Finding the Peak in the Mass-Stack: Rapid and Accurate Detection of Virulence and Resistance in Clinical Routine Diagnostics with MALDI-TOF Mass Spectrometry

Inauguraldissertation zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der Universität Basel
von

Aline Cuénod

2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von:

Prof. Adrian Egli
Prof. Urs Jenal
Prof. Sylvain Brisse

Basel, 22.02.2022

Prof. Dr. Marcel Mayor, Dean of the Faculty

Table of Contents

1 Abbreviations

aa	Amino Acid
AAHC	Antibiotic-Associated Haemorrhagic Colitis
AIEC	Adherent Invasive *Escherichia coli*
AMR	Antimicrobial Resistance
CHCA matrix	alpha-Cyano-4-hydroxycinnamic matrix
ANI	Average Nucleotide Identity
bp	Base pairs
CI	Confidence Intervall
CRP	C-reactive Protein
Da	Dalton
DAEC	Diffusely Adherent *Escherichia coli*
EAEC	Enteroaggregative *Escherichia coli*
EHEC	Enterohemorrhagic *Escherichia coli*
EIEC	Enteroinvasive *Escherichia coli*
EPEC	Enteropathogenic *Escherichia coli*
EQA	External Quality Assessment
ESBL	Extended-Spectrum Beta-Lactamases
ESKAPE	*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter*, identified by the WHO as priority pathogens for the development of new antibiotic treatment options
ETEC	Enterotoxigenic *Escherichia coli*
ExPEC	Extraintestinal Pathogenic *Escherichia coli*

GLM	Generalised Linear Model
GLMM	Generalised Mixed-effect Model
bGWAS	Bacterial Genome Wide Association Study
HGT	Horizontal gene transfer
ICU	Intensive Care Unit
IQR	Inter Quartile Range
m/z	Mass to charge ratio
MALDI-TOF	Matrix Assisted Laser Desorption Ionization-Time of Flight
MB	Megabase (i.e. million base pairs)
MDR	Multi-Drug Resistance
MLST	Multi-Locus Sequence Typing
MRSA	Methicillin Resistant *Staphylococcus aureus*
MS	Mass Spectrometry
MSQ	Mass Spectral Quality
PCR	Polymerase Chain Reaction
rMLST	Ribosomal Multi Locus Sequence Typing
ST	Sequence Type
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
UPEC	Uropathogenic *Escherichia coli*
USB	University Hospital Basel
spp.	Species
subsp.	Subspecies
UTI	Urinary Tract Infection

WGS	Whole genome Sequencing
WHO	World Health Organisation

2 Summary

Infectious diseases are amongst the most common causes of morbidity and death. Moreover, the global increase of antimicrobial resistance (AMR) threatens to undo earlier achievements in modern medicine. Accurate and fast bacterial identification in clinical diagnostics is key, as it forms the basis of a tailored treatment. The most commonly used tool for bacterial species identification in clinical routine diagnostics is Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS). MALDI-TOF MS is fast, accurate, and cost-effective. However, three key problems remain: (i) not all clinically relevant bacterial species can reliably be identified using MALDI-TOF MS; (ii) virulent strains within a species are not routinely identified; and (iii) the time from sample collection until an AMR profile is available still requires 48-72 hours. How can these critical aspects in diagnostics be overcome? And what would be the potential impact for the patient?

In this thesis, I aimed to increase the resolution of bacterial identification by MALDI-TOF MS in clinical routine diagnostics with a focus on the genus *Klebsiella* (**Chapter I**) and the species *E. coli* (**Chapter II**) and to predict AMR from MALDI-TOF mass spectra, using machine learning approaches (**Chapter V**).

In the first part of my thesis (**Chapter I**), I established a ribosomal marker-based approach to distinguish the species within the genus *Klebsiella* with MALDI-TOF MS. Next, I applied this to a large, international dataset of mass spectra (n=33,160) and AMR profiles (n=7,876) to identify species-specific trends in AMR profiles. Further, I linked the species classification to clinical outcomes compiled from a single healthcare centre (n=957 clinical cases). I found that strains of the *K. oxytoca* complex to be significantly more likely causes invasive infections than strains of the *K. pneumoniae* complex.

To anticipate the course of an infection, it is necessary to know which bacterial strains have the potential to cause life-threatening diseases, such as sepsis. In the second project of my thesis (**Chapter II**), I, therefore, isolated over 1,000 *E. coli* strains from urinary tract- and bloodstream infections and compiled patient characteristics and outcomes of the respective clinical cases (n=831). Applying a bacterial genome wide association study (bGWAS), I substantiated *papGII* as an important, patient-independent bacterial factor for causing invasive infections (bacteraemia). However, I could not identify MALDI-TOF MS peaks specific for *E. coli* strains carrying this virulence factor and rapid sequence amplification-based diagnostics might be more suitable and faster.

MALDI-TOF mass spectral quality (MSQ) is crucial for accurate species identification. While analysing MALDI-TOF mass spectra from different healthcare centres, I observed differing numbers of detected ribosomal marker masses. MSQ is currently not precisely defined nor

regularly assessed in diagnostic laboratories. I, therefore, sought to identify mass spectral features, which can be used to precisely describe MSQ and identify simple protocols yielding the highest MSQ for varying bacterial strains (**Chapter III**).

Further, I identified a large heterogeneity of MSQ from mass spectra acquired in 36 international diagnostic laboratories, mainly driven by a few particularly well or poorly performing MALDI-TOF MS devices and likely linked to sample preparation practices (**Chapter IV**). Applying the simple protocols identified in **Chapter III** improved MSQ for previously poorly performing devices/laboratories. The resolution of MALDI-TOF MS based bacterial identification can be improved with a high MSQ in clinical routine diagnostics. A standardised MSQ will likely benefit direct phenotype prediction by supervised classification algorithms (i.e. machine learning).

To assess the potential of machine learning based analysis approaches to predict AMR from MALDI-TOF mass spectra, we compiled an extensive dataset of over 300,000 routinely acquired MALDI-TOF mass spectra with matching AMR data from four healthcare centres (**Chapter V**). We yielded accurate predictions for important species and clinically relevant antibiotic drugs: for Ceftriaxone (as an indicator for ESBL) for *E. coli* and *K. pneumoniae,* we yield an area under the receiver operating characteristic curve (AUROC) of 0.74 for both and for Oxacillin (as an indicator for MRSA) for *S. aureus* with an AUROC of 0.80. The classification was most accurate if the classifiers were trained on spectra of a single species acquired at the same centre and in close temporal proximity to the test set.

Overall, my thesis showed (i) the potential of MALDI-TOF MS to identify bacteria with higher resolution, (ii) that AMR can accurately be predicted from MALDI-TOF mass spectra and (iii) high MSQ is essential to translate these advances into clinical routine diagnostics.

3 Introduction

Microbes shape all forms of life. Eight percent of the human genome originates from ancient viral infections (1,2). Interactions with microbes can benefit animal and human hosts - resulting in a symbiotic relationship (3). However, microbes can also be the source of infection resulting in local inflammation and tissue damage induced by the host's immune response. The host-microbe interactions are a major driver of (co-)evolution (4). Infectious diseases have been a burden on humanity since time immemorial and can be caused by eukaryotic parasites (5), fungi (6), yeasts (7), viruses (8), and bacteria (9).

Up until the discovery of vaccines and antibiotic drugs, many bacterial infectious diseases, such as tuberculosis (10), whooping cough (11), and diphtheria (12) were life-threatening. Community acquired pneumonia (13), meningitis (14), and sepsis (15) remain a major threat to human health in many places around the globe. In the last few decades, and before the pandemic of COVID-19, deaths caused by infectious diseases have globally decreased (16,17), mainly because of vaccines (18), improved diagnosis (19), and antibiotic treatment (20). However, at the same time, the prevalence of antimicrobial resistant (AMR) bacterial strains is increasing globally (21,22). With it, the number of deaths attributed to infectious disease is predicted to increase again dramatically (23). This situation may challenge the achievements of modern medicine. A recent study by Cassini et al. showed that in 2015, 671'689 infections with antibiotic-resistant bacteria accounted for an estimated 33'110 attributable deaths and 874'541 disability adjusted life-years in the European Union (DALYs). In particular, the burden was highest in infants (aged <1 year) and people aged 65 years or older and had increased since 2007 (24). An international consortium (the *Antimicrobial Resistance Collaborators*) estimated 4.95 million deaths in 2019 globally were associated with bacterial AMR, 1.27 million deaths of which were directly attributed to the resistance (25). O'Neill estimated that the global burden of AMR will result in approximately 10 million deaths per year by 2050, exceeding deaths by cancer and resulting in economic costs of >10 trillion US Dollars (23). Why do AMR and hypervirulent bacterial strains evolve and spread so rapidly? And how can this profoundly worrisome development be prevented?

Resistance or virulence can be developed either by point mutations (26) or by acquiring resistance and virulence genes (27). Point mutations arise *de novo* (28) and are confined to vertical transfer within a lineage, the acquisition of AMR and virulence genes is associated with horizontal gene transfer. This can drive their spread between different lineages (29) or even between bacterial species (30–33). Within a bacterial species, the genes can be assigned to the core- or the accessory genome: On one side, certain genes are shared amongst all strains of a species, this part is called the core genome (34); On the contrary,

other genes occur only in one or a subset of the strains within the species, this part is called the accessory genome (35). Together, the core and accessory genome form the pan-genome, comprising all genes described in strains of this species (36). The size of the accessory genome of a species correlates with the lifestyle of a pathogen (37). Some pathogens have specialised life cycles with human cells as their specific niches (38), such as the intracellular, obligate human pathogens *Chlamydia trachomatis* (37) and *Mycobacterium tuberculosis* (39). They are understood to have a closed pan-genome, with no accessory genome (40,41).

In contrast, other species can occupy various ecosystems and easily migrate between niches (42). Such versatile lifestyles are reflected by a large accessory genome, where a single strain can quickly adapt to new environments by adding beneficial functions from a large gene pool (43,44). *Klebsiella* spp. and *Escherichia coli* (*E. coli*) strains, both considered opportunistic pathogens within the family of *Enterobacteriaceae*, are known for their genome plasticity, which is mainly driven by the frequent exchange of mobile genetic elements (43,45). This rapid adaptation of gene repertoires can lead to heterogeneous phenotypes in closely related strains, affecting clinically relevant traits such as virulence and antimicrobial resistance (43). Especially during evolutionary bottlenecks such as antibiotic drug exposure, AMR genes may rapidly accumulate in the accessory genome (46,47). Thereby fluctuations in the accessory genome are pivotal for the variation of pathogenicity within a species. Can we readily identify this intra-species diversity using microbiological diagnostics?

The main task of diagnostics is to identify the infection's causative agents and determine the pathogen-specific antibiotic resistance profile. The timely assessment of species and AMR profile allows (i) to estimate the virulence potential of an infection and (ii) to determine the most effective and optimal treatment (48). Kumar et al. showed that a delay in effective antibiotics results in a direct increase in mortality (49). Considering the phenotypic heterogeneity of closely related bacterial species and strains, this identification must be as precise as possible. Ideally, it enables the detection of hypervirulent or resistant strains and thereby anticipate the clinical course and what treatments have to be initiated. The current key challenges in microbiological routine diagnostics are: (i) the low taxonomic resolution, as not all clinically relevant bacterial species can be distinguished using routine diagnostic approaches (50–52) and (ii) the lack of readily available information about a pathogen's virulence (53) and AMR profile. Currently, culture-based classical phenotypic methods for AMR identification may take 48-72 hours (54). How can these important challenges in clinical routine diagnostics be overcome?

My thesis aims to assess (i) whether the taxonomic resolution of bacterial identification in clinical routine diagnostics can be increased by optimising currently used methods (**Chapter I** and **Chapter II**), (ii) what clinical implication an increased resolution has (**Chapter I** and **Chapter II**), (iii) whether there are technical limitations to an increased resolution (**Chapter**

III), (iv) how technical limitations can be overcome in clinical routine diagnostics (**Chapter IV**), and (v) whether the turn-around time for the first evidence on AMR can be accelerated using new analytical approaches (**Chapter V**).

3.1 Bacterial species identification in clinical diagnostics

Rapid and precise bacterial species identification is the first and pivotal step in tailoring the appropriate treatment (55). Therefore, understanding the diagnostic workflow is key (**Figure 1**). As the first step in clinical microbiology diagnostics, patient materials are collected. Depending on the clinical question, the samples are (i) either cultured on various growth media (56) or (ii) subjected to rapid Polymerase Chain Reaction (PCR) assays directly from the patient material. Both methods aim to substantially increase elements or the whole pathogen to increase the sensitivity of the detection method. While the identification via PCR is rapid, it is limited to targeted genes and is impossible to follow up with phenotypic analysis, such as antimicrobial resistance profiles. Therefore, the culture of bacterial infections is still the most commonly used approach. After several hours to days of incubation on solid media, visible colonies are subjected to species identification using a Matrix-Assisted Laser Desorption Ionization-Time of Flight (MALDI-TOF) Mass Spectrometry (MS) system (56). Time from sample collection to species identification requires, in most cases, around 24 hours using MALDI-TOF MS. If the identified species is of clinical relevance, an antimicrobial resistance profile is performed using various automated or manual methods such as broth microdilution, disc diffusion, or MIC gradient strips (57). The turn-around time in bacterial diagnostics is critical, as faster reporting is associated with improved clinical outcomes (58). Recent advances in sample preparation methods enable positive blood cultures to be subjected to direct MALDI-TOF MS measurements, using a specific protocol to remove host material such as erythrocytes and leukocytes. Applying this workflow yields a species identification almost 24 hours before visible colonies can be picked from subcultured solid agar plates (48,59). Other advances aim to increase the taxonomic resolution of identification in clinical routine diagnostics. In the last few years, an increasing number of modern diagnostic laboratories have incorporated WGS in their diagnostic workflow and research repertoire to assess questions related to hospital epidemiology, unclear species assignments, AMR, and bacterial virulence (60). New technologies, such as WGS and MALDI-TOF MS, benefit from standardisation in the clinical routine diagnostic workflows. Standardisation is necessary to increase a methods reproducibility. High reproducibility is required to transition from a research tool to diagnostics. A standard tool to ensure high reproducibility in clinical routine diagnostics are external quality assessments (EQA). EQAs are used for comparing technologies and understanding the impact of variation in the operating protocols. Commonly,

EQAs focus mainly on the correct species identification and AMR profiling of selected samples without instructing on the laboratory methods to use (61,62). However, a more research-driven EQA may help to also profoundly understand the impact of a new technology at higher resolution and provide a deeper insight into the available new method.

Despite the highly efficient diagnostic workflow, the time from sample collection to AMR profile can take up to 48 hours (54). Accelerating these turn-around times would be highly desirable, especially in a hospital context, where critically ill patients are treated. For this reason, substantial efforts are made to develop in-house or commercial methods, trying to reduce the time to identification and time to AMR profile. However, speed often comes with a lack of resolution. To resolve this problem, a major remaining question is, can MALDI-TOF MS and a high-resolution method such as whole genome sequencing (WGS) be combined?

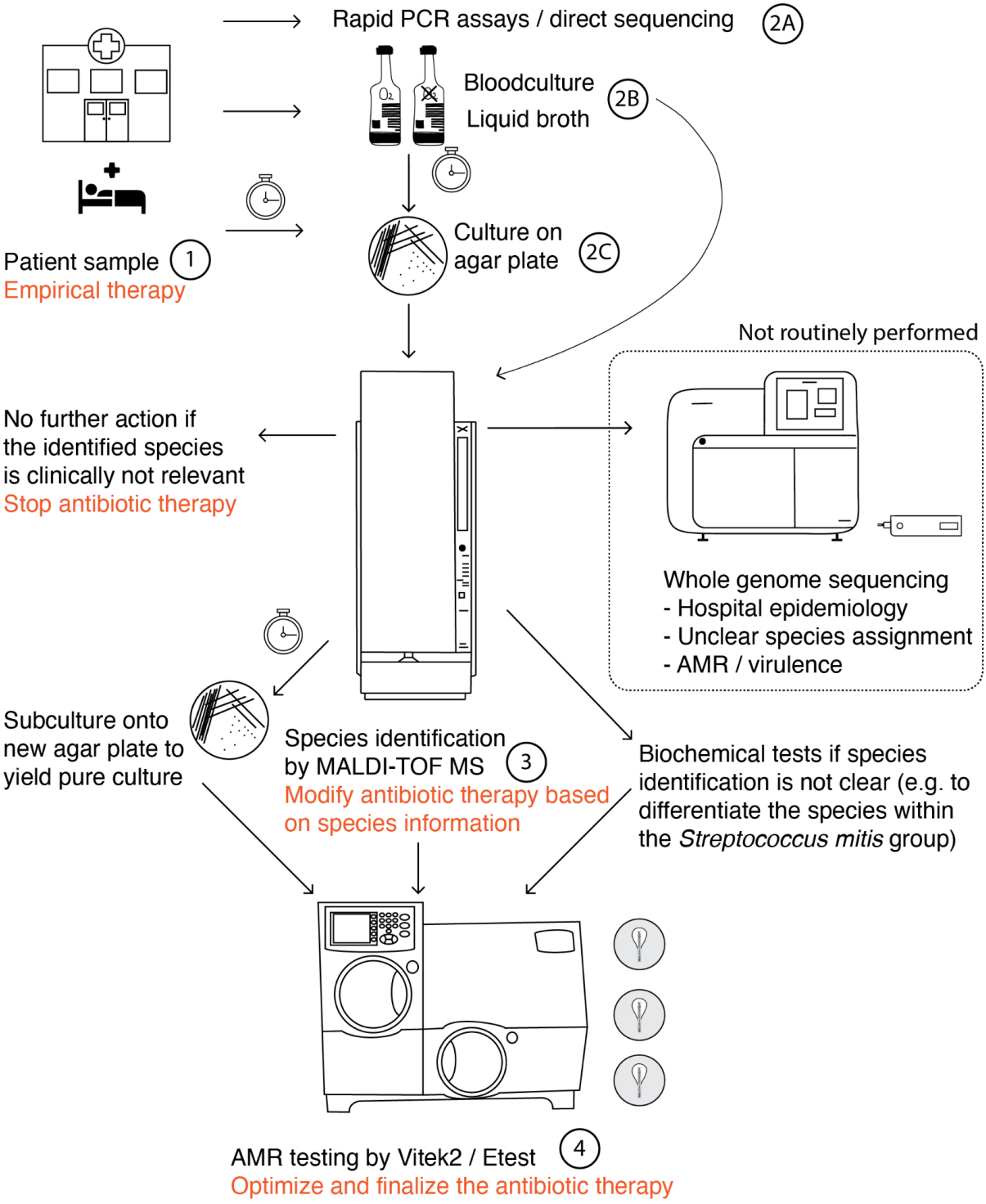

Figure 1: Overview of the workflow in modern clinical, microbiological diagnostic laboratories for positive blood cultures, e.g. in sepsis diagnostics. 1: Collection of the patient sample in blood culture flasks; Processing of positive blood cultures with 2A: Rapid diagnostic assays, e.g. panel PCR, 2B: MALDI-TOF MS directly from positive blood culture flasks, 2C: Sub-culturing on agar plate material from the positive blood culture; 3: Species identification using MALDI-TOF MS; 4: Determination of the AMR profile. 'PCR': Polymerase Chain Reaction, 'AMR': Antimicrobial Resistance, 'MALDI-TOF MS': Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry

3.2 MALDI-TOF MS for microbial species identification

The following section is adapted from:

Aline Cuénod and Adrian Egli. "Advanced Applications of MALDI-TOF MS – Typing and Beyond." Book Chapter 9 *Application and Integration of Omics-Powered Diagnostics in Clinical and Public Health Microbiology*, edited by Jacob Moran-Gilad and Yael Yagel, 153–73. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-62155-1_9.

3.2.1 MALDI-TOF MS functions and workflow

MALDI-TOF MS has become the most commonly used tool for microbial species identification in clinical routine diagnostics and has widely replaced species identification assays based on biochemical properties (63). MALDI-TOF MS assigns the bacterial species within minutes from cultured isolates (64) and has been shown to yield accurate results (65,66). How did MALDI-TOF MS become so successful, and how does it work?

Briefly, the routine MALDI-TOF MS works as follows: a small amount of bacterial material (from a single colony) is transferred onto a MALDI-TOF MS target plate, using a plastic inoculation needle or a wooden toothpick. The sample is then overlaid with 1 µl of formic acid (70% for laboratories using the microflex Biotyper system and 25% for other MALDI-TOF MS devices) air-dried and coated with 1 µl of alpha-Cyano-4-hydroxycinnamic (CHCA) matrix before measurement (**Figure 2**). Alternative sample preparation protocols recommend overlaying the sample directly with matrix, without pre-treatment with formic acid (67). The matrix serves as an energy absorbent organic compound, co-crystallising with the microbial sample. After the matrix has dried, the target plate is inserted into the mass spectrometer, where a vacuum is established. Subsequently, a laser beam (e.g. nitrogen or YAC laser) is fired on the sample-matrix crystals, evaporating and ionising the analytes in the sample. The ions are accelerated in electrostatic potential, in a manner dependent on the ions' mass and charge. Ions with a low mass/charge (m/z) ratio are accelerated more efficiently and reach a higher velocity, whereas ions with a higher m/z ratio fly comparably slower. After acceleration, the ions enter a flight tube without an electrostatic field, where they are separated according to their m/z ratio and detected at the end of the tube. The m/z ratio of an ion is determined by measuring the time required to travel through the flight tube. A MALDI-TOF mass spectrum is generated according to this time of flight (TOF) information and the strength of the recorded signal (in millivolt). The mass range in which peptides are detected in clinical routine diagnostics is 2,000 and 20,000 Daltons. Although many aspects of this workflow are standardised, the technical and biological reproducibility of MALDI-TOF MS is variable both

within and between different laboratories (68). At the end of this process, a mass spectrum is generated, containing approximately 50-150 peaks corresponding to the most abundant proteins. This information in itself is not helpful for a clinical assessment of an infection but has to be analysed to identify species.

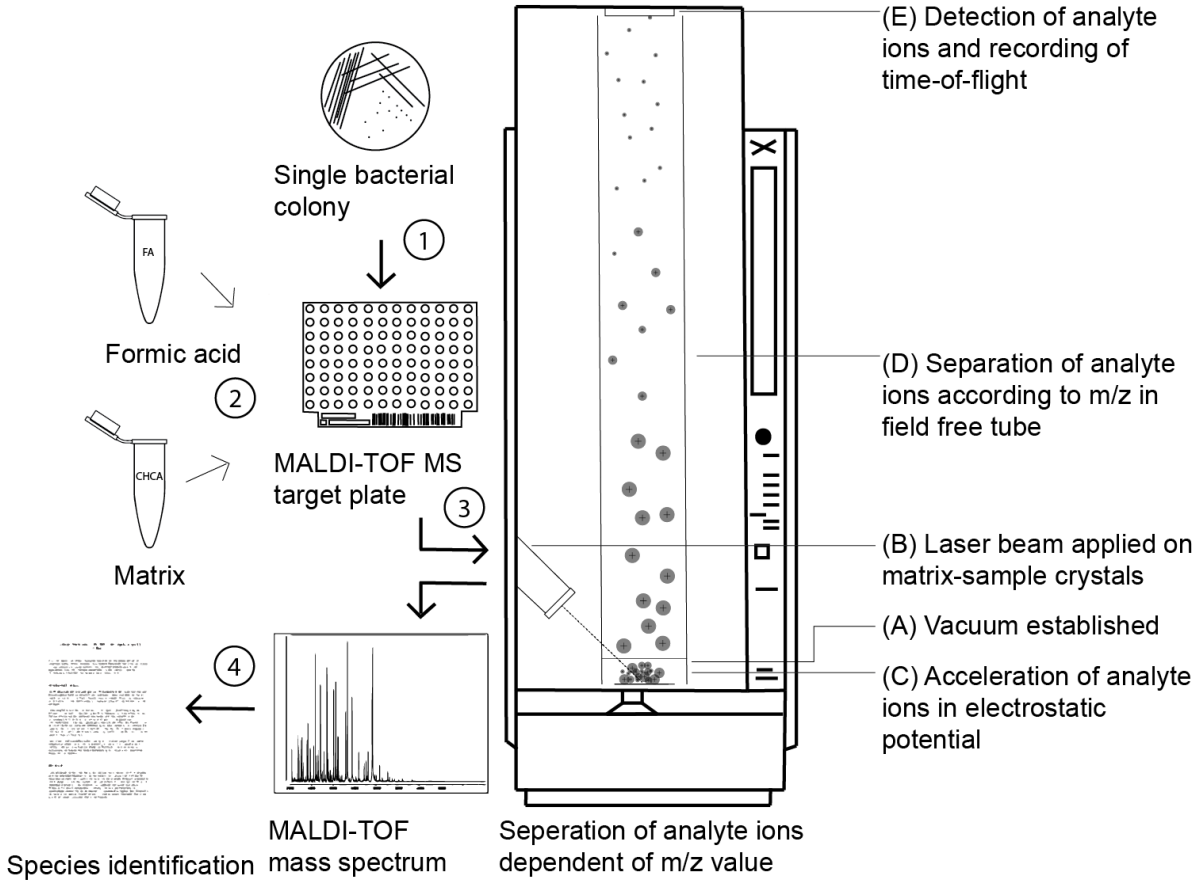

Figure 2: Schematic representation of the MALDI-TOF MS function and workflow. Processes outside the MALDI-TOF MS device are assigned numbers, inside the mass spectrometer are assigned letters. 1. A single bacterial colony is transferred onto a target plate, 2. overlaid with formic acid and / or Matrix solution, and 3. inserted into the device, where (A) a vacuum is established and (B) a laser shot is applied to the matrix-sample crystals. (C) The bacterial cells break open, and the analyte ions are accelerated in a manner dependent on their mass and their charge. (D) The analyte ions are separated in a field-free flight tube at the end of which they are (E) detected and recorded. 4. The recorded mass spectrum is compared to a reference database, yielding a species identification.

3.2.2 Two analytic approaches allow species identification from MALDI-TOF mass spectra

Different inherent properties of the MALDI-TOF mass spectra are used to identify species. In any case, the location and intensity of the peaks are compared to a reference database of bacterial reference strains. There are mainly two analytic approaches for species identification

from MALDI-TOF mass spectra. The two approaches differ in the kind of data deposited in the database. On the one hand, similarity-based databases contain a MALDI-TOF mass spectrum per bacterial strain, i.e. all detected signals, regardless of their protein identity (63). On the other hand, marker-based databases contain a list of masses per bacterial strain that correspond exclusively to known phylogenetic marker proteins (69). These two approaches are not mutually exclusive since the masses of known proteins are also present in the entire MALDI-TOF mass spectra. Nevertheless, both approaches have different advantages and disadvantages, briefly outlined in the following sections.

3.2.2.1 Similarity-based species identification by MALDI-TOF MS

Similarity-based databases contain a complete collection of all MALDI-TOF mass spectrum signals detected per bacterial strain. The assumption behind similarity-based approaches is that strains of each microbial taxon produce unique patterns when measured by MALDI-TOF MS, so-called "peptide mass fingerprints". Ideally, these fingerprints should assign measured strains unambiguously to single taxa. The two most widely used commercial systems in clinical microbiology diagnostics are the MALDI Biotyper (Bruker Daltonics, Bremen, Germany) and the VitekMS (bioMérieux, Marcy l'Étoile, France), which both rely on similarity-based approaches. Few similarity-based innovations, such as the SARAMIS database, additionally offer the possibility to weight specific peaks as more important for identification, reducing ambiguity in identification and increasing resolution (70,71).

This similarity-based approach has multiple advantages: (i) it yields accurate identification on genus level and for many important bacterial species (72,73); (ii) the databases are straightforward to set up: as no information other than the correct species assignment has to be considered, no extensive computing power is required; and (iii) also peaks of unknown protein identity are considered which can increase the resolution of identification.

However, similarity-based approaches have multiple important disadvantages: (i) Each database entry requires the acquisition of MALDI-TOF mass spectra from cultured strains, which puts a limit to the completeness of species that can be covered; (ii) varying culture conditions and growth media might reduce similarity scores, even of identical clones; and (iii) comparison to spectra based databases can yield ambiguous results and defining identification thresholds and their interpretation is not intuitive.

3.2.2.2 Marker-based bacterial identification by MALDI-TOF MS

In marker-based identification approaches, the acquired MALDI-TOF mass spectra are compared against databases of known masses, whose protein identity is often known and specific to bacteria of interest. Marker-based approaches have been applied to distinguish closely related bacterial species (74), clinically relevant lineages within a species (75,76),

bacterial strains carrying a particular antimicrobial resistance factor (77) or mobile genetic element (78). These approaches are often used as second-line identifications (74,79) when the approximate taxonomic affiliation of a bacterial strain has already been identified using similarity-based approaches.

Discriminatory marker masses are (i) either identified empirically by comparing the MALDI-TOF mass spectra of the group of interest to mass spectra from closely related strains or (ii) by predicting discriminatory mass shifts from genomic data (69). The former has the advantage of only interpreting signals that are measurable. Like similarity-based analyses, however, bacterial strains and MALDI-TOF mass spectra must be available, which are limited in their accessibility. The latter has the advantage of being able to use genomic databases, which are significantly more diverse and accessible.

A large proportion of the peaks reproducibly detected in MALDI-TOF mass spectra correspond to ribosomal subunit proteins (80). Ribosomal subunit proteins are among the most abundant cytosolic proteins in replicating bacterial cells (81) and are often within the mass range of MALDI-TOF MS. They are rarely post-translationally modified, apart from N-terminal methionine loss, which can be accounted for when predicting the mass from amino acid sequences of these proteins (82). Several studies predicting m/z values from whole-genome sequences have focused on ribosomal subunit proteins and demonstrated an increase in resolution of bacterial identification by MALDI-TOF MS gained by this approach, compared to similarity-based identification approaches (69,76,83–85).

The advantages of a marker-based approach predicting discriminatory peaks from publicly available genomic data are: (i) it yields an increased taxonomic resolution compared to purely similarity-based approaches; (ii) a more extensive and more diverse set of bacterial strains can be included, as surprisingly more genomic than MALDI-TOF data is publicly available; (iii) the specificity of the predicted mass shifts can be assessed on a large scale by accessing the strain diversity in public genome databases; and (iv) the comparison to marker-based databases make uncertainties in identification more apparent. In the case that a crucial marker mass is not detected, no ambiguous identification is assigned.

The disadvantages of a marker-based approach are (i) the increased need for computing power and bioinformatic knowledge and (ii) no consideration of unknown peaks, which might increase the discriminatory power.

Overall when it comes to marker-based approaches it is necessary to evaluate (i) which marker proteins can reproducibly be detected? (ii) What is the sensitivity and specificity of this gained resolution? (iii) What are the clinical phenotypes of the species/lineages which can be identified (**Chapter I**)? Besides species identification, MALDI-TOF MS has been further developed for additional applications, including the determination of antibiotic resistances.

3.2.3 Antimicrobial Resistance Testing Using MALDI-TOF MS

3.2.3.1 Phenotypic Antimicrobial Resistance Testing Using MALDI-TOF MS

MALDI-TOF MS can be used for phenotypic antibiotic susceptibility testing by either (i) detecting the enzymatic hydrolysis of the antibiotic substance (86,87) or (ii) by detecting growth vs no growth of the bacterial strain in the presence of the tested antibiotic (88,89).
The detection of enzymatic hydrolysis is established for cephalosporins and carbapenems. For these, commercial and certified kits are available, yielding results after 30-60 minutes of incubation (86). In these assays a mass shift can be detected in the MALDI-TOF mass spectra specific to the cleavage of the antibiotic beta-lactam ring. Thereby, the beta-lactamase activity can be assessed independently of the bacterial species. However, such approaches fail to detect AMR resulting from mechanisms other than cleavage of the antibiotic, such as porin loss, the action of efflux pumps or alterations of the antibiotic target proteins.
In contrast, growth assays can be applied to all bacterial species and antibiotic combinations, as growth in the presence of an antibiotic substance serves as a universal readout for AMR (88). However, the detection of antimicrobial resistance by MALDI-TOF MS requires detectable concentrations of bacterial cells, which can be heavily influenced by the sensitivity of the analyte ion detector. A detector, which is not sensitive enough, or low-quality mass spectra might lead to false negative results, reporting resistant strains as susceptible. This should be avoided at all costs in a clinical setting (so-called ‘very major error’) (54). Moreover, no unambiguous cut-off values (for instance, on the total intensity, the number of intracellular proteins detected) to discretely distinguish between resistant and susceptible strains have yet been established. Growth assays using MALDI-TOF MS have not yet been introduced to clinical diagnostics. Overall, all currently available phenotypic antimicrobial resistance tests are limited in their ability to detect AMR. Alternative routes for the detection of AMR from routinely acquired MALDI-TOF MS without prior antibiotic incubation need to be developed.

3.2.3.2 Antimicrobial resistance prediction from routinely acquired MALDI-TOF mass spectra

Multiple studies have aimed to distinguish antibiotic susceptible and resistant strains from MALDI-TOF mass spectra without the prior incubation with antibiotic substances (90–92). This has the advantage of providing information on the antimicrobial resistance pattern at the time of the species identification without the need for further data acquisition or hands-on time. Most of these studies have focused on one species - antibiotic combination and identified a single or a few peaks as discriminatory signals. Some discriminatory peaks correspond to protein biomarkers encoded on mobile genetic elements, also encoding the resistance

mechanism. Examples of such biomarkers are PSM-mec encoded on the SCC*mec* cassette, which also encodes *mecA* in *S. aureus*, conferring resistance to methicillin (92) or a hypothetical protein encoded on a plasmid harbouring a carbapenemase, which is circulating within the family of *Enterobacteriaceae* and mainly within *K. pneumoniae* strains (78). More frequently, however, phylogenetic marker peaks have been identified, distinguishing resistant from susceptible lineages within a species and thereby serving as surrogate markers (75,93,94). Single peak identification has a series of serious disadvantages: (i) the association of these peaks to AMR might reflect the local spread of a resistant clone and might not be generalisable; (ii) AMR carried by a lineage or encoded on a mobile genetic element not carrying the peak in question cannot be identified; and (iii) MALDI-TOF mass spectra of poor quality can be noisy or low-signal, leading to incorrect AMR assignments. Therefore, additional methods were explored to overcome these problems (**Chapter V**).

Multiple studies predict antimicrobial resistance from MALDI-TOF mass spectra using supervised classification algorithms (i.e. machine learning) focused on one species- antibiotic combination and focused on a limited amount of mass spectra acquired at one health care centre (94–96). These studies have important drawbacks: (i) they do not provide external validation of their classification algorithms (such as from strains collected at a different location or a different time point), (ii) they often do not assess whether their classifications would have clinical implications, and (iii) often do not evaluate which signals were most important for accurate classification. In considerations of these shortcomings and as MALDI-TOF MS exhibits large inter-laboratory variability (68) is not yet clear how this affects AMR prediction. Currently, AMR prediction from MALDI-TOF mass spectra is not yet routinely achieved.

To evaluate the potential of MALDI-TOF MS for the prediction of AMR, a large-scale statistical analysis of routinely acquired mass spectra from different health care centres is required (**Chapter V**). Further, the impact of an early AMR prediction on clinical practice and antibiotic stewardship needs to be assessed. Especially, as there are several remaining challenges for bacterial identification using MALDI-TOF MS.

3.2.4 Current challenges for MALDI-TOF MS in clinical routine diagnostic

Although MALDI-TOF MS is already a highly valuable method mainly for species identification in modern microbiological diagnostic laboratories, the technique also has important limitations. Most crucially, some clinically relevant bacterial species or genomic features (i.e. AMR) cannot readily and reliably be identified.

The low resolution for some species or features can have various reasons, such as (i) incomplete databases, (ii) close relatedness of the bacterial species of interest, and (iii) the variation in mass spectral quality. While I have addressed the limitation of the database in the

previous chapters and the similarity of the bacterial species or features is inherent, the quality of MALDI-TOF mass spectra quality remains to be improved.

3.2.4.1 MALDI-TOF mass spectral quality

The quality of MALDI-TOF mass spectra can be highly variable and impact the species identification (97,98). The criteria to assess MALDI-TOF mass spectral quality (MSQ) are currently not precisely defined. Consequently, the MSQ is rarely assessed in diagnostic laboratories. EQA focussing on the use of MALDI-TOF MS for microbial species identification have been reported (99), aiming to compare the ability of the participating laboratories to identify a specific set of bacterial strains. However, the resolution of species identification is heavily influenced by the reference database used, and little is known about the impact of the quality of the acquired mass spectra, independent of the database used.

Although there are guidelines on sample preparation and interpretation of species identification results (such as by the MALDI-TOF MS manufacturers or by the *Clinical and Laboratory Standard Institute* (100)), many diagnostic laboratories have developed their own standard operating procedures and varying sample preparation protocols are used. Moreover, only a fraction of diagnostic laboratories regularly assesses the MALDI-TOF mass spectral quality, most often using the score assigned by the species identification database as only readout. The mass spectral quality in diagnostic laboratories is largely unknown. To translate marker-based species identification approaches to clinical routine diagnostics, a high mass spectral quality and the reproducible detection of marker masses is essential.

Therefore, it has to be determined (i) how can MSQ best be described? (ii) which routinely used sample preparation protocols yield the highest MSQ (**Chapter III**)? (iii) what is the current MSQ in routine diagnostic laboratories? And (iv) can it be improved using simple workflow adaptations (**Chapter IV**)? An improved MSQ will enable increased resolution in bacterial identification. In this thesis, I focused on *Klebsiella* spp. and *E. coli*, as these are clinically highly relevant and include virulent (53,101) as well as resistant (102,103) lineages.

3.3 The genus *Klebsiella*

3.3.1 Taxonomic aspects within the genus *Klebsiella*

The genus *Klebsiella* comprises rod-shaped, non-spore forming, Gram negative bacteria and is part of the family of *Enterobacteriaceae*. Currently, 18 species are published, of which ten have been newly described since 2004 (104–112) (**Figure 3**). These species are largely discrete from each other, where the closest average nucleotide identity (ANI) of the type-strain

genomes of two sister species is 96.21% (between *K. grimontii* and *K. pasteurii*), and the furthest is 85.67% (between *R. terrigen*a and *R. electrica*).

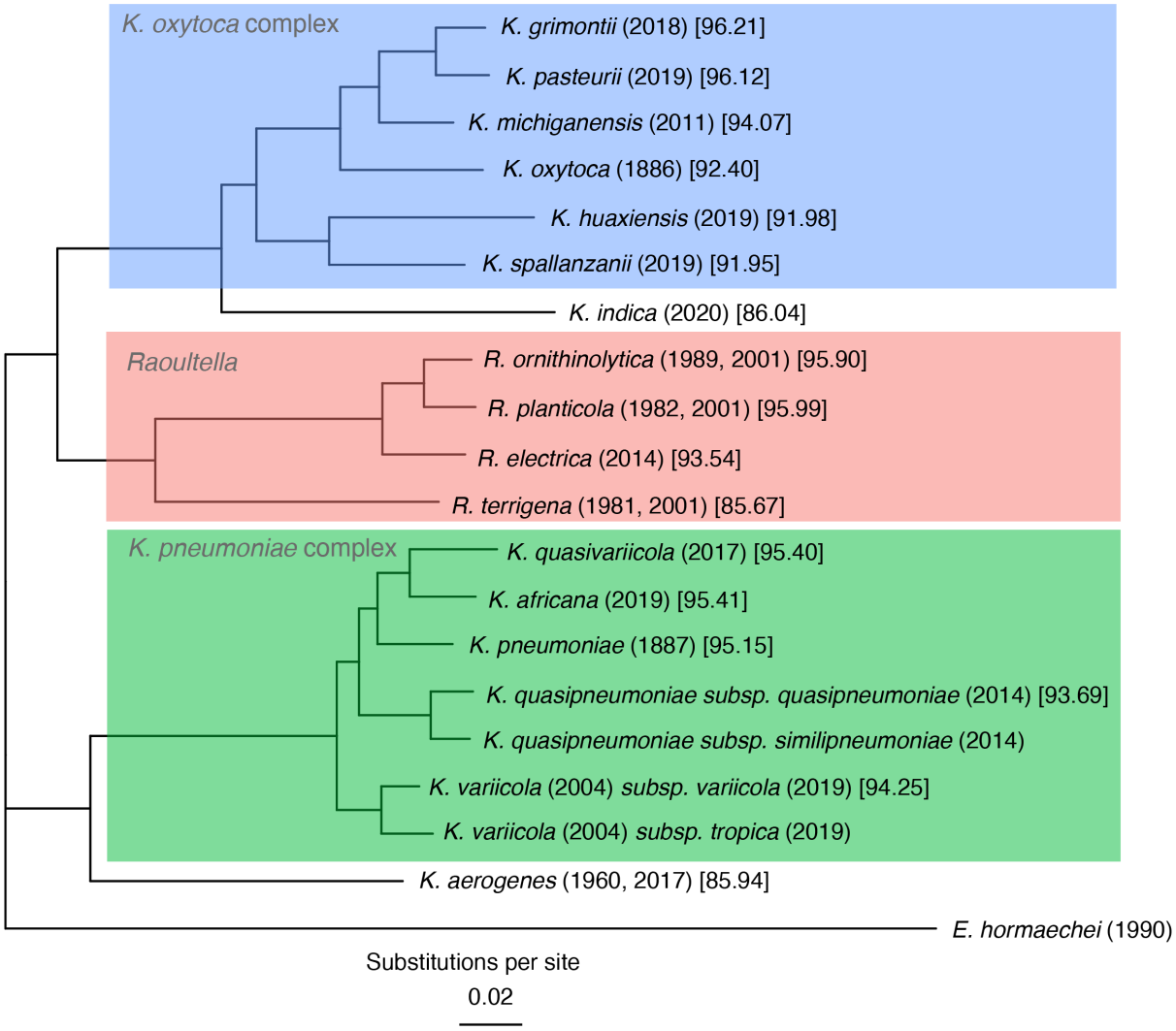

Figure 3: Core genome phylogeny constructed from the reference genomes of species and subspecies within the genus *Klebsiella*, excluding *K. granulomatis*, as no genomic sequence is publicly available for this species. The core genome alignment was compiled by panaroo (40), aligned using mafft (113) from which the tree was constructed using RaxML (114). Three complexes are highlighted: *K. pneumoniae* complex (green), the *K. oxytoca* complex (blue) and *Raoultella* (red). Tip labels consist of the species name, the year in which the species was first described in round brackets and the average nucleotide identity (ANI) to the closest species on the tree in square brackets calculated by fastANI (115).

Within the genus there are three main phylogenetic clusters: (i) The *Klebsiella pneumoniae* complex, (ii) the *Klebsiella oxytoca* complex and (iii) species which have been assigned to the genus "*Raoultella*". This assignment was based on *gyrB* sequences, but is debated, as it would render the genus *Klebsiella* polyphyletic (116).

The *K. pneumoniae* complex currently comprises five species, with two phylogenetic subspecies being described for the *K. variicola* and *K. quasipneumoniae* (104,112). The two subspecies defined within the species *K. pneumoniae* (*K. pneumoniae subsp. rhinoscleromatis* and *K. pneumoniae subsp. ozaenae*) cause distinct clinical phenotypes. However, genomic analyses have clarified that these correspond to clonal complexes within the species rather than forming distinct subspecies (117,118). The *K. oxytoca* group currently comprises six validly published species with no subspecies, and a recent study suggests the presence of three additional species (119). It has been proposed to reunify the genus *Raoultella* with the genus *Klebsiella*, as species of the two genera form one monophyletic clade, with the species assigned as '*Raoultella'* being nested within the species of the genus *Klebsiella* (120).

Two species are not assigned to the three complexes. Firstly *K. aerogenes*, previously classified within the genus *Enterobacter* (121), forms a distinct sister clade to the *K. pneumoniae* complex. Routinely identifying *K. aerogenes* is particularly important, as it carries an intrinsic active AmpC beta-lactamase (122). Secondly, *K. indica* (110) which is most closely related to the *K. oxytoca* complex (119). Moreover, *K. granulomatis* has validly been published (123). As neither a type-strain nor a complete genome sequence of this species is currently publicly available, its genotypic characteristics remain largely unknown. Overall, the *Klebsiella* genus is both genotypically and phenotypically highly variable. While the clinical relevance has been described for a few prevalent *Klebsiella* species, many have only recently been observed, and we are largely unaware of the clinical implications. The question that remains is, can we correlate different *Klebsiella* species with clinical phenotypes?

3.3.2 Clinical importance of the genus *Klebsiella*

Klebsiella spp. strains are found in natural environments (124), in association with plants (106) and colonise a wide range of animal hosts, including humans. Within humans, they are found in the gut, upper respiratory tract, and skin (125). In the human host, *Klebsiella* spp. strains can cause a variety of infections, including pneumonia, urinary tract infections, wound infections and severe diseases such as pyelonephritis, liver abscesses, sepsis and septic shock (103).

3.3.2.1 Virulence

Although all *Klebsiella* spp. have been described in a clinical context, the most frequently isolated and, consequently, the most studied species is *K. pneumoniae* (sensu stricto) (116). *K. pneumoniae* is estimated to cause 80-90% of infections attributed to the *K. pneumoniae* complex (117,126,127). In most cases, *K. pneumoniae* complex strains are considered

opportunistic pathogens, only occasionally causing disease, such as in immunocompromised patients, and are a leading cause of hospital-acquired infections (125,128,129).
In contrast to *K. pneumoniae* complex strains, strains of the *K. oxytoca* complex have been described causing antibiotic-associated haemorrhagic colitis (AAHC), which is linked to the production of the cytotoxins tilivaline and tilimycin (119,130–132). Reports on the clinical presentations of *Raoultella* spp. infections are scarce but seem to resemble those of other *Klebsiella* spp. (133). Little is known about species-specific symptoms within the complexes. A recent paper from Japan did not find significantly different clinical phenotypes between infectious diseases caused by members of the *K. pneumoniae* complex (134). A Swedish study reported a higher mortality rate for bloodstream (135) infections caused by *K. variicola* strains than *K. pneumoniae* strains. A nationwide evaluation of the *K. pneumoniae* group in Norway found that *K. variicola* more rarely showed ESBL (136).
Some hypervirulent *Klebsiella* spp. strains can cause severe, community-acquired infections in immunocompetent hosts, such as liver abscesses (53). Most *K. pneumoniae* liver abscesses are caused by a few clonal groups (118,137,138), most prominently, clonal group 23 (53), which includes strains of ST23, ST26, ST57 and ST63 (53,139).
Hypervirulent strains of the clonal group 23 have been reported to encode specific virulence factors, including the capsular type K1, the siderophores yersiniabactin, salmochelin, aerobactin, the genotoxin colibactin, and *rmpA*, which regulates the hypermucoid phenotype (45,53,117).

3.3.2.2 Antibiotic resistance

The number of multidrug-resistant *Klebsiella* spp. isolates being reported is increasing globally (103). Notably, *K. pneumoniae* is one of the ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter*), which have been identified by the World Health Organisation (WHO) as priority pathogens for the development of new antibiotic treatment options (140). Strains of the *K. pneumoniae* complex are intrinsically resistant against ampicillin, with the SHV, OKP and LEN beta-lactamases responsible in *K. pneumoniae*, *K. quasipneumoniae* and *K. variicola*, respectively (103,141). The species within the *K. oxytoca* complex encode different versions of the OXY beta-lactamase (119). Most AMR genes within *Klebsiella* are carried on mobile genetic elements such as plasmids. Moreover, it is not uncommon for strains within this genus to carry multiple AMR plasmids simultaneously (103). Numerous AMR factors have been discovered first within *K. pneumoniae* before spreading to other *Enterobacteriaceae* species (103,142–146), including certain ESBL and carbapenem resistance genes and plasmid born resistance against colistin. In particular, ST258 and its derivative ST512 are associated with the carriage of KPC plasmids (103,147). *K. pneumoniae,*

therefore, plays a key role in disseminating AMR genes from environmental to clinical settings (148). Within *K. pneumoniae,* multidrug-resistance and hypervirulence are associated with distinct lineages (45) while only a few hypervirulent, multidrug-resistant strains have been reported (149,150). However, as factors determining multidrug resistance and hypervirulence are encoded on mobile genetic elements, this convergence is possible and a major threat to public health (45).

3.3.2.3 Clinical diagnostics of *Klebsiella* spp.

In clinical routine diagnostics, strains within the genus are often identified at the complex level, as multiple species are absent from the commonly used MALDI-TOF MS species identification databases (74,119) and cannot be uniquely identified biochemically (50,111,119). Further, the distinction of the recently described species within each complex is not yet achieved, their virulence phenotypes remain largely unknown. A recent genomic study showed differences in the occurrence of AMR and virulence factors between the species within the two complexes (151). *K. variicola* and *K. grimontii* less frequently harbouring AMR genes than their close relatives within the *K. pneumoniae* complex and the *K. oxytoca* complex, respectively (151). Estimates of the occurrence of the newly described species rely on genomic studies, which can be biased towards the clones of the highest clinical interest, such as multidrug-resistant clones. They might undersample clinically less conspicuous isolates.
To assess the clinical phenotypes of the different *Klebsiella* spp. it needs to be evaluated (i) whether they can be distinguished in the clinical routine diagnostics using MALDI-TOF MS, (ii) how prevalent the newly described *Klebsiella* spp. are in the clinical setting; and (iii) whether there is a difference in their AMR and virulence phenotypes (**Chapter I**).
Within the family of *Enterobacteriaceae*, there are not only large differences in pathogenicity and AMR between the species (interspecies) but also within a species (intraspecies), most prominently within the species *E. coli*.

3.4 The species Escherichia coli

3.4.1 Taxonomic aspects within *Escherichia coli*

E. coli are rod-shaped, non-spore forming, Gram negative bacteria and are part of the genus *Escherichia* within the family of *Enterobacteriaceae*. Within the *E. coli* species, deep branching phylogenetic groups (defined as phylogroups) have been identified (**Figure 4** and **Figure 5**) (152). Based on a few housekeeping genes, eight *E. coli* phylogroups (A, B1, B2, C, D, E, F and G) were initially observed (152,153). A, B1, B2 and D represent the majority of isolated strains (43). The previously described ‘cryptic’ *E. coli* clades (153) have been proposed to be

classified as distinct species, and *E. ruysiae* has been described, encompassing the 'cryptic' *E. coli* clades III and IV (154). On the basis of genomic Mash distances (155) the *E. coli* phylogroups have recently been split up, and 14 phylogroups have been proposed, including two corresponding to strains of the genus of *Shigella* (156). Strains of the genus *Shigella* are intracellular pathogens causing dysentery (157,158). Four *Shigella* spp. have been defined on the basis of serotypes, *S. sonnei*, *S. flexneri*, *S. boydii* and *S. dysenteriae* (158). However, as the *Shigella* spp. strains fall within the *E. coli* phylogeny, and they are now considered as part of the species (159). Three Multi Locus Sequence Typing schemes have been developed for *E. coli*: (i) one scheme ('*Ec*MLST'), hosted at the University of Michigan (USA), which was developed initially for enteropathogenic *E. coli* (160), (ii) one scheme hosted at the Institut Pasteur (Paris, France) (161) and one hosted at the Warwick Medical School (Warwick, UK) (162) (163). Each of these schemes is based on varying combinations of seven different housekeeping genes, of which only *ind* is common to all (163). Of these schemes, the 'Warwick' scheme is most frequently used, as it represents the core genome phylogeny most accurately (163,164). These phylogenetic classifications are based on variations of core genes. They only account for a fraction of the diversity within this species, whose accessory genome is many times bigger than its core genome (44). In species with a large accessory genome and varying gene contents, even in closely related strains, these schemes do not allow a definitive assessment of the clinical phenotype of a strain.

E. coli have an open pan-genome currently including > 55,000 genes (44), with more genes being added constantly with the number of sequenced strains daily increasing. The core genome of this species is comparably small and comprises approximately 1,400 genes, which corresponds to 2.5% of the pan-genome and 30% of genes carried by a single *E. coli* strain (44). Within the species, very rapid changes in gene repertoires have been observed. These changes are mainly driven by the frequent acquisition of various mobile genetic elements (43) such as plasmids or prophages. This variability is reflected by diverging genome sizes of *E. coli* strains ranging from 4.2 - 6 million bp (43). Why is the genetic variability of *E. coli* clinically and diagnostically important?

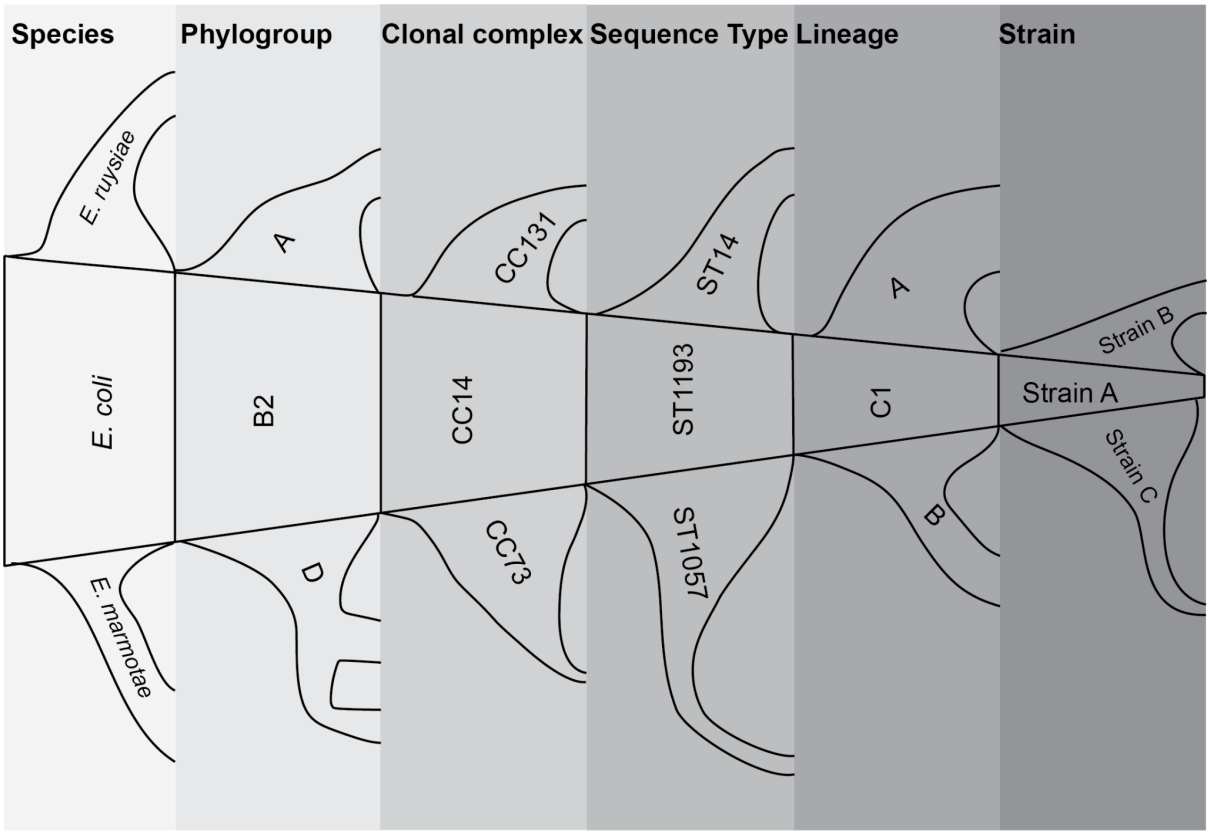

Figure 4: Schematic representation of the phylogenetic classifications of the species *E. coli* and their hierarchy. 'Species': *E. marmotae* and *E. ruysiae* are closely related to *E. coli* and have recently been defined as separate species (154,165); 'Phylogroups': the *E. coli* phylogroups are deep branching clades within the species, which have first been described based on the sequence comparison on four, then eight housekeeping genes (152,153); 'Clonal Complex': Multiple ST which cluster together around a central ST are grouped into the same Clonal Complex; 'Sequence Type' (ST): Strains which share the same allelic profile for the seven housekeeping genes included in typing scheme are assigned the same ST; 'Lineage': closely related strains within an ST; 'Strain': Clones within the same lineage which share a very recent common ancestor.

3.4.2 Clinical importance of *Escherichia coli*

Due to its genetic variability, members of the species *E. coli* show a wide range of host and environmental adaptations, also resulting in a large diversity of clinically relevant phenotypes. *E. coli* are commonly found in soil, freshwater, and part of the healthy human microbiota. Moreover, certain strains of the species, most prominently K12 (166) with its many daughter strains such as DH5α (167), have commonly been used as laboratory strains, making them well-studied model organisms of a generalist species (168,169). Within humans, *E. coli* can be pathogenic, causing intestinal and extraintestinal disease (43). Based on distinct clinical phenotypes and, in some cases, the presence of molecular markers, eight different *E. coli* pathotypes have been described (44,170).

3.4.2.1 *E. coli* pathotypes

Pathotypes which cause diarrhoeal disease are separated into enteropathogenic *E. coli* (EPEC) (171), enterotoxigenic *E. coli* (ETEC) (172), enterohemorrhagic *E. coli* (EHEC) (173), enteroaggregative *E. coli* (EAEC) (174), diffusely adherent *E. coli* (DAEC) (175), adherent invasive *E. coli* (AIEC) (176) and enteroinvasive *E. coli* (EIEC) (177), which often comprise *Shigella* spp. strains. These diarrheagenic pathotypes are each defined by a specific set of virulence factors and cause similar pathology (178). *E. coli* strains, which are isolated from body sites other than the gastrointestinal tract, are termed extraintestinal pathogenic *E. coli* (ExPEC) (179) including uropathogenic *E. coli* (UPEC) (180) isolated from the urinary tract as a sub-category. UPEC implies that these strains are observed causing urinary tract infections or pyelonephritis and can lead to urosepsis. There is no genetic marker that exclusively defines this pathotype (178).

The *E. coli* phylogroups have been associated with distinct pathotypes and clinical phenotypes (44): phylogroups A and B1 are linked to asymptomatic carriage in the human gut and frequently found isolated from freshwater samples (43,181); phylogroup E2 (182) and the *Shigella* (157) phylogroups are associated with intestinal infections and the phylogroups B2, D and F are associated with extraintestinal infection (183), such as urinary tract infections (UTI) and urosepsis (43,156) (**Figure 5**). However, these associations are loose, and no pathotype assignment can be deduced from the phylogroup. Indeed, the phylogroups B1 and B2 include specialised EHEC lineages, and phylogroup C comprises ExPEC strains (44).

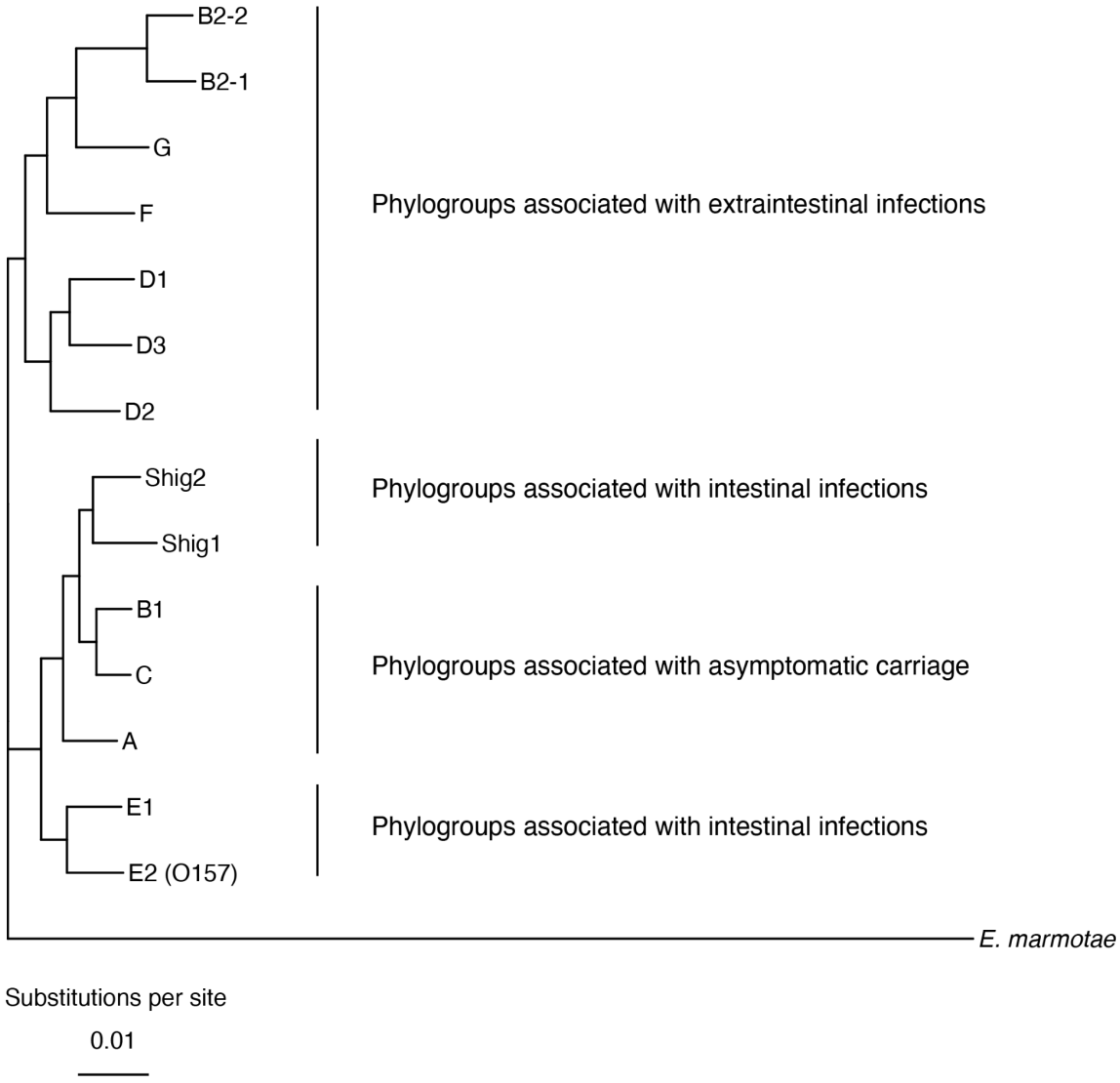

Figure 5: Core genome phylogenetic tree constructed from one reference strain per *E. coli* phylogroup (156) and the type strain of the closely related species *E. marmotae* as outgroup. The core genome alignment was compiled by panaroo (40), aligned using mafft (113) from which the tree was constructed using RaxML (114). The clinical phenotypes to which these phylogroups have been associated are indicated.

3.4.2.2 Urinary tract infections

Urinary tract infections (UTIs) are amongst the most common infections in humans worldwide (184). *E. coli* is the single most common cause (185), responsible for roughly two-thirds of all UTIs (186,187) (**Figure 6A**). Human risk factors include increased age, immunosuppression, functionally and structurally compromised urinary tracts, catheterisation, and behavioural factors, such as sexual activity (188). UTI occurs much more frequently in women than in men. In women, infections are often asymptomatic or mild, whereas UTI in men is uncommon and often associated with predisposing risk factors such as structural abnormalities (188) or

hyperplasia of the prostate (189). Although 10% of women experience one episode of symptomatic UTI each year, host genetic factors, which influence the susceptibility to UTI are still minimally understood (188). Symptomatic UTIs are usually treated for the relief of symptoms.

Due to efficient host defence mechanisms, such as the antibacterial activity of urine, including antibacterial peptides, an acid pH (190), and shearing forces of urine flow, bacterial pathogens require virulence factors to persist in the urinary tract (188,191). Host risk factors presumably reduce this requirement (192). Known virulence factors include polysaccharide capsules, iron capturing systems and adhesins, such as type I fimbriae, AfA adhesins, and pyelonephritis-associated pili (P-pili) (193,194). P-pili consist of multiple subunits, with PapG forming the adhesive tip (**Figure 6B**), which binds to human uroepithelial and kidney cells (195,196). PapG binds to a glycolipid receptor which is expressed on cells of the human urinary tract. These receptors exist in different isoforms, which differ in the number of sugar residues in the glycolipid molecule (197). The minimal receptor is called GbO3. The addition of a single sugar creates GbO4, whereas two additional sugars create GbO5 (197). PapG exists in different isoforms: PapGI to PapGV, which bind with different specificities to the varying receptor isotypes. Of these, *papGII* has been associated with invasive UTI, including pyelonephritis and urosepsis (198). PapGII binds the GbO4 globoseries of glycosphingolipids (GalNAcβ1-3Galα1-4Galβ1-4GlcCer) which is abundant on cells of the upper urinary tract of humans, whereas the PapGIII preferably binds to GbO5 (GalNAcα1-3-GalNAc3Galα1-4Galβ1-4GlcCer) (199,200). It has been shown that, on average, strains of the ExPEC phylogroups harbour a higher number of these virulence factors than strains of the carriage associated phylogroups (198). However, the distribution of these factors is heterogeneous within a phylogroup (44) and the phylogroup membership alone does not allow a definitive conclusion about the virulence potential of an *E. coli* strain. The most common UPEC Sequence Types (STs) isolated worldwide are ST69, ST73, ST95 and the ESBL associated ST131 (201,202).

Despite the knowledge on bacterial virulence and host risk factors, these are often analysed separately and rarely concurrently. In order to understand their interplay and epidemiological dynamics, a concurrent analysis is required (**Chapter II**). Does the ST information allow assessment of the AMR risk of a pathogen?

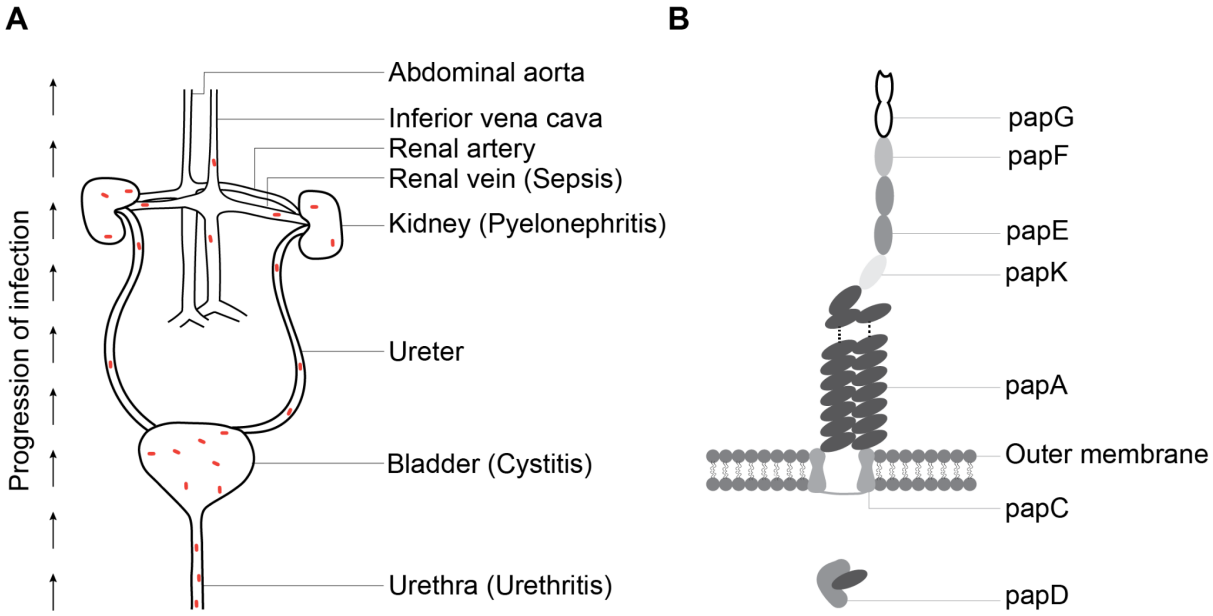

Figure 5: A: Schematic representation of an ascending Urinary Tract Infection (UTI). The name of the infection is given in brackets next to the respective organ name; B: Schematic representation of the *E. coli* pyelonephritis associated pilus (P-Pilus). Both schemes were adapted from Klein et al., *Nature Reviews Microbiology*, 2020 (203)

3.4.2.3 Antibiotic resistance

AMR is frequent in *E. coli* (204) with epidemiological survey studies reporting up to 60% of ExPEC isolates to be resistant against at least three antibiotic classes (21,205,206). The increase of AMR within ExPEC has been driven by a few successful lineages, most prominently ST131 (207,208). ST131 is associated with the carriage of an multi-drug resistant (MDR) plasmid, which encodes a CTX-M extended-spectrum beta-lactamase (209) and additionally carries a metallo-beta-lactamase conferring resistance to carbapenems (206,210). ST131 subclade C moreover encodes fluoroquinolone resistance, which is conferred via point mutations in both the DNA gyrase *gyrA* and the DNA topoisomerase *parC* (206,211,212), which have later been acquired by ST1193, another globally successful UPEC clone (213,214).

3.4.3 Laboratory diagnostics of *Escherichia coli*

E. coli grows well on standard 5% sheep blood agar plates and does not require any specific supplement (215). *E. coli* is commonly identified in microbiological diagnostics using MALDI-TOF MS or by biochemical profiling from single bacterial colonies. MALDI-TOF MS accurately identifies *E. coli* strains as such but does not distinguish the *E. coli* phylogroups and consequently assigns *Shigella* spp. as *E. coli* (216). Given the large heterogeneity in their clinical phenotype, the distinction of strains carrying important virulence factors is required to

assess the virulent potential of an *E. coli* strain. Rapid PCR assays exist to detect the diarrhoea associated pathotypes EIEC (including *Shigella*), EAEC, EPEC, ETEC and STEC (217). However, as no genetic determinants yet uniquely define the ExPEC pathotypes, no such assays exist for the early detection of virulent UPEC clones, which would allow identifying patients at risk for an invasive infection leading to urosepsis. An early distinction of virulent UPEC strains, potentially causing severe disease and commensal *E. coli* strains with only a limited virulence potential, would allow for an early adaptation of clinical treatments, adapt clinical follow-ups, and promote antibiotic stewardship.

It is, therefore, necessary to investigate (i) which *E. coli* virulence and patient characteristics are most important for the invasive progression of a UTI and (ii) whether virulent UPEC clones can be distinguished using MALDI-TOF MS (**Chapter II**).

4 Aims of the thesis

With the work presented in this thesis, we aim to (i) increase the accuracy of bacterial identification, virulence and AMR assessment in clinical routine diagnostics using MALDI-TOF MS and (ii) generate new knowledge about pathogenicity and outcomes of the host-pathogen interaction. More precisely, we focus on improving species and sub-species identification within *Klebsiella* spp. (**Chapter I**) and *E. coli* (**Chapter II**), addressing virulence. Further, we define (**Chapter III**) and improve (**Chapter IV**) MSQ in clinical routine diagnostics. Finally, we investigate the possibilities of detecting AMR using MALDI-TOF MS (**Chapter V**).

The aims of this thesis are to:

1. Distinguish the species within the genus *Klebsiella* using MALDI-TOF MS and assess their clinical phenotypes (**Chapter I**).
   We aim to achieve this by answering the following questions:
   - Do publicly available genomic data give evidence that the different *Klebsiella* spp. differ in virulence and AMR? (Page 46-50)
   - Which ribosomal marker masses can reproducibly be detected in MALDI-TOF mass spectra of routine quality acquired at multiple healthcare centres? (Page 51)
   - What is the sensitivity and specificity of a marker-based analytical approach for MALDI-TOF MS on species and complex levels? (Page 53)
   - Do the *Klebsiella* spp. exhibit varying phenotypic AMR profiles? (Page 55-56)
   - Is there a difference in virulence between the different *Klebsiella* spp.? (Page 56-57)
2. Elucidate which *E. coli* genetic factors and host characteristics have the greatest impact on the progression of UTI using a bacterial genome wide association study (bGWAS) and whether virulent *E. coli* strains can be identified in MALDI-TOF mass spectra (**Chapter II**).
   We aim to achieve this by answering the following questions:
   - Is there a difference between the phylogroup and ST distribution of *E. coli* strains (i) causing UTI vs urosepsis and (ii) isolated from different patient populations? (Page 77)
   - Which bacterial genes and host characteristics are most commonly associated with invasive UTIs? (Page 79-81)
   - Is there a difference in host immune response to *E. coli* strains carrying vs not carrying this virulence factor? (Page 82)

- Can virulent/invasive *E. coli* strains be recognised using MALDI-TOF MS? (Page 81)

3. Identify mass spectral features that can serve as good proxies for mass spectral quality and sample preparation protocols yielding highest mass spectral quality (**Chapter III**).
   We aim to achieve this by answering the following questions:
   - Which MALDI-TOF mass spectral features differ between mass spectra which are correctly classified on the species level and mass spectra which are not correctly identified? (Page 99-101)
   - Using these spectral features as proxies for MSQ: which routinely used sample preparation protocols yield the highest MSQ for unknown samples? (Page 102-106)
   - Are there differences in how different phylogenetic groups yield the highest MSQ? (Page 106-107)
4. Assess the mass spectral quality in routine diagnostic laboratories around the world and test whether it can be improved using simple workflow adaptations (**Chapter IV**).
   We aimed to achieve this by answering the following questions:
   - What MSQ do 36 laboratories in 12 countries achieve when measuring a well-defined set of diverse bacterial strains using their routine laboratory procedure? (Page 127-128)
   - Can these differences be linked to specific protocols and machine setups? (Page 130-131)
   - Can the assessed MSQ in the 36 laboratories be improved using simple workflow adaptations and the same strain set? (Page 132-133)
5. Predict AMR from MALDI-TOF mass spectra acquired in clinical routine diagnostics using machine learning algorithms (**Chapter V**).
   We aim to achieve this by answering the following questions:
   - Is the prediction from MALDI-TOF mass spectra more accurate than the prediction using the species information alone? (Page 145)
   - Which algorithms yield the highest performance? And are there differences between specific bacteria and antibiotics? (Page 147)
   - Is the prediction more accurate when the training set consists of mass spectra of a single bacterial species or of all species combined? (Page 150)
   - How do classifiers perform if the training dataset was acquired at timely or geographical distance from the test dataset? (Page 148-150)

5 Results

Chapter I: Whole-genome sequence-informed MALDI-TOF MS diagnostics reveal importance of *Klebsiella oxytoca* group in invasive infections: a retrospective clinical study

Cuénod *et al. Genome Medicine* (2021) 13:150
https://doi.org/10.1186/s13073-021-00960-5

Genome Medicine

RESEARCH Open Access

Whole-genome sequence-informed MALDI-TOF MS diagnostics reveal importance of *Klebsiella oxytoca* group in invasive infections: a retrospective clinical study

Aline Cuénod[1,2*], Daniel Wüthrich[1,2,3], Helena M. B. Seth-Smith[1,2,3], Chantal Ott[1,2], Christian Gehringer[1,2,4], Frédéric Foucault[5], Roxanne Mouchet[5], Ali Kassim[6], Gunturu Revathi[6], Deborah R. Vogt[7], Stefanie von Felten[7,8], Stefano Bassetti[4], Sarah Tschudin-Sutter[9,10], Timm Hettich[11], Götz Schlotterbeck[11], Christina Homberger[1,2], Carlo Casanova[12], Jacob Moran-Gilad[13,14], Orli Sagi[13,14], Belén Rodríguez-Sánchez[15,16], Franco Müller[17], Martina Aerni[17], Valeria Gaia[18], Helke van Dessel[19], Greetje A. Kampinga[20,21], Claudia Müller[22], Claudia Daubenberger[23,24], Valentin Pflüger[5] and Adrian Egli[1,2*]

* corresponding authors:
Aline Cuénod
Prof. Dr. Dr. Adrian Egli, MD PhD

This manuscript has been published by *BMC Genome Medicine*:

Cuénod et al. "Whole-Genome Sequence-Informed MALDI-TOF MS Diagnostics Reveal Importance of *Klebsiella oxytoca* Group in Invasive Infections: A Retrospective Clinical Study." *Genome Medicine* 13, no. 1 (September 13, 2021): 150. https://doi.org/10.1186/s13073-021-00960-5.

