

Chapter 10

Bacterial Identification at the Serovar Level by Top-Down Mass Spectrometry

Melinda A. McFarland, Denis Andrzejewski and John H. Callahan

Introduction

Members of the *Salmonella enterica enterica* subspecies are the cause of most human salmonellosis and in the USA, most cases are food-borne. *S. enterica enterica* consists of more than 2500 different O and H cell surface antigen combinations, or serovars (FDA 2012). *S. enterica* serovar Typhimurium and *S. enterica* serovar Heidelberg are among the top ten serovars implicated in food-borne *Salmonella* infections (CDC 2014). Although these are distinct serovars, their genomes are 99% similar (data not shown). Species- and subspecies-level assays are generally adequate for clinical diagnostics. However, localization of the source of a food-borne *Salmonella* contamination requires serovar or strain-level specificity.

Pulsed field gel electrophoresis (PFGE) has become the gold standard for molecular subtyping of *Salmonella*, and polymerization chain reaction (PCR)-based assays built around genomic markers are becoming increasingly popular (Wattiau et al. 2011). Differentiating between two highly similar serovars such as *S. Typhimurium* and *S. Heidelberg* requires multiple enzymes and relies on matching to a previously validated standard. Detection methods that require selection of probe-based assays, such as PCR, are limited by probe selection. Changes to untargeted genes and newly acquired genetic material are likely to be missed. More recently, approaches based on whole-genome sequencing (WGS) have been used to address strain identification (Lienau et al. 2011).

Mass spectrometry is a powerful analytical tool that can be used to probe proteins, peptides, lipids, and metabolites produced by bacteria; mass spectrometers are a ubiquitous, sensitive, specific, and inherently multiplexed platform that can potentially be used to identify and differentiate bacteria. A nontargeted mass spectrometry-based method provides a relatively unbiased snapshot of the expressed

M. A. McFarland (✉) · D. Andrzejewski · J. H. Callahan
Spectroscopy and Mass Spectrometry Branch, Center for Food Safety and Applied Nutrition,
US Food and Drug Administration, 5100 Paint Branch Pkwy., College Park, MD 20740, USA
e-mail: melinda.mcfarland@fda.hhs.gov

© Springer International Publishing Switzerland 2016
P. Demirev, T. R. Sandrin (eds.), *Applications of Mass Spectrometry in Microbiology*,
DOI 10.1007/978-3-319-26070-9_10

257

proteins in a wide range of bacterial samples and is amenable to both screening and targeted analysis. This facilitates differentiation of closely related bacteria, as well as the detection of un-sequenced or newly acquired non-synonymous SNPs and plasmid proteins that may be specific to a given strain.

Mass spectrometry is commonly used to identify proteins from the bottom-up, using peptides derived from enzymatic digestion of protein lysates (McCormack et al. 1997). However, the cross genome homology present in bacteria limits the feasibility of differentiation across closely related isolates by bottom-up peptide-based analysis. If an MS/MS spectrum is not generated for the SNP (henceforth, this term will be used to mean non-synonymous or non-silent SNPs)-containing peptide, the presence of that SNP will be missed. If the SNP has not been genomically sequenced or is not present in the searched database, the biomarker will also go undetected. The identification of unknown bacterial lysates lacking fully sequenced genomes may be challenging due to a bias toward those species that are most represented in the database. Consequently, there is a distinct advantage of using intact proteins to detect differences induced by non-synonymous SNPs, as the presence of such mutations would result in measurable differences in the mass of the intact protein, with no need for a sequenced genome.

Intact protein mass spectrometry of bacterial lysates provides an inherently multiplexed measurement of the mass of expressed proteins in their intact state, at a given growth stage (Krishnamurthy and Ross 1996; Fenselau and Demirev 2001; Conway et al. 2001). This is particularly useful because bacteria exhibit fewer overall post-translational modifications (PTMs) and, given a controlled growth state, minimal PTM variability as compared to mammalian systems. Bacterial proteins and their modifications are highly conserved across species. Although protein abundances may vary from serovar to serovar, their masses should be highly conserved. Therefore, for bacterial lysates it is a reasonable assumption that the minimal mass shifts found between closely related bacteria are the result of SNPs (Wilcox et al. 2001; Dieckmann et al. 2008; Arnold and Reilly 1999). These mass-shifted proteins serve as biomarkers for differentiation of bacteria.

