41.7% specific for pathogen detection compared with culture and we could accurately detect antibiotic resistance genes. After confirmatory quantitative PCR and pathobiont-specific gene analyses, specificity and sensitivity increased to 100%. Nanopore metagenomics can rapidly and accurately characterize bacterial LRIs and might contribute to a reduction in broad-spectrum antibiotic use.

LRIs caused at least three million deaths worldwide in 2016 (http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death). They can be subdivided into community-acquired pneumonia (CAP), hospital-acquired pneumonia (HAP), bronchitis, bronchiolitis and tracheitis[1]. Morbidity and mortality rates vary dependent on infection site, pathogen and host factors. In the United Kingdom, CAP accounts for approximately 29,000 deaths per annum, and in the United States, HAP causes approximately 36,000 deaths per annum[2,3]. The most common bacterial CAP pathogens are *Streptococcus pneumoniae* and *Haemophilus influenzae*, and the most common HAP pathogens are *Staphylococcus aureus*, Enterobacteriaceae and *Pseudomonas aeruginosa*[4–6]. However, multiple bacterial and viral pathogens can cause LRIs, which makes diagnosis and treatment a challenge. Respiratory tract infections account for 60% of all antibiotics prescribed in general practice in the United Kingdom[1]. Initial treatment for severe LRIs usually involves empirical broad-spectrum antibiotics. Guidelines recommend that such therapy should be refined or stopped after 2 to 3 days, once microbiology results become available[7,8], but this is often not done if the patient is responding well or the laboratory has failed to identify a pathogen. Such extensive 'blind' use of broad-spectrum antibiotics is wasteful and constitutes poor stewardship, given that many patients are infected with susceptible bacteria or a virus. Antimicrobial therapy disrupts resident gut flora and can contribute to the emergence of resistant bacteria and *Clostridium difficile*[9,10].

Rapid and accurate microbiological diagnostics could enable tailored treatments and reduce overuse of broad-spectrum antibiotics. 'Gold standard' culture and susceptibility testing is too slow, with typical turnaround times of 48–72 h and low clinical sensitivity[4,11]. Molecular methods may help overcome the limitations of culture, as highlighted by the UK Government 5-year AMR action plan and the O'Neill report[12–14], by identifying pathogens and their antibiotic resistance profiles in a few hours, enabling early targeted therapy and supporting antibiotic stewardship. Although nucleic acid amplification tests (including PCR) are rapid and highly specific/sensitive, there are limits on multiplexing[15–19], and there is also a constant need to update PCR-based methods to include emerging resistance genes and mutations[16,20,21].

Metagenomic sequencing-based approaches have the potential to overcome the shortcomings of both culture and PCR, by combining speed with comprehensive coverage of all microorganisms present[22,23]. Next-generation sequencing platforms, such as Ion Torrent and Illumina, are widely used for metagenomics sequencing, but they require the sequencing run to be complete before analysis can begin (although LiveKraken, a recently described method, enables analysis of raw Illumina data before the run ends[24]). Nanopore sequencing (Oxford Nanopore Technologies, ONT) has the advantage of rapid library preparation and real-time data acquisition[25,26]. Nanopore sequencing has been used to identify viral and bacterial pathogens from clinical samples using targeted approaches and in proof-of-concept studies using samples with high pathogen loads, for example, urinary tract infection[26–28].

Respiratory specimens present a difficult challenge for metagenomics sequencing due to variable pathogen load, the presence of commensal respiratory tract flora, and the high ratio of host:pathogen nucleic acids present (up to 10^5:1 in sputum). Nanopore sequencing has previously been used for samples from two bacterial pneumonia patients without host cell/DNA depletion, but the vast majority of reads were of human origin, with only one and two reads aligned to the infecting pathogens, *P. aeruginosa* and *S. aureus*, respectively[29]. It seems likely that a metagenomics method would be improved by introducing host DNA depletion. Although

[1]Bob Champion Research and Educational Building, University of East Anglia, Norwich Research Park, Norwich, UK. [2]Quadram Institute Bioscience, Norwich Research Park, Norwich, UK. [3]CIDR, King's College London, St Thomas' Hospital, London, UK. [4]Microbiology Department, Norwich and Norfolk University Hospital, Norwich, UK. [5]Oxford Nanopore Technologies, Gosling Building, Oxford Science Park, Oxford, UK. [6]Earlham Institute, Norwich Research Park, Norwich, UK. [7]AMRHAI, Public Health England, London, UK. [8]These authors contributed equally: Themoula Charalampous, Gemma L. Kay, Hollian Richardson. *e-mail: justin.ogrady@quadram.ac.uk

Appendix 3

Patent application “Method for digesting nucleic acid in a sample” (PCT/GB2020/052986)
Inventors: Justin O’Grady, Gemma Kay, Themoula Charalampous, Alp Aydin, Riccardo Scotti

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 June 2021 (03.06.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/105659 A1

(51) International Patent Classification:
C12N 15/10 (2006.01)

(21) International Application Number:
PCT/GB2020/052986

(22) International Filing Date:
24 November 2020 (24.11.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1917101.6 25 November 2019 (25.11.2019) GB

(71) Applicant: **UEA ENTERPRISES LIMITED** [GB/GB]; The Registry, University of East Anglia, Norwich Research Park, Norwich Norfolk NR4 7TJ (GB).

(72) Inventors: **O'GRADY, Justin Joseph**; c/o UEA Enterprises Limited, The Registry, University of East Anglia, Norwich Research Park, Norwich Norfolk NR4 7TJ (GB). **KAY, Gemma Louise**; c/o UEA Enterprises Limited, University of East Anglia, Norwich Research Park, Norwich Norfolk NR4 7TJ (GB). **CHARALAMPOUS, Themoula**; c/o UEA Enterprises Limited, University of East Anglia, Norwich Research Park, Norwich Norfolk NR4 7TJ (GB). **AYDIN, Alp**; c/o UEA Enterprises Limited, University of East Anglia, Norwich Research Park, Norwich Norfolk NR4 7TJ (GB). **SCOTTI, Riccardo**; c/o UEA Enterprises Limited, University of East Anglia, Norwich Research Park, Norwich Norfolk NR4 7TJ (GB).

(74) Agent: **NOVAGRAAF UK**; Centurm, Norwich Research Park, Colney Lane, Norwich Norfolk NR4 7UG (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

WO 2021/105659 A1

(54) **Title:** METHOD FOR DIGESTING NUCLEIC ACID IN A SAMPLE

(57) **Abstract:** Provided are methods, compositions and kits for depleting host nucleic acid in a biological sample, said sample having been previously obtained from an animal host.

Appendix 4

Baker DJ, Aydin A, Le-Viet T, Kay GL, Rudder S, de Oliveira Martins L, et al. CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. Genome Med. 2021 Feb 9;13(1):21

Baker *et al. Genome Medicine* (2021) 13:21
https://doi.org/10.1186/s13073-021-00839-5

Genome Medicine

METHOD Open Access

CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes

Dave J. Baker[1†], Alp Aydin[1†], Thanh Le-Viet[1], Gemma L. Kay[1], Steven Rudder[1], Leonardo de Oliveira Martins[1], Ana P. Tedim[1,2], Anastasia Kolyva[1,3], Maria Diaz[1], Nabil-Fareed Alikhan[1], Lizzie Meadows[1], Andrew Bell[1], Ana Victoria Gutierrez[1], Alexander J. Trotter[1,4], Nicholas M. Thomson[1], Rachel Gilroy[1], Luke Griffith[4], Evelien M. Adriaenssens[1], Rachael Stanley[3], Ian G. Charles[1,4], Ngozi Elumogo[1,3], John Wain[1,4], Reenesh Prakash[3], Emma Meader[3], Alison E. Mather[1,4], Mark A. Webber[1,4], Samir Dervisevic[3], Andrew J. Page[1*†] and Justin O'Grady[1,4*†]

Abstract

We present CoronaHiT, a platform and throughput flexible method for sequencing SARS-CoV-2 genomes (≤ 96 on MinION or > 96 on Illumina NextSeq) depending on changing requirements experienced during the pandemic. CoronaHiT uses transposase-based library preparation of ARTIC PCR products. Method performance was demonstrated by sequencing 2 plates containing 95 and 59 SARS-CoV-2 genomes on nanopore and Illumina platforms and comparing to the ARTIC LoCost nanopore method. Of the 154 samples sequenced using all 3 methods, ≥ 90% genome coverage was obtained for 64.3% using ARTIC LoCost, 71.4% using CoronaHiT-ONT and 76.6% using CoronaHiT-Illumina, with almost identical clustering on a maximum likelihood tree. This protocol will aid the rapid expansion of SARS-CoV-2 genome sequencing globally.

Keywords: SARS-CoV-2, Nanopore, Sequencing, NGS, Genome, Genetic, Multiplexing, ARTIC

Background

The COVID-19 pandemic caused by the SARS-CoV-2 virus began late 2019 in Wuhan, China, and has now spread to virtually every country in the world, with tens of millions of confirmed cases and millions of deaths [1]. Key to the control of the pandemic is understanding the epidemiological spread of the virus at global, national and local scales [2]. Whole-genome sequencing of SARS-CoV-2 is likely to be the fastest and most accurate method to study virus epidemiology as it spreads. We are sequencing SARS-CoV-2 as part of the COVID-19 Genomics UK (COG-UK) consortium, a network of academic and public health institutions across the UK brought together to collect, sequence and analyse whole genomes to fully understand the transmission and evolution of this virus [3]. The SARS-CoV-2 genome was first sequenced in China using a metatranscriptomic approach [4]. This facilitated the design of tiling PCR approaches for genome sequencing, the most widely used of which is the ARTIC Network [5] protocol. Consensus genome sequences are typically made publicly available on the Global Initiative on Sharing Avian Influenza Data (GISAID) database [6]. This has enabled real-time public health surveillance of the spread and evolution of the pandemic through interactive tools such as NextStrain [7]. The ARTIC network protocol was designed for nanopore technology (Oxford Nanopore Technologies), enabling rapid genome sequencing for outbreak response. The method was originally capable of testing

* Correspondence: andrew.page@quadram.ac.uk; justin.ogrady@quadram.ac.uk
[†]Dave J. Baker, Alp Aydin, Andrew J. Page and Justin O'Grady contributed equally to this work.
[1]Quadram Institute Bioscience, Norwich Research Park, Norwich NR4 7UQ, UK
Full list of author information is available at the end of the article

Appendix 5

Page AJ, Mather AE, Le-Viet T, Meader EJ, Alikhan NF, Kay GL, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. Microb Genomics. 2021 Jun;7(6)

MICROBIAL GENOMICS

RESEARCH ARTICLE

Page *et al.*, *Microbial Genomics* 2021;7:000589

DOI 10.1099/mgen.0.000589

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management

Andrew J. Page[1,*,†], Alison E. Mather[1,2†], Thanh Le-Viet[1], Emma J. Meader[3], Nabil-Fareed Alikhan[1], Gemma L. Kay[1], Leonardo de Oliveira Martins[1], Alp Aydin[1], David J. Baker[1], Alexander J. Trotter[1,2], Steven Rudder[1], Ana P. Tedim[1,4], Anastasia Kolyva[1,3], Rachael Stanley[3], Muhammad Yasir[1], Maria Diaz[1], Will Potter[3], Claire Stuart[3], Lizzie Meadows[1], Andrew Bell[1], Ana Victoria Gutierrez[1], Nicholas M. Thomson[1], Evelien M. Adriaenssens[1], Tracey Swingler[2], Rachel A. J. Gilroy[1], Luke Griffith[2], Dheeraj K. Sethi[3], Dinesh Aggarwal[5,6,7,8], Colin S. Brown[5], Rose K. Davidson[2], Robert A. Kingsley[1,2], Luke Bedford[9], Lindsay J. Coupland[3], Ian G. Charles[1,2], Ngozi Elumogo[1,3], John Wain[1,2], Reenesh Prakash[3], Mark A. Webber[1,2], S. J. Louise Smith[10], Meera Chand[5], Samir Dervisevic[3], Justin O'Grady[1,2] and The COVID-19 Genomics UK (COG-UK) Consortium

