Identification of Microorganisms by Liquid Chromatography-Mass Spectrometry (LC-MS$^1$) and in silico Peptide Mass Libraries

Peter Lasch$^{1\star}$, Andy Schneider$^1$, Christian Blumenschei$^1$ and Joerg Doellinger$^1$

1 Robert Koch-Institute, ZBS6 - Proteomics and Spectroscopy, Seestraße 10, Berlin, D-13353, Germany

* corresponding author, e-mail: LaschP@rki.de, phone +49-30-18754 2259

2. Running Title: In silico peptide databases for LC-MS$^1$ microbial biotyping

Keywords: Bacteria, Microbiology, Bioinformatics software, Diagnostics, Mass Spectrometry, Identification of Microorganisms, LC-MS$^1$

3. Abbreviations: ACN, acetonitrile; AGC, automatic gain control; DTT, dithiothreitol; CAA, 2-chloroacetamide; FA, formic acid; ID, identifier; MALDI-TOF, matrix-assisted laser desorption/ionization - time–of–flight; MBT, MALDI Biotyper; MS, mass spectrometry; MW, molecular weight; NCE, normalized collision energy; ppm, parts per million; RKI, Robert Koch-Institute; SDS, sodium dodecyl sulfate; SNR, signal-to-noise ratio; SPEED, sample preparation by easy extraction and digestion; STrap, suspension trapping; TFA, trifluoroacetic acid; TSA, tryptic soy agar; TCEP, Tris(2-carboxyethyl)phosphine; UniProtKB, UniProt Knowledgebase
4. ABSTRACT
Over the past decade, modern methods of mass spectrometry (MS) have emerged that allow reliable, fast and cost-effective identification of pathogenic microorganisms. While MALDI-TOF MS has already revolutionized the way microorganisms are identified, recent years have witnessed also substantial progress in the development of liquid chromatography (LC)-MS based proteomics for microbiological applications. For example, LC-tandem mass spectrometry (LC-MS\textsuperscript{2}) has been proposed for microbial characterization by means of multiple discriminative peptides that enable identification at the species, or sometimes at the strain level. However, such investigations can be laborious and time-consuming, especially if the experimental LC-MS\textsuperscript{2} data are tested against sequence databases covering a broad panel of different microbiological taxa.

In this proof of concept study, we present an alternative bottom-up proteomics method for microbial identification. The proposed approach involves efficient extraction of proteins from cultivated microbial cells, digestion by trypsin and LC-MS measurements. Peptide masses are then extracted from MS\textsuperscript{1} data and systematically tested against an \textit{in silico} library of all possible peptide mass data compiled in-house. The library has been computed from the UniProt Knowledgebase covering Swiss-Prot and TrEMBL databases and comprises more than 12,000 strain-specific \textit{in silico} profiles, each containing tens of thousands of peptide mass entries. Identification analysis involves computation of score values derived from correlation coefficients between experimental and strain-specific \textit{in silico} peptide mass profiles and compilation of score ranking lists. The taxonomic positions of the microbial samples are then determined by using the best-matching database entries. The suggested method is computationally efficient – less than two minutes per sample - and has been successfully tested by a test set of 39 LC-MS\textsuperscript{1} peak lists obtained from 19 different microbial pathogens. The proposed method is rapid, simple and automatable and we foresee wide application potential for future microbiological applications.
5. INTRODUCTION
Rapid and reliable identification of pathogenic bacteria is of vital importance in many areas of public health and is relevant also in the food industry and for biodefense. In the context of clinical microbiology, a large variety of very different techniques, among them biochemical, serological, chemotaxonomic, and more recently spectroscopic, spectrometric and genomic tools are routinely utilized. For example, mass spectrometry-based techniques, such as matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) have emerged as invaluable tools for accurate and cost-effective identification of microorganisms in today’s routine clinical microbiology (1-4). The MALDI-TOF MS approach allows obtaining the genus and species identity of unknown samples by matching microbial mass spectra against spectral libraries collected from microorganisms with a known taxonomic identity. While identification is most reliably achieved at the species level, the question of whether MALDI-TOF MS is suitable for identification and discrimination below the species level is still controversially discussed by the scientific community (4-7).

While a large number of studies convincingly demonstrate successful discrimination and identification of pathogenic bacteria by MALDI-TOF MS at the species level, there is also ample evidence for limitations of the taxonomic resolution, particularly at the infraspecies level and when dealing with differentiation of genetically closely related species (8-11). For example, differentiation between Escherichia coli and Shigella (12, 13) or of Bacillus cereus and Bacillus anthracis (14, 15) requires additional measures beyond the standard microbial identification workflow, such as custom reference libraries, higher levels of standardization, and/or sophisticated data analysis concepts (machine learning, etc.). In these studies the reduced discriminatory power of MALDI-TOF MS has been attributed to the restricted m/z range (m/z 2000-20,000) and the dependence from mass patterns produced by a sub-proteome of small, abundant and basic proteins, mainly ribosomal subunit proteins (8). Since the molecular evolution of these proteins is rather slow, ribosomal proteins are supposed to offer only limited taxonomic specificity.
Further limiting factors of the MALDI-TOF MS method are the relatively low resolution, resulting in decreased selectivity and a reduced dynamic sensitivity, i.e. a lowered detectability of protein signals over a wide concentration range (11).

In contrast to MALDI-TOF MS, liquid chromatography-tandem MS (LC-MS\textsuperscript{2}) generally detects large numbers of signals at very high resolution with very high mass accuracy in a single run (11). Shotgun proteomic methods observe proteolytic cleavage products, often tryptic peptides, instead of intact proteins. This enables MS data collection with high analytical sensitivity. Moreover, coupling of mass spectrometry with chromatographic separation (LC) has shown to increase the dynamic sensitivity and allows sensitive detection also of low abundant peptides. Finally, LC-MS\textsuperscript{2} is much less restricted to classes of proteins with specific physicochemical properties. Even though proteomic techniques are still complex, rather cost-intensive and limited for use by well-equipped laboratories, the many advantages of LC-MS have led to an increasing number of activities aiming at evaluating potential applications of LC-MS in microbiology (16-20).

Various groups have used shotgun proteomics for the classification and identification of pathogenic microorganisms. For example, a proteomics-based workflow for bacterial identification has been suggested by Dworzanski which involved construction of a bacterial proteome database from bacterial genomes, LC-MS\textsuperscript{2} data acquisition from digested bacterial cell extracts, identification of tryptic peptides and sequence-to-bacterium assignments (21). The approach has been later utilized to determine the relatedness among strains of \textit{B. cereus} sensu stricto, \textit{B. thuringiensis} and \textit{B. anthracis} by estimating fractions of shared peptides derived from a prototype database (22). LC-MS\textsuperscript{2} has been also used by Tracz and coworkers to identify Biosafety Level 3 bacteria (19). Sequence data from tryptic microbial peptides were obtained and employed for Mascot searches against a database containing concatenated protein sequences derived from microbial genomes. Identification of bacterial species was carried out by
summing up matches from unique and degenerated (shared) peptides found per concatenated sequence; a post-culture analysis time of less than 8 hours has been reported.

Another alternative based on LC-MS has been proposed by Jabbour. Bacterial samples were lysed and subjected to tryptic digestion followed by LC-MS\(^2\) (16). Subsequently, peptides were identified and matched against databases. Bacteria were then identified based on the assessment of unique peptides obtained by an algorithm called BACid.

Comparison between microbial protein sequence data obtained by bottom-up tandem MS and reference databases was also performed by Boulund and colleagues. The proposed analysis pipeline (TCUP) not only returned specific genes of reference genomes that matched with peptide sequences determined by LC-MS\(^2\), but provided also the relative abundances of individual bacteria identified in a given mixed culture (23). In this way, TCUP allowed typing and characterizing pathogenic bacteria from pure cultures and to estimate the relative abundances of individual microbial species from mixed microbial samples (23). In the same year Berendsen and co-workers suggested a generic LC-MS\(^2\) method for the identification of microorganisms from positive blood cultures (18). A LC-MS compatible sample preparation method was developed that enabled accurate identification of bacteria grown in blood culture flasks to species level based on LC-MS\(^2\) bottom-up proteomics, database searches and matching with taxon-specific discriminative peptides.

Advantages of the LC-MS\(^2\)-based approaches for bacterial identification outlined above are the excellent accuracy of identification, high taxonomic resolution, universal applicability to the ever-growing numbers of known microbes and the ability to identify bacteria from mixtures, e.g. in polymicrobial infections. At the other hand the comparatively high computational requirements have to be mentioned. Since the time required for peptide identification correlates with the number of entries contained in sequence databases, computation time can be saved by restricting the size of the database, for example by using genus-specific databases. However, database restrictions contradict the use of shotgun proteomics as an
unbiased approach for microbial identification. Another important limitation of LC-MS\(^2\)-based approaches is the severely reduced accuracy and sensitivity of common search algorithms when extensive protein sequence databases are used. The large search space impedes the identification of true peptide matches within large numbers of similar sequences.