Intact protein mass spectrometry has become a commercially available tool for clinical bacterial differentiation based on the matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) technology (Bizzini and Greub 2010; Clark 2013). However, a mass range generally limited to below 15 kDa and a bias toward ribosomal proteins (Ryzhov and Fenselau 2001) often limit MALDI applications to species- and subspecies-level identifications. The increased mass range, improved reproducibility, and greater number of proteins ionized using an electrospray ionization (ESI)-based platform provide access to a more diverse range of proteins and an increased specificity for differentiation of closely related bacteria (Krishnamurthy et al. 1999; Ho and Hsu 2002; Mott et al. 2010). This approach, known as intact protein chromatography electrospray mass spectrometry, has already been used to identify marker masses that differentiate thermophilic versus non-thermophilic groups of *Cronobacter sakazakii* (Williams et al. 2005) to identify proteins characteristic of specific outbreak strains of *Vibrio parahaemolyticus* (Williams et al. 2004), and to differentiate closely related species within the enterobacteriaceae family (Mott et al. 2010; Everley et al. 2008).

The addition of online “top-down” MS/MS fragmentation of the intact proteins provides identification of the proteins containing measured mass differences (Cargile et al. 2001; Lee et al. 2002; Fagerquist et al. 2006; Wynne et al. 2010; McFarland et al. 2014). By identifying which of the most highly expressed bacterial proteins are conserved and which contain amino acid differences, we can differentiate between samples, validate genomically predicted SNPs for sequenced genomes, and for un-sequenced species, determine whether a mass shift in a specific protein represents a novel, and possibly virulent, mutation. This provides a direct link back to genome-sequencing data, facilitating gene-specific marker and sequence validation at an expressed protein level.

The combination of intact protein chromatography ESI-MS with top-down mass spectrometry facilitates the identification of proteins that result from expressed serovar-specific non-synonymous SNPs. This approach is based on deconvoluted ESI-MS generated intact protein expression profiles (Williams et al. 2002) to facilitate rapid differentiation between samples, combined with top-down identification of proteins for marker confirmation. Application of this methodology as a screening method would require sequencing only expression profile masses that show a mass shift when compared to a reference strain, and such an analysis can be done without prior selection of biomarker proteins and without a sequenced genome. Knowledge of which protein sequences are variable across serovars provides a common link to genome sequencing and phylogenetic strain-typing efforts.

Methods

Bacterial Strains

Salmonella enterica enterica serovar Typhimurium strain LT2 and *S. Heidelberg* strain A39 bacterial strains used in the study were obtained from the stock culture collection of the Food and Drug Administration (FDA)/Center for Food Safety and Applied Nutrition. Bacteria were grown for 24 h at 37 °C on lysogeny broth agar plates (Teknova, Hollister, CA). For the multi-isolate study, 36 semi-blinded *Salmonella* isolates from food-borne outbreaks investigated by the FDA were cultured overnight on tryptic soy agar plates. Cell isolates were collected in a 1.5-mL sample tube and washed twice with sterile water and resuspended in 0.5-mL of 70% ethanol to facilitate sterilization of bacteria (Williams et al. 2003) as well as minimize protease activity. The approximate cell concentration is 8×10^{10} cfu/mL.

Extraction of Cellular Proteins

The sample tube containing bacterial cells suspended in 70% ethanol was centrifuged at 9800xg for 5 min. The ethanol solution was removed, and 1.0 mL of a 50:49:1 extraction solution consisting of acetonitrile, high-performance liquid

chromatography (HPLC)-grade water, and formic acid was added and the tube was vortexed to resuspend the cells. The 1.0 mL suspension was transferred to a Barocycler® FT500 pulse tube (Pressure Biosciences, Inc., Boston, MA) along with an additional 0.4 mL of extraction solution and was capped. The Barocycler NEP 3229 was pressure cycled 24 times at 44 °C starting at 35,000 psi for 15 s and then at 0 psi for 10 s. The pulse tube contents were transferred to a 1.5-mL low-binding sample tube and centrifuged at 9800 \times g for 20 min to pellet the cellular debris. A portion of the supernatant was transferred to an autosampler tube for LC-MS analysis.

HPLC of Intact Proteins

Intact proteins were separated by reverse-phase HPLC using an Agilent (Palo Alto, CA) 1100 system fitted with two ProSphere P-HR (W.R. Grace, MD) 2.1 mm i. d. \times 15 cm columns connected in series. Two microliters of the protein extract were injected into the column at an oven temperature of 50 °C and a flow rate of 200 μ L/min. Mobile phase A was 95 % HPLC-grade water and mobile phase B was 95 % acetonitrile, both with 5 % acetic acid. The gradient was as follows: 0–5 min 90 % A, hold for 1 min, 70 min 50 % A, 80 min 10 % A, 92 min 10 % A, and 94 min 90 % A. Identical separation methods were used in-line with both instrument platforms to retain consistent retention times across platforms. For the multi-isolate study, all conditions were the same, except proteins were separated on a Kinetex C8 (Phenomenex, Torrance, CA) 1.7 μ m, 100A, 15 cm column, with mobile phase A 98 % HPLC-grade water and mobile phase B 98 % acetonitrile, both with 2 % formic acid.

