Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection

Themoula Charalampous[1]*, Gemma L. Kay[1,2]*, Hollian Richardson[1]*, 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^]

[1] Bob Champion Research and Educational Building, University of East Anglia, Norwich Research Park, Colney Ln, Norwich, UK, NR4 7UQ

[2] Quadram Institute Bioscience, Norwich Research Park, Colney Ln, Norwich, UK, NR4 7UA

[3] CIDR, King's College London, St Thomas' Hospital, Westminster Bridge Road, London, UK, SE1 7EH

[4] Microbiology Department, Norwich and Norfolk University Hospital, Conley Ln, Norwich, UK, NR4 7GJ

[5] Oxford Nanopore Technologies, Gosling Building, Oxford Science Park, Edmund Halley Rd, Oxford, UK, OX4 4DQ

[6] Earlham Institute, Norwich Research Park, Conley Ln, Norwich, UK, NR4 7UZ

[7] AMRHAI, Public Health England, 61 Colindale Ave, London, NW9 5EQ

^Correspondence to justin.ogrady@quadram.ac.uk

*These authors contributed equally to this work: TC, HR, GLK

Editors summary

Nanopore sequencing coupled with a metagenomics framework that effectively removes human DNA from samples enables rapid bacterial LRI diagnosis.

ABSTRACT

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 optimised method (6 hours from sample to result) was 96.6% sensitive and 41.7% specific for pathogen detection compared to culture and we could accurately detect antibiotic-resistance genes. After confirmatory qPCR and pathobiont-specific gene analyses, specificity and sensitivity increased to 100%. Nanopore metagenomics can rapidly and accurately characterise bacterial LRIs and might contribute to a reduction in broad-spectrum antibiotic use.

INTRODUCTION

Lower respiratory infections (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 UK, CAP accounts for approx. 29,000 deaths per annum and in the US HAP causes approx. 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 pathogen, 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 UK [1]. Initial treatment for severe LRIs usually involves empirical broad-spectrum antibiotics. Guidelines recommend that such therapy should be refined or stopped after two to three 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 hours 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 sequencting 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 e.g. urinary tract infection [26-28].

Respiratory specimens present a difficult challenge for metagenomics sequencing, owing 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 commercial kits and published methods are available for this purpose (which include differential lysis, human DNA removal and microbial DNA enrichment methods [30-33]), they do not perform well in complex respiratory samples and better methods are needed [34].

We present an optimised clinical nanopore metagenomics framework for bacterial LRIs that can remove up to 99.99% of host nucleic acid from clinical respiratory samples, and enables pathogen and antibiotic resistance gene identification within six hours.

RESULTS

Pilot method development

A pilot method was tested on respiratory samples from 40 patients with suspected bacterial LRI. This method was 91.2% sensitive (95% CI; 75.2-97.7%) and 100% specific (95% CI; 54.07-100%), *not* counting additional organisms in culture-positive samples as false positives (Table 1), and took 8 hours to perform (Figure 1). Up to 99.9% or ~10^3 fold (median 352-fold, interquartile range 144-714; maximum 1024-fold) of host DNA was removed using saponin depletion, as measured by qPCR. Microorganisms, including potential respiratory pathogens (Online methods), were identified in real-time using ONT's 'What's In My Pot?' (WIMP) pipeline. Additional pathogens, not reported by microbiological culture, were detected in 5/40 samples: *Moraxella catarrhalis* was detected in P8; *Escherichia coli* in P14; *H. influenzae* in P22 and P30; *Klebsiella pneumonia*e and *M. catarrhalis* in P29 (Table 1).

Organisms cultured using routine clinical microbiology were not detected in 3/40 sequenced samples. 2/3 samples were mixed infections, where one of the two pathogens present was missed by our pilot method – specifically, *S. pneumoniae* and *H. influenzae* were not detected in P3 and P37 respectively. *S. aureus* was not detected in the third sample, P34.

Metagenomics protocol optimisation

We sought to increase sensitivity (8.8% false negative rate) by improving bacterial cell lysis. A sample pre-treatment step was introduced (bead-beating or an enzyme cocktail, Online methods) to optimise cell lysis. Two culture-positive sputa were used for optimisation experiments, one containing *S. aureus* (Gram-positive) and one containing *P. aeruginosa* (Gram-negative). Neither pre-treatment 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. 4-fold, and bead-beating by 21-fold, compared with the pilot method, as determined by 16S qPCR (Supplementary Table 1a). The increased bacterial DNA 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. We included bead-beating in the optimised method. Removal of the second DNase treatment and reducing the number of washes shortened the host DNA removal protocol from 90 min to 50 min, without affecting efficiency (Supplementary Table 1a). Additional time was saved by reducing the library preparation PCR extension time from six to four minutes. Comparison of the microbial community profile (organisms with ≥0.5% classified reads) between libraries produced with four and six minute extension times showed only minor differences in the abundance of minor members of the community and a small reduction in average read length for the *S. aureus* sample (<600bp) (Supplementary Table 1b). Altogether these changes reduced metagenomic library preparation to 2.5 hours with an overall turnaround time of less than four hours before DNA sequencing.

Limit of detection

The limit-of-detection (LoD) of the optimised method was determined using uninfected 'normal respiratory flora' (NRF) sputum samples (high and low commensal bacterial backgrounds in triplicate) spiked with serial ten-fold dilutions of *S. aureus* and *E. coli* cultures at known cell densities. Each replicate was defined as positive for the spiked 'pathogen' if present at ≥1% classified microbial reads (low quality read alignments with a WIMP assignment q-score <20 were removed from the analysis). The LoD (≥2/3 replicates positive) 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 (Supplementary Table 2a). The LoD was lower (10^3 *S. aureus* and *E. coli*) in sputum samples with a lower bacterial background (Supplementary Table 2b). Hence, the LoD of the method ranges from 10^3-10^5 CFU/ml, however, different levels of background commensal/human DNA could potentially result in different LoDs.

Mock community detection

Our optimised method was tested in triplicate on a panel of common respiratory pathogens spiked into an NRF sputum sample (~10^3-10^6 CFU/pathogen) to determine whether the saponin human DNA depletion method led to inadvertent loss of any bacterial DNA. We observed no bacterial DNA loss (average ΔCq <1) for any organisms (*E. coli, H. influenzae, K. pneumoniae, P. aeruginosa, S. aureus* and *S. maltophilia*) tested except *S. pneumoniae* where there was a 5.7-fold loss, (average ΔCq 2.52) between depleted and undepleted samples (Supplementary Table 3).