With this proof-of-concept study, we introduce an alternative, easy-to-use and computational less demanding approach for microbial identification. The proposed method is based on bottom-up proteomics as the analytical technique and involves acquisition of LC-MS data from pure microbial cultures. MS\(^1\) data are extracted and tested against a database compiled in-house using public protein databases (UniProtKB) with currently more than 12,000 strain-specific *in silico* mass profiles. We demonstrate that the MS\(^1\) information can be used for rapid and accurate taxonomic identification, at least at the species level, and discuss possibilities to combine the suggested analysis pipeline with known MS\(^2\)-based analysis methods in microbiology.

### 6. EXPERIMENTAL PROCEDURES

**Microbial strains:** The performance and accuracy of the proposed method for microbial identification was tested using 19 well-characterized bacterial strains which were predominantly obtained from established strain collections such as DSM (Deutsche Sammlung von Mikroorganismen), ATCC (American Type Culture Collection) and NCTC (National Collection of Type Cultures). Strains E 125, E 131 and E153 of *Burkholderia thailandensis* originated from the strain collection at the Robert Koch-Institute (RKI) (24). An overview of the microbial strains and species utilized is given in Tab. 1.

**Cultivation:** Bacteria were streaked under sterile conditions on solid culture media by an inoculation loop and incubated for 48 hours. Strains prepared by the STrap protocol (see below and Tab. 1) were cultivated according to species-specific cultivation requirements on tryptic soy agar (TSA, ReadyPlate TSA ISO, Merck Life Science) or Middlebrook agar produced in-house. Cells from these cultures were
harvested by scraping; approximately 10 µL of microbial material was then washed three times by 1 mL of ice-cold phosphate buffered salt solution and centrifuged at 4000 g at 4°C for 5 min. All other bacteria were cultured on Casein-Soy-Peptone (Caso, Oxoid, Wesel, Germany) agar plates under aerobic conditions for 24 h at 37°C and further processed according to the protocol Sample Preparation by Easy Extraction and Digestion (SPEED, see below and Tab. 1 for details).

**Experimental design and statistical rationale**

In this proof of concept study we tested the proposed MS1-based identification workflow by proteomic data from 19 different bacterial strains from which 39 RAW files were collected. The *Burkholderia* subset (see below and supporting information) included biological and technical replicate spectra. LC-MS measurements were shuffled in most cases in such a way, that technical replicates of the same sample were not measured consecutively.

**Sample preparation by suspension trapping (STrap):** Cells were suspended in 5 % sodium dodecyl sulfate (SDS), 20 mM dithiothreitol (DTT), 50 mM Tris/HCl buffer, pH 7.6 (sample/buffer 1:10 (v/v)), incubated at 95°C for 10 min and further sonicated for 15 cycles à 30 s at high intensity level and 4°C using Bioruptor®Plus (Diagenode, Liege, Belgium). Protein concentrations were determined by measuring the tryptophan fluorescence at an emission wavelength of 350 nm using 295 nm for excitation with an Infinite® M1000 PRO microplate reader (Tecan, Männedorf, Switzerland) according to (25). Samples were further processed using S-Trap™ mini filters according to the STrap sample preparation protocol (26) and the manufacturer’s instructions (Protifi, Huntington NY, USA).

**Sample preparation by the SPEED protocol:** SPEED protocol-based preparation of microbial cells was carried as outlined in (27). Briefly, bacterial cells were suspended in trifluoroacetic acid (TFA, Uvasol® for spectroscopy, Merck, Darmstadt, Germany) in a sample/TFA ratio of 1:4 (v/v) and heated at 70°C for 3 min. Acid extracts were then neutralized with 2M TrisBase using the tenfold volume of TFA solution and further incubated at 95°C for 5 min after adding tris (2-carboxyethyl)phosphine (TCEP) to a final
concentration of 10 mM and 2-chloroacetamide (CAA) to a final concentration of 40 mM. Protein concentrations were determined by turbidity measurements at 360 nm (1 AU = 0.79 µg/µL) using a GENESYS™ 10S UV-Vis Spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and adjusted to 0.25 µg/µL using a 10:1 (v/v) mixture of 2M TrisBase and TFA. The solution was afterwards diluted 1:5 (v/v) with water. Proteins were digested for 20 h at 37°C using Trypsin Gold, Mass Spectrometry grade (Promega, Fitchburg, USA) at an enzyme/protein ratio of 1:50 (w/w).

**Peptide desalting:** Peptides were desalted using 200 µL StageTips™ packed with three Empore™ SPE Disks C18 (3M Purification, Inc., Lexington, USA) according to (28) and concentrated using a vacuum concentrator. Samples were resuspended in 20 µL 0.1 % formic acid (FA) and peptides were quantified by measuring the absorbance at 280 nm using a Nanodrop 1000 device (Thermo Fisher Scientific, Rockford, IL, USA).

**Nano-LC-MS measurements:** Nano-LC tandem MS (nLC-MS²) measurements of pathogenic bacteria were carried out within the scope of different proteomics projects. Desalted digests were analyzed on an EASY-nanoLC 1200 device (Thermo Fisher Scientific, Bremen, Germany) coupled online to a Q Exactive™ Plus mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Two different LC setups were used. 1 µg peptides of samples #1 – 13 were loaded on a Acclaim™ PepMap™ trap column (20 mm x 75 µm i.d., 100 Å, C18, 3 µm; Thermo Fisher Scientific, Bremen, Germany) at a flow rate of 3 µL/min and were subsequently separated on a 200 cm µPAC™ column (PharmaFluidics, Ghent, Belgium) using a 160 min gradient of 3 to 28 % acetonitrile (ACN) in 0.1 % FA at 300 nL/min flow rate. Column temperature was kept at 50°C using a butterfly heater (Phoenix S&T, Chester, PA, USA). In contrast, 1 µg peptides of samples #14 – 21 were directly loaded and separated on a 50 cm Acclaim™ PepMap™ column (75 µm inner diameter, i.d., 100 Å C18, 2 µm; Thermo Fisher Scientific, Bremen, Germany) using a linear 120 min gradient of 3 to 28 % ACN in 0.1 % FA at 200 nL/min flow rate using a column temperature of 40°C. The mass spectrometer was operated in a data-dependent manner in the m/z range of 300 – 1,650. Full scan
spectra were recorded with a resolution of 70,000 using an automatic gain control (AGC) target value of $3 \times 10^6$ with a maximum injection time of 20 ms. Up to the 10 most intense $2^+ – 5^+$ charged ions were selected for higher-energy c-trap dissociation (HCD) with a normalized collision energy (NCE) of 25 %. Fragment spectra were recorded at an isolation width of 2 Th and a resolution of 17,500@200 m/z using an AGC target value of $1 \times 10^5$ with a maximum injection time of 50 ms. The minimum AGC MS$^2$ target value was set to $1 \times 10^4$. Once fragmented, peaks were dynamically excluded from precursor selection for 30 s within a 10 ppm window. Peptides were ionized using electrospray with a stainless steel emitter, i.d. 30 µm, (Proxeon, Odense, Denmark) at a spray voltage of 2.0 kV and a heated capillary temperature of 275°C.

**Extraction of MS$^1$ data from MS$^2$ data sets, pre-processing of MS$^1$ data:** This study entirely relied on MS$^1$ data. Peptide feature detection was carried out using the Minora algorithm of the Proteome Discoverer software v. 2.2.0388 (Thermo-Fisher Scientific) with default settings. Peptide feature text files were then processed by `LCMS-Biotyping.vvf`, an in house-written method compiled for obtaining the following parameters from the peptide features: MS$^1$ peak positions (in m/z units) with the respective ion charge state, normalized abundance and signal-to-noise ratio (SNR). These parameters were stored in tab-separated text files and imported by a custom-designed Matlab function `readlcmsstxtfile` (Matlab, The Mathworks, Natick, USA). As part of the `parseuniprot` toolbox (see below) this function supports import of peptide feature text files obtained from LC-MS$^1$ data and performs data pre-processing, including molecular weight (MW) determination by considering charge states, detecting and removing peak entries originating from oxidized peptides (mass shift +15.99491 Da) as well as from peptides with deamidated glutamine or asparagines residues (mass shift +0.98402 Da). Spectral pre-processing involved furthermore partially removing (underweighting) of low intensity and low MW peaks; based on the principle that the relevance of a specific peak for subsequent identification analysis co-varies with its intensity and MW values (see below). As the result of pre-processing, experimental LC-MS$^1$ data of a
given sample is collapsed into a single MS\textsuperscript{1} mass peak list, which contains the filtered peptide mass data. Such data are in the following referred to as experimental LC-MS\textsuperscript{1} peak lists, or – after conversion into continuous spectra - as LC-MS\textsuperscript{1} test spectra.

