Development and application of clinical metagenomics for the diagnosis and characterisation of lower respiratory infections

Themoula Charalampous

Norwich Medical School

Faculty of Medicine and Health Sciences

University of East Anglia, Norwich, UK

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy, September 2020

Abstract

Lower respiratory tract infections (LRTIs) are the largest infectious cause of death globally according to the WHO. Rapid diagnosis and appropriate management are key to control. The gold standard for diagnosis, microbiological culture, is too slow to provide clinicians with the necessary information to treat patients appropriately and reduce morbidity and mortality.

In recent years, there has been growing interest in the application of clinical metagenomics (CMg) for the characterisation of pathogens in clinical samples. CMg has the potential to be faster and more comprehensive than culture, capable of detecting any pathogen (bacteria, viruses and fungi) in a single test within hours. This would transform the field of clinical microbiology, ensuring patients received the appropriate antibiotics sooner, while concurrently providing pathogen genomic surveillance data. However, the development and implementation of rapid CMg has been challenging, mainly due to the high ratio of human:pathogen DNA present in clinical samples, resulting in high sequencing cost and long turnaround times. In this study, we developed and optimised a rapid CMg pipeline for the diagnosis and characterisation of LRTIs that overcomes these challenges.

The pipeline includes: a simple and highly efficient saponin-based host DNA depletion step, automated microbial DNA extraction, rapid library preparation, low-input nanopore sequencing and real-time identification of microorganisms and resistance genes. The pipeline was developed, tested and optimised using excess respiratory samples from suspected LRTI patients. The optimised pipeline was then evaluated in a clinical trial comparing three technologies for the rapid diagnosis of hospital acquired (HAP) and ventilator associated (VAP) pneumonia (the INHALE trial). The pipeline was also evaluated for the rapid characterisation of *Legionella* spp. in respiratory samples for outbreak investigation applications in collaboration with Public Health England.

The developed CMg test had a turnaround time of six hours for the identification of bacterial pathogens and resistance genes with high specificity and sensitivity. A reduction in sensitivity was observed when applying the method for the detection of pathogens in suspected HAP/VAP patients and for samples containing *Legionella* spp. Reduced performance was related to the difference in how respiratory samples from the intensive care unit are cultured and testing of old freeze thawed samples, respectively.

CMg demonstrates great potential for replacing culture for the diagnosis of LRTIs, however, further optimisation is required to enable concurrent detection of viruses and improved automation is required to allow successful clinical implementation.

Table of Contents

Acknowledgements

Firstly, I would like to thank my supervisory team, Dr. Justin O’Grady, Dr. Richard Leggett, Dr. Daniel Turner and Prof. John Wain for their support and help. I would also like to thank the University of East Anglia and Oxford Nanopore Technologies for funding my PhD studentship.

I am especially grateful to Dr. O’Grady for the constant help, teaching and guidance provided throughout my PhD studies. Thank you very much, for being a great teacher and for providing me with the knowledge and skills that allowed me to complete my studies but also start my scientific career.

I would also like to especially thank Dr. Gemma Kay, who was always there to support and help me numerous times throughout this journey both personally and professionally and for also taking the time on reading/correcting my thesis. Additionally, I would like to thank Dr. Leggett for reading and improving the bioinformatics section of my thesis.

I would also like to acknowledge, Sara Grundy from the clinical microbiology laboratory Norwich and Norfolk for her constant help with sample collection. Dr. Meera Chand and Jessica Day from Public Health England for their help with providing samples and help for the *Legionella* spp. study.

Last but not least I would like to give a big thank you to my my mum, dad, sister and my boyfriend for their constant support and love. I would have never been able to finish my studies without you.

List of Figures

1. Introduction

1.1 Overview of Lower Respiratory Tract Infections

1.2 Next Generation Sequencing

3. Results

3.1 Development, optimization and testing of a clinical metagenomics pipeline with a host depletion method for the diagnosis of LRTIs

D, *mea* gene alignment of depleted versus undepleted during 2 h of sequencing. E, *E. coli* genome coverage of depleted versus undepleted during 2 h of sequencing. F, *bla*$_{\mathrm{TEM}}$, *sul1* and *dfr*A17 gene alignment of depleted versus undepleted during 2 h of sequencing. In C-F, three independent clinical samples were analysed (examples of a Gram-positive and a Gram-negative are represented)

List of Tables

1. Introduction

1.1. Overview of Lower Respiratory Tract Infections

1.4 Microbial bioinformatics

2. Methodology

2.3. Microbiological investigation of respiratory samples

3.2 Implementation of the clinical metagenomics pipeline

3.3 Application of clinical metagenomics for public health

1. Introduction

1.1 Overview of Lower Respiratory Tract Infections

Lower respiratory infections (LRTIs) are the cause of three million deaths worldwide and were recently characterized by WHO as the deadliest communicable disease (https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death). Infections of the upper respiratory tract affect the nose, sinuses, pharynx and larynx. LRTIs, on the other hand, are infections of the trachea, bronchi, and lung parenchyma. A uniform definition of LRTI does not exist, however from an epidemiological point of view, the most important LRTI infections are bronchitis, bronchiolitis and pneumonia (1-3).

Pneumonia is then subdivided into community-acquired pneumonia (CAP) and nosocomial pneumonia (4, 5). CAP is considered the most common type of pneumonia and develops in individuals who have not previously been hospitalized or were in healthcare facilities and initial diagnosis is based on LRTI-related symptoms (cough, chest pain etc.) along with high fever (>38ºC) (2).

Nosocomial pneumonia includes hospital-acquired pneumonia (HAP) and ventilator-associated pneumonia (VAP) (6). HAP usually develops after 48-72 hours from initial hospital admission and VAP (the least common type of nosocomial pneumonia but with the highest mortality rate) is generally associated with the intensive care unit (ICU) (7). VAP develops 48-72 hours after patient intubation (early-onset is often defined as after two days of intubation and late-onset VAP after six) (8).

Previously published guidelines (in both the EU and USA) have introduced the term ‘health care-associated pneumonia’ (HCAP) in order to describe similar but different infections to the ones usually observed in community-related LRTIs. However, the most recent published guidelines do not consider this term as clinically relevant, hence HCAP will not be used in this study (5, 7).

Atypical pneumonias also fall under the umbrella of LRTIs. This term was initially used to described viral CAPs but recently has been used to describe LRTIs caused by certain 'not-as-common' respiratory pathogens (9).

1.1.1 Epidemiology and aetiology of LRTIs

LRTIs can be caused by various pathogenic agents, bacterial and/or viral as reported by many studies (2, 10, 11). Etiology also differs amongst the different types of LRTIs. Bronchitis and bronchiolitis (commonly observed in children) are mainly caused by viral agents such as Influenza virus and respiratory syncytial virus (RSV) respectively. Infectious agents are also dependent on host's factors (e.g. age, patient's immune system condition). For example, RSV has also been associated with bronchitis cases in elderly patients (6, 12).

The main causative agent of CAP is *Streptococcus pneumoniae* (pneumococcus) and this has been proven by numerous epidemiological studies carried out in the UK, Europe and the USA (1, 8, 13-15). This is followed by the Gram-negative bacteria and viruses (see Figure 1.1), although with the increase in sensitivity of diagnostic tests more viral CAP cases have been reported (2) including influenza, rhinovirus and human coronavirus. Viral aetiologies are mainly associated with pediatric cases (>75% vs 25% adults CAP cases). Conversely, *S. pneumoniae* infections are mostly associated with older patients (5, 13).

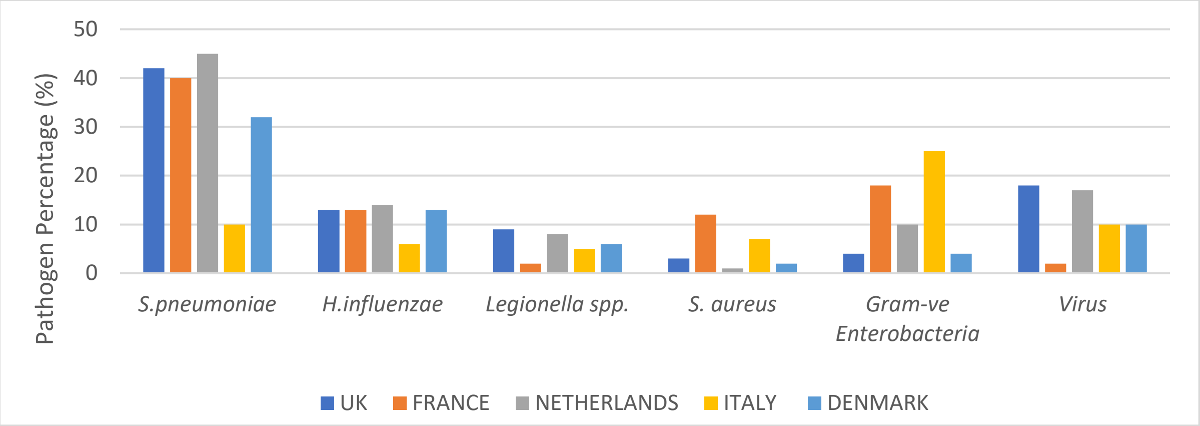


Figure 1.1: Main causative agents of community-acquired pneumonia observed in selected European countries (12).

The main aetiologies of HAP are aerobic Gram-negative bacilli (such as *Escherichia coli* and *Klebsiella* spp.), *Pseudomonas aeruginosa*, *Acinetobacter* spp. and Gram-positive cocci such as *Staphylococcus aureus* (5, 16, 17). Early-onset VAP has been associated with community-related pathogens such as *S. aureus* and *Haemophilus influenzae*. Late-onset VAP conversely is associated more, with polymicrobial infections (18, 19) and with multi-drug resistant (MDR) HAP pathogens such as methicillin-resistant *S. aureus* (MRSA) and extended spectrum beta-lactamase (ESBL) producing Gram-negative bacteria (*Enterobacter* spp., *Klebsiella* spp.). These pathogens in addition to *Enterococcus faecium*, are known as the ESKAPE pathogens (20). ESKAPE pathogens are responsible for 80% of all VAP and HAP cases (20). The SENTRY Antimicrobial Surveillance Program (21) also reported similar findings, with *S. aureus* being the main HAP and VAP bacterial pathogen reported in the United States (36%) and Europe (23%) followed by *P. aeruginosa* (19.75% cases in the United States and 20.8% in Europe) and the other members of the ESKAPE pathogens (22) (Figure 1.2). Other Gram-negative bacteria such as *Stenotrophomonas maltophilia and Moraxella* spp. have also been reported less as cases of HAP/VAP (10, 22).

Other non-bacterial pathogens of VAP and HAP include respiratory viruses such as, influenza (A and B) and parainfluenza (in adult patients), cytomegalovirus (CMV) (reported in ICU patients (23), and mimivirus (associated with prolonged ventilated ICU patients (24-26)) and fungal organisms such as *Aspergillus* and *Candida* spp. It is thought that *Candida* colonization in pneumonia patients suggests the existence of an underlying bacterial infection rather than the cause of the pneumonia. *Candida* is also associated with higher mortality rates and poor patient outcomes (27, 28). A small number of aspergillosis cases (3%) have been reported in late-onset VAP. Aspergillosis (caused by *Aspergillus* spp.) is mostly associated with severely ill patients (29, 30) (Figure 1.2).

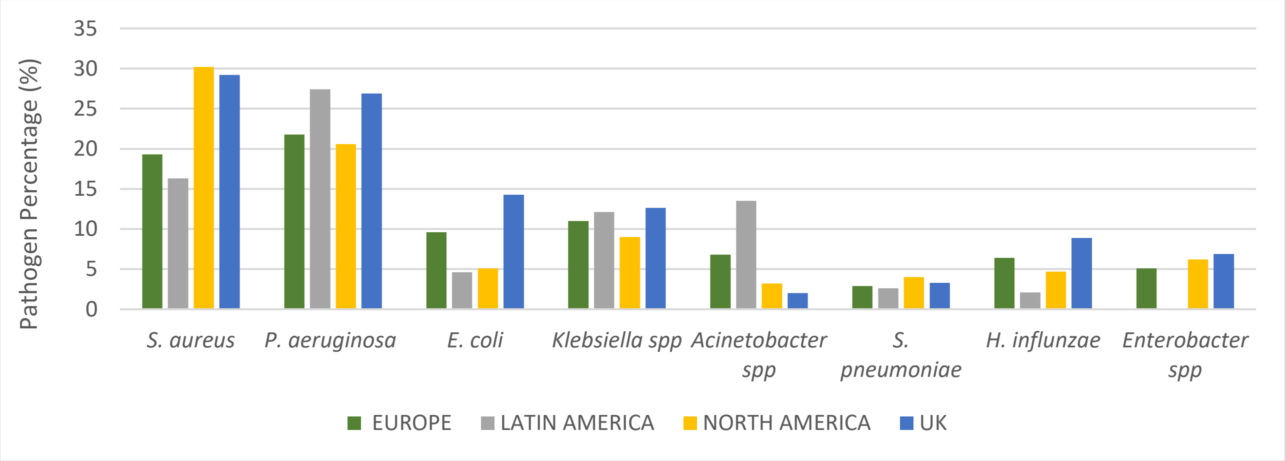


Figure 1.2. Overall frequency of isolated pathogens from patients with nosocomial pneumonia in different geographical regions (22).

Atypical pneumonia is mainly caused by bacterial organisms that cannot be identified using the standard methods (e.g. culture or Gram stain). Most cases of atypical pneumonia have been caused by *Mycoplasma pneumoniae*, *Chlamydophila pneumoniae* and *Legionella pneumophila* (referred to as Legionnaire's disease or legionellosis) (31).

Legionella spp. reside in aquatic habitats and water distribution systems hence, the main transmission source is through inhaling contaminated aerosols (32). Many water distribution systems, such as hot tubs, industrial/domestic plumbing systems and cooling towers, are contaminated with *Legionella* spp. and have been connected to legionellosis outbreaks. Legionellosis or Legionnaires' disease is also associated with <7% of severe CAP (sCAP) cases (33, 34) and 2-9% CAP cases. The European Legionnaires' Disease Surveillance Network, between 2011 and 2015, reported 30,532 cases (of which 92.3% were confirmed) of Legionnaires' disease from which: 67% were community-acquired associated with reported outbreaks such as the 2014 reported in Portugal and the 2012 outbreak reported in Edinburgh, 24% were travel-associated; and 7% were related with health-care (32, 35). *Legionella pneumophila* is the most commonly detected species of the *Legionella* genus, with sequence types (ST) belonging to Lp1 serogroup associated with the majority of legionellosis cases in the US (35), in England and Wales (36).

Legionnaires disease is less frequently caused by *Legionella longbeachae* (associated with contaminated soil) and other *Legionella* spp. such as *Legionella micdadei* and *Legionella bozemanae* have been associated with infections in immunocompromised patients (32).

1.1.2 Clinical features and diagnosis of LRTIs

Patients with bronchitis or bronchiolitis are typically presented with a cough and this is the main symptom observed in healthy adults. Cough can be persistent for a few weeks (depending on the infectious agent) and can be productive/non-productive with the production of mucoid sputum (1, 4). In the majority of cases bronchiolitis is viral and is self-limiting and initial antibiotic treatment is not necessary. However, the production of purulent sputum is often associated with bacterial infections. Additionally, if a clinician suspects that bronchitis has progressed to pneumonia, a chest radiograph will be requested to exclude the presence or absence of a pulmonary infiltrate (PI). However, in some paediatric cases, when the clinical features are more severe, hospital admission and ventilation might be necessary (1).

The main clinical features of CAP, HAP and VAP, are cough, dyspnoea/tachypnoea, increased pulse (>100), persistent fever (> 4 days) and the presence of a pulmonary infiltrate (5). A pulmonary infiltrate can only be confirmed with a chest radiograph and according to the guidelines published in 2011 by ESCMID, a chest radiograph should always be performed in suspected cases of pneumonia (4, 5, 16). For HAP and VAP cases additional physical examination such as lateral or posterior chest radiograph can be requested to determine the progress of the infection (29). The infection's progression is also monitored by routine blood counts and chemistry (6, 15).

In addition to these tests for patients with suspected pneumonia, a C-reactive protein (CRP) test can be performed. Pneumonia is likely present if CRP level= >100 mg/L and is highly unlikely if CRP levels = <20 mg/L (5). In addition to the CRP test, erythrocyte sedimentation rate (ESR) can be performed to determine inflammation levels – this test measures the rate red blood cells (RBC) settle in a test tube (higher rate suggests presence of inflammation). However, the diagnostic value of such tests is debated amongst clinicians (5, 37). The diagnostic value of the CRP test is also debated - accepted by some studies (38, 39) and rejected by others (40).

The main symptoms for atypical pneumonia such as Legionnaire's disease include headache, low-grade fever, cough, chills and malaise, meaning it can be easily mistaken as pneumococcal pneumonia due to symptoms resembling mild CAP (41). Also, in patients with suspected legionellosis and pneumonia (confirmed by a chest radiograph), neurological abnormalities and gastrointestinal manifestations (such as diarrhoea and vomiting) caused by persistent headaches are observed. Other less common symptoms of legionellosis are myalgia or arthralgia and chest pain (observed in <50% of patients) (38-40).

If the clinician suspects infection, they will request microbiological tests (e.g. culture, gram stains) to identify the causative agent (described in 1.1.4). The typical respiratory sample used for testing is expectorated sputum, as it is easily collected. The use of expectorated sputum is recommended by the UK and European guidelines) for culture testing and investigation sensitivity (5, 42). However, in some cases, such as previous treatment failure or if the patient cannot produce sputum, an alternative is required. Bronchoalveolar lavage (BAL), endotracheal aspirates (ETA) and protected specimen brush (PSB) can be collected instead. These are considered 'cleaner' samples, as they are collected invasively, meaning contamination from the upper respiratory microbiota is minimised. Invasive techniques are used only is certain cases, such as in unresolved pneumonia cases or for severely ill patients as invasive techniques require expertise and are costly (5).

1.1.3 Empiric antibiotic treatment

The European Respiratory Society and European Society for Clinical Microbiology and Infectious Diseases (ESCMID) has published guidelines to be followed by clinical routine practices for the management of LRTIs (5). The guidelines recommend for all patients with confirmed pneumonia (by x-ray) should be given empirical antibiotic therapy (5). The chosen treatment is determined by the patient's history and risk factors (5, 16) (Table 1.1). The UK guidelines according to the 'start smart and then focus' report published by PHE for antimicrobial stewardship, recommends the start of empirical antibiotic treatment as soon as possible only in patients with sepsis or severe infections (43). Microbiological cultures should be available prior to initiating empirical therapy when possible (in cases where patients will not be at risk).

Table 1.1: Recommended criteria for determining empiric antibiotic therapy according to NICE guidelines (42).

Criteria	**Additional information**
Severity of patient	For severely-ill or pregnant patients, advice from a specialist is recommended for choosing empirical antibiotic therapy
Co-morbidity	A co-morbidity may be related with the causative agent
Residence	Similar microbial patterns are observed by patients residing at the same nursing homes
Patient's Infection History	Patient could have been previously infected with resistant organisms
Microbial and resistance patterns locally and regionally	Allows a possible prediction of microbial aetiology and (if any) resistance
Toxicity of antimicrobial agents	Assessed for each patient individually
Risk factors for an immunosuppressed system	Immunosuppressed patients are at a higher risk of infections by opportunistic pathogens

ESCMID guidelines, in cases of CAP, recommend for penicillin (amoxicillin) or tetracyclines as the first-choice broad spectrum antibiotics. In cases of penicillin allergy, macrolides are recommended as a good alternative if macrolide-resistance is low in the country/region. If there is high resistance to first choice antibiotics then quinolones such as levofloxacin or moxifloxacin are recommended (44-46). For severe cases of hospitalised patients with CAP (sCAP) the combination of macrolides with beta-lactams (or with antipseudomonal cephalosporins only if a *Pseudomonas* infection is high) is recommended. If the risk of infection by ESBLs Gram negative enterobacteria is high then ertapenem is recommended, only if the risk of a *Pseudomonas* infection is low/excluded (47-49). These recommendations are also in agreement with the UK guidelines (published by the National Institute of Health Care and Excellence - NICE), where the use of amoxicillin or tetracyclines or macrolides are recommended for the empirical treatment for suspected CAP cases (42). Although for severe cases of pneumonia, quinolones (levofloxacin) are recommended as the initial treatment (42). In CAP cases when *Legionella* infection is suspected then the recommended first line of antibiotic treatment is quinolones (levofloxacin) or macrolides (50, 51).

Treatment of nosocomial pneumonia is determined based on the risk of multi-drug resistant (MDR) bacterial infection and if it is a late-onset (i.e. VAP) case according to the NICE guidelines (42). If the risk is high or it is a late-onset VAP case then the broad-spectrum antibiotics of choice are cephalosporins (ceftazidime for Gram-negative bacteria) or beta-lactamase inhibitors with penicillins (piperacillin with tazobactam) (active against both Gram-negative including *P. aeruginosa* and Gram-positive bacteria (52)). For MRSA suspected infections, dual therapy is recommended and the antibiotics of choice are glycopeptides (teicoplanin or vancomycin) and oxazolidinones (linezolid) (42). Antibiotics in these cases should be administered intravenously instead (42). However, if the patient is at a low risk for

MDR bacterial infection or it is an early-onset VAP case, antibiotics should be given orally. Penicillin (co-amoxiclav) is the first-choice of antibiotics recommended, followed by tetracyclines (doxycycline), or sulfonamides (co-trimoxazole) by the NICE guidelines. Quinolones (levofloxacin) are only recommended if the previous options are not suitable (42).

Recent guidelines published by both the American Thorasic Society and IDSA (Infectious Diseases Society of America) also focus on the need for targeted antibiotic therapy and the limiting of empiric antibiotic treatment (53). The consequences of inappropriate empirical antibiotic therapies have been clearly highlighted throughout the years (54-56). Overuse of broad-spectrum antibiotics has resulted in increased levels of antimicrobial resistance and consequently higher mortality rates, especially in patients with nosocomial pneumonia (54-58). Alvarez-Lerma *et al.* showed that, in 214 (43.7%) of 490 VAP cases no improvement was observed with initial therapy and a change in antibiotic treatment was necessary (59). Other studies have observed a reduction in inappropriate initial therapy and drug-related costs when utilising computer algorithms based on epidemiological data (such as local antibiotic resistance patterns) and patient-information (i.e. the patient’s history and microbiology results) gathered from the relevant local microbiology labs and ICU (60, 61). Using a targeted antibiotic treatment is not only beneficial for the patient’s health/outcome but also reduces hospital and drug related costs. However, a targeted antibiotic treatment can only be chosen once the aetiology is known.

The antimicrobial stewardship treatment (AMS) algorithm by PHE, states that reviewing antibiotic prescription can only happen after 48-72 hrs due to the current choice of diagnostic method for LRTIs (43). At the moment in the UK, testing is based on traditional microbiological culture, meaning results are only available after 48-96 hours or longer (62).

1.1.4 Microbiological diagnosis and targeted antibiotic treatment

The current ‘gold standard’ method for bacterial and fungal identification from respiratory samples (sputum and tracheal aspirates) is semi-quantitative culture (42). Other microbiological tests include pneumococcal and legionella antigen detection tests (63, 64) (see table 1.2 of all diagnostic approaches). Expectorated sputum is the most common respiratory sample collected, despite the fact sputum cultures have low sensitivity and specificity due to the carry-over of contaminants from the upper respiratory tract (URT) (63). Hence, it is recommended, to collect sputum samples early in the morning, which are considered to contain pooled secretions from the lower tract concentrated overnight and are more likely to contain pathogenic agents(65).

Sputum and mucoid samples are firstly sputasol-treated (1:1 ratio of sample to sputasol added) and plated (previously sample is diluted with water) on blood agar, chocolate agar, and MacConkey agar (media inhibiting the growth of Gram-positive bacteria). For sterile samples (e.g. BALs) and samples from ICU and/or immunocompromised patients, 10 µL of undiluted primary sample is also plated on Cysteine–lactose–electrolyte-deficient (CLED) agar and sabouraud dextrose agar (SAD) agar (65). For *Legionella*-suspected samples, 100 µL of undiluted sputasol-treated primary sample is plated on *Legionella*-selective media, buffered charcoal yeast extract (BCYE) supplemented with anisomycin or cefamandole, which restricts the growth of lung microbiota and promotes the growth of *Legionella* spp. (65). Samples from patients with suspected *Legionell*osis, are plated on Buffered Charcoal Yeast Extract (BCYE), Buffered polymyxin anisomycin (BMPA) and Buffered Charcoal Yeast Extract with Cefamandole (BCY-C) are incubated for maximum of 10 days to enable the growth of *Legionella* spp. After 4 days of incubation the plates are initially examined for growth followed by re-confirmation at 10 days (6, 62, 66-68).

Other microbiological tests such as antigen and serological tests are also performed due to their rapidity and high sensitivity. Antigen tests are mostly used to identify viral aetiologies but are also recommended for bacterial infections such as *L. pneumophila* serogroup 1 using urine samples (63, 69, 70). Also, any Legionella species reported in clinical samples, must be isolated and sent for identification and serogrouping (65). Urinary antigen are also recommended for rapid identification of *S. pneumoniae* (5, 6, 16, 42, 63, 71). Antigen tests (targeting galactomannan glycoprotein) using BALs or serum samples with PCR-based tests are also recommended for screening of patients with high risk of fungal infection (*A. fumigatus*) (65). However, positive results from these tests cannot distinguish active infections (65).

Table 1.2. Diagnostic approaches followed based on suspected LRTI pathogen (65)

Predicted pathogen	**Sample type**	**Rapid test**	**Conventional test**	**Incubation Time**
S. pneumoniae *H. influenzae* *S. aureus* *M. catarrhalis*	Sputum or BAL	Gram stain	Culture on Chocolate agar + Bacitracin disc* or incorporated in the medium	40-48 hrs
S. pneumoniae	Urine	Antigen test	-	-
B. cepacia complex	Sputum or BAL	-	*B. cepacia* selective agar	5 days
Enterobacteria *Pseudomonas*	Sputum or BAL	Gram stain	Culture on CLED or MacConkey agar	40-48 hrs
S. aureus	Sputum or BAL	Gram stain	Culture on Mannitol Salt / Chromogenic Agar	40-48 hrs
Fungal pathogens	Respiratory samples	Antigen**	Culture on Sabouraud agar	Up to 5 days
Legionella spp.	Urine	Antigen test	-	-
	Respiratory samples	PCR test	Culture on Legionella selective agar	Up to 10 days
Viruses	Sputum	Antigen and PCR tests	-	-

*Chocolate agar with bacitracin is used for the isolation of *M. catarrhalis* and *S. pneumoniae*

**Galactomannan antigen testing for *A. fumigatus* using BAL or serum samples

The AMS treatment algorithm published by PHE recommends to stop antibiotic treatment when appropriate - there is no evidence of bacterial infection (negative culture findings) and patient is improving (42, 43). Also, intravenous antibiotics should be changed to oral antibiotics. Antibiotic treatment can also be refined when appropriate, following the microbiological findings and susceptibility profiles, and narrow-spectrum antibiotics should be chosen (43) (see Table 1.3). Targeted therapy can be either combination or monotherapy, although in the UK, monotherapy is recommended and combination therapy should not be used routinely (42, 43). ESCMID guidelines however, recommend combination therapy as the initial targeted treatment in severe cases, followed by monotherapy after 3-5 days and if the patient is improving (Table 1.5) (5, 16). When MDR Gram-negative bacteria (GNB) are identified (including ESBL and resistant *P. aeruginosa*) then carbapenems are recommended for refined treatment but when there is known/past carbapenem resistance then treatment should be refined following susceptibility profiles and national guidelines (72). For example, the use of polymixins (colistin) in combination with aminoglycosides or tigecyline for carbapenamase-resistant GNB is recommended (72).

Table 1.3. Recommended antibiotic treatment based on identified respiratory pathogens and susceptibility/resistance (5, 72)

Identified pathogen	Recommended treatment
Resistant *S. pneumoniae*	Levofloxacin or Vancomycin or Teicoplanin
MSSA	Flucloxacillin or Cephalosporin II or Clindamycin or Levofloxacin
MRSA	Vancomycin or Teicoplanin or Linezolid
Ampicillin-resistant *H. influenzae*	Aminopenicillin and b-lactamase inhibitor or Levofloxacin
Legionella spp.	Macrolide or Levofloxacin
Carbapenem-resistant *A. baumannii*	Ampicillin and sulbactam or Polymixins
Quinolone- or Carbapenem-resistant *P. aeruginosa*	Avibactam and Ceftazimide or Polymixins
MDR-GNB including resistant *P. aeruginosa*	Carbapenems or Ceftazimide/avibactam

Although culture-based tests can successfully isolate pathogens and identify susceptibility profiles, various challenges and difficulties still remain. Several studies have highlighted the problems that arise form culture-based diagnostics (10, 62, 73, 74). Culture has a slow turnaround time (>48 hrs) and has low clinical sensitivity, as >30% of culture tests fail to identify the causative agent (10, 75). Failure in identifying the etiology in hospitalized patients increases to 50% (76) when infection is polymicrobial. Sub-optimal results and slow diagnosis delay the design of tailored treatment which increases the use of broad-spectrum antibiotics, promoting the emergence of antibiotic resistance. In 2017, the WHO released a list of 12 'priority' resistant bacteria that are currently the greatest threat to humanity and seven bacteria on the list are important respiratory pathogens (https://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed). In fact, in the north-western hemisphere, the majority of prescribed antibiotics are for the treatment of

LRTIs (77). For example, in the UK, 60% of all antibiotics prescribed are for respiratory infections (63). Additionally, according to CDC, in the US, 46% prescribed antibiotics are for respiratory illnesses in urgent care centres and 25% in emergency departments (78). These numbers demonstrate the significant overuse of empirical antibiotics in respiratory infection and the likely contribution this has on the emergence of antibiotic resistance (79). Delayed appropriate antibiotic treatment leads to poor patient outcomes, meaning hospitalisation is prolonged therefore increasing hospital-related costs (80). In Europe, the annual hospital-related costs associated with LRTIs were €10.1 billion (80), rising to at least €17 billion in the US (81), with inpatient care accounting for more than half of the costs. The increasing rate of antibiotic resistance in respiratory pathogens and the substantial economic burden of LRTIs demonstrate the need for changes in the diagnostic approach of LRTIs.

1.1.5 Diagnosis using molecular-based techniques

Rapid diagnosis would reduce turnaround time, allowing tailored antibiotic treatment to be started earlier. This will eventually lead to reduction in costs for the NHS, reduced antimicrobial resistance and improved patient outcomes. Molecular methods, as discussed by the UK Government 5-year AMR action plan and the O'Neill report (82-84), have the potential to overcome the limitations of culture-based diagnostics, as pathogens and the associated antibiotic resistance would be identified in a few hours, allowing early targeted therapy and better antibiotic stewardship.

PCR-based techniques offer increased sensitivity and specificity compared to culture and enable easier detection of polymicrobial infections, hard-to-culture bacteria and atypical respiratory pathogens such as *L. pneumophila* (10, 85). Common and atypical respiratory pathogens and some selected resistance genes are the chosen targets for designing these assays (86). DNA extracted directly from primary respiratory samples (BAL, ETA, expectorated sputum) is used as an input for the reaction (87, 88).

Initially PCR-based diagnostic tests could only target one or two pathogens or resistances. Over the years the number of targets have increased massively due to technology advances (such as the use of TaqMan sequence-specific probes) and now multiplex PCR assays can detect up to tens of gene targets (88). However, the use of PCR-based tests for diagnostics has been challenging mainly due to low-quality templates (89). Low-quality templates are consisted of PCR-inhibitory compounds coming from the extraction procedures. However, this challenge was overcome by the development of sample-in answer-out technologies (90, 91), which are rapid, easy-to-use, require minimal sample handling and are not inhibited by contaminants present in the samples.

Numerous diagnostics companies have also developed instruments and assays for pathogen and related antimicrobial resistance genes detection, targeting respiratory infections. Currently there are two PCR-based sample-in answer-out machines that are commercially available and provide panels for the diagnosis of LRTIs that are FDA-approved, the BIOFIRE® FILMARRAY® Pneumonia Panel plus (BioMérieux SA) and the Unyvero P55 Pneumonia Cartridge (Curetis AG). The BIOFIRE FILMARRAY panel allows the semi-quantitative identification of 18 bacteria, nine viruses and seven antibiotic resistance targets (https://www.biomerieux-diagnostics.com/biofire-filmarray-pneumonia-panel.) The Unyvero P55 Cartridge can identify 19 respiratory bacteria and 10 resistances and can provide a semi-quantitative information (91) (see Table 1.4 for all bacteria and resistance gene targets utilised by these two systems).

Table 1.4. Bacterial targets and antimicrobial resistance gene markers used by two commercially-available PCR platforms

PCR platform	Bacterial targets	Antimicrobial resistance markers
BIOFIRE® FILMARRAY®	*Acinetobacter calcoaceticus-baumannii* complex *S. marcescens* *Proteus* spp. *K. pneumoniae* group *E. aerogenes* *Enterobacter cloacae* *E. coli* *H. influenzae* *Moraxella catarrhalis* *P. aeruginosa* *S. aureus* *S. pneumoniae* *Klebsiella oxytoca* *Streptococcus pyogenes* *Streptococcus agalactiae* *L. pneumophila* *Mycoplasma pneumoniae* *Chlamydia pneumoniae*	*mecA/C* and MREJ KPC NDM Oxa48-like CTX-M VIM IMP
UNYVERO P55	*Acinetobacter* spp. *Chlamydophila pneumoniae* *Citrobacter freundii* *E. cloacae* complex *E. coli* *H. influenzae* *K. oxytoca* *K. pneumoniae* *Klebsiella variicola* *L. pneumophila* *Moraxella catarrhalis* *Morganella morganii* *M. pneumoniae* *Proteus* spp. *P. aeruginosa* *Serratia marcescens* *S. aureus* *Stenotrophomonas maltophilia* *Streptococcus pneumoniae*	KPC NDM Oxa-23 Oxa-24/40 Oxa-48 Oxa-58 VIM CTX-M *mecA* TEM

Since their release, many clinical and comparative studies have evaluated these tests to compare their outputs against culture or in-house-developed PCR assays (92, 93). Gadsby *et al.* compared the Unyvero P55 Pneumonia Cartridge (Curetis AG) and an in-house PCR assay against culture for the diagnosis of bacterial respiratory infections. The Unyvero P55 was 56.9% sensitive and 58.5% specific and their in-house PCR assay was 63.2% sensitive and 54.8% specific when compared against culture on 74 BAL samples(91). The authors concluded that the tested assays would only benefit clinical microbiology as an additional test to culture and not as the primary test (91). Another recent study evaluated the BIOFIRE FILMARRAY by testing 1,682 samples (846 BALs and 836 sputa). The test was 100% sensitive for 15/22 pathogenic targets and 10/24 targets in BALs and sputa respectively and was 87.2% specific (94). These two platforms were also compared against routine microbiology in a recent study where 644 surplus respiratory samples from patients with HAP and VAP were tested. Enne *et al.* concluded that the Unyvero 55 had a higher concordance with culture but the FILMARRAY had a higher clinical sensitivity and fewer major discordances with culture (95).

Despite the major advantages PCR assays can offer, such as rapidity and increased sensitivity, challenges, remain. PCR tests are limited on their multiplexing targets, meaning only a number of bacterial targets and resistance genes can be included in an assay. Also, there is a constant need for updating the sequence target of PCR primers to be able to detect newly identified point mutations or resistance genes. Another concern with PCR tests is distinguishing pathogens from lung commensals belonging to the same genus, and therefore, increasing the numbers of false positive samples resulting in the unnecessary treatment of patients (96-99). A example of this, is the commensal bacteria of the *Streptococcus* genus, which can be misclassified as *S. pneumoniae* due to high genetic similarities (99, 100).

An agnostic, non-targeted molecular approach such as metagenomic sequencing-based diagnostics could overcome the many of the challenges observed for culture and PCR, due to its rapidity coupled with comprehensiveness and the unbiased approach of identifying all microbes present in a sample (101, 102).

1.2 Next Generation Sequencing

The incredible advancements in sequencing technology since 1977, when Sanger sequencing was developed, to the beginning of the 21st century when next generation sequencing (NGS) platforms were developed, is unquestionable (103, 104). A good illustration of these advancements are the differences between the first human genome sequenced with the latest sequenced genomes. It took 14 years and $3 billion to sequence the first human genome back in 2004, compared to only ~48 hrs using the NovaSeq (Illumina) in 2018. The MinION (Oxford Nanopore Technologies) was used to sequence the human genome in 2018 which only cost thousands of dollars (105-107).

Sanger sequencing utilises dideoxy nucleotides (dNTPs), which are incorporated at the end of a DNA strand during an amplification cycle, giving a unique pattern, which will then be translated into the DNA sequence. For decades this sequencing technology was the only one available and due to the low error rates and cost is still used today, mainly for specific amplicon confirmation (104, 108). Sanger sequencing is still considered by reference laboratories as the “gold standard” for molecular typing of *L. pneumophila* (Day, J 2019, personal communication October 11). The major advancements in sequencing technology led to the development of NGS and third generation sequencing (refer to reference (109) for a comprehensive review on the history of sequencing). In the past two decades, the NGS industry was dominated by Illumina, Ion Torrent and Pacific Biosciences (PacBio) technologies, but the interest in nanopore sequencing developed by ONT has been growing significantly since its introduction in 2014 (110). Illumina is a sequencing-by-synthesis technology and is based on the incorporation of dNTPs at the last position on the DNA strand during DNA synthesis. Once the incorporated dNTP is determined with specific fluorophore excitation, is then removed enzymatically and DNA synthesis can continue. The current Illumina platforms are the (iSeq, MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq) and enable high-throughput short read (single or paired end) sequencing (50-600bp long) (109). Conversely, the PacBio platforms (RSII and Sequel - average read length of 30kb (111)) and ONT platforms (longest reads >2Mb) allow high-throughput long-read sequencing. For comprehensive description of PacBio sequencing see (109).

1.2.1 Nanopore sequencing

Nanopore sequencing was first introduced in the 1980s, but ever since has been constantly developing. Unlike sequencing-by-synthesis approaches, nanopore-based sequencing directly 'reads' the DNA sequence. Briefly, during nanopore sequencing, a dsDNA fragment is enzymatically unwound and then ssDNA is translocated through a nanopore protein/pore embedded in a membrane. The membrane is immersed in a charged solution, and as the ssDNA passes through the pore, a unique change in the current is created by each of the four bases, which is then used to identify the sequence of the DNA strand (107, 112, 113) (for a comprehensive review on the history of nanopore sequencing see (114)).

In 2014, MinION, the first commercially available nanopore sequencer, was launched by ONT. Initially, MinION sequencing quality and yield restricted its applications and it was mainly used for amplicon and small genome sequencing. Significant improvements in accuracy and yield coupled with the introduction of additional technologies such as PromethION have made it suitable for most sequencing applications and it has even been used to sequence *E. coli* in space

(115). Nanopore sequencing initially, had low sequencing accuracy (~60-70%) due to the pores (R6 followed by R7) and basecalling technology used, when compared to other sequencing technologies e.g. Illumina sequencing is 99% (116). However, since then ONT has managed to improve the single-read error rate to ~10% (116). Improvements to single-read accuracy was mainly due to: i) improving the release of R9 pores, followed by R.9.4 and R9.4.1 pores (117) (Figure 1.3) and ii) the evolution in basecalling softwares from event-based basecalling to raw signal-based basecalling (Albacore v2.0.1) (116, 118).

Additionally, in March 2019 ONT announced the release of the R10 pore which has a longer barrel allowing dual-reading of each nucleotide (Figure 1.3) - this has improved the sequencing of homopolymers and increased consensus accuracy to QS50 vs QS40 with the R9.4.1 chemistry (119). The most recent pore chemistry is R10.3, which is soon-to-be commercially available, and promises similar sequencing yield to the widely used R9.4.1 pore whilst improving accuracy (https://londoncallingconf.co.uk/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store). Nanopore sequencing does not have a read length limitation and can potentially sequence any length DNA molecule presented to it. Hence it is considered a long-read sequencing technology (longest reads >2Mb). Long-read sequencing technology provides advantages over short read technologies.

Whilst short-read sequencing technologies are more accurate and are cost-effective for high-throughput sequencing, long-read sequencing provides advantages that can never be offered by short reads. Long reads can span big regions of a DNA sequence as they can be megabases-long (longest read from nanopore sequencing is > 2Mb). Such long reads provide a number of advantages such as: more accurate genome assemblies and improved mapping confidence (120), identification of the host by mapping flanking regions of a chromosomal resistance gene (121-123) and can be used to identify genome rearrangements (124).

Also, a major advantage of nanopore sequencing (Flongle, MinION, GridION, PromethION) is that data are produced and available for analysis in real-time (101, 110, 125-127) making nanopore sequencing the fastest sequencing technology currently available on the market. This feature has been utilised to provide rapid answers for diagnostic purposes and to characterise outbreaks. For example, Votintseva *et al.* showed *Mycobacterium tuberculosis* could be detected from sputum in 7.5hrs (128), Schmidt *et al.* developed a 4hr-diagnostic pipeline for urinary tract infections (UTIs) (129) and Quick *et al.* characterised the recent Ebola outbreak in West Africa (130).

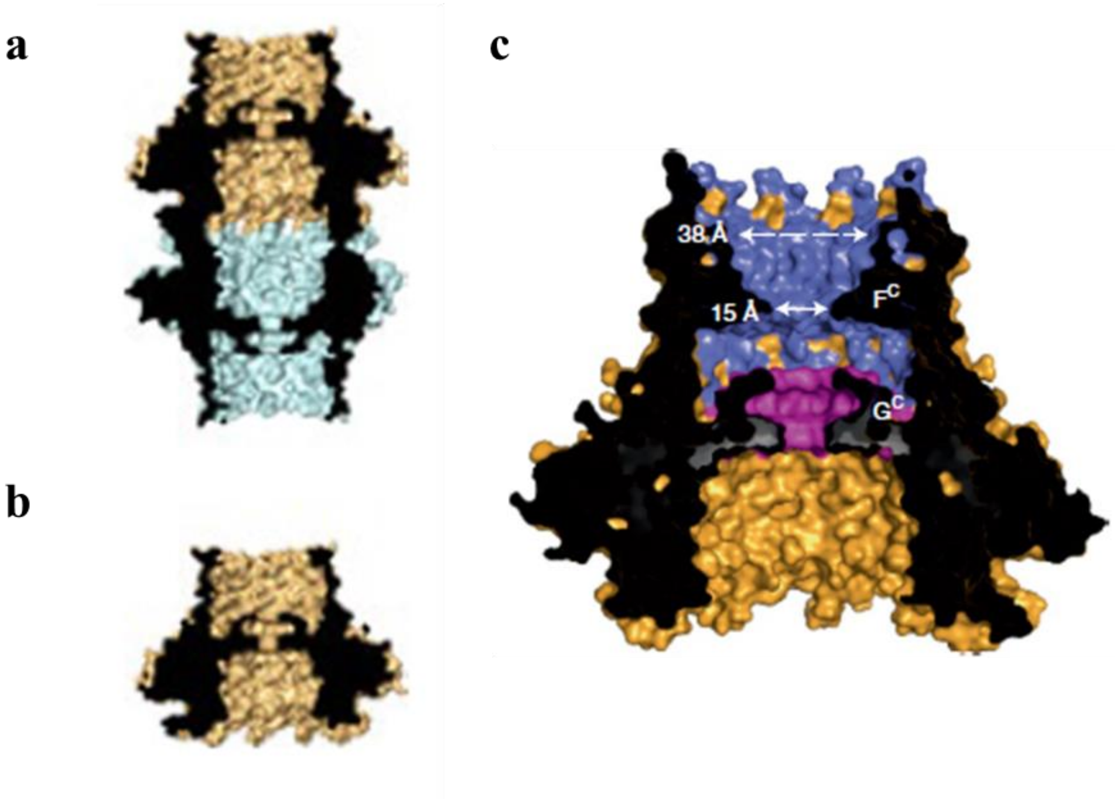


Figure 1.3: Schematic representation of the barrels present within the two latest nanopore chemistries and cryogenic-electromagnetic (EM) structure of R.10 pore chemistry. Schematic representation of the complexes present in R10 pore (A) and R9.4.1 pore (B) chemistries. Cryogenic-EM structure of R.10 pore chemistry represented by different colours. Figure adapted by ref (131).

1.3 Clinical metagenomics

Metagenomic sequencing-based approaches can combine speed and comprehensive coverage of all microorganisms present in a clinical sample, providing the potential to replace culture and PCR (especially for LRTIs) while also providing whole genome sequence data useful for public health microbiology applications. Metagenomics is defined as the identification and characterisation of all genomes present directly from the primary sample (environmental, clinical or food) (132, 133). When metagenomics is used for the characterisation of pathogens for diagnostic or epidemiological purposes directly from clinical samples, it is defined as clinical metagenomics (CMg) (134, 135).

There are numerous advantages of clinical metagenomics over other sequencing-based approaches designed to characterise the microbial composition of clinical samples, such as 16/18S rRNA sequencing or amplicon-based sequencing. Sequencing targeting the ribosomal rRNA (16S rRNA for bacteria or 18S rRNA for fungi) can only provide bacterial or fungal (not viral) identification information reliably down to the genus and cannot provide any information on antimicrobial resistance (122). Amplicon-based sequencing approaches suffer from many of the same drawbacks as PCR, utilizing primers targeting only a known set of pathogens and/or resistance genes. These targeted approaches cannot detect the breath of pathogens and resistances associated with respiratory infections or provide whole pathogen genomes. Conversely, CMg allows simultaneous identification of all microbes and associated resistance genes present in a clinical sample. Also, metagenomic data can provide whole pathogen genome data if sufficient genome coverage is recovered. This information can be used for both diagnostics (pathogen identification+AMR gene detection) and for public health applications (genome assemblies for outbreak characterisation and surveillance) (122).

1.3.1 Clinical metagenomics applied for diagnostics

Current diagnosis of LRTIs still relies on culture despite the poor sensitivity and long turnaround times. Clinical metagenomics can replace current diagnostics and, in recent years, the interest in applying CMg for diagnostic purposes has increased. There are two approaches to diagnostic CMg: i) deep-metagenomic sequencing of genetic material (human and all microorganisms) present in clinical samples and ii) sequencing of genetic material directly from a clinical sample after microbial enrichment or human depletion is performed (102).

For the first CMg approach, sample processing includes DNA extraction of both human and microbial genetic material present in the sample (102). Next, sequencing of the extracted DNA and RNA (reverse transcribed to cDNA) is done followed by data analysis (102, 136, 137). Wilson *et al*, used untargeted deep metagenomic sequencing directly from a cerebrospinal-fluid sample from a critically ill 14-old boy and identified a *Leptospira* infection (138). Another example was recently demonstrated by Wilson *et al.* (137). This multicenter study, evaluated the diagnostic efficiency of CMg for meningitis and encephalitis. Their pipeline included DNA and RNA extraction (both human and microbial) directly from CSF samples, followed by DNA and cDNA sequencing on an Illumina HiSeq instrument in rapid-run mode and data analysis (137). Metagenomics was concordant with clinical testing for 19 cases and identified 13 cases not previously identified by traditional diagnostics. Diagnostic output from metagenomic sequencing provided guidance for treatment for 7/13 cases identified by metagenomics only. Their pipeline however, failed to report 26 confirmed infections and failing of the 8/26 was due to low pathogenic titre in the samples. Pipelines utilising a CMg-based approach can also detect host infection biomarkers (transcriptome analysis) (139, 140).

Sensitivity of these CMg pipelines is related to the amount of human and microbial (non-pathogenic) background present in clinical samples – higher background, lower sensitivity (102). This limitation was recently documented by Pendleton *et al.* that applied CMg for the diagnosis two confirmed cases of bacterial pneumonia without any human depletion or microbial enrichment strategy. This resulted in the majority of sequencing reads being of human origin (99%) and only three pathogenic reads were identified (one for *P. aeruginosa* and two for *S. aureus)* (141). In clinical samples the ratio to human:microbial DNA is very high, therefore, only with deep sequencing, enough genome coverage (for high-titre pathogens) can be recovered for pathogen identification to be possible (102). Deep sequencing results in slow turnaround times (>2 days) and high sequencing cost and is not a suitable approach for replacing culture. This approach is only beneficial as a 'last-resort-test' for cases where other tests (e.g. PCR, culture) have failed to identify a pathogen.

The second CMg approach utilises human cell/DNA depletion or microbial/pathogen cell/DNA enrichment prior to sequencing to improve sensitivity and reduce turnaround time and cost (122, 142). The Pendleton *et al.* (141) study highlighted the importance of human depletion or microbial enrichment for the successful implementation of clinical metagenomics for the diagnosis of LRTIs. If CMg is ever going to be implemented for the routine diagnosis of LRTIs, turnaround time must be <8 hours (before second dose of antibiotics), cost must be reasonable (<$200 per test, similar to the cost of Filmarray and Unyvero multiplex tests) and sufficient genome coverage is required to identify pathogens and detect any drug resistance genes. CMg pipelines that utilise human depletion or microbial enrichment strategies coupled with rapid sequencing technology have the necessary characteristics for implementation. Hence, a CMg pipeline based on enrichment/depletion was considered more suitable for the aims of this PhD. The various steps involved in such a pipeline are introduced in the following sections.

1.3.2 Clinical metagenomics for public health applications

CMg can be applied beyond the field of diagnostics. It has been specifically applied in molecular epidemiology and outbreaks (143-145), but has the potential to be utilised for diagnosis and public health applications simultaneously. Molecular epidemiological studies aim to provide answers beyond diagnosis, for example, characterizing the causes of infectious diseases and their distribution (i.e. hospital transmission or infectious outbreaks) (146, 147). The majority of studies utilizing sequencing for molecular epidemiology have used whole genome sequencing (WGS) or amplicon sequencing from bacteria isolated from clinical samples (148-150). CMg can be used for epidemiological studies directly from the primary specimen without the need of growing and isolating the pathogen first as generated data can be used to assemble whole genomes (122, 143) or sequence-based typing (used for linking *L. pneumophila* outbreaks) (151).

Loman *et al.,* used CMg with Illumina sequencing and was able to investigate the 2011 outbreak of the Shiga-toxin producing *E. coli* (STEC). After sequencing 45 fecal samples from patients with diarrhea, the authors were able to recover the strain's genome from 26 samples with a >1x genome coverage and in 10 samples with a >10x coverage (152). Greninger *et al.* (2017) and collaborators were able to implement changes to infection control procedures during a flu outbreak observed in a children's hospital. Genomic assemblies of parainfluenza 3 virus (HPIV3) were generated from 8/13 samples (all from the three flu cases) and phylogenetic clustering confirmed a suspected transmission pattern due to genetic similarities being observed in 2/3 cases (153).

However, due to the constant improvement of nanopore sequencing in recent years, more CMg studies utilizing real-time nanopore sequencing for public health microbiology have been made available. Most recently, nanopore metagenomic sequencing was used in real-time for the characterization of the Lassa fever outbreak in Nigeria 2018 (145). Kafetzopoulou *et al.* (2018)

identified phylogenetic similarities with strains known to be transmitted through zoonotic hosts. This information was able to rule out the possibility of human-to-human transmission and led to a direct response from Nigerian authorities to design the most appropriate public health response (145). Nanopore metagenomic sequencing was also used during the current COVID-19 pandemic for the identification of SARS-CoV-2. Moore *et al.* was able to identify SARS-CoV-2 and characterize the lung microbiome directly from swabs with a turnaround time of 8 hrs (154).

