Taylor & Francis
Taylor & Francis Group

Future Microbiology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ifmb20

Bacterial discrimination by Fourier transform infrared spectroscopy, MALDI-mass spectrometry and whole-genome sequencing

Rachel McGalliard, Howbeer Muhamadali, Najla AlMasoud, Sam Haldenby, Valeria Romero-Soriano, Ellie Allman, Yun Xu, Adam P Roberts, Steve Paterson, Enitan D Carrol & Royston Goodacre

To cite this article: Rachel McGalliard, Howbeer Muhamadali, Najla AlMasoud, Sam Haldenby, Valeria Romero-Soriano, Ellie Allman, Yun Xu, Adam P Roberts, Steve Paterson, Enitan D Carrol & Royston Goodacre (2024) Bacterial discrimination by Fourier transform infrared spectroscopy, MALDI-mass spectrometry and whole-genome sequencing, Future Microbiology, 19:9, 795-810, DOI: 10.2217/fmb-2024-0043

To link to this article: https://doi.org/10.2217/fmb-2024-0043

© 2024 The Authors

View supplementary material

Published online: 29 May 2024.

Submit your article to this journal

Article views: 680

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ifmb20

Research Article

Future MICROBIOLOGY

Bacterial discrimination by Fourier transform infrared spectroscopy, MALDI-mass spectrometry and whole-genome sequencing

Rachel McGalliard*[,‡,1,7], Howbeer Muhamadali[‡,2,3], Najla AlMasoud[4], Sam Haldenby[5], Valeria Romero-Soriano[5], Ellie Allman[6], Yun Xu[2,3], Adam P Roberts[6], Steve Paterson[5], Enitan D Carrol[§,1,7] & Royston Goodacre[§,2,3]

[1]Department of Clinical Infection, Microbiology & Immunology, University of Liverpool Institute of Infection, Veterinary & Ecological Sciences, Ronald Ross Building, 8 West Derby Street, Liverpool, UK
[2]School of Chemistry, Manchester Institute of Biotechnology, University of Manchester, Manchester, UK
[3]center for Metabolomics Research, Department of Biochemistry, Cell & Systems Biology, Institute of Systems, Molecular & Integrative Biology, University of Liverpool, Biosciences Building, Crown Street, Liverpool, UK
[4]College of Science, Princess Nourah Bint Abdulrahman University, Department of Chemistry, Riyadh, 11671, Saudi Arabia
[5]center for Genomic Research, University of Liverpool, Mersey Bio Building, Crown Street, Liverpool, UK
[6]Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool, UK
[7]Department of Infectious Diseases, Alder Hey Children's NHS Foundation Trust, Eaton Road, Liverpool, UK
*Author for correspondence: rmcg@liverpool.ac.uk
‡Authors contributed equally
§Authors contributed equally

Aim: Proof-of-concept study, highlighting the clinical diagnostic ability of FT-IR compared with MALDI-TOF MS, combined with WGS. **Materials & methods:** 104 pathogenic isolates of *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Streptococcus pyogenes* and *Staphylococcus aureus* were analyzed. **Results:** Overall prediction accuracy was 99.6% in FT-IR and 95.8% in MALDI-TOF-MS. Analysis of *N. meningitidis* serogroups was superior in FT-IR compared with MALDI-TOF-MS. Phylogenetic relationship of *S. pyogenes* was similar by FT-IR and WGS, but not *S. aureus* or *S. pneumoniae*. Clinical severity was associated with the zinc ABC transporter and DNA repair genes in *S. pneumoniae* and cell wall proteins (biofilm formation, antibiotic and complement permeability) in *S. aureus* via WGS. **Conclusion:** FT-IR warrants further clinical evaluation as a promising diagnostic tool.

Plain language summary: We tested a technique (FT-IR) to identify four different, common bacteria from 104 children with serious infections and compared it to lab methods for diagnosis. FT-IR was more accurate. We tested if it could identify subtypes of bacteria, which is important in outbreaks. It was able to subtype two species, but not the two other species. However, it is a much faster and cheaper technique than the gold standard. It may be useful in certain outbreaks. We also investigated the trends between genes and the length of hospital stay. This can support further laboratory research. As a fast, low-cost test, FT-IR warrants further testing before it is applied to clinical labs.

First draft submitted: 13 February 2024; Accepted for publication: 21 March 2024; Published online: 23 April 2024

Keywords: diagnostics • Fourier transformed infrared (FT-IR) spectroscopy • matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) • pediatric pathogens • serious bacterial infections (SBI) • whole-genome sequencing (WGS)

Despite advances in modern medicine, invasive bacterial infections remain a cause of significant morbidity and mortality [1–3]. Traditional techniques, such as phenotypic and serological tests on cultured organisms, are still predominantly used in clinical settings for microbiological identification. Culturing methods, which is considered the gold standard diagnostic technique has low sensitivity [4,5], is time consuming (usually between 24 and 72 h depending on the growth rate of the micro-organism), and laborious. This introduces delay in clinical decision-making, and limits the ability to feedback information on pathogen identification and resistance profiles. Molecular

Taylor & Francis
Taylor & Francis Group

biology approaches such as polymerase chain reactions (PCR) of 16S ribosomal DNA or pathogen-specific PCR are much more rapid than traditional culture, but are less widely available, more expensive, and require specialised instruments while cannot provide full antibiotic susceptibility profiles [6].

Bacterial subtyping allows the determination of clonal and phylogenetic relatedness of bacterial strains. Subtyping is necessary for outbreak investigations and targeting control measures in real time but its implementation to clinical environments is limited due to the following factors [7,8]. Current methods are expensive, time consuming and technically demanding, with whole-genome sequencing (WGS) being the gold standard; allowing for the detection of virulence factors, antimicrobial resistance genes, and single nucleotide polymorphism (SNP) or extrachromosomal element analysis [9,10]. A rapid, user-friendly and low-cost typing method that allows real time typing or complements WGS would reduce turnaround times in clinical practice.

Molecular fingerprinting approaches through spectroscopy-based techniques are a growing area of interest [11,12]. Fourier transform infrared (FT-IR) spectroscopy, a vibrational spectroscopic technique, is a low cost, rapid (<60 s per sample) and high-throughput fingerprinting technique which is widely used within life sciences including food safety [13–17], environmental [18–20] and biotechnology [21–24] industries. Consumables are relatively low cost and reusable. The spectrum from the specimen is obtained by the absorption of infrared light by various molecular bonds resulting in different vibrational modes, which reflects the overall (bio-) chemical composition of a sample, while multivariate analysis approaches are employed to discriminate between the samples, and identify the molecular species contributing to such discrimination [25]. Nonetheless, this technique does have some limitations, most notably strong water absorbance, which can be superseded by drying the sample, or subtracting the water signal. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a widely used tool in bacterial identification [26,27]. Through detecting the mass-to-charge ratio (m/z) of the various abundant proteins and peptides in a whole cell sample and providing spectra, this is compared with an existing reference database of organisms' under the same conditions to type the isolate [28,29]. MALDI-TOF MS has also been shown to detect antibiotic resistance promptly in certain bacterial species [30,31]. However, while MALDI-TOF MS is generally favored toward the detection of peptides and proteins, FT-IR analyses is considered a more wholistic approach providing insight into the metabolic fingerprint of the organism, and allowing for the detection of a wider range of biochemical diversity within the cells [32].

In this study, we compared FT-IR and MALDI-TOF MS spectral fingerprints, combined with multivariate statistical analysis and WGS approaches for the accurate classification of four species across 104 clinical isolates causing invasive disease in children in critical care. We also compared these data with clinical metadata outcomes. This is a proof of concept study to apply FT-IR diagnostics to clinical pathogenic samples, predominantly Gram positives, rather than ecological Gram negatives or lab strains.

Materials & methods

Sample collection

The isolates were obtained from children admitted to Alder Hey Children's Hospital, as part of a larger multi-center study of children with life-threatening bacterial infections, details of which have been published elsewhere [3]. Written informed consent was gained during the study period to use microbiology samples for future research (REC reference: 11/LO/1982). Age, weight, length of stay, need for ventilation and PICU admission were available in all patients. Clinical data (observation recordings and blood results) were available in 28 patients, those admitted to PICU, and were used to calculate the pediatric Index of Mortality 2 (PIM2) and pediatric Risk of Mortality (PRISM) severity scores according to published literature [33,34]. A total of 104 bacterial samples from community-acquired pathogenic isolates of blood culture (69), joint aspirate (15) cerebrospinal fluid (9), pleural (3) and purulent samples collected from an aseptically-obtained surgical specimen (11), were isolated and identified using analytical profile identification (API) techniques (bioMérieux). Individual colonies were isolated and frozen on Protect Multipurpose beads (Thermo Fisher Scientific Inc. UK) at -80 °C.

Bacterial growth conditions

All chemicals and reagents were purchased from Sigma Aldrich (Sigma Aldrich, UK) unless otherwise stated. All microorganisms in this study were sub-cultured three-times on blood agar medium, to ensure purity and phenotypic stability, before being used for inoculum preparation. *Neisseria meninigitidis*, *Streptococcus pneumoniae*, and *Streptococcus pyogenes* (group A *Streptococcus*) were inoculated onto a defibrinated horse blood (Fisher Scientific UK LTD) agar plate and incubated at 37 °C in 5% CO_2 conditions for 24 h. *Staphylococcus aureus* was inoculated

Taylor & Francis
Taylor & Francis Group

onto horse blood agar plates and incubated at 37 °C for 24 h. Three biological replicates of all individual isolates were cultured.