**Compilation of the in silico peptide mass database:** Assembly of the *in silico* library of microbial peptide mass data was done utilizing information from the UniProt Knowledgebase (UniProtKB). In particular, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL text files, *uniprot_sprot_bacteria.dat.gz* (size: 211,465,997 byte, date: Dec 10, 2018) and *uniprot_trembl_bacteria.dat.gz* (60,987,643,246 byte, Dec 09, 2018) were downloaded from ftp://ftp.uniprot.org and unpacked. Extraction of relevant data i.e. of protein sequences and metadata was carried out by means of *parseuniprot*, a collection of specifically developed Matlab-based functions (The Mathworks). The *parseuniprot* toolbox is available from the principal author of this study on request and will be publicly available after publication of this study.

Compilation of the *in silico* database and identification analysis was carried using a dual processor Dell Precision T7500 workstation (2× Intel 3.46 GHz x5690 CPU with 6 cores each) that was equipped with 160 GB RAM, a Samsung 512 GB solid-state drive (850 Pro) and additional 20 TB hard drive space. A 64-bit version of Microsoft Windows 7 operating system and Matlab-based software developed in-house (*parseuniprot*, MicrobeMS) running under a 64-bit version of Matlab R2014a (The Mathworks, Natick, USA) were utilized. MicrobeMS and *parseuniprot* required the Statistics (v. 9.0), Bioinformatics (v. 4.4) and Parallel Computing (v. 6.4) toolboxes of Matlab. The output of the *parseuniprot* toolbox i.e. the *in silico* library of microbial peptide mass data can be stored as a so-called *pkf* file. Files of this type are compatible with MicrobeMS, the identification software utilized (29); a detailed description of the *pkf* file format is provided in the MicrobeMS wiki (30).

**Bacterial Identification analysis:** For bacterial identification analysis we used the MicrobeMS toolbox which has been initially developed in the context of biotyping pathogenic microorganisms by MALDI-TOF MS (10, 15). This Matlab-based toolbox can be freely downloaded from the MicrobeMS website...
In silico peptide database for LC-MS1 microbial biotyping (compiled version for Microsoft Windows 64-bit, registration required). Within the scope of the present study the MicrobeMS toolbox was specifically adapted in order to meet the larger computational and higher memory requirements.

For identification analyses LC-MS1 peak lists (see above) were first imported via the muf data format that is specific for MicrobeMS (30). Inter-spectral distances between experimental LC-MS1 peak lists and in silico peptide mass profiles were then obtained utilizing the function compare with DB of MicrobeMS. Bar coded MS1 test spectra were constructed from LC-MS1 peak lists using MW bins of a relative width of 1.2 ppm. As distance metrics, variance-scaled Pearson's product momentum correlation coefficients (Pareto scaling 0.25) were selected, whereby data between 2000 and 5500 Da served as inputs. The calibration range factor was set to a value of 2 giving the total number of calibration factor variations of 125, see MicrobeMS wiki for details (30). In MicrobeMS correlation coefficient-based inter-spectral distances are converted to score values between 0 and 1000. This is achieved on the basis of linear scaling, whereby a score of 1000 can only be achieved if the LC-MS1 test and a given in silico database profile match entirely (identity). A score of zero, on the other hand, is obtained only when any correlation is absent.

In MicrobeMS the score values determined between a test spectrum and the strain-specific in silico mass profiles are arranged in a ranking list and the best matching database entries are used to determine the taxonomic identity of the strain investigated. This approach is not new and has been used for many years in infrared, Raman, or MALDI-TOF MS identification software solutions such as Bio-Rad's KnowItAll, Bruker's MALDI Biotyper, or Biomerieux's Saramis / VITEK MS. In the current implementation of MicrobeMS, the score ranking list is provided in a HTML format where the top 30 best-matching in silico database records are displayed for each LC-MS1 test spectrum analyzed (30), see also supporting information, SI.
7. RESULTS

**Overview of the identification analysis workflow:** In this proof of concept study we present a computational pipeline which is suitable for identification of pathogenic microorganisms from bottom-up mass spectrometry (MS) data. An overview of the proposed sequence of analysis steps is presented in Fig. 1. First steps of the proposed approach include cultivation of pathogenic bacteria under standardized conditions followed by sample preparation using validated protocols. Microbial materials are subsequently characterized by means of nanoLC-MS (LC-MS\(^1\)). First steps of analysis involve extraction and pre-processing the MS\(^1\) data. LC-MS\(^1\) test spectra obtained are then systematically compared against entries of an *in silico* database, which has been compiled beforehand from the totality of microbial protein entries present in the UniProt Knowledgebase, i.e. in UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases. The output of this comparison is a ranking list of score values derived from inter-spectral correlation, or distance values. Score lists provide information on the taxonomic identity of the microorganism under investigation.

**Compilation of an in silico database of microbial peptide mass data:** Compilation of a database of strain-specific peptide mass profiles from microbial proteome data constituted the main challenge in this proof of concept study. Development of such a database required analyzing the complete bacterial UniProtKB/Swiss-Prot library with reviewed and manually annotated proteins and the UniProtKB/TrEMBL proteome data (unreviewed, computationally analyzed proteins). Both mutually exclusive databases were merged and the protein sequence information and scattered metadata were extracted, further processed, re-sorted and stored in a format the microbial identification software can read. The specified data analysis steps were carried out by means of the `parseuniprot` toolbox, a Matlab-based analysis pipeline specifically developed at RKI. In addition to supplementary command-line Matlab functions, e.g. for merging the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases or for creating and modifying taxonomy white lists (see below), the `parseuniprot` toolbox comes with three major functions, (i)
readdat, (ii) resort and (iii) modfeat which are consecutively executed and whose analysis steps build upon each other.

Fig. 2 schematically illustrates the sequence of the analysis steps readdat, resort and modfeat. The first function parseuniprot/readdat has been employed for data import from a UniProtKB file format and for generating a first Matlab structure array. In this array each protein is mapped to an array element whereby each element comprises specific fields that contain protein sequence and metadata as well as protein, proteome and taxonomic identifiers (IDs). In order to reduce the overall data amount and to exclude entries with unclear taxonomic assignment from further analyses parseuniprot/readdat has also support for “whitelisting” organisms. The white list was obtained from UniProtKB non-redundant proteomes and initially contained 22,529 taxon entries with genus, species and strain designation, protein numbers, and the corresponding UniProtKB taxonomic and proteome IDs. Redundant proteomes are defined as proteomes in which all or nearly all protein sequences are highly similar or identical to an existing proteome from the same species (31). The raw white list was additionally scanned for keywords like 'unidentified', 'candidatus', 'sp.', 'uncultured', etc. for subsequent removal of respective entries so that the final white list included more than 13,000 different strain entries.

The data structure produced by parseuniprot/readdat served as input for the second function, parseuniprot/resort and contained altogether sequences and metadata from 47,537,746 microbial proteins. Subsequent analyses involved disregarding strain entries with less than 1200 proteins, identification/assignment of posttranslational modification (PTM) sites such as cleavage of N-terminal methionine and in silico digestion by trypsin using the rule “cleave at the carboxyl termini of lysine or arginine, except when either is followed by proline”. No missed cleavages were allowed. Afterwards, exact peptide molecular weight (MW) determination was conducted from amino-acid sequences of tryptic peptides whereas it was considered that carbamidomethylcysteines (160.030647 Da) are present in CAA treated samples instead of cysteine (103.009185 Da). Although principally available through
Matlab’s Bioinformatics Toolbox, PTM profiling and MW determination routines had to be re-written for performance reasons and, after careful testing, integrated into the parseuniprot toolbox. All calculated MW values of tryptic peptides were then re-indexed according to proteome IDs. As a result, each proteome or taxon endpoint in question was associated with the respective peptide entries which permitted compiling strain-specific peptide mass lists after sorting, filtering and cleansing. Peak lists are then converted to bar-coded spectra, or profiles using MW bins of a predefined relative width. Spectra are strain-specific and are in the following referred to as in silico peptide mass profiles, or in silico spectra. Each profile contains the totality of in silico peptides of a given microbial strain no matter if peptides are taxon-specific (i.e. unique) with regard to MS²-based workflows, or not. In this context it is important that a combination of peptide features can be highly specific, even if the individual mass features it contains are not specific in itself.

Data was allowed for use in further analyses in cases where the peptide MW was larger than 780 Da and if the profiles contained more than 15,000 and less than 280,000 peptide entries. After application of these criteria, the in silico database for LC-MS¹ based microbial identification contained 723,820,940 individual peptide entries sorted in altogether 12,044 strain-specific mass profiles which corresponds to approximately 60,000 peptide peaks per profile.

