Journal of proteome research

pubs.acs.org/jpr

Article

A Proteogenomic Resource Enabling Integrated Analysis of *Listeria* Genotype−Proteotype−Phenotype Relationships

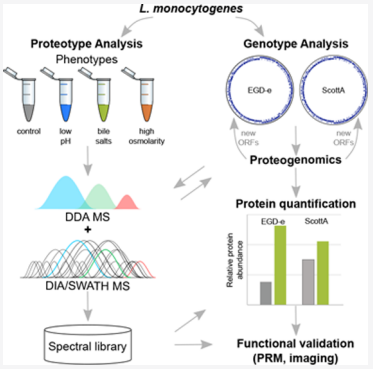
Adithi R. Varadarajan, Sandra Goetze, Maria P. Pavlou, Virginie Grosboillot, Yang Shen, Martin J. Loessner, Christian H. Ahrens,* and Bernd Wollscheid*

Cite This: *J. Proteome Res.* 2020, 19, 1647−1662

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: *Listeria monocytogenes* is an opportunistic foodborne pathogen responsible for listeriosis, a potentially fatal foodborne disease. Many different *Listeria* strains and serotypes exist, but a proteogenomic resource that bridges the gap in our molecular understanding of the relationships between the *Listeria* genotypes and phenotypes via proteotypes is still missing. Here, we devised a next-generation proteogenomics strategy that enables the community to rapidly proteotype *Listeria* strains and relate this information back to the genotype. Based on sequencing and *de novo* assembly of the two most commonly used *Listeria* model strains, EGD-e and ScottA, we established two comprehensive *Listeria* proteogenomic databases. A genome comparison established core- and strain-specific genes potentially responsible for virulence differences. Next, we established a DIA/SWATH-based proteotyping strategy, including a new and robust sample preparation workflow, that enables the reproducible, sensitive, and relative quantitative measurement of *Listeria* proteotypes. This reusable and publicly available DIA/SWATH library covers 70% of open reading frames of *Listeria* and represents the most extensive spectral library for *Listeria* proteotype analysis to date. We used these two new resources to investigate the *Listeria* proteotype in states mimicking the upper gastrointestinal passage. Exposure of *Listeria* to bile salts at 37 °C, which simulates conditions encountered in the duodenum, showed significant proteotype perturbations including an increase of FlaA, the structural protein of flagella. Given that *Listeria* is known to lose its flagella above 30 °C, this was an unexpected finding. The formation of flagella, which might have implications on infectivity, was validated by parallel reaction monitoring and light and scanning electron microscopy. *flaA* transcript levels did not change significantly upon exposure to bile salts at 37 °C, suggesting regulation at the post-transcriptional level. Together, these analyses provide a comprehensive proteogenomic resource and toolbox for the *Listeria* community enabling the analysis of *Listeria* genotype−proteotype−phenotype relationships.



KEYWORDS: *Listeria, proteotype, proteogenomics, smORFs, DIA, PRM, EGD-e, ScottA*

INTRODUCTION

Listeria monocytogenes is a highly adaptable environmental bacterium that can exist both as plant saprophyte and as animal pathogen.[1] The Gram-positive, rod-shaped, facultative anaerobic bacterium is the causative agent of listeriosis.[2,3] Although the incidence of listeriosis is relatively low compared to other common foodborne diseases, it is associated with one of the highest mortality rates (20−30%).[4] *Listeria* strains are categorized into at least 14 serotypes;[5] among these, three (1/2a, 1/2b, and 4b) are responsible for the majority of clinical cases. EGD-e, a widely used model system of serotype 1/2a, is the serotype most frequently recovered from foods or food-processing plants. In contrast, ScottA is a widely used model system of serotype 4b, which causes the majority of human epidemics.[5] Infection by *L. monocytogenes* usually occurs after digestion of contaminated foods and in individuals with impaired cell-mediated immunity. The elderly, immunosuppressed patients, pregnant women, and neonates are particularly susceptible.[2] An infection may lead to meningitis, sepsis, or, by crossing the placenta, infection of the fetus and subsequent abortion.[5] Upon ingestion, *L. monocytogenes* must resist multiple stresses encountered in the gastrointestinal (GI) tract, including variation in pH, osmolarity, and bile salts.[6,7] Notably, survival in the GI tract is a prerequisite to establish a successful infection in the host.[8]

L. monocytogenes has served as a key bacterial model system to study host pathogen interaction,[9] and studies of *L. monocytogenes* have led to the discovery of several new concepts in biology.[10] These included the discovery of unconventional mechanisms regulating bacterial gene expression, including the first RNA thermosensor regulating virulence,[11] the excludon concept,[12] and the discovery of an atypical member of the CRISPR family devoid of *cas* genes.[10]

Received: December 11, 2019
Published: February 24, 2020

Additionally, research on *L. monocytogenes* has contributed to a better understanding of the structure and dynamics of the host cell cytoskeleton with the discovery of the first actin nucleator in eukaryotic cells (the Arp2/3 complex)[13,14] and the elucidation of a novel role for clathrin in actin polymerization.[10] Finally, analysis of *L. monocytogenes* has been instrumental in characterizing naive-to-memory CD8 T cell generation and differentiation.[15]

The number of 'omics data sets that study different aspects of *L. monocytogenes* biology has increased exponentially in recent years. These data sets include a growing number of Illumina-based fragmented genomes[16] as well as complete genome sequences,[17] the latter of which provide an optimal basis for comparative genomics and functional genomics studies. Such comparisons have helped to identify *Listeria* virulence factors and regions associated with pathogenicity.[18,19] Moreover, analyses of complete genomes enabled detailed investigations of genes transcribed under clinically relevant conditions[12] and enabled identification of novel protein coding genes and the correct protein N-termini/start sites through proteogenomics.[20] Yet, changes in transcript abundances upon perturbation often do not correlate with abundance changes of the corresponding protein products,[21,22] and transcriptional data alone does not provide important functional information about post-transcriptional regulation such as protein modifications[23] or changes in protein interaction networks or cell-surface remodeling of the host following an infection.[24] Such information can, however, be obtained using state-of-the-art proteotype profiling.

The proteotype is defined as the state of the proteome at a particular time.[25] A proteome therefore consists of many proteotypes. The proteotype concept takes the dynamic nature of the proteome into account and extends it to the organization of proteins and their coexisting proteoforms in time and space.[25] Mass spectrometry-based proteotype profiling has matured recently through technological and methodological advances allowing for increased depth of proteome coverage[26−29] and sample throughput. This launches the next generation (next-gen) proteomics era of comprehensive and quantitative proteotype profiling. Recognizing the unmet need to integrate these data sets and to enable meta-analysis in a user-friendly manner, pioneers in the *Listeria* field developed the interactive Listeriomics Web site (https://listeriomics.pasteur.fr). At the time of publication, it contained 83 *Listeria* genome, 426 transcriptome, and 76 proteome data sets.[30] Notably, the majority of proteomic studies (qualitative and quantitative) were based on 2D gel studies, so modest numbers of proteins have been quantified thus far.[31−37] Similarly, the workflows employed to date have not provided the depth required for quantitative systems-level characterizations at the protein level.

In the present study, we set out to generate and validate two proteogenomic resources to enable analysis of genotype−proteotype−phenotype relationships in *L. monocytogenes* strains. By relying on the *de novo* assembled genomes of the model strains ScottA and EGD-e, we generated a mass spectrometry-based toolbox using next-gen DIA/SWATH workflows to enable the sensitive, repetitive, and quantitative interrogation of *L. monocytogenes* proteotypes. DIA/SWATH-MS (for data independent acquisition/sequential window acquisition of all theoretical mass spectra) is a recently introduced proteotyping technology based on mass spectrometry, which, compared to data-dependent acquisition (DDA) MS strategies, enables more sensitive generation of comprehensive proteotype data sets. The dynamic range of DIA/SWATH-MS is in excess of 4 orders of magnitude,[38] which matches the dynamic range of the proteotypes expected for our *L. monocytogenes* strains. Recently, proteogenomic studies on *Streptococcus pyogenes*[39] and *Mycobacterium tuberculosis*[40] utilizing the DIA/SWATH technology have illustrated its impact, which led to new biological insights with respect to invasiveness of clinical isolates and dormancy and resuscitation, respectively.

We applied our newly developed proteogenomic resources and toolbox to investigate how *L. monocytogenes* cells cope with stress encountered during passage through the upper GI tract, a prerequisite for systemic infection of the host. We uncovered evidence for the unexpected expression of flagella upon exposure to bile salts at 37 °C, a condition mimicking the duodenum. The comprehensive proteogenomic resource and toolbox we established here will enable further analyses of the *Listeria* genotype−proteotype−phenotype relationships.

EXPERIMENTAL PROCEDURES

Bacterial Strains and Growth Conditions

Bacterial strain EGD-e (serovar 1/2a) was derived from strain EGD, originally isolated from guinea pigs and used in studies of cell-mediated immunity,[19] and differs quite substantially from EGD.[41] ScottA is a clinical strain (serovar 4b) that was isolated during the Massachusetts listeriosis outbreak in 1983.[42] Cultures were grown to stationary phase in brain heart infusion (BHI) at 30 °C with shaking and then diluted 1:10 in BHI. Diluted cultures were incubated at 37 °C until the OD_{600} was 1. Cells were washed once with PBS and resuspended in the same volume of selected growth medium. Samples were incubated at 37 °C for 1 h with shaking, the medium was removed by centrifugation, and cell pellets were frozen until MS analysis. Three media were prepared to resemble conditions encountered in different parts of the upper GI tract[43] with buffered peptone water serving as control medium (BPW: 1% (w/v) peptone, 0.5% (w/v) NaCl, 0.35% (w/v) Na_2HPO_4, 0.15% (w/v) KH_2PO_4, pH 7.2). Low pH medium (stomach) was BPW, pH 4, 1000 units/mL pepsin. Bile salts medium (duodenum) was BPW, pH 7, 0.3% (w/v) bile salts. High osmolarity medium (jejunum) was BPW, pH 8, 0.3 M sucrose. Buffer composition was optimized based on the outcome of viability measurements. To assess viability, seven serial dilutions of cells were prepared, and 10 μL of each sample was spotted on agar plates. Plates were incubated at 30 °C overnight, and colony counting was performed for the dilution where colonies were well separated from each other.

Genome Sequencing, Assembly, and Annotation

Genomic DNA was prepared from overnight cultures of EGD-e and ScottA using the Sigma GenElute kit. An insert library was prepared, size selected with BluePippin (fragments >10 kb), and sequenced on the PacBio RSII sequencing platform (1 SMRT cell per strain; P6−C4 chemistry). Initial preprocessing steps (read quality control, preassembly) and *de novo* genome assembly were carried out using HGAP3[44] as described in detail before.[45] Subsequently, terminal repeats were trimmed, and the contigs were circularized and polished for two rounds using the PacBio preassembled reads. The EGD-e chromosome was aligned to the closest NCBI reference (GenBank accession: NC_003210) to adjust the start position according to the reference genome. For ScottA, the circularized

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

contig was start-aligned to the *dnaA* gene as described.[46] Both strains were also sequenced using Illumina MiSeq (2 × 300 bp paired end reads); raw fastq reads were mapped to the respective PacBio contigs using BWA-MEM v 0.7.12.[47] The final, high-quality genome sequences were submitted to NCBI GenBank and annotated with NCBI's PGAP 4.0.[48]

Functional Annotation

In addition to the NCBI annotation, all protein coding genes were also annotated using Interproscan v 5.30−69.0[49] (restricted to hits with an e-value below 1e-5), adding information on Gene Ontology (GO) classification, protein domains, patterns, protein families, profiles, etc. Furthermore, we extracted functional annotations for the protein sequences using eggnog-mapper (v 1.0.3),[50] which transfers functional annotation from the orthologous proteins present in EggNOG 4.5.[51] Prophage sequences were identified and annotated using PHASTER; putative prophages have lower scores than those predicted to be intact.[52] A GO enrichment analysis of the unique proteins, differentially expressed proteins, and proteins in the spectral library of each strain against all protein-coding genes of the respective strain was performed using the topGO package.[53] A Fisher's exact test with a *p*-value cutoff of 0.01 (0.05 for unique and differentially expressed proteins) was applied to identify significantly enriched GO categories across all three domains (i.e., BP, MF, and CC).[54]

Comparative Genomics

A comparison of our two *de novo* assembled, complete genomes was carried out using Roary (v 3.7.0)[55] with standard parameters (minimum identity for blastp set to 50%, no paralog splitting). From the gene presence/absence output table, we extracted the core and strain-specific protein-coding gene clusters (Table S3). The unique gene clusters and encoded proteins were subsequently used to identify the subset of proteotypic peptides that allowed quantification of the subset of proteins specific to each strain.

