analytical chemistry

Article

pubs.acs.org/ac

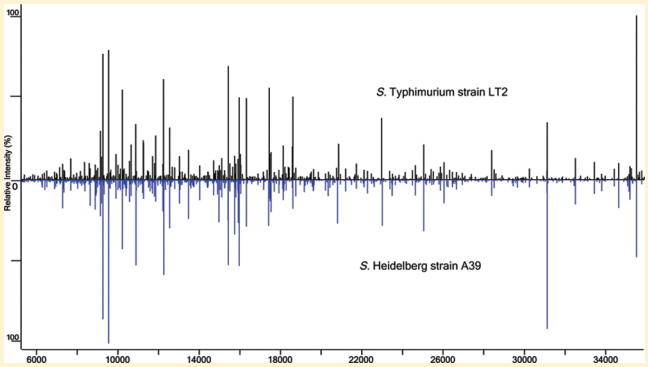
Platform for Identification of *Salmonella* Serovar Differentiating Bacterial Proteins by Top-Down Mass Spectrometry: *S.* Typhimurium vs *S.* Heidelberg

Melinda A. McFarland,* Denis Andrzejewski, Steven M. Musser, and John H. Callahan

Spectroscopy and Mass Spectrometry Branch, Center for Food Safety and Applied Nutrition, United States Food and Drug Administration, 5100 Paint Branch Parkway, College Park, Maryland 20740, United States

Supporting Information

ABSTRACT: Intact protein expression profiling has proven to be a powerful tool for bacterial subspecies differentiation. To facilitate typing, epidemiology, and trace-back of Salmonella contamination in the food supply, a minimum of serovar level differentiation is required. Subsequent identification and validation of marker proteins is integral to rapid screening development and to determining which proteins are subject to environmental pressure. Bacterial sequencing efforts have expanded the number of sequenced genomes available for single-nucleotide polymorphism (SNP) analyses, but annotation is often missing, start site errors are not uncommon, and the likelihood of expression is not known. In this work we show that the combination of intact protein expression profiles and top-down liquid chromatography−mass spectrometry (LC−MS/MS) facilitates the identification of proteins that result from expressed serovar specific nonsynonymous SNPs. Combinations of these marker proteins can be used in assays for rapid differentiation of bacteria. LC−MS generated intact protein expression profiles establish which bacterial protein masses differ across samples and can be determined without prior knowledge of the sample. Subsequent top-down LC−MS/MS is used to identify expressed proteins and their post-translational modifications (PTM), identify serovar specific markers, and validate genomic predicted orthologues as expressed biomarkers.



Identification of bacterial species is a moving target. The generation time of bacteria in its host is less than 24 h, and a single mutation or convergence can make a drastic difference in virulence or expression. Food safety efforts require serovar and strain level specificity for trace-back of bacterial contamination to its source. Detection methods that require selection of probe based assays are limited by probe selection. A nontargeted mass spectrometry based method provides a high-throughput and relatively unbiased snapshot of the expressed proteins in a wide range of bacterial samples and is amenable to both screening and targeted analysis. Such an inherently multiplexed technique facilitates differentiation of closely related bacteria as well as the detection of unsequenced or newly acquired nonsynonymous single-nucleotide polymorphisms (SNPs) and plasmid proteins that may be specific to a given strain.

Members of the *Salmonella enterica enterica* subspecies are the cause of most human salmonellosis. In the United States, most cases are foodborne. *S. enterica enterica* consists of more than 2500 different O and H cell surface antigen combinations or serovars.[1] *S. enterica* serovar Typhimurium and *S. enterica* serovar Heidelberg are among the top 10 serovars implicated in foodborne *Salmonella* infections.[2] Recent phylogenetic and multilocus sequence tag (MLST) analyses[3] confirm that the chosen strains are members of two-closely related serovars.

Species and subspecies level assays are generally adequate for clinical diagnostics. However, localization of the source of a foodborne contamination requires serovar or strain level specificity. Pulsed field gel electrophoresis (PFGE) has become the gold standard for molecular subtyping of *Salmonella*, and PCR based assays built around genomic markers are becoming increasingly popular.[4] However, differentiating between two highly similar serovars such as *S.* Typhimurium and *S.* Heidelberg requires multiple enzymes or PCR targets and relies on matching to a previously validated standard. Changes to untargeted genes and newly acquired genetic material are likely to be missed.

Intact protein mass spectrometry of bacterial lysates offers an inherently multiplexed measure of the mass of soluble proteins in their intact state at a given growth stage.[5,6,7] This is particularly useful because bacteria exhibit fewer overall 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, particularly within the same phyla. Consequently, while protein

Received: February 4, 2014
Accepted: June 4, 2014
Published: June 4, 2014

abundances may vary from serovar to serovar, there should be very little difference in their masses. Therefore, for bacterial lysates it is a reasonable assumption that the minimal mass shifts found between closely related bacteria (strain or serovar level) are the result of SNPs (henceforth this term will be used to mean nonsynonymous or nonsilent SNPs), and novel masses may indicate plasmid insertions or proteins that have undergone a significant change in expression level.[8−10] These mass shifted proteins serve as biomarkers for differentiation of bacteria, here at the serovar level.