The third function of the analysis pipeline, parseuniprot/modfeat, was designed for feature selection, specifically to allow filtering of presumptive non-specific peptide mass entries with the objective to increase the accuracy of subsequent identification analysis. In the present implementation of the parseuniprot/modfeat function selected peptide MW entries are removed on the basis of the following considerations: MW data of a given in silico profile should ideally be taxon-specific, i.e. should preferentially be entirely absent, or at least less represented in the other database profiles. To this end, an algorithm has been tested which served the purpose to determine throughout the whole database the relative MW frequency of each bin (i.e. relative occupancy of bins). Subsequently, frequency values
are extracted for all non-empty bins of the first *in silico* profile. Frequencies values are ranked in descending order and MW bin values of the first profile above a certain frequency threshold are set to zero. The procedure has to be repeated for each profile of the *in silico* database. Tests revealed that the overall accuracy of identification increased if peptide MW data above the 90th frequency percentile are disregarded from further analyses. In this way, ~10% of the MW data with presumably non-specific information are removed from the *in silico* database. The adjusted *in silico* spectra and selected metadata were subsequently stored in the *pkf* file format, see above.

For the future, it is planned to test and implement more advanced feature selection approaches. Genetic algorithms, for example, could be advantageously applied to identify combinations of taxon-specific peptide markers and could thus contribute to improve the accuracy of microorganism identification. Fig. S1 shows a screenshot with the graphical user interface of the current version of the *parseuniprot* toolbox.

**Spectral pre-processing:** Pre-processing of raw experimental data aims at increasing the robustness and accuracy of subsequent quantitative or classification analysis (32, 33). In the context of the present study, the strategy of pre-processing LC-MS$^1$ spectra was inspired by the following ideas:

Firstly, the number of experimentally determined MS$^1$ peaks usually varied between 60,000 and 90,000 per strain. It is reasonable to assume that relevant fractions of these peaks carry non-specific information with some of them arising from chemically modified peptides. For the sake of simplicity, we have proposed that the intensity of such peaks is lower on average compared to intensities from unmodified peptides. Thus, underweighting low intensity features from MS$^1$ spectra was assumed to have a positive impact on the accuracy of identification. Secondly, high MW peptides are thought to be somewhat more specific with regard to pathogen identification than short peptides with a lower MW. An important objective of data pre-processing was therefore to eliminate low intensity peaks in low MW regions at a higher rate than in high MW regions. Thirdly, peaks from chemically modified peptides, i.e.
from oxidized or deamidated species, are not represented in the *in silico* database and should be thus identified and removed from experimental data.

The results of pre-processing MS1 data are exemplarily illustrated in Fig. 3. As shown by data of the LC-MS1 test spectrum of *Enterococcus faecalis* DSM 20371, pre-processing preferentially removed low intensity peaks in the low MW region. The upper row of Fig. 3 illustrates MS1 peak density functions of raw (blue bars) and of pre-processed data (red bars) whereas the x-axes encode the log10-scaled intensity (left), or the MW value (right). The blue shaded area indicates the lower and upper bounds of the MW region utilized for identification analysis by MicrobeMS (see below). More data exemplifying strategies of spectral pre-processing are provided in the supporting information (see Figs. S2 and S3).

**MS1-based microbial identification:** The strategy of microorganism identification analysis is outlined in Fig. 4. Pre-processed MS1 peak list data are systematically compared with *in silico* peptide mass profiles derived from genome sequences using pattern matching based on Pearson's correlation coefficients. Coefficients are converted to scores and ranked in descending order. The results of identification analysis by means of score ranking lists are given in Tab. 2. This table summarizes information on the *a priori* taxonomic identity, i.e. the genus, species and strain assignments of the microorganisms tested (cf. column 2) and the sample preparation technique applied (SPEED for samples #1 - #13 and STrap for samples #14 - #21, see column 1). Furthermore, columns 3, 4 and 5 list the three top-ranked bacteria as determined by MicrobeMS: Column 3 details their genus, species and strain identities while columns 4 and 5 depict the corresponding proteome IDs and the correlation-derived score values, respectively. Test results summarized in Tab. 2 generally suggest a high level of identification accuracy and demonstrate that LC-MS1 data from microbial samples can be meaningfully and successfully queried against *in silico* peptide mass profiles. As an illustration, identification was accurate at the genus and species level in 20 out of 21 cases and in 5 cases there was even an agreement at the strain level. Furthermore, complete matching between the species identity of the test sample and three top scored
database entries has been determined in 9 out of 21 instances. In view of the facts that only some of the characterized strains are contained in the \textit{in silico} database and that not all microbial species are represented by three or more database entries, these first findings point out the large potential of MS\textsuperscript{1} data analysis for rapid microbial identification. Detailed analysis of the \textit{Burkholderia pseudomallei} group test subset which comprised 5 different strains of \textit{Burkholderia thailandensis} and one strain of \textit{Burkholderia oklahomensis} revealed further excellent test results (cf. samples \#8 - \#13): almost perfect accuracy was obtained from this sample set considering the facts that (i) \textit{Burkholderia pseudomallei} and \textit{Burkholderia mallei} are the phylogenetically closest relatives of the latter two species (34) and (ii) that the current version of the \textit{in silico} library contained only one \textit{in silico} spectrum of \textit{B. oklahomensis} and two profiles from strains of \textit{B. thailandensis}.

The complete results of identification tests of the \textit{Burkholderia pseudomallei} group subset consisting of altogether 24 LC-MS\textsuperscript{1} data sets from biological and technical replicates are given in the supporting information (Fig. S4).

A closer examination of test results of samples \#1 (\textit{Bacillus velezensis}), \#17 (\textit{Bacillus cereus}) and \#18 (\textit{Burkholderia cepacia}) showed that genus and species assignments of the two top scored bacteria were accurate whereas bacteria listed at position 3 were either close relatives thereof, or strains with unclear species assignment. For example, in case of sample \#1, \textit{Bacillus vallismortis} was found at position 3. \textit{B. vallismortis} is like \textit{B. velezensis} a member of the \textit{Bacillus subtilis} group (35). In the instances of samples \#17 and \#18 either a close relative of \textit{B. cereus} was ranked at position 3 - \textit{Bacillus thuringiensis} is a member of the \textit{B. cereus} group, cf. (36-38) - or \textit{Burkholderia reimsis}, better known as \textit{Burkholderia} sp. BE51 a bacterium with incomplete species description was listed at position 3, see (34, 39) for details.

Identification results of samples \#6 (\textit{Citrobacter freundii}) and \#21 (\textit{Escherichia coli}) reveal that two out of the three top scored bacteria were correctly identified, including both top hits. Inspection of data from sample \#21 suggested, however, that differentiation between \textit{E. coli} and strains of \textit{Shigella} is not
straightforward, an observation which is substantiated by the special taxonomy of *E. coli* and *Shigella* and additionally reflected by findings made by other techniques, such as MALDI-TOF MS (12, 40), LC-MS²-based approaches (18) or whole genome sequencing, WGS (41). In the instance of sample #21, we found 25 strain entries of *E. coli* and 5 entries from 3 different *Shigella* species (*S. sonnei*, 3 × *S. boydii* and *S. dysenteriae*) among the top 30 entries present in the score ranking list (data not shown). This and the finding of insignificant distances among scores in the ranking list suggest that the current version of the MS¹ analysis pipeline might not always provide the required reliability of identifying *E. coli* or *Shigella*. This observation is supported by taxonomy data. For example, it has been stated that *Shigella* strains can be viewed from a genetic perspective as subpopulations within *E. coli* (12, 41, 42) and some studies even recommend re-classification of *Shigella* and *E. coli*, e.g. (42, 43).

True misidentification was observed only in a single instance, see sample #14. In this example *Yersinia pseudotuberculosis* DSM 8992 has been identified as *Yersinia pestis* SCPM-O-DNA-17 (I-2457, top hit); strains of *Y. pseudotuberculosis* were, however, ranked at positions 2 and 3. As in the previous example, the analysis of the taxonomy is helpful to understand the particular test result. On a genomic level *Y. pestis* is known to be highly similar to the enteric pathogen *Y. pseudotuberculosis* (44, 45). In fact, *Y. pestis* can be considered a clone of *Y. pseudotuberculosis* which has evolved only recently (45, 46). Moreover, the top-scored strain of *Y. pestis* is a member of the subspecies *Y. pestis* ssp. *ulegeica* (47). Strains of *Y. pestis* ssp. *ulegeica* belong to a branch which is from a phylogenetic point of view more closely related to the ancestor *Y. pseudotuberculosis* than members from other branches of *Y. pestis*, including recent strains from *Y. pestis* ssp. *pestis* (44, 48).