Abstract

The COVID-19 pandemic has spread rapidly throughout the world. In the UK, the initial peak was in April 2020; in the county of Norfolk (UK) and surrounding areas, which has a stable, low-density population, over 3200 cases were reported between March and August 2020. As part of the activities of the national COVID-19 Genomics Consortium (COG-UK) we undertook whole genome sequencing of the SARS-CoV-2 genomes present in positive clinical samples from the Norfolk region. These samples were collected by four major hospitals, multiple minor hospitals, care facilities and community organizations within Norfolk and surrounding areas. We combined clinical metadata with the sequencing data from regional SARS-CoV-2 genomes to understand the origins, genetic variation, transmission and expansion (spread) of the virus within the region and provide context nationally. Data were fed back into the national effort for pandemic management, whilst simultaneously being used to assist local outbreak analyses. Overall, 1565 positive samples (172 per 100000 population) from 1376 cases were evaluated; for 140 cases between two and six samples were available providing longitudinal data. This represented 42.6% of all positive samples identified by hospital testing in the region and encompassed those with clinical need, and health and care workers and their families. In total, 1035 cases had genome sequences of sufficient quality to provide phylogenetic lineages. These genomes belonged to 26 distinct global lineages, indicating that there were multiple separate introductions into the region. Furthermore, 100 genetically distinct UK lineages were detected demonstrating local evolution, at a rate of ~2 SNPs per month, and multiple co-occurring lineages as the pandemic progressed. Our analysis: identified a discrete sublineage associated with six care facilities; found no evidence of reinfection in longitudinal samples; ruled out a nosocomial outbreak; identified 16 lineages in key workers which were not in patients, indicating infection control measures were effective; and found the D614G spike protein mutation which is linked to increased transmissibility dominates the samples and rapidly confirmed relatedness of cases in an outbreak at a food processing facility. The large-scale genome sequencing of SARS-CoV-2-positive samples has provided valuable additional data for public health epidemiology in the Norfolk region, and will continue to help identify and untangle hidden transmission chains as the pandemic evolves.

DATA SUMMARY

Raw reads are deposited in the European Nucleotide Archive, and individual accession numbers are listed in Table S1 (available in the online version of this article). Consensus sequences are deposited with GISAID (Global Initiative on Sharing All Influenza Data), and individual accession numbers are listed in Table S1. The data can be visualized in MicroReact https://beta.microreact.org/project/8HsL3eyA16Wsjtymhvbb Ea-cog-uk-2021-03-18-uk-sars-cov-2. Bioinformatics pipelines used to process the data are available from https://github.com/quadram-institute-bioscience/ncov2019-artic-nf/tree/qib which were adapted from https://github.com/connor-lab/ncov2019-artic-nf. Additional metadata, trees and alignments are available from COG-UK https://www.cogconsortium.

Appendix 6

Ravi A, Troncoso-Rey P, Ahn-Jarvis J, Corbin KR, Harris S, Harris H, et al. Linking carbohydrate structure with function in the human gut microbiome using hybrid metagenome assemblies [Internet]. bioRxiv; 2021 [cited 2022 Aug 16]. p. 2021.05.11.441322. Available from: https://www.biorxiv.org/content/10.1101/2021.05.11.441322v2

New Results

Linking carbohydrate structure with function in the human gut microbiome using hybrid metagenome assemblies

Anuradha Ravi, Perla Troncoso-Rey, Jennifer Ahn-Jarvis, Kendall R. Corbin, Suzanne Harris, Hannah Harris, Alp Aydin, Gemma L. Kay, Thanh Le Viet, Rachel Gilroy, Mark J. Pallen, Andrew J. Page, Justin O'Grady, Frederick J. Warren

doi: https://doi.org/10.1101/2021.05.11.441322

Abstract

Background Complex carbohydrates that escape digestion in the small intestine, are broken down in the large intestine by enzymes encoded by the gut microbiome. This is a symbiotic relationship between particular microbes and the host, resulting in metabolic products that influence host gut health and are exploited by other microbes. However, the role of carbohydrate structure in directing microbiota community composition and the succession of carbohydrate-degrading microbes is not fully understood. Here we take the approach of combining data from long and short read sequencing allowing recovery of large numbers of high quality genomes, from which we can predict carbohydrate degrading functions, and impact of carbohydrate on microbial communities.

Appendix 7

Li T, Garcia-Gutierrez E, Yara DA, Scadden J, Davies J, Hutchins C, et al. An optimised protocol for detection of SARS-CoV-2 in stool. BMC Microbiol. 2021 Sep 6;21:242

Li *et al. BMC Microbiology* (2021) 21:242
https://doi.org/10.1186/s12866-021-02297-w

BMC Microbiology

METHODOLOGY ARTICLE Open Access

An optimised protocol for detection of SARS-CoV-2 in stool

Tianqi Li[1,2†], Enriqueta Garcia-Gutierrez[1†], Daniel A. Yara[1], Jacob Scadden[1], Jade Davies[1], Chloe Hutchins[1], Alp Aydin[3], Justin O'Grady[3], Arjan Narbad[1], Stefano Romano[1*] and Lizbeth Sayavedra[1*]

Abstract

Background: SARS-CoV-2 has been detected in stool samples of COVID-19 patients, with potential implications for faecal-oral transmission. Compared to nasopharyngeal swab samples, the complexity of the stool matrix poses a challenge in the detection of the virus that has not yet been solved. However, robust and reliable methods are needed to estimate the prevalence and persistence of SARS-CoV-2 in the gut and to ensure the safety of microbiome-based procedures such as faecal microbiota transplant (FMT). The aim of this study was to establish a sensitive and reliable method for detecting SARS-CoV-2 in stool samples.

Results: Stool samples from individuals free of SARS-CoV-2 were homogenised in saline buffer and spiked with a known titre of inactivated virus ranging from 50 to 750 viral particles per 100 mg stool. Viral particles were concentrated by ultrafiltration, RNA was extracted, and SARS-CoV-2 was detected via real-time reverse-transcription polymerase chain reaction (RT-qPCR) using the CDC primers and probes. The RNA extraction procedure we used allowed for the detection of SARS-CoV-2 via RT-qPCR in most of the stool samples tested. We could detect as few as 50 viral particles per 100 mg of stool. However, high variability was observed across samples at low viral titres. The primer set targeting the N1 region provided more reliable and precise results and for this primer set our method had a limit of detection of 1 viral particle per mg of stool.

Conclusions: Here we describe a sensitive method for detecting SARS-CoV-2 in stool samples. This method can be used to establish the persistence of SARS-CoV-2 in stool and ensure the safety of clinical practices such as FMT.

Keywords: FMT, COVID19, RT-qPCR, Stool, Clinical-test

Background

The global pandemic caused by SARS-CoV-2, poses an imminent threat to the global population. From December 2019 until the 22nd of June 2021, the number of confirmed cases stands at 179 million and rising, leading to an unprecedented challenge on health systems internationally. SARS-CoV-2 causes severe acute respiratory syndrome - infecting human cells by binding to the receptor angiotensin converting enzyme 2 (ACE2). ACE2 is an inflammation regulator expressed by epithelial cells located in the lung, liver, and gastrointestinal tract. It has been reported that gastrointestinal symptoms, such as diarrhoea, nausea, and vomiting, may be observed in up to 61 % of cases [1]. These gastrointestinal symptoms may be linked to the severity of the COVID-19 disease based on viral load and the degree of viral replication in the gut [2–5]. SARS-CoV-2 RNA has been detected in patient stool both during infection and after patients have apparently recovered – indicated by a lack of viral detection from nasal swab [6]. Viable SARS-CoV-2 has been isolated from stool samples [6–8], which suggests that there is a potential risk of faecal-oral

* Correspondence: Stefano.Romano@quadram.ac.uk; Lizbeth.Sayavedra@quadram.ac.uk
Tianqi Li and Enriqueta Garcia-Gutierrez are shared first-authorship.
[1]Gut Health and Microbes, Quadram Institute Bioscience, Norwich Research Park, Norwich, UK
Full list of author information is available at the end of the article

Glossary

AIDS	Acquired Immunodeficiency Syndrome
AMR	Antimicrobial Resistance
ARDB	Antibiotic Resistance Genes Database
AST	Antibiotic Susceptibility Testing
BAL	Bronchoalveolar Lavage
BCRE	Bob Champion Research and Education
BLAST	Basic Local Alignment Search Tool
BLT	Bead Linked Transposase
CAP	Community Acquired Pneumonia
CARD	Comprehensive Antibiotic Resistance Database
CDC	Centers for Disease Control and Prevention
CDER	Center for Drug Evaluation and Research
CFE	Centre for Excellence
CFU	Colony Forming Units
CLIMATE	Clinical Metagenomics and Antimicrobial resistance
CLIMB	Cloud Infrastructure for Microbial Bioinformatics
CLSI	Clinical and Laboratory Standard Institute
CMg	Clinical Metagenomics
Co-amoxiclav	Amoxicillin with Clavulanic acid
Co-trimoxazole	Trimethoprim with Sulfamethoxazole
COG-UK	COVID-19 Genomics UK Consortium
CoronaHiT	Corona High Throughput
COVID-19	Coronavirus Disease 2019
CPE	Carbapenemase-producing Enterobacteriaceae
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeat
C_T	Cycle Threshold
DDBJ	DNA Data Bank of Japan
DTT	Dithiothreitol
ENA	European Nucleotide Archive
ERS	European Respiratory Society
EU/EEA	European Union/European Economic Area
EUCAST	European Committee On Antimicrobial Susceptibility
FMT	Faecal microbiota transplantation

GISAID	Global Initiative on Sharing Influenza Data
GUI	Graphical User Interface
GXL	PrimeSTAR GXL Polymerase (Takara Bio)
HAP	Hospital Acquired Pneumonia
HCAP	Healthcare-Associated Pneumonia
HIV	Human Immunodeficiency Virus
HL-SAN	Heat Labile Salt Active Nuclease
HSV	Herpes Simplex Virus
ICU	Intensive Care Unit
ITS	Internal Transcribed Spacer
IUPAC	International Union of Pure and Applied Chemistry
LAMP	Loop-mediated isothermal amplification
LAT	LongAmp Taq Polymerase (New England Biolabs)
LFT	Lateral Flow Tests
LOD	Limit Of Detection
LRTI	Lower Respiratory Tract Infection
MDR	Multi-Drug Resistance
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
MIC	Minimum Inhibitory Concentration
MLST	Multi Locus Sequence Typing
MNase	Micrococcal Nuclease
mRNA	Messenger RNA
MRSA	Methicillin Resistant *Staphylococcus aureus*
MSSPE	Metagenomic Sequencing with Spiked Primer Enrichment
NAAT	Nucleic Acid Amplification Test
NGS	Next Generation Sequencing
NICE	National Institute for Health and Care Excellence
NNUH	Norfolk and Norwich University Hospital
NPA	Nasopharyngeal Aspirates
NPI	Non-Pharmaceutical Interventions
NRF	Normal Respiratory Flora
OECD	Organisation for Economic Co-operation and Development
ONT	Oxford Nanopore Technologies
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction

PLC	Phospholipase C
PMA	Propidium Monoazide
QC	Quality Control
QIB	Quadram Institute Bioscience
R&D	Research and Development
RASE	Resistance-Associated Sequence Elements
REACT	Real-time Assessment of Community Transmission
RECOVERY	Randomised Evaluation of COVID-19 Therapy
RGI	Resistance Gene Identifier
RPB004	Rapid PCR Barcoding kit
RPM	Rotations Per Minute
RPM	Reads Per Million
rRNA	Ribosomal RNA
RSV	Respiratory Syncytial Virus
RT	Reverse Transcriptase
RT-qPCR	Reverse Transcription qPCR
SAGE	Scientific Advisory Group for Emergencies
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SISPA	Sequence-Independent Single-Primer Amplification
SMRT	Single Molecule
SNP	Single Nucleotide Polymorphism
ssRNA	Single-Stranded RNA
SURPI	Sequence-based Ultrarapid Pathogen Identification
TB	Tuberculosis
UCL	University College London
UKHSA	UK Health Security Agency
URTI	Upper Respiratory Tract Infection
UTI	Urinary Tract Infections
UV	Ultraviolet
VOC	Variants of Concern
WGS	Whole Genome Sequencing
WHO	World Health Organisation
WIMP	What's In My Pot
ZOI	Zone of Inhibition

References

1. The top 10 causes of death [Internet]. [cited 2021 May 31]. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

2. Li Y, Nair H. Trends in the global burden of lower respiratory infections: the knowns and the unknowns. Lancet Infect Dis [Internet]. 2022 Aug 11 [cited 2022 Aug 13];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(22)00445-5/fulltext

3. WHO Coronavirus (COVID-19) Dashboard [Internet]. [cited 2021 May 31]. Available from: https://covid19.who.int

4. Conducting Passages | SEER Training [Internet]. [cited 2021 May 31]. Available from: https://training.seer.cancer.gov/anatomy/respiratory/passages/

5. Pneumonia - Symptoms and causes [Internet]. Mayo Clinic. [cited 2021 May 31]. Available from: https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204

6. Bronchitis - Symptoms and causes [Internet]. Mayo Clinic. [cited 2021 May 31]. Available from: https://www.mayoclinic.org/diseases-conditions/bronchitis/symptoms-causes/syc-20355566

7. Thomas M, Bomar PA. Upper Respiratory Tract Infection. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 [cited 2021 May 31]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK532961/

8. Forum of International Respiratory Societies, European Respiratory Society. The global impact of respiratory disease. 2017.

