Optimization of a Top-Down Proteomics Platform for Closely Related Pathogenic Bacterial Discrimination

Mathieu Dupré, Magalie Duchateau, Christian Malosse, Diogo Borges-Lima, Valeria Calvaresi, Isabelle Podglajen, Dominique Clermont, Martial Jean-Pierre Rey, Julia Chamot-Rooke

Journal of proteome research

pubs.acs.org/jpr

Article

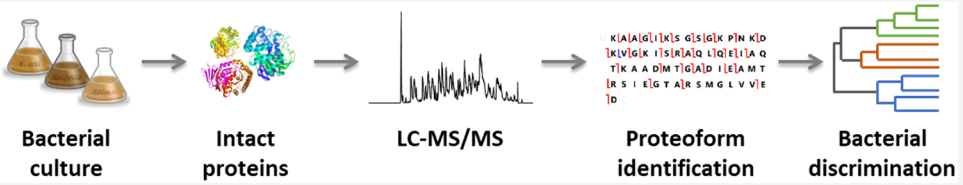
Optimization of a Top-Down Proteomics Platform for Closely Related Pathogenic Bacterial Discrimination

Mathieu Dupré, Magalie Duchateau, Christian Malosse, Diogo Borges-Lima, Valeria Calvaresi, Isabelle Podglajen, Dominique Clermont, Martial Rey, and Julia Chamot-Rooke*

Cite This: https://dx.doi.org/10.1021/acs.jproteome.0c00351

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The current technique used for microbial identification in hospitals is matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). However, it suffers from important limitations, in particular, for closely related species or when the database used for the identification lacks the appropriate reference. In this work, we set up a liquid chromatography (LC)−MS/MS top-down proteomics platform, which aims at discriminating closely related pathogenic bacteria through the identification of specific proteoforms. Using *Escherichia coli* as a model, all steps of the workflow were optimized: protein extraction, on-line LC separation, MS method, and data analysis. Using optimized parameters, about 220 proteins, corresponding to more than 500 proteoforms, could be identified in a single run. We then used this platform for the discrimination of enterobacterial pathogens undistinguishable by MALDI-TOF, although leading to very different clinical outcomes. For each pathogen, we identified specific proteoforms that could potentially be used as biomarkers. We also improved the characterization of poorly described bacterial strains. Our results highlight the advantage of addressing proteoforms rather than peptides for accurate bacterial characterization and qualify top-down proteomics as a promising tool in clinical microbiology. Data are available *via* ProteomeXchange with the identifier PXD019247.

KEYWORDS: *top-down proteomics, proteoforms, enterobacteria, pathogen, characterization, discrimination*

INTRODUCTION

In the last decade, the development of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) for rapid microbial identification has revolutionized the field of clinical microbiology.[1,2] By comparing the spectral profile obtained from the ionization of intact proteins from a bacterial colony to reference spectra, species identification can be achieved.[3] This approach is now used in many hospitals for routine identification of bacterial pathogens, as it is faster, more accurate, and less expensive than conventional phenotypic or genotypic methods. However, MALDI-TOF MS suffers from significant limitations. Some bacteria remain difficult to be identified, either because they do not give a specific profile or because the database lacks the appropriate reference.[4] In addition, the discriminatory power of the technique is often insufficient for reliably differentiating subspecies within species or clones within subspecies.[5] Therefore, there is an unmet diagnostic need for innovative analytical approaches allowing efficient and more accurate bacterial identification. Only subtle modifications, such as single amino acid change, are expected to differentiate proteins in closely related species. Therefore, bottom-up proteomics (BUP), which is based on the analysis of peptides generated upon protein enzymatic digestion, cannot provide the required level of accuracy to achieve this goal.[6−8]

To overcome these problems, an attractive alternative is to focus on intact proteins using top-down proteomics (TDP). In TDP, intact proteins are directly separated by liquid chromatography (LC) and sequenced by high-resolution MS.[9−11] This approach eliminates the additional complexity and uncertainty brought by the enzymatic digestion, which leads to loss of protein information. The major advantage of TDP is its ability to address protein variations and characterize all forms of a protein (proteoforms) arising from alternative splicing, allelic variation, or post-translational modification (PTM).[12,13] As in MALDI-TOF MS, intact proteins are

Received: May 25, 2020
Published: September 15, 2020

analyzed, but here, the MS/MS spectra bring an additional layer of information.

Few studies have already shown the added value of TDP for the analysis of microbial proteomes.[14−17] For instance, Ansong *et al.* clearly demonstrated that measuring bacterial proteomes at the intact protein level can bring crucial insights into biological mechanisms that would be impossible to obtain with the bottom-up technology.[18] In their study, they used a single-dimension LC−MS/MS TDP approach on the Gram-negative bacterial pathogen *Salmonella enterica enterica* Typhimurium and identified 563 proteins and 1,665 proteoforms. The authors also reported the differential utilization of bacterial protein S-thiolation in response to infection-like conditions. This study shows that a very high level of details can be obtained when bacterial strains are analyzed by TDP. Other bacterial proteomes that have been studied using TDP include *Escherichia coli*,[19] *Shewanella oneidensis*,[20] *Pseudomonas aeruginosa*,[21] *Novosphingobium aromaticivorans*,[22] and *Enterobacter sakazakii*.[23] As pointed out in these papers, TDP is of high interest for the characterization of PTMs and can also provide evidence for incorrect or missing protein annotations in protein sequence databases. Interestingly, using top-down LC−MS/MS, the differentiation between *Salmonella* Typhimurium and *Salmonella* Heidelberg was also shown to be possible by the identification of proteins that result from serovar-specific nonsynonymous coding single-nucleotide polymorphism (cSNP).[24] Another interesting application is the use of TDP for the phylogenetic classification of unsequenced organisms.[25] Taken together, these results highlight the potential of TDP for a deep characterization of bacterial pathogens. Therefore, our objective here is to set up a simple and robust LC−MS/MS TDP platform which aims at discriminating closely related pathogenic bacteria through the identification of specific proteoforms. This platform does not intend, as it is, to replace MALDI-TOF MS in hospital settings but could potentially help to overcome some of its limitations.

We considered that identifying the highest number of proteins and proteoforms (with the best sequence coverage possible) was highly desirable to achieve this goal. We thus used an Orbitrap Fusion Lumos that combines features such as electron transfer dissociation (ETD) fragmentation and intact protein mode,[26] which enables an efficient analysis of intact proteins in an LC time scale.[27] *E. coli* (K12) was chosen as a simple bacterial model for the optimization of the different steps of the pipeline: sample preparation, LC separation, MS and MS/MS conditions, and data analysis. We show here that our optimized platform allows the robust identification of specific proteoforms, allowing the discrimination of enterobacterial species that are difficult to differentiate with MALDI-TOF MS.

EXPERIMENTAL SECTION

Chemicals and Reagents

LB-Miller (Luria Bertani broth medium) was prepared by the medium preparation platform of Institut Pasteur. Dulbecco's phosphate-buffered saline (PBS) (1×, Gibco), formic acid (FA), phenylmethanesulfonyl fluoride (PMSF), and ethylenediaminetetraacetic acid (EDTA) were purchased from Thermo Fisher. Ammonium bicarbonate (AB), urea, and glass beads (acid-washed) were purchased from Sigma-Aldrich. Ethanol (70%), methanol (MeOH), and acetonitrile (ACN) were purchased from Carlo-Erba. RapiGest was purchased from Waters.