LC-MS and Data Analysis

The HPLC was interfaced to a Q-TOF Premier (Waters, Beverly, MA) mass spectrometer. The instrument was operated at 3.0 kV capillary voltage, 100 °C source temperature, 150 °C desolvation temperature, desolvation gas 600 L/h, and scanning from 550 to 2000 Da in 1.0 s in single reflectron mode. Data were collected using MassLynx software version 4.1 (Waters, Beverly, MA).

MS Data Analysis

Automated analysis of full-scan (MS) data was performed with ProTrawler6 (previously named Retana) and custom software (BioAnalyte, Inc., Portland, ME). Its function is to automatically process sequential complex, multiply charged mass

spectra obtained during ESI-LC-MS analysis and produce a text file containing the binned uncharged protein mass, retention time, and intensity of all proteins deconvoluted from the LC-MS run. A detailed explanation of the approach has been published (Williams et al. 2002). Briefly, spectra are summed in 30 s windows. In version 6 of ProTrawler the summed spectrum from each time window is baseline subtracted and de-noised using the proprietary ReSpect™ algorithm (Positive Probability, Shrewsbury, UK). The resultant spectrum is deconvoluted using maximum entropy deconvolution. After generating a protein mass/abundance list for each time window, ProTrawler then bins the data for each time window, determines the time range over which a given mass occurs, and calculates an abundance-weighted time centroid for the mass, which is used to represent the retention time. Masses corresponding to multimers and adducts are also removed. Abundances are then normalized to the summed intensity. The resulting text file contains a cumulative list of all the intact protein masses, abundances, and retention times, of which the mass and abundance information can be represented graphically as mass versus intensity, similar to a traditional mass spectrum. The retention time is also included in the output so that proteins of similar mass can be distinguished based on the retention time.

Top-Down LC-MS/MS

Online intact protein separation was the same as for the Q-TOF LC-MS (above) for consistent protein retention times across platforms. For LC-MS/MS the eluent flow was split to a flow rate of 350 nL/min via the TriVersa NanoMate (Advion BioSciences, Ithaca, NY) chip-based nanospray source and analyzed with a LTQ-Orbitrap XL (Thermo Fisher, San Jose, CA) mass spectrometer. The instrument was operated in a top-three data dependent mode, with both MS spectra and collision-induced dissociation (CID) MS/MS spectra acquired at 60,000 resolving power in the Orbitrap. CID collision energy was operated at 15%. Each MS spectrum was composed of three microscans, and each MS/MS spectrum was the average of 10 microscans. To facilitate the analysis of intact proteins, the instrument was operated with the HCD gas off and the delay before image current detection shortened to 5 ms.

Top-Down Data Analysis

ProSightPC 2.0 (Zamdborg et al. 2007) was used to search MS/MS spectra against a protein sequence library of UniprotKB Swiss-Prot and TrEMBL protein sequence entries for the *Salmonella* Typhimurium fully sequenced strain LT2 or a custom-made *S. Heidelberg* database from fully sequenced strain SL476 (as of the time of this work a fully sequenced A39 genome was not available). Neutral mass deconvoluted precursor and fragment mass lists were generated with the Xtract algorithm

(Thermo Fisher, San Jose, CA) option within ProSightPC 2.0. The precursor mass tolerance was 1000 Da, and the fragment ion tolerance was 20 ppm for the monoisotopic mass. Only disulfide bonds were included as a modification in the primary search. PTMs were inferred from mass differences relative to the theoretical mass. Modifications were subsequently validated by manual addition of the proposed modification followed by re-assignment of fragment ions and rescoring via the sequence gazer option in ProSightPC. Modifications were considered valid if there was an increase in matched fragment ions upon inclusion of the predicted modification. A secondary search was also performed that included the most commonly inferred PTMs as confirmation of the amended modification as the top-scoring identification. Only proteins identified with ProSight *e*-values better than $1e^{-5}$ for a minimum of three MS/MS spectra were considered valid identifications.

Results and Discussion

The power of intact protein analysis is that the mass of the protein is measured with functional modifications intact. This is ideally suited for bacterial proteins because, unlike mammalian systems, bacterial lysates from similar species appear to exhibit highly reproducible and conserved PTMs under similar growth conditions. Although protein abundances may vary, there should be few differences in their masses. Therefore, for bacterial lysates grown under the same conditions, it is reasonable to assume that a small number of mass shifts found across serovars are SNPs, and novel masses are insertions or proteins that have undergone a significant change in the expression level. These mass-shifted proteins serve as markers for differentiation of bacteria at the species, subspecies, and serovar levels.