Intact protein mass spectrometry has become a commercially available tool for clinical bacterial differentiation based on matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) technology,[11,12]. These products rely on spectral library searches. While this method is faster than LC−MS/MS, commercial libraries offer a limited mass range of 2−20 kDa, typically less than 15 kDa. In a complex bacterial sample, ribosomal proteins are preferentially observed with MALDI ionization. This bias toward small ribosomal proteins[13] often limits utility to species and subspecies identification, although serovar typing examples have been shown on noncommercial platforms.[14] The increased mass range, sample to sample reproducibility, and greater number of proteins ionized using an electrospray ionization (ESI) based platform provide access to a more diverse range of proteins, potentially providing greater specificity for differentiation of closely related bacteria.[15−17] This approach has already been used to identify marker masses that differentiate thermophilic vs nonthermophilic groups of *Cronobacter sakazakii*,[18] to identify proteins characteristic of specific outbreak strains of *V. parahaemolyticus* and guide the development of PCR probes[19] and to differentiate closely related species within the enterobacteriaceae family.[17,20] In addition, the approach has been shown to quantify protein expression differences by using certain housekeeping proteins as internal standards.[21]

The addition of online "top-down" MS/MS fragmentation of the intact proteins subsequently provides identification of the proteins (genes) containing said mass differences.[22−25] 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 nonsequenced species, determine if a mass shift in a specific protein represents a novel, and possibly virulent, mutation. This information provides a direct link back to genome sequencing data, providing gene specific marker and sequence validation at an expressed protein level. Because the majority of proteins are conserved across bacterial intact protein expression profiles of *Salmonella* serovars, the accurate mass and retention times can be used to identify proteins in subsequent analyses. Only the masses that differ as compared to a standard strain will need to be reanalyzed by MS/MS to confirm identity, making the time frame of this experiment feasible as a screening platform.

In this article, we demonstrate a methodology for combining intact protein liquid chromatography (LC) ESI-MS with top-down LC−ESI-MS/MS to facilitate the identification of proteins that result from expressed serovar specific nonsynonymous SNPs. This approach is based on the power of inherently multiplexed deconvoluted ESI-MS generated intact protein expression profiles[26] to facilitate rapid differentiation between samples, combined with top-down protein identification for marker confirmation. We demonstrate the efficacy of this platform in the identification of strain differentiating proteins using two reference strains from closely related Salmonella serovars. A number of proteins, homologous but differing due to SNPs, can be easily identified in the two strains. On-the-fly identification of strain differentiating proteins would facilitate querying of bacterial genome repositories without prior selection of biomarker proteins. In addition, we show that information generated by this approach can be used to verify genomic information and demonstrate examples of misannotation due to start site errors. Knowledge of which protein sequences are variable across serovars provides a common link to genomic sequencing and phylogenetic strain typing efforts.

METHODS

Bacterial Strains. *Salmonella enterica enterica* serovar Typhimurium strain LT2 and *S.* Heidelberg strain A39 strains used in the study were obtained from the Food and Drug Administration/Center for Food Safety and Applied Nutrition stock culture collection. Both serovars are members of the Salmonella reference A collection (SARA).[27] Bacteria were grown for 24 h at 37 °C on LB agar plates (Teknova, Hollister, CA). Cell isolates were collected in a 1.5 mL Eppendorf tube, washed twice with sterile water, and resuspended in 0.5 mL of 70% ethanol to facilitate sterilization of bacteria[28] as well as minimize protease activity. All work prior to suspension in 70% ethanol was performed in a biosafety cabinet. The approximate cell concentration was 8×10^{10} cfu/mL. Bacterial cells were lysed within 24 h of harvesting.

Extraction of Cellular Proteins. The sample tube containing bacteria cells suspended in 70% ethanol was centrifuged at 9800*g* for 5 min. The ethanol solution was removed and 1.0 mL of a 50:49:1 extraction solution consisting of acetonitrile (J.T. Baker, Phillipsburg, N.J.), HPLC-grade water (J.T. Baker), and formic acid (Sigma-Aldrich Chemical Co., St. Louis, MO) was added. The suspension was transferred to a Barocycler FT500 pulse tube (Pressure Biosciences, Inc. Boston, MA) along with an additional 0.4 mL of extraction solution. Tubes were transferred to the Barocycler NEP 3229 pressure cycling device and pressure cycled 24 times starting at 35 000 psi for 15 s then 0 psi for 10 s at 44 °C. Pulse tube contents were transferred to a 1.5 mL tube and centrifuged at 9800*g* for 20 min to pellet the cellular debris. A portion of the supernatant was transferred to an autosampler tube for LC−MS analysis. The remaining lysate was stored at −24 °C.