Safety Considerations

All bacterial cultures and lysis have been performed in BSL2 laboratory at the CIP.

Bacterial Cell Culture and Lysis

Four *Salmonella enterica enterica* strains (serotypes Enteritidis, Typhimurium, Newport, and Muenchen), one *Shigella sonnei* strain, two *Shigella flexneri* strains (serotypes 2a and 3), and five *E. coli* strains (O157:H7, O157:H7 with Shiga-toxin 1 (*stx1*) and Shiga-toxin-2 genes, O26:H11 with *stx1* and *eae* genes, O26:H11 with *eae* gene, and MG155 K12) were obtained frozen from the Collection of Institut Pasteur (identifiers are listed in Table S1 in the Supporting Information). The bacteria were first cultured overnight at 37 °C in the LB medium. Subcultures in fresh LB were then performed in 40 mL Falcon tubes for ∼4 h and were harvested at the late exponential growth phase to obtain a bacterial density measured at O.D. 600 nm of 2.0−2.5 corresponding to $\sim 2 \times 10^9$ cells/mL. The LB medium was thus removed and the cells were treated with 70% ethanol to inactivate the pathogens and washed three times with PBS. After centrifugation, cell pellets were recovered in lysis buffer and were transferred into homogenization microtubes (BeadBug unfilled tubes, Sigma-Aldrich) for lysis. In this work, eight lysis buffers were screened: B1: PBS (1×), B2: AB 100 mM, B3: H_2O/ACN/FA 80:10:10 (v/v/v), B4: H_2O/ACN/FA 15:50:35 (v/v/v), B5: RapiGest 0.05% (w/w), B6: Urea 8M, B7: Urea 4M, and B8: Urea 2M. Protease inhibitors (PMSF, 1 mM; and EDTA, 1 mM) were added in all buffers. Cell lysis was performed by mechanical disruption using glass beads with a high-speed benchtop homogenizer (FastPrep-24-5G instrument, MP Biomedicals). Lysis was carried out according the following parameters: a speed of 6 m/s, three cycles of 30 s, and 180 s of cooling time between two cycles. Microtubes were then centrifuged at 16,000*g* for 10 min at 4 °C to remove the cell debris and the supernatant was kept (step repeated twice). Bacterial lysates were then transferred into protein LoBind tubes (Eppendorf), and the samples were aliquoted and stored at −80 °C. Protein concentration was measured using the Micro BCA Protein Assay Kit (Thermo Fisher), and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analyses were performed using 10−20% Tris-Tricine Ready Gels (Bio-Rad).

Protein Desalting

Different experimental protocols were tested for protein desalting before MS analysis: dialysis (with a 3.5 kDa cutoff), ultrafiltration (3, 50, or 100 kDa Amicon), and solid-phase extraction. All experimental details can be found in the Supporting Information.

Bottom-Up Proteomics

Label-free BUP analyses were performed to evaluate the efficiency of the protein sample preparation from bacterial cell lysates. Liquid protein digestion, MS acquisition, data processing, and statistical analyses were conducted as described in the Supporting Information.

LC Separation of Intact Proteins

For reverse-phase nano-LC, a Dionex Ultimate 3000 system, equipped with a trap column coupled to an analytical column, was used. In this work, six different chromatographic

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

conditions were evaluated (Table S2): LC1: an EASY-Spray PepSwift Monolithic PS-DVB column (200 μm × 25 cm), 1 μL·min^{-1}; LC2: a ProSwift Monolithic RP-4H column (100 μm × 50 cm), 1 μL·min^{-1}; LC3: a ProSwift Monolithic RP-5H column (100 μm × 50 cm), 1 μL·min^{-1}; LC4: an in-house packed PLRP-S (5 μm particles of 1000 Å pore size, Agilent) column (75 μm × 25 cm), 0.3 μL·min^{-1}; LC5: an in-house packed C4 (5 μm porous spherical particles of 300 Å pore size, Reprosil) column (75 μm × 60 cm), 0.5 μL·min^{-1}; and LC6: an in-house packed C4 (3.6 μm porous spherical particles of 200 Å pore size, Phenomenex) column (75 μm × 60 cm), 0.5 μL·min^{-1}. Solvent A consisted of 98% H_2O, 2% ACN, and 0.1% FA, and solvent B consisted of 20% H_2O, 80% ACN, and 0.1% FA. The following gradient (gradient 1) was used for LC1, LC3, and LC4; LC5 and LC6: 2.5% B from 0 to 4 min; 15% B at 5.6 min; 50% B at 124 min; 99% B from 126 to 131 min; and 2.5% B from 132 to 150 min. For LC2, the gradient (gradient 2) was then slightly adjusted (10% B at 5.6 min and 40% B at 124 min). The same stationary phase was used for the trap and the analytical columns when possible. All LC information is summarized in Table S2. For all LC−MS analyses, 1−1.5 μg of intact protein sample was injected.

Mass Spectrometry

An Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific) fitted with a nano-electrospray ionization source was used for all experiments. All experiments were performed using the intact protein mode at 2 mTorr as ion routing multipole pressure. All spectra were acquired in the profile mode. Several MS parameters were tested for optimization (Table 1): the MS and MS/MS resolution set at 15k, 30k, 60k, and 120k resolving power (at m/z 400); type of fragmentation: higher-energy collision dissociation (HCD) with normalized collision energies (NCEs) of 15, 20, and 25%; ETD with reaction times of 5, 10, and 15 ms; and electron-transfer/higher-energy collision dissociation (EThcD) with 10 ms of reaction time combined with NCEs of 5, 10, or 15% for supplemental activation (SA); number of microscans (μscans) (3, 6, 12, and 24 in MS and adjusted corresponding number in MS/MS); automatic gain control (AGC) target; and data-dependent acquisition parameters (top N or top speed). The various combinations are listed in Table S3A and discussed in the Results and Discussion section.

Table 1. List of All Investigated Parameters and Tested Values

parameters	tested values
AGC target (ms)	100; 250; 500
charge-state exclusion	yes; no
DDA mode	top N: 2, 4, 6; top speed: 5 s
μscans (MS)[a]	1; 2; 3; 4; 6; 12; 24
MS resolution	15k; 30k; 60k; 120k
μscans (MS/MS)[a]	2; 3; 4; 6; 12
MS/MS resolution	30k; 60k; 120k
MS/MS mode	ETD (@5 ms; @10 ms); HCD (@15NCE; @20NCE; @25NCE); EThcD (@10 ms@5NCE; @10 ms@10NCE)
precursor selection range (m/z range)	600−900; 900−1200; 600−1200; 500−1750
source fragmentation (V)	0; 15

[a]Adjusted according to the resolution settings to maintain the duty cycle constant. Please note that not all combinations have been tested (see Table S3A).

