

# **Mass Spectrometry-Based Proteomic Analysis of Selected Bacteria from the Human Gut Microbiome**

**Dissertation**

zur Erlangung des Doktorgrades  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Jerome Genth

Kiel, 2024



Erster Gutachter (First reviewer):

Prof. Dr. Andreas Tholey

Zweiter Gutachter (Second reviewer):

Prof. Dr. Ruth Schmitz-Streit

Tag der Disputation (Date of defense):

16.10.2024

Zum Druck genehmigt (Approved for publication):

16.10.2024



*„Who you are is defined by what you're willing to struggle for.”*

**- Mark Manson**





Human gut bacteria live in a highly dynamic environment with frequent changes such as variations in pH or nutrient availability. To ensure their survival and functionality under these fluctuating conditions, they constantly adapt their proteomes by adjusting the abundance of essential proteins. However, the specific proteomic changes in individual human gut microbiome (HGM) members under different *in vitro* conditions remain largely unexplored. This study employs mass spectrometry-based proteomic analysis to address this gap on the selected members of the human gut microbiome.

Initially, the reproducibility and accuracy of bottom-up label-free quantification (LFQ) and tandem mass tag (TMT)-based quantification were assessed in quantifying proteomic changes of *Bacteroides thetaiotaomicron* induced by sucrose and glucose. Both methods achieved comparable results, indicating that this bacterium primarily alters the abundance of proteins involved in the machinery required to utilize the provided carbon sources.

The LFQ approach was then applied to examine proteomic changes in *B. thetaiotaomicron*, *Blautia producta*, and *Bifidobacterium longum* in response to different environmental pH levels. Distinct and concurrent alterations in pathway-related and stress-associated proteins were identified, including histidine biosynthesis in *B. producta*, nitrogen metabolism in *B. thetaiotaomicron*, and inositol carbohydrate metabolism in *B. thetaiotaomicron* and *B. producta*. The comparative analysis of bottom-up and top-down proteomics in *B. producta* demonstrated their ability to quantify the same proteins, whether differentially abundant or not.

The third project focused on identifying novel proteins, specifically short open reading frame-encoded peptides (SEP), in *B. producta*. A proteogenomic approach was used to analyze their presence under various cultivation conditions, including different media (BHI and YCFA), pH levels, and supplemented factors (yeast extract, SCFAs, and LPS). Stringent validation criteria ensured accurate identification of SEP, and biochemical predictions explored their potential functional roles. The combined bottom-up and top-down proteomics analyses identified a total of 45 SEP, including previously reported SEP (BP1 to BP14). This study demonstrated that the production of certain SEP in *B. producta* is influenced by specific environmental factors rather than solely by interspecies interactions, as previously suggested.

The last project established and optimized an isolation protocol for the LC-MS-based proteomic analysis of extracellular vesicles. Bottom-up proteomic analysis of outer membrane vesicles (OMVs) isolated from *E. coli* under different culture conditions and growth phases quantified OMV marker proteins and classified the subcellular topological distribution of the

*E. coli* OMV proteomes. Nanoparticle tracking analysis validated OMV nanoparticles, and an *in vitro* wound healing assay with human colonic Caco-2 cells suggested a dose-dependent inhibitory effect of OMVs on wound closure. Various sample preparation protocols were evaluated, and the integration of non-ionic detergents for OMV lysis and proteolytic digestion improved protein and peptide identifications, making the improved in-solution digestion protocol the preferred method for future analyses.

In most projects, a proteoform-directed top-down analysis, including a discovery-based open modification search, was applied. This analysis identified several potential post-translational modifications (e.g.,  $\beta$ -methylthio-aspartic acid on *B. thetaiotaomicron* ribosomal protein S12 and seryl-phosphorylations on *B. producta* HPr proteins) and neo-termini (potential artificial cleavage events, initiator methionine excisions, and alternative initiation sites).

Menschliche Darmbakterien leben in einer hochdynamischen Umgebung mit häufigen Veränderungen wie Schwankungen im pH-Wert oder der Nährstoffverfügbarkeit. Um die Überlebensfähigkeit und Funktionalität unter diesen wechselhaften Bedingungen zu gewährleisten, ist eine kontinuierliche Anpassung des Proteoms erforderlich. Allerdings sind die spezifischen proteomischen Veränderungen einzelner Mitglieder des menschlichen Darmmikrobioms (HGM) unter verschiedenen *in vitro*-Bedingungen weitgehend unerforscht. Zur Schließung dieser Forschungslücke, wurden in der vorliegenden Arbeit unter Verwendung von massenspektrometrischen Proteomanalysen ausgewählte Mitglieder des menschlichen Darmmikrobioms untersucht.

In einer ersten Studie wurden die Reproduzierbarkeit und Genauigkeit der markierungsfreien Quantifizierung (LFQ) und der Tandem-Mass-Tag (TMT)-basierten Quantifizierung untersucht. Ziel war es, proteomische Veränderungen in *Bacteroides thetaiotaomicron* zu analysieren, die durch Saccharose und Glukose induziert wurden. Beide Methoden erzielten vergleichbare Ergebnisse, was darauf hinweist, dass dieses Bakterium hauptsächlich die Menge an Proteinen verändert, die für die Verwertung der bereitgestellten Kohlenstoffquellen erforderlich sind.

Unter Verwendung der LFQ-Methode wurden proteomische Veränderungen in *B. thetaiotaomicron*, *Blautia producta* und *Bifidobacterium longum* als Reaktion auf unterschiedliche pH-Werte der Umgebung untersucht. Dadurch konnten unterschiedliche und gleiche Veränderungen in Stoffwechselweg-bezogenen und Stress-assoziierten Proteinen identifiziert werden, einschließlich der Histidinbiosynthese in *B. producta*, des Stickstoffstoffwechsels in *B. thetaiotaomicron* und des Inositolstoffwechsels in *B. thetaiotaomicron* und *B. producta*. Die vergleichende Analyse von *Bottom-up* und *Top-down* Proteomik in *B. producta* weist darauf hin, dass beide Methoden dieselben Proteine quantifizieren können, unabhängig davon, ob diese signifikant unterschiedlich häufig vorkommen oder nicht.

Das dritte Projekt konzentrierte sich auf die Identifizierung neuer Proteine, insbesondere kurzer offener Leserahm-kodierter Peptide (SEP), in *B. producta*. Ein proteogenomischer Ansatz wurde verwendet, um die Translation von SEP unter verschiedenen Kulturbedingungen, einschließlich verschiedener Medien (BHI und YCFA), pH-Werte und ergänzten Faktoren (Hefeextrakt, SCFAs und LPS), zu analysieren. Strenge Validierungskriterien gewährleisteten die genaue Identifizierung von SEP, und durch

biochemische Vorhersagen wurden potenzielle funktionelle Rollen untersucht. Die kombinierten *Bottom-up* und *Top-down* Proteomik-Analysen identifizierten insgesamt 45 SEP, einschließlich zuvor berichteter SEP (BP1 bis BP14). Durch diese Studie konnte gezeigt werden, dass die Produktion bestimmter SEP auch durch spezifische Umweltfaktoren beeinflusst wird und nicht nur, wie bisher angenommen, durch interspezifische Interaktionen.

Innerhalb des letzten Projektes wurde ein Isolationsprotokoll für die LC-MS-basierte Proteomanalyse extrazellulärer Vesikel etabliert und optimiert. Die *Bottom-up* Proteomanalyse von *E. coli*-Membranvesikeln (OMVs) unter verschiedenen Kulturbedingungen und Wachstumsphasen ermöglichte die Quantifizierung der OMV-Markerproteine und die Klassifizierung der subzellulären topologischen Verteilung der *E. coli* OMV-Proteome. Nanopartikel-Tracking-Analysen validierten OMV-Nanopartikel, und ein *in vitro* Wundheilungsassay mit menschlichen Caco-2-Zellen deutete auf eine dosisabhängige Hemmwirkung von OMVs auf den Wundverschluss hin. Die Evaluation verschiedener Probenvorbereitungsprotokolle ergab, dass die Integration nichtionischer Detergenzien zur OMV-Lyse und proteolytischen Verdauung zu einer deutlichen Steigerung der Protein- und Peptididentifikation führte. Das verbesserte In-Lösung-Protokoll stellte die bevorzugte Methode für zukünftige Analysen dar.

In den meisten Projekten wurde eine Proteoform-gerichtete *Top-down*-Analyse, einschließlich einer entdeckungsbasierten offenen Modifikationssuche, angewendet. Durch diese Analysen konnten mehrere potenzielle posttranslationale Modifikationen (z.B.  $\beta$ -Methylthio-Asparaginsäure auf dem *B. thetaiotaomicron* Ribosomprotein S12 und Seryl-Phosphorylierungen auf *B. producta* HPr-Proteinen) und Neo-Termini (mögliche künstliche Spaltungsereignisse, Abspaltung des initialen Methionins und alternative Initiationsstellen) identifiziert werden.

## DANKSAGUNG

---

Zunächst möchte ich meinem Betreuer, Prof. Dr. Andreas Tholey, meinen aufrichtigen Dank aussprechen. Nicht nur dafür, dass er mir die Gelegenheit gegeben hat in seiner Gruppe zu arbeiten, sondern auch für seine endlose Begeisterung für die Wissenschaft. Ich danke ihm für seine Hilfe bei der Vorbereitung von Präsentationen und dem Verfassen wissenschaftlicher Arbeiten. Er war stets freundlich und geduldig in der Beantwortung all meiner Fragen und bereit bei Herausforderungen zu helfen. Ohne seine große Unterstützung hätte ich niemals das erreicht, was ich erreicht habe.

Ein herzliches Dankeschön geht auch an Prof. Dr. Ruth Schmitz-Streit, die tapfer das Abenteuer einging, das Zweitgutachten meiner Dissertation anzufertigen. Ich hoffe dies war keine zu große Achterbahnfahrt!

Mein aufrichtiger Dank gilt auch Prof. Dr. Jan Rupp, Dr. Simon Graspeutner und Kathrin Schäfer, die bakterielle Kulturen durchführten und mikrobielle Lysate bereitstellten, die in dieser Studie untersucht wurden. Ein Teil dieser Arbeit wäre ohne ihre Hilfe unmöglich gewesen. Generell möchte ich jedem Mitglied der DFG-Forschungseinheit 5042 "miTarget" danken, insbesondere Eike Zell für ihre administrative Unterstützung.

Ich hatte das außergewöhnliche Glück, mit meinen Kollegen am Institut für Experimentelle Medizin zusammenzuarbeiten. Ihre Unterstützung war unentbehrlich für den Erfolg dieser Arbeit. Besonders möchte ich Dr. Christian Treitz, Dr. Liam Cassidy, Dr. Mohammad Abukhalaf und Dr. Tomas Koudelka für ihre herausragenden Fähigkeiten beim Betrieb von Massenspektrometer und ihrer Hilfe bei der Datenanalyse danken. Ein herzliches Dankeschön geht auch an Dr. Phillip Kaulich, Dr. Stephanie Bilke, Patrick Kaleja, Max Steinbach und Theo Matzanke für ihre ständig gute Laune, inspirierenden Diskussionen und experimentellen Details. Es war stets eine Freude, mit euch zusammenzuarbeiten! Darüber hinaus danke ich Britta Steer für ihre Unterstützung bei der Zellkultur.

Ich möchte auch Prof. Dr. Ottmar Janßen für seine Anleitung bei Nanosight-Messungen und Priv.-Doz. Dr. Marcus Lettau für die Unterstützung bei der Analyse von Nanosight-Daten danken.

Ein riesiges Dankeschön an meine Eltern, meine große Schwester und meine Großeltern, die immer an mich geglaubt haben. Ihr seid die Besten und die wahren Helden meiner Dissertation!

Ein fettes Dankeschön an meine Freunde, die mich größtenteils seit meiner Geburt oder der Schulzeit auf diesem Weg begleitet haben. Danke für unzählige Momente, endlose Motivation und Unterstützung, wenn die Forschung mal wieder zum Verzweifeln war. Ihr seid die besten Freunde, die ich mir wünschen kann! Cheers auf euch!



# **LIST OF PUBLICATIONS AND CONFERENCE CONTRIBUTIONS**

---

## **List of publications**

Genth J, Kaleja P, Treitz C, Schäfer K, Graspentner S, Rupp J, Tholey A. (2022)

The intracellular proteome of the gut bacterium *Bacteroides thetaiotaomicron* is widely unaffected by a switch from glucose to sucrose as main carbohydrate source. *Proteomics*, 22(22), 1–6.

Genth J, Schäfer K, Cassidy L, Graspentner S, Rupp J, Tholey A. (2023)

Identification of proteoforms of short open reading frame-encoded peptides in *Blautia producta* under different cultivation conditions. *Microbiology Spectrum*, 11(6), e0252823

## **Conference contributions**

Label-Free and Isobaric Labeling Approaches for Quantitative Analysis of the *Bacteroides thetaiotaomicron* Proteome

3. International RTG Symposium, May 2021, online

*Blautia producta* Displays Great Plasticity of Short Open Reading Frame-encoded Peptides under Different Environmental Conditions

54th Annual Conference of the DGMS, May 2023, Dortmund, Deutschland

pH-Induced Alterations in the Proteomes of Three Major Human Gut Bacteria

International Symposium of the RU miTarget, June 2023, Kiel, Deutschland



# TABLE OF CONTENTS

---

Abstract .....	i
Zusammenfassung .....	iii
Danksagung .....	v
List of Publications and Conference Contributions .....	vii
Table of Contents .....	ix
<b>I    GENERAL INTRODUCTION .....</b>	<b>1</b>
<b>II    GENERAL METHODS .....</b>	<b>17</b>
<b>III    PROTEOMIC ANALYSIS OF <i>B. THETA</i> IOTAOMICRON.....</b>	<b>37</b>
<b>IV    INFLUENCE OF PH ON BACTERIAL PROTEOMES .....</b>	<b>63</b>
<b>V    PROTEOGENOMIC ANALYSIS OF <i>B. PRODUCTA</i> .....</b>	<b>103</b>
<b>VI    OUTER MEMBRANE VESICLES ANALYSIS .....</b>	<b>121</b>
Bibliography .....	I
List of Abbreviations .....	XXIII
List of Figures .....	XXV
List of Tables .....	XXVIII
Appendix.....	XXIX

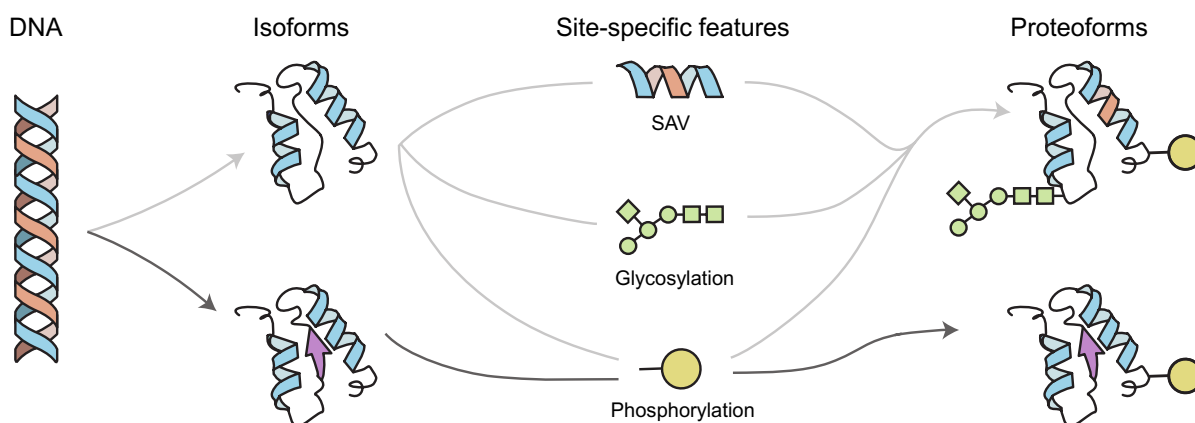


<b>1</b>	<b>The Dynamic Nature of the Proteome.....</b>	<b>2</b>
1.1	Proteome Variation.....	3
<b>2</b>	<b>Principles of Mass Spectrometry .....</b>	<b>5</b>
2.1	Online LC-MS .....	5
2.2	Mass-to-charge Analysis .....	5
2.3	Tandem Mass Spectrometry and Fragmentation Analysis.....	7
<b>3</b>	<b>Analysis of Mass Spectrometry Data.....</b>	<b>8</b>
3.1	Data Acquisition Techniques.....	9
3.2	Peptidoform and Protein Identification .....	10
3.3	Quantitative Techniques in Proteomics .....	11
<b>4</b>	<b>Inflammatory Bowel Disease .....</b>	<b>12</b>
4.1	The Human Gut Microbiome as a Therapeutic Target.....	12
4.2	Proteomics in IBD Research .....	14
<b>5</b>	<b>Objective of the Thesis .....</b>	<b>15</b>

# 1 The Dynamic Nature of the Proteome

The genome contains all genetic information and typically remains largely stable within a specific cell. In contrast, the proteome, representing the complete set of proteins (Wasinger et al., 1995), is constantly changing. This constant alteration of the proteomic landscape is characterized by continuous protein synthesis, degradation, and modification. The ability of cells to adjust their protein composition continuously in response to changing needs and influences, both internal and external, is a key aspect of cellular flexibility and crucial for a cell to effectively carry out its functions.

The fundamental process of transcribing DNA (deoxyribonucleic acid) into mRNA (messenger ribonucleic acid) molecules and then translating them into amino acid sequences results in a wide variety of proteins. Each protein is assembled from the repertoire of 22 proteinogenic amino acids, with each amino acid imparting unique physico-chemical characteristics to the resulting protein. The term 'protein' is essentially a generic term referring to a canonical amino acid sequence (Cassidy et al., 2023). With an increased understanding of the proteome, it becomes evident that the traditional notion of 'one-gene, one-protein, one-function' is inadequate to fully comprehend the complexity of the proteome (Carbonara et al., 2021; Cassidy et al., 2023). Recognizing this complexity, terms such as 'protein species' (Schlüter et al., 2009) and 'proteoforms' (Smith and Kelleher, 2013) have emerged, with 'proteoforms' gaining widespread acceptance (Carbonara et al., 2021; Marx, 2024). The proteoform concept includes a broad spectrum of molecular variations resulting from co- and post-translational modifications (PTMs) and sequence variants, which go beyond transcription and translation errors (FIGURE I-1). Understanding the function of diverse proteoforms is essential for comprehending cellular dynamics and processes (Marx, 2024).



**FIGURE I-1 | Sources of Proteome Complexity.** Isoform variation of the same gene combined with site-specific changes generate a variety of proteoforms. Abbreviation: SAV (single amino acid variation). Adapted with permission from (Aebersold et al., 2018).

## 1.1 Proteome Variation

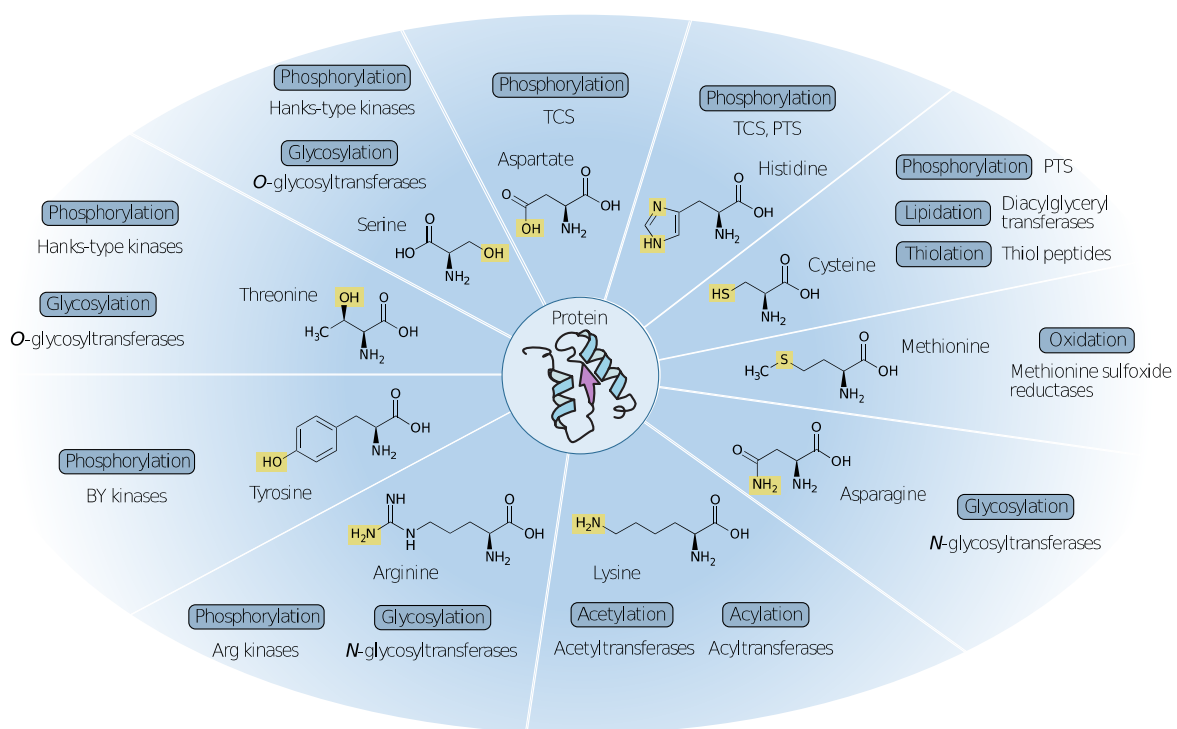
Organisms have developed many effective strategies to diversify the proteome without increasing the size of the genome, using mechanisms that generate multiple proteins from a single gene.

**Gene- and Transcript-level Variation** – Transcriptional read-through, a fundamental mechanism, allows RNA polymerases to bypass termination signals and transcribe multiple operons within a single mRNA molecule (Wade and Grainger, 2014). In response to environmental cues, bacteria dynamically exhibit transcriptional read-through to adapt to changing conditions (Junier and Rivoire, 2016). In eukaryotes, transcriptional complexity arises not only from transcript elongation but also from mRNA splicing, during which internal exons are removed. Variations in transcripts, such as ribosomal frameshifting, can lead to translation initiation and termination at different sites within a single mRNA molecule (Atkins et al., 2016), resulting in polypeptides with sequences differing from the main open reading frame (ORF) (Korniy et al., 2019). In particular, alternative translation initiation and synthesis of N-terminally truncated polypeptides are associated with the development of certain diseases (Bogaert et al., 2020). Additionally, RNA editing, involving processes such as nucleotide deamination of adenosine into inosine or nucleotide insertions and deletions, can alter mRNA sequences and lead to the formation of distinct mRNA isoforms (Knoop, 2011).

**Translation-level Variation** – Co- and post-translational modifications can rapidly modify protein properties and functions. Currently, over 200 types of PTMs or biological and chemical modifications, primarily targeting specific amino acid residues (FIGURE I-2), have been documented (Creasy and Cottrell, 2004; Montecchi-Palazzi et al., 2008). For example, phosphorylation and lipidation, can redirect proteins to specific cellular locations or modulate their interactions with other molecules, thereby significantly influencing various physiological processes such as transcription, translation, and metabolic functions (Jiang et al., 2018; Macek et al., 2019). Additionally, other PTMs such as glycosylation, can influence protein folding and stability (Jayaprakash and Surolia, 2017). Controlled proteolysis, by endopeptidases, can precisely cleave proteins, converting inactive zymogens into their biologically active forms (Neurath and Walsh, 1976). Zymogens typically contain inhibitory propeptides at their N-termini, ensuring both the protection of protein function and their directed transport to specific cellular compartments. Upon reaching their destination, the propeptide is removed, activating the catalytic activities of the protein. For example, cathepsins, a family of lysosomal proteases, achieve their optimal functionality within the highly acidic environment of lysosomes upon activation (Jordans et al., 2009). This selective, compartment-specific activation isolates potentially detrimental reactions, thereby safeguarding the cell. In contrast, exopeptidases that

primarily remove terminal amino acids play a significant role in protein degradation processes. Both proteolytic processes possess the potential to generate truncated proteoform variants, capable of modulating protein activities and potentially influencing the development of specific diseases (Bogaert et al., 2020).

The analysis of PTMs requires precise mass spectrometry (MS) analysis because isobaric PTMs share similar molecular masses and, consequently, comparable mass-to-charge ratios. This similarity increases the risk of misidentifications due to measuring errors (Kim et al., 2016). For example, trimethylation ( $C_3H_6$ , 42.047 Da) and acetylation ( $C_2H_2O$ , 42.011 Da) on lysine residues, commonly found on histone proteins, have remarkably similar masses, differing by only 0.036 Da. In the case of a typical 1 kDa tryptic peptide (Fricker, 2015), this corresponds to a difference of 36 ppm (parts per million). While modern high-resolution MS mass analyzers have mitigated this problem for peptides, challenges remain for intact proteins. For a 15 kDa core histone, this mass difference is 2.4 ppm, highlighting the critical role of accurate mass measurements in determining the correct proteoform.



**FIGURE I-2 | Protein Modifications in Bacteria.** Overview of the most commonly occurring protein post-translational modifications, corresponding amino acid residue, and corresponding modifying enzymes. Reactive groups on amino acid side chains are highlighted. PTS, phosphotransferase system; TCS, two-component system. Adapted with permission from (Macek et al., 2019).

## 2 Principles of Mass Spectrometry

Mass spectrometry is a powerful analytical technique used for identifying molecules based on their mass-to-charge ratio ( $m/z$ ).

### 2.1 Online LC-MS

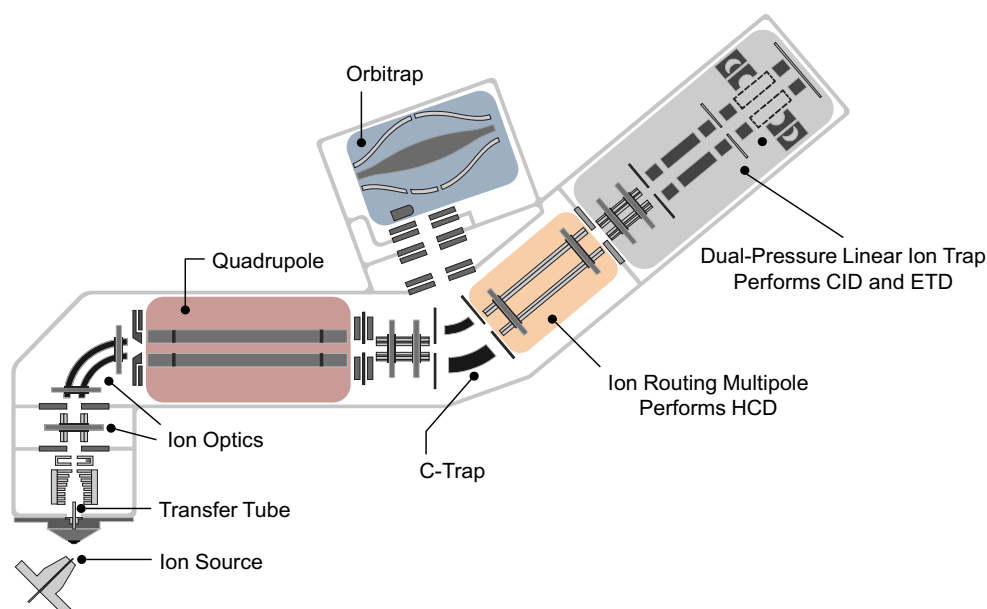
The preferred method in proteomics involves the online coupling of a separation technique, such as liquid chromatography (LC), with MS. Peptide and protein mixtures are typically separated based on their hydrophobic interactions with a non-polar stationary phase, employing reversed-phase high-performance liquid chromatography (RP HPLC). Commonly used microparticulate columns consist of silica beads with covalently bonded hydrophobic chains, such as octadecyl alkane chains ( $C_{18}$ ) for peptides and butyl alkane chains ( $C_4$ ) for protein mixtures. This results in species with differing hydrophobicity eluting at different times due to their affinities for the stationary phase. In contrast, monolithic columns feature a continuous porous structure for the separation of peptides or proteins. Retention, a metric that quantifies how long an analyte remains on the column, and resolution, the ability to distinguish adjacent peaks, are determined by the balance between the analyte's hydrophobic interactions with the stationary phase and its solubility in the mobile phase. The gradual increase in the concentration of organic solvent, typically acetonitrile, in the mobile phase initiates elution. The utilization of an acidic mobile phase, often containing an ion-pairing modifier such as formic acid or TFA, can enhance chromatographic separation, aid in controlling retention times, improve peak shapes, and improve the overall detection efficiency and performance of the LC-MS system (García, 2005; Lenčo et al., 2022).

Electrospray ionization (ESI) is a widely used technique for transferring analytes from the liquid phase to the gas phase in online LC-MS setups. It enables the gentle vaporization of molecules without causing significant fragmentation (Kearle and Verkerk, 2009). During the ESI process, the analyte solution is directed through a conductive capillary (referred to as the emitter), to which a voltage is applied. In positive ion mode, the preferred polarity for the analysis of proteins and peptides, ions are converted into protonated molecular ions  $[M+nH]^{n+}$  with various charge and protonation states. Protonation primarily occurs at the free amino terminus and the basic side-chain functionalities of arginine, lysine, and histidine residues.

### 2.2 Mass-to-charge Analysis

Various mass analyzers and detectors, each with unique characteristics such as mass accuracy, speed, sensitivity, and resolution (the ability to distinguish closely spaced  $m/z$  ratios), have been developed for measuring ion  $m/z$  ratios. High-resolution and accurate mass analyzers allow for the identification of characteristic isotope patterns, particularly in peptides

containing naturally occurring stable, heavy isotopes such as  $^{13}\text{C}$  and  $^{15}\text{N}$ . Mass analyzers operate on diverse principles: some (e.g., linear or quadrupole ion traps) utilize different electric currents to manipulate ions for selective isolation, trapping, and fragmentation, while others (e.g., time-of-flight or Orbitrap analyzers) accelerate ions towards a detector in an electric field to measure their time-of-flight or radial oscillation frequencies to determine  $m/z$  (Savaryn et al., 2016). Today, hybrid instruments combine different analyzers allowing flexibility in experimental design. A notable example of such hybridization is the Fusion Lumos Tribrid mass spectrometer, integrating quadrupole, Orbitrap, and dual-pressure linear ion trap analyzers (FIGURE I-3). Before entering the mass spectrometer, ions can be introduced into a high-field asymmetric waveform ion mobility spectrometry (FAIMS) module, to separate them based on ion mobility variations in the presence of high and low electric fields (Guevremont, 2004). Acting as a mass filter, FAIMS can eliminate singly charged contaminant ions and enhance the depth of proteome coverage by applying multiple compensation voltages (CVs) (Swearingen and Moritz, 2012; Kaulich et al., 2022a). After entering the Fusion Lumos Tribrid mass spectrometer, ions pass through ion optics, which focus and direct them toward the quadrupole. The quadrupole then acts as a mass filter, allowing to select precursor ions of interest based on their  $m/z$  values (Savaryn et al., 2016). Ions within a specific isolation window are collected, stored in the ion-routing multipole, and then transferred through the C-trap into the Orbitrap. Fragment spectra can be acquired in either the Orbitrap or the ion trap mass analyzer.



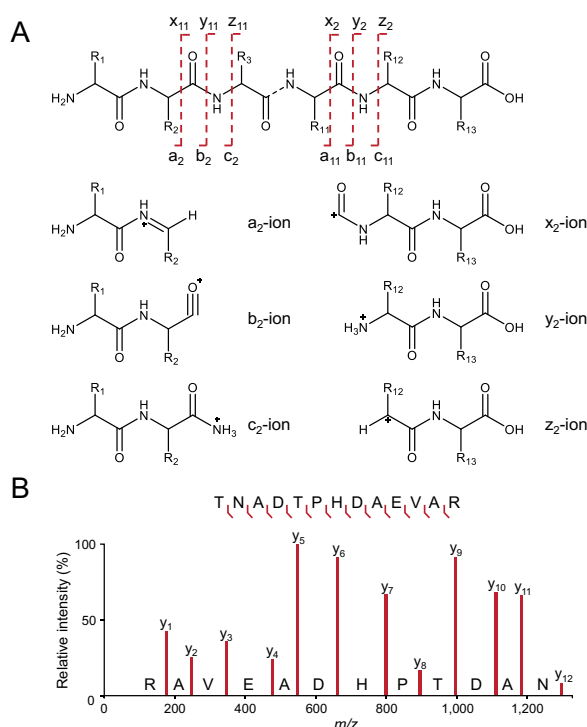
**FIGURE I-3 | The Orbitrap Fusion Lumos Tribrid Mass Spectrometer.** After ionization, ion optics focus and direct gas phase analyte ions toward the quadrupole. In the quadrupole, ions are selectively filtered based on their mass-to-charge ratio ( $m/z$ ) within a specific isolation window. The selected precursor ions are then directed to the ion-routing multipole or the linear ion trap mass analyzer for fragmentation. Fragment ion spectra are then acquired in either the Orbitrap or the linear ion trap. Adapted from <http://planetorbitrap.com>.

## 2.3 Tandem Mass Spectrometry and Fragmentation Analysis

Analyzing samples via tandem mass spectrometry (MS/MS) with data-dependent acquisition (DDA) involves obtaining precursor MS<sup>1</sup> spectra, fragmenting isolated ions, and determining the  $m/z$  values of resulting fragment ions in the MS<sup>2</sup> spectra. This process provides essential insights into the amino acid sequence of the peptide (Biemann, 1992), facilitating the differentiation of isobaric peptides with distinct amino acid sequences or PTMs (Kim et al.,

2016).

Tribrid instruments offer various ion activation methods, with collision-induced dissociation (CID) (Hunt et al., 1986) or higher-energy collisional dissociation (HCD) (Olsen et al., 2007) being the most common. CID and HCD involve collisions with inert gases (e.g., nitrogen, helium, or argon), resulting in charge-directed fragmentation, driven by the kinetic energy of ionizing protons, which mostly results in cleavage of the peptide backbone. In the Fusion Lumos, CID is performed in the high-pressure collision cell of the linear ion trap, while HCD is conducted separately in the ion routing multipole. This separation allows for enhanced resolution and improved mass accuracy measurements of fragment ions by enabling higher kinetic



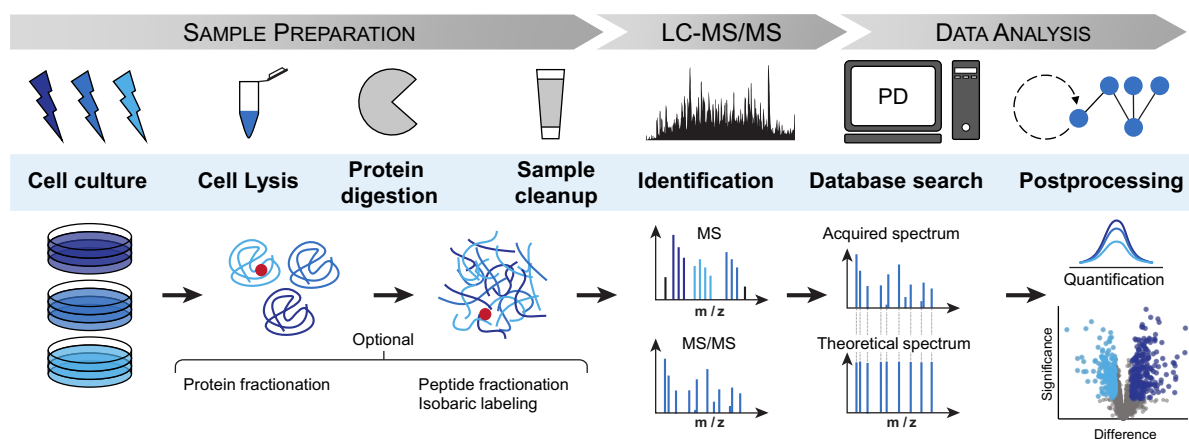
**FIGURE I-4 | Fragment Ion Nomenclature. (A)** Roepstorff–Fohlmann–Biemann nomenclature for fragment ions (Roepstorff and Fohlman, 1984; Biemann, 1992). **(B)** Determination of peptide sequence using y-ion series.

energies and shorter impact times during fragmentation (Michalski et al., 2012). Another important method is the ETHcD approach, combining electron transfer dissociation (ETD) (Syka et al., 2004) in the linear ion trap with subsequent transfer of precursors and product ions to the collision cell for HCD fragmentation. Fragmentation of precursor ions by ETD results in radical-driven fragmentation of peptide bonds while limiting the neutral loss of labile groups and preserving PTMs (Syka et al., 2004). Fragment ions resulting from peptide dissociation are influenced by several factors, including the mass and charge state of the precursor ion, amino acid composition, and adjacent residues at the backbone cleavage site (Reid et al., 2001; Tabb et al., 2003; Haverland et al., 2017). Typically, the charge is distributed between both fragments, with fragment ions retaining their charge either at the C-terminus (x-, y-, and z-ions) or the N-terminus (a-, b-, and c-ions) according to the Roepstorff-Fohlmann-Biemann nomenclature (Roepstorff and Fohlman, 1984; Biemann, 1992) (FIGURE I-4).

### 3 Analysis of Mass Spectrometry Data

Unlike genomics and transcriptomics, which can amplify DNA or RNA using techniques like polymerase chain reaction, respectively, proteomics lacks a comparable method for amplifying proteins. This limitation necessitates careful attention to the proteomics workflow to maximize information from the potentially limited starting material.

In general, a proteomics workflow can be divided into three major steps (FIGURE I-5). The initial step (i) involves sample preparation, which includes isolating the protein mixture from the studied biological sample. The complexity of samples can be reduced and proteome coverage enhanced by employing protein or peptide pre-fractionation techniques (Zhang et al., 2010). Following sample cleanup, the second step (ii) involves online separation and MS measurement of analytes, typically using reversed-phase LC-MS/MS. In principle, peptide or protein masses can be determined from MS<sup>1</sup> spectra by utilizing the  $m/z$  value and spacing between isotope peaks. The corresponding amino acid sequences are then determined based on the MS<sup>2</sup> mass difference between the fragments. With modern mass spectrometers generating thousands of spectra in a short time, manual annotation becomes impractical, necessitating an automated approach for peptide and protein identification. Thus, the final step (iii) involves employing various software applications and computational tools to identify peptides and proteins, along with performing post-processing tasks like quantification, statistical testing, and metabolic mapping to evaluate significant differences between samples. In this study, Proteome Discoverer served as the primary software tool for proteomic data analysis. Notably, Proteome Discoverer demonstrated improved performance compared to the widely used software MaxQuant, particularly in terms of quantification yield, dynamic range, and reproducibility for label-free quantification (Palomba et al., 2021).



**FIGURE I-5 | Generalized Bottom-up Proteomics Workflow.** Proteins are extracted from cells treated with various stimuli. Sample complexity can be reduced through protein prefractionation, peptide fractionation, or isobaric labeling after enzymatic digestion. Following sample clean-up using solid-phase extraction, samples are separated by reversed-phase liquid chromatography (LC) and analyzed by mass spectrometry (MS). Analytes are identified from MS/MS spectra using Proteome Discoverer (PD) database matching. Post-processing includes quantification, statistical testing, and metabolic mapping to assess significant differences between samples. (Adapted with permission from (Mergner and Kuster, 2022)).

### 3.1 Data Acquisition Techniques

In the field of MS-based proteomics, the methods used for data acquisition are crucial for determining the depth and accuracy of proteome analysis. Three primary acquisition methods are commonly employed: data-dependent acquisition (DDA), data-independent acquisition (DIA), and target acquisition, which focuses on identifying and quantifying specific sets of peptides (Aebersold and Mann, 2016).

DDA-MS, widely employed for protein identification and quantification, operates by (automatically) selecting precursor ions for fragmentation (Pappireddi et al., 2019). However, the stochastic nature of precursor ion selection in DDA-MS can pose challenges, hindering the identification of the same set of peptides across replicate analyses of the same sample (Chapman et al., 2014). This can result in a bias towards more abundant peptides and lead to a higher incidence of missing values, thus impacting reproducibility in large-scale experiments (Pappireddi et al., 2019). To minimize redundancy and mitigate missing values, dynamic exclusion can be applied to prevent the reselection of the same precursor ion over a defined time (Chapman et al., 2014). Additionally, while it is feasible to impute missing data, doing so may alter statistical calculations and compromise quantitative accuracy (Gardner and Freitas, 2021). In contrast to DDA-MS, DIA-MS acquires fragment ion spectra for all the precursors contained within predetermined isolation windows (Chapman et al., 2014; Pappireddi et al., 2019). This approach significantly enhances the detectable dynamic range and reproducibility in protein identification between technical replicate experiments (Chapman et al., 2014; Aebersold and Mann, 2016). However, a major limitation of DIA-MS is the need to generate a study-specific spectral library for data processing, which generally demands higher starting sample amounts and extensive instrument time (Aebersold and Mann, 2016).

A novel approach, wide window acquisition (WWA), was developed to combine the strengths of both DDA and DIA methods. WWA introduces a wide isolation window ( $\geq 4$   $m/z$ ) into the data-dependent precursor selection, facilitating the co-fragmentation of precursor ions near the selected ones (Matzinger et al., 2024). This process generates chimeric spectra containing multiple peptides akin to those observed in DIA runs. This innovative strategy not only enhances peptide identification rates but also increases proteome coverage by capturing low-abundance peptides that might otherwise remain elusive (Truong et al., 2023). The potential of WWA is further enhanced through the integration of innovative artificial intelligence (AI)-driven search algorithms like CHIMERYs, marking a revolutionary advancement in proteome analysis (Matzinger et al., 2023, 2024; Truong et al., 2023). By maximizing proteome coverage, addressing challenges associated with missing values, and identifying low-abundance peptides, the combination of WWA with CHIMERYs holds significant potential for advancing the understanding of complex systems. In summary, this novel approach has the potential to overcome the limitations of traditional DDA and DIA methods.

## 3.2 Peptide and Protein Identification

In the field of proteomics, two main analytical approaches are employed to characterize the proteome: bottom-up analysis, which generally identifies peptides, and top-down analysis, which characterizes proteoforms.

**Bottom-up** – Peptide-centric proteomics, often referred to as shotgun or bottom-up proteomics (BUP), involves characterizing peptides generated from protein digestion using specific endopeptidases such as trypsin. Trypsin specifically cleaves internal peptide bonds C-terminal to lysine and arginine residues, except when followed by proline. This specificity generates two positive charges due to the basic amine side groups at the N- and C-terminus of the resulting peptides which enhances both detection and fragmentation in tandem mass spectrometry (Biemann, 1992; Gershon, 2014). Despite peptide MS combined with chromatographic separation techniques like LC, dominating proteomics for decades (Carbonara et al., 2021), it often faces challenges in achieving comprehensive protein sequence coverage. This limitation arises as only a small fraction of the generated peptides can be reliably identified, partly due to factors such as ionization efficiency or co-isolation. Furthermore, while BUP can identify peptides and infer protein groups, it inherently loses valuable proteoform information, making discrimination between highly similar proteins and elucidation of posttranslational modifications difficult (Nesvizhskii and Aebersold, 2005). While attempts have been made to obtain proteoform data indirectly from peptide identifications, such approaches remain limited (Bludau et al., 2021).

**Top-down** – Typically proteoform analysis is performed by top-down proteomics (TDP), an innovative approach that directly examines intact proteins, offering a comprehensive exploration of protein diversity and its connections to biological processes (Aebersold et al., 2018; Smith and Kelleher, 2018). In theory, TDP has the potential to address the longstanding challenge of "protein inference" that affects BUP (Nesvizhskii and Aebersold, 2005; Tholey and Becker, 2017). However, identifying large proteoforms (>30 kDa) in TDP analysis is challenging. Co-eluting proteoforms of different charge states can share overlapping  $m/z$  ranges, which can impede accurate proteoform assignment via spectral deconvolution (Basharat et al., 2023). Deconvolution involves decharging and deisotoping to calculate monoisotopic masses (Jeong et al., 2020). Additionally, larger proteoforms generally exhibit lower ionization efficiencies than smaller peptides or proteoforms, resulting in decreased detection sensitivity. This lower sensitivity can lead to reduced fragmentation efficacy and fewer informative product ions for sequence determination. Consequently, TDP analysis has primarily been successful in identifying abundant, lower molecular weight proteoform species in less complex samples (Durbin et al., 2016).

In response to these limitations, prefractionation strategies have been developed to reduce sample complexity. These strategies involve separating proteins based on size, charge, or hydrophobicity prior to mass spectrometric analysis. Molecular weight-based prefractionation methods, such as Gel Elution Liquid Fraction Entrapment Electrophoresis (GELFrEE), efficiently partition the proteome into liquid fractions by continuous elution gel electrophoresis (Tran and Doucette, 2008). The Passively Eluting Proteins from Polyacrylamide gels as Intact species (PEPPI) technique, which efficiently recovers proteins from conventional SDS-PAGE gels, provides another effective approach (Takemori et al., 2020). Further two-dimensional chromatography (Kaulich et al., 2024), molecular weight cut-off filters (Yang et al., 2020), solid-phase extraction (Petruschke et al., 2020; Cassidy et al., 2021a), and organic solvent depletion (Cassidy et al., 2019) can also reduce sample complexity. In addition, gas-phase separation methods, such as FAIMS, improve proteoform separation and reduce spectral complexity by selectively introducing specific ions into the mass spectrometer (Fulcher et al., 2021; Gerbasi et al., 2021; Kaulich et al., 2022a). Collectively, these techniques offer multiple ways to overcome sample complexity.

### 3.3 Quantitative Techniques in Proteomics

Within quantitative proteomics, two prominent comparative strategies have gained popularity: labeling approaches and label-free approaches. Label-based quantitative proteomic methods, including tandem mass tagging (TMT) (Thompson et al., 2003) and isotope tags for relative and absolute quantification (iTRAQ) (Wiese et al., 2007), introduce isobaric tags at the N-termini and lysine side chains of proteins and peptides. Upon fragmentation, reporter ions with distinct  $m/z$  ratios are released and serve as quantifiable isobaric markers reflecting the abundance of the corresponding peptides from which they originate (Rauniyar and Yates, 2014). Stable isotope labeling by amino acids in cell culture (SILAC) (Ong et al., 2002) differs from TMT and iTRAQ by utilizing the metabolic incorporation of isotopes, typically  $^{13}\text{C}$ -arginine or  $^{15}\text{N}$ -lysine, during cell culture (Pappireddi et al., 2019). The multiplexing capability of labeling approaches reduces experimental variability and minimizes potential bias by subjecting all samples to the same conditions throughout the workflow, thereby enabling accurate comparison of protein abundances across multiple samples (Schulze and Usadel, 2010; Bantscheff et al., 2012).

In contrast, label-free quantification (LFQ) methods, often based on  $\text{MS}^1$  precursor intensities or peptide-to-spectrum matches (PSMs) counts, typically require less sample manipulation. Further, LFQ methods are not limited to a specific number of isobaric channels, which can restrict experimental design (Bantscheff et al., 2012; Pappireddi et al., 2019). However, LFQ demands a robust and consistent MS operation, as each LFQ sample is independently analyzed by LC-MS/MS. To address small differences in sample loading and chromatographic

gradient variability, a practical strategy involves splitting measurements into consecutive analytical batches, each potentially encompassing distinct technical variations (Čuklina et al., 2021; Phua et al., 2022). This practice, along with randomizing samples (Burger et al., 2021), can mitigate machine drift and batch effects thus improving reproducibility (Burger et al., 2021; Čuklina et al., 2021; Phua et al., 2022). Combined with advances in bioinformatics, notably in LC-MS feature alignment (Cox et al., 2014), and the incorporation of AI-driven data processing algorithms alongside WAA, LFQ techniques have significantly improved in efficiency over the last few years (Matzinger et al., 2023, 2024; Truong et al., 2023).

However, most quantitative proteomic methods come with a fundamental drawback - their inherent reliance on comparisons. Variations in factors like amino acid composition, charge state, peptide length, or PTMs lead to significant differences in ion intensities, even when these peptides originate from the same protein (Dupree et al., 2020). Therefore, accurate quantification relies on comparing peptides with identical  $m/z$  ratios, acquired under consistent LC-MS/MS operating parameters. Consequently, quantitation techniques based on relative MS measurements inherently involve comparing two or more samples. Conversely, absolute quantification demands stringent experimental design and appropriate data analysis, often involving comparisons to a spiked "known" standard, as exemplified by isotope-labeled synthetic AQUA (Absolute Quantification) peptides (Gerber et al., 2003). While empirical relationships, like iBAQ (intensity-based absolute quantification), calculated by dividing protein intensities by the number of theoretically observable tryptic peptides (Schwanhäusser et al., 2011), suggest absolute quantification, they are still intrinsically based on relative intensities.

## 4 Inflammatory Bowel Disease

### 4.1 The Human Gut Microbiome as a Therapeutic Target

Inflammatory bowel disease (IBD) is a family of chronic inflammatory disorders that affect the digestive system and is characterized by symptoms such as abdominal pain, diarrhea, and rectal bleeding (Wang et al., 2023b). The two primary forms of IBD are Crohn's disease and ulcerative colitis (Abdulla and Mohammed, 2022). While historically more prevalent in Northern Europe, North America, and China, the incidence of IBD has increased in many previously uncommon regions such as Southern Europe and several developing countries (Wang et al., 2023b). The development of IBD is the result of a complex interplay between genetic predispositions, innate immune factors, and environmental influences (Abdulla and Mohammed, 2022). Although multiple environmental triggers likely contribute to IBD, the exact cause remains uncertain. Research indicates that the microbial population residing in our guts, known as the microbiome, plays a crucial role in the severity and progression of colitis in IBD patients. Factors such as diet (Dolan and Chang, 2017), antibiotic usage (Rook et al., 2015),

smoking (Yadav et al., 2017), and alcohol consumption (White et al., 2022) can significantly alter the microbiome and influence IBD prevalence. Furthermore, modern lifestyle changes such as improvements in hygiene standards and the widespread consumption of sterilized industrialized foods can result in reduced microbial exposure in early life, which has been linked to an elevated risk of IBD (Rook et al., 2015). Targeting and correcting imbalances in the gut microbiome may be crucial for preventing and treating IBD (Abdulla and Mohammed, 2022). Particularly, maintaining the balance of microorganisms capable of degrading mucin, which can influence host-microbe interactions and immune responses, is essential for intestinal health and reducing the risk of IBD (Png et al., 2010). However, the exact contribution of microbiome imbalance to IBD pathogenesis, whether primary or secondary, remains a topic of ongoing research.

The gut microbiome can generally be classified into three enterotypes based on the relative abundance of *Bacteroides*, *Prevotella*, or *Ruminococcus* genera (Arumugam et al., 2011). While *Bacteroides* species are typically regarded as commensal bacteria promoting immune tolerance and gut homeostasis, certain species like *B. fragilis* or *B. vulgatus* have been implicated in IBD pathogenesis (Wexler, 2007). Consequently, variations in the *Firmicutes/Bacteroidetes* (F/B) ratio have been associated with both IBD (decreased F/B ratio) and obesity (increased F/B ratio) (Magne et al., 2020; Stojanov et al., 2020). Conversely, within a harmonious gut ecosystem, *B. thetaiotaomicron* plays a crucial role in the catabolism of complex carbohydrates and dietary fiber, facilitating the digestion of components that are otherwise indigestible by human enzymes (Xu and Gordon, 2003; Cantarel et al., 2009).

To date, several *Ruminococcus* species have been reclassified as *Blautia* (Liu et al., 2008), representing a significant proportion of the total microbiota, ranging from 2.5% to 16% (Zoetendal et al., 2002). *Blautia*'s role in gastrointestinal health is multifaceted and complex, with both positive and negative associations impacting health (Liu et al., 2021b). While some studies suggest a negative correlation between *Blautia* abundance and IBD (Chen et al., 2014), others indicate a positive correlation with irritable bowel syndrome (IBS) (Rajilić-Stojanović et al., 2011). Most studies on *Blautia* have primarily focused on the genus level, with comprehensive investigations at the species or even strain level remaining scarce. However, among the diverse species within this genus, *B. producta* stands out for its remarkable probiotic properties, including the ability to inhibit the proliferation of antibiotic-resistant pathogens like vancomycin-resistant enterococci (VRE) without adversely affecting other commensal bacteria (Caballero et al., 2017; Kim et al., 2019).

*Bifidobacteria*, while not classified within the three enterotypes of the gut microbiome, play a crucial role in maintaining gut health (Sadeghpour Heravi and Hu, 2023). They are particularly abundant in the infant's gut and contribute significantly to its development (Turroni et al., 2012). Despite a decline in their presence during childhood and adolescence, influenced by factors such as the introduction of solid foods, lifestyle changes, puberty, and antibiotic use,

*Bifidobacteria* continue to contribute substantially to intestinal integrity, preventing harmful bacterial colonization, and modulating the immune system throughout life (Turroni et al., 2012; Avershina et al., 2013). One prominent species within the *Bifidobacteria* genus is *B. longum*, which is present in individuals of all ages (Turroni et al., 2009). Due to its health benefits in promoting gut microbiota balance and offering protection against and treatment of IBD, *B. longum* is commonly utilized as a probiotic (Sadeghpour Heravi and Hu, 2023).

## 4.2 Proteomics in IBD Research

Exploring the function of the human gut microbiome has been largely based on *in silico* analyses of genetic data. These studies have investigated various microbial activities, such as amino acid and B-vitamin biosynthesis (Magnúsdóttir et al., 2015; Ashniev et al., 2022), inositol utilization (Weber and Fuchs, 2022), and the identification of gene clusters responsible for producing small molecules like non-ribosomal peptides and polyketides (Donia et al., 2014). Through these efforts, the understanding of the microbiome's complex functionality has significantly expanded. While genomics provides insights into genetic information, it does not provide information about corresponding protein products and their abundance within biological systems. Therefore, proteomics has emerged as an essential tool in IBD research (Longo et al., 2020; Fabian et al., 2023). However, most proteomic studies have predominantly focused on specific tissues, such as intestinal epithelial cells, as well as body fluids like serum and stool samples collected from both IBD patients and healthy individuals (Fabian et al., 2023).

The advancements in anaerobic culture techniques (Hungate, 1969), the development of specific culture media (Duncan et al., 2002), and targeted phenotypic culture systems, known as "culturomics" (Browne et al., 2016), have facilitated the cultivation of a wide array of gut bacteria. These progressions have enabled the cultivation of bacterial isolates, including those previously considered 'uncultivable' (Lagier et al., 2016), and have led to the development of small co-culture systems, such as the "simplified model system of the human gut microbiome" (SIHUMIx) (Krause et al., 2020). Although there remains a significant gap in understanding the functions and phenotypes of many indigenous gut bacteria, these advances have enabled the analysis of numerous human gut bacteria.

To date, most studies of the human gut microbiome have focused on characterizing and analyzing the entire protein complement of a microbial community, rather than solely focusing on individual organisms. This approach, known as metaproteomics, has facilitated the discovery of distinct proteomic abundance profiles based on gastrointestinal locations (Lichtman et al., 2016), as well as distinct bacterial protein signals associated with Crohn's disease and ulcerative colitis (Lehmann et al., 2019).

Although studies on single bacterial isolates in response to diverse *in vitro* conditions are limited, they have yielded diverse proteomic findings. For instance, proteomic analyses were utilized to validate *B. thetaiotaomicron*-specific protein requirements under various growth conditions (Liu et al., 2021a), observe the selective packing of acidic glycosidases and proteases into *Bacteroides* outer membrane vesicles (Elhenawy et al., 2014), and characterize proteins required for mucin degradation by human gut bacteria (Crouch et al., 2020). However, how proteomic profiles change in single members of the microbiota at IBD initiation, progression, and during therapeutic interventions remains largely unknown. To improve functional knowledge of the gut microbiome in the context of IBD, it is essential to generate detailed proteomic profiles.

## 5 Objective of the Thesis

This study aims to analyze the proteomic adaptations of selected bacteria of the human gut microbiome in response to varying environmental conditions, focusing on the specific objectives:

**Evaluation of Quantitative Proteomic Approaches (Chapter III)** – Assess the reproducibility and accuracy of bottom-up label-free quantification (LFQ) and tandem mass tag (TMT)-based quantification methods for detecting proteomic changes induced by sucrose and glucose in *B. thetaiotaomicron*.

**Exploration of pH-Dependent Proteomic Alterations (Chapter IV)** – Analyze proteomic changes in *B. thetaiotaomicron*, *B. producta*, and *B. longum* in response to different environmental pH levels.

**Identification of Short Open Reading Frame-Encoded Peptides (Chapter V)** – Employ a proteogenomic approach to identify and analyze the translation of previously undiscovered SEP in *B. producta* under various cultivation conditions.

**Optimization of OMV Isolation Protocol (Chapter VI)** – Establish and optimize a protocol for the LC-MS-based proteomic analysis of extracellular vesicles, specifically outer membrane vesicles (OMVs) isolated from *E. coli*.

**Integration of Proteoform-Directed Analysis (Chapters III to IV)** – Apply a proteoform-directed top-down analysis, including a discovery-based open modification search, to identify potential post-translational modifications and neo-termini in HGM members.



## II GENERAL METHODS

---

<b>1</b>	<b>General Materials.....</b>	<b>18</b>
1.1	Chemicals and Reagents .....	18
1.2	LC-column Packing .....	18
<b>2</b>	<b>Protein and Peptide Sources.....</b>	<b>19</b>
2.1	Cultivation of Human Gut Bacteria .....	19
2.2	Cultivation of <i>E. coli</i> and OMV Isolation .....	20
2.3	Caco-2 Wound Healing Assay .....	22
<b>3</b>	<b>Sample Preparation for LC-MS Analysis.....</b>	<b>23</b>
3.1	Proteome Clean-up and Digestion .....	23
3.2	Molecular Weight-based Prefractionation .....	25
3.3	SDC Removal by Phase-transfer .....	26
3.4	Solid-phase Extraction .....	27
3.5	Tandem Mass Tag Labelling .....	27
3.6	Protein Gel and Staining .....	29
<b>4</b>	<b>Mass Spectrometry Data Acquisition .....</b>	<b>30</b>
4.1	Bottom-up LC-MS Measurements.....	30
4.2	Top-down LC-MS Measurements .....	31
4.3	MALDI-TOF Measurements .....	32
<b>5</b>	<b>Data Processing and Analysis .....</b>	<b>32</b>
5.1	Database Search.....	32
5.2	Genome and Proteome Sequence-based Predictions .....	34
5.3	Functional <i>in silico</i> Analysis.....	34
5.4	Functional and Statistical Data Analyses .....	35

## 1 General Materials

### 1.1 Chemicals and Reagents

Deionized water (18.2 MΩ/cm) was obtained using an Arium611 VF system (Sartorius). Complete protease inhibitor cocktail was purchased from Roche Diagnostics. Pierce Coomassie and BCA protein assay kits (both Thermo) were used for protein quantification. Single-pot, solid-phase-enhanced (SP3) bead-based purification was performed using Sera-Mag SpeedBead carboxylate-modified magnetic particles (GE Life Sciences). Sequencing grade modified trypsin was purchased from Promega. Lyophilized protein standards of cytochrome C (*Equus caballus*), myoglobin (*Equus caballus*), beta-casein (*Bos taurus*), carbonic anhydrase (*Bos taurus*), bovine serum albumin (*Bos taurus*), and alcohol dehydrogenase (*Saccharomyces cerevisiae*) were purchased from Sigma-Aldrich. HeLa and cytochrome C digest were purchased from Thermo Fisher. Synthetic peptides were purchased from JPT Peptide Technologies GmbH. Additional chemicals required for various stages of media preparation, sample handling, and LC-MS/MS analysis were purchased from various suppliers, including Sigma-Aldrich, Merck, and Serva.

### 1.2 LC-column Packing

Frits for column packing were prepared by cross-linking Kasil (potassium silicate solution) with formamide (Cortes et al., 1987). Kasil 1, a 29.1% (w/w) potassium silicate solution in water, was solubilized at 80°C and 2000 rpm for 1 hour. Equal parts of Kasil 1 and a 25% w/v formamide solution in water were combined to create the final frit solution. After mixing and centrifuging at 21,000 g for 5 minutes at 20°C, tub fused silica capillaries (360 µm OD x 75 or 150 µm ID) were gently pressed onto a glass microfiber filter (GF/C, Whatman), which had been previously soaked with 2 µL of the frit solution (Maiolica et al., 2005). Capillaries were incubated at 85°C for 20 hours for polymerization. Using a high-pressure bomb loader, capillaries were packed with PLRP-S beads (5 µm, 1000 Å), which were collected from a PLRP-S column (4.6 x 50 mm, Agilent). The collected beads were washed with 50% and then 100% acetonitrile (ACN), dried at 70°C and suspended in methanol (60 mg/ml). After allowing the material to settle by gravity for 20-30 minutes and mixing for 1 minute and sonication, columns were packed under low-speed stirring (400-500 rpm) with continuous mechanical tapping (Kovalchuk et al., 2019). The bomb containing the capillary was pressurized to 100 bar with nitrogen immediately after mounting the capillary to prevent passive filling of the capillary with the solvent. After the packing process was completed, the pressure was slowly released for 10 minutes to prevent bubble formation within the column. The freshly packed columns were connected to an Ultimate 3000 system and flushed with 95% ACN at a flow rate of 600

nl/min for 30 minutes to compress the sorbent bed. The columns were then cut to the desired length, and connections were made using ZIRCOFIT UHPLC fittings with a 1/16" 13-mm bore (MS Wil). This resulted in two different types of columns: pre-columns (150  $\mu\text{m}$  x 4 cm) and analytical columns (75  $\mu\text{m}$  x 17 cm).

## 2 Protein and Peptide Sources

Details on the number of bacterial cell culture replicate and specific culture conditions for particular experiments are described in the experimental procedure sections of the corresponding chapters.

### 2.1 Cultivation of Human Gut Bacteria

The cultivation of *Bacteroides thetaiotaomicron* VPI-5482, *Blautia producta* ATCC 27340, and *Bifidobacterium longum* NCC 2705 was performed by Kathrin Schäfer (Department of Infectious Diseases and Microbiology, University of Lübeck, UKSH Lübeck; chair: Prof. Dr. Jan Rupp). Yeast extract, casein, and fatty acid (YCFA) medium (Duncan et al., 2009) (TABLE II-1) or brain-heart infusion (BHI) medium adjusted to different pH values (pH 6.0, pH 7.0, and pH 8.0) and supplemented with different carbon sources (27.8 mM glucose or sucrose) or various growth supplements (lipopolysaccharide (LPS), SCFA or yeast extract) were utilized for bacterial cultivation. Bacterial cells were cultured at 37°C under strict anaerobic conditions in an anoxic chamber (H35, Don Whitley Scientific Limited) containing 85% (v/v) N<sub>2</sub>, 10% (v/v) CO<sub>2</sub> and 5% (v/v) H<sub>2</sub>. A single colony was transferred to 5 ml of YCFA or BHI medium, incubated overnight and 0.1% (v/v) was used for inoculum.

**TABLE II-1 | Modified YCFA Medium**

COMPONENTS	[g/l]	COMPONENTS	[mg/l]	COMPONENTS	[g/l]
Casitone	10	Biotin	0,02	Acetic acid	1900
Yeast extract	2,5	Folic acid	0,10	Propionic acid	700
Carbon Source	5	Pyridoxine hydrochloride	0,05	<i>iso</i> -Butyric acid	90
MgSO <sub>4</sub> x 7 H <sub>2</sub> O	0,45	Thiamine-HCl x 2 H <sub>2</sub> O	0,05	<i>n</i> -Valeric acid	100
CaCl <sub>2</sub> x 2 H <sub>2</sub> O	0,90	Riboflavin	0,05	<i>iso</i> -Valeric acid	100
K <sub>2</sub> HPO <sub>4</sub>	0,45	Nicotinic acid	0,05		
KH <sub>2</sub> PO <sub>4</sub>	0,45	D-Calcium pantothenate	0,001		
NaCl	0,90	Vitamin B12	0,05		
Resazurin	0,01	<i>p</i> -Aminobenzoic acid	0,05		
Distilled water	4	Lipoic acid	10		
NaHCO <sub>3</sub>	1				
L-Cysteine HCl	0,1				
Hemin	0,02				

## 2.2 Cultivation of *E. coli* and OMV Isolation

**Growth Media** – Lysogeny broth (LB) medium (1% tryptone, 0.5% yeast extract, 1% NaCl, pH 7.0, Sigma-Aldrich) was prepared according to the manufacturer's instructions. Agar plates were prepared by adding 1% agar and autoclaving at 121°C for 15 minutes. M9 culture media were prepared as indicated in TABLE II-2 and TABLE II-3 and sterilized by autoclaving at 121°C for 12 minutes. Glucose and acetate solutions were autoclaved separately, and heat-sensitive components such as biotin and thiamine were filter-sterilized (0.2 µm filter) and added to the media afterward.