My contributions:

- Conceptualisation of the study
- Coordination of collaborators
- Compilation of >30,000 MALDI-TOF mass spectra from 8 healthcare centres and >7,500 AMR profiles from 2 healthcare centres
    - Communication and coordination of the data collection
    - Data curation
- Characterisation of 50 *Klebsiella* spp. strains:
    - Acquisition of MALDI-TOF mass spectra
    - Acquisition of AMR profiles
    - Biochemical profiling
- Part of the bioinformatic analysis of whole genome sequences and MALDI-TOF mass spectra. Contributions to:
    - Pre-processing of MALDI-TOF mass spectra (peak picking, calibration)
    - Species identification from MALDI-TOF mass spectra
    - Validation of the species identification from MALDI-TOF mass spectra
    - Species identification of the *Klebsiella* spp. assemblies
    - Screen *Klebsiella spp.* assemblies for AMR factors
- Visualisation of the data (Figure 3, Figure S1, Figure S3, Figure S5)
- Writing of the original manuscript.

I presented this project as Posters at the following international conferences:

- European Congress of Clinical Microbiology and Infectious Diseases (*ECCMID*), 2019 (Amsterdam, the Netherlands)
- American Society for Microbiology Microbe (*ASM Microbe*) Conference 2019 (San Francisco, USA)

Note: The following part contains the full publication

The Supplementary material can be accessed via the following link:
https://doi.org/10.1186/s13073-021-00960-5

Whole-genome sequence-informed MALDI-TOF MS diagnostics reveal importance of *Klebsiella oxytoca* group in invasive infections: a retrospective clinical study

Aline Cuénod[a,b,*], Daniel Wüthrich[a,b,c], Helena M.B. Seth-Smith[a,b,c], Chantal Ott[a,b], Christian Gehringer[a,b,d], Frédéric Foucault[e], Roxanne Mouchet[e], Ali Kassim[f], Gunturu Revathi[f], Deborah R. Vogt[g], Stefanie von Felten[g,h], Stefano Bassetti[d], Sarah Tschudin-Sutter[i,j], Timm Hettich[k], Götz Schlotterbeck[k], Christina Homberger[a,b], Carlo Casanova[l], Jacob Moran-Gilad[m,n], Orli Sagi[m,n], Belén Rodríguez-Sánchez[o,p], Franco Müller[q], Martina Aerni[q], Valeria Gaia[r], Helke van Dessel[s], Greetje A. Kampinga[t,u], Claudia Müller[v], Claudia Daubenberger[w,x], Valentin Pflüger[e], Adrian Egli[a,b,*]

[a] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland

[b] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland

[c] Swiss Institute for Bioinformatics, Basel, Switzerland

[d] Division of Internal Medicine, University Hospital Basel, Basel, Switzerland

[e] Mabritec AG, Riehen, Switzerland

[f] Aga Khan University Hospital, Nairobi, Kenya

[g] Department of Clinical Research, University of Basel and University Hospital Basel, Basel, Switzerland

[h] Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Zurich, Switzerland

[i] Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland.

[j] Department of Clinical Research, University of Basel, Basel, Switzerland

[k] Division of Instrumental Analytics, School of Applied Sciences (FHNW), Muttenz, Switzerland

[l] Institute for Infectious Diseases, University of Bern, Bern, Switzerland

[m] Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva

[n] Soroka University Medical Center, Beer Sheva, Israel

[o] Hospital General Universitario Gregorio Marañón, Madrid, Spain.

[p] Servicio de Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain. Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain

[q] Labor Team W AG, Goldach, Switzerland.

[r] Servizio di microbiologia EOLAB, Ente Ospedaliero Cantonale, Bellinzona, Switzerland

[s] Department of Medical Microbiology, Maastricht University Medical Center, Maastricht, the Netherlands.
[t] Department of Medical Microbiology & Infection prevention, University of Groningen, Groningen, the Netherlands
[u] University Medical Center Groningen (UMCG), Groningen, the Netherlands
[v] Laborgemeinschaft 1, Zurich, Switzerland.
[w] Swiss Tropical and Public Health Institute, Basel, Switzerland.
[x] Department of Sciences, University of Basel, Basel, Switzerland.

* Corresponding authors:
Aline Cuénod
Applied Microbiology Research
Department of Biomedicine
University of Basel
Hebelstrasse 20
4031 Basel
aline.cuenod@stud.unibas.ch

Adrian Egli, MD PhD
Clinical Bacteriology & Mycology
University Hospital Basel
Petersgraben 4
4031 Basel
adrian.egli@usb.ch

Abstract:

Background: *Klebsiella* spp. are opportunistic pathogens which can cause severe infections, are often multi-drug resistant and a common cause of hospital acquired infections. Multiple new *Klebsiella* species have recently been described, yet their clinical impact and antibiotic resistance profiles are largely unknown. We aimed to explore *Klebsiella* group- and species-specific clinical impact, antimicrobial resistance (AMR) and virulence.
Methods: We analysed whole genome sequence data of a diverse selection of *Klebsiella* spp. isolates, and identified resistance and virulence factors. Using the genomes of 3,594 *Klebsiella* isolates, we predicted the masses of 56 ribosomal subunit proteins and identified species-specific marker masses. We then re-analysed over 22,000 MALDI-TOF mass spectra routinely

acquired at eight healthcare institutions in four countries looking for these species-specific markers. Analyses of clinical and microbiological endpoints from a subset of 957 patients with infections from *Klebsiella* species were performed using generalized linear mixed-effects models.

Results: Our comparative genomic analysis shows group- and species-specific trends in accessory genome composition. With the identified species-specific marker masses, eight *Klebsiella* species can be distinguished using MALDI-TOF MS. We identified *K. pneumoniae* (71.2%;n=12,523), *K. quasipneumoniae* (3.3%;n=575), *K. variicola* (9.8%;n=1,717), *"K. quasivariicola"* (0.3%;n=52), *K. oxytoca* (8.2%;n=1,445), *K. michiganensis* (4.8%;n=836), *K. grimontii* (2.4%;n=425), and *K. huaxensis* (0.1%;n=12). Isolates belonging to the *K. oxytoca* group*,* which includes the species *K. oxytoca, K. michiganensis* and *K. grimontii,* were less often resistant to 4th generation cephalosporins than isolates of the *K. pneumoniae* group, which includes the species *K. pneumoniae, K. quasipneumoniae*, *K. variicola* and *"K. quasivariicola"* (Odds Ratio=0.17, p<0.001, 95% Confidence Interval [0.09,0.28]). Within the *K. pneumoniae* group, isolates identified as *K. pneumoniae* were more often resistant to 4th generation cephalosporins than *K. variicola* isolates (Odds Ratio=2.61, p=0.003, 95% Confidence Interval [1.38,5.06])*. K. oxytoca* group isolates were found to be more likely associated with invasive infection to primary sterile sites than *K. pneumoniae* group isolates (Odds Ratio=2.39, p=0.0044, 95% Confidence Interval [1.05,5.53]).

Conclusions: Currently misdiagnosed *Klebsiella* spp. can be distinguished using a ribosomal marker-based approach for MALDI-TOF MS. *Klebsiella* groups and species differed in AMR profiles, and in their association with invasive infection, highlighting the importance for species identification to enable effective treatment options.

Keywords: MALDI-TOF MS, *Klebsiella spp.*, invasive infections, antimicrobial resistance, species identification

Background

Klebsiella spp. are opportunistic pathogens, resident as respiratory and intestinal microbiota, and are commonly isolated during severe infections such as sepsis, pneumonia, and pyelonephritis (218,219). Particularly hypervirulent strains of *K. pneumoniae*, which have been linked to specific capsular factors resulting in a muco-viscous phenotype, cause pyogenic liver abscesses and sepsis (220,221). In addition, the number of multi-drug resistant (MDR) isolates is increasing globally, carrying plasmids encoding for extended spectrum beta-lactamase (ESBL) or carbapenemase genes (222,223). *K. pneumoniae* is part of the ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*,

Acinetobacter baumannii, *Pseudomonas aeruginosa*, *Enterobacter*) which were classified by the WHO in 2017 as critical priority pathogens for research and development of new antibiotic treatment modalities (140).

The taxonomy of the genus *Klebsiella* has been in flux for the past few years, divided into two main groups: the *K. pneumoniae* and the *K. oxytoca* group. Recently, several new species have been described within the *K. pneumoniae* group, which includes: *K. pneumoniae* (sensu stricto), *K. quasipneumoniae* (224), *K. variicola* (106), *"K. quasivariicola"* (225) and *K. africana* (226). The *K. oxytoca* group comprises *K. oxytoca* (sensu stricto)*, K. michiganensis* (107), *K. grimontii* (109), *K. spallanzanii* and *K. pasteurii* (105). *K. huaxensis* (108), is more closely related to the *K. oxytoca* group than to the *K. pneumoniae* group, but forms a distinct clade (**Figure 1A**). Most of the *Klebsiella* spp. have been observed within a clinical context (105,108,109,225,227,228). The *K. oxytoca* group has been reported to cause antibiotic-associated haemorrhagic colitis in neonates, and hospital acquired infections such as pneumonia and urinary tract infections (UTI) (130,229). Previous studies have analysed the population structure of a subset of *Klebsiella* spp. (141,225,230). The gained knowledge from these studies could however not yet be translated to clinical routine diagnostics and the clinical relevance of these recently described *Klebsiella* spp. is still unclear.

The newly described *Klebsiella* spp. are not yet identified with routine hospital based diagnostic procedures. Biochemical reaction profiling cannot distinguish between all *Klebsiella* spp. (51,227), neither is 16S rRNA a good sequencing target for species distinction within *Enterobacteriaceae* (231). The most widely used technology for bacterial species identification in microbiological routine diagnostics is Matrix Assisted Laser Desorption Ionization - Time Of Flight Mass Spectrometry (MALDI-TOF MS) (63). The two commonly used commercial databases (MALDI Biotyper (MALDI Biotyper Compass Library, Revision E (Vers. 8.0, 7854 MSP, RUO) Bruker Daltonics, Bremen, Germany) and VitekMS DB (v.3.2, bioMérieux, Marcy-l'Étoile, France) allow spectral identification of *K. pneumoniae, K. variicola* and *K. oxytoca.* Importantly, *"K. quasivariicola"*, *K. quasipneumoniae*, *K. africana, K. michiganensis, K. grimontii*, *K. pasteurii, K. spallanzanii,* and *K. huaxensis* are currently not included in these databases, and strains of these species are wrongly identified as either *K. pneumoniae* or *K. oxytoca* using MALDI-TOF MS (227). Fortunately, recent developments show that a distinction of *Klebsiella* spp. is possible in routine diagnostics, using Fourier-transform infrared spectrometry (232) and MALDI-TOF MS using alternative databases (233).

Ribosomal subunit proteins are suitable as phylogenetic protein markers for MALDI-TOF mass spectra, as they are highly abundant in replicating cells and of relatively low molecular weight (234,235). Combinations of ribosomal subunit protein derived masses allow the separation of sub-lineages within *Escherichia coli* (83) and *Streptococcus agalactiae* (76) by MALDI-TOF MS.

The aim of our study was to investigate the clinical presentation and distribution of AMR and virulence across the genus *Klebsiella*. Furthermore, using whole genome sequences we aimed to develop a ribosomal subunit based MALDI-TOF MS scheme to robustly distinguish between *Klebsiella* spp. in clinical routine and to apply this on a large international dataset.

Methods

An outline of the study method is given in **Additional file 1: Figure S1**.

Ethics

Bacterial strains have been collected in clinical routine diagnostics. The collection of bacterial strains and their analysis for diagnostic assay development do not fall under the Swiss human research act and no ethical approval nor consent to participate from patients was required. The analysis of patient demographic and clinical outcome data was approved by the 'Ethikkommission Nordwest- und Zentralschweiz' (EKNZ) (BASEC-Nr. 2016-01899 and 2018-00225) for patients who did not reject the hospitals general research consent. Patients who did reject the hospital's general consent were excluded from all analyses which include patient demographic and clinical outcome data.

Bacterial isolates and whole genome sequencing (WGS)

261 *Klebsiella* spp. isolates were collected from various tissue sources (see **Additional file 2: Table S1** for more details) at three routine diagnostic laboratories in Switzerland including the University Hospital of Basel (USB; Basel, Switzerland), Mabritec AG (Riehen, Switzerland), and Labor Team W AG (LTW; Goldach, Switzerland)). Isolates were grown on Columbia 5% Sheep Blood Agar (bioMérieux, Marcy-l'Étoile, France) and DNA was extracted using the QIACube with the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany). After quality control of the DNA by Tapestation (Agilent, Santa Clara, USA), tagmentation libraries were generated as described by the manufacturer (Nextera XT kit, Illumina, San Diego, USA). The genomes were sequenced under 24x multiplexing using a 2 × 300 base pairs V3 reaction kit on an Illumina MiSeq instrument reaching an average coverage of approximately 60-fold for all isolates. 11 isolates, covering reference and clinical isolates of 6 species were additionally sequenced on a PacBio Sequel at the Functional Genomics Center Zurich (FGCZ, ETH Zurich, Switzerland).

All available whole genome assemblies designated as *Klebsiella* spp. were downloaded from NCBI in December 2017 (n=3,047), representing members of the species *K. pneumoniae, K. quasipneumoniae, K. variicola, "K. quasivariicola", K. oxytoca, K. michiganensis, K. grimontii,*

K. huaxensis, *K. aerogenes* and three species of the genus *Raoultella* (*R. ornithinolytica*, *R. planticola* and *R. terrigena*). An additional selection of publicly available *K. pneumoniae* whole genome sequences were included from NCBI SRA, which was sampled to maximize diversity (n=286) (141). Two sets of *Klebsiella* spp. genomes were used for this study: first, a total of n=3,594 publicly available genome sequences including the 3,333 described above, and the 261 sequenced at the USB, were used to in silico predict ribosomal protein masses. The species identity of these genome sequences was determined by comparison to the type-strains of *K. pneumoniae, K. quasipneumoniae*, *K. variicola*, "*K. quasivariicola*"*, K. oxytoca*, *K. michiganensis*, *K. grimontii, K. huaxensis R. ornithinolytica, R. planticola and R. terrigena* using Average Nucleotide Identity (ANIm) and a threshold of 96%. Second, a computationally more manageable subset of these genomes (n=999) was used for comparative genomic analyses, selected to represent the largest genomic diversity between and within species, and geographically. This subset included all assemblies of *K. quasipneumoniae*, *K. variicola*, *"K. quasivariicola"*, *K. oxytoca*, *K. michiganensis*, *K. grimontii,* and *K. huaxensis*. For *K. pneumoniae*, only strains sequenced at USB and the previously published, diverse set of sequences (141) were included. To avoid bias introduced by outbreak strains, we excluded genomes which shared ANIm values > 99.9% with another genome in the collection, resulting in a final dataset of n=548 genomes. Both datasets, including accession numbers and those of the short and long reads sequenced for this study, can be found in **Additional file 2: Table S1**. *K. africana*, *K. pasteurii,* and *K. spallanzanii* were not included in this analysis as the species were not published at the time of the analysis and are extremely rare in clinics.

Comparative genomic analysis

WGS data was quality controlled using FastQC (236) and MetaPhlAn (v2.0) (237) and adaptors were trimmed using Trimmomatic (238). Genome assemblies were created using Unicycler (v0.4.4) (239). Prokka (1.12) (240) was used for annotation. Orthologous groups were built using Roary (v3.10.2, option: -i 90) (241). The resulting core-genome alignment was used for the construction of a phylogenetic tree using FastTree (v2.1) (242,243). The sizes of the core- and pan-genomes were calculated using a python script () (244).

The O-loci and K-loci were determined using KLEBORATE (v0.3.0) (245–247). The genomes were investigated for the presence of known virulence loci (those included in KLEBORATE and the cytotoxin tilivalline (131)) and AMR determinants (via KLEBORATE). Potential plasmids were detected by comparing the genomes to the PlasmidFinder database (248) using abricate (249). Genomic analyses were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel.

Scripts generating figures from the output of these tools were deposited on GitHub (244).

In silico prediction of ribosomal subunit protein masses from WGS data

The molecular weight of 56 ribosomal subunits were predicted in silico as described [26]: Tblastn (v 2.2.31+) was used to extract the amino acid sequences of 56 ribosomal subunits from 3,594 *Klebsiella* spp. assemblies. Full ribosomal subunit sequences were retained when start and stop codons were identified and the length was within the median ±3 codons. The subunits L1, L2 and S12 were not found in over 90% of the genomes and were therefore excluded from further analysis. The ribosomal subunit protein S1 was also excluded because S1-like domains are found in proteins unrelated to the ribosome [40]. The masses of the ribosomal subunit proteins alleles were predicted including the N-end rule to account for possible methionine loss at the N-terminus (250). The mass of subunit L33 was corrected by 15 Daltons to account for post translational methylation (251).

Definition of species-specific MALDI-TOF MS marker masses

A diverse selection of bacterial isolates (n=50) representing at least eight isolates of *K. pneumoniae*, *K. variicola*, *K. oxytoca*, *K. michiganensis,* and *K. grimontii,* whole genome sequenced for this study, were used to validate the detection of the predicted marker masses in MALDI-TOF mass spectra. These represent the most common species within the *K. pneumoniae* and the *K. oxytoca* group (141,222,230). MALDI-TOF mass spectra of these 50 *Klebsiella* isolates were acquired on four MALDI-TOF MS systems in different laboratories, including one microflex Biotyper (Bruker Daltonics, Bremen, Germany) at the USB (Basel, Switzerland), one Axima Confidence (Shimadzu, Ngoyo, Japan) at Mabritec AG (Riehen, Switzerland) and two VitekMS devices (bioMérieux, Marcy-l'Étoile, France) at the Laborgemeinschaft 1 (LG1) (Zürich, Switzerland) and the Scuola universitaria professionale della Svizzera italiana (Bellinzona, Switzerland). The *Klebsiella* isolates were measured on each system in quadruplicate using direct smear method and overlaid with formic acid (70% for the spectra acquired on the Microflex Biotyper and 25% for all other machines) and cyano-4-hydroxycinnamic acid (CHCA) matrix solution. MALDI-TOF mass spectra acquired on the VitekMS (bioMérieux, Marcy-l'Étoile, France) (n=400 spectra) were output as .mzml files containing a list of peaks per spectrum. For MALDI-TOF mass spectra acquired on the Axima Confidence (n=200 spectra) peak picking was performed using the Launchpad Software (v2.8, Shimadzu, Ngoyo, Japan) (parameters: scenario: 'Advanced'; peak width: 80 chans; smoothing method: 'Average'; smoothing filter width: 500 chans; peak detection method: 'Threshold-Apex'; threshold type: 'dynamic'; threshold offset: 0.025 mV; threshold response 1.25x). MALDI-TOF mass spectra acquired on a microflex Biotyper (n=200 spectra) were output as fid-files and peak picking was performed in the flexAnalyses software (v3.4)

(parameters: peak detection: 'Centroid'; signal to noise threshold: 2; relative intensity threshold: 0%; minimal intensity threshold: 600; maximal number of peaks: 300; peak width: 4 m/z; peak height: 90%; baseline subtraction: 'TopHat'; smoothing algorithm: 'Savitzky Golay'; smoothing width: 2 m/z; smoothing cycles: 10). All MALDI-TOF mass spectra were internally calibrated with the conserved masses 4365.3 m/z, 6383.5 m/z, 7158.7 m/z, 7244.5 m/z, 10286.1 m/z and a tolerance range of 1,000 ppm, using the R-packages MALDIQuant and MALDIQuantForeign (252). All spectra were exported in ASCII format and interrogated for the ribosomal subunit protein allele masses predicted from the respective WGS data. : We used the following criteria to select the ribosomal target proteins for subsequent species identification, in order to maximise discriminatory power and reproducibility: ribosomal subunit protein masses in the mass range from 3,000-13,000 Da (L36, S22, L34, L30, L32, L33, L35, L29, L31, S21, L27, S20, S15, S19, L25, S14, L21, L18) were selected if they were detected with a reproducibility >80% in a least one center, with the exception of ribosomal subunit L28, which had a maximal reproducibility of 57%. Additionally, ribosomal subunit protein masses in the higher mass range from 13,000-15,000 Da (L19, S13, L20, S8, L17, S9) were included, as they were detected with a reproducibility of at least 35% in at least one center. Ribosomal subunits with a predicted molecular weight >15'000 Da were not included for further analysis. The bacterial species was assigned for which most marker masses could be detected in the acquired mass spectrum. If in a spectrum an equal number of marker masses from different *Klebsiella* species were found, the spectrum was assigned a multi-species ID (e.g. *K. michiganensis / K. oxytoca*) and labelled as "Multispecies ID only".

Classification of MALDI-TOF mass spectra acquired in routine microbiology diagnostics

Routinely acquired MALDI-TOF mass spectra (n=33,160) from eight international healthcare institutions from four countries were analysed: the Soroka Medical Centre (SMC; Beer Sheva, Israel); the Hospital General Universitario Gregorio Marañón (HGUM; Madrid, Spain); the Servizio di microbiologia EOLAB, Ente Ospedaliero Cantonale (Bellinzona, Switzerland); the LG1, (Zurich, Switzerland); the LTW (Goldach, Switzerland); the USB (Basel, Switzerland); the Maastricht University Medical Center (MUM; Maastricht, the Netherlands); and the University Medical Center Groningen (UMCG; Groningen, the Netherlands). The MALDI-TOF mass spectra were processed as described above. Each routinely acquired spectrum was interrogated for the presence of the reproducibly detected ribosomal protein subunit derived mass combinations, with an accepted error range of 300 ppm. 10,814 MALDI-TOF mass spectra were from duplicated bacterial isolates and excluded from further analysis, leading to a final number of 22,346 MALDI-TOF mass spectra representing unique bacterial isolates.

Phenotypic profiling

The same collection of *Klebsiella* spp. strains (n=50), which were used to define species-specific marker masses, were subjected to biochemical profiling on a Vitek2 (bioMérieux, Marcy-l'Étoile, France) using the GN ID card and the API50ch panel (bioMérieux, Marcy-l'Étoile, France). Primary metabolites of the same strains were measured and analysed as described in **Additional file 3: Supplementary Methods.** Additionally, 11 strains were subject to fatty acid profiling as described in **Additional file 3: Supplementary Methods.** These 11 strains included reference strains of the species *K. pneumoniae, K. quasipneumoniae, K. variicola, K. oxytoca, K. michiganensis*, one clinical isolate for each of the species *K. pneumoniae, K. variicola, K. oxytoca, K. michiganensis*, and two clinical isolates of the species *K. grimontii.*

Antimicrobial resistance determination

AMR profiles of isolates associated with the retrospectively analysed spectra were accessed through the laboratory information systems of the USB and the LTW (n= 7,876). The accessed AMR profiles were measured in clinical routine diagnostics from January 2015 to June 2018 using either microdilution methods (Vitek2, AST-N242 GN Cards, bioMérieux), MIC strip tests (Liofilchem, Roseto degli Abruzzi, Italy), or disc diffusion tests (ThermoFisher Scientific, Waltham, USA). Breakpoints were interpreted as susceptible or resistant according to the current EUCAST Breakpoint table (v6.0 – 8.1) (253).

Retrospective assessment and statistical analysis of clinical data

We assessed the relative distribution of *Klebsiella* spp. with regards to laboratory and country of isolation (n = 22,346, including spectra from eight laboratories). We examined the association of *Klebsiella* groups and species with resistance to antibiotic classes using logistic regression (post hoc analyses; n = 7,876, including spectra from two laboratories).

Patient demographic and clinical data from patients with *Klebsiella* spp. infections were retrospectively accessed via the USB clinical information system in a case report form for a subset of clinical cases (n = 957). Inclusion criteria were: patients for which at least one isolated bacterial colony was identified as *Klebsiella* spp. by MALDI-TOF MS collected between January 2015 and June 2018 at the USB and who did not reject the hospital's general research consent form, as approved by the ethical committee. The USB is a tertiary healthcare center with more than 750 beds in a low endemic region for ESBL-producing bacteria (254).

The clinical outcomes of 957 patients with *Klebsiella* infections were analysed and included all-cause mortality within 30 days from *Klebsiella* spp. diagnosis as primary endpoint and

secondary endpoints: (i) time to death after *Klebsiella* spp. diagnosis in days, (ii) admission to an intensive care unit (ICU), (iii) invasive infection to sterile sites (including the bloodstream, deep tissues, and cerebrospinal fluids), (iv) length of hospital stay in days, (v) the number of medical disciplines involved to manage the specific case (as a surrogate marker for case complexity), and (vi) whether the infection was mentioned in the patient letters. Clinical outcomes were examined for an association with distinct *Klebsiella* spp., age, sex, immunosuppression (defined as a dose equivalent of 20 mg prednisone / day or mentioning of immunosuppression in the patient notes), Charlson Comorbidity Index (CCI) (255), resistance to 3rd generation cephalosporins, and antibiotic treatment (defined as at least one dose of antimicrobial agent at hospital entry or during the hospital stay). Binary outcomes were analysed using generalized linear mixed models (GLMM) with binomial error distribution. Count outcomes (number of medical disciplines involved) were analysed using GLMM with Poisson error distribution. Time to death within hospital and length of hospital stay (time to discharge) were considered competing risks and jointly analysed by a competing risks model. For further detail see **Additional file 3: Supplementary Methods.**

Results

Comparative genomic analyses of *Klebsiella* spp.

To determine the core-genome of the genus *Klebsiella*, from the 3,594 *Klebsiella* spp. isolates with available WGS, a selection of 548 isolates was made, reflecting between- and within-species diversity. The genus core-genome comprises n=1,171 genes, which are genes shared between 99% of these 548 isolates. The core-genome based phylogeny (**Figure 1A**) clearly shows the *Klebsiella* phylogeny as previously described (222,230,256), with subspecies within *K. quasipneumoniae (K. quasipneumoniae* subsp. *quasipneumoniae* and *K. quasipneumoniae* subsp. *similipneumoniae*) as well as within *K. variicola* (*K. variicola* subsp. *variicola* and *K. variicola* subsp. *tropica*) can be distinguished. These species also contain a diverse array of strains, in contrast to *K. pneumoniae*, which is more homogeneous when comparing these 1,171 core genes. Interestingly, *K. grimontii* includes two deeply branching sub clades, which have not yet been described as subspecies (**Figure 1A**).

The pan-genomes of the *K. pneumoniae* group and the *K. oxytoca* group were determined by investigating the number of unique orthologous clusters within our genome data set (**Figure 1B**). A larger pan- to core-genome ratio could suggest adaptation to diverse environments, whereas a smaller pan- to core-genome ratio could, in a clinical context, reflect adaptation to the human host or even site-specific infections. Both groups show an open pan-genome, with the number of unique orthologous clusters increasing as more genomes are added to the analysis. There seems to be a larger increase in the pan-genome of the *K. oxytoca* group with

strains added, although fewer genomes have been sampled to date. At a species level, pan-genome sizes of those within the *K. pneumoniae* group are relatively similar, whereas *K. michiganensis* within the *K. oxytoca* group shows a larger pan-genome, although this may reflect sampling bias (**Additional file 1: Figure S2**).

Plasmid complements are known to vary widely between *Klebsiella* isolates, and can carry accessory genes involved in AMR and pathogenicity (45,257,258). We detected a lower median number of plasmid replicons per isolate within the *K. oxytoca* group (median=2, interquartile range (IQR) = 0-4) compared to isolates of the *K. pneumoniae* group (median=4, IQR=2-4) (**Figure 1C**). The lowest median number of plasmids was detected for *K. grimontii* isolates (median=1, IQR=0-5), and the highest median number of plasmids detected in *"K. quasivariicola"* (median=7, IQR=5-10) isolates. We also observed that within the *K. pneumoniae* group, the median count of plasmids detected was lower in *K. variicola* isolates (median=3, IQR=0-5) than in isolates of the other species of the *K. pneumoniae* group (median=4, IQR=3-6 for *K. pneumoniae*; median=5, IQR=2-7 for *K. quasipneumoniae* and median=7, IQR=5-10 for *"K. quasivariicola"*) (**Figure 1C**). No group specificity could be detected in plasmid replicon profiles (**Figure 1D**). The two replicons known to be particularly related to virulence, plasmids KpVP-1 and KpVP-2, (IncHI1B_pNDM-MAR and IncFIB(K)_Kpn3 respectively), were detected in isolates from all *Klebsiella* spp. (*K. pneumoniae* n=162/218, 73.3%; *K. quasipneumoniae* n=61/83, 73.5%; *K. variicola* n=56/109, 51.4%; *"K. quasivariicola"* n=4/5, 80.0%; *K. oxytoca* n=18/41, 43.9%; *K. michiganensis* n=23/54, 42.6%; *K. grimontii* n=16/37, 43.2%) with the exception of *K. huaxensis*, for which only a single genome was available for this study.

From long read genome assemblies for a subset of 11 isolates sequenced as part of this study, the plasmids assembled for nine isolates agreed with the replicon findings. In *K. quasipneumoniae* DSM-2811T we detected a plasmid of 4,267 bp with 99.1% identity and 71% coverage to the *K. pneumoniae* plasmid pB1019, which was not identified by PlasmidFinder. A previously undescribed plasmid of 3,596 base pairs was identified within *K. grimontii* 606641-17, which does not carry any known virulence or resistance factors, showing that there is further diversity of *Klebsiella* spp. plasmids to discover.

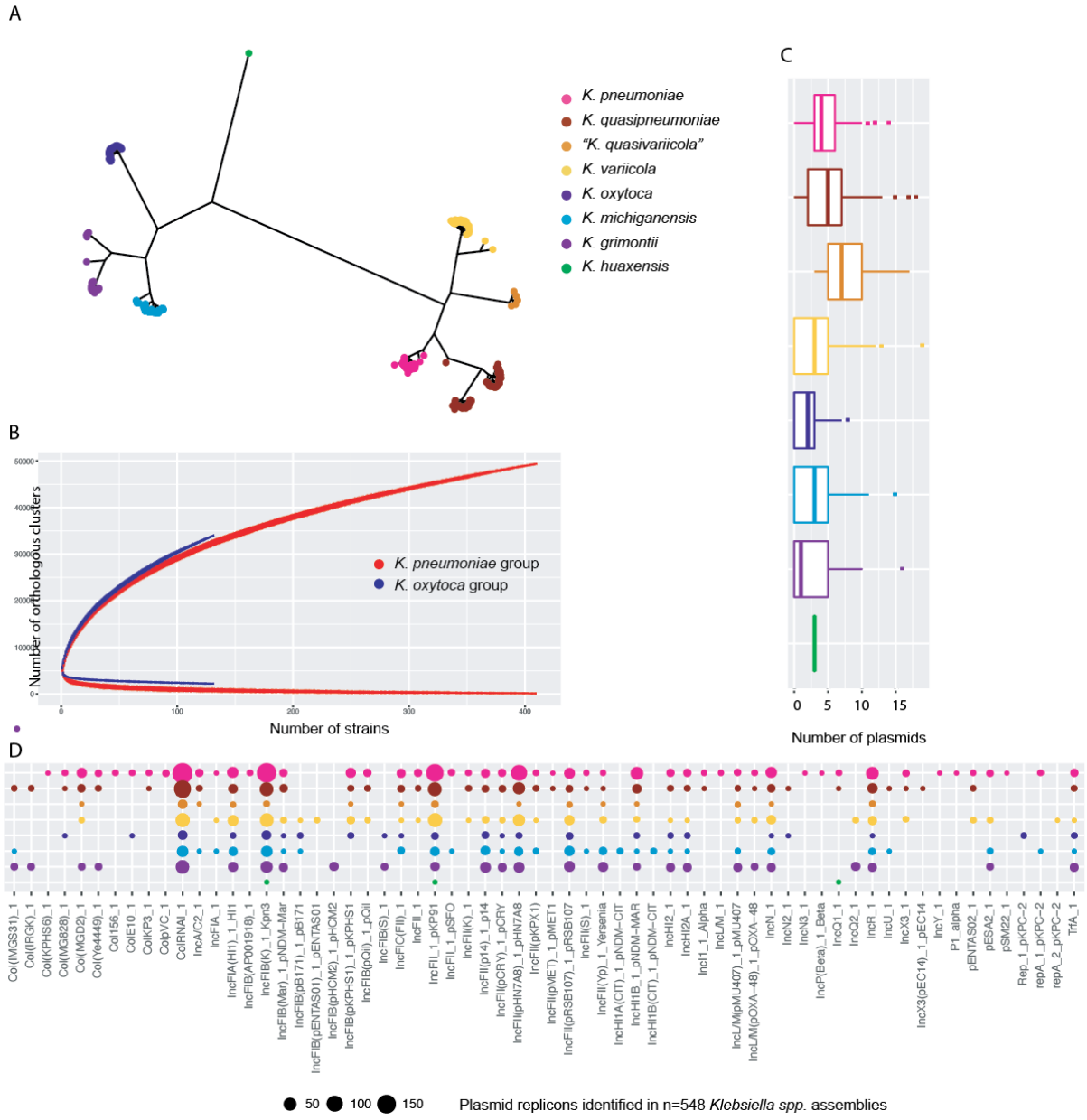

Figure 1. Genomic content of isolates across the *Klebsiella* genus. A. Core-genome phylogeny of the genus *Klebsiella* including isolates from *K. pneumoniae* (n=218), *K. quasipneumoniae* (n=83), *"K. quasivariicola"* (n=5), *K. variicola* (n=109), *K. oxytoca* (n=41), *K. michiganensis* (n=54), *K. grimontii* (n=37) and *K. huaxensis* (n=1) based on 1,171 core genes. The species colour key is used throughout the figure and paper. **B.** Pan- (upper) and core- (lower) gene accumulative curves comparing *K. pneumoniae* group and *K. oxytoca* group. **C.** Number of plasmid replicons identified in each isolate, per species. Boxes indicate the IQR with the median displayed as middle lines. **D.** Plasmid replicons identified by PlasmidFinder in all isolates (n=548), shown per *Klebsiella* spp.

Virulence factors of *Klebsiella* spp.

Virulence factors were investigated by comparing 548 *Klebsiella* spp. genomes against known databases and virulence factors. Genes encoding the iron-chelating siderophores aerobactin

and salmochelin were detected in a minority of isolates (n=18/548, 3.3% and n=21/548, 3.8% of isolates, respectively), only within *K. pneumoniae* and *K. quasipneumoniae*, often co-occurring within isolates. The siderophore yersiniabactin is more prevalent in isolates of the *K. oxytoca* group (n=111/132, 84.1%) than in the *K. pneumoniae* group (n=52/415, 12.5%). The *kfu* operon, encoding an iron transport system, was detected in isolates of all species except *K. oxytoca* (n=333/548; 60.8% of all isolates and n=0/41; 0% within the species *K. oxytoca*).

The genes involved in allantoin metabolism, which enable *Klebsiella* spp. to assimilate nitrogen from this metabolic intermediate and increase its virulence in certain infection sites, were detected in isolates of all species (n=132/548; 24.1%), notably in all *K. oxytoca* (n=41/41; 100%) and *K. grimontii* (n=37/37; 100%) isolates. The distribution of bacterial toxin operons encoding microcin, colibactin and tilivalline was also examined. The complete microcin operon was detected in a few *K. pneumoniae* genomes (n=8/218; 3.7%) and in one *K. michiganensis* genome (n=1/54; 1.9%), whereas colibactin was detected solely in *K. pneumoniae* (n=11/218; 5.0%). In contrast, the complete tilivalline operon was exclusively identified in isolates within the *K. oxytoca* group (n=64/132; 48.5%) particularly in isolates of *K. oxytoca* (n=29/41; 59.2%) and *K. grimontii* (n=30/37; 81.1%) (**Figure 2A**).

The regulator of the mucoid phenotype *rmp* genes were detected in *K. pneumoniae* (n=14/218; 6.4%) and *K. quasipneumoniae* (n=1/83; 1.2%). Ecotin has been described as being able to modulate the host immune response (259), and was detected in isolates belonging to all species (n=445/548; 81.2%).

We found a high diversity of K-loci in our dataset, with 114 capsule types. KL107 was found to be the most prevalent K-locus type across all *Klebsiella* spp. (107/548; 19.5%), with the exception of *K. huaxensis* where only one genome was included (KL46; **Figure 2B**). KL1 and KL2, which are associated with a muco-viscous phenotype and frequently detected in hypervirulent strains, were exclusively detected in isolates of species within the *K. pneumoniae* group; KL1 was detected in isolates of *K. pneumoniae* (n=4/218; 1.8%) and *K. quasipneumoniae* (n=4/83; 4.8%); whereas KL2 was detected in isolates of *K. pneumoniae* (n=10/218; 4.6%) and *K. variicola* (n=1/109; 0.9%). Within the O-loci we found less diversity, with 17 types. Type O1v1 is the most common in isolates of *K. pneumoniae* (n=53/218, 24.3%), *K. michiganensis* (n=41/54; 75.9%), and *K. grimontii* (n=36/37; 97.2%), whereas most isolates of *K. oxytoca* (n=22/41; 53.7%) carry the O-locus OL104. The most common combination of K- and O-loci per species were as follows: KL107-O2v1, KL107-OL101 and KL64-O1v1 in *K. pneumoniae* (each 7/218; 3.2%), KL48-O5 in *K. quasipneumoniae* (n=4/83; 4.8%) and KL107-O5 in *K. variicola* (n=6/109; 5.5%). All five *"K. quasivariicola"* genomes studied carry unique combinations of K- and O-loci. Within the *K. oxytoca* group, the most common combination of K- and O-loci is carried by a bigger proportion of the isolates

compared to the *K. pneumoniae* group species: KL68-O5 in *K. oxytoca* (n=7/41; 17.1%), KL107-O1v1 in *K. michiganensis* (n=11/54; 20.4%) and *K. grimontii* (n=14/37; 41.2%) were the most frequently detected combinations.

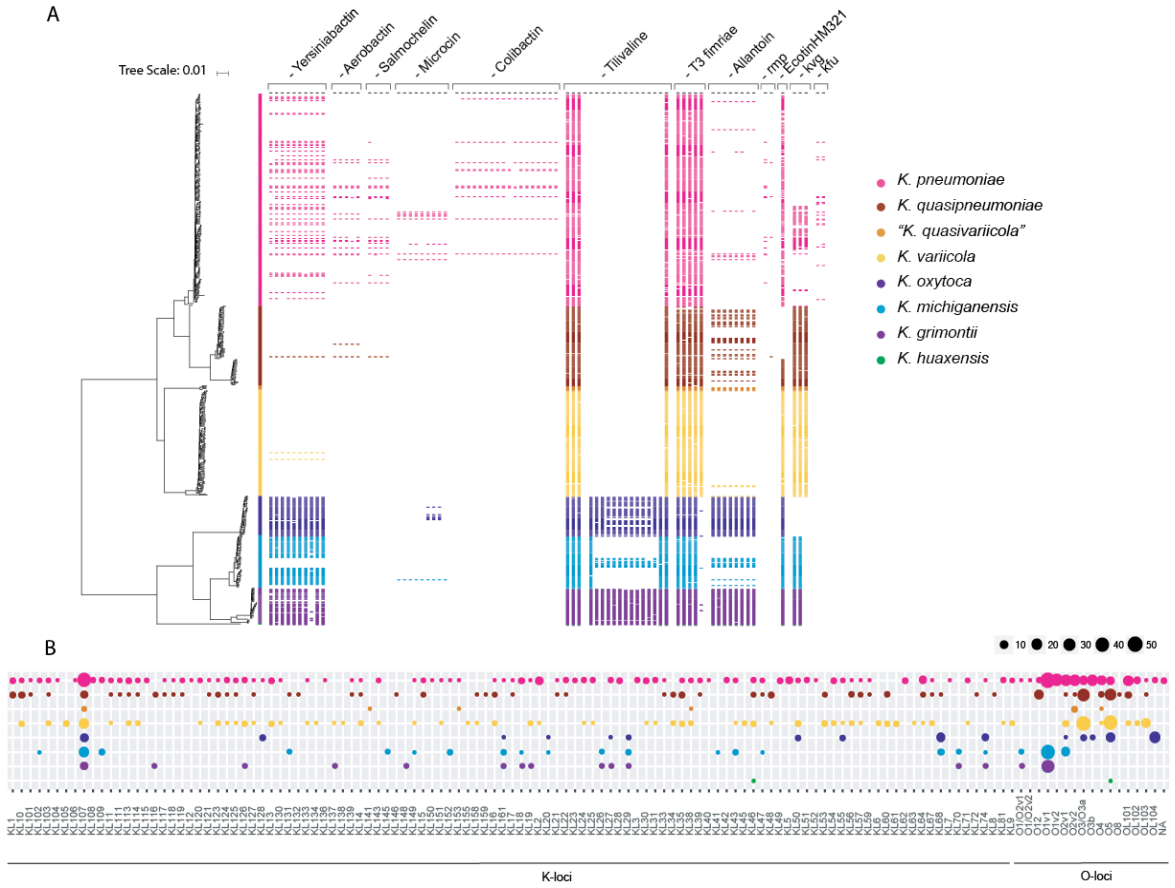

Figure 2. Virulence factors across the genus *Klebsiella*. A. Core-genome phylogeny of 548 *Klebsiella* spp. genomes (left, in line with Figure 1A) with identified virulence related genes shown per isolate (right), coloured by species. **B.** Polysaccharide (K-locus, left) and lipopolysaccharide (O-locus, right) predicted serotypes of isolates grouped by species.

Antimicrobial resistance genes in *Klebsiella* spp.

We examined the occurrence of AMR genes in *Klebsiella* spp. (n=548) (**Additional file 1: Figure S3**). We detected the chromosomally encoded *AmpH* for strains of the *K. pneumoniae* group and a beta-lactamase of the *LEN* family in *K. variicola* isolates, which both confer low level resistance to beta-lactam antibiotics (260). Further, we detected the chromosomally encoded beta-lactamase genes *bla*$_{OXY1}$-*bla*$_{OXY8}$ in genomes within the *K. oxytoca* group, for which each species carries distinct variants, as described (261).

Fewer isolates of the *K. oxytoca* group were found to carry ESBL genes (8/132; 6.1%) than isolates of the *K. pneumoniae* group (125/415; 30.1%). Within the *K. pneumoniae* group, fewer

isolates of the species *K. variicol*a were found to carry ESBL genes (14/109; 12.8%), compared to isolates within *K. pneumoniae* (78/218; 35.8%), *K. quasipneumoniae* (32/83; 38.6%) and *"K. quasivariicola"* (1/5; 20.0%). We observed a higher number of *K. quasipneumoniae* isolates encoding carbapenemases (12/83; 14.4%) compared to *K. pneumoniae* (15/218; 6.9%), *K. variicola* (9/109; 8.6%) and *"K. quasivariicola"* (0/5; 0%). Within the *K. oxytoca* group, we detected the highest frequency of ESBL and carbapenemases in *K. michiganensis* (6/54; 11.1% and 5/54; 9.3%, respectively) compared to *K. oxytoca* (both in 1/41; 2.4%) and *K. grimontii* (both in 2/37; 5.4%).

Klebsiella spp. identification in routine diagnostics

Given the group- and species-specific trends in accessory genome composition and content, an accurate species identification may have an important clinical impact. For the strains included in this study, fatty acid analysis, GC-MS, and a panel of biochemical reactions were unable to identify a characteristic feature that could be used to distinguish unambiguously between the included *Klebsiella* spp. (**Additional file 4: Table S2**, **Additional file 1: Figure S4**, **Additional file 5: Table S3**). As such, in order to find a robust and accurate way to distinguish *Klebsiella* spp. based on MALDI-TOF mass spectra, we used ribosomal subunit proteins as species-specific MALDI-TOF MS markers. To do this, we first *in silico* predicted protein masses of the 56 ribosomal subunit proteins from 3,594 genome drafts.

Only proteins with a mass between 2,000-20,000 Daltons can be detected in MALDI-TOF mass spectra, and due to the intrinsic measurement error of MALDI-TOF MS not all predicted masses can be distinguished. Additionally, not all ribosomal subunit proteins can be equally ionized, and therefore detected, in similar proportions. Therefore, to determine the practical value of our approach, we examined which of these potential marker masses can reproducibly be detected in MALDI-TOF mass spectra of routine quality. Fifty isolates, representing the species *K. pneumoniae* (n=10), *K. variicola* (n=10), *K. oxytoca* (n=8), *K. michiganensis* (n=12), *and K. grimontii* (n=10), all with species identification confirmed by WGS, were analysed in quadruplicate on four different MALDI-TOF MS systems in different laboratories, resulting in the generation of 800 spectra. The 25 ribosomal subunits L17, L18, L19, L20, L21, L25, L27, L28, L29, L30, L31, L32, L33 (methylated), L34, L35, L36, S8, S9, S14, S13, S15, S19, S20, S21 and S22 were subsequently included as valid target proteins for identification of *Klebsiella* spp. The reproducibility of these marker masses varied between the different laboratories (**Additional file 6: Table S4**). We found a complete set of these 25 target proteins in 3,360 assemblies and therefore based our further analyses on these predicted mass profiles (**Additional file 7: Table S5**).

We assessed the distribution of the *in silico* predicted masses of these target proteins in a representative dataset of 464 genomes, comprising those which were included in the

comparative genomic analysis and were complete for these 25 target proteins. We identified multiple group- and species-specific alleles for *K. pneumoniae, K. variicola, "K. quasivariicola", K. oxytoca* and *K. grimontii* (**Figure 3**). No marker mass uniquely identifying *K. michiganensis* and *K. quasipneumoniae* was detected. In order to unambiguously separate *K. michiganensis* and *K. quasipneumoniae* from closely related species, a combination of marker masses needs to be detected (e.g. S15 at 10,078 Da and L25 at 10,636 Da for *K. michiganensis*, **Figure 3**).

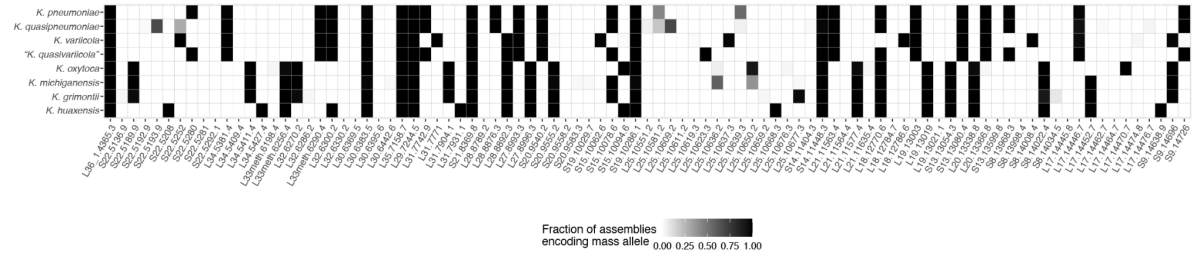

Figure 3: Distribution of the *in silico* predicted masses of 25 target proteins encoded by the species *K. pneumoniae* (n=208)*, K. quasipneumoniae* (n=60)*, K. variicola* (n=68)*, "K. quasivariicola"* (n=3)*, K. oxytoca* (n=37)*, K. michiganensis* (n=51)*, K. grimontii* (n=36) and *K. huaxensis* (n=1).

Based on the differential masses of the 25 target proteins, we can distinguish 110 distinct *Klebsiella* spp. ribosomal mass profiles (**Additional file 8: Table S6**) that allow us to distinguish *K. pneumoniae, K. quasipneumoniae, K. variicola, "K. quasivariicola", K. oxytoca, K. michiganensis, K. grimontii,* and *K. huaxensis*. These 110 distinct ribosomal mass profiles were subsequently included in an in house developed reference database for species identification. We assessed the specificity and sensitivity of this approach, using the spectra (n=800) of the same diverse *Klebsiella* spp. isolates (n=50). The species identification by MALDI-TOF MS was compared to the species identity as assigned by WGS of the identical isolate. If essential target proteins could not be detected in a MALDI-TOF mass spectrum and the acquired mass profile did not allow unique species identification, this spectrum was labelled as “Multispecies ID only”. **Table 1** shows the evaluation of the species identification of these spectra, resulting from comparison of the acquired MALDI-TOF mass spectra with the 110 ribosomal marker mass profiles. The identification was evaluated on two levels: (i) whether the assignment to *K. pneumoniae* or *K. oxytoca* group was correct and (ii) whether the correct species within each group could be assigned. Species identification using these marker mass profiles resulted generally in accurate identification and provided better species identification within the *K. oxytoca* group than the currently used commercially available databases Microflex Biotyper Database (Bruker Daltonics flexControl v.3) and the VitekMS Database (v.3), as these databases do not include *K. grimontii* and *K. michiganensis*.

Table 1. Evaluation of the species identification of 50 *Klebsiella* spp. isolates by MALDI-TOF mass spectra using ribosomal marker mass profiles. Species identification of 800 MALDI-TOF mass spectra using 110 marker mass profiles was compared to the species identification assigned using WGS data. Specificity and sensitivity were computed on two levels: (i) whether the assignment to *K. pneumoniae* group or *K. oxytoca* group was correct (= 'Group level') and (ii) whether the correct species within each group could be assigned (= 'Species Level').

Species identification by WGS	**Identification of MALDI-TOF mass spectra using Marker Mass profiles based on 25 pre-defined ribosomal subunit proteins**			
	Group level		**Species level**	
	Sensitivity [%]	**Specificity [%]**	**Sensitivity [%]**	**Specificity [%]**
K. pneumoniae	98.8	100	70.0	100
K. variicola	98.1	100	93.1	100
K. oxytoca	99.2	100	80.5	99.4
K. michiganensis	100	100	44.3	100
K. grimontii	99.4	100	70.6	100

The sensitivity and specificity of the approach (**Table 1**) were computed based on the identification of MALDI-TOF mass spectra acquired on all four MALDI-TOF MS systems. Specificity and sensitivity on group level were >98% using marker mass profiles, for all tested species on all four MALDI-TOF MS systems, reflecting a low probability of false positive results using this approach.

Sensitivity on the species level varied between the MALDI-TOF MS systems, especially within the *K. oxytoca* group. For the species *K. oxytoca,* sensitivity at the species level ranged from 37.5-100% between the four MALDI-TOF MS systems, for *K. michiganensis* from 8.3-70.8%, and for *K. grimontii* from 12.5-95.0%. Species identification within the *K. oxytoca* group requires the detection of marker masses in a high mass range and variations in sensitivity between the MALDI-TOF MS systems could potentially be linked to MALDI-TOF mass spectral quality (262).

Classification of international routinely acquired MALDI-TOF mass spectra

Using the 110 species-specific marker mass profiles, we retrospectively analysed 22,346 spectra derived from bacterial isolates from eight healthcare centers in four countries (dataset

(a), **Additional file 1: Figure S1 and Additional file 1: Figure S5**). All spectra had previously been used to diagnose isolates as *Klebsiella* spp. in routine diagnostic laboratories. Re-analysing the spectra using the newly compiled mass profile database, we attempted to categorise all spectra first into the two groups and then to one of the eight species in the database (**Figure 4A and B**).

A subset of the samples (n=427; 1.9%) were identified as *Raoultella* spp. or *K. aerogenes* and excluded from further analysis. 85 samples (0.3%) could only be identified to the genus level. The remaining 21,834 samples (97.7%) were categorised as *K. pneumoniae* group or *K. oxytoca* group. Of these, a higher proportion of samples could be categorised to the species level within the *K. pneumoniae* group (n=14,867/17,555; 84,7%), than within the *K. oxytoca* group (n=2,718/4,249; 64.0%), which reflects the difficulty to reproducibly detect all required species-specific marker masses in the *K. oxytoca* group. The proportion of samples which could not be categorised to the species level varied by healthcare center and MALDI-TOF MS system and ranged from 9.8% (n=40/407 samples) to 30.5% (n=439/1,440 samples) within the *K. pneumoniae* group and from 22.8% (n=349/2,560 samples) to 84.8% (n=217/256 samples) within the *K. oxytoca* group. The remaining 17,585 samples (78.69%) could unambiguously be identified to the species level. Of these, across all centres we identified: *K. pneumoniae* (n=12,523; 71.2%), *K. quasipneumoniae* (n=575; 3.3%) *K. variicola* (n=1,717; 9.8%), *"K. quasivariicola"* (n=52; 0.3%), *K. oxytoca* (n=1,445; 8.2%), *K. michiganensis* (n=836; 4.8%), *K. grimontii* (n=425; 2.4%), and *K. huaxensis* (n=12; 0.1%).

Interestingly, we observed different frequencies of the two *Klebsiella* groups and species depending on the originating healthcare center, possibly reflecting a geographical trend of pathogenic *Klebsiella* spp. distribution (**Figure 4A**): The proportion of isolates belonging to the *K. oxytoca* group was higher in more northern regions, a finding which requires further investigation. The proportion of *Klebsiella* groups and species was also found to vary depending on the patient material from which it was isolated (**Figure 4B**). Isolates of the *K. oxytoca* group were least abundant in urinary tract samples and most abundant in samples of the gastro-intestinal tract samples.

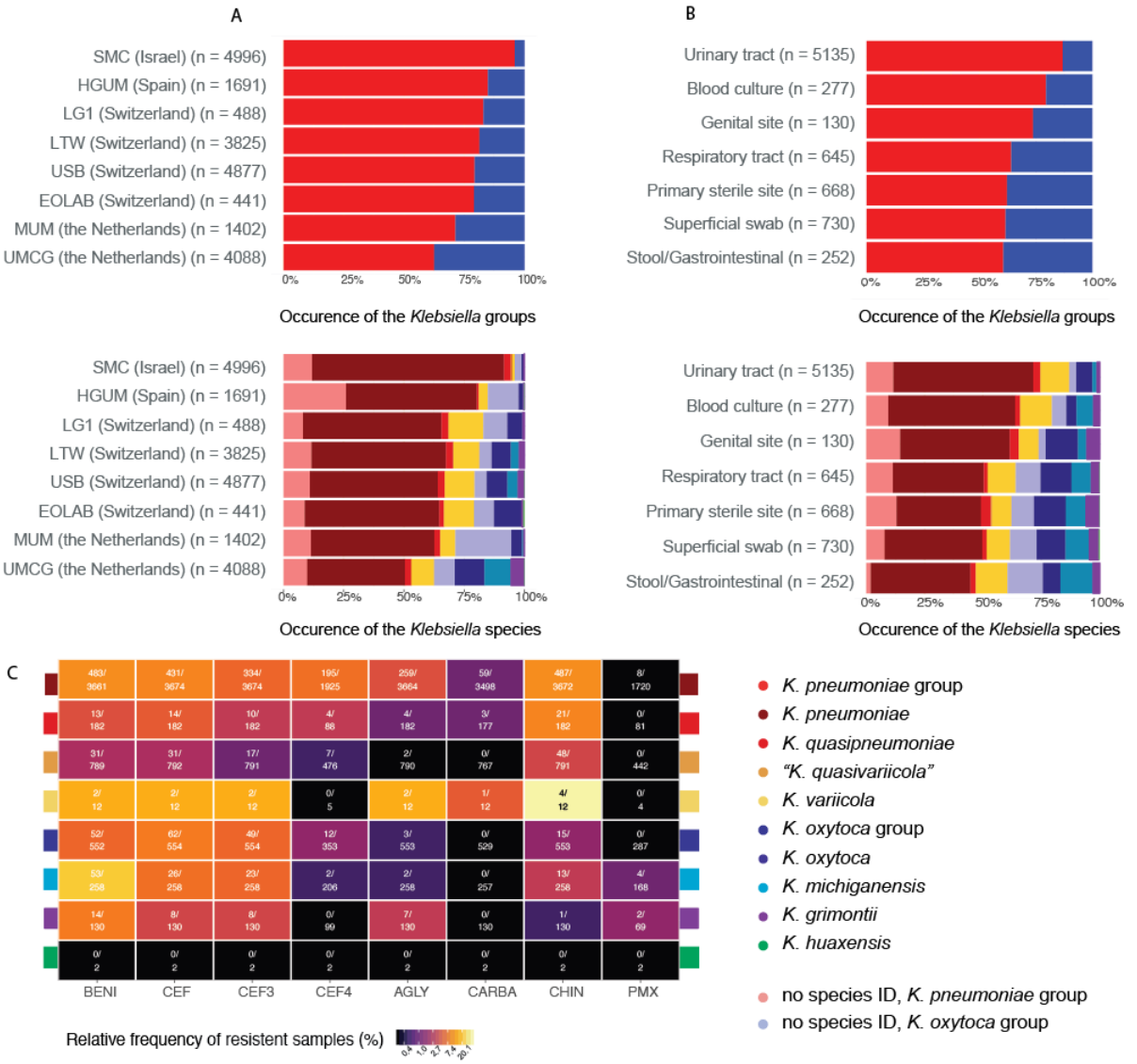

Figure 4. Occurrence of *Klebsiella* spp. in clinical settings, as determined by ribosomal marker MALDI-TOF MS method. A. Occurrence of *Klebsiella* groups and species in eight healthcare centers from Israel (n=1), Spain (n=1), Switzerland (n=4) and the Netherlands (n=2), sorted by increasing occurrence of the *K. oxytoca* group. **B.** Occurrence of *Klebsiella* groups and species in patient samples from various isolation sites. 'Primary sterile sites' includes deep wounds, aspirates and deep tissues, 'Respiratory tract' includes sputum, bronchoalveolar lavage and tracheal secretion, 'Superficial swabs' includes swabs from superficial wounds and skin infections. **C.** Antibiotic resistance of *Klebsiella* spp. (BENI = beta-lactams with beta-lactamase Inhibitors, CEF3 = 3rd generation cephalosporins, CEF4 = 4th generation cephalosporins, AGLY = aminoglycosides, CARBA = carbapenems, CHIN = quinolones, PMX = polymyxins). Please note that the colour grading is on log-scale.

AMR profiles

Information on phenotypic AMR and isolation source was available for 7,876 samples from two healthcare centers in Switzerland (USB and LTW), (dataset (b), **Figure 4C**, **Additional file 1: Figure S1 and Additional file 1: Figure S5**) for which spectra had been analysed.

Isolates of the *K. oxytoca* group were more likely to be resistant against penicilines including beta-lactamase inhibitors and 3rd generation cephalosporins (OR 2.79, p < 0.001, 95% CI [1.70, 4.63]; OR 2.45, p = 0.005, 95% CI [1.31, 4.58], respectively) but less often resistant to 4th generation cephalosporins and aminoglycosides (OR 0.17, p<0.001, 95% CI [0.09, 0.28]; OR 0.22, p<0.001, 95% CI [0.12, 0.35]) than isolates of the *K. pneumoniae* group (**Additional file 9: Tables S7 - S9**). Within the *K. oxytoca* group we found *K. oxytoca* to be less resistant to penicilines including beta-lactamase inhibitors, than *K. michiganensis* (OR 0.57, p < 0.001, 95% CI [0.42, 0.75]) (**Additional file 9: Table S7**). Within the *K. pneumoniae* group, isolates identified as *K. pneumoniae* were more resistant to penicilines including beta-lactamase inhibitors, 3rd and 4th generation cephalosporins as well as to aminoglycosides than *K. variicola,* (OR 2.11, p<0.001, 95% CI [1.42, 3.18]; OR 2.61, p<0.003, 95% CI [1.38,5.06] and OR 5.80, p<0.001, 95% CI [2.40,20.04], respectively) (**Additional file 9: Tables S8-S10**).

Clinical endpoints

Data from patient charts of 957 clinical cases at the USB were reviewed and analysed on multiple clinical endpoints (dataset (c), **Additional file 1: Figure S1 and Figure S5**). In order to examine the clinical phenotype of the *Klebsiella* groups and species, independent of their AMR burden, we corrected our model for resistance against 3rd generation cephalosporins, which is associated with production of ESBL. Clinical outcomes and explanatory variables are summarized in **Additional file 10: Tables S11-12.**

We found no evidence for *Klebsiella* group- or species-specific association with our primary 30-day mortality endpoint (**Additional file 10: Table S13**). As a general finding, female patients seemed to have better outcomes than male patients: all cause 30 day mortality was less likely for female patients (OR 0.60, p=0.012, 95% CI [0.40, 0.89]), female patients were less likely to be affected by invasive infection of sterile sites (OR 0.54, p=0.002, 95 % CI [0.37, 0.79]) and to be admitted to an ICU (OR 0.63, p=0.009, 95 % CI [0.45, 0.89]) (**Table 2, Additional file 10: Tables S13 - S14**). Furthermore, increasing CCI and increasing age seemed to be associated with higher 30-day mortality (OR 1.16, p = 0.40, 95% CI [1.01,1.34], and OR 1.36, p<0.001, 95% CI [1.24,1.49], respectively), whereas antibiotic treatment at entry or during hospitalization was associated with higher odds for ICU admission (OR 4.41, p<0.001 and 95% CI [2.10,8.93]) (**Table S15 – S16**).

Strikingly, isolates of the *K. oxytoca* group were more likely to be involved in invasive infection compared to isolates of the *K. pneumoniae* group (OR 2.39, p = 0.044, 95% CI [1.05,5.53]). As we corrected in our model for resistance against 3rd generation cephalosporins, and as the *K. oxytoca* group is not associated with a higher burden of AMR, we hypothesise that this increased association to invasive infections is independent of AMR and might reflect increased virulence of this group.

We found no evidence for *Klebsiella* group- or species-specific associations with the remaining clinical outcomes (**Additional file 10: Tables S14 – S17**).

Table 2. Odds ratio estimates for invasive infection using the generalized linear mixed-effects model (GLMM). n = 732 complete observations with 162 events**.** OR = Odds Ratio; CI = Confidence Interval; CCI = Charlson Comorbidity Index

	OR	95 % CI	Z	p-value
***K. oxytoca* group *vs. K. pneumoniae* group**	2.39	[1.05,5.53]	2.01	0.044
K. oxytoca* vs. *K. michiganensis	0.75	[0.42,1.34]	-0.97	0.33
K. oxytoca vs. K. grimontii	0.64	[0.30,1.36]	-1.15	0.252
K. pneumoniae vs. K. variicola	1.08	[0.69,1.65]	0.35	0.724
K. pneumoniae vs. K. quasipneumoniae	0.64	[0.35,1.15]	-1.44	0.149
CCI	1.14	[1.04,1.25]	2.95	0.003
Age (centered, 10 years increase)	0.84	[0.75,0.95]	-2.78	0.005
Female (ratio)	0.54	[0.37,0.79]	-3.17	0.002
Immunosuppression	0.5	[0.29,0.86]	-2.5	0.012
Resistance to 3rd generation cephalosporins	1.31	[0.44,3.74]	0.5	0.618
Antibiotic treatment at entry or during hospitalisation	3.11	[1.44,6.49]	2.92	0.003

Discussion

We have described a MALDI-TOF MS method allowing the identification of clinically important and currently often misdiagnosed *Klebsiella* spp. and applied it to an international dataset of over 22,000 unique bacterial isolates from microbiological routine laboratories. While species-specific MALDI-TOF MS patterns within the genus *Klebsiella* have previously been described (224,232,263), their discriminatory power has not yet been assessed in large routinely acquired mass spectral datasets.

Using our ribosomal marker-based approach we are able to separate eight species of the genus *Klebsiella*: *K. pneumoniae*, *K. quasipneumoniae, K. variicola, "K. quasivariicola"*, *K. oxytoca*, *K. michiganensis*, *K. grimontii*, and *K. huaxensis*. This higher phylogenetic resolution power represents an important step forward in clinical diagnostics as *"K. quasivariicola", K.*

michiganensis, K. grimontii, and *K. huaxensis* are currently not found in commonly used databases. Mass spectral quality plays an important role in distinguishing the species within the *K. oxytoca* group, as the species-specific peaks lie in a high mass range with m/z values above 10,000. Moreover, in order to unambiguously identify the species *K. michiganensis*, a unique combination of marker masses in a higher mass range needs to be detected, posing an additional challenge for identification. The inability to detect these in many spectra decreases sensitivity. We evaluated our approach by computing sensitivity and specificity values for five clinically important *Klebsiella* spp.. A limitation of the current study is that the most recently described *Klebsiella* spp.*, K. africana*, *K. pasteurii,* and *K. spallanzanii* were not included in the analysis. Identifying species specific marker masses for these, and a similar evaluation for the less frequently observed *Klebsiella* species would be desirable in future. A recent study introduced a web-based tool which also uses other core proteins than ribosomal subunit proteins to distinguish between the *Klebsiella* spp. [26]. Combining these and our ribosomal marker masses could potentially increase the resolution of MALDI-TOF mass spectral identification even below the species level.

Marker-based approaches are independent of the MALDI-TOF MS system used. Therefore, we were able to assess the occurrence and clinical phenotype of important *Klebsiella* spp. in international clinical laboratories using MALDI-TOF MS systems from different manufacturers. While *K. pneumoniae* remains the most commonly detected species, we also detected isolates from each of the species, which are currently not routinely identified by common MALDI-TOF MS databases, including *K. quasipneumoniae*, *"K. quasivariicola", K. michiganensis, K. grimontii,* and *K. huaxensis* from a variety of patient material.

Our data provide evidence that infections with isolates of the *K. oxytoca* group were more likely to be invasive than infections with isolates of the *K. pneumonia* group, highlighting the clinical importance of this under-appreciated group. The clinical cases analysed in this study were treated at the same healthcare center in a low endemic region for ESBL-producing bacteria. Further studies from different regions would be needed in order to confirm these results in other epidemiological backgrounds.

Due to the higher prevalence of the *K. pneumoniae* group infections, a larger absolute number of invasive infections are caused by isolates of the *K. pneumoniae* group than by isolates of the *K. oxytoca* group. Thus, infections with a *K. oxytoca* group isolate tend to be less frequent, but more severe, a finding which could potentially be linked to the increased frequency of certain virulence factors, such as the siderophore yersiniabactin, genes involved in allantoin metabolism and the cytotoxin tilivallin. Yersiniabactin and genes involved in the allantoin metabolism are well established virulence factors of *K. pneumoniae* (264,265) and their frequent occurrence in assemblies of the *K. oxytoca* group has been observed before (230,266). Based on the data acquired in this study we describe an indirect link between the

association of *K. oxytoca* group strains and invasive infection and the occurrence of these virulence factors in genomes of the same species. However, in order to assess whether these factors actually increase the virulence of *K. oxytoca* group strains, functional studies are needed.

Bloodstream infections caused by *K. variicola* have been described as causing a higher mortality than bloodstream infections caused by *K. pneumoniae* (135), a finding that we could not confirm in our dataset.

Resistance to 3rd generation cephalosporins is most often conferred by ESBLs, which were enriched in the analysed genomes of *K. pneumoniae* group compared to the *K. oxytoca* group (**Additional file 1: Figure S3**). Within the *K. pneumoniae* group isolates for which spectra were analysed, we observed a higher proportion of isolates resistant to 3rd and 4th generation cephalosporins and aminoglycosides in *K. pneumoniae* compared to *K. variicola*. The lower proportion of resistant isolates within *K. variicola* has previously been described (267) and is also in line with the lower frequency of ESBL genes detected in *K. variicola* genome sequences, compared to *K. pneumoniae* genome sequences. As genes encoding AMR are often carried on plasmids, this may be linked to the lower median number of plasmids detected in isolates of *K. variicola.*

These findings need to be explored in other epidemiological situations with higher ESBL and carbapenemase burdens to examine their broader relevance. With this, a rapid and accurate species identification by MALDI-TOF MS may have an important impact on antibiotic stewardship and treatment decisions.

Conclusions

Based on systematic comparison of WGS and *in silico* ribosomal subunit protein mass prediction, we present a MALDI-TOF MS based analytical approach to distinguish eight *Klebsiella* species which can be applied in clinical routine diagnostic laboratories. In this study we identified species-specific AMR and virulence patterns within the genus *Klebsiella* and uncovered an increased association of the *K. oxytoca* group with invasive infection to primary sterile sites.

List of abbreviations

AMR: antimicrobial resistance

WGS: whole genome sequencing

MDR: multi-drug resistant

MALDI-TOF MS: Matrix Assisted Laser Desorption Ionization - Time Of Flight Mass Spectrometry

ANI: Average Nucleotide Identity
CHCA: cyano-4-hydroxycinnamic acid
ICU: intensive care unit
CCI: Charlson Comorbidity Index
GLMM: generalized linear mixed models
IQR: Interquartile Range
OR: Odds Ratio

Declarations

Ethics approval and consent to participate

Bacterial strains have been collected in clinical routine diagnostics. The collection of bacterial strains and their analysis for diagnostic assay development do not fall under the Swiss human research act and no ethical approval nor consent to participate from patients was required. The analysis of patient demographic and clinical outcome data was approved by the 'Ethikkommission Nordwest- und Zentralschweiz' (EKNZ) (BASEC-Nr. 2016-01899 and 2018-00225) for patients who did not reject the hospitals general research consent. Patients who did reject the hospital's general consent were excluded from all analyses which include patient demographic and clinical outcome data.
All analyses performed in this study were in accordance with the Helsinki declaration and its later amendments.

Consent for publication

Not applicable.

Availability of data and materials

Whole genome sequences acquired for this study have been uploaded to Genebank (https://www.ncbi.nlm.nih.gov/genbank/). Accession numbers of these and previously published WGS data used in this study can be found in **Additional file 2: Table S1**.

Software code generating figures from the genomic analysis is available on GitHub () (244)

Competing interests

VP, FF and RM are employees of the company Mabritec AG, Riehen, Switzerland, which commercializes ribosomal marker-based approaches in MALDI-TOF MS data analyses for identification of microorganisms.

FM, MA (Labor Team W AG) and CM (Laborgemeinschaft 1) are employed by private diagnostic laboratories.
The remaining authors declare that they have no competing interests.

Funding

This study was supported by a “Personalized Health” at ETHZ (D-BSSE) and University of Basel grant (PMB-03-17) granted to AE and a Doc.Mobility Fellowship by the Swiss National Science Foundation (P1BSP3-184342) granted to AC.

Authors' contributions

Conceptualization: AC, DW, HSS, CO, CG, FF, RM, GR, AK, DRV, SvF, SB, STS, TH, GS, CH, CC, JMG, OS, BRS, FM, MA, VG, HvD, GK, CM, CD, VP, AE
Data Curation: AC, DW, HSS, CO, CG, FF, RM, AK, DRV, SvF, SB, STS, SH, GS, CH, CC, JMG, OS, BRS, FM, MA, VG, HvD, GK, CM, VP
Formal Analysis: AC, DW, HSS, FF, DRV, SvF, TH, CC, VP
Funding Acquisition: CD, VP, AE
Investigation: AC, CO, CG, RM, TH, CC, VP
Methodology: AC, RM, FF, CH, VP, AE
Project Administration: AC, CD, VP, AE
Resources: GR, AK, SB, STS, JMG, OS, BRS, FM, MA, VG, HvD, GK, CM, AE
Software: AC, DW, HSS, FF, DRV, SvF
Supervision: CD, VP, AE
Validation: AC, DW, HSS, CO, CG, FF, RM, GR, AK, DRV, SvF, SB, STS, TH, GS, CH, CC, JMG, OS, BRS, FM, MA, VG, HvD, GK, CM, CD, VP, AE
Visualization: AC, DW, HSS, DRV, SvF, TH
Writing – Original Draft Preparation: AC
Writing – Review & Editing: AC, DW, HSS, CO, CG, FF, RM, GR, AK, DRV, SvF, SB, STS, TH, GS, CH, CC, JMG, OS, BRS, FM, MA, VG, HvD, GK, CM, CD, VP, AE
All authors read and approved the final manuscript.