Data Analysis

All data were processed with ProSight PC v4.1 (Thermo Scientific) and Proteome Discoverer v2.4 (Thermo Scientific) using the ProSight PD 3.0 node. Spectral data were first deconvoluted and deisotoped using the cRAWler algorithm. The spectra were then searched using a three-tier search tree with searches against the appropriate Uniprot XML database (detailed in Table S4 in the Supporting Information). The proteoform databases were created using the database manager application within ProSight PC v4.1. Potential initial methionine cleavage and N-terminal acetylation, as well as known cSNPs and PTMs, were included, resulting in databases in a ProSight Warehouse File (.pwf) format. Search 1 consists of a ProSight Absolute Mass search with an MS1 tolerance of 10 ppm and an MS2 tolerance of 5 ppm. Search 2 is a ProSight Biomarker search with an MS1 tolerance of 10 ppm and an MS2 tolerance of 5 ppm. Search 3 is a ProSight Absolute Mass search performed with an MS1 tolerance of 10,000 Da and an MS2 tolerance of 5 ppm. Identifications with E-values better than 1×10^{-10}(−log(E-value) = 10) and between 1×10^{-10} and 1×10^{-5} were considered confident and medium hits, respectively. A 1% proteoform spectrum match (PrSM)-level FDR was employed.[28] Full chromatogram deconvolution was also performed using the sliding window deconvolution method and Xtract deconvolution algorithm in BioPharma Finder v3.2 (Thermo Scientific). Briefly, an intact protein analysis method was used, with chromatograms scanned from 10 to 150 min, sliding windows merge tolerance for components set at 10 ppm, charge-state ranges defined from +5 to +50, and the minimum number of detected charge states to produce a component designated as 3.

RESULTS AND DISCUSSION

In order to develop a TDP workflow dedicated to bacterial proteome analysis, we first optimized the MS parameters for intact proteins using an *E. coli* K12 lysate prepared in H_2O/ACN/FA 80:10:10 (v/v/v) (buffer B3 in the Experimental Section) already used as an internal intact protein standard mixture in our lab. Different options for LC separation were then studied. Finally, we searched for the best sample preparation conditions. To compare the different experimental conditions tested, all TDP data were processed with the same ProSight PD 3.0 workflow and the following results were used as indicators: number of proteins, proteoforms and PrSM identified, number of informative MS/MS spectrum, highest and mean PrSM scores (reported as −log P-score), mean matched PrSM ions, highest and mean PrSM MH+ values, and identification rate (Table S3).

Mass Spectrometry

Thirty-nine different LC−MS/MS conditions (Table S2A) combining different MS parameters were performed in duplicates using the LC1 conditions (see Table 1). EThcD is not the default fragmentation mode used in standard proteomics experiments, but it is widely used for intact protein fragmentation.[29] We therefore adopted it as a benchmark method in the first part of this study. Data obtained from duplicates were processed both separately (Table S3A) and together into ProSight PD, leading to a single result file per condition (Table S3B).

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

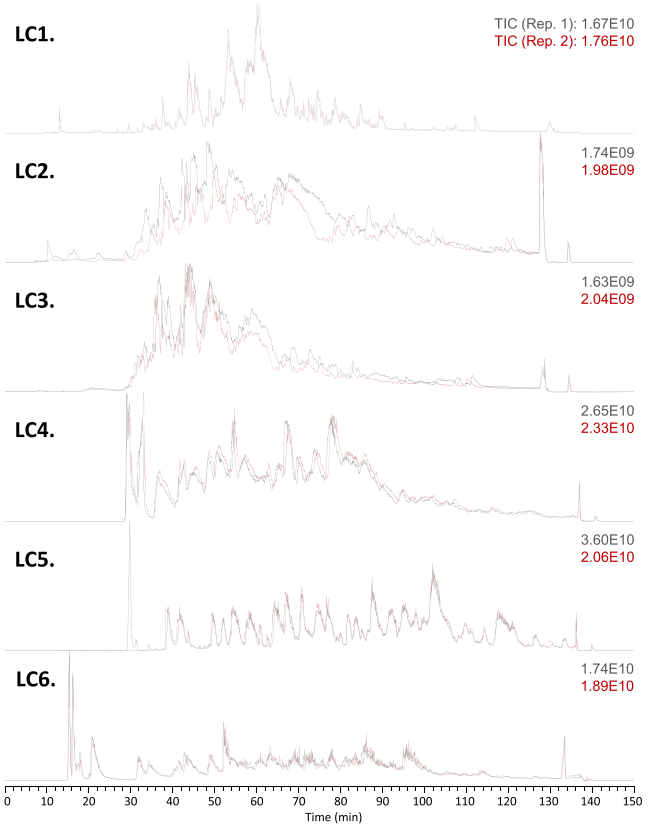


Figure 1. TIC obtained for the analysis of an *E. coli* K12 lysate in duplicate (gray and red) with different LC conditions.

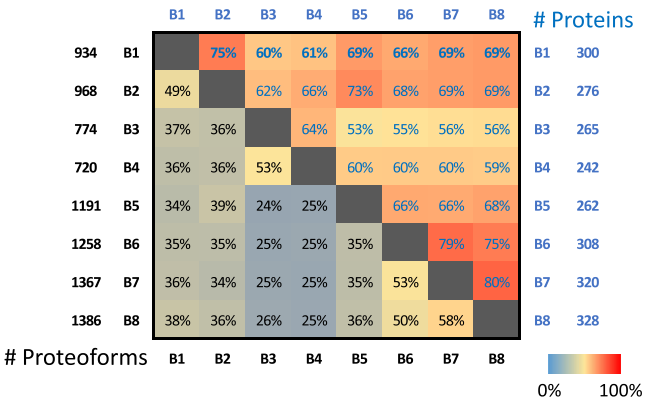


Figure 2. Number of proteins (blue) and proteoforms (black) identified in *E. coli* K12 by TDP using different lysis buffers. In the table are reported the percentages (%) of protein or proteoform similarity between buffers.

Resolution and Number of μscans

In TDP, large ions are analyzed and therefore mass resolving power is a crucial parameter.[10,11] High-resolution settings (typically > 60k, 120k, and 240k) are essential for resolving overlapping isotope patterns of intact protein ions, while low-resolution settings (15k and 30k) enable to achieve a more sensitive analysis and allow the detection of higher mass proteins. Scan speed and consequently transient length are directly related to the resolution settings and have a significant influence on sensitivity. For the number of μscans, while a single μscan is typically used in the BUP experiment, 3−10 μscans are often used for intact protein analysis.[30,31]

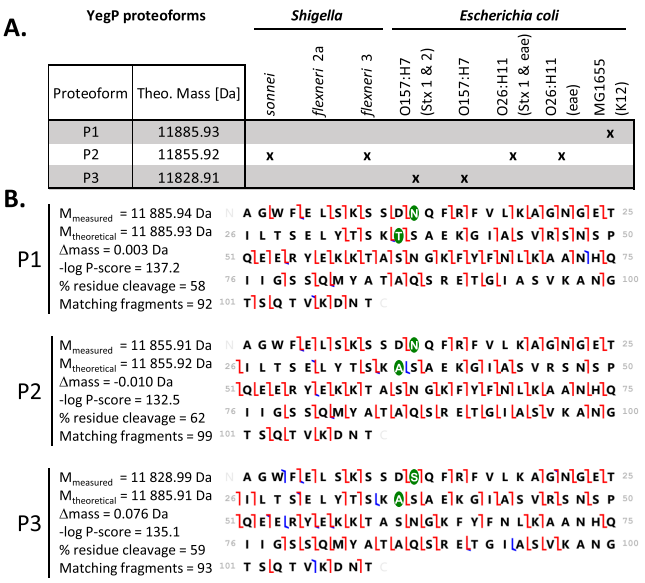


Figure 3. (A) Presence of discriminating proteoforms (P1−P3) of the YegP protein in *Shigella* and *E. coli* strains. (B) Sequence and information on P1−P3: P1 in *E. coli* K12, P2 in *S. sonnei*, and P3 in *E. coli* 0157:H7 with *Stx1* and *Stx2* genes. YegP proteoforms differ only by one or two amino acids (in green), either at position 12 (N in P1 and P2 and S in P3) or at position 36 (T in P1 and A in P2 and P3). The annotated MS/MS spectrum is provided in Figure S11.