**TABLE II-2 | 100X Trace Elements Solution**

COMPONENT	[g/l]	CONCENTRATION
EDTA	5 g/l	13.4 mM
FeCl <sub>3</sub> -6H <sub>2</sub> O	0.83 g/l	3.1 mM
ZnCl <sub>2</sub>	84 mg/l	0.62 mM
CuCl <sub>2</sub> -2H <sub>2</sub> O	13 mg/l	76 µM
CoCl <sub>2</sub> -2H <sub>2</sub> O	10 mg/l	42 µM
H <sub>3</sub> BO <sub>3</sub>	10 mg/l	162 µM
MnCl <sub>2</sub> -4H <sub>2</sub> O	1.6 mg/l	8.1 µM

**TABLE II-3 | M9 Mineral Medium**

VOLUME	COMPONENT	COMPONENT	CONCENTRATION
100 ml	M9 salt solution (10X)	Na <sub>2</sub> HPO <sub>4</sub> -2H <sub>2</sub> O	33.7 mM
		KH <sub>2</sub> PO <sub>4</sub>	22.0 mM
		NaCl	8.55 mM
		NH <sub>4</sub> Cl	9.35 mM
20 ml	Carbon source	Glucose or acetate	15 mM or 45 mM
1 ml	MgSO <sub>4</sub> (1M)	MgSO <sub>4</sub>	1 mM
0.3 ml	CaCl <sub>2</sub> (1M)	CaCl <sub>2</sub>	0.3 mM
1 ml*	Biotin (1 mg/ml)	Biotin	1 µg
1 ml*	Thiamin (1 mg/ml)	Thiamin	1 µg
10 ml*	Trace elements (100X)	Trace elements	1X

\* 0.22-µm filter sterilization

**Cultivation** – A single colony of *Escherichia coli* K-12 strain MG1655 was cultured in 100 ml of LB medium for 18 h at 37°C and 150 rpm in a non-baffled 250 ml shake flask. Afterward, the culture medium was removed by centrifugation at 7,000 g for 3 min at 4°C, and cells were washed twice with filter-sterilized M9 minimal medium. Then, cells were inoculated with 1/100 dilutions of 18 h pre-culture at an initial OD<sub>600</sub> of 0.1 in M9 media containing either 15 mM glucose or 45 mM acetate. The cultures were incubated aerobically at 37°C, 150 rpm in baffled 1 l shake flasks with 300 ml of media. During the cultivation culture samples were taken to track cell growth via OD<sub>600</sub> measurements.

**Isolation of Outer Membrane Vesicles (OMV)** – OMV isolation from *E. coli* supernatant was performed using the ExoBacteria OMV Isolation Kit (System Biosciences). Cultivations were transferred into sterile 50 ml centrifuge tubes and centrifuged at 8,000 g for 20 min at 4°C. The supernatant was transferred to a new sterile 50 ml centrifuge tube and spun again at 8,000 g for 20 min at 4°C. The supernatant was filter-sterilized (0.2 µm filter) and the resulting cell-free supernatant was used to isolate OMVs. The binding column was prepared by adding 1 ml of OMV binding resin and equilibrating it with 10 ml of OMV binding buffer. After equilibration, the binding buffer was allowed to completely flow through the column. Subsequently, the bottom of the column was sealed, and 20 ml of supernatant was added. The top of the column was sealed, and the unit was placed on a rotating rack for 30 min at 4°C to allow for mixing and binding of the OMVs to the resin. After 30 min, the top and bottom of the column were opened to allow the supernatant to flow through the resin. Depending on the experiment, loading of supernatant was repeated up to 2 additional times so that a total of 40 ml or 60 ml of culture supernatant, was incubated with the OMV binding resin. After OMV binding, the supernatant was allowed to flow through the column, and the resin was washed with 15 ml OMV binding buffer per 20 ml of loaded supernatant. Afterward, the bottom of the column was sealed, and 1.5 ml OMV elution buffer was added. Columns were allowed to incubate at 20°C for 2 min with gentle agitation every 30 s, after which the bottom of the column was unsealed and the OMV isolate was collected. Samples were frozen at -80°C until further analysis.

**Nanoparticle Tracking Analysis** – For nanoparticle tracking analysis (NTA) a NanoSight NS300 system (NanoSight Ltd), equipped with a 488 nm laser and a high sensitivity digital camera system sCMOS (scientific complementary metal oxide semiconductor camera) was employed. Videos were acquired and analyzed using the NTA software (version 3.3), with minimum track length and blur setting, all set to automatic. The camera shutter was set manually in dependency of the particle intensity and camera gain was set to 366. Camera levels were set to 14 to 15 and the detection threshold was set to 5, to reveal small particles. The ambient temperature was maintained at 25°C. Samples were administered and recorded under controlled flow, using the NanoSight syringe pump and script control system. For each sample, six videos of 60 seconds duration were recorded, with a 10-second delay between recordings, generating six replicate histograms that were averaged. Lastly, the hydrodynamic diameters and particle size distributions were analyzed by the software using the Stokes-Einstein equation. A summary of the complete list of parameters used in the nanoparticle tracking analysis is provided in TABLE II-4. The instrument was calibrated prior to each experimental run using standardized nanoparticle dilutions purchased from the manufacturer.

**TABLE II-4 | Parameters for Nanoparticle Tracking Analysis**

PARAMETER	SETTING
Instrument	NanoSight NS300
NTA Version	NTA 3.3 Dev Build 3.3.301
Diluent	Water (1:1)
Camera Type	sCMOS
Camera Level	14-15
Laser Type	Blue 488
Slider Shutter	1200 - 1260
Slider Gain	366
Frames/Sec	25
Frames	1498
Temperature	25.0 °C
Viscosity	Water 0.889 cP
Syringe Pump Speed	40
Detect Threshold	5
Blur Size	Auto
Max Jump Distance	Auto

## 2.3 Caco-2 Wound Healing Assay

Human Caco-2 cells were cultured in collaboration with Britta Steer (Systematic Proteome Research & Bioanalytics, Institute for Experimental Medicine, University of Kiel; chair: Prof. Dr. Andreas Tholey). Cells were cultured in Roswell Park Memorial Institute (RPMI) medium (Gibco-Invitrogen) supplemented with 10% fetal calf serum (FCS; Gibco-Invitrogen) and 1% penicillin/streptomycin in a 5% CO<sub>2</sub>, 95% humidity environment at 37°C. Cells were passaged weekly upon reaching 80% confluence. Wound healing assays were conducted using  $\mu$ -dishes with inserts (Ibidi GmbH) at a density of  $4 \times 10^5$  cells/cm<sup>2</sup>. After reaching a confluence of 70-80%, cells were serum-starved overnight using serum-deprived medium (0.1% FCS), the insert was removed, and cell layers were washed twice with phosphate-buffered saline (PBS). Cells were then incubated with 2 ml of 0.1% FCS serum-deprived medium containing different concentrations of *E. coli* OMVs (10, 50, and 100  $\mu$ g/ml), OMV elution buffer, LPS *E. coli* O55:B5 (1  $\mu$ g/mL, Sigma-Aldrich), transforming growth factor  $\beta$  (TGF $\beta$ ) (5 ng/mL, Sigma-Aldrich) and medium only. The migration process into the cell-free gap of approximately 500  $\mu$ m was measured by taking microscopic photographs at 0, 12, and 30 hours using a digital camera on an inverted microscope at 10x magnification. The wound area was quantified using a wound healing plugin for ImageJ (Suarez-Arnedo et al., 2020). The percentage of wound closure was calculated relative to the medium control (arbitrarily assigned as 100%) based on the area measured immediately after insert removal ( $A_{t=0h}$ ), as well as 12 and 30 hours after incubation ( $A_{t=\Delta h}$ ) (eq. 1).

$$\text{Wound closure\%} = \left[ \frac{A_{t=0h} - A_{t=\Delta h}}{A_{t=0h}} \right] \times 100\% \quad (\text{eq. 1})$$

### 3 Sample Preparation for LC-MS Analysis

Details regarding the quantities of protein and peptide used, the number of technical replicas, and different MS parameters for specific experiments are described in the experimental procedure sections of the corresponding chapters. Protein concentrations were determined in at least triplicate using the Pierce Coomassie or BCA Protein Assay Kit according to the manufacturer's instructions.

#### 3.1 Proteome Clean-up and Digestion

**Human Gut Bacteria Samples** – Cell lysis of human gut bacteria was performed by Kathrin Schäfer (Department of Infectious Diseases and Microbiology, University of Lübeck, UKSH Lübeck; Chair: Prof. Dr. Jan Rupp). Culture media was removed by centrifugation at 2,100 g for 10 min at 4°C, washed twice with water, and centrifuged again. Bacterial cells were suspended in lysis buffer (6 M GndHCl (guanidine hydrochloride), 100 mM HEPES (4-2-hydroxyethyl-1-piperazineethanesulfonic acid), 20 mM NaCl and 1x cOmplete protease inhibitor, pH 7.5). Cells were lysed using ten cycles of freeze-thawing (30 s, -80°C in an ethanol bath followed by thawing in a sonication bath for 30 s). After centrifugation at 21,000 g for 20 min at 4°C, supernatants were collected and cell debris was washed twice with lysis buffer, centrifuged and supernatants were pooled. Disulfide bridges were reduced and alkylated using 12 mM tris(2-carboxyethyl)phosphine (TCEP) and 40 mM 2-chloroacetamide (CAA) for 1 hour at 25°C and 800 rpm. The samples were then precipitated using 9x volume ethanol at -20°C. After 16 hours of incubation at -20°C, the precipitates were centrifuged at 21,000 g for 10 min at 4°C and washed twice with cold ethanol. Residual ethanol was evaporated in a fume hood. Precipitates were suspended in 0.5 M GndHCl, 12.5 mM HEPES, or 100 mM triethylammoniumbicarbonat (TEAB) (all pH 8.5), digested by adding trypsin at a 1:40 enzyme to substrate ratio, and incubated overnight at 37°C on a shaker at 800 rpm.

***E. coli* samples** – 1 mg of *E. coli* cells were processed using the sample preparation by simple extraction and digestion (SPEED) method (Doellinger et al., 2020), with TFA added at a 1:4 (v/v) ratio of sample to TFA. Samples were incubated for 5 minutes at 20°C and neutralized with 2 M Tris base using 8-fold the volume of TFA used for lysis. Aliquots of 50 µg protein were reduced and alkylated by incubation in 10 mM TCEP and 40 mM CAA at 95°C for 5 minutes. Samples were diluted 1:5 with water and proteins were digested with trypsin at a 1:50 enzyme-to-substrate ratio for 20 hours at 37°C.

**OMV samples** – 40 µg of OMV samples were processed using different sample preparation protocols, including in-solution, on-bead SP3 (Hughes et al., 2019), and a modified version of the on-membrane filter-assisted sample preparation (FASP) method (Manza et al., 2005; Wiśniewski et al., 2009).

**In-solution Digestion** – OMV samples were lysed by ten cycles of freeze-thaw (30 s, -80°C in an ethanol bath followed by thawing in a sonication bath for 30 s) followed by adjustment to 100 mM TEAB using 1 M TEAB. Alternatively, 40 µg OMV samples were lyophilized before the addition of 50 µl of lysis buffer (2% (w/v) sodium deoxycholate (SDC), 0.1 M TEAB, pH 8.5). Samples were then incubated at 95°C for 5 minutes to inactivate proteases, reduced with 10 mM dithiothreitol (DTT) at 56°C for 1 hour, and alkylated with 50 mM iodoacetamide (IAA) in the dark at 20°C for 30 minutes. For samples containing SDC, the concentration was adjusted to 0.5% SDC with 0.1 M TEAB (pH 8.5) and trypsin was added at a 1:40 (w/w) enzyme-to-substrate ratio. The digestion reaction was performed overnight at 37°C on an orbital shaker at 800 rpm.

**On-bead Digestion** – For OMV lysis, sodium dodecyl sulfate (SDS) or SDC was added to a final concentration of 1% or 2% (w/v), respectively. Subsequently, samples were incubated at 95°C for 5 minutes before being reduced using 10 mM DTT for 1 hour at 56°C and alkylated with 50 mM IAA in the dark for 30 minutes at 20°C. SP3 beads were prepared by mixing equal amounts of hydrophilic and hydrophobic beads, were washed twice with water, and resuspended in water to a final concentration of 20 µg/µl. For each sample, 20 µl of beads were added, and protein binding was induced by adding ethanol to a final concentration of 50% (v/v) and incubating for 5 minutes at 25°C at 800 rpm. The beads were immobilized using a magnetic rack, the supernatant was removed, and the beads were washed twice with 200 µl of 80% ethanol. Beads were resuspended in digestion buffer (100 mM TEAB, pH 8.5) containing trypsin at a 1:40 (w/w) enzyme-to-substrate ratio and incubated overnight at 37°C on a shaker at 800 rpm. Enhanced digestion was performed by adding 0.5% (w/v) SDC or 0.001% (w/v) dodecyl-β-D-maltosid (DDM) to the digestion buffer. The digests were acidified to pH 2 to 3 by adding trifluoroacetic acid (TFA), and the beads were pelleted by centrifugation. The supernatants were stored on ice until sample cleanup. For SDC digests, samples were additionally processed using a modified phase transfer protocol.

**On-membrane Digestion** – For OMV lysis, a final concentration of 1% or 2% (w/v) of SDS or SDC, respectively, was added. The samples were incubated at 95°C for 5 minutes, reduced with 10 mM DTT for 1 hour at 56°C, and loaded onto Amicon centrifugal filter units (30K molecular weight cutoff, Millipore). After centrifugation at 12,000 g for 15 min at 20°C (all buffer exchanges were performed by centrifugation under identical conditions), SDS-lysed samples

were washed with 400 µl of SDS removal solution (8 M urea, 0.1 M TEAB, pH 8.5), while SDC-lysed samples were washed with 400 µl of SDC-solution (0.5% SDC, 0.1M TEAB, pH 8.5). Alkylation was carried out in the dark for 30 minutes at 20°C using 50 mM IAA in the respective solution. Samples were washed twice with the respective solution and SDS-lysed samples were additionally washed three times with digestion buffer (100 mM TEAB, pH 8.5). Trypsin was added at a 1:40 (w/w) enzyme-to-substrate ratio and incubated overnight at 37°C on a shaker at 600 rpm in a digestion buffer or SDC solution. Peptides were collected by centrifugation and acidified with TFA before sample clean-up by solid-phase extraction or phase transfer.

### 3.2 Molecular Weight-based Prefractionation

**Depletion of High Molecular Weight Proteins** – 400 µg precipitated proteins were processed using a high mass protein depletion procedure (Cassidy et al., 2019). For the acidic depletion, proteins were suspended in 80 µl of 210 mM NaCl and 0.1% TFA, resulting in a final concentration of 50 mM NaCl. For the basic depletion, proteins were suspended in 80 µl of 420 mM TEAB, resulting in a final concentration of 100 mM TEAB. The samples were vortexed to ensure solubilization and 3.2x volume of ACN was added, with the addition of 0.1% TFA for the acidic depletion. After incubation for 1 h at 20°C at 1300 rpm, samples were centrifuged at 21,000 g for 20 min at 20°C. The supernatant was then transferred to a new tube and dried using vacuum centrifugation.

**Gel-eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE)** – 500 µg precipitated proteins were suspended in 200 µl of 1% (w/v) SDS solution, mixed with 5× GELFrEE sample buffer and 10 mM DTT before being incubated for 10 min at 50°C at 600 rpm (Toby et al., 2019). Electrophoresis was conducted with the GELFrEE 8100 system (Expedeon) using 12% GELFrEE cartridges (Abcam) following the manufacturer's protocol. In total 12 fractions with a volume of 150 µl were collected and cleaned up using chloroform/methanol water precipitation. Each fraction was mixed with four volumes of methanol, and vortexed, and then one volume of chloroform was added and vortexed again. After phase separation at 10,000 g for 5 min at 20°C, the upper aqueous methanol layer was discarded. Next, four volumes of methanol were added and the sample was vortexed again. After a second centrifugation at 10,000 g for 15 min at 20°C, the supernatant was discarded and the protein pellet was washed twice with methanol before being air-dried and stored at –20°C until measurement.

**Strong Cation Exchange (SCX)** – Two aliquots of 450 µg total protein lysate (with a sample volume of approximately 150 to 200 µl) were incubated at 75°C and 600 rpm for 10 minutes to ensure that any proteolytic enzymes were inactivated. The samples were then acidified with

5% formic acid (FA) to a 13.3-fold dilution. After centrifugation at 21,000 g for 20 min at 20°C, solid phase extraction was performed to extract the LMWP (chapter II.3.4), followed by SCX to reduce the abundance of low-charged ions (Cassidy et al., 2021a). SCX chromatography was performed using an off-line Dionex Ultimate 3000 HPLC system (ThermoFisher Scientific), equipped with a PolySulfoethyl A column (300 Å, 5 µm, 2.1 × 200 mm). After conditioning and equilibration with solvent A (30% ACN, 5 mM KH<sub>2</sub>PO<sub>4</sub>, pH 2.7), solid-phase extracted proteins and peptides were dissolved in 80 µl solvent A and loaded onto the column. The chromatographic separation was accomplished using a 90-minute gradient at a constant flow rate of 250 µl/min with three solvents (TABLE II-5), namely solvent A, solvent B (30% ACN, 350 mM KCl, 5 mM KH<sub>2</sub>PO<sub>4</sub>, pH 2.7), and solvent C (30% ACN, 800 mM NaCl, 5 mM KH<sub>2</sub>PO<sub>4</sub>, pH 2.7). The elution was divided into four fractions (F0, F1, F2, and F3), each containing peptides with different charges (TABLE II-6). Following fractionation, the samples were lyophilized and suspended in 5% FA for solid-phase extraction (SPE) desalting.

**TABLE II-5 | SCX Chromatographic Separation Scheme**

TIME (MIN)	ELUENT A (%)	ELUENT B (%)	ELUENT C (%)
0-10	100%	0%	0%
10-40	80%	20%	0%
40-50	70%	30%	0%
50-55	50%	50%	0%
55-60	0%	100%	0%
60-70	0%	100%	0%
70-75	0%	0%	100%
75-90	100%	0%	0%

**TABLE II-6 | SCX Fraction Collection Scheme**

FRACTION	CHARGED SPECIES	ELUTION TIME (MIN)
F0	Singly and low charged species	0-5
F1	Primarily +2 charged species	5-35
F2	Primarily +3 charged species	35-55
F3	+4 and higher charged species	55-75

### 3.3 SDC Removal by Phase-transfer

A modified phase-transfer protocol was utilized for the removal of SDC from peptide samples (Masuda et al., 2008). Initially, ethyl acetate was added to the sample at a ratio of 1:1 (v:v). After vigorous mixing for 1 minute, samples were centrifuged at 16,000 g for 2 minutes at 20°C, followed by careful aspiration of the upper ethyl acetate layer. This extraction step was repeated three times. To eliminate left-over amounts of ethyl acetate, vacuum centrifugation was performed.

### 3.4 Solid-phase Extraction

Depending on the protein or peptide quantities, different sorbent weights were employed for sample processing via SPE. Organic solvent-free samples were acidified to pH 2 to 3 with neat TFA before loading. Samples that contained organic solvents were dried by vacuum centrifugation and reconstituted in a loading solution (3% ACN, 0.05% TFA).

**Peptide Samples** – For digest of more than 100 µg protein, 50 mg Sep-Pak C<sub>18</sub> 1cc RP (reversed-phase) SPE cartridges (Waters) were used. Peptide quantities below 100 µg were cleaned up using pipette tips (Pierce C<sub>18</sub> tips 100 µl), whereas amounts smaller than 10 µg were desalted using 10 µl tips (both Thermo Fisher Scientific). Volumes of solvents were adapted to sorbent weights using typically 1 ml, 150 µl, and 20 µl for 50 mg cartridges and 100 µl, and 10 µl for ZipTips, respectively. Sorbent conditioning was performed using 100% ACN followed by equilibration with loading solution. After slow loading, bound peptides were washed twice with loading solvent and elution was performed step gradually increasing the concentration of ACN in 0.1% TFA, with three different concentrations used: 50%, 70%, and 100%. Samples were dried by vacuum centrifugation, and stored at - 80°C until further processing.

**Protein Samples** – Protein samples of around 1 mg were processed using 200 mg Sep Pak C<sub>18</sub> 3cc RP SPE cartridges (Waters). Protein samples below 1 mg (e.g., SCX fractionated samples) were processed using 50 mg Sep-Pak C<sub>18</sub> 1cc RP SPE cartridges (Waters). Volumes of solvents were adapted to sorbent weights using typically 3 ml or 1 ml for 200 mg or 50 mg cartridges, respectively. Sorbent materials were conditioned with 100% ACN and equilibrated with loading solvent (5% FA). After sample loading and two washes with loading solvent, elution was performed by sequentially increasing the concentration of ACN in 0.1% TFA at two different concentrations: 70% and 100%. Samples were dried using vacuum centrifugation and stored at -80°C until further processing.

### 3.5 Tandem Mass Tag Labelling

**TMT Labelling** – Peptides were labeled with TMT-6-plex (Thermo Fisher Scientific) (TABLE II-7), using a reduced TMT to peptide ratio (Innovation and Zecha, 2019). In detail, 80 µg peptides were dissolved in 100 mM of TEAB (pH 8.5) and labeled in a 4:1 TMT: peptide ratio with one of the six corresponding TMT stock dissolved in 100% anhydrous ACN. The peptide-TMT mixture was incubated for 1 h at 25°C at 700 rpm until the labeling reaction was stopped by the addition of 5% hydroxylamine to a final concentration of 0.4% and incubated for 15 min at 25°C and 700 rpm.

TABLE II-7 | TMT-Labeling Scheme

SAMPLE	TMT-REAGENT
Glucose biological replica A	TMT-126
Sucrose biological replica A	TMT-127
Glucose biological replica B	TMT-128
Sucrose biological replica B	TMT-129
Glucose biological replica C	TMT-130
Sucrose biological replica C	TMT-131

**Evaluation of Under- and Overlabeling** – To estimate the degree of underlabeling, raw files were searched with TMT as a variable modification on lysine residues and peptide N-termini. Peptide sequences that were modified with TMT on all lysine side chains and free peptide N-termini were counted as ‘fully labeled’. Peptides that did not bear any TMT label were classified as ‘not labeled’, whereas peptides that contained at least one TMT label but were not fully labeled were classified as ‘partially labeled’. To determine over-labeling, TMT was set as fixed on lysine residues and peptide N-termini and, in addition, as a variable modification on histidine, serine, threonine, or tyrosine residues. Peptides that were identified to be labeled with TMT on at least one serine, threonine, or tyrosine were counted as ‘over labeled’. The labeling efficiency was then calculated by the number of labeled N-termini ( $n_{ti}$ ) and lysine residues ( $n_{ki}$ ) and the total number of peptide N-termini ( $n_{tt}$ ) and lysine residues ( $n_{kt}$ ) (eq. 2).

$$100\% \times (n_{ti} + n_{ki}) / (n_{tt} + n_{kt}) \quad (\text{eq. 2})$$

**Mixing and Evaluation of the Six Channels** – A small aliquot of all six TMT-labeled peptides were mixed equally, dried using a vacuum centrifuge, and suspended in loading solution. After sample clean-up by SPE the combined sample was analyzed by LC-MS/MS, and the reporter ion intensities for each TMT channel were obtained using Proteome Discoverer. To correct for any potential bias introduced during TMT labeling, the median reporter ion intensity is calculated for each channel. A correction factor was calculated by dividing the median reporter ion intensity for each channel by the median reporter ion intensity for the channel with the highest median value (eq. 3).

$$\text{Correction factor} = 1 / (MI / MI_{\text{Max}}) \quad (\text{eq. 3})$$

The correction factors were used to correct the mixing and achieve a median reporter ion intensity of 1 for all six TMT channels. After mixing the sample was reanalyzed to confirm that the median intensities for each channel were close to 1:1:1:1:1:1.

**Peptide Fractionation** – TMT-labeled peptides were fractionated on a pH 10 (basic) reversed phase (bRP) column (Gemini-C18, 250 mM x 3 mm, 3  $\mu$ m, Phenomenex) using an off-line Dionex Ultimate 3000 HPLC system (Thermo Fisher Scientific). Samples were cleaned up using SPE prior to fractionation (chapter II.3.4). Peptide separation was performed according to a previously published, detailed protocol (Delmotte et al., 2007) and is described briefly in the following. After conditioning and equilibration with solvent A (72 mM triethylamine adjusted to pH 10 with acetic acid), peptides dissolved in 50  $\mu$ L of solvent A were loaded onto the column. Peptides were washed with 5% solvent B (ACN, 72 mM triethylamine, 3.5 ml acetic acid) for 10 min until being sequentially eluted over 50 min at a gradient from 5% B to 55% B at a flow rate of 200  $\mu$ L/min. Sample fractions were collected every minute, resulting in a total of 56 fractions of 200  $\mu$ L each. The elution was monitored using a UV detector at 254 nm and 280 nm. The fractions were dried by vacuum centrifugation, dissolved in 200  $\mu$ L loading solution, and pooled into seven fractions (TABLE II-8), following a previously described fractionation scheme (Stephanowitz et al., 2012).

**TABLE II-8 | Fraction Pooling Scheme**

POOL	FRACTIONS
1	1, 8, 15, 22, 29, 36, 43, 50
2	2, 9, 16, 23, 30, 37, 44, 51
3	3, 10, 17, 24, 31, 38, 45, 52
4	4, 11, 18, 25, 32, 39, 46, 53
5	5, 12, 19, 26, 33, 40, 47, 54
6	6, 13, 20, 27, 34, 41, 48, 55
7	7, 14, 21, 28, 35, 42, 49, 56

### 3.6 Protein Gel and Staining

Protein samples were separated using 1mm thick vertical separating gels (12%) and stacking gels (4%) using the Mini-PROTEAN Tetra Cell-Package apparatus. For sample preparation, the samples were mixed 1:1 (v/v) with reducing sample buffer (62.5 mM Tris-HCl, pH 6.8, 10% glycerol, 2% SDS, 0.005% Bromophenol Blue and 5% 2-mercaptoethanol), incubated at 95°C for 5 minutes, cooled on ice, and transferred to the wells of the stacking gel. The Roti Mark Standard marker (Carl Roth) was used to estimate the molecular weight of the proteins. Gel documentation was performed by scanning with a calibrated flatbed scanner (ViewPix900, Epson). Analytical gels were stained with Coomassie brilliant blue or silver stained (Shevchenko et al., 1996).

## 4 Mass Spectrometry Data Acquisition

Depending on the sample type and research objectives, varying LC-MS parameters, database parameters, and data processing strategies were employed. Specific experimental approaches are described in the experimental design sections of the respective chapters. In general, the mass spectrometers were operated in positive ionization and DDA mode. The MS acquisition program was initiated with a 5-minute delay. Before analysis, samples were dissolved in loading solution, and insoluble material was removed through centrifugation.

### 4.1 Bottom-up LC-MS Measurements

Nanoflow LC-ESI-MS/MS measurements were conducted using a Dionex Ultimate 3000 HPLC system online coupled to either a Q-Exactive Plus (QE-plus) or Q-Exactive HF-X (HF-X) mass spectrometers (all Thermo Fisher Scientific).

**Liquid Chromatography** – Generally, 1 to 1.5 µg protein digest was concentrated and washed onto a trap column (75 µm × 2 cm, 2 µm C<sub>18</sub> resin, 100 Å; Acclaim PepMap100, Thermo Fisher Scientific) for 5 min with 2% ACN and 0.05% aqueous TFA at a flow rate of 30 µl/min. Subsequently, peptides were separated on an analytical column (75 µm x 50 cm, 2 µm C<sub>18</sub> resin, 100 Å; Acclaim PepMap100, Thermo Fisher Scientific) at 300 nl/min and separated within 60 or 120 min using linearly increasing gradients of LC solvent B (80% ACN, 0.1% FA) in LC solvent A (0.1% FA). The full proteome was separated using 5 to 50% LC solvent B for 120 min gradients. In contrast, lower complex OMV or LMWP samples were separated using 5 to 50% LC solvent B for 60 min gradients. The linear gradient was followed by a sharp increase to 90% solvent B for 5 min, an isocratic 9 min washing step, and finally a column equilibration with 5% B for 12 min.

**Mass Spectrometry** – MS<sup>1</sup> spectra were recorded in the Orbitrap from 300 to 1800 *m/z*, at a resolution of 60K or 70K, using an automatic gain control (AGC) target value of 3e6 (HF-X/QE-plus) charges and a maximum injection time (maxIT) of 50 to 100 ms, depending on expected sample complexity and peptide abundance. Acquisition of MS<sup>2</sup> spectra were obtained in the Orbitrap at 15K (HF-X) or 17.5K (QE-plus) resolution after HCD fragmentation at normalized collision energy (NCE) of 27 for label-free or 35 for TMT-labelled samples. For TMT-labelled samples the first mass was fixed to 100 *m/z* with a scan range up to 2000 *m/z*, applying a narrower isolation window of 1.5 *m/z* to reduce co-isolation. Otherwise, the scan range was set to 200 to 2000 *m/z* with an isolation window of 2.0 *m/z*. The number of MS<sup>2</sup> spectra was limited by a top10 (10 most intense peptide fragments) method. Dynamic exclusion was

adjusted according to gradient length (20 to 40 s), with ions of unassigned, +1, and >+8 charge states excluded, and lock mass (445.12003  $m/z$ ) enabled.

## 4.2 Top-down LC-MS Measurements

Nanoflow LC-ESI-MS measurements were conducted using a Dionex Ultimate 3000 HPLC system online coupled to a Fusion Lumos Tribrid (Lumos) mass spectrometer (both Thermo Fisher Scientific). Depending on the applied method, the mass spectrometer was equipped with the field asymmetric ion mobility spectrometry (FAIMS) pro interface.

**Liquid Chromatography** – Generally, 1 to 1.5  $\mu\text{g}$  protein preparations were concentrated and washed onto a trap column (5 mM x 0.33 mm, 5  $\mu\text{m}$  C<sub>4</sub> resin, 300 Å; PepMap300, Thermo Fisher Scientific) for 5 min with 2% ACN and 0.05% aqueous TFA at a flow rate of 30  $\mu\text{l}/\text{min}$ . Subsequently, proteins were separated on an analytical column (75  $\mu\text{m}$  x 50 cm, 2.6  $\mu\text{m}$  C<sub>4</sub> resin, 150 Å; Accucore, Thermo Fisher Scientific) at 300 nl/min and separated within a 60-, 90- or 140-min linear gradient of increasing LC solvent B (80% ACN, 0.1% FA) in LC solvent A (0.1% aqueous FA). Additionally, GELFrEE samples were separated on a self-packed PLRP-S pre-column (150  $\mu\text{m}$  x 4 cm, 5  $\mu\text{m}$  PLRP-S resin, 1000 Å), before being separated on an analytical column (75  $\mu\text{m}$  x 17 cm, 5  $\mu\text{m}$  PLRP-S resin, 1000 Å). LMWP samples, GELFrEE fractions, and SCX fractions of *B. thetaiotaomicron* were separated using a 90-minute gradient from 15% to 55% B. In contrast, the LMWP samples of *B. producta* were separated using a 60-minute gradient from 15% to 60% B. The linear gradient was followed by a sharp increase to 98% solvent B for 2 min, an isocratic wash step for 13 min, and finally a column equilibration with 5% B for 15 min.

**Mass Spectrometry** – A 15 Volt source-induced dissociation was applied to favor protein ion desolvation and the RF Lens was set at 30%. The acquisition was performed in “peptide mode”, according to the recommendations of vendors for the mass range below ca. 20–30 kDa.

**Dual CV FAIMS Method** – Acquisition of MS spectra were obtained using two methods including two different FAIMS compensation voltages (CVs). The first method involved combining -60 V and -50 V, while the second method used -40 V and -30 V. Different microscan settings were used for each method, with only 2 microscans being utilized for -60 V, and 4 microscans being used for all other CVs. MS<sup>1</sup> spectra were recorded in the Orbitrap from 500 to 1800  $m/z$ , at a resolution of 120K, using an AGC target value of 200% and a maxIT of 246 ms. Both methods obtained MS<sup>2</sup> spectra with a 60K resolution (3 s cycle time) after collision-induced dissociation (CID) with an NCE of 25%. MS<sup>2</sup> spectra were acquired within 4

microscans in a scan range of 500  $m/z$  to 2000  $m/z$ , using a normalized AGC target of 400% and a maxIT of 250 ms. Only precursors with a charge state between 4 and 50 or undetermined charge states were selected with enabled dynamic exclusion (exclude after 2 times, 60 s duration) plus and minus 2.5  $m/z$ .

**Mutli-CV FAIMS Method** – Four different CVs (-60, -50, -40, -20 V) with different MS<sup>1</sup> and MS<sup>2</sup> settings were applied, based on (Kaulich et al., 2022a). Within a cycle time of 3 s MS<sup>2</sup> spectra were acquired with an isolation window of 5  $m/z$ , 50K resolution, 400% AGC target, 250 ms injection time, and for fragmentation CID with an NCE of 30% was utilized. Only precursors with a charge state between 4 and 50 or undetermined charge states were selected with enabled dynamic exclusion (n= 2, 60 s). Settings for CVs -60 and -50 V: resolution 60K/50K (MS<sup>1</sup>/MS<sup>2</sup>), maximum injection time 118/125 ms, microscans 2/2, AGC target 200%/400%. Settings for CVs -40, -20 V: resolution 120k/60k, maximum injection time: 246/250 ms, 4/4 microscans.

### 4.3 MALDI-TOF Measurements

Matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometry was performed using a TOF/TOF 5800 mass spectrometer equipped with the 4000 series explorer software (both AB Sciex). The mass spectrometer accumulated 400 laser pulses at an intensity of 65%. Spectra were acquired within a mass range of 100  $m/z$  to 1200  $m/z$ . Prior to analysis, the instrument was calibrated using a six-peptide solution to calibrate the target  $m/z$  range. Samples were prepared by mixing with varying volumes (v/v) of a CHCA matrix solution (3 mg/ml  $\alpha$ -cyano-4-hydroxycinnamic acid in 70% acetonitrile, 0.1% TFA) and spotted as 1  $\mu$ L drops onto a 384-well Opti-TOF MALDI Insert (AB Sciex).

## 5 Data Processing and Analysis

Details on the specific search parameters for each experiment and the version of Proteome Discoverer (PD) used are available in the experimental procedures section of each chapter.

### 5.1 Database Search

Depending on the sample type, tandem mass spectra were searched and compared to the UniProt reference proteome of *Bacteroides thetaiotaomicron* vpi-5482, *Blautia producta* ATCC 27340, *Bifidobacterium longum* NCC 2705 or *Escherichia coli* K12 (UP000001414 / 4,782 proteins; UP000515789 / 5,372 proteins; UP000000439 / 1,725 proteins; UP000000625 / 4,450 proteins, accessed between September 2020 and August 2022) or the integrated

proteogenomics database (iPtgxDB) for *Blautia producta* ATCC 27340 (accessed November 2021) (Omasits et al., 2017). In addition, common repositories of adventitious proteins were included in the searches.

**Bottom-Up Searches** – Peptide identification and quantification experiments were performed using PD (v.2.2, v.2.5, or v.3.0; Thermo Fisher Scientific) utilizing its built-in search engine SequestHT, INFERYS rescoring, or CHIMERYS algorithm. These included trypsin as the proteolytic enzyme with up to two or four missed cleavage sites allowed, carbamidomethylation of cysteine (57.021 Da) as fixed modification, oxidation of methionine (15.995 Da) as variable modifications, a precursor tolerance of 10 ppm and a fragment ion tolerance of 0.02 Da. For TMT experiments, TMT6 was employed as the fixed modification on lysine and peptide N-termini. To assess under/over labeling, variable TMT6 modifications were added as search parameters for lysine and peptide N-termini, as well as fixed modifications for histidine, serine, threonine, and tyrosine. Additional searches, including phosphorylation as variable modifications on serine, histidine, threonine, tyrosine, or arginine (79.966 Da), and semi-tryptic searches, were customized based on the sample type. Retention-time alignment was applied for samples necessitating MS<sup>1</sup> intensity-based quantification. All results were adjusted to 1% PSM, peptide, and protein FDR, employing a target-decoy approach using reversed protein sequences and posterior error calculation by the percolator algorithm (Käll et al., 2008).

**Top-Down Searches** – Proteoform identification and quantification were performed using PD (v.2.5.0.400; Thermo Fisher Scientific) utilizing its built-in ProSightPD 4.1 or 4.2 nodes (Proteinaceous Inc.). These included the high/high cRAWler node in combination with Xtract for deconvolution. The annotated proteoform search, with a maximum of three proteoform spectrum matches (PrSMs) per precursor, a minimum of three matched fragments, and no delta M mode, was utilized to identify full-length proteoforms. Truncated proteoforms were detected using the subsequence search node, with a maximum of one PrSM per precursor and a requirement of at least six matched fragments. Searches, utilized a precursor and fragment mass tolerance of 10 ppm, with variable modifications acetylation (42.011 Da) and formylation (27.995 Da) at the proteoform N-terminus. Additional searches, specifically designed to search for fixed dehydro-modifications on cysteine (-1.008 Da) to identify potential disulfide bridges and open-modification searches with a 500 Da precursor window, were customized based on the sample type. For label-free quantification, raw data from multi-CV measurements were filtered using Freestyle v.1.6 based on FAIMS CVs, and the resulting filtered data was saved as distinct .raw files, effectively dividing them into four fractions. Quantification was carried out employing the High-Resolution Feature Detector node with the sliding window deconvolution algorithm, utilizing an average retention time width of 0.33 min

with mass features needed to be present in only 1% of the total data files. All results were subjected to an FDR correction for both PrSMs and proteoforms, with a threshold of 1%.

### 5.2 Genome and Proteome Sequence-based Predictions

Protein sequences were retrieved from UniProt and genomic sequences were obtained from the European Nucleotide Archive (Leinonen et al., 2011). Unless otherwise noted, default parameters were used for all prediction tools.

Protein sequences were annotated using WebMGA (Web Services for metagenomic analysis) (Wu et al., 2011) that used the COG (Clusters of Orthologous Genes) database for functional annotation (Galperin et al., 2021). For that, RPSBLAST was run on the prokaryotic NCBI COG database with an E-value cut-off of 0.001. Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology and the prediction of pathway-specific metabolic functions, as well as the reconstruction of KEGG pathways, were facilitated through the utilization of BlastKOALA (KEGG Orthology And Links Annotation) (Kanehisa et al., 2016). This analysis was performed at the genus level using BLASTp to search and compare the data with a non-redundant dataset of pangenome sequences.

Genomic sequences were subjected to subsystem annotation using the Rapid Annotation using Subsystem Technology (RAST) server and analyzed using the SEED viewer (Overbeek et al., 2014). A classic RAST (v. 2.0) annotation search along with FIGfam (release 70) was used for curation of the genomic data. As the automatic annotation process may run into problems, such as overlapping gene pairs or overlapping RNAs these errors and frameshifts were fixed automatically (even if that requires deleting some gene candidates). Debug statements were turned on and gaps were backfilled, allowing RAST to blast large gaps for missing genes. RAST facilitated the classification of genes into predefined subsystems, which represent groups of genes involved in specific biological processes or pathways. Potential protein-coding genes were cross-referenced to coding sequences (CDS) from the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016) and linked back to UniProt accession. The integration of PATRIC (Pathosystems Resource Integration Center) was used to display and analyze RAST-annotated genomes to further investigate genome properties.

### 5.3 Functional *in silico* Analysis

Physicochemical properties such as the isoelectric point (pI) and grand average of hydropathy (GRAVY) score were calculated using ProtParam with default settings for pK values (Gasteiger et al., 2005). Phobius was used for the prediction of protein localization using a posterior probability  $\geq 0.5$  (Käll et al., 2004). Potential antimicrobial peptide (AMP) activity and their

functional targets were assessed using AMPfun (a probability score of >0.5 indicates potential AMP activity, while a score <0.5 indicates non-AMP activity) (Chung et al., 2020). Disulfide bridges were predicted using SCRATCH (Cheng et al., 2005) and functional domains and motifs were predicted using NCBI's Conserved Domains search (v. 3.20) (Wang et al., 2023a). Polysaccharide utilization loci (PULs) were assigned using the Polysaccharide Utilization Loci DataBase (PULDB) (Terrapon et al., 2015). The direction of enzymatic reactions was checked using the ExplorEnz enzyme database (<https://www.enzyme-database.org/>) (McDonald et al., 2007) and the MACiE database (Mechanism, Annotation, and Classification in Enzymes) (Holliday et al., 2005).

## 5.4 Functional and Statistical Data Analyses

**LFQ Data Normalization** – Total protein and peptide concentrations were determined and normalized by BCA assay prior to LC-MS/MS. The data normalization pipeline consisted of the following steps: (I) Data cleanup by removing proteins from potential contaminants and those with low or medium confidence levels. (II) Total intensity normalization was performed by median intensity normalization. (III) The raw and normalized intensity data were tested for normal distribution and Pearson correlation using Matplotlib in Python. (IV) Removal of proteins identified in only one out of three or two out of five biological replicates. (V) Calculate the median of all biological replicates.

**TMT Data Normalization** – Reporter ion intensities were corrected for signal interference by subtracting the percentage of interference from the measured reporter ion intensities (Savitski et al., 2013). For each PSM of the same TMT channel reporter ion intensities were normalized to one and the normalized median intensities were subtracted by the corresponding isolation interference. Occurring negative values due to recalculation were replaced by the minimum positive value in each channel. Only spectra with >50% isolation interference were used for relative quantification and subjected to normalization as described for LFQ.

**Differential Analyses and Statistical Rationale** – The Perseus software suite (v.1.6.14.0) was utilized to perform functional 1D enrichment analyses, Fisher's exact test, and two tails Welch's or Student's t-tests (Tyanova et al., 2016). Statistical tests were corrected for multiple testing applying a permutation-based or Benjamini-Hochberg FDR calculation based on the p-value distribution at 1 or 5% (Benjamini and Hochberg, 1995). Differentially abundant proteins were identified with Log<sub>2</sub> fold change  $\pm$  0.485. Significant differences were assessed using ANOVA with Dunnett's multiple comparison test to identify specific groups or conditions that were significantly different from a control group. Statistically significant differences were considered when \* (p < 0.05); \*\* (p < 0.01); \*\*\*\* (p < 0.0001). Direction pathway analysis (DPA)

was performed using the directPA package in R (v.4.2.2) to perform test statistics in two-dimensional space with a modified Pearson correlation test, to identify concordantly higher, lower, and discordantly abundant proteins (Yang et al., 2014). By specifying eight different directions, DPA was used to perform COG and KEGG pathway analysis on the selected directions, with a p-value  $\leq 0.05$  considered significant. Cleavage site specificity analysis was performed using iceLogo motifs via the standalone version of iceLogo (v.1.2) (Colaert et al., 2009). Prior to analysis, full tryptic peptides, peptides in which initiator N-terminal methionine excision, shared peptides, identical peptide sequences with multiple modifications, and peptides with canonical C-terminus were removed to eliminate false-positive cleavage sites. The calculated cleavage specificities were corrected for the natural abundance of the corresponding amino acids in the organism's proteome. Venn diagrams were created using Venny (Oliveros, 2007). The sequence coverage of top-down data was calculated using the protti package (Quast et al., 2022) in R (v.4.2.2). UpSet plots were generated using the UpSetplot and Matplotlib package (Lex et al., 2014) in Python (v. 3.11.1). Annotation of shotgun proteomics mass spectrometry data was performed using the Interactive Peptide Spectral Annotator (Brademan et al., 2019). Protein crystal structure predictions were generated using AlphaFold Colab v2.3.0 (Jumper et al., 2021) and visualized using PyMOL (Schrödinger, 2002). Experimental design workflows were created using BioRender.com.

**iBAQ Calculation** – iBAQ values were obtained by dividing a protein's total non-normalized intensity by the number of theoretically observable tryptic peptides between 7 and 30 amino acids with up to 2 missed cleavages (eq. 4).

$$\text{iBAQ} = \sum \text{intensity} / \# \text{theoretical peptides} \quad (\text{eq. 4})$$

To obtain the relative iBAQ value (riBAQ) for each protein, the iBAQ value was normalized by dividing by the sum intensity of all iBAQ values (eq. 5).

$$\text{riBAQ} = \text{iBAQ} / \sum \text{iBAQ} \quad (\text{eq. 5})$$

To estimate the relative abundance of histidine-containing proteins, the riBAQ values were scaled to 100% and multiplied by the absolute number of histidine residues in each protein (eq. 6).

$$\text{riBAQ}_{(\text{His})} = \text{riBAQ} \times \# \text{Histidine residue} \quad (\text{eq. 6})$$

# III PROTEOMIC ANALYSIS OF *B. THETA IOTAOMICRON*

---

<b>1</b>	<b>Introduction and Summary .....</b>	<b>39</b>
<b>2</b>	<b>Experimental Design .....</b>	<b>41</b>
	2.1 Comparison of Label-Free and TMT-Based Quantification .....	41
	2.2 Proteoform-Directed Analysis.....	42
<b>3</b>	<b>Results.....</b>	<b>43</b>
	3.1 Evaluation of TMT Labeling Efficiency and Sample Consistency .....	43
	3.2 Comparison of LFQ and TMT Proteomic Analyses.....	45
	3.3 Carbohydrate-Dependent Protein Abundance .....	47
	3.4 Proteoform-Directed Top-Down Analysis .....	50
	3.5 Analysis of Proteoform Termini .....	51
	3.6 Discovery-Based Open Modification Search.....	54
<b>4</b>	<b>Discussion and Conclusion.....</b>	<b>57</b>
	4.1 Future Direction for Quantitative Analysis .....	57
	4.2 Proteoform-Directed Analysis of <i>B. thetaiotaomicron</i> 's Proteome .....	59

Parts of the following chapter have been published in “*The intracellular proteome of the gut bacterium Bacteroides thetaiotaomicron is widely unaffected by a switch from glucose to sucrose as main carbohydrate source*” Genth et al., Proteomics, 22(22), 1–6, (2022).

#### **Supplementary material for (Genth et al., 2022)**

Additional supplementary information's is freely available for download at the publisher's website <https://www.doi.org/10.1002/pmic.202200189>. The MS proteomics raw data and complete Proteome Discover search results have been deposited to the ProteomeXchange Consortium (<http://www.proteomexchange.org/>) via the PRIDE (Vizcaíno et al., 2014) partner repository with the data set identifier PXD033704.

## 1 Introduction and Summary

The human gut microbiota significantly enhances the nutritional value of human diets by breaking down macromolecules and activating key intestinal genes to facilitate nutrient absorption (Ecklu-Mensah et al., 2022). Studies in germ-free mice have demonstrated that microbial colonization reduces the required caloric intake for weight maintenance by 30% and induces rapid changes in body fat composition (Wostmann et al., 1983; Bäckhed et al., 2004). Dietary choices can influence the composition of the gut microbiota and the immune response, both of which are key factors in the development and progression of inflammatory bowel disease (IBD) (Dolan and Chang, 2017). Individuals with IBD commonly develop increased sensitivity or intolerance to certain foods, leading them to avoid specific dietary components (Ballegaard et al., 1997; Zallot et al., 2013).

Notably, the adoption of Western dietary habits, often associated with high levels of processed foods and refined carbohydrates, significantly increases the risk of gastrointestinal inflammatory disorders, including IBD (Ng, 2014; Khademi et al., 2021). However, it is important to acknowledge that dietary interventions can also be utilized to promote healthy gut microbial functions (Ecklu-Mensah et al., 2022).

Within the human gut, *Bacteroides* represents one of the most abundant genera of bacteria (Arumugam et al., 2011). *Bacteroides thetaiotaomicron*, constituting approximately 6% of the total gut microbiota, plays a crucial role in reinforcing the mucosal barrier, maintaining immune response homeostasis, and processing nutrients (Zocco et al., 2007). Its complex repertoire of glycosylhydrolases enables the metabolism of a wide range of otherwise indigestible dietary polysaccharides and host-derived glycans in the human gut (Xu and Gordon, 2003).

The impact of dietary sugars on the competitive dynamics within the human gut microbiota has great consequences. Prolonged consumption of a high-sugar diet can significantly alter gut microbial diversity (Do et al., 2018; Alasmar et al., 2023), leading to a displacement characterized by reduced levels of *Bacteroidetes*, similar to the dysbiosis observed in IBD (Frank et al., 2007). This dietary shift prompts bacteria to employ various adaptive mechanisms, which can include the application of specific carbohydrate transport and utilization systems or virulence genes that mediate toxin production or immune evasion (Poncet et al., 2009).

While monosaccharides, such as glucose and fructose, may arrest the colonization of *B. thetaiotaomicron* in the gut by suppressing the expression of colonization-related proteins (Townsend et al., 2019), increased glucose intake in mice promotes mucolytic bacteria, including *B. fragilis* (Khan et al., 2020). This promotion could potentially compromise the integrity of the protective intestinal mucosal barrier, a critical factor in initiating intestinal inflammation (Png et al., 2010). The contradictory findings highlight the need for further research on this prominent genus and its sugar-microbiota interactions.

This chapter, prompted by the observed correlation between increased sugar consumption and the potential development of IBD (Khademi et al., 2021), aims to evaluate proteomic changes in *B. thetaiotaomicron* in response to different sugar sources.

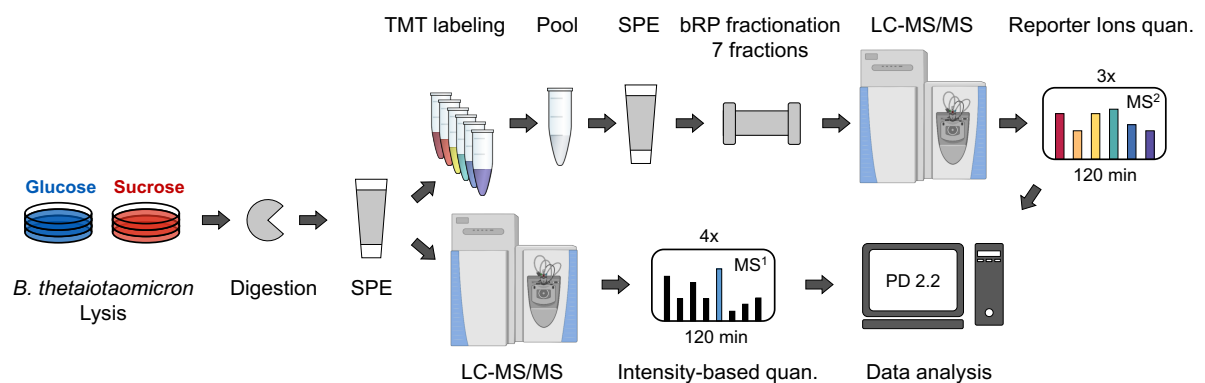
#### **Aim of this study:**

- Utilize two quantitative bottom-up proteomics analyses – label-free quantification (LFQ) and isobaric labeling-based quantification using TMT – to evaluate proteomic changes induced by the presence of sucrose and glucose.
- Perform a comparative analysis of quantitative results to determine the optimal methodology for future proteomic quantitative analysis.
- Identify proteins whose abundance levels are influenced by the type of carbohydrate.
- Apply various sample preparation techniques to deplete the high-molecular-weight proteome and increase the coverage of the low-molecular-weight proteome.
- Conduct a proteoform-directed top-down analysis and employ a discovery-based open modification search to identify post-translational modifications and potential proteolytic cleavage events.

## 2 Experimental Design

### 2.1 Comparison of Label-Free and TMT-Based Quantification

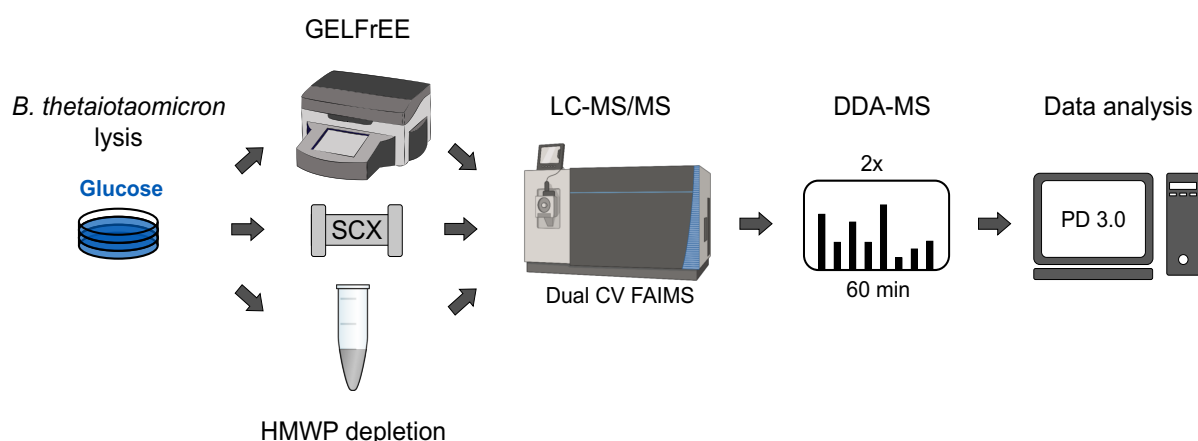
LFQ and TMT-based quantification were compared using intracellular proteome samples from *B. thetaiotaomicron* (FIGURE III-1). Cultures were grown in triplicates in YCFA medium using the monosaccharide glucose and the disaccharide sucrose as carbon sources, allowing them to reach the stationary phase up to 60 hours (chapter II.2.1). After cell lysis using freeze-thawing and protein clean-up using ethanol precipitation (chapter II.3.1), samples were digested using trypsin at a 1:40 enzyme to substrate ratio and subjected to solid-phase extraction (chapter II.3.4). For LFQ analysis, peptides were resuspended in loading solution and analyzed by LC-MS. In contrast, the TMT-based workflow involved isobaric labeling using TMT-6-plex with a reduced TMT-to-peptide ratio (Innovation and Zecha, 2019). Labeled peptides were then separated using basic reversed-phase liquid chromatography (bRP) and concatenated into seven pools (FIGURE III-1). Details about TMT labeling and bRP are described in chapter II.3.5. Samples were separated online using reversed-phase chromatography, employing a gradient of 120 minutes (chapter II.4.1). Following separation, samples were analyzed using the Q-Exactive HF mass spectrometer, with specific parameters varying between the two measurements (chapter II.4.1). Non-labeled peptides were measured in quadruplicates with an isolation window of 2.0  $m/z$  and HCD fragmentation employing an NCE of 27%. In contrast, TMT-labeled peptides were measured in triplicate with an isolation window of 1.5  $m/z$  and HCD fragmentation with an NCE of 35%. The acquired raw data were searched and compared to the *B. thetaiotaomicron* reference proteome using PD 2.2 software (chapter II.5.1). Data were filtered by removing contaminants and requiring proteins to be identified with at least two peptides, one being unique. Quantification required quantitative values from all three biological replicates, and PSMs had to meet a minimal interference threshold (<50%) for isobaric-labeled peptides.



**FIGURE III-1 | Experimental Design for the Comparison of Label-free and TMT-based Quantification.** *B. thetaiotaomicron* cultures, grown in triplicate in YCFA medium supplemented with either glucose or sucrose, were processed using a label-free workflow or subjected to a TMT labeling workflow, including peptide fractionation.

## 2.2 Proteoform-Directed Analysis

To identify proteoforms of *B. thetaiotaomicron*, cultures grown in the presence of glucose were subjected to GELFrEE separation (Toby et al., 2019), SCX fractionation (Cassidy et al., 2021a) and acidic and basic depletion of the high-molecular-weight proteome (HMWP) (Cassidy et al., 2019) (FIGURE III-2). Details about the sample preparations are described in chapter II.3.2. Samples were separated using a 60 min gradient and analyzed using the dual CV FAIMS method (chapter II.4.2). The acquired raw data were searched and compared to the *B. thetaiotaomicron* reference proteome using PD 3.0 software (chapter II.5.1). Additional database search with a 500 Da precursor ion mass tolerance, was performed to evaluate possible PTMs and chemical modifications. Manual curating mass shifts using PSI-MOD (Montecchi-Palazzi et al., 2008) and UniMod (Creasy and Cottrell, 2004) databases allowed the elucidation of mass shifts. Identified proteoforms were required to have a minimum C-Score of 40 (Leduc et al., 2014).



**FIGURE III-2 | Experimental Design for Top-Down Proteoform-Directed Analysis.** Cultures of *B. thetaiotaomicron*, grown in triplicate in YCFA medium supplemented with glucose, were subjected to GELFrEE separation, SCX fractionation, and HMWP depletion.

### 3 Results

#### 3.1 Evaluation of TMT Labeling Efficiency and Sample Consistency

The applied TMT labeling protocol utilized a TMT to-peptide ratio of 4:1 (w/w) compared to the manufacturer's recommended 20-fold excess (Innovation and Zecha, 2019). Despite expecting some degree of reagent hydrolysis, the applied TMT to peptide ratio still resulted in an excess of labeling reagent. Nevertheless, only a small fraction of peptides (less than 0.3%) remained unlabeled or partially labeled, where the  $\epsilon$ -amino group of lysine or the peptide N-terminus was not labeled. Consequently, the labeling efficiency was determined to be greater than 99.8%, indicating highly effective labeling (TABLE III-1).

TABLE III-1 | Effectiveness of TMT labeling

SAMPLE	TMT-REAGENT
Total peptides	8186
N-term labeled peptides	8164
Lysine-containing peptides	5199
$\epsilon$ -N-labeled lysine peptides	5197
Label efficiency	<b><u>99.82%</u></b>

Analysis of amino acid side chain over labeling revealed that approximately 6.9% of the peptide population contained at least one TMT-labeled histidine, serine, threonine, or tyrosine residue. Notably, serine and histidine residues were predominantly affected, accounting for 38.9% and 32.8% of the overlabeled peptides, respectively (FIGURE III-3). In contrast, threonine and tyrosine residues were less abundant, together accounting for 28.3% of the overlabeled peptides (FIGURE III-3).

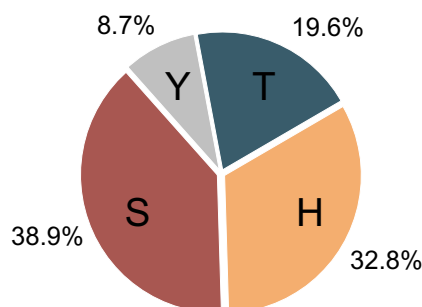
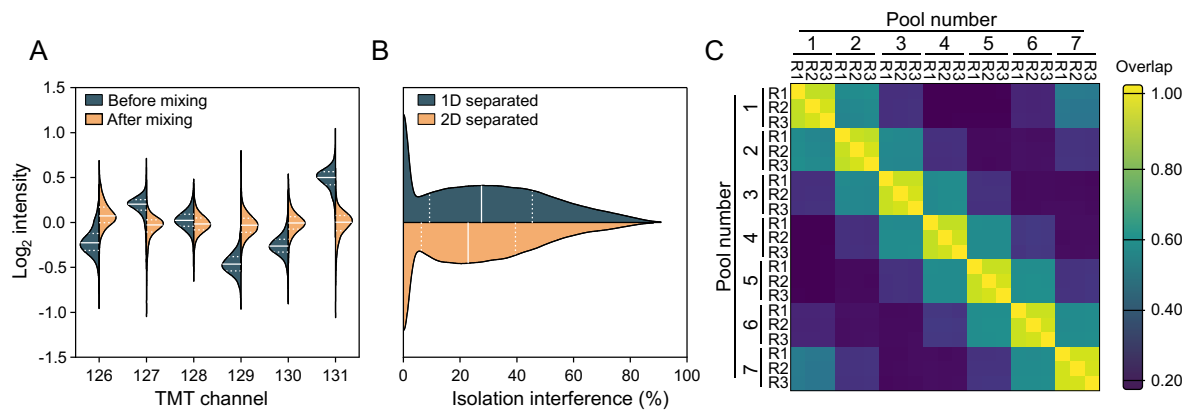


FIGURE III-3 | **Overview of TMT Overlabeling.** Distribution of amino acid side chain over labeling of histidine, serine, threonine, and tyrosine side chains.

Systematic bias from mixing imperfections in the six TMT channels was evaluated by combining small sample volumes of each channel, measuring the pool by LC-MS, and evaluating median protein abundances. Adjusting the mixing effectively equalized intensities across all channels (FIGURE III-4A), effectively minimizing variations in protein abundance.



**FIGURE III-4 | Evaluation of Mixing and Two-dimensional Separation.** (A) Distributions of Log<sub>2</sub> TMT intensities for all six TMT channels before and after mixing. (B) Isolation interference before and after two-dimensional chromatographic separation (2D). (C) Pearson correlation heatmap representing the distance matrix of identified PSMs between the seven individual pools and their three respective technical replicates.

The application of two-dimensional chromatographic separation reduced the occurrence of co-isolation interference above 50% by 3.1% and below 30% by 4.6%, respectively (FIGURE III-4B). Analysis of Pearson correlation coefficients revealed that approximately 30% of the PSMs were unique to a single fraction, while the majority was also identified in either the preceding or subsequent fractions (FIGURE III-4C). Notably, the use of triple injections resulted in an increase of  $11.4\% \pm 1.6\%$  in non-redundant PSM identifications, indicating the advantages of employing multiple injections to increase the potential the peptide identifications and overall protein sequence coverage. Further examination of the total number of PSMs per pool indicated that peptides derived from heat shock proteins and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were the most abundant (TABLE III-2).

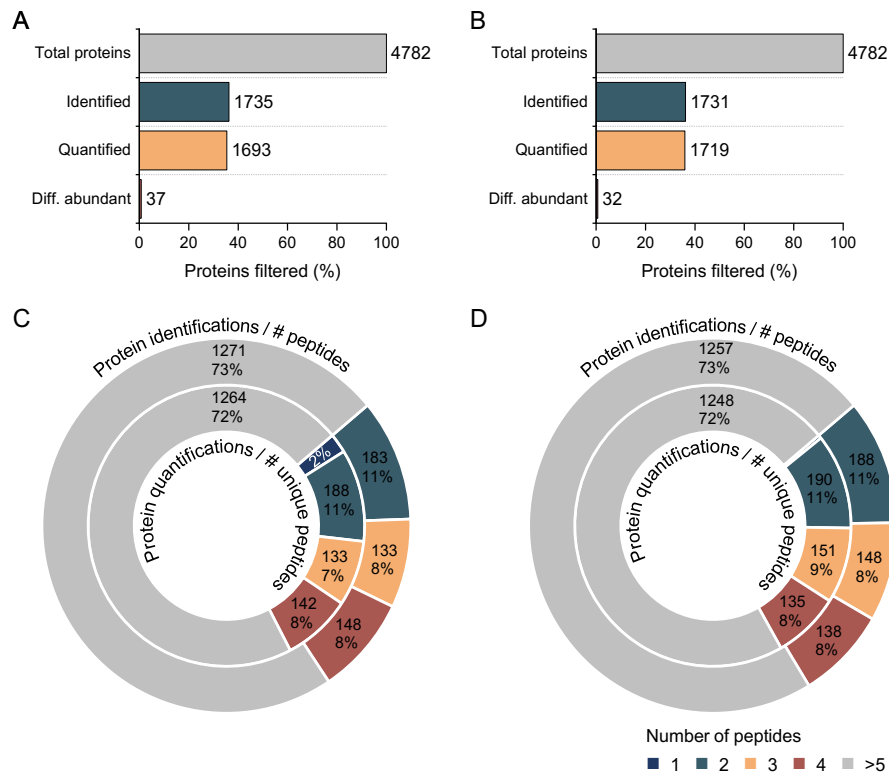
In conclusion, the selected fractionation and pooling approach effectively improved the reduction of co-isolation interference and repetitive measurements.

**TABLE III-2 | Top 10 Peptides Across the Seven Pools.** Abbreviation: GAPDH: glycerinaldehyd-3-phosphat-dehydrogenase.

		$\Sigma$ PSMs PER POOL							
PEPTIDE SEQUENCE	PROTEIN NAME	1	2	3	4	5	6	7	$\Sigma$ 1-7
[K].VE...TK.[N]	Small heat shock	34	17	21	56	29	95	92	344
[K].FQ...DK.[E]	Heat shock protein	20	19	21	31	35	89	62	277
[K].NV...IK.[R]	60 kDa chaperonin	29	28	53	81	36	17	23	267
[K].AG...VK.[V]	GAPDH	42	85	47	22	17	15	16	244
[R].ID...EK.[K]	Heat shock protein	56	61	28	24	21	29	23	242
[K].GI...TR.[T]	GAPDH	12	48	65	26	20	18	18	207
[K].GT...DR.[G]	60 kDa chaperonin	22	12	70	48	25	10	17	204

### 3.2 Comparison of LFQ and TMT Proteomic Analyses

The LFQ and TMT proteomic analyses identified comparable numbers of proteins (1735 and 1731, respectively) and quantified similar numbers across all three biological replicates in at least one group (1693 and 1719, respectively) (FIGURE III-5A-B). Overall, both methodologies covered about 35% of the predicted *B. thetaiotaomicron* proteome (FIGURE III-5A-B). On average, protein identification using both quantitative methods were supported by 10 peptides, with over 73% of identifications and 72% of quantifications supported by at least five peptides (FIGURE III-5C- D). Utilizing multiple peptides for protein quantification generally enhances the accuracy and confidence of quantitative proteomic analysis by mitigating the risk of measurement errors or outliers (Bantscheff et al., 2012).