Acknowledgements

We thank Josiane Reist for MALDI-TOF MS measurements (University of Basel). We thank Magdalena Schneider, Christine Kiessling, Elisabeth Schultheiss, Rosa-Maria Vesco, and Clarisse Straub for the DNA extraction, library preparations and sequencing of the bacterial isolates (all University Hospital Basel). Furthermore, we thank the Functional Genomics

Center Zurich for the PacBio sequencing, and sciCORE for the computational infrastructure provided.

Additional Files

- Additional file 1.pdf
    - Figures S1: Schematic representation of the workflow of the project
    - Figures S2: Gene accumulation curves for species of the *K. pneumoniae* group (A) and the *K. oxytoca* group (B)
    - Figures S3: Genes associated with AMR detected in *Klebsiella spp.*
    - Figure S4: Partial least squares discriminant analysis (PLS-DA) score plot containing primary metabolites measured of five *Klebsiella spp.*
    - Figures S5: Species identity of datasets (a), (b) and (c) included in the statistical analysis
- Additional file 2.xlsx
    - Table S1: Strains included in the study and where sequence data can be accessed
- Additional file 3.pdf
  Supplementary methods for:
    - Measurement of primary metabolites
    - Fatty acid analysis
    - Statistical analysis of clinical outcome data
- Additional file 4.xlsx
    - Table S2: Cellular fatty acid composition of 11 Klebsiella spp. strains
- Additional file 5.xlsx
    - Table S3: Biochemical reaction and AMR profiles of a diverse set of Klebsiella spp. Strains
- Additional file 6.xlsx
    - Table S4: Reproducibility of detection for the predicted ribosomal subunits in MALDI-TOF mass spectra acquired in 4 different centers.
- Additional file 7.xlsx
    - Table S5: Binary table displaying the predicted ribosomal subunits mass variants and whether this variant was predicted from an assembly or not
- Additional file 8.xlsx
    - Table S6: Binary table of the protein marker masses which can reproducibly be detected in MALDI-TOF MS spectra
- Additional file 9.pdf

Odds ratio estimates comparing *Klebsiella* groups and species for resistance to different antibiotic classes:

- Table S7: Penicillins with beta-lactamase Inhibitors
- Table S8: 3rd generation cephalosporins
- Table S9: 4th generation cephalosporins
- Table S10: Aminoglycosides

- Additional file 10.pdf

  Summary data compiled for the statistical analysis of clinical endpoints
  - Table S11: Summary of outcome variables for the clinical data set.
  - Table S12: Summary of explanatory variables for the clinical data set.

  Statistical analyses examining differences in clinical outcome between the *Klebsiella* groups and species:
  - Table S13: Odds ratio estimates from the generalized linear mixed-effects model (GLMM) for all cause death within 30 days from diagnosis
  - Table S14: Odds ratio estimates from the generalized linear mixed-effects model (GLMM) for ICU admission
  - Table S15: Hazard ratio estimates from cause-specific hazards Cox proportional hazards model for time to death within hospital after diagnosis with hospital discharge as competing event
  - Table S16: Estimates of the multiplicative effects from the Poisson generalized linear mixed-effects model (GLMM) for the number of medical disciplines involved
  - Table S17: Odds ratio estimates from the generalized linear mixed-effects model (GLMM) for the mentioning of the infection in the patient letter

Chapter II: Bacterial genome wide association study substantiates *papGII* of *E. coli* as a patient independent driver of urosepsis

Aline Cuénod[1,2], Jessica Agnetti[1,2], Helena Seth-Smith[1,2,3], Tim Roloff[1,2,3], Sarah Tschudin-Sutter[4,5], Stefano Bassetti[6], Nicholas R. Thomson[7,8], Adrian Egli[1,2]

[1] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland [2] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland [3] Swiss Institute for Bioinformatics, Basel, Switzerland [4] Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland [5] Department of Clinical Research, University of Basel, Basel, Switzerland [6] Division of Internal Medicine, University Hospital Basel, Basel, Switzerland [7] Parasites and Microbes, Wellcome Trust Sanger Institute, Hinxton, UK [8] Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK

Manuscript in preparation

My contributions:

- Conceptualisation of the study
- Acquisition of funding (Doc.Mobility Fellowship by the Swiss National Science Foundation (P1BSP3-184342))
- Collection of bacterial strains in routine diagnostics
- Whole genome sequencing of bacterial strains
- Acquisition of MALDI-TOF mass spectra
- Bioinformatic analysis
- Statistical analysis of clinical data
- Data visualisation (all figures)
- Writing of the original manuscript

A first part of this project was accepted for *Paper Poster* presentation at the following international conferences:

- Royal Society Meeting on the population structure of *Escherichia* 2020 (Newport Pagnell, United Kingdom)
- The European Congress of Clinical Microbiology and Infectious Diseases (*ECCMID*), 2020 (Paris, the France)
- Annual Conference of the United Kingdom Microbiology Society, 2020 (Edinburgh, United Kingdom)

All three events were cancelled due to the pandemic of SARS-CoV-2

The project is *in review* for presentation at following international conferences:

- European Congress of Clinical Microbiology and Infectious Diseases (ECCMID) 2022, Lisbon (Portugal)
- American Society for Microbiology Microbe (ASM Microbe) Conference 2022 (Washington DC, USA)

Note: The following part contains the full manuscript.
The Supplementary material can be found in Appendix I of this thesis.

Bacterial genome wide association study substantiates *papGII* of *E. coli* as a patient independent driver of urosepsis

Aline Cuénod[a,b], Jessica Agnetti[a,b], Helena Seth-Smith[a,b,c], Tim Roloff[a,b,c], Sarah Tschudin-Sutter[d,e], Stefano Bassetti[f], Martin Siegemund[g], Christian H. Nickel[h], Tim Keys[i], Valentin Pflüger[j], Nicholas R. Thomson[k,l], Adrian Egli[a,b]

[a] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland

[b] Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland

[c] Swiss Institute for Bioinformatics, Basel, Switzerland

[d] Infectious Diseases and Hospital Epidemiology, University Hospital Basel and University of Basel, Basel, Switzerland.

[e] Department of Clinical Research, University of Basel, Basel, Switzerland

[f] Division of Internal Medicine, University Hospital Basel, Basel, Switzerland

[g] Intensive Care Unit, University Hospital Basel, Basel, Switzerland

[h] Emergency Medicine, University Hospital Basel, Basel, Switzerland

[i] Institute of Microbiology, ETH Zurich, Zurich, Switzerland

[j] Mabritec AG, Riehen, Switzerland

[k] Parasites and Microbes, Wellcome Trust Sanger Institute, Hinxton, UK

[l] Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK

Keywords

Escherichia coli, Urinary Tract Infection, Invasiveness, GWAS, *papGII*

Abstract

Background: Urinary tract infections (UTI) are amongst the most common bacterial infections worldwide, often caused by uropathogenic *Escherichia coli*. Multiple bacterial virulence factors or patient characteristics have been linked individually to a progressive more invasive infection. In this study, we aim to identify pathogen- and patient-specific factors that drive the progression to urosepsis by concurrently analysing bacterial and host characteristics.

Methods: We analysed 1079 *E. coli* strains isolated from 831 clinical cases with UTI and/or bacteraemia by whole genome sequencing (Illumina). Sequence Types (ST) were determined via srst2 and capsule loci via kaptive. First, we compared isolates from urine and blood to confirm clonality. Second, we performed a bacterial Genome Wide Association Study (bGWAS) (pyseer v1.3.9) using bacteraemia as primary clinical outcome. Clinical data was collected by an electronic patient chart review. Finally, we concurrently analysed the association of the most important bGWAS hit and important patient characteristics with the clinical endpoint bacteraemia using a generalised linear model (GLM).

Results: Our patient cohort had a median age of 75.3 years (range: 18.00 - 103.1) and was predominantly female (580/831, 69.8%). The bacterial phylogroups B2 (60.5%; 503/831) and D (16.2%; 102/831), which are associated with extraintestinal infection, represent the majority of the strains in our collection, many of which encode a polysaccharide capsule (77.6%; 642/831). Most commonly observed STs were: ST131 (12.6%; 105/831), ST69 (11.2%; 93/831), and ST73 (10.1%; 84/831). Of interest, in 12.4% (13/105) of cases the *E. coli* pairs in urine and blood were only distantly related. In line with previous bGWAS studies, we identified the gene *papGII* (p-value < 0.001), which encodes the adhesin part of the *E. coli* P-pilus, to be associated with 'bacteraemia' in our bGWAS. In our GLM, correcting for patient characteristics, *papGII* remained highly significant (Odds ratio = 5.95, 95% Confidence Interval = [3.44,7.88], p-value < 0.001). We also observed a stronger local and system inflammatory response with higher leukocyte levels in urine samples and higher C-reactive protein levels in *papGII* positive vs. negative cases (median 730/uL vs. 222/uL, p-value < 0.001 and 114ng/mL vs. 41ng/mL p < 0.001).

Conclusion: The *E. coli* gene *papGII* seems to be a patient independent risk factor for the development of progression from UTI to bacteraemia.

Introduction

Up to 60% of women suffer from at least one symptomatic urinary tract infection (UTI) in their lifespan and 10% of women experience symptomatic UTI each year (188). *Escherichia coli* (*E. coli*) is the most common cause of community and hospital acquired UTI (186). Most UTIs cause only mild symptoms and no long-term effects are expected after the infection is cured. Some UTI, however, progress to cause invasive infections such as pyelonephritis, urosepsis and septic shock (206,268). These invasive infections are associated with high mortality, morbidity, reduction of quality of life in survivors and high healthcare costs (269,270). Therefore, understanding the burden of disease and identifying factors associated with a more severe disease outcome is critical for the patient management, healthcare related costs and antibiotic stewardship.

Strains of the species *E. coli* have diverse lifestyles and while they can be part of the healthy human gut microbiota and are found in soil and freshwater samples, they can also be pathogenic and cause infections. In line with the ability to survive in a vast variety of environments, the species of *E. coli* has an open pan-genome of more than ten times the size of an average genome (44). Only approximately one third of the genes encoded by a single strain are part of the core genome of this species (271). This genomic flexibility and the large gene pool explain the broad range of niches *E. coli* can be associated with. Within *E. coli,* multiple clearly distinct and deep branching phylogenetic groups (defined as phylogroups) have been identified (153). Originally, and on the bases of few housekeeping genes, eight *E. coli* phylogroups (A, B1, B2, C, D, E, F and G) were defined (153), of which A, B1, B2 and D are most frequently isolated (43). On the bases of genomic Mash distances these phylogroups have recently been suggested to be split up, resulting in 14 phylogroups, including two corresponding to strains of the genus *Shigella* (156). For some phylogroups, an association to clinical phenotypes has been observed. Phylogroups A and B1 are associated with asymptomatic carriage in the gut while phylogroups B2, D and F predominantly cause extra intestinal infections (44,156). However, there are large variations in clinical phenotypes within each phylogroup. This is linked to the presence or absence of virulence and antimicrobial resistance (AMR) genes, which can differ between closely related strains (43). Consequently, UTIs can be caused by *E. coli* strains from multiple different phylogroups (272). Well-studied uropathogenic *E. coli* (UPEC) virulence factors include iron uptake systems, capsular polysaccharides, immune modulators, fimbriae and pili (198,273,274). Genes encoding these clinically important bacterial factors are associated with globally successful UPEC clones such as Sequence Types (ST)131, ST69, ST73, ST95 (198,201,206).

A recent study suggests that these virulent UPEC lineages emerged after the independent horizontal acquisition of *papGII* (198). The gene *papGII* encodes one of multiple variants

(PapGI - V) of the adhesin tips of Pap pili, and binds to the globoseries of glycosphingolipids, more specifically Gb5 (GalNAcα1-3-GalNAc3Galα1-4Galβ1-4GlcCer) (199,200) of human uroepithelial and kidney cells (195,196) and can modulate the host immune response (275). Amongst other virulence factors, *papGII* is known to play an important role in the progression of UTI to invasive infection (198) and, from the perspective of invasive infection, is associated with the urinary and the intestinal tracts as ports of entry for *E. coli* bacteraemia (199). Another bacterial factor which has been associated with invasive UTI is *iuc*, which is essential for the biosynthesis of the iron uptake system aerobactin (198).

However, the clinical course of UTI and subsequent pyelonephritis and urosepsis is not only shaped by these bacterial factors, but also by host factors such as an effective immune response, host genetics, comorbidities, age, and gender (188). Hence the clinical importance of any bacterial factor cannot be definitively assessed without correcting for important patient characteristics. Despite the knowledge on UPEC virulence- and patient risk factors, these two interconnected aspects are often analysed separately and their interaction is rarely considered (199,276).

In this study, we aim to identify pathogen- and patient-specific factors that drive the progression from UTI to bacteraemia and urosepsis by concurrently analysing genomic bacterial data and host characteristics from patients with urinary tract or bloodstream infections.

Methods

Ethics

The collection and analysis of strains and patient data was approved by the 'Ethikkommission Nordwest- und Zentralschweiz' (EKNZ) (BASEC-Nr. 2019-00748).

Bacterial isolate collection and whole genome sequencing

E. coli isolates (n= 1079) were prospectively collected at the University Hospital Basel from urine (n= 791), blood culture (n=287) and deep tissue samples (n=1) from 04/2018 to 02/2020. Urine was cultured on 5% sheep blood agar plates (Becton Dickinson, Franklin Lakes, New Jersey, USA) and Chrom ID plate (Becton Dickinson, Franklin Lakes, New Jersey, USA) for a maximum of 48h. Blood cultures were incubated in aerobic and anaerobic flasks for a maximum of six days using the Virtuo system (bioMérieux, Marcy-l'Étoile, France). Bacterial stains have been identified as *E. coli* in routine diagnostics using the microflex Biotyper MALDI-TOF MS system (Bruker Daltonics, Bremen, Germany). Isolates were grown on Columbia 5% Sheep Blood Agar (bioMérieux, Marcy-l'Étoile, France) and DNA was extracted using the QIACube with the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany). After quality control of the DNA by Tapestation (Agilent, Santa Clara, USA), tagmentation libraries were generated as described by the manufacturer (Illumina DNA Prep Kit, Illumina, San Diego, USA). The genomes were sequenced using a 2 × 300 base pairs V3 reaction kit on an Illumina MiSeq, or using a 2 x 150 base pairs on an Illumina NextSeq instrument, reaching an average coverage of 74-fold for all isolates.

Phenotypic antimicrobial susceptibility testing

Antimicrobial susceptibility testing (AST) was performed using the Vitek2 system (*Enterobacteriaceae* AST Card, bioMérieux, Marcy-l'Étoile, France) or using strip diffusion Etests (bioMérieux, Marcy-l'Étoile, France) and the measurements were interpreted according to EUCAST clinical breakpoints (v.9.0) into susceptibility categories 'Susceptible', 'Intermediate' or 'Resistant'.

Definition of 'Invasiveness'

E. coli strains were defined as 'invasive' if a blood culture of the same clinical case was diagnosed positive for *E. coli* within seven days before or after the isolation of the respective strain.

Comparative genomic analysis

We trimmed the rawreads using Trimmomatic (v0.38) (238), generated assemblies using Spades (277) via unicycler (v0.3.0b) (239), and polished them using pilon (v1.23) (278). We examined the following features to ensure the quality of the sequence data and assembly: average read quality (median = 98.43; range = [56.42, 99.49]) and depth (mean = 74.49X; range = [4.7X, 328.8X]), %G+C content (median = 50.61; range=[50.3-51.1]), genome size (median = 5.05 megabases (MB); range=[4.38-5.85]). The purity of the sample was assessed using MetaPhlan (237) and genomes were annotated using prokka (240). Bacterial species identification was confirmed using ribosomal Multi Locus Sequence Typing (rMLST) (279), where three strains were identified as *Escherichia marmotae* and one as *Escherichia ruysiae*. These were excluded from further analysis. We screened all assemblies for the occurrence of previously described UPEC virulence (EcVGDB, (198)) and resistance factors (NCBI, (280)) using abricate (v0.8.10) (https://github.com/tseemann/abricate) and >95% coverage and >95% identity thresholds. In order to assign the *E. coli* phylogroups, we calculated the mash (v2.2) (155) distance to the previously described medoid reference genomes of each phylogroup (156), except for phylogroup C, where an alternative reference genome was used (GCF_001515725.1), as the published medoid reference genome clustered within phylogroup B1. We chose an alternative reference genome for phylogroup C from the Microreact project (https://microreact.org/project/10667ecoli/c38356ec) belonging to phylogroup C according to 'PCR Phylogroup' and 'Mash-Screen-Phylogroup' and having the highest 'Total Score' and 'Sequence Score'. We assigned the phylogroup of the closest reference genome to each queried genome using a cut-off of 0.04 mash distance. We identified the *E. coli* capsule types using fastKaptive (v0.2.2) (281). The well-known Capsule Types K1 and K5 correspond to fastKaptive assignment 'KX03' and 'KX29', respectively, which was identified by running K1 (CP003034.1) and K5 (CP022686.1) reference genomes through fastKaptive. We determined the O- and H antigens as well as the Multi Locus Sequence Type (MLST) via srst2 (v0.2.0) (282,283). We assessed the Average Nucleotide Identity between strains using fastANI (v1.32) (115). We used panaroo ('sensitive' mode) (v1.2.7) (40) to identify the core genome which we then aligned using mafft (v7.467) (113) and used RaxML (`GTRCAT` approximation) (v8.2.8) (114) to construct a phylogenetic tree from this alignment. For pairwise comparison of strains isolated from the same patient sample, we used the variant caller Freebayes (v1.2.0) (284) via snippy (v4.3.6) (https://github.com/tseemann/snippy).

Bacterial Genome Wide Association Study

In order to identify bacterial factors which are associated with invasive infection, we used pyseer (v1.3.9) (285). We used unitigs as inputs, which represent non-redundant sequence

elements of variable length and which we had previously constructed via unitig-counter (v1.0.5) (286) and 'bacteraemia' as clinical endpoint. A minor allele frequency threshold of 0.1% was used. One strain per clinical case (n=831) was included in the analysis. If multiple strains per clinical case had been isolated, we chose strains isolated from blood culture samples over strains isolated from urine samples, as these caused the invasive infection. If there were multiple strains isolated from the same material, we chose the strain isolated at the earliest time point. We used random effects to correct for population structure ('-lmm' mode) by providing a similarity-matrix acquired from the core genome phylogeny which was previously constructed (see above). Unitigs were mapped against the genome annotations. In order to identify the *papG* variant against which the unitigs mapped, we compared these to the genes *papGI*, *papGII*, *papGIII*, *papGIV* and *papGV* (198) using fastANI (v1.32) (115).

Patient data collection

Patient demographic and clinical data from patients with *E. coli* infections were retrospectively accessed via the hospital's clinical information system in a case report form. The University Hospital Basel is a tertiary healthcare centre with more than 750 beds in a low endemic region for ESBL-producing bacteria (22). Inclusion criteria were: patients for which at least one isolated bacterial colony was isolated from urine or blood culture samples and identified as *E. coli* by MALDI-TOF MS (Bruker Daltonics, Bremen, Germany) or using biochemical assays on the Vitek2 (bioMérieux, Marcy-l'Étoile, France); no negative statement for the hospital's general research consent as approved by the ethical committee. We summarised patient demographic and clinical data into units of clinical cases, which are defined as a unique hospital stay and included data collected from the same patient between the hospital entry and the hospital exit date.

Generalised Linear Models

The clinical outcomes analysed in this study included (i) invasive infection to the bloodstream (defined as at least one *E. coli* positive blood culture sample) (ii) experience of typical UTI symptoms, (iii) admission to an intensive care unit (ICU) and (iv) all-cause mortality within 30 days of *E. coli* diagnosis. Clinical outcomes were examined for an association with *papGII* carriage of the infecting *E. coli* strain, age, sex, immunosuppression (defined as a dose equivalent of 20 mg prednisone / day or mentioning of immunosuppression in the patient notes), and Charlson Comorbidity Index (CCI) (255). If multiple strains per clinical case had been isolated, we chose *papGII* carriage of strains isolated from blood culture samples over *papGII* carriage of strains isolated from urine samples and if multiple strains had been isolated of the same material, we chose the *papGII* carriage of the strain isolated at the earliest time

point. In order to compare the effect size between the included variables, we scaled and centred the numerical variables 'age' and 'CCI'. All outcomes were binary and were analysed using generalised linear models (GLM) with binomial error distribution. Statistical analyses were performed in R (v 3.7).

MALDI-TOF MS

For a subset of *E. coli* isolates (n=317), MALDI-TOF mass spectra were acquired. This subset was chosen to represent strains encoding *papGII* (n=83) and strains which did not encode *papGII* (n=234), as well as representative isolates of the phylogroups A (n=19), B1 (n=28), B2-1 (n=48), B2-2 (n=141), C (n=7), D1 (n=45), D2 (n=5), D3 (n=13), E1 (n=2) and E2 (n=1). Each strain was measured in quadruplicate on two MALDI-TOF MS devices including a Microflex Biotyper 'smart' (Bruker Daltonics, Bremen, Germany) and an Axima Confidence (Shimadzu, Ngoyo, Japan) using direct smear method and overlaying with 1µl formic acid (25%) and 1µl cyano-4-hydroxycinnamic acid (CHCA) matrix solution.

Mass spectra acquired on the Axmina Confidence were exported as 'mzXml' and mass spectra acquired on the microflex Biotyper as 'fid' files and both were further processed in R using the packages MALDIQuant and MALDIQuantForeign (252): Mass spectra were trimmed to a mass range of 4000 - 20,000, the intensity was transformed ('sqrt') and smoothed (method="SavitzkyGolay", halfWindowSize=20), the baseline was removed (method="SNIP", 40 and 160 iterations for spectra acquired on the microflex Biotyper or the Axima Confidence, respectively) and the intensity was calibrated (method="median") before peaks were detected ("SuperSmoother", halfWindowSize= 20, SNR=2). Peaks were externally calibrated using 23 conserved masses (4364.4 Da, 5095.8 Da, 6371.5 Da, 6446.3 Da, 6541.7 Da, 7273.4 Da, 7288.8 Da, 8499.9 Da, 9006.4 Da, 9704.3 Da, 10430.2 Da, 11564.2 Da, 11580.4 Da, 11735.4 Da, 12769.5 Da, 13133.1 Da, 13540.9 Da, 14126.4 Da, 14875.2 Da, 15281.0 Da, 15768.9 Da, 17603.2 Da, 17711.4 Da) and 1000ppm tolerance in both directions.

Results

Characterisation of bacterial strains

We assessed the phylogroup distributions in our dataset, comprising one *E. coli* isolate per clinical case with invasive (n=251) or non-invasive (n=580) *E. coli* infection (n=831). We observed a large fraction (76.5%; 636/831) belonging to the ExPEC associated phylogroups B2-2 (43.6%; 362/831), B2-1 (17.0 %; 141/831), D1 (12.3%; 102/831), D2 (1.4%; 12/831), D3 (3.8%; 25/831), and F (2.3%; 19/831). Only a smaller fraction of strains belonged to phylogroups associated with colonisation A (6.2%; 50/831), B1 (10.8%; 90/831) and/or intestinal infection E2 (0.2%; 2/831) (**Figure 1A**). We observed 77.6% (642/831) of the strains

carrying genes encoding capsular polysaccharides. Strains in ExPEC-associated phylogroups most often encoded a capsule (95.8%; 633/661) whereas strains in phylogroups associated with colonisation rarely did (6.4%; 9/140). The most common capsule assignments were 'KX03' (25.7%; 165/642) and 'KX29' (17.9%; 115/642), which correspond to the well-known capsule types K1 and K5 (**Figure 1A**).

We next compared the phylogroups in terms of virulence and resistance factors. Phylogroup A strains carried the fewest virulence factors (median = 205.5, Interquartile range (=IQR) = [194.25; 217.75]), while phylogroup B2-2 strains carried the most virulence factors (median = 275.0; IQR = [259.25, 288.0]) (**Figure 1B**).

The phylogroups A, B1, B2-2, D2, D3 and F carried a median of one resistance gene and of the phylogroups which cover more than 10 strains, phylogroup C carried the most resistance genes (median = 5; IQR=[1,7.5]) followed by B2-1 a (median = 2; IQR=[1,4]) resistance genes (**Figure 1C**).

STs which comprise globally successful UPEC clones (198,201,206,287) were enriched in our strain set, namely ST131 (12.6%; 105/831), ST69 (11.1%; 93/831), ST73 (10.1%; 84/831), and ST95 (5.3%; 44/831) being the four most prevalent STs (**Figure 1D**). We further screened our strains for the carriage of *papGII* and the *iuc* operon, which have previously been identified as an important factor in the progression of UTI (198,288). We detected *papGII* in 20.9% of strains (174/831) and in the ExPEC associated phylogroups B2-1 (20.6%; 29/141), B2-2 (27.1%; 98/362), D1 (33.3%; 34/102), D3 (16.7%; 2/12) and F (57.9%; 11/19). *papGII* frequently occurs in blocks of closely related strains (198) (**Figure 1D**). The *iuc* operon often co-occurs with *papGII*, but is more common. The complete *iuc* operon was detected in 28.6% (238/831) of all strains and in the phylogroups B2-1 (52.5%; 74/141), B2-2 (31.8%; 115/362), C (8.7%; 2/23), D1 (30.4%; 31/102), D3 (16.7%; 2/12) and F (47.4%; 9/19).

Phenotypic ceftriaxone resistance, which is often assessed to screen for the carriage of Extended Spectrum β-Lactamase (ESBL), was rare in our dataset (9.7%; 72/740), except in strains of ST131 (51.6%, 48/93) (**Figure 1D**).

We did not record any phenotypic resistance against meropenem (0/740), and phenotypic resistance against fosfomycin and nitrofurantoin was rare (1.5%; 8/550 and 1.1%; 6/549) and distributed throughout the phylogenetic tree (**Figure S1**). Phenotypic resistance against ciprofloxacin was more prevalent (21.2%; 125/589) and occurred most often in ST1193 strains (100%; 18/18) and ST131 strains (67.7%; 63/93) (**Figure S1**).

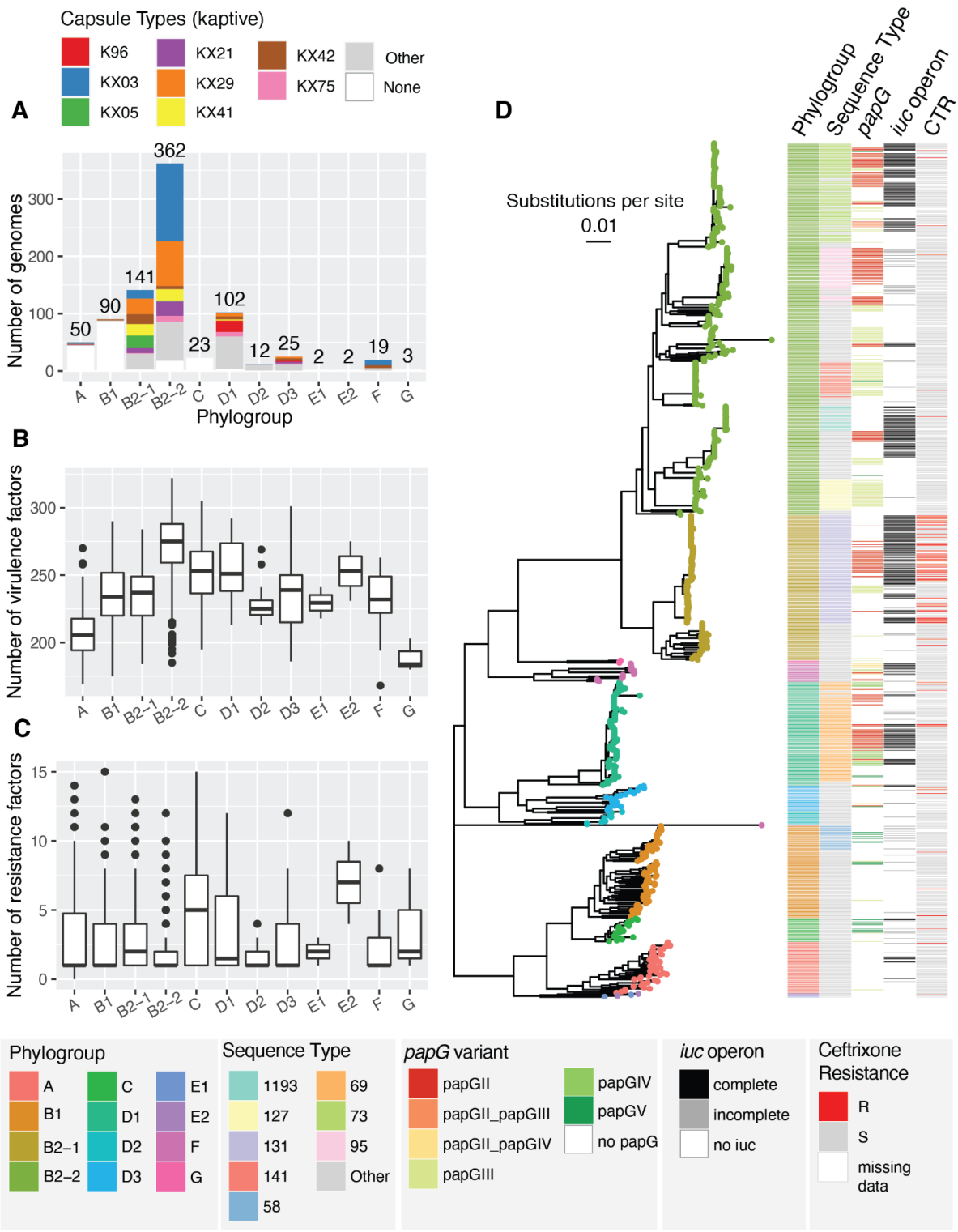

Figure 1: Genomic characterisation of the *E. coli* strains collected for this study (one strain per clinical case, n=831). **A:** Frequency distribution of the *E. coli* phylogroups and their respective capsule types; **B:** Number of virulence factors detected per *E. coli* phylogroup; **C:** Number of genes associated with antimicrobial resistance per *E. coli* phylogroup; **D:** Core genome phylogeny, phylogroup assignment, Sequence Type (ST) (8 most frequent ST are coloured, rarer STs in grey), *papG* variant, occurrence of the *iuc* operon and phenotypic ceftriaxone resistance.

Patient factors

In order to assess which patient characteristics impact the progression of a UTI, we reviewed 831 clinical case charts. Patients included in this study had a median age of 75.3 years (IQR=[63.4,83.0]) and a median CCI of 2 (IQR=[1,3]) and were more often female (69.3%, 576/831) than male (30.7%, 255/831) and 10.6% (86/815) were immunosuppressed (**Figure 2**, **Table 1**). We observed a larger fraction of strains belonging to carriage-associated phylogroups A, B1 and C (121/576, 21.0% vs. 42/255, 16.5%) and of rare ST (302/576, 52.4% vs. 100/255, 39.2%) isolated from female patients compared to male patients, whereas male patients more frequently carried strains of ST131 than female patients (49/255, 19.2% vs. 56/576, 9.7%) (**Figure S2**). While the age of male patients peaks around 80 years, there are two peaks in the age of female patients, one around 80 years, the other around 30 years (**Figure 2**). Female patients younger than 40 years (n=54) were less frequently infected with strains from carriage associated phylogroups than female patients older than 40 years (7/54, 13% vs. 114/522, 21.8%). The most frequent ST isolated from female patients younger than 40 were fromsST95 (8/54, 14.8%), ST69 (8/54, 14.8%) and ST127 (5/54, 9.5%), all of which occurred in lower frequency in female patients older than 40 (22/522, 4.2%; 55/522, 10.5% and 13/522, 2.5%, respectively) (**Figure S3**).

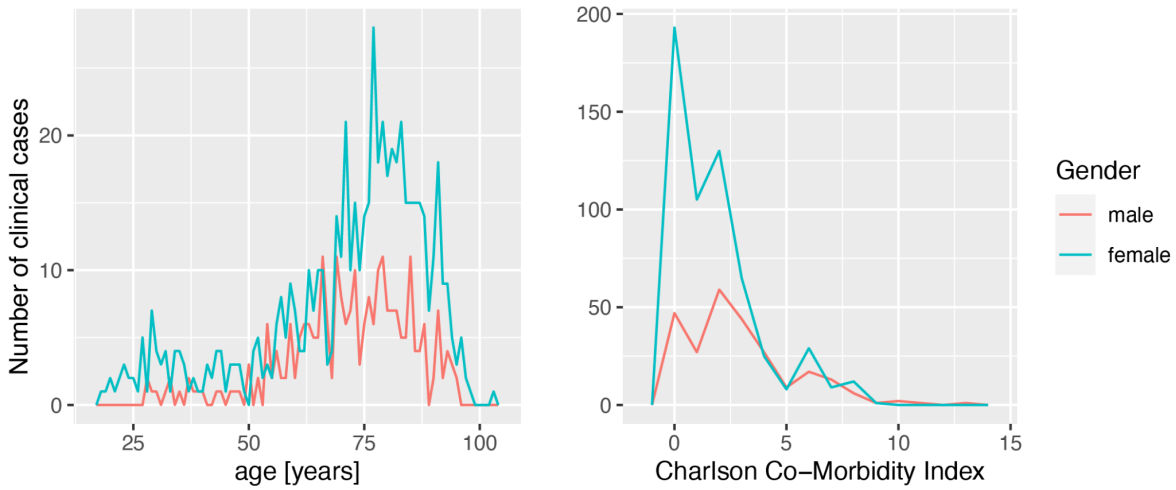

Figure 2: Frequency distribution of age [years] and the sum of the Charlson Co-morbidity Index for male (red) and female (turquoise) clinical cases (n=831) included in this study.

In 31.7% of cases (264/831) at least one blood culture sample tested positive for *E. coli* (='invasive infection'). The composition of phylogroups isolated from invasive and non-invasive infection was similar with 19.3% (51/264) and 19.8% (112/567) being caused by carriage-associated phylogroups A, B1 and C. The three most common STs were the same

in both invasive and non-invasive infection, namely ST69 (39/264, 14.8% and 54/567, 9.5%), ST73 (33/264, 12.5% and 54/567, 9.0%) and ST131 (34/264, 12.9% and 71/567, 12.5%). Rare STs, which overall occurred less than 20 times in our strain collection, caused a smaller fraction of invasive infections than non-invasive infections (109/264, 41.3% vs. 293/567 51.7%) (**Figure S4**).

In 29.6% (232/783) of cases, the patient experienced typical UTI symptoms, 23.1% (191/827) were administered to the ICU and 9.6% (66/685) died within 30 days of the *E. coli* diagnosis (**Table 1**).

Table 1: Patient characteristics and outcome variables of clinical cases (n=831) with invasive (n=255) or non-invasive (n=576) *E. coli* UTI. 'Invasive infection' was defined as at least one positive *E. coli* blood culture, whereas 'non-invasive UTI' was defined as at least one *E. coli* positive urine sample, with no blood culture being tested positive for *E. coli*.

Variable	All cases (n=831)	Cases with invasive infection (n=264)	Cases with non-invasive infection (n=567)
Age [years] (median, [IQR]))	75.3 [63.4,83.0]	75.7 [63.3, 83.4]	74.3 [63.6, 82.3]
Gender = male (n;%)	255/831 (30.7)	122/264 (46.2)	33/567 (23.5)
CCI (median [IQR])	2 [0,3]	2 [0,3]	2 [0,3]
Immunosuppression = T (n; %)	86/815 (10.6)	45/213 (21)	42/516 (8.1)
Typical UTI symptoms	232/783 (29.6)	74/245 (30.2)	158/538 (29.4)
ICU admission	191/827 (23.1)	60/250 (24.0)	131/577 (22.7)
30-day all-cause mortality	66/685 (9.6)	27/209 (12.9)	39/476 (8.2)

Within host genetic diversity

In 12.4% (13/105) of cases where *E. coli* strains isolated from urine and blood cultures of the same patient were available, we found two phylogenetically distant strains with ANI values under 99.9% (**Figure S5A**). In order to compare the strain diversity within urine and blood culture samples, we picked 10 colonies for single colony sequencing each from urine and blood culture samples of three additional cases. In 1/3 urine samples, we observed phylogenetically distant strains, sharing under 99.9% ANI and displaying 71,193 - 80,916 single nucleotide variants (SNV) in pairwise comparisons. When pairwise examining closely

related strains from the same sample (over 99.9 ANI), we identified a higher number of SNVs (median=3, IQR=[2,5]) for strains co-isolated from urine samples than for strains co-isolated from blood culture samples (median=2, IQR=[0,2], p-value = 0.00033, Mann Whitney U test (**Figure S2B**).

Bacterial Genome Wide Association Study (GWAS)

In order to identify bacterial factors promoting the progression of UTI, we performed a GWAS, including one strain per clinical case and using 'bacteraemia' as clinical endpoint. 'Bacteraemia' was defined as at least one positive *E. coli* blood culture, compared to 'non-invasive UTI' which was defined as at least one *E. coli* positive urine sample, with no blood culture being tested positive for *E. coli* within 7 days. We identified three unitigs of the gene *papG* as being significantly associated with 'bacteraemia' (**Figure 3**). All significant *papG* unitigs showed the highest ANI values to the *papG* variant *papGII*.

Multiple genes (n=28) showed association to invasive infection just below the significance threshold and with a log10(p-value) above 4. (**Supplementary Table S2**). Seven of these genes (*papA, papC_3*, *papE*, *papH*, *papK*, *papJ*) encode P-pilus proteins.

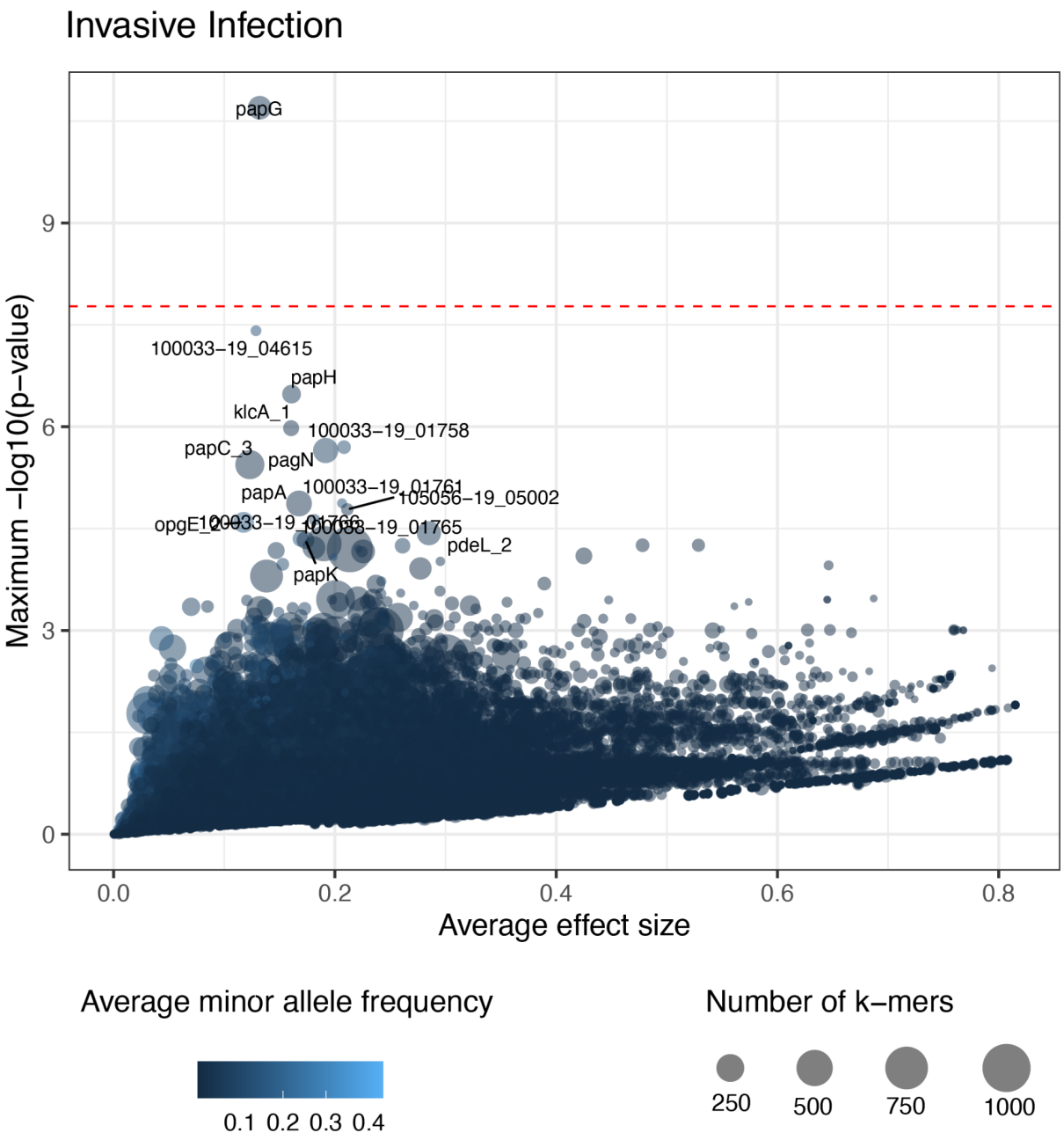

Figure 3: Significance level and average effect size of all genes with mapping unitigs. Red dashed line: significance threshold corrected for multiple testing (log10(1.68E-08)). Genes are labelled if a log10(p-value) > 4.3 was identified.

papGII is associated with invasive infection when correcting for patient characteristics

In order to assess the impact of pathogen and patient-specific factors on the progression of UTI, we concurrently analysed these in a generalised linear model, using 'bacteraemia' as primary endpoint and 'typical UTI symptoms', 'admission to ICU' and '30-day all-cause mortality' as secondary endpoints. As bacterial factors we included *papGII* carriage as the

most prominent hit identified in our GWAS. Further, we included phenotypically tested resistance against ceftriaxone as a bacterial factor in our model.

E. coli strains encoding *papGII* were significantly more likely to be involved in invasive infection (OR 5.91, 95% CI = [3.44,7.88], p-value < 0.001) (**Figure 4**). Clinical cases which carried a *papGII* positive *E. coli* strain developed invasive infection in 55.2% (96/174), whereas 25.3% (166/657) of *papGII* negative clinical cases developed invasive infection. *papGII* thereby has a sensitivity of 36.7%, a specificity of 86.6%, a positive predictive value of 56.3% and a negative predictive value of 74.7%.

Although not significant, and to a lesser extent, we observe a tendency of *papGII* to be more likely than *E. coli* strains which do not encode *papGII* to be involved in (i) clinical cases with typical UTI symptoms (OR 1.45, 95% CI = [0.98,2.15], p-value = 0.063) (**Figure S6A**) and (ii) clinical cases involving an ICU stay (OR 1.44, 95% CI = [0.95,2.18], p-value = 0.085) (**Figure S6B**). Strains encoding *papGII* were isolated from clinical cases for which a significantly higher concentration of C-reactive Protein (CRP) and higher leukocyte counts in blood samples (**Figure S7**), and significantly higher leukocyte and erythrocyte counts in urine samples were detected (**Figure S8**), compared to blood and urine samples from *papGII* negative infected cases. Urine samples of cases for which a *papGII* positive strain was isolated, showed a tendency towards a higher bacterial cell count (median = 7190/uL, IQR = [1401,12818] vs. median = 5216/uL, IQR=[757,11185], p-value = 0.056, Mann Whitney U test) and were less frequently tested positive for nitrite than urine samples of cases for which a *papGII* negative strain was isolated (**Figure S7**), although these differences were not significant (34.0% 54/159 vs. 42.6%; 246/577, p-value = 0.06, Chi-squared test).

We found no evidence of a *papGII* specific association with 30-day mortality (p-value = 0.22) (**Figure S6C**). We observed clinical cases with immunosuppression (OR 2.84, 95% CI = [1.63,4.95], p-value < 0.001) and male patients (OR 3.80, 95% CI = [2.58,5.58], p-value < 0.001) being more likely to be associated with invasive infection (**Figure 4**). Patients 40 years and older were more often infected with *papGII* negative strains (612/764, 80.1%) compared to patients younger than 40 years (45/67, 67.2%) (p-value = 0.02, Chi-squared test), whereas there was no evidence for varying *papGII* frequencies between male and female patients (p-value = 0.35, Chi-squared test), or between immunosuppressed and immunocompetent patients (p-value = 0.47, Chi-squared test) (**Figure S9**).

Older patients were less likely to experience typical UTI symptoms (OR 0.81, 95% CI = [0.68,0.95], p-value = 0.011) (**Figure S6A**) and more likely to die within 30 days of diagnosis (OR 1.96, 95% CI = [1.27,3.02], p-value = 0.002). Immunosuppressed patients (OR 3.11, 95% CI = [1.45,6.70], p-value = 0.004) and patients with an increased CCI (OR 1.68, 95% CI = [1.30,2.16], p-value < 0.001) were more likely to die within 30 days of the *E. coli* diagnosis (**Figure S6C**).

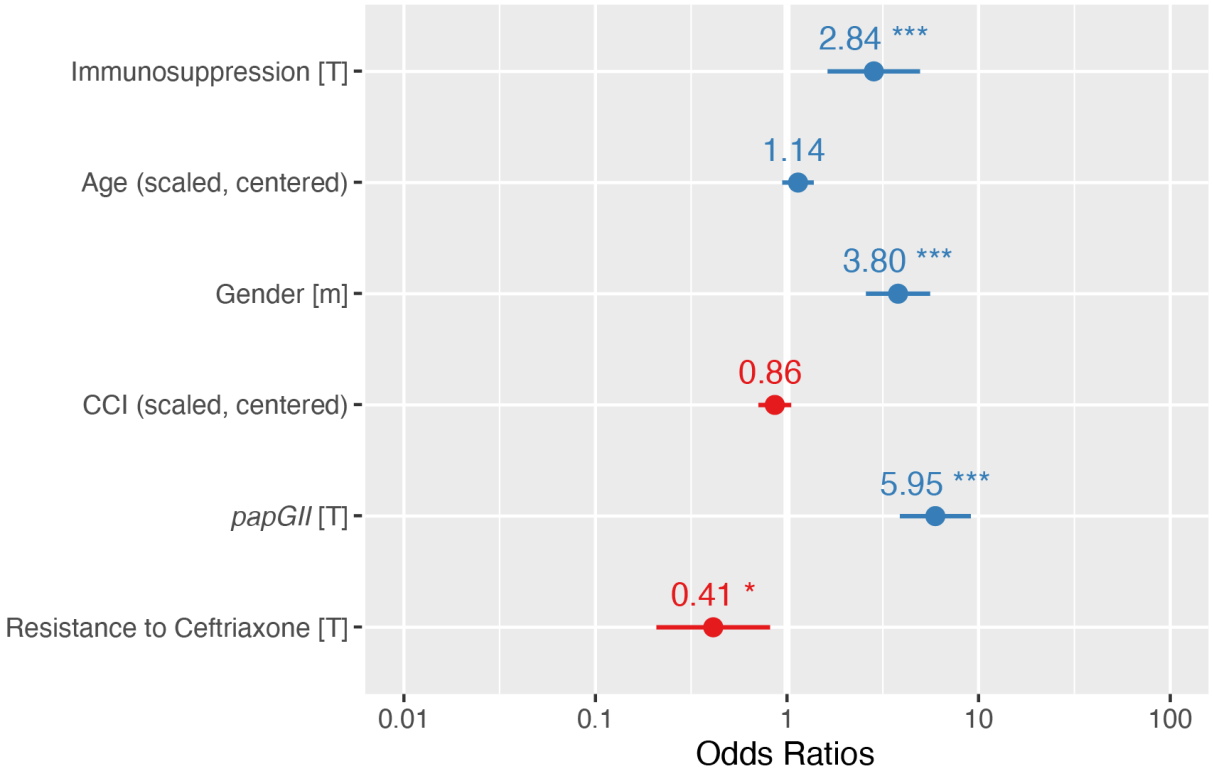

Figure 4: Odds ratio estimates with 95% confidence intervals for invasive infection using the generalised linear model (GLM). *n* = 752 complete observations with 251 events. CCI = Charlson Comorbidity Index

No evidence for *papGII* specific peak in MALDI-TOF mass spectra

Our results and previous studies highlight the importance of *papGII* in invasive *E. coli* infection. Its early detection in clinical diagnostics would be desirable and give evidence towards a severe progression of an ongoing UTI. MALDI-TOF MS is the most widely used tool for bacterial species identification in routine diagnostics and we therefore aimed to assess whether there are *papGII* specific signals in MALDI-TOF MS spectra. We did not identify any *papGII* specific signal, nor a peak being associated with the absence of any *papG* variant (**Figure S10**).

As *papGII* carriage is associated with the ExPEC phylogroups B2-1, B2-2, D1, D2, D3 and F, we further investigated whether the *E. coli* phylogroups can be distinguished by MALDI-TOF MS. We identified a mass shift from around 9711 Da to 9739 Da, which uniquely distinguishes the phylogroups B2-1, B2-2 and F from all other phylogroups. This mass shift has previously been identified as two different mass alleles of the acid stress chaperone HdeA. From MALDI-TOF mass spectra (**Figure S11**) and predicted from genomic data (**Figure S12**) we observe the HdeA mass allele at 9712 Da in the ExPEC associated phylogroups B2-1, B2-2 and F and the mass allele at 9740 Da in the phylogroups A and B1 which are associated with a

colonisation clinical phenotype (**Figure S11** and **Figure S12**). However, the ExPEC associated phylogroups D1, which also carries *papGII*, and D3 also encoded HdeA with a mass of 9740 Da. Strains of the phylogroup D2 were observed to encode HdeA of either 9712 Da or 9740 Da (**Figure S11** and **Figure S12**).

Discussion

In this study we have concurrently analysed bacterial genomic factors and patient characteristics and highlighted the importance of *papGII* in invasive UTI infections. Consistent with previous studies, we identified *papGII* in globally successful UPEC lineages (198). *papGII* often co-occurs with *iuc,* and rarely occurred in strains which were resistant to ceftriaxone, except within the globally disseminating ESBL UPEC lineage ST131.

We observed over 75% of our strains and over 95% of our ExPEC phylogroup strains carrying genes which encode for polysaccharide capsules. This is a substantially larger fraction than has previously been reported in *E. coli* RefSeq sequences, where under 25% of assemblies carried the *kpc* locus (281). Capsular polysaccharides facilitate Gram negative bacteria to evade the innate host immune response (289), and the *E. coli* capsule K1 and K5 have been shown to do so by molecular mimicry, because identical polysaccharides are present on human cells (290). The increased proportion of polysaccharide capsules in our collection gives further evidence on the importance of capsules for *E. coli* to infect the human urinary tract. However, neither the data collected in this study, nor those from previous studies assessing *E. coli* bloodstream infection (198,199) suggested an association of these capsule loci with severe progression of UTI to invasive infection.

In a subset of samples, we identified the *E. coli* strain isolated from the urine sample being unrelated to the *E. coli* strain isolated from the bloodstream from the same clinical case. Possible explanations for this observation are (i) a multi-strain infection of the urinary tract or (ii) a different port of entry to the bloodstream than the urinary tract. In a deeper analysis and in line with previous studies (202,291), we identified strains of different phylogroups isolated from the same urine sample which supports explanation (i). However, as we examined multiple colony picks from the same urine sample only for a small number of cases, further studies are required to assess the within-host *E. coli* strain diversity infecting the urinary tract and their dynamics in disease progression.

In line with a previous study (198) we identified *papGII* as being associated with invasive infection in our bacterial GWAS. However, there was no evidence that *iuc* is associated with invasive infection in our data.

The crucial role of *papGII* in invasive infections is further supported by our GLM, where we concurrently correct our analysis for important patient characteristics as well as resistance

against ceftriaxone. Our analysis also identified the increased likelihood of male patients developing invasive infection, which can presumably be explained by the increased occurrence of uncomplicated UTI in female patients (292).

We did not consider other portals of entry than the urinary tract, which has been determined as the entry point in 50-60% of *E. coli* sepsis (199), and which certainly is a limitation of the current study as this might influence the clinical outcomes. However, as *papGII* binds to human uroepithelial and kidney cells (196,275), its impact on patient outcome might be even larger when exclusively examining patients with urosepsis, compared to all bloodstream infections.

We examined whether there are MALDI-TOF MS peaks which are specific for *E. coli* strains carrying *papGII* and which could be used to detect these virulent UPEC clones in clinical routine diagnostics. The *papGII* protein weighs 37,667 Daltons (293) and lies beyond the mass range of MALDI-TOF MS devices routinely used for bacterial species identification, and unfortunately, no alternative peak was observed which could serve as a surrogate marker. We did, however, observe a previously described mass shift (294) of the acid stress chaperone *HdeA* which is specific for the ExPEC phylogroups B2-1, B2-2 and F, but fails to discriminate the ExPEC phylogroup D1-3 from the carriage associated phylogroups A, B1 and C. Although in this simple analysis no single signal was observed which unambiguously identifies *papGII* positive strains from MALDI-TOF mass spectra, more elaborate statistical analysis might still reveal a combination of peaks or intensity patterns allowing for the identification of such strains.

Conclusions

By performing a bacterial GWAS on genomes of 831 isolates, and by correcting for important patient characteristics, we highlight the importance of *papGII* in invasive UPEC infection. As no *papGII* specific signal was observed in MALDI-TOF mass spectra, rapid sequence based diagnostic assays represent a more promising approach for the early detection of this important factor, allowing the anticipation of disease progression.

Acknowledgments

We thank Magdalena Schneider, Christine Kiessling, Elisabeth Schultheiss, Rosa-Maria Vesco, Clarisse Straub, Josiane Reist, Olivia Grüninger, Daniela Lang and Diana Albertos-Torres for the excellent technical assistance with strain collection, library preparations and sequencing of the bacterial isolates (all University Hospital Basel). We thank Dr. Deborah R. Vogt (Department of Clinical Research, University of Basel and University Hospital Basel, Basel, Switzerland) for consultations regarding the GLM analysis and Dr. Michael Biggel

(University of Zurich) for consultation regarding the GWAS analysis. We thank Dr. Fanny Wegner for valuable feedback on this manuscript. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel.

Supporting information captions

Figure S1: Core genome phylogeny of 831 *E. coli* strains. Columns represent (from left to right): the assigned phylogroup, the sequence type, phenotypic resistance against Ceftriaxone, Meropenem, Fosfomycin, Nitrofurantoin and Ciprofloxacin

Figure S2: Distribution of *E. coli* phylogroups (left) and Sequence Types (ST) (right) in male (n=255) (upper row) and female (n=576) (lower row) patients.

Figure S3: Distribution of *E. coli* phylogroups (left) and Sequence Types (ST) (right) in in female patients (n=576) younger than 40 years (n=54) (upper row) and older than 40 years (n=522) (lower row) patients.

Figure S4: Distribution of *E. coli* phylogroups (left) and Sequence Types (ST) (right) in invasive infections (n=251) (upper row) and non-invasive infections (n=580) (lower row) patients.

Figure S5: Within host genetic diversity of *E. coli* strains isolated from the same clinical cases **A:** core genome phylogeny of *E. coli* strains (n=225), isolated from the same clinical case (n=105), coloured by phylogroup. The numbers correspond to the case identifier and strains were only labelled, if they exhibited < 99.9% Average Nucleotide Identity to the strain isolated from the same clinical case. **B:** Single Nucleotide Variants (SNV) of 10 picked isolates from three cases, either from urine or blood culture samples.

Figure S6: Odds ratio estimates with 95% confidence intervals for **A:** Typical urinary tract infection symptoms (n = 783 complete observations with 232 events); **B:** Admission to the intensive care unit (n = 827 complete observations with 191 events); **C:** 30-day all-cause mortality (n = 685 complete observations with 66 events); using generalised linear models (GLM). CCI = Charlson Comorbidity Index

Figure S7: Bacterial cell count (**A**), nitrite status (**B**), leucocyte count divided by bacterial cell count (**C**) and erythrocyte count divided by bacterial cell count (**D**) measured in urine samples of cases, for which a *papGII* positive or a *papGII* negative *E. coli* strain was isolated from a urine or a blood culture sample

Figure S8: C-reactive protein concentration (**A**) and leucocyte count (**B**) measured in blood samples of cases, for which a *papGII* positive or a *papGII* negative *E. coli* strain was isolated from a urine or a blood culture sample

Figure S9: Relative occurrence of *papGII* in isolates from patients younger (n=67) vs. older (n=764) than 40 years, in male (n=255) vs. female (n=576) patients and in patients which were immunosuppressed (n=86) vs. patients which were not immunosuppressed (n=729).

Figure S10: Occurrence of MALDI-TOF mass peaks in spectra acquired from *E. coli* strains encoding no *papG* gene, encoding a *papG* variant other than *papGII* and encoding *papGII*. Each strain was measured in quadruplicate either on a Microflex Biotyper device, or an Axmina Confidence device. Masses are only depicted if detected in > 30% or < 25% of spectra for one or more of the groups.

Figure S11: Occurrence of MALDI-TOF mass peaks in spectra acquired from *E. coli* strains of different phylogroups. Each strain was measured in quadruplicate either on a Microflex Biotyper device, or an Axmina Confidence device. Phylogroups for which less than five strains were available (E1, E2 and G) were excluded from the plot. Masses are only depicted if detected in > 50% or < 25% of spectra for one or more of the groups.

Figure S12: Core genome phylogeny of the *E. coli* strains collected for this study (one strain per clinical case, n=831). Phylogroup assignment, Sequence Type (ST) (eight most frequent ones coloured, rarer STs in grey), *papG* variant, mass of HdeA, predicted from the amino acid sequence

Table S2: Genes which showed a non-significant tendency towards being associated with invasive infection (log10(p-value) > 4, but below the significance threshold).

Chapter III: Factors associated with MALDI-TOF mass spectral quality of species identification in clinical routine diagnostics

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 16 March 2021
doi: 10.3389/fcimb.2021.646648

Factors Associated With MALDI-TOF Mass Spectral Quality of Species Identification in Clinical Routine Diagnostics

Aline Cuénod[1,2*], *Frédéric Foucault*[3], *Valentin Pflüger*[3] and *Adrian Egli*[1,2]

* corresponding authors:
Aline Cuénod

This manuscript has been published in *Frontiers in Cellular and Infection Microbiology*:

Cuénod, Aline et al. “Factors Associated With MALDI-TOF Mass Spectral Quality of Species Identification in Clinical Routine Diagnostics.” *Frontiers in Cellular and Infection Microbiology* 11 (2021). https://doi.org/10.3389/fcimb.2021.646648.

My contributions:

- Conceptualisation of the study
- Acquisition of MALDI-TOF mass spectra using multiple sample preparation protocols
- Bioinformatic analysis of whole genome sequences (amongst others) and MALDI-TOF mass spectra
- Data visualisation (all figures)
- Writing of the original manuscript

Note: The following part contains the full manuscript

The Supplementary material can be accessed via the following link:

https://www.frontiersin.org/articles/10.3389/fcimb.2021.646648/full#supplementary-material

Factors associated with MALDI-TOF mass spectral quality of species identification in clinical routine diagnostics

Aline Cuénod[a,b], Frédéric Foucault[c], Valentin Pflüger[c], Adrian Egli[a,b]

[a] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland

[b] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland

[c] Mabritec AG, Riehen Switzerland

Abstract

Background: An accurate and timely identification of bacterial species is critical in clinical diagnostics. Species identification allows a potential first adaptation of empiric antibiotic treatments before the resistance profile is available. Matrix assisted Laser Desorption Ionization Time of Flight mass spectrometry (MALDI-TOF MS) is a widely used method for bacterial species identification. However, important challenges in species identification remain. These arise from (i) incomplete databases, (ii) close relatedness of species of interest, and (iii) spectral quality, which is currently vaguely defined.

Methods: We selected 47 clinically relevant bacterial isolates from 39 species, which can be challenging to identify by MALDI-TOF MS. We measured these isolates under various analytical conditions on two MALDI-TOF MS systems. First, we identified spectral features, which were associated with correct species identification in three different databases. Considering these features, we then systematically compared spectra produced with three different sample protocols. In addition, we varied quantities of bacterial colony material applied and bacterial colony age.

Results: We identified (i) the number of ribosomal marker peaks detected, (ii) the median relative intensity of ribosomal marker peaks, (iii) the sum of the intensity of all detected peaks, (iv) a high measurement precision, and (v) reproducibility of peaks to act as good proxies of spectral quality. We found that using formic acid, measuring bacterial colonies at a young age, and frequently calibrating the MALDI-TOF MS device increase mass spectral quality. We further observed significant differences in spectral quality between different bacterial taxa and optimal measurement conditions vary per taxon.

Conclusion: We identified and applied quality measures for MALDI-TOF MS and optimized spectral quality in routine settings. Phylogenetic marker peaks can be reproducibly detected and provide an increased resolution and the ability to distinguish between challenging species

such as those within the *Enterobacter cloacae* complex, *Burkholderia cepacia* complex, or viridans streptococci.

Keywords

MALDI-TOF MS, Quality Control, Standardisation, species identification, microbial diagnostics

Introduction

Matrix assisted Laser Desorption Ionization Time of Flight mass spectrometry (MALDI-TOF MS) has revolutionised microbial diagnostics. Due to its minimal hands-on and turn-around time, low costs, and high accuracy it has become the method of choice for bacterial species identification in clinical diagnostics (295,296). Multiple studies have highlighted the potential of MALDI-TOF MS to identify virulent or resistant bacterial sub-lineages within a species (297,298). Despite these potential applications, important challenges remain for routine diagnostics, such as the inability to properly differentiate clinically relevant taxonomic groups, such as the species within the *Burkholderia cepacia* complex (299), the *K. pneumoniae* complex (300) or viridans *s*treptococci (301). Challenges in species identification arise from (i) incomplete databases, (ii) close relatedness of the bacterial species of interest, and (iii) poor spectral quality.

Species identification through commonly used MALDI-TOF MS systems is based on the comparison of unknown spectra to reference spectra databases through pattern matching. MALDI-TOF mass spectra consist of peaks from highly abundant, intracellular proteins including ribosomal subunit proteins, which are present in high copy numbers in replicating bacterial cells (235,302). With the abundance of bacterial whole genome sequences, reference databases comprising predicted ribosomal subunit masses have become an alternative to pattern based microbial identification in MALDI-TOF MS. The mass of ribosomal subunits can be directly calculated from genomic sequences, as they are relatively conserved and rarely post-translationally modified. Their potential to serve as MALDI-TOF MS biomarkers has been applied to clinically relevant phylogenetic groups (76,85,303), and multiple databases using marker masses predicted from genomic data are now available (304–306). A ribosomal marker based approach has successfully been applied to distinguish between subspecies and clonal complexes within species such as *Streptococcus agalactiae* and *Escherichia coli* (76,307,308).

When MALDI-TOF MS was first applied for microbial species identification (234) and in its first years in routine diagnostics, samples were processed using a protein extraction protocol (309). However, as high accuracies in species identification have been reported using a much

simpler procedure, applying bacterial colonies directly onto the MALDI-TOF MS target plate has become the standard sample preparation protocol in routine diagnostics (310,311). Although the Clinical and Laboratory Standard Institute (Pennsylvania, USA) has published a guideline on bacterial identification by MALDI-TOF MS (100), the definition criteria of spectral quality remain vague. Many diagnostic laboratories have developed their own Standard Operating Procedures for sample preparation and interpretation of species identification by MALDI-TOF MS. Currently it is already well established that MALDI-TOF mass spectral quality is influenced by the amount of bacterial colony material added to the target plate, the age of the bacterial colony, as well as the sample preparation protocol used (312–314). However, there is a clear lack of an optimal and standardised sample preparation and data analysis workflow. Criteria defining the spectral quality will help to compare differences in preparation and analytical workflows. Closing this gap will substantially increase the reproducible detection of phylogenetic marker peaks in MALDI-TOF mass spectra acquired and thereby improve species identification in routine diagnostics.

The purpose of this study is to (i) identify quantitative spectral features suitable to define spectral quality, (ii) compare the influence of sample protocols for bacterial identification by MALDI-TOF MS, and (iii) raise awareness for the potential of an increased resolution of MALDI-TOF MS for subtyping and associated limitations.

We have selected 47 clinically relevant bacterial isolates from 39 species and measured these under various conditions on two different MALDI-TOF MS systems. First, we identified spectral features, which positively correlate with correct species identification. Considering these, we systematically compared spectral quality produced with different sample protocols, with varying amounts of bacterial colony material applied, and with varying bacterial colony age.

Materials & Methods

Bacterial Isolates

We selected 47 clinically relevant bacterial isolates from public and in-house strain collections. The included 39 species can be challenging to identify using MALDI-TOF MS, either because intracellular proteins cannot be ionised easily due to cell wall composition (e.g. *Corynebacterium spp.*), or because of their close relatedness to another bacterial species (e.g. *Klebsiella oxytoca* / *Klebsiella michiganensis*; *Shigella* / *Escherichia coli*).

The bacterial isolates were assigned to 8 phylogenetic groups (**Table 1**). For the strains in each group, we expect both comparable spectral features (e.g. total number of peaks detected) and lysis characteristics, respectively. For the evaluation of species identification, the group '*Streptococcus*' was further split up into 'viridans streptococci' and 'other streptococci' as the former group are of special interest in clinical routine diagnostics.

Table 1. Strains included in this study. Strains were either retrieved from in-house or commercial strain collections. Strains which are assigned the same 'Group' are expected to respond similarly to varying sample protocols, quantities of bacterial colony material applied and varying bacterial colony age.

#	Species	Strain collection	Internal strain number	Group	NCBI / ENA Accession Number
01	*Klebsiella pneumoniae*	in-house	602149-19	*Enterobacteriaceae*	SAMN16951201
02	*Klebsiella oxytoca*	in-house	708776-17	*Enterobacteriaceae*	SAMN12212273
03	*Klebsiella grimontii*	in-house	132656-17	*Enterobacteriaceae*	SAMN12212117
04	*Klebsiella michiganensis*	in-house	401065-17	*Enterobacteriaceae*	SAMN12212153
05	*Klebsiella aerogenes*	in-house	717657-17	*Enterobacteriaceae*	SAMN12212322
06	*Klebsiella variicola*	in-house	717892-17	*Enterobacteriaceae*	SAMN12212293
09	*Escherichia coli*	in-house	807627-2-16	*Enterobacteriaceae*	SAMN16951202
10	*Escherichia coli*	in-house	807628-3-16	*Enterobacteriaceae*	SAMN16951203
11	*Escherichia coli*	in-house	804255-13	*Enterobacteriaceae*	SAMN16951204
12	*Escherichia coli*	in-house	805237-12	*Enterobacteriaceae*	SAMN16951205
13	*Shigella flexneri*	in-house	300666-18	*Enterobacteriaceae*	SAMN16951206
14	*Shigella flexneri*	in-house	301552-18	*Enterobacteriaceae*	SAMN16951207
15	*Shigella sonnei*	commercial	DSMZ-5570	*Enterobacteriaceae*	SAMN16951208
16	*Shigella sonnei*	in-house	301974-17	*Enterobacteriaceae*	SAMEA104430192
35	*Enterobacter sichuanensis*	in-house	403902-15	*Enterobacteriaceae*	SAMN16951209
36	*Enterobacter hormaechei*	commercial	ATCC-49162	*Enterobacteriaceae*	SAMN16951210
37	*Enterobacter asburiae*	commercial	ATCC-35956	*Enterobacteriaceae*	SAMN16951211
38	*Enterobacter ludwigii*	commercial	DSMZ-15213	*Enterobacteriaceae*	SAMN16951212
07	*Listeria monocytogenes*	in-house	107373-13	*Listeria*	SAMN16951213
08	*Listeria monocytogenes*	in-house	O1910-17	*Listeria*	SAMN16951214

17	*Burkholderia cepacia*	in-house	208050-16	*Burkholderia*	SAMN16951215
18	*Burkholderia contaminans*	in-house	O-13	*Burkholderia*	SAMEA54114418
19	*Burkholderia multivorans*	in-house	O-10	*Burkholderia*	SAMEA54118168
20	*Burkholderia cenocepacia*	in-house	O-3	*Burkholderia*	SAMEA54110668
21	*Bordetella bronchiseptica*	in-house	502474-16	*Bordetella*	SAMN16951216
22	*Bordetella pertussis*	commercial	ATCC-9797	*Bordetella*	SAMN16951217
23	*Bordetella parapertussis*	commercial	ATCC-53893	*Bordetella*	SAMN16951218
24	*Streptococcus pneumoniae*	in-house	144265-17	*Streptococcus*	SAMN16951219
25	*Streptococcus infantis*	in-house	131226-17	*Streptococcus*	SAMN16951220
26	*Streptococcus gordonii*	commercial	ATCC-33399	*Streptococcus*	SAMN16951221
27	*Streptococcus gallolyticus*	in-house	PRA0000041	*Streptococcus*	SAMN16951222
28	*Streptococcus lutetiensis*	commercial	DSMZ-15350-TS	*Streptococcus*	SAMN16951223
29	*Streptococcus pseudopneumoniae*	in-house	610886-17	*Streptococcus*	SAMN16951224
30	*Streptococcus equinus*	commercial	ATCC-9812	*Streptococcus*	SAMN16951225
31	*Streptococcus dysgalactiae*	in-house	STO0000159	*Streptococcus*	SAMN16951226
32	*Streptococcus dysgalactiae*	in-house	602125-13	*Streptococcus*	SAMN16951227
39	*Staphylococcus aureus*	in-house	351358-18	*Staphylococcus*	SAMN16951228
40	*Staphylococcus schweitzeri*	commercial	DSM-28300-TS	*Staphylococcus*	SAMN16951229
41	*Staphylococcus argenteus*	commercial	DSMZ-28299-TS	*Staphylococcus*	SAMN16951230
42	*Corynebacterium amycolatum*	commercial	ATCC-700206	*Actinobacteria*	SAMN16951231
43	*Corynebacterium urealyticum*	commercial	DSMZ-7109	*Actinobacteria*	SAMN16951232
44	*Gardnerella vaginalis*	commercial	ATCC-14018	*Actinobacteria*	SAMN16951233
45	*Winkia neuii*	in-house	STO0000012	*Actinobacteria*	SAMN16951234
46	*Actinomyces israelii*	commercial	ATCC-10048	*Actinobacteria*	SAMN16951235

47	*Pasteurella multocida*	commercial	ATCC-11039	Gram negative Anaerobes	SAMN16951236
33	*Bacteroides fragilis*	in-house	609216-11	Gram negative Anaerobes	SAMN16951237
34	*Bacteroides fragilis*	in-house	600609-16	Gram negative Anaerobes	SAMN16951238

Whole Genome Sequencing

Isolates were grown on Columbia 5% Sheep Blood Agar (bioMérieux, Marcy-l'Étoile, France) and DNA was extracted using the QIACube with the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany). After quality control of the DNA by Tapestation (Agilent, Santa Clara, USA), tagmentation libraries were generated as described by the manufacturer (Nextera Flex kit, Illumina, San Diego, USA). The genomes were sequenced under 24x multiplexing using a paired end 150 base pairs V3 reaction kit on an Illumina NextSeq500 instrument (Illumina) reaching an average coverage of approximately 60-fold for all isolates. The resulting raw reads were and assembled using Spades (v3.13) (277) via Unicycler (v0.3.0b) (239) using default settings. All accession numbers can be found in **Table 1.** Species identification of all strains was performed by comparing genomic sequences to bacterial type strains using Average Nucleotide Identity (ANIm) (315) and via the TrueBac ID database (316). For strains of the genus *Bordetella* we used ribosomal Multi Locus Sequence Typing for additional confirmation of the species identity (279).

In silico prediction of ribosomal subunit protein masses from WGS data

The molecular weight of 56 ribosomal subunits were predicted as previously described (76,85). Briefly, tblastn (v 2.2.31+) was used to extract the amino acid sequences of 56 ribosomal subunits from whole genome assemblies. The most frequent post translational modifications (82), specifically N-terminal methionine loss (317) and methylation, were considered for subsequent prediction of the monoisotopic molecular weights of the ribosomal subunit proteins. For the ribosomal subunit protein L33, we added 15 Daltons to the predicted molecular weight for the genera *Enterobacter*, *Escherichia*, *Shigella*, *Klebsiella* and *Pasteurella*, accounting for a single methylation of these proteins.

Spectra quality variables

All scripts used in the course of this study can be accessed via GitHub (https://github.com/appliedmicrobiologyresearch/MALDI-TOF-mass-spectral-quality-study).

We queried each spectrum for the following features: (i) number of peaks, (ii) peak with the highest m/z value, (iii) m/z value of the peak at the 90th percentile, (iv) fraction of peaks with a m/z value > 10,000, and (v) sum of the intensity of all detected peaks. As the highest peak often corresponds to technical artefacts, we included the m/z value of the peak at the 90th percentile for further analysis.

Furthermore, we queried each spectrum for the presence and intensity of ribosomal marker peaks predicted from the genomic sequence of the respective strain using an 800 ppm error range. If multiple peaks were detected in this error range, the one with the highest intensity and lower measurement error was considered for further analysis. Bacterial strains encode variable number of ribosomal markers in the mass range of 2,000 - 20,000 Da. We therefore normalised the number of detected marker peaks by dividing through the number of predicted ribosomal marker peaks in the MALDI-TOF MS mass range, when comparing between the bacterial taxa.