9. Causes of Pneumonia | CDC [Internet]. 2021 [cited 2021 May 31]. Available from: https://www.cdc.gov/pneumonia/causes.html

10. Yap V, Datta D, Metersky ML. Is the Present Definition of Health Care–Associated Pneumonia the Best Way to Define Risk of Infection with Antibiotic-Resistant Pathogens? Infect Dis Clin North Am. 2013 Mar 1;27(1):1–18.

11. Hospital-acquired pneumonia (non COVID-19) - Symptoms, diagnosis and treatment | BMJ Best Practice [Internet]. [cited 2021 May 31]. Available from: https://bestpractice.bmj.com/topics/en-gb/3000092

12. Chalmers J, Campling J, Ellsbury G, Hawkey PM, Madhava H, Slack M. Community-acquired pneumonia in the United Kingdom: a call to action. Pneumonia. 2017 Oct 5;9(1):15.

13. Shoar S, Musher DM. Etiology of community-acquired pneumonia in adults: a systematic review. Pneumonia. 2020 Oct 5;12(1):11.

14. Torres A, Blasi F, Peetermans WE, Viegi G, Welte T. The aetiology and antibiotic management of community-acquired pneumonia in adults in Europe: a literature review. Eur J Clin Microbiol Infect Dis. 2014 Jul 1;33(7):1065–79.

15. Alimi Y, Lim WS, Lansbury L, Leonardi-Bee J, Nguyen-Van-Tam JS. Systematic review of respiratory viral pathogens identified in adults with community-acquired pneumonia in Europe. J Clin Virol. 2017 Oct 1;95:26–35.

16. Viasus D, Vecino-Moreno M, Hoz JMDL, Carratalà J. Antibiotic stewardship in community-acquired pneumonia. Expert Rev Anti Infect Ther. 2017 Apr 3;15(4):351–9.

17. Recommendations | Pneumonia (community-acquired): antimicrobial prescribing | Guidance | NICE [Internet]. NICE; [cited 2022 Aug 14]. Available from: https://www.nice.org.uk/guidance/ng138/chapter/Recommendations

18. Pássaro L, Harbarth S, Landelle C. Prevention of hospital-acquired pneumonia in non-ventilated adult patients: a narrative review. Antimicrob Resist Infect Control. 2016 Nov 14;5(1):43.

19. Micek ST, Chew B, Hampton N, Kollef MH. A Case-Control Study Assessing the Impact of Nonventilated Hospital-Acquired Pneumonia on Patient Outcomes. Chest. 2016 Nov;150(5):1008–14.

20. Wagner A, Turner D, Enne VI, Baldan R, Russell C, Livermore D. Health Economics of Nosocomial Pneumonia in UK Intensive Care Units (ICU): an exploratory study. In 2020 [cited 2021 Jun 20]. Available from: https://ueaeprints.uea.ac.uk/id/eprint/76807/

21. Antimicrobial resistance [Internet]. [cited 2022 Aug 14]. Available from: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

22. Hong HL, Hong SB, Ko GB, Huh JW, Sung H, Do KH, et al. Viral Infection Is Not Uncommon in Adult Patients with Severe Hospital-Acquired Pneumonia. PLOS ONE. 2014 Apr 21;9(4):e95865.

23. WHO publishes list of bacteria for which new antibiotics are urgently needed [Internet]. [cited 2022 Aug 14]. Available from: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

24. Wagner AP, Enne VI, Livermore DM, Craig JV, Turner DA. Review of health economic models exploring and evaluating treatment and management of hospital-acquired pneumonia and ventilator-associated pneumonia. J Hosp Infect. 2020 Dec 1;106(4):745–56.

25. Nhu NTK, Lan NPH, Campbell JI, Parry CM, Thompson C, Tuyen HT, et al. Emergence of carbapenem-resistant Acinetobacter baumannii as the major cause of ventilator-associated pneumonia in intensive care unit patients at an infectious disease hospital in southern Vietnam. J Med Microbiol. 2014 Oct;63(Pt 10):1386–94.

26. Ahl J, Tham J, Walder M, Melander E, Odenholt I. Bacterial aetiology in ventilator-associated pneumonia at a Swedish university hospital. Scand J Infect Dis. 2010 Jul 1;42(6–7):469–74.

27. Charles MP, Easow JM, Joseph NM, Ravishankar M, Kumar S, Sivaraman U. Aetiological agents of ventilator-associated pneumonia and its resistance pattern – a threat for treatment. Australas Med J. 2013 Sep 30;6(9):430–4.

28. Golia S, K.T. S, C.L. V. Microbial Profile of Early and Late Onset Ventilator Associated Pneumonia in The Intensive Care Unit of A Tertiary Care Hospital in Bangalore, India. J Clin Diagn Res JCDR. 2013 Nov;7(11):2462–6.

29. Luyt CE, Bréchot N, Chastre J. What role do viruses play in nosocomial pneumonia? Curr Opin Infect Dis. 2014 Apr;27(2):194–9.

30. Recommendations | Pneumonia (hospital-acquired): antimicrobial prescribing | Guidance | NICE [Internet]. NICE; [cited 2021 Jun 21]. Available from: https://www.nice.org.uk/guidance/ng139/chapter/Recommendations#treatment-for-adults-young-people-and-children

31. Antimicrobial Resistance: Tackling the Burden in the European Union [Internet]. AMR Insights. 2019 [cited 2021 Jun 9]. Available from: https://www.amr-insights.eu/antimicrobial-resistance-tackling-the-burden-in-the-european-union/

32. O'Neill J. Tackling Drug-Resistant Infections Globally: final report and recommendations. AMR Rev. 2016 May 19;

33. O'Neill J. Antimicrobial Resistance: Tackling a Crisis for the Future Health and Wealth of Nations. AMR Rev. 2014 Dec 11;

34. Wunderink RG, Srinivasan A, Barie PS, Chastre J, Dela Cruz CS, Douglas IS, et al. Antibiotic Stewardship in the Intensive Care Unit. An Official American Thoracic Society Workshop Report in Collaboration with the AACN, CHEST, CDC, and SCCM. Ann Am Thorac Soc. 2020 May;17(5):531–40.

35. Shapiro DJ, Hicks LA, Pavia AT, Hersh AL. Antibiotic prescribing for adults in ambulatory care in the USA, 2007-09. J Antimicrob Chemother. 2014 Jan;69(1):234–40.

36. Crowther GS, Wilcox MH. Antibiotic therapy and Clostridium difficile infection – primum non nocere – first do no harm. Infect Drug Resist. 2015 Sep 15;8:333–7.

37. O'Neill J. Rapid Diagnostics: Stopping unnecessary use of antibiotics. AMR Rev. 2015 Oct 13;

38. Liapikou A, Cillóniz C, Torres A. Emerging strategies for the noninvasive diagnosis of nosocomial pneumonia. Expert Rev Anti Infect Ther. 2019 Jun 25;17(7):523–33.

39. SMI B 57: investigation of bronchoalveolar lavage, sputum and associated specimens [Internet]. GOV.UK. [cited 2022 May 11]. Available from: https://www.gov.uk/government/publications/smi-b-57-investigation-of-bronchoalveolar-lavage-sputum-and-associated-specimens

40. UK SMI B 40: investigation of specimens for Mycobacterium species [Internet]. GOV.UK. [cited 2022 May 11]. Available from: https://www.gov.uk/government/publications/smi-b-40-investigation-of-specimens-for-mycobacterium-species

41. SMI B 31: investigation of specimens other than blood for parasites [Internet]. GOV.UK. [cited 2022 May 11]. Available from: https://www.gov.uk/government/publications/smi-b-31-investigation-of-specimens-other-than-blood-for-parasites

42. SMI S 2: Pneumonia [Internet]. GOV.UK. [cited 2022 May 11]. Available from: https://www.gov.uk/government/publications/smi-s-2-pneumonia

43. Enne VI, Aydin A, Baldan R, Owen DR, Richardson H, Ricciardi F, et al. Multicentre evaluation of two multiplex PCR platforms for the rapid microbiological investigation of nosocomial pneumonia in UK ICUs: the INHALE WP1 study. Thorax [Internet]. 2022 Jan 12 [cited 2022 May 15]; Available from: https://thorax.bmj.com/content/early/2022/01/12/thoraxjnl-2021-216990

44. Teixeira PJZ, Seligman R, Hertz FT, Cruz DB, Fachel JMG. Inadequate treatment of ventilator-associated pneumonia: risk factors and impact on outcomes. J Hosp Infect. 2007 Apr;65(4):361–7.

45. Jenkins SG, Schuetz AN. Current Concepts in Laboratory Testing to Guide Antimicrobial Therapy. Mayo Clin Proc. 2012 Mar;87(3):290–308.

46. Okamoto K, Gotoh N, Nishino T. Pseudomonas aeruginosa Reveals High Intrinsic Resistance to Penem Antibiotics: Penem Resistance Mechanisms and Their Interplay. Antimicrob Agents Chemother. 2001 Jul;45(7):1964-71.

47. EUCAST: Disk diffusion methodology [Internet]. [cited 2022 May 15]. Available from: https://www.eucast.org/ast_of_bacteria/disk_diffusion_methodology/

48. Brown DFJ, Wootton M, Howe RA. Antimicrobial susceptibility testing breakpoints and methods from BSAC to EUCAST. J Antimicrob Chemother. 2016 Jan 1;71(1):3–5.

49. Humphries RM, Abbott AN, Hindler JA. Understanding and Addressing CLSI Breakpoint Revisions: a Primer for Clinical Laboratories. J Clin Microbiol. 2019 May 24;57(6):e00203-19.

50. EUCAST: Clinical breakpoints and dosing of antibiotics [Internet]. [cited 2022 May 21]. Available from: https://www.eucast.org/clinical_breakpoints/

51. Khan ZA, Siddiqui MF, Park S. Current and Emerging Methods of Antibiotic Susceptibility Testing. Diagnostics. 2019 May 3;9(2):49.

52. Trotter AJ, Aydin A, Strinden MJ, O'Grady J. Recent and emerging technologies for the rapid diagnosis of infection and antimicrobial resistance. Curr Opin Microbiol. 2019 Oct 1;51:39–45.

53. Polymerase Chain Reaction | Elsevier Enhanced Reader [Internet]. [cited 2022 May 21]. Available from: https://reader.elsevier.com/reader/sd/pii/S0022202X1536139X?token=E2869C7318E22D477E5486B6991F15A37F0EEB95CD17B6F6EB0F1B2F7AE2739D94BBC54F6C29081C25E13B21B09C7FE4&originRegion=eu-west-1&originCreation=20220521140513

54. Nolte FS, Gauld L, Barrett SB. Direct Comparison of Alere i and cobas Liat Influenza A and B Tests for Rapid Detection of Influenza Virus Infection. J Clin Microbiol. 2016 Nov;54(11):2763–6.