Intact Protein Expression Profiles

To facilitate nontargeted SNP discovery, the intact accurate mass, retention time, and relative abundance of proteins from the soluble fraction of bacterial lysates are measured and compared using LC-MS. Figure 10.1a shows a representative total ion current chromatogram from a 90-min LC-MS analysis of an intact bacterial protein lysate. Mass spectra were summed in 30-s windows, and each window was deconvoluted using ProTrawler6 software (Williams et al. 2002). Unlike mass spectra of peptides, intact proteins produce broad charge state distributions, effectively splitting the ion current generated for a given protein over multiple structural conformations (Fig. 10.1b). The elution profile of each protein is 1.5 min wide on average, further distributing the ion current, as well as greatly increasing the likelihood of multiple co-eluting proteins. Consequently, software is necessary to deconvolute each spectrum (or summed spectra) (Fig. 10.1c) and merge consecutive abundances into a single protein mass and intensity. The result (Fig. 10.1d)

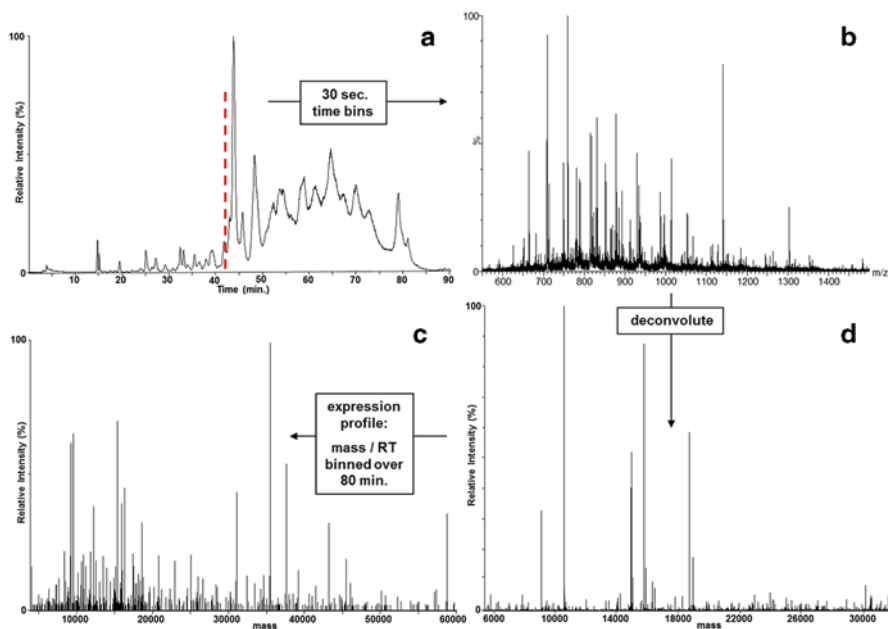


Fig. 10.1 Intact protein expression profile generation. ProTrawler software was used to deconvolute and reconcile all MS scans from the chromatogram into a single mass, retention time, and abundance profile. **a** Representative chromatogram from a 90-min LC-MS analysis of a *S. enterica* strain LT2 intact protein lysate. **b** Mass spectra were summed into 30-s bins across the chromatogram. **c** The resultant spectra at each time interval was deconvoluted to produce a series of neutral mass peak lists consisting of mass, retention time, and intensity. **d** Bins were merged into a single profile based on mass and retention time tolerance. The result is an intact protein expression profile that visually simplifies the assessment of protein differences between lysates. (Reprinted with permission from McFarland et al. 2014. Copyright 2015 American Chemical Society)

is an intact protein expression profile or mass map that represents the masses and intensities of all proteins detected across the chromatogram. This approach has the visual simplicity of a MALDI spectrum but with the greater information content provided by chromatographically resolved ESI spectra. The increase in the number of detectable masses provided by an extended mass range and improved ionization of proteins yields a greater capacity for differentiation as compared to MALDI-MS. The power of our method is the visualization of all proteins detected in an LC-MS experiment in a single spectrum, thus providing a quicker and more complete assessment of differences when compared to relying solely on LC-MS/MS protein or peptide identifications to assess changes between samples (Everley et al. 2008). Intact protein expression profiles facilitate rapid assessment of differential proteins as possible biomarkers and offer a larger dynamic range as compared to chromatographic alignment alone.