Single-Tube Sample Preparation for DIA/SWATH MS Analysis

A 1 mL culture of approximately 10e[9] bacteria was used as starting material yielding roughly 100 μg of total protein. Cell pellets were reconstituted in 500 μL of 50 mM ammonium bicarbonate buffer, and 5 μg of phage endolysin Ply511 was added. Endolysin was produced in *E. coli* and purified by affinity chromatography as described earlier.[56] The amount and time of endolysin incubation was optimized to allow complete lysis (based on OD_{600} measurements). Samples were incubated under fast end-to-end rotation for 15 min at 4 °C and then sonicated at maximum amplitude for 10 s three times or until the viscosity of water was reached. Samples were centrifuged for 10 min at maximum speed to remove debris, and a BCA protein assay was performed to assess the total protein concentration. For further preparation, 100 μg of total protein per sample was used. Proteins were denatured by heat (80 °C for 15 min) and addition of 0.1% acid-cleavable detergent. Proteins were reduced with 10 mM TCEP for 30 min at 25 °C and alkylated with 20 mM IAA for 30 min at 25 °C in the dark. Proteins were digested with LysC (1:300) for 3 h at 37 °C followed by trypsin digestion (1:100) for 16 h at 37 °C. Upon digestion, samples were acidified with 0.1% TFA and precipitated detergent was removed by centrifugation. Samples were desalted via mixed cation exchange chromatography and eluted in 1 mL 5% NH_4OH/90% MetOH. Peptides were dried and reconstituted in 60 μL 5% ACN, 0.1% FA with addition of iRT standard (1:10 v/v).

LC-MS/MS of DDA Runs

Peptides were analyzed on an Orbitrap QExactive Plus mass spectrometer (Thermo Scientific) equipped with a nano-electrospray ion source (Thermo Scientific) and coupled to a nanoflow high pressure liquid chromatography (HPLC) pump with an autosampler (EASY-nLC II, Proxeon). Peptides were separated on a reversed-phase chromatography column (75 μm inner diameter PicoTip Emitter, New Objective) that was packed in-house with a C18 stationary phase (Reprosil Gold 120 C18 1.9 μm, Dr. Maisch). Peptides were loaded onto the column with 100% buffer A (99.9% H_2O, 0.1% FA) at 800 bar and eluted at a constant flow rate of 200 nL/min with a gradient of buffer B (99.9% ACN, 0.1% FA) and a subsequent wash step with 90% buffer B. For the analysis of cell lysates, 3 μg of peptides were separated on a 50 cm heated column with a 120 min linear gradient of 5−35% B, followed by a 10 min gradient to 50% B and a 5 min gradient to 90% B. Between batches of runs, the column was cleaned with two steep consecutive gradients of ACN (10−98%). The MS was operated in DDA mode, with an automatic switch from MS to MS/MS scans. High-resolution MS scans were acquired in the Orbitrap (120,000 resolution, automatic gain control target value 2×10^5) within a mass range of 395 to 1500 m/z. The 20 most intense precursor ions (Top20) were fragmented using higher-energy collisional dissociation (HCD) to acquire MS/MS scans in the Orbitrap (30,000 resolution, intensity threshold 2.5×10^4, target value 2×10^5, isolation window 2 m/z). Dynamic exclusion was set to 30 s. Instrument performance was checked by regular quality control measurements using a yeast lysate and the iRT retention time peptide kit (Biognosys).

Database Search and Spectral Library Construction

The RAW files were processed with Proteome Discoverer software, v 2.1 (http://planetorbitrap.com/proteome-discoverer) using the RefSeq protein databases of the *de novo* assembled EGD-e and ScottA strains with the iRT peptide sequences added. As a first step, the raw files were analyzed with MaxQuant "dependent peptide" settings[57] to identify the most prominent post-translational modifications, which were then used as variable modifications in a second search, thereby limiting the overall search space. The main or second workflow consisted of SequestHT[58] and Amanda[59] search nodes coupled with Percolator.[60] The following search parameters were used for protein identification: (i) peptide mass tolerance set to 10 ppm; (ii) MS/MS mass tolerance set to 0.02 Da; (iii) fully tryptic peptides with up to two missed cleavages were allowed; (iv) carbamidomethylation of cysteine was set as fixed modification, methionine oxidation and protein N-term acetylation were set as variable modifications. Percolator was set at max delta Cn 0.05, with target FDR strict 0.01 and target FDR relaxed 0.05. The spectral libraries were generated in Spectronaut v 11 (Biognosys) using standard parameters including 0.01 peptide spectrum match (PSM) FDR and a Best N-filter with min three and max six fragment ions per peptide.

DIA/SWATH MS Sample Acquisition and Data Analysis

HRM calibration peptides (Biognosys) were spiked into the DIA samples according to the manufacturer's instructions. The samples were analyzed on the same LC-MS system as the DDA

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

runs using identical LC parameters. The mass range m/z 375−1200 was divided into 20 variable windows based on density as described previously.[61] The MS was operated in DIA mode with an automatic switch between MS to MS/MS scans. High-resolution MS scans were acquired in the Orbitrap (35,000 resolution, automatic gain control target value 5×10^6) within a mass range of 400 to 1220 m/z. DIA scans preceded an MS1 full scan in the Orbitrap (35,000 resolution, intensity threshold 3×10^6) with a stepped NCE 22.5, 25, 27.5. Instrument performance was regularly checked as described above for DDA measurements. All DIA data were analyzed directly in Spectronaut v 10 (Biognosys) with standard settings (dynamic peak detection, automatic precision nonlinear iRT calibration, interference correction, and cross run normalization (total peak area enabled)). All results were filtered for a q-value of 0.01 (equal to an FDR of 1% on the peptide level). All other settings were set to default.

Integrated Proteogenomics Search Databases (iPtgxDBs)

iPtgxDBs were created for EGD-e and ScottA by combining the NCBI's RefSeq protein annotation with Prodigal predictions, an *ab initio* gene predictor,[62] and a modified six-frame translation.[45] Proteomics data from the DDA runs were searched against the iPtgxDB FASTA file of each strain individually using MS-GF+ v 2017.01.13[63] to identify evidence for novel ORFs, alternative protein start sites, and SAAVs compared to the reference genome sequence. The search was performed in target-decoy mode, with a precursor mass tolerance of 10 ppm, full-trypticity, maximum precursor charge of 4, carbamidomethylation of cysteine as fixed, and methionine oxidation, asparagine deamidation, and protein N-term carbamidomethylation as variable modifications. The search results were filtered for a PSM level FDR of 0.05%, which ensured an estimated protein level FDR below 1%. In addition, we assessed the proteotypicity of the identified peptides using an in-house version of the original PeptideClassifier,[64] further extended to support proteogenomics in prokaryotes.[45] Only peptides that uniquely mapped to one protein (class 1a) were considered for protein identification. Moreover, following an earlier recommendation, we filtered the unambiguous peptides with an additional, variable PSM cutoff:[65] We required two PSMs/peptide for RefSeq annotated proteins; three PSMs per peptide for Prodigal predictions or N-terminal extensions to RefSeq proteins, and four PSMs per peptide for novel *in silico* predicted proteins. Data were overlaid on top of the GFF file and visualized in a genome browser.

Statistical Data Evaluation

All proteomics experiments on EGD-e and ScottA were performed in biological triplicates, except for EGD-e bile, where we could only quantify duplicates. DIA mapping data were searched against the in-house generated spectral libraries using Spectronaut, and the list of quantified spectral features (fragment ions/peptide sequences) was retrieved. In MSstats3 (v 3.12.2),[66] which is often used for downstream Spectronaut data processing, the features were log-transformed, and then subjected to median normalization. For feature summarization, the Tukey's median polish algorithm was applied. Protein fold changes and their statistical significance were tested using at least five features per protein. Tests for significant changes in protein abundance across conditions were based on a family of linear mixed-effects models. *P*-values were multiple testing corrected to control the experiment-wide FDR at a desired level using the Benjamini-Hochberg method. Proteins were considered differentially expressed if they showed a fold-change of 2 or higher and an adjusted p-value of 0.05 or lower.

Validation via PRM MS

Peptides were separated by reversed-phase chromatography on a high-pressure liquid chromatography (HPLC) column (75 μm inner diameter, New Objective) that was packed in-house with a 15 cm stationary phase (ReproSil-Pur C18-AQ, 1.9 μm) and connected to a nanoflow HPLC combined with an autosampler (EASY-nL1000). Peptides were loaded onto the column with 100% buffer A (99.9% H_2O, 0.1% FA) and eluted at a constant flow rate of 300 nL/min with a 90 min stepped gradient 3−25% buffer B (99.9% ACN, 0.1% FA) and 25−50% B. Mass spectra were acquired in PRM on an Orbitrap Fusion Tribrid Mass Spectrometer (Thermo Fisher Scientific). Spectra were acquired at 15,000 resolution (automatic gain control target value $5.0 \times 10e^4$); peptide ions in the mass range of 340−1400 were monitored. Stepped HCD collision energy was set to 27 (±5)%, maximum injection time to 22 ms. Monitored peptides and results were uploaded to PanoramaWeb (see Data Access).

Flagella Staining for Visualization under Light Microscopy

Flagella of *L. monocytogenes* were stained with Ryu stain using a wet-mount technique. Ryu stain was prepared before every experiment by mixing 1 part solution II with 10 parts solution I. Solution I (mordant) was 10 mL of 5% aqueous solution of phenol, 2 g of tannic acid, and 10 mL of saturated aqueous solution of aluminum potassium sulfate-12 hydrate. Solution II (stain) was 12 g crystal violet in 100 mL of 95% ethanol. For staining, cells were grown in the desired conditions, and 3 μL of culture was transferred on a glass slide and covered with a coverslip leaving small air spaces around the edge. Slides were incubated 10 min at 25 °C for bacterial cells to adhere, and 10 μL of Ryu stain was applied at the edge of the coverslip. The stain was left to mix with the cell suspension by capillary action. Slides were incubated for 10 min at 25 °C and examined under the microscope at 100× (oil).

Flagella Visualization Using Scanning Electron Microscopy

Bacteria were cultured as described above, and 0.2 mL of suspension was applied for 20 min on 12 mm coverslips covered with 15 nm of carbon and coated with poly(L-lysine). After washing twice with PBS, cells were fixed with 2.5% glutaraldehyde in PBS at room temperature before immediately transferring the samples on ice for 2 h. Samples were then processed in a Pelco Biowave Pro+ tissue processor with use of microwave energy and vacuum. Briefly, the fixed samples underwent a second fixation step in 2.5% glutaraldehyde, before being washed and postfixed in 1% OsO_4, followed by 1% uranyl acetate in water and dehydration by successive immersion in increasing concentrations of ethanol and finished by critical point drying out of ethanol. The dried coverslips were mounted on SEM aluminum stubs and sputter-coated with 5 nm of platinum/palladium. SE images were recorded at 2 kV in a Zeiss Gemini 1530 FEG.

Nucleic Acid Extraction, Purification, and cDNA Synthesis

Bacterial nucleic acids were extracted using a phenol−chloroform protocol adapted from a previous report.[67] Briefly, the bead-beating step was carried out with 500 μL of 0.1 mm zirconia/silica beads in a Mixer Mill MM301. Lysis was performed in two rounds of 4 min each with 5 min rest on ice. Nucleic acids were precipitated by addition of 0.1 volume 3 M

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

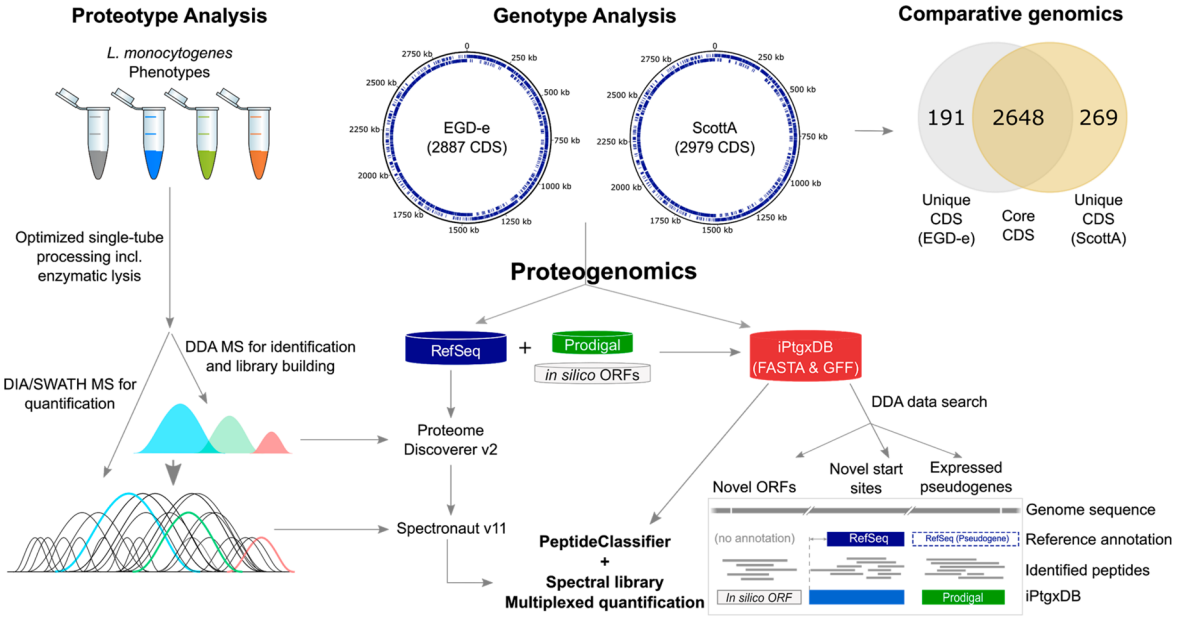


Figure 1. Overview of our next-gen proteogenomics workflow for *L. monocytogenes*. A sample preparation method was developed that allows rapid, reproducible single-tube reactions for lysis, digestion, desalting, and subsequent proteotype profiling with DDA- and DIA-MS (left panel). In parallel, the genomes of both EGD-e and ScottA were *de novo* assembled into complete, high-quality genome sequences, and their RefSeq annotations were obtained from the NCBI's prokaryotic genome annotation pipeline (PGAP)[48] (middle upper panel). Comparative genomics identified both shared core gene clusters and gene clusters unique to EGD-e and ScottA (right upper panel). Moreover, an *ab initio* gene prediction based on Prodigal and a modified *in silico* (six-frame translation) annotation (see Experimental Section) were integrated with the RefSeq annotation to obtain a minimally redundant iPtgxDB[45] for EGD-e and for ScottA. DDA-based proteomics data were searched against the publicly available iPtgxDBs (https://iptgxdb.expasy.org) to obtain proteogenomic evidence for novel open reading frames (ORFs), novel start sites, and expressed pseudogenes (right lower panel). For the proteotype analysis, proteomics data obtained from DDA mode were searched against the RefSeq annotations, and spectral libraries were generated with Spectronaut. These publicly available resources were then used to analyze and quantify proteins obtained from DIA mode.