**Influence of sample preparation on the accuracy of identification**: In this paper, two different methods of microbial sample preparation, STrap (26) and SPEED (27), were utilized. Although the size of the test sample set does not allow a statistically valid comparison, this proof of concept study provides some
preliminary indications with regard to the influence of sample preparation on the accuracy of pathogen identification.

As given by Tab. 2, the collected data involve 13 microbial samples processed by means of the SPEED method (samples #1-#13) while 8 samples were prepared using the STrap protocol (samples #14-#21). Two of the microbial samples, *Mycobacteroides abscessus* DSM 44196 and *Pseudomonas aeruginosa* ATCC 27853 were processed by either method, samples #4 and #7 (SPEED) and samples #16 and #19 (STrap).

Results from differently processed samples revealed that top score values of STrap samples varied between 110 and 160 whereas identification tests with SPEED processed samples led in 12 of 13 instances to scores > 170, among them 9 cases with scores > 200 (identical parameters of identification analysis function). Among samples prepared by the SPEED method, only one case with a score below 170 was observed (#7, *M. abscessus*, score of 144). It should be noted, however, that higher scores of SPEED processed samples do not necessarily indicate an improved accuracy of identification: Data from *M. abscessus* and *P. aeruginosa* which are both prepared by the STrap and by the SPEED protocol demonstrate in both instances accurate identification at the species level and for *M. abscessus* even at the strain level. Although SPEED sample processing resulted in somewhat higher score values - 144 (SPEED) vs. 129 (STrap) in case of *M. abscessus* and 198 vs. 115 in *P. aeruginosa* - the relative distance to the first false hit varied only slightly.

Therefore, it can be stated that higher score values were determined for SPEED samples, whereby identification accuracy is thought to benefit only slightly from this advantage. However, it should be pointed out that the data set used is rather small, so a more comprehensive test set would be required to support further conclusions.
8. DISCUSSION
In this proof-of-concept study, we evaluated the principal applicability of shotgun LC - mass spectrometry (LC-MS\(^1\)) for microbial identification and explored the taxonomic resolution of the proposed method. To this end, a database of strain-specific \textit{in silico} peptide MW profiles was constructed from UniProtKB resources and queried by experimental LC-MS\(^1\) test spectra obtained from microorganisms grown in pure cultures. The results of these queries are summarized in score ranking lists which were helpful to obtain insights into the taxonomic identity of the bacteria studied. It can be stated that the suggested approach is generally (38 of 39 cases) suitable for identifying bacteria at the genus and species level, and sometimes even at strain level. Despite these encouraging results, it should be also noted that in a single instance an ambiguous identification result was found. Targeted screening for combinations of taxon-specific peptide features (feature selection) and a better quality of the underlying protein sequence database, particularly of UniProtKB/TrEMBL (database curation), are suggested as potential starting points to further improve the accuracy of the proposed workflow.

\textit{Computational considerations:} With a size of the final \textit{in silico} database of 1.44 GB, disk space requirements were only moderate. At the other hand, the computational time required for compilation from UniProtKB resource data was substantial: Approximately 30 h were needed using a standard PC workstation. It should be noted, however, that the \textit{in silico} database only needs to be constructed at longer intervals, for example every two months, so that the high computational requirements and long database construction times would become less important in a routine setup.

Similarly, determination of score values based on Pareto-scaled correlation coefficients between pre-processed MS\(^1\) test spectra and \textit{in silico} peptide MW profiles are computationally intensive tasks and require powerful hardware. The Matlab implementation therefore involved parallel programming with fully vectorized code and multicore support, which helped reducing analyses time down to less than 2 minutes per LC-MS\(^1\) test spectrum (Dell Precision T7500 workstation). Precondition for short...
computation times is a sufficient amount of installed memory; approx. 96 GB RAM is considered the minimum requirement. Furthermore, within the scope of the present study the MicrobeMS toolbox has been ported from the original Windows 7 64-bit Matlab version to Linux (Debian 8.11) and tested on one of the RKI’s bioinformatics server (8× Intel Xeon E7-4890 v2@2.8 GHz, 120 cores, 1 TB RAM). This led to shorter analysis times (~40 s per experimental spectrum) whereby tests utilizing more than 12 cores (15, 20, 40 and 60 cores) revealed only minor reductions of computational times (data not shown). Obviously identification analysis did not significantly benefit from utilization of more than 12 cores.

**Comparison with other MS methods for microbial identification**: Basic aspects of the workflow of the proposed method are schematically illustrated in Figs. 1 and 4. It is to be noted that the MS\(^1\)-based analytical method resembles in many aspects the well-known MALDI-TOF MS identification technique. However, despite the many similarities with the MALDI-TOF MS approach, such as the need to cultivate, or utilization of spectral databases and score ranking lists, there is one important difference: The proposed method does not require spectral libraries that have to be compiled by collecting mass spectra from cultivated pathogens of known taxonomic identity. The *in silico* database used in this study has been ultimately computer generated from genome data which results in an impressive number of bacterial strains - more than 12,000 - contained already in its first version. Of note that this number exceeds the number of strains and species entries present in the current version of Bruker’s MALDI Biotyper (MBT) CE *in vitro* diagnostics (CE-IVD) database. The MBT CE-IVD database has been developed over more than a decade and currently contains in version V.9 (2019) altogether 8326 MSP entries from bacteria and yeasts (49). This fact clearly illustrates that the suggested LC-MS\(^1\) method directly benefits from the ever-increasing growth of genomic data: Database coverage of both, newly discovered and already known microbiological taxa will improve automatically with each update of the UniProt Knowledge Base without the need for additional measurements.
The significantly higher number of signals present in LC-MS data constitutes another important difference to MALDI-TOF MS. While MALDI-TOF MS usually detects limited numbers (< 150) of predominantly high-abundance proteins with housekeeping functions, such as basic ribosomal proteins, or nucleic acid-binding proteins (50-53), usually in the m/z region between 2 and 20 kDa, LC-MS shotgun proteomics usually detects more than 50,000 peptides from a single microbial sample. However, further studies are required to clarify whether the enormous increase of information contained in the LC-MS\(^1\) data will indeed lead to a better taxonomic resolution.

Current drawbacks of the LC-MS method are, above all, higher instrument costs and a relatively low dissemination of LC-MS equipment and analysis concepts in clinical or food microbiology laboratories. However, it is to be expected that future technological developments will help to reduce experimental efforts and it is anticipated that dedicated LC-MS\(^1\) systems could be helpful to reduce costs and thus improve dissemination of LC-MS technology.

Compared to published LC-MS\(^2\) methods to identify and classify microbial pathogens, the proposed approach offers a number of important advantages. Firstly, collection of high-quality LC-MS\(^1\) peptide data can be done with considerable less time efforts compared to LC-MS\(^2\) proteomics measurements. In the present study, however, the experimental data originate from 120 or 160 min gradient LC-MS\(^2\) measurements carried out within the context of other project objectives. However, it is known from a large number of published studies and from own experience that LC-MS\(^1\) measurements of microbial samples can be performed within 10 minutes and provide sufficiently large numbers of peptide signals. This fact and the possibility to fully automate sample preparation by the SPEED sample preparation protocol points towards a better sample throughput of the LC-MS\(^1\) method compared to LC-MS\(^2\) approaches and could even rival with the speed of MALDI biotyping.

Secondly, the computational requirements are significantly reduced due to the fact that identification of peptides and/or proteins is not necessary. Under our conditions computational time requirements were
In silico peptide database for LC-MS1 microbial biotyping

negligible (less than 2 minutes). Thirdly, the method does not rely on specifically identifying, i.e. discriminating (“unique”) peptides. This is important because the number of such peptides tends to diminish with the ever-growing number of peptides contained in future database versions. Furthermore, the simplicity of the proposed microbial identification method allows defining scores in a straightforward manner and to adapt a well-established and even FDA-approved principle in the field of MS-based microbial diagnostics (score ranking lists). An in-depth literature search and internal discussions led us to the conclusion that score definition from LC-MS² data by using unique peptides constitutes a rather complex problem for which no universally accepted solution has yet been presented.

Major disadvantages compared to LC-MS²-based identification are the need to work with microbial cells from pure cultures due to the higher requirements of sample purity. Although polymicrobial samples were not tested by us, it is reasonable to assume that the presented spectra correlation analysis approach is not applicable to the study of other than pure samples. For the same reasons as MALDI-TOF MS, the proposed concept of MS¹ data analysis does not allow reliable identification of individual microorganisms from complex microbial mixtures, such as faeces, or natural biofilm habitats. In addition, the method does not offer a direct way for detecting database errors, such as missing entries or only incomplete database coverage. While the absence of unique peptides in MS²-based identification analyses would indicate database gaps, MS¹ score ranking lists do not necessarily provide direct evidences for database errors or poor (incomplete) coverage. However, such gaps are likely to be rare events, due to the diversity and large numbers of microbial strains contained in the present and future in silico database versions.

