Mol Cell Proteomics Papers in Press. Published on October 4, 2019 as Manuscript TIR119.001559

Fast and accurate bacterial species identification in urine specimens using LC-MS/MS mass spectrometry and machine learning

Florence Roux-Dalvai[1], Clarisse Gotti[1], Mickaël Leclercq[2], Marie-Claude Hélie[3], Maurice Boissinot[3], Tabiwang N. Arrey[5], Claire Dauly[5], Frédéric Fournier[1], Isabelle Kelly[1], Judith Marcoux[1], Julie Bestman-Smith[6], Michel G. Bergeron[3,4] and Arnaud Droit[1,2]

[1] Proteomics platform, CHU de Québec – Université Laval Research Center, Québec City, Québec, Canada

[2] Computational Biology Laboratory, CHU de Québec – Université Laval Research Center, Québec City, Québec, Canada

[3] Infectiology Research Center, CHU de Québec – Université Laval Research Center, Québec City, Québec, Canada

[4] Département de microbiologie-infectiologie et d'immunologie, Faculté de médecine, Université Laval, Québec City, Québec, Canada

[5] Thermo Fisher Scientific, Bremen, Germany

[6] Laboratoire de microbiologie-infectiologie, CHU de Québec-Université Laval, pavillon Hôpital de l'Enfant-Jésus, Québec City, Québec, Canada

Correspondance

arnaud.droit@crchuq.ulaval.ca

RUNNING TITLE

LC-MS/MS and machine learning for bacterial identification in urine

ABBREVIATIONS

DDA	Data Dependent Acquisition
DIA	Data Independent Acquisition
LC-MS/MS	Liquid Chromatography tandem Mass Spectrometry
MALDI-TOF	Matrix Assisted Laser Desorption Ionization – Time Of Flight
PRM	Parallel Reaction Monitoring
SRM	Selected Reaction Monitoring
UTI	Urinary tract infection(s)
Cfr	*Citrobacter freundii*
Ecl	*Enterobacter cloacae*

Eco	*Escherichia coli*
Efa	*Enterococcus faecalis*
Kae	*Klebsiella aerogenes*
Kox	*Klebsiella oxytoca*
Kpn	*Klebsiella pneumoniae*
Pae	*Pseudomonas aeruginosa*
Pmi	*Proteus mirabilis*
Sag	*Streptococcus agalactiae*
Sau	*Staphylococcus aureus*
Sep	*Staphylococcus epidermidis*
Sha	*Staphylococcus haemolyticus*
Smi	*Streptococcus mitis*
Ssa	*Staphylococcus saprophyticus*

ABSTRACT

Fast identification of microbial species in clinical samples is essential to provide an appropriate antibiotherapy to the patient and reduce the prescription of broad-spectrum antimicrobials leading to antibioresistances. MALDI-TOF-MS technology has become a tool of choice for microbial identification but has several drawbacks: it requires a long step of bacterial culture prior to analysis (24h), has a low specificity and is not quantitative. We developed a new strategy for identifying bacterial species in urine using specific LC-MS/MS peptidic signatures. In the first training step, libraries of peptides are obtained on pure bacterial colonies in DDA mode, their detection in urine is then verified in DIA mode, followed by the use of machine learning classifiers (NaiveBayes, BayesNet and Hoeffding tree) to define a peptidic signature to distinguish each bacterial species from the others. Then, in the second step, this signature is monitored in unknown urine samples using targeted proteomics. This method, allowing bacterial identification in less than 4h, has been applied to fifteen species representing 84% of all Urinary Tract Infections. More than 31000 peptides in 190 samples were quantified by DIA and classified by machine learning to determine an 82 peptides signature and build a prediction model. This signature was validated for its use in routine using Parallel Reaction Monitoring on two different instruments. Linearity and reproducibility of the method

were demonstrated as well as its accuracy on donor specimens. Within 4h and without bacterial culture, our method was able to predict the predominant bacteria infecting a sample in 97% of cases and 100% above the standard threshold. This work demonstrates the efficiency of our method for the rapid and specific identification of the bacterial species causing UTI and could be extended in the future to other biological specimens and to bacteria having specific virulence or resistance factors.

INTRODUCTION

The identification of the bacterial species or strain present in a biological sample is essential in many fields of microbiology. Epidemiology, for instance, tracks the spreading of microorganisms related to infectious diseases; food safety laboratories ensure the distribution of pathogen-free products to the consumers; environmental bacteria have a strong impact on maintaining the equilibrium of ecosystems; and clinical laboratories require fast diagnosis methods to provide appropriate treatment to patients with a bacterial infection. However, standard methods for the identification of pathogens requires a time-consuming bacterial culture followed by another long step of immunological or biochemical tests of varying duration and cumbersomeness (1-3). During this period, typically of 24 to 48h but could extend to weeks, patients received broad spectrum antimicrobial treatments. While this strategy is efficient to release the infection for the majority of cases, it is also known to have a strong impact on the development of antimicrobial resistance. Indeed, among patient urines tested for UTI, a large proportion are found not infected. For others, the prescription of broad-spectrum

antibiotics, rather than species-specific antibiotics, might lower the efficiency of the therapy (4, 5). But in all cases, this misuse of antibiotics increases the emergence of multi-drug resistant bacteria. (6-9). Therefore, there is a need for the development of fast and robust methods for bacterial identification, in order to improve therapy and guide rational use of antibiotics. Indeed, identification within few hours would allow to wait for the analysis result before initiating the treatment and then reduce the over-prescription of antibiotics to non-infected patients. But also, the knowledge of bacterial species could permit the early use of targeted narrow-spectrum antibiotics thus limiting the selection of resistant species in the overall population.

Genotyping methods, which are based on the sequencing of partial (16S small subunit ribosomal [rRNA] gene sequencing) or entire genomes (Whole Genome Sequencing) of the microorganisms contained in a sample, are promising since they do not require bacterial culture and can be applied to complex samples containing several species (10, 11). However, the cost and the time required to get identification by sequencing methods preclude their use in routine laboratories. In addition, if 16S rRNA sequencing can provide a quite rapid identification (typically 24 hours), the high conservation of 16S gene sequences across bacterial families and species often limits the precision of identification to the genus level (12, 13). By contrast, Whole Genome Sequencing is able to provide an efficient species and even strain typing, but the cost and the time required to get the results is strongly extended by the sequencing itself and by the data analysis. Moreover,

this analysis requires expert scientific knowledge to provide a confident genome assembly as well as large computing resources (14, 15).

In the past few years, Matrix-Assisted Laser Desorption Ionization – Time Of Flight Mass Spectrometry (MALDI-TOF MS) analysis of microbial proteins has made a breakthrough in routine labs for bacterial identification (16-19). This fast, inexpensive, and automatable technology can replace the conventional phenotype-based methods, hence reducing the time required to get an identification from 2 or 4 days to less than 50 hours. For those reasons, two mass spectrometers, the Biotyper (Bruker) and the Vitek-MS (Shimadzu-BioMérieux), have been approved for clinical use by health governmental organizations of most countries including the United States Food and Drug Administration (FDA) in 2013 (20). In the typical workflow, bacterial colonies isolated by culture are submitted to fast sample preparation (typically, a treatment with formic acid and ethanol) prior to acquisition of protein mass spectra that are used to interrogate a spectral database providing a confidence score for the bacterium identification, an information a physician can use to diagnose the infection.

Despite its numerous advantages, bacterial identification by MALDI-TOF MS has several drawbacks: i) it requires a lengthy culture step to isolate bacterial colonies, since the detection is based on a comparison with spectral database acquired on pure colonies. For the same reason, it is not able to identify polymicrobial infections (*i.e.*: when several species are present in the same sample) without analyzing several types of colonies visually selected on the culture plate; ii) because of the minimal sample preparation, the information contained in the spectra is restricted to the most abundant molecules, thus

limiting the specificity of the method and its capability to identify certain species or subspecies and; iii) it is not quantitative, a potentially important information for certain specimens where pathogens need to be distinguished from the normal microbiota, or when a certain level of infection needs to be reached to necessitate antibiotherapy.

To overcome the above mentioned issues, several studies have tried to improve MALDI-TOF bacterial identification (21). For instance, Clark and colleagues refined the specificity of the method to identify *Escherichia coli* pathotypes by examining specific peaks in the spectra (22). Other investigators have tried to improve the specificity using trypsin digestion which allows the accession to a larger set of molecules and the generation of a Peptide Mass Fingerprint of the bacterial subspecies (23). Several studies skip the culture step to provide a faster identification, especially in the case of sepsis where MALDI-TOF acquisition is performed directly from a positive blood culture sample (24, 25). However, it has been shown that sample preparation methods, which are not homogenous from lab to lab, can influence the rate of correct identification of certain microorganisms (26). Although these studies could improve the standard workflow, they are limited by the sensitivity and the specificity of MALDI-TOF mass spectrometer. Therefore, recent studies have investigated the possibility of using LC-MS (Liquid Chromatography - Mass Spectrometry) methods which, because of their high sensitivity and specificity, have replaced MALDI-TOF MS in most research laboratories. Wang and colleagues used the LC-MS approach to identify biomarkers of five major bacterial species in bronchoalveolar lavage specimen (27) and performed strain typing for *Acinobacter baumanii* (28), Karlsson R et *al.* used it for proteotyping within the mitis group of *Streptococcus* genus (29) and

Cheng *et al.* also used LC-MS/MS in Selected Reaction Monitoring (SRM) mode to target specific peptides of the flagella to type *Escherichia coli* at strain level (30). Bioinformatics tools have also been developed to help in the identification of bacteria from 'bottom up' proteomics data (*i.e.* trypsin-digested proteins). These methods were able to reach 89 to 98.5% correct classification rates at the species level but these values have only been demonstrated after a step of bacterial growth (31, 32).