1.3.3 Host depletion and enrichment strategies

For a rapid CMg pipeline, one of the most important steps during sample preparation, is pathogen enrichment or host depletion. Purulent sputum is the most commonly collected respiratory sample and is considered one of the most challenging clinical sample types, due to its complex sample matrix, which consists of mucus, leukocytes, pathogenic and commensal organisms (122). Normally, a purulent sputum has a variable pathogenic load and leukocytes - approximately 10^6/ml (based on our experience). Approximately one bacterial cell has 1000-fold less DNA (~5 fg for a typical bacterium) than a human cell (6.6 pg). So for example, a sputum sample that contains 10^6 pathogens, means only ~1 read out of every 1000 sequenced reads produced will be of pathogen origin as the ratio of human DNA:pathogen DNA would be 10^3:1 (122). Hence, prior to sample sequencing, host depletion or pathogen enrichment should always be performed during sample preparation in order to reduce the ratio of human:pathogen DNA (Figure 1.4).

Enrichment strategies allow the selection of one or more pathogenic agent/s present in a sample (122). Pathogen specific enrichment is possible when the infecting organism is known e.g. in patients with suspected tuberculosis (TB) and the target is *M. tuberculosis* (Mtb). An approach for selecting the pathogen/s is by targeting the unique properties of the outer envelope of the pathogen (for example via antibody binding with the mycolic acids present on the cell wall of *M. tuberculosis*). Also, enrichment strategies using a cell-based approach are able to capture microbes un-selectively by using magnetic nanoparticles (MNPs) using different ligands (155). For example, amino-glycan-functionalised MNPs were used for the rapid selection of pathogenic bacteria in food samples (156). In another example, MNPs using aptamers specifically designed to bind *S. aureus* and *E. coli* enabled the detection at low-levels (10 cfu) in animal blood (157).

Another approach for enriching specific pathogens, at the nucleic acid level, is by using capture bait probes. These probes utilise streptavidin-conjugated magnetic beads that carry nucleotides that can be up to 120bp long and are designed to hybridise with several genomic regions and/or genes e.g. AMR genes (158). At the moment there is a number of commercially available capture-based methods which are widely used (158) such as: i) Illumina's Nextera Rapid Capture Custom Enrichment (NRCCE) Kit (https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-rapid-capture-custom-enrichment.html) designed to capture selected genes of ≤15 Mb long and ii) SureSelectQXT Reagent Kit from Agilent (https://www.agilent.com/en/product/next-generation-sequencing/hybridization-based-next-generation-sequencing-ngs/dna-seq-reagents-kits-library-preparation-kits/sureselectqxt-reagent-kits-232861) that allows the hybridisation of customised genetic targets in 3.5 hrs. A number of studies have utilised these capture-based methods coupled with a customised gene panel for pathogen enrichment and have demonstrated promising results (158-161). Nucleic-acid-based methods can also be used for pan-microbial enrichment. Deng *et al*. developed a spiked primer-based enrichment strategy, which when coupled with metagenomic sequencing detected 14 viral pathogens and increased genome coverage (mean 47%) directly from plasma samples (162). The sequences for spiked primers are designed using pre-existing viral genomes as a reference and an algorithm developed by Deng and collaborators (162). PCR-tiling approaches have also been used successfully mainly for the enrichment of viral genomes such as Ebola (130), Zika virus (163) and recently SARS-CoV-2 virus (164).

The main caveat of enrichment strategies, however, is their limited target panel/s and although they can be efficient for certain applications (such as monomicrobial infectious diseases e.g. TB) they would not be beneficial for LRTIs. Respiratory infections can be caused by multiple organisms (bacterial, fungal, viral) and can also be mixed infections, therefore a targeted enrichment strategy is not always useful and a host depletion approach is more beneficial. Numerous studies have shown that in order for CMg to be successfully applied as a LRTI diagnostic tool, efficient host depletion is essential (122, 141).

Currently, there are a number of commercially available human DNA depletion kits, including the HostZERO Microbial DNA kit by Zymo Research, the NEBNext Microbiome DNA Enrichment Kit by New England Biolabs, the QIAamp DNA Microbiome Kit by Qiagen and the MolYsis Basic5 kit by Molzym. The NEBNext Microbiome DNA Enrichment Kit uses immunomagnetic separation to capture and remove human DNA by utilising differences between human and bacterial DNA, specifically targeting the highly methylated human DNA (DNA methylation is rare in microbial species) (165). The MolYsis Basic5 and QIAamp DNA Microbiome kit, utilises differential cell lysis through a chaotropic buffer that lyses human cells but not microbial cells, followed by DNase treatment for the digestion of cell-free host DNA (166).

Numerous studies have evaluated these kits and have produced variable outcomes in terms of the efficiency of human depletion but also with regards to the loss of microbial DNA (128, 165, 167-169). An example of this, is the clinical metagenomics pipeline developed by Votintseva *et al*. which could provide an accurate Mtb diagnosis within 44 hours (whereas conventional TB diagnostics provide results within weeks) (128). This pipeline incorporated the MolYsis Basic5 kit which was followed by Mtb DNA extraction and metagenomic sequencing. From 37 culture-positive samples tested, the correct species was identified in 35 and first-line antibiotics were predicted for 24/37 samples, which were 96% concordant with reference laboratory reports. Due to the sufficient removal of host nucleic acid, >90% of the mycobacterial genome was recovered in 21/37 samples. However, in 14/37 samples a high number of contaminant reads were observed and <12% genome coverage was achieved after a 16-hr MiSeq run (128).

Thoendel *et al.* developed a clinical metagenomics pipeline for the identification of prosthetic joint infections (PJIs) using the MolYsis Basic5 kit for host depletion, followed by DNA extraction, whole genome amplification (WGA) and sequencing using the Illumina HiSeq, 2x 250 cycles rapid mode (although not mentioned, turnaround would be >60 hrs due to the chosen sequencing mode) (166). In total, 408 samples (infected n=213 and non-infected n=195) were tested with this pipeline and metagenomic sequencing was concordant with culture for 109/115 culture positive PJI samples. In culture-negative PJI samples, potential pathogens were identified by metagenomics in 43/98 and additional pathogens were identified in 11/115 culture-positive PJI samples. During this study, they observed a loss of *P. aeruginosa* after host depletion which is a noticeable limitation, they also did not test if the chaotropic agents had an adverse effect on common PJI pathogens (166). Schmidt *et al.* (129) demonstrated the diagnosis of UTIs in 4hrs by combining a simple host depletion strategy (using differential centrifugation followed by MolYsis) with the real-time analysis on the MinION. Despite early failures observed in the pipeline (i.e. depletion failure or poor quality of flow cells) pathogens were correctly identified by metagenomics. Also, CMg accurately detected antibiotic resistance, reporting 51 resistance genes from the spiked pathogen compared to 55 resistance genes reported after Illumina sequencing (129).

Studies that focus on the rapid diagnosis of LRTIs utilizing rapid CMg pipelines are still limited and although the studies discussed above were not applied for LRTI diagnostics, they provide evidence of the benefits that host depletion can provide for rapid diagnosis.

1.3.4 Saponin-based host depletion

Saponins are found naturally on plant cell walls as surface glycosides and mainly act as defensive molecules, although, their exact role in plants is not completely understood (170, 171). These chemical compounds are mainly used as soap but in recent years their importance in the pharmaceutical industry has grown as they have numerous biological activities (170).

The chemical structure of saponins consists of a hydrophilic domain (a sugar moiety often glucose) and lipophilic domain (known as sapogenin or sapogenol) and their classification is dependent on the number of sugar chains they have; i.e monodesmosidic (one sugar chain attached on C3), diplodesmosidic (two sugar branches at C3 and C8). Here, I will focus on the activity of monodesmosidic saponins as there are most relevant to differential cell lysis.

Monodesmosidic saponins can cause disruption of the biological membrane mainly by inducing pore formation or by increasing the permeability of the membrane (170, 171). Numerous studies have investigated the mechanisms behind this activity and concluded that the sugar chain of the saponin interacts with cholesterols present on the biological membranes (see (171) for a comprehensive review on the biological action of saponins). In one mechanism the sugar moiety of the saponin forms hydrophilic interactions with sterols, inducing the formation of three-dimensional tubule-like structures and a new lipid phase of the biological membrane (171). This eventually leads to membrane rearrangements, increasing the permeability of the biological membrane. In a different mechanism, the sugar chain interacts with the cholesterols inducing the formation of large micelles-like aggregations (10 nm long) increasing diffusion in and out of the cytoplasm (171, 172). Lytic activity of monodesmosidic saponins is reported to be more active than diplodesmosidic saponins (170, 172).

Hence, saponins have been of great interest in research due to their effects on biological membranes such as hemolysis (170, 173, 174) and membrane-permeabilizing. Saponins have been used in cancer research due to their ability to inhibit cell proliferation but also induce lysis of cancer cells (175, 176). However, it was reported that their therapeutic usage might be limited, as they also induce red blood cell lysis due to their high affinity for cholesterols (173). Orjih *et al.* (174) utilized the haemolytic activities of saponins in order to improve microscopic detection of malaria parasites directly from blood samples. After saponin treatment 20-6000 haemolyzed parasites per field were detected, in contrast to 1-15 parasites detected only in blood samples with no saponin treatment.

In a number of studies, saponins have also been used for differential lysis of human cells, aiming to facilitate the design of culture-free rapid diagnostic methods. Zelenin *et al.* designed a microfluidic-based device which utilized a saponin-based rapid selective lysis of human cells directly from blood samples (177). Their device was designed to allow rapid diagnosis of BSIs without the need of culture, and was consisted of two phases: 1) Selective host cell lysis using saponin and 2) Osmotic shock to lyse damaged host cells. The microfluidic device consisted of three inlets, each used for loading the sample, the saponin and the water for the osmotic shock, and one outlet where cell debris and viable bacteria would be collected from. Also, the device has a herringbone-like structure to allow mixing of the samples for optimal blood cell lysis. Mixing was monitored by imaging, through all the different stages, as blood was mixed with a fluorescein solution (177). Diluted blood samples (by 4-fold) spiked with known Gram-negative and Gram-positive bacterial cell densities were used to test device's efficiency for host cell lysis and bacterial viability. After selective host cell lysis with 1% saponin, all blood cells were lysed but not white blood cells (determined via flexible flow cytometry) whilst 100% of the spiked viable bacteria were recovered (cell viability determined via plating) at all tested cell concentrations (10^4- 10^6). Their microfluidic device would accelerate culture-independent testing such as molecular testing of blood samples and is a ready-to-use device as it needs minimal handling.

Anscombe *et al.* further optimized and adapted the saponin-based method published by Zelenin *et al.* to allow whole genome amplification of pathogens directly from sterile clinical samples to enable in-depth pathogen/s characterization (178). This version of the saponin method was coupled with bacterial isolation and bacterial DNA amplification (with multiple displacement amplification) followed by NGS and was performed directly on pathogen spiked horse blood. DNA sequencing from saponin-treated samples recovered 92% of the spiked pathogenic genomes, whilst only 7% of total reads were human. Turnaround time of the altered saponin method (from sample processing to DNA sequencing, not including sequencing duration) was 3.5 hours. However, despite the promising outcomes, this pipeline was not tested on human clinical samples and only *S. aureus* and *E. coli* were spiked in horse blood for testing (178).

1.3.5 Microbial DNA extraction and sequencing

Once human cells/DNA are depleted or microbial cells are enriched, DNA is extracted from the sample (Figure 1.4). Majority of pipelines utilise chemical combined with enzymatic lysis. Consideration needs to be taken on the microbial DNA extraction method used, to ensure efficient lysis of hard-to-lyse microorganisms (such as *S. aureus* and *Aspergillus spp*. which are important HAP/VAP pathogens). Many studies, have highlighted the importance of adding mechanical lysis (i.e. beat-beading) (179, 180) to ensure efficient lysis of all microbes. (Figure 1.4). A less-efficient lysis will lead to underrepresentation of hard-to-lyse bacteria and a biased representation of the microbial community present in the sample. Additional automated DNA extraction and purification steps can be included or this can be done manually using spin columns or commercially available DNA purification kits. Automated extraction is preferred as it is less laborious, is standardised, can be rapid and this is the approach typically used in clinical microbiology laboratories. The MagNA Pure Compact (Roche), MaxWell (Promega) and QIAcube (QIAGEN), are examples of automated systems which offer rapid DNA extraction/purification (181-183) (27 min for 8 samples on Compact, 36 min for 48 samples on Maxwell and 90 min for 96 samples on QIAcube).

Prior to metagenomic sequencing, the extracted DNA is converted into a sequencing library - the choice of library preparation is based upon the sequencing technology (Figure 1.4). As turnaround time is very important, nanopore sequencing is the obvious choice as is the only real-time sequencing technology currently on the market. Also, it is more cost-effective for low-throughput sequencing in comparison with other sequencing platforms. This is important as it is not feasible to batch test (96 samples) respiratory samples from patients with e.g. HAP or VAP to make your CMg pipeline cost-effective.

Once the sequencing platform is chosen then the choice for the appropriate library preparation kit can be made. When the starting DNA quantity is low, a PCR-based library preparation would be preferred in order to amplify the starting material to a quantity sufficient for metagenomic sequencing. For example, respiratory samples, after host depletion typically contain <2ng/µL DNA which is insufficient for PCR-free library preparation (122).

There are various kits available from ONT, such as the 'Rapid PCR barcoding kit' (SQK-RPB004 – previously known as SQK-RLB001) which allows preparation and multiplex metagenomic sequencing of ≤12 samples from low input DNA concentrations (<10 ng as starting material) (122). This workflow has been successfully used by numerous studies aiming to sequence low biomass samples (122, 184, 185). During this workflow initially, the extracted DNA is enzymatically fragmented and tagged with specifically-designed adapters. Then primers complementary to the tagmented adapter are used to PCR amplify the DNA (Figure 1.4). Alternatively, if high concentrations of input DNA are available then the PCR-free 'Rapid Barcoding kit' (SQK-RBK004) can be used instead which allows the preparation of ≤12 samples in 10 min, using the same steps as the RPB004 kit. However, this kit is not suitable for sequencing of low biomass samples, as it requires a minimum of 400 ng input DNA. Illumina also offers the Nextera XT DNA Library Preparation' kit that also allows preparation and sequencing of low-biomass samples (1 ng of input DNA) which also utilises a fragmentation/tagmentation step followed by PCR amplification. This kit has also been used successfully by several studies for sequencing of low biomass samples (136, 162).

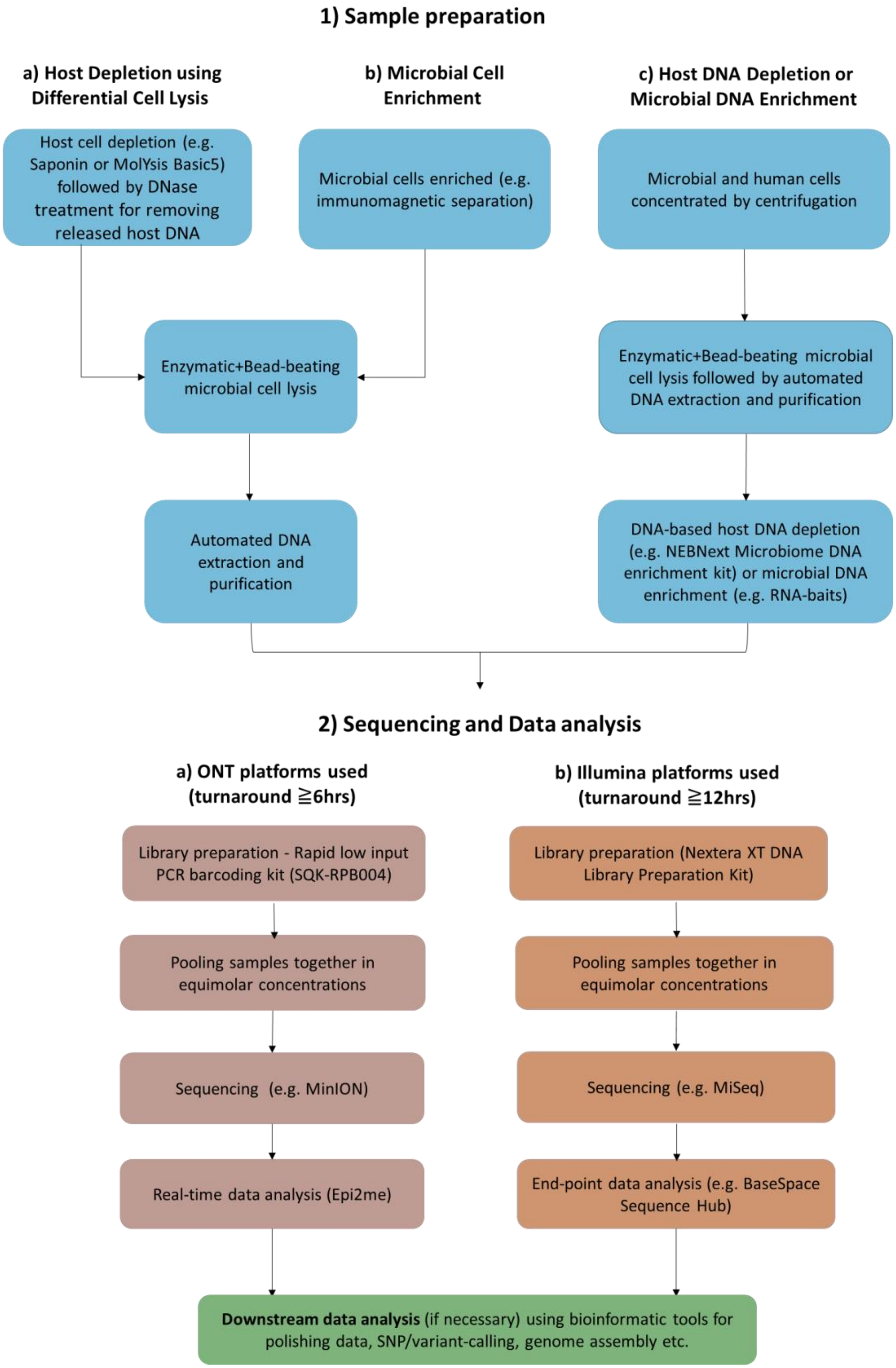

1) Sample preparation
a) Host Depletion using Differential Cell Lysis
b) Microbial Cell Enrichment
c) Host DNA Depletion or Microbial DNA Enrichment
Host cell depletion (e.g. Saponin or MolYsis Basic5) followed by DNase treatment for removing released host DNA
Microbial cells enriched (e.g. immunomagnetic separation)
Microbial and human cells concentrated by centrifugation
Enzymatic+Bead-beating microbial cell lysis
Enzymatic+Bead-beating microbial cell lysis followed by automated DNA extraction and purification
Automated DNA extraction and purification
DNA-based host DNA depletion (e.g. NEBNext Microbiome DNA enrichment kit) or microbial DNA enrichment (e.g. RNA-baits)
2) Sequencing and Data analysis
a) ONT platforms used (turnaround ≧6hrs)
b) Illumina platforms used (turnaround ≧12hrs)
Library preparation - Rapid low input PCR barcoding kit (SQK-RPB004)
Library preparation (Nextera XT DNA Library Preparation Kit)
Pooling samples together in equimolar concentrations
Pooling samples together in equimolar concentrations
Sequencing (e.g. MinION)
Sequencing (e.g. MiSeq)
Real-time data analysis (Epi2me)
End-point data analysis (e.g. BaseSpace Sequence Hub)
Downstream data analysis (if necessary) using bioinformatic tools for polishing data, SNP/variant-calling, genome assembly etc.

Figure 1.4: A schematic workflow showing examples of the steps a rapid CMg pipeline should include. Figure taken from (122) .

1.4 Microbial bioinformatics

Following metagenomic sequencing, a large amount of data is generated which is used to profile the microbial community computationally. However choosing the right approach for data analysis is challenging and is dependent on the aim of the study (186).

ONT provide a set of automated online bioinformatics pipelines, but many users opt for a customised pipeline utilizing offline tools, as this provides increased flexibility. Typically, the first steps in a bioinformatics pipeline involve basecalling the raw FAST5 files produced by the instrument, followed by quality control filtering in which low-quality data and adaptor sequence are removed (186). For the ONT sequencing platforms MinKNOW software (provided by ONT) is used to control the hardware (e.g. MinION, GridION). During sequencing, raw data (in the form of current measurements), are captured and stored by MinKNOW in real time into the raw FAST5 format. MinKNOW can also base-call raw FAST5 data in real-time and convert to basecalled FAST5 or FASTQ reads (187). As previously discussed, ONT is the only sequencing technology that offers real-time data analysis. For users who do not need real-time analysis, ONT provides the offline basecalling tools Guppy, Scrappie and Flappie (188). Guppy uses GPUs to improve execution speed but is offered only to ONT customers (138). Scrappie and Flappie (replacement of Scrappie) are described by ONT as technology demonstrators, for trialling new features that will later be applied to Guppy (187, 189). Scrappie allows direct basecalling of the raw signal (189) and Flappie uses a Connectionist Temporal Classification (CTC) for assigning bases (187). Other offline basecalling tools are offered by third parties, most notably Nanocall (190), DeepNano (191) and Chiron (192).

After base-calling, an additional quality score (QC) and/or read length filtering step can be added to remove poor quality reads. These steps can be beneficial for downstream analysis, such as genome assembly (122, 193).

1.4.1 Bioinformatics pipelines for Clinical metagenomics diagnostics

Following basecalling and filtering of the raw sequencing data, a wide range of downstream analysis tools are available. For diagnostic purposes, the next step usually involves microorganism identification and classification, hopefully, to genus and species level. Additionally, sequence data can provide information on antibiotic resistance (if any) relevant to the identified pathogen/s (122, 123). Prior to microbial classification and AMR gene detection, an additional step may be included to computationally remove any human reads that still remain after host depletion(123, 194). This can be performed using alignment-based tools (e.g. minimap2) to remove reads that map to the human reference genome (GRCh38.p13 latest human assembly - https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39).

Tools for taxonomic classification from metagenomic data utilise a number of algorithmic approaches. One approach involves aligning microbial reads to reference databases or aligning contigs (microbial reads assembled *de novo* into contiguous sequences) to reference databases. Other strategies include analysing *k*-mer (short sequences of size k) content, mapping marker genes or protein sequences translated from the DNA sequences (194).

Table 1.5 provides an overview of common tools for metagenomic classification. BLAST, has for a number of years, remained the gold standard sequence classification tool. However, it is relatively slow due to its computational complexity. MetaPhlan (195) aligns clade-specific marker genes against sequences in order to identify taxa (194). Although not as sensitive as alignment-based tools, a number of classifiers rely on *k*-mer-based classification leading to much improved speed and reducing the computational power required (194). Additionally, the choice of k (length of the sequence) is very important for the sensitivity and specificity of *k*-mer-based classifiers. Long *k*-mers can result in lower sensitivity as exact matches for certain *k*-mers may not be identified, possibly due to sequencing errors or actual differences in the sequenced data, whereas short *k*-mers can result in multiple matches per *k*-mer (possibly false) leading to reduced specificity (196).

Kraken, (197) is a *k*-mer-based classifier which, uses a novel algorithm to identify precise *k*-mer alignment from the Kraken database. *K*-mer/s without a match from the database, are not classified and are discarded (197). Kraken utilises an in-memory *k*-mer database (all compressed in a table which is used for identification of exact matches), which needs a large amount of RAM, meaning it cannot be run on typical desktop or laptop computers (194). In contrast, Centrifuge utilises a memory efficient index scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index, which enables it to index 4000 bacterial genomes in around 4.3 Gb of RAM. It can provide accurate and rapid results from large metagenomic data and can be performed using a conventional desktop computer (198).

A recent study by Ye et al used simulated datasets to evaluate the performance of 20 recently-developed benchmarked metagenomic classifiers, including *k*-mer based classifiers, protein-based classifiers and classifiers utilising marker genes (199). DNA *k*-mer based classifiers using long *k*-mers provided more precise results with better abundance estimates when compared with protein- or marker-gene-based classifiers. However, it was highlighted that a caveat of recently-developed classifiers, is the trade-off between fast classification speed and specificity. To achieve fast classification, current algorithms reduce the number of candidate hits, for example *k*-mer classifiers only seek for exact sequence matches of length k (typically k=31). As previously mentioned, although not as sensitive as BLAST, these classifiers are preferred as they are faster and require less computing requirements (199).

Classification of pathogens from metagenomic data can also be performed using metagenome-assembled genomes (MAGs) (200). For this approach, common overlaps of the sequenced reads are identified to build contigs resulting in the construction of a metagenomic assembly (an assembly of multiple microbial genomes). Once the assembly is constructed, it is mapped against a reference database, to identify alignments of contigs against known-sequenced genomes. The use of contigs instead of shorter sequenced reads can improve the accuracy of pathogen classification but construction of MAGs comes with many limitations. The main difficulties of a metagenome assembly arise from the different abundances of microbial genomes present in metagenomic samples and the presence of many closely related species or strains. If sequencing depth is not high enough for low abundance organisms, then assemblies will be fragmented and classification may fail (200). Pathogen classification using MAGs is still in its infancy and more evaluation studies should be carried out prior to implementing this approach for CMg (201).

Table 1.5: Tools used for taxonomic classification for metagenomic data. Table adapted from (194).

Tool	Outline	Source
BLAST+	Aligner able to use nucleotide or translated-nucleotides as input – provides sensitive results. MegaBLAST (part of the BLAST+) can handle longer DNA sequences and classification of highly similar sequences	https://blast.ncbi.nlm.nih.gov
Centrifuge	Taxonomic classifier providing rapid results using the memory efficient FM index scheme and provides results in a Kraken-like format	http://ccb.jhu.edu/software/centrifuge/
Kaiju	Fast metagenomic classifier using protein sequences as an input	https://github.com/bioinformatics-centre/kaiju
Kraken	Fast taxonomic classifier for metagenomics by identifying exact *k*-mer matches against a compressed database containing multiple genomes	https://ccb.jhu.edu/software/kraken/
MetaMaps	Read-based aligner specifically designed for long reads using a two-stage analysis procedure	https://www.nature.com/articles/s41467-019-10934-2
MetaPhlAn 2	Taxonomic classifier using marker genes as an input	https://bitbucket.org/biobakery/metaphlan2

Once a pathogen (if any) is detected and classified, then relevant AMR genes can be identified (122). This step is necessary to help determine the choice of antibiotics used for the patient's treatment. The most common approach involves identifying alignments of single reads against an antimicrobial resistance gene database. An alternative approach is using whole genome (single or metagenomic) assemblies, which can increase the accuracy of resistance gene identification (200). However, if sufficient sequencing depth is not achieved then whole genomes or accurate MAGs cannot be constructed and AMR genes cannot be identified. Nonetheless, the chosen approach should be able to identify all resistance mechanisms, i.e. acquired resistance genes (including variants) and chromosomal resistance genes (including mutational resistance) (122).

Identifying the resistance gene host is difficult from metagenomic data. However, using long-read rather than short-read sequencing technology, can identify the origin of chromosomal resistance genes, by identifying the pathogen based on the genomic flanking regions (122, 123). Plasmid borne resistance genes are particularly difficult to associate with their host, although long-read sequencing makes it easier to assemble plasmids and potentially provide some information about their likely host. However, this approach is likely to fail when considering promiscuous plasmids.

Perhaps the most challenging part of CMg is trying to predict pathogen phenotype (resistance and/or susceptibility) from metagenomic data. This is because gene expression, permeability and efflux all make resistance and susceptibility more complex that just presence or absence of resistance genes. A number of groups are working on this problem. Brinda *et al.,* developed RASE, a tool able to predict resistance and susceptibility of pathogens from nanopore metagenomic data in real time by identifying a pathogen's closest relatives (143). This technique when used with *k*-mer matching, was able to determine resistance within four hours from sample collection for *S. pneumoniae* directly from respiratory metagenomic sequencing data. Ruppé *et* *al.*, took an alternative approach by using a knowledge-based algorithm for antimicrobial resistance genes detection (202). The algorithm was used on WGS data from Enterobacterales and predictions were confirmed by disc diffusion. The algorithm correctly predicted 963 susceptibilities and 257 resistances (202).

An example, of a mature CMg analysis pipeline is the sequence-based ultrarapid pathogen identification (SURPI), an automated computational tool developed by Naccache *et al.* (203). SURPI+ was developed for metagenomic data analysis such as pathogen identification and classification. Initially, the SUPRI+ pipeline offers data pre-processing, followed by human read removal and microbial classification via alignment against the NCBI Nucleotide (NT) database. Results are then visualised in a user-friendly graphical interface and are readily available for interpretation (136).

EPI2ME, despite not being clinically validated like SURPI, is also another example of a mature tool with a graphical interface. Users of nanopore sequencing have access to EPI2ME, which consists of a desktop agent for uploading reads and a user-friendly web interface that offers a number of pipelines for real-time data analysis including pathogen identification and AMR gene detection. The Antimicrobial Resistance pipeline within EPI2ME allows pathogen identification by combining the WIMP (What's In My Pot) pipeline and antimicrobial resistance gene detection with ARMA. WIMP uses the Centrifuge classifier (described before), to map reads into the RefSeq database for identification of bacteria, viruses and fungi (122). ARMA identifies AMR genes by mapping reads to the 'CARD' (Comprehensive Antimicrobial Resistance Database) database (204). Reads are aligned using minimap2 and any alignments reported are only over ONT-chosen default thresholds (described in 2.8.2). ARMA also offers the 'clinically relevant' parameter which currently reports only acquired and chromosomal resistance genes, but not resistance mutations/SNPs. This feature was designed by ONT and collaborators to allow rapid identification of clinically relevant AMR genes and exclude reporting of resistance genes that would not provide useful information for designing antibiotic treatment (122).

Another example of an automated tool is NanoOK RT, which is designed for microbial classification and AMR gene detection using nanopore data as input (123). NanoOK aligns sequence data against the NCBI nucleotide database for microbial identification and against the CARD database for resistance gene detection in real-time. This tool also provides an additional feature, which aligns the flanking regions of detected chromosomal AMR genes to match the gene's pathogen origin (123).

1.4.2 Databases used for diagnostic purposes

Bioinformatic tools used for microbial classification and antimicrobial resistance gene detection using metagenomic data rely on databases containing either DNA sequences and their translated protein sequences or previously sequenced whole genomes. Therefore, the accuracy and sensitivity of the CMg pipeline relies on the comprehensiveness and accuracy of the database used. The NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy) is the repository for the standard nomenclature and classification for the INSDC, (International Nucleotide Sequence Database Collaboration). The INSDC (205) also contains the main databases such as GenBank, (a DNA sequence database available since 1992), the European Molecular Biology Laboratory (EMBL) and dbSNP (used for single-nucleotide polymorphisms) (206) (Table 1.6).

Microbial classification however, is challenging especially for closely related species in a genus and closely related genera which can lead to assigning the wring taxonomy ID. For example, species of the *Shigella* genus are often misclassified as *Escherichia* (and *vice versa*) due to the high genetic similarity shared between these two genera (>97% similarity). Another example is commensal and pathogenic *Streptococcus* spp. (207). Often even well-established approaches (e.g. Mass spectrometry or PCR-based assays) misclassify non-pathogenic *Streptococcus* species as *S. pneumoniae* (99, 208). Classification can be improved if sufficient sequencing depth and genome coverage are obtained or by increasing microbial classification scores, but a phylogenetic approach is the best strategy for an accurate classification (122). The main molecular phylogenetic approaches (discussed below) either use a gene-by gene comparison (multi locus sequence typing (MLST)) or SNP-based comparison of closely-related genomes (209). A phylogenetic approach is often used for outbreak studies to enable the identification of genetic similarities amongst numerous genomes of the same pathogen, enabling the source of an outbreak and transmission patterns to be determined (209, 210). Additionally, using curated databases such as RefSeq can increase classification accuracy, as additional filters, such as eliminating fragmented assemblies, are used for every new entry (Table 1.6). RefSeq genomes are regularly updated (i.e. when new information/genomes are available), in contrast to GenBank, where even fragmented genomes can be uploaded (211) (Table 1.6).

The chosen database also needs to reflect the purpose of the metagenomic study. This is particularly relevant for databases used for AMR gene detection. Databases used for AMR gene detection in CMg pipelines should only contain clinically relevant genes and their detection would provide useful information for targeted antibiotic treatment. A data-restricted database would make data interpretation easier and reduce computational time for searching through big databases (194). The Resfinder database (212) and the Comprehensive Antibiotic Resistance Database (CARD) are examples of this. CARD provides a comprehensive set of antibiotic resistance gene sequences and their protein sequence along with their targets (213). Equally, the ResFinder database contains an exhaustive range of resistance genes and both acquired resistance genes and chromosomal mutations are now available (213). The ResFinder database is curated and contains characterized and peer-reviewed gene sequences, which makes it a more appropriate choice for providing more accurate matches.

Table 1.6. Number of entries of draft and complete microbial genomes in the most commonly used microbial databases (194)

	Complete genomes			**Draft Genomes**		
	Bacterial species	**Fungal species**	**Viral species**	**Bacterial species**	**Fungal species**	**Viral species**
GenBank	2677	17	0	19078	997	1
RefSeq	2586	7	7073	11217	190	3

1.4.3 Bioinformatic workflows for public health applications

As previously discussed, metagenomic data can also be used to generate whole pathogen genomes which can provide additional information beyond diagnostics (214, 215). Briefly, during the process of assembling genomes, reads are overlapped to create contigs, which are then joined to form scaffolds to later form full or partial genomes (216). The fundamental approaches to generating genome assemblies are overlap–layout–consensus (OLC) and de-bruijn-graph (DBG) (217). The OLC approach involves three steps; during the first step, all reads are overlapped (O) and in the second step a layout (L) of the overlapped-reads is formed which is then used in the third step to form the consensus sequence (C). During the DBG approach, firstly all sequencing reads, are chopped into short *k*-mers which are then used to form the DBG. The DBG-based algorithm then, uses the *k*-mers to infer the genome sequence on the DBG graph generated during the previous step (217).

Additionally, the genome assembly can be generated either by using a reference genome as a guide for assembly (122, 218) or by a *de novo* (i.e. no reference is used to assemble the genome) strategy. When trying to identify novel organisms (not previously sequenced), *de novo* assembly would be most appropriate as through this approach the complete genome of the organism (not previously sequenced) can be reconstructed. Conversely, if the identified organism shares genetic similarities with previously sequenced genomes then a reference-based approach can be used to reconstruct the genome (186).

For the reference-based approach a suitable reference closest to the organism in question should be used. A reference can be chosen based on the most closely related strain identified by metagenomic read classification. However, if a reference-based rather than a *de novo* approach is used then genomic variants would be lost (186). Pandora (https://github.com/rmcolq/pandora) a recently-developed tool, designed for both long and short-read data, although not designed to be used for CMg data, can be used as an alternative approach to identify reference genomes for alignment based genome assemblies where the pathogen is known. This tool uses available genomes in databases to identify conserved regions and nucleotide-level and longer variants to create a pangenome reference graph. The graph is then used to identify which genome shares the highest similarities with the sequenced genome. By using this alternative approach, conserved and novel regions of the sequenced genome can be identified.

Additionally, a number of tools have been recently developed using nanopore long reads for genome assemblies. Currently the main strategies followed for the reconstruction of genomes using long-reads are; the hierarchical method, the hybrid and the direct method - for a review of the approaches used for long-read genome assembly see (219). Canu a long-read assembler (220), uses the hierarchical approach and its pipeline consists of three stages: i) correcting (reads are build into overlaps and 'best' overlaps are selected for correction), ii) trimming (removes unsupported regions of the overlaps) and iii) the assembly stage (identifies any final sequencing errors and creates the best overlap used for contig construction) (220). Canu has been used to generate continuous sequences with high accuracy (220). Alternatively for a quick assembly, Miniasm can be used. Miniasm is an OLC-based *de novo* long-read assembler which implements the overlap ('O') and layout ('L') steps of OLC assemblers but not the consensus step ('C') during which the error rate is corrected by creating a consensus sequence (described before). Miniasm, overlaps mapped raw reads (typically produced by minimap2) for genome assembly and, as it does not have a correction step, it can assemble genomes within minutes. Despite, the possibility of an ultrafast assembly, read selection is performed crudely by Miniasm, meaning sequencing error rate is carried through, from raw data (221). To improve assembly accuracy however, Miniasm can be coupled with long-read correcting tools such as Racon (222). Hybrid assemblies typically provide more accurate assemblies as they benefit from the coverage provided by long reads and the per-base accuracy of short reads (223). Unicycler (224) and SPAdes (225, 226) are hybrid assemblers and both can use long- and short- reads as an input.

1.4.3.1 Molecular typing methods

Once a genome has been assembled, additional analysis is performed which is dependent on the research question. In outbreak studies, transmission patterns and identifying the outbreak source are often paramount. Therefore, steps would be necessary to identify mutations, variants and conserved regions which will help characterise the genomic evolution of the pathogen in question and carry out phylogenetic analysis by microbial typing (122).

Traditional microbial typing used cultivation, as a method to differentiate bacterial species, by identifying a number of phenotypic markers. Molecular typing, initially used a PCR-based approach, during which the DNA sequence of phenotypic markers is amplified which then was used to identify bacterial species. The use of PCR allowed the introduction of sequence-based typing (SBT). SBT-based phylogenetic approaches mostly used two approaches; MLST which targets a number of genes of the genome and SNP-based typing which uses SNPs to identify bacterial strains. The MLST-based approach is then subdivided into core-genome MLST (cgMLST) and whole-genome (wgMLST) (209).

CgMLST uses genes of the core genome to identify genetic similarities between species and initially used a 7-gene target which later then increased to a bigger number of gene loci. During this approach, assembled genomes are aligned to schemes – schemes are consisted of allelic sequences of the targeted genes loci and their associated allelic numbers – in order to identify the allelic profile of the genome in question, leading to further identification of the sequence type (ST). WgMLST, on the other hand, utilises both accessory genes and genes of the core genome/s. Closely-related species are distinguished with higher resolution as genetic similarities with this approach are calculated using a bigger set of gene targets (227, 228). However, various studies reported similar findings between the two approaches. Pearce *et al.*, demonstrated that there was no significant difference in the findings of wgMLST and cgMLST for the typing of *S. enterica* during an Enteritidis outbreak (229). In a practical setting, both approaches should be used - cgMLST to be used initially on a dataset of numerous species, followed by wgMLST typing of closely related strains based on cgMLST-findings (209).

The SNP-based approach utilizes the different SNPs present amongst different strains. The genomes in question are aligned against a reference genome in order to identify SNPs that maybe present in the sequences (209, 230). Once SNP distance is determined (number of SNPs identified between the query genome and reference) then genetic similarity is identified – the smaller the SNP distance the higher the genetic similarity is between the query genome/s and reference sequence. The choice of the reference genome is important for SNP-based typing. The chosen reference needs to cover as many positions as possible of the query genome/s in order to increase likelihood of identifying all SNPs that may be present (209). A distant reference genome, increases the likelihood of identifying inaccurate SNPs, as more differences would be present between query and reference genome. A commonly used approach, is to choose a reference genome from the same serogroup as the query genome. An alternative approach for identifying the reference genome, is to estimate the distance of the 'unknown' genome/s against a dataset of 'known' closely related genomes (231, 232). Molecular typing using any of the three approaches has provided consisted results. The advantage of using cgMLST is that conserved genes are targeted, hence providing an accurate identification at the species level. Strains from the same species however, share high genetic similarity, differing only by a few mutations, making SNP-based typing the best approach for such purpose.

In outbreaks of *Legionnaire's* disease for example, a cgMLST approach is commonly used to identify the sequence type (ST). An allelic profile is used from seven housekeeping genes conserved amongst *L. pneumophila* (233). The seven gene targets used are, *flaA*, *pilE*, *asd*, *mip*, *mompS*, *proA* and *neuA.* Once the allelic profile is determined (i.e. the allele of each gene is identified) then the ST is automatically identified (233-235). As previously discussed, the majority of SBT utilises data from WGS-based studies and studies utilising CMg data for molecular typing are limited.

Metagenomic sequencing has been used to identify *Legionella* spp. from environmental samples in various studies (236-238) but limited attempts have been made to detect *Legionella* spp. from clinical samples. CMg was only recently tested directly on spiked sputum samples with mock communities, consisting of different quantities of human, *L. pneumophila* and three other bacterial species DNA (151).

1.5. Aims and objectives of the study

1.5.1 Aim

The overall aim of my study was to develop, optimize and evaluate a CMg pipeline that could be applied for the rapid diagnosis of lower respiratory infections (including pathogen identification and AMR gene detection) and for public health purposes (such as rapid identification and molecular typing of *Legionella pneumophila*).

1.5.2 Objectives

- To develop and optimize a method that would deplete ≥ 99.9% of human DNA from respiratory samples (sputum, BAL or endotracheal tube aspirates).
- To combine the host depletion method with an efficient DNA extraction, followed by low input library preparation and nanopore sequencing.
- Optimise a bioinformatic pipeline for analysing the nanopore metagenomic data and identifying of respiratory pathogens and associated resistance genes.
- To make the CMg method as rapid, simple and cost-effective as possible.
- To evaluate the analytical performance of the method using spiked respiratory samples.
- To evaluate the clinical performance of the method for the diagnosis of LRTIs.
- To assess whether the method could be utilised to characterise *Legionella* spp. in respiratory samples to guide public health interventions.

2.Methodology

2.1 Ethical approval for sample collection

Ethical approval for the collection of all excess respiratory samples was provided by the UCL Infection DNA Bank (REC reference 12/LO/1089). Excess respiratory samples from patients with suspected lower respiratory infections (persistent (productive) cough, bronchiectasis, CAP/HAP, cystic fibrosis and exacerbation of chronic obstructive pulmonary disease (COPD, emphysema/chronic bronchitis)) were collected for the development and optimization of the human depletion method. Excess respiratory samples were initially used to develop and evaluate the first version of the CMg pipeline (refer to as pilot samples and pilot study), then additional optimization and testing was done for the optimized and final version of the pipeline and samples used for this are referred to as streamline samples (described in 3.1). The optimized CMg pipeline was also implemented in the INHALE trial to evaluate the diagnostic performance of the pipeline for HAP and VAP – described in 3.2. For this excess respiratory samples from patients with VAP/HAP were used.

For all samples (pilot, streamline and INHALE) microbiology results were collected (describing the pathogen(s) identified by routine microbiology and their antibiotic susceptibility profiles) and no patient identifiable information was collected, hence informed consent was not required.

The optimised CMg pipeline was also tested for the characterisation of *Legionella* spp. – this study is described in 3.3 and is referred to as the *Legionella* study and samples are referred as *Legionella* samples. For the implementation of CMg for the *Legionella pneumophila* study excess *Legionella*-positive samples from the Respiratory and Vaccine Preventable Bacteria Reference Unit (RVPBRU), Public Health England (PHE) were used. Ethical approval was not required, as for the *Legionella* study excess samples were collected and no patient identifiable information was collected. Microbiological and molecular-based typing results were collected only, such as the identified *Legionnella* species and sequencing type (ST).

2.2 Sample collection and storage

Excess respiratory samples (sputa, ETA, BAL) were collected and stored at 4 °C prior to testing for clinical metagenomics, after routine microbiology was performed (described in section 2.3) at the Norfolk and Norwich University Hospitals (NNUH) Microbiology Department. Samples were deemed as either positive (contain one or more bacterial pathogen(s)) or as negative samples (NRFs (normal respiratory samples, NG (no growth) and NSG (no significant growth)) by clinical microbiology. To develop and optimize the CMg pipeline, 24 sputum samples were used and 40 samples, (comprising 34 sputa, four BALs and two ETAs) were used to test the CMg method. The CMg pipeline was further tested on additional 41 samples (comprising of 38 sputa, one BAL and 2 ETAs). For the INHALE study 73 fresh respiratory samples were used (comprising of 32 sputa, 9 BALs, 29 endotracheal tube (ETT) exudates, 2 tracheostomy tube exudates and 1 tracheostomy exudate) and for the *Legionella* study 48 excess frozen samples (38 sputum samples, 9 BAL samples and 1 pus sample from muscular abscess) were used.

2.3. Microbiological investigation of respiratory samples

2.3.1. Routine testing for suspected lower respiratory tract infections (NNUH Clinical Microbiology)

Sputum and ETAs were treated with sputasol (Oxoid-SR0233) in a 1:1 ratio and were incubated for a minimum of 15 min at 37 °C. Sputasol-treated samples (10 µL) were added into 5 ml of sterile water and mixed, making the limit of detection (LoD) of culture ~5 x 10^5 CFU/ml. Then, each sample was streaked onto blood, chocolate and cysteine lactose electrolyte deficient (CLED) agar (10 µL per plate). For samples coming from the intensive care unit (ICU), 10 µL sputasol-treated sample was plated with no water dilution. BALs were not sputasol treated like other respiratory sample types. Instead these samples (total volume of sample) were centrifuged to concentrate microbial cells for a minimum of 10 min at 3000 rpm. Then samples were plated directly onto the agar plates and no further dilution occurred prior to plating. Depending on the source of the sample and clinical information, other agar plates were also used, such as: sabouraud, mannitol salt and *Burkholderia cepacia* selective agar. All inoculated agar plates were incubated at 37 °C overnight and then examined for growth with the potential for re-incubation up to 48 hours. If any significant organism was grown, then antibiotic susceptibility testing by agar diffusion using EUCAST methodology was performed. The laboratory's Standard Operating Procedure is based on the Public Health England UK Standards for Microbiology Investigations B 57: Investigation of bronchoalveolar lavage, sputum and associated specimens (65).

2.3.2 Routine testing for the identification and isolation of *Legionella* spp. (PHE, Colindale)

All samples from patients suspected with *Legionella* infection were cultured as described below:

All sputum samples were initially sputasol-treated (1:1 ratio) and centrifuged for 15 min at 1000 rpm. After centrifugation the supernatant was removed and the pellet re-suspended in 1 ml of sterile water. Additionally, 250 µL was heat-treated for 30 min at 50 °C to kill human cells. Plating out was on the following media:

- Buffered Charcoal Yeast Extract (BCYE) – 100 µL neat and heat-treated sample
- Buffered polymyxin anisomycin (BMPA) – 100 µL of neat and heat-treated sample
- Buffered Charcoal Yeast Extract with Cefamandole (BCY-C) – 100 µL of neat sample

Diluted (1:100) samples (both neat and heat-treated) were also plated out on BCYE and BCY-C agar. All plates were then incubated statically at 35-37 °C for a maximum of 10 days. After 4 days of incubation the plates are initial examined for growth followed by re-confirmation at 10 days.

DNA was extracted from all samples for additional PCR-based testing. DNA was extracted using 200 µL of the non-heated sample and added to 200 µL Bacterial-lysis buffer and 20 µL of proteinase K followed by heat treating at 65 °C for 10 mins and at 95 °C for 30 min. The lysed sample was then processed on the MagNA Pure Compact 2.0 automated instrument using DNA_BacteriaV2 program.

The DNA extract was then used for triplex PCR targeting the *mip*, *gfp* and *wzm* genes of *L. pneumophila* (only on culture-positive samples (≥1 cell identified)). For non-*pneumophila* culture-positive samples a PCR targeting the 16S rRNA and *mip* gene was performed on the DNA extract and amplicon (amplicons targeting the *mip* region only) are then subjected to

Sanger sequencing for confirmation of the cultured species (see Table 2.1A for all primer sequences and gene targets used). A nested PCR targeting the seven housekeeping genes (*flaA, pilE, asd, mip, mompS, proA, neuA*) was performed on the DNA extract on all culture-negative samples to obtain a profile on *L. pneumophila* that failed to grow on plates (see Table 2.1B for all primer sequences used for the nested PCR).

In addition to these tests, sequence based typing (SBT) was also done on *L. pneumophila* colonies identified in culture-positive samples. For SBT *L. pneumophila* colonies were isolated and DNA was extracted via a chelex extraction and then SBT with Sanger sequencing was performed. Briefly for the chelex extraction a loopful of *L. pneumophila* culture was emulsified in ~1 mL of sterile distilled water, followed by a 5 min centrifugation at 12,000 xg. Then, supernatant was discarded and pellet was resuspended in 200 µL of Instagene matrix (BIO-RAD cat no 732-6030). Sample was incubated at 56 °C for 30 min and 100 °C for 8 min and after that was centrifuged at 12,000 xg for 5 min. 20 µL of the supernatant was transferred to a clean tube and 180 µL of TE buffer was added and the extract is then subjected to SBT. A negative extraction control (using sterile water) was always performed in parallel to monitor any contamination.

Table 2.1A: PCR Primer sequences and gene targets used for routine testing of *Legionella* spp.

Organism	Gene target	Forward primer (5'-3')	Reverse primer (5'-3')	Probe (5'-3')	Reference
L. pneumophila	*mip*	GAAGCAATGGCTAAAGGCATGC	GAACGTCTTTCATTTGYTGTTCGG	HEX - CGCTATGAGTGGCGCTCAATTGGCTTTA - BHQ1	(239)
	wzm	CAAAGGGCGTTACAGTCAAACC	GACAAACACCCCAACCGTAATCA	FAM - CTTGGGATTGGGTTGGGTTATTTAACTCC - BHQ1	(240)
	gfp	CCTGTCCTTTTACCAGACAACCA	GGTCTCTCTTTTCGTTGGGATCT	TxRed - TACCTGTCCACACAATCTGCCCTTTCG – BHQ2	(241)
***Legionella* spp.**	16S rRNA	AGGCTAATCTTAAAGCGCC	CCTGGCTCAGATTGAACG	FAM-CGGTGAGTAACGCGTAGGAATATGG-BHQ1	(242)

Table 2.1B: PCR Primer sequences and gene targets used for the nested PCR assay for routine testing of *L. pneumophila*

	Gene target	**Forward primer (5’-3’)**	**Reverse primer (5’-3’)**	**Reference**
First round	*flaA*	TATGCGTGAGCT TTCCGTTC	CCATTAATCGTTAAG TTG TAGG	(243)
	pile	CGTTGGAATCGGCTTG TC	CGCATTGGCAGAGG AATCTA	
	Asd	CCCTGGAAGTGA ATCCTCAT	TTGCAGTATTTC AGCGATCTGT	
	Mip	TGAAGATGAAAT TGGTGACTGC	AATAGGTCCGCC AACGCTAC	
	mompS	TTGACCATGAGT GGGATT GG	TGGATAAATTATCCA GCC GGACTTC	
	proA	CCGCTTCTCCAACCAA TG A	CACTCAACATAC CGCAACCA	
	neuA	CCTTGCAGTCGTCTTG TT GT	TTTCTGTTAGAGCCC AAT CG	
Second round	*flaA*	TGTAAAACGACGGCC AGT GCG TAT TGCTCAAAA TACTG	CAGGAAACAGCTAT GACC GGTATCACCTGCGGT TCC A	
	pile	TGTAAAACGACGGCC AGT CAC AAT CGG ATG GAA CAC AAA CTA	CAGGAAACAGCTAT GACC GCTGGCGCACTCGGT ATC T	
	Asd	TGTAAAACGACGGCC AGT CCC TAA TTG CTC TAC CAT TCA GAT G	CAGGAAACAGCTAT GACC CGAATGTTATCTGCG ACT ATCCAC	
	Mip	TGTAAAACGACGGCC AGT GCT GCA ACC GAT GCC AC	CAGGAAACAGCTAT GACC CATATGCAAGACCTG AGGGAAC	
	mompS	TGTAAAACGACGGCC AGT GACATCAATGTGAAC TGG	CAGGAAACAGCTAT GACC CAGAAGCTGCGAAA T CAG	
	proA	TGTAAAACGACGGCC AGT GATCGCCAATGCAAT TAG	CAGGAAACAGCTAT GACC ACCATAACATCAAA A GCC	
	neuA	TGTAAAACGACGGCC AGT CCGTTCAATATGGGGC TT CAG	CAGGAAACAGCTAT GACC CGATGTCGATGGATT CAC TAATAC	

2.3.3 Research laboratory culture growth conditions (UEA)

All bacterial isolates (*Haemophilus influenzae, Klebsiella pneumoniae*, *Stenotrophomonas maltophilia, Streptococcus pneumoniae, Escherichia coli, Staphylococcus aureus, Pseudomonas aeruginosa*) and fungal isolate (*Candida albicans*) were grown aerobically at 37 ˚C overnight, either with orbital shaking at 180 rpm or statically with 5% $CO_{2,}$ in 10ml of an appropriate liquid growth medium (i.e. luria broth – LB, tryptic soy broth – TSB or brain heart infusion – BHI from Thermo Fisher Scientific). Organism specific growth conditions are detailed in Table 2.2.