FT-IR analysis

Biomass from the overnight grown samples was harvested and centrifuged at 4000 × *g* for 10 min at 4 °C using a Sigma 1–16PK microcentrifuge. The supernatant was discarded and the biomass washed twice using sterile physiological saline solution (0.9% NaCl) to remove any residual compounds from the medium, as previously reported [22,35]. All washed samples were resuspended in saline solution and normalized according to their OD at 600 nm [19].

Samples were spotted as 20 µl aliquots onto FT-IR silicon 96-well plates and heated to dryness at 55 °C for ~30 min. Due to the large number of samples (n = 318; 104 bacteria grown in triplicate – biological replicates – plus QCs (see below)) analyzed in this study, the bacterial suspensions were spotted onto six separate FT-IR silicon plates and analyzed on three different days. An *S. aureus* sample with the highest biomass yield was selected and spotted on the last column of every FT-IR plate (Supplementary Figure 1, yellow spots) to be used as a quality control (QC) sample, to account for any day-to-day instrumental variation and alignment of all the data combined. All FT-IR spectral data were collected in absorbance mode in the Mid-IR range (4000–600 cm^{-1}) on a Bruker Equinox 55 infrared spectrometer (Bruker Optics Ltd, Coventry, UK), as 64 co-adds with 4 cm^{-1} resolution, in triplicate (machine replicates) from separate regions of each of the sample spots [36]. FT-IR spectral data were scaled by applying the extended multiplicative signal correction (EMSC) algorithm [37,38], followed by replacement of the CO_2 peaks with a trend (2400–2275 cm^{-1}) [24].

MALDI-TOF-MS analysis

Aliquots (40 µl) of the normalized bacterial cell pellets were re-suspended in 180 µl of 0.1% trifluoroacetic acid (TFA). Sinapinic acid (10 mg) was dissolved in 500 µl of 2% TFA and 500 µl of acetonitrile, this was followed by mixing an equal volume 10 µl of matrix and bacterial strains. This mixture was then vortexed for 3 s, then 2 µl was spotted onto a stainless steel MALDI plate and then allowed to be air dried for 60 min at room temperature. The bacterial strains were analyzed using an AXIMA-Confidence MALDI-TOF-MS (Shimadzu Biotech, Manchester, UK), equipped with a nitrogen-pulsed UV laser at a wavelength of 337 nm. The laser was set at 135 mV laser power with 80 acquired profiles and each profile containing 20 shots, linear TOF mode and positive ionization mode was used. The mass-to-charge (m/z) ranged between 1000 and 14,000. The spectra were collected using a circular raster pattern. The MALDI-TOF-MS device was calibrated using a protein mixture purchased from (Sigma-Aldrich). Three biological replicates were analyzed for each strain and three analytical replicates were performed for each strain.

whole-genome sequencing analysis

Genomic DNA for *S. aureus* was purified using the Puregene DNA purification kit (Qiagen, Crawley, UK) according to the manufacturer's instruction. Genomic DNA for *S. pneumoniae* and *S. pyogenes* were extracted according to Promega Wizard Genomic DNA Purification kit (Promega UK) with the following modifications; mucoid *S. pneumoniae* were washed twice, then all streptococci were resuspended in 100 µl TE buffer (Invitrogen; Thermo Fisher Scientific) and incubated with 50 µl of 3000 U/ml mutanolysin (Sigma Aldrich) in TE buffer and 150 µl of 20 mg/ml lysozyme (Sigma Aldrich) in TE buffer at 37 °C for 60 min.

After DNA extraction, samples were cleaned up using AMPure beads (1.8×). A total of 50 ng of DNA per sample was used to start library preparation, following the NEBNext FS DNA Library Prep Kit protocol with Inputs <100 ng, in half volume reactions (fragmentation time 12 min, 1:10 adaptor dilution, seven PCR cycles). Size distribution of the final libraries was assessed using the fragment analyzer and equimolar pooling was performed. After quantification by qPCR, sequencing was performed on one lane of a NovaSeq S1 flow cell (2 × 150 bp read configuration).

Genotypic data for *N. meningitidis* were exported from the Meningitis Research Foundation Meningococcus Genome Library which contains draft genomes for all English, Welsh and Northern Irish invasive disease isolates received by the Public Health England Meningococcal Reference Unit (PHE MRU) since July 2010 [39]. Quality and adapter-filtered reads for each sample were assembled using SPAdes version 3.15.3 [40] for *S. aureus*, *S. pneumoniae* and *S. pyogenes*. Assemblies were filtered to only include those that (1) have an expected size of +/-30% of the expected genome size (2) 90%+ genome completeness and lower than 10% duplication levels, according to BUSCO

v 4.1.4 [41] and (3) fewer than 10% of reads are assigned to taxa other than expected genus, using MetaPhlAn version 2.9.21 [42]. Multi-locus Sequence Typing profiles and allele sequences were obtained from pubmlst.org and allele sequences were aligned to assemblies using Bowtie2 version 2.3.5.1 [43]. The allele that aligned best for each locus was selected and the sequence type was determined by comparing detected alleles against the database profiles. Genes were predicted using PROKKA version 1.14.6 [44]. Predicted genes were used to reconstruct the pan- and core genome across samples, using Panaroo version 1.2.3 [45]. Phylogenetic estimation was carried out using the core genome sequences generated by Panaroo as input for IQ-TREE version 2.0.3 [46], with 1000 bootstrap replicates using the GTR model. Antimicrobial resistance genes were identified by interrogating genome assemblies with RGI version 5.1.0 [47]. Prokka-predicted genes were translated to protein sequences and used as queries for BLAST searches (version 2.9) [48] against the Virulence Factor Database (VFDB) [49], and BLAST score ratios (BSRs) were calculated for each hit by dividing the hit score by the maximum possible score. SNPs and small indels were detected using Snippy version 4.6.0, with reads being aligned against the NGAS638 (*S. pyo*), Hu17 (*S. pne*) or NCTC8325 (*S. aur*) reference genome. Unitigs, which are non-overlapping unique sequences, were utilized as units of genomic variation in association analyses. By leveraging the unitig-counter tool [50], unitigs were gathered across the genome assemblies. This method captures both the presence or absence of unique genomic regions (including parts of the accessory genome) and distinct SNP variations within each sample set. Subsequently, associations between genomic variations and metadata were identified with Pyseer v1.3.10 [51], using linear mixed models (lmm). This enabled a view of how both gene content and SNP variations related to metadata. To control for population structure, a similarity matrix was derived from a midpoint-rooted core gene phylogeny and provided to Pyseer. The thresholding method provided by the count_patterns.py script from Pyseer, was used to determine an appropriate significance threshold based on the data, to correct for multiple testing. Significant variants were annotated with gene information using the Pyseer accessory script, annotate_hits_pyseer. Based on annotated gene names, genes with exact matching names were identified among samples and extracted from the pan-genome. Nucleotide and translated multiple alignments were carried out on these using Clustal Omega, including the unitig sequence [52]. Variants at the unitig locus were derived from the alignments using a custom Python script.

Statistical analysis

All collected data were analyzed using MATLAB version 2016 (The Mathworks Inc., Natwick, USA). All pre-processed FT-IR spectral data were subjected to principal component analysis (PCA) [53], to reduce the dimensionality of the data, followed by discriminant function analysis (PC-DFA), which uses *a priori* knowledge of the experimental class structure to reduce within-class variance while increasing between-class variance. The class structure here were defined based on each isolate as a separate class and not the bacterial species as a whole, comprising of 104 classes in total, and thus the results are considered semi-supervised. Finally, partial least squares-discriminant analysis (PLS-DA) [54] coupled with bootstrapping validation [55] was employed to generate a classification model using the FT-IR spectral data.

For the available metadata in 28 subjects, the variables with >30% missing values were also removed and the remaining missing values were imputed using K-NN imputation algorithm. O2-PLS model was built to find correlation between FT-IR data and the corresponding meta data. Through joint loadings, the variables in the metadata which are most likely to be correlated with FT-IR data were selected and PLS-R models were built to assess the statistical significance in the correlation between the two types of data. The PLS-R models were validated by using double cross-validation coupled with 1000 permutation test.