HPLC of Intact Proteins. Intact proteins were separated by reversed-phase high performance liquid chromatography (HPLC) using an Agilent (Palo Alto, CA) 1100 system fitted with two ProSphere P-HR (W.R. Grace, MD) 2.1 mm i.d. × 15 cm columns connected in series for improved separation and held at 50 °C. A volume of 2 μL of protein extract were injected at a flow rate of 200 μL/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 to 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 obtain consistent retention times across platforms.

Mass Spectrometry. For LC−MS, the HPLC was interfaced to a Q-TOF Premier (Waters, Beverly, MA) mass spectrometer using positive ion electrospray ionization. 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. Total acquisition time was 90 min. Data

dx.doi.org/10.1021/ac500786s | Anal. Chem. 2014, 86, 6879−6886

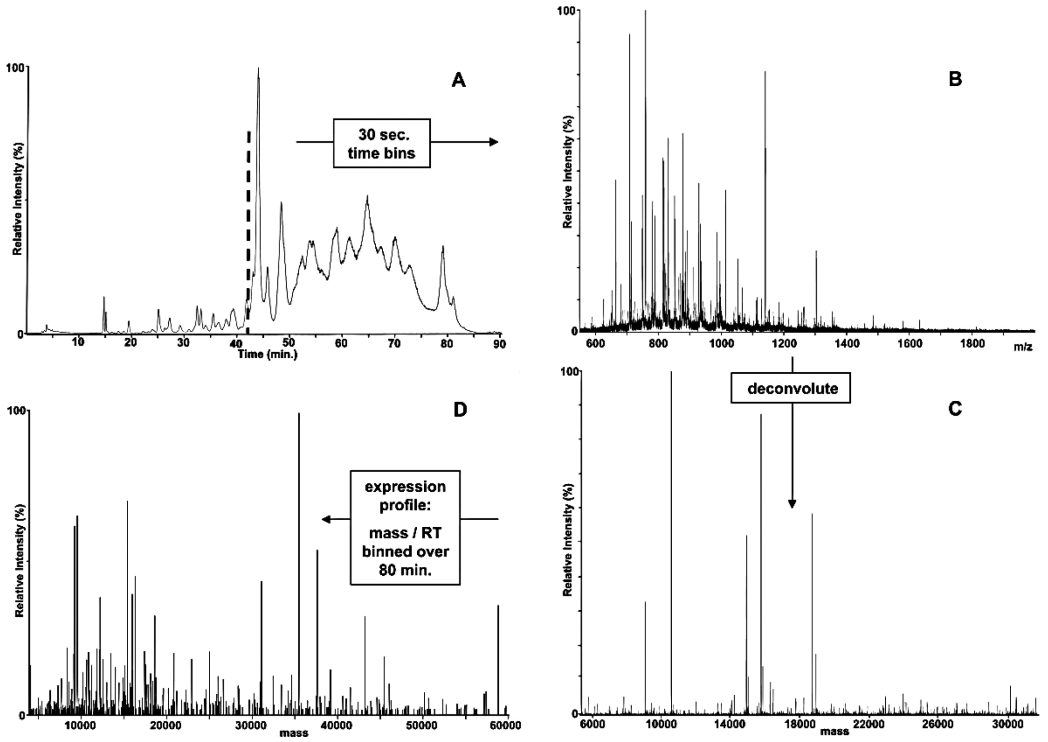


Figure 1. Intact protein expression profile generation. ProTrawler software was used to deconvolute and reconcile all MS scans from the chromatogram into a single mass, retention time, and abundance profile. (A) Representative chromatogram from an LC−MS analysis of an intact protein lysate of *S.* Typhimurium. (B) Mass spectra were summed into 30 s bins across the chromatogram. (C) The resultant spectra at each time interval were 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.

was collected using MassLynx software version 4.1 (Waters, Beverly, MA).

For LC−MS/MS, the HPLC was interfaced to an LTQ-Orbitrap XL (Thermo Fisher, San Jose, CA) mass spectrometer. The eluent flow was split to a flow rate of 350 nL/min via a TriVersa NanoMate (Advion BioSciences, Ithaca, NY) chip based nanospray source. The instrument was operated in 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 set to 15%. Each MS spectrum was the average of three microscans, and each MS/MS spectrum was the average of 10 microscans. Typical MS/MS acquisition time for 10 microscans was 12 s, with an average chromatographic peak width of 1.5 min. To facilitate the analysis of intact proteins, the instrument was operated with the higher-energy collisional dissociation (HCD) gas off and the delay before image current detection shortened to 5 ms. Identifications shown in this work are the result of four LC−MS/MS replicate analyses of each serovar.