Optimised method testing

The optimised method was then tested on 41 respiratory samples from patients with suspected bacterial LRIs. A maximum of 10^4 fold depletion of human DNA (median 600-fold; interquartile range 168-1156 fold; maximum 18,054 fold) was observed between depleted and undepleted samples, as measured by qPCR (Table 2). The overall sensitivity of the optimised method for the detection of respiratory pathogens was 96.6% (95% CI, 80.4-99.8%) and specificity was 41.7% (95% CI, 16.5-71.4%), *not* counting additional organisms in culture-positive samples as false positives (Table 2). The turnaround time from sample to result was approx. 6 hours, including 2 hours MinION sequencing (Supplementary Table 4).

The pathogenic organism reported by routine microbiology was detected 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 2). Up to two potentially pathogenic bacteria were also observed in seven samples reported as NRF/no significant growth (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. Only one pathogenic organism reported by routine microbiology was not detected using the optimised method i.e. S9. This was reported as a mixed infection with *P. aeruginosa* and *E.*

coli, whereas only *E. coli* was detected by metagenomics. There were three other mixed infections reported by routine microbiology, S27, S38 and S41, and both organisms were detected in all three samples using the optimised method.

Confirmatory qPCR was used to establish the presence or absence of the missed/additional pathogens detected by metagenomics in 16 samples (1 sample with a missed pathogen, 15 samples with additional pathogen/s; total of 19 pathogens) and in matched controls i.e. an equal number of samples with no evidence of the pathogen by culture or metagenomics (Supplementary Table 5). This analysis was performed on DNA extracted from samples that did not undergo the depletion process, to rule out depletion as a potential cause of missed/additional pathogen detection. The majority of additional pathogens detected by metagenomics (12/19) were confirmed by qPCR, which increased the specificity of the optimised method to 50% (95% CI, 21.09-78.91% - *not* counting additional organisms in culture-positive samples as false positives (n=2, S5 positive for *K. pneumoniae,* likely *k-mer* mis-classification of *K. oxytoca.* S41 positive for *E. coli,* likely laboratory/kit contamination)). qPCR was negative for *P. aeruginosa* (S9) increasing the sensitivity to 100% (95% CI, 88.06-100%).

Species-specific gene analysis was performed on all samples positive for pathobionts (potentially pathogenic organisms which may reside as commensals in the lung), i.e. *H. influenzae* and *S. pneumoniae,* which can have closely related non-pathogenic species present in the lungs (18 samples containing 20 pathobionts). This confirmatory analysis was used to identify *k-mer* mis-classification of commensal reads as pathogen reads by WIMP. Samples containing >1 *H. influenzae* (*siaT*) or *S. pneumoniae* (*ply*) specific gene alignments were considered positive for that organism. The pathobiont-specific gene analysis confirmed the absence of *H. influenzae/S. pneumoniae* in 5/18 samples (also negative by qPCR - see previous paragraph) and resulted in metagenomics test sensitivity of 100% (95% CI, 88.06-

100%) and specificity of 100% (95% CI, 73.54-100%) compared to the culture+qPCR gold standard (Supplementary Table 6).

Antibiotic resistance

The samples tested using the optimised method had little antibiotic resistance, based upon routine testing (Supplementary Table 7). Across the 33 cultivated organisms, just 43 instances of resistance and intermediate resistance were recorded (Supplementary Table 7), with some of these likely reflecting single underlying mechanisms. Sequencing identified 183 resistance genes across the 41 specimens (with multiple inclusions when ARMA identified multiple variants of e.g. *bla*$_{\mathrm{TEM}}$).

Among the 183 resistance genes, 26 were inherent to the species cultivated (e.g. *oqxA/B* for *K. pneumoniae* or *bla*$_{\mathrm{OXA\text{-}50}}$ in *P. aeruginosa*), leaving 157, of which 24 matched the phenotype seen (Table 3). These comprised of *mecA* 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*$_{\mathrm{TEM}}$ variants spread recorded across two amoxicillin-resistant *E. coli* (S1 and S35 and two amoxicillin-resistant *H. influenzae* (S18 and S36). A caveat regarding this is that although ARMA flagged multiple *bla*$_{\mathrm{TEM}}$ genes, it did not flag *bla*$_{\mathrm{TEM\text{-}1}}$, which was the likeliest variant, given (i) that it is considerably the most prevalent type and (ii) that the isolates remained susceptible to oxyimino- cephalosporins whereas many of the variants flagged should encode extended-spectrum variants. Depending on their strength of expression *bla*$_{\mathrm{TEM}}$ or *bla*$_{\mathrm{OXY}}$ may have explained non-susceptibility to penicillin/β-lactamase inhibitor combinations in Enterobacteriales (4/183 genes), but expression is not quantified by ARMA. A *bla*$_{\mathrm{TEM4}}$ gene (1/183) was also found in a ceftazidime- and piperacillin/tazobactam- resistant *P. aeruginosa* (S37); this could explain the phenotype but is unlikely in this species, where β-lactam resistance most often reflects up-regulation of chromosomal *ampC* or efflux. There were

14/183 genes where any associated resistance could not be confirmed because no relevant drug(s) was tested by the clinical laboratory e.g. *tet* genes were identified in several samples (S2, S8, S9, S16, S30, S35, S38 and S39) but tetracycline was not tested against the isolates cultured. Sixteen genes detected by ARMA did not match the phenotype of isolates cultured, which remained susceptible to relevant antibiotics, and 42 genes were unlikely to be from species grown by the laboratory. Finally, multiple genes (56/183) likely originated from the normal flora: thus *tet(M)* and *bla*$_{\text{TEM-4,}}$ each was found in 8/12 NRF/NSG specimens whilst *mefA* and *mel* were each found in 9/12, as well as in many where the isolates grown were unlikely to have hosted these genes.

There were nine samples where phenotypic resistances remained unexplained by resistance genes found by ARMA. This included two amoxicillin-resistant *M. catarrhalis* (S8 and S26), where the BRO β-lactamase genes were likely to be responsible but were not represented in the ARMA database. The remaining seven 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 but lacking any acquired β-lactamase gene.