Results

Overview of diagnostic techniques

All pathogenic isolates from four species were included in this study (n = 104). Notably, 69 isolates were obtained from blood cultures including six central line associated bacterial infections, 15 from joint aspirations, nine from cerebrospinal fluid, four were surgical wounds and seven from other sites. The results of all 104 samples, illustrating the information collected from the two different analytical techniques applied in this study are displayed in Supplementary Figure 2. The FT-IR spectra (Supplementary Figure 2A) display typical vibrational features found in most bacterial samples, hence are qualitatively very similar but there are subtle quantitative differences. Hence, multivariate statistical approaches are important to interrogate these differences to allow for the differentiation of the bacterial species. PCA scores plot of all the FT-IR spectral data (Supplementary Figure 3A) displayed three separate clusters for the QCs, according to the day the analysis was carried out. Therefore, the QC spectral data was

Taylor & Francis
Taylor & Francis Group

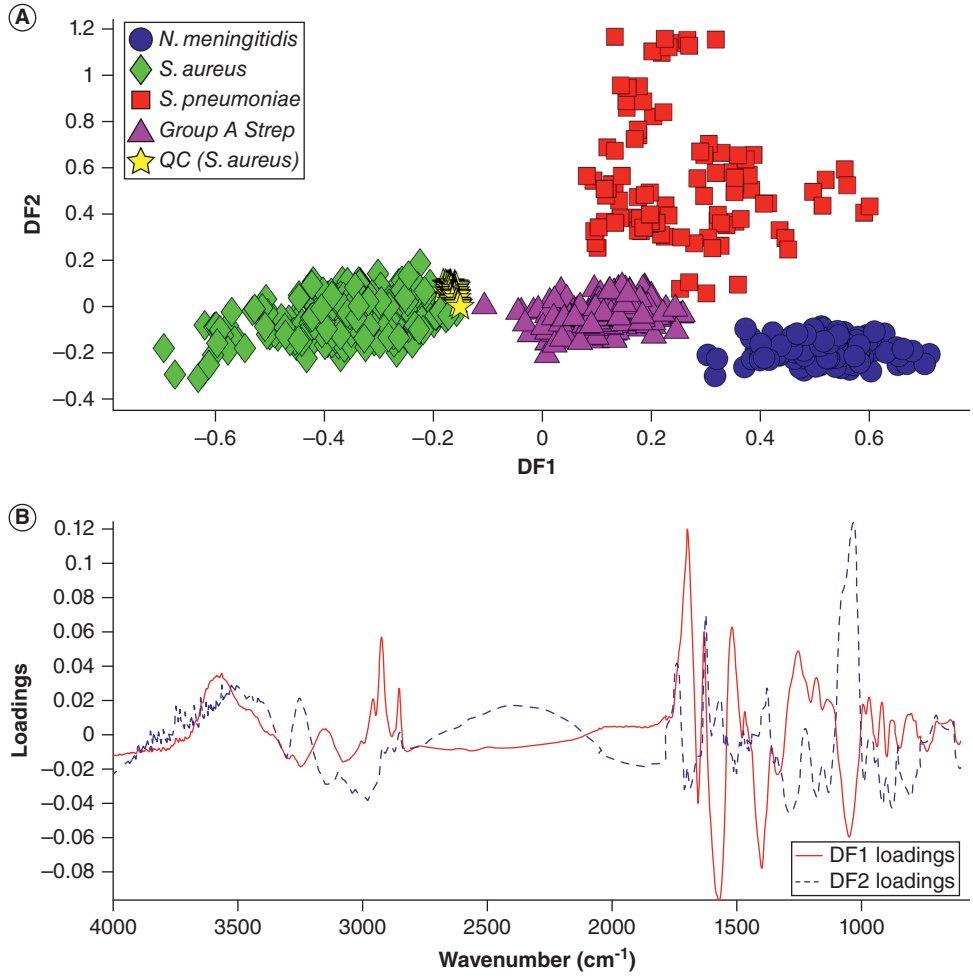

Figure 1. PC-DFA scores plot of all the combined FT-IR spectral data. PC-DFA scores plot of all the FT-IR spectral data combined, using the first five PCs accounting for 91.1% of the TEV **(A)**. DF1 and DF2 loadings plot of the FT-IR data, displaying the most significant vibrational bands contributing to the clustering pattern above **(B)**.

used to align the whole dataset and remove any contributing instrumental variation, that may have been present over the 3-day period that these data were collected. The PCA scores plot of the aligned data (Supplementary Figure 3B) clearly demonstrated the successful alignment of the data and removal of instrumental variation. Although according to the PCA scores plot of the aligned data (Supplementary Figure 3B), *S. aureus* samples are separated from all other samples according to PC1 axis with a total explained variance (TEV) of 40.5%, all other isolates are inseparable and are clustered together. To discriminate between these isolates further PCA was followed by discriminant function analysis (PC-DFA) using the *a priori* knowledge of the experimental class structure. However, as each of the isolates were assigned to an independent class (104 classes in total), it can be said that PC DFA was carried out in a semi-supervised manner, and that the resulting clustering in PC-DFA represents the natural variation among these bacteria.

The PC-DFA scores plot of all FT-IR spectral data (Figure 1A), displayed clear separation of *S. aureus* isolates from all other species according to DF1 axis, while *N. meningitidis* clustered on the opposite side of the DF1 axis, and the *S. pneumoniae* and group A *Streptococcus* isolates clustered closer together in the middle of the plot. It is also worth noting that the QC samples which are of *S. aureus* origin, despite being assigned an independent class, have also clustered with all other *S. aureus* isolates on the negative side of DF1 axis (Figure 1B), which is as expected. According to the PC-DFA loadings plots (Figure 1B), the main vibrational bands contributing to the separation of the isolates are assigned to fatty acids at 2961 and 2924 cm^{-1} (CH_3 and CH_2 asymmetric stretching, respectively), esters at 1740 cm^{-1} (C=O, stretching), amide I at 1655 cm^{-1} (C=O), amide II at 1545 cm^{-1} (combination of C-N

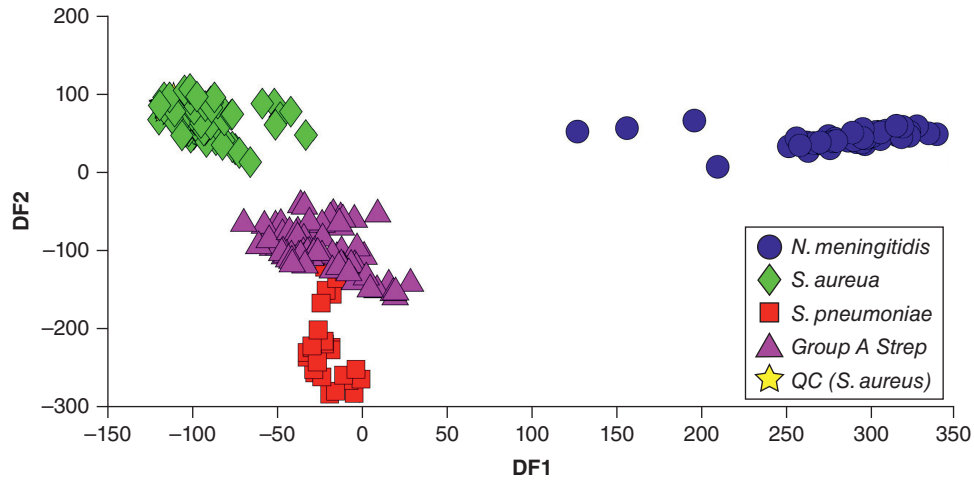

Figure 2. PC-DFA scores plot of the MALDI-TOF-MS spectral data of all the isolates (104 in total) combined, using the first 15 PCs accounting for 87.7% of the TEV. QC clustered by arrow, behind *S. aureus* (obscured).

Table 1. The prediction accuracy of the bacterial species using FT-IR spectral data.

Parameter	*N. meningitidis*	*S. aureus*	*S. pneumoniae*	*S. pyogenes*
N. meningitidis	98.6%	0.0%	0.0%	1.4%
S. aureus	0.0%	100.0%	0.0%	0.0%
S. pneumoniae	1.5%	0.0%	98.3%	0.2%
S. pyogenes	0.0%	0.0%	0.1%	99.7%
FT-IR overall	99.6%			

Table 2. The prediction accuracy of the bacterial species using MALDI-TOF-MS spectral data.

Parameter	*N. meningitidis*	*S. aureus*	*S. pneumoniae*	*S. pyogenes*
N. meningitidis	99.5%	0.1%	0.1%	0.3%
S. aureus	0.6%	97.1%	1.8%	0.5%
S. pneumoniae	1.6%	4.0%	84.4%	10.0%
S. pyogenes	0.6%	2.9%	0.9%	95.6%
MALDI-TOF MS overall	95.8%			

stretching and N-H bending), and other vibrational bands in the fingerprint region [21]. The PC-DFA scores plot of the MALDI-TOF-MS data collected from all the samples (Figure 2), displayed very similar clustering pattern, where *S. aureus* and *N. meningitidis* were completely separated according to DF1 axis, and the *S. pneumoniae* and group A *Streptococcus* isolates were clustered in the middle of the plot. The QC samples in this case, were also clustering with other *S. aureus* isolates, which is in agreement with the FT-IR findings.

In order to generate classification models, so as to identify these bacteria from their spectral fingerprints PLS-DA was used and 1000 test sets generated by bootstrapping were generated and the overall results of these models are presented in Tables 1 & 2 for FT-IR and MALDI-TOF-MS data respectively. High accuracies (>95% overall) in predictions were obtained on both data sets while FT-IR data appeared to have slightly better discriminant power compared with that of MALDI-TOF-MS.