Data Analysis. Automated analysis of full-scan (MS) Q-TOF data was performed with ProTrawler6 (previously named Retana), custom software (BioAnalyte, Inc., Portland, ME). A detailed explanation of the approach has been published.[26] Briefly, spectra are summed in 30 s windows. In version 6 of ProTrawler, the summed spectra from each time window are baseline subtracted and denoised using the proprietary ReSpect algorithm (Positive Probability, Shrewsbury, U.K.). The resultant spectrum is deconvoluted using maximum entropy deconvolution. After generating a protein mass/abundance list for each time window, ProTrawler bins the data for each time window, determines the time range over which a given mass is detected, and calculates an abundance-weighted time centroid for the mass, which is used as the retention time. Masses corresponding to charge state dimers and trimers (multimers) and sodium adducts are removed, and their corresponding intensity is folded back into the primary mass. A peak must be less abundant than its primary peak to be considered a multimer or adduct. Abundances are then normalized to the summed intensity. The resulting text file contains a cumulative list of all intact protein masses, abundances, and retention times, from which the mass and abundance information can be represented graphically as mass versus intensity, similar to a traditional mass spectrum. Retention time is included in the output to distinguish proteins of similar mass. At the time of this work, ProTrawler software is available only for Q-TOF Premier data. A new version applicable to multiple instruments and vendors is forthcoming.

For top-down data analysis, ProSightPC 2.0[29] was used to search MS/MS spectra against a 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). Each database contains the predicted sequence with and without initiator methionine and with and without signal peptides predicted at greater than 50% likelihood by SignalP.[30] Neutral mass deconvoluted precursor and fragment mass lists were generated with the Xtract algorithm (Thermo Fisher, San Jose, CA) option within ProSightPC 2.0. Precursor mass

dx.doi.org/10.1021/ac500786s | *Anal. Chem.* 2014, 86, 6879−6886

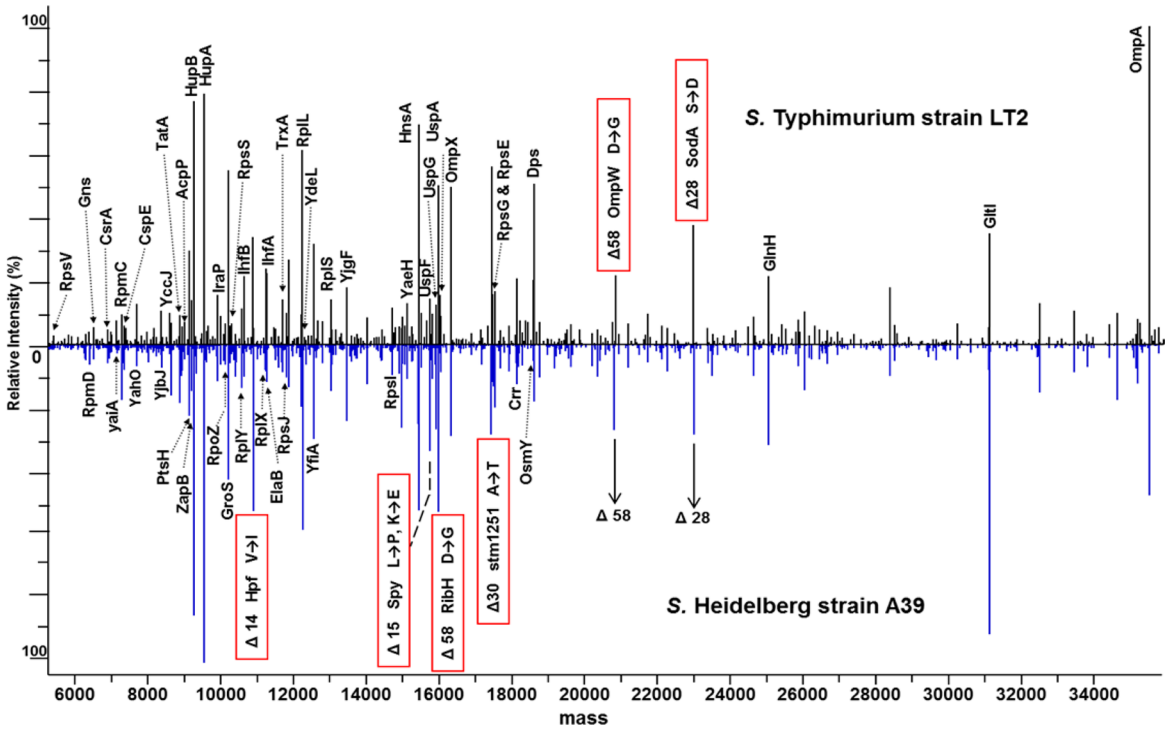


Figure 2. Comparison of intact protein expression profiles for *S.* Typhimurium strain LT2 and *S.* Heidelberg strain A39. A subset of identified proteins is labeled with protein names. Six highlighted masses contain serovar specific SNP related mass differences. Amino acid changes are noted from *S.* Typhimurium to *S.* Heidelberg.

tolerance was 1000 Da and fragment ion monoisotopic mass tolerance was 20 ppm. 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 reassignment of fragment ions and rescoring via the sequence gazer option in ProSightPC. Modifications were considered valid if there was an increase in the number of matched fragment ions upon inclusion of the predicted modification. A secondary search was also performed that included the most commonly detected PTMs as confirmation of the amended modification as the top-scoring identification. Only proteins identified with ProSight e-values better than 1×10^{-5} for a minimum of three MS/MS spectra were considered valid identifications. ProSight reports monoisotopic masses, however, centroid masses were used for tables to facilitate comparison to ProTrawler generated protein expression profiles acquired on the lower resolution Q-TOF Premier. Theoretical average masses were calculated using Protein Prospector (prospector.ucsf.edu).