Taking the advantages of sensitivity and specificity from nanoscale LC-MS/MS technology, and based on these previous studies, we developed a new pipeline using modern proteomics (DIA – Data Independent Acquisition mode) and machine learning algorithms to identify biomarkers able to speciate a set of bacteria of interest in urine specimens. This strategy is based on two steps (Figure 1): i) a training step, that enables to define a peptidic signature for the bacteria of interest and ii) an identification step where the signature is monitored by targeted proteomics to get the identification of bacteria in the infected samples.

Once the training step has been developed, the second step can be performed in routine laboratories on multiple samples and with any type of mass spectrometer working in PRM (Parallel Reaction Monitoring) or SRM (Selected Reaction Monitoring) modes.

This pipeline has been applied to the 15 bacterial species most frequently found in Urinary Tract Infections (UTI). Indeed, urine is the most common clinical specimen with hundreds of samples analyzed each day in most clinical laboratories. Moreover, UTI is one of the most frequent types of infection in humans: it has been demonstrated that 50 to 60 % of

women in western countries will have at least one UTI in their lifetime (33). As reported by statistics of the Enfant-Jésus hospital in Québec City, which analyzes 300 urine specimens each day on average, 68.2% of these samples are infected by the same 4 bacterial species (*Escherichia coli, Streptococcus agalactiae, Klebsiella pneumoniae* and *Enterococcus faecalis*) and 15 species are responsible for more than 84% of all UTI (Supplementary Figure 1). According to literature reports, these are the most frequently found species in UTI (33, 34).

Our original method enables to define a peptidic signature which, when monitored by targeted proteomics, is able to detect which of the 15 bacterial species is present in the urine sample, in less than 4 hours, without any bacterial culture. We also demonstrated that the peptidic signature is transferable to other laboratories and to other mass spectrometers. In addition, we compared the efficiency of our method to the MALDI-TOF standard workflow.

EXPERIMENTAL PROCEDURES

Bacterial culture and counting

Bacterial strains were obtained from the Culture Collection of Centre de Recherche en Infectiologie of Université Laval (CCRI, Québec, Canada).

The bacterial strains used and their corresponding culture conditions are listed in supplementary methods. Semi-log broth bacterial culture calibrated to 0,5 MacFarland suspension were prepared and used for spectral libraries generation or for urine

inoculation. In parallel, they were counted by incubation of 100uL of serial dilutions on blood agar plate (Supplementary Methods).

Urine collection and bacterial inoculation

To mimic urinary tract infections, 1 to 200 µL of semi-log broth culture suspension, corresponding to an estimated final amount of $1x10^4$ to $1x10^6$ CFU/mL, were spiked into 10mL of urine obtained from six different healthy volunteers. The exact concentration of the inoculated cultures was determined in parallel by culture on agar plates as described above.

Moreover, urine specimens from 27 patients were collected at the Microbiology-Infectiology laboratory of Enfant-Jésus hospital of CHU de Québec (Québec, Canada) after few microliters had been used for standard MALDI-TOF analysis. The specimens were kept on ice during the transportation (< 1h) and used immediately.

The consent of all donors was obtained as described in the ethical approval of the Comité d'Éthique de la Recherche of CHU de Québec – Université Laval (recording number 2016-2656).

Sample preparation for spectral libraries

For the generation of bacterial spectral libraries, bacteria from 1mL of semi-log broth bacterial cultures were pelleted by centrifugation at 10,000 x *g* for 15 minutes, the supernatant was discarded and the pellet was washed three times with 1mL of 50mM Tris and centrifuged in the same conditions. The final pellet was frozen dried and stored at -20°C.

Pellets were then resuspended with 50mM of ammonium bicarbonate and 600 units of mutanolysin (Sigma-Aldrich, cat no. M9901) were added to help bacterial lysis by digestion of cell wall peptidoglycan. After a 1-hour incubation at 37°C, 0.5% sodium deoxycholate (SDC) and 20mM dithiothreitol (DTT) (final concentrations) were added and bacterial inactivation was performed by heating 10min at 95°C. Lysis was achieved

by sonication for 15min with a Bioruptor® system (Diagenode), with cycles of 30s ON/30s OFF, high level. A final centrifugation at 16,000 x *g* during 15min was performed to remove cell debris, and protein concentration in the supernatant was measured using a Bradford assay.

Prior to proteolytic digestion, SDC concentration was adjusted to 1% and 120 µg of proteins from each bacterial culture were digested by the addition of trypsin (Promega) in a 1:50 (enzyme:protein) ratio, during 1 hour at 58°C. Trypsin reaction was then stopped by acidification with 350µL of 5% formic acid (FA), which also leads to precipitation of the SDC. After centrifugation at 16,000 x *g* for 5 min, the supernatant was collected, the peptides were purified on Oasis HLB cartridge 10mg (Waters) and vacuum-dried.

The pellet was resuspended in 10mM Ammonium bicarbonate pH10 and an equivalent of 110µg of peptides were fractionated on an Agilent 1200 Series System HPLC equipped with Agilent extend C_{18} (1.0mm x 150mm, 3.5µm) column. Peptides were loaded at 1mL/min of solvent A (10mM ammonium bicarbonate pH10) and eluted by the addition of solvent B (90% acetonitrile, 10% ammonium bicarbonate pH10) with a gradient 5 to 35% solvent B during 60 min and 35 to 70% solvent B during 24 min. Fractions were collected in a 96 well plates at 1 min intervals and finally pooled in rows into 8 fractions which were vacuum-dried.

Each fraction was resuspended in 2% acetonitrile (ACN) / 0.05% trifluoroacetic acid (TFA) at 0.2 µg/µL and 1X iRT peptides (Biognosys) were added. An equivalent of 1µg of peptides was injected on LC-MS/MS system for each fraction of each bacterial species.

Preparation of urine samples

Urine specimens (10 mL), either from patients or artificially inoculated from healthy urine, were treated the same way: human cells were initially pelleted by low speed centrifugation for 5 min at 1,000 x *g*, and the supernatant was high speed centrifuged for 15 min at 10,000 x *g* in order to collect bacteria. Bacterial pellets were then washed

with 1mL of 50mM Tris and centrifuged again in the same conditions, another cycle of wash and centrifugation was added and the resulting pellet was frozen dried.

Protocols for protein extraction, trypsin digestion and peptide purification are described above in the 'Sample preparation for spectral libraries' section and were modified as follows: for each sample, 50 units of mutanolysin was used, 250 ng of trypsin was added for the digestion which was then stopped with 1µL of 100% FA and peptides were purified with StageTips (35) containing C18 reverse phase (3M Empore C18 Extraction Disks). Samples were resuspended in 10µL of 2% ACN, 0.05% TFA and 1X iRT peptides (Biognosys) were added. Half of the final volume was injected on LC-MS/MS system.

LC-MS/MS acquisitions

Samples were analyzed by nanoLC/MS using a UltiMate™ 3000 NanoRSLC system (ThermoScientific, Dionex Softron GmbH, Germering, Germany) coupled to an Orbitrap Fusion Tribrid – ETD mass spectrometer (ThermoScientific, San Jose, CA, USA, Instrument Control Software version 2.0) installed in CHU de Québec - Université Laval research center (Québec, Canada). Peptides were trapped at 20 µL/min in loading solvent (2% acetonitrile, 0.05% TFA) on a µ-Precolumn, 300 µm i.d x 5mm, C18 PepMap100, 5 µm, 100Å (Thermo Fisher Scientific) for 5 minutes. Then, the pre-column was switched online with a PepMap100 RSLC, C18 3µm, 100Å, 75µm i.d. x 50cm column (Thermo Fischer Scientific) and the peptides were eluted with a linear gradient from 5-40% solvent B (A: 0.1% formic acid, B: 80% acetonitrile in 0.1% formic acid) in 90 minutes, at 300 nL/min flow rate.

For faster measurements, a Q-Exactive HF-X (Thermo Scientific, San Jose, CA, USA, Instrument Control Software version 2.9) was coupled to a UltiMate™ 3000 RSLCnano system (Thermo Scientific, Germering, Germany) operated in capillary flow chromatography (installed in Thermo Fisher Scientific mass spectrometers factory in Bremen, Germany). Peptides were loaded onto a µ-Precolumn, 300 µm i.d x 5mm, C18 PepMap100, 5 µm, 100Å (Thermo Fisher Scientific) at a flow rate of 50 µL/min for a min, loading solvent (2% ACN, 0.05% TFA). Then, the pre-column was switch online with

a PepMap100 RSLC, C18 2µm, 100Å, 150µm i.d. x 15cm column (Thermo Fischer Scientific). The peptides were eluted with a linear gradient from 6-60% solvent B (A: 0.1% formic acid, B: 80% acetonitrile in 0.1% formic acid) in 34 minutes, at 1 µL/min flow rate. Mass spectrometer parameters settings in DDA, DIA and PRM modes on both instruments are described in the Supplementary methods.

Peptides libraries generation

Proteome Discoverer 2.1.0.81 (Thermo Fischer Scientific) was used to search DDA raw files against Uniprot bacterial databases (databases are listed in Supplementary Methods). Peak lists were generated with the Spectrum Selector node (default parameters) of Proteome Discoverer and searched using Mascot search engine version 2.5 (MatrixScience). Parameters were set for trypsin enzyme digestion specificity with two possible missed cleavages, methionine oxidation, asparagine and glutamine deamidation were set as variable modifications, and mass search tolerance were 10 ppm and 0.6 Da for MS and MS/MS respectively. Peptides were then validated based on target/decoy search using Percolator software with a Delta Cn parameter >0.05 for PSM filtering (36). Only high confidence peptides (FDR<1% at peptide level) were finally considered .