Differentiation according to serotypes & strains

Differentiation according to serotypes and strains was performed and compared with diagnostic techniques. Comparing PHE MRU sequencing data for *Neisseria* samples, PC DFA scores plot of both FT-IR and MALDI-TOF-MS spectral data (Figure 3) displayed clear separation of a non-meningitidis species (yellow star) from all other *Neisseria* isolates. However, the PC-DFA scores plot of the FT-IR spectral data (Figure 3A), also displayed the separation of *N. meningitidis* serotype W135 from all other *Neisseria* isolates according to DF2 axis and clustering of

Taylor & Francis
Taylor & Francis Group

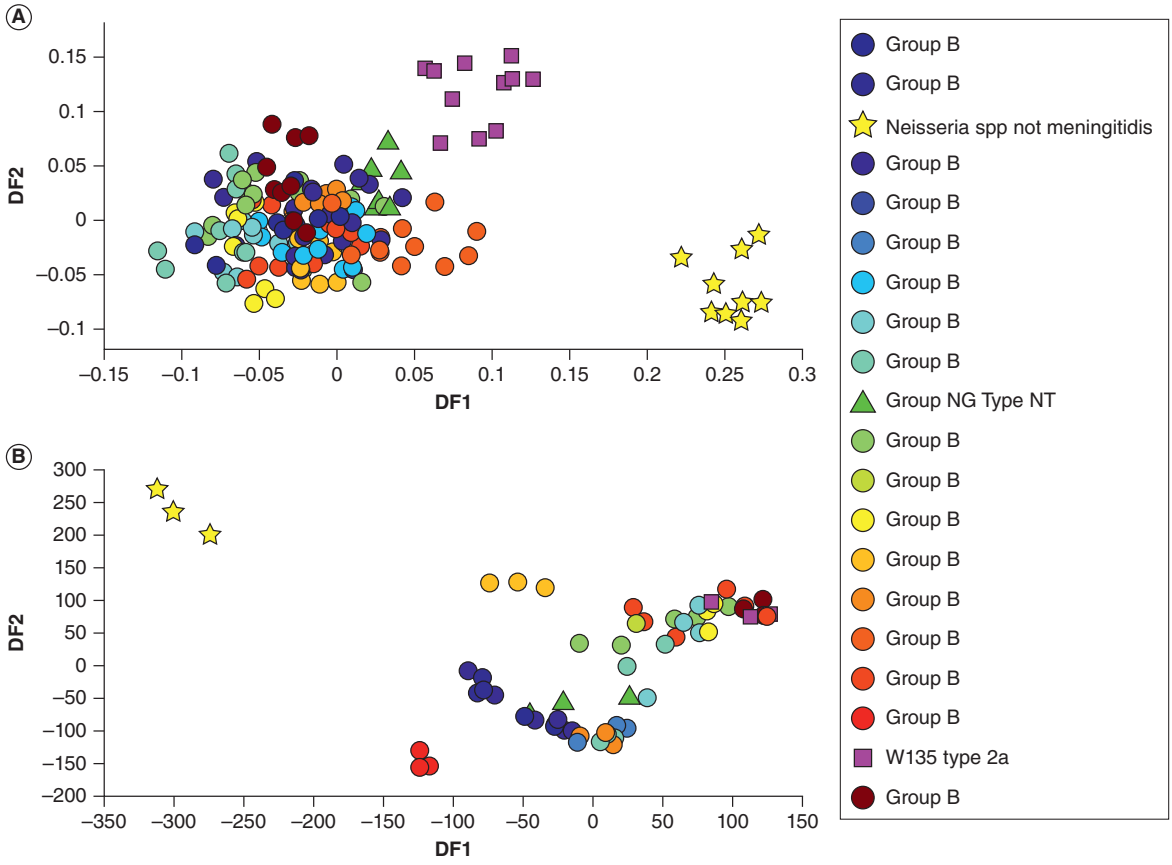

Figure 3. PC-DFA scores plot of the *Neisseria* spp. only. Using the FT-IR **(A)** and MALDI-TOF-MS **(B)** spectral data.

N. meningitidis group NG type NT. To investigate this potential for subtyping further, WGS was undertaken in the remaining three species. Phylogenetic relatedness derived from the core genome of each species was compared with the dendrograms created based on FT-IR spectra. Group A *Streptococcus* dendrograms were very similar between FT-IR and WGS (Supplementary Figure 4C). Although *S. pneumoniae* showed similar clustering of isolates with both methods, branching within these dendrograms were not congruent (Supplementary Figure 4B) and *S. aureus* showed no correlation between WGS and FT-IR phylogenies (Supplementary Figure 4A). FT-IR clustered *N. meningitidis* and group A *Streptococcus* in a similar manner to WGS.

Antimicrobial resistance & virulence factors

From the patient samples, there were 25 AMR genes identified for *S. aureus*; four Group A *Streptococcus*; seven *S. pneumoniae*. There were 78 virulence factors genes (VFs) identified for *S. aureus*; 39 Group A *Streptococcus*; 30 *S. pneumoniae*. Heatmaps were generated for each species for AMRs (Supplementary Figure 5) and VFs (Supplementary Figure 6) alongside clinical outcomes.

Gene variant association analysis

Association analyses were carried out to identify links between sequence variants (unitigs) and corresponding metadata. For *S. aureus*, 48 variants were found to be significantly associated with length of hospital stay, after controlling for multiple comparisons. Details on those described here are presented in Supplementary Table 1. Among these variants, some genes encoded known VFs (FnbA/B) or cell wall proteins (Ebh, SraP) while others were not characterized or found in intergenic regions. A significant unitig that conferred non-synonymous SNPs (G445E, G446D, T447A, T448S relative to the unitig) among strains in *mutL*, a mismatch repair gene, was associated with increased length of stay. SraP is a surface-exposed serine-rich repeat glycoprotein that is required for the pathogenesis of human infective endocarditis via its ligand-binding region adhering to human platelets [56]. A total of 12 strains lacking one of the significant unitigs associated with SraP contain a L1524S amino acid

substitution, relative to those that contain it. Gene *ebh* encodes a cell wall protein that affects flucloxacillin and complement mediated killing and several non-synonymous substitutions and frameshift variants were associated with reduced duration of hospitalization [57]. FnbA and FnbB are cell wall proteins that are implicated in *S. aureus* biofilm formation and unitigs conferring non-synonymous variants in these genes were associated with increased length of stay [58].

For *S. pneumoniae*, 120 variants were found to be significantly associated with the need for non-invasive ventilation or PICU length of stay, and details on those described here are presented in Supplementary Table 2. In contrast with the *S. aureus* variants, these were typically more conservative, conferring fewer amino acid substitutions across strains. Among these variants, some genes are reported to be involved in DNA repair mechanisms such as *dnaE2* (DNA polymerase), *uvrC* (excision repair), *recG* (dsDNA translocase) and *dnaJ* (heat shock protein 70). SNPs in *dnaE2* and *uvrC* were associated with lower non-invasive ventilation days and a variant in recG was associated with length of non-invasive ventilation days (Figure 4A & B). While none of the *uvrC* variants were predicted to impact on protein coding, non-synonymous variants were identified in *dnaE2* and *recG*. Additionally, *adcA* encodes a zinc ABC transporter substrate-binding lipoprotein and truncation of this product is linked to hyperencapsulation and resistance to complement killing [59]. Variants in these gene (coding sequence A9454G and A962G) were significantly associated with ventilation days and PICU length of stay. Absence of *adcC*, another part of the zinc transporter, was also associated with significantly higher non invasive ventilation days. (Figure 4C–E). Unlike *S. aureus*, the *S. pneumoniae* core gene phylogeny does not suggest that samples containing significant unitigs are closely related.

No significant associations were identified for *S. pyogenes*.

Severity analysis

PLS-R modeling on age, sex, weight and length of stay in hospital, which was available for all samples, did not show significant correlation with FT-IR data. Laboratory data were available in 28 patients, those admitted to PICU, and PIM2 and PRISM severity scores were derived. These isolates with severity metadata were correlated with the corresponding FT-IR data using O2-PLS model and joint loading plots. The significant variables were selected from the joint loading plot (Supplementary Figure 7) which included PRISM total score, partial pressure of oxygen (PaO_2), plasma potassium and lactate levels. Furthermore, age and weight had poor correlation with FT-IR data. PLS-R models were created to assess the statistical significance in the correlation between the two types of data for the four selected significant variables (Figure 5). This was validated using double cross-validation coupled with 1000 permutation tests. All variables studied showed statistically significant correlation with FT-IR but PRISM scores showed the strongest correlation with FT-IR data (Figure 5C), especially up to PRISM score values of 10, also demonstrated by Q^2Y, which indicates the predictability of the model of 0.40. Notably the two severity scores used, PIM2 and PRISM had poor correlation, with a Spearman ranked correlation co-efficient 0.17. Similarly, PRISM scores correlated well with the FT-IR data, however PIM2 scores did not. We explored if this was confounded by the species of bacteria but both severity scores showed similar trends of patients with *S. aureus* having high scores, conversely *S. pneumoniae* having low scores (Supplementary Figure 8). This suggests that the observed correlation between FT-IR and PRISM scores is unlikely to be related to bacterial species. High (>7) and low (≤ 7) PRISM scores, according to bioinformatics cut off, were overlaid on PC-DFA plots of FT-IR spectrum and bacterial species (Supplementary Figure 9). This demonstrates clear species separation alongside trends in severity scores, such that invasive *S. aureus* had predominantly high PRISM scores.

Discussion

This proof of concept analysis of invasive, clinical, pathogenic bacterial strains shows distinct clustering of the investigated isolates down to species level using both MALDI-TOF-MS and FT-IR techniques. This study employed a novel approach that combines the use of FT-IR with both MALDI-TOF as a diagnostic tool in benchmarking, and in the comparison with WGS, which serves as the gold standard, as the ground truth of species and subtypes, but not clinically applicable due to high cost and longer turn around times and complex analysis.