Tracing back to the source of a *Salmonella* contamination requires a minimum of serovar-level differentiation. Serovar differentiation is not currently possible on commercially available MALDI-based clinical bacterial typing platforms. *Salmo-*

nella enterica enterica Typhimurium and Heidelberg are closely related serovars that have both been implicated in food-based outbreaks (CDC 2014). Recent phylogenetic and MLST analysis (Bell et al. 2011) confirm that the chosen strains are members of two closely related serovars. Figure 10.2 shows a mirrored comparison of the LC-MS generated intact protein expression profiles of these serovars. Each profile is the result of deconvolution and binning of mass, abundance, and retention time from a representative 90-min LC-MS run. As is expected by the extreme homology across the *Salmonella* species and the similarity of these two serovars, the mass maps look nearly identical, with differences occurring in only a small number of detectable masses.

One can readily observe that the majority of masses detected are conserved across serovars. The observed mass shifts likely represent protein products of SNP-containing genes that differentiate *S. enterica* serovar Typhimurium strain LT2 from *S. Heidelberg* strain A39 and are likely biomarkers for serovar identification. No protein sequencing is required to determine the presence of mass shifts and/or novel masses, and markers do not need to be known prior to analysis.

Top-Down Protein Identification

It has been previously shown that comparisons of intact protein expression profiles are sufficient to differentiate two bacterial serovars (Williams et al. 2004, 2005; Everley et al. 2008). Although the presence of a differential pattern is sufficient for grouping a serovar with a set of previously run samples, it does not readily facilitate identification of uncharacterized strains and provides little to link the result with complementary assays such as targeted PCR probes or genome sequencing. Confirmation of the identity of differential masses as orthologs is necessary to validate the protein as a viable biomarker. The second stage of this method is the addition of top-down MS/MS identification of proteins to the existing LC-MS separation method (Fig. 10.2; McFarland et al. 2014). Proteins maintain the same elution profile but now the most abundant proteins are identified. The recent introduction of faster instruments with improved data-dependent selection increases the number of proteins identified in a single run.

Protein identifications in Fig. 10.2 are represented by the protein name, as assigned for the reference genome of *S. Typhimurium* strain LT2. A complete list of identified proteins and a detailed description of PTM assignments can be found in McFarland et al. (McFarland et al. 2014). Although, in general, the highly conserved protein sequences of related bacterial strains make strain typing challenging, it also means that the vast majority of fragment ions match across proteomes. Searching top-down MS/MS spectra does not require the strict precursor mass accuracy of bottom-up proteomics. In this work, the precursor mass error was permitted to be 1000 Da to account for unpredicted signal peptides and unknown PTMs, such

While a high-throughput top-down approach identifies fewer proteins and SNPs than a typical bottom-up survey, we gain independence from the need for a strain-specific sequenced genome. Comparison of intact protein expression profiles by mass, retention time, and relative abundance is sufficient for determination of masses that differ across serovars. Reproducible SNP identification in a bottom-up experiment would require the sequenced genome, such that the novel SNP must be present in the searched database. Identification of a SNP-containing peptide that is not in the database would require de-novo sequencing of unassigned peptides. Peptide SNP identification by spectral similarity alignment may be possible, but knowledge of the full degree of genetic drift is difficult without knowledge of the mass of the intact protein because complete peptide sequence coverage is rarely achieved. An obvious strength of the intact protein-based methodology presented here is that any differences as compared to proteins in a reference strain are readily apparent.

Proteogenomics

Maintaining a protein's intact mass while still being able to identify the protein to a homologous protein sequence is also advantageous for proteogenomic-based reconciliation of the mass spectrometric detection of expressed proteins with genome sequencing data. This provides a direct link to complementary genome-based methods as well as a mechanism for the detection of genome sequencing errors. For example, protein ElaB identified in *S. Typhimurium* strain LT2 has a theoretical mass of 418 Da greater than its measured mass. The identity of the measured mass was confirmed by CID fragmentation, with 21 y-ions identified. No b-type fragment ions were identified, and the measured mass differs from the theoretical mass as stated (Fig. 10.3a). The assigned e-value of 3.5×10^{-20} confirms confident protein identification, and the absence of b-ions points to a mass discrepancy at the N-terminus. The measured mass of the same protein in *S. Heidelberg* strain A39 does reconcile with its measured mass (after cleavage of the initiator methionine), strongly suggesting that the large mass discrepancy is not due to an unpredicted PTM. Alignment of the *S. Typhimurium* strain LT2 theoretical protein sequences with that of the same protein from another sequenced *S. Typhimurium* strain (strain U288) shows that the mass discrepancy lies at the translational start site of the protein (Fig. 10.3b). Confirmation of a sequencing start site error is seen in Fig. 10.3c. Removal of the erroneous amino acids increases the precursor mass accuracy to less than 3 ppm and results in the identification of a string of N-terminal containing b-type fragment ions. Identification of protein sequences combined with an intact mass measurement provides a unique link to genome sequencing and phylogenetic stg efforts. As the use of high-throughput genome sequencing annotation pipelines increases, validation of start site errors will minimize the propagation of start site errors through multiple genomes.