Table 2.2: Growth conditions for microbial cultures used for mock community and LoD experiments

Pathogen	**Plate**	**Broth**	**Conditions (°C)**	**Shaking**
Haemophilus influenzae	Blood	TSB	37 with 5% CO2	NO
Klebsiella pneumoniae	Blood	TSB	37	YES
Stenotrophomonas maltophilia	Blood	TSB	37	YES
Streptococcus pneumoniae	Blood	BHI	37 with 5% CO2	NO
Escherichia coli	LB	LB	37	YES
Staphylococcus aureus	LB	LB	37	YES
Candida albicans	LB	LB	37	YES
Pseudomonas aeruginosa	CLED	TSB	37	YES

2.4 Clinical sample and microbial DNA extraction and purification

Cell pellets from clinical and bacterial samples were resuspended in bacterial lysis buffer (Roche UK- 4659180001) (380 µL pilot samples or 400 µL for streamline, INHALE and Legionella samples after bead-beating) and 20 µL of proteinase K (>600 mAu/ml) (Qiagen -19133) was added for microbial DNA extraction. Samples were then incubated for 5 min (for streamline, INHALE and Legionella samples) or 10 min (for pilot samples) at 65 °C shaking at 800 RPM. Following this, samples used for the *Legionella* study were subjected to a heat-killing step, which involved a 30 min incubation at 95 °C.

DNA was then purified using the Roche MagNAPure Compact DNA_bacteria_V3_2 protocol (MagNA pure compact NA isolation kit I, Roche UK - 03730964001) on a MagNA Pure Compact machine (Roche UK - 03731146001). Briefly, during the purification, DNA binds to magnetic glass particles, which are then pelleted using a magnet. Cells debris is then removed using multiple washes. Finally, DNA is separated from the magnetic glass particles using high temperature and eluted into 50 µL elution buffer.

2.5 DNA quantification and quality control

The quality and quantity of extracted DNA was assessed as follows.

2.5.1 DNA quantification

DNA quantification was performed using the high sensitivity dsDNA assay kit (Thermo Fisher - Q32851) on the Qubit 3.0 Fluorometer (Thermo Fisher - Q33226) or the Broad Range (BR) dsDNA assay kit (Thermo Fisher - Q32850). In brief, 199 µL of the working solution (199 µL of

Quant-iT™ dsDNA HS buffer and 1 µL of Qubit® dsDNA HS Reagent or 199 µL of Quant® dsDNA BR buffer and 1 µL of Qubit® dsDNA BR Reagent) and 1 µL of sample DNA was used for sample quantification. For the fluorometer's calibration, 190 µL of the working solution was used per standard and 10 µL of either standard 1 (Qubit® dsDNA HS Standard #1 or Qubit® dsDNA BR Standard #1) or standard 2 (Qubit® dsDNA HS Standard #2 or Qubit® dsDNA BR Standard #2) was used. After DNA and standards were added in the working solution, samples were vortexed briefly and incubated at RT for 2 min in the dark. Then standards and samples were quantified using the dsDNA High Sensitivity or the dsDNA Broad Range assay program.

2.5.2 DNA fragment size and quality analysis

DNA quality and fragment size were assessed using the TapeStation 2200 (Agilent Technologies - G2964AA) automated electrophoresis platform with the Genomic ScreenTape (Agilent Technologies - 5067-5365 and a DNA ladder (200 to >60,000 bp, Agilent Technologies - 5067-5366). This step was mainly used to the test the quality of: i) products after the library preparation PCR, ii) MinION libraries prior to sequencing and iii) to test DNA extracts after MagNA Pure extractions from the host-depletion optimisation experiments.

In brief, 1 µL of template DNA or ladder were added to 10 µL of sample buffer (Agilent Technologies - 5067-5365). Samples were vortexed briefly and placed in the TapeStation. Gels were visualized using the TapeStation analysis software (Agilent Technologies - G2999AA)

2.6 Quantitative polymerase chain reaction assays

Quantitative polymerase chain reaction (qPCR) was used throughout this study to quantify human and microbial DNA before and after host depletion.

Controls were run with every qPCR assay (detailed in sections 2.6.1 and 2.6.2), this included a template negative (where the template was replaced with PCR-grade H_2O) and a process negative (where the template was water processed in the same way as clinical samples - including differential cell lysis with saponin and microbial extraction) control.

All qPCRs were performed using the LightCycler® 480 system (LightCycler® 480 Instrument II cat no 05015278001 Roche).

The Roche master mixes used in this study utilize the FastStart Taq Polymerase modified from the thermostable Taq DNA polymerase for a hot-start PCR. High temperature-activated polymerases prevent non-specific primer binding, hence inhibiting non-targeted amplification providing higher specificity and sensitivity of the reactions.

Sequences of primers and probes (and their gene targets) used in this study can be found in Table 2.3.

2.6.1 Probe based qPCR assays

All probe-based qPCR master mixes consisted of the 2x master mix (LightCycler® 480 Probes Master cat no 04707494001 Roche). This master mix utilizes the FastStart Taq Polymerase (modified from the thermostable Taq DNA polymerase), for a hot-start PCR.

Probe-based qPCR was performed to quantify human and microbial (*Candida albicans, Escherichia coli, Haemophilus influenzae, Klebsiella pneumoniae, Pseudomonas aeruginosa,* *Moraxella catarrhalis, Staphylococcus aureus, Streptococcus pneumoniae* and *S. pyogenes*) DNA. The qPCR master mix contained (total volume of 20 µL):

- 3.6-6.6 µL of PCR-grade H_2O
- 10 µL master mix (2x)
- 0.5 µL 10 µM forward primer (final conc. 0.25 µM)
- 0.5 µL 10 µM reverse primer (final conc. 0.25 µM)
- 0.4 µL 10 µM hydrolysis probe (final conc. 0.2 µM)
- 2-5 µL template DNA
- PCR-grade H_2O to make up the 20 µL total volume

The qPCR conditions were as follows:

- pre-incubation: 95 °C 5 min

- amplification: 95 °C 30 sec

55 °C 30 sec } 40 cycles

72 °C 30 sec

- final extension: 72 °C for 5 min.

For all LoD experiments, 45 cycles were used instead to provide accurate Cq measurements at >35 Cq. The qPCR conditions described above were used throughout this study, except for confirmatory qPCR, for which reaction conditions were taken from Fukumoto *et al.* (98) which were:

- pre-incubation: 95 °C 15 min

- amplification: 94 °C 15 sec
60 °C 1 min } 40 cycles

2.6.2 SYBR green based qPCR assay

All SYBR Green based qPCR assays consisted of the 2x SYBR Green master mix (LightCycler® 480 SYBR Green I Master cat no: 04 707 516 001 Roche) and were used to detect and quantify universal bacterial and *Stenotrophomonas maltophilia* DNA before and after host depletion. Universal bacterial DNA detection was achieved using the 16S rRNA V3-V4 gene fragment and the 23S rRNA gene was used for *S. maltophilia* detection.

For all SYBR green based qPCR assays, the master mix consisted of (total volume of 20 µL):

- 6 µL of PCR-grade H_2O
- 10 µL master mix (2x)
- 1 µL 10 mM forward primer (final conc. 0.5 µM)
- 1 µL 10 mM reverse primer (final conc. 0.5 µM)
- 2 µL template DNA

PCR conditions were as follows:

- pre-incubation: 95 °C 5 min,

- amplification: 95 °C 30 sec
55 °C 30 sec } 40 cycles
72 °C 30 sec

- final extension: 72 °C 5 min

- Melt curve analysis was performed at 95 °C for 5 sec then 65 °C for 1 min (ramping to 95 °C at 0.03 °C/s in continuous acquisition mode) followed by cooling to 37 °C.

Table 2.3: qPCR Primer sequences and gene targets

Organism	Gene target	Forward primer (5’-3’)	Reverse primer (5’-3’)	Probe (5’-3’)	Reference
Human	RNA polymerase A	TGAAGCCGTGCGGAAGG	ACAAGAGAGCCAAGTGTCG	[6FAM]TACCACGTCATCTCCTTTGATGGCTCCTAT[BHQ1]	Designed in house
Universal Bacterial	16S rRNA gene V3-V4 fragment	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC	Sybr Green Master Mix	(244)
E. coli	*cyaA*	CGATAATCGCCAGATGGC	CCTAAGTTGCAGGAGATGG	[6FAM]TAGAGCGCCTTCGGTGTCGGT[BHQ1]	Designed in house
H. influenzae	*omp P6*	AGCGGCTTGTAGTTCCTCTAACA	CAACAGAGTATCCGCCAAAAGTT	[6FAM]CGATGCTGCAGGCAATGGTGCT[BHQ1]	(98)
K. pneumoniae	*mdh*	CGGGCGTAGCGCGTAA	GATACCCGCATTCACATTAAACAG	[6FAM]CCCGGCATGGATCGTTCCGA [BHQ1]	(98)
M. catarrhalis	*copB*	GGTGAGTGCCGCTTTTACAAC	TGTATCGCCTGCCAAGACAA	[6FAM]TGCTTTTGCAGCTGTTAGCCAGCCTAAG[BHQ1]	(98)
P. aeruginosa	*oprL*	AGCCTTCCTGGTCCCCTTAC	CCTAATGAACCCCAGTGTATAAGTTTG	[6FAM]TGAACTGACGGTCGCCAACGGTT[BHQ1]	(98)
S. aureus	*Eap*	ACTGTAACTTTGGCACTGG	GCAGATACCTCATTACCTGC	[6FAM]ATCGCAACGACTGGCGCTA[BHQ1]	Designed in house
S. maltophilia	23S rRNA	GCCGAAAGCCCAAGGTTT	CGACTTTCGTCCTCGCCTTA	Sybr Green Master Mix	(98)
S. pneumoniae	*Ply*	GCTTATGGGCGCCAAGTCTA	CAAAGCTTCAAAAGCAGCCTCTA	[6FAM]CTCAAGTTGGAAACCACGAGTAAGAGTGATGAA[BHQ1]	(98)
S. pyogenes	*sdaB*	GGRACACGTACCCAAAATGTAGGA	TCTTGAGCTCTTTGTTCGGTRTAG	[6FAM]CGTGACCAAAAAGGCGGCATGC[BHQ1]	(98)

C. albicans	5.8S rRNA	GGTTTGGTGTTGAGCAATACGA	AAGCGATCCCGCCTTACC	[6FAM]TGGGTTTGCTTGAAAGACGGTAG[BHQ1]	(245)

2.7 Library preparation and MinION sequencing

Sequencing libraries for singleplex and multiplex runs were prepared using the ONT low-input kits rapid kits (SQK-RLI001, SQK-RLB001, SQK-RPB004) with some modifications to the manufacturer's instructions (modifications are detailed throughout section 3.1).

The manufacturer's instructions for singleplex sequencing with the SQK-RLI001 kit were as follows:

Fragmentation/Tagmentation reaction:

- FRM: 2.5 µL
- Template DNA: ≤ 7.5 µL (≥ 10 ng)
- Nuclease free water (NFW): <7.5 µL (to make up the 10 µL volume)

The reagents were mixed by gentle flicking of the tube and were incubated at 30 °C for 1 min and at 75 °C for 1 min.

PCR reaction was then performed on the tagmented DNA according to the manufacturer's instructions:

- 14 µL of nuclease free water (NFW)
- 1 µL primer
- 25 µL of 2x Long Amp Taq Polymerase (New England Biolabs – M0533S)
- 10 µL tagmented DNA

The recommended PCR reaction conditions were:

Initial denaturation 95 °C for 3 min, cycling conditions 14 cycles: denaturation at 95 °C for 15 sec, annealing at 56 °C for 15 sec, elongation at 65 °C for 6 min and final extension at 65 °C for 6 min.

Multiplexed sequencing libraries were prepared using multiple iterations of the ONT rapid barcoding kit SQK-RLB001 and SQK-RPB001. This section will only describe the manufacturer's instructions, modifications tested on these kits are described in section 3.1.

Initially, the tagmentation/fragmentation reaction was set up as follows:

- FRM: 1 µL
- Template DNA: ≤ 4 µL (= 5 ng)
- NFW: < 4 µL (to make up the 5 µL volume)

The reagents were mixed by gentle flicking of the tube and were incubated at 30 °C for 1 min and at 75 °C for 1 min (SQK-RLB001) or at 80 °C (SQK-RPB004) for 1 min.

Then PCR reaction was set up for each sample separately as per manufacturer's instructions:

- 20 µL of nuclease free water (NFW)
- 1 µL of rapid barcode primer (RPB1-12A)
- 25 µL of 2x Long Amp Taq Polymerase (New England Biolabs – M0533S)
- 4 µL template DNA

The recommended PCR cycling conditions were the same as the SKQ-RLI001 kit described above.

Following the PCR reaction/s for singleplex and multiplex libraries, amplicon products were then subjected to a 0.6x AMPure XP (Beckman Coulter-A63881) bead wash. During the wash, a 0.6:1 ratio of beads to DNA was added to the amplified DNA (i.e. 60 µL of beads and 100 µL

PCR product). Samples were mixed by pipetting and were incubated on the Hula Mixer (parameters: rotation speed 15 & timer 10; reciprocal tilting turning angle 45° & timer 5; vibration turning angle 5° & timer 5) for 5 min. Then samples were spun down (pulse) and were placed on a magnetic tube rack for 5 min to pellet the beads. The clear solution was then carefully removed and beads were washed with 500 µl ethanol (70%) for 30 sec. Ethanol was carefully removed and the wash was repeated. After the two ethanol washes, ethanol was carefully removed and tubes were left to air-dry on the magnetic rack with the cap open for 2-3 min. 14 µL of the elution buffer (10 µl 50 mM NaCl, 10 mM Tris.HCl pH8.0) was added and incubated at room temperature for 5 min to elute DNA. Tubes were then placed back on the magnetic rack for 5 min to separate beads from the DNA. The eluted DNA was transferred to a clean tube and was prepared for MinION sequencing (described below).

The MinION flow cell (either R 9.4.1 (FLO-MIN106 Oxford Nanopore Technologies) / R 9.5 / R. 9.4) was then inserted into the MinION device and the dry quality control (QC) step was done, by double clicking on “check flowcell” on the MinKNOW GUI (ONT). During the dry QC, the MUX scan begins, during which the flow cells pores are assessed and divided into four groups.

After dry QC, the MinION library was prepared for sequencing as follows:

1 µL Rapid Adapter (RAD for SQK-RLI001, RPR for SKQ-RLB001 and RPD for SQK-RPB004) was added into 10 µL of bead-washed PCR products (consisting of 50-300 fmol). Solution was then mixed by gentle flicking and was incubated at room temperature for 5 min.

After adapter ligation the library was prepared for MinION sequencing. The following reagents were thawed and mixed, then added in the following order with a final volume of 75 µL and mixed gently by pipetting (SQK-RPB004):

- 34 µL of sequencing buffer (SQB)
- 25.5 µL of loading beads (LB) which were mixed just before use
- 4.5 µL nuclease free water (NFW)
- 11 µL of adapted DNA library

(Libraries prepared using SQK-RLI001 and SQK-RLB001 used 35 µL of RBF (not SQB) and 3.5 µL of NFW instead).

The library was then stored on ice, until it was loaded into the flow cell. The flow cell was then primed according to the manufacturer's instructions, prior to loading the library. Firstly, the bubble was removed from the flow cell, by using a P1000 pipette, the tip was inserted into the priming port and a small volume of buffer was removed (<50 µL). Then through the same port 800 µL of the priming mix (consisting of 1ml of Flush Buffer (FB) mixed with 30 µL Flush Tether (FLT) for SQK-RPB004; 480 µl of running buffer (RBF) and 520 µl of NFW for SQK-RLI001 and SQK-RLB001) was added, with care to avoid the introduction of any air bubbles.

After 5 min, the SpotON sample port was gently opened and 200 µL of priming mix was added in the priming port as described above. Then the mixed DNA library was added via the SpotON sample port in a dropwise fashion in order to ensure all the library was loaded into/on to the flow cell.

After the library was loaded, a new experiment was set up, in MinKNOW. Experiment name was added, kit and sequencing duration was selected before the sequencing run was started.

Once the flow cells temperature had reached 34 °C, a MUX scan was performed automatically, known as the wet QC, to check the pores after addition of the library. The MinION was run for 24-48 hrs.

2.8 Bioinformatics analysis

Initial sequence processing was performed using the MinKNOW software (versions 1.4 - 18.12.9). This software was used to operate the MinION device but also allowed: i) raw data acquisition in real-time in FAST5 format and ii) basecalled raw data in real-time (raw FAST5 files converted to base called FAST5 or FASTQ files). However, for this study, offline base calling was performed using Albacore (versions 1.2.2-2.3.4) or Guppy (versions 2.1.3-3.2.1 (Guppy was used for base calling raw data for all *Legionella* samples only as Albacore was by this time discontinued)), offline tools provided by ONT.

The command used to operate Albacore was:

```
"c:/Program Files/OxfordNanopore/ont-albacore/read_fast5_basecaller.exe" -f
FLO-MIN106 -k sequencing-kit-number --barcoding -o fastq --input
'path_to_input_folder' -s 'path_to_output' -r -t 4
```

The command used to operate Guppy was:

```
guppy_basecaller --input_path 'path_to_input_folder' --recursive --save_path 'path_to_output_folder' --
flowcell FLO-MIN106 --kit sequencing-kit-number
```

The output format used for downstream analysis was FASTQ reads. For the pilot study the first 24 000 reads were used for this analysis, for the optimised and INHALE study the data produced within the first 2 hours of sequencing were analysed and for the *Legionella* study data after 24 hrs of sequencing were used.

2.8.1 Human read removal

Human reads were filtered out from FASTQ files using minimap2 (v2.6-2.10) to align to the human hg38 genome (GCA_000001405.15 “soft-masked” assembly) prior to EPI2ME and downstream analysis for pilot and streamline samples. Only unassigned (non-human) reads were exported to a bam file using Samtools (-f 4 parameter) and were converted back to FASTQ format using bam2fastx. These FASTQ files were processed for pathogen identification using WIMP (1.137-3.3.1), antibiotic resistance gene detection with ARMA (1.136-1.1.5) and for downstream offline data analysis (described in 2.8.4)

2.8.2 Real-time pathogen identification

The EPI2ME desktop agent provided by ONT (versions 2.47-2.59.1896509) was used for initial data analysis for pilot, streamline and INHALE samples. The Antimicrobial Resistance pipeline, available on EPI2ME, was used for pathogen identification and antimicrobial resistance (AMR) gene detection (AMR gene detection is described below in 2.8.3). For pathogen identification, this pipeline utilizes WIMP which enables microbial identification including: bacteria, viruses, fungi, archaea and human reads (described in 1.4.1).

Parameters used for pathogen identification were:

- Minimum basecalling quality score: 7 (default of EPI2ME)
- WIMP alignment q-score: >19 (available in the csv file)
- Bacterial classified reads to be ≥1% of microbial reads

2.8.3 Real-time AMR gene detection

For the detection of antibiotic resistance genes for streamline and INHALE samples, ARMA (Antimicrobial Resistance Mapping Application – versions 1.136-1.1.5) was used. ARMA is also part of the Antimicrobial Resistance pipeline (described before in 1.4.1). AMR genes are identified by mapping reads using minimap2 against the 'CARD' database (204). Alignments over ONT-chosen default thresholds (>75% accuracy and >40% horizontal coverage) are only reported.

Antibiotic resistance genes were recorded if >1 gene alignment was present using the 'clinically relevant' parameter (described in 1.4.1) available in ARMA (rev. 1.1.5). As previously described this feature was designed by ONT in a collaboration with David Livermore, Vicky Enne and

Justin O’Grady to allow rapid identification of clinically relevant AMR genes and exclude reporting of resistance genes that would not provide useful information for designing antibiotic treatment.

Comprehensive manuals explaining WIMP and ARMA are publicly available on the ONT website (https://nanoporetech.com/EPI2ME-amr).

2.8.4 Offline data analysis

Offline tools were used for the downstream analysis of the sequencing data. Downstream analysis included, pathogen identification using Centrifuge(198) and Supernatant for the *Legionella* study using default parameters (described below) and genome assemblies using 2 hrs and 48 hrs of sequencing data of streamline samples and 24 hrs of sequencing data for the *Legionella* study (described below). Offline data analysis was also performed for species specific gene analysis and timepoint analysis using 2 hrs of sequencing data.

2.8.4.1 Bacterial genome assembly

Reference-based genome assemblies were generated from metagenomic data as follows: Firstly, using Albacore, raw FAST5 reads were basecalled to FASTQ reads. Reads shorter than 2000 bp and reads with a quality score <7 were filtered out using the Fastq-to-Fastq script within the Fast5-to-Fastq tool (https://github.com/rrwick/Fast5-to-Fastq). Porechop was then used to remove sequencing adapters located in the middle and/or at the end of DNA sequences and for multiplex runs, re-identification of barcodes was performed using the –b parameter (v0.2.3) (https://github.com/rrwick/Porechop). Next, minimap2(246) (v2.6-2.10) was used to map reads to a reference-genome (the reference-genome chosen was the strain of the pathogen with the most aligned reads reported by the EPI2ME AMR pipeline), using the default parameters for nanopore data (-a -x map-ont). Finally, Canu (220, 247)was used to generate a genome assembly of the aligned reads, using the default parameters (v1.6). Comparison of the assemblies was performed using BLAST Ring Image Generator (BRIG) (248).

Raw sequencing data (FAST5 reads) generated from the *Legionella* study (described in 3.3) were initially basecalled to FASTQ reads using Guppy and were used for *de novo* or reference-based genomic assemblies. For the *de novo* approach, initially *Legionella* spp. reads were classified using the basecalled FASTQ reads with Centrifuge (centrifuge_index_oct2018 was used with default parameters). Then using Supernatant, Centrifuge-classified *Legionella* spp. reads with a centrifuge score >300 were extracted and used for genomic assemblies with Canu (described above) and for pathogen identification.

For the reference-based approach, basecalled FASTQ reads were mapped against a concatenated reference containing all complete genomes of *L. pneumophila* available on NCBI using minimap2 as described above. Aligned reads were then used to generate a genome assembly using Canu as described above.

2.8.4.2 Species-specific gene analysis and timepoint analysis

Species-specific gene alignments were performed throughout this study, to confirm the presence or absence of organisms identified by the metagenomic analysis which were not previously reported by culture. Genes used for this analysis are specific for the bacterial species in question and were chosen from a literature search of targets used in peer-reviewed qPCR assays for pathogen/s in interest (Table 2.4). For this analysis, reads from the first two hours of sequencing for the streamline samples (after human DNA removal for the streamline samples only) and the INHALE samples were aligned to species-specific genes.

For streamline samples this analysis was carried out for samples positive for *H. influenzae* or *S. pneumoniae* by metagenomics only (culture-negative for these pathogens). For INHALE samples this analysis was done for any pathogen identified by metagenomics but not identified by culture or either of the two multiplex PCR platforms.

This analysis was performed only for pathogens that had read numbers above the chosen thresholds for pathogen identification (described in 2.8.2). Minimap2 was used to generate alignments as described above and the number of mapped reads were visualized using qualimap. If a sample contained >1 alignment of the specific gene tested, then it was considered as a true positive sample for that pathogen.

Table 2.4: Species-specific genes and their targets

Pathogen	**Gene**	**Encoded Protein**	**Accession number**
Escherichia coli	*cyaA*	Adenylate cyclase	NC_000913.3
Streptococcus agalactiae	*cfb*	CAMP-factor	NC_004116.1
Streptococcus pneumoniae	*ply*	pneumolysin	NC_003098.1
Haemophilus influenzae	*siaT*	Sialic acid TRAP transporter permease	DQ054471.1
Staphylococcus aureus	*eap*	Extracellular adherence protein	AGY90050.1
Stenotrophomonas maltophilia	*smeT*	transcriptional regulator of SmeDEF efflux pump	AY450955.1
Pseudomonas aeruginosa	*oprL*	Peptidoglycan-associated protein	Z50191.1
Klebsiella pneumoniae	*mdh*	Malate dehydrogenase	ACI09474.1
Klebsiella oxytoca	*pheX*	Polygalacturonase	AAL49975.1
Klebsiella aerogenes	*atpD*	F-ATPase b-subunit	AX110938.1

2.8.4.3 Multi-locus sequence typing analysis

A *Legionella pneumophila* typing scheme containing 2837 *L. pneumophila* ST and all known alleles IDs of the seven *L. pneumophila* housekeeping genes - *flaA*, *pile*, *asd*, *mip*, *mompS*, *proA*, *neuA*, was used for the MLST analysis (scheme generated by Natalie Groves, PHE).

Metagenomic sequencing data from *L. pneumophila* positive samples were used for multi-locus sequence typing (MLST) analysis either using Mlst (https://github.com/tseemann/mlst) or Krocus (249). For the Mlst tool (v.2.x), genome assemblies generated (*de novo* or reference-based described in 2.8.4.1) were used to determine the pathogen's sequence type (ST) using default parameters. For Krocus, *L. pneumophila* FASTQ reads either classified by Centrifuge or identified by mapping against the concatenated reference (both described in 2.8.4.1) were used for sequence-based typing (SBT) with default parameters.

2.9 Declaration of contribution

In this study everything, including samples processing and data analysis was carried out by myself, Themoula Charalampous, except for the following:

- Antibiotic resistance gene analysis for streamline samples was done by Professor David Livermore.
- Time-point analysis of streamline sample set (S1 and S16) was done by Dr. Gemma Kay (described in 2.8.4.2)
- Human depletion and sequencing of 23 samples included in the *Legionella* study were performed by Jessica Day (PHE).
- The molecular pipeline (described in 2.8.4.3) including Supernatant (described in 2.8.4) used for MLST analysis for the *Legionella* study was designed by Natalie Groves (PHE) and data analysis for some of the samples was performed by Graeme Smith (PHE and Viapath).

3.1 Results

3.1 Development, optimization and testing of a clinical metagenomics pipeline with a host depletion method for the diagnosis of LRTIs

A purulent sputum sample typically contains about 1 million leukocytes per mL and a pathogen load ranging anywhere from 10^3-10^9 CFU/mL. Therefore, at best, the human:pathogen DNA ratio is approx. 1:1 and at worst is approx. 10^3:1. Hence, as previously discussed, in order for implementation of clinical metagenomics to be feasible in terms of cost and time, host depletion is necessary, to allow detection of pathogens and resistance genes in a rapid timeframe using metagenomic sequencing.

According to the literature saponin had been mostly used for red blood cell (RBC) lysis, as it forms pores by interacting with cholesterols present on RBCs. Recently the lytic abilities of saponins have also been tested on other human cells such as leukocytes. Various saponin-based methods have been developed for the lysis of human cells followed by the depletion of human DNA, however the method originally developed by Zelenin *et al.* (177) and modified by Anscombe *et al.* (178) appeared the most efficient and was chosen for optimisation and testing in this study.

3.1.1 The effect of different saponin concentrations and incubation times on host depletion

Firstly, we aimed to test the saponin-based host depletion as published to assess its depletion efficacy on respiratory samples. The saponin-based depletion method was performed as follows: Sputasol-treated sputum samples (250 µL) were centrifuged at 8000 xg for 5 min, after which the supernatant was carefully removed and the pellet resuspended in 250 µL of PBS. Saponin (Tokyo Chemical Industry- S0019) was added to a final concentration of 1.43% (100 µL of 5 % saponin), mixed well and incubated at room temperature (RT) for 3 min to promote host cell lysis. Following this incubation, 350 µL of water and 10.5 µL of 5 M NaCl was added to deliver an osmotic shock, lysing the damaged host cells. Samples were next centrifuged at 4000 xg for 5 min, with the supernatant removed and the pellet resuspended in 43 µL of PBS. 5 µL of 10X Turbo DNase buffer (ThermoFisher – AM2238) was added with 2 µL Turbo DNase (ThermoFisher – AM1907) and incubated for 15 min at 37 °C. Finally, the host-DNA depleted samples were washed three times with decreasing volumes of PBS (250 µL, 100 µL, 50 µL). After each wash, the sample was centrifuged at 6000 xg for 3 min, the supernatant was discarded and the pellet was resuspended in PBS. This was followed by nucleic acid extraction and purification was followed as described in section 2.4. DNA quantification and quality control was performed after extraction (described in 2.5). Also, host depletion and bacterial loss/gain were monitored using qPCR assays (described in 2.6). These steps were always performed unless otherwise stated.

This version of saponin-based protocol by Anscombe *et al.* (178) was initially tested on four excess sputum samples. Undepleted controls (where DNA extraction was performed after initial spin without host depletion), were included to determine the level of host depletion (sputum samples used for the optimization of the host depletion are referred to as test samples – T). Host depletion was observed at ≥5.8-fold in T1 and T2 but was <2-fold in T3 and T4 samples (Table

3.1). To further increase this level of host depletion, increased saponin concentration (10%) and various saponin incubation times (3, 5 or 10 min) were tested. Also the following modifications were made to the protocol described above:

- 400 µL of sputasol-treated sputum sample were processed instead of 250 µL to increase microbial yield
- 200 µL of saponin was used instead of 100 µL in order to be 1:1 ratio of sample to saponin

A ~5 fold of host depletion was observed with 4.44% saponin as the final concentration but the saponin concentration which gave the greatest host depletion (>74-fold) with no bacterial loss, was 2.22% saponin final concentration (Table 3.2). No significant microbial loss was observed with either of the saponin incubation times (5 or 10 min). However, some bacterial loss was observed in the samples tested with a 4.44% final saponin concentration. Hence, for all further host depletion experiments, the longer incubation time (10 min) with 2.22% saponin was chosen, in order to ensure full digestion of the free human DNA.

Table 3.1: Human and bacterial DNA qPCR results for sputum samples processed with the original saponin-based protocol.

Sample	**Human qPCR assay (Cq)**	**Human DNA depletion (ΔCq)**
T1-Undepleted control	20.61	2.92 (7.6 fold)
T1-Depleted	23.53	
T2-Undepleted control	20.11	2.56 (5.8 fold)
T2-Depleted	22.67	
T3-Undepleted control	23.28	0.25 (1.18)
T3-Depleted	23.53	
T4-Undepleted control	27.84	0.66 (1.58)
T4-Depleted	28.50	

Table 3.2: Human and bacterial DNA qPCR results for sputum samples processed with different saponin concentrations and incubation times.

Sample	**Human qPCR assay (Cq)**	**Human DNA depletion (ΔCq)**	**16S rRNA gene V3-V4 fragment qPCR assay (Cq)**	**Bacterial DNA loss after host depletion (ΔCq)**
Undepleted control	23.52	-	21.93	-
Original saponin depletion (1.43% + 3 min)	24.28	0.76 (1.7 fold)	22.79	0.86 (1.81)
Saponin depletion (2.22% + 5 min)	29.82	6.3 (78.8 fold)	22.28	0.35 (1.27)
Saponin depletion (2.22% +10 min)	29.72	6.2 (73.5 fold)	22.08	0.15 (1.1)
Saponin depletion (4.44% + 5 min)	26.78	3.26 (9.57 fold)	25.89	3.96 (15.5 fold)
Saponin depletion (4.44% + 10 min)	25.92	2.4 (5.27 fold)	23.16	1.23 (2.34 fold)

3.1.2. Optimisation of the nuclease treatment

The nuclease treatment was optimized to increase host DNA depletion efficiency and remove digested/degraded host DNA. Initial method development utilized Turbo DNase but previous research (within the O’Grady group) had shown HL-SAN DNase was more efficient and robust when using clinical samples. Hence, the HL-SAN DNase was tested against the Turbo DNase in two sputum samples (T5 and T6) and was further tested on three additional sputasol-treated sputum samples (T7, T8 and T9) (Table 3.3). For the HL-SAN DNase treatment the following conditions were followed:

- 5 µL of HL-SAN DNase was added with 100 µL of PBS and 100 µL of HL-SAN buffer (5.5 M of NaCl and 100 mM $MgCl_2$ in 50 ml of H_2O) and samples were incubated at 37 °C for 15 min shaking at 800 RPM.

Also, along with the change of DNase, only one PBS wash was performed instead of three to streamline the method as follows:

- 300 µL of PBS was added and centrifuged at 6000 xg for 3 min

DNA extraction and purification was followed as described in 2.4 (for pilot samples).

The use of HL-SAN resulted in better depletion compared to the Turbo DNase in both samples (Table 3.3). Bacterial loss (7-fold) was observed in T5 but very little loss was observed in the other samples. The removal of host nucleic acid was variable amongst the three remaining sputa - 209.38 fold depletion of human DNA was observed in T7 but only 8.6 fold difference was observed in T9 (Table 3.3). Despite the variable results, due to the improved efficiency of host DNA depletion observed, the HL-SAN DNase was chosen and used for all further DNase treatments.

Table 3.3: Human and bacterial DNA qPCR results of processed samples with the HL-SAN and Turbo DNase.

Sample	Human qPCR assay (Cq)	Human DNA depletion (ΔCq)	16S rRNA gene V3-V4 fragment qPCR assay (Cq)	Bacterial DNA loss/gain after host depletion (ΔCq)
T5-Undepleted control	26.56	-	32.19	-
T5-Depleted (Turbo)	32.57	6.01 (64.4 fold)	35	2.81 (7 fold)
T5-Depleted (HL-SAN)	35	8.44 (347.3 fold)	35	2.81 (7 fold)
T6-Undepleted control	24.04	-	21.48	-
T6-Depleted (Turbo)	25.77	1.73 (3.31 fold)	22.4	0.92 (1.89 fold)
T6-Depleted (HL-SAN)	29.34	5.3 (39.4 fold)	22.1	0.62 (1.53)
T7-Undepleted control	24.13	7.71 (209.38 fold)	23.92	1.2 (2.29)
T7-Depleted (HL-SAN)	31.84		22.72	
T8-Undepleted control	29.67	5.33 (40.2 fold)	22.96	0.04 (1.02)
T8-Depleted (HL-SAN)	35		23	
T9-Undepleted control	27.08	3.11 (8.6 fold)	22.64	0.57 (1.48)
T9-Depleted (HL-SAN)	30.19		23.21	

Next, we aimed to further improve host depletion without affecting bacterial cells. Therefore, we optimised different steps of the depletion method as follows:

- After initial 15 min DNase treatment, 2 µL of HL-SAN DNase was added and sample was incubated for a further 15 min at 37 °C with shaking at 800 rpm. The second DNase treatment was added to improve digestion of human nucleic acid.

- The three PBS washes (with higher volumes) were re-introduced after the DNAse treatment as based on experiments above we believed the addition of more than one washing step would facilitate to the removal of more host nucleic acid. For this step sample was washed three times with decreasing volumes of PBS (300 µL, 150 µL, 50 µL). After each PBS wash, sample was centrifuged at 6000 xg for 3 min, supernatant was carefully removed and pellet was re-suspended in decreasing volumes of PBS. DNA was extracted and purified as described in 2.4 (for pilot samples).

These changes to the depletion protocol were initially tested separately on two sputum samples - T10 and T11 (Table 3.4). Efficiency of host depletion was compared against the version of the method used for T7-T9 samples (with HL-SAN DNase). The addition of the three washing steps had a minor improvement in host depletion in T11 (25.8 fold versus 8.8 fold with the previous version) but it was less efficient in T10 when compared with the previous version (78.2 fold versus 184.8 fold). The extended DNAse treatement also showed a similar trend in the two samples tested when compared with the older version – an improvement was observed in T11 (53.82 fold vs 8.8) but not in T10 (153.27 fold vs 184.8 fold) (Table3.4). We then combined these two steps and tested on two more sputum samples (T12 and T13) to see if their combination could increase the removal of host nucleic acid (Table 3.4). The combination of the extended DNase treatment and the washing steps increased host depletion up to 99.9% or ~10^3 fold without any bacterial loss. In T12 a >349.7 fold of host depletion was recorded and a 1871.5 fold depletion was observed in T13 (Table 3.4). This version of the method was chosen to move forward with.

Table 3.4: Human and bacterial DNA qPCR results for respiratory samples processed with an extended DNase treatment and washing steps

Sample	Human qPCR assay (Cq)	Human DNA depletion (ΔCq)	16S rRNA gene V3-V4 fragment qPCR assay (Cq)	Bacterial DNA loss/gain after host depletion (ΔCq)
T10-Undepleted control	24.85	-	19.53	-
T10-Depleted	32.38	7.53 (184.8. fold)	20.64	1.11 (2.15 fold)
T10-Depleted (2 DNase treatments)	32.11	7.26 (153.27 fold)	20.63	1.1 (2.14 fold)
T10-Depleted (3 washes)	31.14	6.29 (78.2 fold)	20.73	1.2 (2.3 fold)
T11-Undepleted control	23.85	.	23.58	.
T11-Depleted	26.99	3.14 (8.8 fold)	23.53	0.05 (1.03)
T11-Depleted (2 DNase treatments)	29.60	5.75 (53.82 fold)	23.20	0.38 (1.3)
T11-Depleted (3 washes)	28.54	4.69 (25.8 fold)	22.66	0.92 (1.9)
T12-Undepleted control	26.55	>8.45 (>349.7 fold)	22.02	0.26 (1.19)
T12-Depleted (combined*)	>35		22.28	
T13-Undepleted control	22.69	10.87 (1871.5 fold)	21.57	0.59 (1.5)
T13-Depleted (combined*)	33.56		22.16	

*combined= two DNase treatments and three washed were added in the pipeline

3.1.3 Optimization of low input library preparation to enable sequencing of low biomass clinical samples

Clinical samples after host depletion have very low DNA concentrations (often <0.01 ng/µL), as the majority of human nucleic acid is removed and mainly only microbial DNA remains. When this research was being performed, the library preparation kits available by ONT required a high amount of input DNA (>200 ng). However, in early 2017 ONT released a low input kit (that followed a similar principal as the Nextera XT DNA library prep kit) that required a minimum 10 ng of DNA, which enabled sequencing of low-biomass samples. The first version of the protocol, which enabled low-input DNA singleplex sequencing (SQK-RLI001), included a tagmentation step, during which DNA was enzymatically fragmented and tagged simultaneously. Long range PCR was then performed using a primer complementary to the tag added during tagmentation step.

Samples T10 (3.86 ng/µL) and T11 (2.96 ng/uL) were prepared for singleplex sequencing using the RLI001 kit to test if sequencing was possible after host depletion, according to the manufacturer’s instructions (as described in 2.7) except PCR cycles were increased from 14 to 20 to increase yield and sensitivity. Sequencing was successful for both samples that were processed with the altered SQK-RLI001 workflow producing >1.2 million reads for T10 and >1.8 million reads for T11 (>99% were passed reads in both samples) after 48 hrs of sequencing (Table 3.5). The proportion of microbial reads was high, indicating successful host depletion.

Table 3.5: Sequencing data of samples prepared with the SQK-RLI001 kit

Sample No	**Total reads produced after 48 hrs**	**Total passed[*] reads after 48 hr**	**Total failed[*] reads after 48 hr**	**Total human pass reads**	**Total microbial pass reads**	**Average Length (Kb)**
T10-Depleted (2 DNase treatments)	1,281,019	1,278,326 (99.78%)	2,693 (0.2%)	3,393	1,028,110 (80.25%)	1.1
T11-Depleted (3 washes)	1,838,272	1,837,968 (99.98%)	304 (0.016%)	7,499	1,051,858 (57.2%)	1.4

*passed reads had ≥7 quality score (Q score) and failed reads had Q score of <7.

Singleplex metagenomic sequencing of clinical samples is not cost-effective as the flowcell cost is high (min £400). Sequencing multiple samples on a single flowcell, therefore, would significantly decrease overall cost. Shortly after the release of RLI001, ONT released the SQK-RLB001 kit, which allowed multiplex sequencing (up to 12) of low-biomass samples. It works using the same principal as the RLI001 kit, with the addition of barcodes to the primers used to amplify the tagmented library. The barcodes are produced with click chemistry at the 5' ends to enable rapid sequencing adapter attachment.

We then tested the SQK-RLB001 kit for multiplex sequencing of respiratory samples. For initial testing, DNA from depleted samples T14, T15, T16 was prepared using the both RLI001 and RLB001. DNA concentrations of PCR products using the multiplex kit were between 0.6-2.56 ng/µL in contrast to the singleplex kit which gave yields of 7-23 ng/µL using the same amount of input DNA (≤10 ng) (Table 3.6).

Table 3.6: DNA concentrations of samples prepared with the SQK-RLI001* and SQK-RLB001* kits

Sample	**DNA concentration pre-PCR (ng/µL)**	**DNA concentration post-PCR with SQK-RLI001(ng/µL)**	**DNA concentration post-PCR with SQK-RLB001 (ng/µL)**
T14	6.9	7.72	1.56
T15	0.218	9.44	0.672
T16	10.7	23.6	2.56

*SQK-RLB001 is the multiplex kit and SQK-RLI001 is the singleplex kit

Hence, in order to increase the sensitivity of the multiplex PCR reaction using the rapid barcoding kits (SQK-RLB1001 and later SQK-RPB004), the following changes to the protocol were made:

- **the number of cycles was increased from 20 to 25**
- **2.5 µL of tagmentation enzyme (FRM) instead of 1 µL was added and volume of the tagmentation reaction increased from 4 µL to 10 µL (as in RLI001).**
- **the volume of the PCR reaction was doubled (50 µL of the 2x PCR mix, 2 µL of barcode primer, 10 µL tagmented DNA and 38 µL water) to reduce inhibition caused by the sputum DNA.**
- **a bead-based DNA washing step was introduced prior to library preparation (the same bead wash described in 2.7 except using 1.2x beads to DNA volume) again to reduce PCR inhibition. DNA was eluted in 15 µL water.**

The increase in the number of PCR cycles and reagent volume was first tested separately (using DNA from sputum samples T17, T18, T19) and then these changes were combined with the bead-wash and were tested on five clinical samples (T20, T21, T22, T23,T24). Increasing the number of cycles and reagent volume improved the yield and PCR sensitivity (a 3.6 fold increase in DNA yield was observed in T18) but not in all samples tested (Table 3.7). However, the addition of the bead-washing in combination with the increased number of PCR cycles, FRM and PCR reaction volumes showed the biggest improvement in the multiplex PCR reaction. The optimized method was tested against the SQK-RL001 singleplex PCR on five host-depleted respiratory samples. DNA concentrations of PCR products were between 43-58 ng/µL for the multiplex PCR and 1-18 ng/µL for the singleplex PCR using the same amount of input DNA (Table 3.8).

Table 3.7: DNA concentrations pre and post PCR using SQK-RLB001[*] and SQK-RLI001 kits[^]

Sample	**DNA concentration pre-PCR (ng/µL)**	**DNA concentration post-PCR with SQK-RLI001(ng/µL)**	**DNA concentration post-PCR with SQK-RLB001 (ng/µL)**
T17	12.6	19	1.01
T18	10.3	27.2	98.6
T19	2.28	3.86	0.88

[*]with increased number of cycles and reagents volume
[^]SQK-RLB001 is the multiplex kit and SQK-RLI001 is the singleplex kit

Table 3.8: DNA concentrations of samples prepared with the SQK-RLI001* and SQK-RLB001^ kits+

Sample	DNA concentration pre-PCR (ng/µL)	DNA concentration post-PCR with SQK-RLI001(ng/µL)	DNA concentration post-PCR with SQK-RLB001 (ng/µL)
T20	0.228	3.54	45.6
T21	0.598	1.15	46
T22	2.84	18.6	58.6
T23	1.19	9.1	43.8
T24	3.1	1.05	52.8

*no bead-wash, ^with bead-wash, +SQK-RLB001 is the multiplex kit and SQK-RLI001 is the singleplex kit.

3.1.4 Testing of the pilot clinical metagenomics pipeline

The performance of the optimized host depletion and low-input multiplex library preparation method, i.e. the pilot clinical metagenomics (CMg) pipeline (Figure 3.1), was evaluated in a pilot study on 40 respiratory samples from patients with suspected bacterial LRTIs.

Respiratory samples (400 µL) were centrifuged at 8000 xg for 5 min, after which the supernatant was carefully removed and the pellet resuspended in 250 µL of PBS. Saponin (Tokyo Chemical Industry- S0019) was added to a final concentration of 2.22% (200 µL of 5% saponin), mixed well and incubated at room temperature (RT) for 10 min to promote host cell lysis. Following this incubation, 350 µL of water was added and incubation was continued at RT for 30 s, after which 12 µL of 5 M NaCl was added to deliver an osmotic shock, lysing the damaged host cells. Samples were next centrifuged at 6000 xg for 5 min, with the supernatant removed and the pellet resuspended in 100 µL of PBS. HL-SAN buffer (5.5 M NaCl and 100 mM $MgCl_2$ in nuclease-free water) was added (100 µL) with 5 µL HL-SAN DNase (25,000 units, Articzymes - 70910-202) and incubated for 15 min at 37 °C with shaking at 800 RPM for host DNA digestion. An additional 2 µL of HL-SAN DNase was added to the sample, which was then incubated for a further 15 min at 37 °C with shaking at 800 RPM. Finally, the host-DNA depleted samples were washed three times with decreasing volumes of PBS (300 µL, 150 µL, 50 µL). After each wash, the sample was centrifuged at 6000 xg for 3 min, the supernatant discarded and the pellet resuspended in PBS. After the final wash step of the host depletion, nucleic acid extraction purification was followed as described in section 2.4.

Library preparation was then followed either for singleplex using the SQK-RL001 (described in 2.7, but with 20 cycles were used instead of 14) or for multiplex runs using the SQK-RLB001 (described in 2.7) but with applying the final changes tested (the addition of the 1.2x bead wash after DNA extraction and the increase of PCR cycles and reaction volume - described in 3.1.3). After the PCR reaction, adapter ligation, preparation of library for sequencing and MinION sequencing was followed as described in 2.7.

Data analysis was performed using ~24,000 reads, including base-calling of raw data, human read removal and real-time pathogen detection using the WIMP pipeline as described in 2.8.1-2.8.3. The WIMP parameters described in 2.8.2 were applied for pathogen detection, as initial analysis revealed that thresholds were necessary to improve the accuracy of metagenomic pathogen detection. The chosen parameters for WIMP pathogen identification were used: i) to remove misidentified reads introduced through the pipeline or ii) to remove reads arising from barcode cross-talk in the multiplexed runs.

The pilot CMg pipeline (Figure 3.1) was tested on 40 respiratory samples (34 culture-positive samples and 6 culture-negative samples), from patients with suspected bacterial LRIs previously tested by clinical microbiology (described in 2.3.1). Up to 99.9% or ~10^3 fold (median 352-fold, interquartile range 144-714; maximum 1024-fold) of host nucleic acid was removed using saponin depletion described above, as measured by qPCR (described in 2.6) and the overall turnaround time from sample to result (pathogen identification) was eight hours.

CMg detected the correct pathogen in 31/34 culture-positive samples tested. This included single bacterial infections (27/28) and samples with mixed bacterial infections (4/6). Single bacterial infections reported correctly by metagenomics were: five coliform infections (P1, P5, P6, P7 and P11), two *P. aeruginosa* infections (P22 and P32), seven *H. influenzae* (P8, P9, P24,P25, P27, P29 and P35), six *S. aureus* infections (P15, P16, P23, P39 including two MRSA cases in P10 and P38), two *K. pneumoniae* infections (P12 and P21), three *S. pneumoniae* infections (P30, P33 and P36), one *E. coli* infection (P13) and one *M. catarrhalis* infection (P26) (Table3.9). Mixed bacterial infections correctly identified by metagenomic sequencing were: *K. pneumoniae* and *E. cloacae* in P14 and two *H. influenzae* and *S. pneumoniae* infections confirmed in P28 and P40. Metagenomics was also in agreement with routine microbiology for all of the six culture-negative samples (P2, P4, P17, P18, P19 and P20) as no additional pathogens were identified above our chosen thresholds.

Three pathogens in 3/34 sequenced positive samples were missed by metagenomic sequencing. These included mixed infections in 2/3 samples, where one of the two pathogens present was not detected by the pilot method – specifically, *S. pneumoniae* in P3 and *H. influenzae* in P37 were missed and a reported *S. aureus* missed in P34 (Table 3.9).

In 5/40 sequenced samples additional potential pathogens were detected, but were not previously reported by microbiological culture. These included, *H. influenzae* detected in P22 and P30; *M.*

catarrhalis in P8; *E. coli* in P14 and *K. pneumonia*e and *M. catarrhalis* in P29 (Table 3.9). Based on these results the pilot pipeline was 91.2% sensitive (95% CI; 75.2-97.7%) and 100% specific (95% CI; 54.07-100%) when additional organisms identified in culture-positive samples were *not* considered as false positives (Table 3.9).

Table 3.9: Pilot metagenomic pipeline output compared to routine microbiology culture results.

Sample	**Pathogen cultured by microbiology**	**Pathogen identified from metagenomic pipeline**
P1	Coliform*	*P. mirabilis*
P2	NRF	None
P3	*P. aeruginosa* *S. pneumoniae*	*P. aeruginosa*
P4	NRF	None
P5	Coliform*	*E. coli*
P6	Coliform*	*K. pneumoniae*
P7	Coliform*	*S. marcescens*
P8	*H. influenzae*	*H. influenzae*
		M. catarrhalis
P9	*H. influenzae*	*H. influenzae*
P10	MRSA	MRSA
P11	Coliform*	*E. coli*
P12	*K. pneumoniae*	*K. pneumoniae*
P13	*E. coli*	*E. coli*
P14	*K. pneumoniae* *E. cloacae*	*K. pneumoniae*
		E. cloacae
		E. coli
P15	*S. aureus*	*S. aureus*
P16	*S. aureus*	*S. aureus*
P17	NRF	None
P18	NRF	None
P19	NRF	None
P20	NRF	None
P21	*K. pneumoniae*	*K. pneumoniae*

P22	*P. aeruginosa*	*P. aeruginosa*
		H. influenzae
P23	*S. aureus*	*S. aureus*
P24	*H. influenzae*	*H. influenzae*
P25	*H. influenzae*	*H. influenzae*
P26	*M. catarrhalis*	*M. catarrhalis*
P27	*H. influenzae*	*H. influenzae*
P28	*S. pneumoniae* *H. influenzae*	*S. pneumoniae*
		H. influenzae
P29	*H. influenzae*	*H. influenzae*
		K. pneumoniae
		M. catarrhalis
P30	*S. pneumoniae*	*S. pneumoniae*
		H. influenzae
P31	*E. aerogenes* *S. aureus*	*E. aerogenes*
		S. aureus
P32	*P. aeruginosa*	*P. aeruginosa*
P33	*S. pneumoniae*	*S. pneumoniae*
P34	*S. aureus*	
P35	*H. influenzae*	*H. influenzae*
P36	*S. pneumoniae*	*S. pneumoniae*
P37	*H. influenzae* Coliform*	
		K. oxytoca
P38	MRSA	MRSA
P39	*S. aureus*	*S. aureus*
P40	*H. influenzae* *S. pneumoniae*	*H. influenzae*
		S. pneumoniae

*Coliform not further identified by culture.