RESULTS

Intact Protein Expression Profiles. The intact accurate mass, retention time, and relative abundance for proteins from the soluble fraction of bacterial lysates were measured and compared using LC−MS. In order to minimize sample loss and increase long-term reproducibility, sample preparation involves as little postculture sample manipulation as possible and no pre-MS fractionation other than solubilization in extraction buffer.[28] Figure 1A shows a representative total ion current chromatogram from an 80 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.[26] Unlike mass spectra of peptides, intact proteins produce broad charge state distributions, splitting the signal into multiple peaks (Figure 1B). The elution profile of each protein is 1.5 min wide on average, further distributing the signal, as well as greatly increasing the likelihood of multiple coeluting proteins. Consequently, software is necessary to deconvolute each spectrum (Figure 1C) and merge consecutive abundances into a single protein mass and intensity. The result (Figure 1D) 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 sensitivity and complexity provided by chromatographically resolved ESI spectra.

Figure 2 shows a mirrored comparison of the LC−MS generated intact protein expression profiles of *S.* Typhimurium strain LT2 and *S.* Heidelberg strain A39. Each profile is the result of deconvolution and binning of mass, abundance, and retention time from a representative LC−MS run. While masses of bacterial proteins are detectable up to 70 kDa with the Q-TOF, the spectral range has been truncated here to represent the mass range routinely achievable for MS/MS on the Orbitrap XL, which was the instrument used for top-down protein identification. Bacterial proteins were monitored out to 80 kDa using the Q-TOF. Masses greater than 40 kDa were detected with limited intensity and diminished reproducibility, making them poor candidates for nontargeted differential analysis of intact proteins.

For bacteria, visual analysis by expression profiles provides a more rapid and complete means for comparative analysis versus reliance on LC−MS/MS protein or peptide identifications.[20] As is expected by the extreme homology across Salmonella species and the similarity of these two serovars, the mass maps in Figure 2 appear nearly identical, with differences occurring in

dx.doi.org/10.1021/ac500786s | *Anal. Chem.* 2014, 86, 6879−6886

only a small number of detectable masses. The observed mass shifts likely represent protein products of SNP-containing genes that differentiate strain LT2 from strain A39 and, therefore, are potential biomarkers for serovar identification. Protein sequencing is not 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 by inspection[19,18,20] and to differentiate multiple bacteria by PCA or hierarchical cluster analysis.[31] While the presence of a differential pattern is sufficient for grouping a serovar with a set of previously analyzed samples, it does not readily facilitate identification of uncharacterized strains and often provides insufficient data to link the result with complementary assays such as targeted PCR probes or genome sequencing. The second stage of this method is the addition of online top-down MS/MS identification of proteins to the existing LC−MS separation. Proteins maintain the same elution profile but now the most abundant proteins are identified. Advances in ProTrawler deconvolution software to include higher resolution instrumentation will facilitate LC−MS protein expression profiling and MS/MS identification of marker proteins on a single platform.

Protein identifications in Figure 2 are represented by protein name, as assigned for the reference genome of *S.* Typhimurium strain LT2. A complete list of protein identifications and post-translational modifications can be found in supplemental Table 1 in the Supporting Information. No post-translational modifications were included in the initial search of MS/MS spectra against the corresponding database. Instead, the presence of modifications was inferred from the mass difference between the theoretical and measured masses. This approach was taken to minimize arbitrary assignment of N-terminal modifications. Most mass differences corresponded to common modifications of N-terminal methylation, formylation, and acetylation (supplemental Table 1 in the Supporting Information). Modifications were validated by performing a protein specific assignment of fragment ions. In all cases the number of b-type fragment ions increased upon inclusion of the modified protein N-terminus in the database search. In addition, a secondary database search was performed that included the most commonly inferred PTMs as confirmation of the amended modification as the top-scoring identification. The proposed PTMs detected in this work were observed in both serovars.

Most observed masses show no discernible mass difference between the two Salmonella strains analyzed. Because we are able to readily identify the most abundant masses by top-down MS/MS fragmentation, we can confirm that those proteins that do produce serovar specific mass shifts between *S.* Typhimurium and *S.* Heidelberg are indeed products of the same gene. Interestingly, none of the SNP-containing proteins found in this work contains N-terminal modifications other than initiator methionine or signal peptide cleavage. Alignment of the in-silico predicted protein sequences confirms the presence of an amino acid change resulting from a nonsynonymous SNP (Figure S-1 in the Supporting Information).