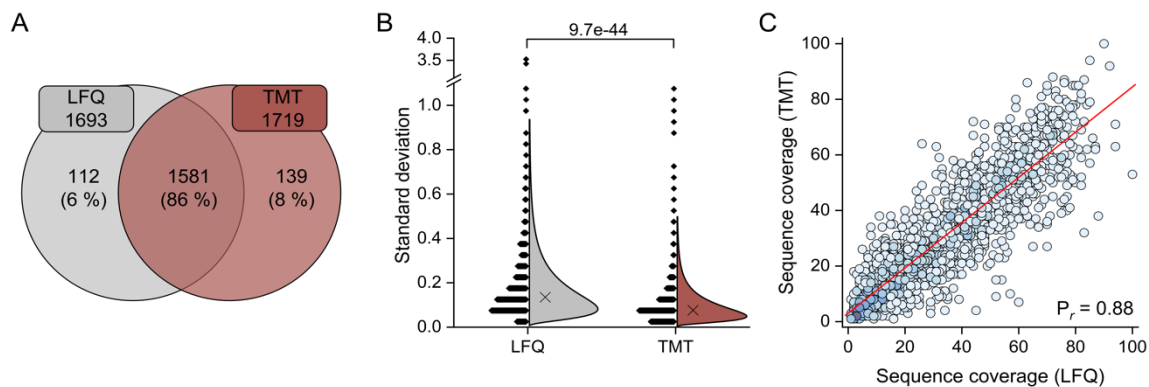


**FIGURE III-5 | Protein and Peptide Identification Overview.** A total number of proteins encoded in the genome, identified proteins, identified proteins, quantified proteins (quantitative values in all three biological replicates of at least one group), and differentially abundant proteins for (A) LFQ or (B) TMT. Pie chart illustrating the number of peptides supporting protein identifications and unique peptides supporting quantitation of the filtered protein groups for (C) LFQ or (D) TMT.

The majority of quantified proteins (86%) were detected using both methods, with unique quantifications observed in 6% of proteins with LFQ and 8% with TMT (FIGURE III-6A).

While LFQ included an additional technical replicate, the multiplexed TMT analysis still exhibited significantly higher reproducibility (standard deviation of 0.076) compared to LFQ (standard deviation of 0.135) (FIGURE III-6B). Both methods achieved comparable protein sequence coverage, with LFQ at 37% and TMT at 33%. Moreover, they exhibited a high

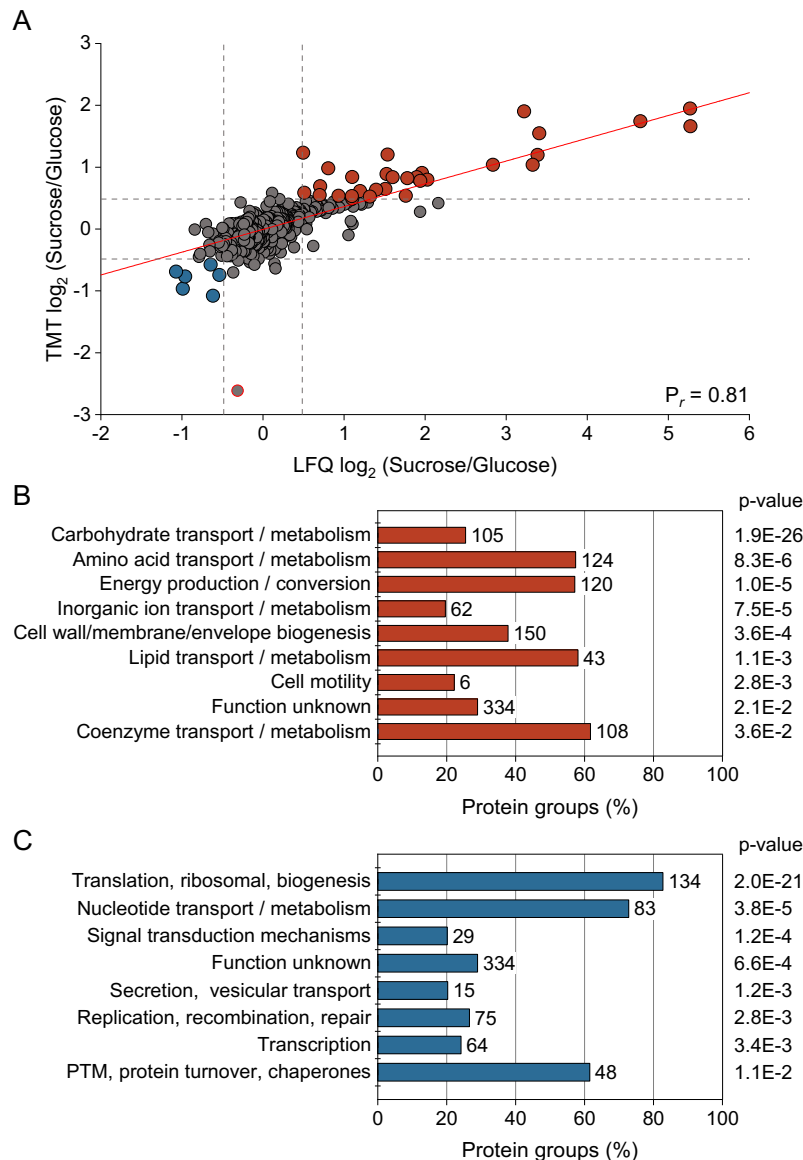
Pearson correlation coefficient of 0.88 (FIGURE III-6C), indicating consistent protein sequence coverage for identical proteins. Overall, both analyses demonstrated a high level of comparability in their quantitative data.



**FIGURE III-6 | Comparison of Quantified Proteins for LFQ and TMT. (A)** Venn diagram illustrating the overlap and complementarity of quantified proteins. **(B)** Violin plot showing the distribution of protein standard deviations and the significant difference in medians (two-sample Student's t-test). **(C)** Correlation of sequence coverage, where the red line indicates the Pearson correlation ( $P_r$ ).

Directional analysis (Yang et al., 2014) of the 1,581 shared quantified proteins identified 32 proteins with higher abundance in the presence of sucrose and 6 proteins with higher abundance in the presence of glucose in both the LFQ and TMT datasets ( $p \leq 0.05$ , FIGURE III-7A and TABLE A-1). Although a single  $\text{Na}^+/\text{H}^+$  antiporter exhibited opposing abundance changes between LFQ and TMT, it did not reach statistical significance and was treated as an outlier. The absence of significantly discordant proteins and the high Pearson correlation coefficient of 0.81 indicated a strong concordance between LFQ and TMT in capturing the proteomic response to changing carbohydrate conditions (FIGURE III-7A).

The directional pathway analysis on proteins of higher abundance in the presence of sucrose revealed significant enrichment ( $p < 0.01$ ) in carbohydrate transport and metabolism, energy production and conversion, as well as various transport and metabolism categories, including amino acid, inorganic ion, lipid, and coenzyme transport and metabolism (FIGURE III-7B). These findings suggest the active involvement of *B. thetaiotaomicron* in sucrose uptake and processing for energy production. Conversely, analysis of proteins of higher abundance in the presence of glucose revealed a predominant enrichment of proteins involved in transcription and translation processes (FIGURE III-7C). This suggests a cellular response favoring gene expression and protein synthesis in the presence of glucose compared to the presence of sucrose. However, among the 6 proteins with higher abundance in the presence of glucose, four were domain-containing proteins, one was a translocase protein (Q8A0B9), and a D-ribitol-5-phosphate phosphatase (Q8A947) (TABLE A-1). None of these proteins are directly involved in translation or transcription processes.

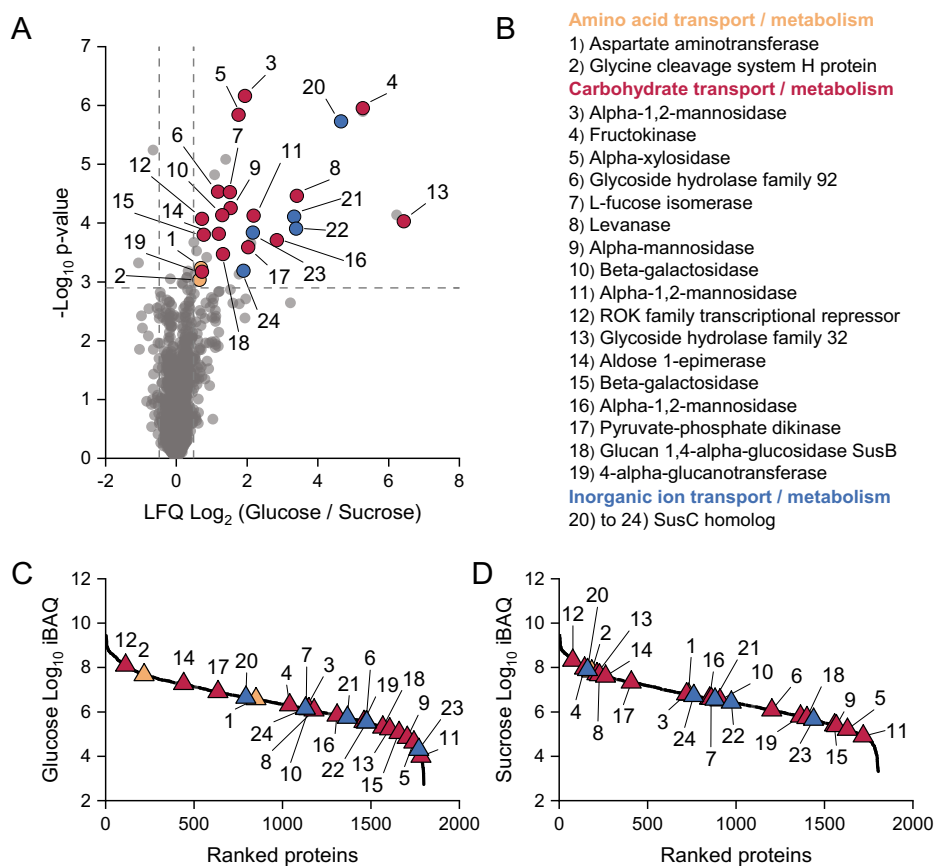


**FIGURE III-7 | Analysis of the Proteomic Response between LFQ and TMT.** (A) Comparison of  $\log_2$  ratios of sucrose versus glucose for LFQ and TMT approaches. Proteins exhibiting higher abundance in the presence of either sucrose (32 proteins, red dots) or glucose (6 proteins, blue dots) are highlighted. The significance ( $p \leq 0.05$ ) was determined using a modified Pearson's correlation test (Yang et al., 2014). The red line represents the Pearson correlation ( $P_r$ ), excluding one outlier. Direction pathway analysis of proteins of higher abundance in the presence of (B) sucrose or (C) glucose. Both graphs display the number of selected proteins in each COG category, the percentage of proteins in the specified category, and the corresponding p-values.

### 3.3 Carbohydrate-Dependent Protein Abundance

Statistical analysis (two-sided Welch's t-test with Benjamini-Hochberg FDR correction for multiple testing,  $q \leq 0.05$  and  $\log_2$  fold change of  $\pm 0.485$ ) identified 37 differentially abundant proteins for LFQ (FIGURE III-5A) and 32 for TMT (FIGURE III-5B). Visualization of the LFQ results revealed differential abundance of proteins associated with carbohydrate metabolism (red dots), amino acid transport and metabolism (yellow dots), and inorganic ion transport and

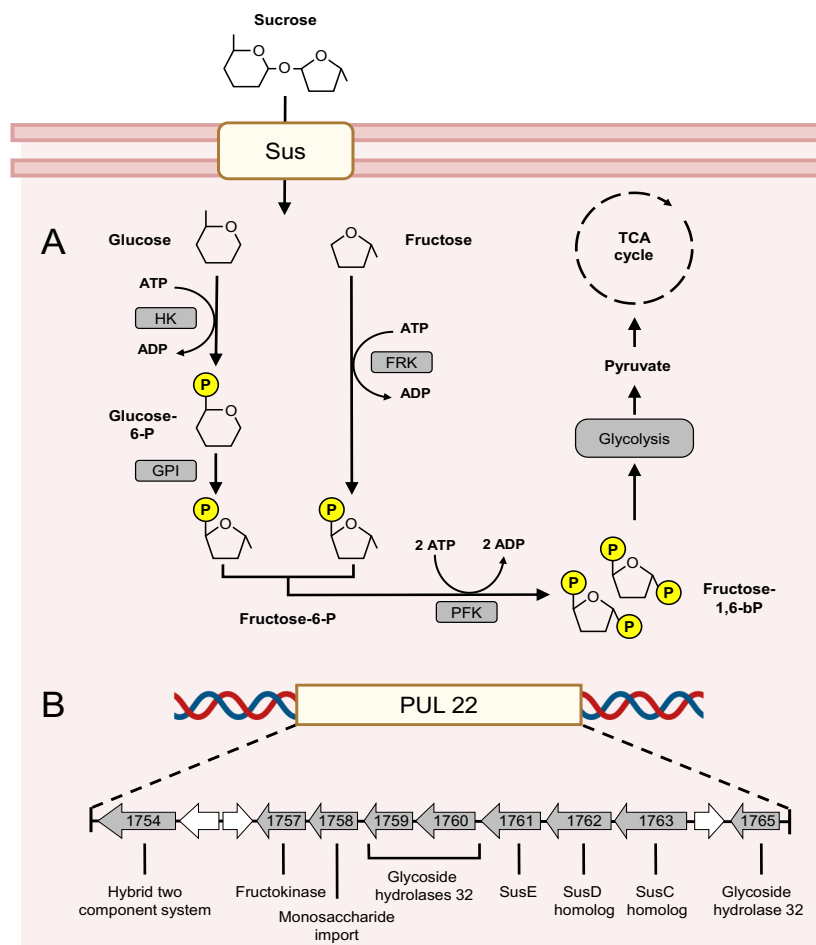
metabolism (blue dots) (FIGURE III-8A, B). Notably, certain proteins such as aldolase 1-epimerase (Q8AAU2), pyruvate-phosphate dikinase (Q8AA21), fructokinase (Q8A6W9), glycine cleavage system H protein (Q8A4S8), and ROK family transcriptional repressor (Q8A4V4) consistently exhibited high intensity-based absolute quantification (iBAQ) values, regardless of the carbon source used in cultivation (FIGURE III-8C-D). These values, calculated by dividing a protein's total non-normalized intensity by the number of theoretically observable tryptic peptides for that protein, indicate the relative absolute abundance of proteins. These results suggest consistent abundance levels of certain proteins and may indicate their potential relevance in diverse cellular processes under different culture conditions.



**FIGURE III-8 | Overview of LFQ Quantitative Results.** (A) Volcano plot with dashed vertical lines represents  $\text{Log}_2$  cutoffs and the dashed horizontal line represents a q-value of 0.05 (two-sided Welch's t-test with Benjamini-Hochberg FDR correction for multiple testing). (B) Proteins involved in amino acid, carbohydrate, or inorganic ion transport and metabolism according to COG annotations. iBAQ intensities of proteins identified in the presence of (C) glucose or (D) sucrose.

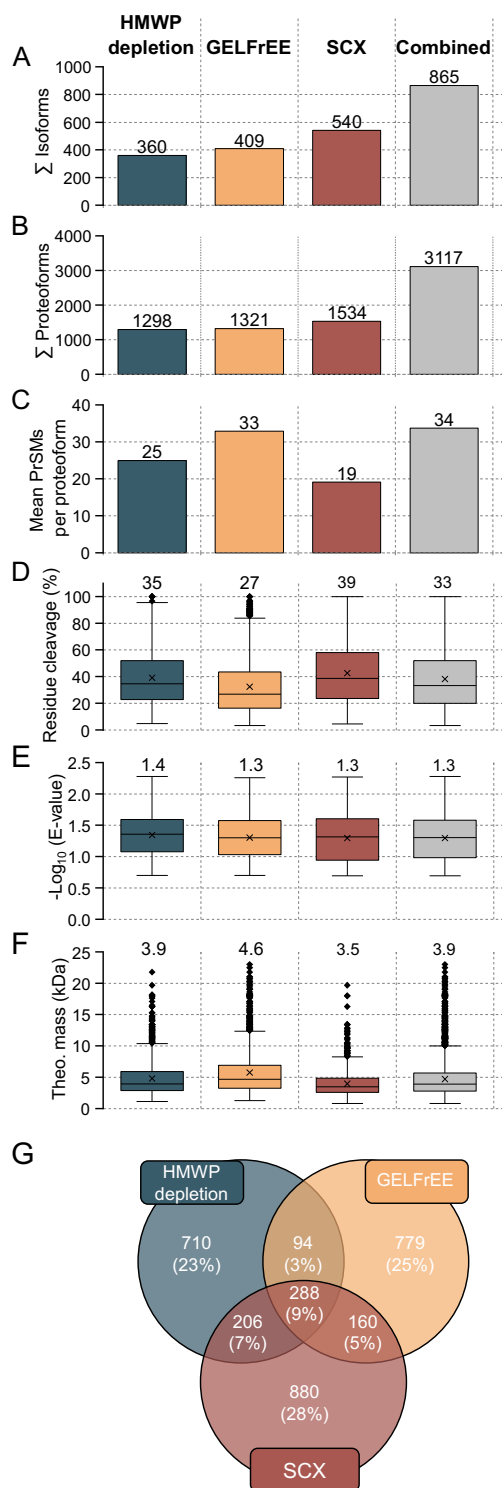
Additionally, an increased abundance of proteins associated with fructan metabolism was identified, suggesting a specialized adaptation for the utilization of fructose-based oligosaccharides. This adaptation was facilitated by the starch utilization system (Sus), encoded within the polysaccharide utilization locus (PUL) 22, which includes eight open reading frames (FIGURE III-9B) (Martens et al., 2008; Sonnenburg et al., 2010). The regulation of this operon is mediated by a hybrid two-component system, positioned adjacent to the PUL, enhancing the efficiency of fructose-based carbohydrate utilization (Sonnenburg et al., 2010).

Periplasmic hydrolysis of sucrose breaks it into its constituent monosaccharides, glucose, and fructose, which are then imported into the cytoplasm and directed toward central metabolic pathways (FIGURE III-9A). Glucose, a preferred energy source in numerous microorganisms, can access central metabolic pathways like glycolysis through the Embden-Meyerhof pathway. This pathway generates pyruvate, ATP, and precursor metabolites for the tricarboxylic acid (TCA) cycle. Unlike other carbohydrates, fructose could also directly enter the Embden-Meyerhof pathway for energy generation, without the need for additional energy-consuming conversion steps (FIGURE III-9A). This metabolic pathway confers an advantage, enhancing the efficiency of fructose utilization across various bacteria, including *B. thetaiotaomicron*.



**FIGURE III-9 | Sucrose Utilization Pathway in *B. thetaiotaomicron*.** (A) Sucrose uptake occurs via the Sus system and upon the cleavage of the O-glycosidic bond, glucose and fructose molecules are channeled into the major glycolytic pathways. The generated pyruvate enters the tricarboxylic acid (TCA) cycle for further processing. (B) Polysaccharide utilization locus (PUL) 22 encodes for the Sus system. Enzymes essential to this catabolic process are highlighted in gray and represented by the following abbreviations: HK (hexokinase), GPI (glucose-6-phosphate isomerase), FRK (fructokinase), and PFK (phosphofructokinase).

### 3.4 Proteoform-Directed Top-Down Analysis



**FIGURE III-10 | Top-down Methodology Comparison.** Number of (A) isoforms, (B) proteoforms, (C) mean PrSMs per proteoform. Distribution of (D) residue cleavage, (E)  $-\log_{10}$  (E-value) and (F) theoretical proteoform mass. (G) Proteoform overlap between different methods.

To increase the coverage of the low-molecular-weight proteome, three distinct methods were employed. These methods included two HMWP depletion techniques, GELFrEE fractionation, and SCX fractionation. The number of identifications per method ranged from 360 to 540 protein groups (FIGURE III-10A) and 1,298 to 1,534 proteoforms (FIGURE III-10B), respectively. Notably, all methods displayed comparable identification metrics, as demonstrated by the mean PrSMs per proteoform (FIGURE III-10D), residue cleavage (FIGURE III-10D),  $-\log_{10}$  (E-value) (FIGURE III-10E), and the size distribution of the identified proteoforms (FIGURE III-10F).

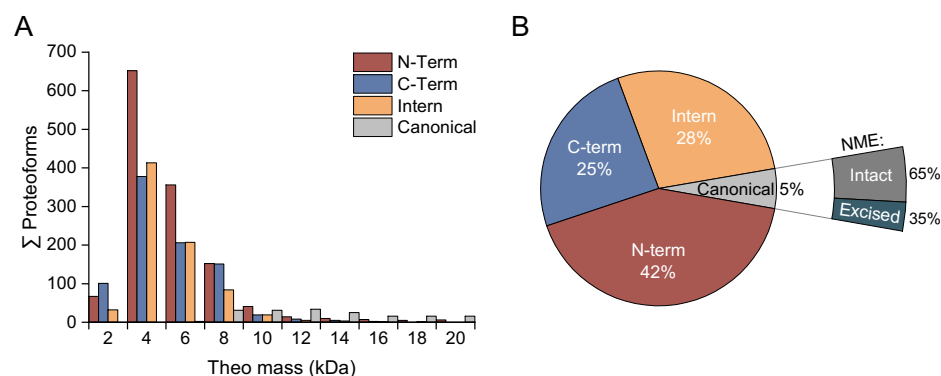
The overlap between the different methods ranged from 3% to 9%, and the majority of proteoforms were uniquely identified by individual methods, constituting approximately 25% each (FIGURE III-10G). Combining all three database searches resulted in the identification of 865 protein groups (FIGURE III-10A), represented by a total of 3,117 proteoforms (FIGURE III-10B).

Detailed information on the exact identification metrics for each method is provided in Appendix 1.2. In summary, the number of identifications per SCX fraction ranged from 125 to 403 protein groups (FIGURE A-12A) and 185 to 1,280 proteoforms (FIGURE A-12B). The highest number of protein groups and proteoform identifications was achieved in fraction 3, which mostly contains higher charged species (FIGURE A-12A-B). Comparison of the acidic and basic HMWP depletion showed that comparable numbers of

protein groups and proteoform identifications were obtained (FIGURE A-12A-B), with each depletion method contributing a comparable number of unique identifications (FIGURE III-10A-B). A comparison of GELFrEE fractions analyzed using different stationary phases (C<sub>4</sub> and PLRP-S), indicated a generally higher number of protein groups (FIGURE A-12G) and proteoform identifications for the C<sub>4</sub> column (FIGURE A-12H). The observed increase in the average number of PrSMs per proteoform identification of the PLRP-S column (FIGURE A-12I) can be attributed to decreased resolution due to wider elution profiles, resulting in peak broadening and poor separation of adjacent peaks compared to the C<sub>4</sub> column (FIGURE A-12J-K). These deviations are due to differences in column characteristics, including the larger particle size (5.0  $\mu\text{m}$ , 1000  $\text{\AA}$ ), shorter column length (17 cm), and potential dead volume introduced during crimping with ZIRCOFIT UHPLC fittings of the PLRP-S column, as opposed to the commercially purchased C<sub>4</sub> column (2.6  $\mu\text{m}$ , 150  $\text{\AA}$ , 50 cm). These differences may lead to fewer proteoforms being effectively separated in a given timeframe for the PLRP-S column. It is important to note that both columns operated at identical flow rates, gradients, and eluents (for detailed information, refer to chapter II.4.2). Therefore, optimization of gradient parameters, including the design of a nonlinear gradient for PLRP-S columns, provides an opportunity to improve chromatographic separation and increase the number of proteoforms (Trudgian et al., 2014).

### 3.5 Analysis of Proteoform Termini

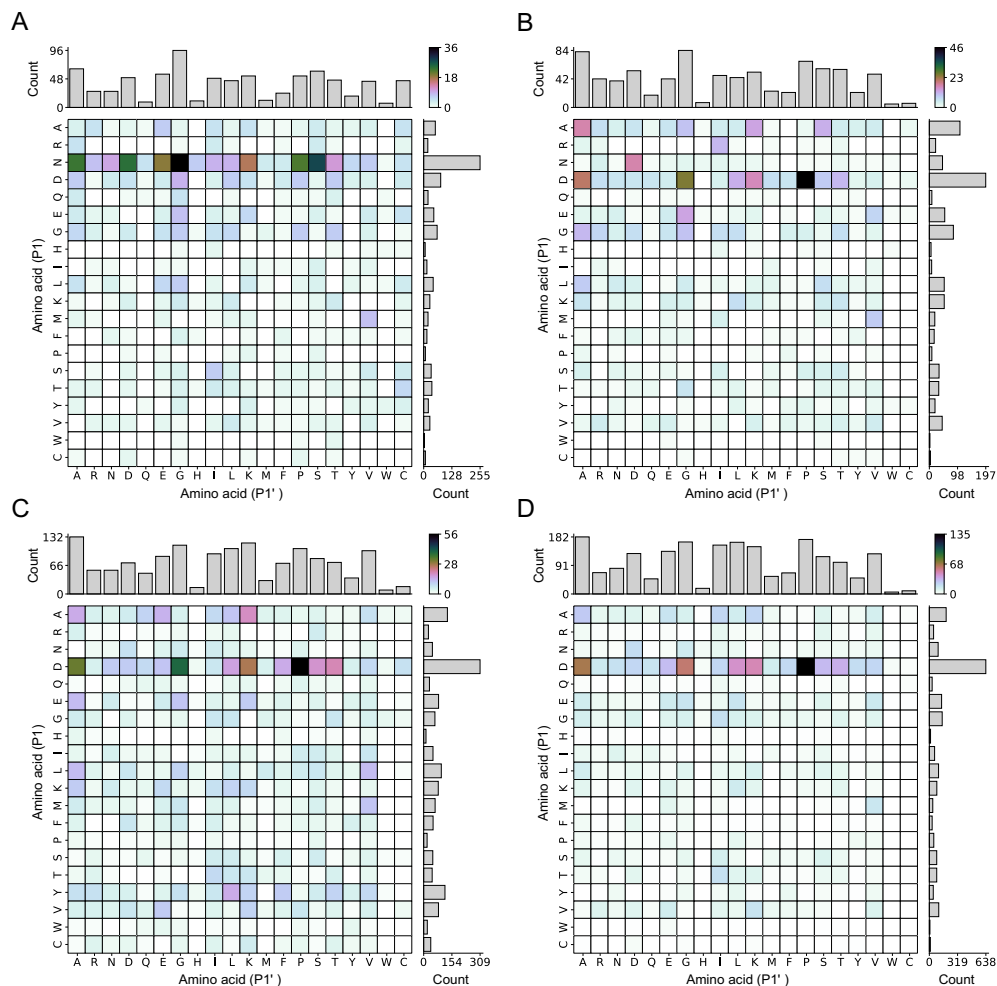
Out of 3,117 identified proteoforms, 2,942 (94%) had molecular weights between 2-10 kDa and consisted mostly of truncated proteoforms (FIGURE III-11A). N-terminal truncation accounted for 42% of the proteoforms, excluding N-terminal methionine excisions (NME), while C-terminal truncation accounted for 25% of the proteoforms (FIGURE III-11B). Notably, the heat shock protein (Q8AAA0) and the 60 kDa chaperonin (Q8A6P8) were prevalent in both TDP and BUP analyses, exhibiting numerous N- and C-terminal truncated proteoforms (38 and 36 respectively, TABLE A-2) and multiple peptides having high PSM counts (TABLE III-2).



**FIGURE III-11 | Neo-termini Analysis of Identified Proteoforms. (A)** Size distribution and **(B)** percentage distribution of identified proteoforms, including N- and/or C-terminal truncated proteoforms with N-terminal initiator methionine excision and intact initiator methionine.



Analysis of the N-terminal residues preceding the cleavage site (P1), and the C-terminal amino acids following it (P1'), revealed a uniform distribution of abundance for the majority of the detected neo-termini (FIGURE III-13). However, the detection of proteoforms with either aspartate at P1 or proline at P1' led to the detection of a prominent Asp-Pro sequence logo in the acidic HMWP depletion, GELFrEE fractionation, and SCX fractionation experiments (FIGURE III-13 B-D). Conversely, in the HMWP deletion performed under basic conditions (pH 8.5), this specific sequence pattern was not observed (FIGURE III-13A). Instead, the cleavage data showed an increased abundance of two distinct proteoforms: one featuring asparagine at P1, and the other featuring either glycine or serine at P1', resulting in noticeable sequence logos, Asn-Gly and Asn-Ser, respectively (FIGURE III-13A). The specific cleavage patterns, which varied depending on the method and pH of the depletion used, suggest that artificial cleavage events may have been introduced during sample preparation or LC-MS/MS analysis.



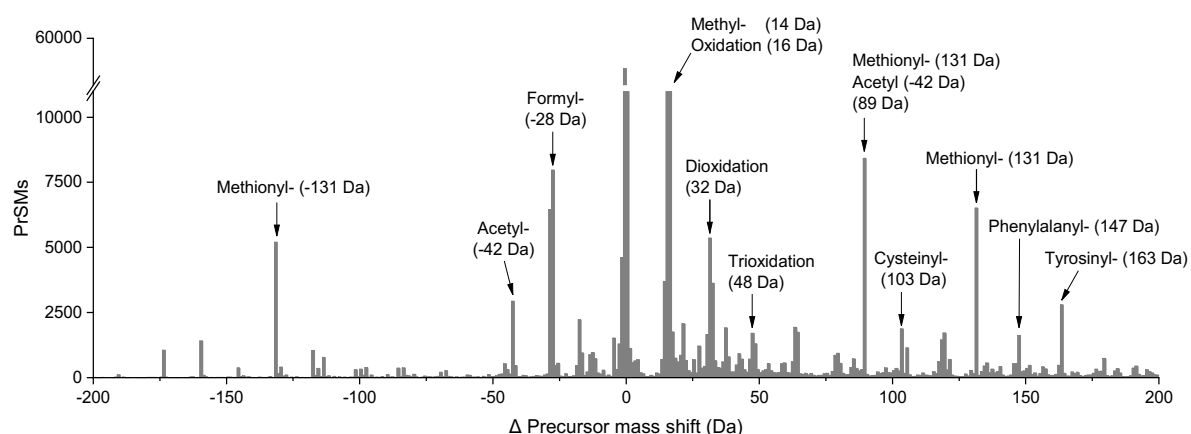
**FIGURE III-13 | Analysis of Proteoform Neo-Termini.** Proteoform analysis for (A) basic HMWP depletion, (B) acidic HMWP depletion, (C) GELFrEE fractionation, and (D) SCX fractionation. Heatmaps illustrate the distribution and intensity of various proteoforms characterized by their N-terminal (P1) and C-terminal (P1') residues, while bar plots summarize the total numbers of proteoforms identified.

### 3.6 Discovery-Based Open Modification Search

During the preparation of biological samples and conducting LC-MS/MS analysis, artificial modifications, non-covalent adduct formation, and artificial truncation events may occur (Schaffer et al., 2021). To address this issue and to identify both known and novel modifications, a discovery-based open modification search was performed. This allowed for the identification of precursor mass shifts without a predefined list of PTMs.

While several mass shifts hinted towards potential interesting PTMs on proteoforms of *B. thetaiotaomicron*, it's crucial to emphasize that these identifications are solely based on MS<sup>1</sup> intensities, and mass shifts may also be attributed to inaccuracies in precursor mass deconvolution (Jeong et al., 2020). Therefore, the following results serve as preliminary indications and warrant further validation. Thousands of peptide spectrum matches (PrSMs) per mass shift were detected, each exhibiting varying degrees of mass accuracy (precursor mass tolerance: 10 ppm). This resulted in a broad distribution of delta precursor mass shifts. Following manual inspection of precursor isotope distribution and fragmentation spectra for potential a PTM, the monoisotopic mass corresponding to entries in the PSI-MOD (Montecchi-Palazzi et al., 2008) or UniMod database (Creasy and Cottrell, 2004) will be reported.

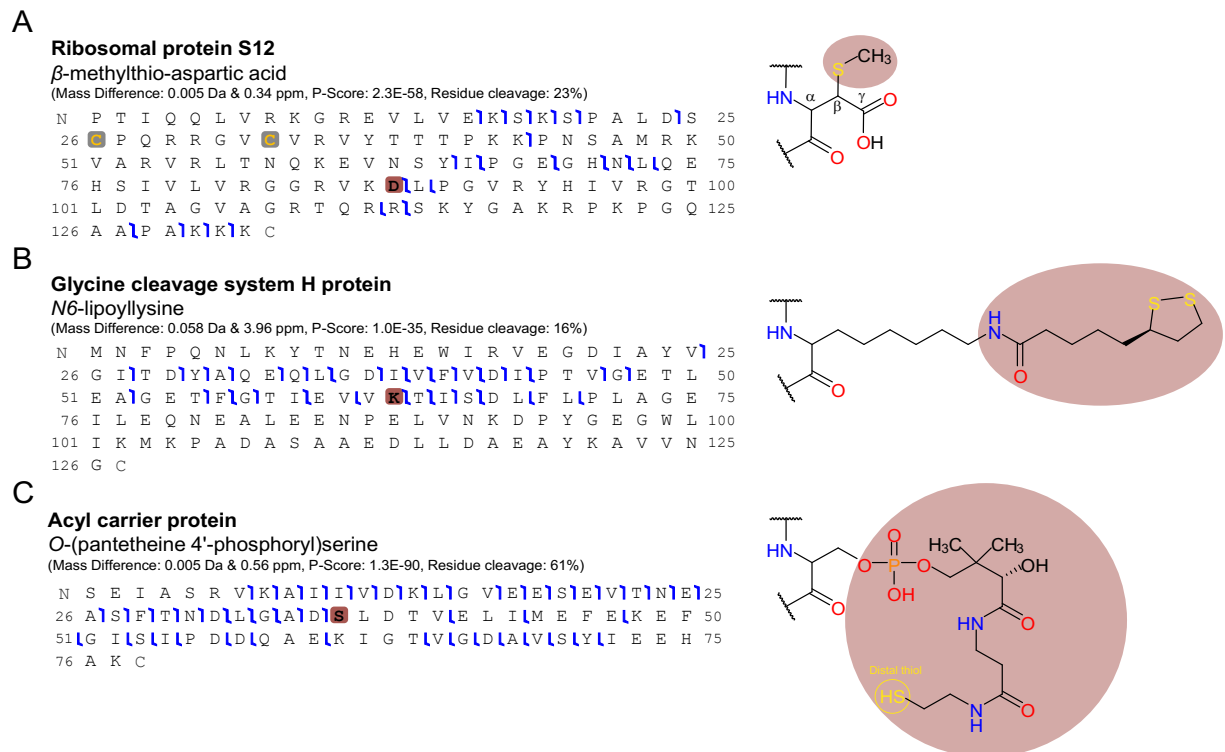
The detected mass shifts included methionine oxidation (+15.994 Da), cysteine dioxidation (+31.988 Da), and combinations of multiple oxidations (FIGURE III-14). These modifications can occur spontaneously during sample preparation and storage (Kaulich et al., 2022b). Additionally, certain mass shifts indicate the absence of specific amino acid residues, such as cysteinyl (+103.009 Da), phenylalaninyl (+147.068 Da), and tyrosinyl (+163.063 Da). Frequent misassignments were often identified as either incorrect annotation (-131.040 Da) or absence (+131.040 Da) of N-terminal initiator methionine residues. Incorrect mass shifts, arising from multiple PTMs or incorrectly assigned modifications, such as the absence of initiator methionine and misassignments of acetylation (-42.010 Da), contributed to prevalent mass shifts of +89.03 Da (FIGURE III-14 and FIGURE A-13A).



**FIGURE III-14 | Distribution of Precursor Delta Mass Shifts.** Arrows highlight potential PTMs with matched monoisotopic masses according to the PSI-MOD (Montecchi-Palazzi et al., 2008) or UniMod (Creasy and Cottrell, 2004) database.

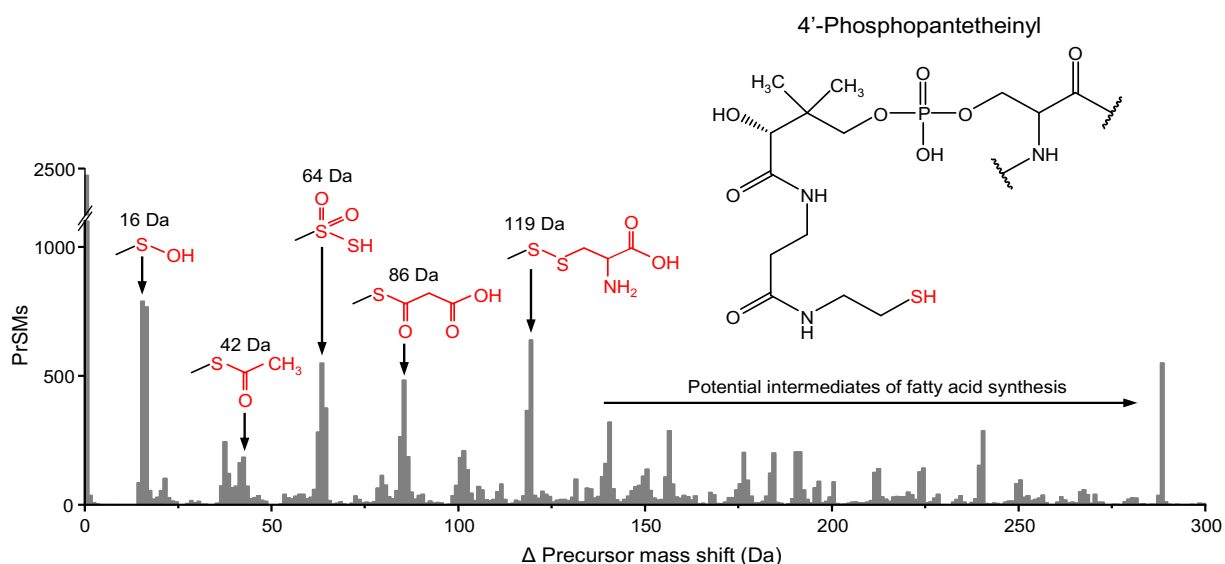
The formation of disulfide bonds in cysteine-containing proteoforms can be indicated by small mass shifts, such as  $-2$  Da and  $-4$  Da. Several potential disulfide bonds have been identified in proteins such as 50S ribosomal protein L32 (Q8A138), which displays a zinc finger motif (Cys-Xaa<sub>2</sub>-Cys-Xaa<sub>9</sub>-Cys-Xaa<sub>2</sub>-Cys) (FIGURE A-13B), and peptidyl-prolyl cis-trans isomerase (Q8A607), featuring an eight-cysteine motif (8CM) (FIGURE A-13C). Although 8CM motifs are present in a large number of fungal extracellular membrane proteins (Kulkarni et al., 2003) and plant defensins (José-Estanyol et al., 2004), their role in bacterial proteins is unclear.

A mass shift of 45.987 Da was detected on ribosomal protein S12 (Q8A472), suggesting a potential  $\beta$ -methylthio-aspartic acid modification (FIGURE III-15A). This PTM has been reported in other bacterial ribosomal protein S12 proteins (Kowalak and Walsh, 1996; Strader et al., 2004, 2011). Furthermore, the potential modification was detected in a BUP search, on the same aspartic acid within a peptide spanning the same sequence (FIGURE A-13D). Proteoforms of the glycine cleavage system H protein (Q8A4S8) exhibited a mass shift of 188.033 Da, potentially indicating an *N*6-lipoyllysine modification (FIGURE III-15B). This PTM is essential for transferring a methylamine group from the P protein to the T protein in the glycine cleavage system, generating CO<sub>2</sub>, NH<sub>3</sub>, and *N*5,*N*10-methylene-tetrahydrofolate (THF) (McCarthy and Booker, 2020).



**FIGURE III-15 | Putative PTMs in *B. thetaiotaomicron* proteins. (A)  $\beta$ -methylthio-aspartic acid on ribosomal protein S12 (Q8A472). (B) *N*6-lipoyllysine on glycine cleavage system H protein (Q8A4S8). (C) O-(pantetheine 4'-phosphoryl)serine on acyl carrier protein (Q8A2E6).**

The largest and most abundant potential PTM discovered was O-(pantetheine 4'-phosphoryl)serine (340,085 Da) on the acyl carrier protein (ACP) (Q8A2E6) (FIGURE III-15C). The prosthetic moiety, 4'-phosphopantetheinyl (Ppant), has a highly reactive distal thiol group that facilitates the covalent transport of various chemical groups, including acyl groups, during fatty acid biosynthesis (Chan and Vogel, 2010). Interestingly, ACP proteoforms exhibiting the Ppant modification showed a wide range of mass shifts that may be caused by different PTMs (FIGURE III-16). Some of these mass shifts may be attributed to common acyl intermediates during fatty acid biosynthesis, such as acetyl (+42,010 Da) and malonyl (+86,000 Da) (Chan and Vogel, 2010). Since fatty acid biosynthesis involves multiple steps to produce full-length chains (typically C<sub>16</sub> or C<sub>18</sub>) (Cronan and Thomas, 2009), the presence of larger mass shifts may indicate longer chain-length intermediates. Furthermore, several mass shifts may indicate thiol-related modifications, such as cysteinylolation (119.004 Da) and sulfur dioxide (SO<sub>2</sub>) addition (+63.961 Da) (FIGURE III-16).



**FIGURE III-16 | Distribution of Precursor Mass Shifts on Acyl Carrier Protein (Q8A2E6).** All indicated mass shifts additionally exhibit the O-(pantetheine 4'-phosphoryl)serine mass shift of 340.085 Da. Arrows highlight potential PTMs with matched monoisotopic masses according to PSI-MOD (Montecchi-Palazzi et al., 2008) or UniMod (Creasy and Cottrell, 2004) databases.

## 4 Discussion and Conclusion

### 4.1 Future Direction for Quantitative Analysis

A comprehensive comparison was performed between two LC-MS-based quantitative proteomics methods using LFQ and isobaric TMT labeling. Both experiments utilized intracellular proteome samples from *B. thetaiotaomicron*, and similar sample processing procedures were employed up to the protein digestion step. The primary objective was to evaluate the performance in terms of precision, data completeness, and quantification accuracy.

The stochastic nature of precursor ion selection in LFQ-DDA presents challenges, potentially leading to missing values that reduce data completeness and statistical reliability (Aittokallio, 2010; Michalski et al., 2011). In contrast, the analysis of multiplexed TMT can reduce the occurrence of missing values as the majority of peptides provide abundance values for each channel during fragmentation. However, co-isolation of peptides can lead to significant MS<sup>2</sup> interference, thus compromising the accuracy of TMT quantification (Sandberg et al., 2014). The two-dimensional separation scheme, combined with the chosen fractionation and pooling approach, reduced peptide co-isolation below 30% by 4.6% (FIGURE III-4B). Although the improvements are modest, they can still enhance the statistical significance and accuracy of reporter ion quantification (Sandberg et al., 2014). Several studies have demonstrated the importance of employing a fractionation strategy and two-dimensional separation prior to MS analysis in quantitative proteomics, especially with isobaric labeling methods (Zhang et al., 2010; Latosinska et al., 2015; Treitz et al., 2016). This approach generally increases protein identification and proteome coverage, outperforming unfractionated acquisition. For example, employing two-dimensional separation with bRP for pre-fractionation significantly increased the identification of bladder cancer proteins from 322 to 1092 using iTRAQ multiple-peptide identifications (Latosinska et al., 2015). Similarly, the authors reported similar numbers using an LFQ approach (910 proteins identified) (Latosinska et al., 2015). Additionally, an extra peptide fractionation step prior to MS analysis, as demonstrated by Patel and colleagues, revealed a comparable number of proteins identified by both iTRAQ and LFQ methods in the proteomic analysis of *Methylocella silvestris* (384 and 425 proteins, respectively) (Patel et al., 2009). Equivalent identification metrics were also reported for two-proteome samples, which included yeast and human cells (SH-SY5Y) (O'Connell et al., 2018), and three-proteome samples comprising human cells (OV3), *E. coli*, and soy proteome (Taverna and Gaspari, 2021).

Similarly, the proteomic analysis performed in this chapter of *B. thetaiotaomicron* in the presence of glucose or sucrose, employing both label-free and TMT-based quantification, identified a similar number of proteins (1735 and 1731, respectively). While the applied isobaric

labeling approach exhibited significantly higher reproducibility compared to LFQ (standard deviation of 0.076 and 0.135, respectively) (FIGURE III-6B), both approaches demonstrated analogous quantitative metrics, including the number of peptides supporting identification and quantification (FIGURE III-5), sequence coverage (FIGURE III-6C), and proteome coverage (FIGURE III-5). The integrative analysis of both proteomic approaches consistently provided similar results with a high Pearson correlation between LFQ and TMT (FIGURE III-7A).

In conclusion, both LFQ-MS<sup>1</sup> and TMT-MS<sup>2</sup> proteomic analyses identified a similar number of differentially abundant proteins (37 for LFQ and 32 for TMT; FIGURE III-5), indicating that these quantification approaches are effective and equivalent in capturing the proteomic response to glucose or sucrose as carbon sources. Overall, the findings suggest that *B. thetaiotaomicron* generally maintains its intracellular proteome composition in response to this dietary shift, with changes primarily observed in proteins involved in utilizing the provided carbon sources. This finding aligns with previous research on non-HGM members like *Lactobacillus sakei*, which also exhibited limited proteomic changes during the transition from glucose to sucrose (Precht et al., 2018). Although additional findings included the presence of proteins from PUL90, PUL91, PUL71, and PUL72, known for targeting glycosidic linkages in mucin O-glycans (Martens et al., 2008) and N-glycans (Cuskin et al., 2015), this occurrence may be attributed to the utilization of yeast extract in the culture medium.

The utilization of distinct MS settings, involving different fragmentation energies for precursor ions and isolation windows, was a consequence of meeting the specific requirements of each quantification method. The decision to employ four technical replicates in LFQ, as opposed to three in TMT, was aimed at mitigating potential variability or outliers, enhancing statistical robustness, and was influenced by the availability of free instrument time. Consequently, a run time of 60 hours was required for LFQ measurements, encompassing 3 biological replicates, each measured in 4 technical replicates, totaling 24 sets of 150-minute segments for the overall MS method. In contrast, TMT-based measurements resulted in a run time of 52.5 hours, encompassing 3 biological replicates, each measured in 3 technical replicates, totaling 21 sets of 150-minute segments for the total MS method.

Based on comparable quantitative results from both methods, the cost-effectiveness and reduced sample preparation steps in LFQ make it a practical option for future studies. Consequently, the LFQ approach will be exclusively employed in upcoming studies that utilize proteomics to study human gut bacteria. To improve reproducibility, future quantitative studies will analyze five bacterial cultures instead of three, while reducing technical replicates from four to three aims to reduce instrument time without compromising data reliability.

## 4.2 Proteoform-Directed Analysis of *B. thetaiotaomicron*'s Proteome

The proteoform analysis of *B. thetaiotaomicron* employed three distinct sample preparation methods, with each method contributing to the identification of approximately 25% of the total proteoforms, thereby enhancing the depth of analysis (FIGURE III-10G). Notably, only 9% (288 proteoforms) were identified by all methods (FIGURE III-10G), indicating that the methods complement each other in capturing proteoform diversity. While the methods differed in the number of protein groups and proteoforms identified (FIGURE III-10A-B), their analytical performance, measured by mean PrSMs per proteoform, residue cleavage, and E-value, was comparable (FIGURE III-10C-F). Together, these methods identified 3,117 proteoforms from 865 protein groups (FIGURE III-10A-B).

Interestingly, the majority (65%) of the canonical proteoforms and proteins (63%) were identified with the initiator methionine retained (FIGURE III-11B and FIGURE III-12D). In the bacterial maturation process of the N-terminus, removal of the N-formyl group by peptide deformylase precedes further modifications, such as NME or N-terminal acetylation (Solbiati et al., 1999; Bienvenut et al., 2015). Given the high energy cost of methionine synthesis, the recycling of initiator methionine is essential for energy conservation (Old et al., 1991).

Methionine aminopeptidase (MAP) plays an essential role by catalyzing the removal of the initiator methionine, ensuring the biological activity and stability of several proteins (Liao et al., 2004). This process is well-conserved across living organisms and is essential for cell function (Wingfield, 2017). Notably, *B. thetaiotaomicron* possesses two MAP genes, and both proteins (Q8AA27 and Q8A497) were identified through BUP analysis.

The observed absence of N-terminal methionine, especially those followed by Ala, Gly, Ser, and Pro at P1' (FIGURE III-12B), aligns with typical MAP kinetics observed in prokaryotic systems like *E. coli* (Frottin et al., 2006; Bienvenut et al., 2015), *Thermus thermophilus* (Suh et al., 2005), and *Shewanella oneidensis* (Gupta et al., 2007). This may indicate comparable substrate cleavage by *B. thetaiotaomicron* MAPs. Despite the general bias against bulky amino acids, two cleavages between Met-Ile and Met-Lys were detected (FIGURE III-12B). Similar exceptions in cleavage patterns at the P1' position have been observed across various bacteria (Bonissone et al., 2013). Given the high frequency of lysine at P1' (30.8%, FIGURE III-12C) and the fact that lysine residues disfavor NME (Frottin et al., 2006), it's likely that these cleavage events are not linked to MAP activity. Further investigations, such as utilizing N-terminal labeling strategies, are necessary to conclusively determine the origin of these proteoforms (Winkels et al., 2022).

The generation of artificial proteoforms can compromise the reliability and biological significance of proteomic analyses. Additionally, it can increase sample complexity, potentially leading to the co-elution of proteoforms with overlapping *m/z* ranges, which can interfere with accurate resolution by spectral deconvolution (Basharat et al., 2023). Therefore, it is crucial to

identify proteoforms originating from non-biological processes. The analysis of proteoform neo-termini of the applied sample methods revealed a normal distribution of most cleavage events (FIGURE III-13), indicating that the majority of proteoform formations are likely generated by endoproteases rather than exopeptidases. Notably, a Asp-Pro sequence logo was consistent in the HMWP acidic depletion, GELFrEE, and SCX fractionation methods, whereas it was absent under mildly basic conditions during HMWP basic depletion (FIGURE III-13). The generation of proteoforms with either aspartate at P1 or proline at P1' may be attributed to artificial cleavage, as acidic conditions and elevated temperatures have been shown to enhance artificial proteolysis of the Asp-Pro peptide bond (Winkels et al., 2022; Kaulich et al., 2024). Further details regarding this particular peptide bond are discussed in chapter IV.4.4.

The absence of the Asp-Pro sequence logo during HMWP depletion under basic conditions is accompanied by a relative increase in cleavage at Asn-Gly and Asn-Ser sites (FIGURE III-13A). These motifs could also be associated with artificial cleavage resulting from asparagine deamidation, a process that is influenced by various factors, including the primary sequences, higher-order structures of proteins, pH, temperature, and solution components (Tyler-Cross and Schirch, 1991). Interestingly, carbonate anions ( $\text{HCO}_3^-$ ), present in the buffer used for HMWP depletion under basic conditions (TEAB, triethylammonium bicarbonate), have been proposed as catalyst ions, mediating the cleavage of Asn-Gly peptide bonds (Kato et al., 2021). The absence of Asn-Gly and Asn-Ser sequence logos in the acidic HMWP depletion may result from a slower deamidation rate under mildly acidic conditions (Liu et al., 2016).

Through the discovery-based open modification search, several proteoforms were identified, exhibiting mass shifts associated with the absence of specific amino acid residues, such as initiator methionine or other N- or C-terminal amino acids encoded in the protein sequence (FIGURE III-14). While not all mass shifts and their corresponding protein sequences have undergone manual verification, potential errors in proteoform annotation or biological factors may contribute to protein sequence alterations. For instance, enzymes like leucyl/phenylalanyl-tRNA-protein transferase can facilitate the non-ribosomal transfer of Leu, Phe, and Met to N-terminal Arg or Lys residues (Dougan et al., 2010). Similarly, enzymes such as bacterial protein transferase from *Vibrio vulnificus* and peptide-amino acyl tRNA ligase from *Pseudomonas syringae* can catalyze the transfer of Arg to acidic N-terminal acceptor residues (Asp and Asn) or Cys to the C-terminus, respectively (Graciet et al., 2006; Zhang and van der Donk, 2019). The post-ribosomal addition of specific amino acids plays an essential role in forming N-terminal degradation signals (N-degrons), marking proteins for proteolytic degradation (Varshavsky, 2011). However, the addition of specific amino acids results in proteoform sequences deviating unpredictably from the genome sequence. While *B. thetaiotaomicron* lacks the genetic information for these enzymes, several other members of the *Bacteroides* genus encode multiple leucyl/phenylalanyl-tRNA-protein transferases.

Further, the proteomic data suggests that proteins in *B. thetaiotaomicron* have been post-translationally modified. Examples include the presence of  $\beta$ -methylthio-aspartic acid on ribosomal protein S12 (Q8A472), N6-lipoyllysine on glycine cleavage system H protein (Q8A4S8), and the occurrence of O-(pantetheine 4'-phosphoryl)serine on ACP (Q8A2E6) (FIGURE III-15). Specifically, the detection of  $\beta$ -methylthio-aspartic acid on Asp89, detected through both top-down and bottom-up proteomic analyses, suggests that *B. thetaiotaomicron* may perform this modification in the presence of glucose. This modification is strictly conserved in all bacterial S12 homologs (Carr et al., 2006) and has been previously observed in proteomic analyses of *E. coli* (Kowalak and Walsh, 1996), *Rhodopseudomonas palustris* (Strader et al., 2004) and *Thermus thermophilus* (Suh et al., 2005). While the functional significance of this modification remains largely unknown, initial indications suggest a potential role in regulating the translation of certain mRNAs (Strader et al., 2011).

Furthermore, several potential thiol-dependent modifications, such as cysteinylolation, which protects thiols from oxidation (Hochgräfe et al., 2007), and sulfur dioxide (SO<sub>2</sub>) addition (Jeong et al., 2011) on ACP-Ppant proteoforms were identified (FIGURE III-16). Thiol modifications are essential for regulating protein functions (Jones and Go, 2011). While typically observed in cysteine-containing proteins, Wang and colleagues identified several thiol oxidative and phosphorylated proteoforms of ACP-Ppant in *E. coli*, which may indicate a broader role for thiol-dependent modifications (Wang et al., 2020).

In conclusion, the identification and curation of mass shifts using established PTM databases reveal numerous PTM-carrying proteoforms and potential sample preparation artifacts. This underlines the significance of conducting comprehensive and unrestricted PTM searches in the field of proteomics.



## IV INFLUENCE OF pH ON BACTERIAL PROTEOMES

---

<b>1</b>	<b>Introduction and Summary .....</b>	<b>64</b>
<b>2</b>	<b>Experimental Design .....</b>	<b>66</b>
	2.1 Bottom-up LFQ Analysis of HGM Proteomes.....	66
	2.2 Top-down LFQ Analysis of <i>B. producta</i> .....	66
<b>3</b>	<b>Results.....</b>	<b>68</b>
	3.1 Proteomic Adaptations to Acidic and Alkaline pH in <i>B. longum</i> .....	68
	3.2 Analysis of the Acidic Response of three HGM Members.....	71
	3.3 Comparison of LMWP-Top-Down and Full-Proteome Bottom-Up Analysis for <i>B. producta</i> .....	79
	3.4 Potential pH-induced Asp-Pro Cleavage .....	84
	3.5 Phosphorylation of HPr proteins.....	87
<b>4</b>	<b>Discussion and Conclusion.....</b>	<b>89</b>
	4.1 Quantitative analysis of the <i>B. longum</i> proteome.....	89
	4.2 Quantitative analysis of the <i>B. thetaiotaomicron</i> proteome .....	92
	4.3 Quantitative analysis of the <i>B. producta</i> proteome.....	93
	4.4 Asp-Pro peptide bond hydrolysis.....	98

## 1 Introduction and Summary

The human gut microbiome (HGM), comprising diverse microbial communities, plays a crucial role in maintaining health. Its composition and balance are influenced by various factors, including growth factors, micronutrients, antimicrobial compounds, and gut pH (Rodionov et al., 2019; Beam et al., 2021; Firrman et al., 2022). While some variability in the pH of the gastrointestinal tract is normal due to factors like diet and microbial activity, significant deviations from typical ranges can alter the microbiome's composition and function, potentially impacting health (Firrman et al., 2022). Generally, the proximal small bowel has lower pH levels (pH 5.5 to pH 7.0) compared to the descending and rectosigmoid colon, which maintains slightly higher pH levels (pH 6.6 to pH 7.5) (Nugent et al., 2001). Notably, increased colonic acidity has been associated with various gastrointestinal diseases, including irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD) (Nugent et al., 2001; Ringel-Kulka et al., 2015).

Microbial responses to varying pH conditions involve physiological and molecular adaptations aimed at maintaining intracellular pH ( $\text{pH}_i$ ) homeostasis. These adaptations can include proton translocation by specialized pumps such as ATP synthase and small ion ( $\text{Na}^+$ ,  $\text{K}^+$ , or  $\text{Ca}^{2+}$ )/ $\text{H}^+$  antiporters (Krulwich et al., 2011). Additionally, enzyme-catalyzed reactions, including processes such as decarboxylation, consume protons, whereas deaminase-catalyzed reactions increase the concentration of alkaline compounds (Krulwich et al., 2011). Protective mechanisms also include changes in lipid composition to reduce proton permeability of the cell, promotion of biofilm formation, adjustment of cell density, and implementation of repair mechanisms to counteract increased damage to macromolecules (Guan and Liu, 2020).

Despite extensive investigations into microbial acid stress responses, knowledge regarding the proteomic adaptations of specific HGM members to acidic stress remains limited. Understanding these proteomic changes could provide valuable insights into how these microbes maintain resilience and adapt to acidic environments. Addressing this research gap could advance understanding of gut microbial dynamics and their potential implications for human health, as well as facilitate the development of targeted therapeutic strategies for gastrointestinal diseases.

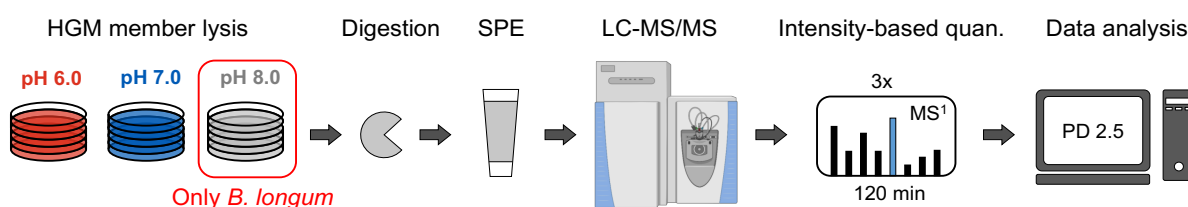
**Aim of this study:**

- To identify proteomic alterations in response to diverse pH conditions (pH 6.0, pH 7.0, and pH 8.0), label-free quantification analyses using bottom-up proteomics were conducted on three HGM species: *Bacteroides thetaiotaomicron*, *Blautia producta*, and *Bifidobacterium longum*.
- Compare protein abundances to identify proteins exhibiting co-abundance in both acidic and alkaline responses, as well as those exhibiting specific abundance under acidic or alkaline pH conditions relative to growth at pH 7.0.
- Identify concurrent and distinct proteomic alterations in pathway-related or stress-associated proteins among the three bacterial species.
- Utilize a top-down proteomics label-free quantification analysis to quantify proteoforms.
- Perform a comparative analysis of bottom-up and top-down quantitative results to evaluate their capacity to quantify the same proteins, whether differentially abundant or not.
- Conduct a discovery-based open modification search to identify potential post-translational modifications.

## 2 Experimental Design

### 2.1 Bottom-up LFQ Analysis of HGM Proteomes

To analyze the proteomic response to cultivation at different pH values in *B. thetaiotaomicron*, *B. producta*, and *B. longum*, the bacteria were cultured in YCFA medium at acidic (pH 6.0) and neutral (pH 7.0) conditions in five biological replicates. *B. longum* was additionally cultivated under alkaline (pH 8.0) conditions (FIGURE IV-1, chapter II.2.1). At mid-stationary phase, cells were harvested and lysed using freeze-thawing and proteins were cleaned up using ethanol precipitation (chapter II.3.1). Intracellular proteomes were isolated, digested using trypsin with a 1:40 enzyme-to-substrate ratio and subjected to solid-phase extraction (chapter II.3.4). All samples were separated online by reversed-phase chromatography with a gradient of 120 minutes and measured in triplicate on the Q-Exactive Plus mass spectrometer (chapter II.4.1). The acquired raw data were searched against the respective reference proteomes using PD 2.5 (chapter II.5.1). Median normalization was used to center the data distribution and compensate for general concentration differences that could be caused by minor variations in sample concentration (chapter II. 5.4). The raw and normalized intensity data were tested for normal distribution and Pearson correlation (chapter II. 5.4). Statistical analysis (two-sided Welch's t-test with Benjamini-Hochberg FDR correction for multiple testing,  $q \leq 0.05$ ) was performed on high-confidence protein identifications (1% FDR) with at least three quantitative values out of the five biological replicates.

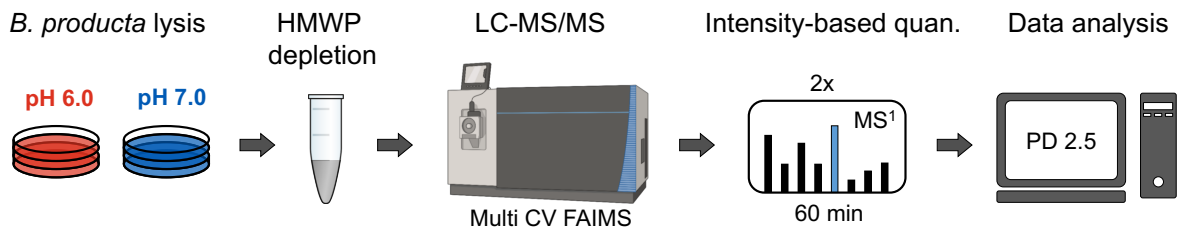


**FIGURE IV-1 | Experimental Design to Analyze the pH-Induced Proteomic Response of HGM Bacteria.** Bottom-up proteomic LFQ analysis of *B. thetaiotaomicron*, *B. producta*, and *B. longum* cultivated in YCFA medium at pH 6.0 and pH 7.0, with additional cultivation of *B. longum* at pH 8.0.

### 2.2 Top-down LFQ Analysis of *B. producta*

To perform label-free quantification of proteoforms in *B. producta*, three of the five biological replicates grown in YCFA medium under different pH conditions (pH 6.0 and pH 7.0) were processed using acidic and basic depletion of the high molecular-weight proteome (FIGURE IV-2) (Cassidy et al., 2019). This approach aimed to enhance the identification of low-molecular-weight proteoforms. Details about the sample preparation are described in chapter II.3.2. Samples were separated by reversed-phase chromatography using a 60 min gradient and

analyzed in duplicates on the Fusion Lumos mass spectrometer, employing the multi-CV FAIMS method which four distinct CVs (-60, -50, -40 and -20, chapter II.4.2) (Kaulich et al., 2022a). The acquired raw data were filtered using Freestyle v.1.6 based on the applied CVs, dividing each LC-FAIMS-MS/MS analysis into four distinct raw files. Subsequently, these files were searched against the *B. producta* reference proteome using PD 2.5 (chapter II.5.1). Quantification was performed utilizing the high-resolution feature detector node with the sliding window deconvolution algorithm with an average retention time width of 0.33 min. All results were subjected to an FDR correction for both PrSMs and proteoforms, with a threshold of 1%. Identified proteoforms were required to have a minimum C-Score of 40 (Leduc et al., 2014). Top-down data processing steps included CV summation, median calculation, and median normalization. The raw and normalized intensity data were tested for normal distribution and Pearson correlation (chapter II. 5.4). Statistical analysis (two-sided Student's t-test with permutation-based FDR correction for multiple testing,  $q \leq 0.05$ ) required at least two quantitative values out of three biological replicates.