The specificity and sensitivity of the developed method for resistance gene detection was not determined as this would have required isolating and sequencing all bacteria (pathogens and commensals) present – a prohibitive task.

Reference-based genome assembly

Two samples containing antibiotic resistant bacteria were chosen as examples to generate reference-based genome assemblies directly from the metagenomic data. This analysis was performed to illustrate that whole pathogen genomes can be generated directly from respiratory samples for public health and infection control applications. Assemblies were generated for an MRSA (S16) and an *E. coli* resistant to amoxicillin, co-amoxiclav and co-trimoxazole (S1). The results were compared with those for undepleted controls after two and 48 hours of sequencing. Within the first two hours of sequencing the human DNA depleted MRSA sample had 47.9x genome coverage with an assembly of 28 contigs (GCA_900660255: longest contig = 478718 and N50=400kbp). Genome coverage increased to 228.7x after 48hrs of sequencing, with a final assembly consisting of 22 contigs (GCA_900660245: longest contig = 481kbp and N50=403kbp). In contrast, the undepleted MRSA sample had an assembly of 69 contigs with 3.9x coverage (GCA_900660235: longest contig = 47kbp and N50=146kbp) after 2hrs and 33 contigs (17.5x coverage) after 48 hours (GCA_900660205: longest contig = 416kbp and N50=263kbp) (Figure 2a).

For the sample positive for resistant *E. coli* there was 33.5x genome coverage within two hours for the depleted sample, with an assembly of 83 contigs (GCA_900660265: longest contig = 437kbp and N50=165kbp). Genome coverage increased to 165.7x after 48 hrs with the final *E. coli* assembly having 72 contigs (GCA_900660275: longest contig = 474kbp and N50=178kbp). The undepleted sample only produced 0.2x coverage after 2hrs, which increased to 1.1x after 48 hrs of sequencing (Figure 2b).

Time-point analysis

Using the same sample set as for genome assembly, data from the first two hours of sequencing were compared over time for depleted samples and undepleted controls to highlight the importance of host depletion for turnaround-time to result. Within 5 min of sequencing the depleted MRSA sample (S16) had 1.6x genome coverage compared with 0.2x coverage for the undepleted control (Figure 2c). The *mec*A gene was not detected in the undepleted sample after 5 min whereas two *mec*A gene alignments were detected in the depleted sample by the same time point (Figure 2d).

The depleted *E. coli* sample (S1) had 5.7x genome coverage within 20 min of sequencing compared to 0.06x for the undepleted control (Figure 2e). This *E. coli* was resistant to amoxicillin (*bla*$_{\mathrm{TEM}}$ gene), co-amoxiclav (possibly owing to *bla*$_{\mathrm{TEM}}$ if strongly expressed) and co-trimoxazole (*sul1* and *dfr*A17 genes). The *bla*$_{\mathrm{TEM}}$ and *dfr*A17 genes were not detected in the undepleted sample within two hours of sequencing and only one alignment was detected for *sul*1. Conversely, all three resistance genes were detected within 20 min of sequencing in the depleted sample and, after two hours, 47 *bla*$_{\mathrm{TEM}}$, 37 *sulf1* and 21 *dfrA17* alignments were detected (Figure 2f).

Discussion

Culture-based diagnostics and susceptibility testing, in use for 70 years [35], have limitations as guides for the appropriate clinical management of acute infections, mainly because of their slow sample-to-result turnaround. Rapid, accurate diagnostics would enable treatment with appropriate antibiotics and improve health outcomes and antimicrobial stewardship alike. We developed a method to prepare respiratory samples for metagenomics sequencing and incorporated it into a nanopore metagenomic sequencing protocol for bacterial pathogen and antibiotic resistance gene identification in LRIs within 6h of sample receipt.

Our metagenomics workflow for respiratory samples includes host DNA depletion, microbial DNA extraction, library preparation, MinION sequencing and real-time data analysis. A pipeline was developed (pilot method) and tested on 40 respiratory samples. We then optimised our method by shortening the depletion protocol, introducing bead-beating for improved microbial lysis, and reducing the library preparation time. Mock community analysis demonstrated that the saponin based human DNA depletion method didn’t inadvertently remove DNA from common respiratory pathogens, except for *S. pneumoniae* (mean 5.8 fold loss – Supplementary Table 3). It is possible that *S. pneumoniae* cells may have lysed during the host DNA depletion process [36] or might have lysed when grown to stationary phase for our mock community experiments. *S. pneumoniae* was correctly identified by metagenomics in five of six culture-positive patients, but it may have been underrepresented in these samples. The time from sample collection to bacterial DNA extraction may be crucial for accurate detection of *S. pneumoniae*.

The LoD of our optimised method (10^3-10^5 cfu/ml) is within the range of culture-based clinical thresholds applied to respiratory samples. Our optimized method was 96.6% sensitive and 41.7% specific compared to culture. Discordant results were investigated using pathogen specific probe-based qPCR assays (Supplementary Table 5) which increased sensitivity (100%) and specificity (50%). Five of seven remaining discordant samples were positive for pathobionts, specifically *H. influenzae* and/or *S. pneumoniae,* by metagenomics. These false positive detections can be caused by misclassification of reads by WIMP, as *k-mer* based read classification can be unreliable at the species level, particularly where species in a genus are highly related or share genes [37, 38]. To overcome this problem we introduced post-hoc pathobiont-specific gene analysis for all *H. influenzae* and/or *S. pneumoniae* positive samples (n=20 pathobionts in 18 samples). This analysis confirmed that the false positive results (n=5) were caused by *k-mer* misclassification and resulted in metagenomics test sensitivity and specificity of 100% compared to culture+qPCR gold standard. This issue highlights the need for new methods to accurately identify bacterial species from metagenomic data[39].