3.1.5 Optimization of the clinical metagenomics protocol

Next we aimed to improve the sensitivity (8.8% false negative rate) of the pilot CMg pipeline. Therefore, we sought to improve bacterial cell lysis to ensure difficult-to-lyse pathogens (e.g. *S. aureus*) were not missed, while refining the method to reduce the turnaround time without affecting clinical sensitivity.

The following lysis methods were tested:

- **a bead-beating step - - pelleted samples after the PBS washes were re-suspended in BLB (500 µL), transferred to a bead-beating tube and bead-beaten at maximum speed for 3 min in a Tissue Lyser bead-beater.**
- **the addition of an enzyme cocktail**

The following changes were made for streamlining the host depletion method:

- **the second DNase treatment was removed and one round of DNase treatment was done instead where 10 µL of HL-SAN DNase was added instead and a single 15 min incubation was carried out with at 37 °C**
- **the number of washes was reduced to two with increasing volumes of PBS (800 µL and 1 ml).**

The lysis methods were tested separately and combined with changes for reducing turnaround time. Two culture-positive sputa, one containing *S. aureus* (Gram-positive) and one containing *P. aeruginosa* (Gram-negative) previously processed by routine microbiology (as described in 2.3.1), were used to test the efficiency of the host depletion method and qPCR results were compared to the pilot method (described in 2.6).

Neither pre-treatment (enzymatic cocktail or bead-beating) affected the bacterial DNA yield in the *P. aeruginosa* sample. The enzyme cocktail increased the amount of bacterial DNA in the *S. aureus* sample by approx. 3-fold, and the bead-beating step by 21-fold, compared with the pilot method (Table 3.10), as determined by 16S rRNA qPCR. The increased bacterial yield in the bead-beaten *S. aureus* sample was likely to have been associated with improved lysis of *S. aureus*, as the pathogen dominated the bacterial community (approx. 80% of reads) present in the sample. Also, changes made to streamline the method, reduced turnaround time of the host depletion from 90 to 50 min without affecting human DNA depletion as compared to the pilot method (Table 3.10). Hence, based on these results, the streamlined host depletion method with bead-beating was used for processing of future clinical samples in the optimised method.

Table 3.10: Comparison of bacterial DNA extraction methods using qPCR.

Sample	**Sample type**	**Microbiology result**	**Condition**	**16S rRNA gene V3-V4 fragment qPCR assay (Cq)**	**Bacterial gain to standard depletion (ΔCq)**	**Human qPCR assay (Cq)**	**New DNA depletion compared to standard depletion (ΔCq)**
O1	Sputum	*S. aureus*	Pilot method	18.22	-	27.34	-
			Streamline depletion + enzyme cocktail*	16.66	1.56 (3 fold)	27.22	0.12
			Streamline depletion + bead beating	13.84	4.38 (21 fold)	28.4	1.04 (2 fold)
O2	Sputum	*P. aeruginosa*	Pilot method	12.06	-	27.92	-
			Streamline depletion + enzyme cocktail	12.65	0.59	27.46	0.46
			Streamline depletion + bead beating	12.18	0.12	27.95	0.03

*enzyme cocktail used was MetaPolyzyme (MERCK – MAC4L)

Turnaround was cut down more by reducing the library preparation PCR extension time from six to four minutes. Sensitivity of the PCR reaction with a four minutes extension was compared against the previously used reaction (six minutes extension) as described in 3.1.3. Microbial communities were also determined (as described in 2.8.2) to investigate if the change in the extension time would affect the microbial community after sequencing. No significant changes in the microbial community profile (organisms with ≥0.5% classified reads) were observed between libraries produced with four and six-minute extension times. The only differences observed were in the abundance of minor members of the community and a reduction in average read length for the *S. aureus* sample (~600 bp) (Table 3.11).

Table 3.11 Comparison of the microbial community with different PCR extension times for MinION sequencing.

Sample	**Sample type**	**Microbiology result**	**PCR extension time (min)**	**Average read length (Kb)**	**Bacterial species from WIMP analysis (≥0.5% of classified reads)**	**Percentage of total classified reads (%)**
O1	Sputum	*S. aureus*	4	2.6	*S. aureus* *S. agalactiae* *S. anginosus* group *S. oralis* *Veillonella parvula*	76.1 7.1 2.3 0.9 0.5
			6	3.2	*S. aureus* *S. agalactiae* *S. anginosus* group *S. oralis*	78.7 6.1 1.4 0.9
O2	Sputum	*P. aeruginosa*	4	2.7	*P. aeruginosa* group *S. oralis* *Pseudomonas* *P. stutzeri* group	87.1 1.1 1.0 0.7
			6	2.4	*P. aeruginosa* group *Pseudomonas* *S. oralis*	88.2 1.0 0.6

The changes (described above) for improving bacterial lysis, streamlining host depletion and reducing duration of the PCR reaction of the library preparation were all combined for the final version of the CMg pipeline (the optimised CMg pipeline). The differences in the final version of the optimised CMg pipeline (Figure 3.1) from the pilot pipeline are summarized below:

i) After the first centrifugation, up to 50 µL of supernatant was left for the saponin treatment so as not to disturb the pellet (final saponin conc. 2.2-2.5%).

ii) One round of DNase treatment was done where the amount of HL-SAN DNase was increased to 10 µL and a single 15 min incubation was carried out with the same conditions as before (37 °C with shaking at 800 RPM)

iii) The number of washes was reduced to two with increasing volumes of PBS (800 µL and 1 mL).

iiii) After the final wash, the pellet was re-suspended in 500 µL of bacterial lysis buffer and bead-beaten for 3 min. The bead-beaten sample was centrifuged at 20,000 xg for 1 min and ~230 µL of supernatant was transferred to a fresh tube for DNA extraction.

In total these alterations reduced the metagenomic library preparation to 2.5 hrs with an overall turnaround time of less than four hours before DNA sequencing. Total turnaround to results was approx. 6 hrs including 2 hrs sequencing and real-time analysis using EPI2ME (described in 2.8.1-2.8.2).

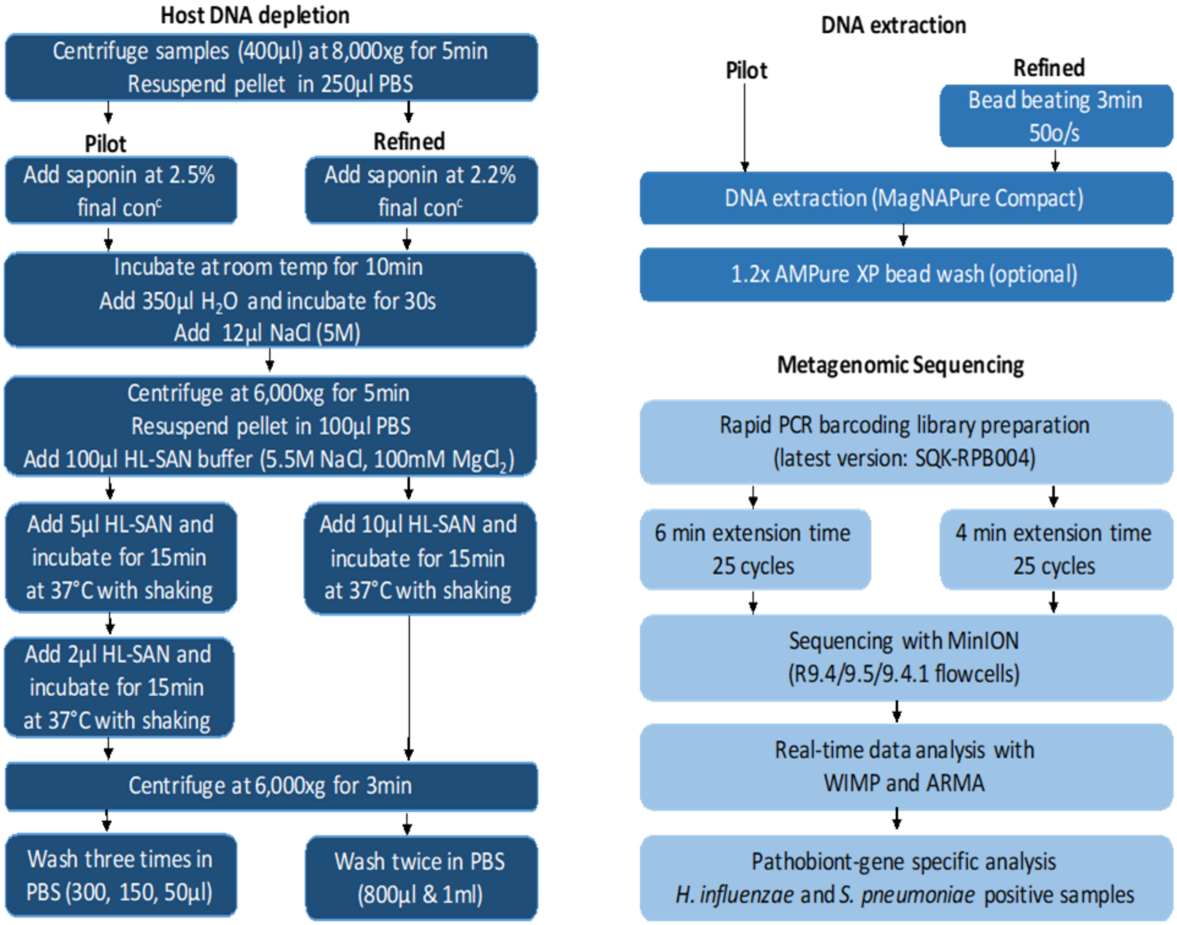

Host DNA depletion
Centrifuge samples (400µl) at 8,000xg for 5min
Resuspend pellet in 250µl PBS
Pilot
Refined
Add saponin at 2.5% final con^c
Add saponin at 2.2% final con^c
Incubate at room temp for 10min
Add 350µl H₂O and incubate for 30s
Add 12µl NaCl (5M)
Centrifuge at 6,000xg for 5min
Resuspend pellet in 100µl PBS
Add 100µl HL-SAN buffer (5.5M NaCl, 100mM MgCl₂)
Add 5µl HL-SAN and incubate for 15min at 37°C with shaking
Add 10µl HL-SAN and incubate for 15min at 37°C with shaking
Add 2µl HL-SAN and incubate for 15min at 37°C with shaking
Centrifuge at 6,000xg for 3min
Wash three times in PBS (300, 150, 50µl)
Wash twice in PBS (800µl & 1ml)
DNA extraction
Pilot
Refined
Bead beating 3min 50o/s
DNA extraction (MagNAPure Compact)
1.2x AMPure XP bead wash (optional)
Metagenomic Sequencing
Rapid PCR barcoding library preparation (latest version: SQK-RPB004)
6 min extension time 25 cycles
4 min extension time 25 cycles
Sequencing with MinION (R9.4/9.5/9.4.1 flowcells)
Real-time data analysis with WIMP and ARMA
Pathobiont-gene specific analysis
H. influenzae and S. pneumoniae positive samples

Figure 3.1: Schematic representation of the optimised and pilot metagenomics pipeline.

3.1.6 Limit of detection experiments

The LoD of the optimised clinical metagenomics pipeline was determined for the detection of one Gram-positive and one Gram-negative bacterium in sputum. Commensal microbial communities vary in composition and abundance in sputum samples and this may affect the sensitivity of detection of pathogens. Hence, while performing LoD experiments, we chose sputum samples from the clinical microbiology lab that tested negative for pathogens (normal respiratory flora, NRF, samples) with different abundance of normal flora (as determined by 16S qPCR) to see how this would affect detection of pathogens spiked at different concentrations. An NRF sample with a confirmed high bacterial background (22 Cq with the 16S rRNA qPCR assay) and a NRF sample with a confirmed low bacterial background (27 Cq with the 16S rRNA qPCR assay) were chosen for spiking. Ten-fold serial dilutions (10^5-10 cfu/ml) of cultured *E. coli* (H141480453) and *S. aureus* (NCTC 6571) were spiked into the chosen NRF sputum samples. The serial dilutions were plated in triplicate on LB agar (described in 2.3.2) to determine colony forming units (CFU). Host depletion and DNA extraction was followed as shown in Figure 3.1. Detection and quantification of bacterial DNA was performed using probe-based qPCR assays (described in 2.6) and MinION sequencing (described in 2.7). Each replicate was defined as positive for the spiked 'pathogen' if present at ≥1% classified microbial reads.

If two of the three replicates were positive for the spiked pathogen, then it was considered positive at that dilution. *S. aureus* was detected in all replicates spiked with 100,000 cells and in 2/3 replicates spiked with 10,000 cells in a high microbial background. Detection of *E. coli* in a high microbial background, however, was only possible in samples spiked with 100,000 cells. Hence, the LoD of the optimised CMg pipeline was determined to be 100,000 (10^5) cells for *E. coli* and 10,000 (10^4) cells for *S. aureus* when in a high bacterial background (Table 3.12A).

The LoD was determined to be lower in sputum samples with a lower bacterial background (Table 3.12B). Detection of both pathogens was possible in the lowest dilution tested (10^3 for *S. aureus* and *E. coli*) with >12% of *S. aureus* classified reads in all samples spiked with 1000 *S. aureus* cells and >4% of *E. coli* reads in all replicates spiked with 1000 *E. coli* cells (Table 3.13B). Hence, the LoD of the optimised CMg pipeline ranges from 10^3-10^5 CFU/mL, however, different levels of background commensal/human DNA could potentially result in different LoDs.

Table3.12A: Sputum sample with a high bacterial background* spiked with Gram-positive and Gram-negative organisms processed with the optimised method to determine limit of detection.

Sample	Replicate	Pathogen	Approx. number of pathogen cells (CFU)	DNA yield (ng/µl)	Total raw read count	Reads mapping to hg38	Non-human reads	Classified from non-human reads	Unclassified from non-human reads	Total number of pathogen reads from total classified reads
SA 10^5	1	*S. aureus*	100,000	27.0	66,528	581 (0.9%)	65,947 (99.1%)	50,922	15,015	12,658 (24.9%)
	2			7.5	54,576	418 (0.8%)	54,158 (99.2%)	42,612	11,524	4,286 (10.1%)
	3			20.6	36,285	221 (0.6%)	36,064 (99.4%)	29,303	6,749	3,882 (13.2%)
SA 10^4	1		10,000	10.9	40,002	263 (0.7%)	39,739 (99.3%)	29,311	10,423	165** (0.6%)
	2			34.6	36,416	312 (0.9%)	36,104 (99.1%)	27,233	8,860	389 (1.4%)
	3			28.6	44,844	416 (0.9%)	44,428 (99.1%)	32,656	11,752	686 (2.1%)
EC 10^5	1	*E. coli*	100,000	26.6	37,222	259 (0.7%)	36,963 (99.3%)	28,559	8,398	1,346 (4.7%)
	2			25.0	50,842	261 (0.5%)	50,581 (99.5%)	38,924	11,637	2,030 (5.2%)
	3			25.6	45,167	208 (0.5%)	44,959 (99.5%)	34,140	10,798	1,092 (3.2%)
EC 10^4	1		10,000	7.8	28,035	119 (0.4%)	27,916 (99.6%)	22,324	5,586	90** (0.4%)
	2			26.2	44,761	292 (0.7%)	44,469 (99.3%)	33,525	10,932	156** (0.5%)
	3			36.6	46,920	331 (0.7%)	46,589 (99.3%)	34,319	12,244	149** (0.4%)

*22Cq bacterial 16S rRNA gene V3-V4 fragment qPCR assay, **number of reads detected was below the ≥1% of classified microbial reads and WIMP assignment q-score ≥20 cut-off required for a sample to be considered positive.

Table 3.12B: Sputum sample with a low bacterial background* spiked with Gram-positive and Gram-negative organisms processed with the optimised method to determine limit of detection.

Sample	**Replicate**	**Pathogen**	**Approx. number of pathogen cells (CFU)**	**DNA yield (ng/µl)**	**Total raw read count**	**Reads mapping to hg38**	**Non-human reads**	**Classified from non-human reads**	**Unclassified from non-human reads**	**Total number of pathogen reads from total classified reads****
SA 10^4	1	*S. aureus*	10,000	5.0	38,213	589 (1.5%)	37,624 (98.5%)	36,271	1,343	23,514 (64.8%)
	2			0.6	8,476	92 (1.1%)	8,384 (98.9%)	8,151	230	5,928 (72.7%)
	3			1.5	14,300	581 (4.1%)	13,719 (95.9%)	12,642	1,074	3,129 (24.8%)
SA 10^3	1		1,000	11.0	61,889	19,276 (31.1%)	42,613 (68.9%)	38,796	3,805	7,222 (18.6%)
	2			3.2	31,360	9,337 (29.8%)	22,023 (70.2%)	19,905	2,115	3,037 (15.3%)
	3			1.8	25,346	2,531 (10.0%)	22,815 (90.0%)	20,6802	2,120	2,499 (12.1%)
EC 10^4	1	*E. coli*	10,000	1.9	29,322	879 (3.0%)	28,443 (97.0%)	26,452	1,983	10,609 (40.1%)
	2			0.8	7,324	185 (2.5%)	7,139 (97.5%)	6,642	493	2,756 (41.5%)
	3			0.9	4,058	93 (2.3%)	3,965 (97.7%)	3,637	327	1,072 (29.5%)
EC 10^3	1		1,000	0.8	2,645	88 (3.3%)	2,557 (96.7%)	2,271	285	110 (4.8%)
	2			1.2	10,380	281 (2.7%)	10,099 (97.3%)	8,736	1,357	384 (4.4%)
	3			1.4	15,016	543 (3.6%)	14,473 (96.4%)	12,663	1,805	555 (4.4%)

*27Cq bacterial 16S rRNA gene V3-V4 fragment qPCR assay. **The number of reads detected for all samples was above the ≥1% of classified reads and WIMP assignment q-score ≥20 required for a sample to be considered positive.

3.1.7 Mock community experiments

Clinical isolates from respiratory samples were used to generate a mock community consisting of *S. pneumoniae, K. pneumoniae, H. influenzae, S. maltophilia, P. aeruginosa* and *C. albicans. E. coli* and *S. aureus* strains were also included (H141480453 and NCTC 6571 respectively). Selected pathogens, were cultured (as described in 2.3.3) and were then spiked into an NRF sample (~10^3-10^6 CFU/pathogen) and then tested in triplicate with the optimised CMg pipeline, to determine if saponin depletion would result in inadvertent lysis of pathogens and loss of their DNA. All spiked samples were processed alongside undepleted controls. qPCR assays (as described in 2.6) were used to determine the relative quantity of each spiked pathogen in depleted and undepleted spiked sputum samples.

Results of the qPCR assays revealed that depleting host nucleic acid (10^3 fold loss) did not result in loss of bacterial DNA for seven out of the eight spiked pathogens, as on average <1 Cq difference was observed between depleted and undepleted samples (Table 3.13). The seven organisms that were not affected by host depletion were: *C. albicans*, *E. coli, H. influenzae, K. pneumoniae, P. aeruginosa, S. aureus* and *S. maltophilia*. The only pathogen affected by host depletion was *S. pneumoniae*, as a 5.8-fold loss was observed (average ΔCq 2.52) between depleted and undepleted samples (Table 3.13).

Table 3.13: Mock community qPCR results in triplicate for spiked NRF samples with and without the optimised method.

qPCR assay	**Sample**	**Triplicate 1 (Cq)**	**Triplicate 2 (Cq)**	**Triplicate 3 (Cq)**	**Average (Cq)**	**Human or microbial loss (ΔCq)**
Human	Undepleted	24.20	24.27	25.27	24.58	9.9 (10^3 fold)
	Depleted	33.97	34.52	34.96	34.48	
C. albicans	Undepleted	26.71	26.48	26.29	26.49	0.04
	Depleted	27.12	25.68	26.80	26.53	
E. coli	Undepleted	23.47	23.53	23.99	23.66	0.12
	Depleted	23.73	23.94	23.68	23.78	
H. influenzae	Undepleted	30.60	30.53	30.55	30.38	0.77
	Depleted	31.55	30.66	31.25	31.15	
K. pneumoniae	Undepleted	29.96	29.78	30.29	30.01	0.05
	Depleted	30.08	30.26	29.85	30.06	
P. aeruginosa	Undepleted	22.78	22.77	23.15	22.90	0.17
	Depleted	23.07	23.14	22.99	23.07	
S. aureus	Undepleted	26.23	26.62	27.99	26.94	0.16
	Depleted	26.19	27.35	26.80	26.78	
S. maltophilia	Undepleted	24.96	24.96	25.66	25.29	0.02
	Depleted	25.56	24.93	25.45	25.31	
S. pneumoniae	Undepleted	25.66	25.70	26.68	26.01	2.52 (5.8 fold)
	Depleted	28.18	28.81	28.62	28.53	

3.1.8 Investigation of *Streptococcus pneumoniae* loss observed in the mock community experiments

The *S. pneumoniae* loss observed in the mock community experiments was further investigated. *S. pneumoniae* positive clinical pilot samples were tested by *S. pneumoniae* qPCR (described in 2.6) to detect any *S. pneumoniae* loss compared to undepleted controls. In four samples (P3, P28, P30 and P33), *S. pneumoniae* loss was observed (minimum ΔCq= 1.7 and maximum ΔCq= 5.84) and no loss was observed in the other 2 samples; P36 and P40 (Table 3.14A).

Table 3.14A: qPCR results of *S. pneumoniae*-positive pilot samples.

Sample	***S. pneumoniae ply* gene qPCR-probe based assay (Cq)**	***S. pneumoniae* DNA loss/gain after host depletion (ΔCq)**
P3-Undepleted control	21.09	1.7 (3.2 fold)
P3-Depleted	22.79	
P28-Undepleted control	19.75	3.17 (9 fold)
P28-Depleted	22.92	
P30-Undepleted control	19.21	5.84 (57.28 fold)
P30-Depleted	25.05	
P33-Undepleted control	21.70	3.64 (12.5 fold)
P33-Depleted	25.34	
P36-Undepleted control	20.7	0.18 (1.13 fold)
P36-Depleted	20.52	
P40-Undepleted control	18.97	0.64 (1.55 fold)
P40-Depleted	18.33	

These results and the mock community results suggested that *S. pneumoniae* cells can be lysed during host depletion, hence we further investigated which step(s) damage the *S. pneumoniae* cell wall.

Initially, the high salt buffer was altered to observe if this was damaging the *S. pneumoniae* cell wall. Using an *S. pneumoniae*-positive sample (SP0), 1 M NaCl for the nuclease buffer (manufacturer's recommended salt concentration for HL-SAN DNase) was compared to the 5.5 M salt buffer (no changes were made to the rest of the optimized protocol). The 1 M salt buffer sample had a 2.77 fold loss (ΔCq= 1.47) and the 5.5 M salt buffer sample had 13.83-fold loss (ΔCq= 3.79), showing that the HL-SAN buffer could potentially lyse (or lead to lysis) *S. pneumoniae* cells (Table 3.14B).

Table 3.14B: qPCR results of a *S. pneumoniae*-positive sample tested with different HL-SAN buffers.

Sample	***S. pneumoniae ply* gene qPCR-probe based assay (Cq)**	***S. pneumoniae* DNA loss/gain after host depletion (ΔCq)**
SP0-Undepletd control	20	-
SP0-Depleted (1M NaCl)	21.47	1.47 (2.77 fold)
SP0-Depleted (5.5M NaCl)	23.79	3.79 (13.83 fold)

Next, each step of the host DNA depletion method was tested to investigate the effect on *S. pneumoniae*. An *S. pneumoniae*-spiked (PMEN1 strain cultured for 24 hrs -as described in 2.3.3) NRF sputum sample was processed in duplicate with the original method (SP1 and SP2 in Table 3.14C) and an altered method where either the saponin treatment (SP3 and SP4 in Table 3.14C) or the osmotic shock (SP5 and SP6 in Table 3.14C) was removed. Duplicate undepleted spiked controls were also included (SP7 and SP8 in Table 3.14C). Each duplicate was compared for *S. pneumoniae* loss compared to the undepleted controls, using probe-based qPCR assay (described in 2.6)

In the duplicates (SP1 and SP2) where the optimised CMg method was carried out with no alterations a 430.5-fold loss of *S. pneumoniae* was observed (Table 3.14C). Loss was increased to 831.7-fold in the two duplicates where the osmotic shock was removed (SP5 and SP6) when compared with the undepleted controls (Table 3.14C). However, some *S. pneumoniae* loss (26.7-fold) was still observed in the duplicates (SP3 and SP4) where the saponin-treatment was not performed but still included DNase treatment with HL-SAN DNase and buffer.

These results suggest that all the main steps (saponin treatment, high salt osmotic shock and possibly HL-SAN buffer) in the host depletion can result in *S. pneumoniae* loss, with the saponin treatment causing the biggest loss. However, this was not a systemic observation as *S. pneumoniae* loss was not recorded in 2/6 culture-positive *S. pneumoniae* samples tested with the pilot CMg pipeline (Table 3.14A).

Table 3.14C: qPCR results of a *S. pneumoniae*-spiked sample tested on different conditions of the optimised method.

Sample	**Spin 8000g/5min, discard/resuspend in PBS**	**5% saponin 10 minutes**	**Osmotic shock**	**Spin, resuspend in HLSAN buffer/PBS and DNAse**	**Two PBS washes**	***S. pneumoniae ply* gene qPCR-probe based assay Cq**	***S. pneumoniae* DNA loss after host depletion (ΔCq)**
SP1	Yes	Yes	Yes	Yes	Yes	23.6	8.75 (430.5 fold)
SP2	Yes	Yes	Yes	Yes	Yes	23.97	
SP3	Yes	No (PBS instead)	Yes	Yes	Yes	19.49	4.74 (26.7 fold)
SP4	Yes	No (PBS instead)	Yes	Yes	Yes	20.05	
SP5	Yes	Yes	No (PBS instead)	Yes	Yes	24.23	9.7 (831.7 fold)
SP6	Yes	Yes	No (PBS instead)	Yes	Yes	25.23	
SP7	Yes	No	No	No	No	15.09	-
SP8	Yes	No	No	No	No	14.97	

3.1.9 Evaluation of the optimised clinical metagenomics pipeline for the diagnosis of bacterial LRTIs

The optimised pipeline was then tested on a set of respiratory samples to determine its clinical sensitivity and specificity compared to clinical microbiology (pathogen and AMR detection).

In total, 41 excess respiratory samples from patients with suspected bacterial LRIs (previously processed by routine microbiology – described in 2.3.1) were collected (described in 2.2) and tested with the optimised method (Figure 3.1). Host depletion was measured by qPCR (described in 2.6) and pathogen and AMR gene detection was determined using 2 hours of sequencing data (described in 2.8.2 and 2.8.3).

A maximum of 10^4-fold depletion of human nucleic acid was reported (in 5/41 samples), but the average was 10^3-fold depletion (median 600-fold; interquartile range 168-1156 fold; maximum 18,054 fold - see Table 3.15). Metagenomic sequencing data also revealed the efficiency of the host depletion as human reads represented <18% of classified reads on average after 2 hrs of sequencing (Table3.16).

No significant loss of bacteria was observed for the majority of the samples but a ≥6-fold bacterial loss was observed in 7/41 samples and bacterial gain was observed 2/41 between depleted samples (Table 3.15).

CMg was concordant with culture for 28/29 culture-positive samples (including 3/28 confirmed mixed infections) tested. Single-bacterial infections that were correctly characterized were: eight *H. influenzae* confirmed samples, five *S. aureus* confirmed samples (including two MRSA-positive samples), four *P. aeruginosa* confirmed samples, two *S. marcescens, M. catarrhalis, E. coli* and *Klebsiella* spp. infections. The three mixed-bacterial infections that were correctly identified were two *H. influenzae* and *S. aureus* infections and one *P. aeruginosa* and *S. aureus* infection (S27, S38 and S41 - Table 3.15). The pathogenic organism reported by routine microbiology was detected by metagenomics together with an additional pathogen (not reported by culture) in eight samples: *K. pneumoniae* in S5, *P. aeruginosa* in S7, *M. catarrhalis* in S14 and S39, *S. pneumoniae* in S8 and S15, *S. aureus* in S29 and *S. pyogenes* in S27 (Table 3.15). Up to two potentially pathogenic bacteria were also observed in seven culture-negative samples (reported as NRF/ NSG) by routine microbiology) i.e. *H. influenzae* and *S. pneumoniae* in S10 and S21; *S. pneumoniae* in S11 and S28; *M. catarrhalis* and *H. influenzae* in S12; *H. influenzae* in S31 and *E. coli* in S32 (Table 3.15).

Only one pathogen reported by routine microbiology in S9 was missed by clinical metagenomics. For this sample metagenomics detected *E. coli* only, whereas culture reported S9 as a mixed infection with *P. aeruginosa* and *E. coli*.

Based on these results, the overall sensitivity of the optimised method for respiratory pathogen detection was 96.6% (95% CI, 80.4-99.8%) and specificity was 41.7% (95% CI, 16.5-71.4%), *not* counting additional organisms detected by metagenomics in culture-positive samples as false positives. The turnaround time from sample to result (including two hours of MinION sequencing) was approximately six hours. (Table 3.15).

Table 3.15. Human and bacterial DNA qPCR results for respiratory samples infected by Gram-negative and Gram-positive bacteria with and without host nucleic acid depletion.

Sample	**Sample type**	**Organism cultured by microbiology**	**Organism identified from metagenomic pipeline**	**Sample treatment**	**Human qPCR assay (Cq)**	**Human DNA depletion (ΔCq)**	**16S rRNA gene V3-V4 fragment qPCR assay (Cq)**	**Bacterial gain/loss to standard depletion (ΔCq)**
S1	ETA	*E. coli*	*E. coli*	Undepleted	22.62	12.38 $(\sim10^{4})$	15.60	0.13
				Depleted	35.00		15.73	
S2	Sputum	*K.pneumoniae*	*K.pneumoniae*	Undepleted	23.73	9.99 $(\sim10^{3})$	15.63	0.02
				Depleted	33.71		15.65	
S3	Sputum	*P. aeruginosa*	*P. aeruginosa*	Undepleted	23.05	9.29 $(\sim10^{3})$	15.46	1.48
				Depleted	32.34		13.98	
S4	Sputum	*S. marcescen*	*S. marcescens*	Undepleted	26.34	9.93 $(\sim10^{3})$	16.96	0.52
				Depleted	36.27		17.48	
S5	Sputum	*K. oxytoca*	*K. oxytoca*	Undepleted	22.96	8.58 $(\sim10^{3})$	12.67	0.64
			K.pneumoniae	Depleted	31.54		12.03	
S6	Sputum	*S. aureus*	*S. aureus*	Undepleted	22.31	9.41 $(\sim10^{3})$	19.11	1.57
				Depleted	31.72		17.54	
S7	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	25.47	9.53 $(\sim10^{3})$	21.44	0.43
			P. aeruginosa	Depleted	35.00		21.87	
S8	Sputum	*M.catarrhalis*	*M.catarrhalis*	Undepleted	22.72	9.17 $(\sim10^{3})$	16.9	0.66
			S.pneumoniae	Depleted	31.89		17.56	
S9	Sputum	*P.aeruginosa* & *E. coli*		Undepleted	23.89	11.11 $(\sim10^{4})$	19.58	3.26
			E. coli	Depleted	35		22.84	
S10	Sputum	NSG	*H. influenzae*	Undepleted	23.46	8.6 $(\sim10^{3})$	14.12	2.39
			S. pneumoniae	Depleted	32.06		16.51	
S11	Sputum	NRF	*S. pneumoniae*	Undepleted	25.77	9.23 $(\sim10^{3})$	17.96	1.92
				Depleted	35.00		19.88	
S12	Sputum	NRF	*H. influenzae*	Undepleted	22.5	8.92 $(\sim10^{3})$	17.61	0.05
			M. catarrhalis	Depleted	31.42		17.56	
S13	Sputum	*S. marcescens*	*S. marcescens*	Undepleted	22.48	7.11 $(\sim10^{2})$	12.77	0.79
				Depleted	29.59		11.98	
S14	Sputum	*S. aureus*	*S. aureus*	Undepleted	23.17	7.68 $(\sim10^{2})$	13.83	0.96
			M. atarrhalis	Depleted	30.85		14.79	

S15	Sputum	*S. aureus*	*S. aureus*	Undepleted	22.66	8.47 (~10^3)	18.73	0.08
			S. pneumoniae	Depleted	31.13		18.65	
S16	Sputum	MRSA	MRSA	Undepleted	25.51	6.43 (~10^2)	15.32	0.24
				Depleted	31.94		15.56	
S17	Sputum	NRF	None	Undepleted	23.51	9.64 (~10^3)	19.55	1.17
				Depleted	33.15		20.72	
S18	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	27.14	7.86 (~10^2)	12.89	2.21
				Depleted	35.00		15.10	
S19	Sputum	NRF	None	Undepleted	22.63	11.18 (~10^3)	19.69	0.69
				Depleted	33.81		19.00	
S20	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	22.44	10.03 (~10^3)	14.99	1.19
				Depleted	32.47		16.18	
S21	Sputum	NRF	*H. influenzae*	Undepleted	24.58	10.42 (~10^3)	16.60	0.82
			S. neumoniae	Depleted	35.00		17.42	
S22	Sputum	NRF	None	Undepleted	22.71	9.22 (~10^3)	14.62	0.39
				Depleted	31.93		15.01	
S23	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	24.82	10.18 (~10^3)	16.80	1.84
				Depleted	35.00		18.64	
S24	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	22.24	10.17 (~10^3)	15.70	1.63
				Depleted	32.41		17.33	
S25	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	25.52	6.26 (~10^2)	16.59	2.67
				Depleted	31.79		19.26	
S26	Sputum	*M.catarrhalis*	*M. catarrhalis*	Undepleted	23.47	11.53 (~10^4)	19.26	0.74
				Depleted	35.00		20.00	
S27	Sputum	*H. influenzae* & *S. aureus*	*H. influenzae*	Undepleted	32.74	2.26 (~5)	23.19	7.92
			S. aureus	Depleted	35.00		15.27	
			S. pyogenes					
S28	Sputum	NRF	*S. pneumoniae*	Undepleted	24.46	10.54 (~10^3)	22.28	2.80
				Depleted	35.00		25.08	
S29	Sputum	*P. aeruginosa*	*P. aeruginosa*	Undepleted	24.05	5.11 (~10^2)	19.81	2.04
			S. aureus	Depleted	29.13		17.77	
S30	BAL	*P. aeruginosa*	*P. aeruginosa*	Undepleted	29.93	5.07 (~33)	22.68	0.00
				Depleted	>35.00		22.68	
S31	Sputum	NRF	*H. influenzae*	Undepleted	21.57	8.26 (~10^3)	19.79	1.65
				Depleted	29.83		21.44	
S32	Sputum	NSG	*E. coli*	Undepleted	25.56	8.68 (~10^3)	15.98	0.47
				Depleted	34.24		16.45	
S33	Sputum	NRF	None	Undepleted	21.73	10.04 (~10^3)	20.69	0.81
				Depleted	31.77		21.50	

S34	Sputum	NSG	None	Undepleted	25.17	5.40 (~10^2)	22.92	0.01
				Depleted	30.57		22.93	
S35	Sputum	*E. coli*	*E. coli*	Undepleted	21.11	5.18 (~10^2)	16.49	0.58
				Depleted	26.29		17.07	
S36	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	22.58	9.70 (~10^3)	16.51	2.00
				Depleted	32.28		18.51	
S37	Sputum	*P. aeruginosa*	*P. eruginosa*	Undepleted	21.56	11.69 (~10^4)	15.25	1.80
				Depleted	33.24		13.45	
S38	Sputum	*S. aureus* & *P.aeruginosa*	*S. aureus*	Undepleted	20.76	6.87 (~10^2)	23.83	3.17
			P. aeruginosa	Depleted	27.63		20.66	
S39	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	23.82	11.18 (~10^3)	14.45	2.79
			M. catarrhalis	Depleted	35.00		17.24	
S40	ETA	MRSA	MRSA	Undepleted	21.69	4.28 (~19)	19.91	1.62
				Depleted	25.97		18.29	
S41	Sputum	*H. influenzae* *& S. aureus*	*H. influenzae* *S. aureus*	Undepleted	20.86	14.14 (~10^4)	16.71	6.85

In addition to these findings, in 29/41 samples metagenomic sequencing also identified additional non-pathogenic organisms that normally reside in the lower respiratory tract. The majority of bacteria identified in the depleted samples were different species of the *Streptococci* genus. Additional organisms identified were *H. parainfluenzae*, *Rothia mucilaginosa*, *Veillonella parvulla, Neisseria sicca* and different species of the *Lactobacillus* genus (see Table 3.16 for a list of additional organisms identified by metagenomics).

Table 3.16: All microorganisms identified in all samples tested using the optimised method and sequencing metadata after 2 hours of sequencing*.

Sample number	**Total raw reads**	**Number of reads minus hg38**	**Human reads**	**Classified reads**	**Unclassified reads**	**Organisms identified from metagenomic pipeline above chosen thresholds (number of pathogenic reads)**
S1	108610	108346	264	107971	364	*Escherichia coli* (91178)
S2	17516	17485	31	17303	182	*Klebsiella pneumoniae* (1692) *Streptococcus parasanguinis*
S3	26641	26257	384	24673	1583	*Pseudomonas aeruginosa* (15817) *Streptococcus oralis* *Streptococcus parasanguinis*
S4	7358	7348	10	6602	743	*Serratia marcescens (*3900) *Pantoea stewartii* *Citrobacter freundii* *Streptococcus parasanguinis* *Salmonella enterica* *Serratia plymuthica*
S5	19888	19865	23	19224	636	*Klebsiella oxytoca (*3254*)* *Citrobacter freundii* *Klebsiella* sp. M5al *Klebsiella pneumoniae* (906) *Klebsiella michiganensis*
S6	16403	13295	3108	13271	24	*Staphylococcus aureus* (11307)

S7	32730	21690	11040	17398	4289	*Streptococcus parasanguinis* *Pseudomonas aeruginosa (1970)* *Veillonella parvula* *Rothia mucilaginosa* *Streptococcus sanguinis* *Haemophilus influenzae (232)* *Haemophilus parainfluenzae* *Neisseria sicca*
S8	49277	44120	5157	35023	9088	*Moraxella catarrhalis (7078)* *Streptococcus parasanguinis* *Streptococcus mitis* *Veillonella parvula* *Streptococcus salivarius* *Rothia mucilaginosa* *Streptococcus pseudopneumoniae* *Streptococcus oralis* *Streptococcus pneumoniae(504)* *Neisseria sicca* *Streptococcus gordonii*
S9	29111	28969	142	28527	437	*Escherichia coli (12974)* *Lactobacillus paracasei* *Lactobacillus casei* *Candida albicans*
S10	56005	51610	4395	49850	1743	*Haemophilus influenzae (35348)* *Streptococcus pseudopneumoniae* *Streptococcus* sp. oral taxon 431 *Streptococcus pneumoniae (571)* *Streptococcus parasanguinis* *Streptococcus mitis*

S11	43088	40944	2144	34012	6927	*Streptococcus parasanguinis* *Streptococcus* sp. A12 *Streptococcus mitis* *Rothia mucilaginosa* *Bifidobacterium longum* *Streptococcus pseudopneumoniae* *Streptococcus pneumoniae (693)* *Streptococcus* sp. I-P16 *Streptococcus* sp. I-G2 *Haemophilus parainfluenzae*
S12	43267	38274	4993	35153	3116	*Haemophilus parainfluenzae* *Moraxella catarrhalis(5573)* *Streptococcus gordonii* *Neisseria sicca* *Haemophilus influenzae(414)* *Streptococcus parasanguinis*
S13	27592	27406	186	26979	425	*Serratia marcescens(22758)*
S14	41154	40622	532	35413	5205	*Staphylococcus aureus (95465)* *Moraxella catarrhalis (8089)* *Rothia mucilaginosa* *Streptococcus constellatus* *Fusobacterium nucleatum* *Fusobacterium periodonticum* *Streptococcus parasanguinis* *Streptococcus anginosus* *Streptococcus intermedius*

S15	37500	36537	963	30473	6058	*Staphylococcus aureus (6597)* *Streptococcus mitis* *Streptococcus oralis* *Streptococcus parasanguinis* *Citrobacter koseri* *Rothia mucilaginosa* *Streptococcus salivarius* *Haemophilus parainfluenzae* *Streptococcus pseudopneumoniae* *Streptococcus pneumoniae (390)* *Prevotella melaninogenica*
S16	85298	85057	241	83750	1301	*Staphylococcus aureus (59392)* *Streptococcus oralis* *Lactobacillus rhamnosus*
S17	25499	23615	1884	19072	4541	*Streptococcus oralis* *Streptococcus salivarius* *Prevotella melaninogenica* *Streptococcus parasanguinis* *Streptococcus* sp. oral taxon 431 *Streptococcus* sp. A12 *Streptococcus* sp. NPS 308 *Streptococcus sp.*FDAARGOS_192
S18	38902	38744	158	38092	650	*Haemophilus influenzae (34396)*

S19	47994	42413	5581	33926	8485	*Streptococcus salivarius* *Streptococcus parasanguinis* *Streptococcus mitis* *Streptococcus* sp.FDAARGOS_192 *Streptococcus oralis* *Streptococcus sanguinis* *Rothia mucilaginosa* *Veillonella parvula* *Streptococcus* sp. oral taxon 431 *Streptococcus gordonii* *Streptococcus constellatus*
S20	46331	43070	3261	42143	920	*Haemophilus influenzae (2827)* *Rothia mucilaginosa* *Streptococcus mitis* *Streptococcus parasanguinis*
S21	45214	45075	139	39288	5780	*Haemophilus influenzae (20161)* *Streptococcus mitis* *Haemophilus parainfluenzae* *Streptococcus parasanguinis* *Streptococcus sp.* oral taxon 431 *Streptococcus salivarius* *Streptococcus oralis* *Streptococcus pneumoniae (408)*
S22	36853	36194	659	27003	9183	*Streptococcus parasanguinis* *Prevotella melaninogenica* *Streptococcus mitis* *Streptococcus sanguinis* *Rothia mucilaginosa* *Veillonella parvula* *Haemophilus parainfluenzae*

						Streptococcus salivarius *Streptococcus* sp. A12 *Streptococcus cristatus*
S23	33140	32933	207	29164	3766	*Haemophilus influenzae (9635)* *Streptococcus parasanguinis* *Streptococcus salivarius* *Streptococcus oralis* *Veillonella parvula* *Streptococcus pseudopneumoniae* *Streptococcus mitis* *Streptococcus* sp.FDAARGOS_192
S24	58752	57669	1083	51978	5686	*Haemophilus influenzae (28443)* *Streptococcus parasanguinis* *Haemophilus parainfluenzae* *Veillonella parvula* *Streptococcus oralis* *Streptococcus salivarius* *Streptococcus* sp. oral taxon 431
S25	36621	35808	813	35716	89	*Haemophilus influenzae (33307)*
S26	38138	36910	1228	33541	3367	*Streptococcus oralis* *Streptococcus mutans* *Moraxella catarrhalis (3831)* *Veillonella parvula* *Streptococcus parasanguinis* *Streptococcus salivarius* *Streptococcus gordonii* *Rothia dentocariosa* *Lactobacillus salivarius*
S27	78311	78064	247	74697	3357	*Haemophilus influenzae (31884)* *Streptococcus pyogenes (19544)* *Staphylococcus aureus (9969)*

						Streptococcus salivarius *Streptococcus parasanguinis*
S28	35804	34699	1105	26717	7980	*Streptococcus mitis* *Streptococcus salivarius* *Streptococcus oralis* *Veillonella parvula* *Streptococcus* sp. A12 *Streptococcus sanguinis* *Rothia mucilaginosa* *Streptococcus* sp. oral taxon 431 *Streptococcus parasanguinis* *Streptococcus pneumoniae (464)* *Streptococcus pseudopneumoniae* *Prevotella melaninogenica* *Streptococcus* sp. I-G2 *Streptococcus* sp. I-P16 *Haemophilus parainfluenzae*
S29	57865	9429	48436	9153	275	*Pseudomonas aeruginosa (688)* *Staphylococcus aureus (1187)*
S30	27371	27217	154	27154	60	*Pseudomonas aeruginosa (21577)*
S31	43823	13463	30360	9567	3895	*Haemophilus influenzae (1644)* *Streptococcus anginosus* *Streptococcus parasanguinis* *Prevotella intermedia* *Tannerella forsythia* *Bifidobacterium longum* *Rothia mucilaginosa* *Streptococcus gordonii* *Veillonella parvula* *Streptococcus oralis* *Campylobacter concisus*

S32	48271	47988	283	46882	1101	*Escherichia coli (35094)*
S33	32085	19181	12904	18722	459	*Streptococcus salivarius* *Clavispora lusitaniae* *Streptococcus parasanguinis* *Streptococcus* sp.FDAARGOS_192 *Rothia mucilaginosa* *Saccharomyces cerevisiae*
S34	40998	35522	5476	35031	489	*Candida albicans* *Enterococcus faecalis*
S35	47893	29765	18128	29663	96	*Escherichia coli (26316)*
S36	33155	32774	381	31170	1603	*Haemophilus influenzae (23245)* *Veillonella parvula* *Streptococcus salivarius* *Streptococcus mitis*
S37	60495	60083	412	59644	432	*Pseudomonas aeruginosa (47138)*
S38	23889	12947	10942	12833	113	*Staphylococcus aureus (1248)* *Pseudomonas aeruginosa (253)*
S39	60316	59972	344	57940	2029	*Moraxella catarrhalis (32133)* *Haemophilus influenzae (14404)* *Streptococcus parasanguinis*
S40	48848	7444	41404	7340	104	*Staphylococcus aureus (6138)* *Streptococcus constellatus* *Streptococcus anginosus* *Streptococcus intermedius*
S41	2320	2129	191	2040	89	*Staphylococcus aureus (744)* *Haemophilus influenzae(343)* *Neisseria sicca* *Escherichia coli (35)*

*above chosen thresholds

3.1.10 Confirmatory analysis of additional and missed pathogens

Confirmatory probe-based qPCR (described in 2.6) was used to confirm the presence or absence of the missed/additional pathogens (described above) detected by the optimised CMg pipeline in 16 samples (1 sample with a missed pathogen, 15 samples with additional pathogen/s detected – including 7 culture-negative samples; total of 19 pathogens) and in matched controls i.e. an equal number of samples that neither culture or metagenomics detected the pathogen (Table 3.17A). Probe-based qPCR was performed on DNA extracts from samples that did not undergo the depletion process (undepleted controls), to eliminate depletion as a potential cause of missed/additional pathogen/s reported. In total 12/19 additional pathogens detected by metagenomics in 16 samples were confirmed by qPCR. This included 5/7 culture-negative (S10, S11,S12, S31,S32) and 5/9 culture-positive samples (S7, S14, S27, S29 and S39) (Table 3.17A). This analysis, increased the specificity of the optimised method to 83.3% (95% CI, 36.5-99.1%) – as culture-negative samples where additional pathogens detected were confirmed by qPCR were now considered as true positive samples and additional organisms detected by metagenomics only in culture-positive samples were *not* counted as false positives. Pathogenic organisms (*not* including pathobionts such as *H. influenzae* and *S. pneumoniae*) identified by metagenomics only and not confirmed by qPCR were: *K. pneumoniae* in S5, likely a *k*-mer mis-classification of *K. oxytoca* reads, and *E. coli* in S41, likely a laboratory/kit contaminant. Also, qPCR was negative for *P. aeruginosa* (S9) increasing the sensitivity of the optimised method to 100% (95% CI, 87.7-100%).

Table 3.17A: Confirmatory qPCR analysis of additional pathogen detected by the optimised clinical metagenomics pipeline.

Discordant results compared to culture	**Sample**	**qPCR gene target**	**qPCR result (Cq)**
Escherichia coli[1]	S32	*cyaA*	22.6
	S41		-
Haemophilus influenzae[1]	S10	*omp P6*	25.4
	S12		-
	S21		32.0
	S31		29.2
Klebsiella pneumoniae[1]	S5	*Mdh*	-
Moraxella catarrhalis[1]	S12	*copB*	23.8
	S14		22.0
	S39		23.5
*Pseudomonas aeruginosa**	S9	*oprL*	-
Pseudomonas aeruginosa[1]	S7	*oprL*	32.9
Staphylococcus aureus[1]	S29	*eap*	32.7
Streptococcus pneumoniae[1]	S8	*ply*	-
	S10		26.1
	S11		32.2
	S15		-
	S21		-
	S28		-
Streptococcus pyogenes[1]	S27	*sdaB*	28.7

[1] Detected by metagenomics and not culture *Detected by culture but not metagenomics

Next, a species-specific gene analysis (described in 2.8.4.2) was performed for all samples positive for *H. influenzae* and *S. pneumoniae.* These two organisms are pathobionts (i.e. potentially pathogenic organisms which may reside as commensals in the lung), which share genetic similarities with commensal species residing in the lungs (there were 18 samples containing 20 pathobionts). This species-specific analysis was used to identify *k*-mer mis-classification of commensal reads as pathogenic reads by WIMP. For this analysis samples containing >1 pathobiont specific gene alignment (*siaT* gene for *H. influenzae* and *ply* gene *S. pneumoniae*) were considered positive for that organism. Pathobiont specific genes were identified in 13/14 *H.influenzae* sample tested and in 2/6 *S. pneumoniae* samples tested. No alignments of either of the genes were identified in 5/18 samples analysed, including three originally culture-negative samples (*H.influenzae* in S12 and *S. pneumoniae* in S21 and S8) and two culture-positive samples where the pathobionts (*S. pneumoniae* in S18 and S15) were reported by metagenomics only. No alignments confirmed *k*-mer mis-classification and the absence of *H. influenzae/S. pneumoniae*. The only remaining culture-negative sample (S28) was also negative by qPCR analysis (see Table 3.17A) and as they were also negative for pathobiont specific genes, resulted in the optimised CMg method being 100% specific (95% CI, 51.7-100%) and 100% sensitive (95% CI, 87.7-100%) when compared to the culture+qPCR gold standard (Table 3.17B).

Table 3.17B. Species-specific gene analysis for *H. influenzae* and *S. pneumoniae* identified by the clinical metagenomics pipeline only.