Here, MS resolutions of 15k, 30k, 60k, and 120k were tested, with 24, 12, 6, and 3 μscans, respectively, to keep the cycle time approximately constant. In MS/MS, 30k, 60k, and 120k resolution settings were tested using 12, 6, and 3 μscans, respectively. Because it was not possible to test all combinations, we first set the MS/MS resolution at 60k and varied the MS resolution (Table S3B). The results clearly indicate that more identifications are obtained with 120k and 60k with, respectively, 173 and 169 proteins identified (307 and 296 proteoforms). We then set the MS resolution at 60k and varied the MS/MS resolution, observing the best results for 120k and 60k (186 and 169 proteins corresponding to 339 and 296 proteoforms, respectively). We thus selected 60k or 120k for MS and MS/MS and varied the number of μscans from 1 to 6. Finally, the best results were achieved using 60k with 2 μscans in MS and 60k with 2 μscans in MS/MS, with 225 proteins and 478 proteoforms.

Top Speed or Top *N*

Initially, data were obtained using a top 4 (with EThcD), and we thus tested top 2 and top 6 acquisition methods, as well as a top speed of 5 s. In general, as shown in Table S3B, higher scores and number of matched ions were observed with top *N* compared to a top speed of 5 s. Among top *N*, top 4 yielded the highest number of proteins and proteoforms identified and was therefore selected.

Fragmentation Mode

In TDP, efficient MS/MS fragmentation techniques are crucial to characterize proteoforms and localize PTMs.[32,33] In this work, we compared the results obtained from the selected activation methods listed in Table 1. Although a similar number of identified proteins was obtained for all the activation methods tested, more significant differences were observed at the proteoform level. HCD with SAs of 20 and 25% provides the largest number of proteoforms (601 and 622,

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

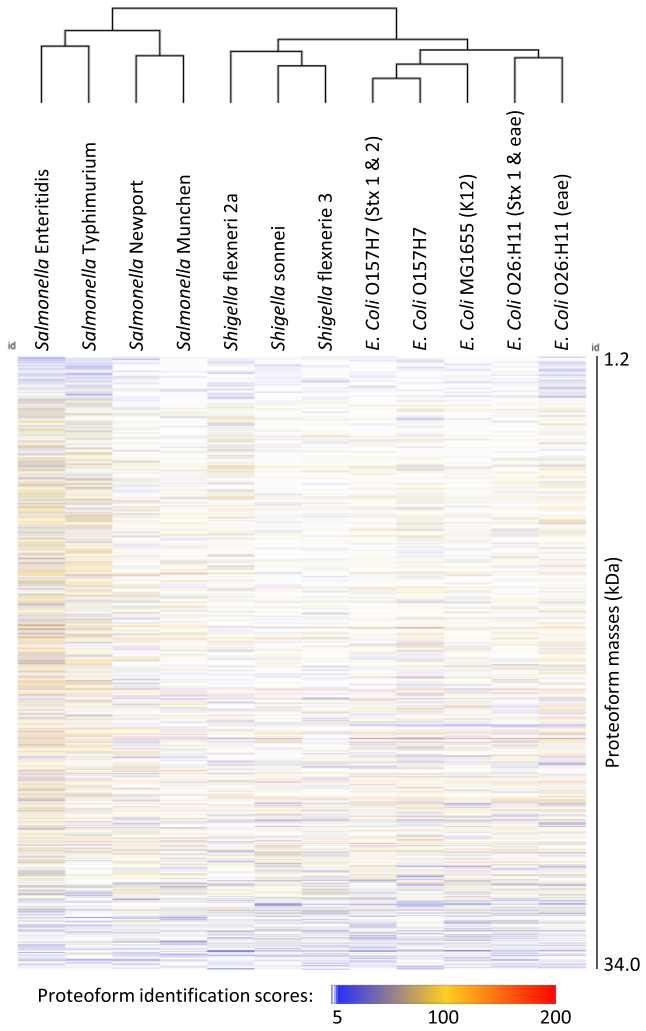


Figure 4. Hierarchical clustering of the 12 enterobacteria using the identified proteoforms and associated scores (data from Table S9). The clustering was performed using the Morpheus software (https://software.broadinstitute.org/morpheus/) with a Pearson correlation-based distance and the complete linkage method.

respectively, Table S3). This result is not surprising because HCD is faster than ETD and EThcD. However, EThcD and ETD led to much higher identification scores than those obtained in HCD, indicative of better sequence coverage, with EThcD generating more fragment ions than ETD. We therefore chose EThcD 10 ms with a SA of 10% that leads to around 500 proteoforms and very good sequence coverage.

Maximum Injection Time

In MS/MS, a high AGC target value is often required to obtain high-quality fragmentation data. We first fixed the AGC target at 5×10^5 and compared the results obtained from experiments acquired using maximum injection times of 100, 250, and 500 ms. Raw data (the number of MS and MS/MS scans) and ProSight results, both in MS (number of proteins and proteoforms) and in MS/MS (identification scores), were found almost identical (see Table S3B), indicating that the species fragmented and identified in our experiments are mainly the most abundant ones. We thus selected a medium value of 250 ms for further investigation.

Other Settings

For precursor selection range, identical results were obtained with the two largest selection windows (500−1750 and 600−1200 m/z), while a decrease of performance was observed with the two other ones (600−900 and 900−1200 m/z) (see Table S2B). This indicates that restricting the m/z selection window for the analysis of highly complex protein mixture leads to a loss of information. We therefore found more useful to keep the largest range (500−1750 m/z) for our optimized method. We then enabled the mass spectrometer to select and fragment only a single charge state per protein in order to reduce the MS/MS information redundancy and increase the number of identifications. Note that in that case, only determined charge states are selected for fragmentation. A significant decrease of PrSM and informative MS/MS spectrum was observed, with lower identification scores, although the number of proteins and proteoforms identified was found almost identical to the best previous experiment (Table S2B). This prompted us to leave this parameter off, in order to ensure more confident identifications. Finally, for source fragmentation, often applied to enhance the last stages of solvent declustering in electrospray, increasing source voltage from 0 to 15 V led to a slight decrease both in proteins and in proteoforms identified (Table S2B) and thus we decided to disable it (0 V).