To maximise the impact on patient management, identification of clinically relevant antibiotic resistance genes as well as the infecting pathogen/s is necessary. In this regard the present pipeline has potential but requires refinement. Both MRSA cases were identified by the presence of *mecA*, with no false positives for this gene. Co-trimoxazole resistance in Enterobacteriaceae was accurately identified with detection of *sul* and *dfr* genes and these were not found in *H. influenzae*, for which resistance is largely mutational [40, 41]. However, genes such as *tet(M), mel, mefA* and *bla*$_{\mathrm{TEM}}$ were found in all samples where no pathogen was grown, suggesting presence in the normal or colonising respiratory flora. To overcome this issue, it will be necessary to associate resistance genes to particular organisms. This can be done by examining flanking sequences [42-45] in the c. 3 kb nanopore reads in cases where a gene is chromosomally inserted (not plasmid-borne resistance genes), as is usual for transposon borne *tet(M)* and *mefA* in streptococci [46-48], including *S. pneumoniae* (Supplementary Figure 1).

Clinical metagenomics data could also be used to assemble pathogen genomes for reference laboratory typing. The quality/depth of the metagenomic data generated by our method could enable monitoring of emergence and patient-to-patient spread of pathogens and antimicrobial resistance directly from clinical samples in real-time [49, 50]. Using PCR for respiratory infection diagnosis must be coupled with microbiological culture, otherwise the link to phenotype is lost, whereas clinical metagenomics could replace routine culture entirely. As viruses are an important cause of LRIs, they can be tested for using PCR, as is current routine practice, or our pipeline could be modified to detect viral nucleic acid by processing the supernatant fraction after centrifugation of the respiratory sample (Figure 1, step 1).

In conclusion, we report the first rapid clinical metagenomics pipeline for the characterization of bacterial LRIs. Pathogens and antibiotic resistance genes can be identified in six hours. With additional sequencing time (up to 48 hrs), it provides sufficient data for public health and infection control applications. Our protocol is being evaluated in a clinical trial (INHALE - http://www.ucl.ac.uk/news/news-articles/1115/181115-molecular-diagnosis-pneumonia) to evaluate the rapid diagnosis of hospital-acquired and ventilator-associated pneumonia in comparison with culture and multiplex-PCR.

Acknowledgements

This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Reference Number RP-PG-0514-20018, JOG, HR), the UK Antimicrobial Resistance Cross Council Initiative (MR/N013956/1, JOG, GLK), Rosetrees Trust (A749, JOG), the University of East Anglia (JOG, TC), Oxford Nanopore Technologies (JOG, TC, AA, DJT), Quadram Institute Bioscience BBSRC Strategic Programme: Microbes in the Food Chain (project number BB/R012504/1, JOG, GLK) and BBSRC grants (BB/N023196/1 and BB/CSP17270/1, RML).

Part of the bioinformatics analysis was run on CLIMB-computing servers, an infrastructure supported by a grant from the UK Medical Research Council (MR/L015080/1).

Author Contributions

The study was devised by JOG, JW, DJT. Laboratory work and data analysis was performed by TC, GLK, AA, HR, RB, DML, RML, JOG. Clinical samples were collected and analysed by CJ, SG, DR. All authors contributed to writing and reviewing the manuscript.

Competing Financial Interests Statement

JOG, RML, GLK and TC received financial support for attending ONT and other conferences and/or an honorarium for speaking at ONT headquarters. JOG, AA and TC received funding and consumable support from ONT for PhD studentships. DJT is a full-time employee and share-option holder of Oxford Nanopore Technologies Ltd. RML and JOG received free flowcells as part of the MAP and MARC programmes.

References

1. NICE. in NICE clinical guideline 69 (Centre for Clinical Practice 2008).
2. Chalmers, J. et al. Community-acquired pneumonia in the United Kingdom: a call to action. Pneumonia 9, 15 (2017).
3. Enne, V.I., Personne, Y., Grgic, L., Gant, V. & Zumla, A. Aetiology of hospital-acquired pneumonia and trends in antimicrobial resistance. Current Opinion in Pulmonary Medicine 20, 252-258 (2014).
4. Carroll, K.C. Laboratory Diagnosis of Lower Respiratory Tract Infections: Controversy and Conundrums. Journal of Clinical Microbiology 40, 3115-3120 (2002).
5. Kollef, M.H. Microbiological Diagnosis of Ventilator-associated Pneumonia. American Journal of Respiratory and Critical Care Medicine 173, 1182-1184 (2006).
6. Moran, G.J., Rothman, R.E. & Volturo, G.A. Emergency management of community-acquired bacterial pneumonia: What is new since the 2007 Infectious Diseases Society of America/American Thoracic Society guidelines. American Journal of Emergency Medicine 31, 602-612 (2013).
7. Garcin, F. et al. Non-adherence to guidelines: an avoidable cause of failure of empirical antimicrobial therapy in the presence of difficult-to-treat bacteria. Intensive Care Medicine 36, 75-82 (2010).
8. Lim, W.S. et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. Thorax 64, iii1 (2009).
9. Burnham, C.A. & Carroll, K.C. Diagnosis of Clostridium difficile infection: an ongoing conundrum for clinicians and for clinical laboratories. Clinical microbiology reviews 26, 604-630 (2013).