FT-IR requires minimal sample processing and spectra can be obtained within minutes and semi-supervised techniques allow rapid throughput analysis of the spectra [60–63]. In comparison with MALDI-TOF-MS, which requires some sample processing and reagent costs (e.g., mixing with a UV absorbing matrix), FT-IR had higher overall accuracy in microbiological identification of 99.6% *versus* MALDI-TOF-MS accuracy of 95.8%. Although both techniques have been available over the last 25 years, MALDI-TOF-MS is now used in many hospital

Taylor & Francis
Taylor & Francis Group

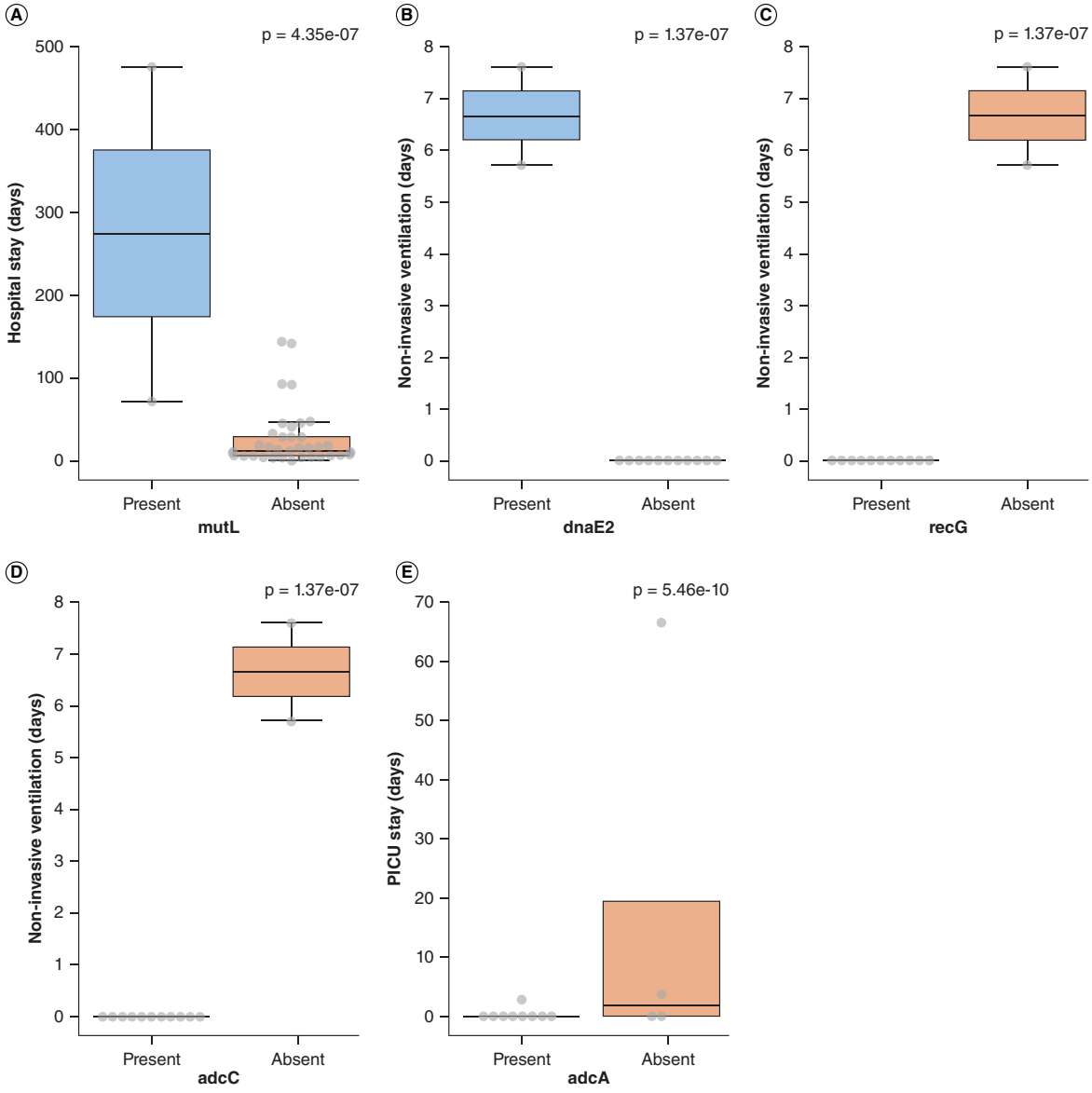

Figure 4. Boxplots of *S. aureus* (A) and *S. pneumoniae* (B–E) showing the presence or absence of significant variants (unitigs) with clinical outcomes on y axis and p-values. *S. aureus* plots for *ebh*, *fnaB/A* and *sraP* were identical to figure **(A)** for *mutL* with length of stay on y-axis. *S. pneumoniae* plots **(B)** *dnaE2* (DNA polymerase) with non invasive ventilation days **(C)** *recG* (dsDNA translocase) **(D)** *adcC* (zinc transporter); and *adcA* (zinc transporter) **(E)** with PICU stay. Details on variant specifics are present in Supplementary Tables 1 & 2.

microbiology laboratories whereas FT-IR has not been as extensively investigated in clinical care [26,27,29]. Wenning *et al.* compared MALDI-TOF MS and FT-IR spectroscopy to differentiate and to identify 93 species of food related bacteria. They found that MALDI-TOF had better species identification, which we did not find in our study, but that FT-IR had higher sensitivity to allow typing of *E. coli*, unlike MALDI-TOF MS [13]. Work by Dinkelacker and colleagues also found that FT-IR performed better than MALDI-TOF MS in phylogenetic identification of *Klebsiella* species compared with whole-genome sequencing but is considerably faster and more affordable which may allow real time rather than retrospective surveillance [64]. Such findings have been the driving force behind the development of specialized instrumentation, such as the IR Biotyper by Bruker with particular focus on industrial and medical applications. Semi-supervised analysis methods can be included within the FT-IR analyzer to provide results without specialized external analysis. Our study suggests that FT-IR has potential for species identification of bacterial pathogens in clinical laboratories.

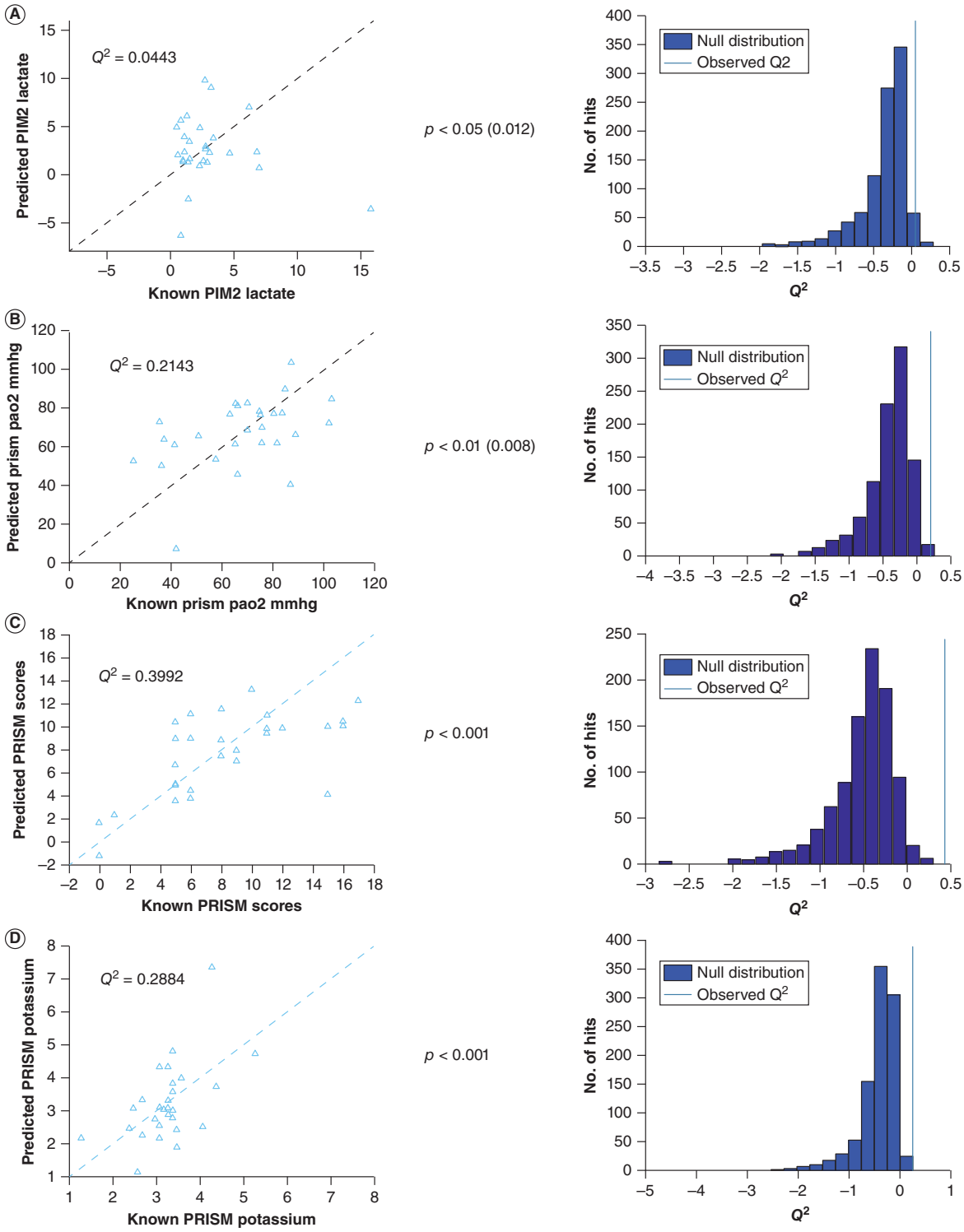

Figure 5. PLS-R model for the four most significant metadata variables correlated with FT-IR, showing correlation and p-values. (A) Lactate; **(B)** PaO_2; **(C)** PRISM total score; **(D)** potassium.