Peptides selection and signal extraction in DIA analyses

For higher confidence and reproducibility in peptide identification, DIA signal extraction was performed on a selected part of the peptides identified in spectral libraries. Only peptides without missed cleavage or potential missed cleavage, having at least 8 amino acids in their sequence without any methionine and cysteine and identified in at least 6 Peptide Spectrum Matches (PSM) were considered. The list of peptides from the 15 bacterial species was then searched with the Unipept software (37, 38) to delete peptide sequences also found in the human proteome, and to associate each peptide to the bacterial proteome it belongs to. Finally, for each bacterium separately, a list of potentially observable peptides was built and imported into Skyline 4.1.0.11796 (39,

40). Shuffle decoy peptides were added to allow further scoring. A spectral library common to the 15 bacterial species was generated using the BiblioSpec 2.0 tool implemented into Skyline using the Mascot .dat files generated from all the individual DDA analyses and a 0.95 cut-off score (Skyline default value) on the Mascot expected value (homology threshold). Retention time predictor was used considering the iRT peptides retention time values. Orbitrap resolving power was set at 30K at 200m/z with a high selectivity extraction. For each precursor (2+ or 3+), only 6 fragments (b or y) were automatically selected within 10 minutes around the predicted RT and their corresponding signal was extracted from the raw files, the signal of precursor masses was not extracted. mProphet algorithm (41) was used within Skyline to score the peaks, considering the decoys and the second best peaks.

Only peaks with a Skyline dot product (dotP) > 0.75 and a *q*-value < 0.01 were considered as quantifiable and for each of them, the peptide areas (*i.e.* the sum of the area under the curve of the 6 most intense fragments) were normalized with the sum of the 10 iRT peptides. The non-quantifiable peaks received a value of 0. Finally, only the best intensity precursor of each peptide was kept to build a final list of peptides with their corresponding area values.

Machine learning

We applied various machine learning models and several feature search approaches to identify a peptidic signature using BioDiscML (42), a tool based on Weka Java library (43).

Briefly, BioDiscML works as follows: during the loading of input data, a sampling is performed to create a test set not used during learning. From the training set, the features, here the peptides, are identified and ranked by their predictive power through information gain ranking for classification. Then, optimal signatures are built using a combination of various stepwise feature selection overall input features and model search approaches. For each iteration on all best ranked features, BioDiscML runs a set of stepwise methods (forward, backward, or a combination of both) using many

machine learning classifiers (e.g. Naïve Bayes, Random Forest) that are evaluated by cross validation procedures (e.g. k-fold, Bootstrapping, repeated holdout, evaluation on test set) and on the test set.

Peptidic signature validation and bacterial identification prediction

After PRM analysis using the Orbitrap Fusion or the Q-Exactive HF-X instrument, the Skyline software 4.1.0.11796 (39, 40) was used to extract the signal of the 82 peptides signature (*i.e*. the sum of the area under the curve of the 6 most intense fragments) in each sample. The peptides were considered as detected if they meet the following Skyline criteria: dotp > 0.85 and average mass error < 10 ppm, or dotp > 0.75 and average mass error < 3 ppm. For each analysis (inoculated urines or patient sample), the list of detected peptides was submitted to the Bayesian Network model trained in the previous section for prediction purposes.

MALDI-TOF analysis

For all MALDI-TOF analyses, the standard procedure of the Enfant-Jésus hospital microbiology laboratory was used. Briefly, 1µL of urine was streaked on blood agar plate and 1µL on Mc Conkey agar plates (Oxoid). The plates were incubated for 18 hours at 35°C. Isolated colonies with homogenous aspect were selected for MS analysis. The non-treated colonies were spotted on MALDI plate with HCCA matrix. MALDI-TOF MS analysis was performed on a Bruker Biotyper instrument using the Flux control version 3.4 (build 135) software and 7311 MSPs database.

Experimental Design and Statistical Rationale

In order to obtain a high quality peptidic signature using machine learning algorithms, 9 high-level and 3 low-level inoculations replicates of each bacterial species were used. 10 non-inoculated urine specimens (biological replicates) were used as control. For the

validation of the method in targeted proteomics (Tier 3 level), four different biological replicates of each bacterial inoculation in urine were monitored in two different analysis conditions. The four non-inoculated urines were used as control. Finally, urine from 27 different patients were used to compare the method to conventional MALDI-TOF analysis. Prediction accuracies were reported.

RESULTS

Our workflow for bacterial identification is composed of two steps: i) a training step which includes the LC-MSMS acquisition of a peptide library on pure bacterial colonies in Data Dependent Acquisition mode followed by Data Independent Acquisition analyses to obtain information on bacterial peptides observability in urine and the generation of a short peptidic signature by machine learning models and ii) an identification step where the signature is monitored in unknown samples by PRM to obtain a bacterial identification through a prediction algorithm (Figure 1).

For the training step, in order to detect minor bacterial peptidic signals in the human proteic background, we used DIA acquisition, on an Orbitrap Fusion instrument operating in nanoflow rate, because of its high sensitivity and its ability to provide a deep coverage of bacterial proteomes by acquisition of all peptides contained in the sample (44, 45). Indeed, in contrast to DDA which uses a full scan MS for the detection of peptide species, the DIA mode, by systematic acquisition of small size windows all along the mass range, improves the dynamic range and, thus, the sensitivity of the analysis. However, the

simultaneous fragmentation of peptides inside this small window generates a complex spectrum which cannot be searched with conventional database search engines .

Acquisition of bacteria spectral libraries

One of the proposed approaches to extract information from the DIA complex spectra is to use spectral libraries previously acquired in DDA mode on the same type of sample and annotated with peptide/protein identifications through a protein database search (44). In our study, we have generated these spectral libraries from pure bacterial colonies in order to be as exhaustive as possible and cover a very wide range of bacterial tryptic peptides, and subsequently be able to extract this specific bacterial peptide information from the DIA complex spectra contaminated with human biological material.

We have generated spectral libraries for the 15 bacterial species of interest. To do so, each species was cultivated separately, proteins were extracted and digested with trypsin as described in the 'experimental procedures' section. The resulting peptides were fractionated by high-pH reversed phase chromatography. For each bacterial species, eight fractions were injected by LC-MS/MS in DDA mode and analyzed through a standard database search pipeline allowing the identification of 10686 to 29558 peptides at 1% FDR corresponding to 810 to 2438 protein groups (Supplementary Tables 1 and 2). As anticipated based on their genome size, gram-positive bacteria generated less protein identifications than the gram-negative. Indeed, there was a good correlation between the number of proteins identified in our study and the genome length (Pearson correlation coefficient r = 0.82) or the protein count predicted from genomic data (Pearson

correlation coefficient r = 0.83) of all those 15 species. Thus, peptide fractionation combined to mass spectrometry analysis on a high resolution and high sensitivity instrument allowed us to cover 22.3 to 48.4 % of the Uniprot reference proteome of each of the 15 species (Supplementary Table 1). Then, the whole list of peptide identifications was refined to filter out: i) the peptides which may not to be reproducible from run to run (*i.e.* cysteine and methionine containing peptides, those containing trypsin missed cleavages, peptides shorter than eight amino-acids), and ii) the less abundant or less ionizable peptides (*i.e.* those having less than six Peptide Spectrum Matches). Finally, we obtained a set of 31096 peptides which, according to their taxonomic affiliations, demonstrated a high redundancy across the 15 species (Supplementary Figure 2a).

This redundancy associated to our reproducibility filters showed that it is not possible to select from these data one or several specific peptides for each bacterial species that could would be further able to specifically sign for the presence of each distinct species in the urine. Indeed, not enough specific peptides are available when working which this large number of bacteria (*i.e.* 15) (Supplementary Figure 2b and 2c).

Thus, we aim to define a set of peptides that could be shared by several species, but which, taken together, form a particular pattern for each bacterial species to be identified. To obtain this 'peptidic signature' our strategy was to use deep proteome coverage combined to machine learning algorithms to obtain this signature.

Data Independent Analysis of artificially inoculated urine replicates

In order to define a peptidic signature of 15 bacterial species in the human urine background, we have generated 12 artificial sample replicates, for each species of our selection, by inoculating urine from healthy volunteers with bacterial culture. Two concentration levels were used set at $1x10^6$ CFU/mL (Colony Forming Unit per milliliter of urine) (n=9) (high level) and below $1x10^5$ CFU/mL (n=3) (low level) approximatively which corresponds to the threshold used by most clinical laboratories for considering a UTI requiring an antibiotherapy. A total of 190 samples were produced, including 'blank' samples corresponding to non-inoculated urine as control. After protein extraction and short trypsin digestion, the resulting peptides were analyzed by LC-MS/MS in DIA mode. In comparison to DDA, DIA analysis enables a deep proteome coverage by reduction of the spectral dynamic range resulting in fewer missing values (46). However, MS/MS spectra acquired in DIA mode are the sum of fragments generated by all precursor peptides selected in the same DIA window. It yields complex spectra where peptides sequences can be deduced by extraction of their specific fragments contained in the spectral libraries previously generated on pure bacterial colonies as described above. To do so, we have used the Skyline software (39, 40) and the list of 31096 selected peptides was used. An additional step of refinement was done to establish, for each species, the list of bacterial peptides to be searched for in the DIA runs. To this purpose, we used the Unipept software (37, 38) that enables to match peptide sequences with all matching taxa in UniProtKB databases. Starting from the non-redundant list of all peptides identified the bacteria (31096 peptides), Unipept was used to confirm in which of the 15 bacterial species these could theoretically be found. Indeed, due to the stochastic effect of DDA

used for library generation, it might be that some peptides belonging to several species had been sequenced by MS/MS in only a subset of them. The Unipept software also helped us to remove peptide sequences shared with the human proteome (57 peptides), hence generating high confidence lists of expressed peptides for each of the 15 species, free of potential human interfering compounds (Supplementary Table 3).