10. Lees, E.A., Miyajima, F., Pirmohamed, M. & Carrol, E.D. The role of Clostridium difficile in the paediatric and neonatal gut — a narrative review. European Journal of Clinical Microbiology & Infectious Diseases 35, 1047-1057 (2016).
11. Cookson, W.O.C.M., Cox, M.J. & Moffatt, M.F. New opportunities for managing acute and chronic lung infections. Nature Reviews Microbiology 16, 111 (2017).
12. Davies, S.C. Chapter 1 Chief Medical Officer's summary. Annual Report of the Chief Medical Officer (2016).
13. Goverment, H. (2019).
14. O'Neill, J. Tackling drug-resistant infections globally: final report and recommendations. 84 (2016).
15. 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. 8, 15479-15488 (2015).
16. Hassibi, A. et al. Multiplexed identification, quantification and genotyping of infectious agents using a semiconductor biochip. Nature Biotechnology (2018).
17. Kais, M., Spindler, C., Kalin, M., Örtqvist, Å. & Giske, C.G. Quantitative detection of Streptococcus pneumoniae, Haemophilus influenzae, and Moraxella catarrhalis in lower respiratory tract samples by real-time PCR. Diagnostic Microbiology and Infectious Disease 55, 169-178 (2006).
18. Kodani, M. et al. Application of TaqMan Low-Density Arrays for Simultaneous Detection of Multiple Respiratory Pathogens. Journal of Clinical Microbiology 49, 2175-2182 (2011).
19. Hayon, J.A.N. 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 165, 41-46 (2002).
20. Buchan, B.W. & Ledeboer, N.A. Emerging Technologies for the Clinical Microbiology Laboratory. Clinical microbiology reviews 27, 783 (2014).
21. Huang, T.-D. et al. Analytical validation of a novel high multiplexing real-time PCR array for the identification of key pathogens causative of bacterial ventilator-associated pneumonia and their associated resistance genes. Journal of Antimicrobial Chemotherapy 68, 340-347 (2012).
22. Chiu, C.Y. & Miller, S.A. Clinical metagenomics. Nature Reviews Genetics (2019).
23. Loman, N.J. et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology 30, 434 (2012).
24. Strauch, B. et al. LiveKraken—real-time metagenomic classification of illumina data. Bioinformatics 34, 3750-3752 (2018).
25. Faria, N.R. et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature 546, 406 (2017).
26. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228 (2016).
27. Greninger, A.L. et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Medicine 7, 99 (2015).
28. Schmidt, K. et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. Journal of Antimicrobial Chemotherapy 72, 104-114 (2017).
29. Pendleton, K.M. et al. Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. American Journal of Respiratory and Critical Care Medicine 196, 1610-1612 (2017).
30. Feehery, G.R. et al. A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA. PLOS ONE 8, e76096 (2013).
31. Hasan, M.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 54, 919-927 (2016).
32. Marotz, C.A. et al. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 6, 42 (2018).

33. Zelenin, S. et al. Microfluidic-based isolation of bacteria from whole blood for sepsis diagnostics. Biotechnology Letters 37, 825-830 (2015).
34. Couto, N. et al. Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Scientific Reports 8, 13767 (2018).
35. McIntosh, J. Emergency Pathology Service. The Lancet 247, 669-670 (1946).
36. Martner, A., Dahlgren, C., Paton, J.C. & Wold, A.E. Pneumolysin Released during Streptococcus pneumoniae Autolysis Is a Potent Activator of Intracellular Oxygen Radical Production in Neutrophils. Infection and Immunity 76, 4079-4087 (2008).
37. Chen, J.H.K. 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).
38. Kutlu, S.S., Sacar, S., Cevahir, N. & Turgut, H. Community-acquired Streptococcus mitis meningitis: a case report. International Journal of Infectious Diseases 12, e107-e109 (2008).
39. Langelier, C. 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 115, E12353 (2018).
40. Eliopoulos, G.M. & Huovinen, P. Resistance to Trimethoprim-Sulfamethoxazole. Clinical Infectious Diseases 32, 1608-1614 (2001).
41. Enne, V.I., King, A., Livermore, D.M. & Hall, L.M.C. Sulfonamide Resistance in Haemophilus influenzae Mediated by Acquisition of sul2 or a Short Insertion in Chromosomal folP. Antimicrobial Agents and Chemotherapy 46, 1934-1939 (2002).
42. Ashton, P.M. et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nature Biotechnology 33, 296 (2014).
43. Orlek, A. et al. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. Frontiers in Microbiology 8 (2017).
44. Xia, Y. et al. MinION Nanopore Sequencing Enables Correlation between Resistome Phenotype and Genotype of Coliform Bacteria in Municipal Sewage. Frontiers in Microbiology 8 (2017).
45. Leggett, R.M. et al. Rapid MinION metagenomic profiling of the preterm infant gut microbiota to aid in pathogen diagnostics. bioRxiv (2017).
46. Roberts, A.P. & Mullany, P. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS Microbiology Reviews 35, 856-871 (2011).
47. Santoro, F., Vianna, M.E. & Roberts, A.P. Variation on a theme; an overview of the Tn916/Tn1545 family of mobile genetic elements in the oral and nasopharyngeal streptococci. Frontiers in Microbiology 5 (2014).
48. Tantivitayakul, P., Lapirattanakul, J., Vichayanrat, T. & Muadchiengka, T. Antibiotic Resistance Patterns and Related Mobile Genetic Elements of Pneumococci and β-Hemolytic Streptococci in Thai Healthy Children. Indian Journal of Microbiology 56, 417-425 (2016).
49. Deurenberg, R.H. et al. Application of next generation sequencing in clinical microbiology and infection prevention. Journal of Biotechnology 243, 16-24 (2017).
50. Greninger, A.L. et al. Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. Journal of Clinical Microbiology 55, 177-182 (2017).

Figure legends

Figure 1: Schematic representation of the metagenomic pipeline with a turnaround time of approx. six hours (optimised) and approx. eight hours (pilot) from sample collection to sample result.

Figure 2: Bacterial genome assembly, genome coverage and antibiotic gene detection with depleted versus undepleted samples.

A: MRSA after 48 hours of sequencing.

B: *E. coli* after 48 hours of sequencing.

C: MRSA genome coverage of depleted versus undepleted during two hours of sequencing*.

D: *mecA* gene alignment of depleted versus undepleted during two hours of sequencing*.

E: *E. coli* genome coverage of depleted versus undepleted during two hours of sequencing*.

F: *bla*$_{\text{TEM}}$, *sul1* and *dfr*A17 gene alignment of depleted versus undepleted during two hours of sequencing*.

*Three independent clinical samples were analysed (an example of a Gram positive and a Gram negative are respresented).