Another important advancement of the suggested LC-MS¹ workflow to bacterial differentiation consists in the possibility for combination with existing LC-MS²-based analysis concepts. Such a hybrid approach would involve LC-MS² measurements of microbial extracts and separate analysis of MS¹ and MS² data. Straightforward correlation analysis between LC-MS¹ test spectra and in silico peptide mass profiles at
the one hand and peptide identification analyses by the help of $\text{MS}^2$ data, ideally from order, family, or genus specific sequence databases at the other hand, could be helpful to reduce analysis times and to improve the accuracy of microbial identification as a whole.

**CONCLUSIONS**

This proof-of-concept study has demonstrated that identification analysis from LC-MS$^1$ data represents a powerful technology that could drive improvements in bacterial identification. The technique utilizes *in silico* libraries generated from publicly available proteome resources and does not require databases of experimental mass spectra. The proposed pipeline is easy to use, computationally efficient and freely available for both Linux and Windows operating systems. The taxonomic resolution of the method is promising, but improvements, such as well-curated databases, application of feature selection methods, better quality checks as well as rigorously conducted tests with large LC-MS$^1$ data sets are needed to answer the question whether the suggested approach can be employed in clinical microbiology in a reliable, effective and useful manner.

**9. ACKNOWLEDGEMENTS**

The authors are thankful to Max Weydmann for conducting LC-MS measurements of the microbial samples prepared by the STrap method.

**10. DATA AVAILABILITY**

The mass spectrometry proteomics data were published under the Creative Commons Attribution Non Commercial 4.0 International (CC-BY-NC) license at Zenodo, an open-access repository developed under the European OpenAIRE program and operated by CERN, (https://zenodo.org/record/3573994) with the digital object identifier (doi) 10.5281/zenodo.3573994 (https://doi.org/10.5281/zenodo.3573994). The *in*
silico database for identification of microorganisms by LC-MS1 data has been also deposited at Zenodo and is available freely under the CC-BY-NC license from https://zenodo.org/record/3573996 (doi: 10.5281/zenodo.3573996).

AUTHOR CONTRIBUTION STATEMENT
PL and JD contributed conception and design of the study; AS and JD collected the data and performed the experiments and first steps of data pre-processing; PL wrote the code of the Matlab toolboxes parseuniprot and MicrobeMS and performed the data processing; PL wrote the first draft of the manuscript; AS, CB and JD wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the final version of the manuscript.

CONFLICT OF INTEREST STATEMENT
AS, JD and PL are the inventors of SPEED and have submitted patent applications related to SPEED.
11. REFERENCES

1. Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P. E., Rolain, J. M., and Raoult, D. (2009) Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry. *Clin. Infect. Dis.* 49, 543-551
2. Nomura, F. (2015) Proteome-based bacterial identification using matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS): A revolutionary shift in clinical diagnostic microbiology. *Biochim. Biophys. Acta* 1854, 528-537
3. Schubert, S., and Kostrzewa, M. (2017) MALDI-TOF MS in the Microbiology Laboratory: Current Trends. *Curr. Issues Mol. Biol.* 23, 17-20
4. Welker, M., Van Belkum, A., Girard, V., Charrier, J. P., and Pincus, D. (2019) An update on the routine application of MALDI-TOF MS in clinical microbiology. *Expert Rev Proteomics* 16, 695-710
5. Sandrin, S. S., and Kostrzewa, M. (2017) MALDI TOF MS profiling of bacteria at the strain level: a review. *Mass Spectrom. Rev.* 32, 188-217
6. Demirev, P., Sandrin, T. R., and (eds) (2016) Applications of Mass Spectrometry in Microbiology - From Strain Characterization to Rapid Screening for Antibiotic Resistance *Springer International Publishing Switzerland* 2016
7. Sauget, M., Valot, B., Bertrand, X., and Hocquet, D. (2017) Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria? *Trends Microbiol.* 25, 447-455
8. Rodrigues, C., Novais, A., Sousa, C., Ramos, H., Coque, T. M., Canton, R., Lopes, J. A., and Peixe, L. (2016) Elucidating constraints for differentiation of major human Klebsiella pneumonias using MALDI-TOF MS. *Eur. J. Clin. Microbiol. Infect. Dis.*
9. Sousa, C., Botelho, J., Grosso, F., Silva, L., Lopes, J., and Peixe, L. (2015) Unsuitability of MALDI-TOF MS to discriminate Acinetobacter baumannii clones under routine experimental conditions. *Front. Microbiol.* 6, 481
10. Lasch, P., Fleige, C., Stammler, M., Layer, F., Nubel, U., Witte, W., and Werner, G. (2014) Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of Enterococcus faecium and Staphylococcus aureus isolates. *J. Microbiol. Methods* 100, 58-69
11. Grenga, L., Pible, O., and Armengaud, J. (2019) Pathogen proteotyping: A rapidly developing application of mass spectrometry to address clinical concerns. *Clinical Mass Spectrometry*
12. Pauw, A., Jonker, D., Roeselers, G., Heng, J. M., Mars-Groenendijk, R. H., Trip, H., Molhoek, E. M., Jansen, H. J., van der Plas, J., de Jong, A. L., Majchrzakiewicz-Koehorst, J. A., and Speksnijder, A. G. (2015) Rapid and reliable discrimination between Shigella species and Escherichia coli using MALDI-TOF mass spectrometry. *Int. J. Med. Microbiol.*
13. He, Y., Li, H., Lu, X., Stratton, C. W., and Tang, Y. W. (2010) Mass spectrometry biotyper system identifies enteric bacterial pathogens directly from colonies grown on selective stool culture media. *J. Clin. Microbiol.* 48, 3888-3892
14. Dybwad, M., van der Laaken, A. L., Blatny, J. M., and Pauw, A. (2013) Rapid Identification of Bacillus anthracis Spores in Suspicious Powder Samples by Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS). *Appl. Environ. Microbiol.* 79, 5372-5383
15. Lasch, P., Wahab, T., Weil, S., Palyi, B., Tomasso, H., Zange, S., Kiland Granerud, B., Drevinek, M., Kokotovic, B., Wittwer, M., Pfuger, V., Di Caro, A., Stammler, M., Grunow, R., and Jacob, D. (2015) Identification of Highly Pathogenic Microorganisms by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry: Results of an Interlaboratory Ring Trial. *J. Clin. Microbiol.* 53, 2632-2640
16. Jabbour, R. E., Deshpande, S. V., Wade, M. M., Stanford, M. F., Wick, C. H., Zulich, A. W., Skowronska, E. W., and Snyder, A. P. (2010) Double-blind characterization of non-genome-sequenced bacteria by mass spectrometry-based proteomics. *Appl. Environ. Microbiol.* 76, 3637-3644
17. Jabbour, R. E., Deshpande, S. V., Stanford, M. F., Wick, C. H., Zulich, A. W., and Snyder, A. P. (2011) A protein processing filter method for bacterial identification by mass spectrometry-based proteomics. *J. Proteome Res.* 10, 907-912
18. Berendsen, E. M., Levin, E., Braakman, R., von der Riet-van Oeveren, D., Sedee, N. J. A., and Pauw, A. (2017) Identification of microorganisms grown in blood culture flasks using liquid chromatography–tandem mass spectrometry. *Future Microbiol.* 0, null
19. Tracz, D. M., McCorrister, S. J., Chong, P. M., Lee, D. M., Corbett, C. R., and Westmacott, G. R. (2013) A simple shotgun proteomics method for rapid bacterial identification. *J. Microbiol. Methods* 94, 54-57

20. Berendsen, E. M., Levin, E., Braakman, R., Prodan, A., van Leeuwen, H. C., and Pauw, A. (2020) Untargeted accurate identification of highly pathogenic bacteria directly from blood culture flasks. *Int. J. Med. Microbiol.* 310, 151376

21. Dworzanski, J. P., Deshpande, S. V., Chen, R., Jabbour, R. E., Snyder, A. P., Wick, C. H., and Li, L. (2006) Mass spectrometry-based proteomics combined with bioinformatic tools for bacterial classification. *J. Proteome Res.* 5, 76-87

22. Dworzanski, J. P., Dickinson, D. N., Deshpande, S. V., Snyder, A. P., and Eckenrode, B. A. (2010) Discrimination and phylogenomic classification of Bacillus anthracis-cereus-thuringiensis strains based on LC-MS/MS analysis of whole cell protein digests. *Anal. Chem.* 82, 145-155