To quantify measurement error, we calculated the mean distance between predicted and detected m/z value of ribosomal marker peaks for each spectrum. In order to estimate reproducibility, we calculated the 'fraction of reproducibly detected peaks'. We defined this as the number of peaks, which were detected in at least three out of four technical replicates using a bin size of 800 ppm, divided by the number of peaks in each spectrum. A more detailed and graphical explanation of the MALDI-TOF mass spectral features analysed in this study can be found in **Supplementary Methods 1**.

Further, we evaluate which of the above MALDI-TOF mass spectral features are good proxies for spectra quality and are associated with a correct species identification. We compared spectra for which the correct species was identified to spectra where the correct species could not be identified. Spectra for which the correct species could not be identified included spectra with wrong species being identified and spectra for which no species identification was possible. Henceforward, we will refer to these collectively as 'incorrectly identified spectra'. We performed this analysis exclusively on species which are covered by all three databases included in this study (**Table S1**) and excluded empty spectra.

In order to assess how spectra quality impacts species identification accuracy, we included spectra acquired using the 'direct smear', the '25% formic acid (FA) overlay' or the 'simple protein extraction' method (see section 'Variation of sample preparation' for details) of the *Enterobacter cloacae complex* (*Enterobacter hormaechei*, *Enterobacter asburiae* and *Enterobacter ludwigii*), the *Burkholderia cepacia complex* (*Burkholderia contaminans*, *Burkholderia multivorans* and *Burkholderia cenocepacia*), and viridans streptococci (*Streptococcus pneumoniae* and *Streptococcus pseudopneumoniae*). We assigned these spectra to three intensity levels, by dividing the sum of the intensities of all detected peaks in three equal parts per group.

MALDI-TOF MS spectra acquisition

All MALDI-TOF mass spectra acquired for this study can be accessed via the Open Science Foundation (https://osf.io/ksz7r/). The bacterial isolates were cultured from Microbank™ freezing beads (Pro-lab Diagnostics, Toronto, Canada) onto 5% Sheep Blood agar plates (bioMérieux, Marcy-l'Étoile, France) and subcultured before MALDI-TOF mass spectra acquisition. Strains were incubated under aerobic conditions at 37 °C except for strains of the species *Bacterioides fragilis, Actinomyces israelii,* and *Winkia neuii,* which were incubated under anaerobic conditions using a whitley A95 anaerobic workstation (Don Whitley Scientific Limited, Bingley, United Kingdom). Strains of the species *Streptococcus pneumoniae*, *Bordetella pertussis,* and *Bordetella parapertussis* were incubated under 5 % enriched CO_2 conditions. All mass spectra were acquired on reusable steel target plates (MBT Biotarget 96 (Bruker Daltonics, Bremen, Germany) and steel target plates (Industrietechnik mab AG (Basel, Switzerland)).

Variation of sample preparation

We cultured the bacterial strains as described above. We prepared the strains under three different short protocols, all of which are frequently used in microbial diagnostics: (i) 'Direct smear' method: using a plastic inoculation needle, we transferred bacterial colonies onto a steel target plate and overlaid each spot with 1 µl α-Cyano-4-hydroxycinnamic (CHAC) matrix (Sigma-Aldrich, St. Louis, USA) and left it to air dry completely before MALDI-TOF MS measurements. (ii) '25 % FA overlay': using a plastic inoculation needle, we transferred bacterial colonies onto a steel target plate and overlaid each spot with 1 µl of 25 % formic acid (Sigma-Aldrich, St. Louis, USA) and left it to air dry completely before applying 1 µl of CHAC matrix onto each spot. The target plates were left to air dry completely before MALDI-TOF MS measurements. (iii) 'Simple protein extraction': we transferred a heaped 1 µl inoculation loop of bacterial colony material into 1 ml PBS, rigorously vortexed, and centrifuged for 5 min at 17,000 x g. We removed the supernatant and added 30 µl 70% formic acid and dissolved the pellet by pipetting up and down. 30 µl acetonitrile (Sigma-Aldrich, St. Louis, USA) were added and the mixture was vortexed before centrifuging for 5 minutes at 17,000 x g. Next, 5 µl of the supernatant were mixed with 25 µl of CHAC matrix before spotting onto the steel target plate. We performed measurements as quadruplicate on a Bruker microflex LT/LS 'smart' (Bruker Daltonics, Bremen, Germany) and a Shimadzu Axima Confidence (Shimadzu, Kyoto, Japan) MALDI-TOF MS device as technical replicates and repeated on three different days with fresh subcultures as biological replicates.

Variation of bacterial colony age

We grew the strains over 1, 2, 3, 4, 5 or 6 nights before preparing them for measurement using the '25% FA overlay' method described above. Each overnight culture corresponds to 18 - 24 hours of incubation time. We performed measurements as quadruplicate on a Bruker microflex LT/LS 'smart' and a Shimadzu Axima Confidence MALDI-TOF MS device.

Variation of bacterial colony material quantity

The amount of bacteria transferred onto the MALDI-TOF MS steel target plate has been shown to impact spectral quality (100). The direct transfer of bacteria onto a steel target plate is difficult to standardise. We therefore decided to measure bacterial suspensions at different dilutions to assess the impact of the amount of bacterial colony material measured on mass spectral quality. We randomly selected the following two strains per phylogenetic group for this experiment: *Enterobacteriaceae:* #07, #08*; Listeria:* #09, #10; *Burkholderia:* #17, #19; *Bordetella:* #21, #23; *Streptococcus:* #26, #27; *Staphylococcus:* #39, #40; *Actinobacteria*: #45, #46; Gram negative anaerobes: #33, #34.

We transferred a heaped 1 µl inoculation loop of bacterial colony material into 200 µl of TMA (1x) buffer (Sigma-Aldrich, St. Louis, USA). Next, 5 µl of the bacterial mixture was diluted in 25 µl of CHAC matrix (1:5 dilution) and spotted onto the target plates. 5 µl of the suspension were transferred into a new tube containing 25 µl CHAC matrix (1:25 dilution). We continued the serial dilution up to a factor of 1:15,625. As the majority of measurements with 1:3,125 and 1:15,625 dilutions yielded empty spectra, these were excluded from further analysis. We measured quadruplicates on two MALDI-TOF MS devices as technical replicates and repeated on three different days as biological replicates.

Variation of time after calibration to assess the impact on measurement precision

We performed the measurements on two microflex Biotyper devices (LH/LS and LH/LS 'smart'). Both devices were calibrated using the Bacterial Test Standard (BTS, Part.-Nr. 8255343, Bruker Daltonics, Bremen, Germany) and steel target plates (Bruker Daltonics, Bremen, Germany).

We used an *E. coli* strain from our strain collection (*E. coli* 805237-12) for these measurements as strains of this species generally yield rich spectra using routine sample preparation. We transferred bacteria onto a steel target plate, overlaid with 1 µl 70% FA, left to air dry completely before applying 1 µl CHAC matrix. Each measurement was performed in quadruplicate on two different target plates and MALDI-TOF MS devices as technical replicates and repeated by picking three different colonies as biological replicates. Spectra were acquired on days 1-7 after calibration on the same target plate which was used for calibration. BTS was measured on the row A of the target plate, measurements on day 1-7

after calibration were measured on rows B, C, D, E, F, G and H, respectively. Both MALDI-TOF MS devices where used for microbial species identification in routine diagnostics over the duration of this experiment with a median of 39 (Interquartile range (IQR): [32, 51]) and 137 (IQR: [123, 173]) of routine measurements per day on the microflex Biotyper LH/LS and the microflex Biotyper LH/LS 'smart', respectively.

MALDI-TOF MS spectra processing

In order to be most comparable to spectra acquired and processed in microbial routine diagnostic, we picked the peaks using default settings by the softwares included in the microflex Biotyper or the Axima Confidence system. Spectra acquired on microflex Biotyper devices were exported as 'fid' files and peak picking was performed in the flexAnalyses software (v3.4) and exported as '.txt' files. Spectra acquired on the Axima Confidence devices were exported as '.mzXml' files. These do already exclusively contain m/z values and intensities of picked peaks and were converted to '.txt' files. We subsequently exclusively worked with the intensity and m/z value of these picked peaks, and did not consider further peak characteristics such as the resolution or the signal to noise ratio of a peak.

We excluded spectra as contaminations for which the identified genus did not match the genus identified by ANIm. Strain 17 and strain 20 are missing in one out of three repetitions of the 'simple protein extraction' protocol, strain 32 is missing from day 6 and of strain 46 only three technical replicates were acquired on the Axima Confidence device using the 'direct smear' method.

Species identification

Each spectrum acquired on a Bruker device was compared to the MALDI Biotyper database (MALDI Biotyper Compass Library, Revision E (Vers. 8.0, 7854 MSP, RUO)) included in the included in the flexControl Software v3 (Bruker Daltonics, Bremen, Germany). Spectra acquired on the Axima Confidence device were analyzed with the VitekMS database (bioMérieux, Marcy-l'Étoile, France) (v3.2). Furthermore, we compared each spectrum to a ribosomal marker based database (PAPMID™ (305), Mabritec AG, Riehen, Switzerland). In this study, we used this marker-based approach as a subtyping module and each spectrum was compared only to a subset of bacterial species. Spectra of the species *Escherichia coli*, *Shigella flexneri*, *Shigella sonnei*, *Streptococcus gordonii*, *Streptococcus gallolyticus*, *Streptococcus lutetiensis*, *Streptococcus equinus*, *Streptococcus dysgalactiae, Corynebacterium amycolatum, Corynebacterium urealyticum, Gardnerella vaginalis, Winkia neuii, Actinomyces israelii, Pasteurella multocida,* and *Bacteroides fragilis* were compared to databases including mass profiles of the respective bacterial family. Spectra of the closely

related *Klebsiella spp.*, *Enterobacter cloacae complex, Listeria spp., Burkholderia cepacia complex, Bordetella spp., Staphylococcus aureus complex*, and viridans streptococci were identified using marker based subtyping modules exclusively including the species of the respective phylogenetic complex. Henceforward, species identification by these subtyping modules and using PAPMID^{TM} database, both based on the detection of ribosomal marker peaks will be referred to as PAPMID^{TM}.

The MALDI Biotyper system classifies species identification according to log scores: mass spectra yielding a log score above 2.0 are assigned the label 'highly confidence identification', whereas spectra with a log score between 1.7 and 2.0 are assigned 'low confidence identification'. Spectra with a log score lower than 1.7 are assigned 'no organism identification possible'. For each spectrum with a log score above 1.7, we evaluated whether the species assigned by the MALDI Biotyper database corresponds to the true species identity determined by whole genome sequence analysis and ANIm. For species with a log score ≥ 2.0 we furthermore evaluated, how many species were assigned a log score ≥2.0.

Similar to the MALDI Biotyper database, the VitekMS database assigns scores to each species classification. Furthermore, each species identification is assigned a *Confidence level* [%] and a *Type of identification*, which is either 'Single Choice' or 'Low Discrimination' and indicates whether the species identification was unambiguous or whether the database could not unambiguously discriminate between two or more species entries. Identifications with an assigned probability, lower than a probability threshold (60%) are not assigned a species label. In this situation, due to low confidence values, the *Type of identification* 'No Identification' is assigned. For spectra above the *Confidence level* threshold, we evaluated the *Type of Identification* and whether the assigned species corresponds to the true species identity of the measured strains.

We compared all spectra in our dataset to a ribosomal marker-based database (PAPMID^{TM}). Marker based species identification tools such as the PAPMID^{TM} database assign scores which correspond to the number of ribosomal marker peaks detected. The bacterial species is assigned for which most marker masses could be detected in a mass spectrum.

If a spectrum matches a maximal number of marker masses for multiple profiles of the same species, an unambiguous single species identification is assigned. If a spectrum matches an equal maximal number of marker masses from different species, multiple species are assigned ('multi-species Identification'). Species identifications with fewer marker peaks detected than the taxon-specific identification threshold, are assigned the label 'No identification possible'. The taxon specific thresholds used in this study were 20 for the species of the *Enterobacter cloacae complex*, 15 for *Klebsiella spp.* and *Escherichia coli / Shigella*, 7 for the species within the *S. aureus* complex, and 10 for all other phylogenetic groups included in this study.

Statistical analysis

We used paired wilcoxon rank tests when comparing spectra acquired from the same strains under different conditions. We excluded spectra of strains which were missing in one of the sets of interest. We used unpaired wilcoxon rank tests when comparing spectra acquired from different strains, e.g. when comparing between different phylogenetic groups or between correctly and incorrectly identified spectra.

When reporting comparisons in the running text, we refer to spectra acquired on the microflex Biotyper if not explicitly stated otherwise and use the nomenclature 'median (lower bound of the IQR , upper bound of the IQR)', throughout the study. Results and summary plots for spectra acquired on the Axima Confidence system can be found in the supplement.

We report the exact p-values when these are > 0.0001 and report use '****' for p-values < 0.0001. All analysis was performed in R (4.0.3) using the ggpubr (4.0) package.

Results

Defining MALDI-TOF mass spectral quality

In order to investigate spectral quality of the different datasets, we first assessed which spectral features are associated with a correct species identification with all databases and therefore suitable as quantitative measures for spectral quality. The spectra analysed here include a range of mass spectral quality, and were acquired using all different sample preparation protocols examined in this study and for the species included in three databases (MALDI Biotyper, VitekMS, PAPMID^{TM}) (**Table S1**).

Five spectral features are good proxies for the correct species identification. In correctly identified spectra (i.e. high spectral quality) over all phylogenetic groups we found an increase in the number of ribosomal marker peaks detected (median = 22 IQR = (18, 25) (same nomenclature used throughout the paper) vs. 13 (6, 20)), their median relative intensity (1.27 (1.02, 1.65) vs. 1.00 (0.77, 1.27)), the sum of the intensity of all detected peaks (1.69×10^6 mV (0.97×10^6 mV, 2.39×10^6 mV) vs. 0.90×10^6 mV (0.27×10^6 mV, 1.62×10^6 mV)) and a decrease in the measurement error (249 ppm (186 ppm, 338 ppm) vs. (289 ppm (213 ppm, 388 ppm)) (all p-values < 0.0001) when compared to incorrectly identified spectra (**Figure 1** and **Figure S1 - S2**). In order to account for reproducibility, we included the fraction of reproducibly detected peaks between technical replicates as fifth quality measure. These five features were henceforth used to evaluate the spectral quality in the dataset.

When comparing correctly to incorrectly identified spectra we observed, over all phylogenetic groups, a small increase in total number of peaks (173 (146, 203) vs. 163 (128, 206), p-value < 0.0001)). However, when comparing within each phylogenetic group, and especially for spectra acquired on the microflex Biotyper, we did not observe a beneficial effect of an

increased total number of peaks (**Figure S1**). Therefore, we did not include the number of peaks as a quality measure. The fraction of peaks > 10'000 Da (30.5 % (23.0 %, 38.2 %) vs. 31.6 % (18.1 %, 42.4 %), p-value = 0.91) and the m/z value at the 90th percentile (15,323 Da (13,304 Da, 16,128 Da) vs. 15,387 Da (11,394 Da, 16,264 Da), p-value = 0.07) were comparable between correctly and incorrectly identified spectra, respectively.

In the following, we evaluated which sample preparation yielded highest quality spectra, over all phylogenetic groups, for unknown samples and within each phylogenetic group separately.

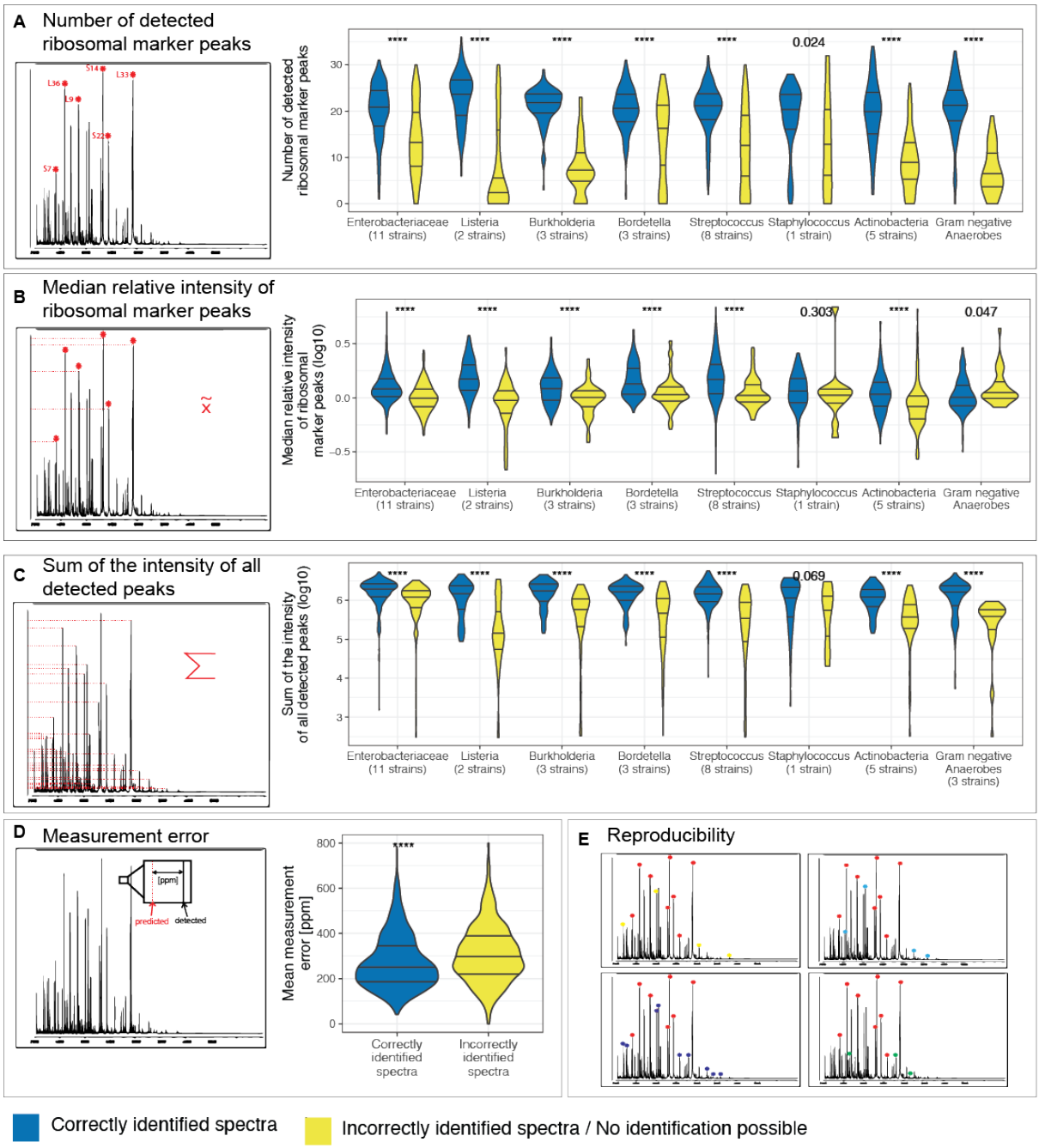

Figure 1: Spectra features compared between correctly spectra and incorrectly identified spectra per phylogenetic group. Species identification was performed by the MALDI Biotyper database and spectra were acquired on a microflex Biotyper. (A) the number of detected ribosomal marker peaks, (B) the median relative intensity of these, (C) the sum of intensity of all detected peaks, (D) measurement error and (E) reproducibility i.e. number of peaks detected in three out of four technical replicates divided by the total number of peaks in a spectrum.

Mass spectral quality improvement with different sample preparation methods

In order to identify the best sample preparation, we first tested three different protocols. Over all phylogenetic groups, we found the '25% FA overlay' method yielded the highest spectral quality.

We observed the median relative intensity of ribosomal marker peaks (1.49 (1.14, 1.91) vs. 1.27 (0.97, 1.73), p-value = 0.025),the sum of the intensity of all detected peaks (2.16×10^6 mV (1.53×10^6 mV, 2.73×10^6 mV) vs. 1.80×10^6 mV (1.22×10^6 mV, 2.36×10^6 mV)) and the fraction of reproducibly detected peaks (74.0% (66.0%, 80.1%) vs. 69.5%, (59.7%, 77.7%)) to be higher for spectra acquired under the '25% FA overlay' method compared to the 'smear' method (p-values < 0.0001). Furthermore, we observed less variation when comparing the number of ribosomal marker peaks detected (22 (19, 25) vs. 22 (16, 25)) for spectra acquired under the '25% FA overlay' method compared to the 'smear' method.

Spectra acquired with the 'simple protein extraction' method yielded overall lower values for these measures ('median relative intensity of the ribosomal marker peaks detected': 1.17 (1.03, 1.37); 'sum of the intensity of all detected peaks': 1.22×10^6 mV, (0.74×10^6 mV, 2.14×10^6 mV); 'number of ribosomal marker peaks detected': 19 (12, 23)) when compared to spectra acquired under the 'smear' method (p-values < 0.0001), except for the fraction of reproducibly detected peaks, where we observed higher values for spectra acquired under the 'simple protein extraction' (73.7%, (65.3%, 82.2)) when compared to spectra acquired under the 'smear' method (p-value < 0.0001). The accuracy of identification by PAPMID^{TM} generally follows quality measures, with the highest fraction of correctly identified spectra under the '25% FA overlay method' (**Figure 1, S3**).

Increased bacterial age decreased spectral quality

In order to assess how the age of a bacterial colony influences mass spectral quality, we measured the strains in our dataset after varying incubation time. We found a younger bacterial colony to be associated with a higher mass spectral quality. Increasing colony age had a negative impact on spectral quality with less ribosomal marker peaks detected (19.5 (17, 22) vs. 22 (20, 24)) and with a lower relative intensity (1.24 (0.96, 1.75) vs. 1.65 (1.34, 2.12)), and a lower fraction of reproducibly detected peaks (69.5% (64.8%, 74.6%) vs. 71.2% (64.6%, 77.6%)) after three days when compared one day incubation time (p-values < 0.0001) (**Figure 2**).

The accuracy of identification by PAPMID^{TM} generally follows quality measures, with an increasing number of spectra not being identified, and decreasing spectral quality over the time period.

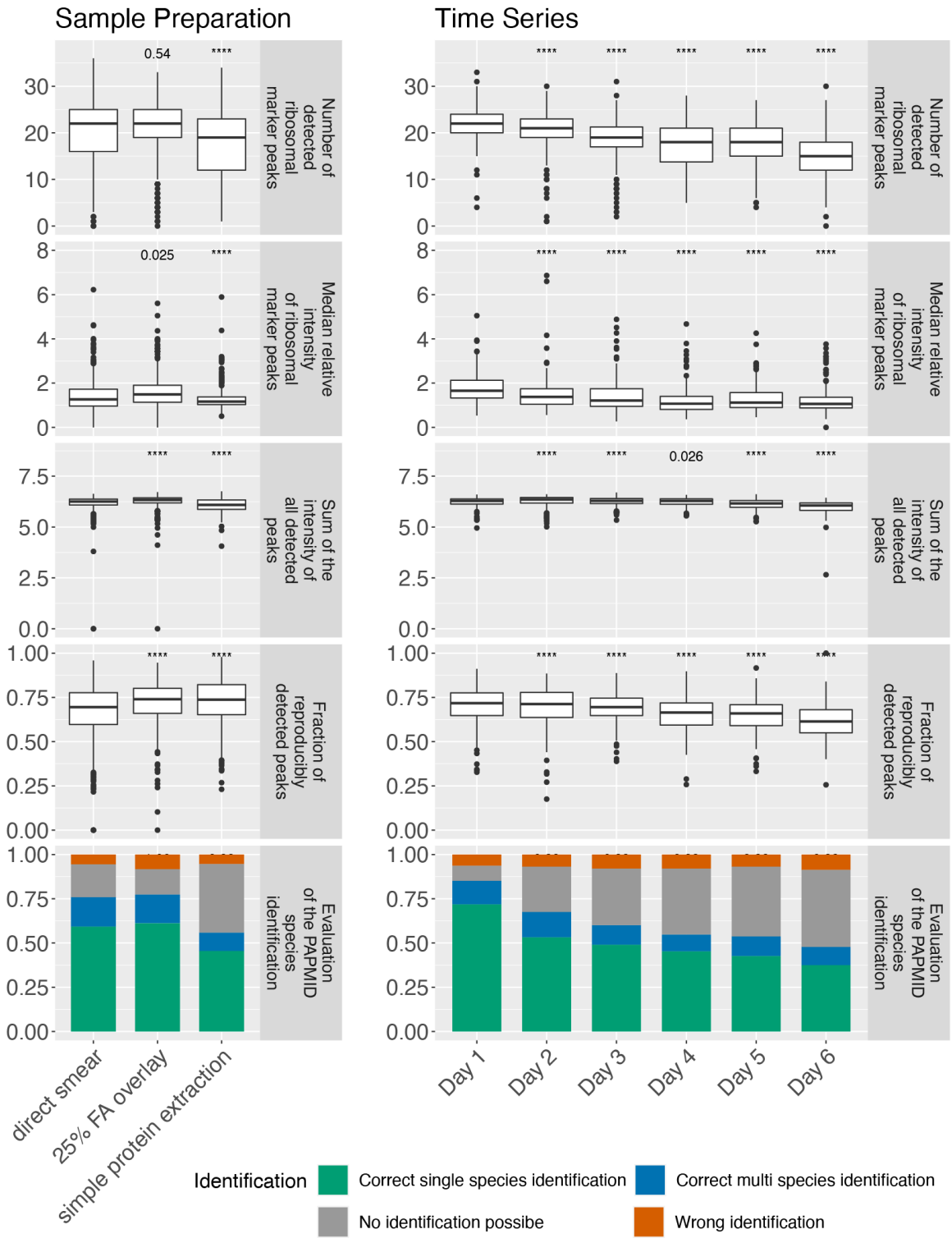

Figure 2: Comparison of different sample preparation protocols and colony ages (A) Comparison of different sample preparation protocols across all 47 bacterial isolates, (including 3 biological replicates and 4 technical replicates for each strain and protocol) and (B) age of the bacterial colony (47 bacterial isolates, 4 technical replicates for each strain and day) for spectra acquired on a microflex Biotyper MALDI-TOF MS system. '****': p-value < 0.0001.

The amount of bacterial colony material applied has a significant impact on spectral quality

In order to identify the best preparation procedure, we tested varying concentrations of the bacterial sample applied to the steel target plate.

Over all phylogenetic groups, we found that diluting the bacterial sample 1:5 did not decrease the number of ribosomal marker peaks detected (22.5 (14, 26) vs. 22 (19, 24)), nor the fraction of reproducibly detected peaks (75.2% (65.0%, 81.4%) vs. 72.4% (65.0%, 80.1%)) when compared to spectra acquired under the '25% FA overlay' method (Figure 3).

However, we observed a decreased median intensity of the ribosomal marker peaks (1.06, (0.94, 1.21) vs. 1.53 (1.16, 2.02)) and a decreased sum of the intensity of all detected peaks (1.17×106 mV (0.50×106 mV, 1.92×106 mV) vs. 2.15×106 mV (1.53×106 mV, 2.64×106 mV)) for 1:5 diluted samples when compared to samples processed using the '25% FA overlay' method (p-values < 0.0001).

Diluting bacterial colony material 1:25 or more generally decreased mass spectral quality ('number of ribosomal marker peaks detected': 21 (8.75, 24); 'median intensity of the ribosomal marker peaks: 1.07 (0.95, 1.29); 'sum of the intensity of all detected peaks': 0.51×106 mV, (0.16×106 mV, 0.85×106 mV)). However, we found taxon specific effects e.g. with Burkholderia yielding highest quality spectra with the highest number of ribosomal marker peaks detected with an additional dilution to 1:25 (**Figure S5** and **S6**).

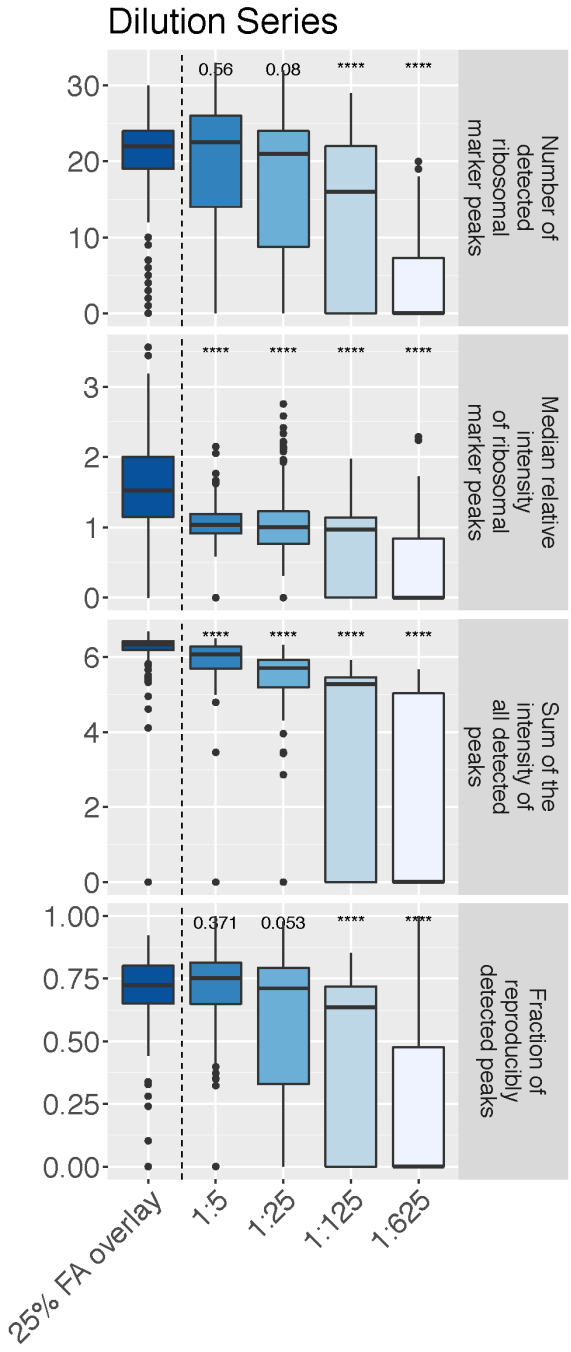

Figure 3: Comparison of the amount of bacterial colony material applied onto a steel target plate for spectral quality (16 bacterial isolates, 3 biological and 4 technical replicates per dilution and strain). Spectra were acquired on a microflex Biotyper MALDI-TOF MS system. '****': p-value < 0.0001.

Calibration is crucial

All MALDI-TOF MS were externally calibrated in a routine setting and the effect of calibration has been previously investigated (318). Here, we tested the impact of time between calibration and the measurement on measurement precision. We found that the measurement error

increased with time after calibration, (Day 1: 194 ppm (166 ppm, 235 ppm) vs. Day 7: 296 ppm (236 ppm, 379 ppm)) (p-value < 0.0001) (**Figure 3**).

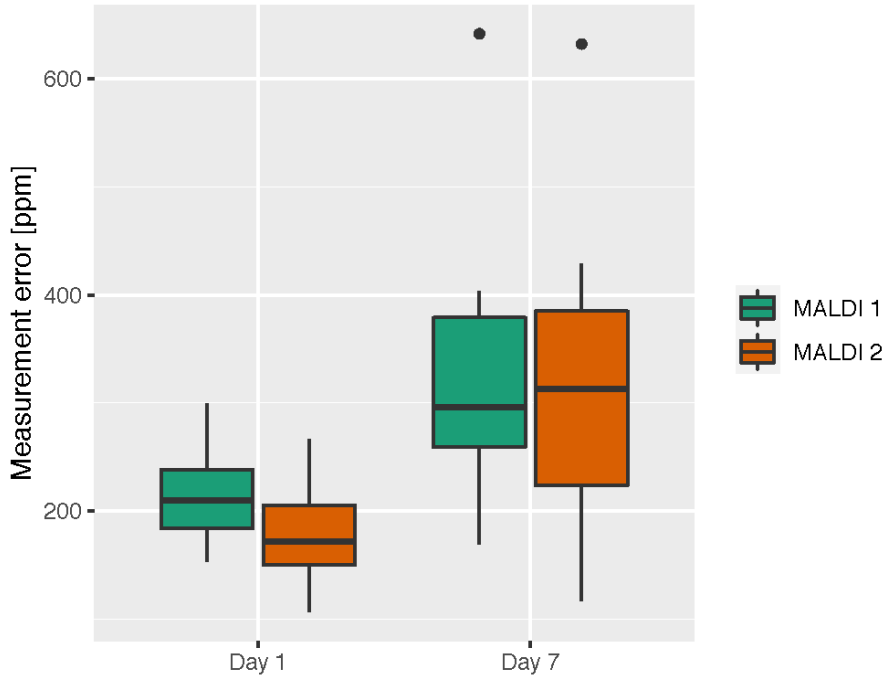

Figure 4: Measurement error for Day 1 and Day 7 after calibration (1 bacterial isolate, 3 biological and 4 technical replicates per day).

Major differences in spectra quality between bacterial taxa

Testing whether the mass spectral quality is sufficient for spectra acquired with the '25% FA overlay' method for all bacterial taxa, we found important differences (**Figure 5, Figure S7**). *Enterobacteriaceae* was the biggest family in our dataset (18 strains) and strains within this family generally yielded rich MALDI-TOF mass spectra with a high fraction of ribosomal marker peaks detected 57.1%, (50.0%, 61.9%)). For statistical analyses, we used *Enterobacteriaceae* as a reference group (**Figure 5, Figure S7**). On both MALDI-TOF MS systems, we found Gram positive bacteria generally yielded lower quality spectra than the Gram negative strains, with a lower fraction of ribosomal marker peaks detected (46.1% (34.7%, 53.6%) vs. 55.0% (50.0%, 61.9%), a lower sum of the intensity of all detected peaks (1.64×10^6 mV (1.11×10^6 mV, 2.39×10^6 mV) vs. 2.38×10^6 (1.91×10^6 mV, 3.05×10^6 mV)) and a lower fraction of reproducibly detected peaks (66.7% (60.2%, 73.9%) vs. 77.7% (72.0%, 82.8%)) (p-values < 0.0001). *Actinobacteria* and streptococci other than viridans streptococci yielded lowest quality MALDI-TOF mass spectra with the lowest fraction of ribosomal marker peaks detected (43.6% (19.8% - 57.1%), 43.2% (33.5% - 48.3%), respectively) and the lowest fraction of reproducibly detected peaks (66.1% (57.2% - 72.4%), 62.3% (58.2% - 69.5%), respectively).

Among the generally lower performing Gram positive bacteria and against the general trend, we detected the highest median fraction of detected ribosomal marker peaks for *Listeria* (58.0% (54.5% - 61.4%)), whereas Gram negative anaerobes yielded the highest fraction of reproducibly detected peaks (79.0%, (75.6% - 83.7%)).

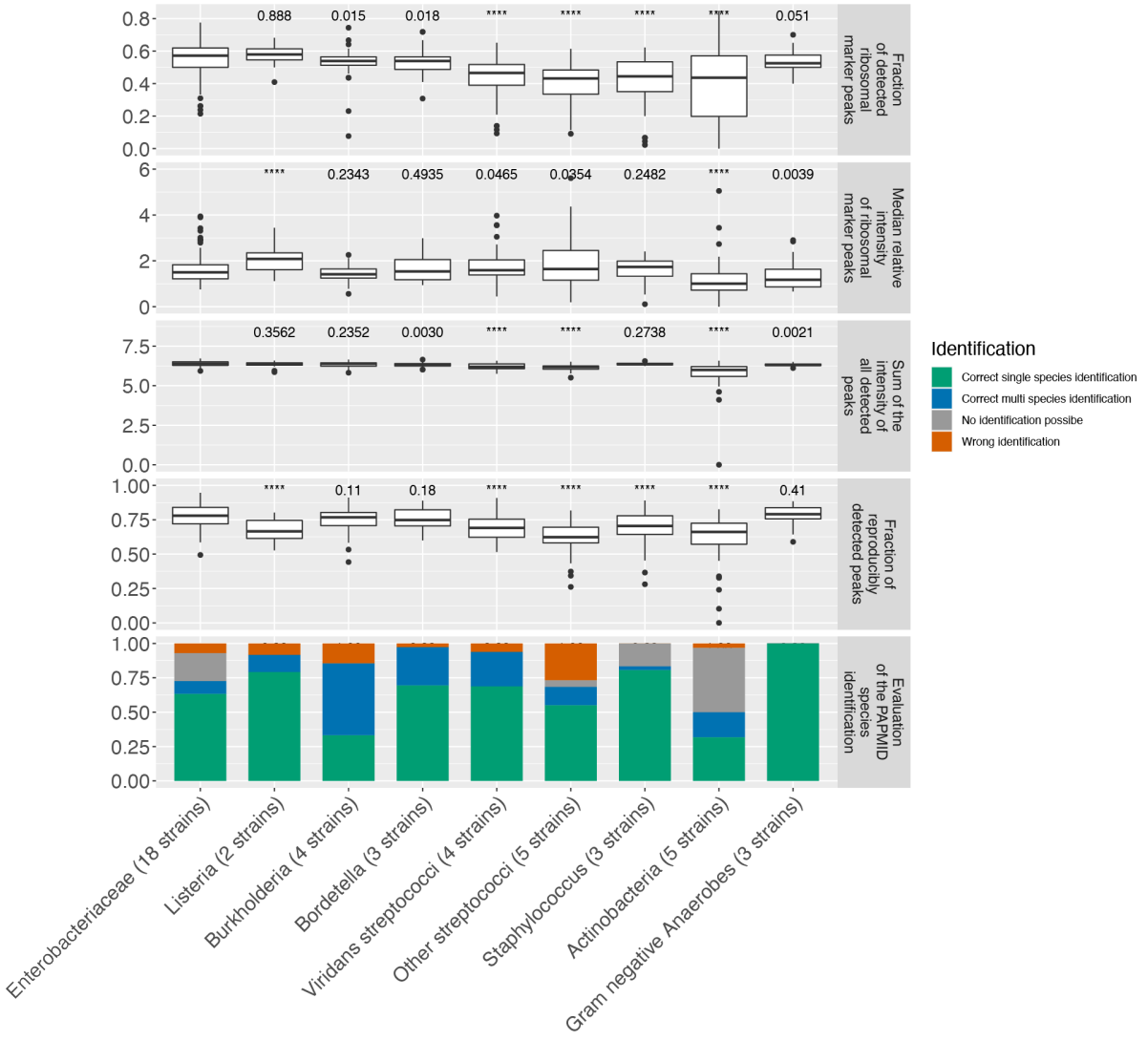

Figure 5: MALDI-TOF mass spectra features and species identification of spectra acquired with the '25% FA overlay' method and after one day of incubation on a microflex Biotyper (3 biological and 4 technical replicates per strain). '****': p-value < 0.0001.

Differences between MALDI-TOF MS databases

In order to evaluate different available databases, we compared spectra acquired on the microflex Biotyper to the MALDI Biotyper database (MALDI Biotyper Compass Library, Revision E (Vers. 8.0, 7854 MSP, RUO)) and spectra acquired on the Axima Confidence system to the VitekMS database (v3.2) for species identification. All spectra compared were

acquired under the '25% FA overlay' method. Please note that, while spectra were compared to the entire latter two databases for species identification, they were compared only to a subset of entries or subtyping modules of the PAPMID™ database.

Neither the MALDI Biotyper nor the VitekMS database cover all species included in this study (**Table S1**). Spectra of strains belonging to species missing in these databases are often wrongly identified as closely related species represented in the database (**Figure 6**). The MALDI Biotyper database covers more species represented in our strain collection than the VitekMS DB (**Table S1**) and more often results in a correct species assignment (**Figure 6**). However, comparison of spectra to MALDI Biotyper databases can lead to ambiguous results with multiple species yielding Scores > 2.0.

We observed the biggest difference between the MALDI Biotyper and the VitekMS database for *Staphylococcus* spectra, including spectra of the species *S. aureus*, *S. argenteus* and *S. schweitzeri*, with correctly identified species in 94.4% of spectra using the MALDI Biotyper database compared to 30.6% using the VitekMS database.

Species identification by the PAPMID™ yielded more often correct single species identification for spectra acquired on the Axima Confidence device than for spectra acquired on the microflex Biotyper device.

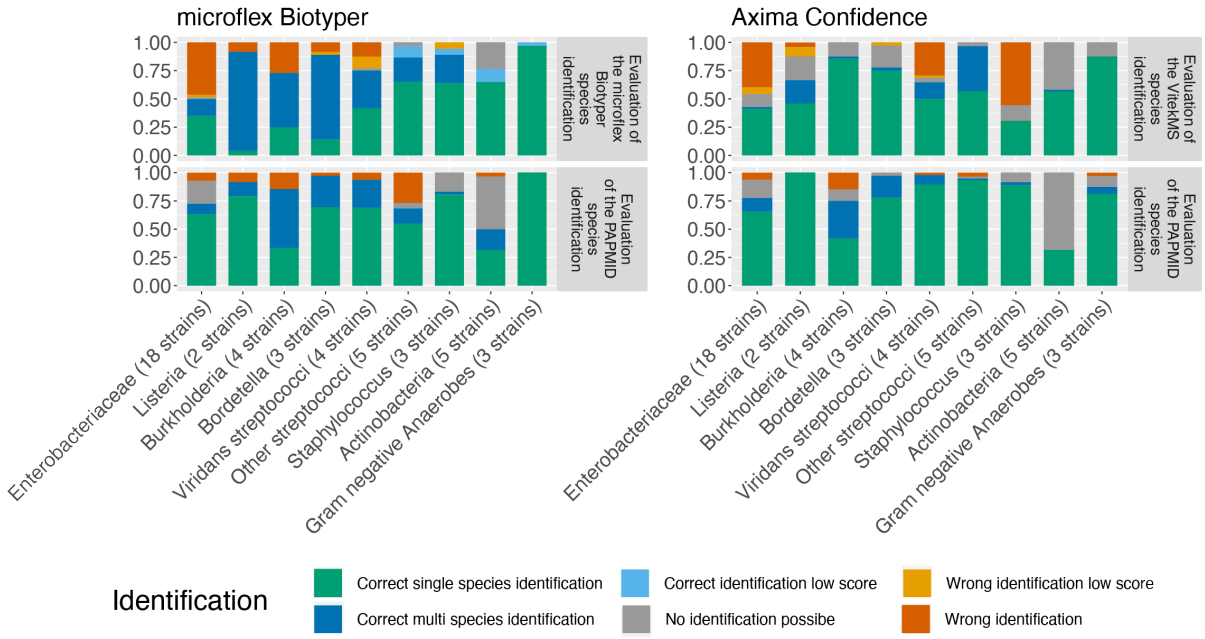

Figure 6: Evaluation of bacterial species identification. MALDI-TOF mass spectra were acquired on the microflex Biotyper (left) and the Axima Confidence (right) with the '25% FA overlay' method and after one day of incubation (3 biological and 4 technical replicates for each strain). Species identification was performed by the three different databases MALDI Biotyper (upper left), the VitekMS (upper right) and PAPMID™ (lower panels). Colour Code: green: correct single species identification: dark blue: correct identification, multiple species above threshold or VitekMS Identification Type 'Low Discriminatory'; light blue: correct identification, MALDI Biotyper log score < 2, grey: no identification possible, yellow: wrong species identified, low security (MALDI Biotyper log score < 2 or VitekMS Identification Type 'Low Discriminatory'), red: wrong species identified, high security (MALDI Biotyper log score > 2, VitekMS Identification Type 'Single Choice' and single species identification using a marker based approach).

Increased mass spectral quality increased species identification accuracy

In order to understand the impact of mass spectral quality on species identification accuracy in more detail we looked at three species complexes, namely the *Burkholderia cepacia complex,* viridans streptococci and the *Enterobacter cloacae complex*, and exclusively considering species which are covered by all three databases examined in this study. We split the spectra of these three complexes into three equal groups according to the sum of the intensity of all detected peaks in each spectrum, using this as a universally applicable proxy for spectral quality. We observed that with increasing intensity, the number of detected ribosomal marker peaks, their median relative intensity, the reproducibility and measurement precision increase, suggesting that these quality features are correlated.

Importantly, we found a larger fraction of correctly identified species with a higher confidence level (MALDI Biotyper log score) with increasing spectral quality (**Figure 7** and **Figure S19**). As an exception, we observed within the *Enterobacter cloacae complex*, an increase of

incorrectly identified spectra using the MALDI Biotyper database with an increasing spectral quality.

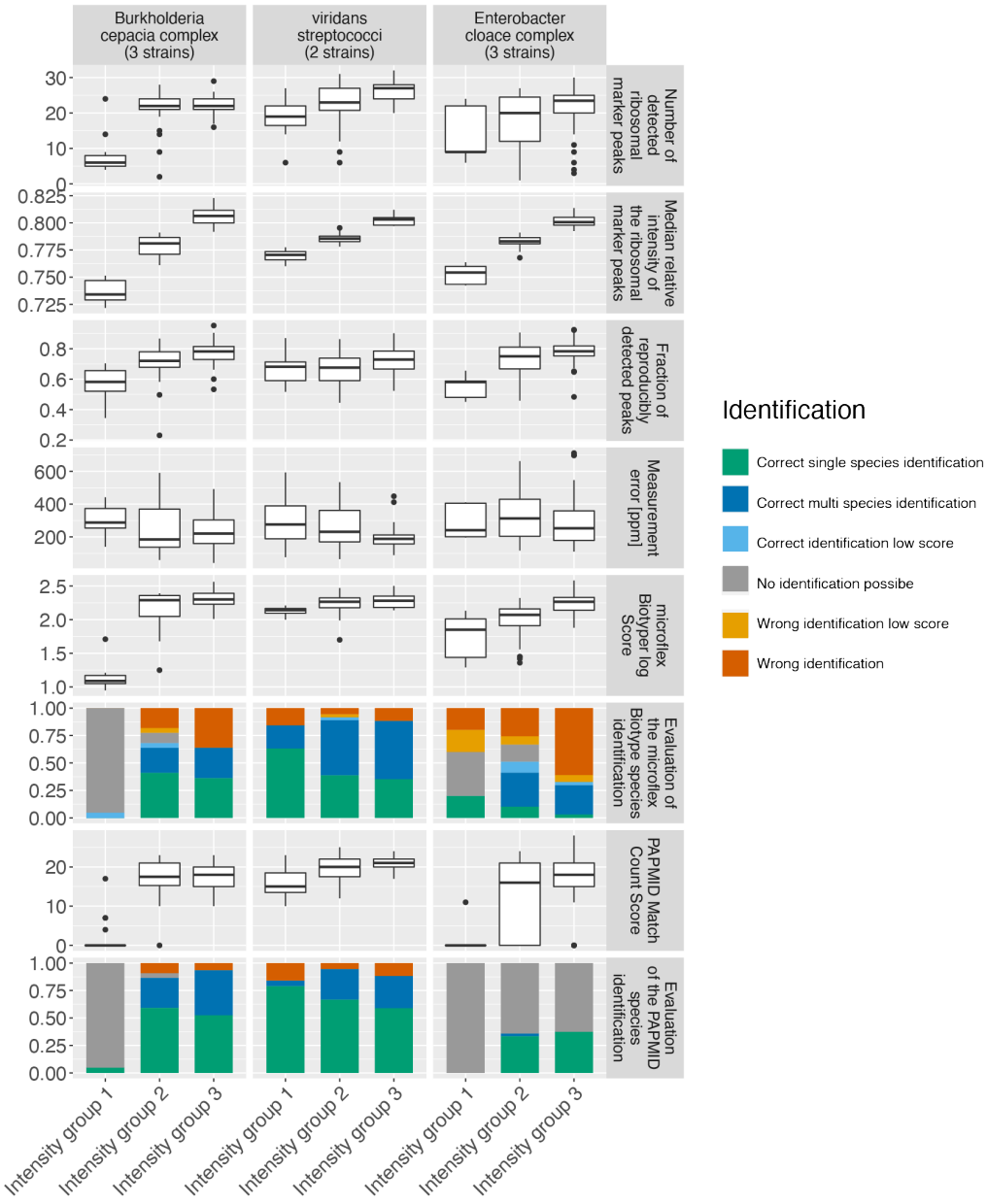

Figure 7: Spectra quality features and evaluation of species identifications grouped by the sum of the intensity of all detected peaks. Spectra acquired on the microflex LT/LH. Colour Code: green: correct single species identification; dark blue: correct identification, multiple species above threshold; light blue: correct identification, MALDI Biotyper log score < 2; grey: no identification possible, yellow: wrong species identified, low security (MALDI Biotyper log score < 2), red: wrong species identified, high security (MALDI Biotyper log score > 2 and single species identification using a marker-based approach).

Taxon-specific sample preparation for highest spectral quality

Following the inherent differences in mass spectral quality between the phylogenetic groups (**Figure 5**) we hypothesise taxon-specific improvement of spectral quality when using different sample preparation, quantity, and age of the bacterial colony. In order to assess these, we have compared the sample preparation conditions evaluated in this study, for each group separately (**Figure S5-S6, S8 - S16**). Here, we suggest optimized taxon-specific sample preparation and handling protocols in order to achieve optimal spectral quality. We summarised the optimal sample preparation and bacterial colony age per group which yielded good quality spectra (**Table 2**).

Table 2: Optimal sample protocol and bacterial age summarised per phylogenetic group

Group	Protocol	Day
Enterobacteriaceae	25% FA overlay	Day 1
Listeria	25% FA overlay / Dilute 1:5	Day 1/2/3/4
Burkholderia	Dilute 1:25	Day 2
Bordetella	25% FA overlay /Dilute 1:5	Day 1/2
Streptococcus	Simple protein extraction	Day 1/2/3/4/5
Staphylococcus	Dilute 1:5 / Simple protein extraction	Day 1/2
Actinobacteria	Simple protein extraction	Day 1
Gram - Anaerobes	Dilute 1:5	Day 1/2

Discussion

In this study we determined MALDI-TOF mass spectra quality features, associated with correct identification and showed that these features can be increased in routine diagnostics by adapting sample preparations protocols.

Comparing the spectra quality yielded by varying sample preparations we found that over all phylogenetic groups and for unknown samples, measuring bacterial samples at a young age and overlaying the sample with 25% formic acid yielded the best quality spectra. As *Enterobacteriaceae* was the biggest group in our dataset, it had the strongest influence on the optimal sample preparation protocols when we analysed all strains congruently. Nonetheless, also when analyzing the impact of different sample preparation protocols for each group

separately, the '25% FA overlay' method was amongst the best performing methods for most phylogenetic groups and with little hands-on time.

Over all phylogenetic groups, we observe the highest mass spectral quality after one overnight culture, followed by a decrease in mass spectral quality with increasing bacterial colony age. However, slower growing bacteria might require a longer incubation time before sufficient bacterial material can be transferred onto a target plate and before entering a phase of exponential growth, where ribosomal proteins are highly abundant (235,302).

We find that, over all phylogenetic groups, diluting the bacterial sample 1:5 does not decrease mass spectral quality and a dilution step can in fact increase the spectral quality for certain taxa.

Overlaying the bacterial colony material with 25% formic acid does not increase spectral quality for all phylogenetic groups and can in fact often be omitted, and these samples can be prepared using the 'direct smear' method when the taxon of an isolate is known. On the other hand, not all phylogenetic groups yielded good-quality spectra even when overlaying the sample with 25% formic acid, most notably *Actinobacteria*. Here, a 'simple protein extraction' might be required to detect intracellular proteins (312).

Summarizing our sample preparation experiments, we encourage laboratories working in routine diagnostics to measure unknown microorganisms after one night of growth, with little bacterial colony material, and overlaying each spot with 25% formic acid. If the spectra acquired using this protocol do not yield satisfying identification results, we furthermore propose the application of taxon-specific protocols. These can also be used to obtain optimal quality mass spectra for subtyping.

To define mass spectral quality, we analysed several spectral features among which we identified the following five as best proxies: (i) number of ribosomal marker peaks detected, (ii) median relative intensity of ribosomal marker peaks, (iii) sum of the intensity of all detected peaks, (iv) measurement precision, and (v) reproducibility of all peaks. The first four were increased in spectra which were correctly identified with all three databases when compared to incorrectly identified spectra. The effect of these features is more pronounced when spectra are acquired on the Axima confidence than on the Microflex Biotyper. Incorrectly identified Axima Confidence spectra appear to be signal poor with a low total number of peaks (**Figure S2**). Incorrectly identified microflex Biotyper spectra can harbour a high number of peaks, but are sparse in ribosomal marker masses and sum of the intensity of all detected peaks, which suggests that these spectra are noisy (**Figure S1**). This is also reflected in the higher number of false positive hits in ribosomal marker masses leading to a higher fraction of wrongly identified microflex Biotyper spectra than Axima Confidence spectra when compared to the PAPMID^{TM} database. As hardware settings, such as the tension of the detector, might affect the total number of peaks, it remains unclear whether the observed trends hold true for all

microflex Biotyper devices. A study involving multiple devices is required to assess this question.

When using the '25% FA overlay' method, we found a median of 76.0% of peaks reproducibly detected in technical replicates of the same sample (**Figure 2**). This measure assesses the reproducibility of picked peaks with which we decided to work with, as they are the bases for species identification. This measure of reproducibility is different from the Pearson correlation, comparing the shapes of two or more spectra (319,320). A reproducible detection of 75% of the picked peaks in a spectrum with 100 peaks, would mean that 75 peaks were detected in at least 3 out of 4 technical replicates of the same measurement. By using optimal sample preparation methods, we can increase the number of reproducibly detected peaks. These reproducibly detectable peaks could potentially be used as marker peaks, additional to ribosomal subunit masses and for spectra identification, further increasing the resolution of this method.

We observed the measurement error to increase with increasing time after calibration and therefore advise for frequent calibration of MALDI-TOF MS devices.

Microbiology taxonomy is in flux and many bacterial species have been newly described or have changed the genus in recent years (321). It is hardly possible for any diagnostic database to be up to date at every moment in time. We would like to emphasise that we have included strains in this study which pose difficulties for bacterial species identification and that bacterial species identification by MALDI-TOF MS is highly accurate in routine diagnostics (314). The challenges posed by the species included in this study are known and also clearly communicated by the MALDI-TOF MS manufacturers by e.g. displaying a warning message indicating which species cannot, or not reliably be distinguished from one another.

As the MALDI Biotyper database covers more of the species included in this study than the VitekMS database, spectral assignment from this database more often results in a correct species identification (**Figure 6**). This is most remarkable for spectra of the *S. aureus* complex, where the MALDI Biotyper database includes all three species (*S. aureus*, *S. argenteus* and *S. schweitzeri*), whereas the VitekMS database lists only *S. aureus* (**Table S1**). However, interpreting the MALDI Biotyper species identification is not always trivial as multiple species can yield a log score > 2, which is used as a threshold for the assignment 'highly confidence identification'.

Importantly, we have shown that an increased spectra quality can increase the accuracy of species identifications by all three databases. However, against the general trend, the number of incorrectly identified spectra increases with increasing spectra quality for species of the *Enterobacter cloacae complex* analysed with the MALDI Biotyper database. A possible explanation could be the MALDI Biotyper database frequently assigning the more frequent sister species *E. cloacae* sensu stricto.

We find *Actinobacteria* yielding the lowest spectra quality of all phylogenetic groups analysed in this study. When comparing spectra of this group to the PAPMID^{TM} database we find less often correctly identified spectra, compared to the other phylogenetic groups. For *Actinobacteria* only few ribosomal marker peaks can be detected, which makes distinction solely based on these, difficult. For this group, species identification using a pattern matching approach, applied by the MALDI Biotyper and the VitekMS database, more often yielded correct results. As it remains unclear which proteins form the basis of this species identification and how these vary between closely related species, it is possible that discrimination between closely related species might be challenging within *Actinobacteria* using a pattern matching approach.

MALDI-TOF mass spectra quality might be influenced by factors not considered in this study including: (i) hardware factors such as the age and intensity of the laser; (ii) the type of MALDI-TOF MS target plates and matrix used; (iii) culturing variables such as the agar media used or the atmosphere in which bacterial isolates are grown; (iv) spectra acquisition settings such as the number of laser shots applied and spectra averaged per measurement and (v) factors considering technical knowledge on acquiring MALDI-TOF mass spectra including regular training of staff and quality control of MALDI-TOF MS measurements. In order to assess and standardise MALDI-TOF mass spectral quality in routine diagnostics, a broader study comparing spectra acquired in multiple laboratories by different personnel is required.

The reliable detection of marker peaks in clinical routine would allow for higher resolution typing based on MALDI-TOF mass spectra, also distinguishing between closely related species e.g. within the *Klebsiella pneumoniae complex*, the *Staphylococcus aureus complex* and within viridans streptococci. An effective standardisation in culture conditions and spectra quality assessment might help the automation process of colony picking and mass spectral acquisition. Using a marker-based approach for identification, we can congruently query spectra acquired on different MALDI-TOF MS systems around the world. Using the potential of routinely generated MALDI-TOF MS data for sub lineage detection would open up new avenues of disease control by tracing the spread of important sub-lineages in real time with little additional effort.

Acknowledgements

We thank Roxanne Mouchet (Mabritec AG, Riehen Switzerland), Doris Hohler (University Hospital of Basel) for excellent technical assistance in acquiring the MALDI-TOF mass spectra and culturing the strains and species identification using a marker-based approach. We want to thank Dr. Tim Roloff, Magdalena Schneider, Christine Kiessling, Elisabeth Schultheiss, Rosa-Maria Vesco, and Clarisse Straub (all University Hospital of Basel) for the DNA

extraction, library preparations and sequencing of the bacterial isolates. We would like to thank Dr. Florian Geier and Dr. Robert Ivanek (Department for Biomedicine, University of Basel) for consultancy considering the statistical analyses. Further, we would like to thank Dr. Martin Welker (bioMérieux, Marcy-l'Étoile, France) for providing us access to the VitekMS database and technical consultancy as well as Ilona Mossbrugger (Bruker Daltonics, Bremen, Germany) for technical consultancy. We thank Dr. Vladimira Hinic, Dr. Helena Seth-Smith, Dr. Kirstine Kobberøe Søgaard (all University Hospital of Basel), Dr. Samuel Lüdin (Mabritec AG) and Vincent Somerville (University of Lausanne) for carefully reading the manuscript and giving valuable feedback. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel, the support from the sciCORE team for the analysis is greatly appreciated.

Chapter IV: Quality of MALDI-TOF Mass Spectra in Routine Diagnostics: Results from an International External Quality Assessment including 36 Laboratories from 12 countries

Aline Cuénod[a,b], Martina Aerni[c], Claudia Bagutti[d], Banu Bayraktar[e], Cynthia Beisert Carneiro[f], Carlo Casanova[g], Alix T. Coste[h], Gilbert Greub[h], Peter Damborg[i], Dick van Dam[j], Mehmet Demirci[k, am], Jaroslav Hrabak[l], Pavel Drevinek[m, am], Olivier Dubuis[s], José Fernandez[ag], Gülen Hürkal Yiğitler[e], Jakub Hurych[m], Thøger Gorm Jensen[n], Géraldine Jost[o], Greetje A. Kampinga[p], Sonja Kittl[q], Christine Lammens[r], Claudia Lang[s], Reto Lienhard[t], Julie Logan[u], Carola Maffioli[v], Ivana Marekovi[w,am], Matthias Marschal[x], Jacob Moran-Gilad[y], Shani Troib[y], Oliver Nolte[z], Michael Oberle[aa], Michael Pedersen[ab], Valentin Pflüger[ac], Sigrid Pranghofer[ad], Julia Reichl[ae], Rob Rentenaar[af], Arnaud Riat[ag], Belén Rodríguez-Sánchez[ah], Elise Willems[ai], Mandy Wootton[aj], Dominik Ziegler[ak], Efe Serkan Boz[al], Adrian Egli[a, b, am]

[a] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland, [b] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland, [c] Labor Team W, Goldach, Switzerland, [d] State Laboratory Basel-Stadt, Basel, Switzerland, [e] University of Health Sciences, Sisli Hamidiye Etfal Teaching and Research Hospital, Istanbul, Turkey, [f] University Hospital Freiburg, Freiburg im Breisgau, Germany, [g] Institute for Infectious Diseases, University of Bern, Bern, Switzerland, [h] Centre hospitalier universitaire vaudois, Lausanne, Switzerland, [i] University of Copenhagen, Department of Veterinary and Animal Sciences, Copenhagen, Denmark, [j] Arts-microbioloog Zuyderland MC, Sittard, the Netherlands, [k] Kirklareli University, Faculty of Medicine, Department of Medical Microbiology, Kirklareli, Turkey, [l] Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Plzen, Czech Republic, [m] Department of Medical Microbiology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic, [n] Department of Clinical Microbiology, Odense University Hospital, Odense, Denmark, [o] Dianalabs, Geneva, Switzerland, [p] Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands, [q] Institute of Veterinary Bacteriology, University of Berne, Berne, Switzerland, [r] Medical Microbiology, University of Antwerp, Belgium, [s] Viollier AG, Allschwil, Switzerland, [t] ADMED Microbiologie, La Chaux de Fonds, Switzerland, [u] Reference Services Division, UK Health Security Agency, London, United Kingdom, [v] MCL laboratories, Niederwangen, Switzerland, [w] University Hospital Centre Zagreb, Zagreb, Croatia, [x] Institute of Medical Microbiology and Hygiene, University of Tübingen, Tübingen, Germany, [y] School of Public Health, Ben Gurion University of the Negev and Soroka University Medical Center, Beer Sheva, Israel, [z] Center

for Laboratory Medicine, St. Gallen, Switzerland, [aa] Cantonal Hospital Aarau, Aarau, Switzerland, [ab] Department of Clinical Microbiology, Hvidovre Hospital, Hvidovre, Denmark, [ac] Mabritec AG, Riehen, Switzerland, [ad] Bioanalytica AG, Lucerne, Switzerland, [ae] Austrian Agency for Health and Food Safety, Vienna, Austria, [af] UMC Utrecht, Utrecht, The Netherlands, [ag] Division of laboratory medicine, Laboratory of bacteriology, University Hospital of Geneva, Geneva, Switzerland, [ah] Hospital General Universitario Gregorio Maranon, Madrid, Spain, [ai] Clinical Laboratory AZNikolaas, Sint-Niklaas, Belgium, [aj] University Hospital of Wales, Cardiff, United Kingdom , [ak] Eurofins Scientific AG, Schönenwerd, Switzerland, [al] University of Health Sciences, Haydarpasa Numune Teaching and Research Hospital, Department of Medical Microbiology, Istanbul, Turkey, [am] European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group for Genomic and Molecular Diagnostics

Manuscript in preparation

My contributions:

- Conceptualisation of the study, shipping bacterial strains
- MALDI-TOF mass spectral quality assessment
- Write feedback reports for each laboratory
- Compile instructions for the second measurements
- Data visualisation (all figures)
- Writing of the original manuscript

I presented the first part of the EQA at the following international conferences:

- iPoster presentation at the *worldmicrobeforum* (joint conference by the *American Society for Microbiology* and the *Federation of European Microbiological Societies*) 2021
- 'Mini oral flash' (3 min presentation) at the European Congress of Clinical Microbiology and Infectious Diseases (ECCMID)
- Oral presentation at the Annual Congress of the Swiss Society for Microbiology 2021

All three events were held online

The complete EQA is *in review* for presentation at following conference:

- European Congress of Clinical Microbiology and Infectious Diseases (ECCMID) 2022, Lisbon (Portugal)

Note: The following part contains the full manuscript
The Supplementary Figure can be found in Appendix II of this thesis.
The Supplementary Tables can be accessed via the following link:
https://drive.google.com/drive/folders/1H7nZMuJNzj3_kp6E5vwIUhc1RCu34NFE?usp=sharing (This link will be deactivated upon publication of the data)

Quality of MALDI-TOF Mass Spectra in Routine Diagnostics: Results from an International External Quality Assessment including 36 Laboratories from 12 countries

Aline Cuénod[a,b], Martina Aerni[c], Claudia Bagutti[d], Banu Bayraktar[e], Cynthia Beisert Carneiro[f], Carlo Casanova[g], Alix T. Coste[h], Gilbert Greub[h], Peter Damborg[i], Dick van Dam[j], Mehmet Demirci[k, am], Jaroslav Hrabak[l], Pavel Drevinek[m, am], Olivier Dubuis[s], José Fernandez[ag], Gülen Hürkal Yiğitler[e], Jakub Hurych[m], Thøger Gorm Jensen[n], Géraldine Jost[o], Greetje A. Kampinga[p], Sonja Kittl[q], Christine Lammens[r], Claudia Lang[s], Reto Lienhard[t], Julie Logan[u], Carola Maffioli[v], Ivana Marekovic[w,am], Matthias Marschal[x], Jacob Moran-Gilad[y], Shani Troib[y], Oliver Nolte[z], Michael Oberle[aa], Michael Pedersen[ab], Valentin Pflüger[ac], Sigrid Pranghofer[ad], Julia Reichl[ae], Rob Rentenaar[af], Arnaud Riat[ag], Belén Rodríguez-Sánchez[ah], Jacques Schrenzel[ag], Elise Willems[ai], Mandy Wootton[aj], Dominik Ziegler[ak], Efe Serkan Boz[al], Adrian Egli[a, b, am]

[a] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland

[b] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland

[c] Labor Team W, Goldach, Switzerland

[d] State Laboratory Basel-Stadt, Basel, Switzerland

[e] University of Health Sciences, Sisli Hamidiye Etfal Teaching and Research Hospital, Istanbul, Turkey

[f] University Hospital Freiburg, Freiburg im Breisgau, Germany

[g] Institute for Infectious Diseases, University of Bern, Bern, Switzerland

[h] Centre hospitalier universitaire vaudois, Lausanne, Switzerland

[i] University of Copenhagen, Department of Veterinary and Animal Sciences, Copenhagen, Denmark

[j] Arts-microbioloog Zuyderland MC, Sittard, the Netherlands

[k] Kirklareli University, Faculty of Medicine, Department of Medical Microbiology, Kirklareli, Turkey

[l] Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Plzen, Czech Republic

[m] Department of Medical Microbiology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic

[n] Department of Clinical Microbiology, Odense University Hospital, Odense, Denmark

[o] Dianalabs, Geneva, Switzerland

[p] Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[q] Institute of Veterinary Bacteriology, University of Berne, Berne, Switzerland

[r] Medical Microbiology, University of Antwerp, Belgium

[s] Viollier AG, Allschwil, Switzerland

[t] ADMED Microbiologie, La Chaux de Fonds, Switzerland

[u] Reference Services Division, UK Health Security Agency, London, United Kingdom

[v] MCL laboratories, Niederwangen, Switzerland

[w] University Hospital Centre Zagreb, Zagreb, Croatia

[x] Institute of Medical Microbiology and Hygiene, University of Tübingen, Tübingen, Germany

[y] School of Public Health, Ben Gurion University of the Negev and Soroka University Medical Center, Beer Sheva, Israel

[z] Center for Laboratory Medicine, St. Gallen, Switzerland

[aa] Cantonal Hospital Aarau, Aarau, Switzerland

[ab] Department of Clinical Microbiology, Hvidovre Hospital, Hvidovre, Denmark

[ac] Mabritec AG, Riehen, Switzerland

[ad] Bioanalytica AG, Lucerne, Switzerland

[ae] Austrian Agency for Health and Food Safety, Vienna, Austria

[af] UMC Utrecht, Utrecht, The Netherlands

[ag] Division of laboratory medicine, Laboratory of bacteriology, University Hospital of Geneva, Geneva, Switzerland

[ah] Hospital General Universitario Gregorio Maranon, Madrid, Spain

[ai] Clinical Laboratory AZNikolaas, Sint-Niklaas, Belgium

[aj] University Hospital of Wales, Cardiff, United Kingdom

[ak] Eurofins Scientific AG, Schönenwerd, Switzerland

[al] University of Health Sciences, Haydarpasa Numune Teaching and Research Hospital, Department of Medical Microbiology, Istanbul, Turkey

[am] European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group for Genomic and Molecular Diagnostics

Keywords

MALDI-TOF MS, Quality Control, Bacterial Species Identification, External Quality Assessment, Standardisation, Diagnostic Performance

Abstract

Objective: MALDI-TOF MS is a widely used method for bacterial species identification. Incomplete databases and mass spectral quality (MSQ) still represent major challenges. Important proxies for MSQ are: number of detected marker masses, reproducibility, and measurement precision. We aimed to assess MSQs across diagnostic laboratories and the potential of simple workflow adaptations to improve it.

Methods: For baseline MSQ assessment, 47 diverse bacterial strains were routinely measured in 36 laboratories from 12 countries, and well-defined MSQ features were used. After an intervention consisting of detailed reported feedback and instructions on how to acquire MALDI-TOF mass spectra, measurements were repeated and MSQs were compared.

Results: At baseline**,** we observed heterogeneous MSQ between the devices, considering the median number of marker masses detected (range = [5, 25]), reproducibility between technical replicates (range = [55%, 86%]), and measurement error (range = [147ppm, 588ppm]). As a general trend, the spectral quality was improved after the intervention for devices which yielded low MSQs in the baseline assessment: for 4/5 devices with a high measurement error, the measurement precision was improved (p-values < 0.001, paired Wilcoxon test); for 6/10 devices, which detected a low number of marker masses, the number of detected marker masses increased (p-values < 0.001, paired Wilcoxon test).

Conclusion: We have identified simple workflow adaptations, which, to some extent, improve MSQ of poorly performing devices and should be considered by laboratories yielding a low MSQ. Improving MALDI-TOF MSQ in routine diagnostics is essential for increasing the resolution of bacterial identification by MALDI-TOF MS, which is dependent on the reproducible detection of marker masses. The heterogeneity identified in this EQA requires further study.

Introduction

MALDI-TOF MS is the most commonly used method for microbial species identification in modern diagnostic laboratories around the world (63) due to its minimal hands-on, short turn-around time for measurements and data analysis, cost-efficiency and high accuracy (322,323). Multiple studies have shown the improved resolution gained by using marker-based analytical approaches (76,83,85,233) compared to pattern matching approaches. This insight has led to the development of marker-based databases for bacterial identification such as the Biotyper subtyping module (Bruker Daltonics, Bremen, Germany), PAPMID^{TM} (Mabritec AG, Riehen, Switzerland) (305), MALDITypeR (233), and Ribopeaks (324). In such approaches, specific peaks of interest, whose presence is associated with a species (325), lineage (308), or even mobile genetic elements (92,326), are queried in the acquired mass spectrum in order to increase specificity and resolution. Many of the peaks, which can be reproducibly detected in MALDI-TOF mass spectra, correspond to protein subunits of the bacterial ribosome (234). The improved resolution of a marker-based identification approach requires a high mass spectral quality (MSQ) in order to reproducibly detect marker peaks. It can be difficult to assess which marker peaks are reproducibly detected in MALDI-TOF mass spectra, as public datasets with matching genomic sequences and technical replicates of isolates from multiple devices are rare.

Despite the success of MALDI-TOF MS for routine microbial species identification, multiple clinically-relevant species cannot be distinguished using pattern matching databases. Possible reasons for this are that (i) the databases are incomplete, (ii) the species of interest resemble closely other species in the databases, and (iii) MALDI-TOF mass spectra are of low quality. We previously compiled a diverse set of 47 bacterial strains, representing 39 species and 15 genera, which are difficult to be identified at a species level for the above-mentioned reasons (262). In this previous publication (262), we defined the following five important spectrum features as good proxies for MSQ: (i) the number of ribosomal marker peaks detected, (ii) the median relative intensity of ribosomal marker peaks, (iii) the sum of the intensity of all detected peaks, (iv) a high measurement precision, and (v) reproducibility of peaks between technical replicates. Applying these MSQ features, we previously assessed

the performance of three different sample preparation protocols (*direct smear*, *formic acid overlay*, and *simple protein extraction*) on different bacterial groups and consequently proposed to use the *formic acid overlay protocol* for unknown samples and *group specific protocols* (i.e. one of the above, depending on the bacterial group measured) for highest MSQ (262). Whether the proposed protocols can effectively increase MSQ in routine settings, when employed by different personnel and on different MALDI-TOF devices has yet to be evaluated. The aim of this study was therefore to (i) assess the MSQ obtained in routine diagnostics, (ii) assess whether there are routine practices associated with an increased MSQ, (iii) assess whether the MSQ can be improved using the protocols proposed, and (iv) compile a reference dataset of MALDI-TOF mass spectra including technical replicates, matching genomic sequences and extensive metadata.

Methods

Design of the External Quality Assessment

Figure 1 provides an overview over the workflow of this study.