Table1: 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

Table 2. Human and bacterial DNA qPCR results for sputum 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 (~10^4)	15.60	0.13
				Depleted	35.00		15.73	
S2	Sputum	*K. pneumoniae*	*K. pneumoniae*	Undepleted	23.73	9.99 (~10^3)	15.63	0.02
				Depleted	33.71		15.65	
S3	Sputum	*P. aeruginosa*	*P. aeruginosa*	Undepleted	23.05	9.29 (~10^3)	15.46	1.48
				Depleted	32.34		13.98	
S4	Sputum	*S. marcescens*	*S. marcescens*	Undepleted	26.34	9.93 (~10^3)	16.96	0.52
				Depleted	36.27		17.48	
S5	Sputum	*K. oxytoca*	*K. oxytoca*	Undepleted	22.96	8.58 (~10^3)	12.67	0.64
			K. pneumoniae	Depleted	31.54		12.03	
S6	Sputum	*S. aureus*	*S. aureus*	Undepleted	22.31	9.41 (~10^3)	19.11	1.57
				Depleted	31.72		17.54	
S7	Sputum	*H. influenzae*	*H. influenzae*	Undepleted	25.47	9.53 (~10^3)	21.44	0.43
			P. aeruginosa	Depleted	35.00		21.87	
S8	Sputum	*M. catarrhalis*	*M. catarrhalis*	Undepleted	22.72	9.17 (~10^3)	16.9	0.66
			S. pneumoniae	Depleted	31.89		17.56	
S9	Sputum	*P. aeruginosa* & *E. coli*		Undepleted	23.89	11.11 (~10^4)	19.58	3.26
			E. coli	Depleted	35		22.84	
S10	Sputum	NSG	*H. influenzae*	Undepleted	23.46	8.6 (~10^3)	14.12	2.39
			S. pneumoniae	Depleted	32.06		16.51	
S11	Sputum	NRF	*S. pneumoniae*	Undepleted	25.77	9.23 (~10^3)	17.96	1.92
				Depleted	35.00		19.88	
S12	Sputum	NRF	*H. influenzae*	Undepleted	22.5	8.92 (~10^3)	17.61	0.05
			M. catarrhalis	Depleted	31.42		17.56	

S13	Sputum	*S. marcescens*	*S. marcescens*	Undepleted	22.48	7.11 (~10^2)	12.77	0.79
				Depleted	29.59		11.98	
S14	Sputum	*S. aureus*	*S. aureus*	Undepleted	23.17	7.68 (~10^2)	13.83	0.96
			M. catarrhalis	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. pneumoniae	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	19.81	2.04

			S. aureus	Depleted	29.13	(~10^2)	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. aeruginosa*	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*	Undepleted	20.86	14.14 (~10^4)	16.71	6.85
			S. aureus	Depleted	35.00		23.56	

Table 3. Resistance genes found by ARMA in relation to pathogens grown: Optimised pipeline (41 samples; 183 genes detected)

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	

ONLINE METHODS

Ethics

This study used excess respiratory samples, after routine microbiology diagnostic tests had been performed, from patients with suspected LRIs such as persistent (productive) cough, bronchiectasis, CAP/HAP, cystic fibrosis and exacerbation of chronic obstructive pulmonary disease (COPD, emphysema/chronic bronchitis). The UCL Infection DNA Bank (REC reference 12/LO/1089) approved use of excess respiratory samples for the study. No patient identifiable information was collected, hence informed consent was not required. The only data collected were routine microbiology results, which detailed the pathogen(s) identified and their antibiotic susceptibility profiles.

Definitions

'Respiratory pathogens' or 'pathogens' are defined in this study as common causes of respiratory infection, in order to differentiate them from commensal organisms. Respiratory pathogens identified in this study were: *E. aerogenes, E. cloacae* complex, *E. coli, H. influenzae, K. oxytoca, K. pneumoniae, M. catarrhalis, P. mirabilis, P. aeruginosa, S. marcescens, S. aureus, S. pneumoniae, S. pyogenes*. A list of all microorganisms identified in all samples tested using the optimised method (above our thresholds) are listed in supplementary table 8. Some of these organisms, not defined as common pathogens here, could be considered pathogens in some clinical contexts.

Routine clinical microbiological investigation

Respiratory samples including sputum, endotracheal secretions and ETAs were treated with sputasol (Oxoid-SR0233) in a 1:1 ratio before being incubated for a minimum of 15 min at 37 °C. Sputasol-treated respiratory samples (10 µl) were inoculated into 5 ml of sterile water and mixed (hence the limit of detection of culture is 10^5 CFU/ml). Following this, 10 µl of sample was streaked onto blood, chocolate and cysteine lactose electrolyte deficient (CLED) agar. BAL samples were not treated with sputasol; instead they were centrifuged to concentrate bacterial cells for a minimum of 10 min at 3000 rpm. BALs did not undergo further dilution and were streaked directly onto the agar plate. Depending on clinical details and the source of the specimen, other agar plates (including sabouraud, mannitol salt and *Burkholderia cepacia* selective agar) were additionally used.

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 [51].

Sample collection and storage

The excess respiratory samples (sputa, ETA, BAL) were collected after culture and susceptibility testing at Norfolk and Norwich University Hospitals (NNUH) Microbiology Department (described above) and stored at 4 °C prior to testing. They were indicated by clinical microbiology to contain bacterial pathogen(s), NRF or to have yielded NSG. Forty samples (n=34 positive and n=6 NRF samples, comprising 34 sputa, four BALs and two ETAs) were used to test the Pilot method and another 41 (n=29 suspected LRI, n=9 NRF and n=3 NSG samples, comprising 38 sputa, one BAL and two ETAs) were used to test the Optimised pipeline.

Pilot method: Host DNA Depletion

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. The saponin-based differential lysis method was modified from previously reported saponin methods [33, 52]. Saponin (Tokyo Chemical Industry- S0019) was added to a final concentration of 2.5 % (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 next was 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.

Pilot method: Bacterial Lysis and DNA Extraction

After the final wash step of the host depletion, the pellet was resuspended in 380 µl of bacterial lysis buffer (Roche UK- 4659180001) and 20 µl of proteinase K (>600mAu/ml) (Qiagen -19133) was added before incubation at 65 °C for 10 min with shaking at 800 RPM (on an Eppendorf Thermomixer). Nucleic acid was then extracted from samples 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).

Optimised method: Host DNA Depletion (Figure 1)

The optimized method sought to improve and shorten some steps. Specifically, after the first 5 min centrifugation at 8000 x g, up to 50 µl of supernatant was left so as not to disturb the pellet (final saponin conc. 2.2-2.5%). Instead of performing two rounds of host DNA digestion, the amount of HL-SAN DNase was increased up to 10 µl and a single incubation of 15 min at 37 °C was carried out with shaking at 800 RPM on an Eppendorf Thermomixer.

Finally, the number of washes was reduced to two with increasing volumes of PBS (800 µl and 1 ml).