sodium acetate and 0.6 volume ice-cold isopropanol. The pellet was resuspended in RNase/DNase-free water and subsequently purified with the AllPrep DNA/RNA kit (Qiagen) according to the manufacturer's instructions. DNA isolated from the control culture grown in BHI at 37 °C served as positive control. DNA and RNA were quantified with the Qubit Fluorometer 3.0. RNA was isolated from all samples, and 0.8 μg were used for cDNA synthesis (in triplicate) with the TaqMan Reverse Transcription Reagents kit according to the manufacturer's instructions using the random hexamer technique. Some reverse transcription reactions were carried out without enzyme as a control for the absence of DNA in the samples.

qPCR

The reaction was performed in a final volume of 10 μL containing 1× SYBR Green Mix, 0.5 μM each of the *flaA* primers (forward: 5′-GCTGGTCTTGCAGTTGTTACTCG-TATG-3′; reverse: 5′-CTAATTGACGCATACGTTGC-AAGATTG-3′) and 1 μL of the diluted DNA template using the SYBRTM Select Master Mix and run on a Rotor-Gene 6000 PCR system with the following program: 50 °C for 2 min, 95 °C for 2 min, followed by 40 cycles at 95 °C for 15 min. Three biological replicates were analyzed with samples analyzed in duplicate by qPCR. Raw data were processed with the LinRegPCR program to determine the Cq values, and Python 3.7.0 was used for analysis and to create plots.

Data Access Note

RAW files from DDA-MS experiments that were used as a basis to develop the spectral libraries and the DIA measurements are available from MASSIVE under ftp://massive.ucsd.edu/MSV000083881/ or from ProteomeXchange PXD014091. Targeted MS experiments can be accessed via Panorama (https://panoramaweb.org/660xuF.url) or from ProteomeXchange at PXD014294. The genome sequences for ScottA and EGD-e are available from NCBI Genbank under accession numbers CP023862 and CP023861, respectively. iPtgxDBs for both strains are available at https://iptgxdb.expasy.org.

RESULTS AND DISCUSSION

Generic, Genomics-Driven Strategy Enabling the Investigation of Genotype−Proteotype−Phenotype Relationships

We selected two widely used *L. monocytogenes* strains, EGD-e and ScottA, as model systems to evaluate the applicability of our genomics-driven next-gen proteomics workflow (Figure 1). EGD-e and ScottA belong to serotypes 1/2a and 4b, respectively, which are responsible for the majority of listeriosis cases (Table S1); ScottA is more invasive than EGD-e.[68] We relied on an integrated workflow to obtain a quantitative profile of the *Listeria* proteotype. This workflow contains four main components: genomics, comparative genomics, proteotype analysis, and proteogenomics (Figure 1).

The first step in our workflow was a genomics analysis. Although a complete NCBI reference genome sequence existed for EGD-e, the NCBI reference genome sequence for ScottA was incomplete and consisted of five contigs.[42] Motivated by our recent finding of significant differences between an NCBI reference genome and the *de novo* assembly

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

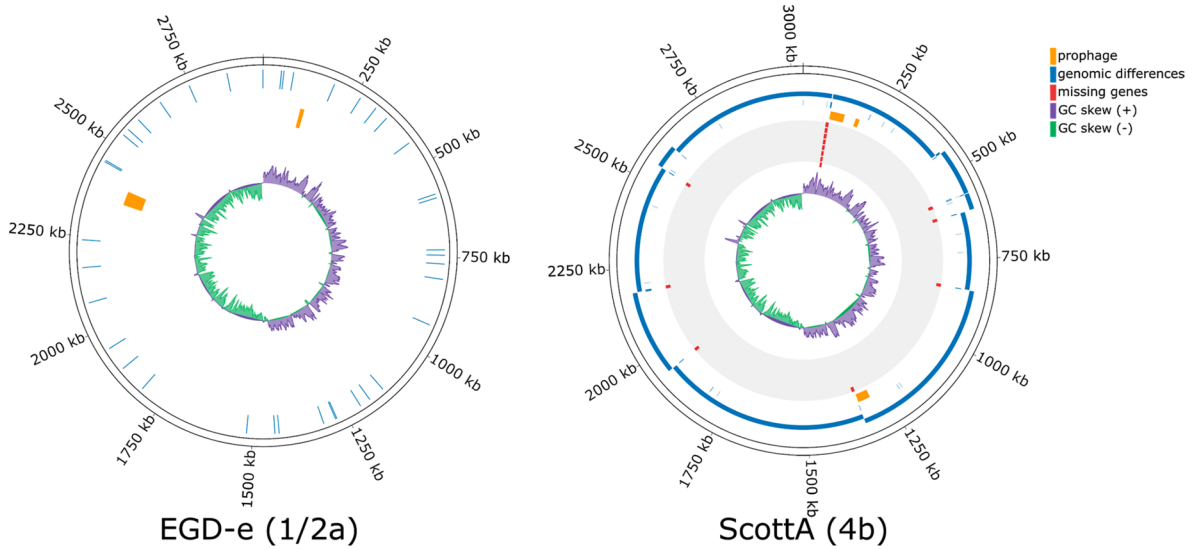


Figure 2. Circular plots showing the *de novo*-assembled genome sequences (outer ring with genomic coordinates) of EGD-e and ScottA and the genomic differences compared to the existing reference sequences, NC_003210 and NZ_CM001159, respectively. Whereas our EGD-e assembly exhibited minor differences compared to the NCBI reference genome (SNVs and INDELS in blue, second circle), there were a number of differences between our complete ScottA assembly and the NCBI reference (incomplete, 5 contigs), which had been assembled using a reference-based assembly approach.[42] The mapping of the 5 contigs and the remaining gaps (blue) are shown in the second circle; prophages (orange) in the third, 14 missing genes (red) in the fourth, and the GC skew (positive, purple; negative, green) in the fifth circle.

of the actual lab strain,[45] and the fact that incomplete short-read-based assemblies can miss important genes as in the case of *Pseudomonas aeruginosa* MPAO1,[69] the parental strain of a widely utilized transposon mutant collection,[70] we sequenced and *de novo* assembled both genomes to create the best possible reference sequence for the two strains, an important aspect for the proteogenomics element. Next, a comparative genomics analysis was carried out to identify core and strain-specific genes, which, upon integration with protein abundance data, might provide clues to explain the different phenotypes of the strains. Third, we performed DDA-MS experiments under relevant conditions to obtain extensive *Listeria* proteotype data sets, which became the basis for the construction of *Listeria* spectral libraries. After establishing a rapid and reproducible sample preparation workflow, we were able to quickly and quantitatively profile *Listeria* under various conditions and perturbations by DIA/SWATH, benefiting from the higher reproducibility of DIA compared to DDA data.[61] Lastly, in addition to the standard proteotype search against generic protein databases such as NCBI RefSeq and Uniprot, a search was carried out against a specialized, integrated proteogenomics search database (iPtgxDB; see Experimental Section),[45] allowing us to identify protein expression evidence for as of yet unannotated small ORFs (smORFs), additional N-terminal start sites, and expressed pseudogenes (Figure 1).

A variant of such a proteogenomics approach recently revealed N-terminal peptides both from internal start sites of annotated *L. monocytogenes* EGD-e proteins and from six novel, unannotated smORFs including Prli42.[20] This 31-amino-acid protein relays oxidative stress signals to the stressosome to activate the general stress-sensing pathway, the sigma B regulon, and represents the long-sought link between stress and the stressosome.[71] Despite the many important functions of smORFs, such small, functional open reading frames are often missed in current genome annotations.[72] Proteogenomics and ribosome profiling have emerged as the most important technologies for comprehensive identification of smORFs.[72] Consequently, we added a proteogenomics element to our generic strategy. This has two benefits: First, proteogenomics can be included in the initial genome annotation, thereby increasing its quality.[45,73] A public Web site (https://iptgxdb.expasy.org/) supports this for newly sequenced genomes,[45] such as isolates from microbiomes or type strains from the Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative, which aims to expand the phylogenetic diversity of completely sequenced prokaryotic genomes.[74] Second, the identification of more comprehensive protein catalogs including functionally relevant smORFs will better enable studies to model systems based on quantitative data of all functional elements.

Complete Genome Sequences of EGD-e and ScottA and Comparative Genomics

An analysis of the repeat complexity of all publicly available, completely sequenced genomes of *L. monocytogenes* strains (status: March, 2018) revealed that almost 95% of the roughly 150 strains are so-called "class I" genomes, which are straightforward to assemble (few repeats, none longer than the rDNA operons of up to 7 kb).[17] In contrast, eight strains also had few repeats overall, but those present were up to 11 kb in length.[17] Using long-read Pacific Biosciences (PacBio) sequencing data (including a BluePippin size selection step; see Experimental Section) and the assembly algorithm HGAP3,[44] we were able to *de novo* assemble one complete chromosome for EGD-e (2.94 Mbp) and one for ScottA (3.03 Mbp) with PacBio read coverages of 260× and 280×, respectively (Figure 2). To correct remaining homopolymer errors and to remove small insertions or deletions (INDELs) in the PacBio data,[75] both strains were also sequenced using the highly accurate short-read Illumina protocol. The Illumina data allowed us to ensure that no additional plasmids could be assembled that might have been lost in the BluePippin size selection step before creating the insert libraries for PacBio sequencing. Additional genome properties, including the number of protein-coding genes predicted by NCBI's PGAP

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

annotation pipeline, are listed in Table S2. Two intact and one putative prophage were identified by PHASTER[52] in ScottA, and two putative prophages were identified in EGD-e (Experimental Section).

Comparing our *de novo* assembled ScottA genome to the NCBI reference (CM001159.1; 5 contigs), we observed a total of 11,953 base pairs (bp) of missing sequence, which affected 14 genes that were completely or partially missed by the earlier, at the time state of the art, reference-based genome assembly.[42] The missing genes included seven hypothetical proteins, three surface proteins, one cell-wall anchor-domain containing protein, and three transposases (Table S3). The genomic differences included 14 insertions, 16 deletions, 93 single-nucleotide variations (SNVs), and 34 variations affecting 2 or more nucleotides (Figure 2). In contrast, only 28 SNVs and 11 single-bp INDELs were observed between our EGD-e assembly and the NCBI reference genome (NC_003210.1) (Figure 2).

The example of ScottA illustrates that a *de novo* assembly strategy is preferable over a reference-based one, which can easily miss important genome sequence differences. Roughly 570 of 9331 bacterial genome assemblies (6.1%) that we recently analyzed[17] were prepared using a reference-based genome assembly strategy; these assemblies thus have to be treated with caution. With current long-read sequencing technologies like Pacific Biosciences[76] and Oxford Nanopore Technologies[77] readily generating sequence reads longer than 20 kb, *de novo* genome assembly has become the preferred approach.

Finally, complete genomes also represent the best basis for comparative genomics studies, as core genes have been missed when more fragmented assemblies based on short Illumina reads were used for comparative genomics.[46] A comparison of the complete genomes of EGD-e and ScottA using Roary[55] revealed a high similarity with 2648 core gene clusters (orthologous proteins), 191 (6.6%) EGD-e-specific, and 269 (9%) ScottA-specific gene clusters (Figure 1 and Table 1).

Table 1. Overview of the Comparative Genomics Results of the Newly Sequenced and Assembled EGD-e and ScottA Genomes Including Core and Strain-Specific Protein-Coding Gene Clusters

	L. monocytogenes EGD-e	*L. monocytogenes* ScottA
Genbank accession #	CP023861	CP023862
Size of the chromosome (bp)	2,944,523	3,030,813
Total number of protein-coding genes	2,887	2,979
Number of strain-specific genes	191	269
Number of core genes	2,648	2,648

Notably, 5 of the 14 genes that were missed in the 5 contigs of the fragmented ScottA genome are indeed core genes that are shared by both strains. A list of core and strain-specific genes is provided in Table S3. These results will enable the *Listeria* research community to further explore differing virulence capabilities and other properties of these two strains. To facilitate such comparisons and integration with other data sets, we also provide a detailed table with the genes of both strains, functional annotations, proteomic abundance evidence, and a reciprocal best BLAST hit comparison against the ListiList EGD-e proteins with identifiers in the form of LmoXXXX (https://listeriomics.pasteur.fr), where X is a number from 0 to 9) in Table S5. Together, these new proteogenomic resources represent a high-quality set of puzzle pieces required for modeling and understanding the *Listeria* life cycle and infectious mode of action.