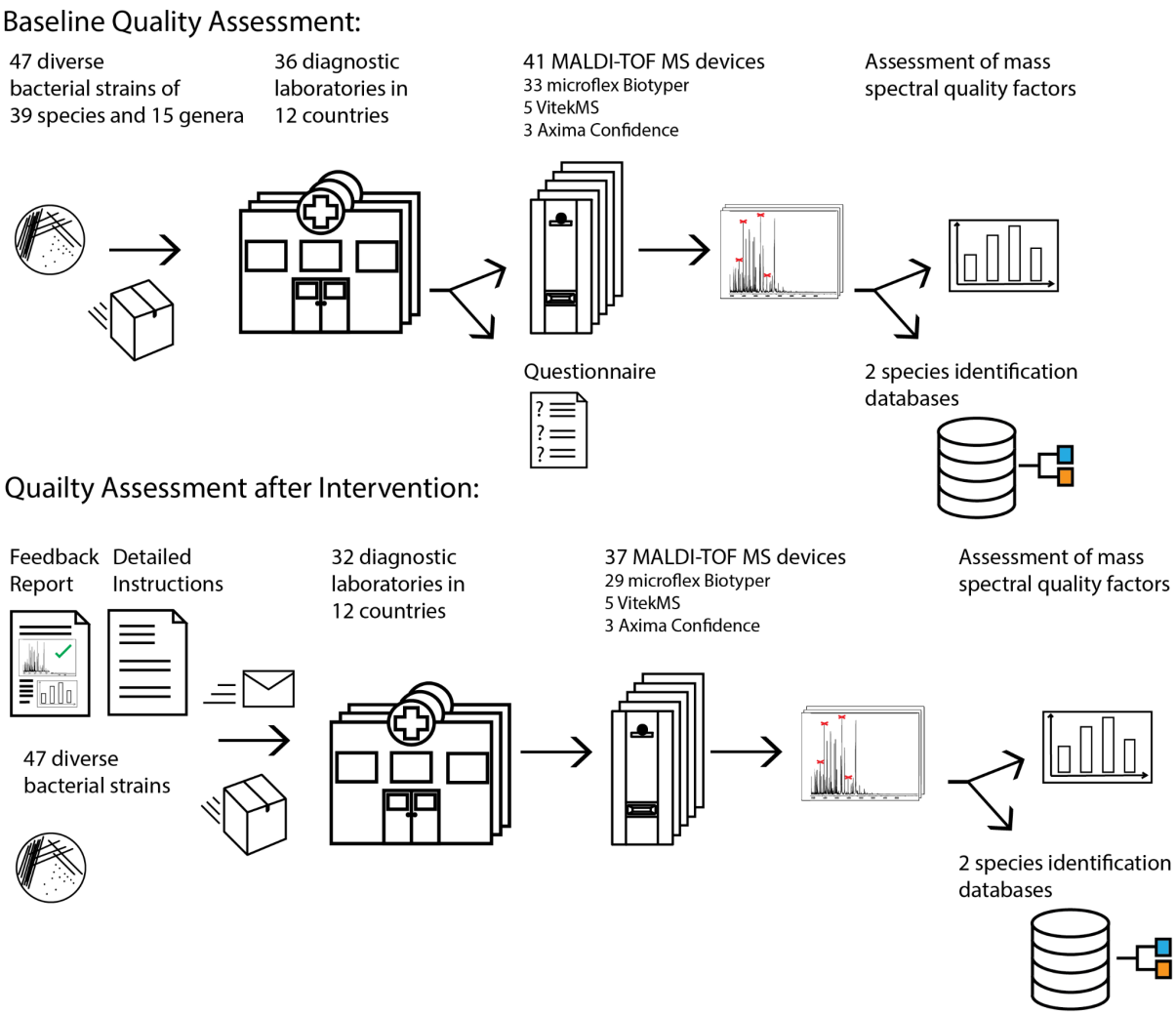

Figure 1: Overview on the workflow of the study. The upper panel shows the baseline quality assessment including 36 participating laboratories and the lower panel shows the post-interventional quality assessment including 32 laboratories using the same bacterial strains.

Bacterial strains

The bacterial strains used in this study have previously been described (262) and whole genome sequences are publicly available (**Table S1**). The masses of the ribosomal subunits have previously been predicted and these, as well as previously acquired MALDI-TOF mass spectra are publicly available, (https://osf.io/ksz7r/).

This set of strains comprises 47 clinically-relevant bacterial isolates from public and in-house strain collections. The included 39 species can be challenging to identify using MALDI-TOF MS, either because intracellular proteins cannot be ionised easily due to cell wall composition

(e.g. *Corynebacterium spp.*), or because of their close relatedness to another bacterial species (e.g. *Klebsiella oxytoca* / *Klebsiella michiganensis*; *Shigella* / *Escherichia coli*). The bacterial isolates were assigned to eight groups, according to rough phylogenetic groupings (**Table S1**). For the strains in each group, we expected comparable spectral features. For the evaluation of species identification, the group '*Streptococcus*' was further split up into 'viridans streptococci' and 'other streptococci', as the former group is of special interest in clinical diagnostics.

Participating laboratories

The baseline quality assessment encompassed 36 laboratories in 12 countries. Five laboratories participated with two MALDI-TOF MS devices, resulting in 41 measurement datasets. Of these, 32 laboratories with 37 devices also acquired MALDI-TOF mass spectra post-intervention, and of these, 28 laboratories with 33 devices acquired mass spectra using the workflow adaptation, consisting of the *formic acid protocol* and the *group specific protocols* (**Figure S1, Table S2**).

Shipment and Culturing of the Bacterial Strains

All strains were cultured on 5% Columbia Sheep Blood Agar plates before shipment in LBM eSwab transport medium (Copan, Brescia, Italy). Strains #33, #34, #45 and #46 require anaerobic conditions to grow, which was indicated to the participating laboratories.

Baseline MALDI-TOF MSQ Assessment

The participating laboratories were asked to culture the bacterial isolates and acquire MALDI-TOF mass spectra according to their routine diagnostic procedures, which may vary between the laboratories. Furthermore, each laboratory was asked to fill out a questionnaire on routine laboratory practice by technical personnel handling the MALDI-TOF MS devices.

Intervention

Each participating laboratory received a detailed feedback report on the MALDI-TOF mass spectra acquired for the baseline quality assessment (example in **Suppl. File 1**.) and instructions on how to acquire MALDI-TOF mass spectra in subsequent measurements of the same strains, aiming to improve the MSQ using a standardised approach (**Suppl. File 2**).
We provided two different sets of protocols: (i) a simple 'generic protocol' for all samples and (ii) group-specific sample preparation protocols, aiming at highest MSQ (262). We proposed the '*formic acid overlay*' protocol as 'Generic protocol for all samples'. Here, bacterial material was transferred onto the target plate using a wooden toothpick or a plastic inoculation needle.

The sample was subsequently overlaid with 1µl formic acid (70% for all laboratories working with a microflex Biotyper system and 25% for all others) and air dried before overlaying with 1µl a-Cyano-4-hydroxycinnamic acid (CHCA) matrix solution. Measurements were performed after 12-18h incubation. Details on the protocols specific for the different bacterial groups can be found in the instructions document (**Suppl. File 2**).

MALDI-TOF MS Spectra Processing

Peaks were picked from raw spectra using default settings by the softwares included in the microflex Biotyper or the VitekMS / Axima Confidence system. Spectra acquired on microflex Biotyper devices were exported as 'fid' files and peak picking was performed in the flexAnalyses software (v3.4) and exported as text files. Spectra acquired on VitekMS / Axima Confidence devices were exported as '.mzXml' (device ID 35, 38), as '.mzml' (device ID 11, 12, 15, 30) or as text files (device ID 09, 43). These already contain exclusively m/z values and intensities of picked peaks, and therefore 'mzml' or 'mzXml' files were converted to text files. We subsequently worked exclusively with the intensity and m/z value of these picked peaks, and did not consider further peak characteristics such as the resolution or the signal to noise ratio of a peak.

We excluded empty spectra (0 peaks assigned using the above described peak picking methods) from further analyses. Further, we excluded spectra considered as contaminations, i.e. for which the identified genus did not match the genus identified from whole genome sequence data with the exception of *E. coli* / *Shigella* and *Raoultella / Klebsiella,* which were regarded as the same genera, respectively, in this study (120,158).

MALDI-TOF MSQ Features

We queried each spectrum for the following features to assess the MSQ: (i) the number of ribosomal marker peaks detected, (ii) the median relative intensity of ribosomal marker peaks, (iii) the sum of the intensity of all detected peaks, (iv) a high measurement precision, and (v) reproducibility of peaks between technical replicates (see **Supplementary Methods** for more detail). As factors (i) - (iii) often correlate (262), we have focused on factors (i), (iv) and (v) in the main text and figures of this study.

Scripts used for spectra evaluation and data visualisation can be accessed via GitHub (https://github.com/acuenod111/MALDI-TOF-MS-EQA).

Databases used for species identification

Each spectrum acquired on a Bruker device was compared to the MALDI Biotyper database (MALDI Biotyper Compass Library, Revision E (Vers. 8.0, 8468 MSP, RUO, Bruker Daltonics,

Bremen, Germany). Spectra acquired on the Axima Confidence device were analysed with the VitekMS database (bioMérieux, Marcy-l'Étoile, France) (v3.2). Furthermore, we compared each spectrum to a ribosomal marker-based database, either PAPMID™ or PAPMID™ subtyping modules, (both Mabritec AG, Riehen, Switzerland, see **Supplementary Methods** for more detail). Both are based on the detection of ribosomal marker peaks and will henceforward be referred to as PAPMID™.

More details about database scores and their interpretations can be found in the **Supplementary Methods**.

Statistical Analysis

We used paired Wilcoxon rank tests when comparing spectra acquired from the same strains and excluded spectra of strains, which were missing in one of the sets of interest. We used unpaired Wilcoxon rank tests (Mann Whitney U tests) when comparing spectra acquired from different strains. The nomenclature 'median (lower bound of the IQR, upper bound of the IQR)' was used when referring to data in the running text throughout the study. All analyses were performed in R (v4.0.3) using the ggpubr (v4.0) and the rstatix (v0.7) package.

Results

Heterogeneity in MSQ across diagnostic laboratories

For the baseline quality assessment, we received 5,035 spectra measured on 41 devices from the 36 participating laboratories. We observed differences between the devices in MSQ considering the number of marker masses detected (e.g. device 7: median=25; interquartile range (IQR)=[20,28] and device 32: median=5, IQR=[3,14]) (**Figure 2A**).

The heterogeneity of MSQ was reflected in varying accuracy in species identification: Over all bacterial strains and using a marker-based species identification, the fraction of spectra, which were correctly and uniquely identified to the species level, ranged from 22.5% (18/80 spectra, device 9) to 78.2% (147/188 spectra, device 35). Strains of the species *Staphylococcus aureus* (99/110 spectra), *Staphylococcus argenteus* (98/108 spectra), *Staphylococcus schweitzeri* (107/112 spectra), *Listeria monocytogenes* (219/230 spectra), *Gardnerella vaginalis* (86/90 spectra) and *Pasteurella multocida* (97/101 spectra) were correctly and uniquely identified using the PAPMID™ database in at least 90% of comparisons, whereas strains of the species *Winkia neuii* (0/115 spectra), *Corynebacterium urealyticum* (8/120 spectra) and *Shigella flexneri* (8/112 spectra, strain #14) were correctly and uniquely identified in less than 10% of comparisons. We observed no difference in MSQ between the different MALDI-TOF MS manufacturers and an increasing accuracy of species identification with increasing MSQ (**Figure 2A**).

The MSQ differed between bacterial groups with a lower MSQ observed for spectra of Gram positive isolates compared to spectra of Gram negative isolates (ribosomal subunits detected: 16, [12, 20] vs. 18, [15, 22], p-value < 0.0001, p-values < 0.0001) (**Figure 2B**).

In order to assess how well the participating devices were calibrated, we compared measurement errors and observed a four-fold difference when comparing the most and least precise measurements (device 11: median=147 ppm, IQR= [109 ppm,192 ppm] vs. device 41: median=588 ppm, IQR=[533 ppm,631 ppm]) (**Figure 2C**).

These differences in MSQ are mainly represented by a few participating laboratories, while most laboratories provided data, which cluster around the overall median (16 (IQR = [13,19]) ribosomal subunits detected and 280 ppm [177 ppm, 426 ppm] in measurement error).

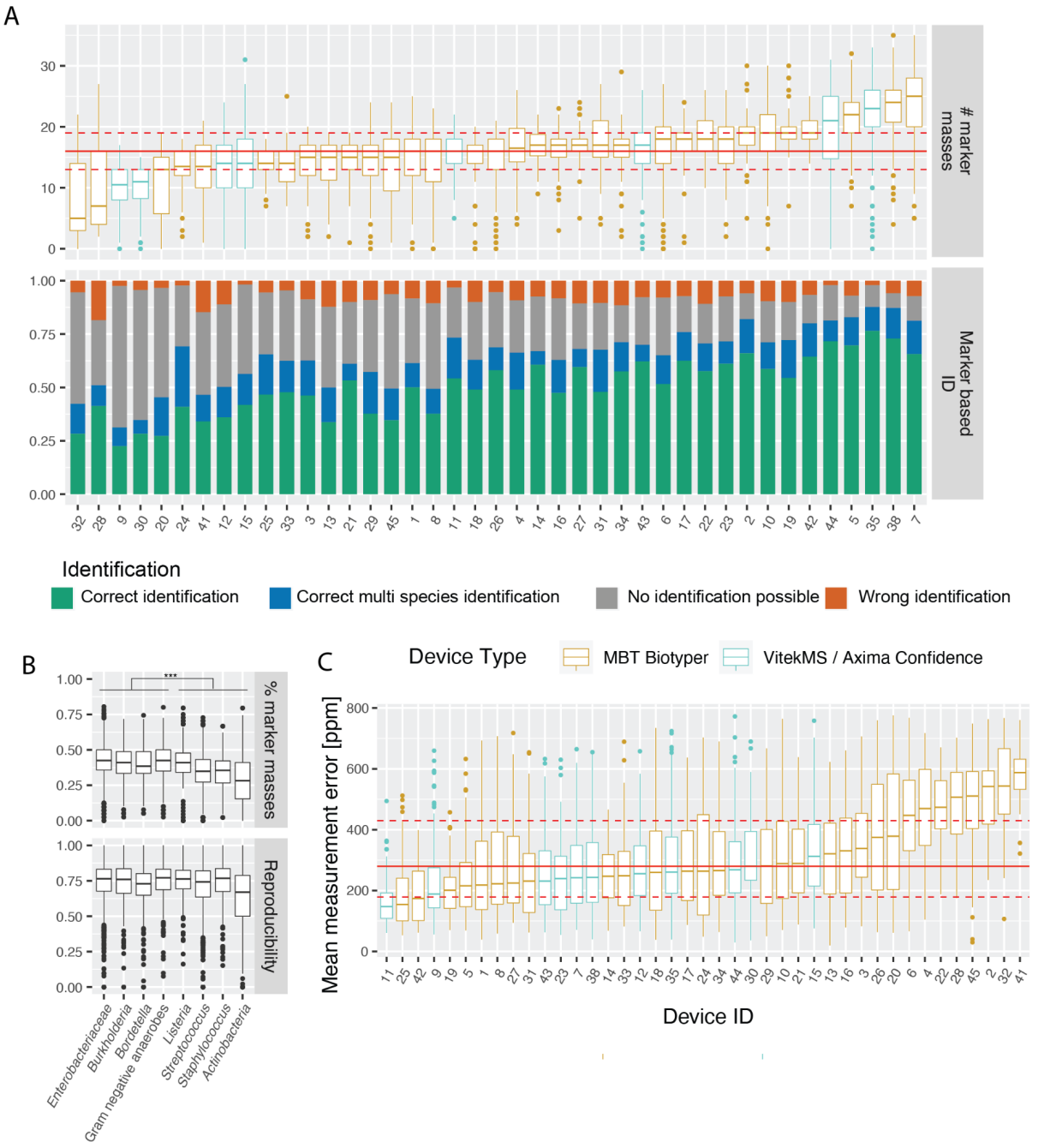

Figure 2: Comparison of MSQ at baseline quality assessment A: Number of ribosomal marker masses detected in MALDI-TOF mass spectra acquired on 41 devices (upper row) and the evaluation of the species identification results using a marker-based approach (lower row). **B:** Relative number of marker masses detected per phylogenetic group (upper row) and reproducibility between technical replicates (lower row). **C:** Measurement errors of the different devices.

Routine laboratory practices are associated with MSQ

Every participating laboratory filled out a questionnaire on laboratory practices (**Suppl. Table 3)**. Linking these answers to the acquired mass spectral data, we aimed to describe which practices are associated with high MSQ. As in each laboratory different combinations of

practices apply, some of which are not reflected in this questionnaire, we cannot identify causative factors of laboratory practices on spectral quality. However, we observed that certain practices correlated with improved MSQ represented by an increased number of ribosomal subunits detected: (a) Acquisition of spectra on steel targets compared to disposable targets (17 [13, 20], vs. 13 [10, 16], p-value < 0.0001), (b) Cleaning steel target plates with '*Methanol-Acetone*' protocol compared to other cleaning protocols (23 [18, 26], vs. 16 [12, 18], p-value < 0.0001) (c) Performing regular hardware services (17 [13, 20], vs. 15 [11, 18], p-value < 0.0001), (d) Working with a MALDI-TOF MS workstation (i.e. for a certain period, one or more member of staff are responsible for all MALDI-TOF MS measurements as opposed to all staff members acquiring MALDI-TOF mass spectra at any timepoint) (17 [14, 21], vs. 16 [12, 18], p-value < 0.0001), (f) Replacing the matrix solution after 7 or less days (17 [13, 20], vs. 15 [11, 17], p-value < 0.0001) and (f) Sub-culturing isolates on agar plates at least once after de-freezing, or culturing the isolates on agar plates from the eSwab transport medium, compared to strains which were measured directly after culturing on agar plates from frozen stocks (17 [13, 20], vs. 11 [6, 15.8], p-value < 0.0001) (**Suppl. Figure 3**).

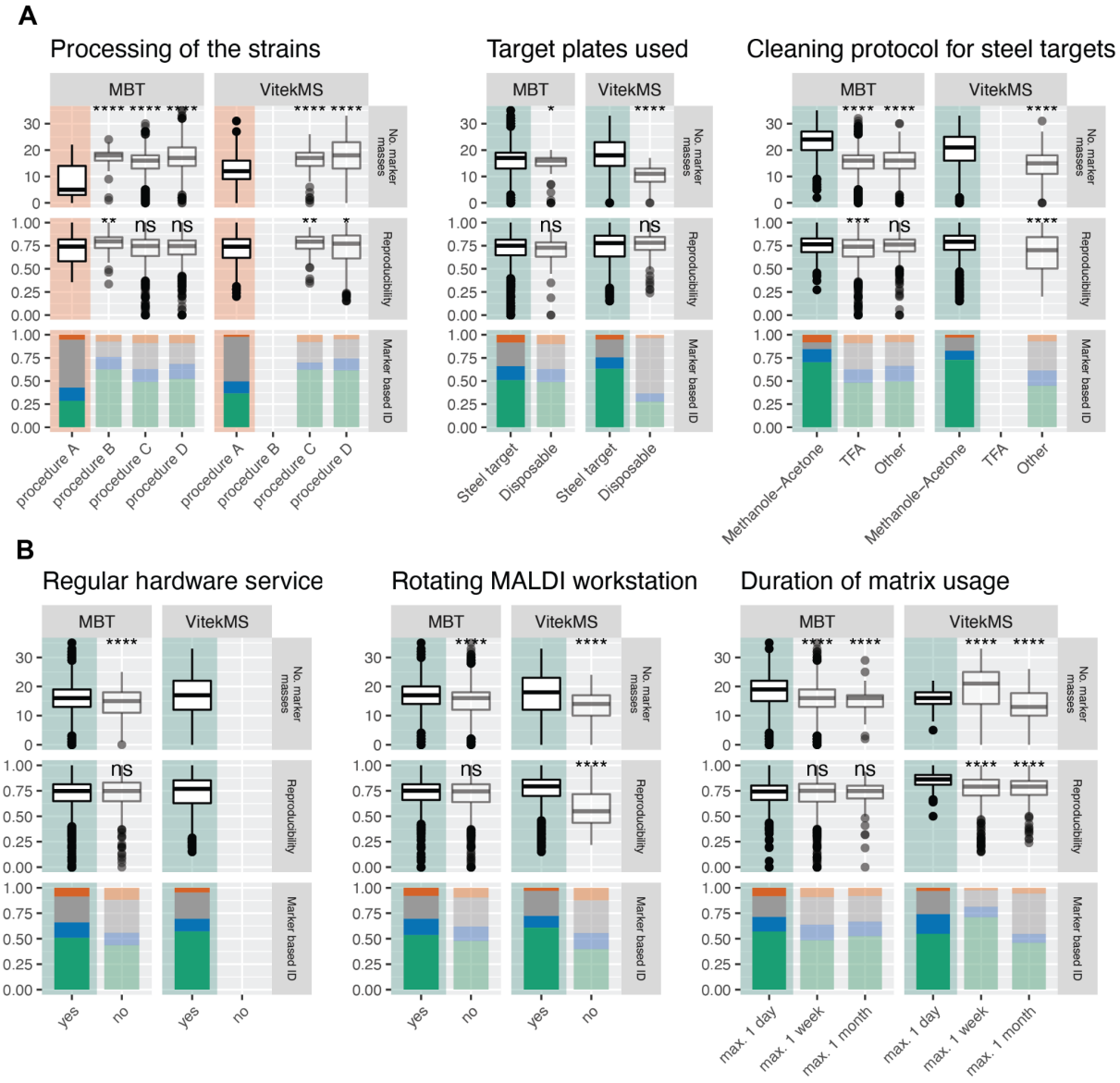

Figure 3: Mass spectra quality features (i) number of detected marker masses and (ii) technical reproducibility of A: strains processed in laboratories using different culturing procedures (procedure A: Streaked out from frozen stock; procedure B: Streaked out from frozen stock, subcultured once; procedure C: Streaked out from eSwab; procedure D: Streaked out from eSwab, subcultured once) (left column), using different target plates (middle column) and using varying cleaning protocols (right column). B: performing hardware services or not (left column), working with a MALDI workstation or not (middle column) and keeping the matrix for varying time in the workflow (right column). Abbreviations: "MBT": microflex Biotyper; "VitekMS": includes VitekMS and Shimadzu devices. "***": p-value <0.001, unpaired wilcoxon-rank test

Impact of standardised protocols on MALDI-TOF MSQ

Calibration

In the next step, we aimed to assess whether adaptations of the sample preparation protocol could improve MALDI-TOF MSQ. We asked all participating laboratories to calibrate the devices before acquiring the second set of MALDI-TOF mass spectra, as calibration has been shown to decrease the measurement error (318). Compared to spectra acquired for the baseline quality assessment, the measurement error was significantly lower for the spectra acquired in this second round in 14/36 devices, no significant change was observed in 11/36 and a significant increase in measurement error was observed on 11/36 devices. When focusing on devices, yielding a measurement error above 500 parts per million (ppm), which is often used as a threshold for marker-based typing applications, we recorded a significant decrease in measurement error in 4/5 cases (**Figure 4**).

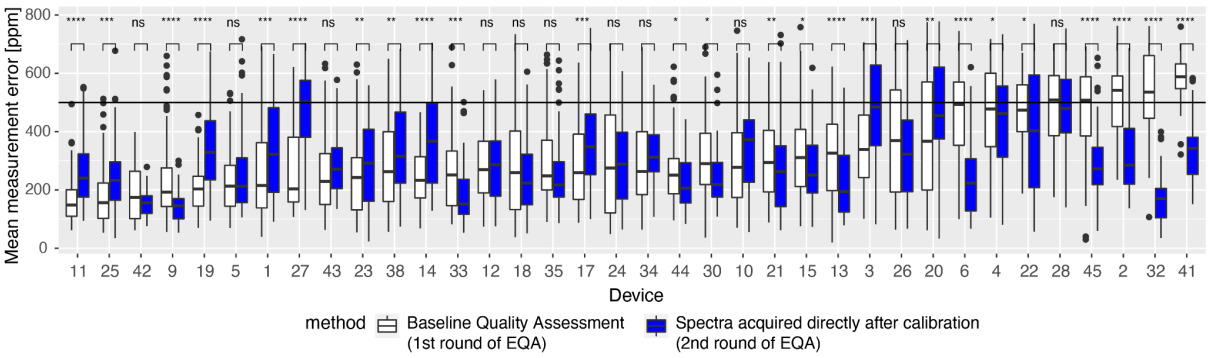

Figure 4. Measurement errors of spectra acquired in the baseline quality assessment (white) and after the intervention, which includes calibrating the device before spectra acquisition (blue). Devices are ordered, according to the median measurement errors recorded for spectra acquired for baseline quality assessment. Statistical comparisons performed using paired Wilcoxon rank tests. 'ppm' = parts per million.

Comparing sample preparation protocols per device

We observed an improved MSQ represented by an increased detection of marker masses using the *formic acid overlay protocol* compared to spectra acquired for baseline quality assessment in 10/36 devices, and a decrease in 17/36 devices. In 9/36 devices there was no significant change. Of the devices for which the median number of marker masses was lower than 15 for baseline acquired spectra (devices 9, 12, 20, 24, 25, 28, 30, 32, 33 and 41), we observed an increase of detected marker masses in 6/10 devices and a decrease in 1/10 devices.

Comparing the *group specific protocols* to the *formic acid overlay protocol*, we observed an increase in the number of marker masses detected in 7/34 devices, and a decrease in 18/34

devices. In 9/34 devices there was no significant change. The accuracy of species identification generally followed the number of marker masses detected (**Figure 5**).

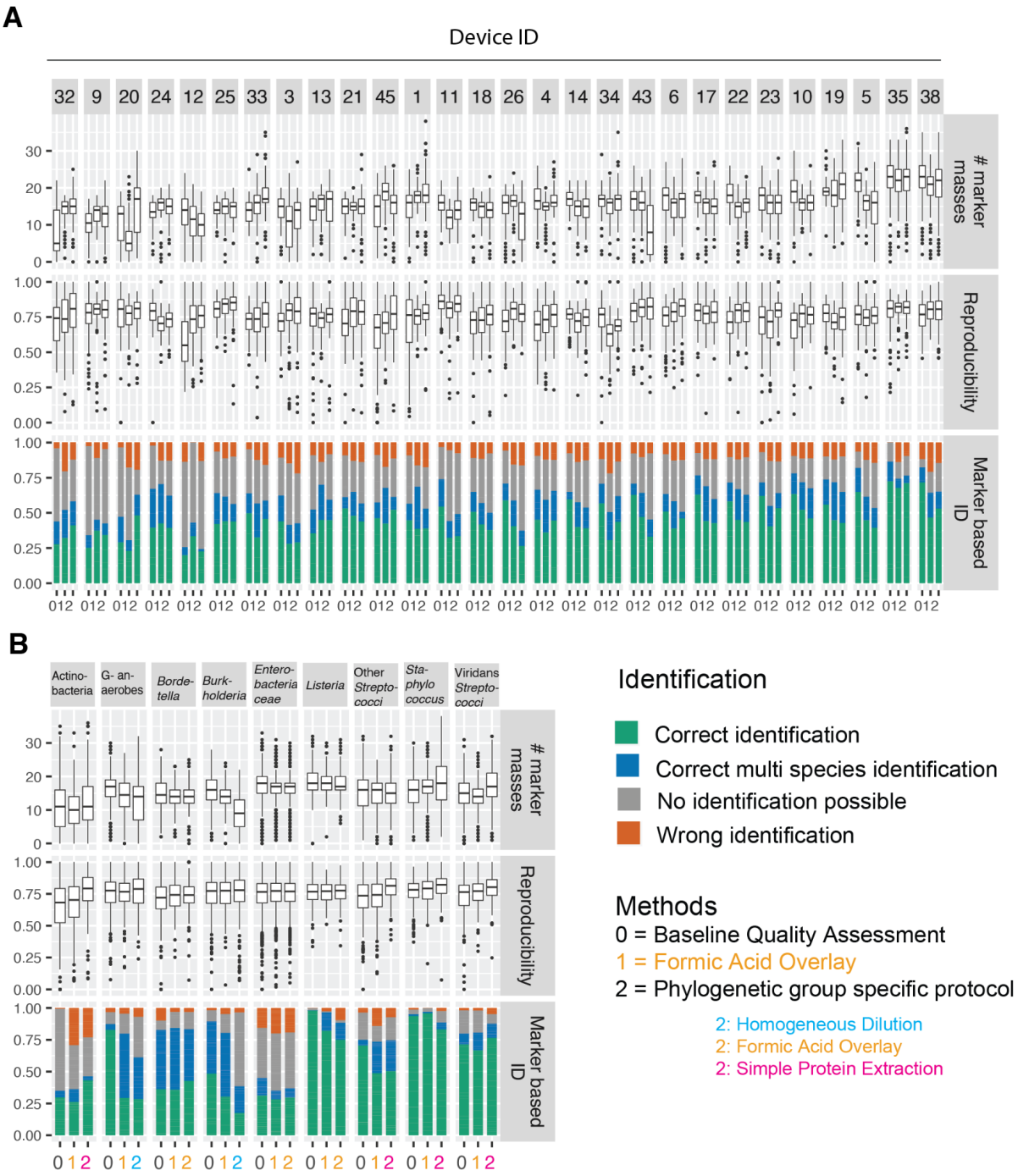

Figure 5 Impact of different sample preparation protocols on MSQ of spectra acquired on 28 MALDI-TOF MS devices (devices on which not all bacterial groups were measured with all three protocols were excluded from this graph). **A:** Number of marker masses detected (upper row), reproducibility between technical replicates (middle row) and evaluation of a marker-based species identification (lower row) for spectra acquired with different methods and on different devices. Devices are ordered according to the number of marker masses recorded in spectra acquired for the baseline quality assessment. **B:** Number of marker masses detected (upper row), reproducibility between

technical replicates (middle row) and evaluation of a marker-based species identification (lower row) for spectra acquired with different methods and from various bacterial groups. The median values for each device are connected with a line.

Sample preparation protocols have varying impact on different bacterial groups

When comparing spectra acquired with the *formic acid overlay protocol* to the routinely acquired spectra per bacterial group, we surprisingly observed an increase of marker masses only in 1/9 bacterial groups, namely *Staphylococcus* (**Figure 5B**).

Overall, we observed a negative impact of the *group-specific protocols* compared to the *formic acid overlay protocol* for many devices, which is mainly driven by two of the bacterial groups - *Burkholderia* and Gram negative anaerobes, for which we observed a decrease in the number of ribosomal subunits detected (14 (12, 16) vs. 9.5 (6, 13), p-value < 0.0001 and 15 (10, 18) vs. 14 (7, 17), p-value = 0.0153, respectively). For these two groups, the phylogenetic group-specific protocol, which we had suggested, required diluting the samples homogeneously in a buffer solution.

We observed a positive effect of using the *simple protein extraction protocols* compared to the *formic acid overlay protocol* for viridans streptococci (number of marker masses detected: 17 (13.75, 20) vs. 14 (12, 17, p-value < 0.0001), *Staphylococcus* (number of marker masses detected: 18 (13, 23), vs. 17 (15, 19), p-value = 0.002) and Actinobacteria (number of phylogenetic marker masses identified: 12 (7, 18) vs. 11 (7, 14, p-value = 0.03, p-value < 0.0001).

We observed the same general trend of the accuracy and resolution of a marker-based species identification following the number of ribosomal marker masses detected. However, this did not hold true for (a) staphylococci, for which we observed a higher number of non-identifiable spectra with a higher median number of ribosomal subunits detected, (b) viridans streptococci, for which an increase in detected ribosomal marker masses did not diminish the number of spectra which could not be identified and (c) actinobacteria, where the number of wrongly identified spectra increased drastically with an insignificantly lower number of ribosomal marker masses being detected.

Discussion

In this EQA, we systematically compared the MALDI-TOF MSQ between 36 routine diagnostic laboratories, using previously defined mass spectral features. EQA on the use of MALDI-TOF MS for microbial species have previously been reported (99), comparing the ability of diagnostic laboratories to identify a defined set of bacterial strains using MALDI-TOF MS. As the species identification results are influenced by the reference database and the MSQ, it is not possible to disentangle these two factors and the MSQ remains largely unknown. We (262)

and others (312) have previously shown how sample preparation adaptations can improve MSQ. Multiple studies examining MALDI-TOF MSQ have been performed on a single device (312,314). In this study, we assessed whether sample preparation protocols, which yielded high quality MALDI-TOF mass spectra in our hands, can increase MSQ more broadly in diverse routine diagnostic laboratories. We thereby compiled a comprehensive dataset of MALDI-TOF mass spectra with up to 250 technical replicates per bacterial strain, with extensive metadata and matching genomic sequences being publicly available.

For the baseline quality assessment, we asked the participating laboratories to culture and measure the strains as they would do in their diagnostic workflow in order to mimic routine spectra acquisition. This makes disentangling methodological effects difficult, but reflects diagnostic reality. The spectra quality observed from these measurements might differ from the spectra quality observed in routine diagnostics, for the following reasons: (i) The participating laboratories knew beforehand that the quality would be assessed, which might have biased the participants towards putting more effort in these measurements, by e.g. repeating measurements, by including an additional calibration of the device or to let most experienced personnel perform the measurements; (ii) these strains were shipped using eSwab transport media and were not cultured directly from patient material; and (iii) it was indicated to grow all strains on standard blood agar plate, whereas in routine diagnostics bacterial colonies might be picked from other media (e.g. chromogenic or MacConkey medium); (iv) the samples were processed outside of the routine workflow and the unusual situation could have decreased MSQ, as special effort was required not to mix up the samples and protocols.

We found a notable heterogeneity between measurements performed on different devices, which was driven by a few, poorly and highly performing devices. The fact that the MSQ from most devices clustered around the overall median highlights the robustness of the methods. When comparing spectra acquired using the standardised protocols included in our intervention to baseline acquired spectra, we found a positive effect of the intervention for devices performing poorly in the baseline quality assessment, whereas the effect of our intervention was often non-significant or even negative for devices performing well in the baseline quality assessment. As the sample preparation protocols were not strictly defined in the baseline quality assessment, it is difficult to identify patterns of sample preparation procedures, which are associated with an increase in MSQ on the basis of the data acquired here.

Hardware factors such as (i) the sort and age (related to strength) of the laser, (ii) the cleanliness of the ion source, and (iii) the tension of the detector can impact MSQ and were not considered in this study. These factors might differ between the baseline quality assessment and the assessment of MSQ after the intervention, as these measurements were

performed several months apart, and their effect on MSQ cannot be accounted for with the data presented in this study. Both protocols proposed in the intervention were tested at the same time with the same hardware settings, and the differences in MSQ between these two protocols can be attributed to sample preparation alone.

As previously described (312) the *simple protein extraction protocol* improves MSQ for Gram positive strains, however the observed effects were modest. For *Burkholderia* and Gram negative anaerobes, we proposed to prepare homogeneous dilutions of the samples, as we have previously reported a positive effect of such dilution steps for strains of these groups (262). However, this protocol did not perform well when tested by the participating laboratories, and we observed a negative effect on MSQ. We hypothesise that this discrepancy might either be the result of (i) differing amounts of bacterial inoculum used, as this was not indicated precisely enough in the instructions document or (ii) differing sensitivities of the MALDI-TOF MS devices used, coming from varying measurement settings.

We observed the accuracy and resolution of a marker-based species identification mainly following the number of detected ribosomal subunits. Possible explanations, for examples where this trend is not followed, could be (i) although more marker masses were detected, no more discriminatory markers were found and the identification could therefore not be improved, or (ii) these spectra were particularly noisy, which led to more false positive marker masses.

We have identified simple workflow adaptations which improve MSQ of poorly performing devices. We propose their usage whenever MSQ is not satisfactory and the desired bacterial identification cannot be achieved. Improving MALDI-TOF MSQ in routine diagnostics is essential for increasing the resolution of bacterial identification by MALDI-TOF MS, which is dependent on the reproducible detection of marker masses. The heterogeneity found in this EQA deserves further study in order to optimise MALDI-TOF MS-based routine identification in clinical laboratories.

Author Contribution

Planning and Conceptualisation of the study: AC and AE

Data acquisition: All authors

Data analysis: AC

Writing of the manuscript: AC and AE

Giving relevant feedback throughout the project and on the manuscript: All authors

Conflict of interest

The authors declare no conflict of interest.

Funding

None

Acknowledgements

We thank Dr. Vladimira Hinic, Dr. Fanny Wegner, Dr. Helena Seth-Smith and Dr. Marco Meola for valuable feedback on this manuscript. We thank Valerie Courtet and Daniel Gander (University Hospital of Basel, Basel, Switzerland); Dr. Orli Sagi and Dr. Boris Khalfin (Soroka University Medical Center, Beersheva, Israel); Manja Ipsen Hanegård and Tony Bønnelycke (University of Copenhagen, Copenhagen, Denmark); Ivana Jozić and Neno Petrić (University Hospital Centre Zagreb, Croatia); Alison Kolaru and Milena Antuskova (Department of Medical Microbiology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic); Stefanie Müller, Simon Feyer and Estelle Gohl (Institute of Veterinary Bacteriology, University of Bern, Bern, Switzerland); Jris Wyss (Institute for Infectious Diseases, University of Bern, Switzerland); Myriam Corthesy (Institute of Microbiology, University Hospital Lausanne, Switzerland); Eliane Haessler, Michael Trouessin and Pierre Glükler (Viollier AG) and Annette Wittmers (University Hospital Freiburg, Freiburg im Breisgau, Germany) for technical support and acquiring MALDI-TOF mass spectra at their respective centre.

Access to data

The filled-out questionnaire (**Table S3**) provides valuable metadata. The bacterial strains included in this study have previously been whole genome sequenced and the raw reads are publicly available (**Table S1**).

Chapter V: Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-021-01619-9

Check for updates

Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning

Caroline Weis[1,2] ✉, Aline Cuénod[3,4], Bastian Rieck[1,2], Olivier Dubuis[5], Susanne Graf[6], Claudia Lang[5], Michael Oberle[7], Maximilian Brackmann[8], Kirstine K. Søgaard[3,4], Michael Osthoff[9,10], Karsten Borgwardt[1,2,11] ✉ and Adrian Egli[3,4,11] ✉

This manuscript has been published in *Nature Medicine*:

Weis, Caroline et al. “Direct Antimicrobial Resistance Prediction from Clinical MALDI-TOF Mass Spectra Using Machine Learning.” *Nature Medicine*, January 10, 2022, 1–11. https://doi.org/10.1038/s41591-021-01619-9.

My contributions:

- Data collection (MALDI-TOF mass spectra and AMR profiles)
- Pre-processing of datasets DRIAMS-A and DRIAMS-B
- Compilation of clinical data
- Retrospective clinical analysis
- Data visualisation (Figure 6)
- Contribution to the writing of the manuscript

Note: The following part contains the full manuscript

Supplementary Material can be accessed via the following link:

https://doi.org/10.1038/s41591-021-01619-9

The complete dataset analysed in this project can be accessed via the following link:

https://doi.org/10.5061/dryad.bzkh1899q

Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning

Caroline Weis (1) (2)*, Aline Cuénod (3) (4), Bastian Rieck (1) (2), Olivier Dubuis (8), Susanne Graf (7), Claudia Lang (8), Michael Oberle (9), Maximilian Brackmann (10), Kirstine K. Søgaard (3) (4), Michael Osthoff (5) (6), Karsten Borgwardt (1) (2)*+, Adrian Egli (3) (4)*+

(1) Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
(2) SIB Swiss Institute of Bioinformatics, Switzerland
(3) Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland
(4) Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland
(5) Division of Infectious Diseases and Hospital Epidemiology, University Hospital and University of Basel, Basel, Switzerland
(6) Department of Internal Medicine, University Hospital Basel and University of Basel, Basel, Switzerland
(7) Department for Microbiology, Cantonal Hospital Baselland, Liestal, Switzerland
(8) Clinical Microbiology, Viollier AG, Allschwil, Switzerland
(9) Institute for Laboratory Medicine, Medical Microbiology, Cantonal Hospital of Aarau, Aarau, Switzerland
(10) Federal Office for Civil Protection, Spiez Laboratory, Proteomics, Bioinformatics and Toxins, Spiez, Switzerland

+ equally contributed (joint last authors)
* corresponding authors:
Caroline Weis
Machine Learning and Computational Biology Lab
ETH Zurich
Mattenstrasse 26
4058 Basel, Switzerland
Email: caroline.weis@bsse.ethz.ch
Phone: +41 61 387 34 39

Prof. Dr. Karsten Borgwardt
Machine Learning and Computational Biology Lab

ETH Zurich
Mattenstrasse 26
4058 Basel, Switzerland
Email: karsten.borgwardt@bsse.ethz.ch
Phone: +41 61 387 34 20

Prof. Dr. Dr. Adrian Egli, MD PhD
Clinical Bacteriology and Mycology
University Hospital Basel
Petersgraben 4
4031 Basel, Switzerland
Email: adrian.egli@usb.ch
Phone: +41 61 556 57 49

Abstract

Early administration of effective antimicrobial treatments is critical for the outcome of infections and the prevention of treatment resistance. Antimicrobial resistance testing enables the selection of optimal antibiotic treatments, but current culture-based techniques can take up to 72 hours to generate results. We have developed a novel machine learning approach to predict antimicrobial resistance directly from MALDI-TOF mass spectra profiles of clinical isolates. We trained calibrated classifiers on a newly-created publicly available database of mass spectra profiles from clinically most relevant isolates with linked antimicrobial susceptibility phenotypes. This dataset combines more than 300,000 mass spectra with more than 750,000 antimicrobial resistance phenotypes from four medical institutions. Validation on a panel of clinically important pathogens, including *Staphylococcus aureus*, *Escherichia coli*, and *Klebsiella pneumoniae*, resulting in AUROC values of 0.80, 0.74, and 0.74 respectively, demonstrated the potential of using machine learning to substantially accelerate antimicrobial resistance determination and change of clinical management. Furthermore, a retrospective clinical case study of 63 patients found that implementing this approach would have changed the clinical treatment of 9 cases, which would have been beneficial in 8 cases (89%). MALDI-TOF mass spectra based machine learning may thus be an important new tool for treatment optimization and antibiotic stewardship.

Introduction

Antimicrobial resistant bacteria and fungi pose a serious and increasing threat to the achievements of modern medicine (327,328). Infections with antimicrobial resistant pathogens are associated with substantial morbidity, mortality, and healthcare costs (24). Rapid treatment with an effective antimicrobial is critical for the outcome of an infection (49,270). However, antimicrobial therapy and dosage need to account for the resistance profiles of presumed pathogens, and also have to consider host-specific factors such as patient age, kidney function, previous medical history, and concurrent medication. Early identification of the microbial species causing an infection can improve targeting of therapeutic options based on e.g., the knowledge of intrinsic resistance mechanisms and local epidemiology of resistance (48,329). However, only a detailed resistance profile permits treatments to be fully optimized. With current culture-based methods, the time from sample collection to resistance reporting can take up to 72 hours, meaning that for a substantial period, a patient may be receiving a too narrow- or too broad-spectrum antimicrobial drug (330,331). To limit the infection-related risk to a patient, broad-spectrum antibiotics are often used. The concept of an optimal selection of an antibiotic drug is an important pillar of antibiotic stewardship and has gained significant attention owing to the global emergence and spread of antibiotic resistant pathogens. A reduction in the time required for a resistance profile to become available will not only substantially improve patient outcomes, but would also align well with other goals of antibiotic stewardship (332), including reducing reliance on precious broad-spectrum antibiotic treatments, reducing unnecessary broad antibiotic use, and thereby combating the development of antibiotic resistance. In addition, rapid information on antimicrobial resistance may help to speed up infection prevention measures such as the isolation or cohorting of patients infected with presumed multidrug resistant pathogens. PCR-based molecular diagnostics may be able to detect single resistance genes directly from patient specimens more rapidly than any culture-based diagnostics. However, such molecular assays are generally narrow-spectrum assays of single gene targets and also suffer from problems with specificity of resistance that is not genetically-mediated (e.g. upregulation of efflux pumps), and the associated costs (333–335).

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry (MS) has proven to be a rapid technology for microbial species identification. In just a few minutes, MALDI-TOF MS can be used to characterise the protein composition of single bacterial or fungal colonies (64,98,336), which are usually available within 24 hours after sample collection (48). MALDI-TOF MS enables precise and low-cost microbial identification, which has led to the technology becoming the most commonly-used method for microbial species identification in clinical microbiology laboratories (337). MALDI-TOF MS has the

potential to move beyond simply identifying an infecting pathogen. Extracting additional information directly from acquired MALDI-TOF mass spectra data may also enable antimicrobial susceptibility testing. Indeed, a recent study used MALDI-TOF mass spectra to detect markers associated with methicillin resistance in clinical samples of *Staphylococcus aureus* (338). However, the absence of a comprehensive catalogue of marker masses for all potential pathogen and drug combinations has made translating such efforts to clinical practice impossible. In this study, we harnessed the full potential of MALDI-TOF MS to predict antimicrobial resistance through machine learning methods. In this context, previous efforts are rare (339,340) and stymied by the lack of large, publicly-available, high-quality benchmark datasets (341,342).

To develop clinically-applicable mass spectra-based antimicrobial resistance prediction approaches, we created the Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra (*DRIAMS)*. *DRIAMS* is a large-scale, publicly-available, high quality collection of bacterial and fungal MALDI-TOF mass spectra derived from routinely-acquired clinical isolates, coupled with the respective laboratory-confirmed antibiotic resistance profile. We used *DRIAMS* to undertake the first large-scale study of the utility of such spectra for antimicrobial resistance prediction, with the aim of improving both patient treatment decisions and antibiotic stewardship.

Results

DRIAMS: Clinical routine database

From 2016 to 2018, we assembled a dataset of MALDI-TOF mass spectra from more than 300,000 clinical isolates from four different diagnostic laboratories in Switzerland. The raw dataset comprises a total of 303,195 mass spectra and 768,300 antimicrobial resistance labels and represents 803 different species of bacterial and fungal pathogens. The dataset was processed and organised in four sub-collections (*DRIAMS-A* to *-D*; **Fig. 1**). *DRIAMS-A,* the largest collection with 145,341 mass spectra, was collected at the University Hospital Basel (Switzerland) and is used for the main analysis presented in this study. *DRIAMS-A* contains resistance labels associated with 71 different antimicrobial drugs, for which the number of spectra and antimicrobial resistance ratios can be found in **Suppl. Tab. 1** and **2**. Importantly, the MALDI-TOF mass spectra in *DRIAMS-A* could be obtained from clinical samples within 24 hours of collection, enabling species identification on a rapid scale as compared to standard phenotypic resistance testing (**Extended Data Fig. 1**). The complete *DRIAMS* database is publicly available at https://doi.org/10.5061/dryad.bzkh1899q.

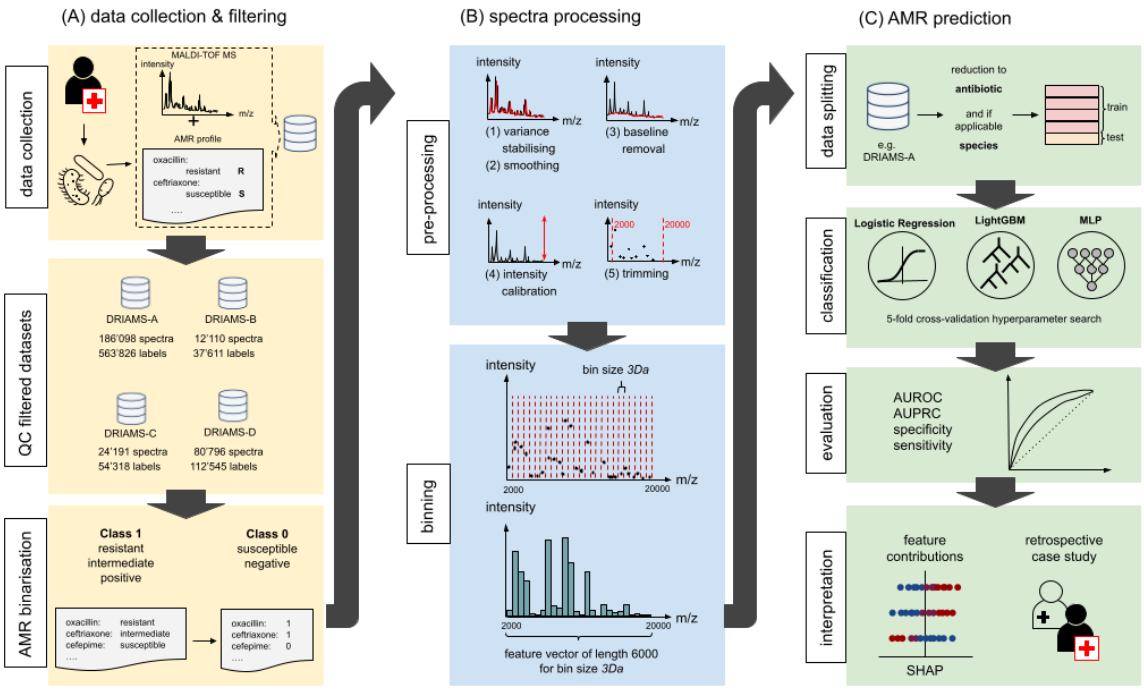

Figure 1: MALDI-TOF MS based AMR prediction workflow. Workflow of MALDI-TOF MS data pre-processing and antibiotic resistance prediction using machine learning. **A.** (i) Data collection: Samples are taken from infected patients, pathogens are cultured, and their mass spectra and resistance profiles are determined. Spectra and resistance are extracted from the MALDI-TOF MS and laboratory information system; corresponding entries are matched and added to a dataset. Samples are filtered according to workstation. (ii) Quality control (QC) filtered datasets: After several quality control steps, the datasets are added to *DRIAMS*. (iii) Antimicrobial resistance (AMR) binarisation: antimicrobial resistance is defined as a binary classification scenario, with the positive class represented by all labels leading to the antimicrobial not being administered, i.e. intermediate or resistant, and positive, while the negative class represents susceptible or negative samples. **B.** (i) Pre-processing: Cleaning of mass spectra. (ii) Binning: Binning spectra into equal-sized feature vectors for machine learning. **C.** (i) Data splitting: For the experiments the samples are subset to only one species. Data is split into 80% training and 20% test, stratified by both antimicrobial class and patient case number. (ii) Classification: Antimicrobial resistance classifiers are trained using a 5-fold cross-validation for hyperparameter search, using the classification algorithms logistic regression, LightGBM, and a deep neural network classifier (MLP). (iii) Evaluation: Predictive performance is measured in metrics commonly used in machine learning (AUROC and AUPRC) and the medical community (specificity and sensitivity). (iv) Interpretation: Interpretation of individual feature contribution to antimicrobial resistance prediction through Shapley values and clinical impact through a retrospective case study on samples from the latest four months of collected data.

Machine learning for MALDI-TOF MS based resistance prediction

To move beyond simple species identification, we pre-processed and binned mass spectra measurement points into fixed bins of 3 Daltons (Da), ranging from 2,000 Da to 20,000 Da, thus obtaining a 6000-dimensional vector representation for each sample. The selected bin size is sufficiently large to adequately represent each spectrum while still remaining computationally tractable (for details see **Methods**). Next, we converted the antimicrobial resistance categories, which are either recorded as *susceptible*, *intermediate*, or *resistant* in the laboratory reports associated with each sample, into a binary label (susceptible vs. intermediate/resistant) (for details see **Methods**). Specifically, we assigned intermediate or resistant samples to the positive class, and susceptible samples to the negative class (in most of the scenarios we considered, the positive class was the minority class). We then split the samples into training and testing datasets, ensuring that all data associated with a specific case was either part of the train dataset, or the test dataset, but not both, while keeping a similar antimicrobial class ratio in both train and test dataset. We used three machine learning approaches for classification, i.e. logistic regression (LR), gradient-boosted decision trees (LightGBM), and a deep neural network classifier (multi-layer perceptron, MLP), to predict resistance to each individual antimicrobial. The three models were selected because they represent different complexity classes of classifiers (for a more in-depth description of these approaches, please see **Methods**). Subsequently, we report the 'area under the receiver operator characteristic curve' (AUROC) and 'area under the precision-recall curve' (AUPRC) as performance metrics. AUROC can be understood as the probability of correctly classifying a pair of samples, i.e. a resistant/intermediate one and a susceptible one; AUPRC quantifies the ability to correctly detect samples from the smaller of the two classes (resistant/intermediate), while minimising false discoveries. Overall, we observed that LightGBM and MLP were the best-performing classifiers in terms of AUROC. **Fig. 1** depicts the workflow from data collection and filtering, to spectra processing, and antimicrobial resistance prediction results.

Species-specific AMR prediction yields high performance

We first sought to determine whether the use of species-specific mass spectra in *DRIAMS-A* would result in high predictive performance. To this end, we performed a focused analysis for three clinically important pathogens: *Staphylococcus aureus, Escherichia coli*, and *Klebsiella pneumoniae*, all of which are on the World Health Organization 'priority pathogens' list (343). For each of the three species, we selected relevant antibiotics to test based on their clinical usage. We then created a *DRIAMS-A* subset for each antibiotic, which we further divided into stratified training and testing data as described above. For each of the three species we chose

one antibiotic resistance as the major scenario of interest; namely oxacillin as a marker for Methicillin-resistant *S. aureus* (MRSA) (344), and ceftriaxone resistance in *E. coli* and *K. pneumoniae* as a marker for resistance against broad spectrum beta-lactam antibiotics e.g., extended spectrum beta-lactamases or carbapenemases. We then trained a classifier using each model for each of the three major species-antibiotic pairs (see **Suppl. Tab. 3**). We analysed to what extent the respective best model was capable of predicting resistance to other antibiotics (see **Fig. 2**), observing high overall performance in both AUROC and AUPRC values; the classifier is therefore capable of providing precise antimicrobial resistance predictions. For *S. aureus*, the prediction of oxacillin resistance reached a high performance with AUROC of 0.80 and AUPRC of 0.49 at a positive (i.e. resistant/intermediate) class ratio of 10.0%. The percentage of positive samples in the test dataset (class ratio) states the AUPRC performance of a random binary classifier and therefore states the baseline to which the classification performance needs to be compared to (see Methods for further details). According to clinical laboratory protocols used in *DRIAMS-A*, for *S. aureus* strains, the reported susceptibilities to beta-lactam antibiotics are inferred from the oxacillin susceptibility test results. We also observed high performance for *E. coli* and *K. pneumoniae*, where the prediction of ceftriaxone resistance reached AUROC values of 0.74 in both species, and AUPRC values of 0.30 and 0.33, at a positive class ratio of 10.0% and 8.2%, respectively. We would expect the generation of such resistance information within 24 hours to have a substantial impact in treatment adaptation and infection prevention management. Overall, this experiment demonstrated that a species-specific classifier can achieve clinically useful prediction performance with significantly faster determination of antibiotic resistance compared to the laboratory standard of phenotypic resistance determination (**Extended Data Fig. 1**). We also analysed to what extent the combination of species identity and mass spectrometry information outperforms predictions based on species identity alone. We analysed AUROC predictive performance for the 42 studied antibiotics (see **Extended Data Fig. 2**). For 31 of them, AUROC values above 0.80 were reached, implying highly accurate predictions. Moreover, for 22 antibiotics, we observed statistically significant improvements in prediction performance using the combined mass spectra in *DRIAMS-A* as compared to using only species information for resistance prediction. The results clearly demonstrate the predictive power of mass spectra based antimicrobial resistance prediction.

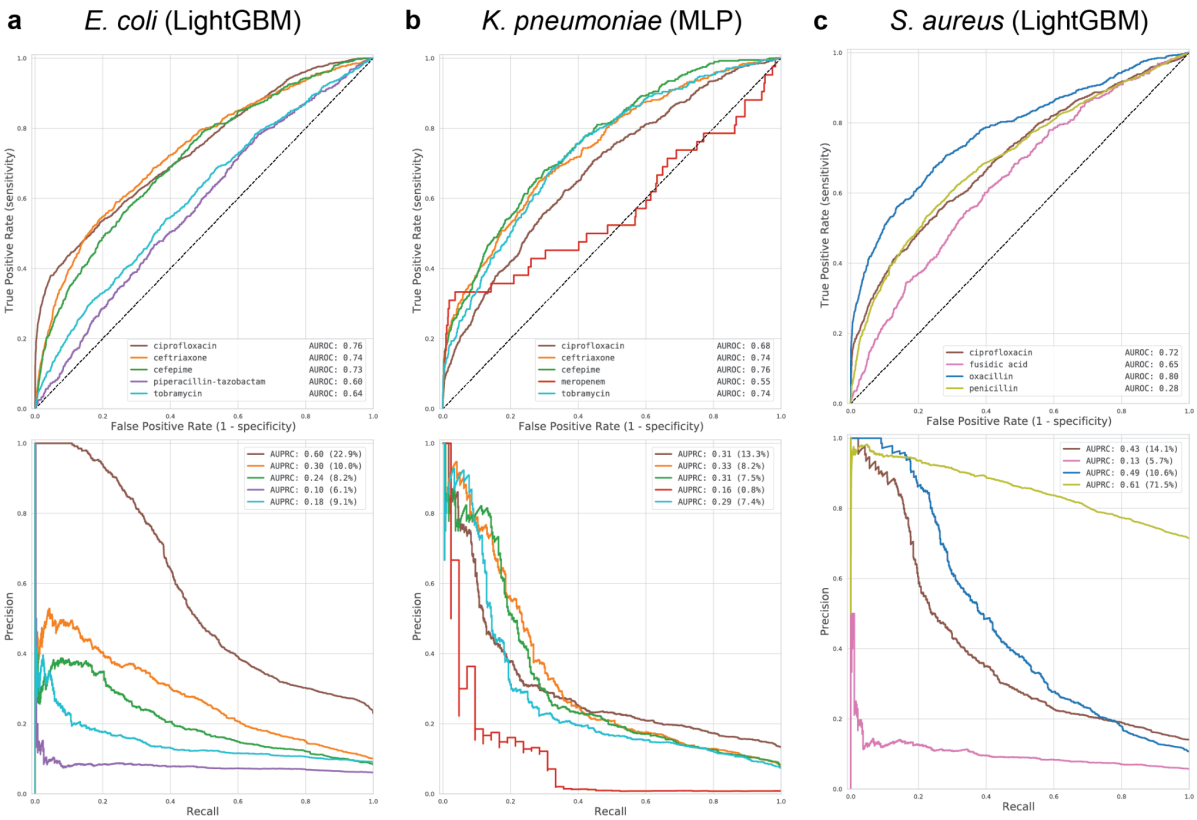

Figure 2: ROC and PR curves of best performance antimicrobial resistance prediction models on *DRIAMS-A.* The curves were created by *appending the scores* while the displayed values stem from *reporting the mean* for 10 different shuffled stratified train–test splits, matching values to the tables. **A. *E. coli* (LightGBM)** For *E. coli*, the best-performing predictor was that for ciprofloxacin, followed by ceftriaxone, critical antibiotics indicating an extended beta-lactamase (ESBL) if resistant. **B. *K. pneumoniae* (MLP)** For *K. pneumoniae*, cefepime exhibited the highest performance of 0.76 AUROC, also indicating an ESBL if resistant. Compared to the other scenarios, its ROC curve has a larger step size, but with over 500 test samples, the sample size is comparable to the other antibiotics. **C**. *S. aureus* (LightGBM) Finally, for *S. aureus*, our model performed best for oxacillin, with an AUROC of 0.78. This is particularly relevant, as for *S. aureus*, the resistance to other beta-lactam antibiotics (including amoxicillin/clavulanic acid and ceftriaxone) is directly derived from oxacillin resistance, indicating a methicillin resistant *S. aureus* (MRSA).

External datasets improve antimicrobial resistance prediction

The use of pre-trained machine learning models could expedite uptake of this approach in clinical laboratories already using MALDI-TOF MS for species identification. As such, we assessed whether predictive performances reached using data from one site (e.g. one specific hospital) are transferable to other sample collection sites. For the datasets *DRIAMS-A* to *-D*, each representing one of our four sites, we divided data associated with each case into train and test datasets as described above, and then trained a predictor before testing on each site. We also compared this site-specific training with predictors trained across all sites. The results

indicate that site-specific training reaches better predictive performance compared to across-site validation. Within the site-specific training, the large *DRIAMS-A* dataset is the or among the best-performing sites (**Fig. 3**).

We further investigated whether we could improve prediction for sites where a large dataset is unavailable by leveraging existing large external datasets such as *DRIAMS-A*. We therefore trained on combinations of training datasets from different sites, including different combinations of the four sites *DRIAMS-A* to *-D*, and tested on a single external site *DRIAMS-B* to *-D*. While the transferability of predictive performance from one site to another is an active area of research in the machine learning field of domain adaptation, a recent study (345) has shown that using empirical risk minimization by learning a single model on pooled data across all training environments often outperforms more complex domain adaptation approaches. The results indicated that the addition of training datasets from other sites to the external site train data was beneficial for validation sites *DRIAMS-B* and *-C* (**Suppl. Tab. 4**). For external validation site *DRIAMS-D*, the best predictive performance was still reached when training exclusively on the site-specific training data. For external validation sites *DRIAMS*-B, the addition of the large *DRIAMS-A* dataset proved most beneficial for scenarios *E. coli* (ceftriaxone) and *K. pneumoniae* (ceftriaxone), while adding more data from *DRIAMS-C* to the training data was beneficial for *S. aureus* (oxacillin).

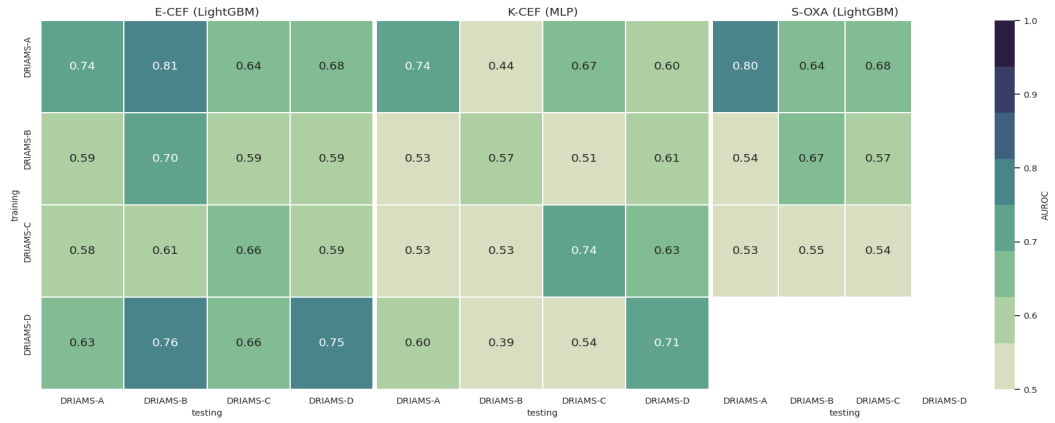

Figure 3: Validation predictive performance of each scenario trained and tested on *DRIAMS-A* to *-D* (AUROC). The results show the mean AUROC performance of 10 random train-test splits. For comparability, the train-test splits are kept the same for each of the respective four *DRIAMS* datasets. The values reported on the top-right (both training and testing *DRIAMS-A*) correspond to the values reported in **Suppl. Tab. 3**. With the exception of *DRIAMS-B E. coli* (ceftriaxone), the highest performance is reached when training is performed on the same site as testing. *DRIAMS-A* and *DRIAMS-D* exhibit the highest transferability with respect to predictive performance, and overall, transferability seems higher in *E. coli* as compared to *K. pneumoniae* and *S. aureus*. Due to the different class ratios between test datasets on different sites, AUROC was chosen to permit comparability. The scenario abbreviations follow **Suppl. Tab. 3**. For S-OXA no *DRIAMS-D* data is available.

Species-stratified learning yields superior predictions

Next, we analysed whether classifiers can improve the predictive performance by training on a large number of samples from multiple species (as opposed to training on samples from a single species). It is known that different species of bacteria can be resistant to a specific antimicrobial through different mechanisms. For example, resistance against beta-lactam antibiotics in Gram-negative bacteria, such as *E. coli*, may be caused by the production of beta-lactamases, such as CTX-M (346), TEM, and SHV (347,348) or carbapenemases, such as OXA-48 (349). Resistance against beta-lactam antibiotics in Gram-positive bacteria, such as *S. aureus*, can be caused by a penicillinase (blaZ), resulting in a resistance only against penicillin (350), or by an alteration within the penicillin-binding protein (PBP2a), resulting in the MRSA phenotype with resistance against multiple beta-lactam antibiotics (351). Hence, pooling spectra across species and predicting antimicrobial resistance using the same model regardless of the species poses a more complex learning task than predicting antimicrobial resistance within one specific species. However, stratification by species reduces the number of samples available for training and might therefore lower predictive performance. We assessed the trade-off between the number of available samples and predictive performance by comparing the performance of (i) a model trained to predict antimicrobial resistance using samples from across all species (*ensemble*) with (ii) a collection of models trained separately for *single* bacterial species (**Fig. 4A**). Each point of the depicted curves corresponds to one classifier, trained with the number of samples specified on the x-axis. The last, i.e. rightmost, point of each curve hence corresponds to the scenario in which all available samples are being used. We observed that training a model for individual species *separately* led to improved performance for all species, despite the reduction in sample size. Notably, all training samples used to reach the last single-species classification results were also included in the training samples for the last ensemble classifier. The last ensemble classifier therefore had access to at least the same amount of information about the respective species as the last single-species classifier. Nevertheless, it never outperformed the single-species classifiers except for oxacillin resistance in *S. aureus*. Furthermore, a few curves reached a plateau, with the single-species classifier increasing more sharply with the last addition of more training samples. This demonstrates the higher complexity of the ensemble prediction task and the benefit of a larger training dataset, which are critical for capturing different resistance mechanisms.

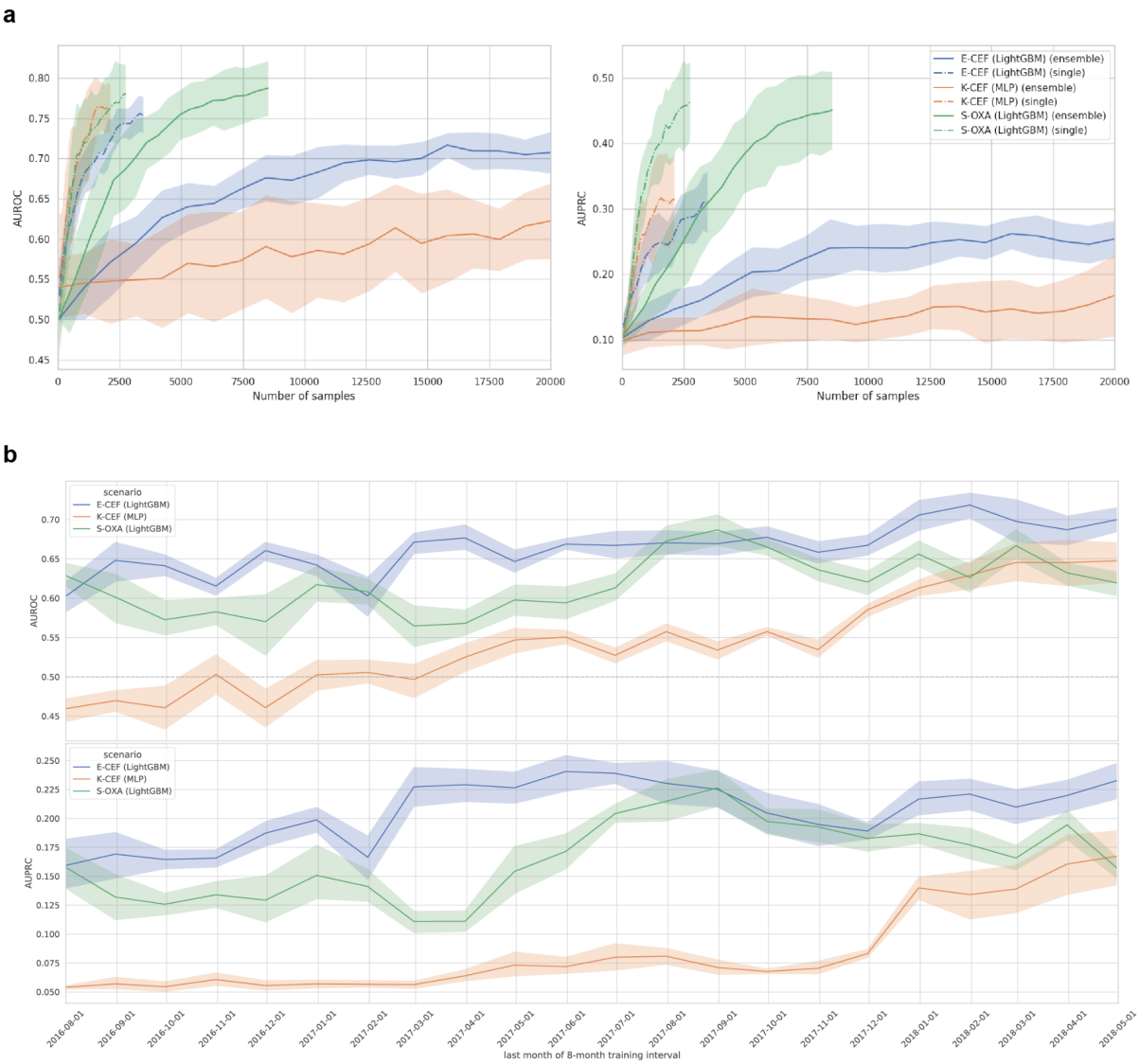

Figure 4: Stability of results with different dataset perturbations. The scenario abbreviations follow **Suppl. Tab. 3. A. Predictive performance with increasing sample size** AUROC and AUPRC as a function of sample size for complete and species-stratified *DRIAMS-A* datasets. Experiments were repeated for ten different shuffled train–test splits. The data are presented as mean values (represented by solid curves) +/- standard deviation (upper and lower envelope curves) of these repetitions. Results are shown for the three major scenarios of interest. With equal sample size, training only on samples from a single species is outperforming training in all scenarios. Even for the datasets containing all available samples from the target species (the rightmost points of each curve), the single-species scenario outperforms the ensemble in both *E. coli* and *K. pneumoniae* (ceftriaxone), while the curves reach a similar predictive performance for *S. aureus* (oxacillin). **B. Temporal validation in *DRIAMS-A* reporting AUROC (upper) and AUPRC (lower) of sliding 8-month training window on fixed test set.** The test dataset is the data collected in the last four months May to the end of August 2018. For *E. coli* (ceftriaxone) (E-CEF) and *S. aureus* (oxacillin) (S-OXA) the predictive performance decreases with increasing temporal distance to the test set, but the fluctuations in the curve are of the same size as the drop over time. The predictive performance for ceftriaxone in *K. pneumoniae* (K-CEF) decreases more continuously and drastically with increasing temporal distance to the test set.

Current samples necessary for accurate resistance prediction

Mass spectra profiles are subject to variations and differences over time, caused by biological differences through the ongoing evolution of the local microbial populations (with new strains being introduced by, for example, travelling), or by technical differences, such as changes after MALDI-TOF MS machine maintenance (such as laser replacement and adjustment of internal spectra processing parameters through machine calibration). To guide and encourage further method development, we wanted to illustrate challenges and limits of mass spectra based antimicrobial resistance prediction. Hence, we studied whether recent samples are necessary, and whether adding more samples collected at older timepoints would increase predictive performance. We fixed the latest four months of data from *DRIAMS-A* as a test dataset, and trained classifiers on data collected within 8-month training windows with increasing temporal distance to the test collection window, simulating the availability of older samples. The training data within the training window was oversampled to match the class ratio in the test data; however, sample sizes could still vary between training windows. We observed a slight *decrease* in performance with increasing temporal distance between training and testing data (**Fig. 4B**) for *E. coli* and *S. aureus;* with a larger decrease for *K. pneumoniae*. We explain this drop by the aforementioned differences that accumulate over time, highlighting the positive effect of having access to recent training samples. **Extended Data Fig. 5** also indicates that the reduction in performance training on older datasets could in fact stem from a lower sample size, as the MALDI-TOF MS technology usage at the *DRIAMS-A* collection site increased over time.

Analysis of feature contributions through Shapley values

Only very few studies have considered full mass spectrum information instead of single peaks for antimicrobial phenotype prediction[19]. We therefore wanted to assess whether predictive performance is primarily driven by only a subset of the peaks, or whether the full spectrum is employed. While this question is partially addressed by the use of feature importance values, their use without additional information can be misleading as their interpretation is highly contingent on the classifier that was employed. Hence, for further analysis, we also calculate the Shapley values, a concept originating from coalitional game theory, which enable the interpretation of model output contributions on both the dataset and per-sample level for each feature (352). **Fig. 5** visualizes the average and per-datapoint Shapley values for the 30 features with the highest average contribution. As the tails of the distribution plots for each feature are coloured with either the highest or lowest feature value, we see that the predictor is using either the presence of a high intensity value (red) or the absence of any measured intensity (blue) for a positive (resistant/intermediate) class prediction. In case of *S. aureus* (oxacillin) for the top four mass-to-charge bins the presence of a peak indicates the positive

(resistant/intermediate) class, while for *E. coli* (ceftriaxone) also the absence of a peak can strongly contribute to a positive class prediction. We further observe that most of the feature bins with the highest average impact are feature bins with a mass-to-charge of less than 10,000 Da (79 out of 90 features bins in **Fig. 5**). Most proteins that are reproducibly detected in MALDI-TOF MS have a weight less than 10,000 Da (262) and the signal indicates their presence or absence. The feature importance distributions over all 6,000 features (**Extended Data Fig. 4**) stemming directly from the classification models indicate that the classifiers utilize the entire range of features.

Most reference studies have focused on oxacillin resistance in S. aureus and have identified peaks that were either used to distinguish between methicillin susceptible S. aureus (MSSA) and MRSA or to distinguish between MRSA sub-lineages (66,77,92,94,353–357). A subset of these discriminatory peaks were identified to correspond to either (i) constitutively conserved housekeeping genes or (ii) other peptides such as stress proteins or low molecular weight toxins (354). The mass of three of the identified proteins, 3007 Da (Delta-toxin), 3891 Da (uncharacterised protein SA2420.1), and 4511 Da (uncharacterised protein SAR1012) can be attributed to highly contributing feature bins (**Fig. 5**, **Suppl. Tab. 5**). A peak at 2415 m/z has previously been identified as MRSA specific (92). This peak corresponds to the peptide PSM-mec (77), which is encoded on a subset of SCCmec cassettes in close proximity to mecA (79,358), which encodes resistance to oxacillin. This peak corresponds to the 83rd highest-ranked feature bin of 2414-2417 m/z (out of 6000 feature bins overall) of our respective classifier.