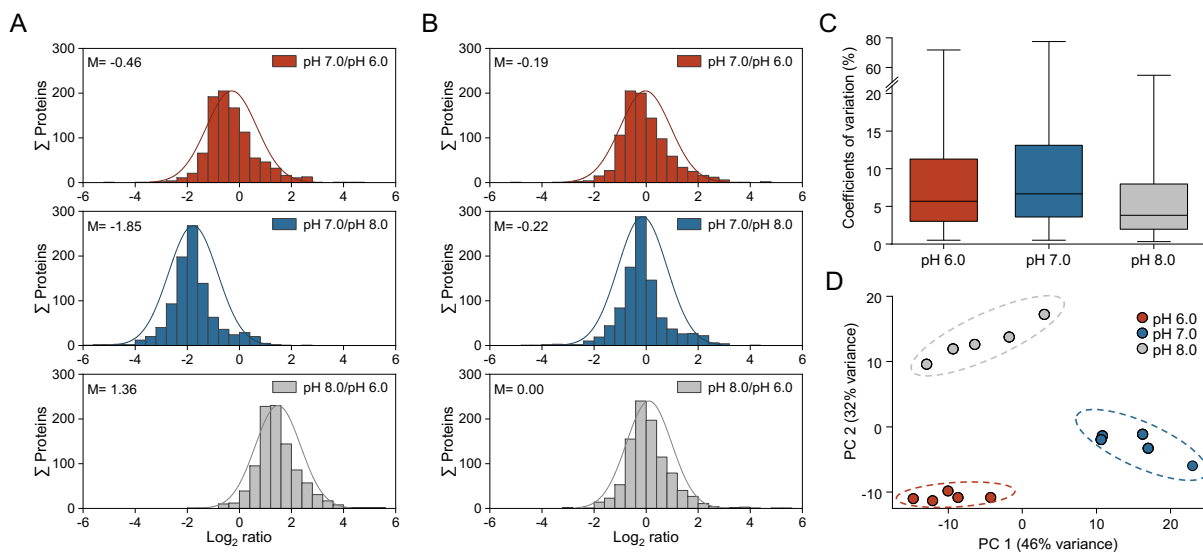


**FIGURE IV-2 | Experimental Design for Top-down LFQ Analysis of the *B. producta* Proteome.** *B. producta* cultivated in YCFA medium at pH 6.0 and pH 7.0 were subjected to acidic and basic HMWP depletion for top-down label-free quantification.

### 3 Results

#### 3.1 Proteomic Adaptations to Acidic and Alkaline pH in *B. longum*

**Data Processing** – Proteomic changes of *B. longum* after cultivation at pH 6.0, pH 7.0, and pH 8.0 were analyzed by BUP-LFQ analysis. A total of 991 proteins, covering 57% of the encoded *B. longum* proteome, were identified. Protein abundance profiles for both biological and technical replicates (FIGURE A-14), as well as culture pH comparisons (pH 7.0/pH 6.0, pH 7.0/pH 8.0, and pH 8.0/pH 6.0), were normalized to a median value of zero (FIGURE IV-3A-B). The datasets showed low variability, with median coefficients of variation ranging from 3.8% to 6.7% (FIGURE IV-3C). Strong concordance among biological and technical replicates is demonstrated by high average Pearson correlation coefficients of 0.95 (FIGURE A-14E). Principal-component analysis (PCA) revealed pH-dependent separation and clustering of biological replicates, with the first two principal components accounting for 78% of the variation (FIGURE IV-3D).



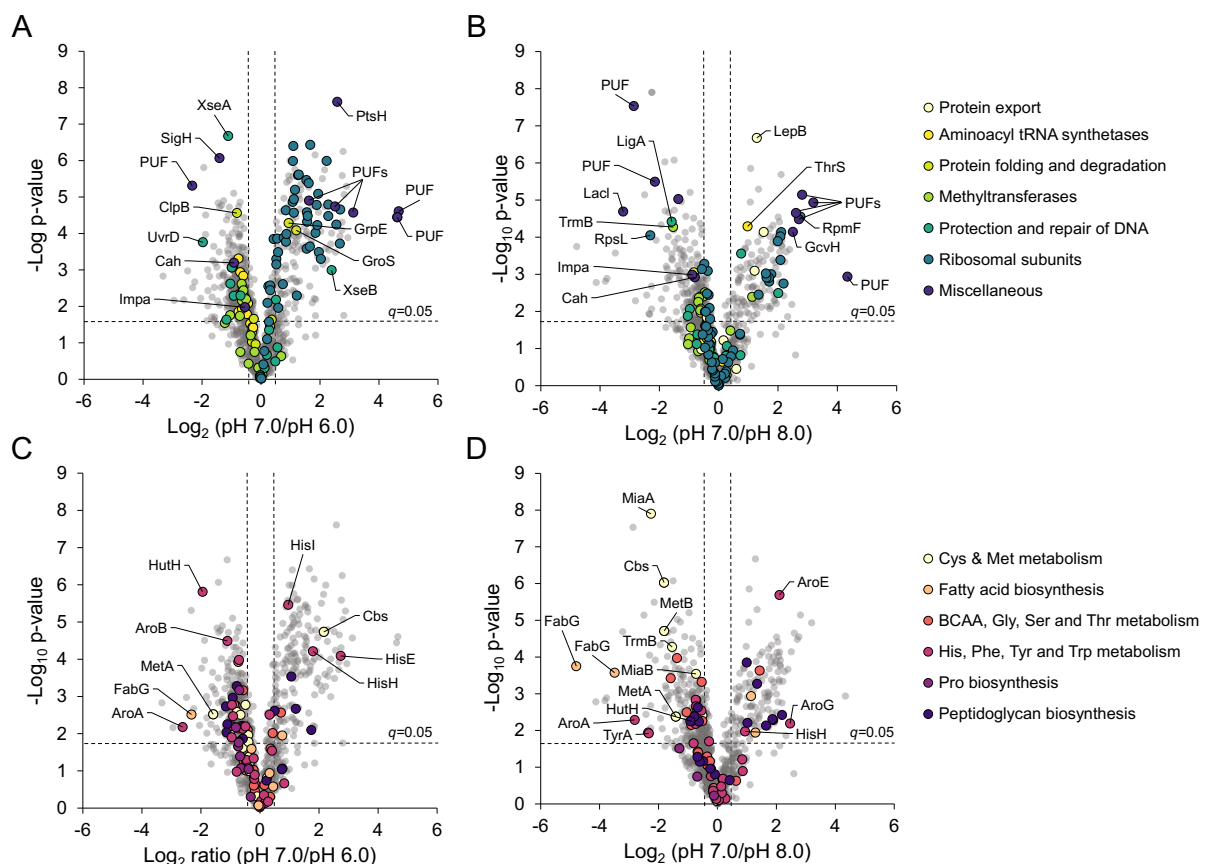
**FIGURE IV-3 | Evaluation of Data Processing of the *B. longum* Proteome.** Distribution of  $\text{Log}_2$  ratios of (A) raw data and (B) median normalized data. Each histogram is overlaid with a Gaussian distribution curve. The median (M) of the datasets is displayed in the upper left corner of each histogram. (C) Box-and-whisker plots of the coefficient of variation of the identified proteomes. Boxes capture the lower and upper quartiles with the median displayed as a horizontal line in the middle; whiskers represent the 1–99 percentile. (D) Principal component (PC) analysis with each circle represents a biological replicate cultivated at the indicated pH.

**Quantitative data** – This study aimed to examine the proteomic response of *B. longum* to acidic (pH 6.0) and alkaline (pH 8.0) culture conditions by comparing protein abundance ratios pH 7.0 vs. pH 6.0 (indicating the acidic response) and pH 7.0 vs. pH 8.0 (indicating the alkaline response). A total of 933 and 935 proteins were quantified for the acidic and alkaline responses, respectively. Subsequent statistical analysis, employing a two-sided Welch's t-test

with Benjamini-Hochberg FDR correction ( $FDR \leq 0.05$ ) and a  $\text{Log}_2$  fold change threshold of  $\pm 0.485$ , identified 444 differentially abundant proteins for the acidic response and 332 for the alkaline response. Changes in protein abundance were detected for proteins associated with cellular processes such as protein export, folding, degradation, DNA protection, and repair, as well as proteins of the translational machinery such as ribosomal subunits and aminoacyl-tRNA synthetases (FIGURE IV-4A-B). Additionally, several proteins involved in amino acid metabolism, fatty acid synthesis, and peptidoglycan biosynthesis were detected as differentially abundant (FIGURE IV-4C-D).

The acidic proteomic response will be later described and discussed (see chapter IV.3.2). Therefore, the following results focus on the alkaline response and functional classes of proteins that had differential abundances in both the acidic and alkaline response and those that had differential abundance at either acid or alkaline pH compared to growth at pH 7.0.

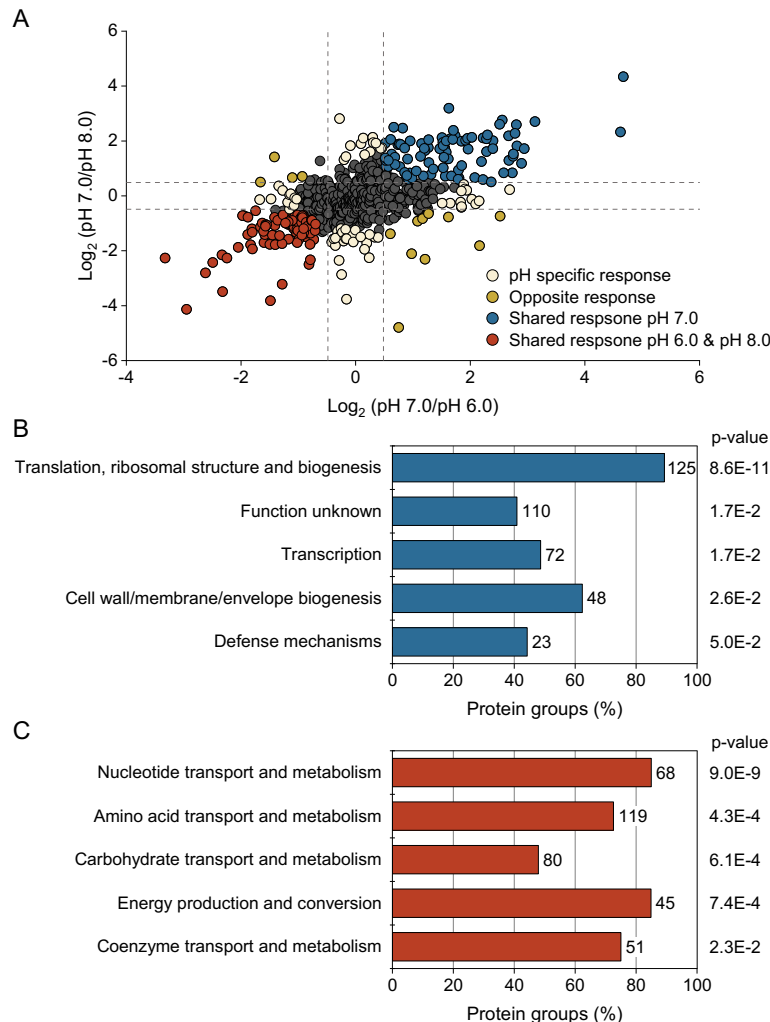
At pH 8.0, a higher abundance of proteins involved in methionine and cysteine biosynthesis or interconversion was observed, including homoserine O-acetyltransferase (MetA; Q8G7A5), cystathionine  $\gamma$ -synthase (MetB; Q8G565), cystathionine  $\beta$ -synthase (Cbs; Q8G564), and methionine synthase (MetE; Q8G651).



**FIGURE IV-4 | *B. longum* Acidic and Alkaline Proteomic Response.** (A-B) Volcano plots of the acidic (pH 7.0/pH 6.0) and (C-D) alkaline (pH 7.0/pH 8.0) response. Proteins are labeled by their respective gene names and are color-coded according to their involvement in cellular processes, with "PUF" denoting proteins of unknown function. Dashed vertical lines represent  $\text{Log}_2$  cutoffs, while the dashed horizontal line corresponds to a q-value of 0.05 (Two-sided Welch's t-test, corrected for multiple testing by Benjamini-Hochberg FDR calculation).

Additionally, two S-adenosyl-L-methionine-dependent enzymes, tRNA-methyltransferases (TrmB; Q8G3T4) and tRNA-methylthiotransferase (MiaB; Q8G4H4), were more abundant at pH 8.0. At pH 7.0, proteins of higher abundance included ribosomal subunits, proteins of unknown function (PUFs; Q8G7V3, Q8G449, Q8G6I8, and Q8G6N5), and histidine synthesis enzymes such as imidazole glycerol phosphate synthase (HisH; Q8G4S6), phosphoribosyl-ATP pyrophosphatase (HisE; Q8G694), and phosphoribosyl-AMP cyclohydrolase (HisI; Q8G6F6) (FIGURE IV-4). Conversely, histidine ammonia-lyase (HutH), which is involved in histidine degradation, and several aminoacyl-tRNA synthetases exhibited higher abundance at both pH 6.0 and pH 8.0 (FIGURE IV-4).

By directional analysis (Yang et al., 2014) of the 932 shared quantified proteins between the acidic and alkaline responses significant variations in protein abundance and pathway changes across different pH environments could be identified. Differential changes in protein abundance ( $p \leq 0.05$ ) were categorized into four groups (FIGURE IV-5A).

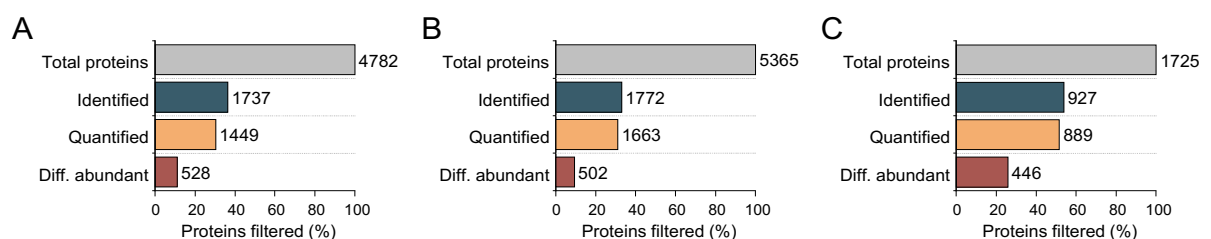


**FIGURE IV-5 | Comparison of the Acidic and Alkaline Response of the *B. longum* Proteome.** (A) Directional analysis with differentially abundant proteins colored based on their change in abundance ( $p \leq 0.05$ ), using a modified Pearson's correlation test (Yang et al., 2014). (B) Directional pathway analysis on proteins with shared abundance at pH 7.0 or (C) shared abundance at pH 6.0 and pH 8.0 using COG annotations. The number and percentage of proteins and the corresponding p-values of each category are shown.

Firstly, 65 proteins (30%) had pH-specific responses, meaning they had differential changes in abundance only at a specific pH condition, either pH 6.0 or pH 8.0. Second, 14 proteins (6%) exhibited unspecific responses, increasing their abundance in one condition but decreasing in another. Examples of these proteins include cystathionine  $\beta$ -synthase (Cbs; Q8G564), nitrogen regulatory protein N-II (GlnB; Q8G738), and 30S ribosomal protein S12 (RpsL; P59162), which increased in abundance under acidic conditions and decreased under alkaline conditions. Third, 71 proteins (32%) maintained a consistent increase in abundance specifically at pH 7.0. Lastly, another 71 proteins (32%) exhibited a shared response at both pH 6.0 and pH 8.0, maintaining a consistent increase in abundance under both acidic and alkaline conditions. The Pearson correlation coefficient between the acidic and alkaline response of 0.59 suggests a moderately complementary proteomic response to acidic and alkaline conditions (FIGURE IV-5A). Directional pathway analysis on proteins with consistent increase in abundance specifically at pH 7.0 revealed significant enrichment ( $p < 0.01$ ) in translational and transcriptional processes, as well as cell wall biogenesis, defense mechanisms, and proteins of unknown function (FIGURE IV-5B). Proteins with a consistent increase in abundance at pH 6.0 and pH 8.0 were enriched in COG categories involved in the transport and metabolism of nucleotides, amino acids, carbohydrates, and coenzymes, as well as proteins associated with energy production and conversion (FIGURE IV-5C). These include 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase (AroA; Q8G5N6) (TABLE A-3), which is essential for catalyzing the formation of EPSP and inorganic phosphate from shikimate-3-phosphate and phosphoenolpyruvate in the shikimate pathway.

### 3.2 Analysis of the Acidic Response of three HGM Members

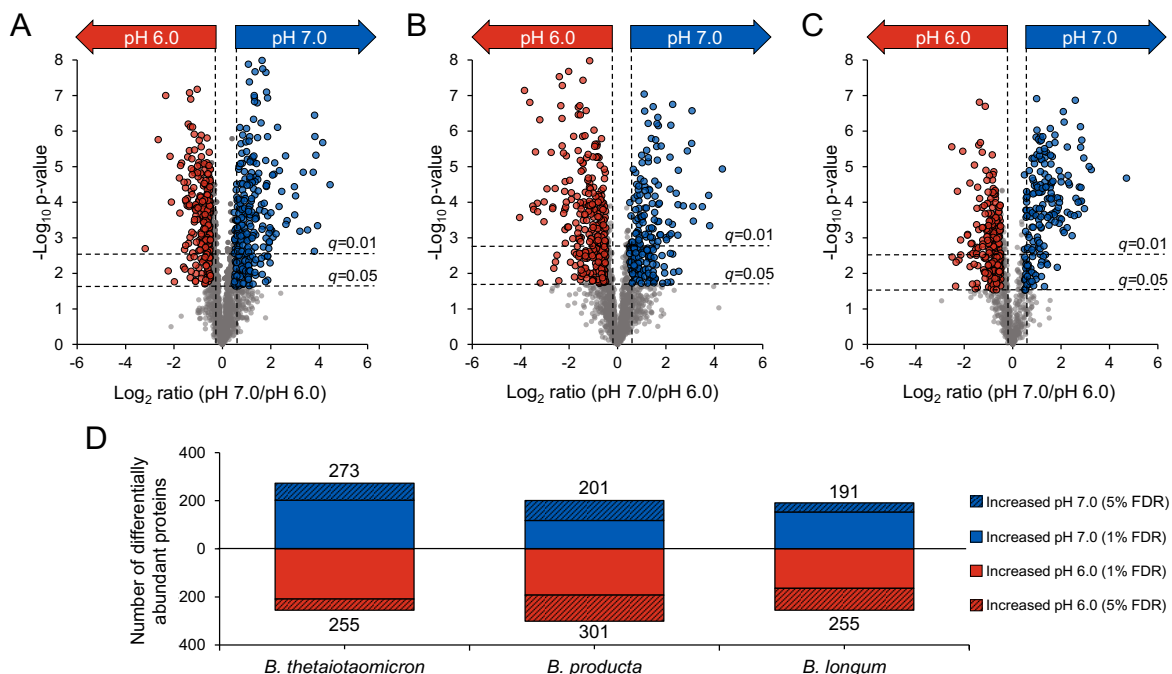
**Data Processing** – A total of 1,497, 1,772, and 927 proteins were identified in the *B. thetaiotaomicron*, *B. producta*, and *B. longum* datasets, respectively (FIGURE IV-6A-C). An overview of the datasets before and after median normalization and the Pearson correlation analysis is provided in the Appendix (FIGURE A-15). Proteins quantified in three out of five replicates of at least one culture condition accounted for 30% (1449 out of 4782) of all encoded proteins for *B. thetaiotaomicron*, 31% (1663 of 5365) for *B. producta*, and 52% (889 of 1725) for *B. longum* (FIGURE IV-6A-C).



**FIGURE IV-6 | Protein Identification Overview.** The total number of proteins encoded in the genome, identified proteins, identified proteins (detected in three of five replicates) and differentially abundant proteins for (A) *B. thetaiotaomicron*, (B) *B. producta* and (C) *B. longum*.

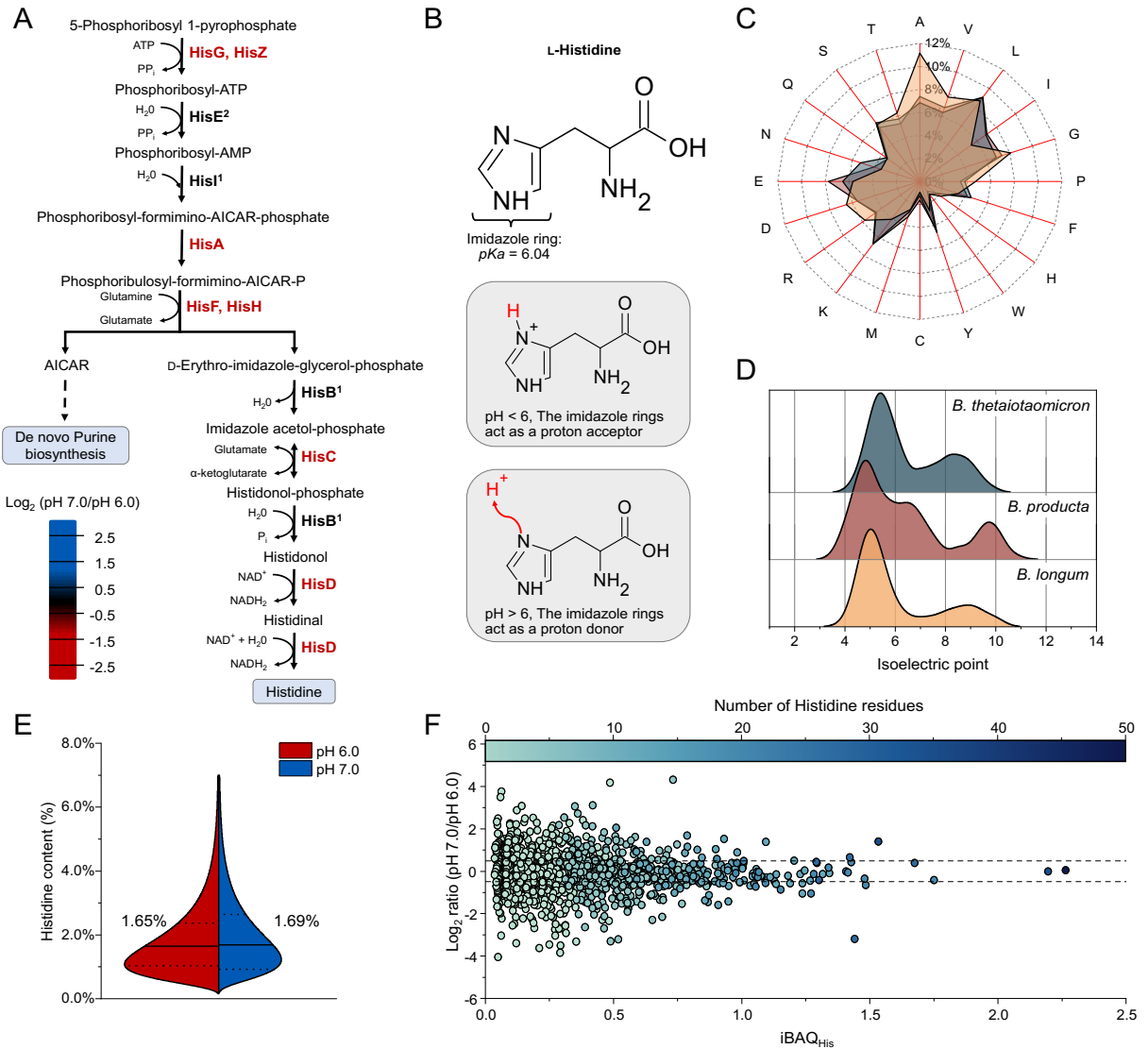
Statistical analysis of the acid pH response (pH 7.0 vs. pH 6.0) of the three bacteria identified 528, 501, and 446 proteins as differentially abundant for *B. thetaiotaomicron*, *B. producta*, and *B. longum*, respectively (FIGURE IV-6A-C).

**Quantitative Data** – Significant proteomic adaptations to acidic culture conditions were observed across all HGM bacteria (FIGURE IV-7A-C). Although a FDR of 5% was chosen as the threshold for statistical significance, the majority of proteins with differential abundance ( $\text{Log}_2$  fold change threshold of  $\pm 0.485$ ) were identified with an FDR of 1% (FIGURE IV-7D). In *B. thetaiotaomicron*, an equal number of differentially abundant proteins were detected with increased abundance at both pH 6.0 and pH 7.0 (FIGURE IV-7D). In contrast, *B. producta* and *B. longum* exhibited a higher number of proteins with increased abundance at pH 6.0 (FIGURE IV-7D). Most significant changes involved proteins involved in metabolic pathways such as carbohydrate utilization, amino acid biosynthesis and degradation, purine and pyrimidine biosynthesis, and peptidoglycan synthesis. Additionally, proteins involved in cellular respiration, transcriptional and translational processes, and the protection and repair of macromolecules exhibited variations in abundance. A summary of differentially abundant proteins and their respective fold changes in the three HGM bacteria is provided in TABLE A-4, and explained in detail in the following sections.



**FIGURE IV-7 | Quantitative Proteome Analysis of Human Gut Bacteria.** Volcano plot for all quantified proteins of (A) *B. thetaiotaomicron*, (B) *B. producta*, and (C) *B. longum* at pH 7.0 vs. pH 6.0. The dashed vertical lines represent  $\text{Log}_2$  ratio thresholds and the dashed horizontal lines represent a q-value of 0.05 or 0.01 (Two-sided Welch's t-test, Benjamini-Hochberg FDR corrected). (D) The total number of differentially abundant proteins quantified with increased abundance at pH 7.0 and pH 6.0 with a q-value of 0.05 or 0.01.

**Amino Acid Metabolism** – All three bacteria showed an increased abundance of proteins involved in the biosynthesis of arginine, isoleucine, and lysine at pH 6.0 (TABLE A-4). While the  $pH_i$  can be maintained by the production of  $NH_3$  and  $CO_2$  through deamidation and decarboxylation reactions (Krulwich et al., 2011), no differential changes in the abundance of amino acid deaminases or decarboxylases were detected in the three HGM bacteria. In *B. producta*, proteins involved in histidine biosynthesis were differentially abundant at pH 6.0 compared to pH 7.0 (FIGURE IV-8A).



**FIGURE IV-8 | Potential Role of Histidine Biosynthesis in *B. producta*  $pH_i$  Maintenance.** (A) Schematic representation of the histidine biosynthetic pathway. Enzymes, represented by their gene names, are colored based on their  $\text{Log}_2$  ratio (pH 7.0/pH 6.0). Abbreviations: AICAR = aminoimidazole carboxamide; Superscript: 1 = not identified; 2 = not quantified. (B) Chemical structure and buffer capacity of L-histidine. (C) Analysis of the amino acid composition of the three bacterial proteomes. (D) Isoelectric point distribution of three bacterial proteomes. (E) Distribution of differentially abundant histidine-containing proteins at pH 6.0 and pH 7.0 with highlighted median histidine amount. (F) Intensity-Based Absolute Quantification multiplied by the total number of histidine residues in each protein ( $iBAQ_{His}$ ) against the  $\text{Log}_2$  ratio (pH 7.0/pH 6.0). Color coding corresponds to the total number of histidine residues.

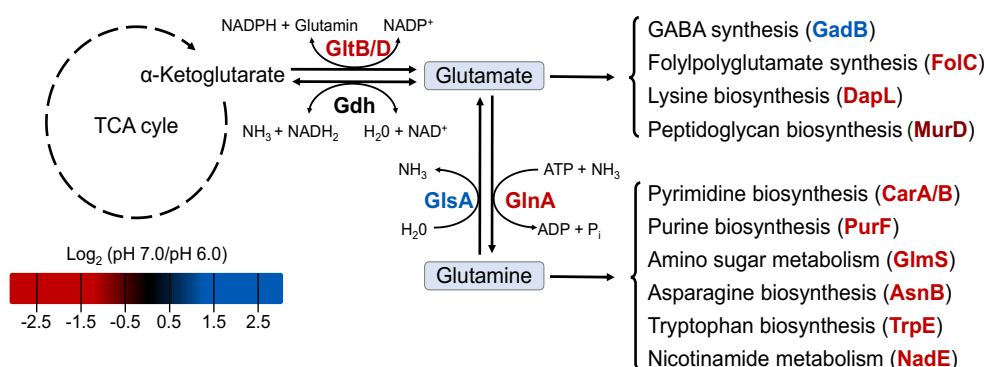
The enzymes catalyzing the initial step, ATP-phosphoribosyltransferase (HisG; A0A4P6LXR4) regulated by HisZ (A0A7G5MP39), and histidinol dehydrogenase (HisD; A0A7G5N0Q3) facilitating the final step, increased up to 11-fold at pH 6.0. This increase may indicate a role for histidine in maintaining pH<sub>i</sub> homeostasis under acidic conditions.

In addition to the carboxyl group ( $pK_a = 1.8$ ) and the amino group ( $pK_a = 9.2$ ) at the  $\alpha$ -carbon atom, histidine is characterized by its imidazole side chain ( $pK_a = 6.0$ ; FIGURE IV-8B). As the pH decreases, the imidazole ring of histidine can function as a proton acceptor, removing hydrogen ions from the solution and thereby increasing the pH. This property allows histidine to act as an effective buffer within approximately  $\pm 1$  pH units of its imidazole side chain's  $pK_a$  (FIGURE IV-8B). Nonetheless, the relatively low concentration of free histidine (68  $\mu$ M) in bacterial cells (Bennett et al., 2009) suggests it may not significantly contribute to overall cellular buffering capacity alone. Histidine residues within proteins can also act as proton acceptors and donors, thereby contributing to intracellular buffering. However, the effectiveness of these residues depends on their exposure to the protein surface to interact with the cytoplasmic environment and effectively buffer excess protons.

The frequency of histidine side chains in proteins was analyzed to assess the potential of protein-bound histidine residues to buffer excess protons entering the cell. With a median frequency of 1.7% histidine residues, *B. producta* showed no significant difference in histidine content compared to *B. thetaiotaomicron* (1.8%) and *B. longum* (2.2%) (FIGURE IV-8C). Additionally, similar bimodal isoelectric point (pI) distributions were observed across the three bacteria, with average pI values of 6.32 for *B. producta*, 6.49 for *B. thetaiotaomicron*, and 6.08 for *B. longum* (FIGURE IV-8D). This indicates that the frequency of protein-bound histidine residues and the overall protein charge distributions are similar among these bacterial species. Further examination of the distribution of differentially abundant histidine-containing proteins in *B. producta* at both pH 6.0 and pH 7.0 revealed a consistent median histidine content of 1.7% (FIGURE IV-8E). Additionally, multiplying intensity-based absolute quantification (iBAQ) values by the total histidine residues (iBAQ<sub>His</sub>) and comparing them to the total number of histidine residues for each protein, did not reveal a preference for a higher abundance of histidine-containing proteins under acidic conditions (FIGURE IV-8F).

Collectively, these results indicate that the abundance of histidine-containing proteins within the *B. producta* proteome remains relatively constant regardless of pH variations between pH 6.0 and pH 7.0. This stability suggests that free histidine, rather than histidine-containing proteins, may serve as a key component of an amino acid-dependent acid tolerance system in *B. producta*.

In *B. thetaiotaomicron*, key enzymes of nitrogen metabolism, which are crucial for the assimilation of nitrogen into cellular processes, such as amino acid or nitrogen-containing compound synthesis, were more abundant at pH 6.0 (FIGURE IV-9). These enzymes include proteins of the GS/GOGAT cycle: glutamine synthetase (GS, GlnA; Q8AAC2) and glutamate synthase (also known as glutamine:  $\alpha$ -oxoglutarate aminotransferase, GOGAT), consisting of a small subunit (GltB; Q8AAB2) and a large subunit (GltD; Q8AAB3). Although glutamate dehydrogenase (Gdh; Q8A6B2), which is crucial for ammonium assimilation or release, was more abundant at pH 6.0, this difference was not statistically significant (FIGURE IV-9). Furthermore, enzymes capable of utilizing glutamine or glutamate as nitrogen sources to catalyze the transfer of an amine to another acceptor molecule also exhibited higher abundance at pH 6.0 (FIGURE IV-9). These enzymes participate in the synthesis of various metabolic intermediates that are essential for cell growth and survival, such as purines, pyrimidines, and peptidoglycan precursors (FIGURE IV-9). Furthermore, proteins involved in  $\gamma$ -aminobutyric acid (GABA) synthesis involving glutaminase (GlsA; Q8A4M8) and glutamate decarboxylase (GadB; Q8A4M9) were more abundant at pH 7.0 (FIGURE IV-9).



**FIGURE IV-9 | Central Nitrogen Metabolism of *B. thetaiotaomicron*.** Glutamate and glutamine metabolism and their role as nitrogen donors for various nitrogen-containing metabolic intermediates. Enzymes, represented by their gene names, are colored based on their Log<sub>2</sub> ratio (pH 7.0/pH 6.0).

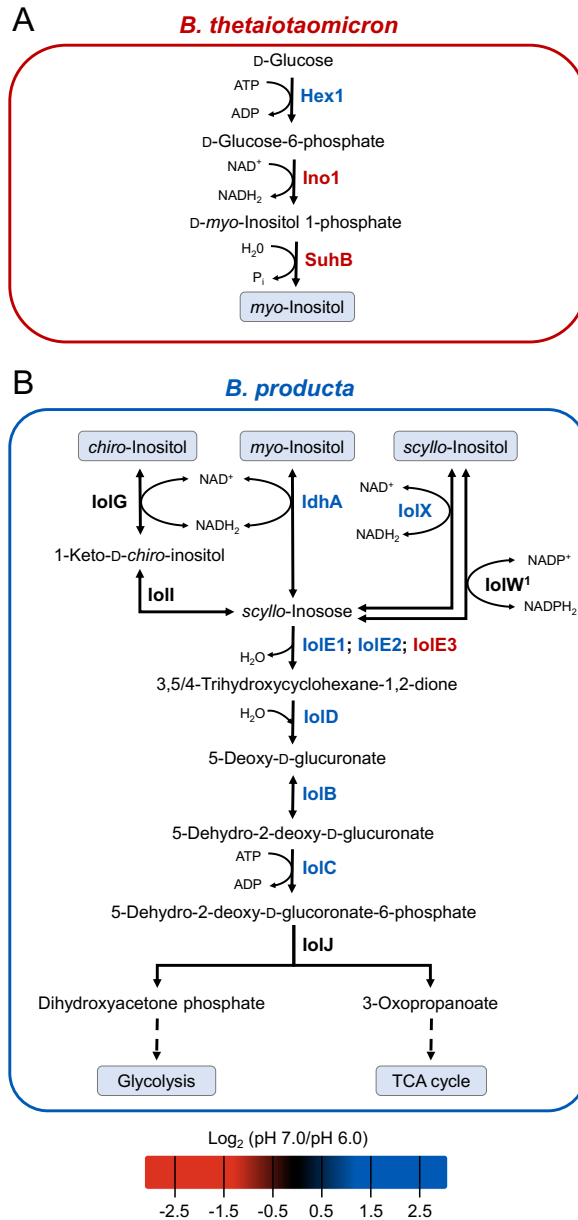
**Cellular Respiration** – Quantitative data were acquired for several key proteins of cellular respiration reactions, such as F- and V-type ATPases, protein electron donors, and acceptors responsible for ATP synthesis, as well as proteins involved in the recycling of reducing equivalents (e.g., NADH and FADH<sub>2</sub>). Specifically, the subunits of Na<sup>+</sup>-translocating NADH dehydrogenase (NQR) of *B. thetaiotaomicron* and the Na<sup>+</sup>-translocating NADH:ferredoxin oxidoreductase (NFO) of *B. producta* were of higher abundance at pH 6.0 (TABLE A-4). These enzyme complexes (NQR and NFO) are essential for the utilization of energy from electron transfer by facilitating the transport of Na<sup>+</sup> from the cytoplasm to the periplasm, thereby creating a sodium ion gradient across the membrane (Mulkidjanian et al., 2008).

The importance of this respiratory mechanism in *Bacteroides* is emphasized by the observation that a single *nqr* deletion mutant of *B. fragilis* was unable to compete with the wild type in the gnotobiotic mouse gut (Ito et al., 2020). Additionally, both *B. producta* and *B. longum* showed a higher abundance of several F<sub>0</sub>F<sub>1</sub>-ATP synthase subunits at pH 6.0 (TABLE A-4).

In contrast, *B. thetaiotaomicron* exhibited a higher abundance of F<sub>0</sub>F<sub>1</sub>-ATP synthase subunits at pH 7.0 (TABLE A-4). Consistent with results in *E. coli*, this observation may suggest that alkaline stress induces ATPase gene expression as a strategy to maintain pH<sub>i</sub> by prioritizing ATP synthesis over hydrolysis (Maurer et al., 2005).

**Carbohydrate Metabolism** – At pH 6.0, the three bacteria exhibited a common proteomic response characterized by increased abundance in key glycolytic enzymes, proteins associated with pyruvate and SCFA metabolism, and proteins involved in glycogen synthesis, polymerization, and branching (TABLE A-4). For *B. thetaiotaomicron*, higher abundance at pH 7.0 of citrate synthase (Q8A616) and isocitrate dehydrogenase (Q8A615) in the oxidative branch of the tricarboxylic acid (TCA) cycle may suggest a potential bifurcation of the TCA cycle into oxidative and reductive branches. This bifurcation plays distinct roles in bacterial metabolic adaptation and has been observed across various growth stages and nutrient conditions for both *B. thetaiotaomicron* and *B. fragilis* (Baughn and Malamy, 2002; Schofield et al., 2018). Metabolite analysis could be helpful in elucidating these potential metabolic adaptations and confirming the presence of a pH-dependent bifurcation in *B. thetaiotaomicron*.

Additionally, *B. thetaiotaomicron* had a higher abundance of inositol-phosphate synthase (Ino1; Q8A7J8) and inositol-monophosphatase (SuhB; Q8A403) at pH 6.0 (FIGURE IV-10A), which suggests an synthesis of *myo*-inositol, a vital precursor for inositol-containing lipids (Heaver et al., 2022; Sartorio et al., 2022). Conversely, *B. producta* had an increased abundance of inositol catabolism-related proteins at pH 7.0, hinting at a potential conversion of inositol into dihydroxyacetone phosphate and 3-oxopropionate (FIGURE IV-10B). These results suggest the utilization of inositol as an alternative carbohydrate energy source. Despite employing various annotation approaches, a potential protein-encoding gene for the *scyllo*-inositol 2-dehydrogenase (IolW) could not be identified. This enzyme is predicted to facilitate the reversible NADP<sup>+</sup>-dependent conversion of *scyllo*-inositol to *scyllo*-inosose (Morinaga et al., 2010). Conversely, only *scyllo*-inositol 2-dehydrogenase (IolX), responsible for the reversible NAD<sup>+</sup>-dependent conversion, could be identified and quantified. Importantly, *B. thetaiotaomicron* and *B. producta* each have unique genetic sets, enabling them to perform either inositol synthesis or degradation.



**FIGURE IV-10 | Inositol Metabolism.** (A) Inositol anabolism in *B. thetaiotaomicron* and (B) inositol catabolism in *B. producta*. Enzymes, represented by their gene names, are colored based on their Log<sub>2</sub> ratio (pH 7.0/pH 6.0). Superscript explanation: 1 = No protein-encoding gene identified.

**Transcriptional and Translational Processes** – All three bacteria exhibited an increased abundance of small and large ribosomal subunits, ribosomal initiation factors, ribosome-binding factor A, ribosome-releasing, and silencing factors at pH 7.0 (TABLE A-4).

In *B. longum*, two site-specific DNA-methyltransferases and eight RNA-methyltransferases showed increased abundance at pH 6.0 (TABLE A-4). These enzymes can participate in maintaining genome integrity, regulating gene expression, and stabilizing transcriptome stability during bacterial stress (Vargas-Blanco and Shell, 2020; Gao et al., 2023). Moreover, both *B. longum* and *B. thetaiotaomicron* had a higher abundance of various aminoacyl-tRNA synthetases at pH 6.0 (TABLE A-4).

Furthermore, the elongation factor-G2 (encoded by *fusA2*) was more abundant at pH 6.0 in both *B. thetaiotaomicron* (Q8A5S1) and *B. producta* (A0A7G5MVL1) (TABLE A-4). The elongation factor G proteins sometimes have chaperone properties (Caldas et al., 2000) and have been previously linked to acid stress (Pérez Montoro et al., 2018). Notably, in *B. thetaiotaomicron*, transcriptional activation of *fusA2* plays a crucial role in regulating protein synthesis during nutrient fluctuations and is essential for successful gut colonization in mice (Townsend et al., 2020). In this context, *fusA2* may serve a similar function in the adaptation of *B. producta*.

**Protection and Repair of Macromolecules** – *B. producta* showed an increased abundance of chaperones at pH 6.0. These include ClpB (A0A7G5N0M9), ClpC (A0A7G5MXB6), DnaK (A0A4P6LWM6), and DnaJ (A0A4P6LZ29) (TABLE A-4). Conversely, in *B. thetaiotaomicron*, ClpB (Q89YY3) and DnaJ (Q8A6R4), along with two heat shock proteins (Q8A6P7 and Q8AAA0), were more abundant at pH 7.0 (TABLE A-4). These chaperone proteins are involved in various cellular processes, including protein synthesis, transport, folding, renaturation, and the removal of damaged proteins, particularly in response to acidic stress (Guan and Liu, 2020). Furthermore, at pH 6.0, a higher abundance of proteins involved in various DNA repair mechanisms, including recombinational DNA repair (RecABCD, RecFOR, RecQ, and RecJ), single-strand DNA binding, and support for the nucleotide excision repair system (such as the UvrABC system, DNA polymerase, and DNA ligase), was observed across all three bacteria (TABLE A-4). Several of these proteins are involved in the SOS response system, a cellular repair system that increases the expression of DNA repair genes to meet the increased demands for DNA repair (Janion, 2001). Notably, the excinuclease subunit UvrA, identified in multiple studies as a key player in repairing acid-induced DNA damage (Hanna et al., 2001; Cappa et al., 2005; Zheng et al., 2018), was more abundant at pH 6.0 in both *B. thetaiotaomicron* (Q8AA87) and *B. longum* (Q8G5D1).

**Cell Wall Morphogenesis** – Proteins involved in the multi-step biosynthesis of cell wall peptidoglycan, including those responsible for the initial synthesis of uridine diphosphate-*N*-acetylglucosamine (UDP-GlcNAc) from fructose-6-phosphate by three enzymes (GlmS, GlmM, and GlmU), as well as its conversion to UDP-*N*-acetylmuramyl-pentapeptide by proteins of the Mur pathway (MurA-F), were more abundant at pH 6.0 in both *B. thetaiotaomicron* and *B. longum* (TABLE A-4). Furthermore, both bacteria had increased abundance of an alanine racemase at pH 6.0, responsible for converting L-alanine to D-alanine and vice versa (TABLE A-4). *B. longum* exhibited higher abundance of D-alanine-D-alanine ligase (Ddl; Q8G7C4) and cyclopropane-fatty-acyl-phospholipid synthase (Cfa; Q8G3T2) at pH 6.0, contributing to the final step of the Mur pathway and the conversion of C<sub>18</sub> phospholipid olefinic fatty acid to cyclopropane fatty acid, respectively (TABLE A-4). Conversely, *B. producta*

had an increased abundance of the cell shape-determining protein (MreB; A0A2S4GRY8), capsule biosynthesis protein (CapA; A0A7G5MP92) (TABLE A-4), and several stage 0 and III sporulation proteins at pH 6.0 (TABLE A-5). CapA facilitates the addition of cell surface capsular polysaccharide that can protect bacteria from immune responses and alter host physiology (Porter and Martens, 2017). While the teichoic acid biosynthesis protein (A0A7G5MS69) was more abundant at pH 7.0, the D-alanyl-lipoteichoic acid biosynthesis protein (DltD; A0A7G5MQG2) increased ninefold at pH 6.0 (TABLE A-4).

Furthermore, several Fts proteins involved in cell elongation, cell cycle control, and cell division were more abundant at pH 6.0 (TABLE A-4). Notably, the cell division protein FtsL (A0A4P6M6W0) showed the highest change in abundance in the dataset, with a 16-fold increase at pH 6.0. Acidic conditions have been reported to reduce the cell length of *E. coli* by modulating the division machinery, specifically the terminal cell division protein FtsN (Mueller et al., 2020). Although these findings originate from a Gram-negative bacterium and may not directly apply to *B. producta*, a Gram-positive bacterium lacking genomic data for FtsN, studies on other Gram-positive bacteria like *S. aureus* and *S. pneumoniae*, which also lack identifiable FtsN homologs, have demonstrated significant size alterations in response to pH variations (Perez et al., 2019; Mueller et al., 2020). These observations suggest the hypothesis that the cellular morphology of *B. producta* might also change in response to pH variations. However, initial electron microscope experiments conducted by Kathrin Schäfer (Department of Infectious Diseases and Microbiology, University of Lübeck, UKSH Lübeck; chaired by Prof. Dr. Jan Rupp) on *B. producta* in response to various culture pH values did not reveal significant changes in cell morphology. Although these preliminary findings do not conclusively rule out the possibility that the cellular morphology of *B. producta* changes under these specific conditions, further research is required to explore this hypothesis.

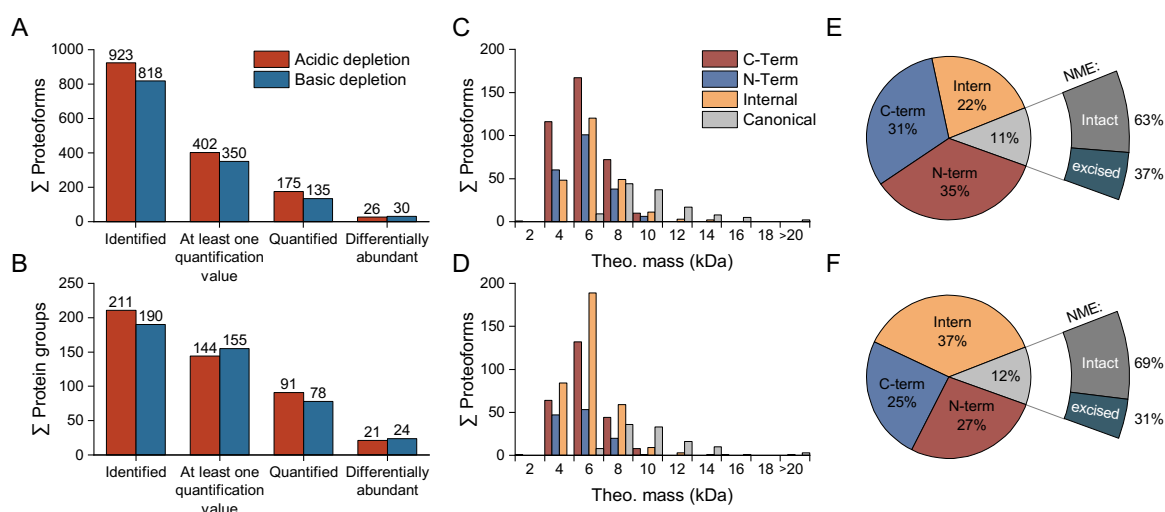
### 3.3 Comparison of LMWP-Top-Down and Full-Proteome Bottom-Up

#### Analysis for *B. producta*

**Data Processing** – Top-down proteomic analysis of acidic and basic HMWP depletions of *B. producta* identified 923 and 818 proteoforms (FIGURE IV-11A), corresponding to 211 and 190 protein groups (FIGURE IV-11B), respectively. The majority of identified proteoforms were between 2-10 kDa in molecular weight and were truncated versions of the canonical proteins (FIGURE IV-11C and D). N-terminal truncation accounted for  $31\% \pm 4\%$  of proteoforms, excluding N-terminal methionine excisions (NME) (FIGURE IV-11B and D). Additionally, C-terminal truncation represented  $28\% \pm 3\%$ , while internal proteoforms resulting from both N- and C-terminal truncation contributed an additional  $30 \pm 8\%$ . This left 12% of identified

proteoforms as full-length canonical proteins, with  $66\% \pm 3\%$  retaining their N-terminal methionine and  $34\% \pm 3\%$  lacking it (FIGURE IV-11E and F).

Interestingly, only 404 and 356 proteoforms had at least one quantitation value after acidic and basic HMWP depletion, respectively (FIGURE IV-11A). Both datasets had a high percentage of missing values, with 38% and 42% missing values after acidic and basic HMWP depletion, respectively. After data processing steps including CV summation, median normalization, and filtering for at least two quantitative values per pH condition, the datasets were reduced to 175 proteoforms after acidic and 135 after basic depletion, leaving many proteoforms unquantified (FIGURE IV-11A).



**FIGURE IV-11 | Identified Proteoforms using the Acidic and Basic HMWP Depletion Method.** Number of identified, quantified, and differentially abundant (A) proteoforms and (B) protein groups. Distribution of identified neo-termini by proteoform size for (C) the acidic and (D) basic HMWP depletion method, and distribution by percentage for (E) the acidic and (F) basic HMWP depletion method.

An overview of the data distribution before and after median normalization and Pearson correlation analysis of biological and technical replicates is provided in the Appendix (FIGURE A-16 and FIGURE A-17). Statistical analysis (two-sided Student's t-test, permutation-based FDR correction,  $FDR \leq 0.05$ , and a  $\text{Log}_2$  fold change of  $\pm 0.485$ ) identified 26 and 30 differentially abundant proteoforms in acidic and basic HMWP depletion experiments, respectively (FIGURE IV-11A and TABLE A-6). In total, 21 proteins were differentially abundant in the acidic HMWP depletion and 24 proteins in the basic HMWP depletion (FIGURE IV-11B). Notably, six proteins had differential abundance changes in both HMWP depletions, with similar abundance changes observed for identical proteoforms or truncated proteoforms of the same protein (TABLE A-7).

**Quantitative data** – The following analysis focused on comparing proteoform-based quantifications from both acidic and basic HMWP depletion methods with the full-proteome peptide-based BUP analysis. While both proteomics-based quantification techniques offer

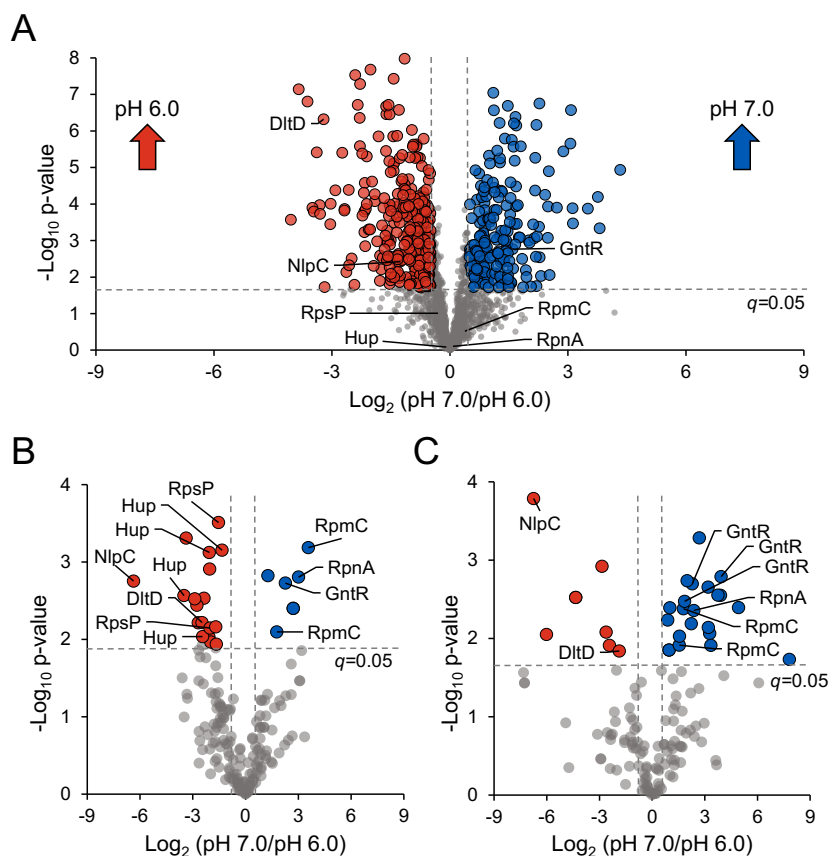
valuable insights, they have different capabilities and limitations. The TDP analysis in this study primarily focused on identifying smaller proteins (<30 kDa), whereas the applied BUP analysis targeted the entire proteome, thus quantifying a broader range of proteins, including larger ones, but it does not identify proteoforms. Detailed variations in quantification, encompassing counts of proteins and proteoforms, are detailed in TABLE A-8 and summarized for both HMWP deletion analysis in TABLE IV-1.

**TABLE IV-1 | Quantitative Results of Top-down Proteomics and Bottom-up Proteomics.** Comparison of proteoform- and protein-level quantitative data acquired by full proteome LFQ bottom-up and acidic and basic HMWP depletion top-down LFQ analysis.

TOP-DOWN QUANTITATIVE RESULTS				
		DIFFERENTIALLY ABUNDANT	NOT DIFFERENTIALLY ABUNDANT	NOT QUANTIFIED
BOTTOM-UP	DIFFERENTIALLY ABUNDANT	13 proteins 16 proteoforms	27 proteins 42 proteoforms	469 proteins N/A
	NOT DIFFERENTIALLY ABUNDANT	22 proteins 33 proteoforms	67 proteins 147 proteoforms	1,089 proteins N/A
	NOT QUANTIFIED	3 proteins 3 proteoforms	20 proteins 33 proteoforms	

While 1,558 proteins were quantified in the BUP analysis but not in the TDP analysis, the TDP uniquely quantified a total of 36 proteoforms from 23 proteins that were absent in the BUP analysis (TABLE IV-1). The majority of these (33 proteoforms from 20 proteins) were not differentially abundant in the TDP analysis (TABLE IV-1). However, three proteoforms from three different proteins, an uncharacterized protein (A0A7G5MR36), an acyl carrier protein (A0A7G5MSW4), and a carbohydrate ABC transporter substrate-binding protein (A0A7G5MNS4), showed differential abundance in the TDP analysis (TABLE IV-1). Further differences between the two quantification methods included 27 proteins (represented by 42 proteoforms) which were detected as differentially abundant in BUP dataset but in the TDP dataset (TABLE IV-1). Conversely, 22 proteins (represented by 33 proteoforms) were differentially abundant in the TDP data but not in the BUP data. These included proteoforms of recombination-promoting nuclease RpnA (RpnA; A0A7G5MZ14) and 50S ribosomal protein L29 (RpmC; A0A4P6LZK9), both of higher abundance at pH 7.0, and proteoforms of DNA-binding protein HU (Hup; A0A2S4GGS2) and ribosomal protein S16 (RpsP; A0A4P6M2Y5), which were more abundant at pH 6.0 (FIGURE IV-12). Notably, all differentially abundant proteoforms of DNA-binding protein HU proteoforms exhibited a canonical N-terminal (TABLE A-6), which has been reported to be relevant for forming a DNA–protein complex (Almarza et al., 2015). This complex protects DNA from endonucleolytic cleavage and oxidative stress damage under acidic pH (Almarza et al., 2015). For 67 proteins, both quantification methods detected no significant difference between pH conditions, while 13 proteins were classified as

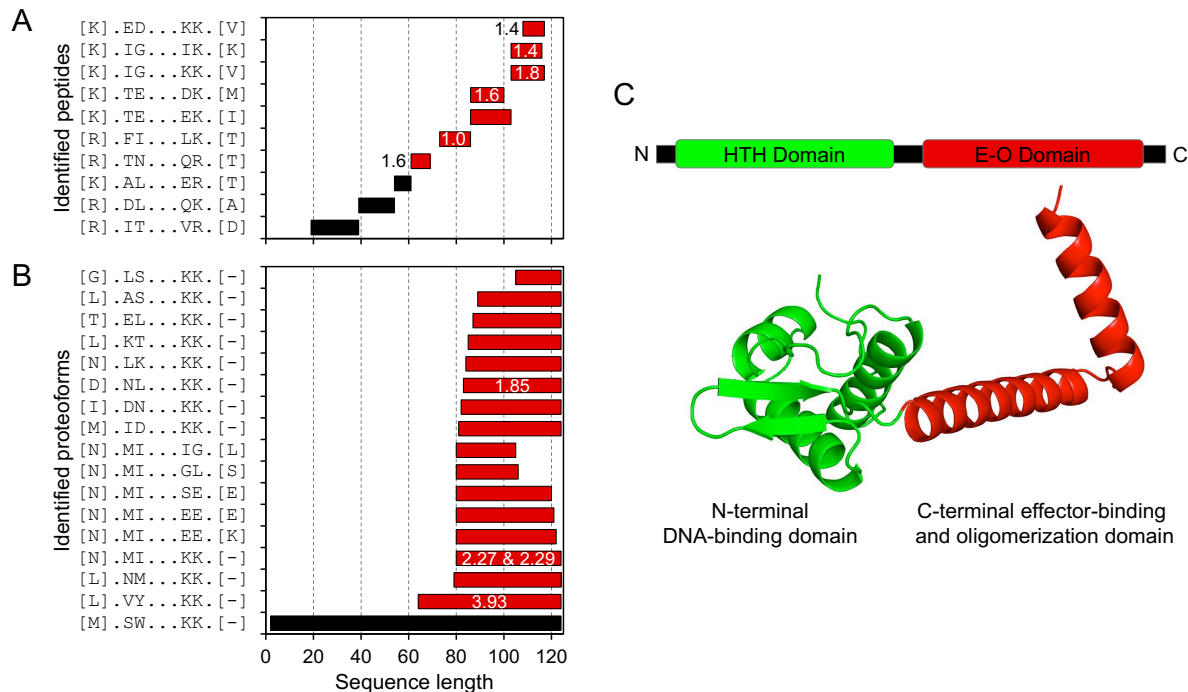
differentially abundant by both TDP and BUP (TABLE IV-1). Of these 13 proteins, each TDP depletion method uniquely quantified 5 proteins, with 3 proteins showing differential abundance in both depletion analyses, including the NlpC/P60 domain-containing protein (NlpC; A0A7G5N0P5), D-alanyl-lipoteichoic acid biosynthesis (DltD; A0A7G5MQG2), and a GntR family transcriptional regulator (GntR; A0A7G5MZW5) (FIGURE IV-12 and TABLE IV-2). Notably, both the BUP and the two TDP proteomics-based quantifications revealed identical abundance changes, with GntR being more abundant at pH 7.0 and NlpC and DltD being more abundant at pH 6.0 (FIGURE IV-12 and TABLE IV-2). Notably, even the different proteoforms of GntR exhibited consistent changes in abundance (TABLE IV-2). Moreover, several proteoforms, including the internal 36-amino acid-long proteoform of DltD and the N-terminal truncated 44-amino acid-long proteoform of GntR, were quantified as differentially abundant in both TDP datasets (TABLE A-7).



**FIGURE IV-12 | Comparison Top-down and Bottom-up Label-free-quantification. (A)** Bottom-up label-free quantitative analysis. **(B)** Top-down label-free quantitative analysis from acidic depletion and **(C)** from basic depletion. Highlighted proteins: D-alanyl-lipoteichoic acid biosynthesis (DltD; A0A7G5MQG2), GntR family transcriptional regulator (GntR; A0A7G5MZW5), DNA-binding protein HU (Hup; A0A2S4GGS2), NlpC/P60 domain-containing protein (NlpC; A0A7G5N0P5), recombination-promoting nuclease RpnA (RpnA; A0A7G5MZ14), 50S ribosomal protein L29 (RpmC; A0A4P6LZK9) and 30S ribosomal protein S16 (RpsP; A0A4P6M2Y5).

**TABLE IV-2 | Overlap of Differentially Abundant Proteins between Bottom-up and Top-down Proteome Analyses.** Top-down Log<sub>2</sub> ratios represent quantified proteoforms, with multiple values corresponding to different quantified proteoforms.

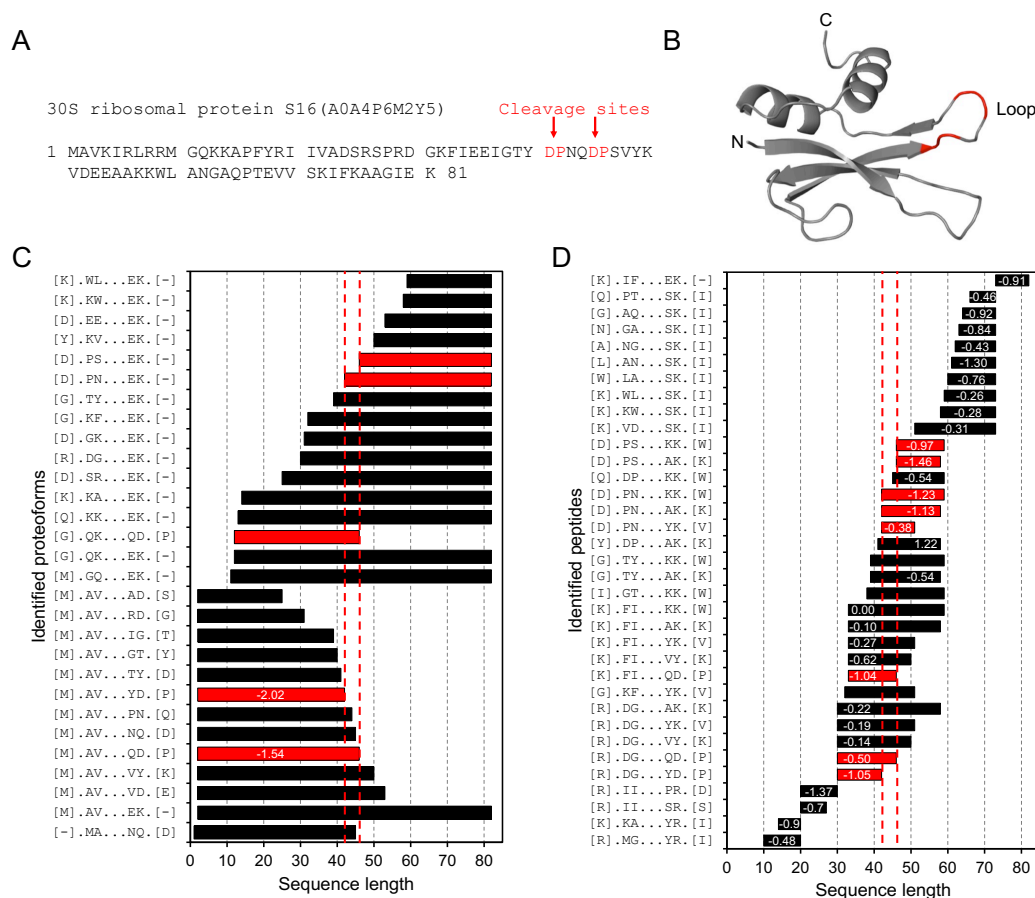
		Log <sub>2</sub> ratio (pH 7.0/pH 6.0)		
PROTEIN NAME	ACCESSION	TDP		BUP
		ACIDIC	BASIC	
NlpC/P60 domain-containing protein	A0A7G5N0P5	-6.4	-6.8	-1.3
D-alanyl-lipoteichoic acid biosynthesis	A0A7G5MQG2	-2.5	-1.9	-3.2
GntR family transcriptional regulator	A0A7G5MZW5	2.3	3.9, 2.3, 1.9	1.6
ABC transporter substrate-binding	A0A7G5MSE0	-2.9	-	-0.5
Uncharacterized protein	A0A7G5N1D2	-2.8	-	1.2
50S ribosomal protein L19	A0A4P6M0Y8	-1.7	-	0.8
XRE family transcriptional regulator	A0A7G5N1G9	-1.7	-	1.0
Uncharacterized protein	A0A7G5N1C6	2.7	-	3.5
DUF1002 domain-containing protein	A0A7G5MNQ5	-	-6.0	1.3
Recombinase RecT	A0A7G5N1F5	-	0.9	-0.8
NADH peroxidase	A0A4P6LUJ8	-	1.0	0.9
Uncharacterized protein	A0A7G5N144	-	1.8	2.3
DUF3502 domain-containing protein	A0A7G5MQI1	-	7.8	0.8



**FIGURE IV-13 | Quantification of the C-terminal Region of the GntR family transcriptional regulator (A0A7G5MZW5).** Distribution of identified (A) peptides and (B) proteoforms with highlighted Log<sub>2</sub> ratios (pH 7.0/pH 6.0). (C) NCBI-CDD and AlphaFold structure prediction revealed an N-terminal HTH (helix-turn-helix) DNA-binding domain and C-terminal effector-binding and oligomerization (E-O) domain.

### 3.4 Potential pH-induced Asp-Pro Cleavage

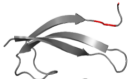
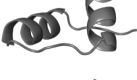


The analysis of neo-termini in TDP datasets revealed multiple proteoforms of the ribosomal protein S16 (A0A4P6M2Y5) with N- or C-terminal Asp-Pro peptide bond cleavages. Further sequence analysis of ribosomal protein S16 revealed two Asp-Pro bonds within a loop structure, that separates the prominent N-terminal  $\beta$ -sheets from the C-terminal  $\alpha$ -helices (FIGURE IV-14A-B). Overall, five proteoforms exhibiting an Asp-Pro cleavage were identified (FIGURE IV-14C), two of which were differentially higher abundant at pH 6.0 in the acid HMWP depletion TDP analysis (FIGURE IV-12B). Although the basic HMWP depletion TDP analysis also detected these two proteoforms, quantitative data could not be acquired. A re-analysis of the BUP data, incorporating semi-tryptic peptides, identified several peptides with the same Asp-Pro cleavage, most of which were of higher abundance at pH 6.0 (FIGURE IV-14D). Further subcellular localization predictions using Phobius suggested that the resulting proteoforms containing C-terminal  $\alpha$ -helices may be directed outside the cytoplasmic space (TABLE IV-3). Additionally, predictions of antimicrobial peptide (AMP) activity using AMPfun suggested that these proteoforms exhibit activity against both Gram-positive and Gram-negative bacteria (TABLE IV-3).