55. The BioFire® FilmArray® Pneumonia Panel [Internet]. BioFire Diagnostics. [cited 2022 May 21]. Available from: https://www.biofiredx.com/products/the-filmarray-panels/filmarray-pneumonia/

56. Jamal W, Al Roomi E, AbdulAziz LR, Rotimi VO. Evaluation of Curetis Unyvero, a Multiplex PCR-Based Testing System, for Rapid Detection of Bacteria and Antibiotic Resistance and Impact of the Assay on Management of Severe Nosocomial Pneumonia. J Clin Microbiol. 2014 Jul;52(7):2487–92.

57. Ozongwu C, Personne Y, Platt G, Jeanes C, Aydin S, Kozato N, et al. The Unyvero P55 'sample-in, answer-out' pneumonia assay: A performance evaluation. Biomol Detect Quantif. 2017 Sep 1;13:1–6.

58. Papan C, Meyer-Buehn M, Laniado G, Nicolai T, Griese M, Huebner J. Assessment of the multiplex PCR-based assay Unyvero pneumonia application for detection of

bacterial pathogens and antibiotic resistance genes in children and neonates. Infection. 2018 Apr 1;46(2):189–96.

59. Edin A, Eilers H, Allard A. Evaluation of the Biofire Filmarray Pneumonia panel plus for lower respiratory tract infections. Infect Dis. 2020 Jul 2;52(7):479–88.

60. Yoo IY, Huh K, Shim HJ, Yun SA, Chung YN, Kang OK, et al. Evaluation of the BioFire FilmArray Pneumonia Panel for rapid detection of respiratory bacterial pathogens and antibiotic resistance genes in sputum and endotracheal aspirate specimens. Int J Infect Dis. 2020 Jun;95:326–31.

61. Ginocchio CC, Garcia-Mondragon C, Mauerhofer B, Rindlisbacher C, Forcelledo L, Fernández J, et al. Multinational evaluation of the BioFire® FilmArray® Pneumonia plus Panel as compared to standard of care testing. Eur J Clin Microbiol Infect Dis. 2021 Aug 1;40(8):1609–22.

62. Evaluation of the Accelerate Pheno System for Fast Identification and Antimicrobial Susceptibility Testing from Positive Blood Cultures in Bloodstream Infections Caused by Gram-Negative Pathogens | Journal of Clinical Microbiology [Internet]. [cited 2022 May 21]. Available from: https://journals.asm.org/doi/full/10.1128/JCM.00181-17

63. Schoepp NG, Schlappi TS, Curtis MS, Butkovich SS, Miller S, Humphries RM, et al. Rapid pathogen-specific phenotypic antibiotic susceptibility testing using digital LAMP quantification in clinical samples. Sci Transl Med. 2017 Oct 4;9(410):eaal3693.

64. Altobelli E, Mohan R, Mach KE, Sin MLY, Anikst V, Buscarini M, et al. Integrated Biosensor Assay for Rapid Uropathogen Identification and Phenotypic Antimicrobial Susceptibility Testing. Eur Urol Focus. 2017 Apr 1;3(2):293–9.

65. Sui W, Zhou H, Du P, Wang L, Qin T, Wang M, et al. Whole genome sequence revealed the fine transmission map of carbapenem-resistant Klebsiella pneumonia isolates within a nosocomial outbreak. Antimicrob Resist Infect Control. 2018 Jun 1;7(1):70.

66. Carlos CC, Masim MAL, Lagrada ML, Gayeta JM, Macaranas PKV, Sia SB, et al. Genome Sequencing Identifies Previously Unrecognized Klebsiella pneumoniae Outbreaks in Neonatal Intensive Care Units in the Philippines. Clin Infect Dis. 2021 Dec 1;73(Supplement_4):S316–24.

67. Deng X, Peirano G, Schillberg E, Mazzulli T, Gray-Owen SD, Wylie JL, et al. Whole-Genome Sequencing Reveals the Origin and Rapid Evolution of an Emerging Outbreak Strain of Streptococcus pneumoniae 12F. Clin Infect Dis. 2016 May 1;62(9):1126–32.

68. Doumith M, Godbole G, Ashton P, Larkin L, Dallman T, Day M, et al. Detection of the plasmid-mediated mcr-1 gene conferring colistin resistance in human and food isolates of Salmonella enterica and Escherichia coli in England and Wales. J Antimicrob Chemother. 2016 Aug;71(8):2300–5.

69. Argimón S, Masim MAL, Gayeta JM, Lagrada ML, Macaranas PKV, Cohen V, et al. Integrating whole-genome sequencing within the National Antimicrobial Resistance Surveillance Program in the Philippines. Nat Commun. 2020 Jun 1;11(1):2719.

70. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing | Journal of

Clinical Microbiology [Internet]. [cited 2022 May 21]. Available from: https://journals.asm.org/doi/full/10.1128/JCM.00666-18

71. Mongkolrattanothai K, Dien Bard J. The utility of direct specimen detection by Sanger sequencing in hospitalized pediatric patients. Diagn Microbiol Infect Dis. 2017 Feb;87(2):100–2.

72. Fida M, Khalil S, Abu Saleh O, Challener DW, Sohail MR, Yang JN, et al. Diagnostic Value of 16S Ribosomal RNA Gene Polymerase Chain Reaction/Sanger Sequencing in Clinical Practice. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021 Sep 15;73(6):961–8.

73. Allicock OM, Guo C, Uhlemann AC, Whittier S, Chauhan LV, Garcia J, et al. BacCapSeq: a Platform for Diagnosis and Characterization of Bacterial Infections. mBio. 2018 Nov 7;9(5):e02007-18.

74. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, et al. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. mBio. 2015 Oct 30;6(5):e01491-15.

75. Jain K, Tagliafierro T, Marques A, Sanchez-Vicente S, Gokden A, Fallon B, et al. Development of a capture sequencing assay for enhanced detection and genotyping of tick-borne pathogens. Sci Rep. 2021 Jun 11;11(1):12384.

76. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted metagenomics. Microbiome. 2018 Jan 15;6(1):11.

77. Bonnet I, Enouf V, Morel F, Ok V, Jaffré J, Jarlier V, et al. A Comprehensive Evaluation of GeneLEAD VIII DNA Platform Combined to Deeplex Myc-TB® Assay to Detect in 8 Days Drug Resistance to 13 Antituberculous Drugs and Transmission of Mycobacterium tuberculosis Complex Directly From Clinical Samples. Front Cell Infect Microbiol [Internet]. 2021 [cited 2022 May 21];11. Available from: https://www.frontiersin.org/article/10.3389/fcimb.2021.707244

78. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019 Jun;20(6):341–55.

79. Voelkerding KV, Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin Chem. 2009 Apr 1;55(4):641–58.

80. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, et al. A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. N Engl J Med. 2008 Mar 6;358(10):991–8.

81. Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life. 2021 Dec 26;12(1):30.

82. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Feb 7;21(1):30.

83. Aigrain L. Beginner's guide to next-generation sequencing. The Biochemist. 2021 Dec 3;43(6):58–64.

84. Jeon SA, Park JL, Kim JH, Kim JH, Kim YS, Kim JC, et al. Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing. Genomics Inform. 2019 Sep 27;17(3):e32.

85. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021 Nov 1;82(11):801–11.

86. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019 Oct;37(10):1155–62.

87. Oxford Nanopore technology update: CTO Clive G Brown unveils latest sequencing chemistry with highest performance to date, Short Fragment Mode and latest methylation performance evaluations [Internet]. Oxford Nanopore Technologies. [cited 2022 Aug 14]. Available from: http://nanoporetech.com/about-us/news/oxford-nanopore-technology-update-cto-clive-g-brown-unveils-latest-sequencing

88. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010 Jun;7(6):461–5.

89. Lewandowska DW, Zagordi O, Zbinden A, Schuurmans MM, Schreiber P, Geissberger FD, et al. Unbiased metagenomic sequencing complements specific routine diagnostic methods and increases chances to detect rare viral strains. Diagn Microbiol Infect Dis. 2015 Oct 1;83(2):133–8.

90. Naccache SN, Peggs KS, Mattes FM, Phadke R, Garson JA, Grant P, et al. Diagnosis of Neuroinvasive Astrovirus Infection in an Immunocompromised Adult With Encephalitis by Unbiased Next-Generation Sequencing. Clin Infect Dis. 2015 Mar 15;60(6):919–23.

91. Greninger AL, Zerr DM, Qin X, Adler AL, Sampoleo R, Kuypers JM, et al. Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. J Clin Microbiol. 2017 Jan;55(1):177–82.

92. Yan Q, Cui S, Chen C, Li S, Sha S, Wan X, et al. Metagenomic Analysis of Sputum Microbiome as a Tool toward Culture-Independent Pathogen Detection of Patients with Ventilator-associated Pneumonia. Am J Respir Crit Care Med. 2016 Sep;194(5):636–9.

93. Ruppé E, Lazarevic V, Girard M, Mouton W, Ferry T, Laurent F, et al. Clinical metagenomics of bone and joint infections: a proof of concept study. Sci Rep. 2017 Aug 10;7(1):7718.

94. PacBio Sequel II Library Prep & Sequencing | DNA Technologies Core [Internet]. [cited 2022 Jun 5]. Available from: https://dnatech.genomecenter.ucdavis.edu/pacbio-library-prep-sequencing/

95. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 2015 Sep 29;7(1):99.

96. Street TL, Sanderson ND, Atkins BL, Brent AJ, Cole K, Foster D, et al. Molecular Diagnosis of Orthopedic-Device-Related Infection Directly from Sonication Fluid by Metagenomic Sequencing. J Clin Microbiol. 2017 Aug;55(8):2334–47.

97. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019 Jul;37(7):783–92.

98. Buytaers FE, Saltykova A, Denayer S, Verhaegen B, Vanneste K, Roosens NHC, et al. Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies. Front Microbiol [Internet]. 2021 [cited 2022 Jun 5];12. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2021.738284

99. Nanopore store [Internet]. Nanoporetech. [cited 2022 Jun 5]. Available from: https://store.nanoporetech.com/uk/devices.html

100. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018 Jan;19(1):9–20.

101. Q-Line [Internet]. Oxford Nanopore Technologies. [cited 2022 Jun 5]. Available from: http://nanoporetech.com/products/qline

102. Zhao F, Bajic VB. The Value and Significance of Metagenomics of Marine Environments. Genomics Proteomics Bioinformatics. 2015 Oct 1;13(5):271–4.

103. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016 Apr 27;8(1):51.

104. Xie F, Duan Z, Zeng W, Xie S, Xie M, Fu H, et al. Clinical metagenomics assessments improve diagnosis and outcomes in community-acquired pneumonia. BMC Infect Dis. 2021 Apr 15;21:352.

105. Yang T, Mei Q, Fang X, Zhu S, Wang Y, Li W, et al. Clinical Value of Metagenomics Next-Generation Sequencing in Bronchoalveolar Lavage Fluid for Patients with Severe Hospital-Acquired Pneumonia: A Nested Case–Control Study. Infect Drug Resist. 2022 Apr 5;15:1505–14.

106. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 2019 Jan 5;29(5):831–42.

107. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, et al. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol. 2019 Apr;4(4):663–74.

108. Barraud O, Ravry C, François B, Daix T, Ploy MC, Vignon P. Shotgun metagenomics for microbiome and resistome detection in septic patients with urinary tract infection. Int J Antimicrob Agents. 2019 Dec 1;54(6):803–8.

109. Schmidt K, Mwaigwisya S, Crossman LC, Doumith M, Munroe D, Pires C, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J Antimicrob Chemother. 2017 Jan 1;72(1):104–14.

110. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, et al. Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. Nat Microbiol. 2020 Mar;5(3):430–42.