Growth under Conditions Mimicking Passage through the Gastrointestinal Tract

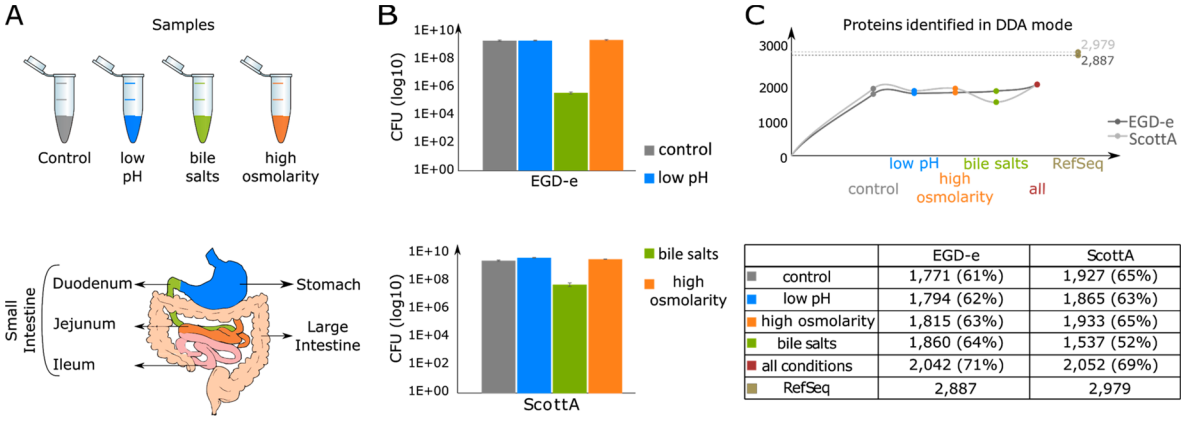
Despite intense research, knowledge of the proteotype adaptations that facilitate survival of *L. monocytogenes* in the GI tract remains incomplete. To cause a systemic infection, the bacteria have to be ingested and travel through the GI tract.[8] The GI tract is a hostile environment, and the bacteria are subjected to mechanical and chemical stresses that differ depending on the compartment: These include acidity (stomach), exposure to bile (duodenum), and exposure to high osmolarity (jejunum) (Figure 3A). To enable a systems-wide, quantitative characterization of the proteins required for survival and adaptation of *L. monocytogenes* under these GI-imposed stresses, we first sought to devise *in vitro* culturing conditions that resemble the different microenvironments of the GI tract. As not all genes will be expressed under one condition, these perturbations were chosen in order to obtain a broad representation of expressed *Listeria* proteins. We reasoned that discovery-driven DDA-based proteotype data from such conditions would allow us to create a comprehensive spectral library, an important element of our next-gen DIA/SWATH-based proteotyping strategy (Figure 1). This was achieved by culturing *L. monocytogenes* cells in three different conditions at 37 °C[43] and in a control condition (buffered peptone water (BPW), pH 7.4) in which the cells did not replicate.

After testing several combinations (Experimental Section, Figure 3A), the following buffer compositions were chosen (Figure 3A): To mimic conditions encountered in the stomach, the pH of the buffer was 4, and the medium included 1000 units/mL of pepsin. To simulate the duodenum, the medium included 0.3% w/v bile salts (pH 7.4). Finally, in order to mimic conditions encountered in the jejunum, the pH of the medium was 8, and it contained 0.3 M sucrose to increase the osmolarity (Figure 3A). The viability of the strains was not affected by the stresses of low pH/pepsin and high pH/high osmolarity (Figure 3B). In contrast, viability was severely compromised by the presence of bile salts, indicating that detergent-like activities are detrimental to survival of bacterial cells. In the presence of bile salts, the viability of ScottA (approximately one log lower than in the control) was affected much less than that of EGD-e (approximately 4 logs lower compared to the control) (Figure 3B). The greater sensitivity of EGD-e to both pH and bile compared to other *L. monocytogenes* strains, including ScottA, has been noted before.[6,7] Although these are clearly model conditions, in the absence of better *in vivo* models, they will lead to a better understanding of the mechanisms by which *Listeria* adapts to the host GI tract, informing the development of novel treatment or prophylactic strategies.

Listeria Proteotype Analysis Using a Single-Tube Workflow and Spectral Libraries

To ensure reproducible, sensitive, and quantitative proteotype measurements both for this study and as a general resource for the *Listeria* research community, a robust sample preparation workflow was needed that included a minimal number of sample preparation steps, while simultaneously enabling a comprehensive protein identification and quantification across

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662



	EGD-e	ScottA
control	1,771 (61%)	1,927 (65%)
low pH	1,794 (62%)	1,865 (63%)
high osmolarity	1,815 (63%)	1,933 (65%)
bile salts	1,860 (64%)	1,537 (52%)
all conditions	2,042 (71%)	2,052 (69%)
RefSeq	2,887	2,979

Figure 3. Discovery-driven DDA-based *Listeria* protein expression data obtained under conditions mimicking passage through the upper GI tract. (A) Sample conditions (control, gray; stomach, blue; duodenum, green; jejunum, orange). (B) Viability of cells in control and three different conditions. (C) Number of proteins identified per condition for each strain, the overall number of identified proteins (red), and the number of proteins annotated in each strain (brown).

Table 2. Numbers of Precursors, Peptides (Confidence Level High), and Proteins (Confidence Level Medium) That Are Covered in the Spectral Libraries of *L. monocytogenes* Strains EGD-e and ScottA in Spectronaut and Peptides That Were Observed for Strain-Specific Proteins

	proteins (strain-specific)	precursors	peptides	proteotypic (per strain)	proteotypic (strain-specific; see text)	proteotypic peptides of 71 strain-specific proteins (expt. observed)
EGD-e	1992 (71)	32,092	23,238	22,431 (96.5%)	3,493 (15.6%)	544 (2.4%)
ScottA	2002 (71)	37,323	23,260	22,682 (97.5%)	3,679 (16.2%)	602 (2.7%)

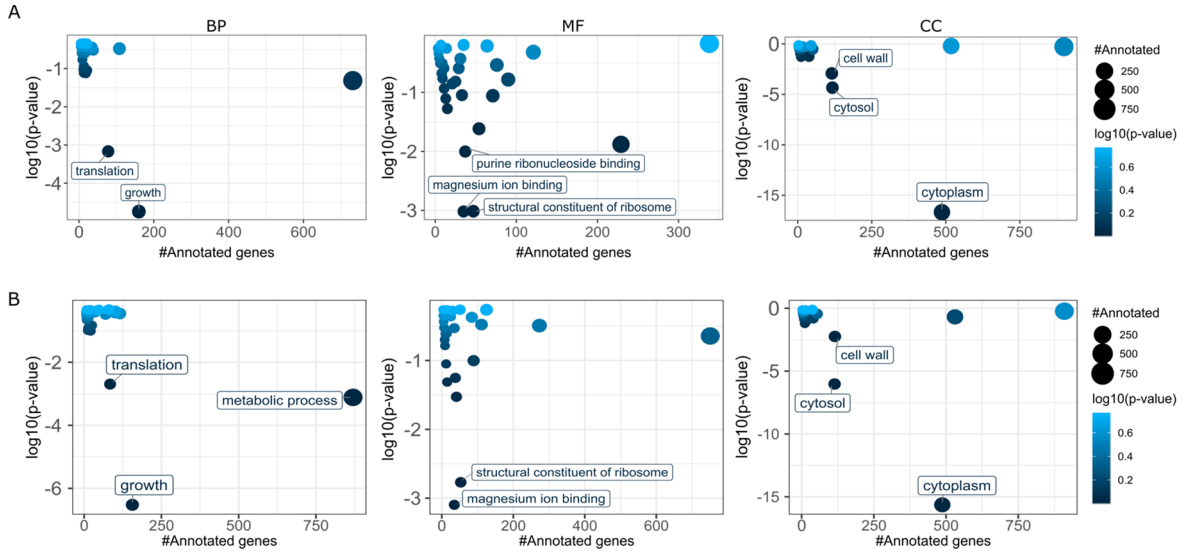


Figure 4. GO enrichment analysis of proteins included in the spectral libraries. GO categories across the three domains biological process (BP), molecular function (MF), and cellular component (CC) were analyzed with a Fisher's exact test (p-value = 0.01). The X-axis shows the number of annotated genes in each GO category, and the Y axis shows the log10 p-value. Significantly enriched GO categories (p-value <0.01) in the spectral library are labeled in the figure. The GO category "cytoplasm" had the most significant e-value. (A) Plots for strain EGD-e. (B) Plots for strain ScottA.

conditions/proteotypes. Typically, bacterial cells are lysed mechanically (i.e., by bead beating) or by use of detergents. However, these methods require additional steps for bead or detergent removal making the workflows more tedious, increasing technical variability, and potentially leading to loss of low abundance proteins. Therefore, an efficient and proteotype analysis-compatible workflow was developed in which all steps from lysis to protein digestion are performed in a single tube (Figure 1). Effective, rapid, and complete cell lysis was achieved by incubating *L. monocytogenes* cells with the bacteriophage endolysin (Ply511; see Experimental Section). Recombinantly produced endolysins applied exogenously to susceptible bacteria display the same lytic properties as their native counterparts.[78] In combination with indirect sonication, combined LysC/trypsin protein cleavage into peptides and desalting using a mixed cationic exchange resin, the sample processing strategy enables rapid proteotyping of *L. monocytogenes*.

To generate spectral libraries, samples were first analyzed in DDA-MS mode. In total, we identified qualitative expression

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

data for between 1700 and 1800 proteins per condition for EGD-e and between 1500 and 1900 proteins per condition for ScottA (Figure 3C), similar to previously reported, extensive MuDPIT data sets from *Listeria*.[36,79] We next used the discovery-driven DDA-based data to generate spectral libraries for DIA/SWATH-based acquisition and reliable protein quantification across conditions. Summaries of the spectral library characteristics for the two strains are depicted in Table 2 and Figure 4.

Each library contained roughly 23,000 peptides that corresponded to approximately 2000 proteins (false discovery rate (FDR) < 1%)). Almost all of the peptides were proteotypic for the given strain as indicated by Spectronaut. Overall, the spectral libraries contained roughly 69% and 67% of the annotated ORFs of EGD-e and ScottA, respectively. To the best of our knowledge, this is the first report of *Listeria*-specific spectral libraries covering such a high proportion of the theoretical proteome. A gene ontology (GO) enrichment analysis of the proteins contained in the spectral libraries showed that the majority of pathways were represented in both of the libraries and that similar GO categories were enriched in both strains across the three domains. Notably, the cellular component (CC) "cytoplasm" was significantly over-represented among proteins in the spectral libraries of both strains (Figure 4), whereas proteins associated with the CC category "transmembrane" were under-represented. This is an expected outcome and can be explained by our focus on the development of a single-tube, rapid and reproducible sample processing workflow. Proteome coverage without a bias against the membrane proteome has only very rarely been achieved and requires elaborate strategies (computational and biochemical and/or subcellular fractionation) and substantial effort.[27] The CC categories "cytosol" and "cell wall" were also over-represented.

The peptides from the two libraries were mapped against the protein-coding sequences of the other genome to identify strain-specific proteotypic peptides and to assess whether some of the strain-specific protein-coding genes were expressed at the protein level. This mapping was done using an in-house tool that extends the original version of PeptideClassifier[64] thereby enabling proteogenomics for prokaryotes.[45] Only 15.6% and 16.2% of peptides from EGD-e and ScottA, respectively, were strain-specific proteotypic, indicating that most peptides could be used to quantify protein abundance in both strains and potentially also in other *Listeria* strains. Additionally, we found that 544 unambiguous peptides of the EGD-e library confirmed expression of 71 of the 191 EGD-e specific protein-coding genes and that 602 ScottA strain-specific proteotypic peptides in the library confirmed expression of 71 of 269 ScottA-specific proteins (Table 2 and Table S3). We provide a GO enrichment analysis of the strain-specific proteins of each strain in Table S4, which will enable a further hypothesis-driven exploration by the community. Together, these efforts resulted in a second new proteogenomic resource that enables fast proteotyping of *L. monocytogenes* using a single-tube processing strategy in combination with a DIA/SWATH-MS-based workflow that benefits from the generated *L. monocytogenes* spectral libraries. These libraries are now publicly available (ftp://massive.ucsd.edu/MSV000083881/). Due to the high sensitivity of this spectral library-based DIA/SWATH MS approach in combination with top-end MS instruments, it is now conceivable to gain proteotype information directly from limited numbers of cells extracted from *in vivo* models of *L. monocytogenes* infection.