**FIGURE IV-14 | Asp-Pro Cleavage in 30S Ribosomal Protein S16 (A0A4P6M2Y5).** (A) The protein sequence exhibits two Asp-Pro cleavage sites (residues 41-42 and 45-46). (B) The structure prediction highlights the position of the Asp-Pro peptide bonds situated in a structural loop. (C) Identified proteoforms and (D) identified peptides with highlighted Log<sub>2</sub> fold changes (pH 7.0/pH 6.0).

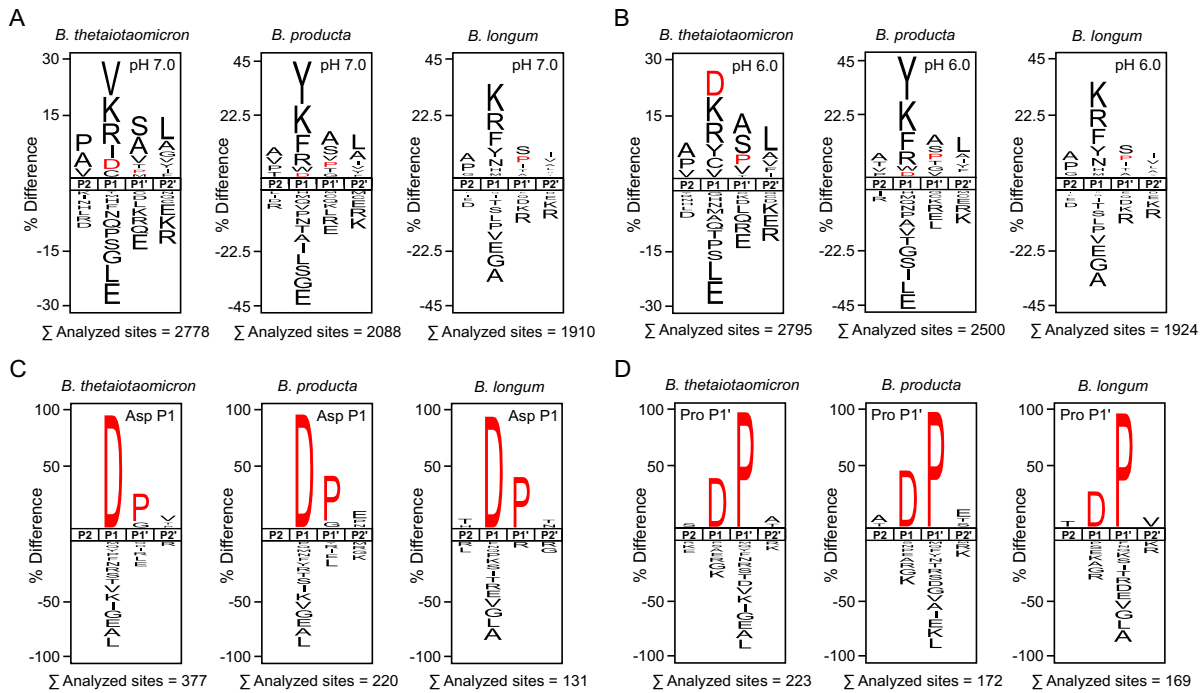
To investigate the predicted AMP activity, five synthetic peptides, each consisting of 22 amino acids, were designed to cover both the N- and C-terminal Asp-Pro cleavage sites (TABLE A-9). Initial testing of the potential AMP activity was conducted at concentrations ranging from 0.01 to 1  $\mu$ M against various members of the human gut microbiota (*B. thetaiotaomicron*, *B. longum*, *A. caccae*, *C. buytricum*, *C. ramosum* and *L. plantrum*). These experiments were performed by Kathrin Schäfer (Department of Infectious Diseases and Microbiology, University of Lübeck, UKSH Lübeck; chaired by Prof. Dr. Jan Rupp) and did not reveal significant AMP activity.

**TABLE IV-3 | Sequence and Structure Predictions of 30S Ribosomal Protein S16.**

	SEQUENCE	FRAGMENT (#AA)	AMP (PREDICTION)	AMP TARGET (PREDICTION)	LOCALIZATION (PREDICTION)	STRUCTURE PREDICTION
First DP motif cleaved	[M] .AV . . . YD . [P]	N-term (40)	AMP (0.757)	Gram-positive (0.550), Gram-negative (0.578)	Cytoplasmic (0.775)	
	[D] .PN . . . EK . [-]	C-term (40)	AMP (0.702)	Gram-negative (0.586)	Non-cytoplasmic (0.681)	
Second DP motif cleaved	[M] .AV . . . QD . [P]	N-Term (44)	AMP (0.7181)	Gram-negative (0.527)	Cytoplasmic (0.774)	
	[D] .PS . . . EK . [-]	C-Term (36)	AMP (0.788)	Gram-positive (0.575), Gram-negative (0.566)	Non-cytoplasmic (0.685)	

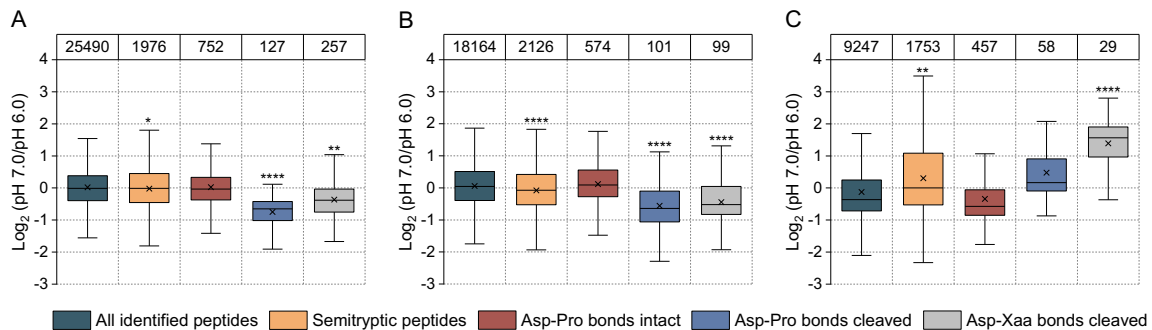
Further investigation was focused on evaluating the possibility of artificially generated Asp-Pro bond hydrolysis during sample preparation, LC-MS measurement, and their potential origin from acidic cultivation.

Analysis of the amino acid frequencies surrounding non-tryptic cleavage sites in the BUP datasets revealed distinct sequence logos among different bacterial species (FIGURE IV-15). Specifically, *B. thetaiotaomicron* and *B. producta* showed an increased relative frequency of Asp at the P1 position and Pro at the P1' position (FIGURE IV-15A). Interestingly, this frequency increased for peptides identified at pH 6.0, particularly for *B. thetaiotaomicron*, suggesting that acidic conditions may promote the potential hydrolysis of Asp-Pro bonds. In contrast, *B. longum* exhibited a significantly increased relative frequency only for Pro at the P1' position. Further analysis focusing on peptides with Asp at the P1 position (FIGURE IV-15C) or Pro at the P1' position (FIGURE IV-15D) confirmed the presence of peptides with Asp-Pro cleavage and also indicated their occurrence for *B. longum*.



**FIGURE IV-15 | Bottom-up Peptide Cleavage Analysis Highlights Asp-Pro Sequence Logos.** Icelogo plot illustrating the relative frequency of P2-P2' amino acids of identified peptides for *B. thetaiotaomicron*, *B. producta* and *B. longum*. **(A)** Identified peptides at pH 7.0 and **(B)** at pH 6.0. **(C)** Identified peptides with Asp at the P1 position and **(D)** with Pro at the P1' position. Amino acids that are significantly enriched (top) or depleted (bottom) ( $p \leq 0.05$ ) are colored black, with Asp and Pro residues highlighted in red.

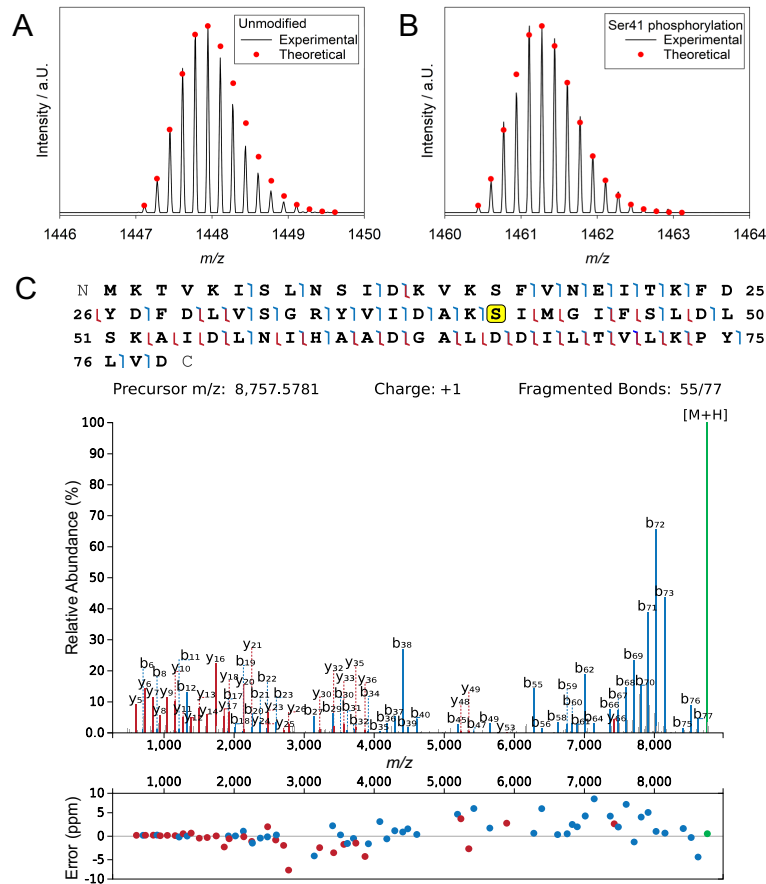
Moreover, significantly higher abundance at pH 6.0 was observed for peptides with cleaved N- or C-terminal Asp-Pro and Asp-Xaa bonds in both *B. thetaiotaomicron* (FIGURE IV-16A) and *B. producta* (FIGURE IV-16B). Conversely, peptides with intact Asp-Pro showed no significant changes in abundance and were distributed around zero (FIGURE IV-16A-B). Additionally, at pH 7.0, peptides with cleaved N- or C-terminal Asp-Xaa bonds exhibited significantly higher abundance in *B. longum* (FIGURE IV-16C).



**FIGURE IV-16 | Peptide Abundance Distribution.** Distribution of all identified peptides, semitryptic peptides (excluding those with intact Asp-Pro, cleaved Asp-Pro and Asp-Xaa bonds), peptides with intact Asp-Pro bonds, peptides with cleaved Asp-Pro and cleaved Asp-Xaa bonds. **(A)** *B. thetaiotaomicron*, **(B)** *B. producta*, and **(C)** *B. longum*. Significant differences were calculated via one-way ANOVA with Dunnett's correction. \* ( $p < 0.05$ ); \*\* ( $p < 0.01$ ); \*\*\*\* ( $p < 0.0001$ ). The number on the top indicates the number of identified peptides.

### 3.5 Phosphorylation of HPr proteins

An open-modification search of the TDP dataset identified several potential seryl-phosphorylations on histidine-containing phosphocarrier proteins (HPr). Overall, a high number of matching fragment ions covering the phosphorylation site, along with a matching distribution of theoretical and experimentally observed peaks of unmodified and phosphorylated HPr proteoforms, was observed (FIGURE IV-17 and FIGURE A-19A-D). Using ProSight Annotator, potential phosphorylation sites were incorporated into HPr protein entries, facilitating a variable search for HPr phosphorylation. Combined with an additional database search of the bottom-up proteomic data that included phosphorylation as a variable modification at serine, histidine, threonine, tyrosine, and arginine residues, several phosphorylated peptides and proteoforms were identified. These findings suggested Ser-41 phosphorylation for HPr (A0A2S4GRU0 and A0A7G5MYS7), Ser-46 phosphorylation for HPr (A0A4V0Z7D2), and Arg-46 phosphorylation for HPr (A0A7G5MP53) (FIGURE IV-18). Although peptides spanning the respective phosphorylation sites of HPr (A0A7G5MP53 and A0A4V0Z7D2) were detected by the bottom-up proteomics analysis, phosphopeptides could only be identified for two HPr variants (A0A2S4GRU0 and A0A7G5MYS7) (FIGURE A-19E-H).



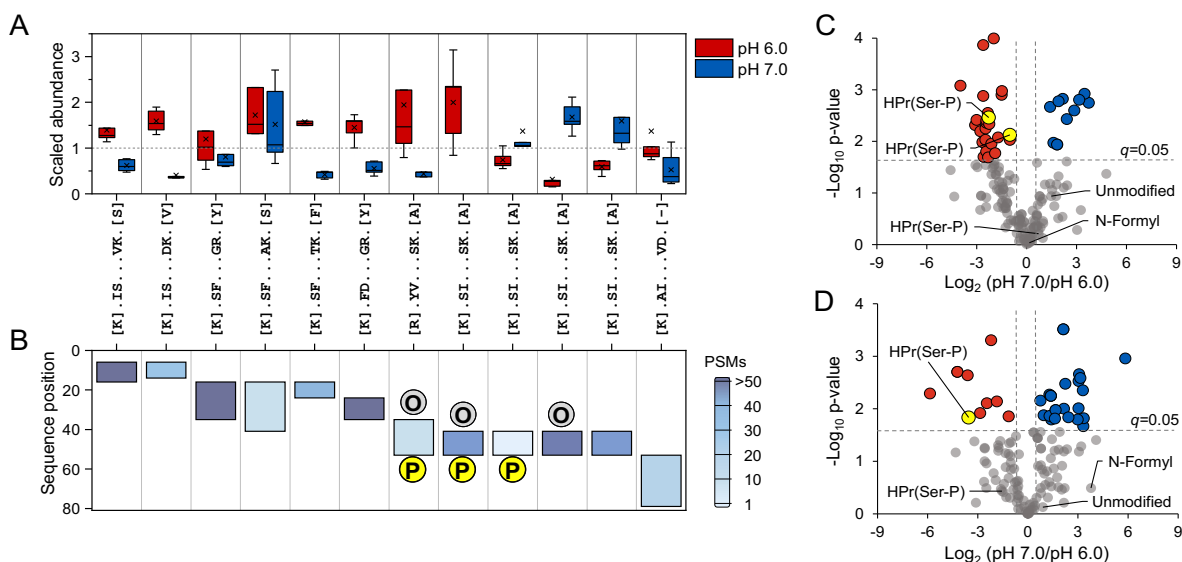
**FIGURE IV-17 | Phosphoserine Modification of HPr (A0A2S4GRU0).** Distribution of experimentally observed and theoretical peaks of identified proteoforms for **(A)** unmodified HPr and **(B)** phosphorylated HPr. **(C)** Identified proteoform sequence and fragment ion spectrum, highlighting the detected phosphorylation site at Ser-41 in yellow. The identified b- and y-ions are highlighted in blue and red, respectively. Dot plot illustrates the mass error in ppm for each observed b- and y-ion, respectively.

#### IV | INFLUENCE OF PH ON BACTERIAL PROTEOMES

				Identified using
A0A4P6M732	8-VNNLIGLHLRPAG-20	42-ANAKSVLSV-50		
A0A7G5MP53	8-ITDPEGIHARPAG-20	42-GDCKRIFGI-50		Top-down proteomics
A0A4V0Z7D2	8-IGISNGLHARPIA-20	42-VNAKSIMGM-50		Top-down proteomics
A0A7G5MYS7	8-LNET---GDVK-20	37-IDAKSILGV-45		Top-down & Bottom-up proteomics
A0A2S4GRU0	8-LNSI---DKVK-20	37-IDAKSIMG-45		Top-down & Bottom-up proteomics
A0A4P6LXB3	8-FSEI---NEIK-20	37-IDAKSILGM-45		
A0A4P6LTJ7	7-FKEV---DEIV-19	36-VDAKSIMGM-44		

**FIGURE IV-18 | Sequence Alignment of HPr Proteins Highlighting Proposed Phosphorylation Sites.** Positions of the two conserved amino acid residues: a histidine (His-15) near the N-terminus and a serine (Ser-46) in the central part of the protein, are highlighted.

The quantitative data of the BUP analysis revealed a differential abundance of HPr (A0A2S4GRU0) at pH 6.0. Moreover, analysis of peptide-level abundance revealed a higher abundance of phosphopeptides at pH 6.0, while unmodified peptides were more abundant at pH 7.0 (FIGURE IV-19A-B). Statistical analysis of the TDP data, which incorporated the variable HPr phosphorylation sites, quantified several phosphorylated proteoforms as differentially abundant at pH 6.0, both by the acidic (FIGURE IV-19C) and the basic HMWP depletion (FIGURE IV-19D). Conversely, unmodified or N-terminal formylated proteoforms did not exhibit differential abundance (FIGURE IV-19C-D). Overall, the observed increase in the abundance of phosphorylated peptides and proteoforms of HPr (A0A2S4GRU0) at pH 6.0 suggests that HPr phosphorylation may represent a direct response to culture acidification.



**FIGURE IV-19 | Quantification of Phosphorylation on HPr (A0A2S4GRU0).** (A) Scaled peptide abundance at pH 6.0 and pH 7.0. Peptides with serine phosphorylation (P) or methionine oxidation (O) are highlighted. One boxplot is absent as the phosphopeptide was exclusively identified at pH 6.0. Proteoform-directed label-free analysis from (B) acidic and (C) basic depletion. Phosphorylated HPr proteoforms with differential abundance at pH 6.0 are highlighted in yellow (Two-sided Student's t-test, Permutation-based FDR corrected, q-value of 0.05).

## 4 Discussion and Conclusion

### 4.1 Quantitative analysis of the *B. longum* proteome

While the pH of the human intestinal environment can become relatively alkaline, reaching levels of up to pH 8.0 primarily due to pancreatic secretions (Nugent et al., 2001), there is a notable gap in research regarding the proteomic response in *B. longum*. Most studies on *B. longum* have centered on its response to acidic stress, comparing parent strains with evolutionary engineered acid-resistant variants such as *B. longum* biotype longum NCIMB 8809 (Sánchez et al., 2007), *B. longum* BBMN68 (Jiang et al., 2016) and *B. longum* JDM301 (Wei et al., 2019). While these studies have identified several proteins associated with acid resistance, there remains, to the best of my knowledge, a gap in research regarding the proteomic analysis of *B. longum* NCC 2705 from the Nestlé Culture Collection (NCC) (Nestlé SA, Lausanne, Switzerland) following cultivation under varying pH conditions. Therefore, this study analyzed the proteomic response of *B. longum* NCC 2705 to acidic, neutral, and alkaline culture conditions.

The variations in pH resulted in differential protein abundances, associated with key cellular processes such as amino acid metabolism, fatty acid synthesis, and peptidoglycan biosynthesis (FIGURE IV-4). Several proteins involved in peptidoglycan synthesis and modification, including those of the Glm and Mur pathways, alanine racemase (Alr; Q8G571), D-alanine-D-alanine ligase (Ddl; Q8G7C4), and cyclopropane-fatty-acyl-phospholipid synthase (Cfa; Q8G3T2), exhibited higher abundance at pH 6.0. The increase in cyclopropane fatty acid composition within cell membranes can render the cell surface more hydrophobic, altering the viscosity and permeability of the membrane, which in turn impedes proton and hydroxyl ion diffusion into the cell (Krulwich et al., 2011). Notably, this also has been suggested to protect *B. longum* BBMN68 from bile acids (An et al., 2014).

Furthermore, at pH 6.0, several aminoacyl-tRNA synthetases exhibited higher abundance. These enzymes are often equipped with proofreading mechanisms such as enzymatically hydrolyzing misactivated aminoacyl-adenylates ('pretransfer' editing) and clearing non-cognate tRNAs using a separate deacylation domain ('post-transfer' editing), which can be crucial for maintaining translational fidelity (Martinis and Boniecki, 2010). Additionally, F<sub>0</sub>F<sub>1</sub>-ATP synthase subunits showed increased abundance at pH 6.0. This protein complex can facilitate the transport of H<sup>+</sup> out of cells at the expense of ATP consumption, generating a transmembrane proton motive force that can counteract a potential drop in intracellular pH (Guan and Liu, 2020). To meet the increased energetic demands, anaerobic bacteria typically employ substrate-level phosphorylation reactions, directly transferring a phosphate group from substrate molecules to ADP or GDP, resulting in the formation of ATP or GTP, respectively (Guan and Liu, 2020).

Notably, the conversion of 1,3-bisphosphoglycerate to 3-phosphoglycerate, a crucial step in the glycolytic pathway catalyzed by phosphoglycerate kinase (Q8G6D6), as well as the conversion of succinyl-CoA to succinate, the only substrate-level phosphorylation step in the TCA cycle, catalyzed by succinyl-CoA synthetase (SucC; Q8G6B4 and SucD; Q8G6B3), exhibited increased abundance at pH 6.0. Particularly, succinyl-CoA synthetase, which can utilize the high proton potential under acidic conditions through proton transfer in the production of ADP or GDP, has been previously associated with *E. coli*'s response to acidic pH (Maurer et al., 2005).

In contrast, at high pH, classic protective mechanisms involve ATPase proton uptake, small ion ( $\text{Na}^+$ ,  $\text{K}^+$ , or  $\text{Ca}^{2+}$ )/ $\text{H}^+$  antiport, and the synthesis and/or uptake of acidic metabolites, especially organic acids and acidic amino acids (Krulwich et al., 2011). However, in this study, no increase in ATPase subunit proteins was observed at pH 8.0, and although *B. longum* possesses multiple  $\text{Na}^+/\text{H}^+$  antiporter homologs (NhaA homolog: Q8G5R2; NhaC homolog: Q8G5A5 and NhaP homolog: Q8G474), they were not detected. Moreover, while alkaline culture conditions did not lead to differential changes in the abundance of amino acid deaminases or in catabolic pathways such as glycolysis or the TCA cycle, which produce mono-, di-, or tricarboxylic acids, it is important to consider that protein abundance does not always correlate directly with functional activity. The specific protein, its regulatory mechanisms, and the context in which it operates are crucial factors. Post-translational modifications, binding partners, cellular localization, and conformational changes can profoundly influence enzymatic activity. In addition, certain proteins can require activation by either proteolysis (in the case of proenzymes) or binding with a cofactor (in the case of apoenzymes) to exhibit catalytic activity. Therefore, even if a protein is differentially abundant, its activity might be modulated by these regulatory mechanisms.

Conversely, proteins involved in the biosynthesis and interconversion of cysteine and methionine exhibited differential abundance at pH 8.0. Although the specific role of this result remains unknown, similar alkaline pH-dependent alterations in proteins related to methionine metabolism have been observed in proteomic analyses of *C. jejuni* (Ramires et al., 2023). Together with the increased abundance of two S-adenosyl-L-methionine-dependent proteins, these alterations may indicate changes in the metabolism of  $\text{C}_1$  units (methyl groups). Particularly noteworthy is the increased abundance of MiaA (Q8CY49) and MiaB (Q8G4H4) at pH 8.0, which catalyze the formation of 2-methylthio- $\text{N}^6$ -isopentenyl adenosine tRNAs and can play a crucial role in enhancing translational fidelity by promoting correct codon-anticodon base-pairing. This process has been associated with the selective translation of mRNAs involved in sporulation or virulence (Edwards et al., 2020). Although factors such as nutrient, phosphate, and iron limitation have been linked to 2-methylthio- $\text{N}^6$ -isopentenyl adenosine tRNA formation (Edwards et al., 2020), the pH-specific modification warrants further investigation.

Overall, the proteomic analysis identified several proteins with increased abundance at pH 7.0 or both pH 6.0 and pH 8.0. An example suggesting potential involvement during acidic and alkaline stress is the alternative sigma factor ( $\sigma^E$ ) (Q8G4M2), which exhibited differentially higher abundance at pH 6.0 and pH 8.0. This factor has been proposed to play an important role during heat stress (Rezzonico et al., 2007), carbohydrate switch (Duboux et al., 2023) in *B. longum* NCC 2705, as well as during bile acid exposure in *B. longum* BBMN68 (An et al., 2014). Homologs of this alternative sigma factor have been reported to regulate the transcription of a wide array of stress response genes, including those involved in pH stress (Helmann, 2002).

Directional analysis (Yang et al., 2014) further revealed that the majority of shared acidic and alkaline response proteins exhibited co-abundances at either pH 7.0 or both pH 6.0 and pH 8.0 (FIGURE IV-5A). The directional pathway analysis revealed that proteins co-abundant at pH 7.0 were primarily associated with fundamental cellular processes such as translational and transcriptional regulation (FIGURE IV-5B). This suggests a prioritization of genetic expression and protein translation under neutral pH conditions, likely reflecting its optimal growth conditions at this pH value (Khaskheli et al., 2015). Conversely, *B. longum* NCC 2705 may reduce protein synthesis and replication during acidic or alkaline conditions, potentially transitioning into a less-proliferative state to prioritize stress response mechanisms (Guo and Gross, 2014). Further, the presence of proteins involved in the transport and metabolism of nucleotides, amino acids, carbohydrates, and coenzymes, as well as those related to energy production and conversion at both pH 6.0 and pH 8.0 indicates a concerted effort to sustain cellular functions despite pH fluctuations (FIGURE IV-5C). This may suggest that *B. longum* NCC 2705 prioritizes energy-generating pathways and metabolic adjustments to support vital cellular processes and maintain metabolic homeostasis under acidic and alkaline stress. Overall, these results indicate that changes in the culture pH, whether to an alkaline or acidic environment, lead to comparable alterations in the proteome *B. longum* NCC 2705, with the bacterium adapting its protein expression profiles to meet the specific challenges posed by different pH conditions.

While other studies investigating the global response to acidic or alkaline pH also suggest that certain genes or proteins exhibit similar responses, the majority of alterations are reported at one of the two pH extremes (Ramires et al., 2023). For instance, in *B. subtilis*, a higher number of genes showed increased expression under alkaline conditions compared to those under acidic conditions (Wilks et al., 2009). In contrast, *E. coli* exhibited a higher abundance of differentially abundant proteins at acidic pH compared to alkaline pH (Maurer et al., 2005). While studies across species may indicate similar adaptations to cope with pH alterations by modulating multiple transporters (such as ATPase or  $H^+$  translocation), amino acid metabolism (amino acid decarboxylases/deaminases), or the generation of carboxylic acids (via glycolysis

or the TCA cycle), specific components can still vary to some extent, even among different strains of the same bacteria (Krulwich et al., 2011).

This study suggests that *B. longum* translates similar proteins to some extent to minimize changes at either acidic or alkaline pH.

## 4.2 Quantitative analysis of the *B. thetaiotaomicron* proteome

Bacterial central nitrogen metabolism plays an essential role in providing primary nitrogen donors for various metabolic molecules, including nucleotides and other amino acids, through transamination reactions. This metabolic pathway involves only three key metabolites (glutamate, glutamine, and  $\alpha$ -ketoglutarate) and three enzymes (GS, GOGAT, and GDH). Notably, in *B. thetaiotaomicron*, GS and GOGAT and enzymes crucial for transamination, which utilize glutamine or glutamate as nitrogen sources, exhibited differentially higher abundance at pH 6.0 (FIGURE IV-9).

Glutamate and glutamine are among the most abundant intracellular amino acids in bacterial cells, with concentrations reaching 96 mM and 3.8 mM, respectively (Bennett et al., 2009). Glutamate is essential for maintaining intracellular potassium levels and electrical potential during osmotic shock (McLaggan et al., 1994; Yan et al., 1996). While the specific role of potassium in bacteria remains elusive, extracellular  $K^+$  (along with  $Na^+$ ) can activate the sensor kinase phosphotransferase (EvgS), leading to the expression of a broad array of acid-resistance genes (Eguchi and Utsumi, 2014). Additionally, glutamate synthesis may be vital for restoring  $pH_i$ , particularly in the presence of weak acids like acetate (Roe et al., 1998).

Many bacteria rely on an acid-resistance system that utilizes glutamate or glutamine to survive in highly acidic environments (Lin et al., 1995; Kanjee and Houry, 2013; Lu et al., 2013; Pennacchietti et al., 2018; Ikeyama et al., 2020). Typically, these systems involve a glutaminase (GlsA) and the antiporter GadC, which facilitate the import of glutamine and the synthesis and export of either glutamate or GABA (Pennacchietti et al., 2018). Similarly, *B. thetaiotaomicron* employs pH-regulated GABA production as a protective mechanism against acid stress, with maximum GABA production occurring at pH 3.1 (Otaru et al., 2021). This pH-dependent GABA production may be attributed to *B. thetaiotaomicron*'s glutamate decarboxylase (GadB), which exhibits optimal activity at pH 3.6 and is regulated by pH-dependent conformational changes (Liu et al., 2023).

While existing literature points towards a central role of GABA production in the acid tolerance of *B. thetaiotaomicron*, proteomic results of this study showed a higher abundance of both glutaminase and glutamate decarboxylase at pH 7.0 (FIGURE IV-9). Similarly, a study of alkaline-grown *E. coli* indicated that inducing the GAD system may protect *E. coli* from acidification resulting from anaerobic fermentation (Blankenhorn et al., 1999). Although produced GABA could potentially be converted into succinate using GABA transaminase

(GabT) (Stancik et al., 2002), such genetic information is missing in *B. thetaiotaomicron*. Further research is needed to fully understand whether the GAD system in *B. thetaiotaomicron* serves broader functions beyond its role in acid tolerance.

Inositol, a carbocyclic sugar polyalcohol with six hydroxyl groups, can exist in nine stereoisomers, with *myo*-inositol (*cis*-1,2,3,5- *trans*-4,6-cyclohexanehexol) and D-*chiro*-inositol (*cis*-1,2,4-*trans*-3,5,6-cyclohexanehexol) being the most biologically active and abundant stereoisomers (Thomas et al., 2016). In this discussion, these different stereoisomers will simply be referred to as inositol. In *B. thetaiotaomicron*, enzymes involved in the endogenous synthesis of *myo*-inositol from glucose 6-phosphate using inositol-phosphate synthase (Q8A7J8) and inositol-monophosphatase (Q8A403) were differentially higher abundant at pH 6.0 (FIGURE IV-10A). The inositol-phosphate synthase in *B. thetaiotaomicron* is encoded within an operon (BT1522–1526), which plays a crucial role in the synthesis of inositol-containing lipids such as phosphatidylinositol (PI) and inositol phosphoceramide (IPC) (Heaver et al., 2022; Sartorio et al., 2022).

Data from human cohorts, including healthy and IBD patients, showed a negative correlation between *Bacteroides* sphingolipids and inflammation, highlighting their essential role in maintaining gut homeostasis and symbiosis (Brown et al., 2019). Loss of inositol lipid production alters capsule expression and antimicrobial peptide resistance in *B. thetaiotaomicron*, resulting in reduced bacterial fitness in a gnotobiotic mouse model (Heaver et al., 2022). Therefore, inositol-containing lipids appear to be vital for *B. thetaiotaomicron* to preserve membrane and capsule structure (Brown et al., 2019; Heaver et al., 2022). Although other enzymes involved in inositol-containing lipid synthesis could not be identified, the results of this study may indicate a potential role of inositol-containing lipid synthesis as a mechanism to enhance bacterial fitness under acidic stress conditions.

### 4.3 Quantitative analysis of the *B. producta* proteome

**Bottom-up Proteomic Analysis** – Several enzymes involved in inositol degradation were differentially more abundant at pH 7.0 compared to pH 6.0 (FIGURE IV-10B). While this catabolic pathway has been extensively characterized in aerobic bacteria (Magasanik, 1953; Yebra et al., 2007; Yoshida et al., 2008), analogous studies on anaerobic metabolism remain limited. Genetic studies suggest that only 16.6% of human gut bacteria possess inositol catabolic gene clusters, enabling them to degrade inositol into the proposed main end products: propionate and acetate (Bui et al., 2021; Weber and Fuchs, 2022). This anaerobic degradation of inositol into these SCFAs could play a crucial role in gut homeostasis and host health (Parada Venegas et al., 2019).

Inositol can be acquired through two main mechanisms: synthesizing it *de novo* from glucose 6-phosphate, a capability *B. producta* lacks genetically, or importing free inositol from the environment using inositol transporters (Reynolds, 2009). Although 35 putative transporter genes of the major facilitator superfamily are annotated, responsible for Na<sup>+</sup>-or H<sup>+</sup> coupled inositol symport, they have not yet been experimentally linked to any transport activities in *B. producta*.

Yeast extract, a component of the growth medium, contains various inositol-containing phospholipids, such as mannose-(inositol-P)<sub>2</sub>-ceramide (Steiner et al., 1969). These phospholipids might have contributed to the observed increased abundance in proteins involved in inositol degradation (FIGURE IV-10B). However, it remains unclear whether these inositol-containing phospholipids solely induce the genetic expression and translation of the *B. producta* inositol catabolic gene cluster, or whether they are hydrolyzed outside the cell to release inositol, which can then be imported and metabolized. The potential degradation of these phospholipids would require specific enzymes capable of hydrolyzing phosphodiester bonds. This could potentially involve phospholipase C (EC 3.1.4.11), which cleaves phospholipids to release inositol-phosphates, and monophosphate phosphatase (EC 3.1.3.25), which hydrolyzes inositol-phosphate into inositol and inorganic phosphate (P<sub>i</sub>). While *B. producta* genome annotations suggest the presence of potential phospholipase C/D domain-containing proteins (A0A7G5MTX5 and A0A7G5MU68), they have not been identified. Additionally, the monophosphate phosphatase remains unannotated.

Additionally, the proteomic results revealed that proteins associated with histidine biosynthesis exhibited higher abundance at pH 6.0 compared to pH 7.0 (FIGURE IV-8A). Histidine can further be converted into biologically active amines such as histamine (decarboxylated histidine) and ergothioneine (a 2-thiourea derivative of histidine). Especially the histidine decarboxylase system has been linked to bacteria's ability to survive in acidic conditions (Diaz et al., 2020). However, both the histidine decarboxylase system and the pathway for ergothioneine synthesis are likely absent in *B. producta* due to the lack of necessary enzyme-encoding genes.

The analysis of protein-bound histidine and the abundance of histidine-containing proteins showed comparable results at both pH 6.0 and pH 7.0 (FIGURE IV-8E-F), suggesting a potential role for free histidine in buffering and regulating pH<sub>i</sub>, rather than its incorporation into proteins. Given the high energy costs of synthesizing aromatic amino acids like histidine, which do not originate from glycolysis or the TCA cycle, precise regulation is crucial (Akashi and Gojobori, 2002; Winkler and Ramos-Montañez, 2009; Kaleta et al., 2013). Typically, feedback inhibition by AMP, ADP, and histidine links the initial catalytic histidine biosynthesis step involving ATP-phosphoribosyltransferase (HisG; A0A4P6LXR4) and its regulatory subunit (HisZ; A0A7G5MP39) to cellular energy levels (Winkler and Ramos-Montañez, 2009). The expensive

synthesis of histidine, which consumes other amino acids and metabolites, can be disadvantageous during acid stress conditions. Therefore, it can be hypothesized that *B. producta* also possesses a histidine uptake system to assimilate extracellular histidine, to avoid the high energy cost associated with amino acid biosynthesis. Although histidine uptake in *B. producta* has not been studied to my knowledge, studies in other Gram-positive bacteria suggest the adaptation of common ABC transporters for histidine uptake (Vitreschak et al., 2008). To experimentally validate the importance of histidine biosynthesis or the uptake of histidine in the acid stress response of *B. producta*, several follow-up experiments can be performed. For instance, transposon sequencing can be used to compare the bacterial fitness of wild-type *B. producta* against a single-gene disruption library grown in media with and without histidine supplementation under acid stress. This approach can identify genes required or detrimental for the growth of *B. producta* under acid stress conditions. While single knock-out studies in *E. coli* suggested that most amino acid transport and metabolism genes, including histidine biosynthetic genes, are non-essential (Baba et al., 2006), transposon sequencing studies in *S. aureus* revealed a crucial gene (SAUSA300\_0846) encoding a histidine transporter (Beetham et al., 2024). The histidine transporter function was confirmed by measuring the uptake of radio-labeled histidine in both wild-type *S. aureus* and the mutant strain, where histidine uptake was largely lost in the mutant. In the absence of exogenous histidine, the wild-type strain exhibited an additional 5-hour lag phase, indicating its reliance on histidine for growth and its ability to adapt and synthesize histidine. Growth experiments conducted at pH 4.3 and pH 7.2, with or without histidine, confirmed the importance of histidine transport via the SAUSA300\_0846 transporter for *S. aureus* growth at acidic pH. Notably, the mutant strain exhibited a more than 200-fold increase in the expression of histidine biosynthesis genes under acid stress conditions.

Currently, to the best of my knowledge, their study stands as the pioneering exploration of a potential histidine-dependent acid tolerance system. Their meticulous validation of the potential involvement of histidine in the acid resistance of *S. aureus* sets an excellent example and should serve as a benchmark for future research into the potential histidine-dependent acid tolerance system of *B. producta*.

**Top-down Proteomic Analysis** – In addition to the conventional proteome analysis using a BUP approach, a TDP approach was employed to characterize and quantify individual proteoforms. To enhance the detection of the low-molecular-weight proteome, two depletion methods and LC-FAIMS-MS<sup>2</sup> analysis with internal CV stepping (Kaulich et al., 2022a) were utilized to boost sensitivity and increase the number of proteoforms. Recent studies have shown that utilizing FAIMS with internal CV stepping not only doubles the number of quantified proteoforms but also maintains quantification accuracy (Kline et al., 2023). While a newer version of Proteome Discoverer (v.3.0) was used by Kline and colleagues, allowing for the

specification of multiple CVs per file in the spectrum selector node, database searches in this study were conducted using Proteome Discoverer (v.2.5.0.400). This version does not support multiple CVs per file. Consequently, the resulting data files had to be sliced into subsets filtered by the applied CVs before database searching (Leipert et al., 2023).

The TDP analysis of acidic and basic HMWP depletions of *B. producta* quantified 175 proteoforms (19% of the 923 identified proteoforms) and 135 proteoforms (17% of the 818 identified proteoforms), respectively (FIGURE IV-11A and B). The low number of quantified proteoforms can be attributed to the significant number of missing values observed (38% and 42%), which is comparable to other TDP-LFQ studies that reported 43% (Leipert et al., 2023) or 54.6% (Ntai et al., 2016). Interestingly, the number of missing values can depend on the applied CVs, which ranged in the proteoform analysis of *Caenorhabditis elegans* from 23% (CV -20) to 43% (CV -50) (Leipert et al., 2023). Notably, Kline and colleagues quantified 499 *E. coli* proteoforms, which represents only 29% of the 1,719 identified proteoforms, potentially indicating a similar occurrence of missing values in their analysis (Kline et al., 2023). The higher number of missing values in a proteoform-centric analysis can be attributed to broader isotopic envelopes and wider charge state ranges compared to a peptide-centric analysis (Basharat et al., 2023). Consequently, co-eluting proteoforms of different charge states can share overlapping  $m/z$  ranges even when they have highly distinct masses. This phenomenon can interfere with resolving and accurately quantifying proteoforms by spectral deconvolution algorithms, such as Xtract deconvolution employed by the ProSightPD node in Proteome Discoverer. While multi-dimensional protein fractionation strategies can mitigate proteoform co-elution (Cassidy et al., 2021a; Kaulich et al., 2024), the increased complexity can pose challenges for LFQ analysis in correctly determining proteoform abundances across numerous fractions. Therefore, future advancements in deconvolution algorithms are essential, particularly those that possess the capability to resolve and quantify co-eluting proteoforms or acquire precursors with minimal interference. Such advancements are crucial for mitigating data loss and ensuring the comprehensive characterization of proteomic samples, ultimately advancing TDP-LFQ analysis (Jeong et al., 2020, 2022; Basharat et al., 2023).

Although the TDP analysis primarily targeted the LMWP (<30 kDa), it detected differential abundance of 38 proteins (TABLE IV-1). Specifically, 21 proteins exhibited differential abundance in the acidic depletion analysis and 24 in the basic depletion analysis (FIGURE IV-11B). Six proteins showed differential abundance in both depletion methods (TABLE A-7). Importantly, both depletion methods allowed comparable quantification of abundance changes for identical proteoforms of the same protein (TABLE A-7).

A comparison of the quantification results of BUP and TDP revealed the non-differential abundance of 67 proteins and the differential abundance of 13 proteins (TABLE IV-1). Among these 13 proteins, three (GntR; A0A7G5MZW5, NlpC; A0A7G5N0P5, and DltD; A0A7G5MQG2) exhibited differential abundance in both TDP depletion analyses and the BUP

analysis (FIGURE IV-12 and TABLE IV-2). Notably, at pH 6.0, the D-alanyl-lipoteichoic acid biosynthesis protein (DltD; A0A7G5MQG2), responsible for catalyzing the D-alanylation of lipoteichoic acid, exhibited increased abundance. D-alanylation of teichoic acids can mask the negative charge of the cell membrane, enhancing growth and survival in low pH environments (Boyd et al., 2000; Wu et al., 2022).

While BUP quantification aims to measure the abundance of peptides derived from a given protein, it often fails to capture proteoform variations due to limitations in peptide-to-protein inference. Estimations by Ntai and colleagues suggest that around 40% of proteoform-level dynamics in abundant, low-molecular-weight (<30 kDa) proteins remain undetected by BUP (Ntai et al., 2016). Despite these limitations, both peptide- and proteoform-level quantification of GntR indicate a higher abundance at pH 7.0 in the C-terminal protein region (FIGURE IV-13). Similarly, BUP analysis of the DNA-binding protein HU (Hup; A0A2S4GGS2) indicated higher abundance of several peptides from the N-terminal region at pH 6.0, potentially crucial for DNA protection under acidic stress (Almarza et al., 2015). However, not all peptides were quantified, and some also showed increased abundance at pH 7.0. In contrast, proteoform analysis revealed a differentially increase in abundance of DNA-binding protein HU proteoforms spanning the N-terminal region at pH 6.0 (TABLE A-6). This further emphasizes the importance of integrating TDP analyses to identify potential biologically proteoform abundances resulting from variant expression, proteolytic truncation, or changes in PTM stoichiometry.

Interestingly, many of the differentially abundant proteoforms from the same protein had abundance changes primarily characterized by truncation, rather than other PTMs (TABLE A-6). However, the protein sequence database of *B. producta* used for this analysis contained only a few automatically UniProt-annotated PTMs, with most proteins lacking information on potential or validated modifications. Consequently, during the Proteome Discoverer database search, proteoforms with potential PTMs that were not annotated in the UniProt database remained elusive. Therefore, an additional discovery-open modification search was utilized to identify potential PTM-carrying proteoforms. This approach successfully identified and quantified multiple seryl-phosphorylations on HPr proteins (FIGURE IV-18 and FIGURE IV-19). Although these findings require further validation, the application of a discovery-open modification search may be considered to complement future analysis, especially for bacterial species with limited UniProt-annotated PTMs.

Overall, the combination of BUP and TDP analysis identified potential seryl-phosphorylations on HPr (A0A2S4GRU0, A0A7G5MYS7, and A0A4V0Z7D2), along with one arginyl-phosphorylation on HPr (A0A7G5MP53) (FIGURE IV-18). The presence of phosphopeptides with missed cleavages near the proteolytic site (P1') (FIGURE A-19E and G) could potentially strengthen the identification of phosphorylation sites, as nearby phosphorylation sites (P1', P2', and P3') can interfere with tryptic cleavage (Dickhut et al., 2014; Gershon, 2014). Although the arginyl-phosphorylation was detected with 68 PrSMs, the absence of fragment ions around

the modification site, the lack of detected phosphopeptides, and the limited literature on arginyl-phosphorylation in HPr proteins make its PTM localization uncertain with the currently available data.

Moreover, sequence analysis of *B. producta*'s HPr proteins revealed notable differences from other members of this protein family. While typical HPr proteins feature two highly conserved amino acid residues, a histidine (His-15) near the N-terminus and a serine (Ser-46) in the central part of the protein, serving as phosphoryl group acceptors (Meadow et al., 1990; Brochu and Vadeboncoeur, 1999; Casabon et al., 2006), the majority of *B. producta*'s HPr proteins lack five amino acids near the N-terminus, including the conserved His-15. This deficiency potentially makes them incapable of forming doubly phosphorylated HPr(His~P)(Ser~P) (FIGURE IV-18).

Typically, HPr phosphocarrier proteins are involved in carbohydrate phosphorylation during transport into bacterial cells via the phosphotransferase system (Meadow et al., 1990). Depending on their phosphorylation state, HPr proteins can also act as coregulators of the catabolite global regulator CcpA (Homeyer et al., 2007), which controls the catabolite repression/activation of up to 10% of total genes (Poncet et al., 2004). This can provide a direct link to the metabolic state of the bacterial cell and a regulatory role in the quorum sensing of the cell (Ha et al., 2018). The combined quantitative data of the BUP and the TDP analysis showed that the HPr protein (A0A2S4GRU0), their phosphorylated peptides as well as their phosphorylated proteoforms were differentially more abundant at pH 6.0 (FIGURE IV-19). In contrast, unmodified peptides covering the same protein sequence and unmodified or formylated proteoforms were higher abundant at pH 7.0 (FIGURE IV-19). Notably, this observation was independent of the applied depletion method for the TDP analysis (FIGURE IV-19C-D). The observed increase in HPr abundance and seryl-phosphorylation under acidic growth conditions aligns with prior research demonstrating a similar trend of increased HPr protein abundance and HPr(Ser-P) formation in response to culture acidification (Casabon et al., 2006; Heunis et al., 2014). Hence, the results of this study further emphasize the significant influence of culture pH on the phosphorylation status of HPr proteins.

#### 4.4 Asp-Pro peptide bond hydrolysis

Notably, peptides with N- or C-terminal Asp-Pro bond cleavages were detected for all three bacteria (FIGURE IV-15C-D). Specifically, for *B. thetaiotaomicron*, there was an increased relative frequency of Asp at the P1 position and Pro at the P1' position for peptides identified at pH 6.0 (FIGURE IV-15B), suggesting that acidic conditions may promote the hydrolysis of Asp-Pro bonds. This potential biological response is further supported by a significant increase in peptides exhibiting N- or C-terminal Asp-Pro and Asp-Xaa cleavages for *B. thetaiotaomicron* (FIGURE IV-16A) and *B. producta* (FIGURE IV-16B). In contrast, the proteome of *B. longum* did

not show an Asp-Pro sequence logo (FIGURE IV-15A-B) and neither significant changes in the abundance of peptides with N- or C-terminal Asp-Pro cleavages (FIGURE IV-16C).

Several factors, such as the peptide sequence, protein folding, temperature, and pH value, can affect the hydrolysis of peptide bonds (Marcus, 1985; Li et al., 2009). The Asp-Pro bond, in particular, exhibits increased susceptibility to peptide bond hydrolysis under acidic conditions and elevated temperatures (Marcus, 1985; Li et al., 2009). The proposed acidolysis of Asp-Pro bonds suggests that the  $\beta$ -carboxyl group of Asp initiates a nucleophilic attack on the carbonyl carbon of the amide bond, facilitating intraresidue cyclization, forming an unstable cyclic anhydride intermediate, which may cause a break in the polypeptide chain (Piszkiewicz et al., 1970). This reaction requires the presence of an adjacent protonated amide nitrogen. Generally, the enhanced rate of Asp-Pro cleavage can be attributed to the greater basicity of the nitrogen atom as part of proline's cyclic structure, which increases its basicity ( $pK_a$  of 10.6) compared to the primary amine groups of other amino acids ( $pK_a$  of 8.7 to 9.9), thus increasing its nucleophilicity and facilitating faster hydrolysis of Asp-Pro bonds under acidic conditions (Piszkiewicz et al., 1970).

The importance of such a labile peptide bond can be crucial for planning and performing proteomic sample preparation and subsequent LC-MS measurement, especially for N- and C-terminomics experiments. For example, acidification using pure TFA for cell lysis and high concentrations of Tris for neutralization such as applied in the SPEED protocol, should be avoided (Doellinger et al., 2020). Furthermore, Tris-based buffers exhibit significant pH changes upon temperature change compared to other buffer systems such as sodium phosphate buffer solutions, which are generally more resistant to acidification upon temperature change (Kolhe et al., 2010). The choice of reducing agent can also be critical, as TCEP can cause a significant drop in pH compared to dithiothreitol (DTT) (Scheerlinck et al., 2015). Additionally, different buffers are required depending on the protease used for digestion. For example, pepsin or neoprosin necessitates highly acidic conditions (pH 1.5 to pH 2.5) for optimal enzymatic activity, a condition that has been demonstrated to significantly increase the hydrolysis of Asp-Pro bonds (Schröder et al., 2017). Moreover, most in-solution workflows involve SPE for sample clean-up, typically involving organic solvents with TFA or FA, which can also increase Asp-Pro bond cleavage (Winkels et al., 2022). Generally, prolonged exposure to acidic conditions or heating should be avoided (Kaulich et al., 2024), and gentler methods such as freeze-drying or desiccation to evaporate organic solvents after SPE are preferable. Alternatively, different sample and clean-up protocols, such as SP3 (Hughes et al., 2019) or FASP (Manza et al., 2005; Wiśniewski et al., 2009), can be applied, which often do not require an additional SPE step. Typically, proteomic LC-MS analysis usually employs an acidic mobile phase and a thermostat-controlled column oven with elevated temperatures for optimal chromatographic separation, controlled retention times, and reduced pressure of the LC system (García, 2005; Lenčo et al., 2022). However, these conditions, in combination with

prolonged in-column residence time, can induce in-column peptide hydrolysis of Asp-Pro bonds (Lenčo et al., 2021).

Despite the significant enrichment and abundance of peptides with Asp-Pro and Asp-Xaa cleavage at pH 6.0 for *B. thetaiotaomicron* and *B. producta*, the possibility of artificial peptide bond hydrolysis during sample preparation and LC-MS analysis cannot be ruled out. While employing the same sample preparation steps and LC-MS setup for the full proteome analysis of the three HGM bacteria, slight variations, such as the total sample volume after SPE and extended vacuum centrifuge evaporation times after SPE sample clean-up, may have occurred. Additionally, the potential for the observed results to be generated *in vivo*, reflecting a biological effect of acidic cultivation, cannot be disregarded. Further experiments using a proteomic sample preparation workflow designed to limit the artificial generation of Asp-Pro cleavage in combination with an N-terminomics approach may clarify the origin of these peptide formations (Winkels et al., 2022).

While Asp-Pro cleavage is reported to be artificial in most studies, selective cleavage of Asp-Pro bonds has been reported to be required for correct folding (Patrick and Egland, 2019), intermolecular protein cross-linking forming an Asp-Lys isopeptide bond (Osička et al., 2004) or covalently linking chondroitin sulfate forming a protein-glycosaminoglycan-protein complex (Zhuo et al., 2004). Moreover, several studies have reported *in vivo* cell-compartment and pH-specific autocatalytic cleavage between Asp-Pro bonds of several proteins exhibiting a Gly-Asp-Pro-His (GDPH) sequence, which is commonly found in von Willebrand factor type D domains. For example, autocatalytic cleavage at body temperature has been reported for several human proteins before being secreted into the intestinal mucus. This includes GDPH cleavage of IgGFC-binding protein (FCGBP) or MUC2 mucin, which is triggered by the lower pH of the endoplasmic reticulum or the Golgi apparatus, respectively (Lidell et al., 2003; Ehrencrona et al., 2021). Similarly, autocatalytic GDPH cleavage of proH3 precursors occurs during passage through the low pH of the Golgi complex (Thuveson and Fries, 2000). While the exact function of GDPH cleavage is still a topic of ongoing investigations, it may be an important factor for covalent protein or carbohydrate cross-linking, especially in the mucus (Lidell et al., 2003). For instance, alterations in the interactome of proteins within the arterial extracellular matrix, potentially contributing to protein aggregation cascades, are suggested by results from the enrichment of Asp-Pro cleavage products of NOTCH3, associated with disease-affected brain tissue (Lee et al., 2023).

The results of this study indicate a potential acidic-induced Asp-Pro cleavage of 30S ribosomal protein S16 (A0A4P6M2Y5; FIGURE IV-14), with C- and N-terminal fragments exhibiting potential AMP activities (TABLE A-9). Although preliminary AMP testing against members of the HGM did not confirm such activity, further experiments with other members of the gut microbiota may be necessary to determine the presence of potential AMP activity.

Similar to the proteoform concept, which suggests that different protein species obtained through processes like truncation or other post-translational modifications can have distinct functions (Jungblut et al., 2016), several ribosomal proteins have been identified with "moonlighting" functions, wherein a single protein resulting from one gene serves multiple roles in a cell or organism, including exhibiting antimicrobial activities (Hurtado-Rios et al., 2022). For example, *Bacillus tequilensis* isolated from healthy human feces can secrete ribosomal protein L1 as an antimicrobial molecule (Ghoreishi et al., 2023), while *Lactobacillus salivarius*, isolated from the feces of four-month-old human infants, secretes ribosomal proteins L27 and L30 with antimicrobial activities (Pidutti et al., 2018). Notably, several antimicrobial peptides from ribosomal proteins L30, L39, S19, and S30 have been identified in cellular extracts of human non-inflamed colonic mucosa, indicating a diverse presence of ribosomal antimicrobial peptides in human colonic mucus (Howell et al., 2003; Tollin et al., 2003; Antoni et al., 2013). Whether *B. producta* ribosomal protein S16 may function as a moonlighting protein with antimicrobial activity remains to be elucidated.



## **V    PROTEOGENOMIC ANALYSIS OF *B. PRODUCTA***

---

<b>1</b>	<b>Introduction and Summary .....</b>	<b>105</b>
<b>2</b>	<b>Experimental Design .....</b>	<b>107</b>
	2.1 Culture Condition Effects on SEP Production .....	107
	2.2 Peptide and Proteoform Validation.....	108
<b>3</b>	<b>Results.....</b>	<b>110</b>
	3.1 Identification and Validation of Translated SEP .....	110
	3.2 SEP Proteoform Diversity.....	113
	3.3 Impact of Cultivation Conditions on SEP Translation.....	115
	3.4 Biochemical Predictions of Identified SEP .....	116
<b>4</b>	<b>Discussion and Conclusion.....</b>	<b>117</b>
	4.1 Challenges and Advances in SEP Identification .....	117
	4.2 Future Directions of SEP Research .....	119

Parts of the following chapter have been published in “Identification of proteoforms of short open reading frame-encoded peptides in *Blautia producta* under different cultivation conditions.” Genth et al., Microbiology Spectrum, 11(6), e0252823, (2023).

**Supplementary material for (Genth et al., 2023)**

Additional supplementary information's is freely available for download at the publisher's website <https://doi.org/10.1128/spectrum.02528-23>. The MS proteomics raw data and complete Proteome Discover search results have been deposited to the ProteomeXchange Consortium (<http://www.proteomexchange.org/>) via the PRIDE (Vizcaíno et al., 2014) partner repository with the data set identifier PXD041979.

## 1 Introduction and Summary

MS-based proteomics is a powerful tool for verifying the existence and translation of proteins within the proteome. However, the complexity of proteomes requires careful experimental design and stringent data evaluation, especially for detecting and characterizing low-abundance small proteins. These small proteins, which typically comprise 100 or fewer amino acids (Slavoff et al., 2013), are increasingly recognized as significant cellular components, even originating from genomic regions conventionally categorized as noncoding. This missing part of the proteome is commonly referred to as the "ghost proteome." (Cardon et al., 2021).

Recent advancements in genome sequencing and open reading frame (ORF) prediction algorithms have revealed an increasing number of unannotated protein-coding short open reading frames (sORFs) within bacterial genomes (Sberro et al., 2019). Direct detection of their translation products, commonly referred to as sORF-encoded peptides (SEP) (Scheidler et al., 2019; Cassidy et al., 2021b; Schlesinger and Elsässer, 2022), has revealed a diverse array of cellular functions, including glucose uptake, cell division, peptidoglycan synthesis, stress responses, virulence, and sporulation (Storz et al., 2014; Khitun et al., 2019; Yadavalli and Yuan, 2022). These peptides or proteins are also termed sProteins (Miravet-Verde et al., 2019; Petruschke et al., 2021) or microproteins, especially when they harbor functional domains (Andrews and Rothnagel, 2014).

To capture the protein-coding potential of prokaryotic genomes comprehensively, specialized databases have been developed. These databases employ various approaches, such as *in silico* six-frame translations of the entire genome sequence (Castellana and Bafna, 2010; Krug et al., 2013), translation of RNA sequencing (RNA-Seq) data into possible reading frames (Slavoff et al., 2013), or an integrated approach that consolidates annotations and predictions from diverse sources (Omasits et al., 2017). An exemplary integrated approach, represented by the "integrated proteogenomics search database" (iPtgxDB) (Omasits et al., 2017), takes into account alternative start sites and strategically reduces redundancy. Consequently, iPtgxDB has gained remarkable popularity in recent years (Bartel et al., 2020; Varadarajan et al., 2020; Petruschke et al., 2021; Hadjeras et al., 2023).

With the growing interest in the human gut microbiome, there has been a notable focus on identifying SEP. Through the utilization of a simplified model system of the human gut microbiome (SIHUMIx) (Krause et al., 2020) and various depletion/enrichment protein extraction methods, several SEP have been successfully discovered (Petruschke et al., 2021). Among these newly identified SEP, some seem to be uniquely synthesized within the SIHUMIx community. This discovery includes several previously uncharacterized SEP originating from *B. producta*, namely BP3, BP5, BP8, BP11, and BP12. However, the authors also noted the influence of varying growth and cultivation conditions on protein identification, emphasizing the need for further research to fully understand their function and relevance in the microbiome.

**Aim of this study:**

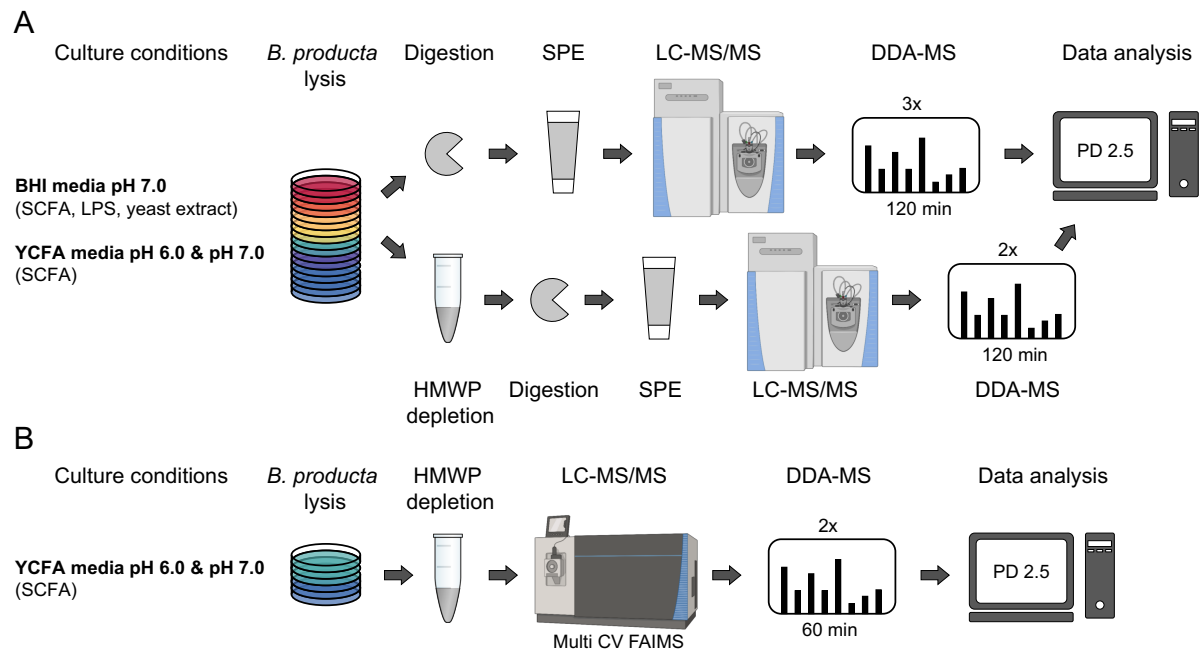
This study is driven by the observation of SEP in single cultures of *B. producta*, seeking to validate the presence of both previously reported and novel SEP.

- Evaluate the presence of SEP in *B. producta* under various culture conditions, including variations of media (BHI and YCFA), pH levels (pH 7.0 and pH 6.0), and supplemented factors (yeast extract, SCFAs, and LPS).
- Conduct bottom-up proteomic analyses, including full-proteome analysis and depletion of the high-molecular-weight proteome, to enhance coverage of the low-molecular-weight proteome.
- Employ a proteoform-directed analysis utilizing top-down proteomics to identify post-translational modifications, detect potential cleavage events, and explore alternative initiation.
- Apply stringent validation criteria to ensure the accurate identification of sORF-encoded peptides and proteoforms.
- Utilize biochemical predictions of the identified SEP to explore their potential functional roles.

## 2 Experimental Design

### 2.1 Culture Condition Effects on SEP Production

To assess the impact of culture conditions on SEP production, *B. producta* was cultured under seven different culture conditions (chapter II.2.1). These conditions included variations in culture media (BHI and YCFA), pH levels (pH 7.0 and pH 6.0), and supplemented factors (yeast extract, SCFAs, and LPS). At mid-exponential growth, cells were lysed using freeze-thawing, and proteins were cleaned up using ethanol precipitation (chapter II.3.1). Samples were subjected to full-proteome analysis and acidic and basic high-molecular-weight proteome (HMWP) depletion to increase the coverage of the low-molecular-weight proteome (LMWP) (Cassidy et al., 2019). Samples were digested using trypsin at a 1:40 enzyme-to-substrate ratio and subjected to solid-phase extraction (chapter II.3.4). Each culture condition was analyzed with two biological replicates for both bottom-up full-proteome and LMWP analyses, except for YCFA pH 6.0 conditions, which were exclusively analyzed using five biological replicates for bottom-up full-proteome analysis (FIGURE V-1A). Bottom-up proteomic samples were separated online using reversed-phase chromatography with a gradient of 120 minutes and measured on the Q-Exactive Plus mass spectrometer (chapter II.4.1). Full-proteome analysis included three technical replicates, while LMWP samples were analyzed in two technical replicates (FIGURE V-1A).



**FIGURE V-1 | Experimental Design for Assessing the Impact of Culture Conditions on SEP Production.** *B. producta* was cultured under various conditions including variations in culture media (BHI and YCFA), pH levels (pH 7.0 and pH 6.0), and supplemented factors (yeast extract, SCFAs, and LPS). **(A)** Bottom-up proteomic analysis of the full-proteome or the low-molecular-weight proteome. **(B)** Top-down proteomic analysis of the low-molecular-weight proteome. Acidic and basic depletion of the high-molecular-weight proteome was performed accordingly to (Cassidy et al., 2019).

Additionally, LMWP samples from cultivation in YCFA media at pH 6.0 and pH 7.0 were analyzed by top-down proteomics (FIGURE V-1B). These samples were separated online using reversed-phase chromatography with a gradient of 60 minutes and measured in two technical replicates on the Fusion Lumos Tribrid mass spectrometer (chapter II.4.2).

The acquired raw data were searched and compared to the iPTgxDB of *B. producta* (Omasits et al., 2017). Bottom-up proteomics searches were conducted using Proteome Discoverer 2.5, allowing up to four missed cleavages with semi-tryptic and full-tryptic specificity (Bartel et al., 2020). To enhance the reliability of SEP identification, each peptide had to be uniquely identified within its respective MS run. Top-down proteomic searches were performed using Proteome Discoverer 2.5.0.400 (chapter II.5.1).

## 2.2 Peptide and Proteoform Validation

To ensure the reliability and accuracy of identifications, several established criteria were employed, guiding a multi-stage validation process to verify the presence of SEP. The initial step involved applying a size filter ( $\leq 100$  amino acids) (Bartel et al., 2020), which reduced the dataset to specifically target small proteins for subsequent validation. For BUP and TDP, SEP identifications generated by Prodigal and ChemGenome predictions required a minimum of at least three PSMs/PrSMs. Conversely, *in silico* ORF predictions demanded a more stringent threshold of at least four PSMs/PrSMs (Nesvizhskii, 2014).