111. Lewandowski K, Xu Y, Pullan ST, Lumley SF, Foster D, Sanderson N, et al. Metagenomic Nanopore Sequencing of Influenza Virus Direct from Clinical Respiratory Samples. J Clin Microbiol [Internet]. 2020 Jan [cited 2022 May 22];58(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6935926/

112. Song P, Chen S, Tan X, Gao Y, Fu J, You Z, et al. Metagenomic Analysis Identifying a Rare Leishmania Infection in an Adult With AIDS. Front Cell Infect Microbiol

[Internet]. 2021 [cited 2022 May 22];11. Available from: https://www.frontiersin.org/article/10.3389/fcimb.2021.764142

113. De R. Metagenomics: aid to combat antimicrobial resistance in diarrhea. Gut Pathog. 2019 Oct 14;11(1):47.

114. Zhao R, Yu K, Zhang J, Zhang G, Huang J, Ma L, et al. Deciphering the mobility and bacterial hosts of antibiotic resistance genes under antibiotic selection pressure by metagenomic assembly and binning approaches. Water Res. 2020 Nov 1;186:116318.

115. Charalampous T, Alcolea-Medina A, Snell LB, Williams TGS, Batra R, Alder C, et al. Evaluating the potential for respiratory metagenomics to improve treatment of secondary infection and detection of nosocomial transmission on expanded COVID-19 intensive care units. Genome Med. 2021 Nov 17;13(1):182.

116. Manenzhe RI, Dube FS, Wright M, Lennard K, Mounaud S, Lo SW, et al. Characterization of Pneumococcal Colonization Dynamics and Antimicrobial Resistance Using Shotgun Metagenomic Sequencing in Intensively Sampled South African Infants. Front Public Health [Internet]. 2020 [cited 2022 May 22];8. Available from: https://www.frontiersin.org/article/10.3389/fpubh.2020.543898

117. Ruppé E, d'Humières C, Armand-Lefèvre L. Inferring antibiotic susceptibility from metagenomic data: dream or reality? Clin Microbiol Infect [Internet]. 2022 May 10 [cited 2022 May 22]; Available from: https://www.sciencedirect.com/science/article/pii/S1198743X22002294

118. Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. EMBO Mol Med. 2020 Mar 6;12(3):e10264.

119. Stalder T, Press MO, Sullivan S, Liachko I, Top EM. Linking the resistome and plasmidome to the microbiome. ISME J. 2019 Oct;13(10):2437–46.

120. Břinda K, Callendrello A, Ma KC, MacFadden DR, Charalampous T, Lee RS, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. Nat Microbiol. 2020 Mar;5(3):455–64.

121. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014 Nov 12;12(1):87.

122. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G, the Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology, et al. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. Arch Pathol Lab Med. 2017 Feb 7;141(6):776–86.

123. Belda J, Leigh R, Parameswaran K, O'Byrne PM, Sears MR, Hargreave FE. Induced sputum cell counts in healthy adults. Am J Respir Crit Care Med. 2000 Feb;161(2 Pt 1):475–8.

124. Choi SH, Hong SB, Hong HL, Kim SH, Huh JW, Sung H, et al. Usefulness of Cellular Analysis of Bronchoalveolar Lavage Fluid for Predicting the Etiology of Pneumonia in Critically Ill Patients. PLoS ONE. 2014 May 13;9(5):e97346.

125. Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. BMC Res Notes. 2019 Feb 27;12(1):106.

126. Wolff T, Veit M. Influenza B, C and D Viruses (Orthomyxoviridae). In: Bamford DH, Zuckerman M, editors. Encyclopedia of Virology (Fourth Edition) [Internet]. Oxford: Academic Press; 2021 [cited 2022 May 28]. p. 561–74. Available from: https://www.sciencedirect.com/science/article/pii/B9780128096338215057

127. Campbell AP, Chien JW, Kuypers J, Englund JA, Wald A, Guthrie KA, et al. Respiratory Virus Pneumonia after Hematopoietic Cell Transplantation (HCT): Associations between Viral Load in Bronchoalveolar Lavage Samples, Viral RNA Detection in Serum Samples, and Clinical Outcomes of HCT. J Infect Dis. 2010 May 1;201(9):1404–13.

128. Tsou TP, Shao PL, Lu CY, Chang LY, Kao CL, Lee PI, et al. Viral load and clinical features in children infected with seasonal influenza B in 2006/2007. J Formos Med Assoc. 2012 Feb 1;111(2):83–7.

129. Couto N, Schuele L, Raangs EC, Machado MP, Mendes CI, Jesus TF, et al. Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Sci Rep. 2018 Sep 13;8(1):13767.

130. Deng X, Achari A, Federman S, Yu G, Somasekar S, Bártolo I, et al. Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. Nat Microbiol. 2020 Mar;5(3):443–54.

131. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018 Feb 27;6:42.

132. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. Nat Rev Microbiol. 2021;19(8):514–27.

133. Shi Y, Wang G, Lau HCH, Yu J. Metagenomic Sequencing for Microbial DNA in Human Samples: Emerging Technological Advances. Int J Mol Sci. 2022 Jan;23(4):2181.

134. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol. 2016 Apr;54(4):919–27.

135. Fong W, Rockett R, Timms V, Sintchenko V 2020. Optimization of sample preparation for culture-independent sequencing of Bordetella pertussis. Microb Genomics. 6(3):e000332.

136. Kuznetsov AV, Veksler V, Gellerich FN, Saks V, Margreiter R, Kunz WS. Analysis of mitochondrial function in situ in permeabilized muscle fibers, tissues and cells. Nat Protoc. 2008 Jun;3(6):965–76.

137. Krawczyk E, Suprynowicz F, Sudarshan S, Schlegel R. Membrane Orientation of the Human Papillomavirus Type 16 E5 Oncoprotein. J Virol. 2009 Dec 1;84:1696–703.

138. Thapa R, Ray S, Keyel PA. Interaction of Macrophages and Cholesterol-Dependent Cytolysins: The Impact on Immune Response and Cellular Survival. Toxins. 2020 Aug 19;12(9):531.

139. Horz HP, Scheer S, Huenger F, Vianna ME, Conrads G. Selective isolation of bacterial DNA from human clinical specimens. J Microbiol Methods. 2008 Jan;72(1):98–102.

140. Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, et al. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. PloS One. 2013;8(10):e76096.

141. Glassing A, Dowd SE, Galandiuk S, Davis B, Jorden JR, Chiodini RJ. Changes in 16s RNA Gene Microbial Community Profiling by Concentration of Prokaryotic DNA. J Microbiol Methods. 2015 Dec 1;119:239–42.

142. Liu G, Weston CQ, Pham LK, Waltz S, Barnes H, King P, et al. Epigenetic Segregation of Microbial Genomes from Complex Samples Using Restriction Endonucleases HpaII and McrB. PLoS ONE. 2016 Jan 4;11(1):e0146064.

143. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods. 2016 Aug 1;127:141–5.

144. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016 Mar 4;17:41.

145. Marquet M, Zöllkau J, Pastuschek J, Viehweger A, Schleußner E, Makarewicz O, et al. Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. Sci Rep. 2022 Mar 7;12(1):4000.

146. Oechslin CP, Lenz N, Liechti N, Ryter S, Agyeman P, Bruggmann R, et al. Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples. Front Cell Infect Microbiol. 2018 Oct 23;8:375.

147. Ivy MI, Thoendel MJ, Jeraldo PR, Greenwood-Quaintance KE, Hanssen AD, Abdel MP, et al. Direct Detection and Identification of Prosthetic Joint Infection Pathogens in Synovial Fluid by Metagenomic Shotgun Sequencing. J Clin Microbiol. 2018 Sep;56(9):e00402-18.

148. Vijayvargiya P, Jeraldo PR, Thoendel MJ, Greenwood-Quaintance KE, Garrigos ZE, Sohail MR, et al. Application of metagenomic shotgun sequencing to detect vector-borne pathogens in clinical blood samples. PLOS ONE. 2019 Oct 2;14(10):e0222915.

149. Lazarevic V, Gaïa N, Girard M, Leo S, Cherkaoui A, Renzi G, et al. When Bacterial Culture Fails, Metagenomics Can Help: A Case of Chronic Hepatic Brucelloma Assessed by Next-Generation Sequencing. Front Microbiol [Internet]. 2018 [cited 2022 Jun 2];9. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2018.01566

150. Kolb M, Lazarevic V, Emonet S, Calmy A, Girard M, Gaïa N, et al. Next-Generation Sequencing for the Diagnosis of Challenging Culture-Negative Endocarditis. Front Med [Internet]. 2019 [cited 2022 Jun 2];6. Available from: https://www.frontiersin.org/article/10.3389/fmed.2019.00203

151. Yang L, Haidar G, Zia H, Nettles R, Qin S, Wang X, et al. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. Respir Res. 2019 Nov 27;20(1):265.

152. Schumacher SG, Wells WA, Nicol MP, Steingart KR, Theron G, Dorman SE, et al. Guidance for Studies Evaluating the Accuracy of Sputum-Based Tests to Diagnose Tuberculosis. J Infect Dis. 2019 Oct 8;220(Supplement_3):S99–107.

153. Peng J, Lu Y, Song J, Vallance BA, Jacobson K, Yu HB, et al. Direct Clinical Evidence Recommending the Use of Proteinase K or Dithiothreitol to Pretreat Sputum for Detection of SARS-CoV-2. Front Med [Internet]. 2020 [cited 2022 Jun 3];7. Available from: https://www.frontiersin.org/article/10.3389/fmed.2020.549860

154. Shehadul Islam M, Aryasomayajula A, Selvaganapathy PR. A Review on Macroscale and Microscale Cell Lysis Methods. Micromachines. 2017 Mar 8;8(3):83.

155. Oriano M, Terranova L, Teri A, Sottotetti S, Ruggiero L, Tafuro C, et al. Comparison of different conditions for DNA extraction in sputum - a pilot study. Multidiscip Respir Med. 2019 Jan 31;14(1):6.

156. Li X, Bosch-Tijhof CJ, Wei X, de Soet JJ, Crielaard W, van Loveren C, et al. Efficiency of chemical versus mechanical disruption methods of DNA extraction for the identification of oral Gram-positive and Gram-negative bacteria. J Int Med Res. 2020 May 27;48(5):0300060520925594.

157. Vandeventer PE, Weigel KM, Salazar J, Erwin B, Irvine B, Doebler R, et al. Mechanical Disruption of Lysis-Resistant Bacterial Cells by Use of a Miniature, Low-Power, Disposable Device ▿. J Clin Microbiol. 2011 Jul;49(7):2533–9.

158. Hwang KY, Kwon SH, Jung SO, Lim HK, Jung WJ, Park CS, et al. Miniaturized bead-beating device to automate full DNA sample preparation processes for Gram-positive bacteria. Lab Chip. 2011 Oct 11;11(21):3649–55.

159. Zinter MS, Mayday MY, Ryckman KK, Jelliffe-Pawlowski LL, DeRisi JL. Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome. 2019 Apr 16;7(1):62.

160. Jurasz H, Pawłowski T, Perlejewski K. Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives. Front Microbiol. 2021 Oct 20;12:745076.

161. Kjartansdóttir KR, Friis-Nielsen J, Asplund M, Mollerup S, Mourier T, Jensen RH, et al. Traces of ATCV-1 associated with laboratory component contamination. Proc Natl Acad Sci U S A. 2015 Mar 3;112(9):E925–6.

162. Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, Herrera J a. R, et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2019 Oct;25(10):1277–85.

163. Dundas N, Leos NK, Mitui M, Revell P, Rogers B. Comparison of Automated Nucleic Acid Extraction Methods with Manual Extraction. J Mol Diagn JMD. 2008 Aug 1;10:311–6.

164. Colman RE, Suresh A, Dolinger DL, Muñoz T, Denkinger CM, Rodwell TC. Review of automated DNA extraction systems for sequencing-based solutions for drug-resistant tuberculosis detection. Diagn Microbiol Infect Dis. 2020 Oct 1;98(2):115096.

165. Sui H yu, Weil AA, Nuwagira E, Qadri F, Ryan ET, Mezzari MP, et al. Impact of DNA Extraction Method on Variation in Human and Built Environment Microbial

Community and Functional Profiles Assessed by Shotgun Metagenomics Sequencing. Front Microbiol. 2020 May 25;11:953.