MS Summary

The final optimized MS method includes full MS scans acquired at a 60k resolving power (at m/z 400) with a scan range set to 500−1750 m/z, two μscans per MS scan, an AGC target value of 5×10^5, and a maximum injection time of 50 ms. Top 4 ions with an intensity threshold $>1 \times 10^5$ were isolated with 1.2 m/z width, fragmented with EThcD (10 ms, 10%), and then added to a dynamic exclusion window for 60 s. MS/MS scans were acquired at 60k resolving power (at m/z 400), with two μscans, an AGC target value of 5×10^5, and a maximum injection time of 250 ms.

Liquid Chromatography

An efficient online separation of proteoforms is crucial for achieving their characterization.[34] Reversed-phase LC (RPLC) is the most common front-end separation for proteoforms in MS.[11] In this study, we evaluated six RPLC-based columns among the most widely used for intact proteins separation (Table S2). Three are commercial and contain a monolithic polymer (LC1, LC2, and LC3); the other three were packed in-house (LC4, LC5, and LC6). A trap column was used for all experiments using the same phase as the analytical column when possible. The *E. coli* lysate was analyzed in duplicate, using the same LC gradient (gradient 1), except for LC2 for which it was slightly adjusted (gradient 2). The total ion chromatograms (TICs) obtained are plotted in Figure 1.

An intense signal with a TIC higher than 1×10^{10} was obtained for LC1, LC4, LC5, and LC6, whereas weaker intensities were obtained for LC2 and LC3. The best protein separation was achieved for LC4, LC5, and LC6. This is illustrated in Figure S1, where an extracted ion chromatogram was performed on randomly selected charge states of proteins with different molecular weights. Higher intensities were observed with LC4, LC5, and LC6, in particular for proteins eluting at the beginning of the gradient. The differences in protein retention were retrieved in the number of MS/MS scans that are significantly different according to the LC conditions (Table S5). The better the proteins are separated, the higher the number of MS/MS scans. Moreover, increased

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

chromatographic resolution allows reduction of the probability to have coisolation of several precursors and thus multiplexed MS/MS spectra. To assess the ability of the LC conditions to generate the most comprehensive proteoform profile, we deconvoluted the entire LC−MS with the sliding window algorithm of BioPharma Finder software (see Figure S2). The profiles obtained clearly indicate that a higher number of unique molecular weights are obtained for LC1 and LC5, with 701 and 706, respectively, deconvoluted species (single run).

The number of identified proteins and proteoforms was also examined (Table S5). At the protein level, particularly good results were obtained for both LC1 and LC5, with more than 230 proteins combining two technical replicates. At the proteoform level, a higher number (711) was identified with LC5 compared to the other LC conditions. LC1 also provided good numbers with 507 proteoforms identified. For the PrSM identification score, only a small variation between LC conditions was observed. The average mass of the PrSMs identified was also slightly higher with LC1, LC2, and LC3 columns, which is consistent with the results described above. In summary, the best results were obtained with LC1 or LC5. We thus evaluated the reproducibility of both LC1 and LC5 protein elution profiles by superimposing four technical LC−MS replicates (Figure S3). A very high reproducibility was obtained in both cases.

LC5, based on the in-house packed C4 column, was finally selected because it provided the best intact proteoform separation. However, the commercial LC1 column can be used as an alternative with only slightly lower performance.

Sample Preparation

We aimed here at developing the most straightforward and rapid procedure for intact protein sample preparation for TDP analysis. For the bacterial culture step, we chose liquid cultures because they allow a more precise bacterial density measurement and thus reproducibility in sample preparation. LB medium, which is a commonly used nutritionally rich medium, was selected for all cultures. Regarding cell lysis, we avoided reagent-based methods, which lead to high concentrations of salts and detergents that are not MS-compatible and difficult to remove at the intact protein level, and chose to perform a mechanical lysis step using a high-speed benchtop homogenizer (FastPrep-24-5G instrument, MP-Biomedicals). Based on previous experiments, we found this instrument to be able to perform efficient both Gram-negative and Gram-positive bacterial lysis, using small sample volumes, without any sample dilution or contamination. To perform cell lysis and protein extraction, we evaluated eight buffers (B1 to B8 described in the Experimental Section) with different biochemical properties. PBS (B1), AB (B2), and urea (B6-8) were chosen because they are extensively used in biochemistry, in particular, for their efficiency in protein solubilization. PBS and AB can easily be removed using a regular online desalting step prior to LC−MS analysis. In contrast, urea requires an additional buffer-exchange step, which is time-consuming and can lead to protein loss. RapiGest, a surfactant commonly used to enhance the enzymatic digestion in BUP, was selected as it is cleavable under acidic conditions. Finally, buffers containing both FA and ACN were also tested because they are MS-friendly and are routinely used for bacterial identification in MALDI-TOF MS to obtain high-quality and reproducible MS profiles.[35] To evaluate protein extraction efficiency, we prepared two series of triplicate samples per buffer condition that were analyzed in BUP and TDP. The samples of the first series are numbered from 1 to 3 and those of the second series from 4 to 6. Protein lysates from buffers not compatible with LC−MS (B6, B7, and B8) were dialyzed against AB 100 mM prior to protein quantitation, SDS-PAGE, and TDP experiments. In our hands, it was difficult to eliminate the RapiGest surfactant by a simple acidification, and a buffer-exchange step was also applied to B5 samples. Protein concentration was estimated to range from 1 to 3.5 mg/mL on average for all buffers, the lower one being obtained for the two ACN-FA buffers (B3 and B4). The SDS-PAGE analysis indicated similar protein profiles for all buffers, except B3, for which less proteins of mass larger than 37 kDa were observed (Figure S4). Note that in the *E. coli* K12 proteome, half of the proteins has a molecular weight of less than 29.9 kDa (Figure S5).

The 48 lysates were analyzed in label-free BUP (as described in the Supporting Information). From 1,766 to 2,350 proteins on average were identified for the eight extraction buffers, with an average of 2,168 proteins (see Table S6A), out of 2,450 nonredundant proteins when gathering all results (Table S6B). The largest number of proteins was identified for B4 and the lowest for B3 (1,767 proteins). A high reproducibility between replicate intra- and inter-series was observed for all buffers, with less than 2.5% variation in the number of identified proteins for the six replicates per extraction condition. Results gathered in Table S6B,C highlight the high homogeneity of the results obtained with almost all buffers in terms of number of proteins. A statistical analysis was also performed to evaluate the peptide-based quantitative variations (Figure S6). The correlation matrix indicates that B1 (PBS), B2 (AB), and all urea buffers (B6, B7, and B8) lead to a comparable behavior. In conclusion, our results show that similar results are obtained in BUP with all buffers, except B3.

The 48 samples were then analyzed by TDP using our previously optimized LC−MS/MS workflow. For our study, we searched for the most straightforward clean-up method. We thus discarded the GelFREE system,[36] which leads to protein fractions with SDS and requires a protein precipitation step. For buffers not directly compatible with LC−MS (B5-8), we tested a simple cutoff mass filter (Amicon 3 kDa) and a dialysis unit (3.5 kDa). Dialysis units resulted in lower sample loss than the Amicon filter and were therefore chosen to remove urea and RapiGest. To concentrate the low-molecular-weight proteins, we also tested the 50 and 100 kDa cutoff mass filters, which were finally discarded because of a low protein recovery.