These lists and the corresponding spectra were added to Skyline for extracting DIA signals in the 12 replicates of each of the 15 inoculated samples. As retention time calibration peptides (iRT, Biognosys) were added in the DDA and DIA runs (performed with 90 minutes gradients), predicted RT could be used for signal extraction in a small window of 10 minutes, thus limiting the probability for the software to select background peaks. A list of decoy peptides generated by Skyline were also extracted in the same conditions to ensure the calculation of a scoring *q*-value through the mProphet algorithm (41) included in Skyline.

Finally, the peptides were considered as detected if they met the following criteria: mProphet *q*-value < 0.01 and library dot product (dotp) > 0.75. After normalization of the peptide ions area values (i.e the sum of the 6 most intense fragments areas) by the sum of the iRT peptides area values and filtering for the best intensity precursor (when both doubly and triply-charged precursors were detected for the same peptide), the 15 final lists were combined into one, composed of 4319 peptides, (Supplementary Table 4) and submitted to machine learning algorithms to classify the bacteria and identify a short peptidic signature. This computational method has been chosen for its ability to handle large datasets and to perform predictions on them using accurate statistical models (47).

Peptidic signature generation by machine learning

In our study, before training classifiers, the dimension of the list of peptides was reduced by mutual information filter (i.e. Information Gain ranking), which ended with a list of first 1000 best peptides according to their ranking. After training and evaluation by BioDiscML, the peptidic signature was composed of feature subsets found by three models having very high predictive performance (AUC > 98% on test set): (i) 68 features signature found by forward stepwise feature selection optimized by Matthew's correlation coefficient criterion (48) using Naive Bayes classifier (49) with discretization parameter option, (ii) 78 features signature found by forward stepwise feature selection optimized by Area Under the Curve (AUC) criterion using Bayesian Network classifier (50) with AD-Tree parameter option, and (iii) 20 features signature found by forward stepwise feature selection combined with backward stepwise feature elimination optimized by Matthew's correlation coefficient criterion using Hoeffding Tree classifier (51) with default parameters. Since stepwise feature selection tends to remove all correlated features, we retrieved those using Pearson and Spearman correlations having >99% correlation. The choice of keeping the features and correlated features selected by more than one classifier was motivated by the need to have the largest and the most precise signature exempt of noisy features. Having highly correlated features here also mean preserving “backup” peptides in case of missing peptides (for instance at low bacterial concentrations) and thus improve the sensitivity threshold of the

method. The overlap between the three signatures was 10 peptides (Supplementary figure 3a and supplementary table 5).

The obtained feature subsets of the three models were then merged into a list of 106 unique peptides which were manually curated by inspection into Skyline software. Peptides which show uncertain peak picking, those also found in blank samples, as well as pairs of peptides having the same precursor mass due to leucine/isoleucine amino-acids were deleted to obtain a final curated signature, composed of 82 peptides. Eight peptides from this list were observed in all three models (Supplementary Figure 3b and Supplementary Table 5). The intensity values for this 82 peptides signature were then discretized into presence (intensity > 0) or no presence (intensity = 0) of a peptide and was used to train a final Bayesian Network prediction model by automated learning. All new samples were analyzed using this predictive model. This model, trained on only high levels of concentration provided 100% classification accuracy on several *k*-fold cross-validations (*k* = 2, 5, 10) and was able to classify at 84% overall accuracy the low-level concentration samples (3 replicates per bacteria) corresponding to a concentration below the clinical threshold of $1x10^5$ CFU/mL.

In the final signature, 5 to 26 peptides are observable for each bacterium (Figure 2 and Supplementary Figure 4). Even though closely related species, such as *Streptococcus epidermidis* and *Staphylococcus aureus*, or *Klebsiella pneumoniae* and *Escherichia coli*, share up to 75% of common peptides there are always a few peptides to distinguish them (4 and 7 peptides respectively in these two cases). For some very low concentration replicates, a few peptides, found high concentration replicates, were not detected. This

loss affected the ability of the algorithm to predict the bacteria in only 15% of the tested low concentration replicates. Inversely, some false positive peptide detections were also observed, probably due to peak picking errors by Skyline in DIA runs, but they did not interfere with the bacterial prediction, assessing the robustness of the Bayesian Network model. As expected, most of the peptides composing the signature belong to relatively abundant proteins such as ribosomal proteins (*e.g.* 50S ribosomal protein L10, 30S ribosomal protein S5) or enzymes involved in amino acid metabolism (*e.g.* formate acetyltransferase) and glycolysis (*e.g.* GAPDH, pyruvate kinase) (52).

Validation of the signature by targeted proteomics

Since the machine learning algorithm has identified a short list of peptides allowing the discrimination of the 15 bacteria of interest, this list can now be monitored by targeted proteomics which is known to give a better reproducibility of measurements and a better sensitivity in peptide detection and could thus improve the limit of detection of bacterial species in urine (Supplementary Table 6). The information on presence or absence of each of the 82 peptides of the signature is then given to the developed prediction model to obtain a probability of contamination. This step corresponds the Identification step of our pipeline (Figure 1). For this purpose, any type of mass spectrometer designed to perform targeted proteomics in Selected Reaction Monitoring (SRM) or Parallel Reaction Monitoring (PRM) modes can be used.

To validate our peptidic signature, we have initially used the Orbitrap Fusion Tribrid in PRM mode to monitor precursor masses of the 82 peptides signature on samples resulting

from inoculated urines. For this purpose, the four most frequently found bacteria in UTIs (*Escherichia coli, Streptococcus agalactiae, Enterococcus faecalis* and *Klebsiella pneumonia*) were inoculated at 5 different concentrations (from 2.56 x 10^4 to 8.77 x 10^6 CFU/mL) in urine from four different healthy volunteers (Supplementary Table 7). The samples were processed as described in the 'experimental procedures' section and analyzed with a 90-minutes gradient typically used in research laboratories. Only half of the volume of each sample was injected while the other half was kept for validation on other instrument types as further described.

In order to validate the detection or non-detection of each peptide of the signature, the Skyline software was used associated with filtering criteria. Several criteria and values were tested to define the best filtering for the detection of these known samples (Supplementary Table 8). The criteria giving the lower level of wrong prediction (dotP > 0.85 & ppm < 10 or dotP > 0.75 & ppm < 3) was retained and further reapplied for all unknown samples. Finally, the list of detected peptides for each replicate sample was submitted to the Bayesian Network model for evaluation. The probabilities of bacterial identification are shown on Figure 3a and Supplementary Table 9a. In 97% of the cases, the method was able to predict the correct species inoculated in the sample. We have investigated the factors which causes the wrong predictions in our dataset. In most cases, this has been observed on urines samples inoculated at low bacterial concentrations (below the clinical laboratory threshold of 1e5 CFU/mL). In those cases, the signal of some peptides appears to be very low (or undetectable) and Skyline picks a wrong peak. These peaks are then filtered out by our criteria on dotp and mass shift. In all cases but one, the

wrong prediction is reported as a “blank” (i.e. no contamination). Finally, when looking at the data above the 1 x 10^5 CFU/mL threshold commonly used by clinical labs, the accuracy in prediction reaches 100%.

Using the intensities given by Skyline for each peptide, linearity curves have been plotted for each detected peptide of the signature (Figure 3b and Supplementary Figure 5a-d). The median of determination coefficient (R^2) over the 5 concentrations was 0.841, suggesting that the method could be used for the quantification of bacterial contamination in urine samples. Since four biological replicates (i.e. bacterial inoculation in urines coming from four different volunteers) have been analyzed for each bacteria concentration, the reproducibility of the method was evaluated. Scatter plots and Pearson correlation factors were calculated from replicate to replicate (Figure 3c and Supplementary Figure 6a-d). For the same bacteria across various biological replicates, the Pearson correlation factors were 0.894 in average. This good reproducibility again suggests a possible use of the method for bacterial quantification in urine.

Transfer of the signature on different instruments

To demonstrate that the initially designed signature using an Orbitrap Fusion instrument coupled to a nanoflow chromatography system is transferable to other instruments in others labs, we have analyzed the same inoculated urines of healthy volunteers (four different bacteria, five concentrations) in PRM mode on a Q-Exactive HF-X instrument coupled to capillary chromatography in PRM mode. Indeed, chromatography at higher flow rate (1 µL/min) improves the robustness of peptide separation and detection. To

reduce the turnaround time between sample collection and bacterial identification as much as possible, the chromatographic gradient was also reduced between the Orbitrap Fusion and the Q-Exactive HF-X from 90 to 30 minutes. As for the Orbitrap Fusion data, the data collected from the Q-Exactive HF-X were analyzed using Skyline with the same validation criteria and the resulting list of detected peptides was used by the prediction algorithm (Figure 4a, Supplementary Figure 7 and Supplementary Table 9b). Thus, in 94% of the cases, the method allowed the correct prediction of the bacteria initially inoculated in the samples. Errors were found only for some of the two lowest concentrations points and in all cases but one, the sample was predicted as blank.

We could observe that wrong predictions happened more frequently on *S.agalactiae* inoculations. This can be explained by the fact that this bacterium has been inoculated at a slightly lower level than the others but also by the fact that only 4 peptides of the signature are used to predict this species. In all wrong predictions on *S.agalactiae* inoculates only 2 peptides among 4 meet our criteria and the samples are predicted as blank. But when looking at the data above the clinical threshold of $1x10^5$ CFU/mL, the accuracy was significantly improved to reach 100%.