Taylor & Francis
Taylor & Francis Group

The potential of FT-IR spectroscopy to differentiate below the species level has been found in environmental studies including *L. monocytogenes*, *E. coli*, *S. enterica* and *Y. entercolitica* [65–68]. The FT-IR spectra of *Neisseria* and Group A *Streptococcus* in this study, showed high resolution into subgroups that corresponded with WGS phylogenetic analysis. This may provide more rapid and low-cost alternatives to WGS in group A *Streptococcus* outbreaks [69]. However, this is not the case for *S. pneumoniae* and *S. aureus*, potentially due to these species having less conserved core-genomes and FT-IR spectra being based on the 'fingerprint' from the whole organism including capsule. However, this is concordant with the existing literature regarding variable typing potential of FT-IR based on bacterial species, especially *S. aureus* [70–72].

Despite the relatively small sample size in the association analysis, there was evidence of a broad phylogenetic tree of disparate ancestry, rather than clonal strains in the pathogenic isolates of *S. aureus*, Group A *Streptococcus* and *S. pneumoniae* causing serious bacterial infections and multiple AMR and VF genes were identified. Notably fibronectin-binding proteins involved in *S. aureus* biofilm formation (FnbB/A) and cell wall proteins that allow binding to platelet (SraP) or inhibit antibiotic and complement directed killing (Ebh) that had previously been elucidated *in vitro* were associated with non invasive ventilation days, as potential mechanisms *in vivo* [56–58]. While these unitigs were identified as significant and of potential interest, the predominant findings were observed in two closely related strains. Even though the analysis was adjusted for population structure, the specific observations combined with the limited sample size warrants caution during interpretation, nonetheless these isolates were identified from different patients, at different times, with different co-morbidities; hence despite different host factors and immune response, these isolates caused extensive morbidity. Conversely genes in *S. pneumoniae* involved in DNA repair (*uvrC*, *dnaE2*, *recG*, *dnaJ*) were significantly associated with clinical severity rather than cell wall proteins, as expected with an encapsulated pathogen. Several genes involved in the zinc ABC transporter in *S. pneumoniae*, were associated with increasing disease severity and length of stay, which has been previously described *in vitro* through a hyperencapsulated phenotype [59].

Furthermore, this study is the first, to the authors' knowledge to compare microbiology metabolomic fingerprint data with clinical parameter metadata, which shows particular promise with PRISM severity scores. Interestingly, this is not fully accounted for by age, weight or bacterial species. Nonetheless, a limitation of the study is that the severity metadata were only available from those admitted to PICU (28 of 104 individuals) so caution should be taken when interpreting these data.

This study's specific strengths include a large number of clinically prevalent, Gram positive pathogenic isolates showing high reproducibility with FT-IR classification. All the clinical isolates were etiological agents of life-threatening infection, and therefore rapid identification by novel methods could potentially lead to earlier diagnosis and treatment, especially if combined with single cell analysis approaches such as Optical-PhotoThermal Infrared Spectroscopy (O-PTIR) on direct patient samples [73,74]. Limitations of the study are that only a limited range of pathogens, in terms of species, were studied and required purification cultures before analysis. Caution should exercised with association analysis of the pathogen sequencing and clinical outcome, due to relatively small sample size per species but may assist future basic research in *S. pneumoniae* and *S. aureus*. Future work could investigate FT-IR profiles in a wider population of pathogenic bacterial species to obtain a library of spectra as has been trialled in MALDI-TOF [30]. Further work is required to ascertain if antibiotic sensitivity and resistance profiles can be elucidated by FT-IR spectra which would allow more rapid, appropriate treatment and if this can be performed directly on patient samples using single cell techniques. Additionally, the utility of FT-IR analysis in outbreak source identification in nosocomial spread of infection in real time could be investigated [18,75] for specific species and allow targeting of WGS in other species. There is also scope to investigate the role of FT-IR spectroscopy in predicting clinical severity.

Conclusion

In conclusion, FT-IR is a promising, rapid and affordable spectral technique with high reproducibility that allows identification of species of clinical, pathogenic bacteria and distinction of subtypes of certain species. This proof-of-concept study will allow further evaluation of a wider range of pathogenic bacteria with varying antimicrobial resistance, identification of sample contaminants and application to direct patient samples using FT-IR.

Summary points

- FT-IR spectroscopy is a low cost, rapid vibrational technique that does not degrade the sample and provides a 'molecular fingerprint'. This has been studied in other areas of life sciences, for example, environmental contamination, but not in pathogenic clinical isolates.
- A total of 104 clinical pathogens from serious bacterial infections in children from 4 species were studied. This is a proof-of-concept design to evaluate FT-IR diagnostic accuracy, clinical severity and ability to subtype pathogenic bacteria.
- FT-IR spectroscopy diagnostic accuracy was 99.6% compared with the currently used clinical technique of MALDI-TOF MS (95.8%) on purified isolates. Further research is required to assess this directly on patient samples.
- FT-IR analysis clustered *N. meningitidis* and group A *Streptococcus* in a similar manner to gold standard whole-genome sequencing. However, it produced significantly different phylogenetic relationships to WGS for *S. pneumoniae* or *S. aureus* subtypes. This is consistent with other laboratory strain studies and suggests FT-IR has limited application in outbreak settings, but may be beneficial in certain bacterial species as it is much more rapid than WGS.
- *S. aureus* WGS variants were significantly associated with longer length of hospital stay, after controlling for multiple comparisons, including known virulence factors (FnbA/B) and cell wall proteins (Ebh, SraP).
- Comparing *S. pneumoniae* WGS data with clinical metadata, genes involved in DNA repair mechanisms were found to be significantly associated with the need for non-invasive ventilation. Loss of a zinc ABC transporter lipoprotein was associated with a greater PICU length of stay and has been described a hyperencapsulated phenotype. In contrast with the *S. aureus* variants, these were typically more conservative.
- PLS-R modeling of FT-IR spectra with clinical demographics (e.g., age, length of stay) did not show significant correlation. However pediatric risk of mortality (PRISM) severity scores correlated with FT-IR spectra for those admitted to pediatric intensive care.
- FT-IR is a promising technique for clinical microbiology application and warrants further study in a wider range of pathogenic species, with antimicrobial resistance and direct patient specimens before it is routinely applied to clinical workflows.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/fmb-2024-0043

Author contributions

Conceptualization – R McGalliard, E Carrol and R Goodacre. Formal analysis – R McGalliard, H Muhamadali, N AlMasoud, S Haldenby, V Romero-Soriano, E Allman. Software – S Haldenby. Methodology – Y Xu.

Funding acquisition – R McGalliard, ED Carrol. Supervision – ED Carrol, R Goodacre, AP Roberts, S Paterson. Writing – original draft preparation – R McGalliard. Writing – review & editing – all authors.

Acknowledgments

This publication made use of the Meningitis Research Foundation Meningococcus Genome Library (http://www.meningitis.org/research/genome) developed by Public Health England, the Wellcome Trust Sanger Institute and the University of Oxford as a collaboration, which is part funded by Meningitis Research Foundation.