166. Hess JF, Kohl TA, Kotrová M, Rönsch K, Paprotka T, Mohr V, et al. Library preparation for next generation sequencing: A review of automation strategies. Biotechnol Adv. 2020 Jul 1;41:107537.

167. Tegally H, San JE, Giandhari J, de Oliveira T. Unlocking the efficiency of genomics laboratories with robotic liquid-handling. BMC Genomics. 2020 Dec;21(1):1–15.

168. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016 Apr 13;7(1):11257.

169. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007 Mar;17(3):377–86.

170. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2017 Sep 23;20(4):1125–36.

171. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015 Mar 25;16:236.

172. Piro VC, Lindner MS, Renard BY. DUDes: a top-down taxonomic profiler for metagenomics. Bioinforma Oxf Engl. 2016 Aug 1;32(15):2272–80.

173. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.

174. Freitas TAK, Li PE, Scholz MB, Chain PSG. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 2015 May 26;43(10):e69.

175. Müller A, Hundt C, Hildebrandt A, Hankeln T, Schmidt B. MetaCache: context-aware classification of metagenomic reads using minhashing. Bioinformatics. 2017 Dec 1;33(23):3740–8.

176. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015 Oct;12(10):902–3.

177. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011 Jul 27;12(Suppl 2):S4.

178. Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat Commun. 2019 Jul 11;10(1):3066.

179. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019 Mar 4;10(1):1014.

180. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016 Dec;26(12):1721–9.

181. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46.

182. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257.

183. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics. 2013 Sep 15;29(18):2253–60.

184. Bağcı C, Patz S, Huson DH. DIAMOND+MEGAN: Fast and Easy Taxonomic and Functional Analysis of Short and Long Microbiome Sequences. Curr Protoc. 2021;1(3):e59.

185. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 2019 Aug 8;178(4):779–94.

186. Meyer F, Fritz A, Deng ZL, Koslicki D, Lesker TR, Gurevich A, et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. Nat Methods. 2022 Apr;19(4):429–40.

187. Tran Q, Phan V. Assembling Reads Improves Taxonomic Classification of Species. Genes. 2020 Aug 17;11(8):946.

188. Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. Brief Bioinform. 2019 Feb 28;21(2):584–94.

189. Sun Z, Huang S, Zhang M, Zhu Q, Haiminen N, Carrieri AP, et al. Challenges in benchmarking metagenomic profilers. Nat Methods. 2021 Jun;18(6):618–26.

190. Arita M, Karsch-Mizrachi I, Cochrane G, on behalf of the International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021 Jan 8;49(D1):D121–4.

191. RefSeq: NCBI Reference Sequence Database [Internet]. [cited 2022 Jun 15]. Available from: https://www.ncbi.nlm.nih.gov/refseq/

192. Kraken2 [Internet]. [cited 2022 Jun 15]. Available from: https://ccb.jhu.edu/software/kraken2/index.shtml?t=downloads

193. Index zone by BenLangmead [Internet]. [cited 2022 Jun 15]. Available from: https://benlangmead.github.io/aws-indexes/

194. Prokaryotic RefSeq Genomes [Internet]. [cited 2022 Aug 14]. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

195. DRAGEN Metagenomics Pipeline [Internet]. [cited 2022 Jun 11]. Available from: https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-metagenomics-pipeline.html

196. EPI2ME WIMP workflow: quantitative, real-time species identification from metagenomic samples [Internet]. Oxford Nanopore Technologies. 2019 [cited 2022 Jun 11]. Available from: http://nanoporetech.com/resource-centre/epi2me-wimp-workflow-quantitative-real-time-species-identification-metagenomic

197. wf-metagenomics [Internet]. EPI2ME Labs; 2022 [cited 2022 Jun 11]. Available from: https://github.com/epi2me-labs/wf-metagenomics

198. Wissel EF, Talbot BM, Johnson BA, Petit RA, Hertzberg V, Dunlop A, et al. Benchmarking software to predict antibiotic resistance phenotypes in shotgun metagenomes using simulated data [Internet]. bioRxiv; 2022 [cited 2022 Jun 11]. p. 2022.01.13.476279. Available from: https://www.biorxiv.org/content/10.1101/2022.01.13.476279v1

199. Seemann T. ABRicate [Internet]. 2022 [cited 2022 Jun 11]. Available from: https://github.com/tseemann/abricate

200. Panunzi LG. sraX: A Novel Comprehensive Resistome Analysis Tool. Front Microbiol [Internet]. 2020 [cited 2022 Jun 11];11. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.00052

201. Bharat A, Petkau A, Avery BP, Chen JC, Folster JP, Carson CA, et al. Correlation between Phenotypic and In Silico Detection of Antimicrobial Resistance in Salmonella enterica in Canada Using Staramr. Microorganisms. 2022 Jan 26;10(2):292.

202. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021 Jun 16;11(1):12728.

203. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother. 2020 Dec 1;75(12):3491–500.

204. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens | Journal of Antimicrobial Chemotherapy | Oxford Academic [Internet]. [cited 2022 Jun 11]. Available from: https://academic.oup.com/jac/article/72/10/2764/3979530

205. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. PLOS Comput Biol. 2015 Dec 18;11(12):e1004557.

206. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020 Jan 8;48(D1):D517–25.

207. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58(1):212–20.

208. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genomics. 2022 Jan 24;8(1):000748.

209. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. Nucleic Acids Res. 2020 Jan 8;48(D1):D561–9.

210. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D443-447.

211. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2008 Jan;36(Database issue):D190–5.

212. Tournoud M, Ruppé E, Perrin G, Schicklin S, Guigon G, Mahé P, et al. Clinical metagenomics bioinformatics pipeline for the identification of hospital-acquired pneumonia pathogens antibiotic resistance genes from bronchoalveolar lavage samples [Internet]. bioRxiv; 2020 [cited 2022 Aug 16]. p. 2020.02.26.966309. Available from: https://www.biorxiv.org/content/10.1101/2020.02.26.966309v1

213. Seth-Smith HMB, Harris SR, Skilton RJ, Radebe FM, Golparian D, Shipitsyna E, et al. Whole-genome sequences of Chlamydia trachomatis directly from clinical samples without culture. Genome Res. 2013 May;23(5):855–66.

214. Andersson P, Klein M, Lilliebridge RA, Giffard PM. Sequences of multiple bacterial genomes and a Chlamydia trachomatis genotype from direct sequencing of DNA derived from a vaginal swab diagnostic specimen. Clin Microbiol Infect. 2013 Sep 1;19(9):E405–8.

215. Casto AM, Adler AL, Makhsous N, Crawford K, Qin X, Kuypers JM, et al. Prospective, Real-time Metagenomic Sequencing During Norovirus Outbreak Reveals Discrete Transmission Clusters. Clin Infect Dis. 2019 Aug 30;69(6):941–8.

216. Huang AD, Luo C, Pena-Gonzalez A, Weigand MR, Tarr CL, Konstantinidis KT. Metagenomics of Two Severe Foodborne Outbreaks Provides Diagnostic Signatures and Signs of Coinfection Not Attainable by Traditional Methods. Appl Environ Microbiol. 2017 Feb;83(3):e02577-16.

217. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.

218. Poon AFY, Gustafson R, Daly P, Zerr L, Demlow SE, Wong J, et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV. 2016 May;3(5):e231–8.

219. Gardy J, Loman NJ, Rambaut A. Real-time digital pathogen surveillance — the time is now. Genome Biol. 2015 Jul 30;16(1):155.

220. Houlihan CF, Frampton D, Ferns RB, Raffle J, Grant P, Reidy M, et al. Use of Whole-Genome Sequencing in the Investigation of a Nosocomial Influenza Virus Outbreak. J Infect Dis. 2018 Nov 1;218(9):1485–9.

221. Blackburn RM, Frampton D, Smith CM, Fragaszy EB, Watson SJ, Ferns RB, et al. Nosocomial transmission of influenza: A retrospective cross-sectional study using next generation sequencing at a hospital in England (2012-2014). Influenza Other Respir Viruses. 2019;13(6):556–63.

222. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228–32.

223. Ebola outbreak 2014-2016 - West Africa [Internet]. [cited 2022 Jun 29]. Available from: https://www.who.int/emergencies/situations/ebola-outbreak-2014-2016-West-Africa

224. Pollett S, Fauver JR, Maljkovic Berry I, Melendrez M, Morrison A, Gillis LD, et al. Genomic Epidemiology as a Public Health Tool to Combat Mosquito-Borne Virus Outbreaks. J Infect Dis. 2020 Mar 28;221(Supplement_3):S308–18.

225. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017 Jun;12(6):1261–76.

226. Thézé J, Li T, du Plessis L, Bouquet J, Kraemer MUG, Somasekar S, et al. Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. Cell Host Microbe. 2018 Jun 13;23(6):855-864.e7.

227. Grubaugh ND, Ladner JT, Kraemer MU, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017 Jun 15;546(7658):401–5.

228. Faria RN, Quick J, Morales I, Thézé J, Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017 Jun 15;546(7658):406–10.

229. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol. 2021 Mar;19(3):141–54.

230. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021 Jul;19(7):409–24.

231. Menni C, Valdes AM, Polidori L, Antonelli M, Penamakuri S, Nogal A, et al. Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: a prospective observational study from the ZOE COVID Study. The Lancet. 2022 Apr 23;399(10335):1618–24.

232. BIO COVID-19 Therapeutic Development Tracker | BIO [Internet]. [cited 2022 Jun 29]. Available from: https://www.bio.org/policy/human-health/vaccines-biodefense/coronavirus/pipeline-tracker

233. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. N Engl J Med. 2020 Dec 31;383(27):2603–15.

234. Holder J. Tracking Coronavirus Vaccinations Around the World. The New York Times [Internet]. 2021 Jan 29 [cited 2022 Jun 29]; Available from: https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html

235. Welcome — RECOVERY Trial [Internet]. [cited 2021 Jun 22]. Available from: https://www.recoverytrial.net/

236. Dexamethasone in Hospitalized Patients with Covid-19. N Engl J Med. 2021 Feb 25;384(8):693–704.

237. RECOVERY trial finds Regeneron's monoclonal antibody combination reduces deaths for hospitalised COVID-19 patients who have not mounted their own immune response — RECOVERY Trial [Internet]. [cited 2021 Jun 22]. Available from: https://www.recoverytrial.net/news/recovery-trial-finds-regeneron2019s-monoclonal-antibody-combination-reduces-deaths-for-hospitalised-covid-19-patients-who-have-not-mounted-their-own-immune-response-1

238. Results — RECOVERY Trial [Internet]. [cited 2022 Jun 29]. Available from: https://www.recoverytrial.net/results

239. Mahase E. Covid-19: Pfizer's paxlovid is 89% effective in patients at risk of serious illness, company reports. BMJ. 2021 Nov 8;375:n2713.

240. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020 Jan 23;25(3):2000045.

241. CDC. Labs [Internet]. Centers for Disease Control and Prevention. 2020 [cited 2022 Jun 29]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/lab/virus-requests.html

242. Michael-Kordatou I, Karaolia P, Fatta-Kassinos D. Sewage analysis as a tool for the COVID-19 pandemic response and management: the urgent need for optimised protocols for SARS-CoV-2 detection and quantification. J Environ Chem Eng. 2020 Oct;8(5):104306.

243. Pickering S, Batra R, Merrick B, Snell LB, Nebbia G, Douthwaite S, et al. Comparative performance of SARS-CoV-2 lateral flow antigen tests and association with detection of infectious virus in clinical specimens: a single-centre laboratory evaluation study. Lancet Microbe. 2021 Sep 1;2(9):e461–71.

244. Vogels CBF, Breban MI, Ott IM, Alpert T, Petrone ME, Watkins AE, et al. Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. PLOS Biol. 2021 May 7;19(5):e3001236.

245. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe. 2020 Jul 1;1(3):e99–100.