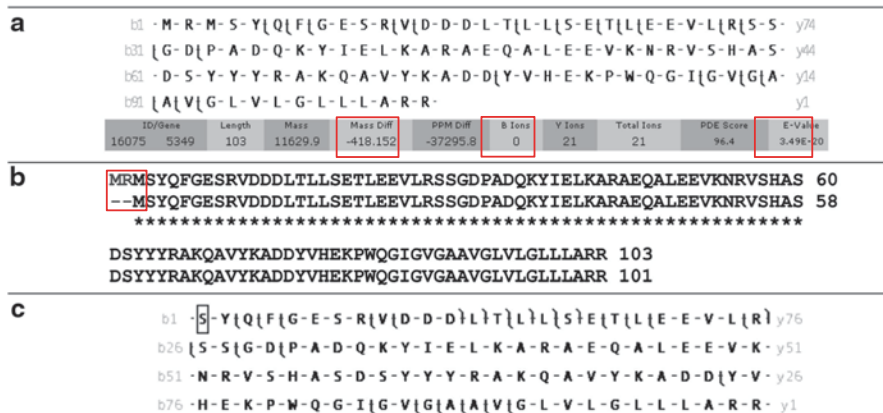


Fig. 10.3 Top-down mass spectrometry to verify genome annotation. **a** The intact mass measured in *S. Typhimurium* strain LT2 for SNP-containing protein ElaB does not agree with the theoretical mass. Genome sequencing predicts a larger mass difference between serovars than is actually expressed. Top-down MS/MS identifies the correct protein but no b-type fragment ions are assigned. **b** Comparison of the predicted protein sequence for strain LT2 against ElaB sequences predicted from other strains (here shown for *S. Typhimurium* strain U288) shows disagreement at the N-terminus. **c** Correction of the N-terminal amino acids in the LT2 sequence results in the additional identification of a substantial sequence tag of b-type ions. (Reprinted with permission from McFarland et al. 2014. Copyright 2015 American Chemical Society)

Multiplexed Serovar Identification of Semi-blinded Isolates

To demonstrate the specificity and scalability of intact protein LC-MS expression profiles for *Salmonella* serovar identification, the method was applied to a semi-blinded study of 36 *Salmonella* isolates originating from food-borne outbreaks (McFarland et al. 2014). Study creators established sample relatedness at the serotype, PFGE, and WGS levels.

Representative LC-MS generated intact protein expression profiles for each serovar are shown in Fig. 10.4. Labeled masses are SNP-containing proteins, SodA, YfeA, and OmpA. Combinations of these markers were sufficient to correctly identify the serovar type for all 36 *Salmonella* isolates, four serovars represented by nine isolates each. Neither the identity of the isolates nor the differentiating protein markers were known in advance. Markers were picked from the resultant LC-MS expression profiles, based on variable masses in abundant proteins. No one marker was sufficient to differentiate all four serovars. As is expected for blind identification, more than one marker is necessary. It is worth noting that top-down identifications of serovar-specific biomarkers did not need to be performed because protein identifications were known from previous top-down work on *S. Heidelberg* and *S. Typhimurium* (McFarland et al. 2014) and were confirmed based on the retention time. Differentiating protein markers were then used to confirm serovar assignments by comparing the measured masses with in-silico protein sequences from publically available protein databases, providing a direct link to genome sequencing data.