A representative CID MS/MS spectra for the $[M + 21H]^{21+}$ ion of a 17.4 kDa putative small heat shock protein (loci stm_1251) fragmented in the linear ion trap and detected at high resolution in the Orbitrap is shown in Figure 3. Here the

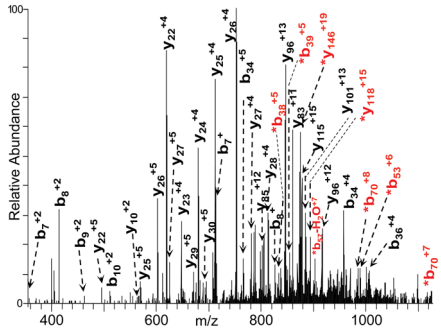


Figure 3. Representative CID spectrum. Top-down LC-MS/MS CID spectrum of the 21+ charge state of a putative molecular chaperone (Q8ZPY6_SALTY, loci stm1251) identified in *S.* Typhimurium. The measured molecular weight in *S.* Typhimurium strain LT2 is 17 420 Da and 17 451 Da in *S.* Heidelberg strain A39. Fragments with an asterix contain a 31 Da mass shift corresponding to the serovar specific single amino acid difference.

SNP is evident in mass shifts of a subset of fragment ions but the dominant fragment ions do not differ. More commonly, the SNP is contained in the middle of the protein and less amenable to site specific fragmentation. However, because we simultaneously detect the mass of the intact protein, we can rely on accurate mass and retention time profiles to confirm that the identified protein does indeed contain a mass shift.

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 genomic sequencing errors. Protein ElaB identified in *S.* Typhimurium strain LT2 has a theoretical mass 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 (Figure 4A). 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 found in Heidelberg strain A39 does reconcile with its measured mass (after cleavage of the initiator methionine), strongly suggesting 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 (Figure 4B). Confirmation of a sequencing start site error can be seen in Figure 4C. Removal of the erroneous amino acids (MR) and cleavage of the remaining N-terminal methionine 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. Two sets of complementary b-type and y-type fragment ions cover the entire length of the protein and help validate the corrected N-terminus. Sequencing errors in the middle of proteins are more difficult to confirm but likely are present. In this data set, start site errors and unannotated signal peptides were the most common annotation flaws. Identification of protein sequences combined with an intact mass measurement provides a unique link to genomic sequencing and phylogenetic strain typing efforts.

dx.doi.org/10.1021/ac500786s | *Anal. Chem.* 2014, 86, 6879−6886

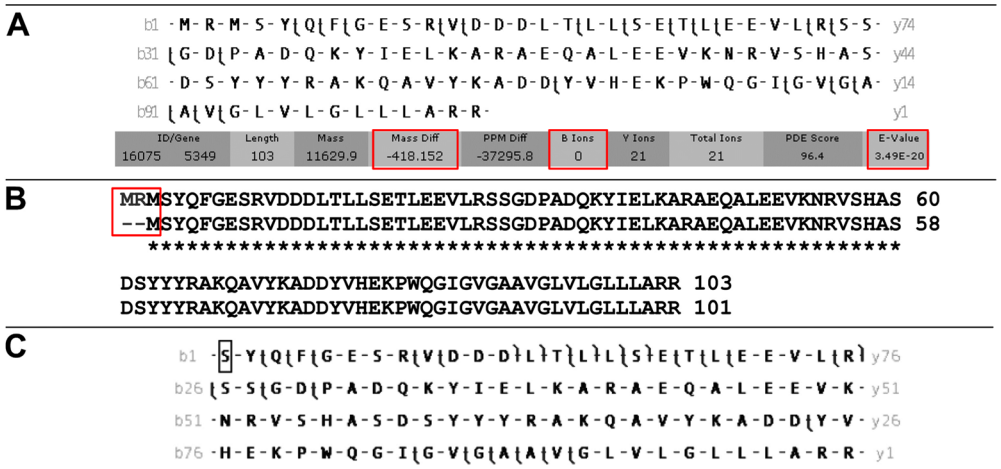


Figure 4. Top-down mass spectrometry to verify genome annotation. (A) The measured mass for protein ElaB in *S.* Typhimurium strain LT2 does not agree with the theoretical mass predicted by genomic sequencing. Top-down MS/MS identifies the correct protein but no b-type fragment ions are assigned. (B) Comparison of the predicted protein sequence against other strains shows disagreement at the N-terminus. (C) Correction of the start site results in identification of a substantial sequence tag of b-type ions.

DISCUSSION

The power of intact protein analysis is that the mass of the protein is measured with functional modifications intact. This is ideally suited to bacterial proteins because, unlike mammalian systems, bacterial lysates from similar species appear to exhibit highly reproducible post-translational modifications under similar growth conditions. The modifications observed in this work are highly conserved across enterobacteria, which facilitates validation of modifications by comparisons to better annotated reference proteomes, such as *Escherichia coli*. While 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 the small number of mass shifts found across serovars are SNPs, and novel masses are insertions or proteins that have undergone a significant change in expression level. These shifts in mass serve as markers for differentiation of bacteria at the serovar level.