Optimised method: Bacterial Lysis and DNA Extraction (Figure 1)

After the final wash, the pellet was re-suspended in 500 µl of bacterial lysis buffer (Roche UK - 4659180001), transferred to a bead-beating tube (Lysis Matrix E, MP Biomedicals - 116914050) and bead-beaten at maximum speed (50 oscillations per second) for 3 min in a Tissue Lyser bead-beater (Qiagen - 69980). This ensured the release of DNA from difficult-to-lyse organisms (e.g. *S. aureus*). The sample was centrifuged at 20,000 xg for 1 min and ~230 µl of supernatant was transferred to a fresh Eppendorf tube. The volume was topped-up with 170 µl of bacterial lysis buffer and 20 µl of proteinase K (>600 mAu/ml, Qiagen - 19133) was added. Samples were then incubated at 65 °C for 5 min with shaking at 800 RPM on an Eppendorf Thermomixer. DNA was extracted from samples 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).

DNA quantification and quality control

DNA quantification was performed using the high sensitivity dsDNA assay kit (Thermo Fisher - Q32851) on the Qubit 3.0 Fluorometer (Thermo Fisher - Q33226). DNA quality and fragment size (PCR products and MinION libraries) 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).

MinION Library Preparation and Sequencing

MinION library preparation was performed according to the manufacturer's instructions for (i) the Rapid Low-Input by PCR Sequencing Kit (SQK-RLI001), (ii) the Rapid Low-Input

Barcoding Kit (SQK-RLB001) or (iii) the Rapid PCR Barcoding Kit (SQK-RPB004) with minor alterations as follows. For single sample sequencing runs using the SQK-RLI001 kit, 10 ng of the MagNA Pure-extracted DNA were used for the tagmentation/fragmentation reaction, where DNA was incubated at 30 °C for 1 min and at 75 °C for 1 min. The PCR reaction was run as per the manufacturer's instructions; however, the number of PCR cycles was increased to 20. For multiplexed runs, SQK-RLB001 and SQK-RPB004 kits were used. A 1.2x AMPure XP bead (Beckman Coulter-A63881) wash was introduced after the MagNA Pure DNA extraction and prior to library preparation for multiplexed runs and DNA was eluted in 15 µl of nuclease-free water. Modifications for the library preparation were i) 10 ng of input DNA and 2.5 µl of FRM were used for the tagmentation/fragmentation reaction and nuclease-free water was used to make the volume up to 10 µl, ii) for the PCR reaction, 25 cycles were used and the reaction volume was doubled. All samples run using the Pilot method used a 6 min extension time, whereas the Optimised method used a reduced extension time of 4 min. When multiplexing, PCR products were pooled together in equal concentrations, then subjected to a 0.6x AMPure XP bead wash and eluted in 14 µl of the buffer recommended in the manufacturer's instructions (10 µL 50 mM NaCl, 10 mM Tris.HCl pH8.0). Sequencing was performed on the MinION platform using R9.4, R9.5 or R9.4.1 flow cells. The library (50-300 fmol) was loaded onto the flow cell according to the manufacturer's instructions. ONT MinKNOW software (versions 1.4-1.13.1) was used to collect raw sequencing data and ONT Albacore (versions 1.2.2-2.1.10) was used for local base-calling of the raw data after sequencing runs were completed. The MinION was run for up to 48 hours with WIMP/ARMA analysis performed on the first six folders (~24,000 reads) for Pilot method samples and the first two hours of data for all Optimised method samples.

Quantitative PCR (qPCR) assays

Probe or SYBR Green based qPCR was performed on samples to detect and quantify human DNA, DNA targets for specific pathogens (*E. coli*, *H. influenzae, K. pneumoniae, M. catarrhalis, P. aeruginosa, S. aureus, Stenotrophomonas maltophilia*, *S. pneumoniae* and *S.*

pyogenes) and the bacterial 16S rRNA V3-V4 gene fragment. All qPCR assays were performed on a Light Cycler® 480 Instrument (Roche). Details of primer sequences and targets can be found in Supplementary Table 9 (oligonucleotides were supplied by Sigma.

For all probe-based qPCR reactions, the master mix consisted of 10 µl LightCycler 480 probe master (2X), 0.5 µl each of reverse and forward primer (final conc. 0.25 µM) and 0.4 µl probe (final conc. 0.2 µM). For all SYBR-Green-based qPCR reactions, the master mix consisted of 10 µl LightCycler 480 SYBR Green I master (2x) and 1 µl of each forward and reverse primer (final conc. 0.5 µM). To the PCR mix, 2 µl of DNA template and nuclease-free water to a total volume of 20 µl were added. The qPCR conditions were: pre-incubation at 95 °C for 5 min, amplification for 40 cycles at 95 °C for 30 sec, 55 °C for 30 sec and 72 °C for 30 sec, with a final extension at 72 °C for 5 min. Melt curves analysis (for SYBR-Green qPCR) was performed at 95 °C for 5 sec, 65 °C for 1 min, ramping to 95 °C at 0.03 °C/s in continuous acquisition mode, followed by cooling to 37 °C. All probe-based confirmatory qPCR used the following conditions: pre-incubation at 95 °C for 15 min, amplification for 40 cycles at 94 °C for 15 sec and 60 °C for 1 min.

Example Limit of detection

The LoD of the Optimised method was determined for the detection of one Gram-positive and one Gram-negative bacteria in sputum using serial dilutions (10 –10^5 cfu/ml) of cultured *E. coli* (H141480453) and *S. aureus* (NCTC 6571) spiked into NRF sputum samples with high and low bacterial commensal backgrounds (as determined by 16S qPCR). The serial dilutions were made in sterile PBS and plated in triplicate on LB agar to determine colony forming units (CFU) per ml. The same dilutions were used to spike an NRF sputum sample for LoD experiments. Detection and quantification of bacterial DNA was performed using probe-based qPCR assays and MinION sequencing.