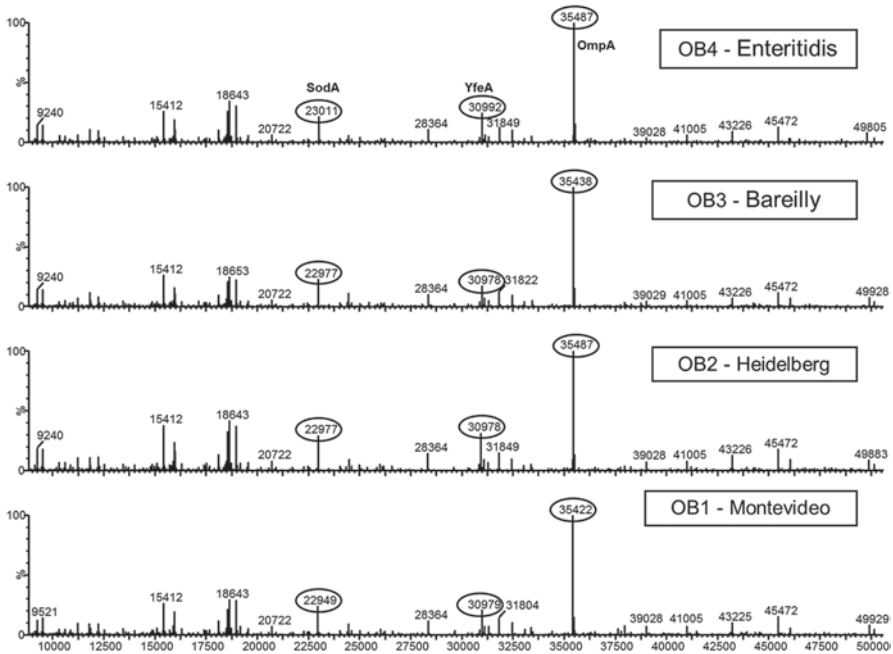


Fig. 10.4 Representative LC-MS generated intact protein expression profile for each serovar. Circled masses are SNP-containing proteins, SodA, YfeA, and OmpA. Combinations of these markers were sufficient to correctly identify serovar type for all 36 semi-blinded isolates. Markers did not need to be selected in advance. LC-MS profiles were acquired and markers were chosen based on the resultant data

LC-MS intact protein expression profiles assigned the correct serovar type for all 36 isolates, as determined by the study key based on PFGE and WGS of the outbreak samples. Mass and abundance profiles generated from triplicate analysis of each strain were used for PCA analysis. Each of the 36 isolates clustered into one of four distinct clusters corresponding to each of the four serovars. Although LC-MS may not provide the strain-level specificity of WGS, LC-MS should offer the same level of specificity as any marker-based method but without the need for preselection of markers. This offers flexibility given that different combinations of markers will be required depending on the serovar in question.

Conclusion

As the speed of whole genome sequencing increases and its cost decreases, strain-level bacterial differentiation will be decided at the genome level, rather than by expressed proteins. While the specificity required for strain-level typing may remain

the purview of phylogenetics, the use of mass spectrometry to track intact protein biomarkers at a serovar level would provide a cheaper, inherently multiplexed screen to determine the value of genetic sequencing. LC-MS/MS analysis not only supplies the detectable masses that differ between two samples (within the upper mass limit of the mass spectrometer) but also the identity of those masses. Knowledge of which gene products contain SNPs or which proteins have been newly transferred to a bacterial strain provides a direct link back to genome sequencing data, providing gene-specific validation at an expressed protein level.

The rapid rate of bacterial evolution translates to a moving target for strain and serovar-differentiating SNP-containing proteins. Any method meant to differentiate across multiple serovars would require a combination of multiple SNP-containing proteins. The advantage of nontargeted expression profiles generated in the method presented here is that any unpredicted changes that occur in the most abundant soluble proteins should be detected. Target marker proteins do not need to be known before sample analysis.

Identification of SNP-containing proteins becomes much quicker once initial identification of the most abundant expression profile masses has been established. Because the majority of the most abundant proteins are conserved across bacterial intact protein expression profiles of *Salmonella* serovars, it is not necessary to identify hundreds of proteins in each new isolate. Most abundant masses can be identified by matching the accurate mass and retention time to existing data from a reference strain. Only the compounds that exhibit a mass difference as compared to a standard strain may need to be analyzed by MS/MS for identity confirmation. This small subset of SNP-containing proteins can then be used to query the rapidly growing number of bacterial genomes as a gene name and intact mass (or mass difference) pair. Instead of comparing each new bacterial expression profile to a mass spectral data repository, we can take advantage of bacterial sequencing and alignment efforts and query for only the expressed proteins that show a change in mass. This targeted analysis would be quicker than whole-genome sequencing and more likely to detect genetic changes than multiplexed PCR or targeted mass spectrometry alone because the biomarkers do not need to be known in advance.

Acknowledgments The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the US Food and Drug Administration.

References

- Arnold RJ, Reilly JP. Observation of *Escherichia coli* ribosomal proteins and their posttranslational modifications by mass spectrometry. *Anal Biochem.* 1999;269(1):105–12.
- Bell RL, Gonzalez-Escalona N, Stones R, Brown EW. Phylogenetic evaluation of the ‘Typhimurium’ complex of *Salmonella* strains using a seven-gene multi-locus sequence analysis. *Infect Genet Evol.* 2011;11(1):83–91.
- Bizzini A, Greub G. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. *Clin Microbiol Infect.* 2010;16(11):1614–9.