Integrated Proteogenomic Databases as a Basis to Relate Proteotype Data Back to the Genotype

In order to make the generated information accessible for public use, we also created an iPtgxDB for each strain. The concept of iPtgxDBs as a "one-stop shop" for a protein search database that combines the benefits of manual curation efforts with the ability to identify missed smORFs by capturing the entire protein-coding potential of a prokaryotic genome has been described previously.[45] Proteotype data from any condition can be searched against the FASTA file, and experimental evidence (peptides or gene expression data, if available) can be integrated with the GFF file provided, thereby allowing users to visualize experimental evidence for novelties (Figure 1). Alternatively, users may simply compare different annotation resources of a genome sequence or even NCBI RefSeq releases, which can differ substantially.

The minimally redundant iPtgxDBs for *L. monocytogenes* EGD-e and ScottA strains contained 65,393 and 67,150 proteins, respectively, and were created by integrating and consolidating annotations from RefSeq,[48] Prodigal,[62] and a modified form of a six-frame translation using the public iPtgxDB web server (https://iptgxdb.expasy.org/iptgxdb/submit/).[45] Metadata on the number of proteins in each annotation source, the progressive increase of annotation clusters, and the overall number of ORFs in the final iPtgxDB are shown in Table S6. Proteotype data measured in the control and three GI-mimicking conditions in DDA-MS mode were searched individually against the iPtgxDB fasta file of each strain (see Experimental Section).

A stringent control of the FDR is particularly relevant for proteogenomics applications, which can lead to corrected genome annotations. We selected a stringent global FDR at the PSM level and relied on an additional class-specific FDR, as advocated before.[65] We thus required more PSMs for purely *in silico* predicted proteins, thereby accounting for the variable credibility of the annotation resources we integrate in our iPtgxDBs. Moreover, for novel proteins implied by a single peptide,[64,80] we ensured that the peptide was proteotypic considering the entire coding potential of the genome,[45] and we explored the E-value distribution for novel hits compared to that of decoy hits. These steps ensure that potentially valid short and lower abundance proteins that generate fewer peptides can be correctly identified,[81] while at the same time keeping the higher error rate for single peptide identifications under control.[82,83] At a stringent peptide-spectrum match (PSM) level FDR (0.05%, resulting in a protein-level FDR well below 1%), we obtained unambiguous peptide evidence for 1907 proteins in EGD-e including 1899 RefSeq proteins and 6 additional novel proteins (Table S7). The novel proteins included two Prodigal-predicted proteins or proteoforms, three *in silico* ORFs, and one protein with an alternative start site. Furthermore, we observed peptide evidence supporting 7 of 28 SNVs in our *de novo* assembly compared to the EGD-e reference (NC_003210). In ScottA, we observed unambiguous peptide evidence for 1910 RefSeq proteins including 4 proteins (3 transposases and 1 cell wall protein) that were missed in the incomplete ScottA reference sequence (Genbank accession: NZ_CM01159; 5 contigs). Additionally, we identified evidence for 6 novel proteins including three Prodigal proteins, one *in silico* ORF, and two alternate protein start sites (Table

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

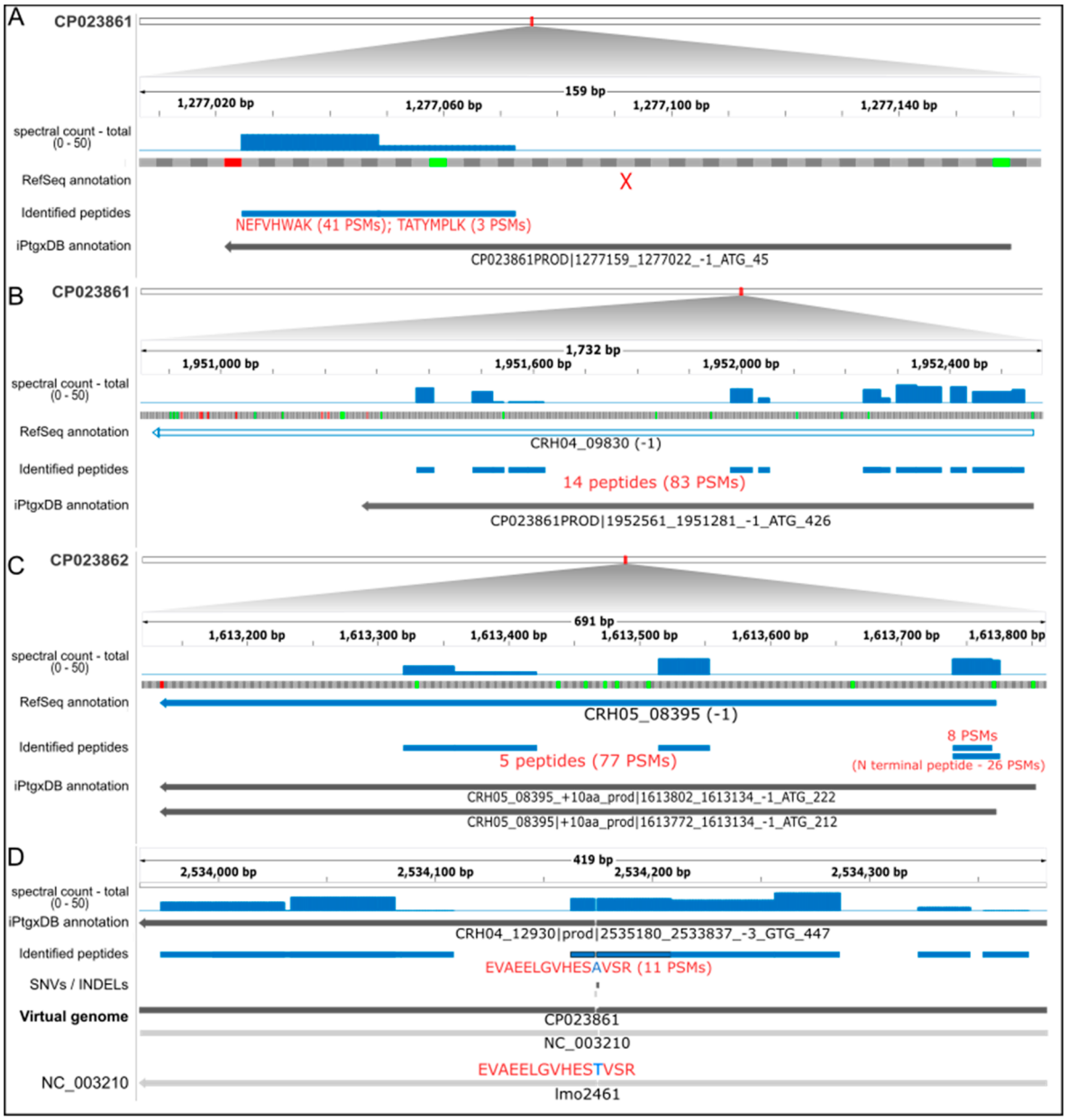


Figure 5. Peptide evidence for novelties identified by proteogenomics. Peptide evidence for (A) a new smORF, (B) an expressed pseudogene, (C) a new start site, and (D) a single amino acid variation (SAAV) uncovered in *L. monocytogenes* strain EGD-e and/or ScottA. Shown are respective genomic localizations. For A–C, the respective accession numbers of our assembly are given on the left above various annotation tracks. The iPtgxDB annotation is shown in dark gray, peptide evidence supporting novelties in blue, and a summary of these peptides and PSMs are shown in red. The sequences of the peptides implying the SAAV are shown in panel (D). In the subfigure in panel (D), two genome sequences are compared as a virtual genome, allowing us to overlay experimental evidence. All novelties were discovered by searching data against the strain-specific iPtgxDB. For simplicity, *in silico* predicted ORFs are not shown.

S7). Further, by adding the peptides that imply novel proteins identified by proteogenomics proteins to the spectral library, we could observe quantification values for the proteins predicted by Prodigal and for alternate start sites across the different conditions and perturbations from both strains using DIA/SWATH MS. This result enabled us to validate the identified novelties independently beyond DDA based data. It also illustrated the utility of the reusable DIA/SWATH MS data.

Identified novelties include a 45-aa hypothetical protein predicted only by Prodigal that was identified in strain EGD-e by 2 peptides and 44 PSMs (Figure 5A). Notably, the same 45-aa protein was also identified with these 2 peptides and 35 PSMs in strain ScottA (Table S7). A BLAST search showed that this smORF is conserved across *L. monocytogenes* strains. The second example is the case of a pseudogene predicted by RefSeq in EGD-e (520 amino acids), which contains an internal C-terminal stop codon. Protein expression evidence was observed for a corresponding smaller Prodigal-predicted protein of 426 amino acids (Figure 5B). Both the Prodigal and Refseq proteins are annotated as formate-tetrahydrofolate ligase. Peptide evidence confirmed expression of the protein to the internal stop codon; this proteoform also contains the P-loop nucleoside triphosphate hydrolase domain that is

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647–1662

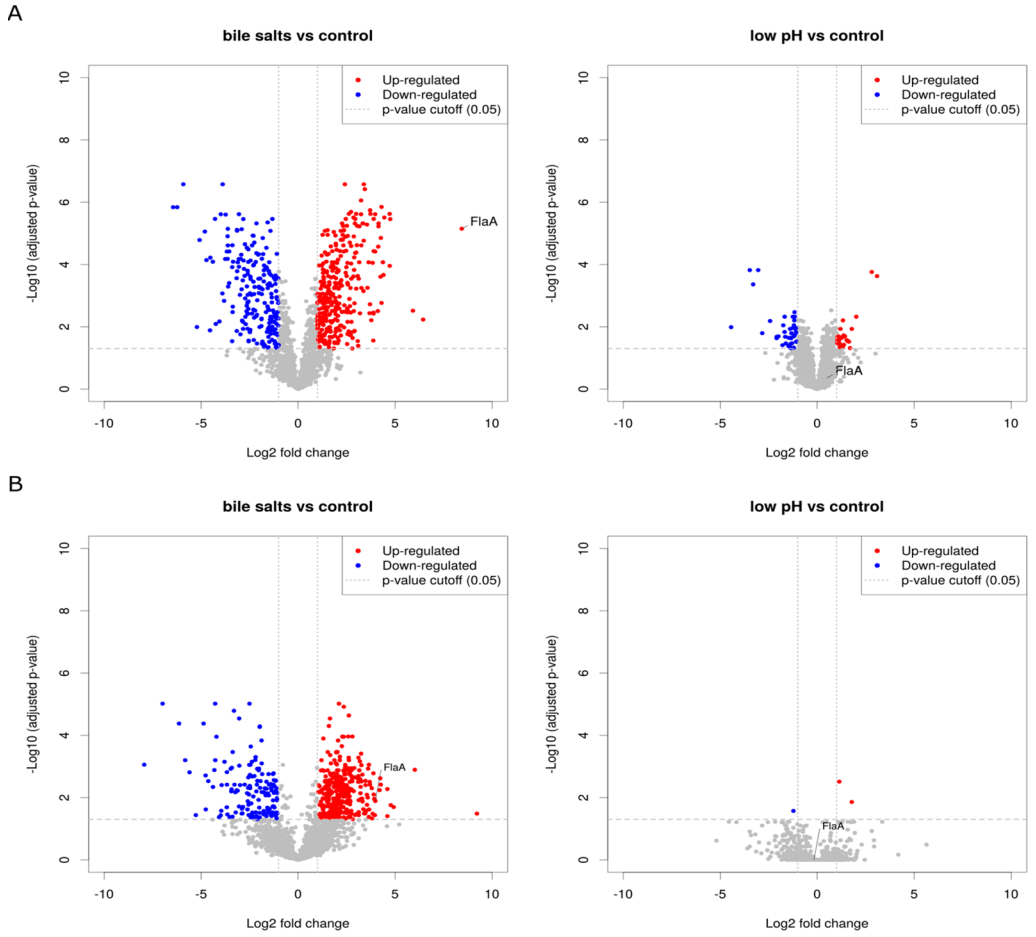


Figure 6. Differential protein abundance in EGD-e and ScottA cells upon exposure to conditions mimicking passage through the upper GI tract. Volcano plots depict the differentially abundant proteins for two conditions, i.e., stomach (low pH) and duodenum (bile salts) compared to control for (A) EGD-e and (B) ScottA. The adjusted *p*-value (multiple testing corrected; see Experimental Section) is shown as a horizontal gray dashed line, the log2 fold change cutoff of 1 as a vertical dashed gray line. Up-regulated proteins are shown in red, down-regulated proteins in blue; FlaA is labeled as one example.

conserved across *L. monocytogenes*. The third example shows yet another Prodigal prediction for a protein that is 10-aa longer than the corresponding RefSeq protein (which is 212 amino acids) of strain ScottA in (Figure 5C); a peptide supporting the longer proteoform was identified with 26 PSMs. The same peptide was also identified in strain EGD-e with 25 PSMs (Table S7). Finally, one peptide (EVAEELGVHES-AVSR, 11 PSMs) supports the nonsynonymous amino acid change (caused by an SNV that results in a threonine to alanine codon change) in the protein annotated in our *de novo* assembly as RNA polymerase factor sigma-54 of 447 amino acids (Figure 5D). Figure S1 shows proteomics evidence (3 peptides, 17 PSMs) for another Prodigal-predicted protein (207 amino acids) annotated as a RpiR family phosphosugar-binding transcriptional regulator; the corresponding Refseq annotation wrongly predicts a pseudogene.