Linearities were also calculated by plotting the intensities of detected peptides across the five bacterial concentrations inoculated. The median of determination coefficients (R^2) of all peptides was 0.803 (Figure 4b and Supplementary Figure 8a-d). In terms of reproducibility, four biological replicates were analyzed on the Q-Exactive HF-X. Scatter plots showed very good reproducibility since the Pearson correlation factors were 0.852 on average across the various biological replicates from the same bacterium at the same

concentration (Figure 4c and Supplementary Figure 9a-d). However, lower Pearson correlation coefficient values were obtained for some peptides of *Streptococcus agalactiae.* Again, this can be explained by the fact that this species has been inoculated at lower concentration than the others (Supplementary Table 7).

Again, the good results in terms of linearity and reproducibility obtained on the Q-Exactive HF-X instrument suggest its potential use for quantification of bacteria in urine.

Validation on patient samples and comparison to MALDI-TOF MS

In order to validate our method in comparison to conventional MALDI-TOF analysis, samples were collected from 27 patients. Aliquots of the samples were analyzed using either our pipeline without any culturing by monitoring the peptidic signature in PRM (nanoscale method) or with the standard MALDI-TOF method after 24h bacterial culture and the predictions of both methods were compared (Figure 5a and Supplementary Tables 9c and 10). Most of the analyzed urines were determined to be not infected (n=7) or infected by *E. coli* (n=9) with both methods, while 4 other samples contained 4 different bacteria among our 15 targeted species. For those 20 patients we found a correlation between MALDI-TOF and LC-MS in 95% of cases. In seven other cases, the LC-MS/MS reported the urine as 'blank' (not infected) while the MALDI-TOF reported an identification marked as 'probable' and limited at the genus level (*i.e.* without species mention). These results might be explained by a low level of contamination (below the level requiring anti-biotherapy) which prevented the LC-MS/MS method without culture

to detect the signature peptides or by a contamination of the bacterial culture by non-pathogens, or infection of urines by species outside of our selection.

Among the 15 bacterial species detectable by our method, many of them have a quite low frequency in UTI and not found in urines tested. In order to validate the detection of these species with our method in comparison with MALDI-TOF, we inoculated one aliquot of healthy urine with each of the 15 bacteria at concentrations ranging from 3.32 x 105 to 5.66 x 106 CFU/mL (Supplementary Table 7) and analyzed them with both pipelines. Our method found the correct inoculated bacterium in 100% of cases, while MALDI-TOF reported 2 errors (Figure 5b and Supplementary Tables 9c and 10). This lack of specificity in MALDI-TOF analysis might also explain some of the miscorrelations observed on patient samples

DISCUSSION

In this study, we developed a new strategy combining proteomics and machine learning for a fast, specific and accurate detection and identification of bacterial species present in urine without the need for time-consuming bacterial culture. We successfully applied our pipeline on the 15 bacterial species most commonly found in UTIs and obtained, in less than 4 hours, high rates of prediction accuracy, especially when looking above the quantitative threshold commonly used by clinical laboratories to consider a urine as infected and requiring anti-biotherapy. These 15 species represent 84% of all UTIs, meaning that, by monitoring this peptidic signature, a wide majority of UTIs, as well as non-contaminated samples could be identified in less than four hours, allowing the

possibility to delay the non-specific antimicrobial treatment of the patients. While the MALDI-TOF technology is able to discriminate thousands of species from pure colonies, our method could be improved in the future to allow the discrimination of more species causing UTIs. In that case, a new peptidic signature could be designed using the same pipeline where machine learning algorithms could be applied on newly acquired DIA data associated to the DIA dataset available with this study.

Additionally, this proof of concept paves the way to the development of new peptidic signatures for the analysis of other types of clinical specimens (bronchoalveolar lavage (BAL), stool, hemoculture…) (53-55), but also for the detection of foodborne or waterborne pathogens (56, 57), to reduce the turnaround time required to obtain a genus- and/or species-specific identification of microorganisms by classical or molecular microbiology methods or MALDI-TOF mass spectrometry. In those cases, the sample preparation might need to be adapted for each type of specimen. However, centrifugation or microfiltration technics could be implemented to concentrate the bacteria and reduce the background proteins concentration. These strategies are widely used to isolate bacteria from water or in dairy industry for milk sterilization (58-60) and they have been applied on clinical samples to enrich for bacterial cells before detection (61-63). Sedimentation on spinning devices have also been reported to isolate bacterial from blood (64). For some of these applications, without any culture, the sensitivity of the method might be too low to detect the bacteria but, it is expected that a short-term culture in a liquid medium might be enough to reach a detectable level (<1 x 10^3 CFU/mL for certain peptides) without the isolation of colonies on a culture plate. In all cases, the

high specificity of the method, due to a fine selection of the signature peptides, leads to a great improvement to what can be obtained with other standard methods such as MALDI-TOF mass spectrometry or 16S rRNA sequencing. This would be particularly valuable for the epidemiological surveillance of specific pathogens, instead of relying on expensive and time-consuming whole genome sequencing (65, 66).

Moreover, the linearity and reproducibility of our method were evaluated and the obtained results suggest that the method could be used for quantification of bacterial cells in urine (for instance by addition of peptidic internal standards during the PRM monitoring). This would be particularly useful since, in the case of urine specimens, a real infection needs to be distinguished from low level bacterial contaminants and this could serve to prevent the inappropriate use of antibiotics (67). This quantification is currently done by a plate counting of the bacterial culture which is a long and inaccurate process. As reported here, once a signature is created, it can be transferred to other laboratories or other LC-MS systems working in Parallel Reaction Monitoring (PRM). Because, only 1.52% of the peptide signals in our PRM data report a signal-to-noise ratio below 10, we can expect that the method could be transferable to Selected Reaction Monitoring (SRM) mode on triple quadrupole (QQQ) instruments as some of those instruments have already been approved as Medical Devices for other applications (68) and are known for their low cost and robustness. However, in that case, it is expected to obtain lower signal-to-noise ratios due to the higher non-specific background in QQQ instruments. The transitions to monitor should also be carefully selected for each peptide of the signature and a fine tuning of the source voltages and collision energy should be performed. But, once this

optimization done, the method could be reused in routine on many samples. Moreover, it is also to note that the identification step could also been performed in DIA mode, since we have demonstrated on *E.coli* inoculates that, despite much more non-specific interferences , prediction rate where similar for both PRM or DIA peptide signature monitoring (data not shown). This finding opens the possibility for monitoring larger signatures (> 100 peptides) even with short gradients. Although the turnaround time to identify bacterial contaminants with this method is short (<4h), the non-parallelizable chromatographic time might limit its use in laboratories analyzing a high number of patient samples every day. Nevertheless, we showed that this time can be reduced from 90 to 30 minutes and could be even more shortened with a better acquisition scheduling, some optimization of the LC gradient or use of high-throughput LC devices (69). Sample preparation time could also be automated and optimized, for instance, the trypsin digestion here done in one hour might be reduced to a couple of minutes as reported in the literature (70, 71). Finally, in a context where the emergence of resistant bacterial strains poses a global public health threat (9, 72, 73), the development of fast methods for bacterial typing becomes essential. Indeed, broad spectrum antimicrobials are commonly prescribed to patients before obtaining the results of the clinical microbiological analysis, a practice that might not be sufficient to control the infection, especially when one considers the risk posed by antibiotic-resistant species and their transmission in the hospital environment or the community (74). For instance, *Staphylococcus aureus* has developed resistance to many antimicrobial drugs including last resort antibiotics and expresses an arsenal of virulence factors (75-77). Or, in

agriculture, the systematic use of antibiotics in farming leads to the selection of resistant bacteria that have been found in commercial food products (78). The accessibility with proteomics methods of the proteins involved in the resistance or virulence processes might constitute a challenge. However, several studies already reported the use of MALDI-TOF and LC-MS/MS to detect changes in the proteome of sensitive versus resistance strains (79-81). Thus, by including specific peptides belonging to proteins involved in resistance or virulence mechanisms in our peptidic signature, we could provide a measure of the risk associated to resistance or virulence and provide additional microbial information a physician could use to prescribe an appropriate antibiotic to a patient, thereby reducing the use of broad-spectrum antibiotics.

Finally, we anticipate that the constant improvement in sensitivity, mass accuracy and acquisition speed of mass spectrometers will contribute in the future to improve the limit and precision of specific bacterial strains detection, making even more relevant the use of LC-MS/MS methods in microbiology.

DATA AVAILABILITY

The raw mass spectrometry data are publicly available on ProteomeXchange (www.proteomexchange.org) upon the following identifiers: PXD013885 (DDA dataset), PXD013888 (DIA dataset) and PXD014970 (PRM dataset).

ACKNOWLEDGMENTS

The authors are grateful to Benjamin Nehmé and Ève Bérubé for assistance in sample preparation, Geneviève Durand for MALDI-TOF analyses, and Anne Gonzalez de Peredo and Luc Bissonnette for critical reading of the manuscript.