Financial disclosure

This study was supported by the Wellcome Trust (ref: ISSF3). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

Taylor & Francis
Taylor & Francis Group

Ethical conduct of research

The samples collected in this study were collected in a previous study with ethical approval (Research Ethics Committee reference: 11/LO/1982) [3]. Written informed consent was gained during the study period to use microbiology samples for future research.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Winters BD, Eberlein M, Leung J, Needham DM, Pronovost PJ, Sevransky JE. Long-term mortalty and quality of life in sepsis: a systematic review*. *Crit. Care Med.* 2010;38(5):1276–1283. doi: 10.1097/CCM.0b013e3181d8cc1d
2. Yende S, Austin S, Rhodes A, *et al.* Long-term quality of life among survivors of severe sepsis. *Crit. Care Med.* 2016;44(8):1461–1467. doi: 10.1097/CCM.0000000000001658
3. Boeddha NP, Schlapbach LJ, Driessen GJ, *et al.* Mortality and morbidity in communityacquired sepsis in Eu-ropean pediatric intensive care units: a prospective cohort study from the European Childhood Life-threatening Infectious Disease Study (EUCLIDS). *Crit. Care* 2018;22(143):1–13. doi: 10.1186/s13054-018-2052-7
4. Martinón-Torres F, Salas A, Rivero-Calle I, *et al.* Life-threatening infections in children in Europe (the EUCLIDS Project): a prospective cohort study. *Lancet Child Adolesc Health* 2018;2(6):404–414. doi: 10.1016/S2352-4642(18)30113-5

• **Original research paper that obtained ethics approval for the clinical samples and metadata used in this paper.**

5. Mancini N, Carletti S, Ghidoli N, Cichero P, Burioni R, Clementi M. The era of molecular and other non-culture-based methods in diagnosis of sepsis. *Clin. Microbiol. Rev.* 2010;23(1):235–251. doi: 10.1128/CMR.00043-09
6. Chisanga M, Muhamadali H, Ellis DI, Goodacre R. Surface-enhanced raman scattering (SERS) in microbiology: illumination and enhancement of the microbial world. *Appl. Spectrosc.* 2018;72(7):987–1000. doi: 10.1177/0003702818764672
7. Sherry NL, Gorrie CL, Kwong JC, *et al.* Multi-site implementation of whole-genome sequencing for hospital infection control: a prospective genomic epidemiological analysis. *Lancet Reg. Health West Pac.* 2022;23:100446. doi: 10.1016/J.LANWPC.2022.100446
8. Mellmann A, Bletz S, Böking T, *et al.* Real-time ge-nome sequencing of resistant bacteria provides precision infection control in an institutional setting. *J. Clin. Microbiol.* 2016;54(12):2874. doi: 10.1128/JCM.00790-16
9. Ibrahim GM, Morin PM. Salmonella serotyping using whole-genome sequencing. *Front Microbiol.* 2018;9(12):2993. doi: 10.3389/FMICB.2018.02993/FULL
10. SenGupta DJ, Cummings LA, Hoogestraat DR, *et al.* Whole-genome sequencing for high-resolution investigation of methicillin-resistant *Staphylococcus aureus* epidemiology and genome plasticity. *J. Clin. Microbiol.* 2014;52(8):2787–2796. doi: 10.1128/JCM.00759-14
11. Byrne HJ, Baranska M, Puppels GJ, *et al.* Spectropathology for the next generation: quo vadis? *Analyst* 2015;140(7):2066–2073. doi: 10.1039/c4an02036g
12. Cialla-May D, Zheng XS, Weber K, Popp J. Recent progress in surface-enhanced Raman spec-troscopy for biological and biomedical applications: from cells to clinics. *Chem. Soc. Rev.* 2017;46(13):3945–3961. doi: 10.1039/c7cs00172j
13. Wenning M, Breitenwieser F, Konrad R, Huber I, Busch U, Scherer S. Identification and differen-tiation of food-related bacteria: a comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry. *J. Microbiol. Methods* 2014;103:44–52. doi: 10.1016/j.mimet.2014.05.011

• **Environmental food-related bacterial identification by FT-IR and MALDI-TOF MS.**

14. Rodriguez-Saona LE, Khambaty FM, Fry FS, Dubois J, Calvey EM. Detection and identification of bacteria in a juice matrix with Fourier transform-near infrared spectroscopy and multivariiate analysis. *J. Food Prot.* 2004;67(11):2555–2559. doi: 10.4315/0362-028x-67.11.2555
15. Ellis DI, Broadhurst D, Kell DB, Rowland JJ, Goodacre R. Rapid and quantitative detection of the microbial spoilage of meat by fourier transform infrared spectroscopy and machine learning. *Appl. Environ. Microbiol.* 2002;68(6):2822–2828. doi: 10.1128/AEM.68.6.2822-2828.2002
16. Muhamadali H, Weaver D, Subaihi A, *et al.* Chicken, beams, and Campylobacter: rapid differentiation of foodborne bacteria via vibrational spectrosco-py and MALDI-mass spectrometry. *Analyst* 2016;141(1):111–122. doi: 10.1039/c5an01945a
17. Ellis DI, Ellis J, Muhamadali H, Xu Y, Horn AB, Goodacre R. Rapid, high-throughput, and quanti-tative determination of orange juice adulteration by Fourier-transform infrared spectroscopy. *Analyt. Methods* 2016;8(28):5581–5586. doi: 10.1039/C6AY01480A
18. Dieckmann R, Hammerl JA, Hahmann H, *et al.* Rapid characterisation of Klebsiella oxytoca isolates from contaminated liquid hand soap using mass spectrometry, FTIR and Raman spectroscopy. *Faraday Discuss* 2016;187:353–375. doi: 10.1039/c5fd00165j

19. Wang H, Hollywood K, Jarvis RM, Lloyd JR, Goodacre R. Phenotypic characterization of She-wanella oneidensis MR-1 under aerobic and anaerobic growth conditions by using fourier transform infrared spectroscopy and high-performance liquid chromatography analyses. *Appl. Environ. Microbiol.* 2010;76(18):6266–6276. doi: 10.1128/AEM.00912-10
20. Wharfe ES, Jarvis RM, Winder CL, Whiteley AS, Goodacre R. Fourier transform infrared spec-troscopy as a metabolite fingerprinting tool for monitoring the phenotypic changes in complex bacterial communities capable of degrading phenol. *Environ. Microbiol.* 2010;12(12):3253–3263. doi: 10.1111/j.1462-2920.2010.02300.x
21. Muhamadali H, Subaihi A, Mohammadtaheri M, *et al.* Rapid, accurate, and comparative differentiation of clinically and industrially relevant microorganisms: via multiple vibrational spectroscopic fingerprinting. *Analyst* 2016;141(17):5127–5136. doi: 10.1039/c6an00883f
22. Muhamadali H, Xu Y, Ellis DI, *et al.* Metabolic profiling of *Geobacter sulfurreducens* during industrial bioprocess scale-up. Löffler FE, editor. *Appl. Environ. Microbiol.* 2015;81(10):3288–3298. doi: 10.1128/AEM.00294-15
23. Wehbe K, Vezzalini M, Cinque G. Detection of mycoplasma in contaminated mammalian cell culture using FTIR microspectroscopy. *Anal. Bioanal. Chem.* 2018;410(12):3003–3016. doi: 10.1007/s00216-018-0987-9
24. Muhamadali H, Xu Y, Ellis DI, *et al.* Metabolomics investigation of recombinant mTNF-α production in Streptomyces lividans. *Microb. Cell Fact.* 2015;14(1):157. doi: 10.1186/s12934-015-0350-1
25. Muhamadali H, Subaihi A, Mohammadtaheri M, *et al.* Rapid, accurate, and comparative differentiation of clinically and industrially relevant microorganisms via multiple vibrational spectroscopic fingerprinting. *Analyst* 2016;141(17):5127–5136. doi: 10.1039/c6an00883f
26. Welker M, Moore ERB. Applications of whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry in systematic microbiology. *Syst. Appl. Microbiol.* 2011;34(1):2–11. doi: 10.1016/j.syapm.2010.11.013
27. Angeletti S. Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology. *J. Microbiol. Methods* 2017;138:20–29. doi: 10.1016/j.mimet.2016.09.003
28. Sauget M, Valot B, Bertrand X, Hocquet D. Can MALDI-TOF mass spectrometry reasonably type bacteria? *Trends Microbiol.* 2017;25(6):447–455. doi: 10.1016/j.tim.2016.12.006
29. P L, D J, R G, T S, J D. Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) for the identification of highly pathogenic bacteria. *TrAC – Trends Analyt. Chem.* 2016;85:103–111. doi: 10.1016/j.trac.2016.04.013
30. Sakarikou C, Ciotti M, Dolfa C, Angeletti S, Favalli C. Rapid detection of carbapenemase-producing Klebsiella pneumoniae strains derived from blood cultures by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS). *BMC Microbiol.* 2017;17(1):54. doi: 10.1186/s12866-017-0952-3
31. Sauget M, Bertrand X, Hocquet D. Rapid antibiotic susceptibility testing on blood cultures using MALDI-TOF MS. Becker K, editor. *PLOS ONE* 2018;13(10):e0205603. doi: 10.1371/journal.pone.0205603
32. Almasoud N, Xu Y, Ellis DI, Rooney P, Turton JF, Goodacre R. Rapid discrimination of *Enterococcus faecium* strains using phenotypic analytical techniques. *Analytical Methods* 2016;8(42):7603–7613. doi: 10.1039/c6ay02326f
33. Slater A, Shann F, Pearson G. PIM2: a revised version of the pediatric Index of Mortality. *Intensive Care Med.* 2003;29(2):278–285. doi: 10.1007/S00134-002-1601-2
34. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score. *Crit. Care Med.* 1988;16(11):1110–1116. doi: 10.1097/00003246-198811000-00006
35. Muhamadali H, Chisanga M, Subaihi A, Goodacre R. Combining Raman and FT-IR spectroscopy with quantitative isotopic labeling for differentiation of *E. coli* cells at community and single cell levels. *Anal. Chem.* 2015;87(8):4578–4586. doi: 10.1021/acs.analchem.5b00892
36. Winder CL, Gordon SV, Dale J, Hewinson RG, Goodacre R. Metabolic fingerprints of *Mycobacterium bovis* cluster with molecular type: implications for genotype-phenotype links. *Microbiology (Reading)* 2006;152(Pt 9):2757–2765. doi: 10.1099/mic.0.28986-0
37. Martens H, Nielsen JP, Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. *Appl. Near-Infrar. Transm. Analys. Powder Mixt.* 2003;75(3):394–404. doi: 10.1021/AC020194W
38. Dhanoa MS, Barnes RJ, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* 1989;43(5):772–777. doi: 10.1366/0003702894202201
39. Hill DMC, Lucidarme J, Gray SJ, *et al.* Genomic epidemiology of age-associated meningococcal lineages in national surveillance: an observational cohort study. *Lancet Infect. Dis.* 2015;15(12):1420–1428. doi: 10.1016/S1473-3099(15)00267-4
40. Bankevich A, Nurk S, Antipov D, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 2012;19(5):455. doi: 10.1089/CMB.2012.0021
41. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31(19):3210–3212. doi: 10.1093/BIOINFORMATICS/BTV351
42. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 2012;9(8):811–814. doi: 10.1038/NMETH.2066