Pathobionts detected by metagenomics pipeline	**Sample**	**Pathobiont specific gene**	**Number of reads aligned to pathobiont specific gene**
Haemophilus influenzae	S7	*siaT*	5
	S10		85
	S12		0
	S18		85
	S20		67
	S21		49
	S23		21
	S24		61
	S25		90
	S27		63
	S31		5
	S36		60
	S39		43
	S41		2
Streptococcus pneumoniae	S8	*ply*	0
	S10		5
	S11		19
	S15		0
	S21		0
	S28		0

3.1.11 Detection of AMR genes using the clinical metagenomics pipeline.

Resistance genes in the respiratory samples were also identified (described 2.8.3) using data from 2 hours of MinION sequencing. Detection of AMR genes was performed to demonstrate that resistance can be determined using rapid CMg. The samples tested using the optimised method were mostly susceptible with little antibiotic resistance, according to routine results (Table 3.18A). Amongst the 33 cultivated organisms, only 43 instances of resistance and intermediate resistance were recorded by culture (described in 2.3.1) with some of these likely reflecting single underlying mechanisms (Table 3.18A). Metagenomic sequencing reported 183 resistance genes across the 41 samples (with multiple genes reported when ARMA identified multiple variants of e.g. $\textit{bla}_{\text{TEM}}$). All genes detected by ARMA are listed in Table 3.18A and are summarized/explained in Table 3.18B.

Table 3.18A: Microbiology antibiogram and ARMA output for all optimised method samples.

Sample	Microbiology culture result	Antibiogram	ARMA Output
S1	*E. 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	TEM-4 sul1 mphA dfrA17 aadA5 ACT-5
S2	*K. 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	oqxB oqxA InuA tetM
S3	*P. aeruginosa*	Gentamicin S, Tazocin S, Ciprofloxacin S, Ceftazidime S, Meropenem S	OXA-50 catB7 mefA mel APH(3’)-llb
S4	*S. marcescens*	Gentamicin S, Co-trimoxazole S, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime S, Amikacin S, Tigecycline S, Tobramycin S, Levofloxacin S, Colistin R, Cefepime S, Minocycline S, Ticarcillin S	AAC(6’)-lc mel oqxB
S5	*K. 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	OXY-4-1 oqxB APH(3')-Ia vgaC
S6	*S. aureus*	Flucloxacillin S, Erythromycin/clarithromycin S, Clindamicin S, Fuscidic acid S, Tetracycline/doxycycline S, Mupirocin S	tet38
S7	*H. influenzae*	Amoxicillin R, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	tetM tetO IsaC
S8	*M. catarrhalis*	Amoxicillin R, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	mefA mel tetM

S9	*P. aeruginosa*	Gentamicin S, Tazocin S, Ciprofloxacin S, Meropenem S, Ceftazidime S	ermC TEM-4 sul1 aadA2 AAC(3’)-lla mphA dfrA12 AAC(3’)-llc tetC
	E. 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	
S10	NSG	ND	mefA tetM mel tetO
S11	NRF	ND	mefA mel TEM-4 tetC
S12	NRF	ND	TEM-4 tetM mel mefA ErmB
S13	*S. marcescens*	Gentamicin S, Co-amoxiclav R, Co-trimoxazole S, Ciprofloxacin S, Meropenem S, Aztreonam S, Ceftazidime S, Ceftriaxone S, Cefuroxime R, Amikacin S, Ertapenem S, Tigecycline I, Tobramycin S, Cefepime S	AAC(6’)-lc oqxB
S14	*S. aureus*	Flucloxacillin S, Erythromycin/clarithromycin S, Clindamicin S, Fuscidic acid S, Tetracycline/doxycycline S, Mupirocin S	tet38 tetM mel tetQ
S15	*S. aureus*	Flucloxacillin S, Erythromycin/clarithromycin S, Clindamicin S, Fuscidic acid S, Tetracycline/doxycycline S, Mupirocin S	tet38 mefA mel tetM tetW TEM-4
S16	*S. aureus/*MRSA	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	tet38 mecA tetM ErmB tetC TEM-4
S17	NRF	ND	tetM mel

			mefA TEM-4 tetC tetQ ErmF CfxA3
S18	*H. influenzae*	Amoxicillin R, Tetracycline/doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	TEM-4 TEM-70 tetM TEM-33 TEM-105 TEM-104 TEM-208
S19	NRF	ND	mel mefA tetM tetW TEM-4 ErmX
S20	*H. influenzae*	Amoxicillin S, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	tetO mefA tetM mel
S21	NRF	ND	mel mefA tetM tet32
S22	NRF	ND	tetM mel tetQ mefA
S23	*H. influenzae*	Amoxicillin S, Tetracycline/doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	tetM mel mefA TEM-4
S24	*H. influenzae*	Amoxicillin S, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	mel tetM mefA tetO tetD tet32 tetW, tetA
S25	*H. influenzae*	Amoxicillin S, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	mefA mel TEM-4

S26	*M. catarrhalis*	Amoxicillin R, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	mel mefA tetO tetM tetC
S27	*H. influenzae*	Amoxicillin S, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole S	ermT tet38 tetM mefA
	S. aureus	Flucloxacillin S, Erythromycin R, Clindamycin R, Fuscidic acid S, Tetracycline S, Mupirocin S	
S28	NRF	ND	mel TEM-4 mefA tetM tetC lsaC
S29	*P. aeruginosa*	Gentamicin S, Tazocin S, Ciprofloxacin S, Meropenem S, Ceftazidime S	OXA-50 TEM-4 catB7 tet38
S30	*P. aeruginosa*	Gentamicin S, Tazocin S, Ciprofloxacin S, Meropenem S, Ceftazidime S	OXA-50 catB7 TEM-4 tetC
S31	NRF	ND	TEM-4 tetC
S32	NSG	ND	TEM-4 AAC(3)-lla AAC(3)-llc vgaC TEM-1
S33	NRF	ND	mel mefA tetC
S34	NSG	ND	tetM lsaA TEM-4 ErmB tetC
S35	*E. 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	TEM-4 TEM-11 TEM-2 TEM-187 ACT-5 tetC

			TEM-67
S36	*H. influenzae*	Amoxicillin R, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole R	mefA ErmB Mel TEM-4 tetM
S37	*P. aeruginosa*	Gentamicin S, Tazocin R, Ciprofloxacin S, Meropenem S, Ceftazidime R	catB7 OXA-50 TEM-4
S38	*P. aeruginosa*	Gentamicin S, Tazocin S, Ciprofloxacin S, Meropenem S, Ceftazidime S	TEM-4 tetC tet38
	S. aureus	Flucloxacillin S, Erythromycin S, Clindamycin S, Fuscidic acid S, Tetracycline S, Mupirocin S	
S39	*H. influenzae*	Amoxicillin S, Doxycycline S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole R	mefA mel TEM-4 tetC
S40	*S. aureus* (MRSA)	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 tet38 tetC TEM-4
S41	*S. aureus*	Flucloxacillin S, Erythromycin S, Clindamycin S, Fuscidic acid S, Tetracycline S, Mupirocin S	TEM-4 tetC
	H. influenzae	Tetracycline S, Amoxicillin S, Ceftriaxone S, Co-amoxiclav S, Ciprofloxacin S, Co-trimoxazole R	

R=resistant, S= sensitive, I= intermediate resistance, ND= Not Done

Among the 183 resistance genes, 26 were inherent to the species identified by culture (e.g. *oqxA/B* for *K. pneumoniae* or *bla*$_{\text{OXA-50}}$ in *P. aeruginosa*), from the remaining 157, 24 resistance genes matched the observed phenotype (Table 3.18B). These included, *mecA* found in both MRSA (S16 and S40), *sul1* and *dfrA12* or *dfrA17* in both co-trimoxazole-resistant *E. coli* (S1 and S9), *aac(3')-IIa* (and *IIc*) in a tobramycin-resistant *E. coli* (S9) and a total of 13 *bla*$_{\text{TEM}}$ variants spread across two amoxicillin-resistant *E. coli* (S1 and S35) and two amoxicillin-resistant *H. influenzae* (S18 and S36). A caveat regarding the identification of *bla*$_{\text{TEM}}$ variants, was that ARMA did not flag *bla*$_{\text{TEM-1}}$, which was the likeliest variant to be present in these samples, given (i) that it is considerably the most prevalent type and (ii) that the isolated organisms remained susceptible to oxyimino- cephalosporins whereas many of the *bla*$_{\text{TEM}}$ variants, flagged by ARMA, should encode extended-spectrum variants. Depending on their strength of expression *bla*$_{\text{TEM}}$ or *bla*$_{\text{OXY}}$ may have explained non-susceptibility to penicillin/β-lactamase inhibitor combinations in Enterobacteriales (4/183 resistance genes recorded), but quantification of gene expression was not possible by ARMA. A *bla*$_{\text{TEM4}}$ gene (1/183) was also found by ARMA in a ceftazidime- and piperacillin/tazobactam- resistant *P. aeruginosa* (S37); which could explain the phenotype reported by routine testing, but is questionable in this species, as β-lactam resistance often results from the up-regulation of chromosomal *ampC* or efflux. There were 14/183 recorded genes where any associated resistance could not be established as the relevant drug(s) was not tested by routine microbiology. An example of these, were the *tet* genes identified in several samples (S2, S8, S9, S16, S30, S35, S38 and S39) and tetracycline was not tested against the isolates cultured. Sixteen genes identified by ARMA did not match the reported phenotype of the cultivated isolates, which remained susceptible to relevant antibiotics according to culture, and 42 genes were unlikely to be from species identified by routine testing.

Finally, multiple genes (56/183) likely originated from the normal lung microbiota. For example, *tet(M)* and *bla*$_{\text{TEM-4}}$ genes, each were found in 8/12 NRF/NSG samples, and *mefA* and *mel* were each found in 9/12 culture-negative samples (Table 3.18B).

There were nine samples where phenotypic resistances could not be explained by resistance genes reported by ARMA. This included two amoxicillin-resistant *M. catarrhalis* (S8 and S26), where BRO β-lactamase genes (most likely to be responsible for the observed phenotype) were not represented in the database utilised by ARMA. The remaining seven out of the nine samples included ampicillin- and co-trimoxazole- resistant *H. influenzae* (S7, S18, S36, S39 and S41), trimethoprim-, ciprofloxacin- gentamicin- and fusidic acid- resistant *S. aureus* (S16) and a *K. pneumoniae* (S2) resistant to both co-amoxiclav and piperacillin/tazobactam where no acquired β-lactamase genes were identified by ARMA (Table 3.18B).

The specificity and sensitivity for resistance gene detection of the optimised method was not calculated as this would have required isolate cultivation and sequencing of all bacteria (pathogens and commensals) present– a prohibitive and unaffordable task for the duration of this PhD research.

Table 3.18B**:** Resistance genes detected by ARMA in relation to pathogens grown for samples processed with clinical metagenomics[*].

ARMA vs. culture result	**No. genes**	**Principal examples**
Gene endogenous in species	26	Mostly efflux components; also *bla*$_{\text{OXA-50}}$, *aph(3')-IIb* and *catB7* from *P. aeruginosa* and *aac(6')-Ic* from *S. marcescens*
Match to observed R	24	Variously including *mecA* in MRSA, *bla*$_{\text{TEM}}$ in Enterobacteriaceae and *H. influenzae,* also *sul1* and *dfr* determinants for *E. coli*
Partial match to observed resistances	4	Instances where *bla*$_{\text{TEM}}$ was found but where MinION flagged an ESBL-encoding variant, usually *bla*$_{\text{TEM-4}}$, but where the phenotype indicated only a classical penicillinase, without oxyimino-cephalosporin resistance
Unlikely match to observed phenotype	1	*P. aeruginosa* with *bla*$_{\text{TEM}}$ resistant to piperacillin/tazobactam and ceftazidime – see text
Possibly present, but relevant drug not tested by clinical lab	14	Commonly (i) where *tet(C*) found but lab tested doxycycline, which is not a substrate for this pump, or (ii) where streptomycin, kanamycin and macrolide determinants were found in gram-negative bacteria but these drugs were not tested, as not relevant to therapy.
Does not match phenotype of isolate	16	Mostly where *bla*$_{\text{TEM}}$ (as *bla*$_{\text{TEM-4}}$) was recorded but the isolate (commonly *H. influenzae*) was susceptible to penicillins as well as cephalosporins, or where *tet(M)* was found together with a tetracycline-susceptible *S. aureus*

Genes unlikely to be from species grown by the laboratory	42	Mostly gram-positive-associated genes when a gram-negative organism was grown, or vice versa: commonly including *tet(M*) and *mefA*
Gene recorded in a specimen with no pathogen grown	56	Mostly *tet*, *mef mel*, *bla*$_{\text{TEM-4}}$ determinants, likely to be associated with normal flora
Total	183	

*183 genes detected from the 41 samples

3.1.12 Impact of host DNA depletion for reference-based genome assembly

Two samples (S1 and S16) containing antibiotic resistant bacteria (confirmed by culture and metagenomics) were chosen as examples to generate reference-based genome assemblies (as described in 2.8.4.1) using the metagenomic data. This analysis was performed to demonstrate that CMg data can be used to generate whole pathogen genomes directly from respiratory samples which could be used for public health and infection control applications.

Reference-based assemblies were generated for an MRSA (S16) and an *E. coli* resistant to amoxicillin, co-amoxiclav and co-trimoxazole (S1). The assemblies were compared with those generated from the undepleted controls after 2 and 48 hours of sequencing to demonstrate the effect of successfully removing host nucleic acid on genome assemblies. For the first two hours of sequencing the depleted MRSA sample had 47.9x genome coverage with an assembly of 28 contigs (GCA_900660255: longest contig = 478718 and N50=400 kbp). After 48hrs of sequencing, genome coverage increased to 228.7x, with a final assembly consisting of 22 contigs (GCA_900660245: longest contig = 481 kbp and N50=403 kbp). In contrast, after 2hrs of sequencing the undepleted control for the MRSA sample had an assembly of 69 contigs with 3.9x coverage (GCA_900660235: longest contig = 47kbp and N50=146 kbp), and 33 contigs (17.5x coverage) after 48 hours of sequencing (GCA_900660205: longest contig = 416 kbp and N50=263kbp) (Figure 3.2A).

For the depleted sample positive for a resistant *E. coli* (S1) there was 33.5x genome coverage with an assembly of 83 contigs (GCA_900660265: longest contig = 437 kbp and N50=165 kbp) within two hours of sequencing. Genome coverage after 48 hrs, was increased to 165.7x with the final *E. coli* assembly having 72 contigs (GCA_900660275: longest contig = 474 kbp and N50=178 kbp). The undepleted sample only had 0.2x coverage after 2hrs, increasing to 1.1x genome coverage after 48 hrs of sequencing (Figure 3.2B). Genome assemblies could not be generated for either of the two chosen timepoints (2 hours and 48 hours) for the undepleted sample.

This analysis highlights the importance of host depletion being incorporated with CMg when pathogen load is low compared to host (an assembly was not possible for S1 without host depletion even after 48 hrs of sequencing) or in time-critical situations (for S16, 47.9x of the MRSA genome was recovered after 2 hours of sequencing).

3.1.13 The effect host DNA depletion has on rapid pathogen identification and AMR gene detection

Using the same sample set (S1 and S16) used for the reference-based genome assemblies, a timepoint analysis (described in 2.8.4.2) was carried out using data from the first two hours of sequencing. Genome recovery and alignments of confirmed resistance genes identified in depleted samples were compared with their undepleted controls over chosen timepoints to highlight the importance of host depletion for rapid results.

Within the first 5 min of sequencing the depleted MRSA sample (S16) had 1.6x genome coverage and was increased to 64.2x after 2 hours of sequencing. In contrast the undepleted control had 0.2x coverage after 5 min of sequencing and 5.8x after 2 hours (Figure 3.2C). Also no *mec*A gene alignments were detected in the undepleted sample in the first 30 min of sequencing, whereas two *mec*A gene alignments were recorded in the depleted sample in the first 5 min of sequencing (Figure 3.2D).

A similar trend was also reported for S1 - the depleted sample had 5.7x genome coverage of *E. coli* after 20 min of sequencing, which increased to 45.6x after 2 hours. The undepleted control however, had 0.06x genome coverage in the first 20 min of sequencing increasing to only 0.2x after 2 hours (Figure 3.2E). This *E. coli* was resistant to amoxicillin (*bla*$_{\text{TEM}}$ gene), co-amoxiclav (possibly owing to *bla*$_{\text{TEM}}$ if gene is strongly expressed) and co-trimoxazole (*sul1* and *dfr*A17 genes). Hence, sequencing data were aligned against these genes for the analysis. No *bla*$_{\text{TEM}}$ and *dfr*A17 gene alignments were detected in the undepleted sample within two hours of sequencing and only one alignment was detected for *sul*1. Conversely, at least one gene alignment was identified for all three resistance genes within 20 min of sequencing in the depleted sample, increasing to 47 *bla*$_{\text{TEM}}$, 37 *sulf1* and 21 *dfrA17* alignments after two hours of sequencing (Figure 3.2F).

These results demonstrate rapid diagnosis, including both pathogen and resistance gene identification is feasible using CMg only when coupled with host depletion strategies.

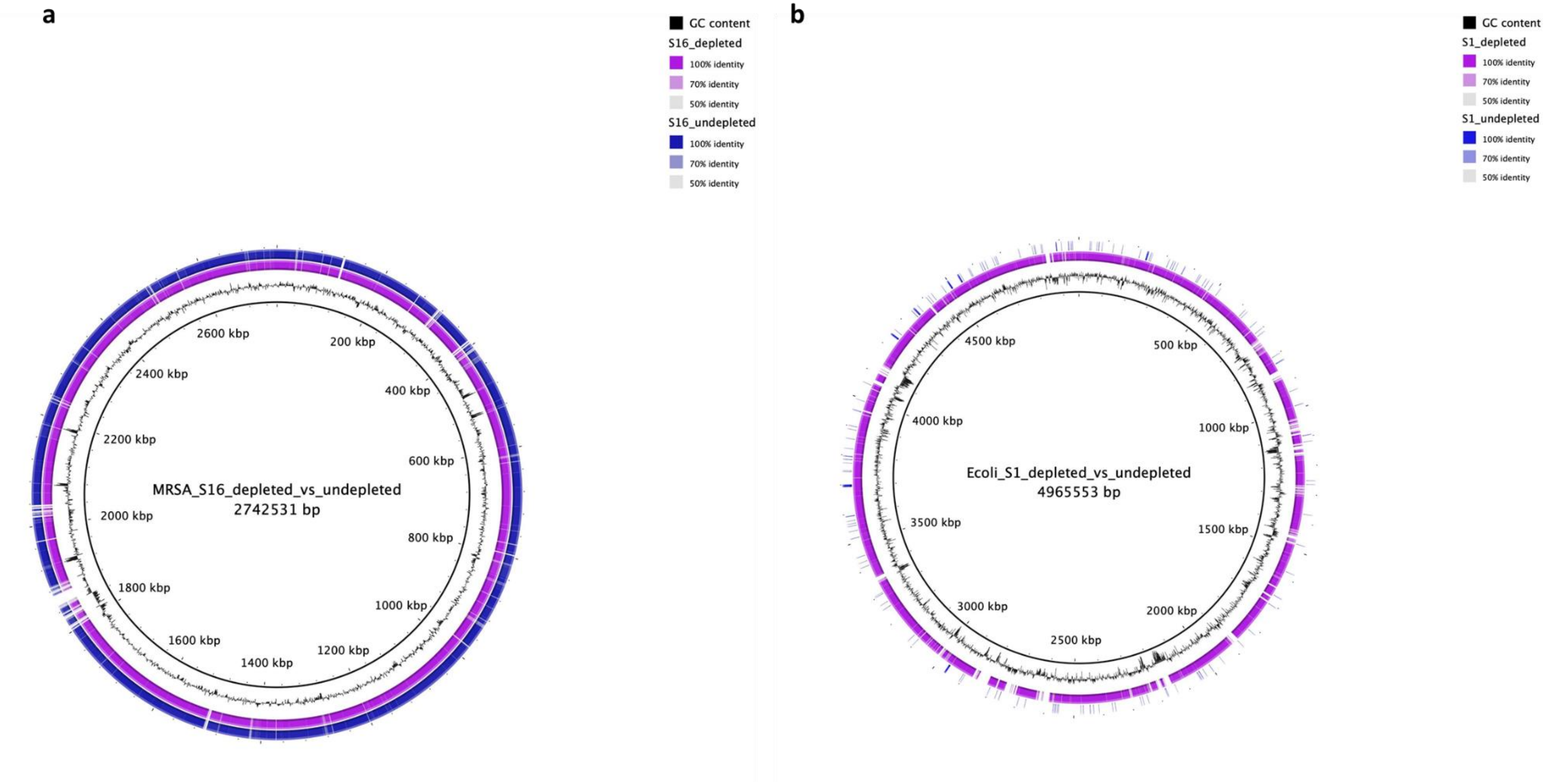


Figure 3.2: Bacterial genome assembly of depleted versus undepleted samples. A, MRSA after 48 h of sequencing. B, *E. coli* after 48 h of sequencing.

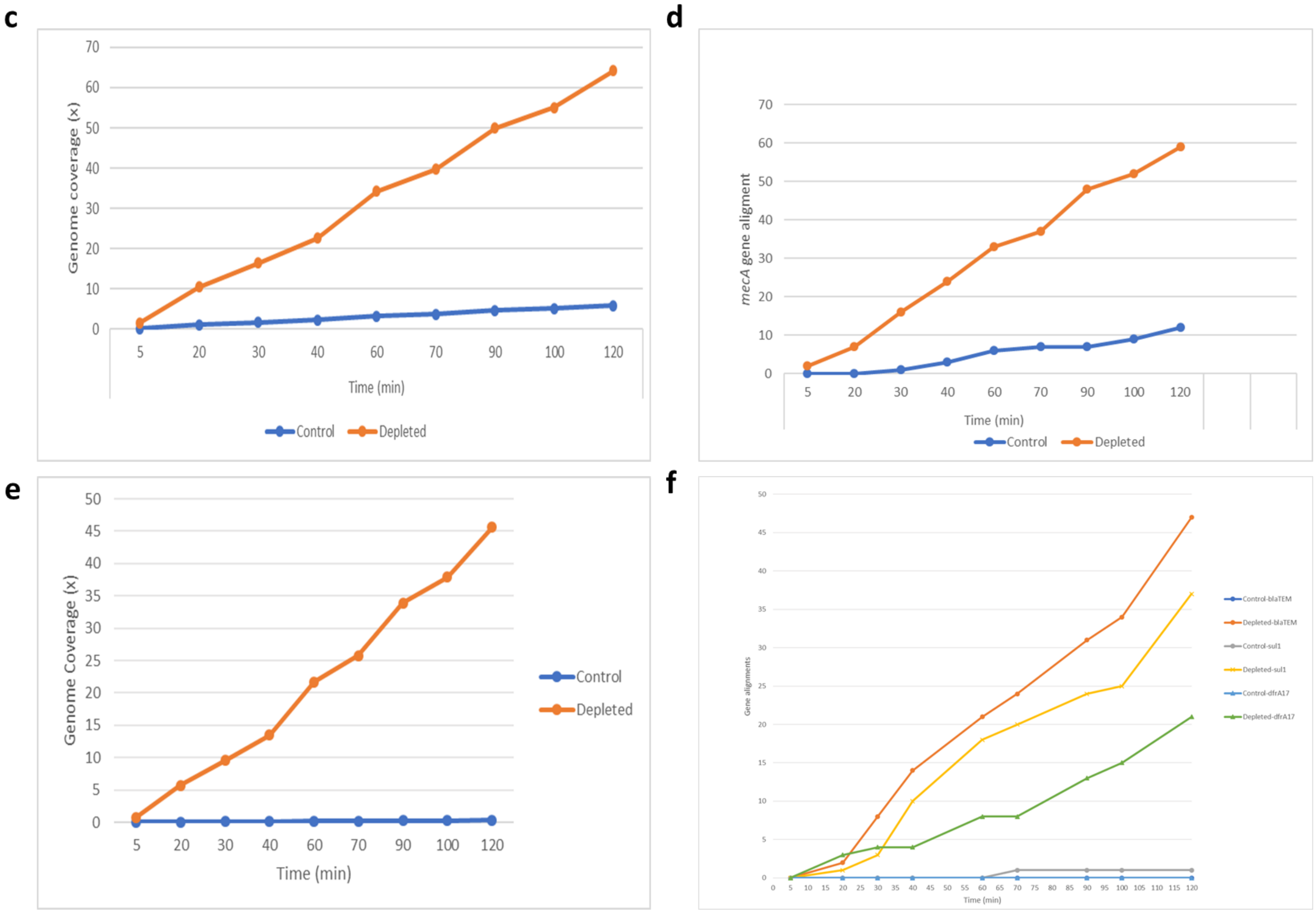


Figure 3.2 (continued): Genome coverage and antibiotic gene detection in depleted versus undepleted samples. C, MRSA genome coverage of depleted versus undepleted during 2 h of sequencing. D, *mecA* gene alignment of depleted versus undepleted during 2 h of sequencing. E, *E. coli* genome coverage of depleted versus undepleted during 2 h of sequencing. F, *bla*TEM, *sul1* and *dfr*A17 gene alignment of depleted versus undepleted during 2 h of sequencing. In C-F, three independent clinical samples were analysed (examples of a Gram-positive and a Gram-negative are represented).

3.2 Implementation of the clinical metagenomics pipeline

The optimised clinical metagenomics pipeline (Figure 3.1) was initially evaluated on a set of respiratory samples in a proof-of-concept study to establish the sensitivity and specificity of the method against microbiological cultures – detailed in 3.1. We then sought to further test the optimised CMg pipeline in a larger prospective study against culture and molecular PCR-based tests as part of the clinical trial entitled - INHALE: Potential of Molecular Diagnostics for Hospital Acquired and Ventilator-Associated Pneumonia in UK Critical Care (https://www.ucl.ac.uk/inhale-project/).

The primary aim of the INHALE trial is to assess the potential, molecular tests would have for the diagnosis of HAP and VAP in UK ICUs. The aim of the first phase of the trial (month 1-24) was to evaluate the performance of three rapid molecular platforms for the diagnosis of HAP and VAP, in terms of pathogen identification and resistance gene detection against culture. Additionally, INHALE aimed to assess the speed, cost, ease of use and implementability of the workflow/platform in the clinical setting. The platform/workflow with the best performance would then progress into a randomised controlled trial versus standard of care (https://www.ucl.ac.uk/inhale-project/project/project-work-plan).

Initially, the three molecular diagnostic tests chosen for evalutation against routine culture were all PCR-based platforms (https://www.ucl.ac.uk/inhale-project/project/project-goals), but halfway through the first phase of the trial, one of the three PCR platforms was not sufficiently developed for testing and was replaced by our optimised CMg pipeline. The three final tests, were the CMg pipeline (122) and the two PCR-based platforms, namely Curetis Unyvero Hospitalised Pneumonia test (https://www.ucl.ac.uk/inhale-project/project/technology-curetis-platform) and BIOFIRE Filmarray Pneumonia panel (https://www.ucl.ac.uk/inhale-project/project/technology-biofire-filmarray-platform).

The diagnostic performance of the three tests was evaluated by using samples collected from ICUs across fifteen UK hospitals: Norwich & Norfolk University Hospitals (NNUH), University College London Hospitals (UCLH), Great Ormond Street Hospital Children's Charity (GOSH), BUPA Cromwell Hospital, Royal Free Hospital, Chelsea and Westminster Hospital NHS Foundation Trust, City Hospitals Sunderland, Dudley Group NHS Foundation Trust, James Paget University Hospitals NHS Foundation Trust, Queen Elizabeth Hospital Kings Lynn NHS Trust, Aintree University Hospital NHS Foundation Trust, Guy’s and St Thomas’ NHS Foundation Trust, North Middlesex University Hospital NHS Trust, Royal Liverpool and Broadgreen University Hospitals NHS Trust and University Hospitals of North Midlands (UHNM) (95). Collected samples were processed at the two main sites of the trial – the Norwich Medical School at University of East Anglia (UEA) and Royal Free Hospitals (RHF). I only tested samples collected for the Norwich sites (which included samples from NNUH, UHNM and James Paget, Queen Elizabeth and Sunderland hospitals), hence only these samples will be presented and discussed in the thesis. The full analysis of all the samples is still underway and is expected to be published by the end of 2020.

3.2.1 Effect of sample freezing of CMg assay performance

A number of samples collected during the INHALE trial had been frozen at -20 °C right away. The optimised CMg pipeline was not previously tested on frozen samples, therefore we had to determine if freezing the samples would have any undesired effect (e.g. damage/lysing of bacterial cells before host depletion, leading to bacterial DNA loss) on pathogen detection and negatively affect the performance of the pipeline. Hence, a set of 13 frozen samples were chosen for testing with the CMg pipeline (Table 3.19).

From the 13 samples processed, two samples (YS030 and YS033) produced <500 reads after 2 hrs of sequencing and were considered sequencing failures. From the remaining samples, clinical metagenomics was in agreement with culture in 5/11 samples, including 4 culture-positive samples (YS051, YS023, YS032, YS022) and 1 culture negative sample (YS031). Clinical metagenomics however, was discordant with culture in the remaining 6 samples analysed (5 culture-positive and 1 culture-negative) as pathogens were completely missed or pathogenic reads were below the chosen thresholds (≥1% of microbial reads and >19 of alignment score - described in 2.8.2). The pathogens missed after metagenomics analysis were: *Pseudomonas* in YS040, *S. pneumoniae* in YS037, *S. marcescens* in YS029, *S. aureus* in YS028 and *S. pyogenes* in YS038. In YS052, *S. aureus* was identified by metagenomics only (Table 3.19). These results suggested that freeze/thawing cycles combined with the host depletion method can damage/lyse bacterial cells eventually affecting pathogen detection. As this would affect the sensitivity of the optimised pipeline, we decided to exclude all samples that were previously frozen for the INHALE study.

Table 3.19: Sequencing data* of frozen INHALE samples processed with clinical metagenomics.

Sample ID	Organism identified from routine microbiology	Organism identified from metagenomic pipeline	Classified reads	Microbial reads	Pathogenic classified reads	Human reads (% of classified reads)
YS051	Gram -ve bacillus	*E. coli*	36,552	10,354	5,449	26,198 (71.7%)
YS052	Negative	*S. aureus*	17,495	2,263	676	15232 (87%)
YS040	Gram -ve bacilli & *Pseudomonas*	*E. coli* *S. marcescens*	55,951	55,230	48,190 1,244	721 (1.3%)
YS023	*H. influenzae*	*H. influenzae* *K. pneumoniae*	36,718	36,415	34,412 1,750	303 (0.8%)
YS033	*Pseudomonas spp.*	*P. aeruginosa*	201	96	56	105 (52.2%)
YS031	Negative		1,184	1,150		34 (2.8%)
YS032	*Streptococcus* group B	*S. agalactiae* *S. aureus* *S. pyogenes*	59,894	59,195	45,424 12,463 3,258	699 (1.2%)
YS029	*E. coli* *S. marcescens*	*E. coli*	5,517	849	352	4,668 (84.6%)
YS030	*E. cloacae* complex		76	51		25 (33%)
YS037	*S. aureus* *S. pneumoniae*	*S. aureus*	61,555	61,065	6,060	490 (0.8%)
YS022	*H. influenzae*	*H. influenzae* *S. agalactiae*	57,529	27,968	965 333	29,561 (51.4 %)
YS028	*S. aureus*		4,154	275		3,879 (93.4%)
YS038	*Streptococcus* group A		6,031	164		5,867 (97.3%)

*After 2 hours of MinION sequencing

3.2.2 NNUH INHALE sample testing

The optimised method was tested on a total of 73 fresh respiratory samples (n= 34 culture-positive confirmed samples and n= 39 culture-negative samples (NRF or NG/NSG/NBG)) collected for the INHALE study. Initial analysis highlighted the need for additional parameters/thresholds to be put in place to accurately and effectively analyse this data to ensure the exclusion of both host depletion and sequencing failures. Therefore, any sample which produced <500 total reads was excluded as sequencing failures and any sample with <10% microbial reads was considered as a host depletion failure. After applying these thresholds, 47 samples remained for analysis (n=26 culture-positive and n=21 culture-negative samples). A processed-negative control rule was then applied to the remaining 47 samples to identify and eliminate contamination and barcode leakage. The same number of pathogenic reads identified in the processed-negative control were removed from each sample on the multiplexed sequencing run. Then pathogen identification was carried out as previously described (section 2.8.2) using two hours of MinION metagenomic sequencing data.

Clinical metagenomics was concordant with culture for 21/26 culture-positive samples (Table 3.20), including three mixed infections as reported by routine microbiology (NS041, NS064, YS018). Single bacterial infections included: *P.aeruginosa* in nine samples (JS015, NS070, NS075, NS078, NS080, QS005, SS004, YS011 and YS013), *S.aureus* in three samples (JS019, NS037 and NS042), *E. coli* in NS043, *Enterobacter cloacae* group in NS044, *S. marcescens* in NS050, *Proteus* spp. in NS057 and *Klebsiella* spp. in two samples (NS063, YS025). The three mixed infections, identified correctly by metagenomics were: an *E. coli* and *S. marcescens* infection in NS041, and *H. influenzae* and *S.aureus* in NS064 and *K. aerogenes* and *H. influenzae* in YS018. Additional pathogens were identified by metagenomics (not by culture) in 14/26 culture-positive samples; *C. freundii* in QS005, *E. coli* in JS015, in NS063 and NS078, *K. oxytoca* in NS041, *H. influenzae* in NS050 and NS057 and QS006, *M.*

morgannii in NS070, *S. agalactiae* in JS019 and NS079, *S. aureus* NS057, *S. maltophilia* in YS013, *S. pneumoniae* in NS042 and *K. pneumoniae* and *E. coli* in NS044. Culture-positive samples including pathogens identified by metagenomics only were not reported as false positive.

Potentially pathogenic bacteria were also identified in nine culture-negative samples (reported as NRF/NSG/NG) by routine microbiology) i.e. *E. aerogenes* in YS020, *E. coli* in JS016, NS067, YS020, YS034, YS044, and NS073, *K. pneumoniae* in YS034, *H. influenzae* in NS048, NS073 and YS020, *P. mirabilis* in NS072, *S. aureus* in NS073 and YS020 and *S. maltophilia* in YS042 (Table 3.20). However, pathogens in five samples were missed by metagenomics (identified by culture), i.e. *E. coli* in NS071 and YS059, *K. pneumoniae* in QS006, *P. aeruginosa* in NS079 and *S. aureus* in YS053 (Table3.20).

Based on these results, the optimised clinical metagenomics pipeline was 80.77% sensitive (95% CI; 60.65% to 93.45%) and 57.14% specific (95% CI; 34.02% to 78.18%) when used as the third test for the INHALE trial. The overall performance of the test is still being evaluated for the INHALE study and the analysis presented in this study is preliminary.

Table 3.20: Metagenomic sequencing data along with pathogens identified by metagenomics and routine testing.

Sample ID	Classified reads with human	Microbial Reads	Organism identified from metagenomic pipeline	Reads mapping to pathogen (% of pathogen reads)	Routine microbiology result
JS007	37,768	36,582	Negative		Negative
JS013	3,831	2,136	Negative		Negative
JS015	26,205	26,097	*E. coli* *P. aeruginosa*	10,462 (40.09%) 12,134 (46.5%)	*P. aeruginosa*
JS016	15,212	2,037	*E. coli*	671 (32.94%)	Negative
JS019	75,397	33,532	*S. aureus* *S. agalactiae*	31,398 (93.63%) 860 (2.56%)	*S. aureus*
NS037	28,447	20,896	*S. aureus*	6,781 (32.45%)	*S. aureus*
NS038	4,650	4,479	Negative		Negative
NS039	18,686	4,821	Negative		Negative
NS041	41,762	41,098	*E. coli* *K. oxytoca* *Serratia marcescens*	9,877 (24.03%) 665 (1.62%) 11,366 (27.65%)	*E. coli* *S. marcescens*
NS042	66,889	17,642	*S. aureus* *S. pneumoniae*	11,490 (65.8%) 541 (3.1%)	*S. aureus*
NS043	34,205	34,171	*E. coli*	27,782 (81.3%/)	*E. coli*
NS044	49,712	49,640	*Enterobacter cloaceae* group *E.coli* *K. pneumoniae*	22,667 (45.82%) 500 (1%) 732 (1.47%)	*E. cloaceae* group

NS047	1,852	1,501	Negative		Negative
NS048	6,646	798	*H. influenzae*	60 (7.52%)	Negative
NS049	7,750	3,644	Negative		Negative
NS050	58,702	58,686	*H. influenzae* *S. marcescens*	1,060 (1.8%) 55,287 (94.21%)	*S. marcescens*
NS057	25,859	23,551	*H. influenzae* *Proteus mirabilis* *S. aureus*	3,425 (14.54%) 13,854 (58.82%) 251 (1.06%)	Proteus spp.
NS063	4,159	1,745	*E. coli* *K. pneumoniae*	28 (1.6%) 209 (11.98%)	*K. pneumoniae*
NS064	61,248	60,488	*H. influenzae* *S. aureus*	55,485 (91.73%) 1,189 (1.96%)	*H. influenzae* *S. aureus*
NS067	10,273	10,138	*E. coli*	249 (2.45%)	Negative
NS069	4,562	4,003	Negative		Negative
NS070	67,389	52,241	*Morganella morganii* *P. aeruginosa*	1,803 (3.45%) 48,647 (93.12%)	*P. aeruginosa*
NS071	57,966	52,872	*S. aureus*	5,143 (9.73%)	*S. aureus* *E. coli*
NS072	39,302	39,116	*P. mirabilis*	36,899 (94.33%)	Negative
NS073	8,883	984	*E. coli* *H. influenzae* *S. aureus*	33 (3.35% 34 (3.4%) 30 (3.05%)	Negative
NS074	51,887	49,867	Negative		Negative
NS075	61,930	57,367	*P. aeruginosa*	54,834 (95.58%)	*P. aeruginosa*
NS077	32,912	10,100	Negative		Negative

NS078	759	647	*E. coli* *P. aeruginosa*	24 (3.71%) 71 (10.97%)	*P. aeruginosa*
NS079	47,491	46,143	*S. pneumoniae*	481 (1.04%)	*P. aeruginosa*
NS080	135,177	132,854	*P. aeruginosa*	125,053 (94.13%)	*P. aeruginosa*
QS005	42,119	42,083	*Citrobacter freundii* *P. aeruginosa*	481 (1.14%) 38,978 (92.62%)	*P. aeruginosa*
QS006	28,655	28,526	*H. influenzae* *P. mirabilis*	3,602 (12.63%) 8,876 (31.11%)	*P. mirabilis* *K. pneumoniae*
SS004	71,198	70,941	*P. aeruginosa*	68,829 (97.02%)	*P. aeruginosa*
YS011	66,541	66,464	*P. aeruginosa*	63,988 (96.27%)	*P. aeruginosa*
YS013	57,839	15,954	*P. aeruginosa* *S. maltophilia*	12,064 (75.62%) 1,435 (9%)	*P. aeruginosa*
YS014	22,898	19,744	Negative		Negative
YS018	118,804	118,284	*Enterobacter aerogenes* *H. influenzae*	12,759 (10.78%) 100,478 (84.94%)	*K. aerogenes* *H. influenzae*
YS020	9,255	6,398	*E. aerogenes* *E. coli* *H. influenzae* *S. aureus*	217 (3.39%) 236 (3.69%) 1,050 (16.41%) 600 (9.37%)	Negative
YS025	44,592	43,320	*K. oxytoca*	2,787 (6.43%)	*K. oxytoca*
YS034	68,899	67,912	*E. coli* *K. pneumoniae*	61,062 (89.91%) 864 (1.27%)	Negative
YS036	60,393	50,502	Negative		Negative

YS042	100,049	99,832	*S. maltophilia*	95,735 (95.9%)	Negative
YS044	2,045	1,212	*E. coli*	188 (15.51%)	Negative
YS053	56,724	56,641	*P. aeruginosa*	52,096 (91.97%)	*P. aeruginosa* *S. aureus*
YS057	15,893	15,851	Negative		Negative
YS059	114,792	114,343	*S. aureus*	113,008 (98.83%)	*S. aureus* *E. coli*

3.2.2.1 Further analysis of the performance of the CMg pipeline

The performance of the CMg test was reassessed using the results from the PCR tests run on the same samples (Biofire and Curetis) and pathogen specific gene analysis. Pathogens identified in six out of the nine false positive samples were also reported by at least one of the PCR-based machines. The organisms identified by both metagenomics and by BIOFIRE were: *E. coli* in JS016, *H. influenzae* in NS048, *E. coli* in NS067, *P. mirabilis* in NS072, *E. coli* and *H. influenzae* in YS020 and *E. coli* and *K. pneumoniae* in YS034 (Table 3.23). Curetis confirmed the presence of *S. maltophilia* in YS042 (*S. maltophilia* is only detectable by Curetis). Samples concordant with the PCR tests were then classified as true positives and not false positives, which increased the specificity to 80% (95% CI; 51.91% to 95.67%) but also increased the sensitivity to 84.38% (95% CI; 67.21% to 94.72%) due to these samples now being considered as true positives (i.e. in total there were 27/32 true positive samples).

Species-specific gene analysis (described in 2.8.4.2) was carried out on the remaining organisms in the false positive samples (YS020, YS044, NS073) in a similar way as previously done after testing the optimised method (specificity of the optimised method initially was 41.7% and then was 100% after additional analysis – described in 3.1.10). Genes selected from the literature (listed in Table 3.21) were used to either confirm the presence of the pathogen or identify contaminants or *k*-mer mis-classification of commensal organisms. Any sample containing >1 specific gene alignment after this analysis, was considered positive for that organism or negative if no gene alignments were detected. Only one gene alignment was detected, *E. aerogenes* (*atpD*) in YS020 (Table 3.21). The detection of *K. aerogenes* by metagenomics was probably due to barcode leakage as this sample was sequenced in a multiplex run containing a culture-positive sample for *K. aerogenes* (YS018 – see Table 3.20). The metagenomics pipeline was 84.38% sensitive (95% CI; 67.21% to 94.72%) and 93.33% specific (95% CI; 68.05% to 99.83%) after this analysis.

Table 3.21: Species-specific gene analysis for additional pathogenic organisms identified.

Pathogens identified from metagenomic pipeline	**Sample ID**	**Gene target**	**Alignments of species-specific genes**
Escherichia coli	YS044	*cyaA*	0
	NS073		0
Haemophilus influenzae	NS073	*siaT*	0
Staphylococcus aureus	NS073	*eap*	0
	YS020		0
Enterobacter aerogenes	YS020	*atpD*	1

3.2.3 Antibiotic resistance analysis of INHALE samples

Clinical metagenomics can detect pathogens but also provide information on antimicrobial resistance (if any) directly from samples taken from ICU patients – this data could potentially be used to guide antimicrobial therapy. Sequence data from the INHALE samples was analysed for the presence of acquired antibiotic resistance genes (as described in 2.8.3) using 2 hours of sequencing data.

The INHALE samples (n=47) were not highly resistant, according to culture. Amongst the 33 organisms tested by routine microbiology for susceptibility/resistance, only 35 instances of resistance (n=34) and intermediate resistance (n=1) were recorded by culture (Table 3.22A). Metagenomic sequencing was able to identify 26 clinically relevant resistance genes across the 47 specimens (all clinically-relevant genes detected by ARMA are listed in Table 3.22A and are summarised/explained in Table 3.22B)

In three culture-positive samples, genes identified by ARMA (n=9 genes) matched the phenotype reported by routine testing (Table 3.22B). These included *bla*$_{\text{TEM}}$ and *bla*$_{\text{SHV}}$ in an ampicillin and cefpodoxime resistant *E. coli* (NS043), *bla*ACT genes for *S. marcescens* (NS041) and *E. cloacae* (NS044), both resistant to amoxicillin-clavulanate and ampicillin. ARMA identified *bla*$_{\text{TEM-4}}$ genes in a meropenem and piperacillin/tazobactam resistant *P. aeruginosa* (JS015 – see Table 3.22B), which could explain the β-lactam resistance reported by routine microbiology (Table 3.22A) but is uncertain in this species as this phenotype is often due to the up-regulation of efflux pumps or chromosomal *amp*C (122).

Clinically-relevant resistance genes (n=14 genes) were identified in nine culture-negative (YS034, YS014, NS073, YS044, NS074, NS077, YS036, YS020 and YS057) and three culture-positive samples (JS019, NS071 and YS018) with susceptible organisms according to routine microbiology (Table 3.22B). Genes found in these samples are likely to be from the normal respiratory microbiota of the lung. ARMA identified resistance genes (n=2) that could not explain the phenotype reported by routine microbiology in two culture-positive samples. These included the *bla*$_{\mathrm{TEM}}$ genes reported both in NS064 and NS078. NS064 was positive for a susceptible *H. influenzae* and a clindamycin and erythromycin resistant *S. aureus, - bla*$_{\mathrm{TEM}}$ genes cannot explain the reported resistance for *S. aureus* (for example *erm* genes confer macrolide resistance in *Staphylococci* (250, 251)). Also, NS078 was positive for a meropenem-resistant *P. aeruginosa* and *bla*$_{\mathrm{TEM}}$ genes do not confer carbapenem resistance, which is primarily caused by efflux-pump upregulation (252) (Table 3.22A).

The phenotype reported by routine microbiology could not be explained by ARMA in five samples. These included fucidic acid resistant *S. aureus* (NS037), meropenem resistant *P. aeruginosa* (NS070, NS075 and YS011) and a penicillin- cephalosporin-resistant *S. marcescens* (NS050). Fucidic acid resistant resistance in *S. aureus* and cephalosporin resistance in *S. marcescens* are primarily due to chromosomal mutations (252) and the AMR pipeline was not designed for SNP detection. Also, as said before, meropenem resistance in *P. aeruginosa* is commonly mediated by to the up-regulation of efflux pumps. As the analysis presented in this thesis is preliminary we did not seek to characterise such resistances. Also, CMg did not detect the resistant organism in three samples (QS006, NS071 and YS059), hence the relevant resistance genes could not be identified.

The specificity and sensitivity of the optimised CMg pipeline for detecting resistance genes could not be calculated, as it would have required isolation, cultivation and sequencing of all bacteria present in all the respiratory samples. This task was beyond the scope of this PhD study.

Table 3.22A. Microbiology antibiogram and ARMA output for all INHALE samples.

Sample ID	Microbiology culture result	Antibiogram	ARMA OUTPUT
NS037	*S. aureus*	Fucidic acid R, clarithromycin S, flucloxacillin S, mupirocin S, tetracycline S	ND
QS006	*P. mirabilis* *H. influenzae*	Amoxicillin-clavulanate S, Aztreonam S, Cefpodoxime S, Cefpodoxime S, Ceftazidime S, Ciprofloxacin S, Co-trimoxazole S, Ertapenem S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S, Amoxicillin R, Cefuroxime R	ND
NS039	NSG	ND	ND
NS038	NSG	ND	ND
QS005	*P. aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S	ND
NS043	*E. coli*	Amoxicillin-clavulanate S, Ceftazidime S, Cefuroxime S, Aztreonam S, Ciprofloxacin S, co-Trimoxazole S, Ertapenem S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S, Ampicillin R, Cefpodoxime R	TEM-4, SHV-129, SHV-13
NS044	*E. colacae* group	Cefpodoxime S, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S, Ampicillin R, Amoxicillin-Clavulanate R	ACT-35, ACT-7, ACT-17, ACT-27, ACT-33
NS047	NRF	ND	ND
NS048	NG	ND	ND
NS050	*S.marcescens*	Aztreonam S, Ceftazidime S, Ciprofloxacin S, Co-trimoxazole S, Ertapenem S, Gentamicin S, Meropenem S, Piperacillin-Tazobactam S, Ampicillin R, Amoxicillin-	ND

		clavulanate R, Cefpodoxime R, Cefuroxime R	
NS049	NSG	ND	ND
JS007	NSG	ND	ND
NS042	*S. aureus*	Fucidic acid R, Clarithromycin S, Clindamycin S Flucloxacillin S, Mupirocin S, Tetracycline S	ND
NS041	*E. coli*	Ampicillin S, Amoxicillin-Clavulanate S, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-Tazobactam S, Meropenem S, Piperacillin-Tazobactam S	ACT-5
	S. marcescens	Ampicillin R, Amoxicillin-clavulanate R, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-Tazobactam S	
NS063	*K. pneumoniae*	Amikacin S, Amoxicillin S, Amoxicillin-clavulanate S, Aztreonam S, Cefixime S, Cefotaxime S, Ceftazidime S, Cefuroxime S, Ciprofloxacin S, Co-trimoxazole S, Ertapenem S , Gentamicin S, Meropenem S, Piperacillin-tazobactam S, Tigecycline S, Tobramycin S, Ampicillin R	ND
JS013	NSG	ND	ND
NS057	*Proteus* spp.	Ampicillin S, Amoxicillin-clavulanate S, Cefpodoxime S, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S	ND
NS067	NRF	ND	ND
NS069	NRF	ND	ND
YS013	*P. aeruginosa*	Ciprofloxacin S, Gentamicin S, Meropenem S	ND
JS015	*P. aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem R, Piperacillin-tazobactam R	TEM-4
JS016	Negative	ND	ND

NS080	*P. aeruginosa*	Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S	ND
JS019	*S. aureus*	Clindamycin S, Doxycycline S, Erythromycin S, Flucloxacillin S, Mupirocin S	TEM-4
YS057	NRF	ND	TEM-4
YS059	*S. aureus*	Clarithromycin S, Co-trimoxazole S, Flucloxacillin S, Tetracycline S.	ND
	E. coli	Amoxicillin-clavulanate S, Aztreonam S, Co-trimoxazole S, Fosfomycin S, Gentamicin S, Amoxicillin R	
YS020	NSG	ND	ACT-4 , TEM-4
NS075	*P. aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem R, Piperacillin-tazobactam S	ND
NS070	*P.aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem R, Piperacillin-tazobactam S	ND
YS036	NG	ND	TEM-4
YS018	*H. influenzae*	Amoxicillin S, Co-trimoxazole S, Tetracycline S	ACT, CMY-98
	K. aerogenes	ND	
YS011	*P.aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem R, Piperacillin-tazobactam S	ND
NS077	NSG	ND	TEM-4
NS074	NSG	ND	TEM-4
YS044	NSG	ND	TEM-4
NS078	*P. aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem R, Piperacillin-tazobactam S	TEM-4
YS042	NSG	ND	ND

NS073	NG	ND	TEM-4
NS064	*H. influenzae*	Ampicillin S, Amoxicillin-Clavulanate S, Cefuroxime S, Ciprofloxacin S, Co-trimoxazole S, Tetracycline S	TEM-4
	S. aureus	Flucloxacillin S, Fusidic acid S, Mupirocin S, Tetracycline S, Clindamycin R, Erythromycin R	
SS004	*P.aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S	ND
YS014	NSG	ND	TEM-4
YS025	*K. oxytoca*	ND	ND
YS034	NRF	ND	TEM-4
NS079	*P. aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Meropenem S, Piperacillin-tazobactam S	ND
YS053	*P. aeruginosa*	Ceftazidime S, Ciprofloxacin S, Gentamicin S, Piperacillin-tazobactam S	ND
	S. aureus	Clarithromycin S, Flucloxacilin S, Tetracycline S	
NS071	*S. aureus*	Clindamycin S, Erythromycin S, Flucloxacilin S,Fusidic acid S, Mupirocin S, Tetracycline S.	mecA
	E. coli	Ertapenem S, Gentamicin S, Meropenem S, Amikacin I, Ampicillin R, Amoxicillin-Clavulanate R, Aztreonam R, Cefotaxime R, Cefpodoxime R, Ceftazidime R, Cefuroxime R, Ciprofloxacin R, Co-Trimoxazole R, Piperacillin-Tazobactam R, Tobramycin R	
NS072	NSG	ND	ND

R=resistant, S= sensitive, I= intermediate resistance, ND= Not Detected

Table 3.22B Resistance genes detected by ARMA in relation to phenotypic resistance of pathogens identified in INHALE samples*.