**Bottom-up Peptide Validation** – Putative non-canonical peptides were required to have a minimum sequence tag consisting of five consecutive b- or y-ions in the MS<sup>2</sup> spectrum (Slavoff et al., 2013), which were confirmed by manual inspection of the respective spectra (Khitun and Slavoff, 2019). In addition, the peptide-centric search engine PepQuery was used to verify the quality of the PSMs and determine whether the matched sequences were distinctive or likely associated with peptides resulting from hypothetical PTMs of the referenced proteome (Wen et al., 2019). The retrieved MS/MS spectra were converted to MGF format using msconvert from ProteoWizard (Adusumilli and Mallick, 2017). The standalone version of Pepquery (v. 2.0.2) was then executed with unrestricted searching for modifications and substitutions of amino acids using the following command line:

```
java -jar pepquery-2.0.2.jar -db [fasta file] -ms [mgf files] -i [peptid list] -aa -hc TRUE -c 4 -maxVar 4 -itol 0.02 -o [output directory]
```

Validated peptides meeting the statistical threshold (p-value  $< 0.01$ ) were filtered and subsequently analyzed using Proteomics Data Viewer (Li et al., 2019b). The validated PepQuery peptides were compared to the *B. producta* proteome (taxid:33035) and the RefSeq database for bacterial species (taxid:2) using NCBI Protein Blast (BLASTp) to exclude peptides

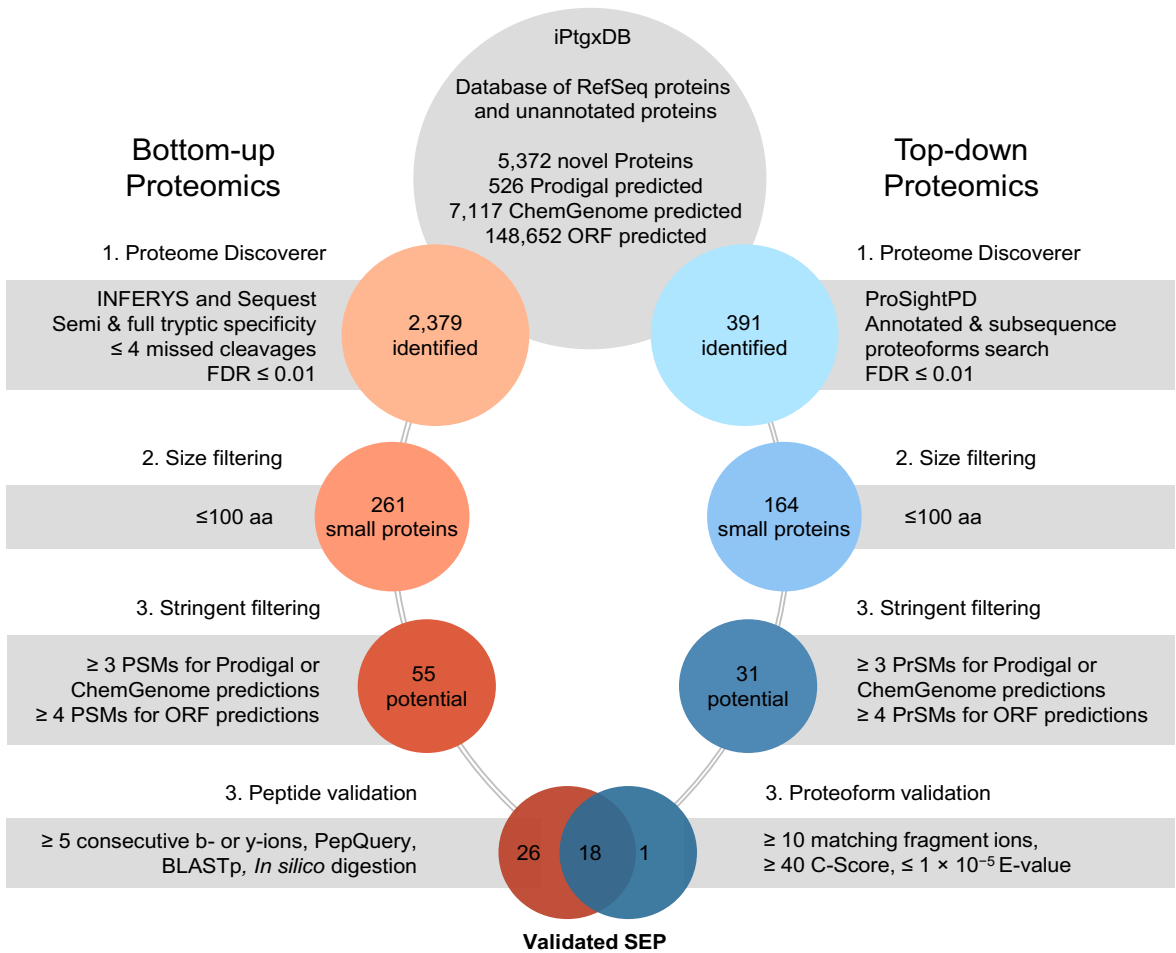
with single amino acid variations. Alignment with the RefSeq database required a minimum sequence identity of 50% and query coverage of 50%, while searching the *B. producta* proteome involved a minimum sequence identity of 90% and complete query coverage (100%). In cases of multiple homologs, the nearest relative match to the query strain was determined with an E-value threshold of  $1 \times 10^{-5}$ . To reduce potential false positives from protein contaminants in the yeast extract or BHI media, an *in silico* digest of the *Saccharomyces cerevisiae*, porcine, and bovine proteomes was performed using MS-Digest (<https://prospector.ucsf.edu/prospector>). The *in silico* digest, allowing for two missed cleavage sites and a minimum peptide length of 7 amino acids, was then compared to the iPtgxDB *in silico* digest results.

**Top-down Proteoform Validation** – SEP proteoforms are required to have a minimum C-Score of 40 (Leduc et al., 2014), along with an E-value smaller than  $1 \times 10^{-5}$  and at least ten matching fragment ions with a mass error of less than 10 ppm (Liu et al., 2013; Sun et al., 2013).

### 3 Results

#### 3.1 Identification and Validation of Translated SEP

The rigorous filtering process, which involved assessing the number of PSMs and validating peptides for the BUP approach, as well as evaluating the number of PrSMs and validating proteoforms for the TDP approach, aimed to ensure accurate identification of SEP in both methodologies. The combination of BUP and TDP identified a total of 45 SEP (FIGURE V-2). These included BP1 to BP14, which had previously been identified (Petruschke et al., 2021). Following the nomenclature introduced by Petruschke and colleagues (Petruschke et al., 2021), 30 novel SEP (BP16-BP46) were identified and reported for the first time at the protein level. Notably, BP46 was exclusively identified using the TDP approach (FIGURE V-2).



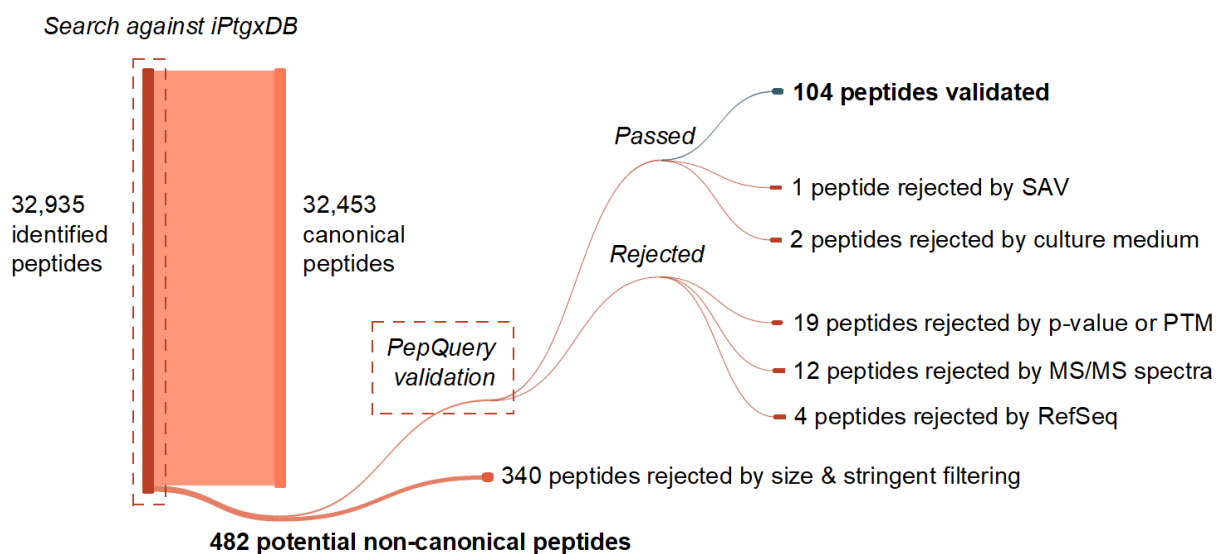
**FIGURE V-2 | Flowchart of the Multistep SEP Verification Process.** Overview of the applied proteogenomic workflow including the use of both bottom-up and top-down data that were searched and compared to the iPtgxDB of *B. producta* to identify novel SEP. After applying a size filter to identify small proteins, stringent filtering, and peptide and proteoform validations, a total of 45 SEP were identified.

**Peptide Validation** – The applied BUP approach identified 2,379 proteins across all seven cultivation conditions, of which 261 proteins (9.1%) had less than 100 amino acids (FIGURE V-2). Among these, 55 proteins passed the PSM filtering criteria, represented by 142 peptides, which were subsequently subjected to the verification process (FIGURE V-2). The validation of these non-canonical peptides using PepQuery led to 107 peptides passing the strict verification criteria (FIGURE V-3). Out of the 35 peptides that did not pass the PepQuery verification process, 12 were not matched by PepQuery to any MS/MS spectra with sufficient quality scores, 4 had superior matches to peptides in the reference database, and 19 were either better matched with reference peptides carrying potential PTMs or did not meet statistical thresholds ( $p$ -value  $< 0.01$ ) matching the non-canonical sequence (FIGURE V-3).

Subsequent BLASTp analysis of the remaining peptides that successfully passed PepQuery identified a single amino acid variation (SAV) in one peptide compared to a RefSeq protein. Considering the potential occurrence of single nucleotide polymorphisms in genetic sequences, which can arise through various mechanisms such as DNA replication errors, there was an increased likelihood that the detected sequence variant may not necessarily be linked to a novel SEP. To ensure the accuracy and reliability of the results, both the peptide and its corresponding SEP were consequently excluded as a precautionary measure.

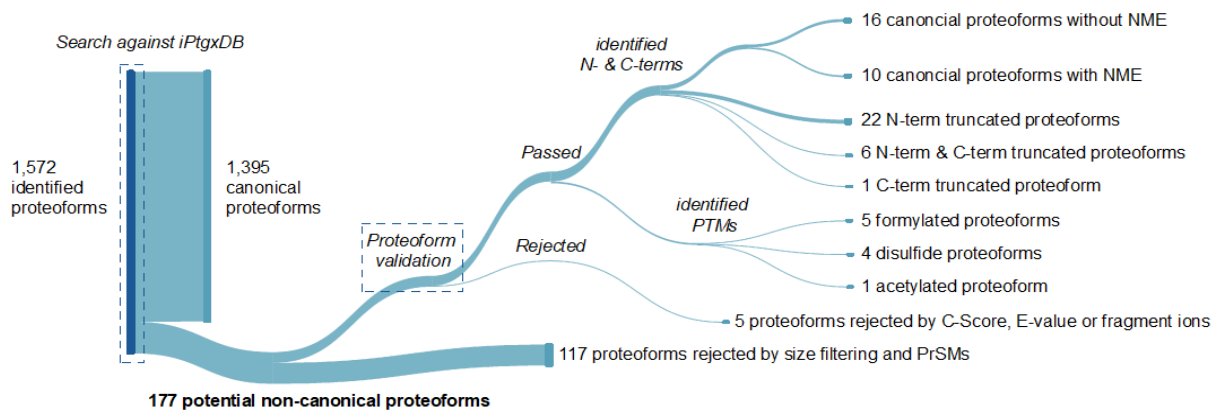
Comparison with different *in silico* digestions, even when treating leucine and isoleucine as equivalent in peptide sequence analysis (FIGURE A-20), led to the exclusion of two additional peptides, but not their respective SEP. Therefore, potential peptide contamination due to the use of protein-containing yeast extract or BHI could be largely excluded.

In conclusion, 104 peptides were validated (FIGURE V-3), which led to the identification of 44 SEP (FIGURE V-2).



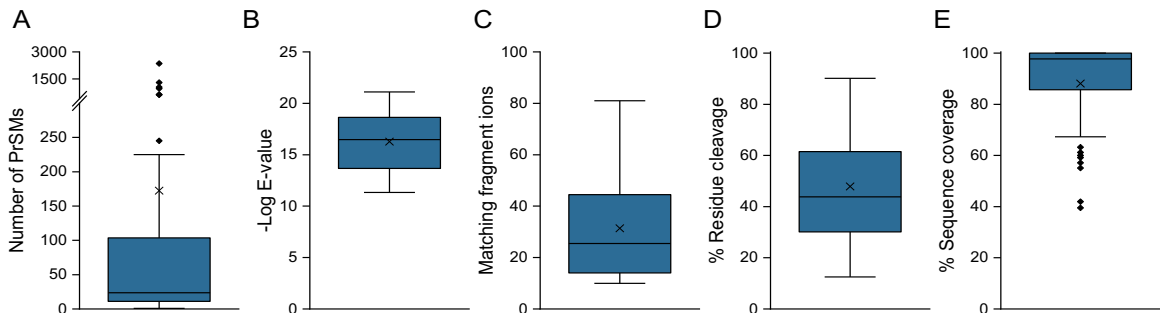
**FIGURE V-3 | Validation of Non-canonical Peptides.** Overview of stringent filtering and PepQuery peptide validation. Abbreviation: SAV (single amino acid variation).

**Proteoform Validation** – A total of 1,572 proteoforms were identified and subjected to a rigorous filtering and validation process, resulting in the identification of 177 potential non-canonical proteoforms (FIGURE V-4). Among these, 60 proteoforms met the criteria for both protein size and number of PrSMs. Additional filtering, which included specific criteria such as C-Score, E-value, and the number of fragment ions, led to the exclusion of 5 proteoforms that did not meet the established criteria (FIGURE V-4). As a result, a total of 55 validated proteoforms (FIGURE V-4), derived from 19 SEP, successfully passed the filtering and validation process (FIGURE V-2).



**FIGURE V-4 | Validation of Non-canonical Proteoforms.** Stringent filtering and proteoform validation, including the distribution of identified N- and C-termini and post-translational modifications (PTMs) on SEP proteoforms.

The average number of PrSMs identified for each SEP proteoform was 24, with some exhibiting considerably higher numbers, up to 2,360 (FIGURE V-5A). The identified SEP proteoforms had a median -Log E-value of 17 (FIGURE V-5B), 26 fragment ions (FIGURE V-5C), and a residue cleavage rate of 44% (FIGURE V-5D). These results, combined with a median sequence coverage of 98% (FIGURE V-5E), collectively provide strong evidence for a high confidence level associated with the identified proteoforms.

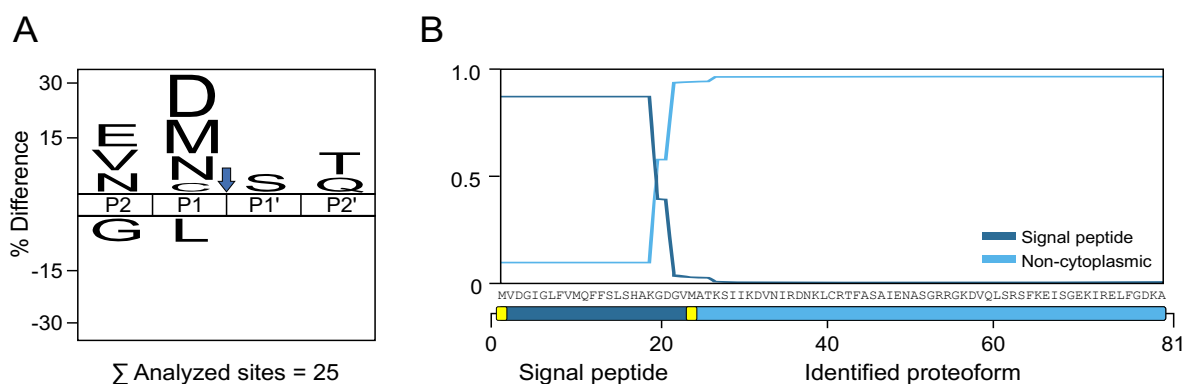


**FIGURE V-5| Metrics of Proteoform Identifications.** (A) Number of PrSMs, (B) -Log E-values, (C) matching fragment ions, (D) residue cleavage, and (E) sequence coverage of proteoform identifications. The box-and-whisker plots illustrate the lower quartile and upper quartile, with the median displayed as a horizontal line and the mean depicted as a cross. Whiskers represent the minimum and maximum values that fall within 1.5 times the interquartile range

### 3.2 SEP Proteoform Diversity

Neo-termini analysis of the identified proteoforms identified 26 canonical proteoforms of which 16 retained their N-terminal methionine and 10 lacked their N-terminal methionine, indicating N-terminal methionine excision (NME) (FIGURE V-4). The remaining proteoforms exhibited either N- or C-terminal truncations, or truncations at both ends (FIGURE V-4). Analysis of the positions surrounding neo-termini sites (P2-P2') indicated an increased specificity for methionine at the P1 position (FIGURE V-6A). While proteoforms resulting from NME were excluded from this analysis, these results may still indicate potential NME by methionine aminopeptidase. The identification of N-terminally methionine-truncated proteoforms, in conjunction with the presence of smaller amino acid residues, such as serine or alanine in the P1' position – known for their increased efficiency in methionine removal (Meinzel et al., 1993) – suggests the possibility of alternative translation initiation.

Such proteoform variants were observed for BP4 and BP7, with N-terminal truncation initiating at methionine position 5, as well as canonical variants with or without N-formylmethionine (FIGURE V-7). Alternative initiation seems likely, given the short N-terminal portion, which is insufficient for a signal peptide (Peng et al., 2019). Similarly, BP14 proteoforms lacked the encoded N-terminal portion of 24 amino acids, including a second methionine at position 24 (FIGURE V-6B). The missing N-terminal portion fell within the typical range of a potential signal peptide (Peng et al., 2019). Phobius analysis (Käll et al., 2004) revealed a potential signal peptide for the missing N-terminal segment (posterior probability: 0.87) and a non-cytoplasmic region for the remaining C-terminal segment (posterior probability: 0.96) (FIGURE V-6B). Whether this represents alternative initiation, signal peptidase cleavage, or another truncation event remains to be elucidated. Understanding these mechanisms and exploring whether these variants play specialized roles in response to specific environmental conditions or stimuli could provide insights into the regulatory processes controlling their expression.



**FIGURE V-6 | Neo-termini Analysis of Identified SEP Proteoforms. (A)** IcelLogo analysis of neo-termini sites (P2-P2'), showing significantly enriched (top) or depleted (bottom) amino acids ( $p \leq 0.05$ ). Proteoforms resulting from N-terminal methionine excision were excluded prior to the analysis. **(B)** BP14 signal peptide prediction by Phobius, highlighting initiator and alternative initiator methionine in yellow.

A

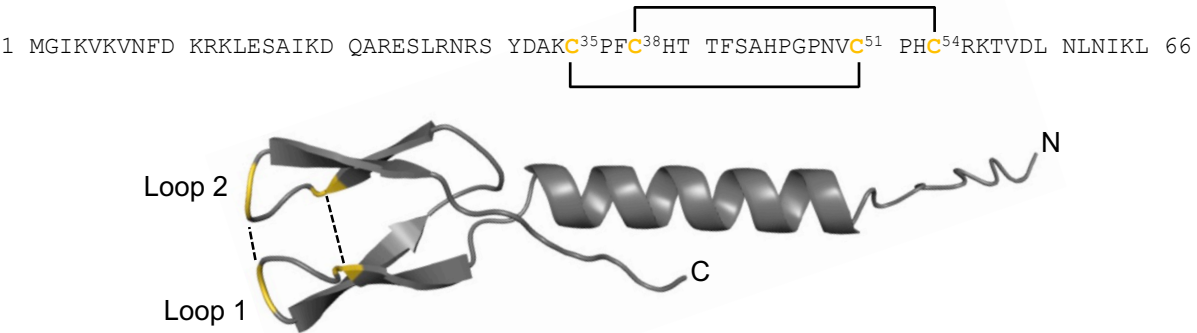
**Canonical** (E-value: 3.7E-21, P-Score: 5.4E-117, Residue cleavage: 87%)  
N M[E]D[N]M[T]D[K]Q[F]K[T]I[L]E[M]F[G]M[I]L[D]G[C]K[D]L[E]E[A]K[K]K[V]E[K]L[L]E[E]Q[K]N[K]S[E]C  
**Formylated canonical** (E-value: 6.5E-20, P-Score: 1.7E-89, Residue cleavage: 67%)  
N M[E]D[N]M[T]D[K]Q[F]K[T]I[L]E[M]F[G]M[I]L[D]G[C]K[D]L[E]E[A]K[K]K[V]E[K]L[L]E[E]Q[K]N[K]S[E]C  
**Alternative initiation - Met<sub>5</sub>** (E-value: 3.3E-20, P-Score: 1.9E-95, Residue cleavage: 71%)  
Missing M[T]D[K]Q[F]K[T]I[L]E[M]F[G]M[I]L[D]G[C]K[D]L[E]E[A]K[K]K[V]E[K]L[L]E[E]Q[K]N[K]S[E]C

B

**Canonical** (E-value: 8.7E-22, P-Score: 2.6E-135, Residue cleavage: 90%)  
N M[N]E[D]M[S]V[F]K[S]Y[L]R[R]L L[Q]D[L]K[D]L[K]E[A]I[K]S[K]E[Y]D[K]A[E]N[M]V[D]K L I[D]D[T]Q[K]G[I]E D[D]C  
**Formylated canonical** (E-value: 1.9E-22, P-Score: 6.4E-81, Residue cleavage: 57%)  
N M[N]E[D]M[S]V[F]K[S]Y[L]R[R]L L[Q]D[L]K[D]L[K]E[A]I[K]S[K]E[Y]D[K]A[E]N[M]V[D]K L I[D]D[T]Q[K]G[I]E D[D]C  
**Alternative initiation - Met<sub>5</sub>** (E-value: 3.2E-21, P-Score: 1.4E-118, Residue cleavage: 81%)  
Missing M[S]V[F]K[S]Y[L]R[R]L L[Q]D[L]K[D]L[K]E[A]I[K]S[K]E[Y]D[K]A[E]N[M]V[D]K L I[D]D[T]Q[K]G[I]E D[D]C

**FIGURE V-7 | Potential Alternative Initiation in SEP Proteoforms.** Proteoforms for (A) BP4 and (B) BP7. For each proteoform, b-ions and y-ions are displayed in blue, while c-ions and z-ions are displayed in red, along with corresponding E-value, P-Score, and residue cleavage information.

Overall, multiple SEP proteoforms were identified, some potentially featuring N-terminal formylation, acetylation, and disulfide bonds (FIGURE V-4). Specifically, proteoforms of BP3, BP4, BP7, BP10, and BP12 exhibited N-terminal formylation, while a proteoform of BP3 showed N-terminal acetylation. In addition to its role in protein synthesis, the N-formyl group may also serve as an indicator for cotranslational membrane insertion, with the N-terminal formyl group frequently retained (Bienvenut et al., 2015). Disulfide bridges were detected for proteoforms of BP12, BP24, and BP46, but the precise assignment of these linkages to specific cysteine residues in BP12 was not feasible with the available data (FIGURE A-21A-B). Secondary structure prediction using Alphafold revealed the presence of  $\alpha$ -helix and  $\beta$ -sheet formations, as well as potential disulfide bonds between Cys<sub>35</sub>-Cys<sub>51</sub> and Cys<sub>38</sub>-Cys<sub>54</sub> for BP12 (FIGURE V-8 and FIGURE A-21C-E).



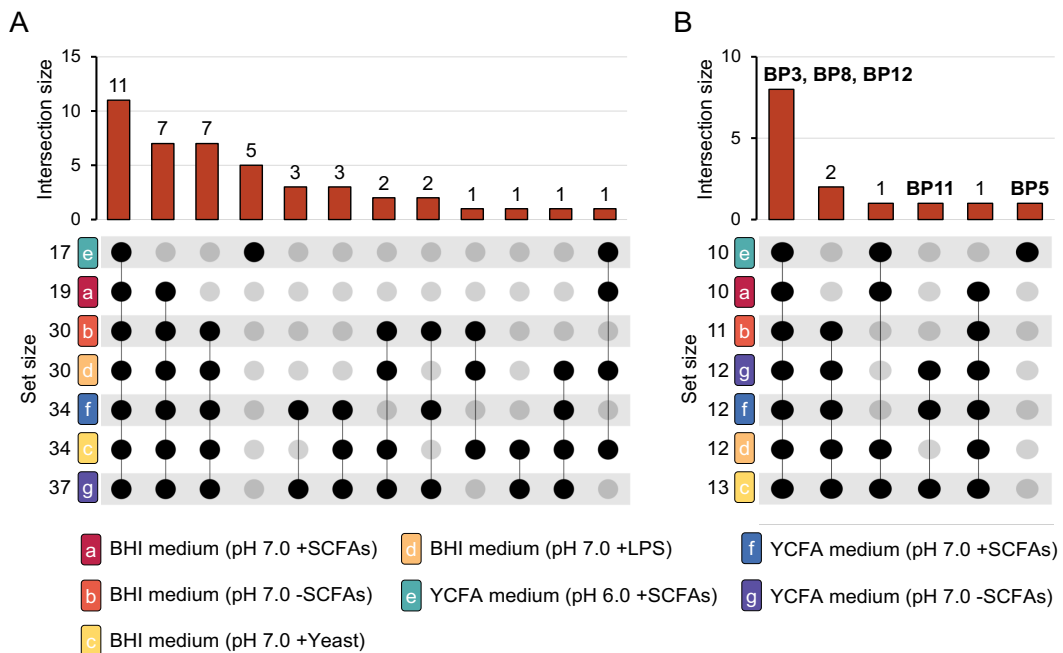
**FIGURE V-8 | Alphafold Structure Prediction for BP12.** The predicted disulfide bridges between Cys<sub>35</sub>-Cys<sub>51</sub> and Cys<sub>38</sub>-Cys<sub>54</sub> are indicated by lines connecting the orange-colored cysteine residues.

### 3.3 Impact of Cultivation Conditions on SEP Translation

Among the 44 SEP identified by BUP, 11 were consistently detectable across all seven growth conditions (FIGURE V-9A). Five of these SEP were exclusively identified under acidic growth conditions, while seven SEPs were consistently present across all conditions except for the acidic YCFA condition (pH 6.0 with the presence of SCFAs). Additionally, specific SEP were exclusively identified in the presence of particular components, such as yeast extract or LPS. These observations point towards the significant impact of environmental variables on SEP production dynamics.

However, not all SEPs were influenced by these external factors. For example, three SEP (BP26, BP36, and BP42) were consistently produced in the YCFA medium (pH 7.0), regardless of the presence or absence of SCFAs, suggesting their independence from these specific factors. Additionally, all previously described SEP except BP15 were identified, including five SEPs (BP3, BP5, BP8, BP11, and BP12) previously detected only in co-culture with other bacteria (SIHUMIx) (Petruschke et al., 2021) (FIGURE V-9B). Their identification in single bacterial cultivation data suggests that their biosynthesis may not be solely reliant on interspecies interactions or communication within the microbiome.

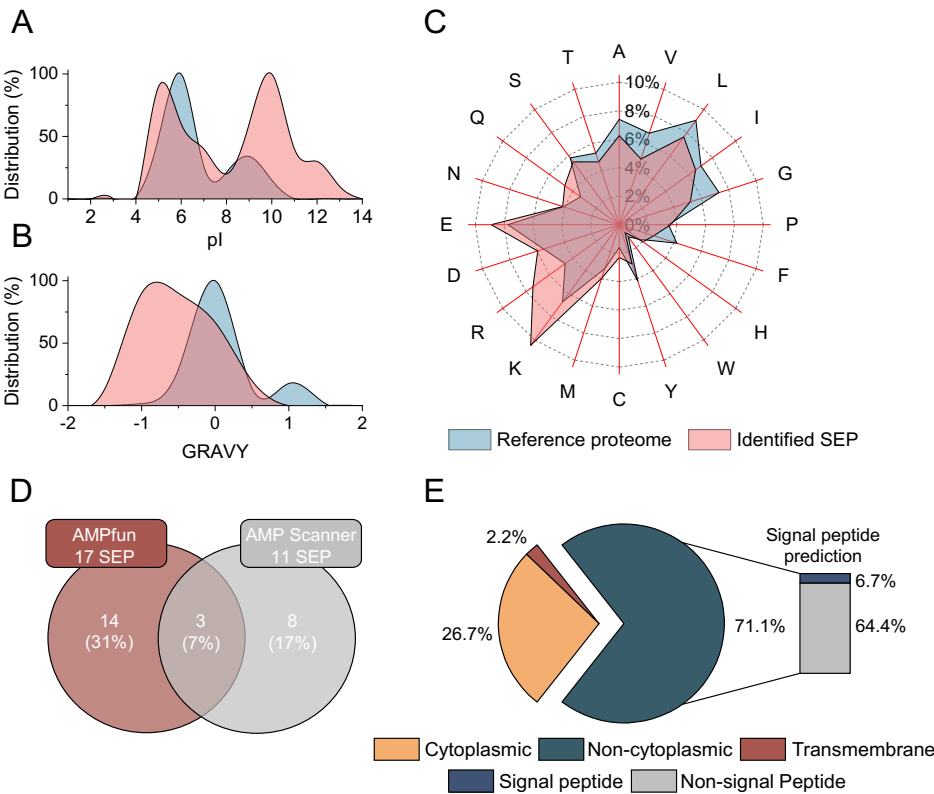
Furthermore, it suggests that the production of multiple SEP might have been influenced by specific bacterial growth and stress factors (FIGURE V-9B). For instance, BP11 could be linked to the presence of yeast extract in the medium at pH 7.0, while BP5 was only detected at pH 6.0 (FIGURE V-9B), indicating a potential involvement in the acid stress response.



**FIGURE V-9 | Presence of SEP across Diverse Culture Conditions.** Bottom-up identifications across the seven different culture conditions of (A) all detected SEP (44 in total) and (B) previously described SEP (14 in total). Highlighted SEP are reported to be exclusively produced within the microbiome community. Set size indicates the total number of SEP identified in each condition, with intersection size representing shared SEP across different conditions. With permission from (Genth et al., 2023).

### 3.4 Biochemical Predictions of Identified SEP

The potential functions of the identified SEP were analyzed by predicting physicochemical properties and by examining sequence homologies based on reference domains, families, specific sites, or motifs. A distinct bimodal distribution of isoelectric point (pI) values was observed, indicating a notable shift of the SEP towards more basic pI values compared to the *B. producta* reference proteome (FIGURE V-10A). Furthermore, examination of the grand average of hydropathy (GRAVY) scores revealed a broader spectrum of hydrophilic properties (FIGURE V-10B), which aligned with the amino acid composition of the SEP characterized by a reduced frequency of hydrophobic amino acids like alanine, leucine, and isoleucine, alongside a higher occurrence of lysine and arginine residues (FIGURE V-10C). One particularly noteworthy finding was the potential antimicrobial peptide (AMP) activity exhibited by 25 SEP, as predicted by AMPfun (Chung et al., 2020) or AMP scanner v.2 (Veltri et al., 2018) (FIGURE V-10D). Among these, BP16, BP34, and BP40 exhibited consistent AMP predictions (FIGURE V-10A). Notably, BP40 also exhibited the potential to target both Gram-positive and Gram-negative bacteria. Additionally, approximately 71% of the identified SEP were predicted to be non-cytoplasmic (FIGURE V-10E). Furthermore, three non-cytoplasmic SEP (BP14, BP37, BP45) were predicted to have signal peptides, hinting at their potential involvement in cell-cell and cell-host communication (Hayes et al., 2010).



**FIGURE V-10 | Biochemical Predictions of Identified SEP.** Distribution of (A) isoelectric points (pI), (B) grand average of hydropathy (GRAVY), and (C) amino acid composition between the *B. producta* reference proteome and the identified SEP. (D) Overlap of antimicrobial peptide (AMP) predictions. (E) Distribution of cellular localizations by Phobius.

## 4 Discussion and Conclusion

### 4.1 Challenges and Advances in SEP Identification

The presence of unannotated protein-coding sORFs were analyzed by validating their respective translated SEP products at the protein level. Utilizing various growth conditions, alongside the application of both BUP and TDP methodologies, a total of 45 SEP were successfully identified (FIGURE V-4). While BUP predominantly detected the majority of SEP primarily due to its higher sensitivity compared to TDP (Cassidy et al., 2023), TDP exclusively identified one SEP (BP46) (FIGURE V-2). Among the 45 SEP identified, 31 SEP (BP16-BP46) were identified for the first time, while 14 SEP (BP1-BP14) had been previously described (Petruschke et al., 2021). Both BUP and TDP approaches complemented each other in the identification of SEP, indicating the strengths of each methodology (FIGURE V-2).

To ensure high-confidence identification of SEP, strict filtering criteria were implemented, aligning with recent recommendations in the field (Chen et al., 2023). Despite 17 out of 44 BUP identifications relying on a single peptide to support the protein-coding potential of the sORFs, these peptides exhibited an average of 69 PSMs. Although single-peptide identifications can be susceptible to increased false discovery rates (Nesvizhskii, 2010; Hadjeras et al., 2023), the majority of MS-based SEP identifications typically rely on a single unique peptide (Slavoff et al., 2013; Cassidy et al., 2019). This is due to the small protein size of SEP, resulting in a limited generation of tryptic peptides suitable for identification. To address this limitation, the use of multiple proteases (Bartel et al., 2020; Kaulich et al., 2021) or the integration of both BUP and TDP (Cassidy et al., 2016, 2021b) has proven effective in significantly improving sequence coverage and identification confidence of SEP. Incorporating TDP not only strengthens the evidence for the existence of SEP but also enables the detection of N- and C-terminal neo-termini and PTMs, achievements often challenging with BUP alone (Tholey and Becker, 2017).

Application of TDP in this study identified 26 canonical proteoforms (FIGURE V-4) and increased the average sequence coverage by 19% compared to BUP. Furthermore, the identification of truncated proteoforms, potentially originating from proteolytic processing or alternative initiation, may suggest the potential impact of structural variations on the biological function of SEP (Melo et al., 2023). The identification of several proteoforms with PTMs suggests that the identified SEP can undergo posttranslational protein processing.

The results of this study, alongside those of Petruschke and colleagues (Petruschke et al., 2021), suggest that the production of certain SEP is not solely dependent on interspecies interactions within the SIHUMIx co-culture system. Instead, specific SEP can also be produced under monoculture conditions and are influenced by various growth conditions and extracellular factors. Factors such as the presence of endotoxins (LPS) or pH variations,

known to play significant roles in gut inflammatory diseases such as IBD (Bai et al., 2016; Candelli et al., 2021), appear to act as stimuli or regulators for the production of specific SEP (FIGURE V-9). Moreover, a high number of identifications independent of culture conditions indicate a universal role for SEP in the survival and growth of *B. producta*. Overall, the results from this study, alongside those of Petruschke and colleagues, suggest a complex interplay between SEP and environmental factors (Petruschke et al., 2021). Nonetheless, it's worth considering why some of these potentially universal SEP were not identified in Petruschke and colleagues' study (Petruschke et al., 2021).

Various factors may have limited the depth and detection capabilities of their proteomic analysis. One critical factor impacting proteomic workflows is the database search. Integrating a large proteogenomic database like SIHUMIX, including eight bacterial strains of interest, poses challenges in proteogenomic studies and subsequent FDR analysis (Nesvizhskii, 2014). This challenge arises due to the increased number of candidates competing for matching to an experimental MS/MS spectrum, increasing the risk of incorrect matches and distinguishing true from false identifications (Nesvizhskii, 2010). Consequently, large proteogenomic database searches may generate fewer non-canonical and total peptides compared to conventional searches with a reference database (Aggarwal et al., 2022).

It's also essential to note that the absence of certain SEP in specific culture conditions does not necessarily imply their complete absence from those conditions. Instead, their presence may be at concentrations falling below the limits of detection of the methods or instruments used for analysis. Furthermore, interfering factors such as co-eluting compounds that compete for ionization can lead to reduced ionization efficiency for the target peptides (Keller et al., 2008), resulting in reduced signal intensities and sensitivity for low-abundance peptides, making identification challenging (Cassidy et al., 2023). To address these challenges, various techniques can be applied such as isolating, enriching, or depleting proteins of interest (Cassidy et al., 2019), applying a second chromatographic separation (Cassidy et al., 2021a), or using gas-phase separation (Swearingen and Moritz, 2012). These techniques can extend the limits of detection for low-abundant species in complex samples.

To address some of the challenges Petruschke and colleagues conducted a comprehensive investigation of various proteomic techniques and approaches. They compared different enrichment strategies (C<sub>8</sub>-cartridge and GELFrEE), global proteomics methods (SP3, FASP, in-gel, and in-solution), and multiple protease cleavage approaches (trypsin and Asp-N) to identify small proteins in the SIHUMIX system (Petruschke et al., 2020). Their investigation provided valuable insights into the strengths and limitations of these techniques. By applying them to the SIHUMIX system, they identified several novel SEP, thereby contributing to the field of human gut microbiome research (Petruschke et al., 2021). They also acknowledged that differences in growth and cultivation conditions could impact the detection of specific SEP.

The results presented in this study complement the work of Petruschke and colleagues, suggesting that the production of certain SEP in *B. producta* is not exclusively reliant on interspecies interactions. Environmental factors can also impact the production of certain SEP in *B. producta*. Although SEP research is complex, and the challenges in methodology are acknowledged, both studies provide unique insights into the intricate nature of SEP biology. Therefore, it is crucial to regard these efforts as complementary, with each study contributing significantly to our understanding of SEP production. A comprehensive approach considering various experimental factors, including growth conditions, sample preparation techniques, and proteomic workflows, will be pivotal in advancing the field and uncovering the full spectrum of SEP.

## 4.2 Future Directions of SEP Research

Given the focus of this study on soluble proteins, it is not surprising that the biochemical properties of the identified SEP exhibited a decreased relative frequency of hydrophobic amino acids and an increased frequency of charged amino acids, particularly lysine and arginine residues (FIGURE V-10C). These charged residues, especially those located at the C-termini of peptides resulting from tryptic digestion, typically enhance proton affinity, ionization efficiency, and CID fragmentation, leading to improved MS sensitivity (Dupree et al., 2020). The small overlap of AMP predictions for three SEP (BP16, BP34, BP40; FIGURE V-10A), can be likely attributed to variations in training sets, physicochemical properties, and amino acid frequencies at each sequence position, thus influencing the prediction model for classifying antibacterial peptides (Gabere and Noble, 2017).

Notably, one SEP (BP27) was predicted to be a short transmembrane protein (FIGURE V-10E). Short membrane-associated SEP are predicted to constitute 35% of the SEP population in the human microbiome (Sberro et al., 2019). Given their potential roles in essential cellular processes such as transport, signaling pathways, cell division, respiration, sporulation, and membrane integrity (Yadavalli and Yuan, 2022), these transmembrane SEP should be targeted for future analysis. Consequently, specialized, MS-compatible membrane protein enrichment protocols should be prioritized in future studies to analyze these membrane-associated SEP (Capri and Whitelegge, 2017; Ahrens et al., 2022; Meier-Credo et al., 2022). Despite their potential significance, one of the major challenges associated with SEP lies in their lack of sequence homology with known proteins. Their small size and compact folding make traditional annotation and structure prediction tools less effective, as these tools typically depend on homology searches (Ahrens et al., 2022). This limits the understanding of the SEP function and emphasizes the need for experimental validation. Functional proteomics emerges as a promising approach to address these challenges and reveal the functional properties, biological roles, and mechanistic contributions of SEP.

Various proteomic methodologies, such as biochemical fractionation of soluble complexes coupled with mass spectrometry (Havugimana et al., 2022), affinity purification mass spectrometry (AP-MS) (Gnanasekaran and Pappu, 2023), or cross-linking mass spectrometry (XL-MS) (Piersimoni et al., 2022), are available to detect native protein complexes within cellular extracts and generate networks of protein-protein interactions (Low et al., 2021). These methodologies can significantly contribute to the identification of SEP within established protein networks, thereby enhancing our understanding of their roles and contributions to cellular functions (Garcia-del Rio et al., 2023; Leblanc et al., 2023). Additionally, genetic systems like the yeast two-hybrid assay can complement these efforts by mapping protein-protein interactions, providing insights into the roles of SEP within the cellular interactome (Mehla et al., 2015). Furthermore, integrating transcriptomic data created using RNA-Seq or ribosome profiling with proteomic analysis of the same biological samples can increase the number of experimentally validated SEP (Guilloy et al., 2023; Hadjeras et al., 2023). This integrated omics approach enables a more comprehensive exploration of the SEP proteome, leading to a better understanding of their biological functions. Future research in this field will likely provide more insight into the significance of these small proteins.

## VI OUTER MEMBRANE VESICLES ANALYSIS

---

<b>1</b>	<b>Introduction and Summary .....</b>	<b>122</b>
<b>2</b>	<b>Experimental Design .....</b>	<b>125</b>
	2.1 MS and Proteolytic Digestion Compatibility Check .....	125
	2.2 Biophysical and Proteomic Characterization .....	125
	2.3 Assessment of Loading Capacity .....	126
	2.4 Optimizing Lysis and Proteolytic Digestion of OMV Preparations .....	126
	2.5 Caco-2 Wound-Healing Assay .....	127
<b>3</b>	<b>Results .....</b>	<b>128</b>
	3.1 Proteomic Workflow Compatibility .....	128
	3.2 Biophysical Characteristics and Kit Loading Capacity .....	129
	3.3 Proteomic Analysis of <i>E. coli</i> OMVs .....	131
	3.4 Improving OMV Lysis and Trypsin-Based Digestion .....	134
<b>4</b>	<b>Discussion and Conclusion .....</b>	<b>139</b>
	4.1 Biophysical and Biological Characteristics of <i>E. coli</i> OMVs .....	139
	4.2 Proteomic Characterization of <i>E. coli</i> OMVs .....	140
	4.3 Optimized OMV Sample Preparation Workflow .....	142

## 1 Introduction and Summary

Extracellular vesicles (EVs) are nanoparticles with a diameter ranging from 50 to 150 nm (van Niel et al., 2018). They play a crucial role in mediating intercellular communication across diverse biological systems by carrying proteins, lipids, nucleic acids, and bioactive molecules (Michel and Gaborski, 2022). Various cell types release these vesicles. For instance, Gram-negative bacteria can release two distinct types into the extracellular milieu: outer membrane vesicles (OMVs) and outer-inner membrane vesicles (O-IMVs) (Pérez-Cruz et al., 2015; Toyofuku et al., 2019). OMV formation possibly occurs via budding, a process where lipid membranes protrude outward, enclosing periplasmic components within a lipid membrane. In contrast, O-IMVs encapsulate both the outer and inner membranes, resulting in the inclusion of cytoplasmic components within the vesicles (Pérez-Cruz et al., 2015; Toyofuku et al., 2019). In the human gut, OMVs can modulate host immune responses, contribute to gut homeostasis, and dynamically interact with diverse host cells, including intestinal epithelial cells and immune cells, as well as components such as the mucus layer (Taheri et al., 2018; Bittel et al., 2021; Juodeikis and Carding, 2022). The cargo molecules transported by OMVs possess biomedical significance and represent an essential aspect of the exoproteome and exometabolome (Juodeikis and Carding, 2022).

Methods for isolating EVs vary according to the study's specific requirements. Commonly used non-specific methods encompass differential ultracentrifugation with or without a sucrose gradient, size exclusion chromatography, and polymer-based precipitation (Abramowicz et al., 2016). These methods typically involve rigorous pre- and post-purification steps to remove non-specifically interacting proteins and contaminants, including high-abundance soluble proteins, e.g., lipoproteins or albumin, which are frequently co-isolated with the vesicles (Li et al., 2019a). Yet, for a mass spectrometry-based analysis of EVs, the choice of appropriate isolation and purification methods remains a major challenge (Abramowicz et al., 2016). Sample quality, including salts, detergents, abundant protein components, and pH, significantly affects downstream proteomic analysis, sometimes more than the capabilities of the MS instrument itself (Annesley, 2003; Keller et al., 2008).

While immunoaffinity interactions targeting surface markers, such as those from the tetraspanin family (CD63, CD9, CD81), can enable more specific isolation of eukaryotic EVs, surface markers for OMVs are relatively rare (Abramowicz et al., 2016). Although these methods offer high specificity, they may exclude certain vesicles lacking the targeted surface markers and potentially introduce bias into the analysis. Alternatively, surface charge-based mechanisms have been explored for EV isolation. Targeting negatively charged phosphatidylserine, lipoteichoic acid, lipopolysaccharides, or other macromolecules that confer a negative charge to EVs has enabled the development of novel strategies (Deregibus et al., 2016; Midekessa et al., 2020). Cationic polymers, such as poly-L-lysine (PLL), can

adhere to exosomes through electrostatic interactions, facilitating their isolation (Kim et al., 2020; Kim and Shin, 2021). With the growing interest in EV research, marked by an increased growth in related publications, diverse isolation methods have been developed (Chen et al., 2022; Liu et al., 2022).

Despite extensive efforts, achieving full agreement on a standardized method for isolating EVs remains a challenge (Théry et al., 2018). Acknowledging the diversity of isolation methods, various commercial kits have been introduced to simplify the isolation process. For instance, the ion-exchange platform ExoCAS-2 utilizes PLL polymer-functionalized magnetic beads for the specific capture of EVs from culture medium or plasma samples (Kim and Shin, 2021). Additionally, the ExoBacteria OMV Isolation Kit employs a precipitation-free ion-exchange column system containing a capture resin and gravity column specifically designed for capturing Gram-negative OMVs. Notably, a commercial kit for capturing Gram-positive OMVs is not yet available. While the consensus on standardized methods remains elusive, isolation kits present a potentially promising solution in specific research contexts.

In addition to the challenges posed by non-standardized methods, confirming isolated particles as true EVs is crucial. The International Society for Extracellular Vesicles has issued guidelines outlining key protocols and steps in EV science (Théry et al., 2018). These guidelines emphasize the importance of several key steps in evaluating EVs, including confirming the presence (or enrichment) of EV markers and absence of non-EV markers, quantifying the abundance of EVs (total particle number and/or protein or lipid content), estimating the purity of vesicle preparations (e.g., by comparing the ratio of nano-vesicle counts to protein concentration), verifying specific EV-associated functions (e.g., through dose-response assessment), and assessing individual extracellular vesicles using various methods (e.g., image-based techniques such as electron microscopy and non-image-based methods like nanoparticle tracking analysis) (Théry et al., 2018).

Commercial isolation kits with predefined protocols, are user-friendly and require no specialized equipment such as an ultracentrifuge, making them suitable for laboratories with limited resources. This study evaluates the compatibility of the ExoBacteria OMV Isolation Kit with mass spectrometry-based sample preparation. Despite previous use in mass spectrometry studies employing paper spray ionization MS (PSI-MS) for measuring metabolites, lipids, and small peptides (2–4 amino acid residues) (Chamberlain et al., 2021), as well as inductively coupled plasma MS (ICP-MS) for measuring the chemical element lanthanum (Fujitani et al., 2022), its compatibility with LC-MS workflows and suitability for sample preparation in proteomic analysis require further evaluation.

### **Aim of this study:**

- Evaluate the compatibility of the OMV elution solution with LC-MS analysis and tryptic digestion.
- Isolate *E. coli* OMVs from different culture conditions (glucose and acetate) at different growth phases (mid-logarithmic, prestationary, stationary, and death phase).
- Validate OMV nanoparticles using nanoparticle tracking analysis and estimation of the purity of vesicle preparations.
- Evaluate the supernatant loading capacity of the ExoBacteria OMV isolation kit.
- Perform proteomic analysis of OMVs, classify the subcellular topological distribution of the *E. coli* OMV proteome, and quantify OMV markers.
- Evaluate the biological activity of OMVs using a Caco-2 wound healing assay.
- Test various sample preparation protocols and integrate non-ionic detergents for OMV lysis and subsequent proteolytic digestion.

## 2 Experimental Design

Unless otherwise stated, all OMV samples were analyzed in three technical replicates on the Q-Exactive Plus mass spectrometer using an LFQ method with a 60-minute gradient for peptide separation, and raw data were analyzed using PD 3.0 utilizing CHIMERYS.

### 2.1 MS and Proteolytic Digestion Compatibility Check

The evaluation of MS compatibility included the detection of polyethylene glycol (PEG) or other polymer contaminations through MALDI MS. Instead of bacterial supernatant, water was processed through a kit blank run (chapter II.2.2). The resulting eluate in OMV elution buffer was subsequently diluted at ratios of 1:10, 1:20, 1:50, and 1:100 using CHCA matrix and analyzed using MALDI MS. The MS spectra were manually examined. Additional LC-MS analysis of two dilutions (1:10 and 1:100) were conducted on the Q-Exactive HF, using cytochrome C digest for retention time and HeLa digest for peptide identification controls.

The influence of the OMV elution solution on tryptic digestion was evaluated by assessing the efficiency of protein digestion for cytochrome C (CytC; 12 kDa), myoglobin (Myo; 17 kDa),  $\beta$ -casein ( $\beta$ -Cas; 24 kDa), carbonic anhydrase (CA; 29 kDa), alcohol dehydrogenase (ADH; 41 kDa), and bovine serum albumin (BSA; 66 kDa). The proteins were dissolved in either OMV elution solution or 100 mM TEAB (pH 8.5) at a concentration of 20  $\mu$ g/ $\mu$ l and combined in equal parts (v/v) to produce a 6-protein or 4-protein mixture (excluding myoglobin and alcohol dehydrogenase). Following reduction and alkylation using 12 mM TCEP and 40 mM CAA, both protein mixtures were digested in either 100 mM TEAB-buffered OMV elution buffer or 100 mM TEAB alone (both at pH 8.5) using trypsin at a protease-to-substrate ratio of 1:50. The digests were then mixed and incubated with SDS sample buffer before being visualized using SDS-PAGE.

### 2.2 Biophysical and Proteomic Characterization

*E. coli* was cultured in duplicates using M9 medium supplemented with either 15 mM glucose or 45 mM acetate and OMVs were isolated at different growth phases, including mid-logarithmic, prestationary, stationary, and death phases. Nanoparticle tracking analysis (NTA) was used to validate the successful isolation of OMV nanoparticles and to characterize their biophysical properties by measuring particle size and concentration. Details about the cultivation of *E. coli*, OMV Isolation, and NTA parameters are described in chapter II.2.2. For the proteomic analysis, OMVs were lysed using 10 freeze-thaw cycles. Proteins were reduced and alkylated (12 mM TCEP and 40 mM CAA), and digested in-solution using 100 mM TEAB-buffered OMV elution buffer (pH 8.5), using trypsin at a protease-to-substrate ratio of 1:40 (20 h, 37°C, 800 rpm).

## 2.3 Assessment of Loading Capacity

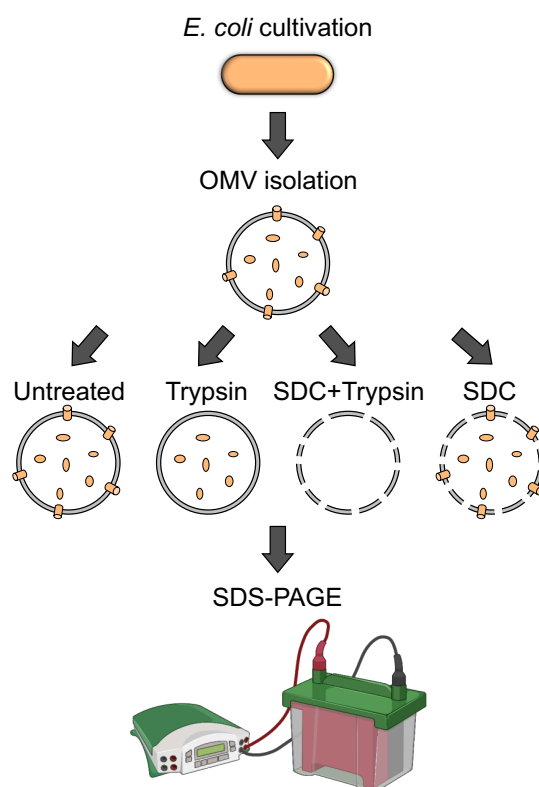
Different amounts of *E. coli* supernatants (20, 40, and 60 ml) were processed using the ExoBacteria OMV Isolation Kit. Supernatants were collected from three independent *E. coli* cultures grown in an M9 glucose medium until the mid-logarithmic growth phase (OD of 1.0). After isolation, the number and size of OMVs in each sample were determined by NTA, and the total protein concentration was measured by BCA assay to calculate a purity ratio (particles/ $\mu$ g protein).

## 2.4 Optimizing Lysis and Proteolytic Digestion of OMV Preparations

**Integration of Non-Ionic Detergent Lysis** – The surface exposure of OMV proteins and potential protein contaminations, as well as the effectiveness of non-ionic detergent lysis using SDC, were assessed (FIGURE VI-1). An untreated reference of the OMV preparations served as a control. The enzyme-treated sample was subjected to tryptic digestion by adjusting OMV preparations to 100 mM TEAB and adding trypsin at a 1:40 enzyme-to-substrate ratio. For the detergent- and enzyme-treated sample, OMV lysis was performed using 2% (w/v) SDC, followed by dilution to 0.5% SDC using 100 mM TEAB prior to proteolytic digestion. The detergent-treated sample was only subjected to OMV lysis using 2% (w/v) SDC. Subsequently, all samples were mixed and incubated with SDS sample buffer before visualization using SDS-PAGE.

### Improvement of Lysis and Proteolytic Digestion of OMVs

Forty microgram protein aliquots of isolated OMVs were processed using on-bead SP3 (Hughes et al., 2019), on-membrane FASP (Manza et al., 2005; Wiśniewski et al., 2009), or in-solution sample preparation workflows. Protocol-specific buffer compositions and final reagent concentrations are listed in TABLE VI-1. Each protocol was performed in triplicates using OMV preparations from three independent *E. coli* cultivations. To prevent the precipitation of SDC in acidic environments (Scheerlinck et al., 2015), pH variations were avoided. Since this necessitated the avoidance of TCEP, reducing and alkylating agents instead



**FIGURE VI-1 | Workflow to Evaluate OMV-associated proteins and Non-Ionic Detergent OMV Lysis.**

comprised 10 mM DTT and 50 mM IAA within a 100 mM TEAB buffer environment. All samples were digested with trypsin at a 1:40 enzyme-to-substrate ratio (20 h, 37°C, 800 rpm), with the addition of 0.01% (w/v) *n*-dodecyl- $\beta$ -D-maltoside (DDM) or 0.5% (w/v) SDC. For SDC digests, samples were additionally processed using a modified phase transfer protocol (Masuda et al., 2008) (chapter II.3.3).

**TABLE VI-1 | Overview of Protocol-specific Sample Processing Steps.** All listed values represent end concentrations for the in-solution digestion (ISD), single-pot, solid-phase-enhanced protein extraction (SP3), or filter-aided sample preparation (FASP) protocol.

PROTOCOL STEP	CONDITIONS	ISD-CLASSIC	ISD-IMPROVED	SP3-SDS-TEAB	SP3-SDS-SDC	SP3-SDS-DDM	SP3-SDC	FASP-SDC	FASP-SDS-SDC
OMV LYSIS	Freeze thawing 2% SDC 1% SDS	✓	✓	✓	✓	✓	✓	✓	✓
RED. & ALK.	10 mM DTT & 50 mM IAA	✓	✓	✓	✓	✓	✓	✓	✓
SDS REMOVAL	8 M urea, 100 mM TEAB								✓
TRYPTIC DIGESTION	0.5% SDC, 100 mM TEAB 0.001% DDM, 100 mM TEAB 100 mM TEAB	✓	✓	✓	✓	✓	✓	✓	✓
SDC REMOVAL	0.5% TFA & 100% ethyl acetate		✓		✓		✓	✓	✓

## 2.5 Caco-2 Wound-Healing Assay

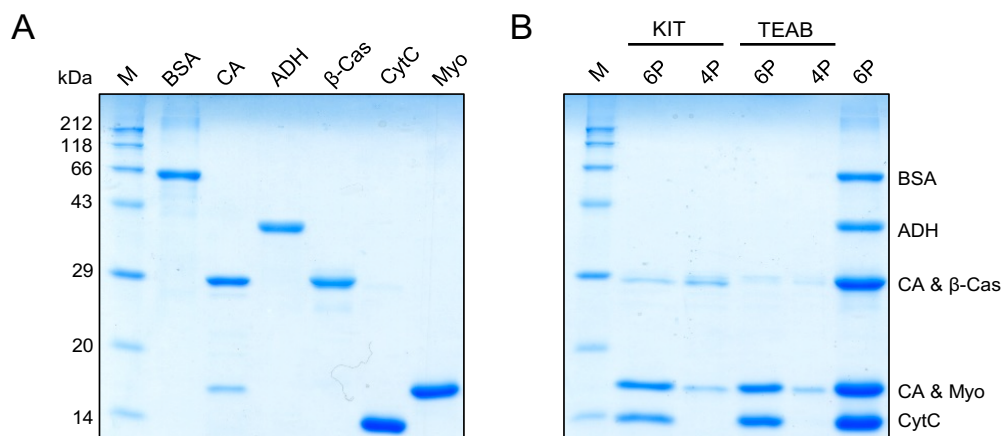
An important feature of OMVs is that the proteins associated with them exhibit various biological activities. Particularly, OMVs produced by *E. coli* can exhibit an inhibitory effect on cell proliferation and induce pro-inflammatory responses in intestinal epithelial cells (Cañas et al., 2016; Patten et al., 2017). To evaluate the potential biological activity of isolated *E. coli* OMVs, a wound-healing assay utilizing the human intestinal epithelial cell line Caco-2 was conducted. Wound closure was measured immediately after removing the insert, as well as at 12 and 30 hours after incubation with increasing OMV concentrations (10, 50, and 100  $\mu$ g/ml). Positive and negative controls included incubation with 5 ng/ml TGF $\beta$  and 1  $\mu$ g/ml LPS, respectively. Cells solely incubated with the medium (0.1% FCS) and the OMV elution buffer served as references for comparison with the treated groups. These controls facilitated the calculation of relative wound closure. Further details on the cultivation of Caco-2 cells and the wound healing assay can be found in chapter II.2.3.

### 3 Results

#### 3.1 Proteomic Workflow Compatibility

Given that the reagents of the OMV kit are undisclosed, conducting a thorough compatibility check, specifically focusing on LC-MS/MS analysis and protein digestion, was essential for its integration into a proteomic workflow. MALDI MS analysis of diluted eluate from a blank kit run revealed no characteristic PEG ion series or other polymeric impurities. While the absence of characteristic impurities indicates compatibility of the OMV kit with mass spectrometry techniques (Keller et al., 2008), it does not guarantee complete compatibility. Additional LC-MS analysis of two dilutions (1:10 and 1:100) using cytochrome C digest for retention time and HeLa digest for peptide identification controls showed no peptide retention time shifts and comparable peptide identification. This indicates that the elution behavior of peptides and peptide detection remained largely unaffected, at least for subsequent LC-MS runs. Monitoring signal stability and background noise revealed a singly charged peak ( $309.125\ m/z$ ) at 53 minutes, with intensities of approximately  $1.8 \times 10^8$  and  $6.4 \times 10^8$  in the 1:100 and 1:10 dilutions, respectively. The observed minimal impact on chromatographic runs and MS analysis confirms the compatibility of the OMV kit with LC-MS analysis.

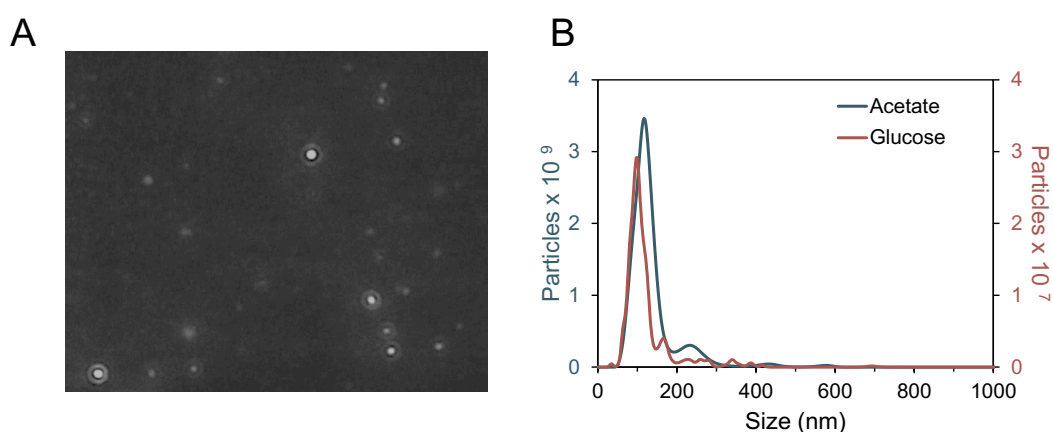
Furthermore, the compatibility of the OMV elution buffer with tryptic digestion was analyzed using protein mixtures consisting of six proteins (6P), and four proteins (4P), the latter excluding myoglobin and alcohol dehydrogenase (FIGURE VI-2A). The mixtures were digested using either TEAB-buffered OMV elution buffer (KIT) or 100 mM TEAB (TEAB) and were evaluated through SDS-PAGE analysis (FIGURE VI-2B). While the comparison of the two different digestion conditions showed slightly reduced protein band intensities with the TEAB-buffered OMV elution buffer (FIGURE VI-2B), the results still confirmed the compatibility of the OMV kit with tryptic digestion, allowing integration of the kit into a proteomic workflow.



**FIGURE VI-2 | SDS-PAGE Analysis of Protein Digestion using the Exobacteria Elution Buffer. (A)** Molecular weight distribution and composition of the six-protein mixture. **(B)** Evaluation of the tryptic digestion of the 6-protein (6P) or 4-protein (4P) mixtures (excluding myoglobin and alcohol dehydrogenase) using TEAB-buffered OMV elution buffer (KIT) or 100 mM TEAB.

### 3.2 Biophysical Characteristics and Kit Loading Capacity

**Biophysical Characteristics** – The isolated OMVs were first plated on LB agar to confirm the absence of bacterial contamination. Subsequently, they were characterized using dynamic light scattering with NTA, which allowed direct, real-time visualization of the isolated OMVs. FIGURE VI-3A illustrates a screenshot from one of the recorded videos. Examination of the isolated OMVs revealed nanoparticles ranging from  $117.2 \pm 2.1$  nm (mode  $\pm$  standard error) for OMVs isolated from acetate supernatant to  $95.3 \pm 4.1$  nm for those isolated from glucose supernatant (FIGURE VI-3B and TABLE VI-2). The average particle concentration was higher for OMVs isolated from supernatant obtained from acetate cultivation ( $9.4 \times 10^{10} \pm 0.9$  particles/ml) compared to those isolated from supernatant obtained from glucose cultivation ( $7.8 \times 10^8 \pm 0.8$  particles/ml). The observed differences in particle concentration between the two supernatants, while providing a general reference, require additional replicates for thorough validation. Furthermore, NTA measurements indicate a relatively narrow size distribution of particles, suggesting a more monodisperse sample rather than a polydisperse one.

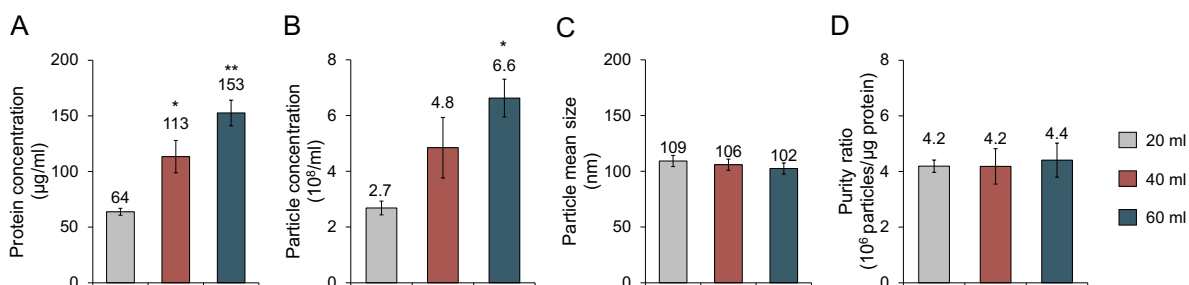


**FIGURE VI-3 | Nanoparticle Tracking Analysis of *E. coli* OMVs.** (A) Representative visualization of OMV particles captured from the recorded video. (B) Particle size distribution of OMVs isolated from *E. coli* supernatant cultured in acetate or glucose M9 medium.