Taylor & Francis
Taylor & Francis Group

43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 2012;9(4):357–359. doi: 10.1038/NMETH.1923
44. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30(14):2068–2069. doi: 10.1093/BIOINFORMATICS/BTU153
45. Tonkin-Hill G, MacAlasdair N, Ruis C, *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* 2020;21(1):180. doi: 10.1186/S13059-020-02090-4
46. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic al-gorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 2015;32(1):268–274. doi: 10.1093/MOLBEV/MSU300
47. Alcock BP, Raphenya AR, Lau TTY, *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48(D1):D517–D525. doi: 10.1093/NAR/GKZ935
48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 1990;215(3):403–410. doi: 10.1016/S0022-2836(05)80360-2
49. Chen L, Yang J, Yu J, *et al.* VFDB: a reference database for bacterial viru-lence factors. *Nucleic Acids Res.* 2005;33(Database issue):D325–D328. doi: 10.1093/NAR/GKI008
50. Jaillard M, Lima L, Tournoud M, *et al.* A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. *PLoS Genet.* 2018;14(11):e1007758. doi: 10.1371/JOURNAL.PGEN.1007758
51. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for mi-crobial pangenome-wide association studies. *Bioinformatics* 2018;34(24):4310–4312. doi: 10.1093/BIOINFORMATICS/BTY539
52. Sievers F, Higgins DG. Clustal omega. *Curr. Protoc. Bioinformatics* 2014;48:3.13.1–3.13.16. doi: 10.1002/0471250953.BI0313S48
53. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom. Intellig. Lab. Syst.* 1987;2(1–3):37–52. doi: 10.1016/0169-7439(87)80084-9

• **Detailed comparison of supervised learning methods used in this paper.**

54. Gromski PS, Muhamadali H, Ellis DI, *et al.* A tutorial review: metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta.* 2015;879:10–23. doi: 10.1016/j.aca.2015.02.012
55. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J. Anal. Test.* 2018;2(3):249–262. doi: 10.1007/s41664-018-0068-2
56. Yang Y-H, Jiang Y-L, Zhang J, *et al.* Structural insights into SraP-mediated *Staphylococcus aureus* adhesion to host cells. *PLoS Pathog.* 2014;10(6):e1004169. doi: 10.1371/journal.ppat.1004169
57. Cheng AG, Missiakas D, Schneewind O. The giant protein Ebh is a determinant of *Staphylococ-cus aureus* cell size and complement resistance. *J. Bacteriol.* 2014;196(5):971. doi: 10.1128/JB.01366-13

• **Research paper exploring hyperencapsulated phenotype in *S. pneumoniae*.**

58. Lei MG, Cue D, Roux CM, Dunman PM, Lee CY. Rsp inhibits attachment and biofilm formation by repressing fnbA in *Staphylococcus aureus* MW2. *J. Bacteriol.* 2011;193(19):5231–5241. doi: 10.1128/JB.05454-11/FORMAT/EPUB
59. Durmort C, Ercoli G, Ramos-Sevillano E, *et al.* Deletion of the zinc transporter lipoprotein AdcAII causes hyperencapsulation of *Streptococcus* pneumoniae associated with distinct alleles of the Type I restriction-modification system. *mBio* 2020;11(2):e00445-20. doi: 10.1128/mBio

•• **Review paper of FT-IR in bacterial diagnostic techniques.**

60. Preisner O, Lopes JA, Guiomar R, Machado J, Menezes JC. Fourier transform infrared (FT-IR) spectroscopy in bacteriology: towards a reference method for bacteria discrimination. *Anal. Bioanal. Chem.* 2007;387(5):1739–1748. doi: 10.1007/s00216-006-0851-1
61. Novais Å, Freitas AR, Rodrigues C, Peixe L. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. *Europ. J. Clin. Microbiol. Infect. Dis.* 2019;38(3):427–448. doi: 10.1007/s10096-018-3431-3
62. Yang H, Shi H, Feng B, *et al.* Protocol for bacterial typing using Fourier transform infrared spectroscopy. *STAR Protoc.* 2023;4(2):102223. doi: 10.1016/j.xpro.2023.102223
63. Zarnowiec P, Lechowicz L, Czerwonka G, Kaca W. Fourier transform infrared spectroscopy (FTIR) as a tool for the identification and differentiation of pathogenic bacteria. *Curr. Med. Chem.* 2015;22(14):1710–1718. doi: 10.2174/0929867322666150311152800
64. Dinkelacker AG, Vogt S, Oberhettinger P, *et al.* Typing and species identification of clinical Klebsiella isolates by Fourier-1 transform infrared (FTIR) spectroscopy and matrix-assisted laser 2 desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry 3 4 Running title: typing of Klebsiella isolates by FTIR and MALDI-TOF Downloaded from. *J. Clin. Microbiol.* 2018;56(11):e00843-18. doi: 10.1128/JCM.00843-18
65. Kuhm AE, Suter D, Felleisen R, Rau J. Identification of Yersinia enterocolitica at the species and subspecies levels by Fourier transform infrared spectroscopy. *Appl. Environ. Microbiol.* 2009;75(18):5809–5813. doi: 10.1128/AEM.00206-09
66. Davis R, Mauer LJ. Subtyping of Listeria monocytogenes at the haplotype level by Fourier trans-form infrared (FT-IR) spectroscopy and multivariate statistical analysis. *Int. J. Food Microbiol.* 2011;150(2–3):140–149. doi: 10.1016/j.ijfoodmicro.2011.07.024
67. Sousa C, Novais Â, Magalhães A, Lopes J, Peixe L. Diverse high-risk B2 and D *Escherichia coli* clones depicted by Fourier Transform infrared spectroscopy. *Sci. Rep.* 2013;3(1):3278. doi: 10.1038/srep03278

68. Cordovana M, Mauder N, Join-Lambert O, *et al.* Machine learning-based typing of *Salmonella enterica* O-serogroups by the Fourier-Transform Infrared (FTIR) spectroscopy-based IR Biotyper system. *J. Microbiol. Methods* 2022;201:106564. doi: 10.1016/J.MIMET.2022.106564

- **FT-IR and WGS comparison for nocosomial bacterial subtyping.**

69. Brouwer S, Rivera-Hernandez T, Curren BF, *et al.* Pathogenesis, epidemiology and control of Group A *Streptococcus* infection. *Nat. Rev. Microbiol.* 2023;21:431–447. doi: 10.1038/s41579-023-00865-7

70. Teng ASJ, Habermehl PE, van Houdt R, *et al.* Comparison of fast Fourier transform infrared spectroscopy biotyping with whole-genome sequencing-based genotyping in common nosocomial pathogens. *Anal. Bioanal. Chem.* 2022;414(24):7179. doi: 10.1007/S00216-022-04270-6

71. Busby EJ, Doyle RM, Leboreiro Babe C, *et al.* Evaluation of matrix-assisted laser desorption ionization-time of flight mass spectrometry for molecular typing of *Acinetobacter baumannii* in comparison with orthogonal methods. *Microbiol. Spectr.* 2023;11(3):e0499522. doi: 10.1128/spectrum.04995-22

72. Aranega-Bou P, Cornbill C, Rodger G, *et al.* Evaluation of Fourier Transform Infrared spectroscopy (IR Biotyper) as a complement to whole-genome sequencing (WGS) to characterise *Enterobacter cloacae*, *Citrobacter freundii* and *Klebsiella pneumoniae* isolates recovered from hospital sinks. *medRxiv* 2023. doi: 10.1101/2023.04.24.23289028

73. Shams S, Lima C, Xu Y, Ahmed S, Goodacre R, Muhamadali H. Optical photothermal infrared spectroscopy: a novel solution for rapid identification of antimicrobial resistance at the single-cell level via deuterium isotope labeling. *Front Microbiol.* 2023;14:1077106. doi: 10.3389/fmicb.2023.1077106

74. Muhamadali H, Chisanga M, Subaihi A, Goodacre R. Combining Raman and FT-IR spectroscopy with quantitative isotopic labeling for differentiation of *E. coli* cells at community and single cell levels. *Anal. Chem.* 2015;87(8):4578–4586. doi: 10.1021/acs.analchem.5b00892

75. Bisognin F, Messina F, Butera O, *et al.* Investigating the origin of *Mycobacterium chimaera* contamination in heater-cooler units: integrated analysis with Fourier Transform infrared spectroscopy and whole-genome sequencing. *Microbiol. Spectr.* 2022;10(6):e0289322. doi: 10.1128/SPECTRUM.02893-22