246. Consortium Partners | COVID-19 Genomics UK Consortium [Internet]. COVID-19 Genomics UK Consortium | UK-Wide Genomic Sequencing. 2021 [cited 2021 Jun 22]. Available from: https://www.cogconsortium.uk/cog-uk/consortium-partners/

247. Real-time Assessment of Community Transmission (REACT) Study | Faculty of Medicine | Imperial College London [Internet]. [cited 2022 Jun 29]. Available from: https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/

248. Mashe T, Takawira FT, Martins L de O, Gudza-Mugabe M, Chirenda J, Munyanyi M, et al. Genomic epidemiology and the role of international and regional travel in the SARS-CoV-2 epidemic in Zimbabwe: a retrospective study of routinely collected surveillance data. Lancet Glob Health. 2021 Dec 1;9(12):e1658–66.

249. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020 Nov;5(11):1403–7.

250. Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, et al. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. Genome Biol. 2021 Jul 1;22(1):196.

251. UK SARS-CoV-2 (2020-02-05⇨2022-02-11) [Internet]. Microreact. [cited 2022 Jun 29]. Available from: https://microreact.org/project/eDRh8N1gx66onGUvPAsQik-uk-sars-cov-2-2020-02-052022-02-11

252. UK completes over 2 million SARS-CoV-2 whole genome sequences [Internet]. GOV.UK. [cited 2022 Jun 29]. Available from: https://www.gov.uk/government/news/uk-completes-over-2-million-sars-cov-2-whole-genome-sequences

253. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 2021 Jan 7;184(1):64-75.e11.

254. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021 May;593(7858):266–9.

255. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021 Jan 8;eabf2946.

256. Covid-19: Christmas rules tightened for England, Scotland and Wales. BBC News [Internet]. 2020 Dec 20 [cited 2021 Jul 4]; Available from: https://www.bbc.com/news/uk-55379220

257. Surge testing to be deployed in Wandsworth and Lambeth [Internet]. GOV.UK. [cited 2021 Jul 4]. Available from: https://www.gov.uk/government/news/surge-testing-to-be-deployed-in-wandsworth-and-lambeth

258. Countries added to red list to protect UK against variants of concern [Internet]. GOV.UK. [cited 2021 Jul 4]. Available from: https://www.gov.uk/government/news/countries-added-to-red-list-to-protect-uk-against-variants-of-concern

259. Covid: Lockdown easing in England to be delayed by four weeks. BBC News [Internet]. 2021 Jun 14 [cited 2021 Jul 4]; Available from: https://www.bbc.com/news/uk-57464097

260. Tracking SARS-CoV-2 variants [Internet]. [cited 2022 Jun 30]. Available from: https://www.who.int/activities/tracking-SARS-CoV-2-variants

261. Wise J. Covid-19: Omicron sub variants driving new wave of infections in UK. BMJ. 2022 Jun 20;377:o1506.

262. Artic Network [Internet]. [cited 2022 Jun 30]. Available from: https://artic.network/

263. Quick J. nCoV-2019 sequencing protocol [Internet]. protocols.io. 2020 [cited 2022 Jun 30]. Available from: https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w

264. Illumina COVIDSeq Test | SARS-CoV-2 NGS test (for the COVID-19 Coronavirus) [Internet]. [cited 2022 Aug 16]. Available from: https://emea.illumina.com/products/by-type/ivd-products/covidseq.html

265. DNA Pipelines R&D, Farr B, Rajan D, Betteridge E, Shirley L, Quail M, et al. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol. 2020 [cited 2021 Jan 22]; Available from: https://www.protocols.io/view/covid-19-artic-v3-illumina-library-construction-an-bgq3jvyn

266. Baker DJ, Aydin A, Le-Viet T, Kay GL, Rudder S, de Oliveira Martins L, et al. CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. Genome Med. 2021 Feb 9;13(1):21.

267. Lambisia AW, Mohammed KS, Makori TO, Ndwiga L, Mburu MW, Morobe JM, et al. Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol. Front Med. 2022 Feb 17;9:836728.

268. Quick J. nCoV-2019 sequencing protocol v2 (GunIt) [Internet]. protocols.io. 2020 [cited 2022 Apr 2]. Available from: https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bdp7i5rn

269. Quick J. nCoV-2019 sequencing protocol v3 (LoCost). 2020 [cited 2021 Jan 22]; Available from: https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye

270. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biol Methods Protoc [Internet]. 2020 Jan 1 [cited 2021 Jun 14];5(1). Available from: https://doi.org/10.1093/biomethods/bpaa014

271. Resende PC, Motta FC, Roy S, Appolinario L, Fabri A, Xavier J, et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms [Internet]. bioRxiv; 2020 [cited 2022 Jun 30]. p. 2020.04.30.069039. Available from: https://www.biorxiv.org/content/10.1101/2020.04.30.069039v1

272. Eden JS, Sim E. SARS-CoV-2 Genome Sequencing Using Long Pooled Amplicons on Illumina Platforms [Internet]. protocols.io. 2020 [cited 2022 Jun 30]. Available from: https://www.protocols.io/view/sars-cov-2-genome-sequencing-using-long-pooled-amp-befyjbpw

273. Brejová B, Boršová K, Hodorová V, Čabanová V, Gafurov A, Fričová D, et al. Nanopore Sequencing of SARS-CoV-2: Comparison of Short and Long PCR-tiling Amplicon Protocols. medRxiv. 2021 May 13;2021.05.12.21256693.

274. Oxford Nanopore launches 'Midnight kit', suitable for low to high-throughput SARS-CoV-2 sequencing, enabling rapid, low-cost, large-scale genomic surveillance of COVID-19 [Internet]. Oxford Nanopore Technologies. [cited 2022 Jun 30]. Available from: http://nanoporetech.com/about-us/news/oxford-nanopore-launches-midnight-kit-suitable-low-high-throughput-sars-cov-2

275. Nasir JA, Kozak RA, Aftanas P, Raphenya AR, Smith KM, Maguire F, et al. A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture. Viruses. 2020 Aug 15;12(8):895.

276. Y Kim B, E. Miller D, Wang J. DNA extraction and Nanopore library prep from 15-30 whole flies v1 [Internet]. [cited 2022 Mar 25]. Available from: https://www.protocols.io/view/dna-extraction-and-nanopore-library-prep-from-15-3-bdfqi3mw

277. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37(7):783–92.

278. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013 Jan 1;41(1):e1.

279. Palmer S, Wiegand AP, Maldarelli F, Bazmi H, Mican JM, Polis M, et al. New Real-Time Reverse Transcriptase-Initiated PCR Assay with Single-Copy Sensitivity for Human Immunodeficiency Virus Type 1 RNA in Plasma. J Clin Microbiol. 2003 Oct;41(10):4531–6.

280. Fukumoto H, Sato Y, Hasegawa H, Saeki H, Katano H. Development of a new real-time PCR system for simultaneous detection of bacteria and fungi in pathological samples. Int J Clin Exp Pathol. 2015 Nov 1;8(11):15479–88.

281. quadram-institute-bioscience/coronahit_guppy [Internet]. Quadram Institute Bioscience; 2020 [cited 2022 Apr 2]. Available from: https://github.com/quadram-institute-bioscience/coronahit_guppy

282. Releases · artic-network/fieldbioinformatics [Internet]. GitHub. [cited 2022 Apr 2]. Available from: https://github.com/artic-network/fieldbioinformatics/releases

283. quadram-institute-bioscience/fieldbioinformatics [Internet]. Quadram Institute Bioscience; 2021 [cited 2022 Apr 2]. Available from: https://github.com/quadram-institute-bioscience/fieldbioinformatics

284. iVar [Internet]. Andersen Laboratory @ Scripps Research; 2022 [cited 2022 Apr 2]. Available from: https://github.com/andersen-lab/ivar

285. ncov2019-artic-nf [Internet]. Quadram Institute Bioscience; 2022 [cited 2022 Apr 2]. Available from: https://github.com/quadram-institute-bioscience/ncov2019-artic-nf

286. Davalieva KG, Efremov GD. A new thermostable DNA polymerase mixture for efficient amplification of long DNA fragments. Prikl Biokhim Mikrobiol. 2010 Apr;46(2):248–52.

287. Scagaire [Internet]. Quadram Institute Bioscience; 2021 [cited 2022 Jul 9]. Available from: https://github.com/quadram-institute-bioscience/scagaire

288. Lu J. Bracken 2.7 abundance estimation [Internet]. 2022 [cited 2022 Jul 9]. Available from: https://github.com/jenniferlu717/Bracken

289. Chrzastek K, Lee DH, Smith D, Sharma P, Suarez DL, Pantin-Jackwood M, et al. Use of Sequence-Independent, Single-Primer-Amplification (SISPA) for rapid detection, identification, and characterization of avian RNA viruses. Virology. 2017 Sep;509:159–66.

290. Pacific Biosciences. Guidelines for Using PacBio Barcodes for SMRT Sequencing [Internet]. [cited 2022 Jul 24]. Available from: https://www.pacb.com/wp-content/uploads/Shared-Protocol-PacBio-Barcodes-for-SMRT-Sequencing.pdf

291. Illumina. Illumina Adapter Sequences [Internet]. 2015 [cited 2022 Jul 24]. Available from: https://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2013/06/illumina-adapter-sequences_1000000002694-00.pdf

292. Oligo synthesis: Coupling efficiency and quality control | IDT [Internet]. Integrated DNA Technologies. [cited 2022 Jul 17]. Available from: https://eu.idtdna.com/pages/education/decoded/article/oligo-synthesis-why-idt-leads-the-oligo-industry

293. Perez-Sepulveda BM, Heavens D, Pulford CV, Predeus AV, Low R, Webster H, et al. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. Genome Biol. 2021 Dec 21;22(1):349.

294. Page AJ, Mather AE, Le-Viet T, Meader EJ, Alikhan NF, Kay GL, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. Microb Genomics. 2021 Jun;7(6).

295. Banham Poultry: Covid outbreak factory loses £4m worth of chicken. BBC News [Internet]. 2020 Sep 23 [cited 2022 Aug 15]; Available from: https://www.bbc.com/news/uk-england-norfolk-54273106

296. Koropatkin NM, Cameron EA, Martens EC. How glycan metabolism shapes the human gut microbiota. Nat Rev Microbiol. 2012 May;10(5):323–35.

297. Ravi A, Troncoso-Rey P, Ahn-Jarvis J, Corbin KR, Harris S, Harris H, et al. Linking carbohydrate structure with function in the human gut microbiome using hybrid metagenome assemblies [Internet]. bioRxiv; 2021 [cited 2022 Aug 16]. p. 2021.05.11.441322. Available from: https://www.biorxiv.org/content/10.1101/2021.05.11.441322v2

298. Zhang Y, Chen C, Song Y, Zhu S, Wang D, Zhang H, et al. Excretion of SARS-CoV-2 through faecal specimens. Emerg Microbes Infect. 2020 Dec;9(1):2501–8.

299. Li T, Garcia-Gutierrez E, Yara DA, Scadden J, Davies J, Hutchins C, et al. An optimised protocol for detection of SARS-CoV-2 in stool. BMC Microbiol. 2021 Sep 6;21:242.

300. Cusini A, Rampini SK, Bansal V, Ledergerber B, Kuster SP, Ruef C, et al. Different Patterns of Inappropriate Antimicrobial Use in Surgical and Medical Units at a Tertiary Care Hospital in Switzerland: A Prevalence Survey. PLOS ONE. 2010 Nov 16;5(11):e14011.

301. Summary of the evidence | Pneumonia (community-acquired): antimicrobial prescribing | Guidance | NICE [Internet]. NICE; [cited 2022 Aug 6]. Available from: https://www.nice.org.uk/guidance/ng138/chapter/summary-of-the-evidence

302. Hawkins RC. Laboratory Turnaround Time. Clin Biochem Rev. 2007 Nov;28(4):179–94.

303. 60-minute diagnostic tests tackling antibiotic resistance to be eligible for £8m Longitude Prize [Internet]. Longitude Prize. 2022 [cited 2022 Aug 6]. Available from: https://longitudeprize.org/press-release/60-minute-diagnostic-tests-tackling-antibiotic-resistance-to-be-eligible-for-8m-longitude-prize/

304. Heravi FS, Zakrzewski M, Vickery K, Hu H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J Microbiol Methods. 2020 Mar 1;170:105856.