As the speed of full genome sequencing increases and its cost decreases, strain level bacterial differentiation will increasingly 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 less expensive, inherently multiplexed screen to determine the accuracy of genetic sequencing. LC–MS/MS analysis reveals both the detectable masses that differ between two samples and 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 wide precursor mass tolerance permissible when searching top-down mass spectra also lends itself to validation and correction of genomic sequencing start sites. Ideally, top-down MS/MS spectra searched against its corresponding sequenced genome should have measured intact masses in agreement with the translated theoretical mass (excluding post-translational modifications). Realistically, a subset of proteins exhibit sequence errors that result from the selection of erroneous genomic sequencing start sites. These start site errors can be confirmed by reassigning MS/MS fragment ions based on the protein sequence beginning at an alternative start site. As the use of high throughput genome sequencing annotation pipelines increases, validation of start sites will minimize the propagation of start site errors through multiple genomes.

The extreme homology across enterobacteria strongly suggests that for a top-down analysis we do not need strain specific fully sequenced genomes to identify SNP-containing bacterial proteins.[25,32] While 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. A fragment ion mass accuracy requirement of 20 ppm[33] provides sufficient specificity to identify sequence tags without an exact precursor mass. Consequently, one can confidently identify enough fragment ions to identify a MS/MS spectrum to a homologous protein while still maintaining the intact mass of the protein. Comparison of the measured intact mass with that of the identified protein readily determines whether the measured protein contains a mass shift. This theory was tested by searching the data from each serovar against the protein database of the other serovar (data not shown). Greater than 90% of the proteins shown here were identified when searched against a different but related protein database (here, *S.* Heidelberg strain A39 data searched against *S.* Typhimurium strain LT2 database). The difference between the measured mass and database mass can be used to assess the likelihood of amino acid differences or post-translational modifications. This approach is particularly helpful in assessing N-terminal modifications, such as formylation and signal peptide cleavage.

Admittedly, the rapid rate of bacterial evolution translates to a moving target for strain and serovar differentiating SNP-containing proteins. The SNPs confirmed in this work (Figure 2) tentatively differentiate *Salmonella enterica* serovar Typhimurium from serovar Heidelberg. All *S.* Typhimurium and *S.* Heidelberg genomes found in the Uniprot knowledge base show 100% alignment within the marker containing sequences used in this work (18–20 *S.* Typhimurium genomes and 11 *S.* Heidelberg genomes, Figure S-2 in the Supporting Information). However, it is not possible to include all existing

dx.doi.org/10.1021/ac500786s | *Anal. Chem.* 2014, 86, 6879–6886

salmonella strains in this analysis. Certainly any method meant to differentiate across multiple serovars would require a combination of multiple SNP-containing proteins. The advantage of the nontargeted intact protein expression profiles generated in the method presented here is that differences as compared to the most abundant soluble proteins in a reference strain should be detected. Consequently, an unknown or unsequenced sample would be amenable to this method.

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 proteins are conserved across intact bacterial protein expression profiles of Salmonella serovars, it is not necessary to identify hundreds of proteins in each new isolate. An accurate mass and retention time match to a previously characterized standard strain will identify most abundant masses. Only the masses that differ 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 against 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 instead 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 full 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.

CONCLUSION

Bacterial sequencing efforts have expanded the number of sequenced genomes available for SNP analyses, but annotation is often missing, start site errors are not uncommon, and the likelihood of expression is not known. In this work we have shown that the combination of LC−ESI-MS generated intact protein expression profiles and top-down LC−ESI-MS/MS facilitates the identification of proteins that result from serovar specific single-nucleotide polymorphisms (SNPs). Combinations of these marker proteins can be used in assays for rapid differentiation of bacteria. LC−MS generated intact protein expression profiles establish which bacterial protein masses differ across samples and can be determined without prior knowledge of the sample and without prior selection of marker proteins. Subsequent top-down LC−MS/MS is used to identify expressed proteins and their post-translational modifications (PTM), identify serovar specific markers, and validate genomically predicted orthologues as expressed biomarkers. Because the mass of the intact protein is measured, protein identification can be made without access to a serovar specific genome. In addition, the resultant data can be used to validate genomic sequencing annotation.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: melinda.mcfarland@fda.hhs.gov.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Rebecca Bell for culture and harvest of bacteria and Errol Strain and Robert Stones for informatics consultation. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Food and Drug Administration.

REFERENCES

(1) U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition. *Bad Bug Book: Foodborne Pathogenic Microorganisms and Natural Toxins*, 2nd ed.; U.S. Government Printing Office: Washington, DC, 2012; pp 9−13.

(2) U.S. Department of Health & Human Services, Centers for Disease Control and Prevention. *Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet Surveillance Report for 2011*; U.S. Department of Health & Human Services: Atlanta, GA, 2012; p 19.