In total, from 172 to 252 proteins (371 to 783 proteoforms) could be identified depending on the buffer used. Merging all runs leads to 474 unique proteins and 3,012 proteoforms. Our top-down analysis revealed that some extraction buffers lead to similar TICs (Figure S7). Indeed, B1 and B2 TICs were comparable, as well as those between B6, B7, B8, and to a less extent B5. From 172 to 252 proteins (371 to 783 proteoforms) (see Table S7A) were identified for the 48 samples, out of a total of 474 nonredundant proteins and 3012 nonredundant proteoforms identified by combining the 48 samples. In terms of number of identifications, the results obtained are highly homogeneous with less than 7% variation for protein identification and 12% for proteoforms for the six replicates per extraction condition. At the protein level, PBS (B1) and urea buffers (B6−B8) provide the highest number of identification (between 300 and 328 considering the sum of the six replicates, Table S7B).

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

We observed that combining the six replicates per buffer allowed the identification of 242−328 proteins, which correspond to only 51−69% of all identified proteins (Table S7C). This percentage is even lower at the proteoform level with 24% to 46% (Table S7D). This shows that, in contrast to BUP, the results obtained in TDP largely depend on the buffer used. Nevertheless, as expected from the similar protein elution profiles obtained for B1 and B2, 75% of the proteins and 49% of the proteoforms identified were identical. B6−B8 also share a similar behavior with more than 75% of their proteins and 50% of their proteoforms in common (Figure 2). For the number of proteoforms, B1−B4 lead on average to 3 times more proteoforms than proteins, although much higher numbers are obtained for the other ones (B5−B8).

To understand this result, we decided to go deeper into proteoform analysis by examining the "type" of proteoforms identified by ProSight (Figure S8). The analysis of our data revealed that many proteoforms obtained for B5−B8 correspond to truncated sequences. Looking at the proteoform molecular mass distribution, we clearly observed that smaller size species were identified with these four buffers, in line with the high number of truncated proteoforms (Table S7 and Figure S9). We therefore concluded that RapiGest and urea conditions were leading to artifactual proteoforms (maybe degradation occurring during the desalting step) and thus decided to discard these buffers for our study.

Note that the largest proteoform that could be identified for all buffers is less than 30 kDa. It is often the case that large proteoforms cannot be identified in TDP LC−MS experiments because of a combination of issues: decreased LC resolution leading to coelution and competition for ionization, high number of charge states and isotopes that spread the signal over many peaks, lower fragmentation efficiency, and so forth.

Finally, B1 (PBS) was selected as the extraction buffer to prepare all bacterial lysates for TDP, as it leads to a high number of proteins/proteoforms and does not require any desalting step.

Application to the Discrimination of Enterobacterial Pathogens

Enterobacteria are small Gram-negative bacteria living mostly in the gut and responsible for a variety of diseases such as respiratory, gastroenteritis, or urinary infections.[37] The enterobacterial family constitutes one of the most diverse bacteria group, including 170 species. Around 25 species represent 95% of clinically relevant strains. A major problem for enterobacterial identification is that some species are so closely related that they are not distinguishable by MALDI-TOF MS, although leading to different clinical outcomes.[38] This is for instance the case for *Escherichia* and *Shigella*, which are also close to *Salmonella*. We therefore decided to use our TDP platform to evaluate its capacity to discriminate these pathogens. Twelve different bacterial strains of *Salmonella*, *Shigella*, and *E. coli* species were chosen (see the Experimental Section for details). Three biological replicates were performed for each one.

In total, considering the three biological replicates, from 261 to 457 proteins were obtained for all bacterial strains (Table S8), corresponding on average to 2−3 times more proteoforms (from 482 to 1,435). This large difference range can be explained by the number of proteoforms present in each database, which varies from 69,838 to 293,069 (Table S4). For instance, for *S. enterica enterica* serotypes Enteritidis and Typhimurium, a very high number of proteoforms (1,435 and 1,049, respectively) is obtained, reflecting the high number of entries. These high numbers are however slightly lower that the ones described in the Ansong *et al.* paper (1,665 proteoforms for 563 proteins). The different experimental conditions used (column length, gradient duration, and amount of sample injected) can easily explain this feature. For *S. enterica enterica* serotype Muenchen, the number of identifications is much smaller (261 proteins and 482 proteoforms) but the database used contains only eight reviewed proteins. This clearly shows the drawbacks of using database search for the analysis of microorganisms.

Nevertheless, we decided to go deeper into data analysis and search for the presence of specific proteoforms either at the species or subspecies level. To do this, we first thought of using the Basic Local Alignment Search Tool (Blast) in Uniprot. However, this is not possible because only protein sequences (and not proteoforms) can be searched. Moreover, the very heterogeneous quality of microbe databases precludes such an approach. We then considered searching all data against a unique database containing all enterobacterial sequences present in Uniprot. This makes sense because enterobacteria share a large number of proteins. However, this unique database would consist of almost 7 million protein entries, thus an exponentially increased number of proteoforms, making the analysis unmanageable by the software and potentially corrupted by too many false positives. We also tried to use the BUP identifications obtained for all strains to create a merged database (data not described in the paper). A combined result file containing 6,580 protein groups corresponding to 70,537 protein IDs was obtained, leading to an unattractable number of theoretical proteoforms by ProSight. We therefore decided to create another database by merging the sequences identified in TDP for the 12 strains after removing duplicates. This database, which contains 1,516 protein entries and 10,425 proteoforms (Table S3), was used for a two-tier search with ProSight PD. We removed the third search (large tolerance on the precursor mass) because the presence of many analogous sequences in the database would induce a high number of false positives.

As expected, a high correlation is observed between *Shigella* and *E. coli* strains, which share many proteoforms (Figure S10A). On the contrary, *Salmonella* species share much less proteoforms with the other two species. At the subspecies level, many proteoforms are found identical. For example, 408 identical proteoforms are detected both in *S. enterica enterica* serotypes Enteritidis and Typhimurium (Figure S10B). From a more general perspective, 110, 138, and 128 common proteoforms could be identified in all *Salmonella* strains, all *Shigella* strains, and all *E. coli* strains, respectively.

Remarkably, several proteoforms are also found specific at the strain level, including some differing only by a single amino acid, as illustrated in Figure 3 and Table S9. These specific proteoforms would have not been easily identified by a BUP approach because they would have required the modified peptide to be identified in each case.

These specific proteoforms clearly show that TDP can be used to discriminate closely related bacterial pathogens that cannot be differentiated with MALDI-TOF MS. This important result is also highlighted in the phylogenetic tree built from all TDP data, in which *Salmonella*, *E. coli*, and *Shigella* species are separated, with *E. coli* and *Shigella* being more closely related than *Salmonella* (Figure 4).

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

In addition, using the merged database also allowed to specifically assign proteins/proteoforms to a given species/strain (Table S9), contributing to expand our knowledge of poorly characterized strains. These data appoint TDP not only as an identification method but also as a powerful tool for microbial proteogenomics. These results highlight the added value of TDP to characterize proteomes in general, and here more specifically bacterial proteomes, with a high level of precision.