305. Ahannach S, Delanghe L, Spacova I, Wittouck S, Van Beeck W, De Boeck I, et al. Microbial enrichment and storage for metagenomics of vaginal, skin, and saliva samples. iScience. 2021 Nov 19;24(11):103306.

306. Chen J, Sun L, Liu X, Yu Q, Qin K, Cao X, et al. Metagenomic Assessment of the Pathogenic Risk of Microorganisms in Sputum of Postoperative Patients With Pulmonary Infection. Front Cell Infect Microbiol [Internet]. 2022 [cited 2022 Aug 11];12. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2022.855839

307. Yager TD, McMurray CT, van Holde KE. Salt-induced release of DNA from nucleosome core particles. Biochemistry. 1989 Mar 7;28(5):2271–81.

308. Elkins MR, Bye PT. Mechanisms and applications of hypertonic saline. J R Soc Med. 2011 Jul;104(Suppl 1):S2–5.

309. Martner A, Skovbjerg S, Paton JC, Wold AE. Streptococcus pneumoniae Autolysis Prevents Phagocytosis and Production of Phagocyte-Activating Cytokines. Infect Immun. 2009 Sep;77(9):3826–37.

310. Pendleton KM, Erb-Downward JR, Bao Y, Branton WR, Falkowski NR, Newton DW, et al. Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. Am J Respir Crit Care Med. 2017 Dec 15;196(12):1610–2.

311. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Turnbaugh P, Garrett WS, Turnbaugh P, Quince C, Gibbons S, editors. eLife. 2019 Sep 10;8:e46923.

312. Emonet S, Lazarevic V, Pugin J, Schrenzel J, Ruppé E. Clinical Metagenomics for the Diagnosis of Hospital-acquired Infections: Promises and Hurdles. Am J Respir Crit Care Med. 2017 Dec 15;196(12):1617–8.

313. Rapid Pathogen Detection by Metagenomic Next-Generation Sequencing of Infected Body Fluids - PMC [Internet]. [cited 2022 Aug 6]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9020267/

314. Silverman JD, Bloom RJ, Jiang S, Durand HK, Dallow E, Mukherjee S, et al. Measuring and mitigating PCR bias in microbiota datasets. PLOS Comput Biol. 2021 Jul 6;17(7):e1009113.

315. Potapov V, Ong JL. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS ONE. 2017 Jan 6;12(1):e0169774.

316. Schrader C, Schielke A, Ellerbroek L, Johne R. PCR inhibitors – occurrence, properties and removal. J Appl Microbiol. 2012;113(5):1014–26.

317. Product comparison [Internet]. Oxford Nanopore Technologies. [cited 2021 Jul 6]. Available from: http://nanoporetech.com/products/comparison

318. Krause M. Flongle DirectRNA Library preparation [Internet]. protocols.io. 2020 [cited 2022 Aug 7]. Available from: https://www.protocols.io/view/flongle-directrna-library-preparation-bcwcixaw

319. Community [Internet]. [cited 2022 Aug 7]. Available from: https://community.nanoporetech.com/posts/flongle-september-update

320. Sanderson ND, Swann J, Barker L, Kavanagh J, Hoosdally S, Crook D, et al. High precision Neisseria gonorrhoeae variant and antimicrobial resistance calling from metagenomic Nanopore sequencing. Genome Res. 2020 Jan 9;30(9):1354–63.

321. UK SMI ID 7: identification of Staphylococcus species, Micrococcus species and Rothia species [Internet]. GOV.UK. [cited 2022 Aug 11]. Available from: https://www.gov.uk/government/publications/smi-id-7-identification-of-staphylococcus-species-micrococcus-species-and-rothia-species

322. Yarbrough ML, Kwon JH, Wallace MA, Hink T, Shupe A, Fraser VJ, et al. Frequency of Instrument, Environment, and Laboratory Technologist Contamination during Routine Diagnostic Testing of Infectious Specimens. J Clin Microbiol. 2018 May 25;56(6):e00225-18.

323. Farnsworth CW, Wallace MA, Liu A, Gronowski AM, Burnham CAD, Yarbrough ML. Evaluation of the Risk of Laboratory Microbial Contamination during Routine Testing in Automated Clinical Chemistry and Microbiology Laboratories. Clin Chem. 2020 Sep 1;66(9):1190–9.

324. Brugger SD, Bomar L, Lemon KP. Commensal–Pathogen Interactions along the Human Nasal Passages. PLOS Pathog. 2016 Jul 7;12(7):e1005633.

325. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci. 2018 Dec 26;115(52):E12353–62.

326. Ferreira-Coimbra J, Sarda C, Rello J. Burden of Community-Acquired Pneumonia and Unmet Clinical Needs. Adv Ther. 2020;37(4):1302–18.

327. Blaschke AJ. Interpreting Assays for the Detection of Streptococcus pneumoniae. Clin Infect Dis. 2011 May 1;52(suppl_4):S331–7.

328. UCL. Inhale Project [Internet]. Inhale Project. 2019 [cited 2022 Aug 16]. Available from: https://www.ucl.ac.uk/inhale-project/inhale-project

329. Gonzales-Siles L, Karlsson R, Schmidt P, Salvà-Serra F, Jaén-Luchoro D, Skovbjerg S, et al. A Pangenome Approach for Discerning Species-Unique Gene Markers for Identifications of Streptococcus pneumoniae and Streptococcus pseudopneumoniae. Front Cell Infect Microbiol [Internet]. 2020 [cited 2022 Aug 11];10. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2020.00222

330. Kilian M, Riley DR, Jensen A, Brüggemann H, Tettelin H. Parallel Evolution of Streptococcus pneumoniae and Streptococcus mitis to Pathogenic and Mutualistic Lifestyles. mBio. 2014 Jul 22;5(4):e01490-14.

331. MRSA, MSSA and Gram-negative bacteraemia and CDI: annual report [Internet]. GOV.UK. [cited 2022 Aug 12]. Available from: https://www.gov.uk/government/statistics/mrsa-mssa-and-e-coli-bacteraemia-and-c-difficile-infection-annual-epidemiological-commentary

332. Trepanier P, Mallard K, Meunier D, Pike R, Brown D, Ashby JP, et al. Carbapenemase-producing Enterobacteriaceae in the UK: a national study (EuSCAPE-UK) on prevalence, incidence, laboratory detection methods and infection control measures. J Antimicrob Chemother. 2017 Feb 1;72(2):596–603.

333. Olsen JE, Christensen H, Aarestrup FM. Diversity and evolution of blaZ from Staphylococcus aureus and coagulase-negative staphylococci. J Antimicrob Chemother. 2006 Mar 1;57(3):450–60.

334. DiPersio LP, DiPersio JR, Frey KC, Beach JA. Prevalence of the erm(T) Gene in Clinical Isolates of Erythromycin-Resistant Group D Streptococcus and Enterococcus. Antimicrob Agents Chemother. 2008 Apr;52(4):1567–9.

335. Goldstein EJC, Murphy TF, Parameswaran GI. Moraxella catarrhalis, a Human Respiratory Tract Pathogen. Clin Infect Dis. 2009 Jul 1;49(1):124–31.

336. Wani AK, Roy P, Kumar V, Mir T ul G. Metagenomics and artificial intelligence in the context of human health. Infect Genet Evol. 2022 Jun 1;100:105267.

337. D'Mello A, Riegler AN, Martínez E, Beno SM, Ricketts TD, Foxman EF, et al. An in vivo atlas of host–pathogen transcriptomes during Streptococcus pneumoniae colonization and disease. Proc Natl Acad Sci. 2020 Dec 29;117(52):33507–18.

338. Unyvero A50 Molecular Diagnostic Platform Curetis [Internet]. Curetis. [cited 2022 Aug 8]. Available from: https://curetis.com/products/unyvero-a50-system/

339. Kralik P, Ricchi M. A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. Front Microbiol [Internet]. 2017 [cited 2021 Jul 6];8. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.00108/full

340. de Nies L, Lopes S, Busi SB, Galata V, Heintz-Buschart A, Laczny CC, et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. Microbiome. 2021 Feb 17;9(1):49.

341. Jennings LC, Anderson TP, Beynon KA, Chua A, Laing RTR, Werno AM, et al. Incidence and characteristics of viral community-acquired pneumonia in adults. Thorax. 2008 Jan 1;63(1):42–8.

342. Rouzé A, Nseir S. Hospital-Acquired Pneumonia/Ventilator-Associated Pneumonia and Ventilator-Associated Tracheobronchitis in COVID-19. Semin Respir Crit Care Med. 2022 Apr;43(2):243–7.

343. Claro IM, Ramundo MS, Coletti TM, Silva CAM da, Valenca IN, Candido DS, et al. Rapid viral metagenomics using SMART-9N amplification and nanopore sequencing [Internet]. Wellcome Open Research; 2021 [cited 2022 Aug 12]. Available from: https://wellcomeopenresearch.org/articles/6-241

344. Greninger AL. A decade of RNA virus metagenomics is (not) enough. Virus Res. 2018 Jan 15;244:218–29.

345. Kesimer M, Scull M, Brighton B, DeMaria G, Burns K, O'Neal W, et al. Characterization of exosome-like vesicles released from human tracheobronchial ciliated epithelium: a possible role in innate defense. FASEB J. 2009 Jun;23(6):1858–68.

346. Elzanowska J, Semira C, Costa-Silva B. DNA in extracellular vesicles: biological and clinical aspects. Mol Oncol. 2021 Jun;15(6):1701–14.

347. Viral Envelope - an overview | ScienceDirect Topics [Internet]. [cited 2022 Aug 15]. Available from: https://www.sciencedirect.com/topics/immunology-and-microbiology/viral-envelope

348. History of COG-UK | COVID-19 Genomics UK Consortium [Internet]. COVID-19 Genomics UK Consortium | UK-Wide Genomic Sequencing. 2021 [cited 2022 Aug 15]. Available from: https://www.cogconsortium.uk/about/about-us/history-of-cog-uk/

349. Raghwani J, du Plessis L, McCrone JT, Hill SC, Parag KV, Thézé J, et al. Genomic Epidemiology of Early SARS-CoV-2 Transmission Dynamics, Gujarat, India. Emerg Infect Dis. 2022 Apr;28(4):751–8.

350. Aggarwal D, Warne B, Jahun AS, Hamilton WL, Fieldman T, du Plessis L, et al. Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission. Nat Commun. 2022 Feb 8;13(1):751.

351. Aggarwal D, Page AJ, Schaefer U, Savva GM, Myers R, Volz E, et al. Genomic assessment of quarantine measures to prevent SARS-CoV-2 importation and transmission. Nat Commun. 2022 Feb 23;13(1):1012.

352. Siddle KJ, Krasilnikova LA, Moreno GK, Schaffner SF, Vostok J, Fitzgerald NA, et al. Transmission from vaccinated individuals in a large SARS-CoV-2 Delta variant outbreak. Cell. 2022 Feb 3;185(3):485-492.e10.

353. Aggarwal D, Myers R, Hamilton WL, Bharucha T, Tumelty NM, Brown CS, et al. The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities. Lancet Microbe. 2022 Feb 1;3(2):e151–8.

354. Tonkin-Hill G, Martincorena I, Amato R, Lawson AR, Gerstung M, Johnston I, et al. Patterns of within-host genetic diversity in SARS-CoV-2. Neher RA, Sawyer SL, Neher RA, Lauring AS, editors. eLife. 2021 Aug 13;10:e66857.

355. GISAID. GISAID Tracking of Variants [Internet]. [cited 2022 Aug 15]. Available from: https://gisaid.org/hcov19-variants/

356. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern [Internet]. [cited 2022 Aug 15]. Available from: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern

357. Amman F, Markt R, Endler L, Hupfauf S, Agerer B, Schedl A, et al. Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale. Nat Biotechnol. 2022 Jul 18;1–9.

358. Genome sequencing SARS-CoV-2 plays a critical role in informing national and international COVID-19 public health responses [Internet]. Quadram Institute. [cited 2022 Aug 12]. Available from: https://quadram.ac.uk/case_studies/genome-sequencing-sars-cov-2-plays-a-critical-role-in-informing-national-and-international-covid-19-public-health-responses/