- Cargile BJ, McLuckey SA, Stephenson JL. Identification of bacteriophage MS2 coat protein from *E-coli* lysates via ion trap collisional activation of intact protein ions. *Anal Chem.* 2001;73(6):1277–85.
- CDC. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet Surveillance Report for 2012 (final report) (trans: U.S. Department of Health and Human Services C). CDC, Atlanta, GA; 2014.
- Clark AE, Kaleta EJ, Arora A, Wolk DM. Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clin Microbiol Rev.* 2013;26(3):547–603.
- Conway GC, Smole SC, Sarracino DA, Arbeit RD, Leopold PE. Phyloproteomics: species identification of Enterobacteriaceae using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J Mol Microbiol Biotechnol.* 2001;3(1):103–12.
- Dieckmann R, Helmuth R, Erhard M, Malorny B. 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.* 2008;74(24):7767–78.
- Everley RA, Mott TM, Wyatt SA, Toney DM, Croley TR. Liquid chromatography/mass spectrometry characterization of *Escherichia coli* and *Shigella* species. *J Am Soc Mass Spectrom.* 2008;19(11):1621–8.
- Fagerquist CK, Bates AH, Heath S, King BC, Garbus BR, Harden LA, Miller WG. Sub-speciating *Campylobacter jejuni* by proteomic analysis of its protein biomarkers and their posttranslational modifications. *J Proteome Res.* 2006;5(10):2527–38.
- FDA. Bad Bug Book: foodborne pathogenic microorganisms and natural toxins. 2nd ed.; 2012. pp. 9–13.
- Fenselau C, Demirev PA. Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrom Rev.* 2001;20(4):157–71.
- Ho YP, Hsu PH. Investigating the effects of protein patterns on microorganism identification by high-performance liquid chromatography-mass spectrometry and protein database searches. *J Chromatogr A.* 2002;976(1–2):103–11.
- Krishnamurthy T, Ross PL. Rapid identification of bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells. *Rapid Commun Mass Spectrom.* 1996;10(15):1992–6.
- Krishnamurthy T, Davis MT, Stahl DC, Lee TD. Liquid chromatography microspray mass spectrometry for bacterial investigations. *Rapid Commun Mass Spectrom.* 1999;13(1):39–49.
- Lee SW, Berger SJ, Martinovic S, Pasa-Tolic L, Anderson GA, Shen YF, Zhao R, Smith RD. Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. *Proc Natl Acad Sci U S A.* 2002;99(9):5942–7.
- Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao GJ, Meng JH, Stones R. Identification of a salmonellosis outbreak by means of molecular sequencing. *N Engl J Med.* 2011;364(10):981–2.
- McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal Chem.* 1997;69(4):767–76.
- McFarland MA, Andrzejewski D, Musser SM, Callahan JH. Platform for identification of *Salmonella* serovar differentiating bacterial proteins by top-down mass spectrometry: *S. Typhimurium* vs *S. Heidelberg*. *Anal Chem.* 2014;86(14):6879–86.
- Meng FY, Cargile BJ, Miller LM, Forbes AJ, Johnson JR, Kelleher NL. Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat Biotechnol.* 2001;19(10):952–7.
- Mott TM, Everley RA, Wyatt SA, Toney DM, Croley TR. Comparison of MALDI-TOF/MS and LC-QTOF/MS methods for the identification of enteric bacteria. *Int J Mass spectrom.* 2010;291(1–2):24–32.
- Ryzhov V, Fenselau C. Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Anal Chem.* 2001;73(4):746–50.

- Wattiau P, Boland C, Bertrand S. Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Appl Environ Microbiol.* 2011;77(22):7877–85.
- Wilcox SK, Cavey GS, Pearson JD. Single ribosomal protein mutations in antibiotic-resistant bacteria analyzed by mass spectrometry. *Antimicrob Agents Chemother.* 2001;45(11):3046–55.
- Williams TL, Leopold P, Musser S. Automated postprocessing of electrospray LC/MS data for profiling protein expression in bacteria. *Anal Chem.* 2002;74(22):5807–13.
- Williams TL, Andrzejewski D, Lay JO, Musser SM. Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. *J Am Soc Mass Spectrom.* 2003;14(4):342–51.
- Williams TL, Musser SM, Nordstrom JL, DePaola A, Monday SR. Identification of a protein biomarker unique to the pandemic O3: K6 clone of *Vibrio parahaemolyticus*. *J Clin Microbiol.* 2004;42(4):1657–65.
- Williams TL, Monday SR, Edelson-Mammel S, Buchanan R, Musser SM. A top-down proteomics approach for differentiating thermal resistant strains of *Enterobacter sakazakii*. *Proteomics.* 2005;5(16):4161–9.
- Wynne C, Edwards NJ, Fenselau C. Phyloproteomic classification of unsequenced organisms by top-down identification of bacterial proteins using capLC-MS/MS on an Orbitrap. *Proteomics.* 2010;10(20):3631–43.
- Zamdborg L, LeDuc RD, Glowacz KJ, Kim Y-B, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* 2007;35:W701–6.