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* and *P. aeruginosa. E. coli* and *S. aureus* strains were also included (H141480453 and NCTC 6571 respectively). Pathogens (*E. coli* and *S. aureus* in 10 ml Luria-Broth and *K. pneumoniae*, *P. aeruginosa* and *S. maltophilia* in 10 ml Tryptic Soy Broth (TSB)) were cultured overnight at 37 °C with shaking at 180 RPM. *H. influenzae* (in 10 ml TSB) and *S. pneumoniae* (in 10 ml Brain Heart Infusion Broth) were cultured statically at 37 °C with 5% CO_2 in an aerobic incubator. Cultured pathogens were then spiked into an NRF sample (~10^3-10^6 CFU/pathogen). The spiked samples were then tested in triplicate with the Optimised method, to determine if saponin depletion resulted in any inadvertent lysis of pathogens and loss of their DNA. All spiked samples were processed alongside undepleted controls. Probe or SYBR Green-based qPCR assays were used to determine the relative quantity of each spiked pathogen in depleted and undepleted spiked sputum samples.

Human read removal

Human reads were removed from basecalled FASTQ files using minimap2 to align to the human hg38 genome (GCA_000001405.15 “soft-masked” assembly) prior to Epi2ME analysis. Only unassigned reads were exported to a bam file using Samtools (-f 4 parameter). Non-human reads were converted back to FASTQ format using bam2fastx. These FASTQ files were processed for pathogen identification using WIMP and antibiotic resistance gene detection with ARMA. Further downstream analysis for genome coverage was performed using minimap2 with default parameters for long-read data (-a -x map-ont) and visualised using qualimap (used for time-point analysis).

Pathogen identification and antibiotic resistance gene detection

The EPI2ME Antimicrobial Resistance pipeline (ONT, versions 2.59.1896509) was used for initial analysis of MinION data for the identification of bacteria present in the sample and any associated antimicrobial resistance genes. Within this pipeline, WIMP (What's in my Pot – rev. 3.3.1) supports the identification of bacteria, viruses, fungi, archaea and human reads and was used for respiratory pathogen identification. WIMP utilises 'Centrifuge', a *k-mer*-based read identification tool based on a Burrows-Wheeler transform and the Ferragina-Manzini index, to identify reads using the RefSeq database [53]. ARMA (Antimicrobial Resistance Mapping Application – rev. 1.1.5) is also included in the Antimicrobial Resistance pipeline. ARMA utilises the CARD database for antibiotic resistance gene detection and identification by aligning input reads using minimap2 (alignments reported at >75% accuracy and >40% horizontal coverage [54]). Full manuals are publicly available for WIMP and ARMA on the ONT website (https://nanoporetech.com/EPI2ME-amr). NanoOK/NanoOK RT [45, 55] are publically available tools which identify microbes and antimicrobial resistance using basecalled nanopore data, providing similar outputs to those from ONTs WIMP and ARMA software.

Initial analysis of respiratory metagenomic data revealed that thresholds would be required to improve the accuracy of results. Thresholds, in terms of number of bacteria per ml of body fluid, are applied in clinical microbiology laboratories for some infections including those of the urinary and respiratory tracts. The same approach was required for metagenomics. The clinical thresholds used for respiratory samples is typically 10^5 pathogens/ml (range 10^3-10^5/ml dependent on sample type) and is achieved by sample dilution [51]. We routinely applied thresholds at ≥1% of classified reads, with a WIMP assignment q-score ≥20 (within .csv files). We chose these thresholds to: censor reads arising from pipeline contaminants; remove barcode leakage between samples on multiplexed runs (ONT's Flongle (https://nanoporetech.com/products/comparison), an adapter for single use flowcells designed for diagnostic applications, should overcome this issue) and; remove low quality WIMP alignments, which result in misclassified reads. Antibiotic resistance genes were reported if >1 gene alignment was present using the 'clinically relevant' parameter within

ARMA. This parameter currently reports resistance genes, acquired and chromosomal, but not resistance mutations/SNPs.

Pathobiont-specific gene analysis

Species-specific gene alignments were performed on samples positive for *H. influenzae* or *S. pneumoniae* by metagenomics (above our thresholds). Reads (after human DNA removal) were aligned to pathobiont-specific genes (*siaT, ply* – chosen from a literature search for species-specific genes in *H. influenzae*[56] and *S. pneumoniae*[15]*,* respectively) using minimap2 with default parameters for long-read data (-a -x map-ont) and the number of mapped reads visualised using qualimap. If a sample contained >1 copy of the specific gene it was considered positive for the species.

Bacterial genome assembly

Genome assembly was performed first using Fast5-to-Fastq to remove reads shorter than 2000 bp and with a mean quality score lower than seven (https://github.com/rrwick/Fast5-to-Fastq). Porechop was used to remove sequencing adapters in the middle and/or the ends of each read, and re-identification of barcodes was carried out for each multiplexed sample (v0.2.3) (https://github.com/rrwick/Porechop). Filtered reads were aligned to a reference genome (chosen based on WIMP classification of pathogen reads) using minimap2 with default parameters for ONT long-read data (v2.6-2.10) [57]. Finally, Canu was used to assemble mapped reads into contigs using this long-read sequence correction and assembly tool (v1.6) [58, 59]. BLAST Ring Image Generator (BRIG) was used for BLAST comparisons of the genome assemblies generated [60].

Data availability

All clinical sample sequence data and assemblies are available via European Nucleotide Achive (ENA) under study accession number PRJEB30781.

References

51. Services, M. UK Standards for Microbiology Investigations. Investigation of bronchoalveolar lavage, sputum and associated specimens. Bacteriology B57, 38 (2018).

52. Anscombe, C., Misra, R.V. & Gharbia, S. Whole genome amplification and sequencing of low cell numbers directly from a bacteria spiked blood model. bioRxiv (2018).

53. Kim, D., Song, L., Breitwieser, F.P. & Salzberg, S.L. Centrifuge: rapid and sensitive classification of metagenomic sequences. bioRxiv (2016).

54. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research 45, D566-D573 (2017).

55. Leggett, R.M., Heavens, D., Caccamo, M., Clark, M.D. & Davey, R.P. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. Bioinformatics 32, 142-144 (2015).

56. Price, E.P. et al. Simultaneous identification of Haemophilus influenzae and Haemophilus haemolyticus using real-time PCR. Future microbiology 12, 585-593 (2017).

57. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, bty191-bty191 (2018).

58. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R. & Phillippy, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv (2016).

59. Koren, S. et al. Complete assembly of parental haplotypes with trio binning. bioRxiv (2018).

60. Alikhan, N.-F., Petty, N.K., Ben Zakour, N.L. & Beatson, S.A. BLAST Ring Image

Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12, 402 (2011).