The iPtgxDBs are publicly available (https://iptgxdb.expasy.org/database/). These databases will support efforts in the *Listeria* community to find proteogenomic evidence for additional novel smORFs, as pioneered with the example of Prli42.[20] This type of data has been instrumental in uncovering novelties in several model organisms including the eroded genome of an obligate plant symbiont.[84] Notably, the integration with global dRNA-seq data sets allowed identification of internal start sites of annotated proteins,[85] similar to the N-terminomics study in *Listeria*.[20]

L. monocytogenes Proteotype Analysis Reveals Adaptation during Stress

Overall, approximately 1700 and 1900 proteins were identified and quantified (for details see Table S8) for EGD-e and ScottA, respectively, which represents a major advance for quantitative proteotype profiling studies in *Listeria*. Table S9 provides a summary over the respective library recovery percentage, data completeness, and median CVs for both strains. Average correlation coefficients of biological replicates were above 0.98 indicating a very good reproducibility of sample analysis. The reproducibility of the sample analysis was also reflected by the median CVs that ranged around 20% (Table S9). Additionally, unsupervised hierarchical clustering revealed clustering of the biological replicates and a distinction across the different conditions tested (Figure S2). Notably, the samples incubated with bile salts (duodenum-mimicking condition) demonstrated lower library recovery and a higher number of missing values. This was a result of the significantly lower amount of starting material due to decreased cell viability (Figure 3B). Figure 6 summarizes the proteins found to be differentially abundant in bile salts and low pH compared to the control condition for strains EGD-e (Figure 6A) and

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

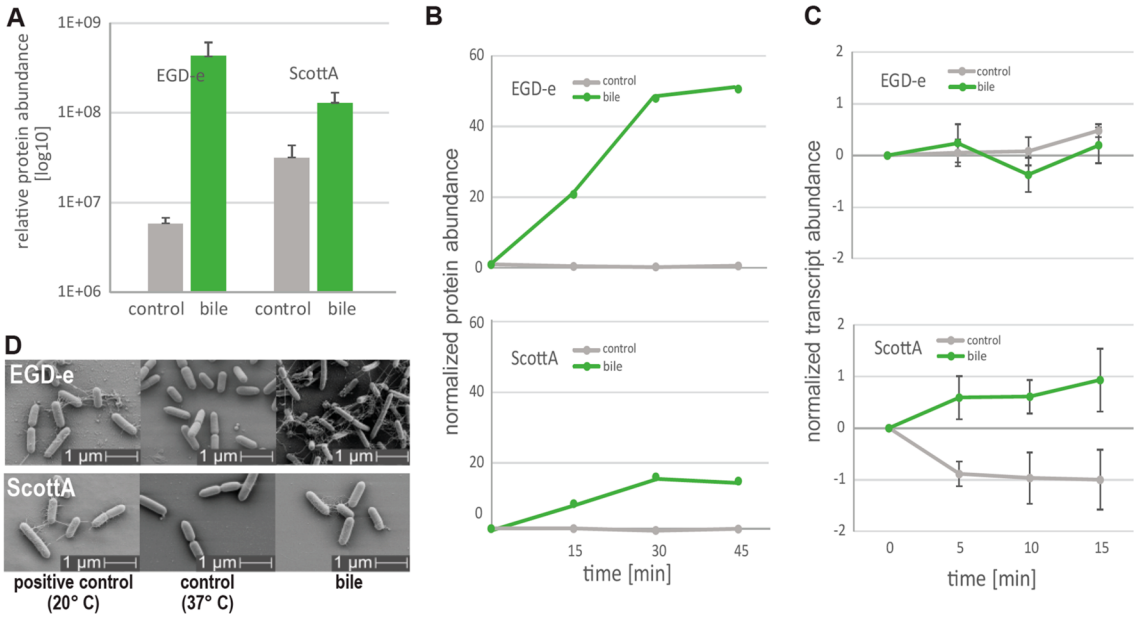


Figure 7. Production of flagella in *Listeria* strains incubated with bile salts at 37 °C. (A) Bar graph of FlaA relative protein abundance based on PRM quantitation. (B) Protein and (C) RNA expression levels of *flaA* assessed in a time-course experiment using PRM and qPCR, respectively. (D) Representative scanning electron microscopy images of EGD-e and ScottA cells grown in indicated conditions.

ScottA (Figure 6B), respectively. Very few proteins were differentially regulated under high osmolarity compared to the control (Table S10). The complete list of differentially regulated proteins under all conditions and the GO categories enriched in these sets of proteins are included in Table S10 and Table S4, respectively. In general, the condition mimicking those found in the stomach (low pH/pepsin) resulted in more quantitative changes compared to changes observed under high osmolarity for both strains (Table S10). Also, more proteins were found to be differentially abundant upon perturbation in EGD-e compared to strain ScottA.

Several studies have investigated the effect of bile salts on *L. monocytogenes* cells, e.g., through gene knockout analyses, and several genes have been associated with resistance to bile (Table S11). The majority of these proteins are present in our spectral libraries, and one of them was specific to the EGD-e genome (CRH04_02340, which corresponds to the product of *pva*). Interestingly, at the proteotype level, only one of these proteins was found to be significantly differentially abundant when bacterial cells from both strains were incubated with bile salts compared to control condition (*bilEB*; EDG-e: CRH04_07370, ScottA: CRH05_07950; Tables S10 and S11). This likely indicates that the simplistic *in vitro* conditions developed only partially reflect the full complexity, e.g., of the gallbladder fluid, which in other studies had been used from sacrificed animals.[86] Another significantly differentially abundant protein specific to EGD-e in bile salts compared to control is CRH04_02210 (Q8Y9U9, *ftsW/rodA/spoVE* family cell cycle protein; Table S10). CRH04_02210 is known to be essential for cell division and peptidoglycan synthesis. In the past, deletion of this gene was found to lead to an abnormal cell shape and a higher susceptibility to antibiotics, which is related to virulence to some extent.[87,88] As for ScottA, CHR05_06055 and CRH05_14955 were found to be specifically up-regulated in bile salts compared to control. These two protein products, i.e., a CDP-glycerol glycerophosphotransferase and glycosyl transferase, respectively, are involved in serovar 4b specific teichoic acid synthesis and glycosylation and were recently reported to play a role in fitness and virulence.[89]

One of the most highly differentially regulated proteins in both strains upon bile salt treatment was flagellin, the structural protein of *Listeria* flagella. This result was unexpected given that flagella are known to be thermoregulated and not formed in temperatures higher than 30 °C.[90] The gene encoding flagellin (*flaA*) is located right in between two predicted operons of flagella and motility associated genes[41] (Table S11). The majority of these genes were also detected at the protein level and are thus represented in the spectral libraries. For ScottA, the products of three additional flagellar hook-associated genes (*flgE*, CRH05_04145; *flgK*, CRH05_04185; *flgL*, CRH05_04190) were found to be up-regulated upon incubation with bile salts. In contrast, in EGD-e, the products of the two component regulatory system (2CRS) *cheY* (CRH04_3620), a chemotaxis response regulator, and cheA (CRH04_03625) were up-regulated in bile and low pH, or bile condition, respectively. Moreover, the list of differentially abundant proteins from all conditions was compared to the list of unique, strain-specific genes to investigate whether strain-specific alterations of the proteotype exist. For EGD-e, nine proteins encoded by the corresponding strain specific genes were differentially abundant across bile (7) and low pH (3) conditions with CRH04_01625 (a DUF1433 domain-containing protein) up in both (Table S10). In contrast, 15 ScottA strain-specific proteins were differentially regulated upon exposure to bile salts, including three transcriptional regulators (CRH05_01760, CHR05_02245, CRH05_11180; Table S10). Together, these results suggest that different mechanisms are used by the strains in response to the bile salt stress.

Flagella Expression in Bile Salt Condition Could Hint at a Possible Escape/Survival Mechanism

The most striking phenotypic change that was observed in both strains based on the DIA data (Figure 6) was the significant increase of flagellin protein FlaA, the structural

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

protein of flagella, upon exposure to bile salts. The increased abundance of FlaA (Figure 6) was verified independently by parallel reaction monitoring mode (PRM) assays (Figure 7A). Additionally, as revealed by a PRM time-course experiment, protein levels of flagellin had increased after 15 min of incubation with bile salts (Figure 7B). A quantitative real-time PCR experiment was performed to assess *flaA* mRNA levels. Given that the FlaA protein levels had increased after 15 min of incubation in bile-containing medium, the qPCR was performed with bacterial cells grown in control and bile conditions for 5, 10, and 15 min. Strikingly, no significant upregulation of *flaA* mRNA was observed (Figure 7C). This suggests that protein production of FlaA is controlled at the post-transcriptional level.

To investigate whether the detected increase of FlaA protein upon bile salt exposure was the result of *de novo* formation of flagella, the cell surfaces of *L. monocytogenes* cultured in control, and bile salt conditions were stained and visualized using scanning electron microscopy (Figure 7D). As expected, the images confirmed flagella expression at 20 °C and the reduction or even absence of flagella at 37 °C. Rather unexpectedly, upon exposure to bile salts, an increase in flagella formation was readily observable also at higher temperatures such as 37 °C, confirming our protein expression data.

The above observations illustrate one valuable example of the additional unique information that can be obtained from quantitative next-gen proteomics but not from gene expression analysis. The presence of flagella under a bile salt stimulus could enable the bacteria to move away from that signal toward the intestinal mucus where they might subsequently lose their flagella and infect the host. Further experiments will be required to elucidate the function of flagella in the duodenum.

CONCLUSION

Here, we devised a generic proteogenomics strategy allowing researchers to investigate genotype−proteotype−phenotype relationships for EGD-e and ScottA, two major models of Listeria serotype 1/2a and 4b strains, which cause most cases of listeriosis. *De novo* assembly of complete genome sequences improved the previous fragmented, reference-based assembly for ScottA and allowed development of comprehensive catalogs of shared and unique genes through a comparative genomic analysis. Moreover, the integration of extensive proteotype data from DDA as well as DIA workflows for both strains provided evidence of missed protein coding genes, expressed pseudogenes, and novel start sites. Using this extensive proteogenomics toolbox, we investigated the *Listeria* proteotype in various states mimicking passage through the upper gastrointestinal (GI) tract. The quantitative analysis of bacterial protein abundance changes from different sections of the mimicked GI tract uncovered functionally relevant protein abundance changes related to bacterial motility that were not detected at the transcript level and that could have an impact on bacterial infectivity. Our proteogenomics resource will allow the *Listeria* community to uncover novel fascinating aspects of *Listeria* biology (or genotype−proteotype−phenotype relationships), some of which may be regulated exclusively at the proteotype level. It could for instance be applied in the context of mechanistic studies aiming at elucidating the molecular basis of pathogenesis in Listeria such as bacterial resistance to environmental stress. This might advance our understanding of Listeria−host interactions and beyond.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00842.

Figure S1. An incorrectly predicted pseudogene in strain ScottA. Figure S2. Hierarchical cluster analysis of biological replicates of EGD-e and ScottA. Table S1. Bacterial strains. Table S2. Overview of genome properties of strains EGD-e and ScottA. Table S6. Summary of annotation clusters in the iPtgxDBs. Table S8. Summary information on precursors, peptides, and protein groups identified for EGD-e and ScottA over all conditions using DIA. Table S9. Summary of library recovery percentage, data completeness, and median CVs for the DIA data set. (PDF)

Table S3. Overview of core genes and genes specific to ScottA and to EGD-e. (XLSX)

Table S4. Gene Ontology (GO) categories enriched among the strain-specific and differentially expressed proteins of each strain (XLSX)

Table S5. Master table. (XLSX)

Table S7. Proteomics evidence for the novelties identified through the proteogenomics search using MS-GF+ of DDA data for L. monocytogenes strains EGD-e and ScottA. (XLSX)

Table S10. List of differentially abundant proteins. (XLSX)

Table S11. Candidate genes for bile resistance and operons for flagellar genes. (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Christian H. Ahrens − *Agroscope, Molecular Diagnostics, Genomics & Bioinformatics, 8820 Wädenswil, Switzerland; Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland*; Email: christian.ahrens@agroscope.admin.ch

Bernd Wollscheid − *Department of Health Sciences and Technology (D-HEST) and Institute of Translational Medicine (ITM), ETH Zürich, 8092 Zürich, Switzerland; Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland;* orcid.org/0000-0002-3923-1610; Email: wbernd@ethz.ch

Authors

Adithi R. Varadarajan − *Department of Health Sciences and Technology (D-HEST), ETH Zürich, 8092 Zürich, Switzerland; Agroscope, Molecular Diagnostics, Genomics & Bioinformatics, 8820 Wädenswil, Switzerland; Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland;* orcid.org/0000-0002-9581-8598

Sandra Goetze − *Department of Health Sciences and Technology (D-HEST) and Institute of Translational Medicine (ITM), ETH Zürich, 8092 Zürich, Switzerland; Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland;* orcid.org/0000-0001-6880-8020

Maria P. Pavlou − *Department of Health Sciences and Technology (D-HEST) and Institute of Translational Medicine (ITM), ETH Zürich, 8092 Zürich, Switzerland*

Virginie Grosboillot − *Department of Health Sciences and Technology (D-HEST) and Institute of Food, Nutrition and Health (IFNH), ETH Zürich, 8092 Zürich, Switzerland*

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662

Yang Shen − *Department of Health Sciences and Technology (D-HEST) and Institute of Food, Nutrition and Health (IFNH), ETH Zürich, 8092 Zürich, Switzerland*

Martin J. Loessner − *Department of Health Sciences and Technology (D-HEST) and Institute of Food, Nutrition and Health (IFNH), ETH Zürich, 8092 Zürich, Switzerland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.9b00842

Author Contributions

M.P.P. performed all experiments except those noted below. V.G. performed qPCR. S.G. performed PRM experiments. A.R.V., M.P.P., S.G., and V.G. analyzed data. Y.S. and M.J.L. supervised V.G. and *Listeria* experiments in the Loessner laboratory. M.P.P. and B.W. conceived the project, C.H.A. conceived genomics and proteogenomics aspects. A.R.V., S.G., M.P.P, C.H.A., and B.W. designed research. C.H.A., A.R.V., S.G., and B.W. wrote and revised the manuscript. All authors commented and provided feedback on the manuscript. The first three authors (A.R.V., S.G., M.P.P.) and the last two authors (C.H.A. and B.W.) contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Ulrich Omasits for bioinformatic support in the early stage of the project and Michael Schmid (both Agroscope) for contributions to the *de novo* genome assemblies. We would like to thank Patrick Studer (ETH Zurich) who helped with culture of the *Listeria* strains and J. R. Wyatt for text editing. We thank Dr. Stephan Handschin at the Scientific Center for Optical and Electron Microscopy of ETH Zurich, for his technical assistance on scanning electron microscopy. B.W. acknowledges support from the Swiss National Science Foundation (SNSF) under grant 31003A_160259, C.H.A. acknowledges support from the SNSF for A.R.V. under grants 31003A-156320 and 188722.