(3) Bell, R. L.; Gonzalez-Escalona, N.; Stones, R.; Brown, E. W. *Infect. Genet. Evol.* **2011**, *11* (1), 83−91.

(4) Wattiau, P.; Boland, C.; Bertrand, S. *Appl. Environ. Microbiol.* **2011**, *77* (22), 7877−7885.

(5) Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996**, *10* (15), 1992−1996.

(6) Fenselau, C.; Demirev, P. A. *Mass Spectrom. Rev.* **2001**, *20* (4), 157−171.

(7) Conway, G. C.; Smole, S. C.; Sarracino, D. A.; Arbeit, R. D.; Leopold, P. E. *J. Mol. Microbiol. Biotechnol.* **2001**, *3* (1), 103−112.

(8) Wilcox, S. K.; Cavey, G. S.; Pearson, J. D. *Antimicrob. Agents Chemother.* **2001**, *45* (11), 3046−3055.

(9) Dieckmann, R.; Helmuth, R.; Erhard, M.; Malorny, B. *Appl. Environ. Microbiol.* **2008**, *74* (24), 7767−7778.

(10) Arnold, R. J.; Reilly, J. P. *Anal. Biochem.* **1999**, *269* (1), 105−112.

(11) Bizzini, A.; Greub, G. *Clin. Microbiol. Infect.* **2010**, *16* (11), 1614−1619.

(12) Clark, A. E.; Kaleta, E. J.; Arora, A.; Wolk, D. M. *Clin. Microbiol. Rev.* **2013**, *26* (3), 547−603.

(13) Ryzhov, V.; Fenselau, C. *Anal. Chem.* **2001**, *73* (4), 746−750.

(14) Dieckmann, R.; Malorny, B. *Appl. Environ. Microbiol.* **2011**, *77* (12), 4136−4146.

(15) Krishnamurthy, T.; Davis, M. T.; Stahl, D. C.; Lee, T. D. *Rapid Commun. Mass Spectrom.* **1999**, *13* (1), 39−49.

(16) Ho, Y. P.; Hsu, P. H. *J. Chromatogr., A* **2002**, *976* (1−2), 103−111.

(17) Mott, T. M.; Everley, R. A.; Wyatt, S. A.; Toney, D. M.; Croley, T. R. *Int. J. Mass Spectrom.* **2010**, *291* (1−2), 24−32.

(18) Williams, T. L.; Monday, S. R.; Edelson-Mammel, S.; Buchanan, R.; Musser, S. M. *Proteomics* **2005**, *5* (16), 4161−4169.

(19) Williams, T. L.; Musser, S. M.; Nordstrom, J. L.; DePaola, A.; Monday, S. R. *J. Clin. Microbiol.* **2004**, *42* (4), 1657−1665.

(20) Everley, R. A.; Mott, T. M.; Wyatt, S. A.; Toney, D. M.; Croley, T. R. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (11), 1621−1628.

(21) Williams, T. L.; Callahan, J. H.; Monday, S. R.; Feng, P. C. H.; Musser, S. M. *Anal. Chem.* **2004**, *76* (4), 1002−1007.

(22) Cargile, B. J.; McLuckey, S. A.; Stephenson, J. L. *Anal. Chem.* **2001**, *73* (6), 1277−1285.

(23) Lee, S. W.; Berger, S. J.; Martinovic, S.; Pasa-Tolic, L.; Anderson, G. A.; Shen, Y. F.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (9), 5942−5947.

(24) Fagerquist, C. K.; Bates, A. H.; Heath, S.; King, B. C.; Garbus, B. R.; Harden, L. A.; Miller, W. G. *J. Proteome Res.* **2006**, *5* (10), 2527−2538.

(25) Wynne, C.; Edwards, N. J.; Fenselau, C. *Proteomics* **2010**, *10* (20), 3631−3643.

(26) Williams, T. L.; Leopold, P.; Musser, S. *Anal. Chem.* **2002**, *74* (22), 5807−5813.

(27) Beltran, P.; Plock, S. A.; Smith, N. H.; Whittam, T. S.; Old, D. C.; Selander, R. K. *J. Gen. Microbiol.* **1991**, *137*, 601−606.
(28) Williams, T. L.; Andrzejewski, D.; Lay, J. O.; Musser, S. M. *J. Am. Soc. Mass Spectrom.* **2003**, *14* (4), 342−351.
(29) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y.-B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. *Nucleic Acids Res.* **2007**, *35*, W701−W706.
(30) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. *J. Mol. Biol.* **2004**, *340* (4), 783−795.
(31) Callahan, J. H.; McFarland, M. A.; Musser, S. M.; Bell, R.; Andrzejewski, D. *Abstr. Pap. Am. Chem. Soc.* **2009**, *238*.
(32) Wynne, C.; Fenselau, C.; Demirev, P. A.; Edwards, N. *Anal. Chem.* **2009**, *81* (23), 9633−9642.
(33) Meng, F. Y.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L. *Nat. Biotechnol.* **2001**, *19* (10), 952−957.