CONCLUSIONS

In this paper, we described the optimization of a TDP platform that can be used for the discrimination of closely related bacterial pathogens. We discussed in detail the four main steps of the workflow: sample preparation (in particular lysis buffer), online LC separation of intact proteins, MS/MS analysis, and database search for proteoform identification. The optimized method allowed us to identify about 220 proteins and 500 proteoforms in a single LC−MS/MS run for *E. coli*, which was used as a bacterial model.

Applied to 12 enterobacterial species belonging to the pathogens *Salmonella*, *E. coli*, and *Shigella*, our TDP platform led to the characterization of bacterial proteins at the proteoform level. Several specific proteoforms could be found for each species, showing that our platform can be used to discriminate closely related bacterial species undistinguishable with MALDI-TOF MS. The main issue we faced was database search, which performed well for reference strains such as *E. coli* K12 but turned out to be highly problematic for less-studied species. Using TDP in a routine manner to characterize bacterial species or discriminate closely related ones will require the development of a new software tool that does not rely on database search. Indeed, as in MALDI-TOF MS, we can envision a tool allowing a comparison of TDP data sets by searching for discriminative MS/MS spectra, eluding identification. This tool, combined to our optimized TDP pipeline, would represent a major step forward in clinical microbiology.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00351.

Figure S1. TICs obtained for the TDP analysis of *E. coli* with the six LC conditions; Figure S2. Deconvoluted mass spectra obtained for the TDP analysis of *E. coli* with the six LC conditions; Figure S3. LC1 and LC5 reproducibility; SDS-PAGE analysis of *E. coli* with the eight buffers; theoretical MW distribution of *E. coli* K12 proteins; statistical analysis of *E. coli* BUP data with the eight buffers; TICs for TDP analysis of *E. coli* with the eight buffers; number of PrSM identified for each buffer and type of search in ProSight PD; distribution of proteoform masses for each buffer; TDP results for all strains using a unique database; and annotated deconvoluted MS/MS spectra obtained for the discriminating proteoforms of the YegP protein (PDF)

Table S1. Enterobacterial strains studied; Table S2. LC conditions; TAble Table S3. TDP results for MS optimization; Uniprot and ProSight PC databases; TDP results for LC optimization; BUP results for sample preparation optimization; TDP results for sample preparation optimization; TDP results for all strains using specific databases; and identified proteoforms for all strains using a merged database (XLSX)

AUTHOR INFORMATION

Corresponding Author

Julia Chamot-Rooke − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France;* orcid.org/0000-0002-9427-543X; Email: julia.chamot-rooke@pasteur.fr

Authors

Mathieu Dupré − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France;* orcid.org/0000-0002-1845-0048

Magalie Duchateau − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France*

Christian Malosse − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France*

Diogo Borges-Lima − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France*

Valeria Calvaresi − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France*

Isabelle Podglajen − *Microbiology Department, Georges Pompidou European Hospital, Assistance Publique-Hôpitaux de Paris, Paris 75015, France*

Dominique Clermont − *Collection of the Institut Pasteur (CIP), Institut Pasteur, Paris 75015, France*

Martial Rey − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00351

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The MS proteomics data have been deposited to the ProteomeXchange Consortium *via* the PRIDE[39] partner repository with the data set identifier PXD019247. This work has been supported by the Institut Pasteur, CNRS, and the European EPIC-XS project number 823839, funded by the Horizon 2020 programme of the European Union. Financial support was also obtained from the Agence Nationale de la Recherche (PathoTOP project ANR-15-CE18-0021) and from the PasteurInnov (AAP PasteurInnov 214 PathoTOP) program. The authors are grateful to Q. Giai-Gianetto for the bottom-up statistical analysis and J. Dhenin for his help on figures.

REFERENCES

(1) Demirev, P. A.; Fenselau, C. Mass spectrometry for rapid characterization of microorganisms. *Annu. Rev. Anal. Chem.* **2008**, *1*, 71−93.

(2) Nassif, X. A revolution in the identification of pathogens in clinical laboratories. *Clin. Infect. Dis.* **2009**, *49*, 552−553.