Sample ID	ARMA versus phenotypic resistance reported by routine	Number of AMR genes detected	Examples
NS041	Matched to phenotypic resistance	9	*bla*$_{\text{TEM}}$ in Enterobacteriaceae and blaACT in *S. marcescens*
NS043			
NS044			
JS015	Unlikely matched to reported phenotype	1	*bla*$_{\text{TEM}}$ in piperacillin-tazobactam resistance *P. aeruginosa*
JS019	Genes reported in culture-negative samples or samples with a susceptible isolate	14	Majority was *bla*$_{\text{TEM}}$ genes that are likely to be related with the lung microbiome
YS018			
YS034			
YS014			
NS073			
YS044			
NS077			
NS074			
YS036			
YS020			
NS071			
YS057			
NS064	Genes reported did not explain phenotypic resistance of the isolate	2	*bla*$_{\text{TEM}}$ genes found in samples with susceptible organisms (e.g. *bla*$_{\text{TEM}}$ for a susceptible *H. influenzae)*
NS078			

*27 genes detected in 47 samples

3.2.4 Concordance of metagenomics with PCR-based machines

INHALE samples (n=47) processed with the optimised CMg pipeline were also processed with two PCR-based platforms (CURETIS and BIOFIRE) for the detection of respiratory pathogens (previously described in 3.2.1). Hence, we compared results generated from metagenomics and BIOFIRE in order to examine the similarities and differences in their performance for diagnosing HAP and VAP. BIOFIRE results were only used for this comparison, as, BIOFIRE was the machine chosen for the INHALE RCT. Therefore, for this analysis when both tests (i.e. CMg and BIOFIRE) were in complete agreement for the detection/absence of pathogen, they were deemed as concordant and discordant if not. Partially concordant samples had at least one pathogen detected by both metagenomics and BIOFIRE (Table 3.23).

Metagenomics was concordant with BIOFIRE for 18/47 samples. Of these, both tests reported one organism in 8/18 samples (NS043, NS048, NS067, NS072, SS004,YS011, YS013, YS025) and two organisms in 1/18 samples (JS015); neither test reported pathogens for 9/18 samples (JS007, JS013, NS047, NS069 NS074, NS077, YS014, YS036 and YS057). The two techniques were discordant for 7/47 samples, from which, 4/7 metagenomics did not report pathogens detected by BIOFIRE (NS038, NS039, NS049 and YS042) and 1/7 BIOFIRE did not detect pathogen reported by metagenomic sequencing (YS044). In 2/7 discordant samples different pathogens were reported by each test; metagenomics reported *S. pneumoniae* in NS079, whereas BIOFIRE reported *E. coli* and *P. aeruginosa* and in NS073, BIOFIRE reported *S. pneumoniae* and CMg detected *E. coli, H. influenzae* and *S. aureus.*

For the remaining 22/47 samples the two platforms were partially concordant, with BIOFIRE reporting more pathogens than metagenomics in 15/22 partially concordant samples; one additional organism was reported in 10/15 samples (QS006, NS050, NS042, NS080, JS016, JS019, YS059, NS075, YS018, and NS064, two additional organisms were reported in 4/15 samples (NS037,NS071, YS034 and YS053) and four additional organisms were reported in 1/15 samples (NS041). In 5/22 partially concordant samples metagenomics detected more pathogenic organisms than BIOFIRE; one additional organism was reported in 3/5 samples (NS057, NS063 and NS078) and two additional organisms were reported in 2/5 samples (NS044 and YS020). In 2/22 partially concordant samples (QS005 and NS070) both metagenomics and BIOFIRE reported the same number of pathogens; i.e. *P. aeruginosa* was reported by both techniques for both samples, but metagenomics reported *Citrobacter freundii* (QS005) and *Morganella morganii* (NS070), whereas BIOFIRE reported *Enterobacter cloacae complex* (QS005) and *E. aerogenes* (NS070) (see Table 3.23 for a list of pathogens reported).

These results show that BIOFIRE detected more potential pathogens than metagenomics, suggesting that BIOFIRE was more sensitive than metagenomics. Overall, pathogens reported by metagenomics (in terms of pathogen detection and/or absence) were more similar to the results reported by culture than BIOFIRE, suggesting that the sensitivity of our CMg test is closer to the sensitivity of culture than the sensitivity of BIOFIRE.

Table 3.23: Comparison of pathogenic organisms detected by metagenomics and BIOFIRE.

Sample ID	**Pathogens identified by metagenomics**	**Pathogens identified by BIOFIRE**	**Concordance (C), Partial Concordance (P), Discordant (D)**
NS037	*Staphylococcus aureus*	*Enterobacter cloacae complex* *Haemophilus influenzae* *Staphylococcus aureus*	P
QS006	*Proteus mirabilis* *Haemophilus influenzae*	*Haemophilus influenzae* *Klebsiella pneumoniae* *Proteus spp.*	P
NS039		*Staphylococcus aureus*	D
NS038		*Serratia marcescens* *Staphylococcus aureus* *Streptococcus agalactiae*	D
QS005	*Pseudomonas aeruginosa* *Citrobacteri freundii*	*Enterobacter cloacae complex* *Pseudomonas aeruginosa*	P
NS043	*Escherichia coli*	*Escherichia coli*	C
NS044	*Enterobacter cloacae group* *Escherichia coli* *Klebsiella pneumoniae*	*Enterobacter cloacae complex*	P
NS047			C
NS048	*Haemophilus influenzae*	*Haemophilus influenzae*	C
NS050	*Serratia marcescens* *Haemophilus influenzae*	*Haemophilus influenzae* *Serratia marcescens* *Staphylococcus aureus*	P
NS049		*Streptococcus pneumoniae*	D
JS007			C
NS042	*Staphylococcus aureus* *Streptococcus pneumoniae*	*Haemophilus influenzae* *Staphylococcus aureus* *Streptococcus pneumoniae*	P
NS041	*Serratia marcescens* *Escherichia coli* *Klebsiella oxytoca*	*Enterobacter cloaceae complex* *Escherichia coli* *Haemophilus influenzae* *Klebsiella oxytoca* *Moraxella catarrhalis*	P

		Serratia marcescens *Staphylococcus aureus*	
NS063	*Escherichia coli* *Klebsiella pneumoniae*	*Klebsiella pneumoniae*	P
JS013			C
NS057	*Proteus mirabilis* *Haemophilus influenzae* *Staphylococcus aureus*	*Haemophilus influenzae* *Proteus spp.*	P
NS067	*Escherichia coli*	*Escherichia coli*	C
NS069			C
YS013	*Pseudomonas aeruginosa* *Stenotrophomonas maltophilia**	*Pseudomonas aeruginosa*	C
JS015	*Pseudomonas aeruginosa* *Escherichia coli*	*Escherichia coli* *Pseudomonas aeruginosa*	C
JS016	*Escherichia coli*	*Escherichia coli* *H. influenzae*	P
NS080	*Pseudomonas aeruginosa*	*Proteus spp.* *Pseudomonas aeruginosa*	P
JS019	*Staphylococcus aureus* *Streptococcus agalactiae*	*Staphylococcus aureus* *Streptococcus agalactiae* *Streptococcus pyogenes*	P
YS057			C
YS059	*Staphylococcus aureus*	*Escherichia coli* *Staphylococcus aureus*	P
YS020	*Staphylococcus aureus* *Klebsiella aerogenes* *Escherichia coli* *Haemophilus influenzae*	*Escherichia coli* *Haemophilus influenzae*	p
NS075	*Pseudomonas aeruginosa*	*Proteus spp.* *Pseudomonas aeruginosa*	P
NS070	*Pseudomonas aeruginosa* *Morganella morgannii*	*Enterobacter aerogenes* *Pseudomonas aeruginosa*	P
YS036			C
YS018	*Haemophilus influenzae* *Enterobacter aerogenes*	*Enterobacter cloaceae complex* *Enterobacter aerogenes* *Haemophilus influenzae*	P
YS011	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	C

NS077			C
NS074			C
YS044	*Escherichia coli*		D
NS078	*Pseudomonas aeruginosa* *Escherichia coli*	*Pseudomonas aeruginosa*	P
YS042	*Stenotrophomonas maltophilia**	*Streptococcus pneumoniae*	D
NS073	*Escherichia coli* *Haemophilus influenzae* *Staphylococcus aureus*	*Streptococcus pneumoniae*	D
NS064	*Haemophilus influenzae* *Staphylococcus aureus*	*Haemophilus influenzae* *Moraxella catarrhalis* *Staphylococcus aureus*	P
SS004	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	C
YS014			C
YS025	*Klebsiella oxytoca*	*Klebsiella oxytoca*	C
YS034	*Escherichia coli* *Klebsiella pneumoniae*	*Enterobacter cloaceae complex* *Escherichia coli* *Klebsiella oxytoca* *Klebsiella pneumoniae*	P
NS079	*Streptococcus pneumoniae*	*Escherichia coli* *Pseudomonas aeruginosa*	D
YS053	*Pseudomonas aeruginosa*	*Escherichia coli* *Pseudomonas aeruginosa* *Staphylococcus aureus*	P
NS071	*Staphylococcus aureus*	*Escherichia coli* *Staphylococcus aureus* *Streptococcus agalactiae*	P
NS072	*Proteus mirabilis*	*Proteus spp.*	C

**S. maltophilia* cannot be detected by BIOFIRE so YS013 is considered as concordant for this analysis

3.3 Application of clinical metagenomics for public health

The respiratory metagenomics pipeline demonstrated excellent performance for the diagnosis of LRTIs (described in 3.1) and for the diagnosis of nosocomial pneumonia for the INHALE trial (described in 3.2). Next, we sought to implement our CMg pipeline for public health applications – could the data generated directly from primary samples using our methods also be used for rapid outbreak detection, surveillance and strain characterization? To investigate this, we tested a set of retrospectively collected respiratory samples (collected from >5 years ago) previously tested for *Legionella* spp. using traditional methods and qPCR at Public Health England (PHE), Colindale. The specific aim was to determine whether we could detect *Legionella* spp. directly in respiratory samples, correctly identify the species present and get enough genome coverage to sequence type the strains without the need for culture. Identification of the sequence type is invaluable for *L. pneumophila* outbreaks as it can identify the source of the infection and identify epidemiological links between environmental and clinical isolates (234). There are clear benefits to being able to sequence type directly from respiratory samples within hours such as: 1) rapidly determining whether the patient is actually infected with *Legionella*, 2) rapidly determining whether the infection is part of an outbreak or is a sporadic case and 3) rapidly determining the common source of infection in outbreak situations. This would lead to faster containment, thereby reducing the size and severity of potential outbreaks. Infected patients would also receive appropriate treatment in a timely fashion, improving patient outcomes and reducing patient morbidity and mortality.

3.3.1 Identification of *Legionella* spp. using clinical metagenomics

Clinical samples n=48, (42 culture-PCR-positive for *Legionella* spp. and 6 culture-PCR-negative - Table 3.24) were processed with the clinical metagenomics pipeline (Figure 3.1) using the sequence data collected after 24 hours. The data was analyzed with Centrifuge (198) using default parameters (centrifuge score of >300) followed by Supernatant (developed by Natalie Groves, PHE), to filter the Centrifuge output, thereby improving the read identification quality (described in section 2.8.4). Samples were considered positive for *Legionella* species (*L. pneumophila*, *L. longbeachae* and *L. sainthelensi*) if supernatant-extracted Legionella reads were >0.01% of microbial reads and were ≥5 reads in total. A sequencing positive control was included in 4/6 multiplex sequencing runs (1 ng of *L. pneumophila* extracted DNA was added) to monitor sequencing failures. Hence, a negative control rule was applied to remove contamination and/or barcode leakage that could be introduced from the sequencing positive control, where the same number of *Legionella* reads observed in the extraction control was removed from the samples tested on the same flowcell. The microbial read proportion threshold for *Legionella* positive samples was set lower than for previous applications due to the relatively low abundance of *Legionella* spp. in some of the samples pre-depletion as determined by qPCR (Cq median value = 28.37, interquartile range 25.35-30; maximum 39.03). The median *Legionell*a classified microbial read proportion in positive samples was 3.03% (interquartile range 0.11%-20.75; maximum= 76%) after removing the contaminant reads, whereas the highest *Legionella* read proportion number observed in culture-negative samples was 0.005% of classified microbial reads. Hence, we determined that the best threshold to discriminate negative and positive samples was >0.01% of *Legionella* spp. reads as a proportion of total microbial reads.

CMg results were concordant with routine diagnostics (PCR and culture) for 30/36 *Legionella pneumophila*-positive samples using the thresholds described above. In 22/30 samples *L. pneumophila* reads were ≥1% of microbial reads (L2, L4, L6, L7, L8, L9, L10, L11, L12, L13, L15, L16, L17, L18, L20, L23, L24, L31, L34, L40, L41 and L42) and in the remaining 8 samples (L1, L3, L5, L14, L32, L35, L36 and L38) *L. pneumophila* reads represented ≥0.01% of microbial reads (Table 3.24). CMg was also in agreement with routine testing for 6/6 samples where non-pneumophila *Legionella* spp. were reported by routine methods. These included *L. sainthelensi* in L30 and *L. longbeachae* in L25, L26, L27, L29 and L30. *Legionella* reads were low in one sample (0.1%, L27) but significantly higher in the other samples (13.3-73.3% of microbial reads – Table 3.24). *Legionella* spp. were not detected by CMg after applying the chosen thresholds in the six PCR-and-culture negative samples i.e.; L33, L37, L39, L45, L47 and L48 (Table 3.24). Additional *Legionella* spp. (≥5 microbial classified reads and ≥0.01% of microbial reads) were detected by CMg in seven *L. pneumophila* positive samples (L12, L25, L26, L27, L29, L30 and L31) that were not reported by culture or PCR (Table 3.24). This was likely due to *k*-mer misclassification (multiple *Legionella* spp. were also reported in the positive controls spiked with *L. pneumophila* DNA), therefore these samples were not reported as false positives (Table 3.24).

The six *L. pneumophila* positive samples missed by CMg (L19, L21, L22, L43, L44 and L46) had higher Cqs than samples correctly identified by CMg (false negative samples mean Cq = 33.24 vs true positive samples mean Cq = 26.6), which indicates that missed samples had very low cell numbers (<100) (Table 3.24). Additional analysis with a *Legionella* qPCR assay (methods section 2.3.2) was carried out on the pre- and post-depleted DNA extract on 24/42 positive samples. *L. pneumophila* DNA loss was reported in 23/24 of the frozen samples processed (median 30.27-fold, interquartile range 15.13-84.56; maximum 7702-fold) suggesting that freeze-thawing lyses *L. pneumophila* cells and the DNA is then lost during the host depletion step of the CMg pipeline (see Table 3.25).

Another issue observed from analyzing the processed extraction controls was the presence of cross-contamination between the sequencing positive control and clinical samples. In the process extraction control 8, *L. pneumophila* was identified (511 reads reported), suggesting contamination occurred during sample processing (barcode leakage is not typically this high). Hence, contaminant and pathogenic reads could not be distinguished in L43, L44 and L46 and these samples were deemed negative.

Based on these results CMg was 85.71% (95% CI: 71.46% to 94.57%) sensitive and 100% (95% CI: 54.07% to 100%) specific for the detection of *Legionella* spp. when compared against routine culture+PCR testing (Table 3.24).

Table 3.24: Clinical metagenomics output on samples tested against routine testing for the *Legionella* study.

Sample number	**No of pass* reads**	**Microbial reads**	***Legionella* spp. reads**	***Legionella* spp. reads+**	**CMg output^**	**PCR + Culture**	**PCR Cq Pre-depletion**
L1	57932	6154	7	7	*L. pneumophila*	*L. pneumophila*	22
L2	22150	9093	5754	5754	*L. pneumophila*	*L. pneumophila*	23
L3	220092	4929	32	32	*L. pneumophila*	*L. pneumophila*	26.3
L4	16334	5107	351	351	*L. pneumophila*	*L. pneumophila*	27
L5	783278	146282	95	95	*L. pneumophila*	*L. pneumophila*	28.75
L6	51322	17151	240	240	*L. pneumophila*	*L. pneumophila*	29.1
Processed extraction control 1	0	0	0	0	N/A	N/A	N/A
L7	39520	25878	5373	5373	*L. pneumophila*	*L. pneumophila*	18.5
L8	9250	5720	660	660	*L. pneumophila*	*L. pneumophila*	29.13
L9	28302	179	64	64	*L. pneumophila*	*L. pneumophila*	29.09
L10	351262	5,862	2187	2187	*L. pneumophila*	*L. pneumophila*	29.09
L11	28542	368	5	5	*L. pneumophila*	*L. pneumophila*	23
L12	79260	52013	10034	10034	*L. pneumophila*	*L. pneumophila*	22
			7	7	*L. longbeachae*		
L13	4277	433	5	5	*L. pneumophila*	*L. pneumophila*	30
Processed extraction control 2	0	0	0	0	N/A	N/A	N/A
L14	85165	30584	164	107	*L. pneumophila*	*L. pneumophila*	30
L15	18166	13295	333	276	*L. pneumophila*	*L. pneumophila*	29
L16	1098	125	101	44	*L. pneumophila*	*L. pneumophila*	27
L17	18897	2213	159	102	*L. pneumophila*	*L. pneumophila*	29.09
L18	19506	476	164	107	*L. pneumophila*	*L. pneumophila*	28.86

Processed extraction control 3	637	71	57	0	N/A	N/A	N/A
Sequencing positive control 1	873460	862195	851458	851401	*L. pneumophila*	N/A	N/A
L19	146932	112269	1	1		*L. pneumophila*	26.91
L20	9015	181	40	40	*L. pneumophila*	*L. pneumophila*	26
L21	4680	827	0	0		*L. pneumophila*	30
L22	605760	512904	20	20		*L. pneumophila*	30
L23	112	75	57	57	*L. pneumophila*	*L. pneumophila*	29
L24	25728	5247	164	164	*L. pneumophila*	*L. pneumophila*	27
Processed extraction control 4	0	0	0	0	N/A	N/A	N/A
Sequencing positive control 2	18464	18442	17792	17792	L. pneumophila	N/A	N/A
			6	6	*L. longbeachae*		
L25	325797	25342	17605	17605	*L. longbeachae*	*L. longbeachae*	24
			1213	1213	*L. pneumophila*		
			295	295	*L. sainthelensi*		
			227	227	*L. fallonii*		
			64	64	*L. spiritensis*		
			136	136	*L waltersii*		
			5	5	*L. oakridgensis*		
L26	339422	1081	144	144	*L. longbeachae*	*L. longbeachae*	17.44
			12	12	*L. pneumophila*		
L27	322839	10063	11	11	*L. longbeachae*	*L. longbeachae*	NR
			7	7	*L. pneumophila*		
L28	70174	154	32	32	*L. longbeachae*	*L. longbeachae*	33.2
L29	373634	182197	34,493	34,493	*L. longbeachae*	*L. longbeachae*	21.87

			1259	1259	*L. pneumophila*		
			734	734	*L. sainthelensi*		
			226	226	*L. waltersii*		
L30	12025	60	44	44	*L. sainthelensi*	*L. sainthelensi*	21
			5	5	*L. pneumophila*		
Processed extraction control 5	0	0	0	0	N/A	N/A	N/A
Sequencing positive control 3	112094	111870	111070	111070	*L. pneumophila*	N/A	N/A
			12	12	*L. sainthelensi*		
			18	18	*L. clemsonensis*		
			11	11	*L. fallonii*		
L31	3600	2856	2122	2111	*L. pneumophila*	*L. pneumophila*	28
			7	7	*L. clemsonensis*		
L32	75949	29135	21	10	*L. pneumophila*	*L. pneumophila*	32
L33	101008	69600	15	4			N/A
L34	79497	8405	175	164	*L. pneumophila*	*L. pneumophila*	26
L35	134524	92347	31	20	*L. pneumophila*	*L. pneumophila*	32.55
L36	80937	1215	20	9	*L. pneumophila*	*L. pneumophila*	25.8
Processed extraction control 6	366	63	11	0	N/A	N/A	N/A
Sequencing positive control 4	86616	86443	85660	85649	*L. pneumophila*	N/A	N/A
			14	14	*L. clemsonensis*		
L37	1777	224	5	0			N/A
L38	78454	44273	32	25	*L. pneumophila*	*L. pneumophila*	26.94
L39	214	32	7	0			N/A
L40	5411	105	20	13	*L. pneumophila*	*L. pneumophila*	30.37
L41	1170	109	21	14	*L. pneumophila*	*L. pneumophila*	22

L42	434577	374	18	11	*L. pneumophila*	*L. pneumophila*	NR
Processed extraction control 7	4373	301	7	0	N/A	N/A	N/A
Sequencing positive control 5	94189	93951	93561	93554	*L. pneumophila*	N/A	N/A
			12	12	*L. clemsonensis*		
L43	44830	19866	259	0		*L. pneumophila*	39.03
L44	190091	40386	371	0		*L. pneumophila*	37.53
L45	575	406	143	0			N/A
L46	18568	10930	111	0		*L. pneumophila*	36
L47	488	165	64	0			N/A
L48	76258	46094	344	0			N/A
Processed extraction control 8	984	560	511	0	N/A	N/A	N/A
Sequencing positive control 6	361034	359583	346391	345880	*L. pneumophila*	N/A	
			68	68	*L. clemsonensis*		

[*]total number of passed reads from 24 hrs after basecalling with guppy and demultiplexing with porechop. [+] Legionella reads after applying the negative control rule. [^]CMg output presented after applying the chosen thresholds. NR=Not Reported.

Table 3.25. qPCR* results of pre- and post-depletion on *L. pneumophila* positive samples.

Sample number	**PCR Cq Pre-depletion**	**PCR Cq Post-depletion**	***L. pneumophila* loss/gain ΔCq (fold loss)**
L1	22	25.13	3.13 (8.75)
L2	23	35.91	12.91 (7702)
L3	26.3	28.29	1.99 (3.97)
L4	27	31.82	4.82 (28.24)
L5	28.75	35.51	6.76 (108.38)
L6	29.1	33.41	4.31 (19.83)
L7	18.5	24.76	6.26 (76.63)
L8	29.13	33.41	4.28 (19.42)
L9	29.09	34.44	5.35 (40.78)
L10	29.09	34	4.91 (30.06)
L11	23	25.75	2.75 (6.72)
L12	22	24.78	2.78 (6.86)
L13	30	27.23	2.77 (6.82)
L14	30	35.67	5.67 (50.91)
L15	29	40	11 (2048)
L16	27	31.82	4.82 (28.24)
L17	29.09	34.44	5.35 (40.78)
L18	28.86	36	7.14 (141.04)
L19	26.91	35.8	8.89 (474.41)

L20	26	31.34	5.34 (40.50)
L21	30	34.11	4.11 (17.26)
L22	30	31.73	1.73 (3.31)
L23	29	33.93	4.93 (30.48)
L24	27	34.7	7.7 (207.93)

**L. pneumophila* qPCR assay described in 2.3.2

3.3.2 Sequence based typing of *L. pneumophila* using clinical metagenomics data

Following identification of *L. pneumophila*, we attempted to further characterize the pathogen using metagenomic data for genome assemblies. For this analysis, *de novo* genome assemblies were initially generated with Supernatant-extracted reads (i.e. *L. pneumophila* and *L. longbeachae*) using Canu (described in 2.8.4.1). In total, assemblies could only be generated for pathogens identified in 5/36 samples reported positive by metagenomics (L2, L7, L12, L25 and L29) using data after 24 hours of sequencing, from which three were assembled genomes of *L. pneumophila* and two of *L. longbeachae.* The *L. longbeachae* reads of L25 were assembled in 74 contigs (longest contig=277 Kbp and n50 length=137 Kbp) and the genome assembly of L29 consisted of 19 contigs (longest contig= 1272 Kbp and n50 length=537 Kbp). The *L. pneumophila* reads were assembled in 246 contigs (longest contig=12871bp and n50 length=3579 bp) for L2, 180 contigs (longest contig= 80323 bp and n50 length=11551 bp) for L7 and 239 contigs (longest contig=47582 bp and n50 length=10877 bp) for L12 (see Table 3.26A).

Assemblies were used for sequence based typing by MLST (https://github.com/tseemann/mlst) for *L. pneumophila*-positive samples only (described in 2.8.4.3). According to routine sequence-based typing (described in 2.3.2), the sequence type (ST) reported for L7 was 1, for L12 was 18 and no ST was reported for L2. Sequence-based typing (SBT) using the *de novo* genome assemblies generated from CMg data could not report a ST, hence comparison against routine SBT could not be done (Table 3.26B).

Failure to identify a sequence type was most likely due to the low coverage observed in the majority of the samples after metagenomic sequencing (<10x genome coverage). Therefore, we attempted to generate assemblies through a reference-based approach, aiming to increase the genome coverage observed with the previous approach. Hence, *L. pneumophila* samples with >1000 *L. pneumophila* classified reads (n=5) after 24 hrs of sequencing were mapped against a concatenated reference containing all complete genomes of *L. pneumophila* available in NCBI using minimap2 (246) (described in 2.8.4.1). Reference-based assemblies were produced from four samples (Table 3.26A), including the three samples, that *de novo* assemblies were generated (L2, L7, L12) and L10 (no ST was reported for L10 by the clinical lab). An improvement was observed in the *L. pneumophila* reference-based assemblies for L2 and L12 when compared to *de novo* assemblies, i.e. L2 consisted of 239 contigs (*de novo* assembly consisted of 246) and longest contig for L12 was 144 kb versus 47 kb with the *de novo* assembly. However, for L7, the number of contigs increased from 180 to 209 for the reference-based assembly and for L10 the reference-based assembly had 61 contigs (longest contig=32362 bp and n50 length=7663 bp) - see Table 3.26A.

SBT was still not possible using the reference-based assemblies and the correct sequencing type or allele ID could not be identified. Only two correct allele IDs were reported for *mompS* (2) and *proA* (5) for L12 when compared to SBT reported by the clinical lab (Table 3.26B).

Table 3.26A: Comparison of *L. pneumophila de* novo and reference-based assemblies.

	***De* novo assemblies**			**Reference-based assemblies**			
Sample ID	L2	L7	L12	L2	L7	L10	L12
Number of contigs	246	180	239	239	209	61	208
Longest contig (bp)	12871	80323	47582	12874	80592	32362	144460
n50 length (bp)	3579	11551	10877	3666	14709	7663	24138

After failing to identify STs by using genome assemblies (either *de novo* or reference-based) we attempted to perform SBT using unassembled reads. We used Krocus (249), a tool which uses basecalled FASTQ reads for multi-locus-sequence based-typing and is specifically designed to tolerate the high single read error-rate of nanopore reads. Krocus was initially tested using mapped-reads from different timepoints (5, 20 and 40 min) of sequencing data of sample S1 (an *E. coli*-positive sample previously processed with the optimised CMg pipeline in the proof of concept study - described in 3.1.13) to identify the minimum genome coverage required by the tool to provide an ST. Krocus reported the correct *E. coli* ST (ST131) using the data after 40 min of sequencing from S1. This suggested that the minimum coverage needed to provide a ST with Krocus is ≥13X from metagenomic data (13.48x genome was recovered in S1 and 414,241 *E. coli* reads were reported after 40 min of sequencing for S1 - see Figure 3.2E).

L12 was the only *Legionella* sample that produced the minimum coverage required (16x coverage after 24 hrs of sequencing) and was the only sample used for this analysis. The correct allele IDs for 5/7 housekeeping genes (*pilE, asd, mip, proA,* neuA) were identified using Krocus when compared with results reported by the clinical laboratory (Table 3.26B) However, as the correct allele IDs for 2/7 housekeeping genes (*flaA* and *mompS*) were not identified, an ST identification was still not possible. Failure to identify a ST using data generated by the CMg pipeline was due to the low genome coverage obtained – this was directly related to the use of historic frozen samples and the loss of bacterial DNA observed after depletion.

Table 3.26B: Sequence based typing (SBT) compared against routine testing of *L. pneumophila* samples.

		L2	**L7**	**L10**	**L12**
SBT- Routine testing	***flaA***	2	1	1	2
	pilE	0	4	4	10
	asd	17	3	3	9
	mip	1	1	0	13
	mompS	9	1	0	2
	proA	4	1	0	5
	neuA	1	1	1	6
	Sequence type	No ST	1	No ST	18
SBT- assembled CMg	***flaA***	Not found	39	Not found	9
	pilE	58	68	Not found	76
	asd	Not found	58	Not found	58
	mip	Not found	89	Not found	91
	mompS	Not found	Not found	Not found	2
	proA	48	2	Not found	5
	neuA	Not found	Not found	Not found	65
	Sequence type	No ST	No St	Not found	Not found
SBT- Krocus	***flaA***	N/A	N/A	N/A	32
	pilE	N/A	N/A	N/A	10
	asd	N/A	N/A	N/A	9
	mip	N/A	N/A	N/A	13
	mompS	N/A	N/A	N/A	63
	proA	N/A	N/A	N/A	5
	neuA	N/A	N/A	N/A	6
	Sequence type	N/A	N/A	N/A	No ST

4. Discussion

LRTIs are considered as the deadliest communicable disease. Current diagnostic methods used for LRTIs are too slow (48-72hours), hence contribute to the overuse of empiric antibiotics in respiratory infections, which increases the emergence of antibiotic resistance in respiratory pathogens (10, 62). Delayed targeted antibiotic treatment, prolongs hospital stays, significantly increasing hospital costs (80, 253). The need for rapid diagnostics to guide appropriate antibiotic therapy, thereby reducing patient morbidity and mortality and the emergence of antimicrobial resistance has been emphasized by the UK government in the 5-year AMR action plan and the O'Neil report (82, 83). Although current FDA/CE-IVD-approved molecular-based tests (such as PCR-based tests) provide results in hours, they are not comprehensive enough to replace current diagnostics (254).

Clinical metagenomics (CMg)-based tests have demonstrated the potential to revolutionise clinical microbiology and overcome challenges of current diagnostic tests (102). The main aim of my PhD was to develop a CMg pipeline that could replace current diagnostics tests and detect pathogens and relevant resistance genes in a rapid turnaround time. The metagenomics pipeline developed in this study was mainly focused on the diagnosis and characterization of LRTIs. We considered, LRTIs as a good starting point for the application of CMg, for two reasons: i) the importance of respiratory infections on the public and economic sector (previously discussed) and ii) pathogen levels present in respiratory samples (>100-1000 cells). Lower levels of pathogens are challenging to detect and discriminate from contamination and requires the use of additional steps in the pipeline prior to sequencing, such as a sensitive whole genome amplification, to amplify low amounts of pathogen DNA (femtogram range) remaining after host depletion.

One of the main shortcomings of CMg-based assays is turnaround time. An example, is the service for the diagnosis of encephalitis/meningitis offered by the UCSF Clinical Microbiology Laboratory (https://nextgendiagnostics.ucsf.edu/our-diagnostic-lab/). With this pipeline, pathogens are detected via very deep sequencing (no host depletion) where only a small fraction of the genome is recovered, meaning results are provided in a slow turnaround (>48 hrs, excluding sample transportation) (136). This CMg pipeline, is not able to replace culture as it can only be used as a last-resort diagnostic for cases where commonly used tests failed to identify the cause of infection. The goal of this study, and of the O'Grady group in general, is for CMg to be used as the primary test for the diagnosis of infectious diseases, replacing culture-based diagnostics.

The pipeline developed in this study, consists of human DNA depletion, microbial DNA extraction, sequencing of low-biomass samples and data analysis. Pathogen and resistance genes can be rapidly detected (~6 hrs), which will enable clinicians to choose the appropriated targeted antimicrobial therapy avoiding a second dose of broad-spectrum antibiotics. The rapid turnaround time achieved with our pipeline was achieved by developing a rapid and inexpensive host depletion method and the use of rapid nanopore library preparation and real-time sequencing.

4.1 Optimisation of the Clinical metagenomics pipeline

4.1.1 Host depletion

As previously discussed, in clinical samples the ratio of human:microbial DNA is high and therefore, rapid pathogen detection in clinical samples with metagenomic sequencing would not be possible without efficient removal of host nucleic acid or enrichment of pathogen DNA prior to sequencing (122). The O'Grady lab has significant expertise in the development of host depletion methods, including differential cell-lysis approaches, and have applied them in urine (129) and blood samples (255). However, for this study we aimed to develop a pipeline that would be cheaper and faster, which would make it easier to implement in a clinical microbiology laboratory setting.

A review of the literature suggested that saponin was a good candidate for depleting human DNA (174) and at the beginning of this PhD study, we found a saponin-based method optimised in whole blood (177, 178, 256). Based on the methods described in these studies, we adjusted and optimised the saponin-based method described in this study, for use in respiratory samples (described in 3.1.1 and 3.1.2) which resulted in a host depletion method capable of removing >99.99% human DNA in respiratory samples (10^4-fold enrichment).

It should be noted that prior to optimising the host depletion method we also observed a high number of depletion failures mainly in sputum samples. Purulent sputum is a complex matrix mainly consisting of mucus and WBCs, making it viscous and hard-to-work with. However, sputum samples and ETAs that were previously treated with sputasol had better depletion rates. This was probably due to sputasol breaking down the mucus and making WBCs more accessible to saponin. We recommend that our saponin-depletion method be coupled with a sputasol-treatment step to avoid depletion failures in respiratory samples (if the sputum remains viscous, a second treatment should be performed). In our method we use a higher saponin concentration and extended duration of the treatment than what was used in previous studies (177, 178) to improve performance in sputum samples. We also knew from previous experience within the O'Grady research lab, with blood samples (257), that the HL-SAN nuclease does not lose efficiency in clinical samples when used with very high salt buffer (>5M) unlike other commonly-used nucleases (such as the Turbo DNase) that lose activity in complex clinical samples such as sputum (in our hands). The high salt concentration is also, likely important for efficient depletion of human DNA, because histones (proteins associated with chromatins) undergo certain rearrangements making the DNA more accessible in stress conditions such as high salt (258). The addition of the 5.5M salt HL-SAN buffer, results in such rearrangements, making the human DNA more accessible for digestion with the nuclease. Therefore, the highly salt-tolerant HL-SAN nuclease is better suited for this application than other nucleases which cannot withstand high salt concentrations.

The optimised version of our pipeline provided more robust depletion rates in comparison to reported results in the literature of the commercially-available host depletion kits or microbial enrichment kits. The MolYsis Kit, that also utilises a differential-lysis approach with chaotropic agents, has been reported to provide 10^4-fold in PJI samples and oral samples (259) but was reported to lyse Gram-ve bacteria such as *Pseudomonas* spp. (166). The NEBnext microbiome kit that promises an efficient microbial enrichment has been reported to lose efficiency in clinical samples such as sputa (165). This kit is part of the pipeline used by the UCSF Clinical Microbiology Laboratory and has been reported to enrich microbes only by ~2.5-fold in CSF samples (136) or <100-fold in other sample types (165).

We did not report any significant bacterial loss with the optimised version of the saponin-based host depletion (average 2.63-fold loss (see 3.1.9) according to the 16S rRNA qPCR assay performed on depleted samples and undepleted controls). On average 86.65% of classified reads were microbial according to MinION sequencing and up to 10^4 of human DNA was depleted according to qPCR results for any of the pathogens tested, with the exception of *S. pneumoniae* (discussed later in this section).

Qiagen also has a saponin-based patent for the differential lysis of human/animal cells (260). In this protocol, 500 µL of 7.73 wt-% of saponin is used and samples are incubated for 30 min followed by nuclease treatment and centrifugation steps to remove lysed cells and digested nucleic acid. Although the depletion rate in depleted samples and undepleted controls was not reported, Ct values of human DNA target in depleted samples were close to the detection limit of the rt-PCR used (reported CT= 31->35) (260). Reported human DNA levels of this patent were similar with human DNA levels reported in samples processed with the optimised method (average Cq=32.5 of human DNA tested with probe-based qPCR assay post-depletion). Our saponin treatment however, is faster (10 min vs 30min) and also it is unclear how the Qiagen method would perform in a complex sample type like sputum (Qiagen method tested on whole blood and swab samples).

Other in-house developed pipelines, such as the one reported by Hasan *et al.* (167), also use a saponin-based depletion method. Their pipeline was consisted of a saponin-based host cell depletion (0.025% final concentration) followed by digestion of nucleic acid with Turbo DNAse and microbial DNA extraction. Satisfactory depletion rates were reported (1.9%-2.1% of relative human DNA quantity in depleted samples compared to 100% quantity in undepleted controls) with a minimal effect in spiked organisms in CSF and nasopharyngeal aspirates (NPA) (167).

However, it is likely that the depletion efficiency of their method would decrease in sputum samples due to the DNase used – in our hands the Turbo DNase lost efficiency in mucoid respiratory samples (discussed earlier).

Rapid, inexpensive host depletion coupled with real-time sequencing should be considered as a key factor in the application of CMg-based assays for the rapid diagnosis of not only LRTIs but other infectious diseases. Our method has been successfully applied in other samples types (such as blood and PJIs) by other members of the O'Grady group (during and after this PhD study) and by or in collaboration with other groups e.g. in urine samples where it enabled the recovery of ≥ 92.8x of the *N. gonorrhoea* genome (261, 262).

4.1.2 Microbial DNA extraction

Approaches for microbial DNA extraction have been extensively investigated in microbiome studies as a non-efficient or biased extraction would lead to a false representation of the microbiome (263). Studies investigating the extraction efficiency of different approaches, like chemical, enzymatic and bead-beating have concluded that the most accurate representation of the microbiome is retrieved when bead-beating is incorporated (263, 264). In a recent study, where chemical and bead-beading based lysis were compared in saliva, the bead-beating extraction provided greater yield of microbial DNA when Gram +ve bacteria were present in the samples (265). We also demonstrated that the best approach for efficient unbiased DNA extraction was to include a bead-beating step combined with a chemical based-extraction.

At the early stages of the study we were satisfied with the results provided by utilising a chemical-based extraction but after evaluating the performance of the pilot method metagenomic sequencing missed an *S. aureus* culture-positive sample. This was resolved with the addition of bead-beating as all *S. aureus* culture-positive samples were correctly identified by metagenomics. Although we mainly focused on the identification of bacterial LRTIs in this study, the addition of the bead-beating step makes this pipeline suitable for the detection of fungal pathogens. While diagnosis of fungal pathogens was not investigated in depth in this study, fungi were identified in a high proportion of samples tested in a study performed by collaborators at Pittsburgh Medical School using our optimised pipeline (185).

For microbial DNA purification, we used an automated system (MagNA Pure Compact 2.0) but we did not test any other automated systems or manual extraction kits. The MagNA Pure was chosen as it is used by clinical microbiology laboratories (e.g. PHE), has a rapid turnaround (25 mins) and automated/standardises the purification step of the pipeline. It is possible that other manual or automated extraction/purification methods may have yielded more or higher quality DNA from sputum samples. Other methods have been tested in the O'Grady lab – typically the automated magnetic bead based methods were better than manual methods and MagNA Pure performed well in comparison. The lab has recently moved to the Promega Maxwell (with the Maxwell® Pure Food Pathogen kit) as it can process more samples and produces a higher yield from sputum, although it takes about 10 minutes longer to run.

4.1.3 Nanopore Sequencing of low-biomass samples

Nanopore sequencing (MinION) was exclusively used for this study. Nanopore sequencing overcomes the main limitations of other sequencing platforms related to implementation in clinical microbiology settings. These are: i) real-time data acquisition and analysis, ii) reduced cost for low-throughput sequencing and iii) small footprint (122). The long reads generated using nanopore sequencing also have some advantages over short reads such as AMR-gene host identification and genome/plasmid assembly. However, nanopore sequencing was still under development and at the beginning of this PhD study we observed poor flowcells and inconsistency in library preparation kit performance. Despite these early difficulties, rapid development of nanopore technology over the course of my PhD resulted in robust products suitable for clinical application. R9.4.1 flow cells have proven to be highly reliable, providing sequencing quality and data yields comparable to other sequencing platforms (120).

At the beginning of this study, all nanopore sequencing kits required at least 1µg of input DNA. As discussed earlier, respiratory samples are low biomass samples after host depletion, hence have low DNA quantities often not detectable with fluorescent-based assays (122). This limitation was overcome when ONT released a rapid PCR-based library preparation (SQK-RLI001) that only required 10ng of input DNA (described in 3.1.3). Another advantage of this workflow was that it allowed amplification and preparation of samples in a simple and fast manner, which makes it easier to implement in a clinical lab. This workflow follows a similar approach used by Illumina’s Nextera XT DNA Library Prep Kit (136) which is used in the CMg diagnostic assays implemented in the UCSF Clinical Microbiology Laboratory. We used this original kit on a small number of depleted respiratory samples and it provided satisfactory results in terms of yield and turnaround.

Shortly after, ONT released the multiplexed version of this kit (SQK-RLB001 and later SQK-RPB004). At the beginning of the study however, performance in respiratory samples wasn’t as good as with the original kit. We worked together with ONT to optimize this (described in 3.1.3). Washing the MagNA Pure extracted DNA, increasing the amount of transposase (FRM) and increasing the PCR reaction volume from 50ul to 100ul improved performance. These steps were introduced to overcome the inhibitory effect of the sputum matrix on these enzymatic reactions. ‘Cleaner’ DNA extracts would mean that the extra washing step would not be necessary. Removal of this step would reduce method complexity and turnaround time. Also, increased reagent volumes might not be necessary if the DNA was cleaner, reducing costs. New extraction methods are constantly evaluated in the O’Grady lab to find a method that produces suitably clean DNA in a short turnaround (discussed earlier).

We sequenced six samples and a negative control per flowcell throughout the study and didn’t investigate increased or decreased multiplex sample numbers. We estimated that sequencing six samples per flow cell would provide a good balance in terms of cost, turnaround time and genome coverage and this worked well in our hands. Samples could be run at a cost of US$130 per sample (122) and allowed enough genome coverage to be recovered for pathogen and resistance gene detection with using only 2 hrs of sequencing data. Cost could be reduced by multiplexing more samples, but this may have an impact on turnaround time. Also, when testing patients for HAP/VAP for example, only a few patients per day might need to be tested, and if you wait for 11 or 12 samples before you test, patients will be waiting days for their results, negating the use of a rapid test. Towards the end of the study, Flongle was made available by ONT. Flongle flowcells costs $90, hence are suitable for testing 1-2 samples at a time. This provides the flexibility in throughput necessary for clinical implementation.

4.1.4 Data analysis for bacterial identification and resistance gene detection

Bioinformatics pipelines used for the analysis of metagenomics data have the difficult task of accurately classifying microbes from a massive database of microbial sequences along with detecting antimicrobial resistance genes. Additionally, in order for the metagenomics pipeline to be implementable in clinical microbiology, the pipeline needs to present the results in an easy-to-interpret format (203). This would allow the biomedical scientist to interpret and report results just as easy if it would be to 'read' a plate. An example of a clinically implemented CMg pipeline is SURPI (266), which was developed and implemented by the Chiu lab at UCSF. The pipeline identifies microbes in a rapid turnaround (11 min to 5 hrs depending on read count analysed) and results are presented in an easy-to-interpret summary (136). In this study we used ONT's EPI2ME Antimicrobial Resistance pipeline for pathogen identification and resistance gene detection, which analyses metagenomic data rapidly and presents results in an easy-to-interpret format. This pipeline combines 'Centrifuge' kmer based tool for read identification using the RefSeq database with 'Minimap2' mapping of reads to the CARD database to identify resistance genes (described in 2.8.2 and 2.8.3).

The quality of the chosen microbial and resistance gene databases has a direct impact on the accuracy of microbial classification and resistance gene detection. Comprehensive databases are usually overpopulated with model organisms such as *E. coli* genomes, which can skew the analysis of closely related species towards the overrepresented species in the database used. This is a particular problem with *S. pneumoniae* and related species in respiratory samples. Databases used for the analysis of metagenomic data should be curated, where incomplete genomes would be removed and representation/addition of all the relevant taxa of pathogens and commensals should be evaluated (266). Curation of databases should be done by experts in the field to remove unrelated sequences that are beyond the scope of the study. Analysis would then be more reliable and faster as unnecessary 'matching' against not-relevant targets will not be possible (102). In this study, a similar approach was used for AMR prediction, where a knowledge-based parameter was used to only report 'clinically relevant' genes in the CARD database (used by ARMA). This eliminated reporting of irrelevant resistance genes and simplified the analysis. For microbial identification however, this was not possible as curating/adapting the RefSeq database was not possible in this study.

It should also be noted that the O'Grady lab is focused on the development and evaluation of wet-lab method development rather than data analysis, hence in this study we did not develop/tested alternative pipelines. We are unsure whether ONT validated the FASTQ Antimicrobial Resistance Pipeline appropriately before release by, for example, using specimens spiked with known organisms or simulated data, to test classification accuracy (267). We validated the pipeline by testing clinical samples with known pathogen and AMR profiles. While the EPI2ME agent performed well in this study, it is very difficult to accurately validate its performance for either pathogen or AMR gene identification in these types of samples due to their complexity. For example, AMR genes in a metagenomic sample can come from any of the bacteria in the sample (pathogen or commensal) whereas the AMR profile provided by the lab is only for the isolated pathogen. Similarly, culture reports pathogens and not commensals, so it is hard to know whether all bacteria identified by the pipeline were truly present in the sample. However, our analysis of additional detections (pathogens reported by metagenomics and not culture) by PCR in the pilot study, showed that all additional detections were truly present in the sample with two exceptions, a *K. oxytoca* reported by culture had *K. oxytoca* and *K. pneumoniae* incorrectly reported by metagenomics (most likely bioinformatics misclassification of reads) and an *E. coli* reported in a mixed-infection sample (likely a laboratory/kit contaminant) – discussed later in this section.

Thresholds for CMg studies or NGS studies are valuable as they help identify false-positive results and increases confidence of reporting accurate results. CMg tests always apply thresholds to their bioinformatics pipeline to remove low-quality reads, barcode cross-talk, reagent and kit contaminants and misalignments occurring from metagenomic classifiers. For example, Miller *et al.,* applied thresholds to their bioinformatics pipeline when validating their CMg test. Firstly, only reads with a high stringency (203) were analyzed and any identified pathogen would only be reported as 'detected' when the RPM-r (reads per million (RPM) ratio) was ≥10 (where the RPM-r was calculated by the reads corresponding to the pathogen in the clinical sample divided by the reads in the negative control (136)). Thresholds are also often applied during routine culture i.e. the sputum sample is diluted with water prior to plating (described in 2.3.1) such that only bacteria present in high concentrations (>10^5/ml) grow – this step is incorporated in some labs to reduce false positives as concentrations of pathogen below this threshold are considered clinically irrelevant. In a similar manner, we also use thresholds to eliminate any reads with a low-quality alignment score (>19 qscore) and only pathogens reads above the chosen threshold (≥1% microbial classified reads) were considered significant for the infection and were reported.

4.2 Evaluation of the pilot and optimised Clinical metagenomics pipeline

4.2.1 Testing of the pilot pipeline

Initial testing of the pilot version of pipeline was carried out using excess respiratory samples collected from the NNUH clinical microbiology lab from community and hospital patients (described in 3.1.4). Forty samples were collected and used to evaluate the performance of the pipeline. Numbers were limited to this number as we considered this was sufficient to determine the initial performance of the pipeline before any required optimisation. The pipeline was 91.2% sensitive and 100% specific. Only six culture-negative samples were tested, which, in hindsight, should have been increased to get a more accurate representation of the specificity. Additionally, the primary sample type used was sputum and the pipeline should have been tested on more BALs, as this sample type is ‘cleaner’ (fewer upper respiratory tract commensals) and it has a lower microbial load. However, collecting BALs was challenging, as sputum is the primary sample collected from patients with a suspected LRTI in the community and at NNUH and BALs are typically only collected from some ventilated patients (268).

The pilot version of the pipeline had a turnaround time of ~8hrs and was reduced to ~6hrs after optimisation (described in 3.1.5). This rapid turnaround is currently superior to the turnaround time reported for other published CMg pipelines which are able to report results within 12-48 hrs or longer (128, 136, 269, 270). As previously discussed, rapid turnaround time is extremely important for patient outcomes and antibiotic stewardship. Ideally, results should be available before antibiotics are prescribed and administered (<1 hour turnaround). However, as it stands now our pipeline can only be used to guide treatment decisions after they have received one dose of empiric therapy, typically 8 hours after first treatment (62). Although there is potential to further reduce turnaround time for this pipeline, this was not further pursued due to time limitations (discussed further in future work).

4.2.2 Limit of Detection

The LoD of the pipeline is sample-dependent as the commensal and pathogen loads are variable in each sample and the ratio of pathogen:commensal DNA is directly related to the LoD of the test. Variability in the efficiency of the host depletion step can also impact on LoD. To test LoD, we used one sputum with a ‘high’ amount commensal background and one with a ‘low’ amount as representatives of the variability we observe in respiratory samples (described in 3.1.6). The analytical LoD of the streamline version of the CMg pipeline was determined to be at 10^3-10^5 cfu/ml. The determined LoD is similar to the LoD applied for culture in NNUH (described in 2.3.1) (271). However, the clinical microbiology lab has different guidelines for samples for ICU patients or for patients with risk factors (such as immunocompromised) - samples from these patients do not get diluted which makes culture more sensitive (estimated LoD of undiluted culture is10^2 cfu/ml – i.e. one colony on a plate streaked using a 10µl loopful of sample). Also, clinical microbiology labs from different NHS trusts have different guidelines. For example, the clinical lab at St Thomas’ does not dilute respiratory samples prior to plating. This variability in clinical microbiology lab testing makes measuring performance of any new diagnostic tests (not just CMg) against culture very challenging.

The estimated LoD of our pipeline is in a similar range to the LoD reported by Zelenin *et al.,* (10^4 cfu/ml in whole blood) but the pipeline used by Anscombe *et al.,* was more sensitive (10 cfu/ml in whole blood). Other pipelines used for the diagnosis of infections of the sterile sites have also reported to have lower detection limits. For example, Miller *et al.*, reported an LoD of 8-10 cfu/ml for bacterial pathogens and 14-313 copies/ml for DNA and RNA viruses respectively in CSF (136). Blauwkamp *et al.*, also reported an LoD of 39-103 molecules of cfDNA µl/plasma. Greninger *et al.*, reported an LoD of 10^5 copies/ml for RNA viruses in whole blood samples (144). These studies use different approaches during library preparation (e.g. a more sensitive WGA (178)) or deep sequencing which is coupled with slow turnaround time (136, 269). Also, LoD of CMg tests in sterile site samples vs non-sterile site specimens are not comparable as commensals DNA competes with pathogen DNA for sequencing reads.