23. Boulund, F., Karlsson, R., Gonzales-Siles, L., Johnning, A., Karami, N., Al-Bayati, O., Ahren, C., Moore, E. R. B., and Kristiansson, E. (2017) Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics. *Mol. Cell. Proteomics* 16, 1052-1063

24. Lasch, P., Stammier, M., Zhang, M., Baranska, M., Bosch, A., and Majzner, K. (2018) FT-IR Hyperspectral Imaging and Artificial Neural Network Analysis for Identification of Pathogenic Bacteria. *Anal. Chem.* 90, 8896-8904

25. Wisniewski, J. R., and Gaugaz, F. Z. (2015) Fast and sensitive total protein and Peptide assays for proteomic analysis. *Anal. Chem.* 87, 4110-4116

26. Zougman, A., Selby, P. J., and Banks, R. E. (2014) Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* 14, 1006-1000

27. Doellinger, J., Schneider, A., Hoeller, M., and Lasch, P. (2020) Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. *Mol. Cell. Proteomics* 19, 209-222

28. Ishihama, Y., Rappsilber, J., and Mann, M. (2006) Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics. *J. Proteome Res.* 5, 988-994

29. Lasch, P. (2020) MicrobeMS: A Matlab Toolbox for Analysis of Microbial MALDI-TOF Mass Spectra. [http://www.microbe-ms.com](http://www.microbe-ms.com), last accessed Jul 01, 2020

30. Lasch, P. (2020) MicrobeMS Wiki: A Matlab Toolbox for Analysis of Microbial MALDI-TOF Mass Spectra. [http://wiki.microbe-ms.com/Main_Page](http://wiki.microbe-ms.com/Main_Page), last accessed Jul 01, 2020

31. UniProtKB (2020) What is a redundant proteome? [https://www.uniprot.org/help/proteome_redundancy_faq](https://www.uniprot.org/help/proteome_redundancy_faq), last accessed Mar 01, 2020

32. Lasch, P. (2012) Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* 117, 100-114

33. Tsai, T. H., Wang, M., and Ressom, H. W. (2016) Preprocessing and Analysis of LC-MS-Based Proteomic Data. *Methods Mol. Biol.* 1362, 63-76

34. Vandamme, P., and Eberl, L. (2018) Burkholderia. *Bergey's Manual of Systematics of Archaea and Bacteria*

35. Fan, B., Blom, J., Klenk, H. P., and Borri, R. (2017) Bacillus amyloliquefaciens, Bacillus velezensis, and Bacillus siamensis Form an "Operational Group B. amyloliquefaciens" within the B. subtilis Species Complex. *Front. Microbiol.* 8, 22

36. Helgason, E., Okstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I., and Kolsto, A. B. (2000) Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis--one species on the basis of genetic evidence. *Appl. Environ. Microbiol.* 66, 2627-2630

37. Carroll, L. M., Wiedmann, M., and Kovac, J. (2020) Proposal of a Taxonomic Nomenclature for the Bacillus cereus Group Which Reconciles Genomic Definitions of Bacterial Species with Clinical and Industrial Phenotypes. *mBio* 11

38. Ehling-Schulz, M., Lereclus, D., and Koehler, T. M. (2019) The Bacillus cereus Group: Bacillus Species with Pathogenic Potential. *Microbiol Spectr* 7

39. Esmaeel, Q., Issa, A., Sanchez, L., Clement, C., Jacquard, C., and Barka, E. A. (2018) Draft Genome Sequence of Burkholderia reimis BE51, a Plant-Associated Bacterium Isolated from Agricultural Rhizosphere. *Microbiol Resour Announc* 7

40. Khot, P. D., and Fisher, M. A. (2013) Novel approach for differentiating Shigella species and Escherichia coli by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.* 51, 3711-3716
41. Chattaway, M. A., Schaefer, U., Tewolde, R., Dallman, T. J., and Jenkins, C. (2017) Identification of Escherichia coli and Shigella Species from Whole-Genome Sequences. J. Clin. Microbiol. 55, 616-623
42. Pupo, G. M., Lan, R., and Reeves, P. R. (2000) Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. U. S. A. 97, 10567-10572
43. Alves, G., Wang, G., Ogurtsov, A. Y., Drake, S. K., Gucek, M., Sacks, D. B., and Yu, Y. K. (2018) Rapid Classification and Identification of Multiple Microorganisms with Accurate Statistical Significance via High-Resolution Tandem Mass Spectrometry. J. Am. Soc. Mass Spectrom. 29, 1721-1737
44. Demeure, C. E., Dussurget, O., Mas Fiol, G., Le Guern, A. S., Savin, C., and Pizarro-Cerda, J. (2019) Yersinia pestis and plague: an updated view on evolution, virulence determinants, immune subversion, vaccination, and diagnostics. Genes Immun. 20, 357-370
45. Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999) Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. U. S. A. 96, 14043-14048
46. Califf, K. J., Keim, P. S., Wagner, D. M., and Sahl, J. W. (2015) Redefining the differences in gene content between Yersinia pestis and Yersinia pseudotuberculosis using large-scale comparative genomics. Microb Genom 1, e000028
47. Kislichkina, A. A., Bogun, A. G., Kadnikova, L. A., Maiskaya, N. V., Solomentsev, V. I., Sizova, A. A., Dentovskaya, S. V., Balakhonov, S. V., and Anisimov, A. P. (2018) Six Whole-Genome Assemblies of Yersinia pestis subsp. microtus bv. ulegeica (Phylogroup 0.PE5) Strains Isolated from Mongolian Natural Plague Foci. Genome Announc 6
48. Kutyrev, V. V., Eroshenko, G. A., Motin, V. L., Nosov, N. Y., Krasnov, J. M., Kukleva, L. M., Nikiforov, K. A., Al’khova, Z. V., Oglodin, E. G., and Guseva, N. P. (2018) Phylogeny and Classification of Yersinia pestis Through the Lens of Strains From the Plague Foci of Commonwealth of Independent States. Front. Microbiol. 9, 1106
49. Dewaele, K., Lagrou, K., Frans, J., Hayette, M. P., and Vernelen, K. (2019) Hospital Laboratory Survey for Identification of Candida auris in Belgium. J Fungi (Basel) 5
50. Ryzhov, V., and Fenselau, C. (2001) Characterization of the protein subset desorbed by MALDI from whole bacterial cells. Anal. Chem. 73, 746-750
51. Pineda, F. J., Antoine, M. D., Demirev, P. A., Feldman, A. B., Jackman, J., Longenecker, M., and Lin, J. S. (2003) Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers. Anal. Chem. 75, 3817-3822
52. Dieckmann, R., Helmuth, R., Erhard, M., and Malorny, B. (2008) Rapid classification and identification of salmonellae at the species and subspecies levels by whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl. Environ. Microbiol. 74, 7767-7778
53. Lasch, P., Jacob, D., Grunow, R., Schwecke, T., and Doellinger, J. (2016) Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF MS for the identification of highly pathogenic bacteria. Trac-Trend Anal Chem 85, Part B, 103–111

12. FOOTNOTES
### 13. TABLES

#### Table 1.

| No | Genus, Species               | Strain name(s)        |
|----|------------------------------|-----------------------|
| 1  | *Acinetobacter baumannii*    | DSM 30007             |
| 2  | *Bacillus cereus*            | DSM 31                |
| 3  | *Bacillus velezensis*        | DSM 23117 / FZB42     |
| 4  | *Burkholderia cepacia*       | ATCC 25416            |
| 5  | *Burkholderia thailandensis*| DSM 13276             |
| 6  | *Burkholderia thailandensis*| E125*                 |
| 7  | *Burkholderia thailandensis*| E131*                 |
| 8  | *Burkholderia thailandensis*| E153                  |
| 9  | *Burkholderia thailandensis*| LMG 20219             |
| 10 | *Burkholderia oklahomensis*  | DSM 21774             |
| 11 | *Citrobacter freundii*       | DSM 30039             |
| 12 | *Enterococcus faecalis*      | DSM 20371             |
| 13 | *Escherichia coli*           | DSM 3871 (K12 W3110 derivative) |
| 14 | *Mycobacteroides abscessus*  | DSM 44196             |
| 15 | *Pseudomonas aeruginosa*     | ATCC 27853            |
| 16 | *Staphylococcus epidermidis* | DSM 1798             |
| 17 | *Staphylococcus aureus*      | DSM 20231 / NCTC 8532 |
| 18 | *Vibrio cholerae*            | NIH 41                |
| 19 | *Yersinia pseudotuberculosis*| DSM 8992             |

Overview of the microbial species and strains used in this study. Abbreviations: DSM – Deutsche Sammlung von Mikroorganismen; ATCC – American Type Culture Collection; NCTC – National Collection of Type Cultures; LMG – Belgian Coordinated Collections of Microorganisms, Universiteit Gent – Laboratorium voor Microbiologie, NIH – National Institute of Health, * (24)
### Table 2.