**TABLE VI-2 | Summary of Nanoparticle Tracking Analysis**

PARAMETER	ACETATE	GLUCOSE
Mode particle size	$117.2 \pm 2.1$ nm	$95.3 \pm 4.1$ nm
D10	$82.9 \pm 3.4$ nm	$79.2 \pm 3.8$ nm
D50	$124.3 \pm 1.7$ nm	$103.4 \pm 2.6$ nm
D90	$204.3 \pm 5.7$ nm	$207.4 \pm 19.7$ nm
Particle concentration	$9.4 \times 10^{10} \pm 0.9$ particles/ml	$7.8 \times 10^8 \pm 0.8$ particles/ml

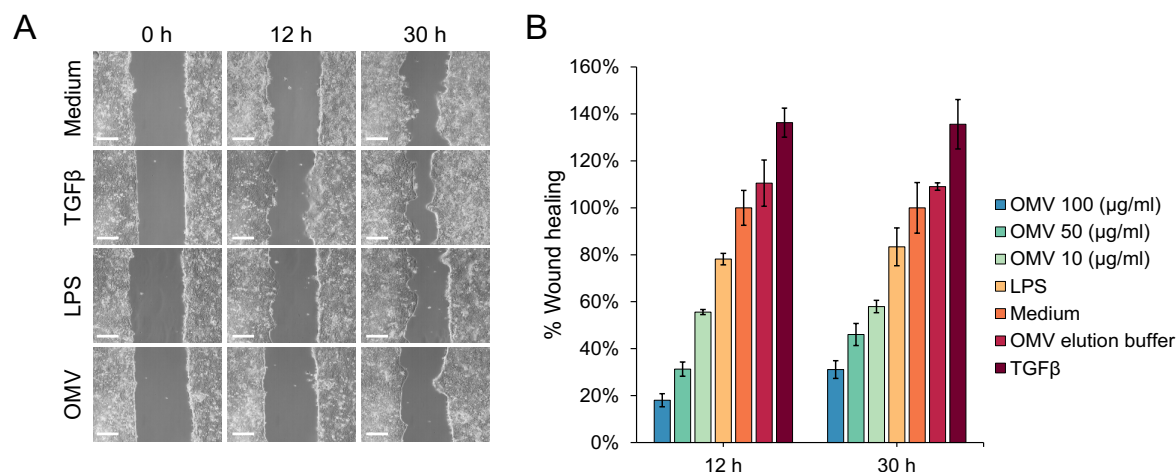
**Loading Capacity** – The assessment of the kit's loading capacity revealed that an increase in the initial supernatant volume from 20 to 40 or 60 ml resulted in a significant doubling of protein concentrations (FIGURE VI-4A) and particle concentrations (FIGURE VI-4B). Notably, the particle size and distribution of particles remained constant at  $106 \pm 3.0$  nm (FIGURE VI-4C and FIGURE A-24), indicating that OMVs maintained their size and distribution even with an increased particle concentration. Furthermore, the purity ratio (the ratio of vesicle counts to protein concentration) of OMV preparations, which considers both protein contamination and loss of OMVs during isolation (Webber and Clayton, 2013), remained relatively constant at  $4.3 \times 10^6 \pm 0.1$  particles/ $\mu$ g of protein (FIGURE VI-4D). Despite variations in the loaded supernatant volume, a comparable particle size and purity ratio could be maintained, indicating the kit's efficiency in handling increased sample volumes.



**FIGURE VI-4 | Loading Capacity Evaluation of the ExoBacteria OMV Isolation Kit.** Different *E. coli* supernatant volumes (20, 40, and 60 ml) were processed with the OMV isolation kit and analyzed using NTA and BCA. **(A)** Protein concentration, **(B)** particle concentration, **(C)** particle mean size, and **(D)** normalized purity ratio (particle/protein ratio). Data represent mean  $\pm$  standard error from 3 independent experiments. Significant differences were calculated via one-way ANOVA with Dunnett's correction. \* ( $p < 0.05$ ); \*\* ( $p < 0.01$ ).

**Biological Activities** – The ability of isolated OMVs to exhibit biological activities was evaluated using an *in vitro* wound closure assay, which examined the migration and proliferation of human colonic Caco-2 cells. The evaluation of wound size reduction after 12 and 30 hours indicated a notable reduction of wound healing compared to untreated control cells (FIGURE VI-5A). The quantitative assessment of wound healing, with the medium arbitrarily set at 100% as a reference, revealed reduced wound closure with increasing concentrations of OMVs. At concentrations of 10  $\mu$ g/ml, 50  $\mu$ g/ml, and 100  $\mu$ g/ml, OMVs exhibited decreasing percentages of wound closure after 12 hours ( $45 \pm 1.1\%$ ,  $68 \pm 3.0\%$ , and  $82 \pm 2.8\%$ , respectively FIGURE VI-5B). Despite statistical analysis using one-way ANOVA with Dunnett's correction, no significant changes in wound healing ( $p < 0.05$ ) were observed. This lack of statistical significance may be attributed to the limited sample size of only three biological experiments. Although the results were not statistically significant, they suggest a possible dose-dependent inhibitory effect of OMVs on wound closure, with higher concentrations having a greater effect. While the OMV elution buffer showed a modest

stimulatory effect on wound closure after 30 hours ( $9 \pm 2.0\%$  FIGURE VI-5B), this effect may be attributed to common buffer components like calcium and phosphoric acid, which can influence the wound healing process (Navarro-Requena et al., 2018; Sim et al., 2022). The negative control (LPS) resulted in a reduction of  $22 \pm 2.5\%$ , while the positive control (TGF $\beta$ ), known for its role in promoting proliferation (Penn et al., 2012), stimulated wound closure by  $36 \pm 6.2\%$ .



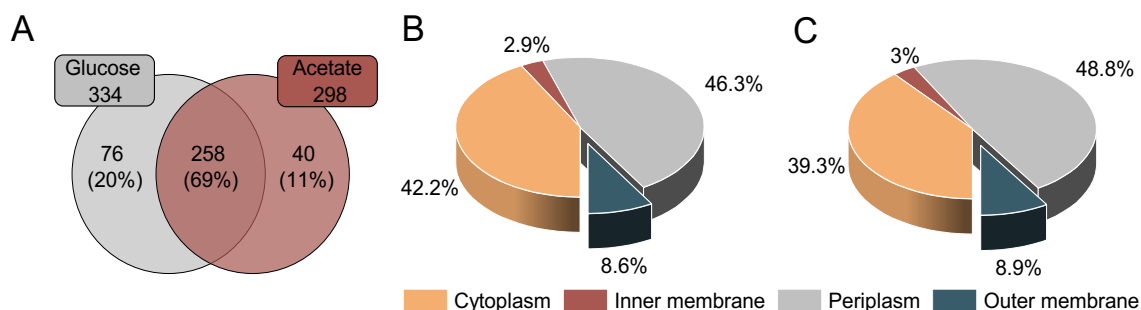
**FIGURE VI-5 | Caco-2 Wound Healing Assay. (A)** Time-lapse microscopy images of wound closure of Caco-2 cells with medium, TGF $\beta$  (5 ng/ml), LPS (1  $\mu$ g/ml), and OMVs (50  $\mu$ g/ml). Images were captured at 0 h, 12 h, and 30 h (10-fold magnification). **(B)** Wound closure after 12 and 30 h incubation with varying OMV concentrations (10, 50, and 100  $\mu$ g/ $\mu$ l), LPS (1  $\mu$ g/ml), OMV elution buffer, medium alone, or TGF $\beta$  (5 ng/ml). Medium alone was arbitrarily assigned as 100%. Data represent mean  $\pm$  standard error from three independent experiments. Bars represent 200  $\mu$ m.

### 3.3 Proteomic Analysis of *E. coli* OMVs

The proteomic changes in *E. coli* OMV protein content under different growth states and culture conditions were analyzed by a bottom-up proteomic analysis. Changes in protein abundance and localization were estimated by annotating subcellular locations using STEPdb 2.0 (Loos et al., 2019), which categorizes proteins into 13 distinct subcellular classes (FIGURE A-23A). Due to the dynamic nature of protein localization and their ability to move to various extracytoplasmic compartments, several proteins may exhibit multiple subcellular locations (FIGURE A-23B). To facilitate annotation, proteins from different subcellular compartments were classified into four distinct subcellular topological groups: the cytoplasm (F1, A, R, and N), the inner membrane (B), the periplasm (I, G, F2, F3, and E), and the outer membrane/extracellular group (H, X, and F4) (FIGURE A-23C). These subcellular topological classifications considered most of the dynamic protein movement, providing robust insights into their cellular localization. For mid-logarithmic growth phases, a total of 334 and 298 proteins were identified in OMVs isolated from supernatant obtained from glucose and acetate cultivation, respectively (FIGURE VI-6A). Among these proteins, 258 (69%) were identified in both isolations, while 76 proteins were identified exclusively under glucose and 40 proteins exclusively under acetate conditions

(FIGURE VI-6A). The relevance of the exclusively identified OMV proteins remains uncertain due to the absence of clear functional or pathway enrichments.

Comparison of the subcellular locations of identified OMV proteins showed that OMVs isolated from supernatant obtained from acetate or glucose cultivation were enriched in periplasmic and cytoplasmic proteins, whereas only 9% of the total proteins were classified as outer membrane or extracellular proteins (FIGURE VI-6B-C). The majority of identified cytoplasmic proteins were associated with ribosomal functions or involved in glycolysis, such as enolase (P0A6P9), glyceraldehyde-3-phosphate dehydrogenase (P0A9B2), and phosphoglycerate kinase (P0A799). However, most other abundant and essential cytoplasmic proteins required for bacterial survival were not detected (Goodall et al., 2018).



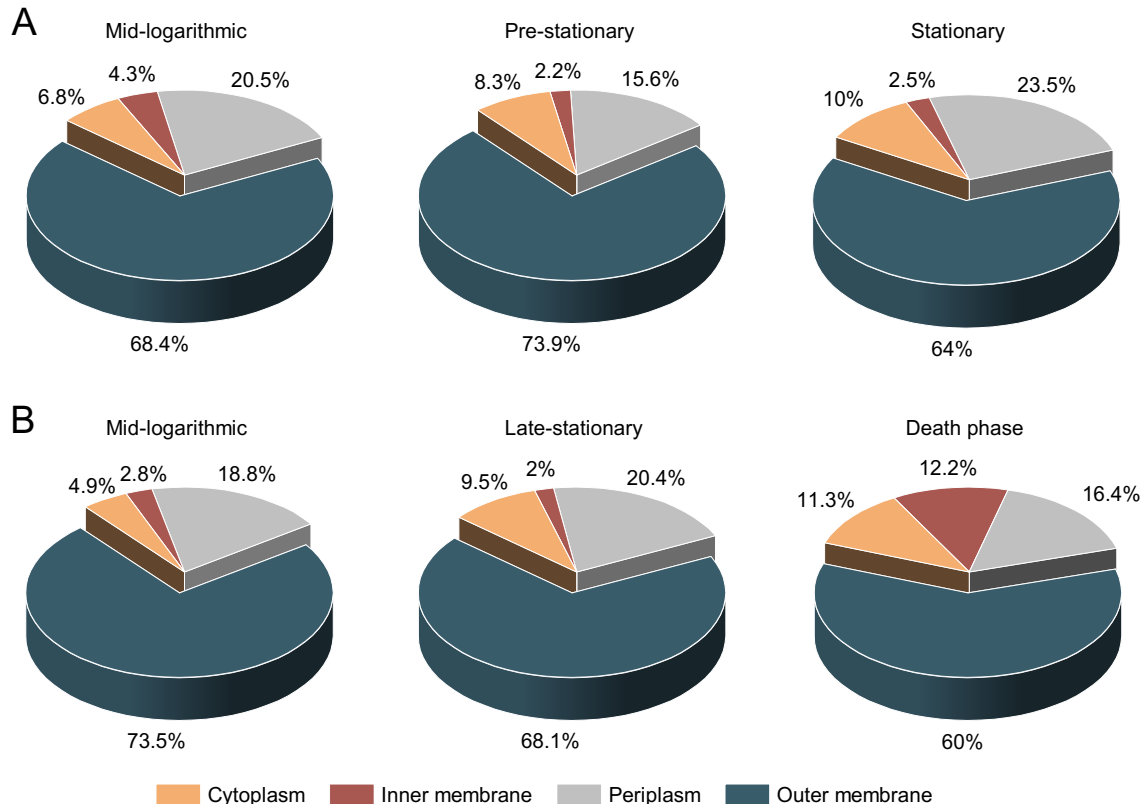
**FIGURE VI-6 | Subcellular Topological Distribution of OMV Proteins based on Total Protein Identifications.** (A) Overlap of proteins identified in OMV isolated at mid-logarithmic growth from glucose and acetate cultivation. (B) Subcellular topological distribution of OMVs isolated at mid-logarithmic growth from glucose and (C) acetate cultivation

Determination of median abundance values from the two independent OMV isolates, each measured in three technical replicates, allowed evaluation of protein abundance within subcellular topological categories. Comparing the distribution based on the median abundance with the distribution based on the total number of identified proteins of each subcellular topological category revealed a distinct pattern (FIGURE VI-7). For OMVs isolated from supernatant obtained during the mid-logarithmic growth phase of glucose cultivation, 68.4% of the median abundance originated from the outer membrane/extracellular, 20.5% from periplasmic, 4.3% from the inner membrane, and 6.8% from cytoplasmic proteins (FIGURE VI-7A). Similarly, OMVs isolated from supernatant obtained during acetate cultivation exhibited a distribution where 73.5% of the median abundance originated from outer membrane/extracellular, 18.8% from periplasmic, 2.8% from inner membrane, and 4.9% from cytoplasmic proteins (FIGURE VI-7B). This observation suggests that, while a substantial number of cytoplasmic proteins were present within the OMVs isolated from supernatant obtained from glucose (143 proteins) or acetate cultivation (119 proteins), their median abundance is notably lower compared to that of the 29 glucose or 27 acetate outer membrane/extracellular proteins. However, it is essential to note that solely relying on median

abundance may underestimate the significance of less abundant proteins which could also play important functional roles or contribute to biological processes. Therefore, both distributions of the subcellular topological categories should be considered complementary, providing distinct insights into protein abundance and diversity.

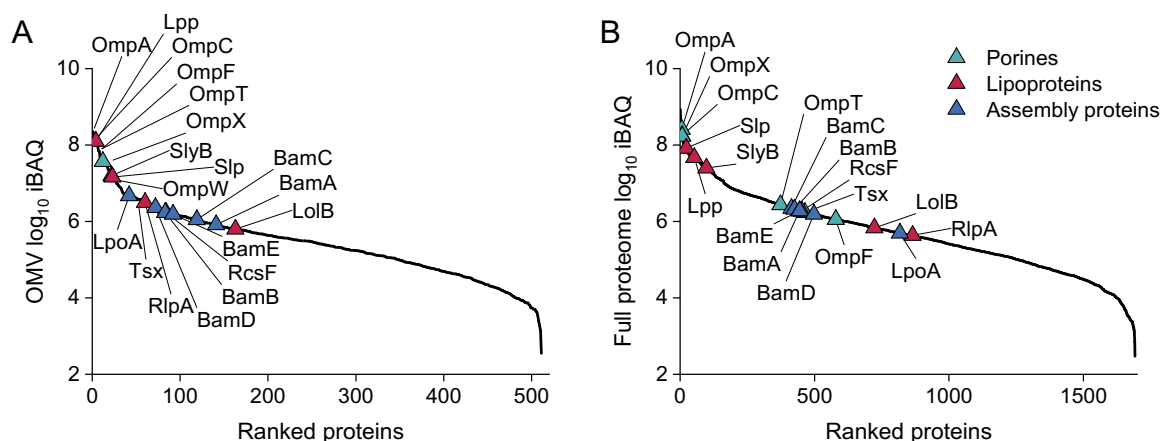
Despite an increase in the overall protein content of the OMV isolates (from 0.23 to 0.83  $\mu\text{g}/\mu\text{l}$ ), the subcellular topological distribution of OMVs isolated from supernatant obtained from glucose cultivation remained stable throughout the mid-logarithmic to stationary phase (FIGURE VI-7A). As cultivation time increased, only a 3.2% increase in the median abundance of cytoplasmic proteins and a 4.8% decrease in outer membrane/extracellular proteins were observed for OMVs isolated from supernatant obtained from glucose cultivation (FIGURE VI-7A). In contrast, OMVs isolated from supernatant obtained from acetate cultivation exhibited a notable 13.5% decrease in the median abundance of outer membrane/extracellular proteins, accompanied by a 6.4% increase in cytoplasmic proteins and a 9.4% increase in inner membrane proteins spanning the mid-logarithmic to death phase (FIGURE VI-7B).

The decrease in outer membrane/extracellular proteins may be attributed to the direct growth inhibition that occurred upon entering the stationary phase (FIGURE A-22). Additionally, the activation of programmed cell death mechanisms may have contributed to the increase in cytoplasmic and inner membrane proteins (Juodeikis and Carding, 2022).



**FIGURE VI-7 | Subcellular Topological Distribution of OMV Proteins based on Median Abundance.** (A) OMVs isolated from supernatant obtained during the mid-logarithmic, pre-stationary, and stationary growth phases of glucose cultivation. (B) OMVs isolated from supernatant obtained during the mid-logarithmic, late-stationary, and death growth phases of acetate cultivation.

Examination of major components of outer membranes (TABLE A-10), suggested as ubiquitous markers for OMV validation (Daleke-Schermerhorn et al., 2014; Hong et al., 2019), revealed a consistent increase in iBAQ values in OMVs isolated from supernatant obtained from glucose cultivation (FIGURE VI-8A). Certain marker proteins, such as OmpT and OmpF, exhibited increased iBAQ values in isolated OMV samples (FIGURE VI-8A) compared to full proteome preparations of *E. coli* (FIGURE VI-8B), which may suggest the selective sorting and encapsulation of certain proteins into OMVs. Interestingly, other marker proteins, such as OmpA, OmpX, and OmpC, consistently showed high iBAQ values for both proteomes.



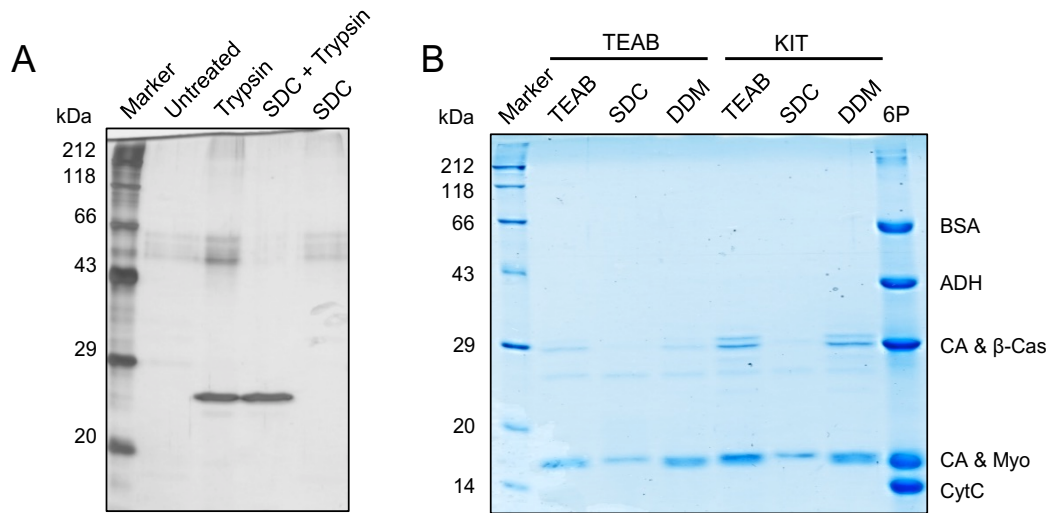
**FIGURE VI-8 | iBAQ Distribution of Potential *E. coli* OMV Protein Markers.** (A) The OMV proteome isolated from the supernatant obtained from glucose cultivation and (B) the complete *E. coli* proteome. Protein names are listed in TABLE A-10.

### 3.4 Improving OMV Lysis and Trypsin-Based Digestion

The impact of the non-ionic detergent SDC on OMV integrity and the susceptibility of OMV proteins to enzymatic degradation were evaluated (FIGURE VI-1). OMV samples subjected to tryptic digestion exhibited protein bands within the 43 to 66 kDa range (FIGURE VI-9A). The application of 2% (w/v) SDC before enzymatic digestion resulted in the reduction of these bands. A control experiment involving only 2% SDC confirmed that the observed bands corresponded to proteins encapsulated within the OMVs. Protein bands with a molecular mass of approximately 23.5 kDa correspond to trypsin itself. Overall, the results indicate the capability of SDC to induce OMV lysis and the subsequent susceptibility of proteins initially protected within OMVs to enzymatic degradation.

Evaluation of SDC and DDM impact on tryptic digestion of the six-protein mixture, compared against a TEAB control, indicated improved tryptic digestion using 0.5% (w/v) SDC (FIGURE VI-9B). Conversely, 0.01% (w/v) DDM exhibited protein bands with similar intensities to those of the TEAB control, suggesting minor differences in protein abundance between the two digestion conditions (FIGURE VI-9B). Especially for TEAB-buffered OMV elution buffer (KIT),

the results indicate an improved trypsin digestion efficiency in the presence of SDC compared to DDM or TEAB.

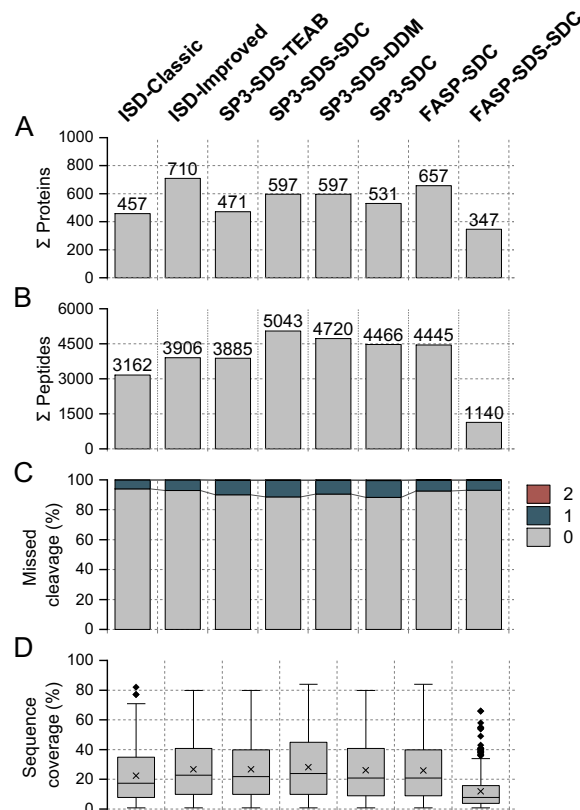


**FIGURE VI-9 | Evaluation of Non-Ionic Detergents on OMV Lysis and Tryptic Digestion.** (A) SDC-Mediated OMV Lysis. From left to right: molecular weight ladder (in kDa), untreated OMV sample, tryptic-treated OMV sample, OMV sample treated with 2% (w/v) SDC followed by tryptic treatment, and OMV sample treated with 2% (w/v) SDC. (B) Non-ionic detergents effect on tryptic digestion using TEAB-buffered OMV elution buffer (KIT) or 100 mM TEAB (TEAB).

To improve trypsin-based digestion of OMV samples, various sample preparation protocols were tested, including on-bead SP3, on-membrane FASP, and in-solution methods. These protocols utilized both SDC and DDM for effective lysis and tryptic digestion of OMVs. Total identified protein groups ranged from 347 to 710 (FIGURE VI-10A), with peptide identifications ranging from 1140 to 5043 (FIGURE VI-10B). The ISD-improved protocol, employing SDC for both OMV lysis and tryptic digestion, exhibited the highest number of protein identifications (FIGURE VI-10A). The SP3 protocol, integrating SDS for OMV lysis and SDC for tryptic digestion (SP3-SDS-SDC), exhibited the most identified peptides (FIGURE VI-10B).

Comparison of different SP3 protocols showed that the integration of SDC or DDM for tryptic digestion resulted in higher protein and peptide identifications compared to detergent-free digestion using only TEAB. (FIGURE VI-10A-B). Despite the potential interference from SDS, which could have affected trypsin's activity and led to incomplete protein digestion (Masuda et al., 2008), less than 3% of the identified peptides had 2 missed cleavages (FIGURE VI-10C). These results indicated efficient removal of SDS and successful protein digestion. An exception was the FASP-SDS-SDC method, which exhibited the lowest number of protein and peptide identifications, as well as the lowest sequence coverage (FIGURE VI-10D), across all three biological replicates. Small leftovers of SDS, which may not have been sufficiently removed, could potentially have interfered with tryptic digestion or led to signal suppression during LC-MS measurements (Rundlett and Armstrong, 1996; Masuda et al., 2008).

Furthermore, sample loss, due to small leftovers remaining in the filter reservoir or the filter itself, might have occurred during sample processing.

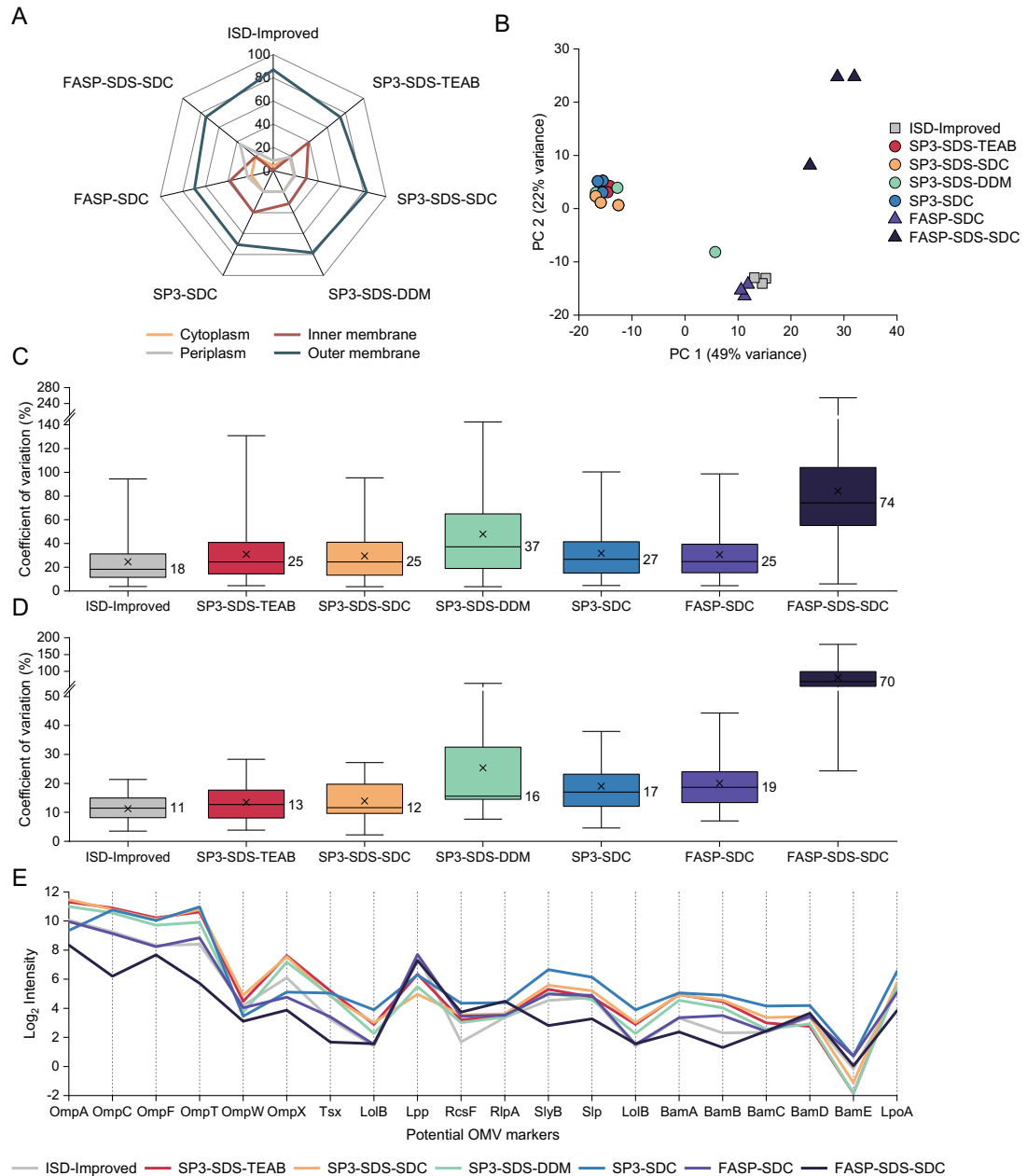


**FIGURE VI-10 | Peptide and Protein Identification Data of Different OMV Sample Preparations.** (A) Protein identifications, (B) peptide identifications, (C) missed cleavage events, and (D) protein sequence coverage. Box-and-whisker plots capture lower quartile and upper quartile with the median displayed as a horizontal line and the mean depicted as a cross; whiskers represent minimum and maximum values that fall within 1.5 times the interquartile range.

Overall, all methods achieved similar sequence coverages of 23% (FIGURE VI-10D), with each method covering largely similar fractions of the *E. coli* proteome (FIGURE A-25). Compared to the detergent-free ISD-classic protocol which employed freeze-thawing for OMV lysis, the integration of SDS or SDC for OMV lysis and SDC or DDM for digestion notably improved peptide and protein identifications, resulting in higher sequence coverage. For this reason, subsequent analysis focused on evaluating these improved methods.

Notable methodological variations were observed in the subcellular topological distribution of OMV proteins, with cytoplasmic proteins ranging from 2.2% to 4.2%, inner membrane proteins from 0.2% to 27.8%, periplasmic proteins from 2.4% to 24%, and outer membrane proteins from 64% to 87% (FIGURE VI-11A). The ISD-improved method exhibited the highest abundance of outer membrane proteins (87%) and the lowest abundance of inner membrane proteins (0.2%) (FIGURE VI-11A). Interestingly, FASP-SDC and SP3-SDC, which also employed SDC for digestion and lysis, showed the highest abundance of inner membrane proteins at 26% and

28%, respectively (FIGURE VI-11A). A principal component analysis (PCA) revealed distinct clusters, particularly for samples derived from the SP3 method (FIGURE VI-11B). This indicates that the magnetic bead-mediated protein pulldown is a method-specific approach to protein extraction, regardless of the applied lysis or digestion conditions.



**FIGURE VI-11 | Variability of Different OMV Sample Preparations.** (A) Subcellular topological distribution based on the relative abundance, (B) principal-component analysis (PCA), and (C) coefficient of variation of identified proteins. (D) Coefficient of variation and (E)  $\log_2$  normalized intensity of quantified potential OMV marker proteins.

Variance among most biological replicates, with exceptions observed in one biological replicate for SP3-SDS-DDM and FASP-SDS-SDC, suggests that the majority of sample preparations had a comparable level of variability (FIGURE VI-11B). Furthermore, correlation analysis of technical replicates indicated that the majority exhibited a high correlation, close to 0.98 (FIGURE A-26). Overall, the ISD-improved protocol exhibited the lowest median coefficient of variation (CV) of 18% (FIGURE VI-11C). The other methods, except for SP3-SDS-DDM and FASP-SDS-SDC had CV values between 25% and 27% (FIGURE VI-11C).

Comparison of median CV values for the potential OMV marker proteins revealed that the ISD-improved protocol exhibited the highest quantitative reproducibility at 11%, closely followed by SP3-SDS-SDC with 12%, and SP3-SDS-TEAB with 13% (FIGURE VI-11D). All sample preparation methods successfully quantified the 20 potential OMV marker proteins, including the low-abundance BamE (FIGURE VI-11E), highlighting the effectiveness of the chosen methodologies in quantifying the abundance of both prominent and less abundant OMV marker proteins.

## 4 Discussion and Conclusion

### 4.1 Biophysical and Biological Characteristics of *E. coli* OMVs

Despite variations in culture conditions or differences in the loaded supernatant, all isolated *E. coli* OMVs exhibited comparable particle sizes ranging from 95 to 117 nm (TABLE VI-2 and FIGURE VI-4C). The isolated OMVs had a relatively narrow size distribution of particles (FIGURE VI-3B and FIGURE A-24), suggesting a more monodisperse sample. While some degree of polydispersity is expected for OMVs due to natural biological variability, the observed uniformity might be indicative of well-isolated and purified OMVs. This may be attributed to the specific conditions under which the OMVs were isolated and purified, ultimately reflecting the effectiveness of the isolation kit used or the particular characteristics of the bacterial strain from which they were derived. Further, although a volume three times that of the recommended kit volumes was used, the observed purity ratio remained relatively constant at  $4.2 \times 10^6$  particles/ $\mu\text{g}$  protein (FIGURE VI-4D). While the OMV isolations did not meet the recommended benchmark of  $3 \times 10^{10}$  particles/ $\mu\text{g}$  protein for high-quality EV preparations (Webber and Clayton, 2013), this deviation could be attributed to multiple factors. Specifically, reference ratios for *E. coli* OMVs isolated from 800 ml to 160 L of supernatant range from  $10^7$ – $10^{10}$  particles/ $\mu\text{g}$  protein (Hong et al., 2019; Bittel et al., 2021; Won et al., 2023), indicating a dependence on the volume of culture supernatant. Notably, the volumes of *E. coli* supernatant used to isolate OMVs in this study only ranged from 20 to 60 ml. Moreover, SDS-PAGE analysis and total ion chromatogram intensities in MS measurements indicated that the calculated protein concentration (as determined by BCA assays) underestimated the actual OMV protein content. Additionally, NTA measurements do not differentiate between vesicles and non-vesicular structures. Consequently, all detected particles were assumed to be vesicles, although this assumption may not have been entirely accurate. Therefore, the presence of protein aggregates, salt crystals, and other components could have resulted in an overestimation of vesicle counts.

Recent publications that utilized the ExoBacteria OMV Isolation Kit to isolate *E. coli* OMVs often lack sufficient information on particle and protein concentrations (Diallo et al., 2022; Kim et al., 2022; Brückner et al., 2023). This absence of comprehensive data on purity assessments limits the determination of purity ratios and complicates the establishment of a conclusive rationale for the observed purity ratio. Addressing these limitations in future research is crucial for improving the reliability and comparability of findings in the field.

Overall, the results of this study suggest that the ExoBacteria OMV Isolation Kit's full potential may not have been fully utilized, suggesting opportunities for improvement. Optimizing the isolation procedure and utilizing larger supernatant volumes could potentially further increase the number of OMV particles and protein quantities.

The *in vitro* wound-closure assay using Caco-2 cells indicated that the wound closure may be inhibited in a dose-dependent manner by *E. coli* OMVs (FIGURE VI-5). This result is consistent with the inhibitory effect of *E. coli* OMVs on the proliferation of intestinal epithelial cells (Cañas et al., 2016; Patten et al., 2017), as well as the dose-dependent interaction with macrophage cells, resulting in their internalization and subsequent induction of reactive oxygen species (ROS) and inflammatory cytokine production (Guangzhang et al., 2023). While specific OMV proteins involved in various cellular processes, such as signal transduction and immune response modulation, such as OmpA (P0A910) and OmpC (P06996), have been identified, it is important to consider the potential introduction of co-isolated components or artifacts during the isolation process. These factors might have influenced the observed effects on wound healing. Therefore, to elucidate the specific contributions of OMVs and assess any potential effects mediated by co-isolated components, future experiments should include comparisons between intact OMVs and detergent-treated samples (György et al., 2011). Additionally, techniques offering high-resolution images of individual OMVs, such as electron microscopy, should be employed to assess both close-up and wide-field morphological features (Théry et al., 2018).

## 4.2 Proteomic Characterization of *E. coli* OMVs

Bacteria continuously adjust their proteome composition throughout all growth stages and in response to environmental changes such as temperature and media composition (Serbanescu et al., 2022; Kratz and Banerjee, 2023). Likewise, environmental shifts can also influence the production of OMVs, their physical properties such as particle size or vesicle charge, and their protein composition (Bai et al., 2014; Orench-Rivera and Kuehn, 2016; Johnston et al., 2023). Although OMVs are generated at all growth stages, the maximum production rate, observed across various bacterial species, consistently occurs during the pre-stationary phase (Orench-Rivera and Kuehn, 2016; Sharif et al., 2021).

Previous quantitative proteome analysis of *E. coli* under acetate and glucose culture conditions revealed significant changes in proteins associated with central carbon metabolism, amino acids, and protein synthesis (Treitz et al., 2016). While changes in protein abundance and subcellular topological distribution of OMV proteomes were observed to be media- and growth-phase-dependent (FIGURE VI-7), specific enrichments among OMV proteins exclusively identified during mid-logarithmic growth under these conditions remained elusive. These findings suggest that, while these specific culture conditions may alter the intracellular proteome composition, they might not necessarily affect the protein composition of *E. coli* OMVs.

A comparison between full proteome and OMV analysis from mid-logarithmic glucose cultures revealed notable differences in subcellular topological identifications. The full proteome

analysis detected cytoplasmic proteins (716 out of 2812), inner membrane proteins (45 out of 970), periplasmic proteins (82 out of 461), and outer membrane/extracellular proteins (14 out of 93). Despite a lower total protein count, the isolated OMVs exhibited an enrichment of periplasmic proteins (157 out of 461) and outer membrane/extracellular proteins (29 out of 93), alongside a reduced number of inner membrane proteins (10 out of 970). Similarly, proteomic studies on *E. coli* OMVs consistently reported an enrichment in outer membrane proteins and the exclusion of inner membrane proteins (Hong et al., 2019; Won et al., 2023). The high relative abundance of outer membrane/extracellular proteins further suggests a selective enrichment of outer membrane proteins (FIGURE VI-7).

The identification of several cytoplasmic proteins (143 out of 2812), accounting for 42.2% of all identified proteins (FIGURE VI-6B), aligns with results from other OMV studies, typically ranging between 25% and 56% (Lee et al., 2007; Bai et al., 2014; Berzosa et al., 2022). Despite their lower abundance compared to outer membrane/extracellular proteins (FIGURE VI-7), their presence in OMV samples is unexpected by definition (Toyofuku et al., 2019). However, it is crucial to recognize that despite the kit's name implying a preference for OMVs, it does not differentiate between distinct subtypes of bacterial extracellular vesicles (OMVs and O-IMVs) during the isolation process (Chamberlain et al., 2021). As a result, it may capture all bacterial extracellular vesicles in the final isolate. The possible formation of O-IMVs, which naturally incorporate cytoplasmic and inner membrane proteins during the vesiculation process (Pérez-Cruz et al., 2015; Toyofuku et al., 2019), could explain their identification. Additional processes such as transertion, involving coupled transcription-translation and protein translocation (Woldringh, 2002; Irastortza-Olaziregi and Amster-Choder, 2021), can associate ribosomal proteins, chaperones, and elongation factors with the inner membrane. Consequently, during the vesiculation of O-IMVs, these proteins may have a higher probability of becoming encapsulated within the vesicles. It is also worth considering the potential influence of contaminating cell fragments or debris from the isolation process. However, this seems unlikely as several rigorous proteomic analyses of OMVs have reported the presence of cytoplasmic proteins (Lee et al., 2007; Bai et al., 2014; Hong et al., 2019; Berzosa et al., 2022; Won et al., 2023). Therefore, it may be reasonable to consider the presence of cytoplasmic proteins in OMV analysis as common, rather than a result of contamination.

The quantitative iBAQ results from both OMV and full proteome samples indicated an enrichment of specific OMV marker proteins in OMV samples, while others consistently maintained similar abundance in both proteomes (FIGURE VI-8). A comparison of different sample preparation methods, detailed in the supplementary material of Doellinger and colleagues (Doellinger et al., 2020), assessed the performance of SPEED, SDS-based FASP, SP3, and urea-based in-solution digestion of the *E. coli* proteome. Analysis of their data revealed a consistent trend of the 20 marker proteins towards higher iBAQ values (FIGURE A-27). Although each sample preparation exhibited a bias towards certain proteins, the overall

consistency across different protocols suggested that the selected marker proteins generally exhibited higher abundance in full proteome analysis of *E. coli*.

Therefore, relying solely on identifying or quantifying OMV marker proteins through proteomic analysis may not effectively distinguish OMVs from full proteome samples. The dynamic and selective nature of protein sorting into OMVs, influenced by various factors (Loos et al., 2019), presents a significant challenge in the proteomics field for establishing robust and specific OMV markers. This challenge is further intensified when attempting to establish reliable proteomic OMV markers across different organisms, as the composition of OMVs frequently varies between bacteria and in response to different environmental conditions (Orench-Rivera and Kuehn, 2016; Juodeikis and Carding, 2022). For comprehensive OMV characterization, it is recommended to employ supplementary techniques such as electron microscopy, NTA measurements, and lipid analysis to complement proteomic analysis (Théry et al., 2018).

### 4.3 Optimized OMV Sample Preparation Workflow

While SDS remains a widely used anionic detergent, its application presents numerous challenges in downstream sample processing, including enzyme inactivation, LC separation disruption, ion suppression, and background noise (Rundlett and Armstrong, 1996; Masuda et al., 2008). Thus, efficient removal of SDS is crucial to ensuring successful MS analysis.

For this reason, non-ionic detergents like DDM and SDC offer multiple advantages over SDS. DDM, for instance, allows high organic solvent elution, eliminating the need for surfactant removal and thus making it compatible with mass spectrometry (Liu et al., 2015). In contrast, SDC co-elutes with tryptic peptides on RP columns, which can lead to ionization suppression and interference with peptide detection. This necessitates its removal through methods such as acid precipitation or phase transfer prior to LC-MS analysis (Masuda et al., 2008). SDC has been demonstrated to lyse EV subpopulations from human cell lines (Osteikoetxea et al., 2015), and several studies have explored various sample preparation methods incorporating non-ionic detergents for mass spectrometry-based proteomic analysis of barley leaves (Wang et al., 2018), HeLa cells (Varnavides et al., 2022) and rat liver mitochondrial samples (Leon et al., 2013). However, to the best of my knowledge, the utilization of SDC for lysis and proteolytic digestion of OMV samples from *E. coli* or other Gram-negative bacteria remains unexplored.

The results of this study demonstrate the effectiveness of 2% (w/v) SDC in lysing *E. coli* OMVs and enhancing tryptic digestion at 0.5% (w/v). Integrating SDC into various sample preparation methods, such as in-solution, on-bead SP3, or on-membrane, notably improved peptide and protein identifications (FIGURE VI-10). However, it's important to acknowledge potential limitations such as sample loss due to protein adherence to SP3 beads or FASP filters, which were not addressed in this study. Evaluating sample recovery efficiencies could further enhance the reliability of the methodology. The advantages of SDC, including its easy removal

prior to LC-MS measurements and absence of associated protein modification artifacts (Masuda et al., 2008; Leipert et al., 2021; Tsai et al., 2021; Varnavides et al., 2022), strongly support its use in future OMV analysis. Therefore, future studies should apply SDC for OMV lysis and integrate it for efficient tryptic digestion.

Compared to the detergent-free ISD-classic method, which employed freeze-thawing for OMV lysis and OMV elution buffer for proteolytic digestion, using SDC for OMV lysis and digestion resulted in a notable increase in protein and peptide identifications (FIGURE VI-10). Among the tested protocols, ISD-improved demonstrated a high abundance of outer membrane proteins (87%), with 710 proteins and 3906 peptides identified, on average 5.5 peptides per protein and a mean sequence coverage of 22.4% (FIGURE VI-10 and FIGURE VI-11A). The FASP-SDC protocol exhibited comparable results, with 657 proteins (8% fewer) and 4445 peptides (13% more), averaging 8.4 peptides per protein and achieving an average sequence coverage of 26.1% (FIGURE VI-10). Notably, it exhibited a higher abundance of inner membrane proteins (26%) compared to the ISD-improved method (0.2%) (FIGURE VI-11A). However, the varying abundance of OMV protein subcellular localizations, influenced by factors such as bacterial species, culture conditions, growth phase, OMV isolation method, sample processing procedures, and bioinformatic filtering (Lee et al., 2007; Bai et al., 2014; Hong et al., 2019; Berzosa et al., 2022; McMillan and Kuehn, 2023; Won et al., 2023), complicates the establishment of a universal standard. Consequently, determining the accuracy of subcellular topological categories for each method remains challenging.

Further analysis of the quantification accuracy revealed that the ISD-improved method exhibited the highest quantitative reproducibility, quantifying 2025 peptides with CV values below 20% (50% of all quantified peptides). In contrast, the FASP-SDC protocol quantified only 1119 peptides with CV values below 20% (26% of all quantified peptides). At the protein level, the ISD-improved method exhibited a 7% lower median CV for all proteins and an 8% lower median CV for the 20 OMV marker proteins compared to the FASP-SDC protocol (FIGURE VI-11).

Considering the potential need for efficiently handling large sample sets and the preference for quick, straightforward sample preparation methods incorporating minimal processing steps, the ISD-improved protocol emerges as the optimal choice. In this study, its simplicity, high identification rate, and robust quantitative reproducibility made it the preferred protocol for future analysis of *E. coli* OMVs.



- Abdulla, M., and Mohammed, N. (2022). A Review on Inflammatory Bowel Diseases: Recent Molecular Pathophysiology Advances. *Biol. Targets Ther.* Volume 16, 129–140.
- Abramowicz, A., Widlak, P., and Pietrowska, M. (2016). Proteomic analysis of exosomal cargo: the challenge of high purity vesicle isolation. *Mol. Biosyst.* 12, 1407–19.
- Adusumilli, R., and Mallick, P. (2017). Data Conversion with ProteoWizard msConvert. *Methods Mol. Biol.* 1550, 339–368.
- Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., et al. (2018). How many human proteoforms are there? *Nat. Chem. Biol.* 14, 206–214.
- Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 347–55.
- Aggarwal, S., Raj, A., Kumar, D., Dash, D., and Yadav, A. K. (2022). False discovery rate: the Achilles' heel of proteogenomics. *Brief. Bioinform.* 23, 1–15.
- Ahrens, C. H., Wade, J. T., Champion, M. M., and Langer, J. D. (2022). A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. *J. Bacteriol.* 204, e0035321.
- Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief. Bioinform.* 11, 253–264.
- Akashi, H., and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci.* 99, 3695–3700.
- Alasmar, R. M., Varadharajan, K., Shanmugakonar, M., and Al-Naemi, H. A. (2023). Early-Life Sugar Consumption Affects the Microbiome in Juvenile Mice. *Mol. Nutr. Food Res.* 67, 1–10.
- Almarza, O., Núñez, D., and Toledo, H. (2015). The DNA-binding protein HU has a regulatory role in the acid stress response mechanism in *Helicobacter pylori*. *Helicobacter* 20, 29–40.
- An, H., Douillard, F. P., Wang, G., Zhai, Z., Yang, J., Song, S., et al. (2014). Integrated Transcriptomic and Proteomic Analysis of the Bile Stress Response in a Centenarian-originated Probiotic *Bifidobacterium longum* BBMN68. *Mol. Cell. Proteomics* 13, 2558–2572.
- Andrews, S. J., and Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15, 193–204.
- Annesley, T. M. (2003). Ion suppression in mass spectrometry. *Clin. Chem.* 49, 1041–1044.
- Antoni, L., Nuding, S., Weller, D., Gersemann, M., Ott, G., Wehkamp, J., et al. (2013). Human colonic mucus is a reservoir for antimicrobial peptides. *J. Crohn's Colitis* 7, e652–e664.
- Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180.
- Ashniev, G. A., Petrov, S. N., Iablokov, S. N., and Rodionov, D. A. (2022). Genomics-Based Reconstruction and Predictive Profiling of Amino Acid Biosynthesis in the Human Gut Microbiome. *Microorganisms* 10, 740.
- Atkins, J. F., Loughran, G., Bhatt, P. R., Firth, A. E., and Baranov, P. V. (2016). Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* 44, 7007–78.
- Avershina, E., Storrø, O., Øien, T., Johnsen, R., Wilson, R., Egeland, T., et al. (2013). Bifidobacterial Succession and Correlation Networks in a Large Unselected Cohort of Mothers and Their Children. *Appl. Environ. Microbiol.* 79, 497–507.
- Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2, 2006.0008.
- Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., et al. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl. Acad. Sci.* 101, 15718–15723.

## BIBLIOGRAPHY

- Bai, J., Kim, S. I., Ryu, S., and Yoon, H. (2014). Identification and Characterization of Outer Membrane Vesicle-Associated Proteins in *Salmonella enterica* Serovar Typhimurium. *Infect. Immun.* 82, 4001–4010.
- Bai, J. P. F., Burckart, G. J., and Mulberg, A. E. (2016). Literature Review of Gastrointestinal Physiology in the Elderly, in Pediatric Patients, and in Patients with Gastrointestinal Diseases. *J. Pharm. Sci.* 105, 476–483.
- Ballegaard, M., Bjergstrøm, A., Brøndum, S., Hylander, E., Jensen, L., and Ladefoged, K. (1997). Self-Reported Food Intolerance in Chronic Inflammatory Bowel Disease. *Scand. J. Gastroenterol.* 32, 569–571.
- Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* 404, 939–965.
- Bartel, J., Varadarajan, A. R., Sura, T., Ahrens, C. H., Maaß, S., and Becher, D. (2020). Optimized Proteomics Workflow for the Detection of Small Proteins. *J. Proteome Res.* 19, 4004–4018.
- Basharat, A. R., Zang, Y., Sun, L., and Liu, X. (2023). TopFD: A Proteoform Feature Detection Tool for Top-Down Proteomics. *Anal. Chem.* 95, 8189–8196.
- Baughn, A. D., and Malamy, M. H. (2002). A mitochondrial-like aconitase in the bacterium *Bacteroides fragilis*: Implications for the evolution of the mitochondrial Krebs cycle. *Proc. Natl. Acad. Sci.* 99, 4662–4667.
- Beam, A., Clinger, E., and Hao, L. (2021). Effect of Diet and Dietary Components on the Composition of the Gut Microbiota. *Nutrients* 13, 2795.
- Beetham, C. M., Schuster, C. F., Kviatkovski, I., Santiago, M., Walker, S., and Gründling, A. (2024). Histidine transport is essential for the growth of *Staphylococcus aureus* at low pH. *PLOS Pathog.* 20, e1011927.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.
- Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* 5, 593–599.
- Berzosa, M., Nemeskalova, A., Calvo, A., Quincoces, G., Collantes, M., Pareja, F., et al. (2022). Oral Immunogenicity of Enterotoxigenic *Escherichia coli* Outer Membrane Vesicles Encapsulated into Zein Nanoparticles Coated with a Gantrez® AN-Mannosamine Polymer Conjugate. *Pharmaceutics* 14, 123.
- Biemann, K. (1992). Mass spectrometry of peptides and proteins. *Annu. Rev. Biochem.* 61, 977–1010.
- Bienvenut, W. V., Giglione, C., and Meinel, T. (2015). Proteome-wide analysis of the amino terminal status of *Escherichia coli* proteins at the steady-state and upon deformylation inhibition. *Proteomics* 15, 2503–2518.
- Bittel, M., Reichert, P., Sarfati, I., Dressel, A., Leikam, S., Uderhardt, S., et al. (2021). Visualizing transfer of microbial biomolecules by outer membrane vesicles in microbe-host-communication in vivo. *J. Extracell. vesicles* 10, e12159.
- Blankenhorn, D., Phillips, J., and Slonczewski, J. L. (1999). Acid- and Base-Induced Proteins during Aerobic and Anaerobic Growth of *Escherichia coli* Revealed by Two-Dimensional Gel Electrophoresis. *J. Bacteriol.* 181, 2209–2216.
- Bludau, I., Frank, M., Dörig, C., Cai, Y., Heusel, M., Rosenberger, G., et al. (2021). Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat. Commun.* 12, 3810.
- Bogaert, A., Fernandez, E., and Gevaert, K. (2020). N-Terminal Proteoforms in Human Disease. *Trends Biochem. Sci.* 45, 308–320.
- Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A., and Pevzner, P. A. (2013). N-terminal Protein Processing: A Comparative Proteogenomic Analysis. *Mol. Cell. Proteomics* 12, 14–28.
- Boyd, D. A., Cvitkovitch, D. G., Bleiweis, A. S., Kiriukhin, M. Y., Debabov, D. V., Neuhaus, F. C., et al. (2000). Defects in D-alanyl-lipoteichoic acid synthesis in *Streptococcus mutans* results in acid sensitivity. *J. Bacteriol.* 182, 6055–6065.

- Brademan, D. R., Riley, N. M., Kwiecien, N. W., and Coon, J. J. (2019). Interactive Peptide Spectral Annotator: A Versatile Web-based Tool for Proteomic Applications. *Mol. Cell. Proteomics* 18, S193–S201.
- Brochu, D., and Vadeboncoeur, C. (1999). The HPr(Ser) Kinase of *Streptococcus salivarius*: Purification, Properties, and Cloning of the hprK Gene. *J. Bacteriol.* 181, 709–717.
- Brown, E. M., Ke, X., Hitchcock, D., Jeanfavre, S., Avila-Pacheco, J., Nakata, T., et al. (2019). Bacteroides-Derived Sphingolipids Are Critical for Maintaining Intestinal Homeostasis and Symbiosis. *Cell Host Microbe* 25, 668–680.e7.
- Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., et al. (2016). Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. *Nature* 533, 543–546.
- Brückner, S., Müller, F., Schadowski, L., Kalle, T., Weber, S., Marino, E. C., et al. (2023). (p)ppGpp and moonlighting RNases influence the first step of lipopolysaccharide biosynthesis in *Escherichia coli*. *microLife* 4, 1–18.
- Bui, T. P. N., Mannerås-Holm, L., Puschmann, R., Wu, H., Troise, A. D., Nijse, B., et al. (2021). Conversion of dietary inositol into propionate and acetate by commensal *Anaerostipes* associates with host health. *Nat. Commun.* 12, 1–16.
- Burger, B., Vaudel, M., and Barsnes, H. (2021). Importance of Block Randomization When Designing Proteomics Experiments. *J. Proteome Res.* 20, 122–128.
- Caballero, S., Kim, S., Carter, R. A., Leiner, I. M., Sušac, B., Miller, L., et al. (2017). Cooperating Commensals Restore Colonization Resistance to Vancomycin-Resistant *Enterococcus faecium*. *Cell Host Microbe* 21, 592–602.
- Caldas, T., Laalami, S., and Richarme, G. (2000). Chaperone Properties of Bacterial Elongation Factor EF-G and Initiation Factor IF2. *J. Biol. Chem.* 275, 855–860.
- Cañas, M.-A., Giménez, R., Fábrega, M.-J., Toloza, L., Baldomà, L., and Badia, J. (2016). Outer Membrane Vesicles from the Probiotic *Escherichia coli* Nissle 1917 and the Commensal ECOR12 Enter Intestinal Epithelial Cells via Clathrin-Dependent Endocytosis and Elicit Differential Effects on DNA Damage. *PLoS One* 11, e0160374.
- Candelli, M., Franza, L., Pignataro, G., Ojetti, V., Covino, M., Piccioni, A., et al. (2021). Interaction between Lipopolysaccharide and Gut Microbiota in Inflammatory Bowel Diseases. *Int. J. Mol. Sci.* 22, 6242.
- Cantarel, B. I., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* 37, 233–238.
- Cappa, F., Cattivelli, D., and Cocconcelli, P. S. (2005). The *uvrA* gene is involved in oxidative and acid stress responses in *Lactobacillus helveticus* CNBL1156. *Res. Microbiol.* 156, 1039–1047.
- Capri, J., and Whitelegge, J. P. (2017). “Full Membrane Protein Coverage Digestion and Quantitative Bottom-Up Mass Spectrometry Proteomics,” in *Methods in molecular biology (Clifton, N.J.)*, (Humana Press Inc.), 61–67.
- Carbonara, K., Andonovski, M., and Coorssen, J. R. (2021). Proteomes Are of Proteoforms: Embracing the Complexity. *Proteomes* 9, 38.
- Cardon, T., Fournier, I., and Salzert, M. (2021). Shedding Light on the Ghost Proteome. *Trends Biochem. Sci.* 46, 239–250.
- Carr, J. F., Hamburg, D.-M., Gregory, S. T., Limbach, P. A., and Dahlberg, A. E. (2006). Effects of Streptomycin Resistance Mutations on Posttranslational Modification of Ribosomal Protein S12. *J. Bacteriol.* 188, 2020–2023.
- Casabon, I., Couture, M., Vaillancourt, K., and Vadeboncoeur, C. (2006). Synthesis of HPr(Ser-P)(His~P) by Enzyme I of the Phosphoenolpyruvate: Sugar Phosphotransferase System of *Streptococcus salivarius*. *Biochemistry* 45, 6692–6702.

## BIBLIOGRAPHY

- Cassidy, L., Helbig, A. O., Kaulich, P. T., Weidenbach, K., Schmitz, R. A., and Tholey, A. (2021a). Multidimensional separation schemes enhance the identification and molecular characterization of low molecular weight proteomes and short open reading frame-encoded peptides in top-down proteomics. *J. Proteomics* 230, 103988.
- Cassidy, L., Kaulich, P. T., Maaß, S., Bartel, J., Becher, D., and Tholey, A. (2021b). Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open reading frame-encoded peptides. *Proteomics* 21, e2100008.
- Cassidy, L., Kaulich, P. T., and Tholey, A. (2019). Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes. *J. Proteome Res.* 18, 1725–1734.
- Cassidy, L., Kaulich, P. T., and Tholey, A. (2023). Proteoforms expand the world of microproteins and short open reading frame-encoded peptides. *iScience* 26, 106069.
- Cassidy, L., Prasse, D., Linke, D., Schmitz, R. A., and Tholey, A. (2016). Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon *Methanosarcina mazei*. *J. Proteome Res.* 15, 3773–3783.
- Castellana, N., and Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteomics* 73, 2124–2135.
- Chamberlain, C. A., Hatch, M., and Garrett, T. J. (2021). Extracellular Vesicle Analysis by Paper Spray Ionization Mass Spectrometry. *Metabolites* 11, 308.
- Chan, D. I., and Vogel, H. J. (2010). Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* 430, 1–19.
- Chapman, J. D., Goodlett, D. R., and Masselon, C. D. (2014). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* 33, 452–470.
- Chen, J., Li, P., Zhang, T., Xu, Z., Huang, X., Wang, R., et al. (2022). Review on Strategies and Technologies for Exosome Isolation and Purification. *Front. Bioeng. Biotechnol.* 9, 811971.
- Chen, L., Wang, W., Zhou, R., Ng, S. C., Li, J., Huang, M., et al. (2014). Characteristics of Fecal and Mucosa-Associated Microbiota in Chinese Patients With Inflammatory Bowel Disease. *Medicine (Baltimore)*. 93, e51.
- Chen, Y., Cao, X., Loh, K. H., and Slavoff, S. A. (2023). Chemical labeling and proteomics for characterization of unannotated small and alternative open reading frame-encoded polypeptides. *Biochem. Soc. Trans.* 51, 1071–1082.
- Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* 33, W72–W76.
- Chung, C.-R., Kuo, T.-R., Wu, L.-C., Lee, T.-Y., and Horng, J.-T. (2020). Characterization and identification of antimicrobial peptides with different functional activities. *Brief. Bioinform.* 21, 1098–1114.
- Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009). Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* 6, 786–787.
- Cortes, H. J., Pfeiffer, C. D., Richter, B. E., and Stevens, T. S. (1987). Porous ceramic bed supports for fused silica packed capillary columns used in liquid chromatography. *J. High Resolut. Chromatogr.* 10, 446–448.
- Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* 13, 2513–2526.
- Creasy, D. M., and Cottrell, J. S. (2004). Unimod: Protein modifications for mass spectrometry. *Proteomics* 4, 1534–1536.
- Cronan, J. E., and Thomas, J. (2009). Bacterial fatty acid synthesis and its relationships with polyketide synthetic pathways. *Methods Enzymol.* 459, 395–433.

- Crouch, L. I., Liberato, M. V., Urbanowicz, P. A., Baslé, A., Lamb, C. A., Stewart, C. J., et al. (2020). Prominent members of the human gut microbiota express endo-acting O-glycanases to initiate mucin breakdown. *Nat. Commun.* 11, 4017.
- Čuklina, J., Lee, C. H., Williams, E. G., Sajic, T., Collins, B. C., Rodríguez Martínez, M., et al. (2021). Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol. Syst. Biol.* 17, e10240.
- Cuskin, F., Lowe, E. C., Temple, M. J., Zhu, Y., Cameron, E. A., Pudlo, N. A., et al. (2015). Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature* 517, 165–169.
- Daleke-Schermerhorn, M. H., Felix, T., Soprova, Z., ten Hagen-Jongman, C. M., Vikström, D., Majlessi, L., et al. (2014). Decoration of Outer Membrane Vesicles with Multiple Antigens by Using an Autotransporter Approach. *Appl. Environ. Microbiol.* 80, 5854–5865.
- Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., and Huber, C. G. (2007). Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: An alternative approach to high-resolution peptide separation for shotgun proteome analysis. *J. Proteome Res.* 6, 4363–4373.
- Deregibus, M. C., Figliolini, F., D'Antico, S., Manzini, P. M., Pasquino, C., De Lena, M., et al. (2016). Charge-based precipitation of extracellular vesicles. *Int. J. Mol. Med.* 38, 1359–1366.
- Diallo, I., Ho, J., Lalaouna, D., Massé, E., and Provost, P. (2022). RNA Sequencing Unveils Very Small RNAs With Potential Regulatory Functions in Bacteria. *Front. Mol. Biosci.* 9, 1–15.
- Diaz, M., del Rio, B., Ladero, V., Redruello, B., Fernández, M., Martin, M. C., et al. (2020). Histamine production in *Lactobacillus vaginalis* improves cell survival at low pH by counteracting the acidification of the cytosol. *Int. J. Food Microbiol.* 321, 108548.
- Dickhut, C., Feldmann, I., Lambert, J., and Zahedi, R. P. (2014). Impact of Digestion Conditions on Phosphoproteomics. *J. Proteome Res.* 13, 2761–2770.
- Do, M. H., Lee, E., Oh, M.-J., Kim, Y., and Park, H.-Y. (2018). High-Glucose or -Fructose Diet Cause Changes of the Gut Microbiota and Metabolic Disorders in Mice without Body Weight Change. *Nutrients* 10, 761.
- Doellinger, J., Schneider, A., Hoeller, M., and Lasch, P. (2020). Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. *Mol. Cell. Proteomics* 19, 209–222.
- Dolan, K. T., and Chang, E. B. (2017). Diet, gut microbes, and the pathogenesis of inflammatory bowel diseases. *Mol. Nutr. Food Res.* 61, 1600129.
- Donia, M. S., Cimermancic, P., Schulze, C. J., Wieland Brown, L. C., Martin, J., Mitreva, M., et al. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* 158, 1402–1414.
- Dougan, D. A., Truscott, K. N., and Zeth, K. (2010). The bacterial N-end rule pathway: expect the unexpected. *Mol. Microbiol.* 76, 545–58.
- Duboux, S., Pruvost, S., Joyce, C., Bogicevic, B., Muller, J. A., Mercenier, A., et al. (2023). The Pleiotropic Effects of Carbohydrate-Mediated Growth Rate Modifications in *Bifidobacterium longum* NCC 2705. *Microorganisms* 11, 588.
- Duncan, S. H., Hold, G. L., Harmsen, H. J. M., Stewart, C. S., and Flint, H. J. (2002). Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* 52, 2141–2146.
- Duncan, S. H., Louis, P., Thomson, J. M., and Flint, H. J. (2009). The role of pH in determining the species composition of the human colonic microbiota. *Environ. Microbiol.* 11, 2112–22.
- Dupree, E. J., Jayathirtha, M., Yorkey, H., Mihasan, M., Petre, B. A., and Darie, C. C. (2020). A critical review of bottom-up proteomics: The good, the bad, and the future of this field. *Proteomes* 8, 1–26.
- Durbin, K. R., Fornelli, L., Fellers, R. T., Doubleday, P. F., Narita, M., and Kelleher, N. L. (2016). Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa. *J. Proteome Res.* 15, 976–982.

## BIBLIOGRAPHY

- Ecklu-Mensah, G., Gilbert, J., and Devkota, S. (2022). Dietary Selection Pressures and Their Impact on the Gut Microbiome. *Cell. Mol. Gastroenterol. Hepatol.* 13, 7–18.
- Edwards, A. M., Addo, M. A., and Dos Santos, P. C. (2020). Extracurricular Functions of tRNA Modifications in Microorganisms. *Genes (Basel)*. 11, 907.
- Eguchi, Y., and Utsumi, R. (2014). Alkali metals in addition to acidic pH activate the EvgS histidine kinase sensor in *Escherichia coli*. *J. Bacteriol.* 196, 3140–3149.
- Ehrencrona, E., van der Post, S., Gallego, P., Recktenwald, C. V., Rodriguez-Pineiro, A. M., Garcia-Bonete, M.-J., et al. (2021). The IgGFC-binding protein FCGBP is secreted with all GDPH sequences cleaved but maintained by interfragment disulfide bonds. *J. Biol. Chem.* 297, 100871.
- Elhenawy, W., Debelyy, M. O., and Feldman, M. F. (2014). Preferential packing of acidic glycosidases and proteases into *Bacteroides* outer membrane vesicles. *MBio* 5, 1–12.
- Fabian, O., Bajer, L., Drastich, P., Harant, K., Sticova, E., Daskova, N., et al. (2023). A Current State of Proteomics in Adult and Pediatric Inflammatory Bowel Diseases: A Systematic Search and Review. *Int. J. Mol. Sci.* 24, 9386.
- Firman, J., Liu, L., Mahalak, K., Tanes, C., Bittinger, K., Tu, V., et al. (2022). The impact of environmental pH on the gut microbiota community structure and short chain fatty acid production. *FEMS Microbiol. Ecol.* 98, 1–9.
- Frank, D. N., St. Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* 104, 13780–13785.
- Fricker, L. D. (2015). Limitations of Mass Spectrometry-Based Peptidomic Approaches. *J. Am. Soc. Mass Spectrom.* 26, 1981–1991.
- Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., et al. (2006). The proteomics of N-terminal methionine cleavage. *Mol. Cell. Proteomics* 5, 2336–2349.
- Fujitani, Y., Shibata, T., and Tani, A. (2022). A Periplasmic Lanthanide Mediator, Lanmodulin, in *Methylobacterium aquaticum* Strain 22A. *Front. Microbiol.* 13, 921636.
- Fulcher, J. M., Makaju, A., Moore, R. J., Zhou, M., Bennett, D. A., De Jager, P. L., et al. (2021). Enhancing Top-Down Proteomics of Brain Tissue with FAIMS. *J. Proteome Res.* 20, 2780–2795.
- Gabere, M. N., and Noble, W. S. (2017). Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics* 33, 1921–1929.
- Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., and Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* 49, D274–D281.
- Gao, Q., Lu, S., Wang, Y., He, L., Wang, M., Jia, R., et al. (2023). Bacterial DNA methyltransferase: A key to the epigenetic world with lessons learned from proteobacteria. *Front. Microbiol.* 14, 1129437.
- Garcia-del Rio, D. F., Cardon, T., Eyckerman, S., Fournier, I., Bonnefond, A., Gevaert, K., et al. (2023). Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome. *iScience* 26, 105943.
- García, M. C. (2005). The effect of the mobile phase additives on sensitivity in the analysis of peptides and proteins by high-performance liquid chromatography-electrospray mass spectrometry. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* 825, 111–23.
- Gardner, M. L., and Freitas, M. A. (2021). Multiple imputation approaches applied to the missing value problem in bottom-up proteomics. *Int. J. Mol. Sci.* 22.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). “Protein Identification and Analysis Tools on the ExPASy Server,” in *The Proteomics Protocols Handbook*, (Totowa, NJ: Humana Press), 571–607.
- Genth, J., Kaleja, P., Treitz, C., Schäfer, K., Graspeuntner, S., Rupp, J., et al. (2022). The intracellular proteome of the gut bacterium *Bacteroides thetaiotaomicron* is widely unaffected by a switch from glucose to sucrose as main carbohydrate source. *Proteomics* 22, 1–6.