The increased occurrence of multidrug resistant E. coli has been attributed to the spread of a few clonal lineages, in particular to sequence type (ST) 131 (359). Previous studies (93,360) have identified ST131 characteristic peaks (8448m/z. 8496m/z, 11783m/z), which can be attributed to feature bins receiving high feature importances and Shapley values by the ceftriaxone E. coli classifier.

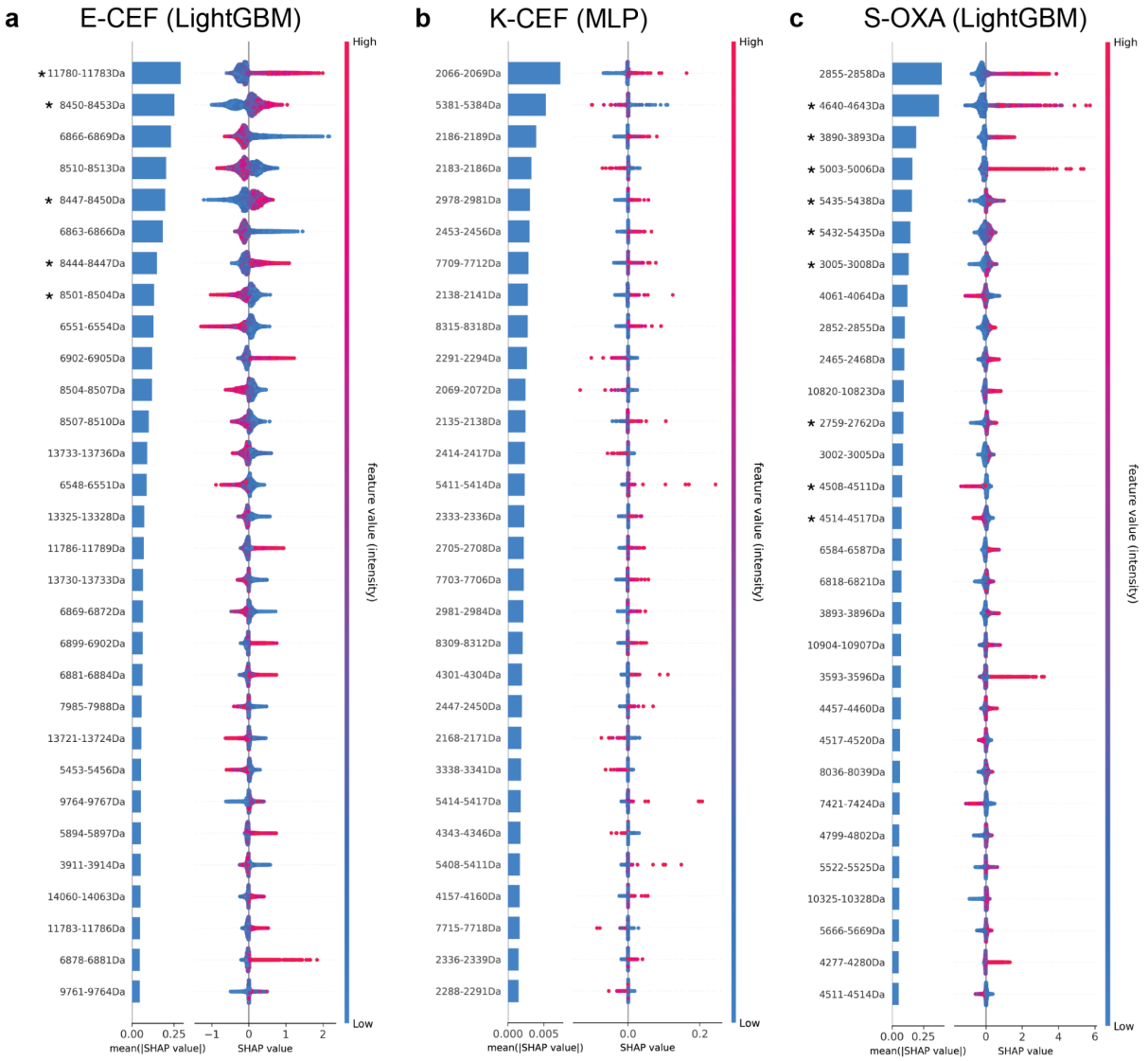

Figure 5: Quantification of feature impact on prediction through analysis of Shapley additive explanations (SHAP) values of the 30 most impactful features. Scenarios are **A.** E-CEF (LightGBM) **B.** K-CEF (MLP) **C.** S-OXA (LightGBM). For each scenario, a barplot on the left indicates the mean Shapley value, i.e. the average impact of each feature on the model output magnitude. The scatterplot on the right indicates the distribution of Shapley values, and their impact on the model output, over all test samples. The colours of each test spectrum (according to the colorbar: blue for low feature values and red for high feature values) indicates the feature value, i.e. the intensity value of the respective feature in the spectrum. The scenario abbreviations follow **Suppl. Tab. 3.** The asterisks marked feature bins containing a previously identified protein peak listed in **Suppl. Tab. 5**.

Retrospective clinical case study

In order to evaluate the clinical benefit of our classifier, we evaluated the antibiotic therapy of patients represented in *DRIAMS-A,* with invasive serious bacterial infections treated between May and August 2018. We reviewed 416 clinical cases that included positive cultures with *E. coli*, *K. pneumoniae,* or *S. aureus* from either blood culture or deep tissue samples. For 63 of

these cases, an infectious diseases specialist (hereafter referred to as clinician) was consulted regarding the antibiotic treatment. Consultation occurred between the species being identified and before the phenotypic antibiotic resistance testing was available (**Extended Data Fig. 3**). For each case, we retrospectively reviewed the recommendations and assessed whether an alternative antibiotic therapy would have been suggested if our classifier had been employed at the time at which the MALDI-TOF mass spectrum was acquired.

In 54 out of 63 clinical cases, the employment of the algorithm would not have changed the suggested antibiotic treatment: in 22 cases the clinician suggested de-escalation of the antibiotic regimen to a more-narrow spectrum antibiotic, in 25 cases to continue the current antibiotic regimen, and in seven cases to escalate the antibiotic treatment to a broader spectrum antibiotic. The classifier reported an accurate prediction of the antibiotic resistance in 51 of these 54 cases, but as the decision on antibiotic treatment can be influenced by multiple factors other than the antibiotic resistance of one bacterial species against one antibiotic agent, such as allergy, these did not change the suggested therapy (**Fig. 6**). In three cases, our algorithm predicted susceptibility, where phenotypic testing revealed resistance to antibiotics. In none of these three cases however, would this incorrect prediction have led to a less effective treatment than suggested without the algorithm: In two of these cases a known MRSA colonization of the patient would have been considered by the clinician, regardless of the prediction of the algorithm. In the third case, *K. pneumoniae* and *E. coli* were both identified in blood culture samples. As there were no indications implying antibiotic resistance, the clinician would have suggested to keep the current antibiotic treatment against *E. coli* with or without the use of the algorithm, and escalation to a broader spectrum antibiotic was only implemented after phenotypic testing.

For nine cases an alternative antibiotic therapy would have been suggested by the clinician with the employment of the classifier at the time of species identification: in seven cases, the classifier would have led to a de-escalation of the antibiotic therapy, while in one case, the employment of the algorithm would have changed the suggested treatment to continue the current antibiotic therapy, where the clinician suggested to escalate to an antibiotic agent with a broader spectrum without the employment of the classifier (**Fig. 6**), and while in one single case, the employment of the algorithm would have led to an unnecessary escalation of the antibiotic therapy due to a false resistance prediction. In summary, for eight out of these nine cases (89%) where the employment of the algorithm would have changed the empiric antibiotic regimen, did the classifier correctly predict susceptibility and this change would have been beneficial and would have promoted antibiotic stewardship, whereas in one case, the wrongly predicted resistance would have led to an unneeded escalation of the antibiotic therapy.

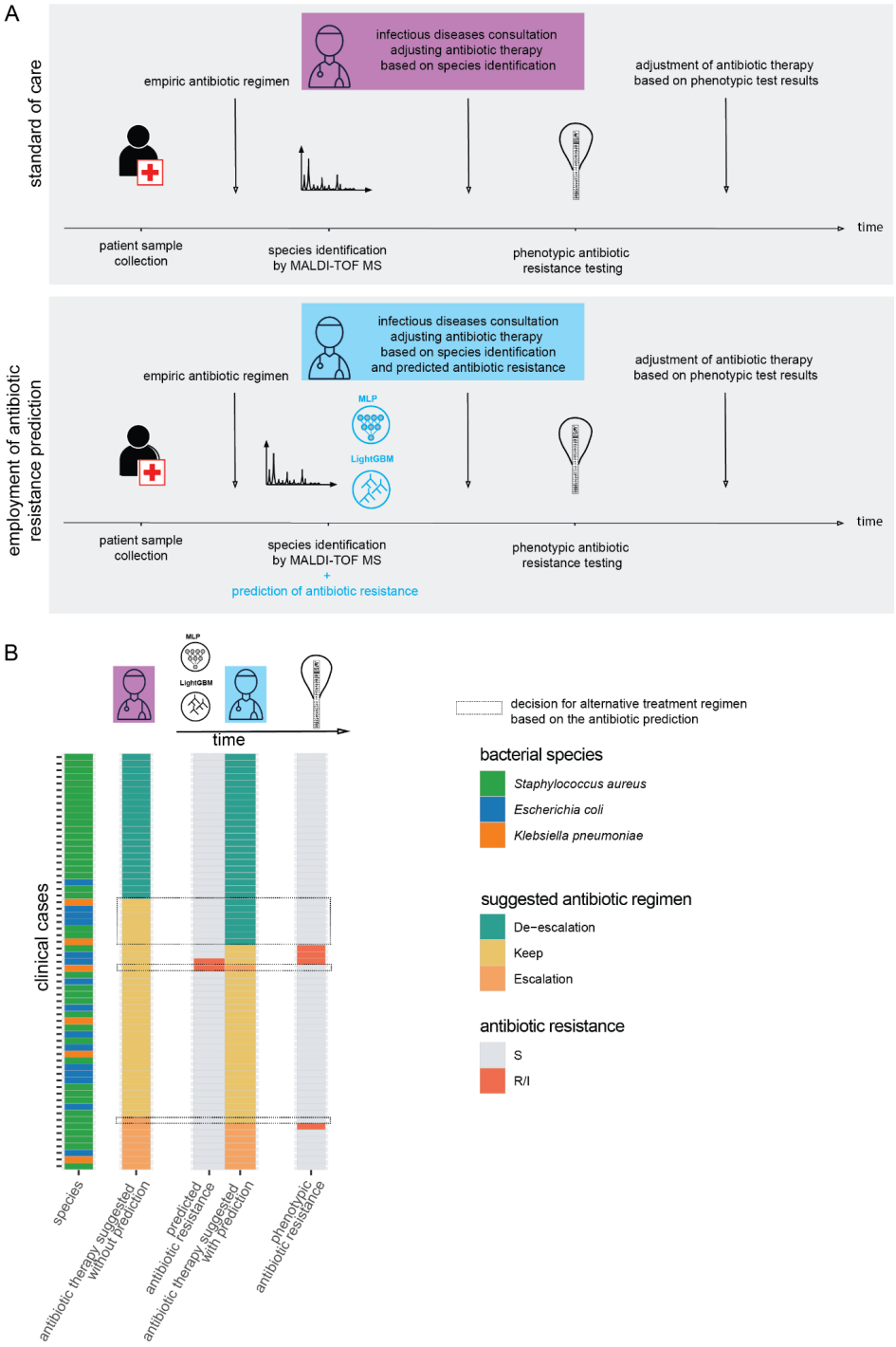

Figure 6: Retrospective clinical case study including 63 cases with invasive bacterial infection. (A) schematic representation of the current standard of care (top row) and the possible employment of our classifiers in the clinical workflow (bottom row). (B) We evaluated the antibiotic regime suggested by a clinician without the employment of the classifier (column 2), which antibiotic resistance the classifier predicted (column 3), the antibiotic treatment suggested considering the predicted antibiotic resistance (column 4) and the phenotypically tested antibiotic resistance. The dashed boxes highlight where the employment of the classifiers would have led to an alternative antibiotic treatment suggestion. ‘de-escalation’: change antibiotic regimen to a more narrow spectrum antibiotic agent, ‘keep’: continue the current antibiotic regimen, ‘escalate’: change antibiotic regime to a broader antibiotic regimen.

Discussion

We have demonstrated that MALDI-TOF mass spectra based antimicrobial resistance prediction from routine diagnostic clinical samples is capable of providing accurate predictions within 24 hours of sample collection. This analysis was made possible by collecting the largest real-world clinical dataset of MALDI-TOF mass spectra and corresponding antimicrobial resistance phenotypes. Overall, we observed high predictive performance using calibrated LightGBM and MLP classifiers trained on individual species–antibiotic combinations, such as ceftriaxone resistance in *E. coli* and *K. pneumoniae* and oxacillin resistance in *S. aureus*, obtaining AUROC values larger than 0.70.

We found that the performance of classifiers trained on mass spectra from one site is not generalisable to mass spectra measured at other sites. This may be influenced by many sources, including (i) different phylogenetic strains, (ii) different prevalence of resistance (i.e. different class ratios), which can impact predictive performance, or (iii) technical variability (97), owing to different machine-specific parameters and settings (i.e. 'batch effects'). Similarly, the closer the time of collection of the training samples is to the time of prediction, the better the predictive power of the trained classifier, likely owing to the same aforementioned reasons. Hence, we would recommend that a clinically-applied classifier should be retrained regularly with the most recent data, originating from its deployment site. In clinical practice such an algorithm may require regular re-certification. Nonetheless, for individual specific species–antibiotic scenarios, our results suggest that even small sample sizes can lead to high predictive performance.

We demonstrate that in order to obtain a classifier at a site with a smaller training dataset, combining the available data with an external dataset, such as *DRIAMS-A,* can increase the training performance). Combining training datasets from different sites increases the sample size, and potentially the coverage of rarer bacterial strains, which improves the predictive performance. However, combining training data originating from different sites also increases the variance in the data, which has the potential to decrease predictive performance. Merging training datasets did not lead to an increased performance on the *DRIAMS-D* test data, indicating its outpatient sample pool creates a dataset dissimilar to the hospital datasets. These results motivate the potential of large-scale MALDI-TOF MS clinical routine dataset acquisition for antimicrobial resistance prediction—combining large datasets could increase the predictive performance on either prediction site. Furthermore, it is worth noticing that all collection sites contributing data to this study are located in a low endemic area for ESBL producing and MRSA bacteria. Future analysis should assess how data from healthcare centers with a higher burden of antibiotic resistant bacteria influence the performance of our classifiers.

We found the predictive performance of classifiers trained on a single species to be higher than that of classifiers trained on multiple species, indicating the higher complexity of predicting multiple resistance mechanisms. This, combined with the general trend of improved performance if more samples are available, indicates the potential benefits of having access to a large database of MALDI-TOF mass spectra. However, many proteins causing resistance are beyond the effective mass range of MALDI-TOF mass spectra. For example, the penicillin-binding protein in *S. aureus* has a mass of approximately 76,400 Da (361), beta-lactamases in *E. coli* and *K. pneumoniae* weigh approximately 30,000 Da (362–365), and the *E. coli* outer membrane porin OmpC weighs approx. 40,300 Da (366). Therefore, we hypothesise that our predictor detects resistance-associated changes in the proteome as well as phylogenetic similarity between resistant vs. susceptible samples. Our results also give an indication at what sample size most information/variance of samples originating from the *DRIAMS-A* collection site are covered by the training data, with sample sizes from 2,500 to 5,000 being required to reach the 'plateau'. We therefore suggest collecting a dataset of at least 2,500 samples when working on MALDI-TOF MS based antimicrobial resistance prediction.

While the antimicrobial resistance classifiers, i.e. LightGBM and the MLP, are trained and predict resistance labels as a black-box system, analysing the contribution of each feature bin to the predictive outcome is of utmost importance to explain the antimicrobial resistance predictor decision-making process in a manner that can be interpreted by the user. We therefore determined the feature importance and the Shapley values of each feature bin and compared the results of the highest-weighted bins to known resistance associated peaks from the literature. The Shapley values indicate that very high or very low feature bin values (corresponding to the presence or absence of a MALDI-TOF mass peak) contribute to the prediction outcome, rather than variations in the feature bin magnitude. This is in line with prior knowledge on MALDI-TOF MS; confirming that the detection of proteins is responsible for the predictive power, rather than confounding signals or noise.

The literature reference comparison confirms the discriminatory potential of single feature bins, contributing substantially to our classifiers and also highlighting their generalisability, as the spectra for these studies were acquired from independent strain collections and on different MALDI-TOF MS devices. Moreover, our classifiers use many more feature bins, for which the discriminatory potential has not previously been identified. An investigation of the protein identity of these yet unknown discriminatory feature bins and their occurrence throughout the respective species would be desirable in the future.

Our retrospective clinical case study shows that our classifier might have a beneficial impact on patient treatment and promote antibiotic stewardship. In 51 out of 63 cases, the algorithm supported the treatment regimen suggested by the clinician. In three cases, the inaccurate prediction by our classifier would not have changed the suggested antibiotic regime, since the

decision is influenced by multiple other factors in addition to the resistance profile towards one antibiotic such as (i) allergies of the patient, (ii) other bacterial species involved in the infection, (iii) patient history including the antibiotic profile of previous isolates, (iv) type of administration of the antibiotic agent. In eight out of 63 cases the accurate prediction by our algorithm would have led to an earlier streamlining of the antibiotic regimen to a more narrow spectrum antibiotic agent. Similar benefits to antibiotic stewardship have been observed when using genotypic assays such as rapid PCR assays (367). These findings exemplify the potential of classifiers to optimize antibiotic treatment and assist antibiotic stewardship efforts using real clinical cases. The evaluation of our classifier in prospective clinical studies, on multiple sites with different prevalence of antimicrobial resistant bacteria, will be necessary to fully evaluate its clinical impact. While the prediction of resistance alone would not be used, the prediction may support clinical decision-making that also considers additional patient-related factors.

In summary, our work demonstrates that MALDI-TOF MS based machine learning can provide novel ways to predict antimicrobial resistance in clinically highly-relevant scenarios. The results demonstrate the benefit of large sample sizes on predictive performance. Further work could build upon these findings and leverage unlabelled (no antimicrobial resistance profile available) MALDI-TOF mass spectra in *DRIAMS* for pre-training a classifier in a semi-supervised fashion before fine-tuning the model on the labelled dataset. In addition to potentially improving the prediction performance, such a training setup could result in a transfer learning scenario to mitigate batch-effects between different collection sites. While these idiosyncratic challenges need to be overcome, there is also a large potential to improve patient treatment.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/
s41591-021-01619-9.

Acknowledgements

We thank Olivia Grüninger, Josiane Reist, Yoann Mahiddine, Daniela Lang, Clarisse Straub, and Magdalena Schneider for excellent technical assistance in accessing the mass spectra and antimicrobial resistance data at the University Hospital Basel. We further thank Belén Rodríguez-Sánchez (Hospital General Universitario Gregorio Marañón, Madrid, Spain), Helena Seth-Smith (USB), Carole Kaufmann (USB), and Valentin Pflüger (MabritecAG,

Riehen, Switzerland) for valuable advice, technical consulting, and feedback throughout the project. We thank Xiao He (ETHZ) for fruitful discussions and we thank Dexiong Chen (ETHZ) and Jacques Schrenzel (HUG) for proofreading the manuscript and providing valuable feedback.

Author Contributions

C.W., B.R., K.B. designed machine learning experiments; C.W., B.R. implemented all experiments of the machine learning analysis; A.E., A.C., K.K.S. organised data collection; A.C., S.G., O.U., C.L., M.O. extracted clinical data; A.C. and M.O. performed retrospective clinical case study; A.C, C.W. implemented pre-processing of datasets *DRIAMS-A*, *DRIAMS-B*; C.W. implemented pre-processing of datasets *DRIAMS-C* and *DRIAMS-D*; M.O., A.E. provided feedback on clinical implications of resistance predictions; C.W., B.R., A.C. designed all display items; C.W., B.R., A.C., K.B., A.E. wrote the manuscript with the assistance and feedback of all the other co-authors. K.B., A.E. conceived and supervised the study.

Competing Interests statement

The authors declare no competing interests.

Methods

Ethical approval

This study received ethical approval through the local ethical review board (EKNZ number: IEC 2019-00729, 2019-01860 and 2019-00748).

Data availability statement

The full datasets generated during and analysed during the current study are available in the Dryad repository, https://doi.org/10.5061/dryad.bzkh1899q.

Code availability statement

All R and Python scripts can be found in https://github.com/BorgwardtLab/maldi_amr under a BSD 3-Clause License.

MALDI-TOF mass spectra acquisition and antimicrobial resistance testing

We collected data from daily clinical routine at ISO/IEC 17025 accredited diagnostic routine laboratories. The study was evaluated by the local ethical committee (IEC 2019-00729). All data used for the machine learning analysis was deidentified prior to analysis. Specifically, all MALDI-TOF mass spectra contained in *DRIAMS-A* to *-D* were acquired at four microbiological

laboratories in Switzerland providing routine diagnostic services for hospitals and private practices. All laboratories use the Microflex Biotyper System by Bruker Daltonics (Bremen, Germany), which is a widely-employed MALDI-TOF MS system in microbiological routine diagnostics both in North America[16] and in Europe[17,18]. The four diagnostic laboratories included in this study are (1) University Hospital Basel-Stadt (*DRIAMS-A*), (2) Canton Hospital Basel-Land (*DRIAMS-B*), (3) Canton Hospital Aarau (*DRIAMS-C*), and (4) laboratory service provider Viollier (*DRIAMS-D*). While Canton Hospitals Basel-Land and Aarau employ the Microflex Biotyper LT/SH System, Viollier uses the Microflex smart LS System. Although these two systems differ in their respective laser gas, they use the same reference spectra database, so we included spectra of both Microflex Biotyper systems. University Hospital Basel-Stadt uses the two Microflex Biotyper systems in parallel. The species of each mass spectrum was identified using the Microflex Biotyper Database (MBT 7854 MSP Library, BDAL V8.0.0.0_7311-7854 (RUO)) included in the flexControl Software (Bruker Daltonics flexControl v.3.4). Similar to the mass spectra, antimicrobial resistance profiles were routinely acquired in the same four microbiological laboratories within the same time frames of the dataset. Resistance categories for bacteria values were determined either using microdilution assays (VITEK® 2, BioMérieux, Marcy-l'Étoile, France), or by minimal inhibitory concentration (MIC) stripe tests (Liofilchem, Roseto degli Abruzzi, Italy), or disc diffusion tests (ThermoFisher Scientific, Waltham, USA). Resistance categories for yeast were determined by using Sensititre Yeast One (Thermofisher). All breakpoint measurements were interpreted to be either susceptible, intermediate, or resistant according to EUCAST and CLSI (2015 M45; 2017 M60) recommendations. The EUCAST versions used were updated with every EUCAST (368) Breakpoints table update and include v6-v8.

Quality control

Empty spectra and calibration spectra were excluded from further analysis. This serves to ensure a similar level of data quality for the different sites.

Matching of MALDI-TOF mass spectra and antimicrobial resistance profiles

MALDI-TOF MS based antimicrobial resistance prediction requires a dataset containing mass spectra and their corresponding resistance labels, in the form of antimicrobial resistance profiles. In order to construct such a dataset, MALDI-TOF MS and resistance profile measurements belonging to the same microbial isolate have to be matched. Since each site in *DRIAMS* stores the mass spectra and their corresponding antimicrobial resistance profiles in separate databases, a matching procedure has to be developed for each site.

We use the term 'laboratory report' for the document used to report laboratory measurement results, including antimicrobial resistance profiles, for each patient within the clinical care. The

species of the specimen is obtained through Bruker Microflex MALDI-TOF MS and added to the laboratory report. This decouples laboratory report entry and the mass spectrum; there is no link required between the spectrum file and the laboratory entry after the species is entered. The antimicrobial resistance profiles obtained in their individual experiments are also added to the laboratory report. The laboratory report entries are commonly identified by codes linking them to a patient, or a unique sample taken from a patient, to which we refer as 'sample ID'. Multiple entries with the same sample ID can exist if several probes were taken from the same patient or several colonies tested from the same probe.

In general, the spectra recorded by the Bruker Microflex systems were labelled with an ambiguous, i.e. non-unique, code corresponding to the non-unique sample ID in the laboratory report. MALDI-TOF mass spectra and their corresponding antimicrobial resistance profiles were stored in separate files. In the clinic, MALDI-TOF MS spectra are never intended to be matched up with the laboratory report entries; therefore, no proper protocols for matching exist. Matching protocols had to be developed uniquely and in an ad-hoc fashion for each labelling system at each institution.

In order to link mass spectra to their antimicrobial resistance profiles, we constructed a *unique* identifier, using the sample ID and the determined genus of a sample. The rationale behind this strategy is that if multiple sample ID entries exist, this is most likely due to multiple genera being present in the patient samples, leading to several measurements. We omitted samples for which we were unable to construct a unique sample ID–genus pair.

Mass spectra were stored without information on the determined species. Hence, for each spectrum, the species and genus label is determined by re-analysing the spectra with the University Hospital Basel-Stadt Bruker library and then matching the spectrum to its corresponding antimicrobial resistance profile using the assigned sample ID and the determined genus. All MALDI-TOF MS systems used in this study were maintained according to the manufacturer's standard and spectra were routinely acquired using the 'AutoXecute' acquisition mode. The genus is used (instead of species), as it allows for some flexibility between the species assigned to a sample in the laboratory report and the Microflex Biotyper. The species label given in the laboratory report can differ from the species assigned to the corresponding MALDI-TOF mass spectrum by the Microflex Biotyper System as additional microbiological tests can give a more accurate label. In what follows, we provide additional details regarding the matching procedure which are specific to each site.

University Hospital Basel-Stadt. Starting in 2015, the spectra were labelled with a 36-position code by the Bruker machine (e.g. 022b130c-6c8c-49b5-814d-c1ea8b2e7f93), which we term 'Bruker ID'. This code is guaranteed to be unique for all spectra labelled from one machine. Each AMR profile is labelled with a 6-digit sample ID, which is unique for samples in one year. Antimicrobial resistance profiles were collected using the laboratory information system. The

laboratory information system includes all entries made for a sample, also entries which have later been corrected and have not been reported nor considered for patient treatment. As such manual corrections are very rare, the uncertainty in antimicrobial resistance labels is limited. For each year (2015, 2016, 2017, and 2018) there are separate antimicrobial resistance profile tables and folders containing all spectrum samples collected during the corresponding year. We lost 40,569 spectra out of 186,098 by following the aforementioned pre-processing routines (*DRIAMS-A*).

Canton Hospital Basel-Land. The antimicrobial resistance profiles and mass spectra are each labelled with a 6-digit sample ID. The genus depicted in each mass spectrum was determined through comparison to the Microflex Biotyper Database (Bruker Daltonics flexControl v.3.4); the genus of each antimicrobial resistance profile was stated in the laboratory report. Mass spectra and antimicrobial resistance profiles were merged using the 6-digit sample ID and the genus information.

Canton Hospital Aarau. Here, the laboratory report contains the 10-digit sample ID, species label, and antimicrobial resistance profiles of measured samples. This software version did not provide a unique 36-character code for each spectrum, but only a 10-digit sample ID that had to be used to match spectra to the antimicrobial resistance profiles from the laboratory. Since the sample ID can be shared by different spectra, it cannot be used to uniquely match a species label to an input spectrum. To circumvent this problem, we divided the spectra in 15 batches, each one only containing unique 10-digit sample IDs. Repeated sample IDs were distributed over the batches. These 15 batches were re-analysed and labelled by the Bruker software, and 15 output files with the given species labelled were created. Through the separation in batches, the certain species label was determined for each spectrum. The label for each spectrum in the batches can be determined, as we only included spectra that already had a label in the lab file. Now, each spectrum file has a combined label made up of its 10-digit sample ID and its species label. If this combined label was found to have a unique match within the lab results file, the AMR profile was assigned to the spectrum, otherwise its antimicrobial resistance profile position remained empty and only the spectrum with its species label was added to the dataset. We ignore all spectra that could not be matched to an entry in the lab results file (such spectra arise from measurements that do not provide AMR information).

Viollier. While all other sites reported AMR labels with either ‘R’, ‘S’, ‘I’, ‘positive’ or ‘negative’ values, samples provided by Viollier are labelled with precursory measurements, namely the minimal inhibitory concentration (MIC) of each antibiotic. We therefore use the breakpoints given the up-to-date EUCAST guidelines (v.9) to convert the MIC values to ‘RSI’ values.

80,796 spectra in the fid file format are present, identified again through a unique 36-character ‘Bruker ID’. The antimicrobial resistance results are identified by a 10-digit sample ID, which

are linked to the Bruker IDs in an additional file, the 'linking file'. The main reasons for loss of data in pre-processing are (1) the antimicrobial resistance results and ID 'linking files' contained significantly fewer entries than fid files present (40,571 and 51,177 respectively) and (2) following advice by the lab personnel, only the 10-digit sample ID could be used for matching to the BrukerID (which contained a longer version of the LabID). Through exclusion of all entries without a unique 10-digit sample ID in both the antimicrobial resistance results and linking files, another significant portion of data was lost. Specifically, there is an overlap of 10,852 filtered entries from the laboratory report file and the linking file. After matching these entries with spectra, 7,771 spectra with 7,720 antimicrobial resistance profiles remained. Spectra without an antimicrobial resistance profile are not used for any supervised learning tasks (such as prediction).

Hospital Hygiene

The hospital hygiene department specifically screens for multidrug-resistant pathogens in order to take actions which prevent nosocomial transmission of these. These samples are cultured primarily on selective media containing antibiotics, enabling the growth of resistant strains only.

Growth media have an impact on the bacteria's proteome and thereby on the MALDI-TOF MS spectrum (369). In order to avoid that our classifiers recognise media specific characteristics in the MALDI-TOF mass spectra from the selective media instead of media independent signatures of non-susceptible bacterial strains, we excluded samples that were collected for the hospital hygiene department from *DRIAMS-A* for further analysis. The individual sample sizes per workstation and their predictive performance from MALDI-TOF mass spectra is given in **Suppl. Tab. 6**.

Patient case identification

For *DRIAMS-A*, a *clinical cas*e was defined as a unique hospital stay, i.e. the timeframe between the hospital entry and exit of a patient. If a patient was treated at the hospital in 2015 and again in 2018, these were defined as two separate cases. For the retrospective clinical analysis, infections with different bacterial species and different patient isolation materials during the same hospital stay were regarded as different entities, as different species might require different antibiotic therapies.

For *DRIAMS-B*, *DRIAMS-C* and *DRIAMS-D*, no information regarding clinical cases was provided and therefore not considered during analysis.

Dataset characteristics

All medical institutions are located in Switzerland. Microbial samples in the University Hospital Basel-Stadt database (i.e. *DRIAMS-A*) mostly originate from patients located in the city of Basel and its surroundings. Such patients visit the hospital for either out- or inpatient treatment. Samples in the Canton Hospital Basel-Land dataset (i.e. *DRIAMS-B*) primarily originate from the town surrounding the City of Basel. Patients from the Swiss Canton Aargau seek medical care at the Canton Hospital Aarau (*DRIAMS-C*). Viollier (*DRIAMS-D*) is a service provider that performs species identification for microbial samples collected in medical practices and hospitals. Samples originate from private practices and hospitals all over Switzerland.

DRIAMS-A to *-D* are datasets that contain data collected in the daily clinical routine. All mass spectra measured in a certain time frame are included. The time frame during which each dataset was collected are as follows:

DRIAMS-A: 34 months (11/2015–08/2018)
DRIAMS-B: 6 months (01/2018–06/2018)
DRIAMS-C: 8 months (01/2018–08/2018)
DRIAMS-D: 6 months (01/2018–06/2018)

Spectral representation

In the *DRIAMS* dataset, we include mass spectra in their raw version without any pre-processing, and binned with several bin sizes. After initial analyses, a bin size of 3 Da was used for all machine learning analyses in this study. This bin size is small enough to allow for separation of mass peaks (for which the exact mass-to-charge position can vary slightly due to measurement noise), while large enough not to impede computational tractability. The spectra are extracted from the Bruker Flex machine in the Bruker Flex data format. The following pre-processing steps are performed using the R package `MaldiQuant` (252) version 1.19: (1) the measured intensity is transformed with a square-root method to stabilize the variance, (2) smoothing using the Savitzky-Golay algorithm with half-window-size 10 is applied, (3) an estimate of the baseline is removed in 20 iterations of the SNIP algorithm, (4) the intensity is calibrated using the total-ion-current (TIC), and (5) the spectra are trimmed to values in a 2,000 to 20,000 Da range. For exact parameter values, please refer to the code.

After pre-processing, each spectrum is represented by a set of measurements, each of them described by its corresponding mass-to-charge ratio and intensity. However, this representation results in each sample having potentially a different dimensionality (i.e. cardinality) and different measurements being generally irregularly-spaced. Since the machine learning methods used in this manuscript require their input to be a feature vector of fixed dimensionality, intensity measurements are binned using the bin size of 3 Da. To perform the

binning, we partition the m/z axis in the range from 2,000 to 20,000 Da into disjoint, equal-sized bins and sum the intensity of all measurements in the sample (i.e. a spectrum) falling into the same bin. Thus, each sample is represented by a vector of fixed dimensionality, i.e. a vector containing 6,000 features, which is the number of bins the m/z axis is partitioned into. We use this feature vector representation for all downstream machine learning tasks.

Antimicrobial resistance phenotype binarization

For the machine learning analysis, the values of antimicrobial resistance profiles were binarised during data input to have a binary classification scenario. The categories are based on EUCAST and CLSI recommendations. For tests that report RSI values, resistant (R) and intermediate (I) samples were labelled as class 1, while susceptible (S) samples were labelled as class 0. We grouped samples in the intermediate class together with resistant samples, as both types of samples prevent the application of the antibiotic. In EUCAST v6-v8, the intermediate category shows higher MIC values but due to safety reasons in clinical practice this was usually counted to resistance in order to have a safety buffer in reaching sufficiently high antibiotic drug concentrations.

Statistical methods

If not otherwise indicated, solid lines and performance metrics displayed in figures and tables refer to the mean performance over the test sets of a 5-fold cross-validation. Shaded areas and numbers added with ± signs refer to the standard deviation across the respective evaluation metric.

The center line of the box plot in **Extended Data Fig. 1** shows the median, and the lower and upper limit show Q1 and Q3 respectively, where Q1 is the first quartile and Q3 is the third quartile. The lower whisker shows the lowest time requirement until diagnostic result, and the upper whisker shows the largest time recorded (excluding outliers).

The bars in **Extended Data Fig. 2** state the mean performance over the test sets of a 10-fold cross-validation. The asterisks marking the bars indicate a statistically significant difference between the reported metrics between all species and species information alone of a two-sided Welch's t-test, without assuming equal population variance, at a significance level of 0.05.

Machine learning methods

For AMR classification, we used a set of state-of-the-art classification algorithms with different capabilities. It included (1) logistic regression, (2) LightGBM (370), a modern variant of gradient-boosted decision trees, and (3) a multi-layer perceptron deep neural network (MLP). For LightGBM, we use the official implementation in the `lightgbm` package, while we use the

scikit-learn package for all other models (371). These models cover a large spectrum of modern machine learning techniques, with logistic regression representing an algorithm from classic statistics, whose training process can be regularised. LightGBM, by contrast, represents a modern variant of tree-based learning algorithms, focussing specifically on good scalability properties while maintaining high accuracy. Finally, MLPs constitute a simple example of deep learning algorithms. While they have the highest complexity in terms of compute resources and data requirements than the aforementioned models, deep learning methods can be effective in uncovering complex relationships between input variables.

For each antibiotic, all samples with a missing AMR profile were removed and the machine learning pipeline was applied to the reduced dataset. Samples were randomly split into a training dataset comprising 80% of the samples and a test dataset with the remaining 20%, while stratifying for the class and the species, and ensuring that a sample with a specific patient case is either part of the train dataset, or the test dataset, but not both. This step ensures that sample measurements of the same infection (that are likely very similar to each other) are not causing information leakage from train to test. This is slightly unusual in standard machine learning setups, which typically only require stratification by a single class label, but crucial for our scenarios to guarantee similar prevalence values. To select an appropriate model configuration for a specific task, we employ 5-fold cross-validation on the training data; in case an insufficient number of samples is available, our implementation falls back to a 3-fold cross-validation on the training dataset to optimize the respective hyperparameters. The hyperparameters are model-specific (see below for more details), but always include the choice of an optional standardisation step (in which feature vectors are transformed to have zero mean and unit variance). To determine the best-performing hyperparameter set, we optimized the area under the ROC curve (AUROC) on the training dataset only. This metric is advantageous in our scenario, as it is not influenced by the class ratio and summarises the performance of correct and incorrect susceptibility predictions over varying classification score thresholds. Having selected the best hyperparameters, we retrain each model on the full training dataset, and use the resulting classifier for all subsequent predictions. Our hyperparameter grid is extensive, comprising, for example, the choice of different logistic regression penalties (L1, L2, no penalty), the choice of scaling method (standardisation or none), and regularisation parameters ($C \in \{10^{-3}, 10^{-2}, \cdots 10^{2}, 10^{3}\}$). For more details, please refer to our code (`models.py`).

We implemented all models in Python and published them in a single package (https://github.com/BorgwardtLab/maldi_amr), which we modelled after scikit-learn, a powerful library for machine learning in Python.

Evaluation metrics

We report AUROC as the main metric of performance evaluation. The datasets of most antibiotics under consideration exhibit a high class imbalance (20 out of 42 antibiotics show a resistant/intermediate class ratio <20% or >80%). AUROC is invariant to the class ratio of the dataset and therefore permits a certain level of comparability between antibiotics with different class ratios. A pitfall of reporting AUROC in the case of unbalanced datasets, however, is that it does not reflect the performance with respect to precision (or positive predictive value). Therefore, the AUROC can be high while precision is low. To account for this bias, we additionally report the area under the precision-recall curve (AUPRC); this metric is not used during the training process, though.

Two other metrics commonly used in clinical research are sensitivity and specificity. Analogous to the ROC curves, we show sensitivity vs. specificity curves to illustrate the tradeoff between both metrics. Please note commonalities to other metrics: *sensitivity*, *recall* and *true positive rate* are synonyms and all correspond to the same metric; specificity is a counterpart to the false positive rate, that is, $true\ positive\ rate = 1 - specificity$.

Connection to confusion matrix All of the metrics we employ here can be derived from the counts within a confusion matrix.

The area under the receiver operating characteristic (AUROC) shows the true positive rate (TP/TP+FN) against the false positive rate (FP/FP+TN). The AUPRC, as well as the AUROC, is traditionally reported on the minority class. In our scenario, however, while the minority class is the resistant class in most cases, this is not consistent and for some antibiotics more samples of the resistant will be present. The precision-recall curve shows the recall (TP/TP+FN) against the precision (TP/TP+FP). The average performance of a random classifier would be 0.5 for AUROC and percentage of samples of the positive (susceptible) class for AUPRC.

Shapley values for interpretability analysis

In order to improve the interpretability of our classifiers, we calculated Shapley values using the shap package. This package directly supports the explanation of many common machine learning techniques. We used the standard algorithms of the shap package to explain the outputs of our logistic regression and LightGBM models. For the MLP, the use of gradient-based explanation techniques turned out to be impossible because of the large memory requirements of the algorithms. We therefore opted to follow common practice and subsample the input data set, reducing it to 50 barycentres, i.e. samples that express most of the variability in the data, via k-means clustering. This enabled us to obtain per-sample Shapley values that contain the relevance of individual features with respect to the overall output of the model.

Funding

This study was supported by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung (K.B.), and the D-BSSE-Uni-Basel Personalised Medicine grant (PMB-03-17, K.B. & A.E.) and a Doc.Mobility fellowship (A.C.) by the Swiss National Science Foundation (P1BSP3-184342).

Additional Publications not directly linked to this thesis

I have contributed to the following publications and projects as a co-author, but these are not directly related to the overall focus of the PhD thesis.

1. Veronika Muigg, **Aline Cuénod**, Srinithi Purushothaman, Martin Siegemund, Matthias Wittwer, Valentin Pflüger, Kristina Schmidt, Maja Weisser, Nicole Ritz, Andreas Widmer, Daniel Goldenberger, Vladimira Hinic, Tim Roloff, Kirstine K. Søgaard, Adrian Egli and Helena M.B. Seth-Smith. "Diagnostic challenges within the *Bacillus cereus*-group: finding the beast without teeth" *In review* in *'New Microbes and new Infection'*

My contributions:

- Analysis of *Bacillus* spp. MALDI-TOF mass spectra using different databases (Table 1)

2. Goldenberger, Daniel, Kirstine K. Søgaard, **Aline Cuénod**, Helena Seth-Smith, Daniel de Menezes, Peter Vandamme, and Adrian Egli. "Cutibacterium Modestum and 'Propionibacterium Humerusii' Represent the Same Species That Is Commonly Misidentified as Cutibacterium Acnes." *Antonie Van Leeuwenhoek*, (May 7, 2021). https://doi.org/10.1007/s10482-021-01589-5.

My contributions:

- Analysis and visualisation MALDI-TOF mass spectra (Figure S1)

3. Goldenberger, Daniel, Karoline Leuzinger, Kirstine K. Sogaard, Rainer Gosert, Tim Roloff, Klaudia Naegele, **Aline Cuénod**, Alfredo Mari, Helena Seth-Smith, Katharina Rentsch, Vladimira Hinić, Hans H. Hirsch, Adrian Egli. "Brief Validation of the Novel GeneXpert Xpress SARS-CoV-2 PCR Assay." *Journal of Virological Methods* 284 (October 1, 2020): 113925. https://doi.org/10.1016/j.jviromet.2020.113925.

My contributions:

- Processing of SARS-CoV-2 samples for measurements on the Cobas system and GeneXpert system (Table 1)

4. Weis, Caroline*, Max Horn*, Bastian Rieck*, **Aline Cuénod**, Adrian Egli, and Karsten Borgwardt. "Topological and Kernel-Based Microbial Phenotype Prediction from

MALDI-TOF Mass Spectra." *Bioinformatics* 36, (July 1, 2020): i30–38. https://doi.org/10.1093/bioinformatics/btaa429.

My contributions:

- Data collection and curation of MALDI-TOF mass spectra and AMR profiles (Table 1)

5. Wüthrich, Daniel, **Aline Cuénod**, Vladimira Hinić, Mario Morgenstern, Nina Khanna, Adrian Egli, and Richard Kuehl. "Genomic Characterization of Inpatient Evolution of MRSA Resistant to Daptomycin, Vancomycin and Ceftaroline." *Journal of Antimicrobial Chemotherapy* 74, no. 5 (May 1, 2019): 1452–54. https://doi.org/10.1093/jac/dkz003.

My contributions:

- In vitro passaging of *S. sureus* strain
- Performing and evaluation of MIC test strips
- Visualisation (Figure 1)

6. Helena MB Seth-Smith, Frank Imkamp, Florian Tagini, **Aline Cuénod**, Rico Hömke, Kathleen Jahn, Anne Tschacher, Peter Grendelmeier, Veronika Bättig, Stefan Erb, Miriam Reinhard, Gottfried Rütimann, Sonia Borrell, Sebastian Gagneux, Carlo Casanova, Sara Droz, Michael Osthoff, Michael Tamm, Ulrich Nübel, Gilbert Greub, Peter M. Keller, Adrian Egli. "Discovery and Characterization of Mycobacterium Basiliense Sp. Nov., a Nontuberculous Mycobacterium Isolated From Human Lungs." *Frontiers in Microbiology* 9 (2018): 3184. https://doi.org/10.3389/fmicb.2018.03184.

My contributions:

- Analysis and visualisation MALDI-TOF mass spectra (Figure S1)

7. Vanni Benvenga, **Aline Cuénod**, Srinithi Purushothaman, Gottfried Dasen, Helena Seth-Smith, Tim Roloff, Adrian Egli. "Molecular epidemiological characterization of a Swiss legacy collection of Methicillin resistant *Staphylococcus aureus* (MRSA)" *Manuscript in preparation*

My contributions:

- Supervision of bioinformatics analysis

7 Discussion

7.1 Rapid and accurate identification of clinically relevant *Klebsiella* spp. and *Escherichia coli* in clinical routine diagnostics

Rapid and accurate assessment of infectious diseases, including profiling of antimicrobial resistance, has been perceived as one of the most important diagnostic challenges of the upcoming years (25). The time delay to effective treatment is critical and associated with morbidity, mortality and healthcare costs (49). Many of the most relevant bacterial species causing infections originate from the family *Enterobacteriaceae* (53,157,372). The most prominent *Enterobacteriaceae* spp. are included in the ESKAPE pathogens, which are responsible for most hospital-acquired infections and are recognised as a critical priority for developing new antibiotic treatments (129,373,374). AMR in *K. pneumoniae* and *E. coli* are attributed each to > 250'000 deaths globally in 2019 (25). Within the **Chapters I** and **II** of this thesis, we explored different aspects related to diagnosing the two *Enterobacteriaceae* taxa *Klebsiella* spp. and *E. coli* - including the links to virulence and AMR at single strain levels.

***Klebsiella* spp.**

In **Chapter I**, we focused on the genus *Klebsiella*, which has gained clinical attention due to the rise of globally successful multidrug-resistant (103) and hypervirulent clones (45). In parallel to the increased research interest, ten new species have been described in the last two decades (104–112). Many of these *Klebsiella* spp. remain poorly characterised. Similarly, only a few species are routinely diagnosed in clinical settings due to the low taxonomic resolution of standard identification workflows. Therefore, we studied the genomic and MALDI-TOF mass spectra based diversity within the *Klebsiella* genus (**Chapter I**) to assess (i) whether genomic variations between the species give evidence for species-specific virulence and AMR patterns , (ii) whether these species can be distinguished by MALDI-TOF MS and (iii) whether the *Klebsiella* spp. have distinct clinical phenotypes. We analysed 3,594 publicly available genomes (including 256 strains sequenced as part of the project) and 33,160 MALDI-TOF spectra, from eight healthcare centres with 7,876 matching AMR profiles and patient data of 957 clinical cases from a single healthcare centre (University Hospital Basel) (325).

Our key findings were that (i) from the genomic data, *Klebsiella* spp. have differing AMR and virulence factors, (ii) nine *Klebsiella* spp. and the genus *Raoultella* are distinguishable using MALDI-TOF mass spectra, (iii) all underdiagnosed species occur in a clinical context, (iv)

phenotypically, the *Klebsiella* spp. exhibit varying AMR profiles, and (v) strains of the *K. oxytoca* group were more likely to be involved in invasive infections than strains of the *K. pneumoniae* group (325). A recent study highlights the pathogenic potential of the *K. oxytoca* complex and further reports the existence of three yet undescribed species (375), suggesting that there is more diversity to uncover within this complex. We observed strains of the *K. oxytoca* complex to be less frequently resistant against 4th generation cephalosporins than strains of the *K. pneumoniae* complex, which is in line with data collated by the Swiss surveillance platform ANRESIS for the past three years (376,377). The data collected for our study suggested that strains of the *K. oxytoca* complex were more frequently resistant against penicillins including beta-lactamase inhibitors and 3rd generation cephalosporins. This finding is not reflected in the data collated by ANRESIS (376,377) and requires further investigation. Possible explanations for this discrepancy are variations in AMR (i) over time, (ii) in different geographical areas, and (iii) from different patient populations. In our data and within the respective complexes, *K. michiganensis* was more frequently resistant against penicillins including a beta-lactamase inhibitor, than *K. oxytoca*. *K. pneumoniae* was more frequently resistant against cephalosporins of the 3rd and 4th generation than *K. variicola*. Both findings are in line with the increased occurrence of AMR genes observed in the respective species (151) from international genomic datasets. These findings cannot yet be compared to Swiss surveillance data, as the species within the complexes are not yet distinguished. Our data indicate the importance of monitoring AMR in the context of detailed species resolution. This would allow to rapidly adapt the antibiotic treatment, based on epidemiological risks of specific well-monitored species.

Since our analysis, further species have been described within the genus (105,110,112) and shown to be distinguishable with MALDI-TOF MS and a marker-based approach (74). Applying a marker-based approach to a large, international and standardised dataset, similarly to what we did in our study, would allow future studies to retrospectively assess the occurrence and clinical phenotype of these newly described species. Such analyses are required to examine whether our observation of the increased virulence of the *K. oxytoca* complex holds true in different epidemiological contexts. It is essential to include the identification of the yet underdiagnosed *Klebsiella* spp. into clinical routine diagnostic workflows to prospectively evaluate their clinical relevance.

Overall, we have observed in our study that there is a large genomic diversity in the genus *Klebsiella*. Many of the newly observed species are clinically relevant and cause severe infections. Moreover, we have shown that MALDI-TOF can distinguish several recently described species, which is relevant as they exhibit distinct clinical phenotypes. Going further, we aimed to assess whether we can use MALDI-TOF MS for the early detection of highly virulent strains within a species.

Escherichia coli

In **Chapter II**, we focused on *E. coli*, which encompasses highly pathogenic and clinically relevant lineages (213,378), as well as less virulent lineages adapted to various environments such as freshwater (43) (**Chapter II**). Phylogenetic typing approaches based on MALDI-TOF MS or other typing methods are highly valuable from a public health perspective (295,379). However, the assignment to a phylogenetic lineage does not allow for a definitive assessment of virulence or antimicrobial resistance, as these can differ between closely related strains (380). This is especially true for *Enterobacteriaceae*, where mobile genetic elements harbouring virulence and resistance factors are frequently exchanged (43,381). Accounting for genetic variations within the *E. coli* phylogeny, we aimed to identify which bacterial genes enable uropathogenic *E. coli* (UPEC) strains to cause invasive infections. We, therefore, sequenced whole genomes of 1,079 bacterial strains from 831 clinical cases at the University Hospital Basel. We assessed (i) which bacterial lineages (phylogroups, sequence types (ST)) are enriched; and (ii) which bacterial genetic factors are significantly associated with invasion to the bloodstream. In line with previous studies (44,198), we confirm the association of phylogroups B2, D and F with UTI. Moreover, we observe an association between particular virulence and AMR genes and these deep branching lineages. We observed a high diversity within each phylogroup and found closely related strains exhibiting variable resistance against beta-lactam antibiotics (**Chapter II Figure 1** and **Appendix I Figure S1**). Varying phenotypes between closely related strains have previously been described (206,208) and underline the necessity for phenotypic characterisation or the direct detection of genetic virulence and resistance factors.

We followed up the discovery of different virulent strains by identifying the genomic loci associated with invasive infection applying a bGWAS approach (285). Multiple virulence factors have been described for ExPEC, amongst them iron acquisition systems (382), capsular polysaccharides (290) and adhesion factors, such as pyelonephritis associated (PAP) pili (383). PapG forms the outermost adhesive tip of these PAP pili and binds to human uroepithelial cells (195,384). Five PapG isoforms have been described so far (PapGI - PapGV) (198). We corroborate the importance of *papGII* for uropathogenic *E. coli* (UPEC) to ascend the human urinary tract and cause invasive infection (**Chapter II**). While two previous studies have also identified *papGII* as an important factor for invasive UTI (198,385), neither had the chance to combine extensive clinical data with bacterial whole genome sequences to substantiate the evidence. The combination of bacterial genomic data and patient characteristics is a clear and important new element - clinical phenotypes such as invasiveness should always be evaluated considering the clinical context and adjusting for host related characteristics. My study, therefore, provides further evidence that *papGII* drives

invasive UTI, independent of host characteristics. While our retrospective study (**Chapter II**) fills this important gap, future prospective studies are required to evaluate the value of *papGII* as a diagnostic marker for invasive UTI. Finally, we assessed if MALDI-TOF enables rapid identification of PapGII. Therefore, we screened mass spectra of 317 UPEC strains for PapGII peaks. However, we did not identify PapGII specific peaks in MALDI-TOF mass spectra, likely because the PapGII protein mass lies beyond the mass range of MALDI-TOF MS (293).

Future studies are required to investigate whether machine learning algorithms (**Chapter V**) can identify patterns of intensity changes which distinguish *papGII* positive from *papGII* negative *E. coli* strains. In the meantime, sequence-based diagnostic assays, such as targeted PCR or loop-mediated isothermal amplification assays, might be promising alternative approaches.

Overall, we have been able to see that (i) the distribution of STs and phylogroups was the same between invasive and non-invasive UTI and that (ii) strains encoding *papGII* were significantly more likely to cause invasive infection, independent of important patient characteristics (**Chapter II**).

To understand the clinical relevance of bacterial infections, patient data and disease outcomes are crucial. Our two studies assessing the virulence of two clinically important taxonomic groups, namely, *Klebsiella* spp. (**Chapter I**) and *E. coli* strains (**Chapter II**), have included patient data spanning a broad range of information. This information included age, gender, immunosuppression and comorbidities. Quantifying comorbidities requires extensive medical know-how, and although standardised guidelines exist (386), established indices can be outdated and lag behind medical progress (387,388). Whilst patient characteristics, including comorbidities, are rarely used in other studies, likely because of restricted access, legally and in terms of effort, they are key to identifying the clinical potential of a bacterial strain.

Moreover, to disentangle the interplay between patient characteristics and bacterial virulence factors more precisely, future analysis should focus on specific patient subpopulations, such as patients carrying a permanent urinary catheter or those with Diabetes mellitus. As for our findings on the increased virulence of *K. oxytoca* complex strains, it will further be important to validate the importance of *papGII* in different epidemiological contexts. To better understand the spread of *papGII* in the *E. coli* population, further analyses focussing on the dynamic evolution of the pathogenicity island and its integration into various chromosomal backgrounds are required. Moreover, future studies could elucidate expression levels of PapGII throughout the progression of invasive UTI, clarifying its role in invasiveness.

In conclusion, **Chapters I** and **II** highlight the importance of i) considering patient characteristics in such analyses; ii) increasing the taxonomic resolution in clinical routine diagnostics, potentially identifying resistance and virulence factors; and iii) having highly reproducible and standardised datasets and analysis approaches.

7.2 MALDI-TOF mass spectral quality in routine diagnostics

Standardised and reproducible workflows and datasets across different locations and time are key for every meta-analysis (389). Whilst analysing *Klebsiella* spp. mass spectra acquired at different healthcare centres (**Chapter I**), we observed major differences in MSQ. This was reflected by varying numbers of ribosomal marker masses being detected, which largely determined the resolution of the species identification. This observation prompted us to disentangle MSQ more closely and to enhance overall spectra quality within a multicentre study (**Chapter III**, **Chapter IV**).

Factors of MSQ

In **Chapter III**, we defined a diverse set of 47 clinically relevant bacterial strains, covering 39 species and 15 genera (262). We used various commonly applied sample preparation protocols to acquire mass spectra on two different MALDI-TOF systems and analysed these with a standardised pipeline. A few reference studies have looked at the MSQ (98,312,390–392). These studies were largely missing data acquired on devices from multiple manufacturers. Further, they were not analysing MSQ in the context of well-defined mass spectral features, which are independent of the identification database used. We identified five mass spectral features as good proxies for MSQ. While some (e.g. technical reproducibility) were previously reported (68), others were less frequently described (e.g. number and relative intensity of ribosomal marker masses). We further focused on the most perspicuous and most comparable MSQ features, namely i) the number of ribosomal marker masses and ii) technical reproducibility. Furthermore, these MSQ features were used to assess which workflow adaptations improve MSQ over all or for individual bacterial taxa.

In line with previous studies, we observed a positive impact of performing a protein extraction or pre-treating the Gram positive bacterial samples with formic acid before the measurement (**Chapter III**) (262,312,390). Whereas we observed MSQ decreasing with bacterial colony age increasing after 24 hours (262), previous studies identified a minimum incubation time of 6-8 hours to yield high MSQ (98), suggesting an ideal incubation time of more than 8, but less than 24 hours for strains of unknown species. While the low inter-laboratory reproducibility (68) and the essential requirement for increased standardisation (393) have been recognised, this knowledge has not yet been used to improve MSQ in routine diagnostics.

In conclusion of **Chapter III**, we identified optimal sample preparation workflows by assessing several MSQ features. This approach should be applied to many diagnostic laboratories to ensure continuous and reproducible results. While some features, such as the total intensity of a mass spectrum, are easily evaluated, others might not be easy to assess for routine diagnostic laboratories, as they are not evaluated by routinely used MALDI-TOF MS software.

Therefore, a standardised ring trial including a broad collection of diagnostic laboratories is essential to ensure high quality and reproducible MALDI-TOF MS.

MALDI-TOF EQA.

In **Chapter IV**, we further analysed and improved the MSQ in routine diagnostics laboratories by an international EQA, including 36 diagnostic laboratories in 12 countries. We aimed to i) assess routine diagnostic MSQ with the feature identified in the previous chapter and ii) assess the improvement after implementing standardised sample preparation protocols. We gathered over 10,000 mass spectra, acquired from the same well-characterised 47 strain set used in **Chapter III**. We identified heterogeneous MSQ between the participating diagnostic laboratories, mainly driven by a few particularly poor or well-performing laboratories. The simple workflow adaptations we proposed improved MSQ for previously poorly performing laboratories but had a limited impact on the average MSQ over all laboratories (**Chapter IV**). This discrepancy is potentially linked to hardware settings and device maintenance, which were not considered in this study. Moreover, to understand the extent and cause of routine MSQ fluctuations, future studies are needed to assess MSQ longitudinally. A regular assessment of MSQ in routine diagnostics would enable early detection of decreasing MSQ.
MALDI-TOF mass spectra comprise a large number of signals, of which only a fraction is required for bacterial species identification (394). We have so far improved MSQ in routine diagnostics for bacterial species identification (**Chapter III** and **IV**). A consistently high MSQ in routine diagnostics likely further improves the machine learning prediction of AMR from MALDI-TOF mass spectra.

7.3 Prediction of Antimicrobial Resistance from MALDI-TOF mass spectra

In **Chapter V**, we examined the potential of machine learning on predicting AMR directly from MALDI-TOF MS data. Previous studies have attempted this (90,95,96,395)**.** However, most of these classifiers were trained on limited datasets of a single species, acquired on a single device, and their transferability into a different laboratory has not yet been assessed. Moreover, prior to our study, there was a clear lack of large, publicly available, high-quality datasets to test the classifiers in different scenarios.
Therefore, we first compiled an extensive data dataset of over 300,000 MALDI-TOF mass spectra with matching AMR profiles acquired in clinical routine diagnostics at four different healthcare centres. Next, we used three machine learning approaches, namely logistic

regression, gradient-boosted decision trees, and a deep neural network classifier, to classify AMR from MALDI-TOF mass spectra in different scenarios. The prediction worked best if the classifiers were trained on spectra of a single species acquired at the same centre and in close temporal proximity to the test set. This observation might be linked to varying MSQ between sites and changing hardware factors over time and further emphasises the need for standardisation. We found that the most discriminatory peaks either had very low or high intensities. This suggests that resistant strains differ from susceptible ones in patterns of presence and absence of biomarker peaks, rather than in varying expression levels of proteins. The AMR factors cannot be detected, as their weight lies outside the MALDI-TOF mass range and they are not expressed under standard laboratory conditions (**Chapter V**). In line with previous studies, we identified biomarker peaks that correspond either (i) to proteins that are directly linked to the AMR gene (78,92,396), or (ii) to phylogenetic marker masses of a lineage that is associated with AMR (93). An example of the first is the peak at 2415 m/z for Methicillin-Resistant *S. aureus* (MRSA), which corresponds to the protein PSM-mec encoded on the same gene cassette as the resistance gene *mecA* (92,396). An example of the latter is the peak at 8,448 m/z, which frequently occurs in spectra of the resistant *E. coli* ST131 (360,396). This peak corresponds to the uncharacterised protein YnfD, which has no known functional connection to AMR (360), but rather serves as a phylogenetic surrogate marker to detect AMR. Such phylogenetic associations to AMR can be transient, reflecting a local spread of a resistant clone (397,398), which offers a further explanation of the performance drop when using temporally or geographically distant spectra to train the classifier. We show that large datasets are required for a successful AMR prediction with over 2,500 mass spectra and AMR measurements needed per species-antibiotic combination to achieve an area under the receiver operating characteristic curve (AUROC) of >= 0.74 for Ceftriaxone resistance in *E. coli*, Ceftriaxone resistance in *K. pneumoniae* and Oxacillin resistance in *S. aureus* (396). Therefore, successful AMR prediction from MALDI-TOF mass spectra requires a recent, sufficiently large and high-quality dataset that has been collected from a similar patient population, ideally at the same hospital, but certainly in close proximity. As meeting this requirement is not possible for smaller laboratories, publishing and constantly updating reference datasets is required to translate this advancement into clinical routine diagnostics. A large, diverse and international dataset might further allow AMR prediction for less frequently isolated species.

Summarising **Chapter III-V** of this thesis, we highlight: (i) the requirement for standardisation of sample preparation and data analysis for MALDI-TOF MS species identification; (ii) the potential of MALDI-TOF MS for high-resolution diagnostics and pathogen surveillance; and (iii) the importance of sharing MALDI-TOF mass spectra and AMR profiles acquired in clinical routine diagnostics.

8 Outlook

8.1 Further developments for MALDI-TOF MS based bacterial identification

With the work presented in this thesis, we highlight the capacity and possibilities of MALDI-TOF MS to (i) increase the taxonomic resolution for bacterial identification using a marker-based approach and to (ii) predict AMR directly from the MS spectra using machine learning. The potential of MALDI-TOF MS for microbiology is not yet fully used, and ample opportunities for further development of microbiological diagnostics considering the four aspects of data analysis, data sharing, diagnostic workflows and hardware technology remain.

Data analysis of MALDI-TOF MS data is currently mainly based on the comparison of spectra of unknown strains to reference spectra databases using a similarity-based approach. On the other hand, the mass of phylogenetic marker proteins can be predicted from genomic sequences (76,85). These mass profiles can subsequently be used as an alternative database for bacterial identification from MALDI-TOF mass spectra (69,74,306). This has multiple advantages compared to using reference mass spectra as database: (i) constitutively expressed proteins are detectable independently of the growth conditions; (ii) as the identity of the examined peaks is known, it is more apparent to estimate uncertainties in identification and (iii) mass profiles can be predicted from genome sequences of strains for which no MALDI-TOF mass profile is yet available, which increases the size and diversity of the database. The latter could potentially enable the identification of unknown bacteria which have not previously been isolated and whose genomic sequences originate from metagenomic sequencing projects.

Data sharing is key for pathogen surveillance and for translating machine learning approaches to clinical routine settings (399). In the infectious disease context, future efforts are required to share MALDI-TOF mass spectra and AMR profiles from different healthcare centres around the world. The data should be published under the FAIR principle (400), making them Findable, Accessible, Interoperable and Reusable (FAIR). Ideally, acquired mass spectra and AMR profiles are shared as they are routinely acquired in a standardised manner (393). The resulting database would be thus constantly updated and quality controlled. Such a database would strongly improve AMR predictions from MALDI-TOF mass spectra using machine learning and allow the transition of this advancement to clinical routine diagnostics. The challenge of building and maintaining such a platform would be administrative and could be met with existing technologies.

Diagnostic workflows need to be adapted to use the potential of AMR prediction from MALDI-TOF mass spectra in clinical diagnostics (**Figure 1**). In a revised diagnostic workflow, AMR will be predicted from MALDI-TOF mass spectra at the time point of species identification. This prediction will not replace phenotypic testing, but will give first evidence on AMR hours before phenotypic test results are available. Routinely acquired MALDI-TOF mass spectra and phenotypic test results will be constantly uploaded into an international database. After the data is curated, it will serve to improve the classification algorithm steadily. This constant update of the classification algorithm is important as the prediction performance decreases with temporally distant data (**Chapter V**). It will further enable the identification of biomarker peaks of locally spreading resistant lineages in real-time. In a revised diagnostic workflow, the MSQ will regularly be assessed and improved (**Chapter III** and **IV**). A constantly high MSQ enables the reproducible detection of more phylogenetic marker peaks, which will in turn enable the distinction of closely related species or even lineages within a species (**Chapter I** and **II**).

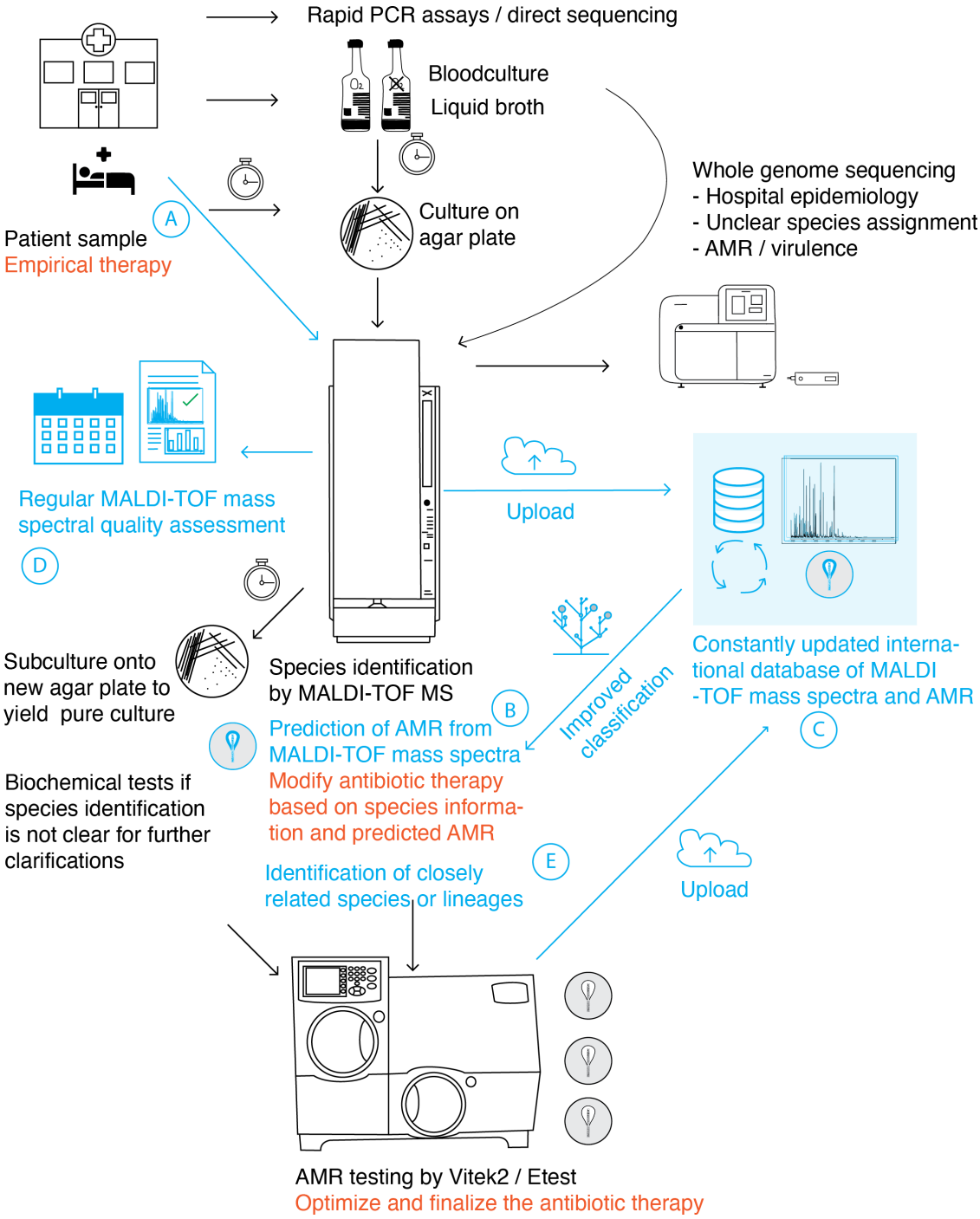

Figure 1: A revised diagnostic workflow, integrating new functionalities of MALDI-TOF MS. Changes to existing workflows are indicated in blue and assigned letters A - E. A: MALDI-TOF measurements directly from the patient sample (e.g. urine sample). B: prediction of AMR from MALDI-TOF mass spectra. C: routinely acquired MALDI-TOF mass spectra and phenotypic AMR measurements are uploaded into an international database. This data is quality controlled and serves as the basis to improve the prediction algorithm. D: The MSQ is regularly assessed, allowing for an early reaction when MSQ decreases. E: With a constantly high MSQ and the reproducible detection of phylogenetic marker masses, the resolution of bacterial identification can further increase, allowing the distinction of closely related species or lineages within a species.

The **technology of MALDI-TOF** MS for bacterial identification is constantly advancing. Currently, routinely employed MALDI-TOF MS detects positively charged ions which originate mainly from intracellular proteins (235). However, a highly promising approach is becoming available in routine diagnostics, where negatively charged ions (i.e. lipids) are measured (401). The negative ion-mode increases the resolution for bacterial typing and allows for the detection of colistin resistance in *E. coli* (402). Other exciting technological developments are the application of MALDI-TOF MS directly from the sample material (403,404) or the increased sensitivity of MALDI-TOF MS to lower bacterial cell counts using microfluidics, thereby decreasing incubation time (405).

Further improvements in hardware used in routine diagnostics could include the use of different matrices (75). These improvements will enlarge the detectable mass range or increase the measurement precision. All of these developments will further improve the resolution of identification using a marker-based analytical approach.

Moreover, recent advances have demonstrated the versatility of MALDI-TOF MS also in non-infectious disease settings by applying it to the identification of viruses (406), mammalian cell lines (407,408), insects (409,410) and truffles (411,412).

8.2 Sequence-based diagnostic assays as alternatives to MALDI-TOF MS

Although MALDI-TOF mass spectrometry has many advantages for bacterial identifications and has revolutionised microbiological diagnostics, it has two intrinsic limitations: (i) bacterial cells have to be present at high concentration to reach the detection threshold, which most often requires several hours of incubation; and (ii) only proteins which are expressed at high levels in laboratory conditions can be detected.

For the rapid detection of specific virulence and AMR markers (217,413), sequence-based diagnostic assays, such as targeted PCR, offer an alternative to MALDI-TOF MS. In targeted, sequence-based diagnostic approaches, a set of known virulence or resistance factors are queried, while factors outside this set remain undetected. Such assays could be used to identify virulent UPEC isolates carrying *papGII*. The diagnostic value of this bacterial biomarker for invasiveness, like of any biomarker (414), further needs to be confirmed in prospective, randomised trials (415). Invasiveness, as a clinical outcome, might have very different dynamics and underlying causes depending on the clinical setting, for example, in a nursing home (416) compared to young people in the emergency ward (417). Therefore, to assess the true diagnostic impact of this biomarker, the clinical scenario of such a trial would need to be precisely defined. Bacterial biomarkers could thereby complement the clinical

assessment of an infection that is currently primarily host driven (418). For a diagnostic assay detecting a bacterial biomarker to have a clinical impact, it needs to provide additional information in a timely manner, requiring short laboratory turn-around times. UTIs are amongst the most frequent infections worldwide and mainly cause mild symptoms (186). The data presented in this thesis (**Chapter II**) gives evidence that few UPECs have the potential to cause invasive disease infection in young and immunocompetent hosts. An assay that detects these virulent UPECs will therefore be required to be cost-effective and with low hands-on time, as only a frequent employment in a routine diagnostic workflow can yield a diagnostic impact. Future assays might allow not only the detection of a biomarker gene, but go a step further and quantify its expression level, yielding a more precise assessment of the progression of an infection (419,420).

The data analysed in this thesis (**Chapter II**) and by others (202), suggest that multiple *E. coli* strains can co-infect the human urinary tract at once. Future studies assessing the within-host genetic diversity of *E. coli* strains, isolated from various body sites, could elucidate the multi-strain infection dynamics. This could have consequences for reconstructing transmission networks (421) and for bacterial diagnostics, which are currently centred around single colony isolates and rarely consider a bacterial community as the causative agent of an infection. To consider this within-host strain diversity, untargeted sequence approaches could offer an alternative to targeted approaches. Shotgun metagenomic sequencing has the potential to differentiate bacterial strains (422–424) and is applied directly to patient samples without the need for incubation (425,426). The rapid advancement of bioinformatic tools and steadily sinking sequencing costs might permit its implementation to clinical routine diagnostics in the coming years (427). However, in line with standardisation for MALDI-TOF MS advocated for in this thesis, high standardisation in sampling and data analysis will be needed to enable the extraction of clinically actionable items from metagenomic sequences in a short time.

9 Conclusion

In view of the steadily increasing frequency of antibiotic-resistant bacteria, rapid assessment of pathogens is of crucial importance to treat an infection effectively, to promote antibiotic stewardship and for pathogen surveillance. In this thesis, we show that (i) MALDI-TOF has the potential to increase in taxonomic resolution, distinguishing between closely related bacterial taxa, and that (ii) AMR can accurately be predicted from MALDI-TOF mass spectra. Moreover, as exemplified throughout this thesis, high MSQ, standardised sample preparation and data analysis workflows as well as FAIR sharing of data are essential in order to translate these achievements to clinical routine diagnostic settings and open up new avenues for infection control.

10 Acknowledgements

Over the course of my PhD, I had the pleasure to work with inspiring and supporting individuals, without whom this thesis would not have been possible. I would like to express my sincere gratitude to:

My supervisor Adrian Egli, for guiding me throughout this thesis, for inspiring discussions, for his wisdom and passion for science, for always being supportive and for giving me the opportunity to conduct exciting research.

Valentin Pflüger for teaching me about MALDI-TOF MS, for his enthusiasm and know-how, scientific discussions, critical feedback and support throughout this thesis.