REFERENCES

1. Bisen PS, Debnath M, Prasad GBKS. Microbes concepts and applications. Hoboken: John Wiley & Sons; 2012. Available from: http://onlinelibrary.wiley.com/book/10.1002/9781118311912.
2. Murray PR, Baron EJ, American Society for Microbiology. Manual of clinical microbiology. 8th ed. Washington, D.C.: ASM Press; 2003.
3. Sharma S, Kaur N, Malhotra S, Madan P, Ahmad W, Hans C. Serotyping and Antimicrobial Susceptibility Pattern of Escherichia coli Isolates from Urinary Tract Infections in Pediatric Population in a Tertiary Care Hospital. J Pathog. 2016;2016:2548517.
4. Buehler SS, Madison B, Snyder SR, Derzon JH, Cornish NE, Saubolle MA, et al. Effectiveness of Practices To Increase Timeliness of Providing Targeted Therapy for Inpatients with Bloodstream Infections: a Laboratory Medicine Best Practices Systematic Review and Meta-analysis. Clin Microbiol Rev. 2016;29(1):59-103.
5. Kollef MH. Broad-spectrum antimicrobials and the treatment of serious bacterial infections: getting it right up front. Clin Infect Dis. 2008;47 Suppl 1:S3-13.
6. Leekha S, Terrell CL, Edson RS. General principles of antimicrobial therapy. Mayo Clin Proc. 2011;86(2):156-67.
7. Adamus-Bialek W, Baraniak A, Wawszczak M, Gluszek S, Gad B, Wrobel K, et al. The genetic background of antibiotic resistance among clinical uropathogenic Escherichia coli strains. Mol Biol Rep. 2018;45(5):1055-65.
8. Davies SC, Fowler T, Watson J, Livermore DM, Walker D. Annual Report of the Chief Medical Officer: infection and the rise of antimicrobial resistance. Lancet. 2013;381(9878):1606-9.
9. WHO. Antimicrobial resistance : global report on surveillance. Geneva: World Health Organization; 2014. xxii, 232 p. p.
10. Woo PC, Lau SK, Teng JL, Tse H, Yuen KY. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect. 2008;14(10):908-34.
11. Balloux F, Bronstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends Microbiol. 2018;26(12):1035-48.
12. Li W, Raoult D, Fournier PE. Bacterial strain typing in the genomic era. FEMS Microbiol Rev. 2009;33(5):892-916.

13. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol. 2017;243:16-24.
14. Field D, Hughes J, Moxon ER. Using the genome to understand pathogenicity. Methods Mol Biol. 2004;266:261-87.
15. Tagini F, Greub G. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. Eur J Clin Microbiol Infect Dis. 2017;36(11):2007-20.
16. Fagerquist CK, Garbus BR, Miller WG, Williams KE, Yee E, Bates AH, et al. Rapid identification of protein biomarkers of Escherichia coli O157:H7 by matrix-assisted laser desorption ionization-time-of-flight-time-of-flight mass spectrometry and top-down proteomics. Anal Chem. 2010;82(7):2717-25.
17. Singhal N, Kumar M, Kanaujia PK, Virdi JS. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. Front Microbiol. 2015;6:791.
18. Angeletti S. Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology. J Microbiol Methods. 2017;138:20-9.
19. Sloan A, Wang G, Cheng K. Traditional approaches versus mass spectrometry in bacterial identification and typing. Clin Chim Acta. 2017;473:180-5.
20. Marko DC, Saffert RT, Cunningham SA, Hyman J, Walsh J, Arbefeville S, et al. Evaluation of the Bruker Biotyper and Vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry systems for identification of nonfermenting gram-negative bacilli isolated from cultures from cystic fibrosis patients. J Clin Microbiol. 2012;50(6):2034-9.
21. Cheng K, Chui H, Domish L, Hernandez D, Wang G. Recent development of mass spectrometry and proteomics applications in identification and typing of bacteria. Proteomics Clin Appl. 2016;10(4):346-57.
22. Clark CG, Kruczkiewicz P, Guan C, McCorrister SJ, Chong P, Wylie J, et al. Evaluation of MALDI-TOF mass spectroscopy methods for determination of Escherichia coli pathotypes. J Microbiol Methods. 2013;94(3):180-91.
23. Gekenidis MT, Studer P, Wuthrich S, Brunisholz R, Drissner D. Beyond the matrix-assisted laser desorption ionization (MALDI) biotyping workflow: in search of microorganism-specific tryptic peptides enabling discrimination of subspecies. Appl Environ Microbiol. 2014;80(14):4234-41.
24. Ferreira L, Sanchez-Juanes F, Munoz-Bellido JL, Gonzalez-Buitrago JM. Rapid method for direct identification of bacteria in urine and blood culture samples by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: intact cell vs. extraction method. Clin Microbiol Infect. 2011;17(7):1007-12.
25. Wuppenhorst N, Consoir C, Lorch D, Schneider C. Direct identification of bacteria from charcoal-containing blood culture bottles using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. Eur J Clin Microbiol Infect Dis. 2012;31(10):2843-50.
26. Jeverica S, Nagy E, Mueller-Premru M, Papst L. Sample preparation method influences direct identification of anaerobic bacteria from positive blood culture bottles using MALDI-TOF MS. Anaerobe. 2018;54:231-5.

27. Wang H, Drake SK, Yong C, Gucek M, Lyes MA, Rosenberg AZ, et al. A Genoproteomic Approach to Detect Peptide Markers of Bacterial Respiratory Pathogens. Clin Chem. 2017;63(8):1398-408.
28. Wang H, Drake SK, Yong C, Gucek M, Tropea M, Rosenberg AZ, et al. A Novel Peptidomic Approach to Strain Typing of Clinical Acinetobacter baumannii Isolates Using Mass Spectrometry. Clin Chem. 2016;62(6):866-75.
29. Karlsson R, Gonzales-Siles L, Boulund F, Svensson-Stadler L, Skovbjerg S, Karlsson A, et al. Proteotyping: Proteomic characterization, classification and identification of microorganisms--A prospectus. Syst Appl Microbiol. 2015;38(4):246-57.
30. Cheng K, She YM, Chui H, Domish L, Sloan A, Hernandez D, et al. Mass Spectrometry-Based Escherichia coli H Antigen/Flagella Typing: Validation and Comparison with Traditional Serotyping. Clin Chem. 2016;62(6):839-47.
31. Jabbour RE, Deshpande SV, Stanford MF, Wick CH, Zulich AW, Snyder AP. A protein processing filter method for bacterial identification by mass spectrometry-based proteomics. J Proteome Res. 2011;10(2):907-12.
32. Boulund F, Karlsson R, Gonzales-Siles L, Johnning A, Karami N, Al-Bayati O, et al. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics. Mol Cell Proteomics. 2017;16(6):1052-63.
33. Foxman B. The epidemiology of urinary tract infection. Nat Rev Urol. 2010;7(12):653-60.
34. Ronald A. The etiology of urinary tract infection: traditional and emerging pathogens. Am J Med. 2002;113 Suppl 1A:14S-9S.
35. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. 2007;2(8):1896-906.
36. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4(11):923-5.
37. Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res. 2012;11(12):5773-80.
38. Mesuere B, Van der Jeugt F, Willems T, Naessens T, Devreese B, Martens L, et al. High-throughput metaproteomics data analysis with Unipept: A tutorial. J Proteomics. 2018;171:11-22.
39. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26(7):966-8.
40. Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ. Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat Protoc. 2015;10(6):887-903.
41. Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods. 2011;8(5):430-5.

42. Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, et al. Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Front Genet. 2019;10:452.
43. Frank E, Hall MA, Witten IH. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". 2016. In: Morgan Kaufmann [Internet]. Available from: www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.
44. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;11(6):O111 016717.
45. Gillet LC, Leitner A, Aebersold R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. Annu Rev Anal Chem (Palo Alto Calif). 2016;9(1):449-72.
46. Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, et al. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. Mol Cell Proteomics. 2017.
47. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. Cell. 2018;173(7):1581-92.
48. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405(2):442-51.
49. John GH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. the Eleventh Conference on Uncertainty in Artificial Intelligence; Montreal, QC, Canada: Morgan Kaufmann Publishers Inc.; 1995.
50. Bouckaert RR. Bayesian Network Classifiers in Weka2004 2004.
51. Hulten G, Spencer L, Domingos P. Mining Time-changing Data Streams. Seventh ACM SIGKDD International conference on knowledge Discovery and Data Mining; San Francisco, CA: ACM; 2001.
52. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions. Proc Natl Acad Sci U S A. 2014;111(23):8488-93.
53. Sung JY, Hwang Y, Shin MH, Park MS, Lee SH, Yong D, et al. Utility of Conventional Culture and MALDI-TOF MS for Identification of Microbial Communities in Bronchoalveolar Lavage Fluid in Comparison with the GS Junior Next Generation Sequencing System. Ann Lab Med. 2018;38(2):110-8.
54. He Y, Li H, Lu X, Stratton CW, Tang YW. Mass spectrometry biotyper system identifies enteric bacterial pathogens directly from colonies grown on selective stool culture media. J Clin Microbiol. 2010;48(11):3888-92.
55. Haigh JD, Green IM, Ball D, Eydmann M, Millar M, Wilks M. Rapid identification of bacteria from bioMerieux BacT/ALERT blood culture bottles by MALDI-TOF MS. Br J Biomed Sci. 2013;70(4):149-55.
56. Elbehiry A, Marzouk E, Hamada M, Al-Dubaib M, Alyamani E, Moussa IM, et al. Application of MALDI-TOF MS fingerprinting as a quick tool for identification and