(3) Singhal, N.; Kumar, M.; Kanaujia, P. K.; Virdi, J. S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* **2015**, *6*, 791.
(4) Sauget, M.; Valot, B.; Bertrand, X.; Hocquet, D. Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria? *Trends Microbiol.* **2017**, *25*, 447−455.
(5) Lasch, P.; Fleige, C.; Stämmler, M.; Layer, F.; Nübel, U.; Witte, W.; Werner, G. Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of Enterococcus faecium and Staphylococcus aureus isolates. *J. Microbiol. Methods* **2014**, *100*, 58−69.
(6) Charretier, Y.; Dauwalder, O.; Franceschi, C.; Degout-Charmette, E.; Zambardi, G.; Cecchini, T.; Bardet, C.; Lacoux, X.; Dufour, P.; Veron, L.; Rostaing, H.; Lanet, V.; Fortin, T.; Beaulieu, C.; Perrot, N.; Dechaume, D.; Pons, S.; Girard, V.; Salvador, A.; Durand, G.; Mallard, F.; Theretz, A.; Broyer, P.; Chatellier, S.; Gervasi, G.; Van Nuenen, M.; Roitsch, C. A.; Van Belkum, A.; Lemoine, J.; Vandenesch, F.; Charrier, J. P. Rapid Bacterial Identification, Resistance, Virulence and Type Profiling using Selected Reaction Monitoring Mass Spectrometry. *Sci. Rep.* **2015**, *5*, 13944.
(7) Hayoun, K.; Gouveia, D.; Grenga, L.; Pible, O.; Armengaud, J.; Alpha-Bazin, B. Evaluation of Sample Preparation Methods for Fast Proteotyping of Microorganisms by Tandem Mass Spectrometry. *Front. Microbiol.* **2019**, *10*, 1985.
(8) Nesvizhskii, A. I.; Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **2005**, *4*, 1419.
(9) Yates, J. R.; Kelleher, N. L. Top Down Proteomics. *Anal. Chem.* **2013**, *85*, 6151.
(10) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445*, 683−693.
(11) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018**, *90*, 110−127.
(12) Smith, L. M.; Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186−187.
(13) LeDuc, R. D.; Schwämmle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; Shaw, J. B.; Martin, M. J.; Vizcaino, J. A.; Alpi, E.; Danis, P.; Kelleher, N. L.; Smith, L. M.; Ge, Y.; Agar, J. N.; Chamot-Rooke, J.; Loo, J. A.; Pasa-Tolic, L.; Tsybin, Y. O. ProForma: A Standard Proteoform Notation. *J. Proteome Res.* **2018**, *17*, 1321−1325.
(14) Gault, J.; Malosse, C.; Machata, S.; Millien, C.; Podglajen, I.; Ploy, M.-C.; Costello, C. E.; Duménil, G.; Chamot-Rooke, J. Complete posttranslational modification mapping of pathogenicNeisseria meningitidispilins requires top-down mass spectrometry. *Proteomics* **2014**, *14*, 1141−1151.
(15) Gault, J.; Ferber, M.; Machata, S.; Imhaus, A.-F.; Malosse, C.; Charles-Orszag, A.; Millien, C.; Bouvier, G.; Bardiaux, B.; Péhau-Arnaudet, G.; Klinge, K.; Podglajen, I.; Ploy, M. C.; Seifert, H. S.; Nilges, M.; Chamot-Rooke, J.; Duménil, G. Neisseria meningitidis Type IV Pili Composed of Sequence Invariable Pilins Are Masked by Multisite Glycosylation. *PLoS Pathog.* **2015**, *11*, No. e1005162.
(16) Gault, J.; Vorontsov, E.; Dupré, M.; Calvaresi, V.; Duchateau, M.; Lima, D. B.; Malosse, C.; Chamot-Rooke, J. Top-Down Proteomics in the Study of Microbial Pathogenicity. *MALDI-TOF and Tandem MS for Clinical Microbiology*; John Wiley & Sons, 2017; pp 493−504.
(17) Randall, E. C.; Bunch, J.; Cooper, H. J. Direct analysis of intact proteins from Escherichia coli colonies by liquid extraction surface analysis mass spectrometry. *Anal. Chem.* **2014**, *86*, 10504−10510.
(18) Ansong, C.; Wu, S.; Meng, D.; Liu, X.; Brewer, H. M.; Deatherage Kaiser, B. L.; Nakayasu, E. S.; Cort, J. R.; Pevzner, P.; Smith, R. D.; Heffron, F.; Adkins, J. N.; Pasa-Tolic, L. Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 10153−10158.
(19) Bunger, M. K.; Cargile, B. J.; Ngunjiri, A.; Bundy, J. L.; Stephenson, J. L., Jr. Automated Proteomics ofE. colivia Top-Down Electron-Transfer Dissociation Mass Spectrometry. *Anal. Chem.* **2008**, *80*, 1459−1467.
(20) Simpson, D. C.; Ahn, S.; Pasa-Tolic, L.; Bogdanov, B.; Mottaz, H. M.; Vilkov, A. N.; Anderson, G. A.; Lipton, M. S.; Smith, R. D. Using size exclusion chromatography-RPLC and RPLC-CIEF as two-dimensional separation strategies for protein profiling. *Electrophoresis* **2006**, *27*, 2722−2733.
(21) Durbin, K. R.; Fellers, R. T.; Ntai, I.; Kelleher, N. L.; Compton, P. D. Autopilot: an online data acquisition control system for the enhanced high-throughput characterization of intact proteins. *Anal. Chem.* **2014**, *86*, 1485−1492.
(22) Wu, S.; Brown, R. N.; Payne, S. H.; Meng, D.; Zhao, R.; Tolić, N.; Cao, L.; Shukla, A.; Monroe, M. E.; Moore, R. J.; Lipton, M. S.; Paša-Tolić, L. Top-Down Characterization of the Post-Translationally Modified Intact Periplasmic Proteome from the Bacterium Novosphingobium aromaticivorans. *Int. J. Proteomics* **2013**, *2013*, 279590.
(23) Williams, T. L.; Monday, S. R.; Edelson-Mammel, S.; Buchanan, R.; Musser, S. M. A top-down proteomics approach for differentiating thermal resistant strains ofEnterobacter sakazakii. *Proteomics* **2005**, *5*, 4161−4169.
(24) McFarland, M. A.; Andrzejewski, D.; Musser, S. M.; Callahan, J. H. Platform for identification of Salmonella serovar differentiating bacterial proteins by top-down mass spectrometry: S. Typhimurium vs S. Heidelberg. *Anal. Chem.* **2014**, *86*, 6879−6886.
(25) Wynne, C.; Fenselau, C.; Demirev, P. A.; Edwards, N. Top-down identification of protein biomarkers in bacteria with unsequenced genomes. *Anal. Chem.* **2009**, *81*, 9633−9642.
(26) Riley, N. M.; Mullen, C.; Weisbrod, C. R.; Sharma, S.; Senko, M. W.; Zabrouskov, V.; Westphall, M. S.; Syka, J. E. P.; Coon, J. J. Enhanced Dissociation of Intact Proteins with High Capacity Electron Transfer Dissociation. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 520−531.
(27) Scheffler, K.; Viner, R.; Damoc, E. High resolution top-down experimental strategies on the Orbitrap platform. *J. Proteomics* **2018**, *175*, 42−55.
(28) LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Shams, D. P.; Thomas, P.; Kelleher, N. L. Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. *Mol. Cell. Proteomics* **2019**, *18*, 796.
(29) Brunner, A. M.; Lössl, P.; Liu, F.; Huguet, R.; Mullen, C.; Yamashita, M.; Zabrouskov, V.; Makarov, A.; Altelaar, A. F. M.; Heck, A. J. R. Benchmarking multiple fragmentation methods on an orbitrap fusion for top-down phospho-proteoform characterization. *Anal. Chem.* **2015**, *87*, 4152−4158.
(30) Ahlf, D. R.; Compton, P. D.; Tran, J. C.; Early, B. P.; Thomas, P. M.; Kelleher, N. L. Evaluation of the compact high-field orbitrap for top-down proteomics of human cells. *J. Proteome Res.* **2012**, *11*, 4308−4314.
(31) Scheffler, K. Top-down proteomics by means of Orbitrap mass spectrometry. *Methods Mol. Biol.* **2014**, *1156*, 465−487.
(32) Greer, S. M.; Brodbelt, J. S. Top-Down Characterization of Heavily Modified Histones Using 193 nm Ultraviolet Photodissociation Mass Spectrometry. *J. Proteome Res.* **2018**, *17*, 1138−1145.
(33) Shliaha, P. V.; Gibb, S.; Gorshkov, V.; Jespersen, M. S.; Andersen, G. R.; Bailey, D.; Schwartz, J.; Eliuk, S.; Schwämmle, V.; Jensen, O. N. Maximizing Sequence Coverage in Top-Down Proteomics By Automated Multimodal Gas-Phase Protein Fragmentation. *Anal. Chem.* **2018**, *90*, 12519−12526.
(34) Capriotti, A. L.; Cavaliere, C.; Foglia, P.; Samperi, R.; Laganà, A. Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics. *J. Chromatogr. A* **2011**, *1218*, 8760−8776.
(35) Drevinek, M.; Dresler, J.; Klimentova, J.; Pisa, L.; Hubalek, M. Evaluation of sample preparation methods for MALDI-TOF MS identification of highly dangerous bacteria. *Lett. Appl. Microbiol.* **2012**, *55*, 40−46.
(36) Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; Marshall, A. G.; Kelleher, N. L. A robust two-dimensional

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX

separation for top-down tandem mass spectrometry of the low-mass proteome. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2183−2191.
(37) Octavia, S.; Lan, R. The Family Enterobacteriaceae. *The Prokaryotes*; Springer: Berlin, Heidelberg, 2014; pp 225−286.
(38) Devanga Ragupathi, N. K.; Muthuirulandi Sethuvel, D. P.; Inbanathan, F. Y.; Veeraraghavan, B. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. *New Microbes New Infect.* **2018**, *21*, 58−62.
(39) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yılmaz, Ş.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2018**, *47*, D442−D450.

https://dx.doi.org/10.1021/acs.jproteome.0c00351
J. Proteome Res. XXXX, XXX, XXX−XXX