Helena Seth-Smith and Vladimira Hinić, for giving invaluable feedback throughout this thesis and for always having an open ear for me.

Nick Thomson at the Sanger Institute, for hosting me in his fascinating group and giving me the chance to learn from him and his team.

Jacob Moran-Gilad, for collaborating throughout the years, for critical feedback and for giving me the chance to visit him and his group in Beer-Sheva.

Karsten Borgwardt, Caroline Weis and Bastian Rieck from the ETHZ for the great collaboration and giving me insights into the world of machine learning.

Birke Mebold for invaluable administrative support.

The Bioinformatics Core Facility, the Scicore-Team and the DBM-IT support team for always helping me out, when I needed it.

The members of my PhD Committee, Prof. Urs Jenal, Prof. Annelies Zinkernagel and Prof. Sylvain Brisse for taking the time to evaluate my work.

All current and past members of the Applied Microbiology Research group and the Infection Biology group at the DBM with whom I had the chance to work with, for the great atmosphere. Special thanks to Yaseen, Jana, Fanny, Alfredo, Marco, Tim, Daniel, Yuki, Steffi, Josi, Olivia, Daniela, Denise, Diana, Laurent, Ann-Kathrin, Giovanni, Vanni, Madlen, Srinithi, Anne, Claudia, Darya, Pascal, Aya.

All members of the Team216 at the Sanger Institute for the great time I had there, and to share their knowledge with me. I especially thank Alyce, Gal, Mimi, Florent, Grace, Kate, Matt, Mat and Sushmita.

The employees at Mabritec, especially Roxanne, Frédéric and Samuel for everything I learned from them and for always helping me out.

Moritz Grubenmann, with whom I did an internship at the Laborgemeinschaft1 and who first introduced me to clinical microbiology and who awakened my passion for bacteria.

To my friends and my family, for love and friendship and to Vincent, for everything.

11 Appendix

Appendix I: Supplementary Material Bacterial genome wide association study substantiates *papGII* of *E. coli* as a patient independent driver of urosepsis

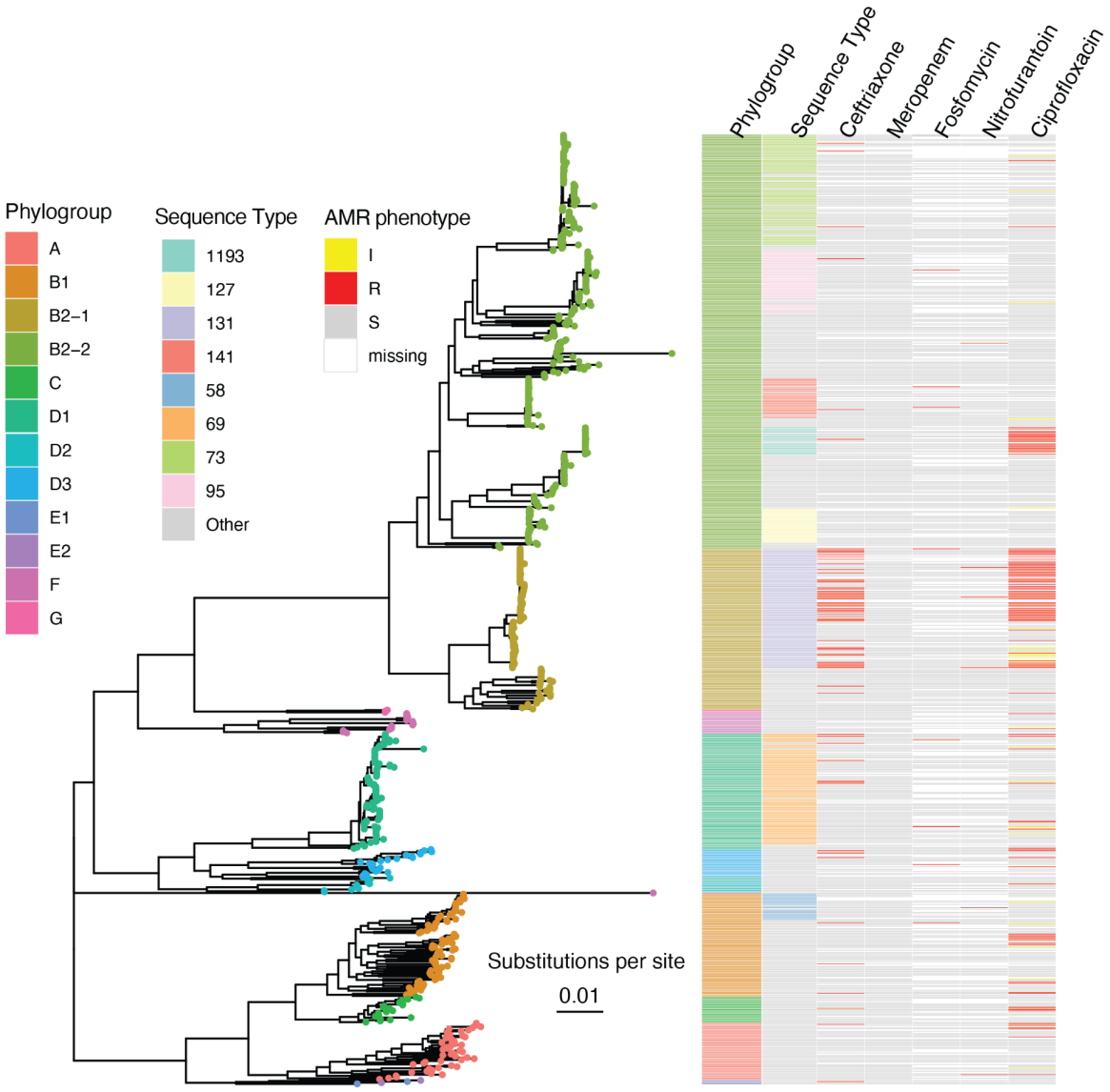

Figure S1: Core genome phylogeny of 831 *E. coli* strains. Columns represent (from left to right): the assigned phylogroup, the sequence type, phenotypic resistance against Ceftriaxone, Meropenem, Fosfomycin, Nitrofurantoin and Ciprofloxacin.

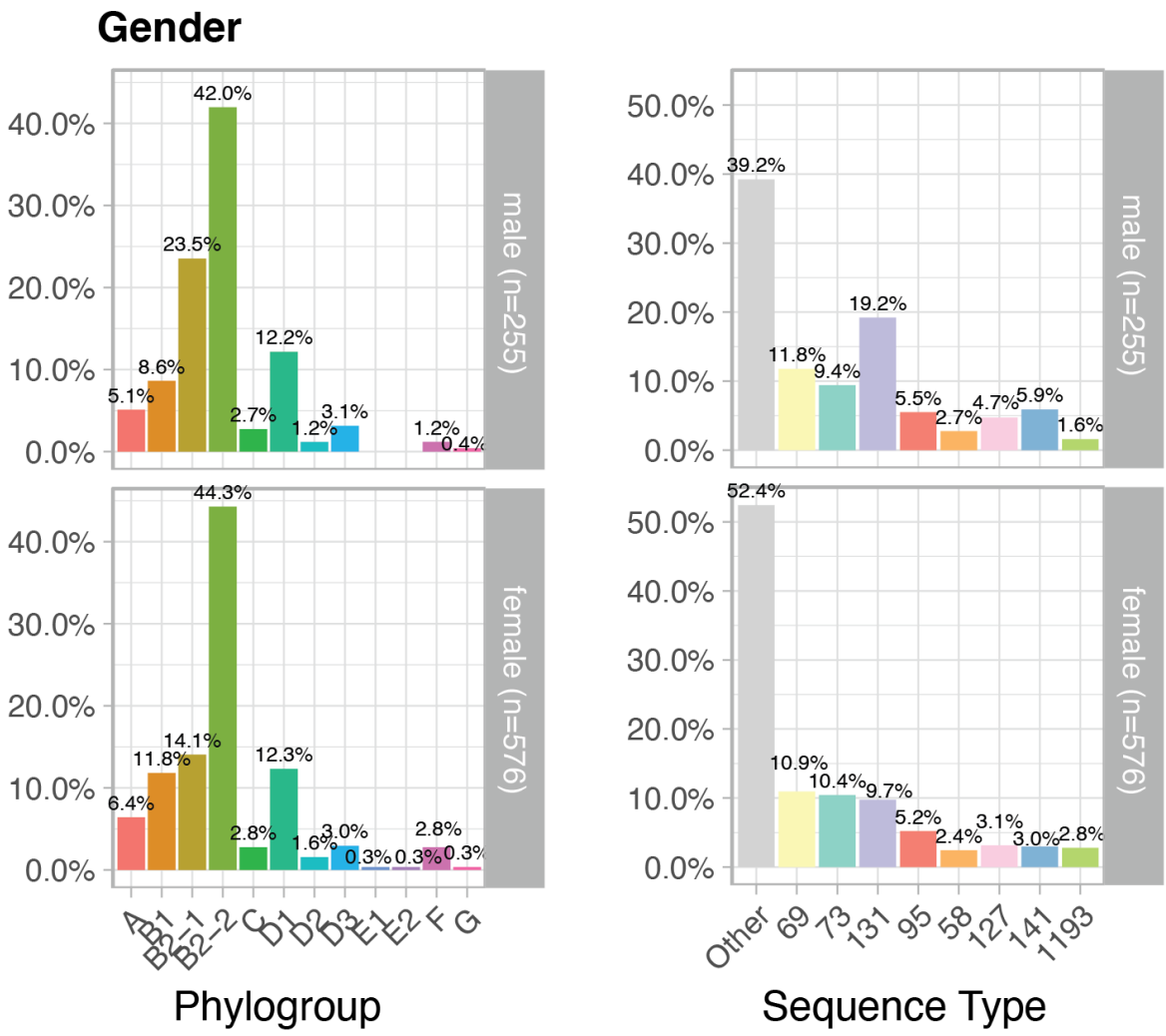

Figure S2: Distribution of *E. coli* phylogroups (left) and Sequence Types (ST) (right) in male (n=255) (upper row) and female (n=576) (lower row) patients.

Table S2: Genes which showed a tendency towards being associated with invasive infection with a -$\log_{10}$(p-value) > 4.Columns: 'Gene': Name of the annotation; 'EggNOG': If no annotation was available, genes were screened using the EggNOG database; 'Function': Assumed function of the genes product; 'Reference': Respective Publication; 'No. of unitigs': How many unitigs were annotated; 'max -$\log_{10}$(p-value): highest value per gene.

Gene	EggNOG	function	References	No.of unitigs	max -$\log_{10}$ (p-value)
papG		part of the P-pili	Uniprot entry: P76364	156	10.7
100033-19_04615	Pilus-assembly fibrillin subunit, chaperone, papJ domain identified	part of the P-pili		7	7.4
papH		part of the P-pili		75	6.5
klcA_1		can be plasmid encoded. Antirestriction enzyme	Serfiotis-Misa et al. *Nucleic Acids Research* (2009)	43	6.0
100033-19_01758	Unknown function, cytoplasmic location	Unknown		23	5.7
pagN		Wrongly annotated, is *tia* (100% similarity and coverage). Tia is a adhesin and invasin, originally described in enterotoxic e. coli, Encoded upstream of papG	Mammarappallil and Elsinghorst *Infection and Immunity* (2000)	188	5.6
papC_3		part of the P-pili		281	5.4
100033-19_01761	"Protein of unknown function (DUF987)"	Unknown		4	4.9
papA		part of the P-pili		202	4.9
105056-19_05002	no orthologues found			15	4.8
100033-19_01765	"Enterobacterial protein of unknown function (DUF957)"	Unknown		10	4.6
opgE_2		Involved in the regulation of osmoregulated periplasmic glucans (OPG). These are sugars in the periplasm of bacteria, which control bacterial motility and the secretion of expopolysaccharides	Bontemps-Gallo et al. *BioMed Research International* (2013)	107	4.6
pdeL_2		Thought of as a "switches individual cells of a population from high to low c-di-GMP signaling states and back", division of labour within a bacterial population	Reinders et al. *bioRxiv* (2021) Sellner et al. *mBio* (2021)	165	4.4
100033-19_01766	Unknown function, Domain of unknown function (DUF4942) identified	Unknown		45	4.4
papK		part of the P-pili		62	4.3
intS_3		prophage integrase. Maybe flanking papGII encoding pathogenicity island?	Uniprot entry: P37326	469	4.3

117416-18_02835	no orthologues found			23	4.3
117416-18_02836	no orthologues found			19	4.3
ompD_3		Outer membrane porine		39	4.2
papE		part of the P-pili		140	4.2
100033-19_01761|cbeA_2	Antitoxin component of a type IV toxin-antitoxin (TA) system	https://www.uniprot.org/uniprot/P76364		1	4.2
gnd		6-phosphogluconate dehydrogenase could be related to an adjacent O-antigen biosynthesis gene cluster passive hitch-hiker of recombination events that determine both LPS antigenic changes and diversifying selection is the third enzyme of the pentose phosphate pathway	Cookson et al. *scientific reports* (2017)	900	4.2
cbeA_2		Antitoxin component of a type IV toxin-antitoxin (TA) system	Uniprot entry: P76364	54	4.2
pimB		wrongly annotated, is colanic acid biosynthesis glycosyltransferase WcaL When introduced to harsh conditions such as low pH, pathogenic Escherichia coli can secrete colanic acid to establish a protective barrier between the organism and the acidic environment This protein is involved in the pathway slime polysaccharide biosynthesis, which is part of Slime biogenesis.	Scott et al. *Biochemestry* (2019) Uniprot entry: P71243	160	4.2
105099-20_04976	function unknown			9	4.2
113856-19_04438	restriction modification system DNA specificity domain identified			52	4.1
100033-19_04622	Homologue of alpA	alpA is a DNA-binding transcriptional activator of the expression of the slpA gene. When overexpressed, leads to suppression of the capsule overproduction and UV sensitivity phenotypes of cells mutant for the Lon ATP-dependent protease	Uniprot entry: P33997 Trempy and Gottesman *Journal of Bacteriology* (1994)	1	4.1
109327-20_05007	no orthologues found			3	4.0

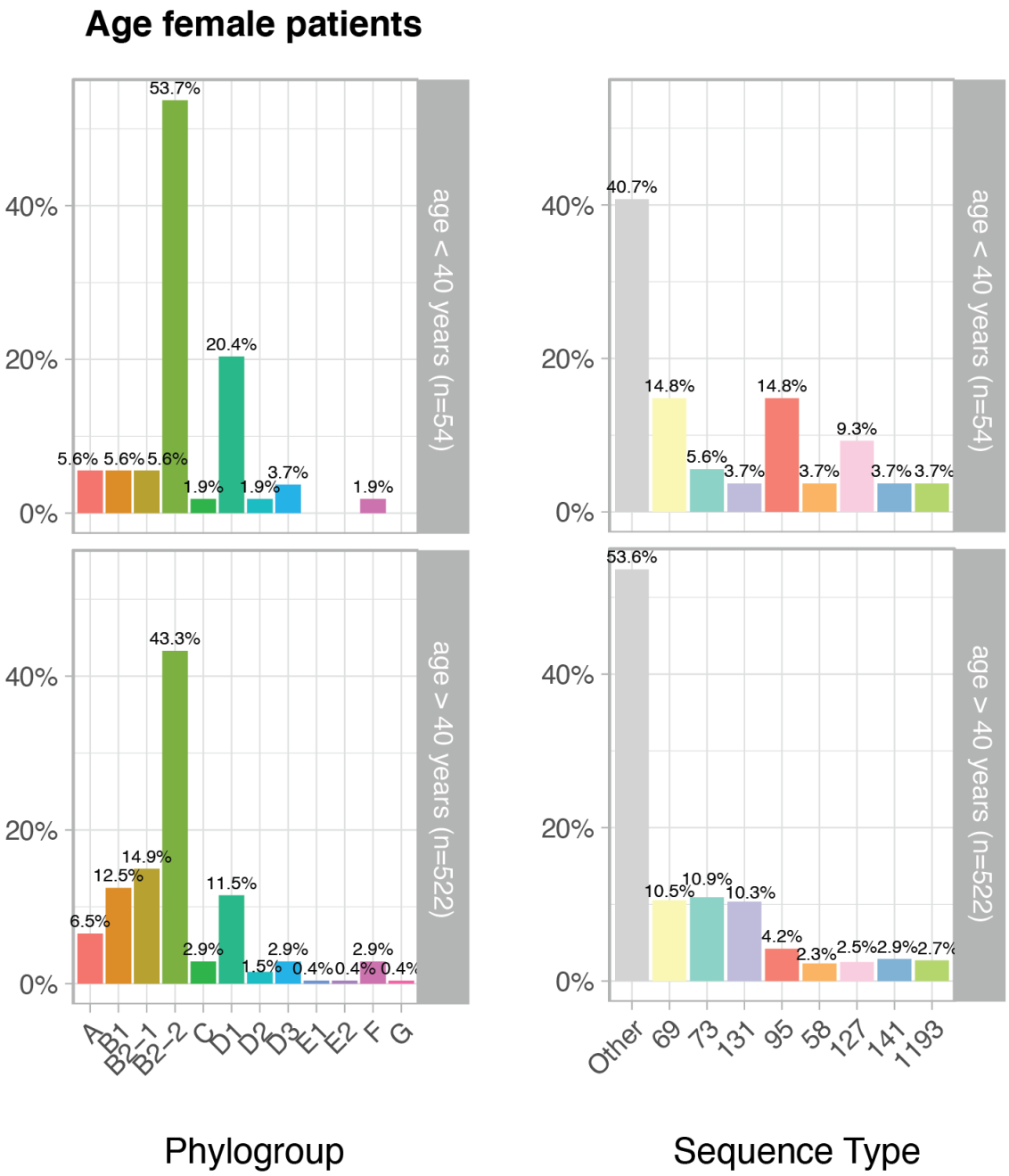

Figure S3: Distribution of E. coli phylogroups (left) and Sequence Types (ST) (right) in in female patients (n=576) younger than 40 years (n=54) (upper row) and older than 40 years (n=522) (lower row) patients.

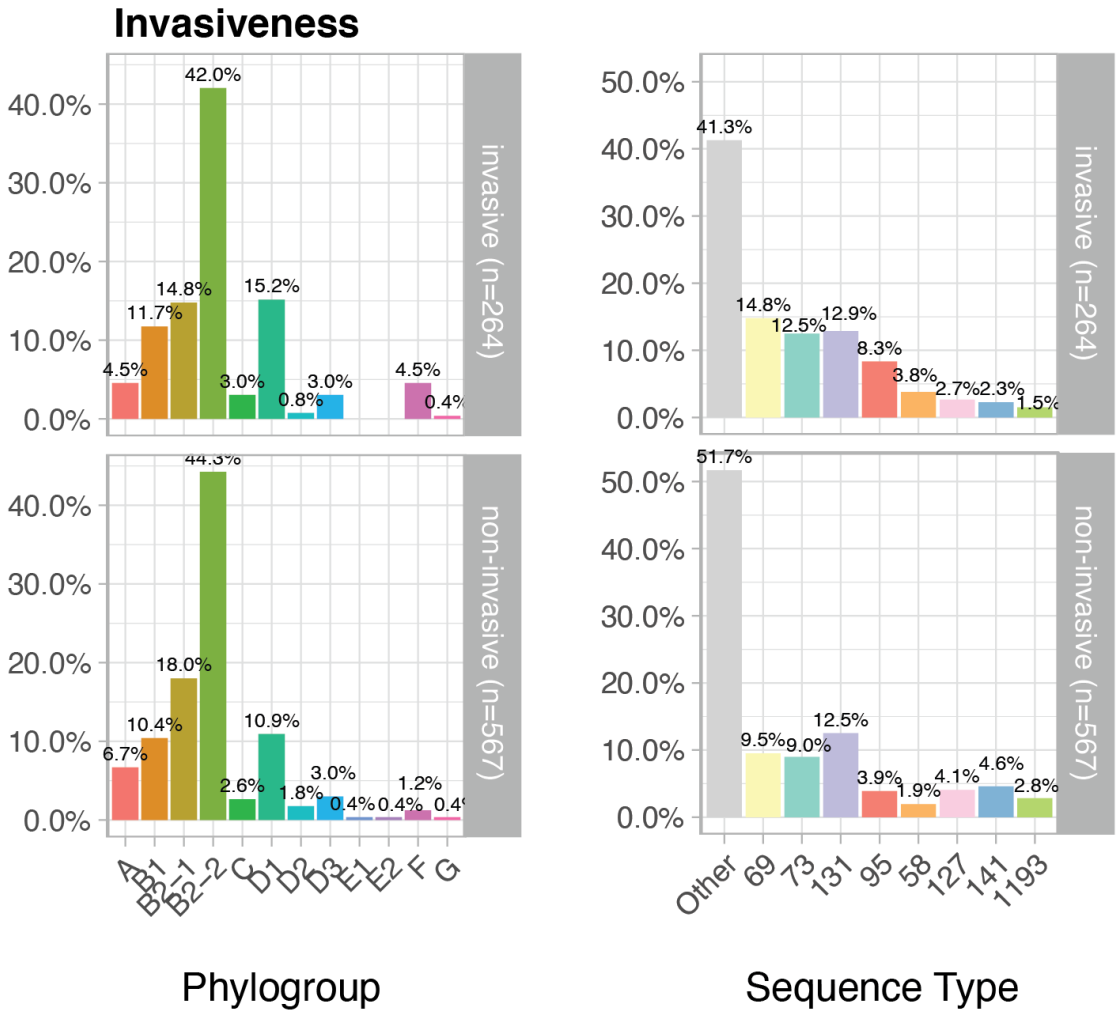

Figure S4: Distribution of E. coli phylogroups (left) and Sequence Types (ST) (right) in invasive infections (n=251) (upper row) and non-invasive infections (n=580) (lower row) patients.

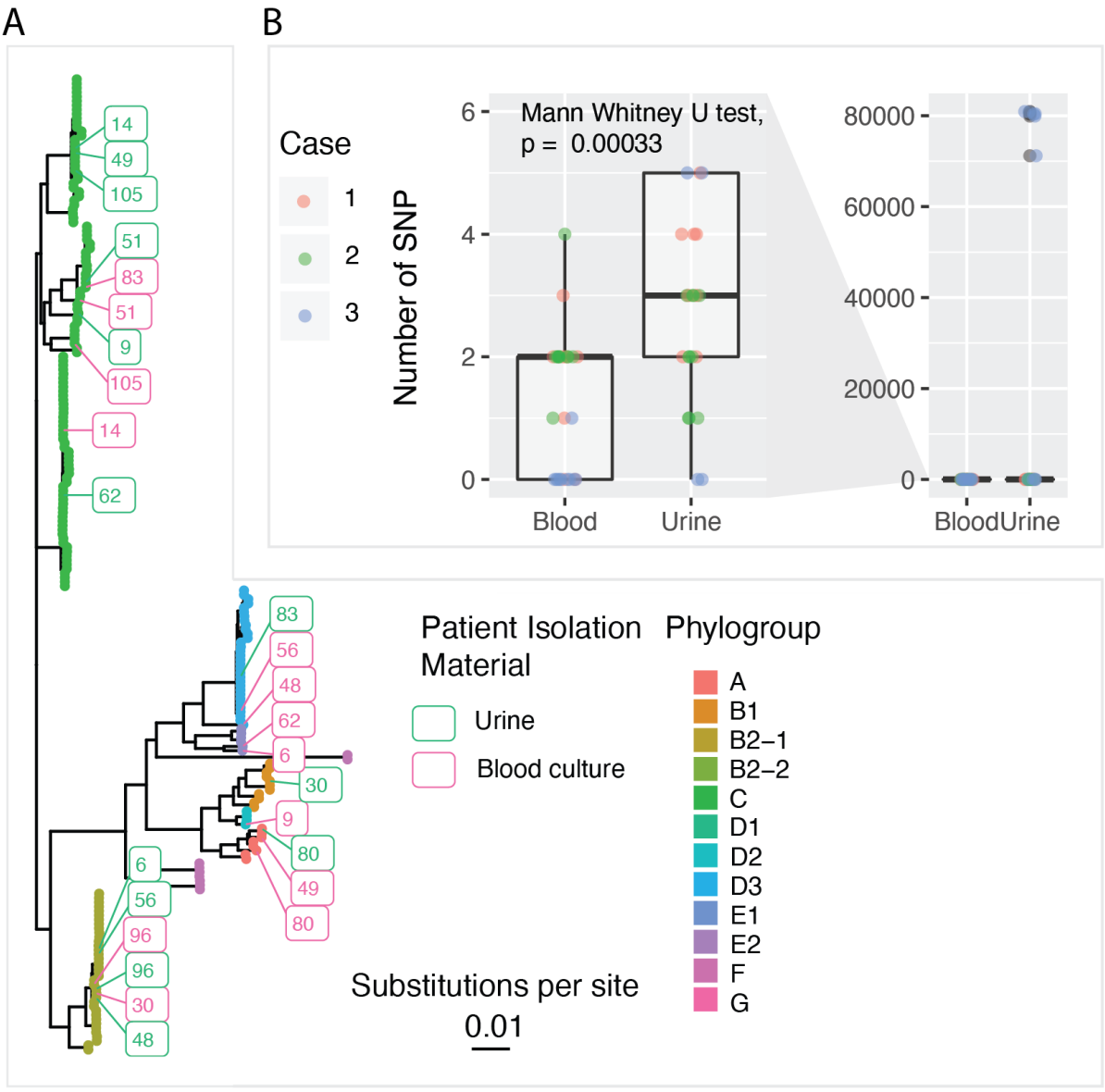

Figure S5: Within host genetic diversity of *E. coli* strains isolated from the same clinical cases A: core genome phylogeny of *E. coli* strains (n=225), isolated from the same clinical case (n=105), colored by phylogroup. The numbers correspond to the case identifier and strains were only labelled, if they exhibited < 99.9% Average Nucleotide Identity to the strain isolated from the same clinical case B: SNV for of 10 picked isolates from three cases, either from urine or blood culture samples.

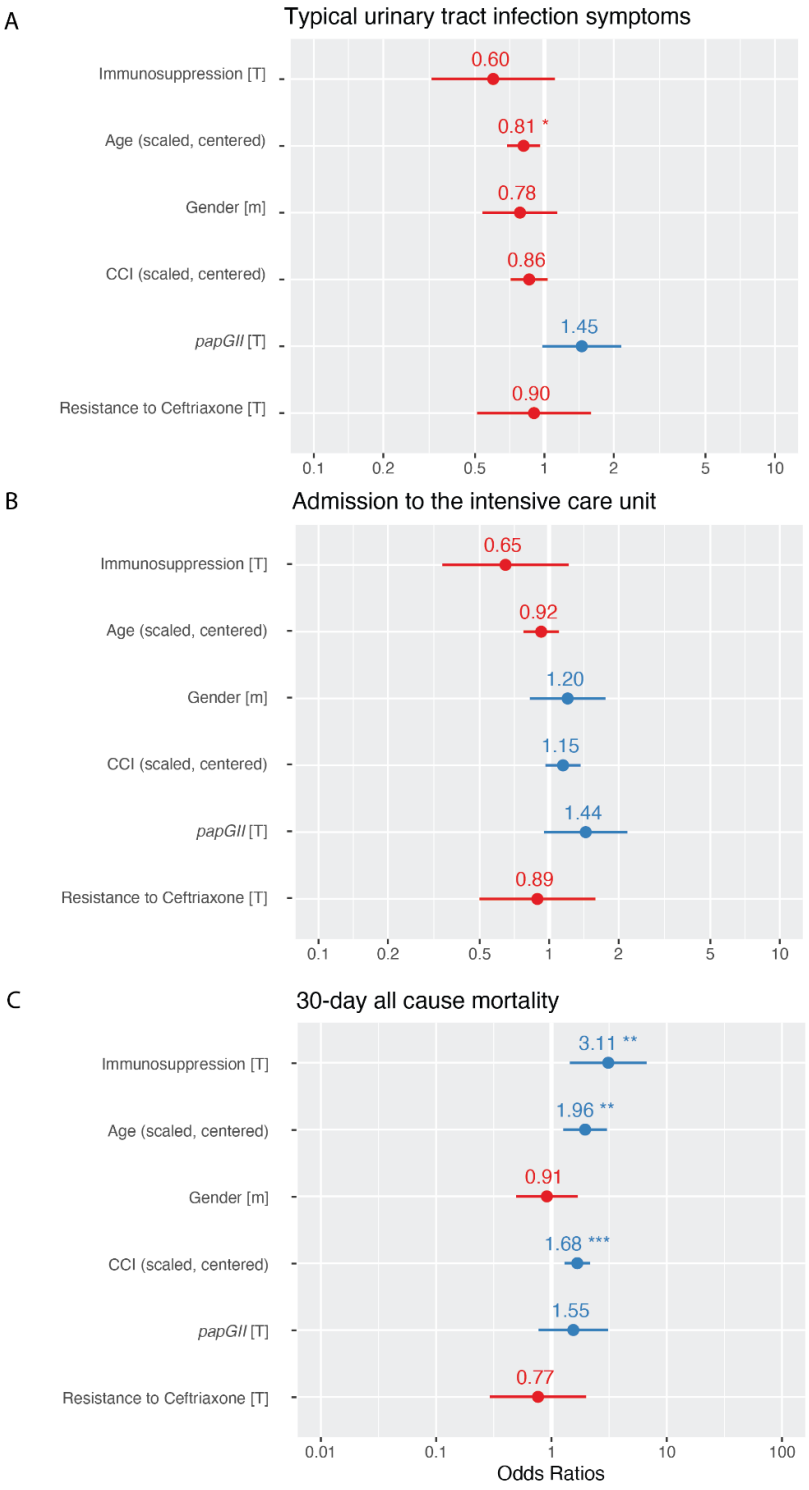

Figure S6: Odds ratio estimates with 95% confidence intervals for **A**: Typical urinary tract infection symptoms (n = 783 complete observations with 232 events); **B**: Admission to the intensive care unit (n = 827 complete observations with 191 events); **C**: 30-day all cause mortality (n = 685 complete observations with 66 events); using the generalised linear model (GLM). OR = odds ratio; CI = confidence interval; CCI = Charlson Comorbidity Index

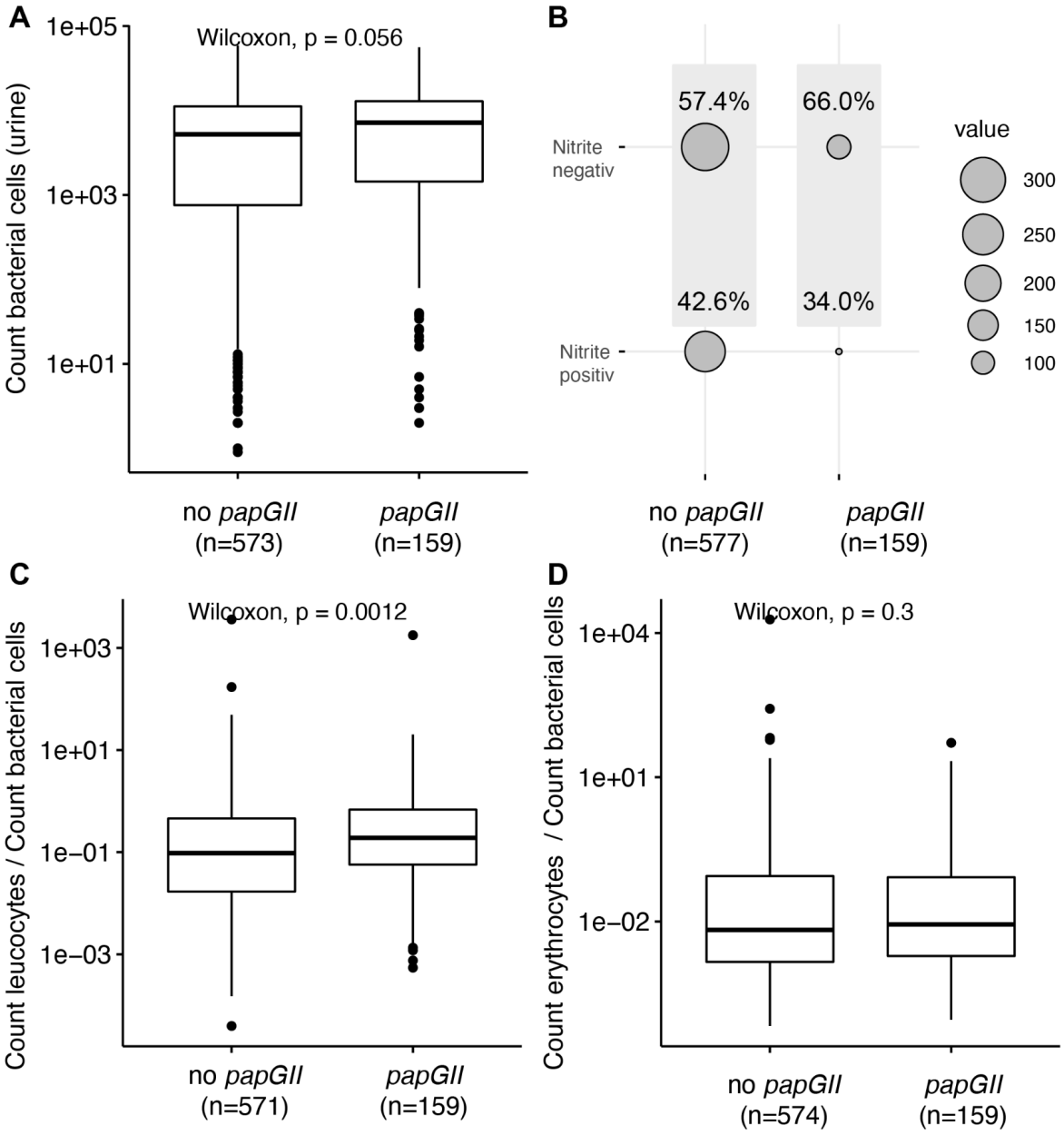

Figure S7: Bacterial cell count (A), nitrit status (B), leucocyte count divided by bacterial cell count (C) and erythrocyte count divided by bacterial cell count (D) measured in urine samples of cases, for which a papGII positive or a papGII negative *E. coli* strain was isolated from a urine or a blood culture samples.

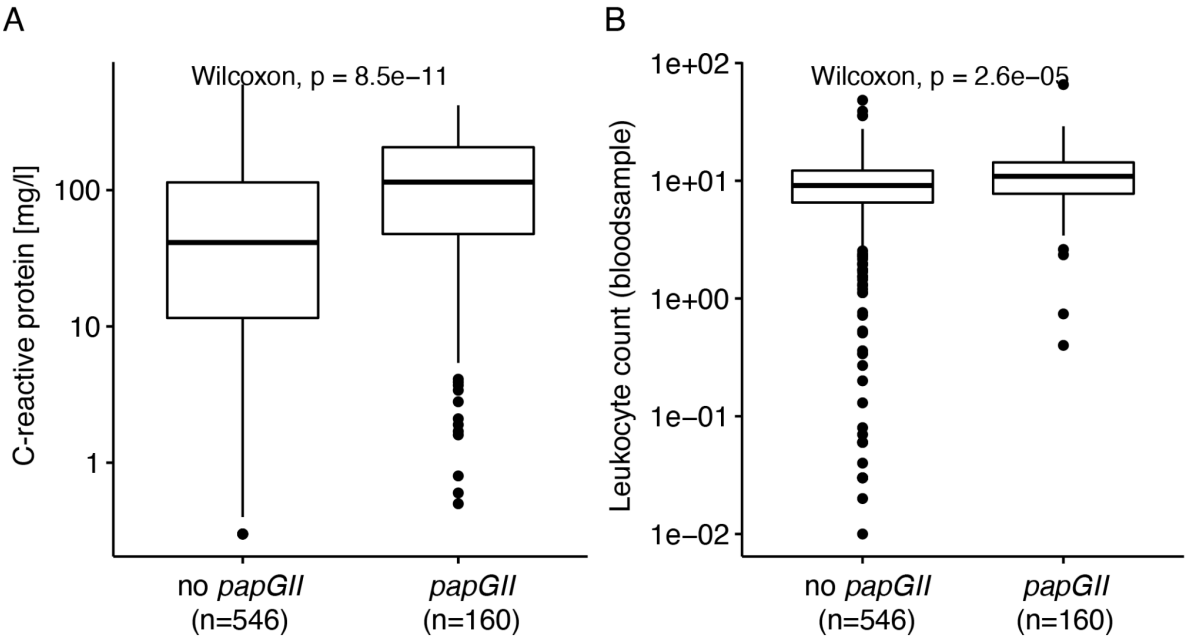

Figure S8: C-reactive protein concentration (A) and leucocyte count (B) measured in blood samples of cases, for which a papGII positive or a *papGII* negative *E. coli* strain was isolated from a urine or a blood culture samples.

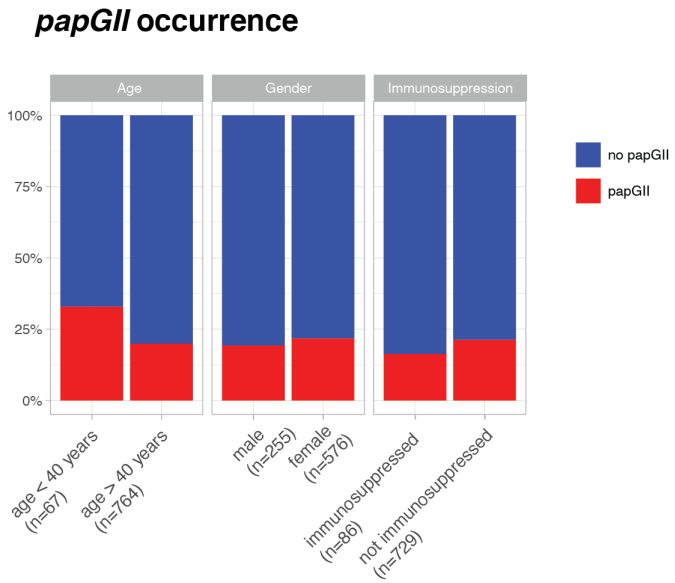

Figure S9: Relative occurrence of *papGII* in isolates from patients younger (n=67) vs. older (n=764) than 40 years, in isolates from male (n=255) vs. female (n=576) patients and in isolates from patients which were immunosuppressed (n=86) vs. patients which were not immunosuppressed (n=729).

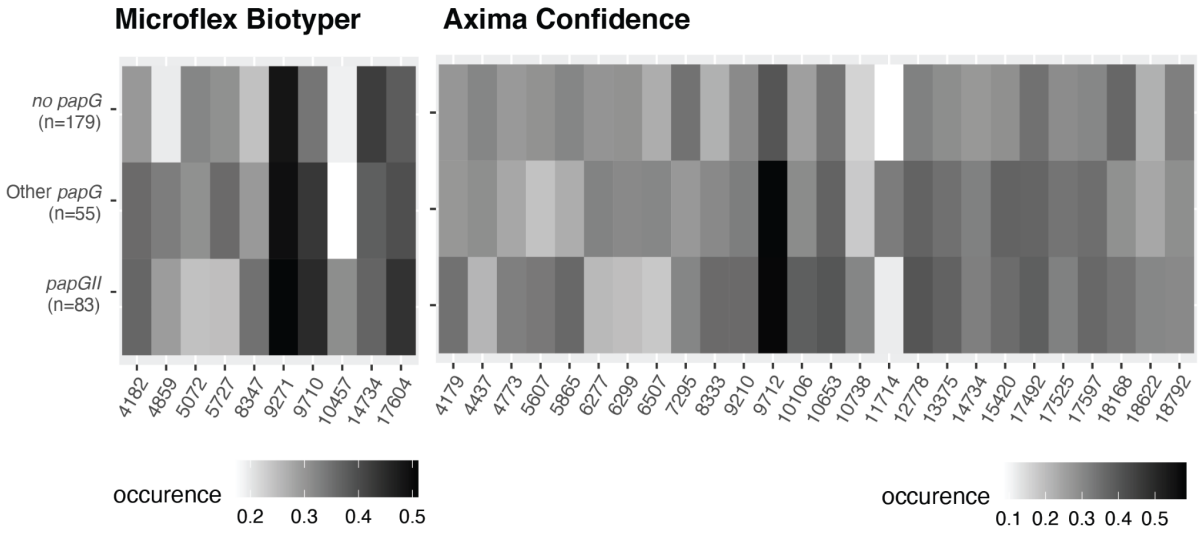

Figure S10: Occurrence of MALDI-TOF mass peaks in spectra acquired from *E. coli* strains encoding no *papG* gene, encoding a *papG* variant other than *papGII* and encoding *papGII*. Each strain was measured in quadruplicate either on a Microflex Biotyper device, or an Axmina Confidence device. Masses are only depicted if detected in > 30% or < 25% of spectra for one or more of the groups.

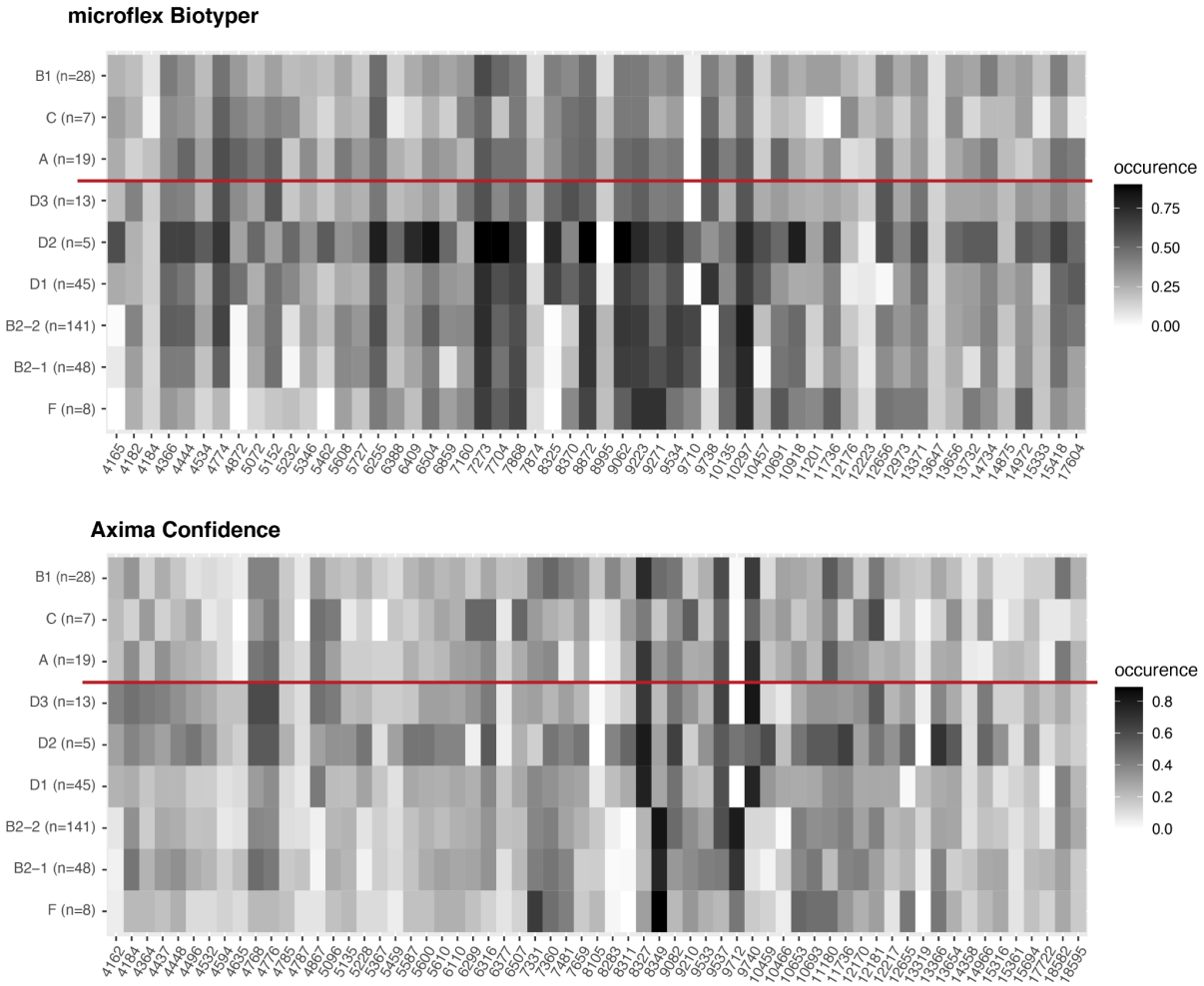

Figure 11: Occurrence of MALDI-TOF mass peaks in spectra acquired from *E. coli* strains of different phylogroups. Each strain was measured in quadruplicate either on a Microflex Biotyper device, or an Axmina Confidence device. Phylogroups for which less than five strains were available (E1, E2 and G) were excluded from the plot. Masses are only depicted if detected in > 50% or < 25% of spectra for one or more of the groups.

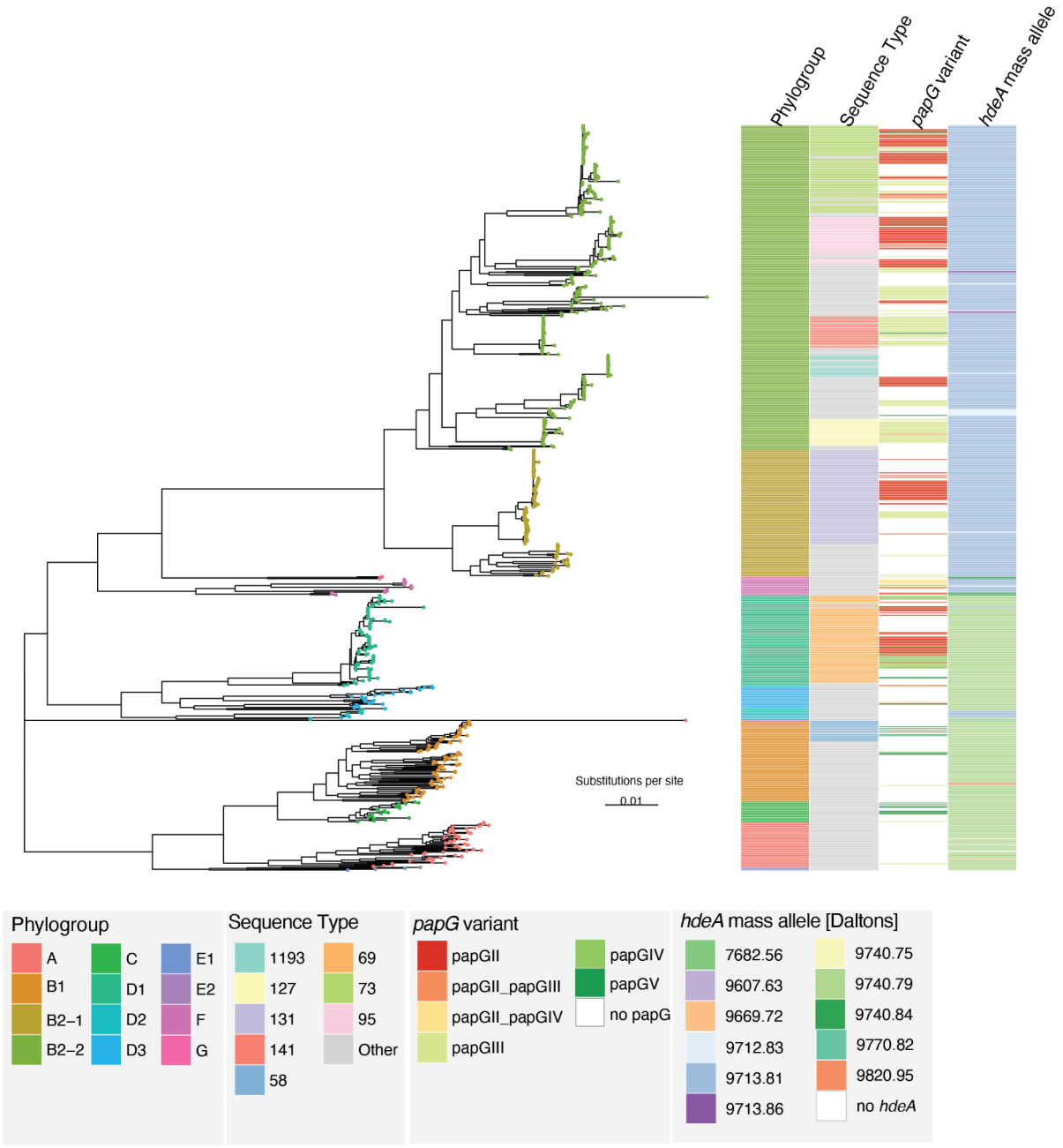

Figure S12: Core genome phylogeny of the *E. coli* strains collected for this study (one strain per clinical case, n=831). Phylogroup assignment, Sequence Type (ST) (eight most frequent ones colored, more rare STs in grey), papG variant, mass of HdeA, predicted from the amino acid sequence

Appendix II: Supplementary Material Quality of MALDI-TOF Mass Spectra in Routine Diagnostics: Results from an International External Quality Assessment including 36 Laboratories from 12 countries

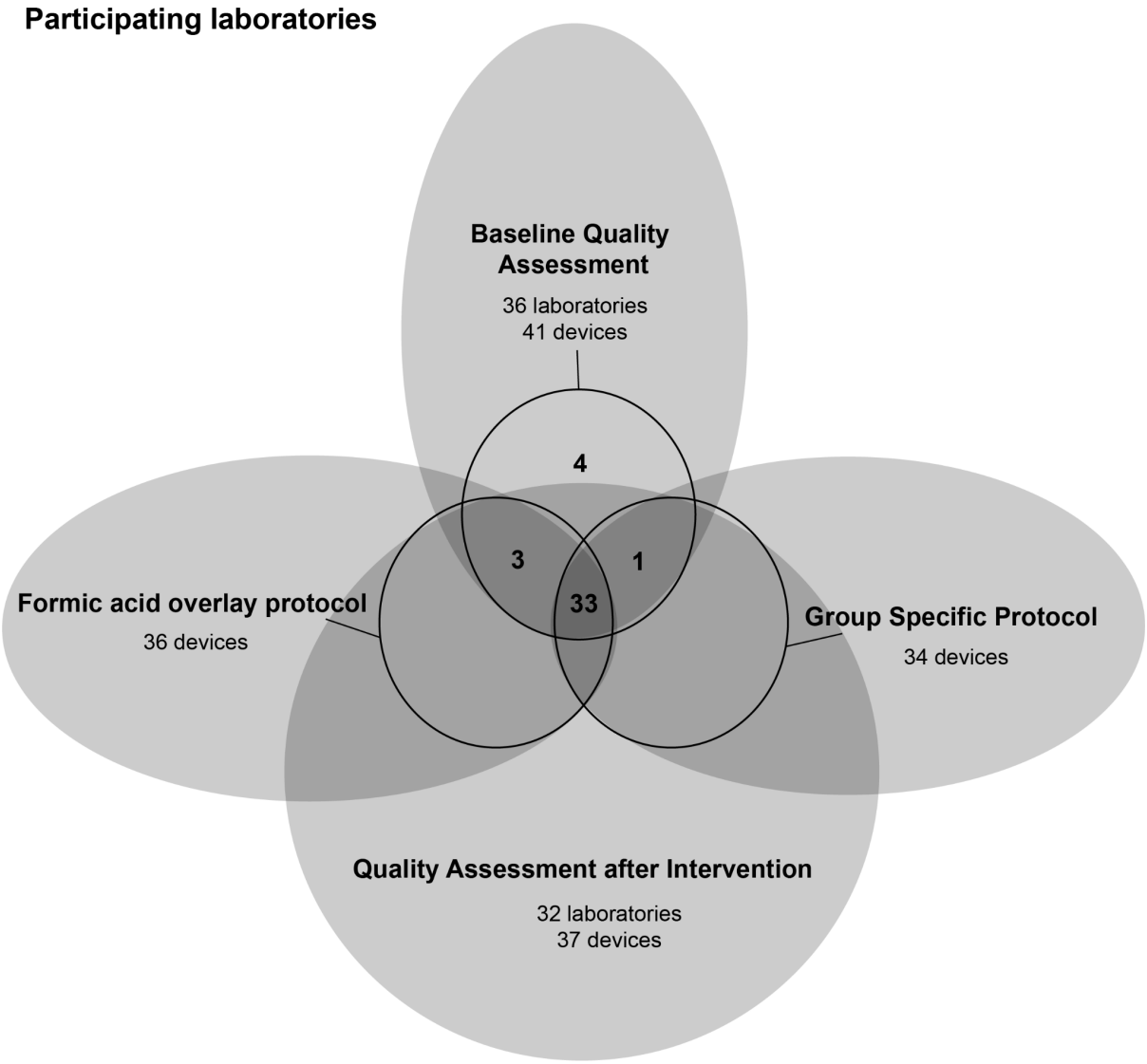

Figure S1: Venn diagram depicting the number of laboratories and devices participating in the different rounds of quality assessment

12 References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921.

2. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature. 2010 Jan;463(7277):84–7.

3. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-Gut Microbiota Metabolic Interactions. Science [Internet]. 2012 Jun 8 [cited 2022 Jan 16]; Available from: https://www.science.org/doi/abs/10.1126/science.1223813

4. Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. McVean G, editor. eLife. 2016 May 17;5:e12469.

5. Ybañez RHD, Ybañez AP, Nishikawa Y. Review on the Current Trends of Toxoplasmosis Serodiagnosis in Humans. Front Cell Infect Microbiol. 2020 May 8;10:204.

6. Marchionni E, Parize P, Lefevre A, Vironneau P, Bougnoux ME, Poiree S, et al. Aspergillus spp. invasive external otitis: favourable outcome with a medical approach. Clinical Microbiology and Infection. 2016 May;22(5):434–7.

7. Jeffery-Smith A, Taori SK, Schelenz S, Jeffery K, Johnson EM, Borman A, et al. Candida auris: a Review of the Literature. Clinical Microbiology Reviews [Internet]. 2017 Nov 15 [cited 2022 Jan 17]; Available from: https://journals.asm.org/doi/abs/10.1128/CMR.00029-17

8. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, et al. The Global Circulation of Seasonal Influenza A (H3N2) Viruses. Science [Internet]. 2008 Apr 18 [cited 2022 Jan 17]; Available from: https://www.science.org/doi/abs/10.1126/science.1154137

9. Turner NA, Sharma-Kuinkel BK, Maskarinec SA, Eichenberger EM, Shah PP, Carugati M, et al. Methicillin-resistant Staphylococcus aureus: an overview of basic and clinical research. Nat Rev Microbiol. 2019 Apr;17(4):203–18.

10. Daniel TM. The history of tuberculosis. Respiratory Medicine. 2006 Nov 1;100(11):1862–70.

11. Chow MYK, Khandaker G, McIntyre P. Global Childhood Deaths From Pertussis: A Historical Review. Clin Infect Dis. 2016 Dec 1;63(Suppl 4):S134–41.

12. Truelove SA, Keegan LT, Moss WJ, Chaisson LH, Macher E, Azman AS, et al. Clinical and Epidemiological Aspects of Diphtheria: A Systematic Review and Pooled Analysis. Clin Infect Dis. 2020 Jul 1;71(1):89–97.

13. McAllister DA, Liu L, Shi T, Chu Y, Reed C, Burrows J, et al. Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis. The Lancet Global Health. 2019 Jan 1;7(1):e47–57.

14. Zunt JR, Kassebaum NJ, Blake N, Glennie L, Wright C, Nichols E, et al. Global, regional, and national burden of meningitis, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology. 2018 Dec 1;17(12):1061–82.

15. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet. 2020 Jan 18;395(10219):200–11.

16. Ritchie H, Roser M. Causes of Death. Our World in Data [Internet]. 2018 Feb 14 [cited 2022 Jan 2]; Available from: https://ourworldindata.org/causes-of-death

17. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet. 2020 Oct 17;396(10258):1204–22.

18. Groenwold RHH, Hoes AW, Hak E. Impact of influenza vaccination on mortality risk among the elderly. European Respiratory Journal. 2009 Jul 1;34(1):56–62.

19. Timbrook TT, Morton JB, McConeghy KW, Caffrey AR, Mylonakis E, LaPlante KL. The Effect of Molecular Rapid Diagnostic Testing on Clinical Outcomes in Bloodstream Infections: A Systematic Review and Meta-analysis. Clin Infect Dis. 2017 Jan 1;64(1):15–23.

20. Singer M, Nambiar S, Valappil T, Higgins K, Gitterman S. Historical and Regulatory Perspectives on the Treatment Effect of Antibacterial Drugs for Community-Acquired Pneumonia. Clinical Infectious Diseases. 2008 Dec 1;47(Supplement_3):S216–24.

21. Alhashash F, Weston V, Diggle M, McNally A. Multidrug-resistant Escherichia coli bacteremia. Emerg Infect Dis. 2013 Oct;19(10):1699–701.

22. Boucher HW, Corey GR. Epidemiology of Methicillin-Resistant Staphylococcus aureus. Clinical Infectious Diseases. 2008 Jun 1;46(Supplement_5):S344–9.

23. O'Neill J. Antimicrobial resistance : tackling a crisis for the health and wealth of nations / the Review on Antimicrobial Resistance chaired by Jim O'Neill. [Internet]. Wellcome Collection. 2014 [cited 2022 Jan 2]. Available from: https://wellcomecollection.org/works/rdpck35v

24. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet Infect Dis. 2019 Jan;19(1):56–66.

25. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. 2022 Jan;S0140673621027240.

26. Fu Y, Zhang W, Wang H, Zhao S, Chen Y, Meng F, et al. Specific patterns of gyr A mutations determine the resistance difference to ciprofloxacin and levofloxacin in Klebsiella pneumoniae and Escherichia coli. BMC Infectious Diseases. 2013 Jan 7;13(1):8.

27. Benz F, Huisman JS, Bakkeren E, Herter JA, Stadler T, Ackermann M, et al. Plasmid- and strain-specific factors drive variation in ESBL-plasmid spread in vitro and in vivo. The ISME Journal. 2021 Mar;15(3):862–78.

28. Woodford N, Ellington MJ. The emergence of antibiotic resistance by mutation. Clinical Microbiology and Infection. 2007 Jan 1;13(1):5–18.

29. Lanza VF, Toro M de, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, et al. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. PLOS Genetics. 2014 Dec 18;10(12):e1004766.

30. Mathers AJ, Peirano G, Pitout JDD. The Role of Epidemic Resistance Plasmids and International High-Risk Clones in the Spread of Multidrug-Resistant Enterobacteriaceae. Clinical Microbiology Reviews [Internet]. 2015 Apr 29 [cited 2022 Jan 19]; Available from: https://journals.asm.org/doi/abs/10.1128/CMR.00116-14

31. Carattoli A, Villa L, Fortini D, García-Fernández A. Contemporary IncI1 plasmids involved in the transmission and spread of antimicrobial resistance in Enterobacteriaceae. Plasmid. 2021 Nov 1;118:102392.

32. Pedersen T, Sekyere JO, Govinden U, Moodley K, Sivertsen A, Samuelsen Ø, et al. Spread of Plasmid-Encoded NDM-1 and GES-5 Carbapenemases among Extensively Drug-Resistant and Pandrug-Resistant Clinical Enterobacteriaceae in Durban, South Africa. Antimicrob Agents Chemother. 2018 May;62(5):e02178-17.

33. Stadler T, Meinel D, Aguilar-Bultet L, Huisman JS, Schindler R, Egli A, et al. Transmission of ESBL-producing Enterobacteriaceae and their mobile genetic elements—identification of sources by whole genome sequencing: study protocol for an observational study in Switzerland. BMJ Open. 2018 Feb 17;8(2):e021823.

34. Tonder AJ van, Mistry S, Bray JE, Hill DMC, Cody AJ, Farmer CL, et al. Defining the Estimated Core Genome of Bacterial Populations Using a Bayesian Decision Model. PLOS Computational Biology. 2014 Aug 21;10(8):e1003788.

35. Sela U, Euler CW, Rosa JC da, Fischetti VA. Strains of bacterial species induce a greatly varied acute adaptive immune response: The contribution of the accessory genome. PLOS Pathogens. 2018 Jan 11;14(1):e1006726.

36. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Current Opinion in Genetics & Development. 2005 Dec 1;15(6):589–94.

37. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017 Mar 28;2(4):1–5.

38. Lovewell RR, Baer CE, Mishra BB, Smith CM, Sassetti CM. Granulocytes act as a niche for Mycobacterium tuberculosis growth. Mucosal Immunol. 2021 Jan;14(1):229–41.

39. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun. 2018 Oct 17;9(1):4306.

40. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biology. 2020 Jul 22;21(1):180.

41. Lees JA, Mai TT, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, et al. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. mBio [Internet]. 2020 Jul 7 [cited 2022 Jan 2]; Available from: https://journals.asm.org/doi/abs/10.1128/mBio.01344-20

42. van Elsas JD, Semenov AV, Costa R, Trevors JT. Survival of Escherichia coli in the environment: fundamental and public health aspects. ISME J. 2011 Feb;5(2):173–83.

43. Touchon M, Perrin A, Sousa JAM de, Vangchhia B, Burn S, O'Brien CL, et al. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLOS Genetics. 2020 Jun 12;16(6):e1008866.

44. Horesh G, Blackwell GA, Tonkin G, Corander J, Heinz E, Thomson NR. A comprehensive and high--quality collection of Escherichia coli genomes and their genes. Microbial Genomics. :15.

45. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. PLOS Genetics. 2019 Apr 15;15(4):e1008114.

46. San Millan A, Toll-Riera M, Qi Q, MacLean RC. Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. Nat Commun. 2015 Apr 21;6(1):6845.

47. Hall JPJ, Brockhurst MA, Harrison E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017 Dec 5;372(1735):20160424.

48. Osthoff M, Gürtler N, Bassetti S, Balestra G, Marsch S, Pargger H, et al. Impact of MALDI-TOF-MS-based identification directly from positive blood cultures on patient management: a controlled clinical trial. Clinical Microbiology and Infection. 2017 Feb 1;23(2):78–85.

49. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med. 2006 Jun;34(6):1589–96.

50. Seki M, Gotoh K, Nakamura S, Akeda Y, Yoshii T, Miyaguchi S, et al. Fatal sepsis caused by an unusual Klebsiella species that was misidentified by an automated identification system. Journal of Medical Microbiology. 62(5):801–3.

51. Fontana L, Bonura E, Lyski Z, Messer W. The Brief Case: Klebsiella variicola—Identifying the Misidentified. Journal of Clinical Microbiology. 2019 Jan 1;57(1):e00826-18.

52. Marín M, Cercenado E, Sánchez-Carrillo C, Ruiz A, Gómez González Á, Rodríguez-Sánchez B, et al. Accurate Differentiation of Streptococcus pneumoniae from other Species within the Streptococcus mitis Group by Peak Analysis Using MALDI-TOF MS. Front Microbiol [Internet]. 2017 Apr 25 [cited 2018 Apr 11];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5403922/

53. Lam MMC, Wyres KL, Duchêne S, Wick RR, Judd LM, Gan Y-H, et al. Population genomics of hypervirulent Klebsiella pneumoniae clonal-group 23 reveals early emergence and rapid global dissemination. Nature Communications [Internet]. 2018 Dec [cited 2019 Jan 25];9(1). Available from: http://www.nature.com/articles/s41467-018-05114-7

54. van Belkum A, Bachmann TT, Lüdke G, Lisby JG, Kahlmeter G, Mohess A, et al. Developmental roadmap for antimicrobial susceptibility testing systems. Nat Rev Microbiol. 2019 Jan;17(1):51–62.

55. Miller JM, Binnicker MJ, Campbell S, Carroll KC, Chapin KC, Gilligan PH, et al. A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiologya. Clinical Infectious Diseases. 2018 Aug 31;67(6):e1–94.

56. Carroll KC, Patel R. Systems for Identification of Bacteria and Fungi. In: Manual of Clinical Microbiology [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2022 Jan 17]. p. 29–43. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1128/9781555817381.ch4

57. Karlowsky JA, Richter SS. Antimicrobial Susceptibility Testing Systems. In: Manual of Clinical Microbiology [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2022 Jan 17]. p. 1274–85. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1128/9781555817381.ch72

58. Galar A, Leiva J, Espinosa M, Guillén-Grima F, Hernáez S, Yuste JR. Clinical and economic evaluation of the impact of rapid microbiological diagnostic testing. J Infect. 2012 Oct;65(4):302–9.

59. Scohy A, Noël A, Boeras A, Brassinne L, Laurent T, Rodriguez-Villalobos H, et al. Evaluation of the Bruker® MBT Sepsityper IVD module for the identification of polymicrobial blood cultures with MALDI-TOF MS. Eur J Clin Microbiol Infect Dis. 2018 Nov;37(11):2145–52.

60. Rossen JWA, Friedrich AW, Moran-Gilad J, ESCMID Study Group for Genomic and Molecular Diagnostics (ESGMD). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin Microbiol Infect. 2018 Apr;24(4):355–60.

61. Deplano A, Dodémont M, Denis O, Westh H, Gumpert H, Larsen AR, et al. European external quality assessments for identification, molecular typing and characterization of Staphylococcus aureus. Journal of Antimicrobial Chemotherapy. 2018 Oct 1;73(10):2662–6.

62. Slotved H-C, Sheppard CL, Dalby T, van der Ende A, Fry NK, Morfeldt E, et al. External Quality Assurance for Laboratory Identification and Capsular Typing of Streptococcus pneumoniae. Sci Rep. 2017 Oct 16;7(1):13280.

63. Angeletti S. Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology. Journal of Microbiological Methods. 2017 Jul 1;138:20–9.

64. Dierig A, Frei R, Egli A. The Fast Route to Microbe Identification: Matrix Assisted Laser Desorption/Ionization—Time of Flight Mass Spectrometry (MALDI-TOF MS). The Pediatric Infectious Disease Journal. 2015 Jan;34(1):97–9.

65. Ling H, Yuan Z, Shen J, Wang Z, Xu Y. Accuracy of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for Identification of Clinical Pathogenic Fungi: a Meta-Analysis. J Clin Microbiol. 2014 Jul 1;52(7):2573–82.

66. Sogawa K, Watanabe M, Sato K, Segawa S, Ishii C, Miyabe A, et al. Use of the MALDI BioTyper system with MALDI–TOF mass spectrometry for rapid identification of microorganisms. Anal Bioanal Chem. 2011 Jun 1;400(7):1905.

67. Khot PD, Couturier MR, Wilson A, Croft A, Fisher MA. Optimization of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry Analysis for Bacterial Identification. J Clin Microbiol. 2012 Dec;50(12):3845–52.

68. Egli A, Tschudin-Sutter S, Oberle M, Goldenberger D, Frei R, Widmer AF. Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass-Spectrometry (MALDI-TOF MS) Based Typing of Extended-Spectrum β-Lactamase Producing E. coli – A Novel Tool for Real-Time Outbreak Investigation. PLoS One [Internet]. 2015 Apr 10 [cited 2018 Mar 9];10(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393243/

69. Kassim A, Pflüger V, Premji Z, Daubenberger C, Revathi G. Comparison of biomarker based Matrix Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) and conventional methods in the identification of clinically relevant bacteria and yeast. BMC Microbiol [Internet]. 2017 May 25 [cited 2018 Aug 29];17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5445374/

70. Ban S, Kasaishi R, Kamijo T, Noritake C, Kawasaki H. An exploratory MALDI-TOF MS library based on SARAMIS superspectra for rapid identification of *Aspergillus* section *Nigri*. Mycoscience. 2021 Jul 20;62(4):224–32.

71. Stephan R, Cernela N, Ziegler D, Pflüger V, Tonolla M, Ravasi D, et al. Rapid species specific identification and subtyping of Yersinia enterocolitica by MALDI-TOF Mass spectrometry. Journal of Microbiological Methods. 2011 Nov;87(2):150–3.

72. Faron ML, Buchan BW, Hyke J, Madisen N, Lillie JL, Granato PA, et al. Multicenter Evaluation of the Bruker MALDI Biotyper CA System for the Identification of Clinical Aerobic Gram-Negative Bacterial Isolates. PLoS One. 2015 Nov 3;10(11):e0141350.

73. Lévesque S, Dufresne PJ, Soualhine H, Domingo M-C, Bekal S, Lefebvre B, et al. A Side by Side Comparison of Bruker Biotyper and VITEK MS: Utility of MALDI-TOF MS Technology for Microorganism Identification in a Public Health Reference Laboratory. PLOS ONE. 2015 Dec 10;10(12):e0144878.

74. Bridel S, Watts SC, Judd LM, Harshegyi T, Passet V, Rodrigues C, et al. Klebsiella MALDI TypeR: a web-based tool for Klebsiella identification based on MALDI-TOF mass spectrometry. Res Microbiol. 2021 Aug;172(4–5):103835.

75. Matsumura Y, Yamamoto M, Nagao M, Tanaka M, Machida K, Ito Y, et al. Detection of extended-spectrum-β-lactamase-producing Escherichia coli ST131 and ST405 clonal groups by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol. 2014 Apr;52(4):1034–40.

76. Rothen J, Pothier JF, Foucault F, Blom J, Nanayakkara D, Li C, et al. Subspecies Typing of Streptococcus agalactiae Based on Ribosomal Subunit Protein Mass Variation by MALDI-TOF MS. Frontiers in Microbiology [Internet]. 2019 Mar 11 [cited 2019 May 17];10. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2019.00471/full

77. Rhoads DD, Wang H, Karichu J, Richter SS. The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in staphylococci. Diagnostic Microbiology and Infectious Disease. 2016 Nov;86(3):257–61.

78. Lau AF, Wang H, Weingarten RA, Drake SK, Suffredini AF, Garfield MK, et al. A Rapid Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry-Based Method for Single-Plasmid Tracking in an Outbreak of Carbapenem-Resistant Enterobacteriaceae. J Clin Microbiol. 2014 Aug;52(8):2804–12.

79. Hu Y, Huang Y, Lizou Y, Li J, Zhang R. Evaluation of Staphylococcus aureus Subtyping Module for Methicillin-Resistant Staphylococcus aureus Detection Based on Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry. Front Microbiol [Internet]. 2019 Oct 31 [cited 2020 Jun 8];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6834645/

80. Fenselau C, Demirev PA. Characterization of intact microorganisms by MALDI mass spectrometry. Mass Spectrom Rev. 2001 Aug;20(4):157–71.

81. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, et al. Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics. 2008 Feb 27;9:102.

82. Arnold RJ, Reilly JP. Observation of Escherichia coli ribosomal proteins and their posttranslational modifications by mass spectrometry. Anal Biochem. 1999 Apr 10;269(1):105–12.

83. Ojima-Kato T, Yamamoto N, Suzuki M, Fukunaga T, Tamura H. Discrimination of Escherichia coli O157, O26 and O111 from Other Serovars by MALDI-TOF MS Based on the S10-GERMS Method. PLOS ONE. 2014 Nov 20;9(11):e113458.

84. Tamura H, Hotta Y, Sato H. Novel accurate bacterial discrimination by MALDI-time-of-flight MS based on ribosomal proteins coding in S10-spc-alpha operon at strain level S10-GERMS. J Am Soc Mass Spectrom. 2013 Aug;24(8):1185–93.

85. Ziegler D, Pothier JF, Ardley J, Fossou RK, Pflüger V, de Meyer S, et al. Ribosomal protein biomarkers provide root nodule bacterial identification by MALDI-TOF MS. Appl Microbiol Biotechnol. 2015 Jul;99(13):5547–62.

86. Anantharajah A, Tossens B, Olive N, Kabamba-Mukadi B, Rodriguez-Villalobos H, Verroken A. Performance Evaluation of the MBT STAR®-Carba IVD Assay for the Detection of Carbapenemases With MALDI-TOF MS. Front Microbiol [Internet]. 2019 [cited 2019 Aug 23];10. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01413/full

87. Cordovana M, Abdalla M, Ambretti S. Evaluation of the MBT STAR-Carba Assay for the Detection of Carbapenemase Production in Enterobacteriaceae and Hafniaceae with a Large Collection of Routine Isolates from Plate Cultures and Patient-Derived Positive Blood Cultures. Microbial Drug Resistance. 2020 Nov 1;26(11):1298–306.

88. Idelevich EA, Sparbier K, Kostrzewa M, Becker K. Rapid detection of antibiotic resistance by MALDI-TOF mass spectrometry using a novel direct-on-target microdroplet growth assay. Clin Microbiol Infect. 2017 Oct 24;

89. Idelevich EA, Storck LM, Sparbier K, Drews O, Kostrzewa M, Becker K. Rapid Direct Susceptibility Testing from Positive Blood Cultures by the Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry-Based Direct-on-Target Microdroplet Growth Assay. J Clin Microbiol. 2018;56(10).

90. Wang H-Y, Lee T-Y, Tseng Y-J, Liu T-P, Huang K-Y, Chang Y-T, et al. A new scheme for strain typing of methicillin-resistant Staphylococcus aureus on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach. Becker K, editor. PLOS ONE. 2018 Mar 13;13(3):e0194289.

91. Feucherolles M, Cauchie H-M, Penny C. MALDI-TOF Mass Spectrometry and Specific Biomarkers: Potential New Key for Swift Identification of Antimicrobial Resistance in Foodborne Pathogens. Microorganisms [Internet]. 2019 Nov 21 [cited 2020 Jun 8];7(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6955786/

92. Josten M, Dischinger J, Szekat C, Reif M, Al-Sabti N, Sahl H-G, et al. Identification of agr-positive methicillin-resistant Staphylococcus aureus harbouring the class A mec complex by MALDI-TOF mass spectrometry. International Journal of Medical Microbiology. 2014 Nov;304(8):1018–23.