clustering of foodborne pathogens isolated from food products. New Microbiol. 2017;40(4):269-78.
57. Dilger T, Melzl H, Gessner A. Rapid and reliable identification of waterborne Legionella species by MALDI-TOF mass spectrometry. J Microbiol Methods. 2016;127:154-9.
58. Totaro M, Valentini P, Casini B, Miccoli M, Costa AL, Baggiani A. Experimental comparison of point-of-use filters for drinking water ultrafiltration. J Hosp Infect. 2017;96(2):172-6.
59. Fernandez Garcia L, Alvarez Blanco S, Riera Rodriguez FA. Microfiltration applied to dairy streams: removal of bacteria. J Sci Food Agric. 2013;93(2):187-96.
60. Brewster JD, Paul M. Short communication: Improved method for centrifugal recovery of bacteria from raw milk applied to sensitive real-time quantitative PCR detection of Salmonella spp. J Dairy Sci. 2016;99(5):3375-9.
61. Bernhardt M, Pennell DR, Almer LS, Schell RF. Detection of bacteria in blood by centrifugation and filtration. J Clin Microbiol. 1991;29(3):422-5.
62. Wilson G, Aitchison LB. The use of a combined enrichment-filtration technique for the isolation of Campylobacter spp. from clinical samples. Clin Microbiol Infect. 2007;13(6):643-4.
63. Fourie J, Loskutoff N, Huyser C. Elimination of bacteria from human semen during sperm preparation using density gradient centrifugation with a novel tube insert. Andrologia. 2012;44 Suppl 1:513-7.
64. Buchanan CM, Wood RL, Hoj TR, Alizadeh M, Bledsoe CG, Wood ME, et al. Rapid separation of very low concentrations of bacteria from blood. J Microbiol Methods. 2017;139:48-53.
65. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017;27(4):626-38.
66. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13(5):435-8.
67. Kwon JH, Fausone MK, Du H, Robicsek A, Peterson LR. Impact of laboratory-reported urine culture colony counts on the diagnosis and treatment of urinary tract infection for hospitalized patients. Am J Clin Pathol. 2012;137(5):778-84.
68. Heaney LM, Jones DJ, Suzuki T. Mass spectrometry in medicine: a technology for the future? Future Sci OA. 2017;3(3):FSO213.
69. Bache N, Geyer PE, Bekker-Jensen DB, Hoerning O, Falkenby L, Treit PV, et al. A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. Mol Cell Proteomics. 2018;17(11):2284-96.
70. Lesur A, Varesio E, Hopfgartner G. Accelerated tryptic digestion for the analysis of biopharmaceutical monoclonal antibodies in plasma by liquid chromatography with tandem mass spectrometric detection. J Chromatogr A. 2010;1217(1):57-64.
71. Kim H, Kim HS, Lee D, Shin D, Shin D, Kim J, et al. Microwave-Assisted Protein Digestion in a Plate Well for Facile Sampling and Rapid Digestion. Anal Chem. 2017;89(20):10655-60.

72. WHO. The evolving threat of antimicrobial resistance : options for action. Geneva: World Health Organization; 2012.
73. Allcock S, Young EH, Holmes M, Gurdasani D, Dougan G, Sandhu MS, et al. Antimicrobial resistance in human populations: challenges and opportunities. Glob Health Epidemiol Genom. 2017;2:e4.
74. Laxminarayan R, Matsoso P, Pant S, Brower C, Rottingen JA, Klugman K, et al. Access to effective antimicrobials: a worldwide challenge. Lancet. 2016;387(10014):168-75.
75. Shorr AF. Epidemiology of staphylococcal resistance. Clin Infect Dis. 2007;45 Suppl 3:S171-6.
76. Lakhundi S, Zhang K. Methicillin-Resistant Staphylococcus aureus: Molecular Characterization, Evolution, and Epidemiology. Clin Microbiol Rev. 2018;31(4).
77. Oliveira D, Borges A, Simoes M. Staphylococcus aureus Toxins and Their Molecular Activity in Infectious Diseases. Toxins (Basel). 2018;10(6).
78. Landers TF, Cohen B, Wittum TE, Larson EL. A review of antibiotic use in food animals: perspective, policy, and potential. Public Health Rep. 2012;127(1):4-22.
79. Park AJ, Krieger JR, Khursigara CM. Survival proteomes: the emerging proteotype of antimicrobial resistance. FEMS Microbiol Rev. 2016;40(3):323-42.
80. Mekonnen SA, Palma Medina LM, Michalik S, Loreti MG, Gesell Salazar M, van Dijl JM, et al. Metabolic niche adaptation of community- and hospital-associated methicillin-resistant Staphylococcus aureus. J Proteomics. 2019;193:154-61.
81. Lin MH, Potel CM, Tehrani K, Heck AJR, Martin NI, Lemeer S. A New Tool to Reveal Bacterial Signaling Mechanisms in Antibiotic Treatment and Resistance. Mol Cell Proteomics. 2018;17(12):2496-507.

Figure 1. **Workflow of the method for bacterial identification.** The workflow is composed of two steps: the "training" step defines of a peptidic signature for the bacteria of interest; the "identification" step uses this signature in routine to identify bacteria in biological samples.

Figure 2. **Heatmap of the peptidic signature corresponding to the 15 most frequently found bacteria in UTI.** Intensity of each of the 82 peptides identified by the machine learning algorithm is represented for the six high-level concentration replicates of urine inoculation for each bacteria of interest. Data are presented with hierarchical clustering in rows and columns.

Figure 3. **Accuracy, linearity and reproducibility of the 'identification' step of the method performed on the four most frequent bacteria in UTI.** (a) Prediction reported by the algorithm after peptidic signature monitoring by PRM associated with its probability (light blue : high probability, dark blue : low probability) for five concentrations corresponding to five inoculation volumes (1, 2, 10, 20 and 100µL or 2, 4, 20, 40 and 200µL) of four bacteria (Eco, Efa, Kpn or Sag) in urine of four different healthy volunteers (A, B, C, and D), dotted red line corresponds to the commonly used clinical laboratories detection threshold of 1 x 10^5 CFU/mL; (b) Linearity curves obtained for 4 peptides of the peptidic signature with the samples across the five tested concentrations, dotted red line corresponds to the commonly used clinical laboratories detection threshold of 1 x 10^5 CFU/mL; (c) Pearson correlation coefficients between two of the four biological samples (*i.e* four different urines of healthy volunteers).

Figure 4. **Accuracy, linearity and reproducibility of the 'identification step' of the method performed in two experimental conditions**: 90 minutes gradient at nanoflow rate with PRM acquisition on an Orbitrap Fusion instrument or 30 minutes gradient at capillary flow rate with PRM acquisition on a Q Exactive HF-X instrument. (a) Right (green) or wrong (red) prediction reported by the algorithm after peptidic signature monitoring by PRM associated to its probability, for five concentrations corresponding to five inoculation volumes (1, 2, 10, 20 and 100µL or 2, 4, 20, 40 and 200µL) of four bacteria (Eco, Efa, Kpn or Sag) in urine of four different healthy volunteers (A, B, C, and D), dotted red line corresponds to the commonly used clinical laboratories detection threshold of 1e5 CFU/mL; (b) Distribution of the determination coefficients of the linearity curves obtained with the same samples across the five tested concentrations, the dotted line represents the average of all values; (c) Distribution of the Pearson correlation coefficients obtained by comparison of two biological replicates with the same samples, the dotted line represents the average of all values.

Figure 5. **Comparison of our fast LC-MS method and the standard MALDI-TOF method**. Prediction reported by the algorithm after peptidic signature monitoring by PRM without bacterial culture (red crosses) or by the MALDI Biotyper analysis after 24h hours bacterial culture (blue circles) on (a) 27 patients urine specimens or (b) 15 inoculations of bacterial species into urine from healthy volunteers.