| Sample no. / Protocol | Genus, species and strain identity | Top database hits | Proteome ID | Scores |
|-----------------------|-----------------------------------|-------------------|-------------|--------|
| #1 SPEED              | Bacillus velezensis DSM 23117 / FZB42 | 1. B. velezensis ATR2 | UP000228760 | 172.1654 |
|                       |                                   | 2. B. velezensis DSM 23117 / FZB42 | UP0000011120 | 165.3249 |
|                       |                                   | 3. B. vallismortis NBF-001 | UP000193819 | 164.1915 |
| #2 SPEED              | Staphylococcus epidermidis DSM 1798 | 1. S. epidermidis DSM 1798 / ATCC 12228 | UP000001411 | 260.842 |
|                       |                                   | 2. S. epidermidis M23864::W2(grey) | UP000004733 | 254.5784 |
|                       |                                   | 3. S. epidermidis ATCC 35984 / RP62A | UP000005531 | 250.1004 |
| #3 SPEED              | Staphylococcus aureus DSM 20231 / NCTC 8532 | 1. S. aureus Newman | UP000006386 | 213.046 |
|                       |                                   | 2. S. aureus USA300 | UP000001939 | 212.0801 |
|                       |                                   | 3. S. aureus NCTC 8325 | UP000008816 | 211.2254 |
| #4 SPEED              | Pseudomonas aeruginosa ATCC 27853 | 1. P. aeruginosa LIM1410 | UP000247072 | 197.6311 |
|                       |                                   | 2. P. aeruginosa LIM1680 | UP000246728 | 195.8522 |
|                       |                                   | 3. P. aeruginosa MH19 | UP0000043988 | 188.7818 |
| #5 SPEED              | Enterococcus faecalis DSM 20371 | 1. E. faecalis CH116 | UP000013839 | 203.0399 |
|                       |                                   | 2. E. faecalis RMC1 | UP000013621 | 202.8773 |
|                       |                                   | 3. E. faecalis EnGen0426 | UP000032191 | 202.6593 |
| #6 SPEED              | Citrobacter freundii DSM 30039 | 1. C. freundii GED7749C | UP000070673 | 178.2909 |
|                       |                                   | 2. E. coli ISC11 | UP000019194 | 172.3027 |
|                       |                                   | 3. C. freundii MRSN 12115 | UP000032554 | 169.3774 |
| #7 SPEED              | Mycobacteroides abscessus DSM 44196 | 1. M. abscessus ATCC 19977 / DSM 44196 | UP000007137 | 144.387 |
|                       |                                   | 2. M. abscessus subsp. abscessus 1135 | UP000184667 | 131.1567 |
|                       |                                   | 3. M. abscessus subsp. bolletii 909 | UP000184881 | 128.8442 |
| #8 SPEED              | Burkholderia oklahomensis DSM 21774 | 1. B. oklahomensis NCTC 13388 | UP000254484 | 217.199 |
|                       |                                   | 2. B. pseudomallei MSHR346 | UP000002031 | 149.3277 |
|                       |                                   | 3. B. pseudomallei 1710b | UP000002700 | 149.2699 |
| #9 SPEED              | Burkholderia thailandensis E153 | 1. B. thailandensis DSM 13276 / LMG 20219 | UP000001930 | 214.043 |
|                       |                                   | 2. B. thailandensis FDAARGOS_243 | UP000235972 | 204.8531 |
|                       |                                   | 3. B. pseudomallei 1710b | UP000002700 | 164.4795 |
| #10 SPEED             | Burkholderia thailandensis E131 | 1. B. thailandensis DSM 13276 / LMG 20219 | UP000001930 | 204.9895 |
|                       |                                   | 2. B. thailandensis FDAARGOS_243 | UP000235972 | 196.8191 |
|                       |                                   | 3. B. pseudomallei 1710b | UP000002700 | 159.7089 |
| #11 SPEED             | Burkholderia thailandensis E125 | 1. B. thailandensis DSM 13276 / LMG 20219 | UP000001930 | 202.7592 |
|                       |                                   | 2. B. thailandensis FDAARGOS_243 | UP000235972 | 195.6465 |
|                       |                                   | 3. B. pseudomallei 1710b | UP000002700 | 161.2297 |
| #12 SPEED             | Burkholderia thailandensis LMG 20219 | 1. B. thailandensis DSM 13276 / LMG 20219 | UP000001930 | 203.342 |
|                       |                                   | 2. B. thailandensis FDAARGOS_243 | UP000235972 | 192.4561 |
|                       |                                   | 3. B. pseudomallei 668 | UP000002700 | 161.0222 |
| #13 SPEED             | Burkholderia thailandensis DSM 13276 | 1. B. thailandensis DSM 13276 / LMG 20219 | UP000001930 | 204.233 |
|                       |                                   | 2. B. thailandensis FDAARGOS_243 | UP000235972 | 194.5373 |
|                       |                                   | 3. B. pseudomallei 668 | UP000002153 | 159.5722 |
| #14 STrap             | Yersinia pseudotuberculosis DSM 8992 | 1. Y. pestis SCPM-O-DNA-17 (I-2457) | UP000239693 | 157.8063 |
|                       |                                   | 2. Y. pseudotuberculosis IP 31758 | UP000002412 | 140.4254 |
|                       |                                   | 3. Y. pseudotuberculosis IP32953 | UP000001011 | 140.2100 |
Identification results of a microbial LC-MS\textsuperscript{1} test data set comprising 19 different strains from 11 genera and 15 species. Samples #1-#13 were prepared by the SPEED proteomic sample preparation protocol developed at RKI (27). The STrap (26) method was utilized for preparing proteomic samples from strains listed at lines 14-21. Of note: two strains, *P. aeruginosa* ATCC 27853 and *M. abscessus* DSM 44196 were prepared and measured twice. Note that samples of the *Burkholderia* test set with closely related microbial strains (#8 - #13) were measured in biological and technical replicates, see supporting information for details.

* refers to a not validly described species (39) (see text for details).
Figure 1.
Overview of the proposed LC-MS\textsuperscript{1} based microbial identification workflow. Pure microbial cultures are prepared and colony material is processed using established sample preparation protocols for shotgun proteomics. Mass spectrometry data are then obtained using LC-MS. MS\textsuperscript{1} data are extracted and pre-processed for subsequent comparison against a library of \textit{in silico} mass profiles obtained from UniProtKB/Swiss-Prot and UniProtKB/TrEMBL protein sequence data. This library is composed of MW pattern, or profiles, each representing a characteristic strain-specific combination of peptide masses whereby peptides may be specific or non-specific in a MS\textsuperscript{2} context. A ranking list of correlation, or inter-spectral distance values (i.e. of scores) is established which provides information on the taxonomic identity of the organism studied.
Figure 2.

Schematic workflow for generating an *in silico* database from UniProtKB/Swiss-Prot and/or UniProtKB/TrEMBL protein sequence data. The Matlab toolbox *parseuniprot* represents a proteomic pipeline in which three main internal functions, *readdat*, *resort* and *modfeat* are consecutively executed. The function *readdat* converts the content from structured text files available from ftp://ftp.uniprot.org into Matlab structure arrays that contain the complete information required to compile the *in silico* databases. Such arrays are subsequently processed by the functions *resort* and *modfeat*; the output of the *parseuniprot* pipeline is a collection of strain-specific *in silico* peptide mass profiles suitable for computer-based comparison (pattern matching) with experimental LC-MS\(^1\) test spectra.
Figure 3.

Pre-processing and feature selection of LC-MS$^1$ data. MS$^1$ peak data were acquired from a culture of Enterococcus faecalis DSM 20371; sample preparation has been carried according to the SPEED sample preparation protocol. (27)

**Top row:** histogram bar chart of log$_{10}$ scaled MS$^1$ peak intensities (left) and the molecular weight (MW) distribution (right) of peaks after feature detection by the Minora algorithm (=original* data, blue bars) and after pre-processing and feature selection by readlcmstxtfile (processed data, red bars).

Total number of peaks in original / processed MS$^1$ data: 82843 / 42559

Number of oxidized / deamidated peptides found and removed: 389 / 329

**Lower row:** ratio between the number of peaks present in processed and in original MS$^1$ data as a function of peak intensity (log$_{10}$ scaled, left), or of the MW (right).

Pre-processing was carried out by readlcmstxtfile, a Matlab function developed in house. This function has been designed to preferentially remove low intensity signals in the low MW region (< 2000 Da). The blue shaded area between 2000 – 5500 Da indicates the MW range used for correlation analysis by MicrobeMS.
Figure 4.

Data analysis workflow for microbial identification based on experimental LC-MS data and *in silico* databases comprising strain-specific peptide mass profiles derived from microbial genomes.