A limitation of our current pipeline for respiratory infections is that it cannot be used for the diagnosis of viral infections. This is particularly important for CAP where a large proportion (>75% paediatric vs 25% adults cases) of disease is caused by viral pathogens rather than HAP or VAP where the pathogens are typically bacterial and fungal. A modification that would enable of our pipeline to be used for diagnosis of viral infections, would be the addition of SISPA (sequence-independent single-primer-amplification) for the amplification of unknown viral genomes (272, 273). During this process random *k*-mers tagged with a known sequence are used as primers for the PCR-based amplification of viral RNA. Recently, this approach has been used to diagnose SARS-CoV-2 directly from nasopharyngeal swabs within 8 hrs (154). Alternatively, multiple displacement amplification (MDA) WGA (including a reverse transcription step for viral RNA) can be used instead, to amplify microbial RNA/DNA prior to sequencing to improve detection limits. More sensitive WGA would also, be required for sterile sample site testing, such as blood stream infections, where pathogens are present in samples at very low levels (1-30 cfu/ml(177)).

4.2.3 Mock community and loss of *S. pneumoniae*

Saponin has previously been reported to lyse some common pathogenic organisms (177, 178), therefore, we sought to investigate whether the saponin-based differential lysis step of the procedure would cause lysis of any common respiratory pathogens in the mock community experiments (described in 3.1.7). Saponin treatment didn't have any negative effects on any of the mock community organism expect for *S. pneumoniae* (5.8-fold loss). Further investigation showed most of the loss occurred during the saponin incubation and from the addition of the HL-SAN buffer (5.5M salt), but other parts of the process were also potentially involved (described in 3.1.8).

The saponin-based step utilises Quillaja saponin and it is known to have lytic effects on cholesterol-containing cell membranes (171, 172). All bacteria have cell walls protecting their inner membrane and *S. pneumoniae* is a Gram-positive bacterium with a thick peptidoglycan wall, therefore saponin should not be capable of lysing this pathogen. A possible explanation for the loss is that when *S. pneumoniae* cells are under stress, the autolysin gene can be expressed, resulting in autolysis (274). Therefore, we hypothesize that the addition of saponin and/or of the HL-SAN buffer may trigger the production of autolysin, causing cell lysis and subsequent degradation of *S. pneumoniae* DNA. Alternatively or additionally, the autolysis may be related to stressed conditions experienced by *S. pneumoniae* when in pure culture or in sputum. In clinical samples, we saw varying loss of *S. pneumoniae* DNA ranging from ΔCq= 1.7-5.84 (see 3.1.8). This would suggest that the loss is not caused by the method itself, but potentially by the conditions the *S. pneumoniae* is stored in and whether this leads to autolysis before host depletion is performed. Hence, the time taken to go from sample collection to host depletion may be very important to preserve *S. pneumoniae* in clinical samples. For the same reasons, *S. pneumoniae* is known to be a fastidious pathogen that is difficult to culture in the clinical microbiology lab. This may no longer be an issue if CMg is implemented in the clinical microbiology lab and fresh samples are tested rather than testing excess samples that are several days old.

Looking at the literature, a similar effect was also reported by Anscombe *et al.* were *S. pneumoniae* loss was observed after human depletion (178). However, Hasan *et al* (167)*.,* demonstrated no loss of *S. pneumoniae* during saponin depletion. The concentration of saponin used in this study was significantly lower (0.025% final concentration) than in our method and Anscombe's method and may be an important factor.

Street *et al.,* which used our saponin-based host depletion method on *Neisseria gonorrhoeae*-spiked and clinical samples reported similar results (261). The pathogen was reported to be more susceptible to lysis in clinical samples after host depletion, as no lysis was observed in spiked samples. The authors stated that pathogenic cells, may have already been damaged pre-depletion, due to the long storage/transferring times (261). Additionally, a limitation of the mock community used in this study is that we did not test the effect that saponin would have on pathogens with no cell wall such as *Mycoplasma pneumoniae.* This pathogen, is an important pathogen for CAP and would most likely lyse during the saponin process due to its physiology (167). More investigations are necessary to determine the effect of saponin on such organisms.

4.2.4 Evaluation of the optimised pipeline

The performance of the optimised version of the pipeline was tested on a similar number of respiratory samples from community patients as the pilot pipeline so a direct comparison of the two versions was possible (122). The optimised pipeline had higher clinical sensitivity than the pilot pipeline (96.6% vs 91.2%) as only one sample was reported as a false negative versus three false negatives with the pilot pipeline. Increased sensitivity is attributed to the optimisation of the microbial extraction by mechanical lysis (discussed before). Additional PCR analysis confirmed the absence of the missed pathogen – suggesting a false positive result from the clinical lab. We used the undepleted control sample to test for the presence of the 'missed' pathogen in case the host depletion led to the loss of the pathogen in the sample. Possible explanations for the false positive culture result are lab contamination or misidentification of the isolate.

The optimised pipeline was less specific than the pilot pipeline (41.7% vs 100%) but that was expected as a higher number of culture-negative sample was tested. Also, our optimised CMg pipeline is more sensitive than respiratory culture due to the sample dilution step applied by the clinical lab (discussed earlier). We investigated additional pathogen findings with confirmatory qPCR and gene specific analysis (described in 3.1.10) to determine whether these findings were 'real'. Indeed, analysis confirmed additional findings in 10/16 samples. Samples where CMg detection of additional pathogens could not be confirmed contained mostly pathobionts *H. influenzae* or *S. pneumoniae*. High false-positivity rates for pathobionts has also being reported in other similar studies (275). Also, as previously mentioned, in one *K. oxytoca*-positive sample, CMg also detected *K. oxytoca* and *K. pneumoniae* – this was a clear example of bioinformatic misclassification of *K. pneumoniae* reads.

The majority of metagenomics classifiers use a *k*-mer-based classification approach (186), which makes accurate calling of closely-related species from sequencing data challenging. There is also the problem of sequence databases being dominated by the pathogenic species in genera such as *Streptococcus* and *Haemophilus* as discussed earlier. This highlights the need for improving/developing bioinformatic tools and sequence databases that accurately call microbes to the species level directly from metagenomic data (275). Perhaps, when pathobionts are identified, their presence should be investigated in the context of the microbial community. For example, if *S. pneumoniae* is reported in a sample where the microbial community is dominated by other non-pathogenic *Streptococci spp.* then *S. pneumoniae* detection may be due to misclassification or may be present, but more likely be commensal rather than pathogenic. Additional analysis should be carried out (e.g. species-specific gene analysis) or more stringent thresholds should be considered (e.g. increasing alignment scores) in such cases.

A diagnostics pipeline needs to report identified pathogens but could also be used for the detection of clinically relevant resistance genes in order to have the biggest impact on patient management. Our CMg pipeline was able to detect resistance genes in some samples that were concordant with clinical microbiology (e.g. *mecA* gene was identified in two MRSA samples) – see 3.1.11. However, this analysis highlighted how challenging it is to confidently report resistant genes from metagenomic data. Resistance genes were reported in samples with susceptible organisms identified or in culture-negative samples. Detected genes were most likely originated from commensal bacteria. A limitation of the EPI2ME analysis pipeline is that the host of the resistance gene cannot be determined. For chromosomal resistance genes, the flanking regions of the long nanopore reads (3kb average in our pipeline) beyond the resistance gene can be used to identify the origin of the gene. This approach was recently demonstrated by

Leggett *et al.,* by using NanoOK-RT, the authors were able to identify a genes' host in gut microbiome samples (123). A caveat of this approach is that it cannot be applied to plasmid-borne resistance genes. We did not seek mutational resistance in this study as that is even more complex in metagenomic data than looking for acquired genes. A recent study by Sanderson *et al.,* has demonstrated that this is possible for *Neisseria gonorrhoeae* in urine using nanopore data (276), however, urine doesn't typically contain commensals (it is much less complex than sputum) and their approach was designed for a single pathogen. Additionally, the new ONT pore chemistry, R10, reduces the single-read error rate of nanopore sequencing down to as low as <1% (119) and this will help improve SNP calling from nanopore metagenomic data.

An alternative approach that can be applied for both acquired and mutational resistance would be to identify the lineage of the pathogen in question (277). In fact, we collaborated with Brinda *et al.,* for the development and evaluation of RASE, a tool that can predict pathogen resistance/susceptibility by identifying the lineage of the pathogen's closest relatives (277). Brinda *et al.,* was able to accurately predict resistance from 5/6 of our *S. pneumoniae* positive metagenomic samples within minutes (277). A limitation of this approach is that it has only been optimised for 2 pathogens (*N. gonorrhoeae* and *S. pneumoniae*) and it is unlikely to work well for certain pathogens i.e. those where the correlation between lineage and AMR isn't strong and good local pathogen genome databases are required for good accuracy (e.g resistance in *P. aeruginosa*). Another approach would be to use a tool based on a knowledge-based algorithm i.e. an analysis tool that only lists the resistance genes that can be found in the pathogen/s that has been identified by metagenomic sequencing. Such a tool tool would be able to exclude irrelevant resistances (typically from commensals) and provide a summary of relevant resistances based on the pathogen identified (202). The O'Grady lab has recently developed such a tool in collaboration with Dr. Andrew Page at Quadram Institute Biosciences (QIB).

The pipeline presented in this study was one of the first to demonstrate a feasible, rapid, cost-effective clinical metagenomics pipeline that could be translated into the clinical microbiology laboratory. Our pipeline is superior to other CMg pipelines primarily due to its rapid turnaround time and low cost, which are related to efficient host depletion. For example, Votintseva *et al.*, report a turnaround time of 7.5 hrs but reproducibility of the fast version of this pipeline was poor (128). Thoendel *et al.*, used CMg for the diagnosis of PJI, but reported bacterial loss when using the MolYsis kit for host depletion (166). In a more recent study, Miller *et al.*, developed and validated a CMg pipeline for the diagnosis of meningitis but is has a slow turnaround as no host depletion is used (136). Langeliel *et al.*,(275) also developed a diagnostics pipeline for the diagnosis of LRTIs, which included metagenomic sequencing of respiratory samples coupled with a novel bioinformatics approach that could separate infectious from non-infectious respiratory illnesses and differentiate pathogens from respiratory commensals. Although good performance of this test was reported (receiver-operating curve (AUC) of 0.80-0.96% for their three bioinformatics models), turnaround time was not reported for their protocol (275). Blauwkamp *et al.*, also developed and validated a CMg pipeline, called the Karius test, that uses cell-free DNA in plasma samples to identify pathogens (269). Although this pipeline is very comprehensive, it has a slow turnaround time (53 hrs), is expensive and has low specificity (63% for the diagnosis of blood stream infection) as it detects cfDNA from any microbe in the body including commensals, gut microbes and pathogens causing unrelated infections (e.g. sore throat) (270). Although, in the pilot study we originally reported low specificity using our pipeline compared to culture, PCR analysis demonstrated additional detections were real (122), raising our specificity to 100%.

As previously mentioned, a limitation of our pipeline is that it was not applied for the detection of viral pathogens. There is potential to modify it for viral diagnostics. Currently, any viruses are lost as centrifugation is used to pellet bacterial cells at several points in the procedure and the supernatant containing viruses is discarded. A second arm of the pipeline could be introduced, during which a second aliquot of the sample (or the supernatant after the first centrifugation step) would processed without centrifugation (122), followed by nuclease treatment of the sample to remove cell free human nucleic acid, viral nucleic acid extraction, cDNA synthesis and pooling with the bacterial DNA before sequencing. This approach is currently being tested and optimized in the O’Grady lab.

4.3 Implementation in the INHALE trial

4.3.1 NNUH INHALE sample testing

The optimised CMg pipeline was implemented in the INHALE trial as a third molecular-based test (alongside 2 PCR based pneumonia panels, Filmarray and Unyvero) for the diagnosis of HAP and VAP in ICU respiratory samples (described in 3.2). Initial analysis of the INHALE samples revealed an increased number of sequencing ‘failures’ and we had to apply additional parameters to remove these samples. The majority of the ‘failures’ were culture-negative samples (NBG/NSG/NG) which produced very few reads. We hadn’t come across such samples during the development of the CMg pipeline as we hadn’t been focussed on the ICU, where no growth samples appear to be more common. These samples revealed the need for a process control to monitor performance of the pipeline. A suitable process control would be a non-pathogenic difficult-to-lyse (probably Gram-positive) bacterium that is never found in the respiratory tract and can be spiked into the clinical sample at a concentration that should always produce sequencing reads in the absence of any respiratory bacteria (but also at a concentration that wouldn’t outcompete low levels of pathogen in a sample) (102). The process control would ensure all steps of the pipeline were successful and no inhibition or microbial loss occurred (267). Without this, it is impossible to tell the difference between the pipeline failing and ‘no growth’ samples. Time constraints meant that I couldn’t develop a process control during my PhD but the O’Grady lab is now working on the development of this control for inclusion in their CMg pipelines.

Miller *et al.*, validated a synthetic CSF matrix positive control for their CMg pipeline which consisted of seven representative pathogens (including bacteria, fungi, DNA and RNA viruses) in known quantities (136). This approach, however, doesn't monitor for individual sample failures, only issues that result in the failure of the entire run. This type of a positive control would be hard to design and validate for respiratory metagenomics, as a synthetic sample mimicking the composition of a respiratory sample is hard to design (consistency and microbial+human load is very variable in respiratory samples).

The sensitivity of the pipeline was decreased when tested on the INHALE sample set compared to the pilot study samples (80.77% versus 96.6%). As previously discussed, this was probably due to the discrepancy between the thresholds applied in ICU respiratory samples by the clinical laboratory and community samples (no dilution of samples prior to culture, making culture more sensitive). Also, in the INHALE sample set, more BALs were processed than the pilot study, which typically contain a lower microbial load (sample diluted in large volumes of saline). Increasing the library prep PCR cycle number to 30 or 35 may be necessary to increase the sensitivity of the pipeline so that BAL samples can be reliably tested.

The majority of missed pathogens in the INHALE sample set (4/5 false negative samples) were reported as bacterial mixed infections by culture. This was probably due to the difference in quantities of the pathogens present in the sample and possibly only a few colonies of the second pathogen were reported by culture, at a concentration below the limit of detection of the pipeline. During the library preparation PCR reaction, the organism present at higher concentration dominates the PCR reaction, leading to poor amplification of the pathogen/s at lower concentration. In fact, in 1/4 cases, sequence reads were detected for the 'missed' pathogen (*S. aureus* in YS053), but they were below the chosen thresholds for pathogen detection. The difference in the concentration of the ‘missed’ pathogens was also reflected by the semi-quantification of the two PCR tests (BIOFIRE and Curetis), as lower quantities of all ‘missed’ pathogens in the four false negative samples were reported.

Additional analysis of the INHALE CMg false positives by the two PCR tests revealed that 6/15 additional pathogens detected were due to contamination arising from common contaminants. Contamination is a major challenge for metagenomic and microbiome studies especially when low-biomass samples are sequenced (278). A study investigating contamination in sequencing and extraction kits reported that contaminant organisms are ubiquitous and were always present in PCR reagents, library preparation kits, water and other reagents (278). Also contamination composition varied amongst the different batches of same kits (278).

We always included a negative process control and thresholds and even then, contamination could not be removed from the INHALE sample set. This highlights the importance of having additional parameter/s dealing with contamination, especially for ICU or sterile samples (where <1000 microbial reads were reported with metagenomic sequencing). A more stringent negative control rule (than the one applied in my analysis) would help remove reads from barcode cross-talk and from real contaminants. For example, common contaminants of the skin microbiome and reagents should be defined at the beginning of the study and not reported when identified (136). However, it isn’t possible to rule out all common contaminants as some e.g. *E. coli,* are important pathogens and in different contexts, e.g. *S. epidermidis* a skin flora bacterium is an important PJI pathogen. For these organisms different parameters should be used.

Also, as it stands now the depletion part of the pipeline consists of various steps which increases the likelihood of introducing contamination, especially in the hands of less-experienced handlers. Sample processing of INHALE samples was performed by different operators which is a likely reason why there was an increase in contamination levels in comparison to the previous samples sets. In fact, it was clear from the data that more experienced (and meticulous) operators produced fewer failed and contaminated datasets. This lack of method robustness is clearly a weakness and further simplification of the pipeline is required to aid implementation (discussed further in future work).

4.3.2 Comparison of metagenomics against BIOFIRE

Comparison of the findings of the CMg pipeline against BIOFIRE (as it was the test progressing to the randomised controlled trial) revealed that CMg was less concordant with BIOFIRE and more concordant with culture (described in 3.2.4). The majority of discordant and partial concordant results were due to BIOFIRE identifying more pathogens than CMg. Our findings are not a surprise, as PCR-based tests are more sensitive than sequencing-based tests (8, 279). However, this raises the question of which organisms are clinically relevant and should be treated? PCR results should not be interpreted without considering other laboratory findings or clinical information. If treatment is only guided based on the PCR output, it may lead to over-diagnosis and over-treatment of patients.

These findings, highlight an advantage CMg has over PCR-based tests, which is its comprehensiveness – CMg would not only detect pathogens but also commensals, meaning the presence of the pathogen can be interpreted in the context of the rest of the microbial community. For example, in our experience a pathogen in a true-positive sample would be dominating the microbial community and will be listed as the most abundant organism in the WIMP report. Conversely, in an NRF sample the top hit is a commensal/s. Reports of metagenomic data, resemble on how culture plates looks – a mix of non-pathogenic bacteria for a negative sample or heavy growth of a pathogen with a few commensal colonies for an infected sample. CMg, however, is more comprehensive than culture, as anaerobes, fastidious organisms, bacteria and fungi are all reported in one test.

Additionally, as CMg does not rely on a pre-defined panel will also detect rare or even novel pathogens. This is particularly topical in relation to the current SARS-CoV-2 pandemic – if a comprehensive bacterial/viral CMg pipeline was in routine use in Wuhan in 2019, the novel virus would have been identified almost immediately after moving to humans and the outbreak may have been stopped before spreading globally. The diagnostic PCRs in the clinical virology labs couldn’t detect the new virus and metagenomics had to be employed to identify the cause of the outbreak (280).

4.4 Characterisation of *Legionella* spp. using clinical metagenomics

Previous CMg studies focused on pathogen identification only as insufficient genome sequence recovered could not reliably detect AMR genes or study the pathogen in more detail (136, 166). Our CMg pipeline can generate whole pathogen genomes due to efficient host depletion using saponin, which can be used to further study the pathogen/s (described in 3.1.13). This capability means that our pipeline should be suitable for not only diagnostics but for public health applications. Public health microbiologists at PHE approached us with an interest in applying CMg to *Legionella* outbreak investigation. Rapid detection and simultaneous genotyping of *Legionella* in suspected outbreak samples would have a major impact in the field. Therefore, we applied our CMg pipeline for the diagnosis and molecular typing of *Legionella pneumophila* directly from respiratory samples (described in 3.3). The sensitivity of the pipeline these samples was lower compared to the proof-of-concept study (85.71% vs 96.6%) even after using less stringent thresholds. This discrepancy was potentially caused by: i) low microbial load of *L. pneumophila* observed in this sample set pre-depletion and ii) the condition of the samples (described in 3.3.1). The samples used in this sample set were frozen samples (some having gone through multiple freeze-thaw cycles) collected over 5 years ago. We previously demonstrated that frozen samples are not ideal for our pipeline (bacteria either lyse or are damaged and possibly lost during the host depletion step)- see 3.2.1. qPCR pre- and post-depletion demonstrated *L. pneumophila* DNA loss in the host depletion step (described in 3.3.1). Hence, we recommend our CMg pipeline only be used only on fresh samples, as this would be the case if implemented. Frozen samples are mainly used in research – clinical samples are never frozen before testing. If samples need to be frozen, they should be frozen with a stability agent so as not to damage microbes upon thawing (281).

Contamination in this sample set was also high and this was due to using a sequencing positive control at a too high concentration. This was extracted *L. pneumophila* DNA which was included as an external positive control at the beginning of the library preparation. The concentration for the control was high, producing many reads, leading to barcode-cross talk and potential cross-contamination. As previously discussed, a process control should have been used – however, as this is not straight forward to develop, the positive control should have at least been something not ever found in the respiratory tract and been used at low concentrations. This study is still on-going and the data and analysis presented in this thesis is preliminary. Therefore, in the future we plan to test fresh samples but also improve the positive control.

Detecting the sequence type (ST) of the *L. pneumophila* positive-samples (described in 3.3.2) was not possible using our CMg pipeline due to low genome coverage. Different strategies were attempted using different inputs: *de novo* assemblies, reference-based assemblies and basecalled FASTQ reads, but none of these approaches were able to identify an ST. The O'Grady lab plans to further investigate if molecular typing of *L. pneumophila* is possible directly from respiratory samples using CMg by processing fresh samples and by applying appropriate thresholds and controls.

4.5 Conclusion

Although application of the pipeline in INHALE and for *Legionella* typing was challenging, we have learned a great deal and know the reasons for reduction of performance compared to the pilot studies (changes in culture processing of ICU samples in INHALE and old frozen samples in the *Legionella* study). We have also learned that our pipeline is too laborious for clinical implementation further work is required to address this (see below). However, the pipeline and data presented in this study have demonstrated that clinical metagenomics has the potential to revolutionise clinical microbiology and replace current tests for the diagnosis of bacterial LRTIs. The CMg pipeline allowed identification of bacterial pathogens and resistance gene detection in ~6hrs. Rapid accurate diagnostics will improve antibiotic stewardship and patient management, which will lead to improved patient outcomes, reduced hospital costs and slow the emergence of AMR.

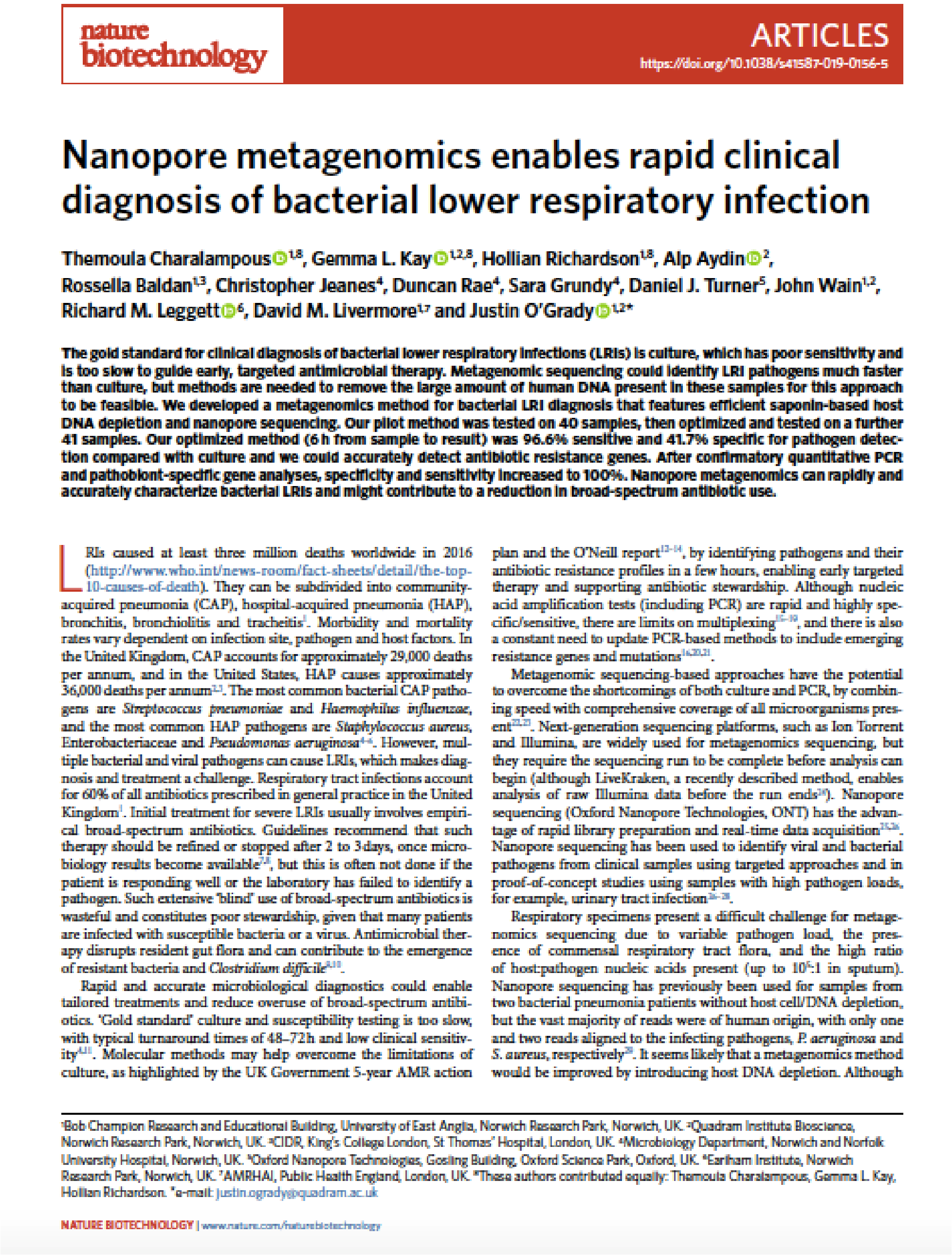
4.6 Future work

- Develop and evaluate a positive process control, suitable for bacteria and DNA and RNA viruses.
- Develop a viral metagenomics arm suitable for the diagnosis of CAP.
- Improve the LoD of the method to make it more reliable for the diagnosis of ICU samples. This can be achieved by improving the sensitivity of the library preparation PCR.
- Further simplify the depletion and shorten the pipeline to make it more robust and implementable. This would likely involve some automation of the process e.g. using a liquid-handler for depletion step and bead-based washing steps.
- Develop and clinically validate a bioinformatics pipeline that will allow an accurate hypothesis-free interpretation of the results.

5. Appendix

PAPER I

Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J, Leggett RM. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nature Biotechnology. 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].

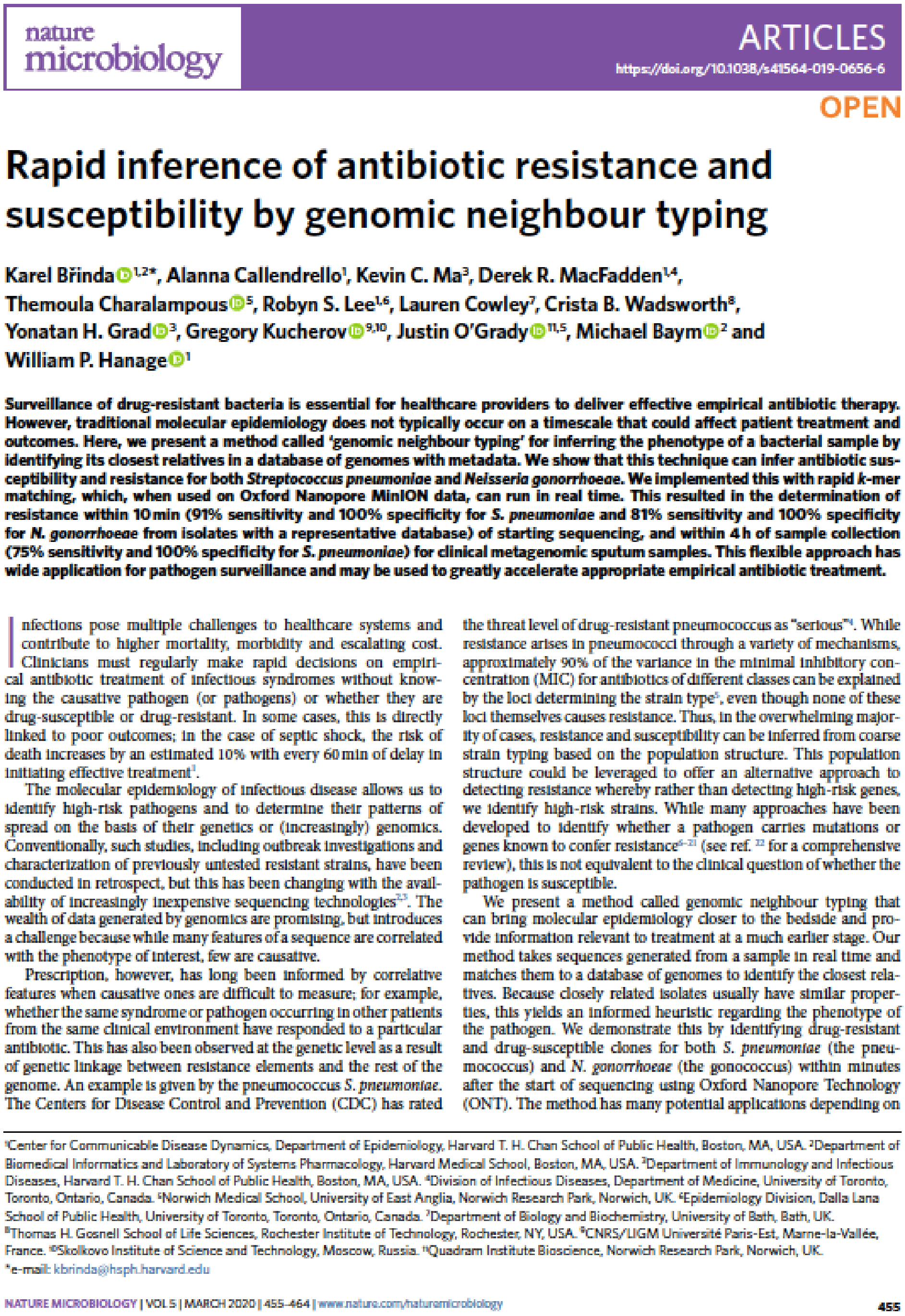
Metagenomic sequencing-based approaches have the potential to overcome the shortcomings of both culture and PCR, by combining speed with comprehensive coverage of all microorganisms present[22,23]. Next-generation sequencing platforms, such as Ion Torrent and Illumina, are widely used for metagenomics 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

PAPER II

Břinda K, Callendrello A, Ma KC, MacFadden DR, Charalampous T, Lee RS, Cowley L, Wadsworth CB, Grad YH, Kucherov G, O'Grady J. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. Nature microbiology. 2020 Mar;5(3):455-64.



nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-019-0656-6

OPEN

Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing

Karel Břinda[1,2,*], Alanna Callendrello[1], Kevin C. Ma[3], Derek R. MacFadden[1,4], Themoula Charalampous[5], Robyn S. Lee[1,6], Lauren Cowley[7], Crista B. Wadsworth[8], Yonatan H. Grad[3], Gregory Kucherov[9,10], Justin O'Grady[11,5], Michael Baym[2] and William P. Hanage[1]

Surveillance of drug-resistant bacteria is essential for healthcare providers to deliver effective empirical antibiotic therapy. However, traditional molecular epidemiology does not typically occur on a timescale that could affect patient treatment and outcomes. Here, we present a method called 'genomic neighbour typing' for inferring the phenotype of a bacterial sample by identifying its closest relatives in a database of genomes with metadata. We show that this technique can infer antibiotic susceptibility and resistance for both *Streptococcus pneumoniae* and *Neisseria gonorrhoeae*. We implemented this with rapid *k*-mer matching, which, when used on Oxford Nanopore MinION data, can run in real time. This resulted in the determination of resistance within 10 min (91% sensitivity and 100% specificity for *S. pneumoniae* and 81% sensitivity and 100% specificity for *N. gonorrhoeae* from isolates with a representative database) of starting sequencing, and within 4 h of sample collection (75% sensitivity and 100% specificity for *S. pneumoniae*) for clinical metagenomic sputum samples. This flexible approach has wide application for pathogen surveillance and may be used to greatly accelerate appropriate empirical antibiotic treatment.

Infections pose multiple challenges to healthcare systems and contribute to higher mortality, morbidity and escalating cost. Clinicians must regularly make rapid decisions on empirical antibiotic treatment of infectious syndromes without knowing the causative pathogen (or pathogens) or whether they are drug-susceptible or drug-resistant. In some cases, this is directly linked to poor outcomes; in the case of septic shock, the risk of death increases by an estimated 10% with every 60 min of delay in initiating effective treatment[1].

The molecular epidemiology of infectious disease allows us to identify high-risk pathogens and to determine their patterns of spread on the basis of their genetics or (increasingly) genomics. Conventionally, such studies, including outbreak investigations and characterization of previously untested resistant strains, have been conducted in retrospect, but this has been changing with the availability of increasingly inexpensive sequencing technologies[2,3]. The wealth of data generated by genomics are promising, but introduces a challenge because while many features of a sequence are correlated with the phenotype of interest, few are causative.

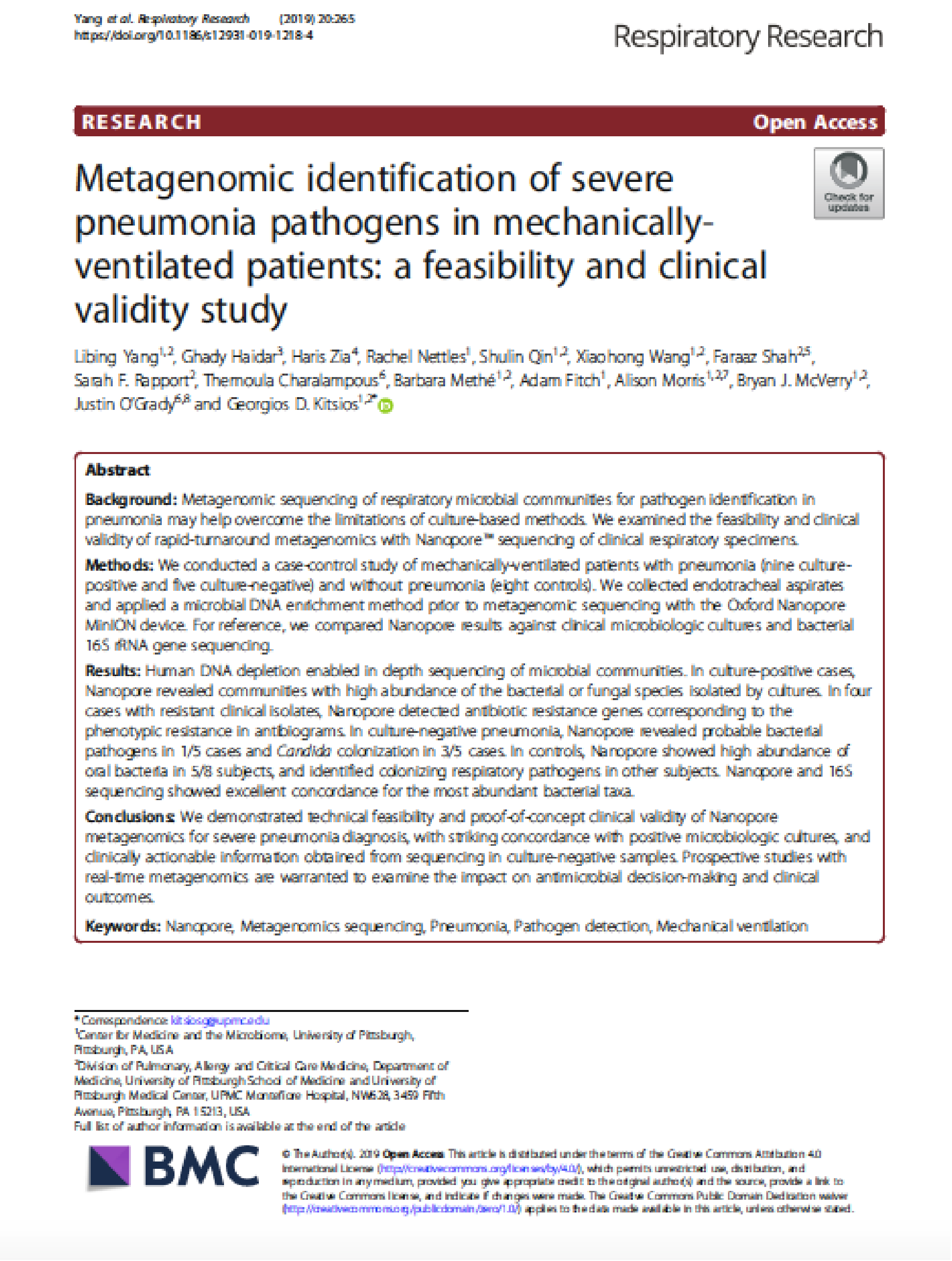
Prescription, however, has long been informed by correlative features when causative ones are difficult to measure; for example, whether the same syndrome or pathogen occurring in other patients from the same clinical environment have responded to a particular antibiotic. This has also been observed at the genetic level as a result of genetic linkage between resistance elements and the rest of the genome. An example is given by the pneumococcus *S. pneumoniae*. The Centers for Disease Control and Prevention (CDC) has rated the threat level of drug-resistant pneumococcus as "serious"[4]. While resistance arises in pneumococci through a variety of mechanisms, approximately 90% of the variance in the minimal inhibitory concentration (MIC) for antibiotics of different classes can be explained by the loci determining the strain type[5], even though none of these loci themselves causes resistance. Thus, in the overwhelming majority of cases, resistance and susceptibility can be inferred from coarse strain typing based on the population structure. This population structure could be leveraged to offer an alternative approach to detecting resistance whereby rather than detecting high-risk genes, we identify high-risk strains. While many approaches have been developed to identify whether a pathogen carries mutations or genes known to confer resistance[6–21] (see ref. [22] for a comprehensive review), this is not equivalent to the clinical question of whether the pathogen is susceptible.

We present a method called genomic neighbour typing that can bring molecular epidemiology closer to the bedside and provide information relevant to treatment at a much earlier stage. Our method takes sequences generated from a sample in real time and matches them to a database of genomes to identify the closest relatives. Because closely related isolates usually have similar properties, this yields an informed heuristic regarding the phenotype of the pathogen. We demonstrate this by identifying drug-resistant and drug-susceptible clones for both *S. pneumoniae* (the pneumococcus) and *N. gonorrhoeae* (the gonococcus) within minutes after the start of sequencing using Oxford Nanopore Technology (ONT). The method has many potential applications depending on

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [2]Department of Biomedical Informatics and Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA. [3]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [4]Division of Infectious Diseases, Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [5]Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK. [6]Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [7]Department of Biology and Biochemistry, University of Bath, Bath, UK. [8]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, USA. [9]CNRS/LIGM Université Paris-Est, Marne-la-Vallée, France. [10]Skolkovo Institute of Science and Technology, Moscow, Russia. [11]Quadram Institute Bioscience, Norwich Research Park, Norwich, UK.
*e-mail: kbrinda@hsph.harvard.edu

PAPER III

Yang L, Haidar G, Zia H, Nettles R, Qin S, Wang X, Shah F, Rapport SF, Charalampous T, Methé B, Fitch A. Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. Respiratory research. 2019 Dec 1;20(1):265.



Yang *et al. Respiratory Research* (2019) 20:265
https://doi.org/10.1186/s12931-019-1218-4

Respiratory Research

RESEARCH Open Access

Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study

Libing Yang[1,2], Ghady Haidar[3], Haris Zia[4], Rachel Nettles[1], Shulin Qin[1,2], Xiaohong Wang[1,2], Faraaz Shah[2,5], Sarah F. Rapport[2], Themoula Charalampous[6], Barbara Methé[1,2], Adam Fitch[1], Alison Morris[1,2,7], Bryan J. McVerry[1,2], Justin O'Grady[6,8] and Georgios D. Kitsios[1,2*]

Abstract

Background: Metagenomic sequencing of respiratory microbial communities for pathogen identification in pneumonia may help overcome the limitations of culture-based methods. We examined the feasibility and clinical validity of rapid-turnaround metagenomics with Nanopore™ sequencing of clinical respiratory specimens.

Methods: We conducted a case-control study of mechanically-ventilated patients with pneumonia (nine culture-positive and five culture-negative) and without pneumonia (eight controls). We collected endotracheal aspirates and applied a microbial DNA enrichment method prior to metagenomic sequencing with the Oxford Nanopore MinION device. For reference, we compared Nanopore results against clinical microbiologic cultures and bacterial 16S rRNA gene sequencing.

Results: Human DNA depletion enabled in depth sequencing of microbial communities. In culture-positive cases, Nanopore revealed communities with high abundance of the bacterial or fungal species isolated by cultures. In four cases with resistant clinical isolates, Nanopore detected antibiotic resistance genes corresponding to the phenotypic resistance in antibiograms. In culture-negative pneumonia, Nanopore revealed probable bacterial pathogens in 1/5 cases and *Candida* colonization in 3/5 cases. In controls, Nanopore showed high abundance of oral bacteria in 5/8 subjects, and identified colonizing respiratory pathogens in other subjects. Nanopore and 16S sequencing showed excellent concordance for the most abundant bacterial taxa.

Conclusions: We demonstrated technical feasibility and proof-of-concept clinical validity of Nanopore metagenomics for severe pneumonia diagnosis, with striking concordance with positive microbiologic cultures, and clinically actionable information obtained from sequencing in culture-negative samples. Prospective studies with real-time metagenomics are warranted to examine the impact on antimicrobial decision-making and clinical outcomes.

Keywords: Nanopore, Metagenomics sequencing, Pneumonia, Pathogen detection, Mechanical ventilation

* Correspondence: kitsiosg@upmc.edu
[1] Center for Medicine and the Microbiome, University of Pittsburgh, Pittsburgh, PA, USA
[2] Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, University of Pittsburgh School of Medicine and University of Pittsburgh Medical Center, UPMC Montefiore Hospital, NW628, 3459 Fifth Avenue, Pittsburgh, PA 15213, USA
Full list of author information is available at the end of the article

PAPER 1V

Charalampous T, Kay GL, OeGrady J. Applying clinical metagenomics for the detection and characterisation of respiratory infections. The Lung Microbiome (ERS Monograph). Sheffield, European Respiratory Society. 2019 Mar 1:35-49.

Chapter 3

Applying clinical metagenomics for the detection and characterisation of respiratory infections

Themoula Charalampous[1], Gemma L. Kay[2] and Justin O'Grady[1,2]

Metagenomics is, by definition, the direct identification and characterisation of all genomes within a sample. When metagenomics is used to characterise pathogens directly from clinical samples, it is then referred to as clinical metagenomics. Clinical metagenomics has the potential to replace conventional methodologies, such as culture, for the diagnosis of infection due to its unbiased approach and the potential for a rapid turnaround time to results. An efficient metagenomics-based pipeline needs to be comprehensive, rapid and cost-effective, and will include: 1) sample preparation, including pathogen DNA enrichment and/or host DNA depletion strategies, 2) library preparation and sequencing, and 3) rapid data analysis. Each of these steps will be discussed (focusing on bacterial infections), with the aim of producing high-quality data while reducing cost and turnaround time. We review the literature on clinical metagenomics for diagnostic and epidemiological applications, and discuss the challenges in applying clinical metagenomics methodologies.

Cite as: Charalampous T, Kay GL, O'Grady J. Applying clinical metagenomics for the detection and characterisation of respiratory infections. *In:* Cox MJ, Ege MJ, von Mutius E, eds. The Lung Microbiome (ERS Monograph). Sheffield, European Respiratory Society, 2019; pp. 35–49 [https://doi.org/10.1183/2312508X.10015618].

@ERSpublications
Recent technological advances have allowed clinical metagenomics to provide rapid, comprehensive characterisation of respiratory infections, to enable targeted antimicrobial therapy, and to provide data for infection control and public health applications. http://ow.ly/MZSK30mLWop

The gold standard for the diagnosis of microbial infection is culture. However, culture has low sensitivity and a long turnaround time, with results being available from 48 h to several weeks after sample collection (*e.g.* up to 6 weeks for the diagnosis of *Mycobacterium tuberculosis*). Initial antimicrobial treatment is empirical, based on guidelines and local or hospital pathogen epidemiology. Suboptimal diagnostics with slow turnaround times lead to: 1) the overuse of broad-spectrum antibiotics, which promotes the emergence of

[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.

Correspondence: Themoula Charalampous, Bob Champion Research and Educational Building, University of East Anglia, Norwich Research Park, Colney Lane, Norwich NR4 7UQ, UK. E-mail: t.charalampous@uea.ac.uk

Copyright ©ERS 2019. Print ISBN: 978-1-84984-101-6. Online ISBN: 978-1-84984-102-3. Print ISSN: 2312-508X. Online ISSN: 2312-5098.

6. Abbreviations

AMR	Antimicrobial Resistance
AMS	Antimicrobial Stewardship
ARMA	Antimicrobial Resistance Mapping Application
BAL	Bronchoalveolar lavage
BCY-C	Charcoal Yeast Extract with Cefamandole
BCYE	Buffered Charcoal Yeast Extract
BHI	Brain Heart Infusion
BLAST	Basic Local Alignment Search tool
BMPA	Buffered Polymyxin Anisomycin
CAP	Community Acquired Pneumonia
CARD	Comprehensive Antibiotic Resistance Database
CDC	Centre for Disease Control and Prevention
CLED	Cysteine–lactose–electrolyte-deficient
CMg	Clinical Metagenomics
CMV	Cytomegalovirus
COPD	Chronic Obstructive Pulmonary Disease
CRP	C-reactive Protein
CSF	Cerebrospinal Fluid
DBG	De Bruijn Graph
DNTPs	Dideoxy Nucleotides
EMBL	European Molecular Biology Laboratory
ESBL	Extended Spectrum Beta-lactamase
ESCMID	European Society of Clinical Microbiology and Infectious Diseases
ESKAPE	*Escherichia coli, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterococcus faecium*
ESR	Erythrocyte Sedimentation Rate
ETA	Endotracheal Aspirates
FDA	US Food and Drug Administration

GOSH	Great Ormond Street Hospital Children's Charity
HAP	Hospital Acquired Pneumonia
HCAP	Health Care Associated Pneumonia
HPIV3	Parainfluenza 3 Virus
ICU	Intensive Care Unit
INSDC	International Nucleotide Sequence Database Collaboration
LB	Luria Broth
LRTI	Lower Respiratory Tract Infection
MAG	Metagenome-Assembled Genome
MDR	Multi Drug Resistant
MDR-GNEB	Multi Drug Resistant- Gram Negative Enterobacteria
MLST	Multi-Locus-Sequence-Typing
MNPs	Magnetic Nanoparticles
MRSA	Methicillin Resistant *Staphylococcus aureus*
MSSA	Methicillin Sensitive *Staphylococcus aureus*
Mtb	*Mycobacterium tuberculosis*
NPA	Nasopharyngeal aspirates
NCBI	National Center for Biotechnology Information
NFW	Nuclease Free Water
NG	No Growth
NGS	Next Generation Sequencing
NICE	National Institute of Health Excellence
NNUH	Norfolk and Norwich University Hospitals
NRCCE	Nextera Rapid Capture Custom Enrichment
NRF	Normal Respiratory Flora
NSG	No Significant Growth
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
PHE	Public Health England
PI	Pulmonary Infiltrate

PJI	Prosthetic Joint Infection
PSB	Protected Specimen Brush
QC	Quality Control
RAD	Rapid Adapter
RASE	Resistance-Associated Sequence Elements
RBC	Red Blood Cells
RCT	Randomized Controlled Trial
RHF	Royal Free Hospitals
RPM	Reads Per Million
RSV	Respiratory Syncytial Virus
RT	Room Temperature
RVPBRU	Respiratory and Vaccine Preventable Bacteria Reference Unit
SAD	Sabouraud Dextrose
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SBT	Sequence Based Typing
sCAP	Severe Community Acquired Pneumonia
SNP	Single-Nucleotide Polymorphisms
ST	Sequence Type
STEC	Shiga-Toxin producing *E. coli*
TB	Tuberculosis
TSB	Tryptic Soy Broth
UCLH	University College London Hospitals
UEA	University of East Anglia
UHNM	University Hospitals of North Midlands
URT	Upper Respiratory Tract
UTIs	Urinary Tract Infections
VAP	Ventilator Associated Pneumonia
WGA	Whole Genome Amplification
WGS	Whole Genome Sequencing
WHO	World Health Organisation
WIMP	What’s In My Pot

7. References

1. ERS. Acute Lower Respiratory Infections. ERS white book. 2018.

2. Feldman C, Shaddock E. Epidemiology of Lower Respiratory Tract Infections in Adults. Expert Rev Respir Med. 2019;13(1):63-77.

3. Greene G, Hood K, Little P, Verheij T, Goossens H, Coenen S, et al. Towards Clinical Definitions of Lower Respiratory Tract Infection (Lrti) for Research and Primary Care Practice in Europe: An International Consensus Study. Prim Care Respir J. 2011;20(3):299-306.

4. Mandell LA, Read RC. Chapter 45 Infections of the Lower Respiratory Tract. Antibiotic and Chemotherapy2010. p. 574-88.

5. Woodhead M, Blasi F, Ewig S, Garau J, Huchon G, Ieven M, et al. Guidelines for the Management of Adult Lower Respiratory Tract Infections - Full Version. Clinical Microbiology and Infection. 2011;17:E1-E59.

6. Carroll KC. Laboratory Diagnosis of Lower Respiratory Tract Infections: Controversy and Conundrums. Journal of Clinical Microbiology. 2002;40(9):3115-20.

7. Muscedere J, Dodek P, Keenan S, Fowler R, Cook D, Heyland D. Comprehensive Evidence-Based Clinical Practice Guidelines for Ventilator-Associated Pneumonia: Prevention. J Crit Care. 2008;23(1):126-37.

8. Hayon JAN, Figliolini C, Combes A, Trouillet J-L, Kassis N, Dombret MC, et al. Role of Serial Routine Microbiologic Culture Results in the Initial Management of Ventilator-Associated Pneumonia. American Journal of Respiratory and Critical Care Medicine. 2002;165(1):41-6.

9. Cunha BA. The Atypical Pneumonias: Clinical Diagnosis and Importance. Clin Microbiol Infect. 2006;12 Suppl 3:12-24.

10. Enne VI, Personne Y, Grgic L, Gant V, Zumla A. Aetiology of Hospital-Acquired Pneumonia and Trends in Antimicrobial Resistance. Current Opinion in Pulmonary Medicine. 2014;20(3):252-8.

11. Ieven M, Coenen S, Loens K, Lammens C, Coenjaerts F, Vanderstraeten A, et al. Aetiology of Lower Respiratory Tract Infection in Adults in Primary Care: A Prospective Study in 11 European Countries. Clin Microbiol Infect. 2018;24(11):1158-63.

12. ERS. Acute Lower Respiratory Infections. The Burden of Lung Disease. European Lung White Book: The European Respiratory Society; 2019.

13. Estimates of the Global, Regional, and National Morbidity, Mortality, and Aetiologies of Lower Respiratory Tract Infections in 195 Countries: A Systematic Analysis for the Global Burden of Disease Study 2015. Lancet Infect Dis. 2017;17(11):1133-61.

14. Quan TP, Fawcett NJ, Wrightson JM, Finney J, Wyllie D, Jeffery K, et al. Increasing Burden of Community-Acquired Pneumonia Leading to Hospitalisation, 1998–2014. Thorax. 2016;71(6):535.

15. Bertsias A, Tsiligianni IG, Duijker G, Siafakas N, Lionis C. Studying the Burden of Community-Acquired Pneumonia in Adults Aged >50 Years in Primary Health Care: An Observational Study in Rural Crete, Greece. NPJ Prim Care Respir Med. 2014;24:14017.

16. Woodhead M, Blasi F, Ewig S, Huchon G, Leven M, Ortqvist A, et al. Guidelines for the Management of Adult Lower Respiratory Tract Infections. European Respiratory Journal. 2005;26(6):1138-80.