■ REFERENCES

(1) Low, J. C.; Donachie, W. A Review of *Listeria Monocytogenes* and Listeriosis. *Vet. J.* **1997**, *153* (1), 9−29.

(2) Vázquez-Boland, J. A.; Kuhn, M.; Berche, P.; Chakraborty, T.; Domínguez-Bernal, G.; Goebel, W.; González-Zorn, B.; Wehland, J.; Kreft, J. *Listeria* Pathogenesis and Molecular Virulence Determinants. *Clin. Microbiol. Rev.* **2001**, *14* (3), 584−640.

(3) Hamon, M.; Bierne, H.; Cossart, P. *Listeria Monocytogenes*: A Multifaceted Model. *Nat. Rev. Microbiol.* **2006**, *4* (6), 423−434.

(4) de Noordhout, C. M.; Devleesschauwer, B.; Angulo, F. J.; Verbeke, G.; Haagsma, J.; Kirk, M.; Havelaar, A.; Speybroeck, N. The Global Burden of Listeriosis: A Systematic Review and Meta-Analysis. *Lancet Infect. Dis.* **2014**, *14* (11), 1073−1082.

(5) Datta, A. R.; Burall, L. S. Serotype to Genotype: The Changing Landscape of Listeriosis Outbreak Investigations. *Food Microbiol.* **2018**, *75*, 18−27.

(6) Olier, M.; Rousseaux, S.; Piveteau, P.; Lemaître, J.-P.; Rousset, A.; Guzzo, J. Screening of Glutamate Decarboxylase Activity and Bile Salt Resistance of Human Asymptomatic Carriage, Clinical, Food, and Environmental Isolates of *Listeria Monocytogenes*. *Int. J. Food Microbiol.* **2004**, *93* (1), 87−99.

(7) Begley, M.; Gahan, C. G. M.; Hill, C. The Interaction between Bacteria and Bile. *FEMS Microbiol. Rev.* **2005**, *29* (4), 625−651.

(8) Gahan, C. G. M.; Hill, C. *Listeria Monocytogenes*: Survival and Adaptation in the Gastrointestinal Tract. *Front. Cell. Infect. Microbiol.* **2014**, *4*, 9.

(9) Cossart, P. Illuminating the Landscape of Host-Pathogen Interactions with the Bacterium *Listeria Monocytogenes*. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (49), 19484−19491.

(10) Rolhion, N.; Cossart, P. How the Study of *Listeria Monocytogenes* Has Led to New Concepts in Biology. *Future Microbiol.* **2017**, *12*, 621−638.

(11) Leimeister-Wächter, M.; Domann, E.; Chakraborty, T. The Expression of Virulence Genes in *Listeria Monocytogenes* Is Thermoregulated. *J. Bacteriol.* **1992**, *174* (3), 947−952.

(12) Toledo-Arana, A.; Dussurget, O.; Nikitas, G.; Sesto, N.; Guet-Revillet, H.; Balestrino, D.; Loh, E.; Gripenland, J.; Tiensuu, T.; Vaitkevicius, K.; et al. The *Listeria* Transcriptional Landscape from Saprophytism to Virulence. *Nature* **2009**, *459* (7249), 950−956.

(13) Dabiri, G. A.; Sanger, J. M.; Portnoy, D. A.; Southwick, F. S. *Listeria Monocytogenes* Moves Rapidly through the Host-Cell Cytoplasm by Inducing Directional Actin Assembly. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87* (16), 6068−6072.

(14) Tilney, L. G.; Portnoy, D. A. Actin Filaments and the Growth, Movement, and Spread of the Intracellular Bacterial Parasite, *Listeria Monocytogenes*. *J. Cell Biol.* **1989**, *109*, 1597−1608.

(15) Khan, S. H.; Badovinac, V. P. *Listeria Monocytogenes*: A Model Pathogen to Study Antigen-Specific Memory CD8 T Cell Responses. *Semin. Immunopathol.* **2015**, *37* (3), 301−310.

(16) Moura, A.; Criscuolo, A.; Pouseele, H.; Maury, M. M.; Leclercq, A.; Tarr, C.; Björkman, J. T.; Dallman, T.; Reimer, A.; Enouf, V.; et al. Whole Genome-Based Population Biology and Epidemiological Surveillance of *Listeria Monocytogenes*. *Nat. Microbiol* **2017**, *2*, 16185.

(17) Schmid, M.; Frei, D.; Patrignani, A.; Schlapbach, R.; Frey, J. E.; Remus-Emsermann, M. N. P.; Ahrens, C. H. Pushing the Limits of *de Novo* Genome Assembly for Complex Prokaryotic Genomes Harboring Very Long, near Identical Repeats. *Nucleic Acids Res.* **2018**, *46* (17), 8953−8965.

(18) Nelson, K. E.; Fouts, D. E.; Mongodin, E. F.; Ravel, J.; DeBoy, R. T.; Kolonay, J. F.; Rasko, D. A.; Angiuoli, S. V.; Gill, S. R.; Paulsen, I. T.; et al. Whole Genome Comparisons of Serotype 4b and 1/2a Strains of the Food-Borne Pathogen *Listeria Monocytogenes* Reveal New Insights into the Core Genome Components of This Species. *Nucleic Acids Res.* **2004**, *32* (8), 2386−2395.

(19) Glaser, P.; Frangeul, L.; Buchrieser, C.; Rusniok, C.; Amend, A.; Baquero, F.; Berche, P.; Bloecker, H.; Brandt, P.; Chakraborty, T. Comparative Genomics of *Listeria* Species. *Science* **2001**, *294* (5543), 849−852.

(20) Impens, F.; Rolhion, N.; Radoshevich, L.; Bécavin, C.; Duval, M.; Mellin, J.; García Del Portillo, F.; Pucciarelli, M. G.; Williams, A. H.; Cossart, P. N-TerminomicsIdentifies Prli42 as a Membrane Miniprotein Conserved in Firmicutesand Critical for Stressosome Activation in *ListeriaMonocytogenes*. *Nat. Microbiol* **2017**, *2*, 17005.

(21) Vogel, C.; Marcotte, E. M. Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses. *Nat. Rev. Genet.* **2012**, *13* (4), 227−232.

(22) Fernández, N.; Cabrera, J. J.; Varadarajan, A. R.; Lutz, S.; Ledermann, R.; Roschitzki, B.; Eberl, L.; Bedmar, E. J.; Fischer, H.-M.; Pessi, G. An Integrated Systems Approach Unveils New Aspects of Microoxia-Mediated Regulation in *BradyrhizobiumDiazoefficiens*. *Front. Microbiol.* **2019**, *10*, 1 DOI: 10.3389/fmicb.2019.00924.

(23) Impens, F.; Radoshevich, L.; Cossart, P.; Ribet, D. Mapping of SUMO Sites and Analysis of SUMOylation Changes Induced by External Stimuli. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (34), 12432−12437.

(24) Kühbacher, A.; Novy, K.; Quereda, J. J.; Sachse, M.; Moya-Nilges, M.; Wollscheid, B.; Cossart, P.; Pizarro-Cerdá, J. Listeriolysin O-Dependent Host Surfaceome Remodeling Modulates *Listeria Monocytogenes* Invasion. *Pathog. Dis.* **2018**, *76* (8), 1 DOI: 10.1093/femspd/fty082.

(25) Röst, H. L.; Malmström, L.; Aebersold, R. Reproducible Quantitative Proteotype Data Matrices for Systems Biology. *Mol. Biol. Cell* **2015**, *26* (22), 3926−3931.