93. Lafolie J, Sauget M, Cabrolier N, Hocquet D, Bertrand X. Detection of Escherichia coli sequence type 131 by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: implications for infection control policies? Journal of Hospital Infection. 2015 Jul;90(3):208–12.

94. Zhang T, Ding J, Rao X, Yu J, Chu M, Ren W, et al. Analysis of methicillin-resistant Staphylococcus aureus major clonal lineages by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry (MALDI–TOF MS). Journal of Microbiological Methods. 2015 Oct;117:122–7.

95. Huang T-S, Lee SS-J, Lee C-C, Chang F-C. Detection of carbapenem-resistant Klebsiella pneumoniae on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach. PLoS One [Internet]. 2020 Feb 6 [cited 2020 Jun 8];15(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7004327/

96. Mather CA, Werth BJ, Sivagnanam S, SenGupta DJ, Butler-Wu SM. Rapid Detection of Vancomycin Intermediate Staphylococcus aureus (VISA) by Matrix Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS). Journal of Clinical Microbiology. 2016 Jan 13;JCM.02428-15.

97. Oberle M, Wohlwend N, Jonas D, Maurer FP, Jost G, Tschudin-Sutter S, et al. The Technical and Biological Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Based Typing: Employment of Bioinformatics in a Multicenter Study. PLOS ONE. 2016 Oct 31;11(10):e0164260.

98. Croxatto A, Prod'hom G, Greub G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiology Reviews. 2012 Mar 1;36(2):380–407.

99. Wittwer M, Lasch P, Drevinek M, Schmoldt S, Indra A, Jacob D, et al. First Report: Application of MALDI-TOF MS within an External Quality Assurance Exercise for the Discrimination of Highly Pathogenic Bacteria from Contaminant Flora. Applied Biosafety. 2012 Jun 1;17(2):59–63.

100. Branda JA, Fritsche TR, Burnham C-A, Butler-Wu S, Doern C, Doing KM, et al. M58-Methods for the Identification of Cultured Microorganisms Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. 2017.

101. von Mentzer A, Blackwell GA, Pickard D, Boinett CJ, Joffré E, Page AJ, et al. Long-read-sequenced reference genomes of the seven major lineages of enterotoxigenic Escherichia coli (ETEC) circulating in modern time. Sci Rep. 2021 Apr 29;11(1):9256.

102. McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, et al. Diversification of Colonization Factors in a Multidrug-Resistant Escherichia coli Lineage Evolving under Negative Frequency-Dependent Selection. mBio. 10(2):e00644-19.

103. Wyres KL, Holt KE. Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones. Trends in Microbiology. 2016 Dec 1;24(12):944–56.

104. Brisse S, Passet V, Grimont PAD. Description of Klebsiella quasipneumoniae sp. nov., isolated from human infections, with two subspecies, Klebsiella quasipneumoniae subsp. quasipneumoniae subsp. nov. and Klebsiella quasipneumoniae subsp. similipneumoniae subsp. nov., and demonstration that Klebsiella singaporensis is a junior heterotypic synonym of Klebsiella variicola. Int J Syst Evol Microbiol. 2014 Sep;64(Pt 9):3146–52.

105. Merla C, Rodrigues C, Passet V, Corbella M, Thorpe HA, Kallonen TVS, et al. Description of Klebsiella spallanzanii sp. nov. and of Klebsiella pasteurii sp. nov. Front Microbiol [Internet]. 2019 [cited 2021 Jan 14];10. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2019.02360/full

106. Rosenblueth M, Martínez L, Silva J, Martínez-Romero E. Klebsiella variicola, a novel species with clinical and plant-associated isolates. Syst Appl Microbiol. 2004 Feb;27(1):27–35.

107. Saha R, Farrance CE, Verghese B, Hong S, Donofrio RS. Klebsiella michiganensis sp. nov., a new bacterium isolated from a tooth brush holder. Curr Microbiol. 2013 Jan;66(1):72–8.

108. Hu Y, Wei L, Feng Y, Xie Y, Zong Z. Klebsiella huaxiensis sp. nov., recovered from human urine. International Journal of Systematic and Evolutionary Microbiology. 2019;69(2):333–6.

109. Passet V, Brisse S. Description of Klebsiella grimontii sp. nov. International Journal of Systematic and Evolutionary Microbiology. 2018;68(1):377–81.

110. Gujarati S, Chaudhari D, Hagir A, Khairnar M, Shouche Y, Rahi P 2020. Klebsiella indica sp. nov., isolated from the surface of a tomato. International Journal of Systematic and Evolutionary Microbiology. 70(5):3278–86.

111. Long SW, Linson SE, Saavedra MO, Cantu C, Davis JJ, Brettin T, et al. Whole-Genome Sequencing of a Human Clinical Isolate of the Novel Species Klebsiella quasivariicola sp. nov. Genome Announc. 2017 Oct 19;5(42):e01057-17.

112. Rodrigues C, Passet V, Rakotondrasoa A, Diallo TA, Criscuolo A, Brisse S. Description of Klebsiella africanensis sp. nov., Klebsiella variicola subsp. tropicalensis subsp. nov. and Klebsiella variicola subsp. variicola subsp. nov. Research in Microbiology. 2019 Apr 1;170(3):165–70.

113. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002 Jul 15;30(14):3059–66.

114. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.

115. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun [Internet]. 2018 Nov 30 [cited 2019 May 8];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6269478/

116. Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae. Nat Rev Microbiol. 2020 Jun;18(6):344–59.

117. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. Genomic surveillance framework and global population structure for *Klebsiella pneumoniae* [Internet]. Genomics; 2020 Dec [cited 2021 May 20]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.12.14.422303

118. Brisse S, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize R, Diancourt L, et al. Virulent clones of Klebsiella pneumoniae: identification and evolutionary scenario based on genomic and phenotypic characterization. PLoS One. 2009;4(3):e4982.

119. Yang J, Long H, Hu Y, Feng Y, McNally A, Zong Z. Klebsiella oxytoca Complex: Update on Taxonomy, Antimicrobial Resistance, and Virulence. Clin Microbiol Rev. 2021 Dec 1;e0000621.

120. Ma Y, Wu X, Li S, Tang L, Chen M, An Q. Proposal for reunification of the genus Raoultella with the genus Klebsiella and reclassification of Raoultella electrica as Klebsiella electrica comb. nov. Research in Microbiology. 2021 Sep 1;172(6):103851.

121. Tindall BJ, Sutton G, Garrity GM. Enterobacter aerogenes Hormaeche and Edwards 1960 (Approved Lists 1980) and Klebsiella mobilis Bascomb et al. 1971 (Approved Lists 1980) share the same nomenclatural type (ATCC 13048) on the Approved Lists and are homotypic synonyms, with consequences for the name Klebsiella mobilis Bascomb et al. 1971 (Approved Lists 1980). International Journal of Systematic and Evolutionary Microbiology. 2017;67(2):502–4.

122. Jacoby GA. AmpC beta-lactamases. Clin Microbiol Rev. 2009 Jan;22(1):161–82, Table of Contents.

123. Carter JS, Bowden FJ, Bastian I, Myers GM, Sriprakash KS, Kemp DJ. Phylogenetic evidence for reclassification of Calymmatobacterium granulomatis as Klebsiella granulomatis comb. nov. Int J Syst Bacteriol. 1999 Oct;49 Pt 4:1695–700.

124. Podschun R, Pietsch S, Höller C, Ullmann U. Incidence of Klebsiella Species in Surface Waters and Their Expression of Virulence Factors. Appl Environ Microbiol. 2001 Jul;67(7):3325–7.

125. Podschun R, Ullmann U. Klebsiella spp. as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors. Clin Microbiol Rev. 1998 Oct;11(4):589–603.

126. Long SW, Olsen RJ, Eagar TN, Beres SB, Zhao P, Davis JJ, et al. Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing Klebsiella pneumoniae Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307. mBio. 2017 Jul 5;8(3):e00489-17.

127. Potter RF, Lainhart W, Twentyman J, Wallace MA, Wang B, Burnham C-AD, et al. Population Structure, Antibiotic Resistance, and Uropathogenicity of Klebsiella variicola. mBio [Internet]. 2018 Dec 18 [cited 2019 Jan 11];9(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6299229/

128. Gorrie CL, Mirceta M, Wick RR, Judd LM, Wyres KL, Thomson NR, et al. Antimicrobial resistant Klebsiella pneumoniae carriage and infection in specialized geriatric care wards linked to acquisition in the referring hospital. Clin Infect Dis [Internet]. [cited 2018 Jan 25]; Available from: https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciy027/4798841

129. Gorrie CL, Mirceta M, Wick RR, Judd LM, Lam MMC, Gomi R, et al. Genomic dissection of the bacterial population underlying *Klebsiella pneumoniae* infections in hospital patients: insights into an opportunistic pathogen [Internet]. Infectious Diseases (except HIV/AIDS); 2021 Dec [cited 2022 Jan 2]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.12.02.21267161

130. Herzog KAT, Schneditz G, Leitner E, Feierl G, Hoffmann KM, Zollner-Schwetz I, et al. Genotypes of Klebsiella oxytoca Isolates from Patients with Nosocomial Pneumonia Are Distinct from Those of Isolates from Patients with Antibiotic-Associated Hemorrhagic Colitis. J Clin Microbiol. 2014 May 1;52(5):1607–16.

131. Schneditz G, Rentner J, Roier S, Pletz J, Herzog KAT, Bücker R, et al. Enterotoxicity of a nonribosomal peptide causes antibiotic-associated colitis. PNAS. 2014 Sep 9;111(36):13181–6.

132. Schneditz G, Rentner J, Herzog KA, Gorkiewicz G, Roier S, Schild S, et al. It takes two to tango: Klebsiella oxytoca overgrowth and production of Tilivalline, a cytotoxin, are the key events causing Antibiotic-Associated Hemorrhagic Colitis. Z Gastroenterol. 2013 May;51(5):A50.

133. Puerta-Fernandez S, Miralles-Linares F, Sanchez-Simonet MV, Bernal-Lopez MR, Gomez-Huelgas R. Raoultella planticola bacteraemia secondary to gastroenteritis. Clinical Microbiology and Infection. 2013 May 1;19(5):E236–7.

134. Imai K, Ishibashi N, Kodana M, Tarumoto N, Sakai J, Kawamura T, et al. Clinical characteristics in blood stream infections caused by Klebsiella pneumoniae,

Klebsiella variicola, and Klebsiella quasipneumoniae: a comparative study, Japan, 2014-2017. BMC Infect Dis. 2019 Nov 8;19(1):946.

135. Maatallah M, Vading M, Kabir MH, Bakhrouf A, Kalin M, Nauclér P, et al. Klebsiella variicola is a frequent cause of bloodstream infection in the stockholm area, and associated with higher mortality compared to K. pneumoniae. PLoS ONE. 2014;9(11):e113539.

136. Fostervold A, Hetland MAK, Bakksjø R, Bernhoff E, Holt KE, Samuelsen Ø, et al. A nationwide genomic study of clinical Klebsiella pneumoniae in Norway 2001-15: introduction and spread of ESBLs facilitated by clonal groups CG15 and CG307. J Antimicrob Chemother. 2021 Dec 22;dkab463.

137. Luo Y, Wang Y, Ye L, Yang J. Molecular epidemiology and virulence factors of pyogenic liver abscess causing Klebsiella pneumoniae in China. Clin Microbiol Infect. 2014 Nov;20(11):O818-824.

138. Liao CH, Huang YT, Chang CY, Hsu HS, Hsueh PR. Capsular serotypes and multilocus sequence types of bacteremic Klebsiella pneumoniae isolates associated with different types of infections. Eur J Clin Microbiol Infect Dis. 2014 Mar;33(3):365–9.

139. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. Multilocus Sequence Typing of Klebsiella pneumoniae Nosocomial Isolates. J Clin Microbiol. 2005 Aug 1;43(8):4178–82.

140. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis. 2018;18(3):318–27.

141. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. PNAS. 2015 Jul 7;112(27):E3574–81.

142. Chaves J, Ladona MG, Segura C, Coira A, Reig R, Ampurdanés C. SHV-1 beta-lactamase is mainly a chromosomally encoded species-specific enzyme in Klebsiella pneumoniae. Antimicrob Agents Chemother. 2001 Oct;45(10):2856–61.

143. Sirot J, Chanal C, Petit A, Sirot D, Labia R, Gerbaud G. Klebsiella pneumoniae and other Enterobacteriaceae producing novel plasmid-mediated beta-lactamases markedly active against third-generation cephalosporins: epidemiologic studies. Rev Infect Dis. 1988 Aug;10(4):850–9.

144. Nordmann P, Cuzon G, Naas T. The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria. Lancet Infect Dis. 2009 Apr;9(4):228–36.

145. Nordmann P, Poirel L, Walsh TR, Livermore DM. The emerging NDM carbapenemases. Trends Microbiol. 2011 Dec;19(12):588–95.

146. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis. 2016 Feb;16(2):161–8.

147. Bowers JR, Kitchel B, Driebe EM, MacCannell DR, Roe C, Lemmer D, et al. Genomic Analysis of the Emergence and Rapid Global Dissemination of the Clonal Group 258 Klebsiella pneumoniae Pandemic. PLOS ONE. 2015 Jul 21;10(7):e0133727.

148. . Wyres KL, Holt KE. Klebsiella pneumoniae as a key trafficker of drug resistance genes from environmental to clinically important bacteria. Current Opinion in Microbiology. 2018 Oct 1;45:131–9.

149. Turton JF, Payne Z, Coward A, Hopkins KL, Turton JA, Doumith M, et al. Virulence genes in isolates of Klebsiella pneumoniae from the UK during 2016, including among carbapenemase gene-positive hypervirulent K1-ST23 and 'non-hypervirulent' types ST147, ST15 and ST383. Journal of Medical Microbiology. 67(1):118–28.

150. Lam MMC, Wyres KL, Wick RR, Judd LM, Fostervold A, Holt KE, et al. Convergence of virulence and MDR in a single plasmid vector in MDR Klebsiella pneumoniae ST15. Journal of Antimicrobial Chemotherapy. 2019 May 1;74(5):1218–22.

151. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat Commun. 2021 Jul 7;12(1):4188.

152. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the Escherichia coli phylogenetic group. Appl Environ Microbiol. 2000 Oct;66(10):4555–8.

153. Clermont Olivier, Christenson Julia K., Denamur Erick, Gordon David M. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environmental Microbiology Reports. 2013 Feb 3;5(1):58–65.

154. van der Putten BCL, Matamoros S, Mende DR, Scholl ER, consortium† C, Schultsz CY 2021. Escherichia ruysiae sp. nov., a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller. International Journal of Systematic and Evolutionary Microbiology. 71(2):004609.

155. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology [Internet]. 2016 Dec [cited 2019 May 8];17(1). Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x

156. Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS, et al. Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. Commun Biol. 2021 Jan 26;4(1):1–12.

157. Kotloff KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, Sansonetti PJ, et al. Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. Bull World Health Organ. 1999;77(8):651–66.

158. Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of Shigella species. PLOS Genetics. 2020 Jul 9;16(7):e1008931.

159. Pettengill EA, Pettengill JB, Binet R. Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological

Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. Front Microbiol. 2015;6:1573.

160. Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature. 2000 Jul;406(6791):64–7.

161. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, et al. Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains. BMC Genomics. 2008 Nov 26;9:560.

162. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol. 2006 Jun;60(5):1136–51.

163. Denamur E, Clermont O, Gordon D. Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. Microbiology. 2015 May 1;161(5):980–8.

164. Sahl JW, Matalka MN, Rasko DA. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl Environ Microbiol. 2012 Jul;78(14):4884–92.

165. Liu S, Jin D, Lan R, Wang Y, Meng Q, Dai H, et al. Escherichia marmotae sp. nov., isolated from faeces of Marmota himalayana. Int J Syst Evol Microbiol. 2015 Jul;65(7):2130–4.

166. Riley M, Abe T, Arnaud MB, Berlyn MKB, Blattner FR, Chaudhuri RR, et al. Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Research. 2006 Jan 1;34(1):1–9.

167. Song Y, Lee B-R, Cho S, Cho Y-B, Kim S-W, Kang TJ, et al. Determination of single nucleotide variants in Escherichia coli DH5α by using short-read sequencing. FEMS Microbiology Letters. 2015 Jun 1;362(11):fnv073.

168. Blount ZD. The unexhausted potential of E. coli. eLife. 2015 Mar 25;4:e05826.

169. O'Brien EJ, Utrilla J, Palsson BO. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. PLOS Computational Biology. 2016 Jun 28;12(6):e1004998.

170. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent Advances in Understanding Enteric Pathogenic Escherichia coli. Clin Microbiol Rev. 2013 Oct;26(4):822–80.

171. Deborah Chen H, Frankel G. Enteropathogenic Escherichia coli: unravelling pathogenesis. FEMS Microbiology Reviews. 2005 Jan 1;29(1):83–98.

172. Qadri F, Svennerholm A-M, Faruque ASG, Sack RB. Enterotoxigenic Escherichia coli in Developing Countries: Epidemiology, Microbiology, Clinical Features, Treatment, and Prevention. Clin Microbiol Rev. 2005 Jul;18(3):465–83.

173. Goldwater PN, Bettelheim KA. Treatment of enterohemorrhagic Escherichia coli (EHEC) infection and hemolytic uremic syndrome (HUS). BMC Medicine. 2012 Feb 2;10(1):12.

174. Kaur P, Chakraborti A, Asea A. Enteroaggregative Escherichia coli: An Emerging Enteric Food Borne Pathogen. Interdiscip Perspect Infect Dis. 2010;2010:254159.

175. Mansan-Almeida R, Pereira AL, Giugliano LG. Diffusely adherent Escherichia colistrains isolated from children and adults constitute two different populations. BMC Microbiology. 2013 Feb 1;13(1):22.

176. Lee JG, Han DS, Jo SV, Lee AR, Park CH, Eun CS, et al. Characteristics and pathogenic role of adherent-invasive Escherichia coli in inflammatory bowel disease: Potential impact on clinical outcomes. PLOS ONE. 2019 Apr 29;14(4):e0216165.

177. van den Beld MJC, Warmelink E, Friedrich AW, Reubsaet FAG, Schipper M, de Boer RF, et al. Incidence, clinical implications and impact on public health of infections with Shigella spp. and entero-invasive Escherichia coli (EIEC): results of a multicenter cross-sectional study in the Netherlands during 2016–2017. BMC Infectious Diseases. 2019 Dec 9;19(1):1037.

178. Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J, Tauschek M. Are Escherichia coli Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? Frontiers in Cellular and Infection Microbiology [Internet]. 2016 Nov 18 [cited 2018 Jul 30];6. Available from: http://journal.frontiersin.org/article/10.3389/fcimb.2016.00141/full

179. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. Clinical Microbiology Reviews [Internet]. 2019 Jun 12 [cited 2022 Jan 17]; Available from: https://journals.asm.org/doi/abs/10.1128/CMR.00135-18

180. Terlizzi ME, Gribaudo G, Maffei ME. UroPathogenic Escherichia coli (UPEC) Infections: Virulence Factors, Bladder Responses, Antibiotic, and Non-antibiotic Antimicrobial Strategies. Front Microbiol [Internet]. 2017 Aug 15 [cited 2018 Jan 24];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5559502/

181. Stoppe N de C, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM, et al. Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates. Front Microbiol. 2017 Dec 21;8:2512.

182. Lim JY, Yoon JW, Hovde CJ. A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O157. J Microbiol Biotechnol. 2010 Jan;20(1):5–14.

183. Hutton TA, Innes GK, Harel J, Garneau P, Cucchiara A, Schifferli DM, et al. Phylogroup and virulence gene association with clinical characteristics of Escherichia coli urinary tract infections from dogs and cats. J Vet Diagn Invest. 2018 Jan;30(1):64–70.

184. Foxman B. Urinary Tract Infection Syndromes. Infectious Disease Clinics of North America. 2014 Mar;28(1):1–13.

185. Bien J, Sokolova O, Bozko P. Role of Uropathogenic *Escherichia coli* Virulence Factors in Development of Urinary Tract Infection and Kidney Damage. International Journal of Nephrology. 2012;2012:1–15.

186. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. Nature Reviews Microbiology. 2015 May;13(5):269–84.

187. Foxman B. The epidemiology of urinary tract infection. Nat Rev Urol. 2010 Dec;7(12):653–60.

188. Bennett JE, Dolin R, Blaser MJ, editors. Mandell, Douglas, and Bennett's infectious disease essentials. Philadelphia, PA: Elsevier; 2017.

189. Thorpe A, Neal D. Benign prostatic hyperplasia. The Lancet. 2003 Apr 19;361(9366):1359–67.

190. Houamel D, Ducrot N, Lefebvre T, Daher R, Moulouel B, Sari M-A, et al. Hepcidin as a Major Component of Renal Antibacterial Defenses against Uropathogenic Escherichia coli. JASN. 2016 Mar 1;27(3):835–46.

191. Shah C, Baral R, Bartaula B, Shrestha LB. Virulence factors of uropathogenic Escherichia coli (UPEC) and correlation with antimicrobial resistance. BMC Microbiology. 2019 Sep 2;19(1):204.

192. Sandberg T, Kaijser B, Lidin-Janson G, Lincoln K, Orskov F, Orskov I, et al. Virulence of Escherichia coli in relation to host factors in women with symptomatic urinary tract infection. Journal of Clinical Microbiology [Internet]. 1988 Aug [cited 2022 Jan 13]; Available from: https://journals.asm.org/doi/abs/10.1128/jcm.26.8.1471-1476.1988

193. Pitout JDD. Extraintestinal Pathogenic Escherichia coli: A Combination of Virulence with Antibiotic Resistance. Front Microbiol [Internet]. 2012 Jan 19 [cited 2018 Jul 30];3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261549/

194. Leimbach A, Hacker J, Dobrindt U. <Emphasis Type="Italic">E. coli</Emphasis> as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity. In: Between Pathogenicity and Commensalism [Internet]. Springer, Berlin, Heidelberg; 2013 [cited 2018 Jul 9]. p. 3–32. (Current Topics in Microbiology and Immunology). Available from: https://link.springer.com/chapter/10.1007/82_2012_303

195. Legros N, Ptascheck S, Pohlentz G, Karch H, Dobrindt U, Müthing J. PapG subtype-specific binding characteristics of Escherichia coli towards globo-series glycosphingolipids of human kidney and bladder uroepithelial cells. Glycobiology. 2019 Oct 21;29(11):789–802.

196. Dodson KW, Pinkner JS, Rose T, Magnusson G, Hultgren SJ, Waksman G. Structural Basis of the Interaction of the Pyelonephritic E. coli Adhesin to Its Human Kidney Receptor. Cell. 2001 Jun;105(6):733–43.

197. Dodson KW, Pinkner JS, Rose T, Magnusson G, Hultgren SJ, Waksman G. Structural basis of the interaction of the pyelonephritic E. coli adhesin to its human kidney receptor. Cell. 2001 Jun 15;105(6):733–43.

198. Biggel M, Xavier BB, Johnson JR, Nielsen KL, Frimodt-Møller N, Matheeussen V, et al. Horizontally acquired papGII -containing pathogenicity islands underlie the emergence of invasive uropathogenic Escherichia coli lineages. Nature Communications. 2020 Nov 24;11(1):5968.

199. Denamur E, Condamine B, Esposito-Farèse M, Royer G, Clermont O, Laouenan C, et al. Genome wide association study of human bacteremia *Escherichia coli* isolates identifies genetic determinants for the portal of entry but not fatal outcome [Internet].

Infectious Diseases (except HIV/AIDS); 2021 Nov [cited 2021 Dec 9]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.11.09.21266136

200. Strömberg N, Marklund BI, Lund B, Ilver D, Hamers A, Gaastra W, et al. Host-specificity of uropathogenic Escherichia coli depends on differences in binding specificity to Gal alpha 1-4Gal-containing isoreceptors. EMBO J. 1990 Jun;9(6):2001–10.

201. Riley LW. Pandemic lineages of extraintestinal pathogenic Escherichia coli. Clin Microbiol Infect. 2014 May;20(5):380–90.

202. McNally A, Alhashash F, Collins M, Alqasim A, Paszckiewicz K, Weston V, et al. Genomic analysis of extra-intestinal pathogenic Escherichia coli urosepsis. Clinical Microbiology and Infection. 2013 Aug 1;19(8):e328–34.

203. Klein RD, Hultgren SJ. Urinary tract infections: microbial pathogenesis, host–pathogen interactions and new treatment strategies. Nat Rev Microbiol. 2020 Apr;18(4):211–26.

204. de Kraker MEA, Jarlier V, Monen JCM, Heuer OE, van de Sande N, Grundmann H. The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System. Clin Microbiol Infect. 2013 Sep;19(9):860–8.

205. Croxall G, Hale J, Weston V, Manning G, Cheetham P, Achtman M, et al. Molecular epidemiology of extraintestinal pathogenic Escherichia coli isolates from a regional cohort of elderly patients highlights the prevalence of ST131 strains with increased antimicrobial resistance in both community and hospital care settings. J Antimicrob Chemother. 2011 Nov;66(11):2501–8.

206. McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, et al. Diversification of Colonization Factors in a Multidrug-Resistant Escherichia coli Lineage Evolving under Negative Frequency-Dependent Selection. mBio [Internet]. 2019 Apr 30 [cited 2020 Aug 12];10(2). Available from: https://mbio.asm.org/content/10/2/e00644-19

207. Kondratyeva K, Salmon-Divon M, Navon-Venezia S. Meta-analysis of Pandemic Escherichia coli ST131 Plasmidome Proves Restricted Plasmid-clade Associations. Sci Rep. 2020 Jan 8;10(1):36.

208. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. PLOS Genetics. 2016 Sep 12;12(9):e1006280.

209. Banerjee R, Johnson JR. A new clone sweeps clean: the enigmatic emergence of Escherichia coli sequence type 131. Antimicrob Agents Chemother. 2014 Sep;58(9):4997–5004.

210. Peirano G, Schreckenberger PC, Pitout JDD. Characteristics of NDM-1-producing Escherichia coli isolates that belong to the successful and virulent clone ST131. Antimicrob Agents Chemother. 2011 Jun;55(6):2986–8.

211. Zakour NLB, Alsheikh-Hussain AS, Ashcroft MM, Nhu NTK, Roberts LW, Stanton-Cook M, et al. Sequential Acquisition of Virulence and Fluoroquinolone Resistance

Has Shaped the Evolution of Escherichia coli ST131. mBio [Internet]. 2016 Apr 26 [cited 2021 Dec 30]; Available from: https://journals.asm.org/doi/abs/10.1128/mBio.00347-16

212. Stoesser N, Sheppard AE, Pankhurst L, Maio ND, Moore CE, Sebra R, et al. Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. mBio [Internet]. 2016 Mar 22 [cited 2021 Dec 30]; Available from: https://journals.asm.org/doi/abs/10.1128/mBio.02162-15

213. Tchesnokova V, Radey M, Chattopadhyay S, Larson L, Weaver JL, Kisiela D, et al. Pandemic fluoroquinolone resistant Escherichia coli clone ST1193 emerged via simultaneous homologous recombinations in 11 gene loci. PNAS. 2019 Jul 16;116(29):14740–8.

214. Wu J, Lan F, Lu Y, He Q, Li B. Molecular Characteristics of ST1193 Clone among Phylogenetic Group B2 Non-ST131 Fluoroquinolone-Resistant Escherichia coli. Frontiers in Microbiology. 2017;8:2294.

215. Elbing K, Brent R. Recipes and tools for culture of Escherichia coli. Curr Protoc Mol Biol. 2019 Jan;125(1):e83.

216. Halimeh FB, Rafei R, Osman M, Kassem II, Diene SM, Dabboussi F, et al. Historical, current, and emerging tools for identification and serotyping of Shigella. Braz J Microbiol. 2021 Dec 1;52(4):2043–55.

217. Torres-Miranda D, Akselrod H, Karsner R, Secco A, Silva-Cantillo D, Siegel MO, et al. Use of BioFire FilmArray gastrointestinal PCR panel associated with reductions in antibiotic use, time to optimal antibiotics, and length of stay. BMC Gastroenterology. 2020 Aug 6;20(1):246.

218. Gorrie CL, Mirčeta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, et al. Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients. Clin Infect Dis. 2017 Jul 15;65(2):208–15.

219. Martin RM, Bachman MA. Colonization, Infection, and the Accessory Genome of Klebsiella pneumoniae. Front Cell Infect Microbiol. 2018;8:4.

220. Siu LK, Yeh K-M, Lin J-C, Fung C-P, Chang F-Y. Klebsiella pneumoniae liver abscess: a new invasive syndrome. The Lancet Infectious Diseases. 2012 Nov;12(11):881–7.

221. Struve C, Roe CC, Stegger M, Stahlhut SG, Hansen DS, Engelthaler DM, et al. Mapping the Evolution of Hypervirulent Klebsiella pneumoniae. mBio [Internet]. 2015 Jul 21 [cited 2017 Dec 5];6(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4513082/

222. David S, Reuter S, Harris SR, Glasner C, Feltwell T, Argimon S, et al. Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. Nat Microbiol. 2019 Nov 1;4(11):1919–29.

223. Grundmann H, Glasner C, Albiger B, Aanensen DM, Tomlinson CT, Andrasević AT, et al. Occurrence of carbapenemase-producing Klebsiella pneumoniae and Escherichia coli in the European survey of carbapenemase-producing Enterobacteriaceae (EuSCAPE): a prospective, multinational study. The Lancet Infectious Diseases. 2017 Feb 1;17(2):153–63.

224. Rodrigues C, Passet V, Rakotondrasoa A, Brisse S. Identification of Klebsiella pneumoniae, Klebsiella quasipneumoniae, Klebsiella variicola and Related Phylogroups by MALDI-TOF Mass Spectrometry. Front Microbiol [Internet]. 2018 [cited 2019 Oct 3];9. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.03000/full

225. Long SW, Linson SE, Saavedra MO, Cantu C, Davis JJ, Brettin T, et al. Whole-Genome Sequencing of a Human Clinical Isolate of the Novel Species Klebsiella quasivariicola sp. nov. Genome Announc. 2017 Oct 19;5(42):e01057-17.

226. Rodrigues C, Passet V, Rakotondrasoa A, Diallo TA, Criscuolo A, Brisse S. Description of Klebsiella africanensis sp. nov., Klebsiella variicola subsp. tropicalensis subsp. nov. and Klebsiella variicola subsp. variicola subsp. nov. Research in Microbiology. 2019 Apr 1;170(3):165–70.

227. Long SW, Linson SE, Ojeda Saavedra M, Cantu C, Davis JJ, Brettin T, et al. Whole-Genome Sequencing of Human Clinical Klebsiella pneumoniae Isolates Reveals Misidentification and Misunderstandings of Klebsiella pneumoniae, Klebsiella variicola, and Klebsiella quasipneumoniae. mSphere [Internet]. 2017 Aug 2 [cited 2017 Oct 27];2(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5541162/

228. Iwase T, Ogura Y, Hayashi T, Mizunoe Y. Complete Genome Sequence of Klebsiella oxytoca Strain JKo3. Genome Announc. 2016 Dec 29;4(6):e01221-16.

229. Leitner E, Zarfel G, Luxner J, Herzog K, Pekard-Amenitsch S, Hoenigl M, et al. Contaminated Handwashing Sinks as the Source of a Clonal Outbreak of KPC-2-Producing Klebsiella oxytoca on a Hematology Ward. Antimicrobial Agents and Chemotherapy. 2015 Jan 1;59(1):714–6.

230. Moradigaravand D, Martin V, Peacock SJ, Parkhill J. Population Structure of Multidrug-Resistant Klebsiella oxytoca within Hospitals across the United Kingdom and Ireland Identifies Sharing of Virulence and Resistance Genes with K. pneumoniae. Genome Biol Evol. 2017 Mar 1;9(3):574–84.

231. Mollet C, Drancourt M, Raoult D. rpoB sequence analysis as a novel basis for bacterial identification. Mol Microbiol. 1997 Dec;26(5):1005–11.

232. Dinkelacker AG, Vogt S, Oberhettinger P, Mauder N, Rau J, Kostrzewa M, et al. Typing and Species Identification of Clinical Klebsiella Isolates by Fourier Transform Infrared Spectroscopy and Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. Journal of Clinical Microbiology. 2018 Nov 1;56(11):e00843-18.

233. Bridel S, Watts SC, Judd LM, Harshegyi T, Passet V, Rodrigues C, et al. Klebsiella MALDI TypeR: a web-based tool for Klebsiella identification based on MALDI-TOF mass spectrometry. Research in Microbiology. 2021 May 15;103835.

234. Anhalt JP, Fenselau Catherine. Identification of bacteria using mass spectrometry. Analytical Chemistry. 1975 Feb;47(2):219–25.

235. Ryzhov V, Fenselau C. Characterization of the Protein Subset Desorbed by MALDI from Whole Bacterial Cells. Anal Chem. 2001 Feb 1;73(4):746–50.

236. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; 2010.

237. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research. 2017 Apr;27(4):626–38.

238. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

239. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology. 2017 Jun 8;13(6):e1005595.

240. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

241. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.

242. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution. 2009 Jul 1;26(7):1641–50.

243. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010 Mar 10;5(3):e9490.

244. Wüthrich D, Cuénod A. Klebsiella-spp [Internet]. 2021 [cited 2021 Aug 23]. Available from: https://github.com/appliedmicrobiologyresearch/Klebsiella-spp

245. Lam MMC, Wyres KL, Judd LM, Wick RR, Jenney A, Brisse S, et al. Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in Klebsiella pneumoniae. bioRxiv. 2018 Jul 25;376236.

246. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of Klebsiella capsule synthesis loci from whole genome data. Microb Genom [Internet]. 2016 Dec 12 [cited 2017 Oct 27];2(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5359410/

247. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive Web: User-Friendly Capsule and Lipopolysaccharide Serotype Prediction for Klebsiella Genomes. Journal of Clinical Microbiology. 2018 Jun 1;56(6):e00197-18.

248. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014 Jul;58(7):3895–903.

249. Seemann T. Abricate. https://github.com/tseemann/abricate;

250. Varshavsky A. The N-end rule pathway and regulation by proteolysis. Protein Sci. 2011 Aug;20(8):1298–345.

251. Wynne C, Fenselau C, Demirev PA, Edwards N. Top-Down Identification of Protein Biomarkers in Bacteria with Unsequenced Genomes. Anal Chem. 2009 Dec 1;81(23):9633–42.

252. Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics. 2012 Sep 1;28(17):2270–1.

253. The European Committee on Antimicrobial Susceptibility Testing. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters, version 6.0-8.1, 2016 - 2018 [Internet]. 2016. Available from: http://www.eucast.org/clinical_breakpoints/

254. Gasser M, Schrenzel J, Kronenberg A. Aktuelle Entwicklung der Antibiotikaresistenzen in der Schweiz. Swiss Medical Forum. 2018 Nov 14;18(46):943–9.

255. Iucif N, Rocha JSY. Study of inequalities in hospital mortality using the Charlson comorbidity index. Rev Saude Publica. 2004 Dec;38(6):780–6.

256. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci U S A. 2015 Jul 7;112(27):E3574–81.

257. Wyres KL, Holt KE. Klebsiella pneumoniae as a key trafficker of drug resistance genes from environmental to clinically important bacteria [Internet]. PeerJ Inc.; 2018 Apr [cited 2019 Feb 15]. Report No.: e26852v1. Available from: https://peerj.com/preprints/26852

258. David S, Cohen V, Reuter S, Sheppard AE, Giani T, Parkhill J, et al. Integrated chromosomal and plasmid sequence analyses reveal diverse modes of carbapenemase gene spread among Klebsiella pneumoniae. PNAS. 2020 Oct 6;117(40):25043–54.

259. Nagy ZA, Szakács D, Boros E, Héja D, Vígh E, Sándor N, et al. Ecotin, a microbial inhibitor of serine proteases, blocks multiple complement dependent and independent microbicidal activities of human serum. PLOS Pathogens. 2019 Dec 20;15(12):e1008232.

260. Hæggman S, Löfdahl S, Paauw A, Verhoef J, Brisse S. Diversity and Evolution of the Class A Chromosomal Beta-Lactamase Gene in Klebsiella pneumoniae. Antimicrob Agents Chemother. 2004 Jul;48(7):2400–8.

261. Granier SA, Leflon-Guibout V, Goldstein FW, Nicolas-Chanoine M-H. New Klebsiella oxytoca β-Lactamase Genes blaOXY-3 and blaOXY-4 and a Third Genetic Group of K. oxytoca Based on blaOXY-3. Antimicrob Agents Chemother. 2003 Sep;47(9):2922–8.

262. Cuénod A, Foucault F, Pflüger V, Egli A. Factors Associated With MALDI-TOF Mass Spectral Quality of Species Identification in Clinical Routine Diagnostics. Front Cell Infect Microbiol [Internet]. 2021 [cited 2021 Mar 17];11. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2021.646648/full

263. Bridel S, Watts SC, Judd LM, Harshegyi T, Passet V, Rodrigues C, et al. Klebsiella MALDI TypeR: a web-based tool for Klebsiella identification based on MALDI-TOF mass spectrometry. bioRxiv. 2020 Oct 13;2020.10.13.337162.

264. Lawlor MS, O'Connor C, Miller VL. Yersiniabactin Is a Virulence Factor for Klebsiella pneumoniae during Pulmonary Infection. IAI. 2007 Mar;75(3):1463–72.

265. Chou H-C, Lee C-Z, Ma L-C, Fang C-T, Chang S-C, Wang J-T. Isolation of a chromosomal region of Klebsiella pneumoniae associated with allantoin metabolism and liver infection. Infect Immun. 2004 Jul;72(7):3783–92.

266. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations. Microbial Genomics [Internet]. 2018 Sep 1 [cited 2019 Jan 11];4(9). Available from: http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000196

267. Brisse S, van Himbergen T, Kusters K, Verhoef J. Development of a rapid identification method for Klebsiella pneumoniae phylogenetic groups and analysis of 420 clinical isolates. Clin Microbiol Infect. 2004 Oct;10(10):942–5.

268. de Kraker MEA, Jarlier V, Monen JCM, Heuer OE, van de Sande N, Grundmann H. The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System. Clin Microbiol Infect. 2013 Sep;19(9):860–8.

269. Arefian H, Heublein S, Scherag A, Brunkhorst FM, Younis MZ, Moerer O, et al. Hospital-related cost of sepsis: A systematic review. Journal of Infection. 2017 Feb 1;74(2):107–17.

270. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. New England Journal of Medicine. 2017 Jun 8;376(23):2235–44.

271. Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, et al. Different evolutionary trends form the twilight zone of the bacterial pan-genome. Microb Genom. 2021 Sep;7(9).

272. Khairy RM, Mohamed ES, Ghany HMA, Abdelrahim SS. Phylogenic classification and virulence genes profiles of uropathogenic E. coli and diarrhegenic E. coli strains isolated from community acquired infections. PLOS ONE. 2019 Sep 12;14(9):e0222441.

273. Lloyd AL, Smith SN, Eaton KA, Mobley HLT. Uropathogenic Escherichia coli Suppresses the Host Inflammatory Response via Pathogenicity Island Genes sisA and sisB. Infect Immun. 2009 Dec;77(12):5322–33.

274. Mobley HL, Jarvis KG, Elwood JP, Whittle DI, Lockatell CV, Russell RG, et al. Isogenic P-fimbrial deletion mutants of pyelonephritogenic Escherichia coli: the role of alpha Gal(1-4) beta Gal binding in virulence of a wild-type strain. Mol Microbiol. 1993 Oct;10(1):143–55.

275. Ambite I, Butler DSC, Stork C, Grönberg-Hernández J, Köves B, Zdziarski J, et al. Fimbriae reprogram host gene expression – Divergent effects of P and type 1 fimbriae. PLOS Pathogens. 2019 Jun 10;15(6):e1007671.

276. Tseng CW, Zhang S, Stewart GC. Accessory Gene Regulator Control of Staphyloccoccal Enterotoxin D Gene Expression. J Bacteriol. 2004 Mar;186(6):1793–801.

277. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012 May;19(5):455–77.

278. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014 Nov 19;9(11):e112963.

279. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology. 2012;158(4):1005–15.

280. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012 Nov;67(11):2640–4.

281. Holt KE, Lassalle F, Wyres KL, Wick R, Mostowy RJ. Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. ISME J. 2020 Jul;14(7):1713–30.

282. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Medicine. 2014 Nov 20;6(11):90.

283. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, et al. In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. Microb Genom [Internet]. 2016 Jul 11 [cited 2018 Jun 15];2(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343136/

284. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-bio] [Internet]. 2012 Jul 20 [cited 2021 Dec 9]; Available from: http://arxiv.org/abs/1207.3907

285. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018 Dec 15;34(24):4310–2.

286. Jaillard M, Lima L, Tournoud M, Mahé P, Belkum A van, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLOS Genetics. 2018 Nov 12;14(11):e1007758.

287. Valenza G, Werner M, Eisenberger D, Nickel S, Lehner-Reindl V, Höller C, et al. First report of the new emerging global clone ST1193 among clinical isolates of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli from Germany. J Glob Antimicrob Resist. 2019 Jun;17:305–8.

288. Denamur E, Condamine B, Esposito-Farèse M, Royer G, Clermont O, Laouenan C, et al. Genome wide association study of human bacteremia *Escherichia coli* isolates identifies genetic determinants for the portal of entry but not fatal outcome [Internet].

Infectious Diseases (except HIV/AIDS); 2021 Nov [cited 2021 Dec 7]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.11.09.21266136

289. Matsuura M. Structural Modifications of Bacterial Lipopolysaccharide that Facilitate Gram-Negative Bacteria Evasion of Host Innate Immunity. Frontiers in Immunology. 2013;4:109.

290. Goh KGK, Phan M-D, Forde BM, Chong TM, Yin W-F, Chan K-G, et al. Genome-Wide Discovery of Genes Required for Capsule Production by Uropathogenic Escherichia coli. mBio. 8(5):e01558-17.

291. Levert M, Zamfir O, Clermont O, Bouvet O, Lespinats S, Hipeaux MC, et al. Molecular and Evolutionary Bases of Within-Patient Genotypic and Phenotypic Diversity in Escherichia coli Extraintestinal Infections. PLOS Pathogens. 2010 Sep 30;6(9):e1001125.

292. Urinary Tract Infections | Mandell, Douglas, and Bennett's Principles… [Internet]. [cited 2020 May 19]. Available from: https://expertconsult.inkling.com/read/mandell-douglas-bennetts-infectious-diseases-8/chapter-74/urinary-tract-infections

293. https://www.uniprot.org/uniprot/Q47450. In: Uniprot, papGII.

294. Sauget M, Valot B, Bertrand X, Hocquet D. Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria? Trends in Microbiology. 2017 Jun 1;25(6):447–55.

295. Rodríguez-Sánchez B, Cercenado E, Coste AT, Greub G. Review of the impact of MALDI-TOF MS in public health and hospital hygiene, 2018. Eurosurveillance. 2019 Jan 24;24(4):1800193.

296. Angeletti S, Ciccozzi M. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry in clinical microbiology: An updating review. Infection, Genetics and Evolution. 2019 Dec 1;76:104063.

297. Wolters M, Rohde H, Maier T, Belmar-Campos C, Franke G, Scherpe S, et al. MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant Staphylococcus aureus lineages. International Journal of Medical Microbiology. 2011 Jan 1;301(1):64–8.

298. Christner M, Trusch M, Rohde H, Kwiatkowski M, Schlüter H, Wolters M, et al. Rapid MALDI-TOF Mass Spectrometry Strain Typing during a Large Outbreak of Shiga-Toxigenic Escherichia coli. PLOS ONE. 2014 Jul 8;9(7):e101924.

299. Fehlberg LCC, Andrade LHS, Assis DM, Pereira RHV, Gales AC, Marques EA. Performance of MALDI-ToF MS for species identification of Burkholderia cepacia complex clinical isolates. Diagn Microbiol Infect Dis. 2013 Oct;77(2):126–8.

300. Dinkelacker AG, Vogt S, Oberhettinger P, Mauder N, Rau J, Kostrzewa M, et al. Typing and Species Identification of Clinical Klebsiella Isolates by Fourier Transform Infrared Spectroscopy and Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. Journal of Clinical Microbiology [Internet]. 2018 Nov 1 [cited 2020 Nov 23];56(11). Available from: https://jcm.asm.org/content/56/11/e00843-18

301. Angeletti S, Dicuonzo G, Avola A, Crea F, Dedej E, Vailati F, et al. Viridans Group Streptococci Clinical Isolates: MALDI-TOF Mass Spectrometry versus Gene Sequence-Based Identification. PLOS ONE. 2015 Mar 17;10(3):e0120502.

302. Fenselau C, Demirev PA. Characterization of intact microorganisms by MALDI mass spectrometry. Mass Spectrometry Reviews. 2001;20(4):157–71.

303. Hotta Y, Teramoto K, Sato H, Yoshikawa H, Hosoda A, Tamura H. Classification of Genus Pseudomonas by MALDI-TOF MS Based on Ribosomal Protein Coding in S10−spc−alpha Operon at Strain Level. J Proteome Res. 2010 Dec 3;9(12):6722–8.

304. Tomachewski D, Galvão CW, de Campos Júnior A, Guimarães AM, Ferreira da Rocha JC, Etto RM. Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. Bioinformatics. 2018 01;34(17):3058–60.

305. Kassim A, Pflüger V, Premji Z, Daubenberger C, Revathi G. Comparison of biomarker based Matrix Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) and conventional methods in the identification of clinically relevant bacteria and yeast. BMC Microbiol [Internet]. 2017 May 25 [cited 2018 Mar 9];17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5445374/

306. Ojima-Kato T, Yamamoto N, Nagai S, Shima K, Akiyama Y, Ota J, et al. Application of proteotyping Strain Solution™ ver. 2 software and theoretically calculated mass database in MALDI-TOF MS typing of <Emphasis Type="Italic">Salmonella</Emphasis> serotype. Appl Microbiol Biotechnol. 2017 Dec 1;101(23–24):8557–69.

307. Matsumura Y, Yamamoto M, Nagao M, Tanaka M, Machida K, Ito Y, et al. Detection of Extended-Spectrum-β-Lactamase-Producing Escherichia coli ST131 and ST405 Clonal Groups by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. J Clin Microbiol. 2014 Apr 1;52(4):1034–40.

308. Lafolie J, Sauget M, Cabrolier N, Hocquet D, Bertrand X. Detection of Escherichia coli sequence type 131 by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: implications for infection control policies? Journal of Hospital Infection. 2015 Jul;90(3):208–12.

309. Patel R. MALDI-TOF MS for the Diagnosis of Infectious Diseases. Clinical Chemistry. 2015 Jan 1;61(1):100–11.

310. Bizzini A, Durussel C, Bille J, Greub G, Prod'hom G. Performance of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Identification of Bacterial Strains Routinely Isolated in a Clinical Microbiology Laboratory. J Clin Microbiol. 2010 May 1;48(5):1549–54.

311. Veen SQ van, Claas ECJ, Kuijper EJ. High-Throughput Identification of Bacteria and Yeast by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry in Conventional Medical Microbiology Laboratories. J Clin Microbiol. 2010 Mar 1;48(3):900–7.

312. Alatoom AA, Cunningham SA, Ihde SM, Mandrekar J, Patel R. Comparison of Direct Colony Method versus Extraction Method for Identification of Gram-Positive Cocci by Use of Bruker Biotyper Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. Journal of Clinical Microbiology. 2011 Aug 1;49(8):2868–73.

313. Veloo ACM, Elgersma PE, Friedrich AW, Nagy E, Winkelhoff AJ van. The influence of incubation time, sample preparation and exposure to oxygen on the quality of the MALDI-TOF MS spectrum of anaerobic bacteria. Clinical Microbiology and Infection. 2014;20(12):O1091–7.

314. Croxatto A, Prod'hom G, Greub G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiol Rev. 2012 Mar 1;36(2):380–407.

315. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009 Nov 10;106(45):19126–31.

316. Ha S-M, Kim CK, Roh J, Byun J-H, Yang S-J, Choi S-B, et al. Application of the Whole Genome-Based Bacterial Identification System, TrueBac ID, Using Clinical Isolates That Were Not Identified With Three Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) Systems. Ann Lab Med. 2019 Nov;39(6):530–6.

317. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, et al. The Proteomics of N-terminal Methionine Cleavage. Molecular & Cellular Proteomics. 2006 Dec 1;5(12):2336–49.

318. Mitchell M, Mali S, King CC, Bark SJ. Enhancing MALDI Time-Of-Flight Mass Spectrometer Performance through Spectrum Averaging. PLOS ONE. 2015 Mar 23;10(3):e0120932.

319. Zhang L, Borror CM, Sandrin TR. A Designed Experiments Approach to Optimization of Automated Data Acquisition during Characterization of Bacteria with MALDI-TOF Mass Spectrometry. PLOS ONE. 2014 Mar 24;9(3):e92720.

320. Oberle M, Wohlwend N, Jonas D, Maurer FP, Jost G, Tschudin-Sutter S, et al. The Technical and Biological Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Based Typing: Employment of Bioinformatics in a Multicenter Study. PLoS One [Internet]. 2016 Oct 31 [cited 2018 Apr 4];11(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5087883/

321. Janda JM. Proposed nomenclature or classification changes for bacteria of medical importance: taxonomic update 5. Diagnostic Microbiology and Infectious Disease. 2020 Jul 1;97(3):115047.

322. Clark AE, Kaleta EJ, Arora A, Wolk DM. Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry: a Fundamental Shift in the Routine Practice of Clinical Microbiology. Clin Microbiol Rev. 2013 Jul;26(3):547–603.

323. Feucherolles M, Poppert S, Utzinger J, Becker SL. MALDI-TOF mass spectrometry as a diagnostic tool in human and veterinary helminthology: a systematic review. Parasites & Vectors. 2019 May 17;12(1):245.

324. Tomachewski D, Galvão CW, de Campos Júnior A, Guimarães AM, Ferreira da Rocha JC, Etto RM. Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. Bioinformatics. 2018 Sep 1;34(17):3058–60.

325. Cuénod A, Wüthrich D, Seth-Smith HMB, Ott C, Gehringer C, Foucault F, et al. Whole-genome sequence-informed MALDI-TOF MS diagnostics reveal importance

of Klebsiella oxytoca group in invasive infections: a retrospective clinical study. Genome Medicine. 2021 Sep 13;13(1):150.

326. Lau AF, Wang H, Weingarten RA, Drake SK, Suffredini AF, Garfield MK, et al. A rapid matrix-assisted laser desorption ionization-time of flight mass spectrometry-based method for single-plasmid tracking in an outbreak of carbapenem-resistant Enterobacteriaceae. J Clin Microbiol. 2014 Aug;52(8):2804–12.

327. WHO. Global action plan on antimicrobial resistance [Internet]. 2015 [cited 2022 Jan 23]. Available from: https://apo.who.int/publications/i/item/global-action-plan-on-antimicrobial-resistance

328. Wise R, Hart T, Cars O, Streulens M, Helmuth R, Huovinen P, et al. Antimicrobial resistance. Is a major threat to public health. BMJ. 1998 Sep 5;317(7159):609–10.

329. Huang AM, Newton D, Kunapuli A, Gandhi TN, Washer LL, Isip J, et al. Impact of rapid organism identification via matrix-assisted laser desorption/ionization time-of-flight combined with antimicrobial stewardship team intervention in adult patients with bacteremia and candidemia. Clin Infect Dis. 2013 Nov;57(9):1237–45.

330. Banerjee R, Teng CB, Cunningham SA, Ihde SM, Steckelberg JM, Moriarty JP, et al. Randomized Trial of Rapid Multiplex Polymerase Chain Reaction–Based Blood Culture Identification and Susceptibility Testing. Clin Infect Dis. 2015 Oct 1;61(7):1071–80.

331. Kommedal Ø, Aasen JL, Lindemann PC. Genetic antimicrobial susceptibility testing in Gram-negative sepsis – impact on time to results in a routine laboratory. APMIS. 2016;124(7):603–10.

332. CDC. Core Elements of Antibiotic Stewardship | Antibiotic Use | CDC [Internet]. 2021 [cited 2022 Jan 23]. Available from: https://www.cdc.gov/antibiotic-use/core-elements/index.html

333. Bourdon N, Bérenger R, Lepoultier R, Mouet A, Lesteven C, Borgey F, et al. Rapid detection of vancomycin-resistant enterococci from rectal swabs by the Cepheid Xpert vanA/vanB assay. Diagn Microbiol Infect Dis. 2010 Jul;67(3):291–3.

334. Huh HJ, Kim ES, Chae SL. Methicillin-resistant Staphylococcus aureus in nasal surveillance swabs at an intensive care unit: an evaluation of the LightCycler MRSA advanced test. Ann Lab Med. 2012 Nov;32(6):407–12.

335. Cury AP, Almeida Junior JN, Costa SF, Salomão MC, Boszczowski Í, Duarte AJS, et al. Diagnostic performance of the Xpert Carba-R™ assay directly from rectal swabs for active surveillance of carbapenemase-producing organisms in the largest Brazilian University Hospital. J Microbiol Methods. 2020 Apr;171:105884.

336. van Belkum A, Welker M, Pincus D, Charrier J-P, Girard V. Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry in Clinical Microbiology: What Are the Current Issues? Ann Lab Med. 2017 Nov;37(6):475–83.

337. Hou T-Y, Chiang-Ni C, Teng S-H. Current status of MALDI-TOF mass spectrometry in clinical microbiology. Journal of Food and Drug Analysis. 2019 Apr 1;27(2):404–14.

338. Kim J-M, Kim I, Chung SH, Chung Y, Han M, Kim J-S. Rapid Discrimination of Methicillin-Resistant Staphylococcus aureus by MALDI-TOF MS. Pathogens. 2019 Nov 1;8(4):E214.

339. Weis C, Horn M, Rieck B, Cuénod A, Egli A, Borgwardt K. Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. Bioinformatics. 2020 Jul 1;36(Supplement_1):i30–8.

340. Vervier K, Mahé P, Veyrieras J-B, Vert J-P. Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. :13.

341. Weis CV, Jutzeler CR, Borgwardt K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. Clinical Microbiology and Infection. 2020 Oct 1;26(10):1310–7.

342. Wang H-Y, Chung C-R, Wang Z, Li S, Chu B-Y, Horng J-T, et al. A large-scale investigation and identification of methicillin-resistant Staphylococcus aureus based on peaks binning of matrix-assisted laser desorption ionization-time of flight MS spectra. Brief Bioinform. 2021 May 20;22(3):bbaa138.

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

344. Centers for Disease Control and Prevention (U.S.). Antibiotic resistance threats in the United States, 2019 [Internet]. Centers for Disease Control and Prevention (U.S.); 2019 Nov [cited 2022 Jan 23]. Available from: https://stacks.cdc.gov/view/cdc/82532

345. Zhang H, Dullerud N, Seyyed-Kalantari L, Morris Q, Joshi S, Ghassemi M. An empirical framework for domain generalization in clinical settings. In: Proceedings of the Conference on Health, Inference, and Learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2021 [cited 2022 Jan 23]. p. 279–90. (CHIL '21). Available from: https://doi.org/10.1145/3450439.3451878

346. Bevan ER, Jones AM, Hawkey PM. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. J Antimicrob Chemother. 2017 Aug 1;72(8):2145–55.

347. Pietsch M, Eller C, Wendt C, Holfelder M, Falgenhauer L, Fruth A, et al. Molecular characterisation of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli isolates from hospital and ambulatory patients in Germany. Vet Microbiol. 2017 Feb;200:130–7.

348. Kim Y-K, Pai H, Lee H-J, Park S-E, Choi E-H, Kim J, et al. Bloodstream infections by extended-spectrum beta-lactamase-producing Escherichia coli and Klebsiella pneumoniae in children: epidemiology and clinical outcome. Antimicrob Agents Chemother. 2002 May;46(5):1481–91.

349. Potron A, Poirel L, Rondinaud E, Nordmann P. Intercontinental spread of OXA-48 beta-lactamase-producing Enterobacteriaceae over a 11-year period, 2001 to 2011. Eurosurveillance. 2013 Aug 1;18(31):20549.

350. Pereira LA, Harnett GB, Hodge MM, Cattell JA, Speers DJ. Real-Time PCR Assay for Detection of blaZ Genes in Staphylococcus aureus Clinical Isolates. Journal of Clinical Microbiology. 2014 Apr 1;52(4):1259–61.

351. Long SW, Olsen RJ, Mehta SC, Palzkill T, Cernoch PL, Perez KK, et al. PBP2a Mutations Causing High-Level Ceftaroline Resistance in Clinical Methicillin-Resistant Staphylococcus aureus Isolates. Antimicrobial Agents and Chemotherapy. 2014 Nov 1;58(11):6668–74.

352. Shapley LS. 17. A Value for n-Person Games. In: 17 A Value for n-Person Games [Internet]. Contributions to the Theory of Games (AM-28), Volume II 307–318; 1953 [cited 2022 Jan 23]. p. 307–18. Available from: https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html?lang=en

353. Camoez M, Sierra JM, Dominguez MA, Ferrer-Navarro M, Vila J, Roca I. Automated categorization of methicillin-resistant Staphylococcus aureus clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. Clinical Microbiology and Infection. 2016 Feb 1;22(2):161.e1-161.e7.

354. Josten M, Reif M, Szekat C, Al-Sabti N, Roemer T, Sparbier K, et al. Analysis of the Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrum of Staphylococcus aureus Identifies Mutations That Allow Differentiation of the Main Clonal Lineages. J Clin Microbiol. 2013 Jun;51(6):1809–17.

355. Østergaard C, Hansen SGK, Møller JK. Rapid first-line discrimination of methicillin resistant Staphylococcus aureus strains using MALDI-TOF MS. International Journal of Medical Microbiology. 2015 Dec;305(8):838–47.

356. Sauget M, van der Mee-Marquet N, Bertrand X, Hocquet D. Matrix-assisted laser desorption ionization-time of flight Mass spectrometry can detect Staphylococcus aureus clonal complex 398. Journal of Microbiological Methods. 2016 Aug;127:20–3.

357. Wolters M, Rohde H, Maier T, Belmar-Campos C, Franke G, Scherpe S, et al. MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant Staphylococcus aureus lineages. Int J Med Microbiol. 2011 Jan;301(1):64–8.

358. Chatterjee SS, Chen L, Joo H-S, Cheung GYC, Kreiswirth BN, Otto M. Distribution and Regulation of the Mobile Genetic Element-Encoded Phenol-Soluble Modulin PSM-mec in Methicillin-Resistant Staphylococcus aureus. PLOS ONE. 2011 Dec 12;6(12):e28781.

359. Ludden C, Decano AG, Jamrozy D, Pickard D, Morris D, Parkhill J, et al. Genomic surveillance of Escherichia coli ST131 identifies local expansion and serial replacement of subclones. Microbial Genomics. 6(4):e000352.

360. Nakamura A, Komatsu M, Ohno Y, Noguchi N, Kondo A, Hatano N. Identification of specific protein amino acid substitutions of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli ST131: a proteomics approach using mass spectrometry. Sci Rep. 2019 Jun 12;9(1):1–8.

361. Beta-lactam-inducible penicillin-binding protein [Internet]. [cited 2022 Jan 23]. Available from: https://www.uniprot.org/uniprot/P07944

362. Beta-lactamase OXA-1 [Internet]. [cited 2022 Jan 23]. Available from: https://www.uniprot.org/uniprot/P13661

363. Beta-lactamase TEM [Internet]. [cited 2022 Jan 23]. Available from: https://www.uniprot.org/uniprot/P62593

364. Beta-lactamase SHV-24 [Internet]. [cited 2022 Jan 23]. Available from: https://www.uniprot.org/uniprot/Q9S169

365. Beta-lactamase CTX-M-1 [Internet]. [cited 2022 Jan 23]. Available from: https://www.uniprot.org/uniprot/P28585

366. Outer membrane porin C [Internet]. [cited 2022 Jan 23]. Available from: https://www.uniprot.org/uniprot/P06996

367. Pickens C, Wunderink RG, Qi C, Mopuru H, Donnelly H, Powell K, et al. A multiplex polymerase chain reaction assay for antibiotic stewardship in suspected pneumonia. Diagn Microbiol Infect Dis. 2020 Dec;98(4):115179.

368. European Committee on Antimicrobial Susceptibility Testing. EUCAST: Clinical breakpoints and dosing of antibiotics [Internet]. [cited 2022 Jan 23]. Available from: https://www.eucast.org/clinical_breakpoints/

369. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol. 2016 Jan;34(1):104–10.

370. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017;9.

371. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–30.

372. Ingle DJ, Ambrose RL, Baines SL, Duchene S, Gonçalves da Silva A, Lee DYJ, et al. Evolutionary dynamics of multidrug resistant Salmonella enterica serovar 4,[5],12:i:- in Australia. Nat Commun. 2021 Aug 9;12(1):4786.

373. Boucher HW, Talbot GH, Bradley JS, Edwards JE, Gilbert D, Rice LB, et al. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. Clin Infect Dis. 2009 Jan 1;48(1):1–12.

374. Pendleton JN, Gorman SP, Gilmore BF. Clinical relevance of the ESKAPE pathogens. Expert Rev Anti Infect Ther. 2013 Mar;11(3):297–308.

375. Yang J, Long H, Hu Y, Feng Y, McNally A, Zong Z. Klebsiella oxytoca Complex: Update on Taxonomy, Antimicrobial Resistance, and Virulence. Clinical Microbiology Reviews [Internet]. 2021 Dec 1 [cited 2021 Dec 22]; Available from: https://journals.asm.org/doi/abs/10.1128/CMR.00006-21

376. ANRESIS. ANRESIS: K. pneumoniae complex [Internet]. ANRESIS. [cited 2022 Jan 20]. Available from: https://www.anresis.ch/antibiotic-resistance/resistance-data-human-medicine/interactive-database-query/

377. ANRESIS. ANRESIS: K. oxytoca complex [Internet]. ANRESIS. [cited 2022 Jan 20]. Available from: https://www.anresis.ch/antibiotic-resistance/resistance-data-human-medicine/interactive-database-query/

378. Pitout JDD, DeVinney R. Escherichia coli ST131: a multidrug-resistant clone primed for global domination. F1000Res [Internet]. 2017 Feb 28 [cited 2019 Oct 8];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333602/

379. Christner M, Dressler D, Andrian M, Reule C, Petrini O. Identification of Shiga-Toxigenic Escherichia coli outbreak isolates by a novel data analysis tool after matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. PLoS One [Internet]. 2017 Sep 6 [cited 2018 Apr 4];12(9). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5587271/

380. Goldstone RJ, Smith DGE. A population genomics approach to exploiting the accessory “resistome” of Escherichia coli. Microbial Genomics [Internet]. 2017 Apr 6 [cited 2022 Jan 15];3(4). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000108

381. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations. Microbial Genomics [Internet]. 2018 [cited 2018 Aug 28]; Available from: http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000196.v1

382. Garcia EC, Brumbaugh AR, Mobley HLT. Redundancy and Specificity of Escherichia coli Iron Acquisition Systems during Urinary Tract Infection. Infect Immun. 2011 Mar;79(3):1225–35.

383. Floyd KA, Moore JL, Eberly AR, Good JAD, Shaffer CL, Zaver H, et al. Adhesive Fiber Stratification in Uropathogenic Escherichia coli Biofilms Unveils Oxygen-Mediated Control of Type 1 Pili. PLOS Pathogens. 2015 Mar 4;11(3):e1004697.

384. Ford B, Verger D, Dodson K, Volkan E, Kostakioti M, Elam J, et al. The Structure of the PapD-PapGII Pilin Complex Reveals an Open and Flexible P5 Pocket. Journal of Bacteriology. 2012 Dec 1;194(23):6390–7.

385. Tseng C-C, Wu J-J, Liu H-L, Sung J-M, Huang J-J. Roles of host and bacterial virulence factors in the development of upper urinary tract infection caused by Escherichia coli. American Journal of Kidney Diseases. 2002 Apr 1;39(4):744–52.

386. Sharma N, Schwendimann R, Endrich O, Ausserhofer D, Simon M. Comparing Charlson and Elixhauser comorbidity indices with different weightings to predict in-hospital mortality: an analysis of national inpatient data. BMC Health Services Research. 2021 Jan 6;21(1):13.

387. Vega HGT la, Castaño-Romero F, Ragozzino S, González RS, Vaquero-Herrero MP, Siller-Ruiz M, et al. The updated Charlson comorbidity index is a useful predictor of mortality in patients with Staphylococcus aureus bacteraemia. Epidemiology & Infection. 2018 Dec;146(16):2122–30.

388. Lambert E, D’Hondt F, Mazzone E, Vollemaere J, Larcher A, Jeugt JVD, et al. Time to Move On: The Impending Need for a New Disease-specific Comorbidity Index for

Bladder Cancer Patients Undergoing Robot-assisted Radical Cystectomy. European Urology Focus. 2021 Jan 1;7(1):139–41.

389. He Q, Barkoff AM, Mertsola J, Glismann S, Bacci S, group (EUpertstrain) C on behalf of the EB expert, et al. High heterogeneity in methods used for the laboratory confirmation of pertussis diagnosis among European countries, 2010: integration of epidemiological and laboratory surveillance must include standardisation of methodologies and quality assurance. Eurosurveillance. 2012 Aug 9;17(32):20239.

390. Bizzini A, Durussel C, Bille J, Greub G, Prod'hom G. Performance of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of bacterial strains routinely isolated in a clinical microbiology laboratory. J Clin Microbiol. 2010 May;48(5):1549–54.

391. Bizzini A, Jaton K, Romo D, Bille J, Prod'hom G, Greub G. Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry as an Alternative to 16S rRNA Gene Sequencing for Identification of Difficult-To-Identify Bacterial Strains. J Clin Microbiol. 2011 Feb;49(2):693–6.

392. Valentine N, Wunschel S, Wunschel D, Petersen C, Wahl K. Effect of culture conditions on microorganism identification by matrix-assisted laser desorption ionization mass spectrometry. Appl Environ Microbiol. 2005 Jan;71(1):58–64.

393. Clark CM, Murphy BT, Sanchez LM. A Call to Action: the Need for Standardization in Developing Open-Source Mass Spectrometry-Based Methods for Microbial Subspecies Discrimination. mSystems [Internet]. 2020 Feb 18 [cited 2022 Jan 6]; Available from: https://journals.asm.org/doi/abs/10.1128/mSystems.00813-19

394. Ojima-Kato T, Yamamoto N, Iijima Y, Tamura H. Assessing the performance of novel software Strain Solution on automated discrimination of Escherichia coli serotypes and their mixtures using matrix-assisted laser desorption ionization-time of flight mass spectrometry. Journal of Microbiological Methods. 2015 Dec 1;119:233–8.

395. Sousa T de, Viala D, Théron L, Chambon C, Hébraud M, Poeta P, et al. Putative Protein Biomarkers of Escherichia coli Antibiotic Multiresistance Identified by MALDI Mass Spectrometry. Biology. 2020 Mar;9(3):56.

396. Weis C, Cuénod A, Rieck B, Dubuis O, Graf S, Lang C, et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nat Med. 2022 Jan 10;1–11.

397. Griffin PM, Price GR, Schooneveldt JM, Schlebusch S, Tilse MH, Urbanski T, et al. Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify vancomycin-resistant enterococci and investigate the epidemiology of an outbreak. J Clin Microbiol. 2012 Sep;50(9):2918–31.

398. Brackmann M, Leib SL, Tonolla M, Schürch N, Wittwer M. Antimicrobial resistance classification using MALDI-TOF-MS is not that easy: lessons from vancomycin-resistant Enterococcus faecium. Clin Microbiol Infect. 2020 Mar;26(3):391–3.

399. Egli A. Digitalization, clinical microbiology and infectious diseases. Clinical Microbiology and Infection. 2020 Oct 1;26(10):1289–90.

400. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018.

401. Larrouy-Maumus G, Clements A, Filloux A, McCarthy RR, Mostowy S. Direct detection of lipid A on intact Gram-negative bacteria by MALDI-TOF mass spectrometry. J Microbiol Methods. 2016 Jan;120:68–71.

402. Furniss RCD, Dortet L, Bolland W, Drews O, Sparbier K, Bonnin RA, et al. Detection of Colistin Resistance in Escherichia coli by Use of the MALDI Biotyper Sirius Mass Spectrometry System. J Clin Microbiol. 2019 Nov 22;57(12):e01427-19.

403. Ferreira L, Sánchez-Juanes F, González-Avila M, Cembrero-Fuciños D, Herrero-Hernández A, González-Buitrago JM, et al. Direct identification of urinary tract pathogens from urine samples by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol. 2010 Jun;48(6):2110–5.

404. Oros D, Ceprnja M, Zucko J, Cindric M, Hozic A, Skrlin J, et al. Identification of pathogens from native urine samples by MALDI-TOF/TOF tandem mass spectrometry. Clinical Proteomics. 2020 Jun 23;17(1):25.

405. Li Y, Wang T, Wu J. Capture and detection of urine bacteria using a microchannel silicon nanowire microfluidic chip coupled with MALDI-TOF MS. Analyst. 2021 Feb 22;146(4):1151–6.

406. Cobo F. Application of maldi-tof mass spectrometry in clinical virology: a review. Open Virol J. 2013;7:84–90.

407. Vogel G, Strauss A, Jenni B, Ziegler D, Dumermuth E, Antz S, et al. Development and validation of a protocol for cell line identification by MALDI-TOF MS. BMC Proceedings. 2011 Nov 22;5(8):P45.

408. Vogel G, Cuénod A, Mouchet R, Strauss A, Daubenberger C, Pflüger V, et al. Functional characterization and phenotypic monitoring of human hematopoietic stem cell expansion and differentiation of monocytes and macrophages by whole-cell mass spectrometry. Stem Cell Res. 2018;26:47–54.

409. Rothen J, Githaka N, Kanduma EG, Olds C, Pflüger V, Mwaura S, et al. Matrix-assisted laser desorption/ionization time of flight mass spectrometry for comprehensive indexing of East African ixodid tick species. Parasit Vectors [Internet]. 2016 Mar 15 [cited 2018 Mar 9];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792108/

410. Müller P, Pflüger V, Wittwer M, Ziegler D, Chandre F, Simard F, et al. Identification of cryptic Anopheles mosquito species by molecular protein profiling. PLoS One. 2013;8(2):e57486.

411. Klein M, Kielhauser R, Dilger T. Optimized rapid and reliable identification of Tuber spp. by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Journal of Mass Spectrometry. 2020;55(12):e4655.

412. El Karkouri K, Couderc C, Decloquement P, Abeille A, Raoult D. Rapid MALDI-TOF MS identification of commercial truffles. Sci Rep. 2019 Nov 27;9(1):17686.

413. Atashi S, Izadi B, Jalilian S, Madani SH, Farahani A, Mohajeri P. Evaluation of GeneXpert MTB/RIF for determination of rifampicin resistance among new tuberculosis cases in west and northwest Iran. New Microbes New Infect. 2017 Jul 13;19:117–20.

414. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol. 2014 Feb;11(2):81–90.

415. Wong HR, Caldwell JT, Cvijanovich NZ, Weiss SL, Fitzgerald JC, Bigham MT, et al. Prospective clinical testing and experimental validation of the Pediatric Sepsis Biomarker Risk Model. Science Translational Medicine [Internet]. 2019 Nov 13 [cited 2022 Jan 21]; Available from: https://www.science.org/doi/abs/10.1126/scitranslmed.aax9000

416. Sloane PD, Ward K, Weber DJ, Kistler CE, Brown B, Davis K, et al. Can Sepsis Be Detected in the Nursing Home Prior to the Need for Hospital Transfer? Journal of the American Medical Directors Association. 2018 Jun 1;19(6):492-496.e1.

417. Ehwerhemuepha L, Heyming T, Marano R, Piroutek MJ, Arrieta AC, Lee K, et al. Development and validation of an early warning tool for sepsis and decompensation in children during emergency department triage. Sci Rep. 2021 Apr 21;11(1):8578.

418. Channon-Wells S, O'Connor D. Host gene signature shows promise to distinguish bacterial and viral infections. The Lancet Digital Health. 2021 Aug 1;3(8):e465–6.

419. Sharma S, Ryndak MB, Aggarwal AN, Yadav R, Sethi S, Masih S, et al. Transcriptome analysis of mycobacteria in sputum samples of pulmonary tuberculosis patients. PLOS ONE. 2017 Mar 10;12(3):e0173508.

420. Khaledi A, Schniederjans M, Pohl S, Rainer R, Bodenhofer U, Xia B, et al. Transcriptome Profiling of Antimicrobial Resistance in Pseudomonas aeruginosa. Antimicrobial Agents and Chemotherapy [Internet]. 2016 May 23 [cited 2022 Jan 21]; Available from: https://journals.asm.org/doi/abs/10.1128/AAC.00075-16

421. Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. PLOS Computational Biology. 2014 Mar 27;10(3):e1003549.

422. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013 Dec;10(12):1196–9.

423. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021 Jun;39(6):727–36.

424. Bickhart DM, Kolmogorov M, Tseng E, Portik DM, Korobeynikov A, Tolstoganov I, et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol. 2022 Jan 3;1–9.

425. Sanabria A, Hjerde E, Johannessen M, Sollid JE, Simonsen GS, Hanssen A-M. Shotgun-Metagenomics on Positive Blood Culture Bottles Inoculated With Prosthetic Joint Tissue: A Proof of Concept Study. Front Microbiol. 2020;11:1687.

426. Geng S, Mei Q, Zhu C, Fang X, Yang T, Zhang L, et al. Metagenomic next-generation sequencing technology for detection of pathogens in blood of critically ill patients. International Journal of Infectious Diseases. 2021 Feb 1;103:81–7.

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