17. Masterton RG, Galloway A, French G, Street M, Armstrong J, Brown E, et al. Guidelines for the Management of Hospital-Acquired Pneumonia in the Uk: Report of the Working Party on

Hospital-Acquired Pneumonia of the British Society for Antimicrobial Chemotherapy. Journal of Antimicrobial Chemotherapy. 2008;62(1):5-34.

18. Giard M, Lepape A, Allaouchiche B, Guerin C, Lehot JJ, Robert MO, et al. Early- and Late-Onset Ventilator-Associated Pneumonia Acquired in the Intensive Care Unit: Comparison of Risk Factors. J Crit Care. 2008;23(1):27-33.

19. Bassetti M, Taramasso L, Giacobbe DR, Pelosi P. Management of Ventilator-Associated Pneumonia: Epidemiology, Diagnosis and Antimicrobial Therapy. Expert Rev Anti Infect Ther. 2012;10(5):585-96.

20. Sandiumenge A, Rello J. Ventilator-Associated Pneumonia Caused by Eskape Organisms: Cause, Clinical Features, and Management. Curr Opin Pulm Med. 2012;18(3):187-93.

21. Sader HS, Castanheira M, Arends SJR, Goossens H, Flamm RK. Geographical and Temporal Variation in the Frequency and Antimicrobial Susceptibility of Bacteria Isolated from Patients Hospitalized with Bacterial Pneumonia: Results from 20 Years of the Sentry Antimicrobial Surveillance Program (1997-2016). J Antimicrob Chemother. 2019;74(6):1595-606.

22. Jones RN. Microbial Etiologies of Hospital-Acquired Bacterial Pneumonia and Ventilator-Associated Bacterial Pneumonia. Clinical Infectious Diseases. 2010;51(Supplement_1):S81-S7.

23. López-Giraldo A, Sialer S, Esperatti M, Torres A. Viral-Reactivated Pneumonia During Mechanical Ventilation: Is There Need for Antiviral Treatment? Frontiers in Pharmacology. 2011;2:66.

24. Costa C, Bergallo M, Astegiano S, Terlizzi ME, Sidoti F, Solidoro P, et al. Detection of Mimivirus in Bronchoalveolar Lavage of Ventilated and Nonventilated Patients. Intervirology. 2012;55(4):303-5.

25. Vincent A, La Scola B, Forel JM, Pauly V, Raoult D, Papazian L. Clinical Significance of a Positive Serology for Mimivirus in Patients Presenting a Suspicion of Ventilator-Associated Pneumonia. Crit Care Med. 2009;37(1):111-8.

26. Guidelines for the Management of Adults with Hospital-Acquired, Ventilator-Associated, and Healthcare-Associated Pneumonia. American Journal of Respiratory and Critical Care Medicine. 2005;171(4):388-416.

27. Hamet M, Pavon A, Dalle F, Pechinot A, Prin S, Quenot JP, et al. Candida Spp. Airway Colonization Could Promote Antibiotic-Resistant Bacteria Selection in Patients with Suspected Ventilator-Associated Pneumonia. Intensive Care Med. 2012;38(8):1272-9.

28. Meersseman W, Lagrou K, Spriet I, Maertens J, Verbeken E, Peetermans WE, et al. Significance of the Isolation of Candida Species from Airway Samples in Critically Ill Patients: A Prospective, Autopsy Study. Intensive Care Med. 2009;35(9):1526-31.

29. Ferrer M, Liapikou A, Valencia M, Esperatti M, Theessen A, Antonio Martinez J, et al. Validation of the American Thoracic Society-Infectious Diseases Society of America Guidelines for Hospital-Acquired Pneumonia in the Intensive Care Unit. Clin Infect Dis. 2010;50(7):945-52.

30. Blot SI, Taccone FS, Van den Abeele AM, Bulpa P, Meersseman W, Brusselaers N, et al. A Clinical Algorithm to Diagnose Invasive Pulmonary Aspergillosis in Critically Ill Patients. Am J Respir Crit Care Med. 2012;186(1):56-64.

31. Nir-Paz R. Atypical Pneumonia (Non-Covid-19)2020 12/08/2020. Available from: https://bestpractice.bmj.com/topics/en-gb/18.

32. Cunha BA, Burillo A, Bouza E. Legionnaires' Disease. The Lancet. 2016;387(10016):376-85.

33. Cilloniz C, Martin-Loeches I, Garcia-Vidal C, San Jose A, Torres A. Microbial Etiology of Pneumonia: Epidemiology, Diagnosis and Resistance Patterns. International journal of molecular sciences. 2016;17(12):2120.

34. Cillóniz C, Ewig S, Ferrer M, Polverino E, Gabarrús A, Puig de la Bellacasa J, et al. Community-Acquired Polymicrobial Pneumonia in the Intensive Care Unit: Aetiology and Prognosis. Critical Care. 2011;15(5):R209.

35. Beauté J, The European Legionnaires' Disease Surveillance N. Legionnaires' Disease in Europe, 2011 to 2015. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2017;22(27):30566.

36. Harrison TG, Afshar B, Doshi N, Fry NK, Lee JV. Distribution of Legionella Pneumophila Serogroups, Monoclonal Antibody Subgroups and DNA Sequence Types in Recent Clinical and Environmental Isolates from England and Wales (2000-2008). Eur J Clin Microbiol Infect Dis. 2009;28(7):781-91.

37. Hopstaken R, Muris J, Knottnerus J, Kester A, Rinkens P, Dinant G. Contributions of Symptoms, Signs, Erythrocyte Sedimentation Rate, and C-Reactive Protein to a Diagnosis of Pneumonia in Acute Lower Respiratory Tract Infection. Br J Gen Pract. 2003;53(490):358-64.

38. Flanders SA, Stein J, Shochat G, Sellers K, Holland M, Maselli J, et al. Performance of a Bedside C-Reactive Protein Test in the Diagnosis of Community-Acquired Pneumonia in Adults with Acute Cough. The American Journal of Medicine. 2004;116(8):529-35.

39. Holm A, Pedersen SS, Nexoe J, Obel N, Nielsen LP, Koldkjaer O. Procalcitonin Versus C-Reactive Protein for Predicting Pneumonia in Adults with Lower Respiratory Tract Infection in Primary Care. Br J Gen Pract. 2007;57(540):555-60.

40. Van der Meer V, Neven AK, van den Broek PJ, Assendelft WJ. Diagnostic Value of C Reactive Protein in Infections of the Lower Respiratory Tract: Systematic Review. Bmj. 2005;331(7507):26.

41. Kirby BD, Snyder KM, Meyer RD, Finegold SM. Legionnaires' Disease: Report of Sixty-Five Nosocomially Acquired Cases of Review of the Literature. Medicine (Baltimore). 1980;59(3):188-205.

42. NICE. Covid-19 Rapid Guideline: Antibiotics for Pneumonia in Adults in Hospital 2020.

43. PHE. Start Smart - Then Focus Antimicrobial Stewardship Toolkit for English Hospitals Public Health England 2015.

44. Salkind AR, Cuddy PG, Foxworth JW. Fluoroquinolone Treatment of Community-Acquired Pneumonia: A Meta-Analysis. Annals of Pharmacotherapy. 2002;36(12):1938-43.

45. Lin T, Lin S, Chen H, Wang C, Wang Y, Chang M, et al. An Open-Label, Randomized Comparison of Levofloxacin and Amoxicillin/Clavulanate Plus Clarithromycin for the Treatment of Hospitalized Patients with Community-Acquired Pneumonia. Chang Gung medical journal. 2007;30(4):321.

46. Dartois N, Castaing N, Gandjini H, Cooper A. Tigecycline Versus Levofloxacin for the Treatment of Community-Acquired Pneumonia: European Experience. Journal of Chemotherapy. 2008;20(sup1):28-35.

47. Kothe H, Bauer T, Marre R, Suttorp N, Welte T, Dalhoff K, et al. Outcome of Community-Acquired Pneumonia: Influence of Age, Residence Status and Antimicrobial Treatment. European Respiratory Journal. 2008;32(1):139-46.

48. Baum Hv, Weite T, Marre R, Suttorp N, Ewig S. Community-Acquired Pneumonia through Enterobacteriaceae and Pseudomonas Aeruginosa: Diagnosis, Incidence and Predictors. European Respiratory Journal. 2010;35(3):598-605.

49. Murcia J, Gonzalez-Comeche J, Marin A, Barberán J, Granizo J, Aguilar L, et al. Clinical Response to Ertapenem in Severe Community-Acquired Pneumonia: A Retrospective Series in an Elderly Population. Clinical Microbiology and Infection. 2009;15(11):1046-50.
50. Garau J, Fritsch A, Arvis P, Read RC. Clinical Efficacy of Moxifloxacin Versus Comparator Therapies for Community-Acquired Pneumonia Caused by Legionella Spp. J Chemother. 2010;22(4):264-6.
51. Dunbar LM, Khashab MM, Kahn JB, Zadeikis N, Xiang JX, Tennenberg AM. Efficacy of 750-Mg, 5-Day Levofloxacin in the Treatment of Community-Acquired Pneumonia Caused by Atypical Pathogens. Curr Med Res Opin. 2004;20(4):555-63.
52. Shah PJ, Ryzner KL. Evaluating the Appropriate Use of Piperacillin/Tazobactam in a Community Health System: A Retrospective Chart Review. P & T : a peer-reviewed journal for formulary management. 2013;38(8):462-83.
53. Dalovisio JR. Overview of Lower Respiratory Tract Infections: Diagnosis and Treatment. The Ochsner journal. 2002;4(4):227-33.
54. Iregui M, Ward S, Sherman G, Fraser VJ, Kollef MH. Clinical Importance of Delays in the Initiation of Appropriate Antibiotic Treatment for Ventilator-Associated Pneumonia. Chest. 2002;122(1):262-8.
55. Ibrahim EH, Sherman G, Ward S, Fraser VJ, Kollef MH. The Influence of Inadequate Antimicrobial Treatment of Bloodstream Infections on Patient Outcomes in the ICU Setting. Chest. 2000;118(1):146-55.
56. Kollef MH. Inadequate Antimicrobial Treatment: An Important Determinant of Outcome for Hospitalized Patients. Clin Infect Dis. 2000;31 Suppl 4:S131-8.
57. Kollef MH, Sherman G, Ward S, Fraser VJ. Inadequate Antimicrobial Treatment of Infections: A Risk Factor for Hospital Mortality among Critically Ill Patients. Chest. 1999;115(2):462-74.
58. Cosgrove SE, Kaye KS, Eliopoulous GM, Carmeli Y. Health and Economic Outcomes of the Emergence of Third-Generation Cephalosporin Resistance in Enterobacter Species. Arch Intern Med. 2002;162(2):185-90.
59. Alvarez-Lerma F. Modification of Empiric Antibiotic Treatment in Patients with Pneumonia Acquired in the Intensive Care Unit. Icu-Acquired Pneumonia Study Group. Intensive Care Med. 1996;22(5):387-94.
60. Evans RS, Classen DC, Pestotnik SL, Lundsgaarde HP, Burke JP. Improving Empiric Antibiotic Selection Using Computer Decision Support. Arch Intern Med. 1994;154(8):878-84.
61. Evans RS, Pestotnik SL, Classen DC, Clemmer TP, Weaver LK, Orme JF, Jr., et al. A Computer-Assisted Management Program for Antibiotics and Other Antiinfective Agents. N Engl J Med. 1998;338(4):232-8.
62. Garcin F, Leone M, Antonini F, Charvet A, Albanèse J, Martin C. Non-Adherence to Guidelines: An Avoidable Cause of Failure of Empirical Antimicrobial Therapy in the Presence of Difficult-to-Treat Bacteria. Intensive Care Medicine. 2010;36(1):75-82.
63. NICE. Respiratory Tract Infections – Antibiotic Prescribing Prescribing of Antibiotics for Self-Limiting Respiratory Tract Infections in Adults and Children in Primary Care. Centre for Clinical Practice 2008
64. NICE. Pneumonia: Diagnosis and Management of Community- and Hospital-Acquired Pneumonia in Adults. Pneumonia NICE guideline: National Institute for Health and Care Excellence; 2014.

65. PHE. Uk Standards for Microbiology Investigations. Investigation of Bronchoalveolar Lavage, Sputum and Associated Specimens. Bacteriology. 2018;B57(3.5):38.

66. Lim WS, Baudouin SV, George RC, Hill AT, Jamieson C, Le Jeune I, et al. Bts Guidelines for the Management of Community Acquired Pneumonia in Adults: Update 2009. Thorax. 2009;64(Suppl 3):iii1.

67. Jarraud S, Descours G, Ginevra C, Lina G, Etienne J. Identification of Legionella in Clinical Samples. Methods Mol Biol. 2013;954:27-56.

68. Pierre DM, Baron J, Yu VL, Stout JE. Diagnostic Testing for Legionnaires' Disease. Annals of Clinical microbiology and antimicrobials. 2017;16(1):59-.

69. Guerrero C, Toldos CM, Yague G, Ramirez C, Rodriguez T, Segovia M. Comparison of Diagnostic Sensitivities of Three Assays (Bartels Enzyme Immunoassay [Eia], Biotest Eia, and Binax Now Immunochromatographic Test) for Detection of Legionella Pneumophila Serogroup 1 Antigen in Urine. J Clin Microbiol. 2004;42(1):467-8.

70. Dirven K, Ieven M, Peeters MF, van der Zee A, De Schrijver K, Goossens H. Comparison of Three Legionella Urinary Antigen Assays During an Outbreak of Legionellosis in Belgium. J Med Microbiol. 2005;54(Pt 12):1213-6.

71. Kobashi Y, Yoshida K, Miyashita N, Niki Y, Matsushima T. Evaluating the Use of a Streptococcus Pneumoniae Urinary Antigen Detection Kit for the Management of Community-Acquired Pneumonia in Japan. Respiration. 2007;74(4):387-93.

72. Hawkey PM, Warren RE, Livermore DM, McNulty CAM, Enoch DA, Otter JA, et al. Treatment of Infections Caused by Multidrug-Resistant Gram-Negative Bacteria: Report of the British Society for Antimicrobial Chemotherapy/Healthcare Infection Society/British Infection Association Joint Working Party†. Journal of Antimicrobial Chemotherapy. 2018;73(suppl_3):iii2-iii78.

73. Enne VI, King A, Livermore DM, Hall LMC. Sulfonamide Resistance in Haemophilus Influenzae Mediated by Acquisition of Sul2 or a Short Insertion in Chromosomal Folp. Antimicrobial Agents and Chemotherapy. 2002;46(6):1934-9.

74. Eliopoulos GM, Huovinen P. Resistance to Trimethoprim-Sulfamethoxazole. Clinical Infectious Diseases. 2001;32(11):1608-14.

75. Bousbia S, Raoult D, La Scola B. Pneumonia Pathogen Detection and Microbial Interactions in Polymicrobial Episodes. Future Microbiol. 2013;8(5):633-60.

76. Waters BM, J. A 2015 Update on Ventilator-Associated Pneumonia: New Insights on Its Prevention, Diagnosis, and Treatment. Curr Infect Dis Rep. 2015;17(8):496.

77. Llor C, Cots JM, López-Valcárcel BG, Arranz J, García G, Ortega J, et al. Interventions to Reduce Antibiotic Prescription for Lower Respiratory Tract Infections: Happy Audit Study. European Respiratory Journal. 2012;40(2):436.

78. CDC. Antibiotic Use in the United States, 2018 Update: Progress and Opportunities. Atlanta, GA: US Department of Health and Human Services; 2019.

79. Cookson WOCM, Cox MJ, Moffatt MF. New Opportunities for Managing Acute and Chronic Lung Infections. Nature Reviews Microbiology. 2017;16:111.

80. Welte T, Torres A, Nathwani D. Clinical and Economic Burden of Community-Acquired Pneumonia among Adults in Europe. Thorax. 2012;67(1):71.

81. Tong S, Amand C, Kieffer A, Kyaw MH. Trends in Healthcare Utilization and Costs Associated with Pneumonia in the United States During 2008–2014. BMC Health Services Research. 2018;18(1):715.

82. Davies SC. Chapter 1 Chief Medical Officer’s Summary. 2016.

83. O'Neill J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. The Review on Antimicrobial Resistance 2016.

84. Goverment H. Tackling Antimicrobial Resistance 2019–2024. The Uk's Five-Year National Action Plan. Department of Health and Social Care; 2019.

85. Ginevra C, Chastang J, David S, Mentasti M, Yakunin E, Chalker VJ, et al. A Real-Time Pcr for Specific Detection of the Legionella Pneumophila Serogroup 1 St1 Complex. Clinical Microbiology and Infection. 2019.

86. Tenover FC. Developing Molecular Amplification Methods for Rapid Diagnosis of Respiratory Tract Infections Caused by Bacterial Pathogens. Clin Infect Dis. 2011;52 Suppl 4:S338-45.

87. Smith MD, Sheppard CL, Hogan A, Harrison TG, Dance DAB, Derrington P, et al. Diagnosis of Streptococcus Pneumoniae Infections in Adults with Bacteremia and Community-Acquired Pneumonia: Clinical Comparison of Pneumococcal Pcr and Urinary Antigen Detection. Journal of Clinical Microbiology. 2009;47(4):1046-9.

88. Templeton KE, Scheltinga SA, van den Eeden WC, Graffelman AW, van den Broek PJ, Claas EC. Improved Diagnosis of the Etiology of Community-Acquired Pneumonia with Real-Time Polymerase Chain Reaction. Clin Infect Dis. 2005;41(3):345-51.

89. Deepak S, Kottapalli K, Rakwal R, Oros G, Rangappa K, Iwahashi H, et al. Real-Time Pcr: Revolutionizing Detection and Expression Analysis of Genes. Current genomics. 2007;8(4):234-51.

90. S K. Clinical Evaluation of the Biofire Filmarray Pneumoniae Panel Plus. Paper Presented At: European Society of Clinical Microbiology and Infectious Disease. European Society of Clinical Microbiology and Infectious Diseases. 2018.

91. Gadsby NJ, McHugh MP, Forbes C, MacKenzie L, Hamilton SKD, Griffith DM, et al. Comparison of Unyvero P55 Pneumonia Cartridge, in-House Pcr and Culture for the Identification of Respiratory Pathogens and Antibiotic Resistance in Bronchoalveolar Lavage Fluids in the Critical Care Setting. European Journal of Clinical Microbiology & Infectious Diseases. 2019;38(6):1171-8.

92. Cercenado E, Marin M, Burillo A, Martin-Rabadan P, Rivera M, Bouza E. Rapid Detection of Staphylococcus Aureus in Lower Respiratory Tract Secretions from Patients with Suspected Ventilator-Associated Pneumonia: Evaluation of the Cepheid Xpert Mrsa/Sa Ssti Assay. J Clin Microbiol. 2012;50(12):4095-7.

93. Lung M, Codina G. Molecular Diagnosis in Hap/Vap. Curr Opin Crit Care. 2012;18(5):487-94.

94. Murphy CN, Fowler R, Balada-Llasat JM, Carroll A, Stone H, Akerele O, et al. Multicenter Evaluation of the Biofire Filmarray Pneumonia/Pneumonia Plus Panel for Detection and Quantification of Agents of Lower Respiratory Tract Infection. Journal of Clinical Microbiology. 2020;58(7):e00128-20.

95. Virve Irene Enne AA, Hollian Richardson, Dewi Rhys Owen, Rossella Baldan, Charlotte Russell, Brenda, Nomamiukor AM, Juliet High, Antony Colles, Federico Ricciardi, Julie Barber, Vanya Gant, David Livermore, Justin O'Grady. Inhale Wp1: An Observational Study Comparing the Performance of Two Multiplex Pcr Platforms against Routine Microbiology for
the Detection of Potential Pathogens in Patients with Suspected Hospital Acquired/Ventilator Associated Pneumonia (Hap/Vap)
across 15 Intensive Care Units (Icus). European Society of Clinical Microbiology and Infectious Diseases. 2018.

96. Kodani M, Yang G, Conklin LM, Travis TC, Whitney CG, Anderson LJ, et al. Application of Taqman Low-Density Arrays for Simultaneous Detection of Multiple Respiratory Pathogens. Journal of Clinical Microbiology. 2011;49(6):2175-82.
97. Hassibi A, Manickam A, Singh R, Bolouki S, Sinha R, Jirage KB, et al. Multiplexed Identification, Quantification and Genotyping of Infectious Agents Using a Semiconductor Biochip. Nature Biotechnology. 2018.
98. 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. 2015;8(11):15479-88.
99. Kutlu SS, Sacar S, Cevahir N, Turgut H. Community-Acquired Streptococcus Mitis Meningitis: A Case Report. International Journal of Infectious Diseases. 2008;12(6):e107-e9.
100. Kais M, Spindler C, Kalin M, Örtqvist Å, Giske CG. Quantitative Detection of Streptococcus Pneumoniae, Haemophilus Influenzae, and Moraxella Catarrhalis in Lower Respiratory Tract Samples by Real-Time Pcr. Diagnostic Microbiology and Infectious Disease. 2006;55(3):169-78.
101. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance Comparison of Benchtop High-Throughput Sequencing Platforms. Nature Biotechnology. 2012;30:434.
102. Chiu CY, Miller SA. Clinical Metagenomics. Nature Reviews Genetics. 2019.
103. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High Throughput Sequencing Methods and Analysis for Microbiome Research. Journal of Microbiological Methods. 2013;95(3):401-14.
104. Loman NJ, Pallen MJ. Twenty Years of Bacterial Genome Sequencing. Nature Reviews Microbiology. 2015;13(12):787-94.
105. International Human Genome Sequencing C. Finishing the Euchromatic Sequence of the Human Genome. Nature. 2004;431:931.
106. International Human Genome Sequencing C, Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. Initial Sequencing and Analysis of the Human Genome. Nature. 2001;409:860.
107. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. Nature Biotechnology. 2018;36:338.
108. Sanger F, Nicklen S, Coulson AR. DNA Sequencing with Chain-Terminating Inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
109. Heather JM, Chain B. The Sequence of Sequencers: The History of Sequencing DNA. Genomics. 2016;107(1):1-8.
110. Eisenstein M. Oxford Nanopore Announcement Sets Sequencing Sector Abuzz. Nature Biotechnology. 2012;30:295.
111. PacBio. Smrt Sequencing: Read Lengths 2018 [Available from: https://www.pacb.com/smrt-science/smrt-sequencing/read-lengths/.
112. Ip C, Loose M, Tyson J, de Cesare M, Brown B, Jain M, et al. Minion Analysis and Reference Consortium: Phase 1 Data Release and Analysis [Version 1; Referees: 2 Approved]. F1000 Research. 2015;4(1075).
113. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. Whole Genome Sequencing and Assembly of a Caenorhabditis Elegans Genome with Complex Genomic Rearrangements Using the Minion Sequencing Device. bioRxiv. 2017.
114. Deamer D, Akeson M, Branton D. Three Decades of Nanopore Sequencing2016. 518-24 p.

115. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. Scientific Reports. 2017;7(1):18022.

116. Rang FJ, Kloosterman WP, de Ridder J. From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy. Genome Biology. 2018;19(1):90.

117. Brown CG. Oxford Nanopore Technologies: "No Thanks, I've Already Got One." Oxford Nanopore Technologies: Oxford Nanopore Technologies; 2016 [Available from: https://www.youtube.com/watch?v=nizGyutn6v4.

118. Brown CG. Nanopore Technologies: Some Mundane and Fundamental Updates Oxford Nanopore Technologies: Oxford Nanopore Technologies; 2018 [Available from: https://www.youtube.com/watch?v=7pIpf-jj-7w.

119. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, et al. Enabling High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or Pacbio Sequencing. bioRxiv. 2020:645903.

120. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and Challenges in Long-Read Sequencing Data Analysis. Genome Biology. 2020;21(1):30.

121. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, et al. Minion Nanopore Sequencing Identifies the Position and Structure of a Bacterial Antibiotic Resistance Island. Nat Biotechnol. 2015;33(3):296-300.

122. 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. Nature Biotechnology. 2019;37(7):783-92.

123. 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. Nature Microbiology. 2020;5(3):430-42.

124. Page AJ, Langridge GC. Socru: Typing of Genome Level Order and Orientation in Bacteria. bioRxiv. 2019:543702.

125. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences Rs Long-Read Sequencing Technology. PLOS ONE. 2012;7(11):e47768.

126. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-Time, Portable Genome Sequencing for Ebola Surveillance. Nature. 2016;530:228.

127. Rhoads A, Au KF. Pacbio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics. 2015;13(5):278-89.

128. Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis Via Whole Genome Sequencing of Direct Respiratory Samples. Journal of Clinical Microbiology. 2017.

129. 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. Journal of Antimicrobial Chemotherapy. 2017;72(1):104-14.

130. 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.

131. Van der Verren SE, Van Gerven N, Jonckheere W, Hambley R, Singh P, Kilgour J, et al. A Dual-Constriction Biological Nanopore Resolves Homonucleotide Sequences with High Fidelity. Nat Biotechnol. 2020;38(12):1415-20.
132. Thomas T, Gilbert J, Meyer F. Metagenomics - a Guide from Sampling to Data Analysis. Microbial Informatics and Experimentation. 2012;2:3-.
133. Wooley JC, Godzik A, Friedberg I. A Primer on Metagenomics. PLOS Computational Biology. 2010;6(2):e1000667.
134. Ruppé E, Greub G, Schrenzel J. Messages from the First International Conference on Clinical Metagenomics (Iccmg). Microbes and Infection. 2017;19(4):223-8.
135. Ruppé E, Schrenzel J. Messages from the Second International Conference on Clinical Metagenomics (Iccmg2). Microbes and Infection. 2018;20(4):222-7.
136. 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;29(5):831-42.
137. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. New England Journal of Medicine. 2019;380(24):2327-40.
138. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. N Engl J Med. 2014;370(25):2408-17.
139. Gliddon HD, Herberg JA, Levin M, Kaforou M. Genome-Wide Host Rna Signatures of Infectious Diseases: Discovery and Clinical Translation. Immunology. 2018;153(2):171-8.
140. 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. Proceedings of the National Academy of Sciences. 2018;115(52):E12353-E62.
141. 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. American Journal of Respiratory and Critical Care Medicine. 2017;196(12):1610-2.
142. 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;127:141-5.
143. 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. Nature Microbiology. 2020;5(3):455-64.
144. 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 Medicine. 2015;7(1):99.
145. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, et al. Metagenomic Sequencing at the Epicenter of the Nigeria 2018 Lassa Fever Outbreak. Science. 2019;363(6422):74.
146. WHO. Epidimiology 2018 [Available from: http://www.who.int/topics/epidemiology/en/.
147. Ioannidis JPA. Genetic and Molecular Epidemiology. Journal of Epidemiology and Community Health. 2007;61(9):757-8.

148. Halachev MR, Chan JZ, Constantinidou CI, Cumley N, Bradley C, Smith-Banks M, et al. Genomic Epidemiology of a Protracted Hospital Outbreak Caused by Multidrug-Resistant Acinetobacter Baumannii in Birmingham, England. Genome Med. 2014;6(11):70.

149. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, Peto T, et al. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. Front Microbiol. 2017;8:182.

150. Phan HTT, Stoesser N, Maciuca IE, Toma F, Szekely E, Flonta M, et al. Illumina Short-Read and Minion Long-Read Wgs to Characterize the Molecular Epidemiology of an Ndm-1 Serratia Marcescens Outbreak in Romania. J Antimicrob Chemother. 2018;73(3):672-9.

151. Cox MJ, Carney S, Cookson W, Chalker VJ, Moffatt M. Developing Metagenomic Methods for Legionella Sequencing. B61 Bacterial Respiratory Infections. American Thoracic Society International Conference Abstracts: American Thoracic Society; 2017. p. A3950-A.

152. Loman NJ, Constantinidou C, Christner M, et al. A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia Coli O104:H4. JAMA. 2013;309(14):1502-10.

153. 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. Journal of Clinical Microbiology. 2017;55(1):177-82.

154. Moore SC, Penrice-Randal R, Alruwaili M, Dong X, Pullan ST, Carter D, et al. Amplicon Based Minion Sequencing of Sars-Cov-2 and Metagenomic Characterisation of Nasopharyngeal Swabs from Patients with Covid-19. medRxiv. 2020:2020.03.05.20032011.

155. Sande MG, Çaykara T, Silva CJ, Rodrigues LR. New Solutions to Capture and Enrich Bacteria from Complex Samples. Medical Microbiology and Immunology. 2020;209(3):335-41.

156. Matta LL, Harrison J, Deol GS, Alocilja EC. Carbohydrate-Functionalized Nanobiosensor for Rapid Extraction of Pathogenic Bacteria Directly from Complex Liquids with Quick Detection Using Cyclic Voltammetry. IEEE Transactions on Nanotechnology. 2018;17(5):1006-13.

157. Wang J, Wu H, Yang Y, Yan R, Zhao Y, Wang Y, et al. Bacterial Species-Identifiable Magnetic Nanosystems for Early Sepsis Diagnosis and Extracorporeal Photodynamic Blood Disinfection. Nanoscale. 2018;10(1):132-41.

158. García-García G, Baux D, Faugère V, Moclyn M, Koenig M, Claustres M, et al. Assessment of the Latest Ngs Enrichment Capture Methods in Clinical Context. Scientific Reports. 2016;6(1):20948.

159. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid Whole-Genome Sequencing of Mycobacterium Tuberculosis Isolates Directly from Clinical Samples. Journal of Clinical Microbiology. 2015;53(7):2230-7.

160. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, et al. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium Tuberculosis Faster Than Mgit Culture Sequencing. Journal of Clinical Microbiology. 2018;56(8).

161. Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, Brehio P, et al. Capturing Sequence Diversity in Metagenomes with Comprehensive and Scalable Probe Design. Nature Biotechnology. 2019;37(2):160-8.

162. 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. Nature Microbiology. 2020.

163. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and Cryptic Transmission of Zika Virus in Brazil and the Americas. Nature. 2017;546:406.

164. COG-UK C. An Integrated National Scale Sars-Cov-2 Genomic Surveillance Network. The Lancet Microbe. 2020.

165. 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.

166. Thoendel MJ, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Identification of Prosthetic Joint Infection Pathogens Using a Shotgun Metagenomics Approach. Clin Infect Dis. 2018;67(9):1333-8.

167. 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. Journal of Clinical Microbiology. 2016;54(4):919-27.

168. Leo S, Gaïa N, Ruppé E, Emonet S, Girard M, Lazarevic V, et al. Detection of Bacterial Pathogens from Broncho-Alveolar Lavage by Next-Generation Sequencing. 2017.

169. 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. Journal of Microbiological Methods. 2016;127:141-5.

170. Lorent JH, Quetin-Leclercq J, Mingeot-Leclercq M-P. The Amphiphilic Nature of Saponins and Their Effects on Artificial and Biological Membranes and Potential Consequences for Red Blood and Cancer Cells. Organic & Biomolecular Chemistry. 2014;12(44):8803-22.

171. Francis G, Kerem Z, Makkar HPS, Becker K. The Biological Action of Saponins in Animal Systems: A Review. British Journal of Nutrition. 2002;88(6):587-605.

172. Sudji RI, Subburaj Y, Frenkel N, García-Sáez JA, Wink M. Membrane Disintegration Caused by the Steroid Saponin Digitonin Is Related to the Presence of Cholesterol. Molecules. 2015;20(11).

173. Bissinger R, Modicano P, Alzoubi K, Honisch S, Faggio C, Abed M, et al. Effect of Saponin on Erythrocytes. International Journal of Hematology. 2014;100(1):51-9.

174. Orjih AU, Cherian P, AlFadhli S. Microscopic Detection of Mixed Malarial Infections: Improvement by Saponin Hemolysis. Med Princ Pract. 2008;17(6):458-63.

175. Waheed A, Barker J, Barton SJ, Owen CP, Ahmed S, Carew MA. A Novel Steroidal Saponin Glycoside from Fagonia Indica Induces Cell-Selective Apoptosis or Necrosis in Cancer Cells. European Journal of Pharmaceutical Sciences. 2012;47(2):464-73.

176. Chen P-S, Shih Y-W, Huang H-C, Cheng H-W. Diosgenin, a Steroidal Saponin, Inhibits Migration and Invasion of Human Prostate Cancer Pc-3 Cells by Reducing Matrix Metalloproteinases Expression. PLOS ONE. 2011;6(5):e20164.

177. Zelenin S, Hansson J, Ardabili S, Ramachandraiah H, Brismar H, Russom A. Microfluidic-Based Isolation of Bacteria from Whole Blood for Sepsis Diagnostics. Biotechnology Letters. 2015;37(4):825-30.

178. Anscombe C, Misra RV, Gharbia S. Whole Genome Amplification and Sequencing of Low Cell Numbers Directly from a Bacteria Spiked Blood Model. bioRxiv. 2018:153965.

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

180. 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. Journal of Clinical Microbiology. 2011;49(7):2533-9.

181. Kessler HH, Mühlbauer G, Stelzl E, Daghofer E, Santner BI, Marth E. Fully Automated Nucleic Acid Extraction: Magna Pure Lc. Clinical Chemistry. 2001;47(6):1124-6.

182. Matijasic BB, Obermajer T, Lipoglavsek L, Grabnar I, Avgustin G, Rogelj I. Association of Dietary Type with Fecal Microbiota in Vegetarians and Omnivores in Slovenia. Eur J Nutr. 2014;53(4):1051-64.
183. McGraw J, Tatipelli VK, Feyijinmi O, Traore MC, Eangoor P, Lane S, et al. A Semi-Automated Method for Purification of Milligram Quantities of Proteins on the Qiacube. Protein Expression and Purification. 2014;96:48-53.
184. Gener AR. Full-Coverage Sequencing of Hiv-1 Provirus from a Reference Plasmid. bioRxiv. 2019:611848.
185. 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. Respiratory Research. 2019;20(1):265.
186. Quince C, W Walker A, T Simpson J, J Loman N, Segata N. Shotgun Metagenomics, from Sampling to Analysis2017. 833-44 p.
187. TECHNOLOGIES ON. Primary Data Analysis 2018 [Available from: https://nanoporetech.com/analyse.
188. Wick RR, Judd LM, Holt KE. Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing. Genome Biology. 2019;20(1):129.
189. Nanoporetech. Scrappie GitHub: Oxford Nanopore Technologies; 2019 [Available from: https://github.com/nanoporetech/scrappie.
190. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. bioRxiv. 2016.
191. Boža V, Brejová B, Vinař T. Deepnano: Deep Recurrent Neural Networks for Base Calling in Minion Nanopore Reads. PloS One. 2017;12(6):e0178751.
192. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: Translating Nanopore Raw Signal Directly into Nucleotide Sequence Using Deep Learning. GigaScience. 2018;7(5).
193. Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. Metagenomics for Pathogen Detection in Public Health. Genome Med. 2013;5(9):81.
194. Breitwieser FP, Lu J, Salzberg SL. A Review of Methods and Databases for Metagenomic Classification and Assembly. Brief Bioinform. 2019;20(4):1125-36.
195. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. Metaphlan2 for Enhanced Metagenomic Taxonomic Profiling. Nat Methods. 2015;12(10):902-3.
196. Noé L, Martin DEK. A Coverage Criterion for Spaced Seeds and Its Applications to Support Vector Machine String Kernels and K-Mer Distances. Journal of Computational Biology. 2014;21(12):947-63.
197. Wood DE, Salzberg SL. Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. Genome Biology. 2014;15(3):R46.
198. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences. bioRxiv. 2016.
199. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 2019;178(4):779-94.
200. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun Metagenomics, from Sampling to Analysis. Nature Biotechnology. 2017;35(9):833-44.
201. Ayling M, Clark MD, Leggett RM. New Approaches for Metagenome Assembly with Short Reads. Briefings in Bioinformatics. 2019.

202. Ruppé E, Cherkaoui A, Charretier Y, Girard M, Schicklin S, Lazarevic V, et al. From Genotype to Antibiotic Susceptibility Phenotype in the Order Enterobacterales: A Clinical Perspective. Clinical Microbiology and Infection. 2019.

203. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A Cloud-Compatible Bioinformatics Pipeline for Ultrarapid Pathogen Identification from Next-Generation Sequencing of Clinical Samples. Genome Res. 2014;24(7):1180-92.

204. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. Card 2017: Expansion and Model-Centric Curation of the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research. 2017;45(D1):D566-D73.

205. Federhen S. The Ncbi Taxonomy Database. Nucleic Acids Research. 2012;40(Database issue):D136-D43.

206. Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database C. The International Nucleotide Sequence Database Collaboration. Nucleic acids research. 2016;44(D1):D48-D50.

207. Wyllie AL, Pannekoek Y, Bovenkerk S, van Engelsdorp Gastelaars J, Ferwerda B, van de Beek D, et al. Sequencing of the Variable Region of Rpsb to Discriminate between Streptococcus Pneumoniae and Other Streptococcal Species. Open Biology.7(9):170074.

208. Chen JHK, She KKK, Wong O-Y, Teng JLL, Yam W-C, Lau SKP, et al. Use of Maldi Biotyper Plus Clinprotools Mass Spectra Analysis for Correct Identification of Streptococcus Pneumoniae and Streptococcus Mitis. Journal of Clinical Pathology. 2015.

209. Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, et al. Typing Methods Based on Whole Genome Sequencing Data. One Health Outlook. 2020;2(1):1-19.

210. Vincent C, Usongo V, Berry C, Tremblay DM, Moineau S, Yousfi K, et al. Comparison of Advanced Whole Genome Sequence-Based Methods to Distinguish Strains of Salmonella Enterica Serovar Heidelberg Involved in Foodborne Outbreaks in Québec. Food Microbiol. 2018;73:99-110.

211. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, et al. Update on Refseq Microbial Genomes Resources. Nucleic acids research. 2015;43:D599-D605.

212. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of Acquired Antimicrobial Resistance Genes. J Antimicrob Chemother. 2012;67(11):2640-4.

213. Xavier BB, Das AJ, Cochrane G, De Ganck S, Kumar-Singh S, Aarestrup FM, et al. Consolidating and Exploring Antibiotic Resistance Gene Data Resources. J Clin Microbiol. 2016;54(4):851-9.

214. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A Draft Genome of Yersinia Pestis from Victims of the Black Death. Nature. 2011;478(7370):506-10.

215. McLean JS, Lombardo M-J, Ziegler MG, Novotny M, Yee-Greenbaum J, Badger JH, et al. Genome of the Pathogen Porphyromonas Gingivalis Recovered from a Biofilm in a Hospital Sink Using a High-Throughput Single-Cell Genomics Platform. Genome Research. 2013;23(5):867-77.

216. Lischer HEL, Shimizu KK. Reference-Guided De Novo Assembly Approach Improves Genome Reconstruction for Related Species. BMC Bioinformatics. 2017;18(1):474.

217. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the Two Major Classes of Assembly Algorithms: Overlap–Layout–Consensus and De-Bruijn-Graph. Briefings in Functional Genomics. 2012;11(1):25-37.

218. Charalampous T, Kay GL, OeGrady J. Applying Clinical Metagenomics for the Detection and Characterisation of Respiratory Infections. The Lung Microbiome (ERS Monograph) Sheffield, European Respiratory Society. 2019:35-49.

219. Koren S, Phillippy AM. One Chromosome, One Contig: Complete Microbial Genomes from Long-Read Sequencing and Assembly. Current Opinion in Microbiology. 2015;23:110-20.

220. Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: Scalable and Accurate Long-Read Assembly Via Adaptive K-Mer Weighting and Repeat Separation. bioRxiv. 2016.

221. Li H. Minimap and Miniasm: Fast Mapping and De Novo Assembly for Noisy Long Sequences. Bioinformatics. 2016;32(14):2103-10.

222. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and Accurate De Novo Genome Assembly from Long Uncorrected Reads. Genome research. 2017;27(5):737-46.

223. Molina-Mora JA, Campos-Sánchez R, Rodríguez C, Shi L, García F. High Quality 3c De Novo Assembly and Annotation of a Multidrug Resistant St-111 Pseudomonas Aeruginosa Genome: Benchmark of Hybrid and Non-Hybrid Assemblers. Scientific Reports. 2020;10(1):1392.

224. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads. PLOS Computational Biology. 2017;13(6):e1005595.

225. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. Hybridspades: An Algorithm for Hybrid Assembly of Short and Long Reads. Bioinformatics. 2016;32(7):1009-15.

226. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al., editors. Assembling Genomes and Mini-Metagenomes from Highly Chimeric Reads. Annual International Conference on Research in Computational Molecular Biology; 2013: Springer.

227. Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, et al. Resolution of a Meningococcal Disease Outbreak from Whole-Genome Sequence Data with Rapid Web-Based Analysis Methods. J Clin Microbiol. 2012;50(9):3046-53.

228. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole Genome-Based Population Biology and Epidemiological Surveillance of Listeria Monocytogenes. Nat Microbiol. 2016;2:16185.

229. Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative Analysis of Core Genome Mlst and Snp Typing within a European Salmonella Serovar Enteritidis Outbreak. Int J Food Microbiol. 2018;274:1-11.

230. Bakker HC, Switt AI, Cummings CA, Hoelzer K, Degoricija L, Rodriguez-Rivera LD, et al. A Whole-Genome Single Nucleotide Polymorphism-Based Approach to Trace and Identify Outbreaks Linked to a Common Salmonella Enterica Subsp. Enterica Serovar Montevideo Pulsed-Field Gel Electrophoresis Type. Appl Environ Microbiol. 2011;77(24):8648-55.

231. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A Primer on Microbial Bioinformatics for Nonbioinformaticians. Clinical Microbiology and Infection. 2018;24(4):342-9.

232. Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, et al. Snapperdb: A Database Solution for Routine Sequencing Analysis of Bacterial Isolates. Bioinformatics. 2018;34(17):3028-9.

233. Gaia V, Fry NK, Afshar B, Lück PC, Meugnier H, Etienne J, et al. Consensus Sequence-Based Scheme for Epidemiological Typing of Clinical and Environmental Isolates of Legionella Pneumophila. Journal of Clinical Microbiology. 2005;43(5):2047.

234. Ratcliff RM. Sequence-Based Identification of Legionella. Methods Mol Biol. 2013;954:57-72.

235. Mentasti M, Fry NK, Afshar B, Palepou-Foxley C, Naik FC, Harrison TG. Application of Legionella Pneumophila-Specific Quantitative Real-Time Pcr Combined with Direct Amplification and Sequence-Based Typing in the Diagnosis and Epidemiological Investigation of Legionnaires' Disease. Eur J Clin Microbiol Infect Dis. 2012;31(8):2017-28.

236. Borthong J, Omori R, Sugimoto C, Suthienkul O, Nakao R, Ito K. Comparison of Database Search Methods for the Detection of Legionella Pneumophila in Water Samples Using Metagenomic Analysis. Frontiers in Microbiology. 2018;9:1272.

237. Llewellyn AC, Lucas CE, Roberts SE, Brown EW, Nayak BS, Raphael BH, et al. Distribution of Legionella and Bacterial Community Composition among Regionally Diverse Us Cooling Towers. PLOS ONE. 2017;12(12):e0189937.

238. Dai D, Rhoads WJ, Edwards MA, Pruden A. Shotgun Metagenomics Reveals Taxonomic and Functional Shifts in Hot Water Microbiome Due to Temperature Setting and Stagnation. Frontiers in Microbiology. 2018;9(2695).

239. Mentasti M, Kese D, Echahidi F, Uldum SA, Afshar B, David S, et al. Design and Validation of a Qpcr Assay for Accurate Detection and Initial Serogrouping of Legionella Pneumophila in Clinical Specimens by the Escmid Study Group for Legionella Infections (Esgli). Eur J Clin Microbiol Infect Dis. 2015;34(7):1387-93.

240. Merault N, Rusniok C, Jarraud S, Gomez-Valero L, Cazalet C, Marin M, et al. Specific Real-Time Pcr for Simultaneous Detection and Identification of Legionella Pneumophila Serogroup 1 in Water and Clinical Samples. Appl Environ Microbiol. 2011;77(5):1708-17.

241. Murphy NM, McLauchlin J, Ohai C, Grant KA. Construction and Evaluation of a Microbiological Positive Process Internal Control for Pcr-Based Examination of Food Samples for Listeria Monocytogenes and Salmonella Enterica. International Journal of Food Microbiology. 2007;120(1):110-9.

242. Gadsby NJ, Helgason KO, Dickson EM, Mills JM, Lindsay DSJ, Edwards GF, et al. Molecular Diagnosis of Legionella Infections – Clinical Utility of Front-Line Screening as Part of a Pneumonia Diagnostic Algorithm. Journal of Infection. 2016;72(2):161-70.

243. Ginevra C, Lopez M, Forey F, Reyrolle M, Meugnier H, Vandenesch F, et al. Evaluation of a Nested-Pcr-Derived Sequence-Based Typing Method Applied Directly to Respiratory Samples from Patients with Legionnaires&#039; Disease. Journal of Clinical Microbiology. 2009;47(4):981.

244. 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 Research. 2013;41(1):e1-e.

245. Schell WA, Benton JL, Smith PB, Poore M, Rouse JL, Boles DJ, et al. Evaluation of a Digital Microfluidic Real-Time Pcr Platform to Detect DNA of Candida Albicans in Blood. Eur J Clin Microbiol Infect Dis. 2012;31(9):2237-45.

246. Li H. Minimap2: Pairwise Alignment for Nucleotide Sequences. Bioinformatics. 2018:bty191-bty.

247. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. Complete Assembly of Parental Haplotypes with Trio Binning. bioRxiv. 2018.

248. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. Blast Ring Image Generator (Brig): Simple Prokaryote Genome Comparisons. BMC Genomics. 2011;12(1):402.

249. Page AJ, Keane JA. Rapid Multi-Locus Sequence Typing Direct from Uncorrected Long Reads Using Krocus. PeerJ. 2018;6:e5233.
250. Feßler AT, Wang Y, Wu C, Schwarz S. Mobile Lincosamide Resistance Genes in Staphylococci. Plasmid. 2018;99:22-31.
251. Schmitz F-J, Sadurski R, Kray A, Boos M, Geisel R, Köhrer K, et al. Prevalence of Macrolide-Resistance Genes in Staphylococcus Aureus and Enterococcus Faecium Isolates from 24 European University Hospitals. Journal of Antimicrobial Chemotherapy. 2000;45(6):891-4.
252. Livermore DM, Winstanley TG, Shannon KP. Interpretative Reading: Recognizing the Unusual and Inferring Resistance Mechanisms from Resistance Phenotypes. J Antimicrob Chemother. 2001;48 Suppl 1:87-102.
253. Tong S, Amand C, Kieffer A, Kyaw MH. Trends in Healthcare Utilization and Costs Associated with Pneumonia in the United States During 2008-2014. BMC Health Serv Res. 2018;18(1):715.
254. Buchan BW, Ledeboer NA. Emerging Technologies for the Clinical Microbiology Laboratory. Clinical Microbiology Reviews. 2014;27(4):783.
255. O'grady JJNN, Wain, John Richard (Norwich Norfolk, GB), Mwaigwisya, Solomon (Norwich Norfolk, GB), Kay, Gemma Louise (Norwich Norfolk, GB), inventor; UEA Enterprises Limited (Norwich Norfolk, GB), assignee. Method for Nucleic Acid Depletion. United States2019.
256. 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. Nature protocols. 2008;3(6):965.
257. John Richard Wain, JJOG, Solomon Mwaigwisya. Method for Nucleic Acid Depletion2016
258. Herrmann C, Avgousti DC, Weitzman MD. Differential Salt Fractionation of Nuclei to Analyze Chromatin-Associated Proteins from Cultured Mammalian Cells. Bio Protoc. 2017;7(6).
259. Horz HP, Scheer S, Huenger F, Vianna ME, Conrads G. Selective Isolation of Bacterial DNA from Human Clinical Specimens. J Microbiol Methods. 2008;72(1):98-102.
260. Johann Kubicek TS, Antje-Katrin Sander, Eva Hänssler, Dominic O'Neil, inventorMethod, Lysis Solution and Kit for Selectively Depleting Animal Nucleic Acids in a Sample. United States of America2016.
261. Street TL, Barker L, Sanderson ND, Kavanagh J, Hoosdally S, Cole K, et al. Optimizing DNA Extraction Methods for Nanopore Sequencing of Neisseria Gonorrhoea Direct from Urine Samples. Journal of Clinical Microbiology. 2019:JCM.01822-19.
262. 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. Journal of Clinical Microbiology. 2017;55(8):2334.
263. Sohrabi M, Nair RG, Samaranayake LP, Zhang L, Zulfiker AH, Ahmetagic A, et al. The Yield and Quality of Cellular and Bacterial DNA Extracts from Human Oral Rinse Samples Are Variably Affected by the Cell Lysis Methodology. J Microbiol Methods. 2016;122:64-72.
264. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. PLOS ONE. 2012;7(3):e33865.
265. Li X, Bosch-Tijhof CJ, Wei X, de Soet JJ, Crielaard W, Loveren Cv, et al. Efficiency of Chemical Versus Mechanical Disruption Methods of DNA Extraction for the Identification of Oral Gram-Positive and Gram-Negative Bacteria. The Journal of international medical research. 2020;48(5):300060520925594-.

266. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G, 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;141(6):776-86.

267. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. Archives of Pathology & Laboratory Medicine. 2017;141(6):776-86.

268. Dubourg G, Abat C, Rolain J-M, Raoult D. Correlation between Sputum and Bronchoalveolar Lavage Fluid Cultures. Journal of Clinical Microbiology. 2015;53(3):994.

269. 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. Nature Microbiology. 2019;4(4):663-74.

270. O'Grady J. A Powerful, Non-Invasive Test to Rule out Infection. Nature Microbiology. 2019;4(4):554-5.

271. Services M. Uk Standards for Microbiology Investigations. Investigation of Bronchoalveolar Lavage, Sputum and Associated Specimens. Bacteriology. 2018;B57(3.4):38.

272. Chrzastek K, Lee D-h, 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;509:159-66.

273. Reyes GR, Kim JP. Sequence-Independent, Single-Primer Amplification (Sispa) of Complex DNA Populations. Mol Cell Probes. 1991;5(6):473-81.

274. Martner A, Dahlgren C, Paton JC, Wold AE. Pneumolysin Released During Streptococcus Pneumoniae Autolysis Is a Potent Activator of Intracellular Oxygen Radical Production in Neutrophils. Infection and Immunity. 2008;76(9):4079-87.

275. 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. Proceedings of the National Academy of Sciences. 2018;115(52):E12353.

276. 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.

277. Břinda K, Callendrello A, Cowley L, Charalampous T, Lee RS, MacFadden DR, et al. Lineage Calling Can Identify Antibiotic Resistant Clones within Minutes. bioRxiv. 2018.

278. 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 Biology. 2014;12(1):87.

279. Miller S, Chiu C, Rodino KG, Miller MB. Point-Counterpoint: Should We Be Performing Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis in the Clinical Laboratory? Journal of Clinical Microbiology. 2020;58(3):e01739-19.

280. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020;382(8):727-33.

281. Ilett EE, Jørgensen M, Noguera-Julian M, Daugaard G, Murray DD, Helleberg M, et al. Gut Microbiome Comparability of Fresh-Frozen Versus Stabilized-Frozen Samples from Hospitalized Patients Using 16s Rrna Gene and Shotgun Metagenomic Sequencing. Scientific Reports. 2019;9(1):13351.