Fig.1

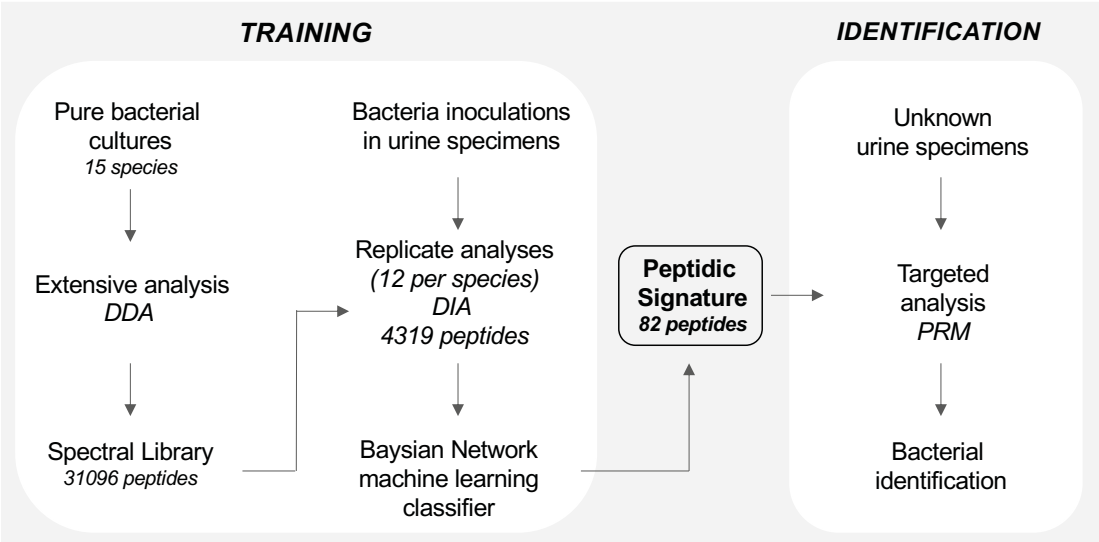

Figure 2

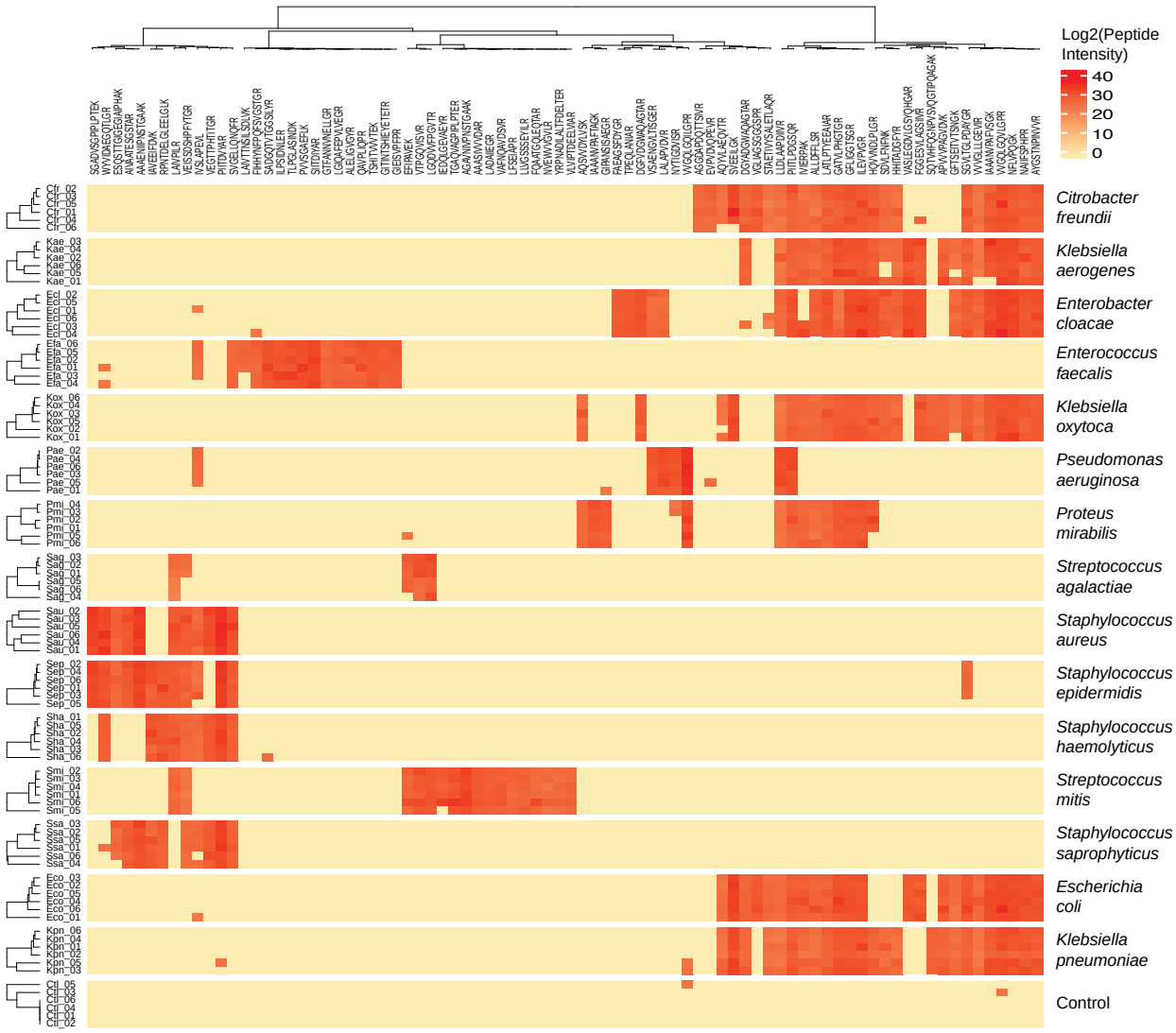

Figure 3

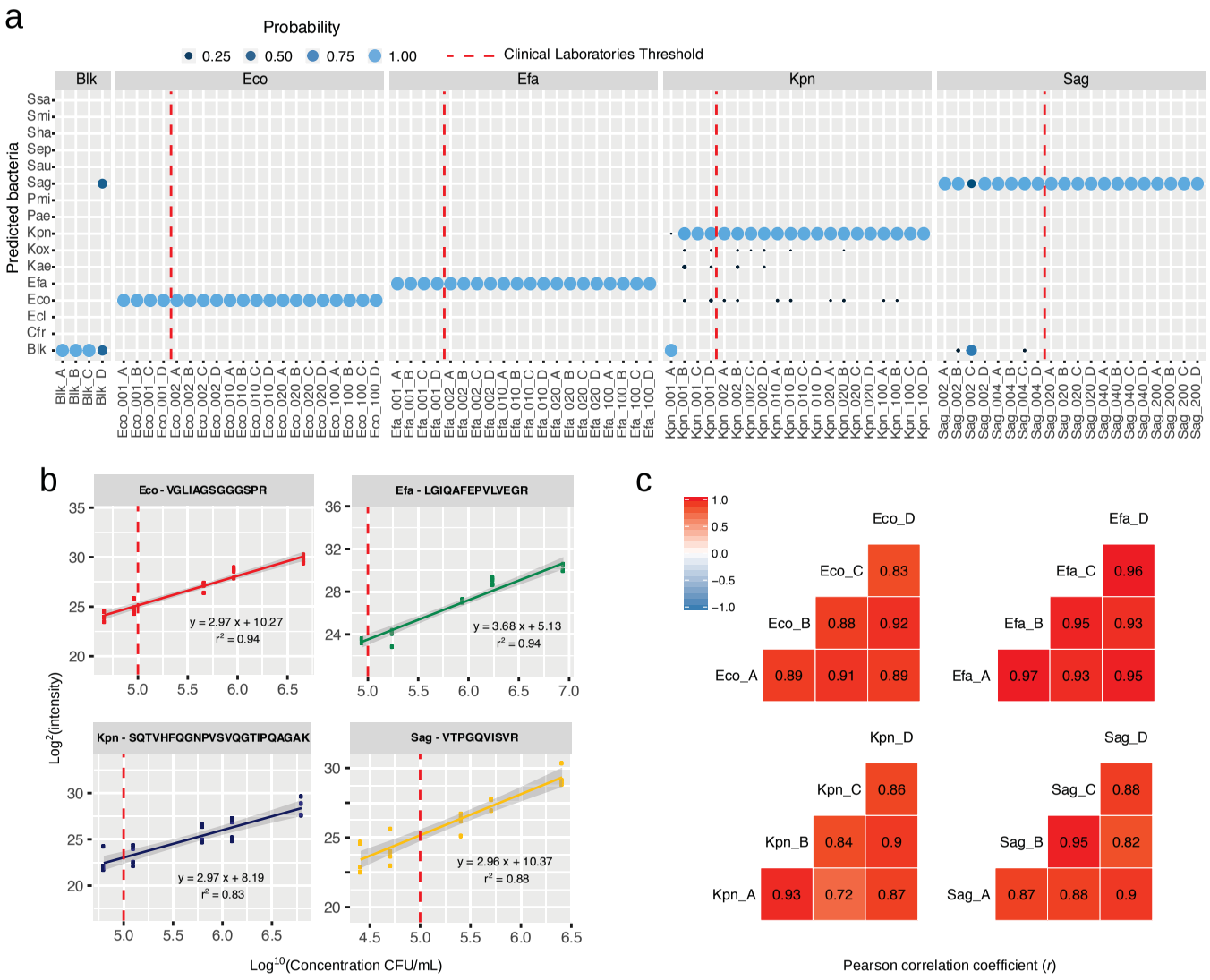

Figure 4

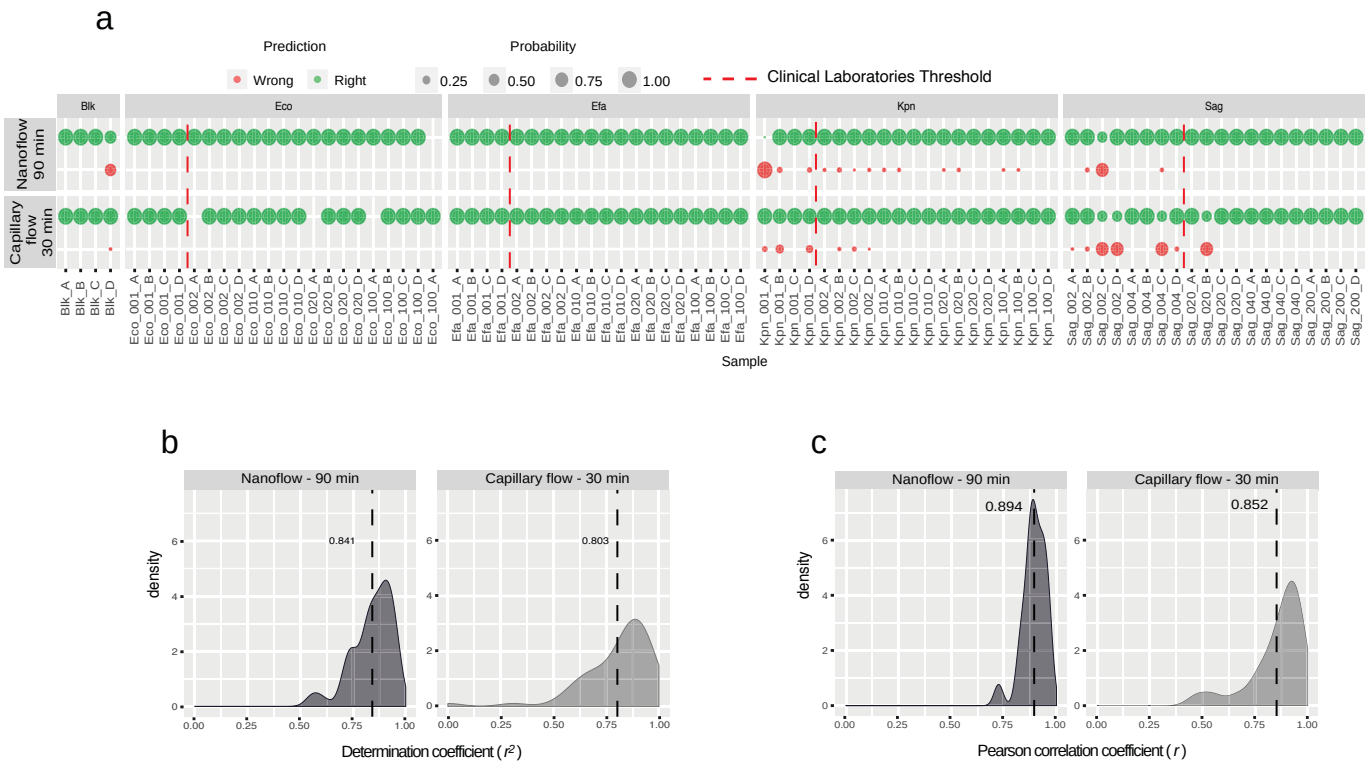

Figure 5