(26) Ahrens, C. H.; Brunner, E.; Qeli, E.; Basler, K.; Aebersold, R. Generating and Navigating Proteome Maps Using Mass Spectrometry. *Nat. Rev. Mol. Cell Biol.* **2010**, *11* (11), 789−801.
(27) Omasits, U.; Quebatte, M.; Stekhoven, D. J.; Fortes, C.; Roschitzki, B.; Robinson, M. D.; Dehio, C.; Ahrens, C. H. Directed Shotgun Proteomics Guided by Saturated RNA-Seq Identifies a Complete Expressed Prokaryotic Proteome. *Genome Res.* **2013**, *23* (11), 1916−1927.
(28) Schmidt, A.; Kochanowski, K.; Vedelaar, S.; Ahrné, E.; Volkmer, B.; Callipo, L.; Knoops, K.; Bauer, M.; Aebersold, R.; Heinemann, M. The Quantitative and Condition-Dependent *Escherichia Coli* Proteome. *Nat. Biotechnol.* **2016**, *34* (1), 104−110.
(29) Aebersold, R.; Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **2016**, *537* (7620), 347−355.
(30) Bécavin, C.; Koutero, M.; Tchitchek, N.; Cerutti, F.; Lechat, P.; Maillet, N.; Hoede, C.; Chiapello, H.; Gaspin, C.; Cossart, P. Listeriomics: An Interactive Web Platform for Systems Biology of *Listeria*. *mSystems* **2017**, *2* (2), 1 DOI: 10.1128/mSystems.00186-16.
(31) Misra, S. K.; MoussanDésiréeAké, F.; Wu, Z.; Milohanic, E.; Cao, T. N.; Cossart, P.; Deutscher, J.; Monnet, V.; Archambaud, C.; Henry, C. Quantitative Proteome Analyses Identify PrfA-Responsive Proteins and Phosphoproteins in *Listeria Monocytogenes*. *J. Proteome Res.* **2014**, *13* (12), 6046−6057.
(32) Malet, J. K.; Impens, F.; Carvalho, F.; Hamon, M. A.; Cossart, P.; Ribet, D. Rapid Remodeling of the Host Epithelial Cell Proteome by the Listeriolysin O (LLO) Pore-Forming Toxin. *Mol. Cell. Proteomics* **2018**, *17* (8), 1627−1636.
(33) Melo, J.; Schrama, D.; Hussey, S.; Andrew, P. W.; Faleiro, M. L. *Listeria Monocytogenes* Dairy Isolates Show a Different Proteome Response to Sequential Exposure to Gastric and Intestinal Fluids. *Int. J. Food Microbiol.* **2013**, *163* (2−3), 51−63.
(34) He, L.; Deng, Q.-L.; Chen, M.-T.; Wu, Q.-P.; Lu, Y.-J. Proteomics Analysis of *Listeria Monocytogenes* ATCC 19115 in Response to Simultaneous Triple Stresses. *Arch. Microbiol.* **2015**, *197* (6), 833−841.
(35) Miyamoto, K. N.; Monteiro, K. M.; da Silva Caumo, K.; Lorenzatto, K. R.; Ferreira, H. B.; Brandelli, A. Comparative Proteomic Analysis of *Listeria Monocytogenes* ATCC 7644 Exposed to a Sublethal Concentration of Nisin. *J. Proteomics* **2015**, *119*, 230−237.
(36) Mata, M. M.; da Silva, W. P.; Wilson, R.; Lowe, E.; Bowman, J. P. Attached and Planktonic *Listeria Monocytogenes* Global Proteomic Responses and Associated Influence of Strain Genetics and Temperature. *J. Proteome Res.* **2015**, *14* (2), 1161−1173.
(37) Aguirre, J. S.; García de Fernando, G.; Hierro, E.; Hospital, X. F.; Espinosa, I.; Fernández, M. Characterization of Damage on *Listeria Innocua* Surviving to Pulsed Light: Effect on Growth, DNA and Proteome. *Int. J. Food Microbiol.* **2018**, *284*, 63−72.
(38) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* **2012**, *11* (6), O111.016717.
(39) Malmström, L.; Bakochi, A.; Svensson, G.; Kilsgård, O.; Lantz, H.; Petersson, A. C.; Hauri, S.; Karlsson, C.; Malmström, J. Quantitative Proteogenomics of Human Pathogens Using DIA-MS. *J. Proteomics* **2015**, *129*, 98−107.
(40) Schubert, O. T.; Ludwig, C.; Kogadeeva, M.; Zimmermann, M.; Rosenberger, G.; Gengenbacher, M.; Gillet, L. C.; Collins, B. C.; Röst, H. L.; Kaufmann, S. H. E.; et al. Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of *Mycobacterium Tuberculosis*. *Cell Host Microbe* **2015**, *18* (1), 96−108.
(41) Bécavin, C.; Bouchier, C.; Lechat, P.; Archambaud, C.; Creno, S.; Gouin, E.; Wu, Z.; Kühbacher, A.; Brisse, S.; Pucciarelli, M. G.; et al. Comparison of Widely Used *Listeria Monocytogenes* Strains EGD, 10403S, and EGD-E Highlights Genomic Variations Underlying Differences in Pathogenicity. *mBio* **2014**, *5* (2), No. e00969−14.
(42) Briers, Y.; Klumpp, J.; Schuppler, M.; Loessner, M. J. Genome Sequence of *Listeria Monocytogenes*ScottA, a Clinical Isolate from a Food-Borne Listeriosis Outbreak. *J. Bacteriol.* **2011**, *193* (16), 4284−4285.
(43) Barbosa, J.; Borges, S.; Magalhães, R.; Ferreira, V.; Santos, I.; Silva, J.; Almeida, G.; Gibbs, P.; Teixeira, P. Behaviour of *Listeria Monocytogenes* Isolates through Gastro-Intestinal Tract Passage Simulation, before and after Two Sub-Lethal Stresses. *Food Microbiol.* **2012**, *30* (1), 24−28.
(44) Chin, C.-S.; Alexander, D. H.; Marks, P.; Klammer, A. A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E. E.; et al. Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data. *Nat. Methods* **2013**, *10* (6), 563−569.
(45) Omasits, U.; Varadarajan, A. R.; Schmid, M.; Goetze, S.; Melidis, D.; Bourqui, M.; Nikolayeva, O.; Québatte, M.; Patrignani, A.; Dehio, C.; et al. An Integrative Strategy to Identify the Entire Protein Coding Potential of Prokaryotic Genomes by Proteogenomics. *Genome Res.* **2017**, *27* (12), 2083−2095.
(46) Schmid, M.; Muri, J.; Melidis, D.; Varadarajan, A. R.; Somerville, V.; Wicki, A.; Moser, A.; Bourqui, M.; Wenzel, C.; Eugster-Meier, E.; et al. Comparative Genomics of Completely Sequenced *Lactobacillus Helveticus* Genomes Provides Insights into Strain-Specific Genes and Resolves Metagenomics Data Down to the Strain Level. *Front. Microbiol.* **2018**, *9*, 63.
(47) Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv* (q-bio.GN) May 26, **2013**, 1303.3997, ver. 2.
(48) Tatusova, T.; DiCuccio, M.; Badretdin, A.; Chetvernin, V.; Nawrocki, E. P.; Zaslavsky, L.; Lomsadze, A.; Pruitt, K. D.; Borodovsky, M.; Ostell, J. NCBI Prokaryotic Genome Annotation Pipeline. *Nucleic Acids Res.* **2016**, *44* (14), 6614−6624.
(49) Jones, P.; Binns, D.-Y.; Chang, H.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* **2014**, *30* (9), 1236−1240.
(50) Huerta-Cepas, J.; Forslund, K.; Coelho, L. P.; Szklarczyk, D.; Jensen, L. J.; von Mering, C.; Bork, P. Fast Genome-Wide Functional Annotation through Orthology Assignment by egg NOG-Mapper. *Mol. Biol. Evol.* **2017**, *34* (8), 2115−2122.
(51) Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M. C.; Rattei, T.; Mende, D. R.; Sunagawa, S.; Kuhn, M.; et al. eggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences. *Nucleic Acids Res.* **2016**, *44* (D1), D286−D293.
(52) Arndt, D.; Grant, J. R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D. S. PHASTER: A Better, Faster Version of the PHAST Phage Search Tool. *Nucleic Acids Res.* **2016**, *44* (W1), W16−W21.
(53) Alexa, A.; Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. *R Package* ver 2.34.0.; 2018.
(54) Alexa, A.; Rahnenfuhrer, J.; Lengauer, T. Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure. *Bioinformatics* **2006**, *22* (13), 1600−1607.
(55) Page, A. J.; Cummins, C. A.; Hunt, M.; Wong, V. K.; Reuter, S.; Holden, M. T. G.; Fookes, M.; Falush, D.; Keane, J. A.; Parkhill, J. Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* **2015**, *31* (22), 3691−3693.
(56) Loessner, M. J.; Wendlinger, G.; Scherer, S. Heterogeneous Endolysins in *Listeria Monocytogenes* Bacteriophages: A New Class of Enzymes and Evidence for Conserved Holin Genes within the Siphoviral Lysis Cassettes. *Mol. Microbiol.* **1995**, *16* (6), 1231−1241.
(57) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-Translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Mol. Cell. Proteomics* **2006**, *5* (5), 935−948.
(58) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid

Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.
(59) Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K. MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *J. Proteome Res.* **2014**, *13* (8), 3679−3684.
(60) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8* (6), 3176−3181.
(61) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinović, S. M.; Cheng, L.-Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; et al. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **2015**, *14* (5), 1400−1410.
(62) Hyatt, D.; Chen, G.-L.; Locascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinf.* **2010**, *11*, 119.
(63) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(64) Qeli, E.; Ahrens, C. H. Peptide Classifier for Protein Inference and Targeted Quantitative Proteomics. *Nat. Biotechnol.* **2010**, *28* (7), 647−650.
(65) Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nat. Methods* **2014**, *11* (11), 1114−1125.
(66) Choi, M.; Chang, C.-Y.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; Vitek, O. MSstats: An R Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Bioinformatics* **2014**, *30* (17), 2524−2526.
(67) Griffiths, R. I.; Whiteley, A. S.; O'Donnell, A. G.; Bailey, M. J. Rapid Method for Coextraction of DNA and RNA from Natural Environments for Analysis of Ribosomal DNA- and rRNA-Based Microbial Community Composition. *Appl. Environ. Microbiol.* **2000**, *66* (12), 5488−5491.
(68) Alonzo, F., 3rd; Bobo, L. D.; Skiest, D. J.; Freitag, N. E. Evidence for Subpopulations of *Listeria Monocytogenes* with Enhanced Invasion of Cardiac Cells. *J. Med. Microbiol.* **2011**, *60*, 423−434.
(69) Varadarajan, A. R.; Allan, R. N.; Valentin, J. D. P.; Castañeda Ocampo, O. E.; Somerville, V.; Pietsch, F.; Buhmann, M. T.; West, J.; Skipp, P. J.; van der Mei, H. C.; et al. An Integrated Model System to Gain Mechanistic Insights into Biofilm Formation and Antimicrobial Resistance Development in Pseudomonas Aeruginosa MPAO1. *bioRxiv* **2020**, DOI: 10.1101/2020.02.06.936690.
(70) Jacobs, M. A.; Alwood, A.; Thaipisuttikul, I.; Spencer, D.; Haugen, E.; Ernst, S.; Will, O.; Kaul, R.; Raymond, C.; Levy, R.; et al. Comprehensive Transposon Mutant Library of Pseudomonas Aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (24), 14339−14344.
(71) Duval, M.; Cossart, P. Small Bacterial and Phagic Proteins: An Updated View on a Rapidly Moving Field. *Curr. Opin. Microbiol.* **2017**, *39*, 81−88.
(72) Storz, G.; Wolf, Y. I.; Ramamurthi, K. S. Small Proteins Can No Longer Be Ignored. *Annu. Rev. Biochem.* **2014**, *83* (1), 753−777.
(73) Jaffe, J. D.; Berg, H. C.; Church, G. M. Proteogenomic Mapping as a Complementary Method to Perform Genome Annotation. *Proteomics* **2004**, *4* (1), 59−77.
(74) Kyrpides, N. C.; Hugenholtz, P.; Eisen, J. A.; Woyke, T.; Göker, M.; Parker, C. T.; Amann, R.; Beck, B. J.; Chain, P. S. G.; Chun, J.; et al. Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. *PLoS Biol.* **2014**, *12* (8), No. e1001920.
(75) Ross, M. G.; Russ, C.; Costello, M.; Hollinger, A.; Lennon, N. J.; Hegarty, R.; Nusbaum, C.; Jaffe, D. B. Characterizing and Measuring Bias in Sequence Data. *Genome Biol.* **2013**, *14* (5), R51.
(76) Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **2009**, *323* (5910), 133−138.
(77) Jain, M.; Koren, S.; Miga, K. H.; Quick, J.; Rand, A. C.; Sasani, T. A.; Tyson, J. R.; Beggs, A. D.; Dilthey, A. T.; Fiddes, I. T.; et al. Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. *Nat. Biotechnol.* **2018**, *36*, 338−345.
(78) Loessner, M. J.; Schneider, A.; Scherer, S. Modified *Listeria* Bacteriophage Lysin Genes (*ply*) Allow Efficient Overexpression and One-Step Purification of Biochemically Active Fusion Proteins. *Appl. Environ. Microbiol.* **1996**, *62* (8), 3057−3060.
(79) Donaldson, J. R.; Nanduri, B.; Burgess, S. C.; Lawrence, M. L. Comparative Proteomic Analysis of *Listeria Monocytogenes* Strains F2365 and EGD. *Appl. Environ. Microbiol.* **2009**, *75* (2), 366−373.
(80) Gupta, N.; Pevzner, P. A. False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule. *J. Proteome Res.* **2009**, *8* (9), 4173−4181.
(81) Grobei, M. A.; Qeli, E.; Brunner, E.; Rehrauer, H.; Zhang, R.; Roschitzki, B.; Basler, K.; Ahrens, C. H.; Grossniklaus, U. Deterministic Protein Inference for Shotgun Proteomics Data Provides New Insights into Arabidopsis Pollen Development and Function. *Genome Res.* **2009**, *19* (10), 1786−1800.
(82) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (11), 2405−2417.
(83) Miravet-Verde, S.; Ferrar, T.; Espadas-García, G.; Mazzolini, R.; Gharrab, A.; Sabido, E.; Serrano, L.; Lluch-Senar, M. Unraveling the Hidden Universe of Small Proteins in Bacterial Genomes. *Mol. Syst. Biol.* **2019**, *15* (2), No. e8290.
(84) Carlier, A. L.; Omasits, U.; Ahrens, C. H.; Eberl, L. Proteomics Analysis of Psychotria Leaf Nodule Symbiosis: Improved Genome Annotation and Metabolic Predictions. *Mol. Plant-Microbe Interact.* **2013**, *26* (11), 1325−1333.
(85) Čuklina, J.; Hahn, J.; Imakaev, M.; Omasits, U.; Förstner, K. U.; Ljubimov, N.; Goebel, M.; Pessi, G.; Fischer, H.-M.; Ahrens, C. H.; et al. Genome-Wide Transcription Start Site Mapping of *Bradyrhizobium Japonicum* Grown Free-Living or in Symbiosis - a Rich Resource to Identify New Transcripts, Proteins and to Study Gene Regulation. *BMC Genomics* **2016**, *17*, 302.
(86) Dowd, G. C.; Joyce, S. A.; Hill, C.; Gahan, C. G. M. Investigation of the Mechanisms by Which *Listeria Monocytogenes* Grows in Porcine Gallbladder Bile. *Infect. Immun.* **2011**, *79* (1), 369−379.
(87) Rismondo, J.; Halbedel, S.; Gründling, A. Cell Shape and Antibiotic Resistance Are Maintained by the Activity of Multiple FtsW and RodA Enzymes in Listeria Monocytogenes. *mBio* **2019**, *10* (4), 1 DOI: 10.1128/mBio.01448-19.
(88) Burke, T. P.; Portnoy, D. A. SpoVG Is a Conserved RNA-Binding Protein That Regulates Listeria Monocytogenes Lysozyme Resistance, Virulence, and Swarming Motility. *mBio* **2016**, 1 DOI: 10.1128/mBio.00240-16.
(89) Sumrall, E. T.; Shen, Y.; Keller, A. P.; Rismondo, J.; Pavlou, M.; Eugster, M. R.; Boulos, S.; Disson, O.; Thouvenot, P.; Kilcher, S.; et al. Phage Resistance at the Cost of Virulence: Listeria Monocytogenes Serovar 4b Requires Galactosylated Teichoic Acids for InlB-Mediated Invasion. *PLoS Pathog.* **2019**, *15* (10), No. e1008032.
(90) Lemon, K. P.; Higgins, D. E.; Kolter, R. Flagellar Motility Is Critical for *Listeria Monocytogenes* Biofilm Formation. *J. Bacteriol.* **2007**, *189* (12), 4418−4424.

https://dx.doi.org/10.1021/acs.jproteome.9b00842
J. Proteome Res. 2020, 19, 1647−1662