- Genth, J., Schäfer, K., Cassidy, L., Graspeuntner, S., Rupp, J., and Tholey, A. (2023). Identification of proteoforms of short open reading frame-encoded peptides in *Blautia producta* under different cultivation conditions. *Microbiol. Spectr.* 11, e0252823.
- Gerbasi, V. R., Melani, R. D., Abbatiello, S. E., Belford, M. W., Huguet, R., McGee, J. P., et al. (2021). Deeper Protein Identification Using Field Asymmetric Ion Mobility Spectrometry in Top-Down Proteomics. *Anal. Chem.* 93, 6323–6328.
- Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci.* 100, 6940–6945.
- Gershon, P. D. (2014). Cleaved and Missed Sites for Trypsin, Lys-C, and Lys-N Can Be Predicted with High Confidence on the Basis of Sequence Context. *J. Proteome Res.* 13, 702–709.
- Ghoreishi, F. S., Roghanian, R., and Emtiazi, G. (2023). Simultaneous Production of Antibacterial Protein and Lipopeptides in *Bacillus tequilensis*, Detected by MALDI-TOF and GC Mass Analyses. *Probiotics Antimicrob. Proteins* 15, 749–760.
- Gnanasekaran, P., and Pappu, H. R. (2023). “Affinity Purification-Mass Spectroscopy (AP-MS) and Co-Immunoprecipitation (Co-IP) Technique to Study Protein–Protein Interactions,” in *Methods in Molecular Biology*, (Humana Press Inc.), 81–85.
- Goodall, E. C. A., Robinson, A., Johnston, I. G., Jabbari, S., Turner, K. A., Cunningham, A. F., et al. (2018). The essential genome of *Escherichia coli* K-12. *MBio* 9, 1–18.
- Graciet, E., Hu, R.-G., Piatkov, K., Rhee, J. H., Schwarz, E. M., and Varshavsky, A. (2006). Aminoacyl-transferases and the N-end rule pathway of prokaryotic/eukaryotic specificity in a human pathogen. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3078–83.
- Guan, N., and Liu, L. (2020). Microbial response to acid stress: mechanisms and applications. *Appl. Microbiol. Biotechnol.* 104, 51–65.
- Guangzhang, C., Fangfang, F., Siqian, D., Xinyi, X., Xiaochuan, B., Yihan, R., et al. (2023). Outer membrane vesicles from *Escherichia coli* are efficiently internalized by macrophage cells and alter their inflammatory response. *Microb. Pathog.* 175, 105965.
- Guevremont, R. (2004). High-field asymmetric waveform ion mobility spectrometry: A new tool for mass spectrometry. *J. Chromatogr. A* 1058, 3–19.
- Guilloy, N., Brunet, M. A., Leblanc, S., Jacques, J.-F., Hardy, M.-P., Ehx, G., et al. (2023). OpenCustomDB: Integration of Unannotated Open Reading Frames and Genetic Variants to Generate More Comprehensive Customized Protein Databases. *J. Proteome Res.* 22, 1492–1500.
- Guo, M. S., and Gross, C. A. (2014). Stress-Induced Remodeling of the Bacterial Proteome. *Curr. Biol.* 24, R424–R434.
- Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R., et al. (2007). Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Res.* 17, 1362–1377.
- György, B., Módos, K., Pállinger, É., Pálóczi, K., Pásztói, M., Misják, P., et al. (2011). Detection and isolation of cell-derived microparticles are compromised by protein complexes resulting from shared biophysical parameters. *Blood* 117, 39–48.
- Ha, J.-H., Hauk, P., Cho, K., Eo, Y., Ma, X., Stephens, K., et al. (2018). Evidence of link between quorum sensing and sugar metabolism in *Escherichia coli* revealed via cocystal structures of LsrK and HPr. *Sci. Adv.* 4.
- Hadjeras, L., Heiniger, B., Maaß, S., Scheuer, R., Gelhausen, R., Azarderakhsh, S., et al. (2023). Unraveling the small proteome of the plant symbiont *Sinorhizobium meliloti* by ribosome profiling and proteogenomics. *microLife* 4, 1–22.
- Hanna, M. N., Ferguson, R. J., Li, Y.-H., and Cvitkovitch, D. G. (2001). *uvrA* Is an Acid-Inducible Gene Involved in the Adaptive Response to Low pH in *Streptococcus mutans*. *J. Bacteriol.* 183, 5964–5973.
- Haverland, N. A., Skinner, O. S., Fellers, R. T., Tariq, A. A., Early, B. P., LeDuc, R. D., et al. (2017). Defining Gas-Phase Fragmentation Propensities of Intact Proteins During Native Top-Down Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* 28, 1203–1215.

## BIBLIOGRAPHY

- Havugimana, P. C., Goel, R. K., Phanse, S., Youssef, A., Padhorny, D., Kotelnikov, S., et al. (2022). Scalable multiplex co-fractionation/mass spectrometry platform for accelerated protein interactome discovery. *Nat. Commun.* 13, 1–8.
- Hayes, C. S., Aoki, S. K., and Low, D. A. (2010). Bacterial Contact-Dependent Delivery Systems. *Annu. Rev. Genet.* 44, 71–90.
- Heaver, S. L., Le, H. H., Tang, P., Baslé, A., Mirretta Barone, C., Vu, D. L., et al. (2022). Characterization of inositol lipid metabolism in gut-associated Bacteroidetes. *Nat. Microbiol.* 7, 986–1000.
- Helmann, J. D. (2002). “The extracytoplasmic function (ECF) sigma factors,” in *Advances in microbial physiology*, 47–110.
- Heunis, T., Deane, S., Smit, S., and Dicks, L. M. T. (2014). Proteomic profiling of the acid stress response in lactobacillus plantarum 423. *J. Proteome Res.* 13, 4028–4039.
- Hochgräfe, F., Mostertz, J., Pöther, D.-C., Becher, D., Helmann, J. D., and Hecker, M. (2007). S-cysteinylation is a general mechanism for thiol protection of Bacillus subtilis proteins after oxidative stress. *J. Biol. Chem.* 282, 25981–5.
- Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O’Boyle, N. M., Murray-Rust, P., Thornton, J. M., et al. (2005). MACiE: A database of enzyme reaction mechanisms. *Bioinformatics* 21, 4315–4316.
- Homeyer, N., Essigke, T., Meiselbach, H., Ullmann, G. M., and Sticht, H. (2007). Effect of HPr phosphorylation on structure, dynamics, and interactions in the course of transcriptional control. *J. Mol. Model.* 13, 431–444.
- Hong, J., Dauros-Singorenko, P., Whitcombe, A., Payne, L., Blenkiron, C., Phillips, A., et al. (2019). Analysis of the Escherichia coli extracellular vesicle proteome identifies markers of purity and culture conditions. *J. Extracell. vesicles* 8, 1632099.
- Howell, S. J., Wilk, D., Yadav, S. P., and Bevins, C. L. (2003). Antimicrobial polypeptides of the human colonic epithelium. *Peptides* 24, 1763–1770.
- Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., and Krijgsveld, J. (2019). Singlepot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* 14, 68–85.
- Hungate, R. E. (1969). “A Roll Tube Method for Cultivation of Strict Anaerobes,” in *Methods in Microbiology*, 117–132.
- Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986). Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 83, 6233–6237.
- Hurtado-Rios, J. J., Carrasco-Navarro, U., Almanza-Pérez, J. C., and Ponce-Alquicira, E. (2022). Ribosomes: The New Role of Ribosomal Proteins as Natural Antimicrobials. *Int. J. Mol. Sci.* 23, 9123.
- Ikeyama, N., Murakami, T., Toyoda, A., Mori, H., Iino, T., Ohkuma, M., et al. (2020). Microbial interaction between the succinate-utilizing bacterium Phascolarctobacterium faecium and the gut commensal Bacteroides thetaiotaomicron. *Microbiologyopen* 9, e1111.
- Innovation, T., and Zecha, R. (2019). TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Mol. Cell. Proteomics* 18, 1468–1478.
- Irastortza-Olaziregi, M., and Amster-Choder, O. (2021). Coupled Transcription-Translation in Prokaryotes: An Old Couple With New Surprises. *Front. Microbiol.* 11, 624830.
- Ito, T., Gallegos, R., Matano, L. M., Butler, N. L., Hantman, N., Kaili, M., et al. (2020). Genetic and Biochemical Analysis of Anaerobic Respiration in Bacteroides fragilis and Its Importance In Vivo. *MBio* 11, e03238-19.
- Janion, C. (2001). Some aspects of the SOS response system--a critical survey. *Acta Biochim. Pol.* 48, 599–610.
- Jayaprakash, N. G., and Surolia, A. (2017). Role of glycosylation in nucleating protein folding and stability. *Biochem. J.* 474, 2333–2347.
- Jeong, J., Jung, Y., Na, S., Jeong, J., Lee, E., Kim, M.-S., et al. (2011). Novel Oxidative Modifications in Redox-Active Cysteine Residues. *Mol. Cell. Proteomics* 10, M110.000513.

- Jeong, K., Babović, M., Gorshkov, V., Kim, J., Jensen, O. N., and Kohlbacher, O. (2022). FLASHIda enables intelligent data acquisition for top-down proteomics to boost proteoform identification counts. *Nat. Commun.* 13, 4407.
- Jeong, K., Kim, J., Gaikwad, M., Hidayah, S. N., Heikau, L., Schlüter, H., et al. (2020). FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Syst.* 10, 213–218.e6.
- Jiang, H., Zhang, X., Chen, X., Aramsangtienchai, P., Tong, Z., and Lin, H. (2018). Protein Lipidation: Occurrence, Mechanisms, Biological Functions, and Enabling Technologies. *Chem. Rev.* 118, 919–988.
- Jiang, Y., Ren, F., Liu, S., Zhao, L., Guo, H., and Hou, C. (2016). Enhanced Acid Tolerance in *Bifidobacterium longum* by Adaptive Evolution: Comparison of the Genes between the Acid-Resistant Variant and Wild-Type Strain. *J. Microbiol. Biotechnol.* 26, 452–460.
- Johnston, E. L., Guy-Von Stieglitz, S., Zavan, L., Cross, J., Greening, D. W., Hill, A. F., et al. (2023). The effect of altered pH growth conditions on the production, composition, and proteomes of *Helicobacter pylori* outer membrane vesicles. *Proteomics*, e2300269.
- Jones, D. P., and Go, Y.-M. (2011). Mapping the cysteine proteome: analysis of redox-sensing thiols. *Curr. Opin. Chem. Biol.* 15, 103–112.
- Jordans, S., Jenko-Kokalj, S., Köhl, N. M., Tedelind, S., Sendt, W., Brömme, D., et al. (2009). Monitoring compartment-specific substrate cleavage by cathepsins B, K, L, and S at physiological pH and redox conditions. *BMC Biochem.* 10, 23.
- José-Estanyol, M., Gomis-Rüth, F. X., and Puigdomènech, P. (2004). The eight-cysteine motif, a versatile structure in plant proteins. *Plant Physiol. Biochem.* 42, 355–365.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.
- Jungblut, P. R., Thiede, B., and Schlüter, H. (2016). Towards deciphering proteomes via the proteoform, protein speciation, moonlighting and protein code concepts. *J. Proteomics* 134, 1–4.
- Junier, I., and Rivoire, O. (2016). Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. *PLoS One* 11, e0155740.
- Juodeikis, R., and Carding, S. R. (2022). Outer Membrane Vesicles: Biogenesis, Functions, and Issues. *Microbiol. Mol. Biol. Rev.* 86, e0003222.
- Kalet, C., Schäuble, S., Rinas, U., and Schuster, S. (2013). Metabolic costs of amino acid and protein production in *Escherichia coli*. *Biotechnol. J.* 8, 1105–1114.
- Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338, 1027–1036.
- Käll, L., Storey, J. D., and Noble, W. S. (2008). Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* 24, i42–i48.
- Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* 428, 726–731.
- Kanjee, U., and Houry, W. A. (2013). Mechanisms of Acid Resistance in *Escherichia coli*. *Annu. Rev. Microbiol.* 67, 65–81.
- Kato, K., Nakayoshi, T., Ishikawa, Y., Kurimoto, E., and Oda, A. (2021). Computational Analysis of the Mechanism of Nonenzymatic Peptide Bond Cleavage at the C-Terminal Side of an Asparagine Residue. *ACS Omega* 6, 30078–30084.
- Kaulich, P. T., Cassidy, L., Bartel, J., Schmitz, R. A., and Tholey, A. (2021). Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides. *J. Proteome Res.* 20, 2895–2903.
- Kaulich, P. T., Cassidy, L., and Tholey, A. (2024). Identification of proteoforms by top-down proteomics using two-dimensional low/low pH reversed-phase liquid chromatography-mass spectrometry. *Proteomics* 24, e2200542.
- Kaulich, P. T., Cassidy, L., Winkels, K., and Tholey, A. (2022a). Improved Identification of Proteoforms in Top-Down Proteomics Using FAIMS with Internal CV Stepping. *Anal. Chem.* 94, 3600–3607.

## BIBLIOGRAPHY

- Kaulich, P. T., Winkels, K., Kaulich, T. B., Treitz, C., Cassidy, L., and Tholey, A. (2022b). MStoDiff: A Tool for the Visualization of Mass Shifts in Deconvoluted Top-Down Proteomics Data for the Database-Independent Detection of Protein Modifications. *J. Proteome Res.* 21, 20–29.
- Kearle, P., and Verkerk, U. H. (2009). Electrospray: From ions in solution to ions in the gas phase, what we know now. *Mass Spectrom. Rev.* 28, 898–917.
- Keller, B. O., Sui, J., Young, A. B., and Whittall, R. M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* 627, 71–81.
- Khademi, Z., Milajerdi, A., Larijani, B., and Esmailzadeh, A. (2021). Dietary Intake of Total Carbohydrates, Sugar and Sugar-Sweetened Beverages, and Risk of Inflammatory Bowel Disease: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. *Front. Nutr.* 8, 707795.
- Khan, S., Waliullah, S., Godfrey, V., Khan, M. A. W., Ramachandran, R. A., Cantarel, B. L., et al. (2020). Dietary simple sugars alter microbial ecology in the gut and promote colitis in mice. *Sci. Transl. Med.* 12.
- Khaskheli, G. B., Zuo, F. L., Yu, R., and Chen, S. W. (2015). Overexpression of Small Heat Shock Protein Enhances Heat- and Salt-Stress Tolerance of *Bifidobacterium longum* NCC2705. *Curr. Microbiol.* 71, 8–15.
- Khitun, A., Ness, T. J., and Slavoff, S. A. (2019). Small open reading frames and cellular stress responses. *Mol. Omi.* 15, 108–116.
- Khitun, A., and Slavoff, S. A. (2019). Proteomic Detection and Validation of Translated Small Open Reading Frames. *Curr. Protoc. Chem. Biol.* 11, e77.
- Kim, H., and Shin, S. (2021). ExoCAS-2: Rapid and Pure Isolation of Exosomes by Anionic Exchange Using Magnetic Beads. *Biomedicines* 9, 28.
- Kim, J., Lee, H., Park, K., and Shin, S. (2020). Rapid and Efficient Isolation of Exosomes by Clustering and Scattering. *J. Clin. Med.* 9, 650.
- Kim, M.-S., Zhong, J., and Pandey, A. (2016). Common errors in mass spectrometry-based analysis of post-translational modifications. *Proteomics* 16, 700–714.
- Kim, S.-I., Ha, J. Y., Choi, S.-Y., Hong, S.-H., and Lee, H.-J. (2022). Use of Bacterial Extracellular Vesicles for Gene Delivery to Host Cells. *Biomolecules* 12, 1171.
- Kim, S. G., Becattini, S., Moody, T. U., Shliaha, P. V., Littmann, E. R., Seok, R., et al. (2019). Microbiota-derived lantibiotic restores resistance against vancomycin-resistant *Enterococcus*. *Nature* 572, 665–669.
- Kline, J. T., Belford, M. W., Huang, J., Greer, J. B., Bergen, D., Fellers, R. T., et al. (2023). Improved Label-Free Quantification of Intact Proteoforms Using Field Asymmetric Ion Mobility Spectrometry. *Anal. Chem.* 95, 9090–9096.
- Knoop, V. (2011). When you can't trust the DNA: RNA editing changes transcript sequences. *Cell. Mol. Life Sci.* 68, 567–586.
- Kolhe, P., Amend, E., and K. Singh, S. (2010). Impact of freezing on pH of buffered solutions and consequences for monoclonal antibody aggregation. *Biotechnol. Prog.* 26, 727–733.
- Korniy, N., Samatova, E., Anokhina, M. M., Peske, F., and Rodnina, M. V. (2019). Mechanisms and biomedical implications of –1 programmed ribosome frameshifting on viral and bacterial mRNAs. *FEBS Lett.* 593, 1468–1482.
- Kovalchuk, S. I., Jensen, O. N., and Rogowska-Wrzesinska, A. (2019). FlashPack: Fast and Simple Preparation of Ultrahigh-performance Capillary Columns for LC-MS\*. *Mol. Cell. Proteomics* 18, 383–390.
- Kowalak, J. A., and Walsh, K. A. (1996).  $\beta$ -Methylthio-aspartic acid: Identification of a novel posttranslational modification in ribosomal protein S12 from *Escherichia coli*. *Protein Sci.* 5, 1625–1632.
- Kratz, J. C., and Banerjee, S. (2023). Dynamic proteome trade-offs regulate bacterial cell size and growth in fluctuating nutrient environments. *Commun. Biol.* 6, 486.

- Krause, J. L., Schaepe, S. S., Fritz-Wallace, K., Engelmann, B., Rolle-Kampczyk, U., Kleinsteuber, S., et al. (2020). Following the community development of SIHUMIx—a new intestinal in vitro model for bioreactor use. *Gut Microbes* 11, 1116–1129.
- Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., and Macek, B. (2013). Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* 12, 3420–3430.
- Krulwich, T. A., Sachs, G., and Padan, E. (2011). Molecular aspects of bacterial pH sensing and homeostasis. *Nat. Rev. Microbiol.* 9, 330–343.
- Kulkarni, R. D., Kelkar, H. S., and Dean, R. A. (2003). An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. *Trends Biochem. Sci.* 28, 118–121.
- Lagier, J.-C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* 1, 16203.
- Latosinska, A., Vougas, K., Makridakis, M., Klein, J., Mullen, W., Abbas, M., et al. (2015). Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis. *PLoS One* 10, e0137048.
- Leblanc, S., Brunet, M. A., Jacques, J.-F., Lekehal, A. M., Duclos, A., Tremblay, A., et al. (2023). Newfound Coding Potential of Transcripts Unveils Missing Members of Human Protein Communities. *Genomics. Proteomics Bioinformatics* 21, 515–534.
- Leduc, R. D., Fellers, R. T., Early, B. P., Greer, J. B., Thomas, P. M., and Kelleher, N. L. (2014). The C-Score: A bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *J. Proteome Res.* 13, 3231–3240.
- Lee, E., Bang, J. Y., Park, G. W., Choi, D., Kang, J. S., Kim, H., et al. (2007). Global proteomic profiling of native outer membrane vesicles derived from *Escherichia coli*. *Proteomics* 7, 3143–3153.
- Lee, S. J., Zhang, X., Xu, G., Borjigin, J., and Wang, M. M. (2023). A midposition NOTCH3 truncation in inherited cerebral small vessel disease may affect the protein interactome. *J. Biol. Chem.* 299, 102772.
- Lehmann, T., Schallert, K., Vilchez-Vargas, R., Benndorf, D., Püttker, S., Sydor, S., et al. (2019). Metaproteomics of fecal samples of Crohn's disease and Ulcerative Colitis. *J. Proteomics* 201, 93–103.
- Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., et al. (2011). The European Nucleotide Archive. *Nucleic Acids Res.* 39, D28–D31.
- Leipert, J., Kaulich, P. T., Steinbach, M. K., Steer, B., Winkels, K., Blurton, C., et al. (2023). Digital Microfluidics and Magnetic Bead-Based Intact Proteoform Elution for Quantitative Top-down Nanoproteomics of Single *C. elegans* Nematodes. *Angew. Chemie - Int. Ed.* 62, 1–5.
- Leipert, J., Steinbach, M. K., and Tholey, A. (2021). Isobaric Peptide Labeling on Digital Microfluidics for Quantitative Low Cell Number Proteomics. *Anal. Chem.* 93, 6278–6286.
- Lenčo, J., Jadeja, S., Naplekov, D. K., Krokhin, O. V., Khalikova, M. A., Chocholouš, P., et al. (2022). Reversed-Phase Liquid Chromatography of Peptides for Bottom-Up Proteomics: A Tutorial. *J. Proteome Res.* 21, 2846–2892.
- Lenčo, J., Šemlej, T., Khalikova, M. A., Fabrik, I., and Švec, F. (2021). Sense and Nonsense of Elevated Column Temperature in Proteomic Bottom-up LC-MS Analyses. *J. Proteome Res.* 20, 420–432.
- Leon, I. R., Schwammle, V., Jensen, O. N., and Sprenger, R. R. (2013). Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol. Cell. Proteomics* 12, 2992–3005.
- Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992.
- Li, J., He, X., Deng, Y., and Yang, C. (2019a). An Update on Isolation Methods for Proteomic Studies of Extracellular Vesicles in Biofluids. *Molecules* 24, 3516.
- Li, K., Vaudel, M., Zhang, B., Ren, Y., and Wen, B. (2019b). PDV: an integrative proteomics data viewer. *Bioinformatics* 35, 1249–1251.

## BIBLIOGRAPHY

- Li, N., Fort, F., Kessler, K., and Wang, W. (2009). Factors affecting cleavage at aspartic residues in model decapeptides. *J. Pharm. Biomed. Anal.* 50, 73–78.
- Liao, Y.-D., Jeng, J.-C., Wang, C.-F., Wang, S.-C., and Chang, S.-T. (2004). Removal of N-terminal methionine from recombinant proteins by engineered *E. coli* methionine aminopeptidase. *Protein Sci.* 13, 1802–10.
- Lichtman, J. S., Alsentzer, E., Jaffe, M., Sprockett, D., Masutani, E., Ikwa, E., et al. (2016). The effect of microbial colonization on the host proteome varies by gastrointestinal location. *ISME J.* 10, 1170–1181.
- Lidell, M. E., Johansson, M. E. V., and Hansson, G. C. (2003). An Autocatalytic Cleavage in the C Terminus of the Human MUC2 Mucin Occurs at the Low pH of the Late Secretory Pathway. *J. Biol. Chem.* 278, 13944–13951.
- Lin, J., In Soo Lee, Frey, J., Slonczewski, J. L., and Foster, J. W. (1995). Comparative analysis of extreme acid survival in *Salmonella typhimurium*, *Shigella flexneri*, and *Escherichia coli*. *J. Bacteriol.* 177, 4097–4104.
- Liu, C., Finegold, S. M., Song, Y., and Lawson, P. A. (2008). Reclassification of *Clostridium coccoides*, *Ruminococcus hansenii*, *Ruminococcus hydrogenotrophicus*, *Ruminococcus luti*, *Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides* gen. nov., comb. nov., *Blautia hansenii* comb. nov., *Blautia hydroge*. *Int. J. Syst. Evol. Microbiol.* 58, 1896–1902.
- Liu, H., Shiver, A. L., Price, M. N., Carlson, H. K., Trotter, V. V., Chen, Y., et al. (2021a). Functional genetics of human gut commensal *Bacteroides thetaiotaomicron* reveals metabolic requirements for growth across environments. *Cell Rep.* 34, 108789.
- Liu, J., Wang, F., Mao, J., Zhang, Z., Liu, Z., Huang, G., et al. (2015). High-sensitivity N-glycoproteomic analysis of mouse brain tissue by protein extraction with a mild detergent of N-dodecyl  $\beta$ -D-maltoside. *Anal. Chem.* 87, 2054–7.
- Liu, S., Moulton, K. R., Auclair, J. R., and Zhou, Z. S. (2016). Mildly acidic conditions eliminate deamidation artifact during proteolysis: digestion with endoprotease Glu-C at pH 4.5. *Amino Acids* 48, 1059–1067.
- Liu, S., Wen, B., Du, G., Wang, Y., Ma, X., Yu, H., et al. (2023). Coordinated regulation of *Bacteroides thetaiotaomicron* glutamate decarboxylase activity by multiple elements under different pH. *Food Chem.* 403, 134436.
- Liu, W., Ma, Z., and Kang, X. (2022). Current status and outlook of advances in exosome isolation. *Anal. Bioanal. Chem.* 414, 7123–7141.
- Liu, X., Hengel, S., Wu, S., Tolić, N., Pasa-Tolić, L., and Pevzner, P. A. (2013). Identification of ultramodified proteins using top-down tandem mass spectra. *J. Proteome Res.* 12, 5830–5838.
- Liu, X., Mao, B., Gu, J., Wu, J., Cui, S., Wang, G., et al. (2021b). *Blautia* —a new functional genus with potential probiotic properties? *Gut Microbes* 13, e1875796.
- Longo, S., Chieppa, M., Cossa, L. G., Spinelli, C. C., Greco, M., Maffia, M., et al. (2020). New Insights into Inflammatory Bowel Diseases from Proteomic and Lipidomic Studies. *Proteomes* 8, 18.
- Loos, M. S., Ramakrishnan, R., Vranken, W., Tsigotaki, A., Tsare, E.-P., Zorzini, V., et al. (2019). Structural Basis of the Subcellular Topology Landscape of *Escherichia coli*. *Front. Microbiol.* 10, 1670.
- Low, T. Y., Syafruddin, S. E., Mohtar, M. A., Vellaichamy, A., A Rahman, N. S., Pung, Y.-F., et al. (2021). Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. *Cell. Mol. Life Sci.* 78, 5325–5339.
- Lu, P., Ma, D., Chen, Y., Guo, Y., Chen, G.-Q., Deng, H., et al. (2013). L-glutamine provides acid resistance for *Escherichia coli* through enzymatic release of ammonia. *Cell Res.* 23, 635–644.
- Macek, B., Forchhammer, K., Hardouin, J., Weber-Ban, E., Grangeasse, C., and Mijakovic, I. (2019). Protein post-translational modifications in bacteria. *Nat. Rev. Microbiol.* 17, 651–664.
- Magasanik, B. (1953). The pathway of inositol dissimilation in *Aerobacter aerogenes*. *J. Biol. Chem.* 205, 1019–26.

- Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., et al. (2020). The Firmicutes/Bacteroidetes Ratio: A Relevant Marker of Gut Dysbiosis in Obese Patients? *Nutrients* 12, 1474.
- Magnúsdóttir, S., Ravcheev, D., de Crécy-Lagard, V., and Thiele, I. (2015). Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. *Front. Genet.* 6, 148.
- Maiolica, A., Borsotti, D., and Rappsilber, J. (2005). Self-made frits for nanoscale columns in proteomics. *Proteomics* 5, 3847–3850.
- Manza, L. L., Stamer, S. L., Ham, A.-J. L., Codreanu, S. G., and Liebler, D. C. (2005). Sample preparation and digestion for proteomic analyses using spin filters. *Proteomics* 5, 1742–1745.
- Marcus, F. (1985). Preferential cleavage at aspartyl-prolyl peptide bonds in dilute acid. *Int. J. Pept. Protein Res.* 25, 542–546.
- Martens, E. C., Chiang, H. C., and Gordon, J. I. (2008). Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe* 4, 447–457.
- Martinis, S. A., and Boniecki, M. T. (2010). The balance between pre- and post-transfer editing in tRNA synthetases. *FEBS Lett.* 584, 455–459.
- Marx, V. (2024). Inside the chase after those elusive proteoforms. *Nat. Methods* 21, 158–163.
- Masuda, T., Tomita, M., and Ishihama, Y. (2008). Phase Transfer Surfactant-Aided Trypsin Digestion for Membrane Proteome Analysis. *J. Proteome Res.* 7, 731–740.
- Matzinger, M., Müller, E., Dürnberger, G., Pichler, P., and Mechtler, K. (2023). Robust and Easy-to-Use One-Pot Workflow for Label-Free Single-Cell Proteomics. *Anal. Chem.* 95, 4435–4445.
- Matzinger, M., Schmücker, A., Yelagandula, R., Stejskal, K., Krššáková, G., Berger, F., et al. (2024). Micropillar arrays, wide window acquisition and AI-based data analysis improve comprehensiveness in multiple proteomic applications. *Nat. Commun.* 15, 1019.
- Maurer, L. M., Yohannes, E., Bondurant, S. S., Radmacher, M., and Slonczewski, J. L. (2005). pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12. *J. Bacteriol.* 187, 304–319.
- McCarthy, E. L., and Booker, S. J. (2020). “The Biosynthesis of Lipoic Acid,” in *Comprehensive Natural Products III*, (Elsevier), 3–23.
- McDonald, A. G., Boyce, S., Moss, G. P., Dixon, H. B., and Tipton, K. F. (2007). ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature. *BMC Biochem.* 8, 14.
- McLaggan, D., Naprstek, J., Buurman, E. T., and Epstein, W. (1994). Interdependence of K<sup>+</sup> and glutamate accumulation during osmotic adaptation of *Escherichia coli*. *J. Biol. Chem.* 269, 1911–1917.
- McMillan, H. M., and Kuehn, M. J. (2023). Proteomic Profiling Reveals Distinct Bacterial Extracellular Vesicle Subpopulations with Possibly Unique Functionality. *Appl. Environ. Microbiol.* 89, 1–37.
- Meadow, N. D., Fox, D. K., and Roseman, S. (1990). The Bacterial Phosphoenol-pyruvate: Glycolose Phosphotransferase System. *Annu. Rev. Biochem.* 59, 497–542.
- Mehla, J., Caufield, J. H., and Uetz, P. (2015). The Yeast Two-Hybrid System: A Tool for Mapping Protein–Protein Interactions. *Cold Spring Harb. Protoc.* 2015, 425–430.
- Meier-Credo, J., Preiss, L., Wüllenweber, I., Resemann, A., Nordmann, C., Zabret, J., et al. (2022). Top-Down Identification and Sequence Analysis of Small Membrane Proteins Using MALDI-MS/MS. *J. Am. Soc. Mass Spectrom.* 33, 1293–1302.
- Meinzel, T., Mechulam, Y., and Blanquet, S. (1993). Methionine as translation start signal: A review of the enzymes of the pathway in *Escherichia coli*. *Biochimie* 75, 1061–1075.
- Melo, R. M., de Souza, J. M. F., Williams, T. C. R., Fontes, W., de Sousa, M. V., Ricart, C. A. O., et al. (2023). Revealing *Corynebacterium glutamicum* proteoforms through top-down proteomics. *Sci. Rep.* 13, 2602.
- Mergner, J., and Kuster, B. (2022). Plant Proteome Dynamics. *Annu. Rev. Plant Biol.* 73, 67–92.

## BIBLIOGRAPHY

- Michalski, A., Cox, J., and Mann, M. (2011). More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC–MS/MS. *J. Proteome Res.* 10, 1785–1793.
- Michalski, A., Neuhauser, N., Cox, J., and Mann, M. (2012). A Systematic Investigation into the Nature of Tryptic HCD Spectra. *J. Proteome Res.* 11, 5479–5491.
- Michel, L. V., and Gaborski, T. (2022). Outer membrane vesicles as molecular biomarkers for Gram-negative sepsis: Taking advantage of nature's perfect packages. *J. Biol. Chem.* 298, 102483.
- Midekessa, G., Godakumara, K., Ord, J., Viil, J., Lättেকivi, F., Dissanayake, K., et al. (2020). Zeta Potential of Extracellular Vesicles: Toward Understanding the Attributes that Determine Colloidal Stability. *ACS Omega* 5, 16701–16710.
- Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., et al. (2019). Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol.* 15, e8290.
- Montecchi-Palazzi, L., Beavis, R., Binz, P.-A., Chalkley, R. J., Cottrell, J., Creasy, D., et al. (2008). The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.* 26, 864–866.
- Morinaga, T., Ashida, H., and Yoshida, K. (2010). Identification of two scyllo-inositol dehydrogenases in *Bacillus subtilis*. *Microbiology* 156, 1538–1546.
- Mueller, E. A., Westfall, C. S., and Levin, P. A. (2020). PH-dependent activation of cytokinesis modulates *Escherichia coli* cell size. *PLoS Genet.* 16, 1–24.
- Mulkidjanian, A. Y., Dibrov, P., and Galperin, M. Y. (2008). The past and present of sodium energetics: May the sodium-motive force be with you. *Biochim. Biophys. Acta - Bioenerg.* 1777, 985–992.
- Navarro-Requena, C., Pérez-Amodio, S., Castaño, O., and Engel, E. (2018). Wound healing-promoting effects stimulated by extracellular calcium and calcium-releasing nanoparticles on dermal fibroblasts. *Nanotechnology* 29, 395102.
- Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 73, 2092–2123.
- Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11, 1114–1125.
- Nesvizhskii, A. I., and Aebersold, R. (2005). Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* 4, 1419–1440.
- Neurath, H., and Walsh, K. A. (1976). Role of proteolytic enzymes in biological regulation (a review). *Proc. Natl. Acad. Sci.* 73, 3825–3832.
- Ng, S. C. (2014). Epidemiology of inflammatory bowel disease: Focus on Asia. *Best Pract. Res. Clin. Gastroenterol.* 28, 363–372.
- Ntai, I., LeDuc, R. D., Fellers, R. T., Erdmann-Gilmore, P., Davies, S. R., Rumsey, J., et al. (2016). Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol. Cell. Proteomics* 15, 45–56.
- Nugent, S. G., Kumar, D., Rampton, D. S., and Evans, D. F. (2001). Intestinal luminal pH in inflammatory bowel disease: possible determinants and implications for therapy with aminosaliclates and other drugs. *Gut* 48, 571–7.
- O'Connell, J. D., Paulo, J. A., O'Brien, J. J., and Gygi, S. P. (2018). Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J. Proteome Res.* 17, 1934–1942.
- Old, I. G., Phillips, S. E. V., Stockley, P. G., and Saint Girons, I. (1991). Regulation of methionine biosynthesis in the enterobacteriaceae. *Prog. Biophys. Mol. Biol.* 56, 145–185.
- Oliveros, J. C. (2007). Venny. An interactive tool for comparing lists with Venn's diagrams. Available at: <https://bioinfogp.cnb.csic.es/tools/venny/index.html>.
- Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* 4, 709–712.
- Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., et al. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Res.* 27, 2083–2095.

- Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., et al. (2002). Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* 1, 376–386.
- Orench-Rivera, N., and Kuehn, M. J. (2016). Environmentally controlled bacterial vesicle-mediated export. *Cell. Microbiol.* 18, 1525–1536.
- Osička, R., Procházková, K., Šulc, M., Linhartová, I., Havlíček, V., and Šebo, P. (2004). A Novel “Clip-and-link” Activity of Repeat in Toxin (RTX) Proteins from Gram-negative Pathogens. *J. Biol. Chem.* 279, 24944–24956.
- Osteikoetxea, X., Sódar, B., Németh, A., Szabó-Taylor, K., Pálóczi, K., Vukman, K. V., et al. (2015). Differential detergent sensitivity of extracellular vesicle subpopulations. *Org. Biomol. Chem.* 13, 9775–9782.
- Otaru, N., Ye, K., Mujezinovic, D., Berchtold, L., Constancias, F., Cornejo, F. A., et al. (2021). GABA Production by Human Intestinal Bacteroides spp.: Prevalence, Regulation, and Role in Acid Stress Tolerance. *Front. Microbiol.* 12, 656895.
- Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–D214.
- Palomba, A., Abbondio, M., Fiorito, G., Uzzau, S., Pagnozzi, D., and Tanca, A. (2021). Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. *J. Proteome Res.* 20, 3497–3507.
- Pappireddi, N., Martin, L., and Wühr, M. (2019). A Review on Quantitative Multiplexed Proteomics. *ChemBioChem* 20, 1210–1224.
- Parada Venegas, D., De la Fuente, M. K., Landskron, G., González, M. J., Quera, R., Dijkstra, G., et al. (2019). Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Front. Immunol.* 10, 277.
- Patel, V. J., Thalassinos, K., Slade, S. E., Connolly, J. B., Crombie, A., Murrell, J. C., et al. (2009). A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches. *J. Proteome Res.* 8, 3752–3759.
- Patrick, M. E., and Eglund, K. A. (2019). SUSD2 Proteolytic Cleavage Requires the GDPH Sequence and Inter-Fragment Disulfide Bonds for Surface Presentation of Galectin-1 on Breast Cancer Cells. *Int. J. Mol. Sci.* 20, 3814.
- Patten, D. A., Hussein, E., Davies, S. P., Humphreys, P. N., and Collett, A. (2017). Commensal-derived OMVs elicit a mild proinflammatory response in intestinal epithelial cells. *Microbiology* 163, 702–711.
- Peng, C., Shi, C., Cao, X., Li, Y., Liu, F., and Lu, F. (2019). Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond. *Front. Bioeng. Biotechnol.* 7, 139.
- Penn, J. W., Grobbelaar, A. O., and Rolfe, K. J. (2012). The role of the TGF- $\beta$  family in wound healing, burns and scarring: a review. *Int. J. Burns Trauma* 2, 18–28.
- Pennacchietti, E., D’Alonzo, C., Freddi, L., Occhialini, A., and De Biase, D. (2018). The Glutaminase-Dependent Acid Resistance System: Qualitative and Quantitative Assays and Analysis of Its Distribution in Enteric Bacteria. *Front. Microbiol.* 9, 2869.
- Pérez-Cruz, C., Delgado, L., López-Iglesias, C., and Mercade, E. (2015). Outer-Inner Membrane Vesicles Naturally Secreted by Gram-Negative Pathogenic Bacteria. *PLoS One* 10, e0116896.
- Perez, A. J., Cesbron, Y., Shaw, S. L., Bazan Villicana, J., Tsui, H.-C. T., Boersma, M. J., et al. (2019). Movement dynamics of divisome proteins and PBP2x:FtsW in cells of *Streptococcus pneumoniae*. *Proc. Natl. Acad. Sci.* 116, 3211–3220.
- Pérez Montoro, B., Benomar, N., Caballero Gómez, N., Ennahar, S., Horvatovich, P., Knapp, C. W., et al. (2018). Proteomic analysis of *Lactobacillus pentosus* for the identification of potential markers involved in acid resistance and their influence on other probiotic features. *Food Microbiol.* 72, 31–38.

## BIBLIOGRAPHY

- Petruschke, H., Anders, J., Stadler, P. F., Jehmlich, N., and von Bergen, M. (2020). Enrichment and identification of small proteins in a simplified human gut microbiome. *J. Proteomics* 213, 103604.
- Petruschke, H., Schori, C., Canzler, S., Riesbeck, S., Poehlein, A., Daniel, R., et al. (2021). Discovery of novel community-relevant small proteins in a simplified human intestinal microbiome. *Microbiome* 9, 55.
- Phua, S.-X., Lim, K.-P., and Goh, W. W.-B. (2022). Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Comput. Struct. Biotechnol. J.* 20, 4369–4375.
- Pidutti, P., Federici, F., Brandi, J., Manna, L., Rizzi, E., Marini, U., et al. (2018). Purification and characterization of ribosomal proteins L27 and L30 having antimicrobial activity produced by the *Lactobacillus salivarius* SGL 03. *J. Appl. Microbiol.* 124, 398–407.
- Piersimoni, L., Kastiris, P. L., Arlt, C., and Sinz, A. (2022). Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein Interactions—A Method for All Seasons. *Chem. Rev.* 122, 7500–7531.
- Piszkiewicz, D., Landon, M., and Smith, E. L. (1970). Anomalous cleavage of aspartyl-proline peptide bonds during amino acid sequence determinations. *Biochem. Biophys. Res. Commun.* 40, 1173–1178.
- Png, C. W., Lindén, S. K., Gilshenan, K. S., Zoetendal, E. G., McSweeney, C. S., Sly, L. I., et al. (2010). Mucolytic Bacteria With Increased Prevalence in IBD Mucosa Augment In Vitro Utilization of Mucin by Other Bacteria. *Am. J. Gastroenterol.* 105, 2420–2428.
- Poncet, S., Mijakovic, I., Nessler, S., Gueguen-Chaignon, V., Chaptal, V., Galinier, A., et al. (2004). HPr kinase/phosphorylase, a Walker motif A-containing bifunctional sensor enzyme controlling catabolite repression in Gram-positive bacteria. *Biochim. Biophys. Acta - Proteins Proteomics* 1697, 123–135.
- Poncet, S., Milohanic, E., Mazé, A., Abdallah, J. N., Aké, F., Larribe, M., et al. (2009). “Correlations between Carbon Metabolism and Virulence in Bacteria,” in *Bacterial Sensing and Signaling*, 88–102.
- Porter, N. T., and Martens, E. C. (2017). The Critical Roles of Polysaccharides in Gut Microbial Ecology and Physiology. *Annu. Rev. Microbiol.* 71, 349–369.
- Precht, R. M., Janßen, D., Behr, J., Ludwig, C., Küster, B., Vogel, R. F., et al. (2018). Sucrose-Induced Proteomic Response and Carbohydrate Utilization of *Lactobacillus sakei* TMW 1.411 During Dextran Formation. *Front. Microbiol.* 9, 2796.
- Quast, J.-P., Schuster, D., and Picotti, P. (2022). protti: an R package for comprehensive data analysis of peptide- and protein-centric bottom-up proteomics data. *Bioinform. Adv.* 2, vbab041.
- Rajilić-Stojanović, M., Biagi, E., Heilig, H. G. H. J., Kajander, K., Kekkonen, R. A., Tims, S., et al. (2011). Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples From Patients With Irritable Bowel Syndrome. *Gastroenterology* 141, 1792–1801.
- Ramires, T., Wilson, R., Padilha da Silva, W., and Bowman, J. P. (2023). Identification of pH-specific protein expression responses by *Campylobacter jejuni* strain NCTC 11168. *Res. Microbiol.* 174, 104061.
- Rauniyar, N., and Yates, J. R. (2014). Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *J. Proteome Res.* 13, 5293–5309.
- Reid, G. E., Wu, J., Chrisman, P. A., Wells, J. M., and McLuckey, S. A. (2001). Charge-State-Dependent Sequence Analysis of Protonated Ubiquitin Ions via Ion Trap Tandem Mass Spectrometry. *Anal. Chem.* 73, 3274–3281.
- Reynolds, T. B. (2009). Strategies for acquiring the phospholipid metabolite inositol in pathogenic bacteria, fungi and protozoa: making it and taking it. *Microbiology* 155, 1386–1396.
- Rezzonico, E., Lariani, S., Barretto, C., Cuanoud, G., Giliberti, G., Delley, M., et al. (2007). Global transcriptome analysis of the heat shock response of *Bifidobacterium longum*. *FEMS Microbiol. Lett.* 271, 136–145.
- Ringel-Kulka, T., Choi, C. H., Temas, D., Kim, A., Maier, D. M., Scott, K., et al. (2015). Altered Colonic Bacterial Fermentation as a Potential Pathophysiological Factor in Irritable Bowel Syndrome. *Am. J. Gastroenterol.* 110, 1339–1346.

- Rodionov, D. A., Arzamasov, A. A., Khoroshkin, M. S., Iablokov, S. N., Leyn, S. A., Peterson, S. N., et al. (2019). Micronutrient Requirements and Sharing Capabilities of the Human Gut Microbiome. *Front. Microbiol.* 10, 1316.
- Roe, A. J., McLaggan, D., Davidson, I., O'Byrne, C., and Booth, I. R. (1998). Perturbation of Anion Balance during Inhibition of Growth of *Escherichia coli* by Weak Acids. *J. Bacteriol.* 180, 767–772.
- Roepstorff, P., and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* 11, 601.
- Rook, G. A. W., Lowry, C. A., and Raison, C. L. (2015). Hygiene and other early childhood influences on the subsequent function of the immune system. *Brain Res.* 1617, 47–62.
- Rundlett, K. L., and Armstrong, D. W. (1996). Mechanism of Signal Suppression by Anionic Surfactants in Capillary Electrophoresis–Electrospray Ionization Mass Spectrometry. *Anal. Chem.* 68, 3493–3497.
- Sadeghpour Heravi, F., and Hu, H. (2023). Bifidobacterium: Host–Microbiome Interaction and Mechanism of Action in Preventing Common Gut-Microbiota-Associated Complications in Preterm Infants: A Narrative Review. *Nutrients* 15, 709.
- Sánchez, B., Champomier-Vergès, M.-C., Collado, M. D. C., Anglade, P., Baraige, F., Sanz, Y., et al. (2007). Low-pH Adaptation and the Acid Tolerance Response of *Bifidobacterium longum* Biotype *longum*. *Appl. Environ. Microbiol.* 73, 6450–6459.
- Sandberg, A., Branca, R. M. M., Lehtiö, J., and Forshed, J. (2014). Quantitative accuracy in mass spectrometry based proteomics of complex samples: The impact of labeling and precursor interference. *J. Proteomics* 96, 133–144.
- Sartorio, M. G., Valguarnera, E., Hsu, F.-F., and Feldman, M. F. (2022). Lipidomics Analysis of Outer Membrane Vesicles and Elucidation of the Inositol Phosphoceramide Biosynthetic Pathway in *Bacteroides thetaiotaomicron*. *Microbiol. Spectr.* 10, 1–13.
- Savaryn, J. P., Toby, T. K., and Kelleher, N. L. (2016). A researcher's guide to mass spectrometry-based proteomics. *Proteomics* 16, 2435–2443.
- Savitski, M. M., Mathieson, T., Zinn, N., Sweetman, G., Doce, C., Becher, I., et al. (2013). Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *J. Proteome Res.* 12, 3586–3598.
- Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., et al. (2019). Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* 178, 1245–1259.e14.
- Schaffer, L. V., Anderson, L. C., Butcher, D. S., Shortreed, M. R., Miller, R. M., Pavelec, C., et al. (2021). Construction of Human Proteoform Families from 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Top-Down Proteomic Data. *J. Proteome Res.* 20, 317–325.
- Scheerlinck, E., Dhaenens, M., Van Soom, A., Peelman, L., De Sutter, P., Van Steendam, K., et al. (2015). Minimizing technical variation during sample preparation prior to label-free quantitative mass spectrometry. *Anal. Biochem.* 490, 14–19.
- Scheidler, C. M., Kick, L. M., and Schneider, S. (2019). Ribosomal Peptides and Small Proteins on the Rise. *ChemBioChem* 20, 1479–1486.
- Schlesinger, D., and Elsässer, S. J. (2022). Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *FEBS J.* 289, 53–74.
- Schlüter, H., Apweiler, R., Holzhütter, H.-G., and Jungblut, P. R. (2009). Finding one's way in proteomics: a protein species nomenclature. *Chem. Cent. J.* 3, 11.
- Schofield, W. B., Zimmermann-Kogadeeva, M., Zimmermann, M., Barry, N. A., and Goodman, A. L. (2018). The Stringent Response Determines the Ability of a Commensal Bacterium to Survive Starvation and to Persist in the Gut. *Cell Host Microbe* 24, 120–132.e6.
- Schräder, C. U., Lee, L., Rey, M., Sarpe, V., Man, P., Sharma, S., et al. (2017). Neprosin, a selective prolyl endoprotease for bottom-up proteomics and histone mapping. *Mol. Cell. Proteomics* 16, 1162–1171.
- Schrödinger, L. (2002). The PyMOL Molecular Graphics System, Version 2.5. 40, 82–92. Available at: <http://citebay.com/how-to-cite/pymol/> (Accessed January 2, 2023).

## BIBLIOGRAPHY

- Schulze, W. X., and Usadel, B. (2010). Quantitation in Mass-Spectrometry-Based Proteomics. *Annu. Rev. Plant Biol.* 61, 491–516.
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global quantification of mammalian gene expression control. *Nature* 473, 337–342.
- Serbanescu, D., Ojkic, N., and Banerjee, S. (2022). Cellular resource allocation strategies for cell size and shape control in bacteria. *FEBS J.* 289, 7891–7906.
- Sharif, E., Eftekhari, Z., and Mohit, E. (2021). The Effect of Growth Stage and Isolation Method on Properties of ClearColi™ Outer Membrane Vesicles (OMVs). *Curr. Microbiol.* 78, 1602–1614.
- Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* 68, 850–858.
- Sim, P., Song, Y., Yang, G. N., Cowin, A. J., and Garg, S. (2022). In Vitro Wound Healing Properties of Novel Acidic Treatment Regimen in Enhancing Metabolic Activity and Migration of Skin Cells. *Int. J. Mol. Sci.* 23, 7188.
- Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., et al. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* 9, 59–64.
- Smith, L. M., and Kelleher, N. L. (2013). Proteoform: A single term describing protein complexity. *Nat. Methods* 10, 186–187.
- Smith, L. M., and Kelleher, N. L. (2018). Proteoforms as the next proteomics currency. *Science* (80-. ). 359, 1106–1107.
- Solbiati, J., Chapman-Smith, A., Miller, J. L., Miller, C. G., and Cronan, J. E. (1999). Processing of the N termini of nascent polypeptide chains requires deformylation prior to methionine removal. *J. Mol. Biol.* 290, 607–614.
- Sonnenburg, E. D., Zheng, H., Joglekar, P., Higginbottom, S. K., Firbank, S. J., Bolam, D. N., et al. (2010). Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* 141, 1241–1252.
- Stancik, L. M., Stancik, D. M., Schmidt, B., Barnhart, D. M., Yoncheva, Y. N., and Slonczewski, J. L. (2002). pH-Dependent Expression of Periplasmic Proteins and Amino Acid Catabolism in Escherichia coli. *J. Bacteriol.* 184, 4246–4258.
- Steiner, S., Smith, S., Waechter, C. J., and Lester, R. L. (1969). Isolation and partial characterization of a major inositol-containing lipid in baker's yeast, mannosyl-diinositol, diphosphoryl-ceramide. *Proc. Natl. Acad. Sci. U. S. A.* 64, 1042–1048.
- Stephanowitz, H., Lange, S., Lang, D., Freund, C., and Krause, E. (2012). Improved two-dimensional reversed phase-reversed phase LC-MS/MS approach for identification of peptide-protein interactions. *J. Proteome Res.* 11, 1175–1183.
- Stojanov, S., Berlec, A., and Štrukelj, B. (2020). The Influence of Probiotics on the Firmicutes/Bacteroidetes Ratio in the Treatment of Obesity and Inflammatory Bowel disease. *Microorganisms* 8, 1715.
- Storz, G., Wolf, Y. I., and Ramamurthi, K. S. (2014). Small Proteins Can No Longer Be Ignored. *Annu. Rev. Biochem.* 83, 753–777.
- Strader, M. B., Costantino, N., Elkins, C. A., Chen, C. Y., Patel, I., Makusky, A. J., et al. (2011). A Proteomic and Transcriptomic Approach Reveals New Insight into  $\beta$ -methylthiolation of Escherichia coli Ribosomal Protein S12. *Mol. Cell. Proteomics* 10, M110.005199.
- Strader, M. B., VerBerkmoes, N. C., Tabb, D. L., Connelly, H. M., Barton, J. W., Bruce, B. D., et al. (2004). Characterization of the 70S Ribosome from Rhodopseudomonas p alustris Using an Integrated “Top-Down” and “Bottom-Up” Mass Spectrometric Approach. *J. Proteome Res.* 3, 965–978.
- Suarez-Arnedo, A., Torres Figueroa, F., Clavijo, C., Arbeláez, P., Cruz, J. C., and Muñoz-Camargo, C. (2020). An image J plugin for the high throughput image analysis of in vitro scratch wound healing assays. *PLoS One* 15, e0232565.

- Suh, M.-J., Hamburg, D.-M., Gregory, S. T., Dahlberg, A. E., and Limbach, P. A. (2005). Extending ribosomal protein identifications to unsequenced bacterial strains using matrix-assisted laser desorption/ionization mass spectrometry. *Proteomics* 5, 4818–4831.
- Sun, L., Knierman, M. D., Zhu, G., and Dovichi, N. J. (2013). Fast top-down intact protein characterization with capillary zone electrophoresis-electrospray ionization tandem mass spectrometry. *Anal. Chem.* 85, 5989–5995.
- Swearingen, K. E., and Moritz, R. L. (2012). High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. *Expert Rev. Proteomics* 9, 505–517.
- Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9528–9533.
- Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R. (2003). Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* 75, 1155–1163.
- Taheri, N., Mahmud, A. K. M. F., Sandblad, L., Fällman, M., Wai, S. N., and Fahlgren, A. (2018). *Campylobacter jejuni* bile exposure influences outer membrane vesicles protein content and bacterial interaction with epithelial cells. *Sci. Rep.* 8, 16996.
- Takemori, A., Butcher, D. S., Harman, V. M., Brownridge, P., Shima, K., Higo, D., et al. (2020). PEPPI-MS: Polyacrylamide-Gel-Based Prefractionation for Analysis of Intact Proteoforms and Protein Complexes by Mass Spectrometry. *J. Proteome Res.* 19, 3779–3791.
- Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 44, 6614–6624.
- Taverna, D., and Gaspari, M. (2021). A critical comparison of three MS-based approaches for quantitative proteomics analysis. *J. Mass Spectrom.* 56, e4669.
- Terrapon, N., Lombard, V., Gilbert, H. J., and Henrissat, B. (2015). Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* 31, 647–655.
- Théry, C., Witwer, K. W., Aikawa, E., Alcaraz, M. J., Anderson, J. D., Andriantsitohaina, R., et al. (2018). Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. *J. Extracell. Vesicles* 7, 1535750.
- Tholey, A., and Becker, A. (2017). Top-down proteomics for the analysis of proteolytic events - Methods, applications and perspectives. *Biochim. Biophys. Acta - Mol. Cell Res.* 1864, 2191–2199.
- Thomas, M. P., Mills, S. J., and Potter, B. V. L. (2016). The “Other” Inositols and Their Phosphates: Synthesis, Biology, and Medicine (with Recent Advances in myo-Inositol Chemistry). *Angew. Chemie - Int. Ed.* 55, 1614–1650.
- Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., et al. (2003). Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* 75, 1895–1904.
- Thuveson, M., and Fries, E. (2000). The Low pH in Trans-Golgi Triggers Autocatalytic Cleavage of Pre- $\alpha$ -inhibitor Heavy Chain Precursor. *J. Biol. Chem.* 275, 30996–31000.
- Toby, T. K., Fornelli, L., Srzentić, K., DeHart, C. J., Levitsky, J., Friedewald, J., et al. (2019). A comprehensive pipeline for translational top-down proteomics from a single blood draw. *Nat. Protoc.* 14, 119–152.
- Tollin, M., Bergman, P., Svenberg, T., Jörnvall, H., Gudmundsson, G. H., and Agerberth, B. (2003). Antimicrobial peptides in the first line defence of human colon mucosa. *Peptides* 24, 523–30.
- Townsend, G. E., Han, W., Schwalm, N. D., Hong, X., Bencivenga-Barry, N. A., Goodman, A. L., et al. (2020). A Master Regulator of Bacteroides thetaiotaomicron Gut Colonization Controls Carbohydrate Utilization and an Alternative Protein Synthesis Factor. *MBio* 11, e03221-19.
- Townsend, G. E., Han, W., Schwalm, N. D., Raghavan, V., Barry, N. A., Goodman, A. L., et al. (2019). Dietary sugar silences a colonization factor in a mammalian gut symbiont. *Proc. Natl. Acad. Sci. U. S. A.* 116, 233–238.

## BIBLIOGRAPHY

- Toyofuku, M., Nomura, N., and Eberl, L. (2019). Types and origins of bacterial membrane vesicles. *Nat. Rev. Microbiol.* 17, 13–24.
- Tran, J. C., and Doucette, A. A. (2008). Gel-eluted liquid fraction entrapment electrophoresis: An electrophoretic method for broad molecular weight range proteome separation. *Anal. Chem.* 80, 1568–1573.
- Treitz, C., Enjalbert, B., Portais, J., Letisse, F., and Tholey, A. (2016). Differential quantitative proteome analysis of *Escherichia coli* grown on acetate versus glucose. *Proteomics* 16, 2742–2746.
- Trudgian, D. C., Fischer, R., Guo, X., Kessler, B. M., and Mirzaei, H. (2014). GOAT - A simple LC-MS/MS gradient optimization tool. *Proteomics* 14, 1467–1471.
- Truong, T., Webber, K. G. I., Madisyn Johnston, S., Boekweg, H., Lindgren, C. M., Liang, Y., et al. (2023). Data-Dependent Acquisition with Precursor Coisolation Improves Proteome Coverage and Measurement Throughput for Label-Free Single-Cell Proteomics. *Angew. Chemie Int. Ed.* 62, e202303415.
- Tsai, C.-F., Zhang, P., Scholten, D., Martin, K., Wang, Y.-T., Zhao, R., et al. (2021). Surfactant-assisted one-pot sample preparation for label-free single-cell proteomics. *Commun. Biol.* 4, 265.
- Turroni, F., Foroni, E., Pizzetti, P., Giubellini, V., Ribbera, A., Merusi, P., et al. (2009). Exploring the diversity of the bifidobacterial population in the human intestinal tract. *Appl. Environ. Microbiol.* 75, 1534–1545.
- Turroni, F., Peano, C., Pass, D. A., Foroni, E., Severgnini, M., Claesson, M. J., et al. (2012). Diversity of Bifidobacteria within the Infant Gut Microbiota. *PLoS One* 7, e36957.
- Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., et al. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* 13, 731–740.
- Tyler-Cross, R., and Schirch, V. (1991). Effects of amino acid sequence, buffers, and ionic strength on the rate and mechanism of deamidation of asparagine residues in small peptides. *J. Biol. Chem.* 266, 22549–56.
- van Niel, G., D'Angelo, G., and Raposo, G. (2018). Shedding light on the cell biology of extracellular vesicles. *Nat. Rev. Mol. Cell Biol.* 19, 213–228.
- Varadarajan, A. R., Goetze, S., Pavlou, M. P., Grosboillot, V., Shen, Y., Loessner, M. J., et al. (2020). A Proteogenomic Resource Enabling Integrated Analysis of *Listeria* Genotype-Proteotype-Phenotype Relationships. *J. Proteome Res.* 19, 1647–1662.
- Vargas-Blanco, D. A., and Shell, S. S. (2020). Regulation of mRNA Stability During Bacterial Stress Responses. *Front. Microbiol.* 11, 2111.
- Varnavides, G., Madern, M., Anrather, D., Hartl, N., Reiter, W., and Hartl, M. (2022). In Search of a Universal Method: A Comparative Survey of Bottom-Up Proteomics Sample Preparation Methods. *J. Proteome Res.* 21, 2397–2411.
- Varshavsky, A. (2011). The N-end rule pathway and regulation by proteolysis. *Protein Sci.* 20, 1298–1345.
- Veltri, D., Kamath, U., and Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 34, 2740–2747.
- Vitreschak, A. G., Mironov, A. A., Lyubetsky, V. A., and Gelfand, M. S. (2008). Comparative genomic analysis of T-box regulatory systems in bacteria. *RNA* 14, 717–735.
- Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32, 223–226.
- Wade, J. T., and Grainger, D. C. (2014). Pervasive transcription: Illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* 12, 647–653.
- Wang, J., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, S., et al. (2023a). The conserved domain database in 2023. *Nucleic Acids Res.* 51, D384–D388.

- Wang, R., Li, Z., Liu, S., and Zhang, D. (2023b). Global, regional and national burden of inflammatory bowel disease in 204 countries and territories from 1990 to 2019: a systematic analysis based on the Global Burden of Disease Study 2019. *BMJ Open* 13, e065186.
- Wang, W.-Q., Jensen, O. N., Møller, I. M., Hebelstrup, K. H., and Rogowska-Wrzesinska, A. (2018). Evaluation of sample preparation methods for mass spectrometry-based proteomic analysis of barley leaves. *Plant Methods* 14, 72.
- Wang, Z., Yu, D., Cupp-Sutton, K. A., Liu, X., Smith, K., and Wu, S. (2020). Development of an Online 2D Ultrahigh-Pressure Nano-LC System for High-pH and Low-pH Reversed Phase Separation in Top-Down Proteomics. *Anal. Chem.* 92, 12774–12777.
- Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., et al. (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* 16, 1090–1094.
- Webber, J., and Clayton, A. (2013). How pure are your vesicles? *J. Extracell. vesicles* 2.
- Weber, M., and Fuchs, T. M. (2022). Metabolism in the Niche: a Large-Scale Genome-Based Survey Reveals Inositol Utilization To Be Widespread among Soil, Commensal, and Pathogenic Bacteria. *Microbiol. Spectr.* 10, e0201322.
- Wei, Y., Gao, J., Liu, D., Li, Y., and Liu, W. (2019). Adaptational changes in physiological and transcriptional responses of *Bifidobacterium longum* involved in acid stress resistance after successive batch cultures. *Microb. Cell Fact.* 18, 156.
- Wen, B., Wang, X., and Zhang, B. (2019). PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res.* 29, 485–493.
- Wexler, H. M. (2007). Bacteroides : the Good, the Bad, and the Nitty-Gritty. *Clin. Microbiol. Rev.* 20, 593–621.
- White, B. A., Ramos, G. P., and Kane, S. (2022). The Impact of Alcohol in Inflammatory Bowel Diseases. *Inflamm. Bowel Dis.* 28, 466–473.
- Wiese, S., Reidegeld, K. A., Meyer, H. E., and Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* 7, 340–350.
- Wilks, J. C., Kitko, R. D., Cleeton, S. H., Lee, G. E., Ugwu, C. S., Jones, B. D., et al. (2009). Acid and Base Stress and Transcriptomic Responses in *Bacillus subtilis*. *Appl. Environ. Microbiol.* 75, 981–990.
- Wingfield, P. T. (2017). N-Terminal Methionine Processing. *Curr. Protoc. Protein Sci.* 88, 6.14.1-6.14.3.
- Winkels, K., Koudelka, T., Kaulich, P. T., Leippe, M., and Tholey, A. (2022). Validation of Top-Down Proteomics Data by Bottom-Up-Based N-Terminomics Reveals Pitfalls in Top-Down-Based Terminomics Workflows. *J. Proteome Res.* 21, 2185–2196.
- Winkler, M. E., and Ramos-Montañez, S. (2009). Biosynthesis of Histidine. *EcoSal Plus* 3.
- Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat. Methods* 6, 359–362.
- Woldringh, C. L. (2002). The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol. Microbiol.* 45, 17–29.
- Won, S., Lee, C., Bae, S., Lee, J., Choi, D., Kim, M., et al. (2023). Mass-produced gram-negative bacterial outer membrane vesicles activate cancer antigen-specific stem-like CD8 + T cells which enables an effective combination immunotherapy with anti-PD-1. *J. Extracell. Vesicles* 12, e12357.
- Wostmann, B. S., Larkin, C., Moriarty, A., and Bruckner-Kardoss, E. (1983). Dietary intake, energy metabolism, and excretory losses of adult male germfree Wistar rats. *Lab. Anim. Sci.* 33, 46–50. Available at: <https://pubmed.ncbi.nlm.nih.gov/6834773/> (Accessed March 7, 2024).
- Wu, H., Zhang, Y., Li, L., Li, Y., Yuan, L., E, Y., et al. (2022). Positive regulation of the DLT operon by TCSR7 enhances acid tolerance of *Lactococcus lactis* F44. *J. Dairy Sci.* 105, 7940–7950.
- Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12, 444.
- Xu, J., and Gordon, J. I. (2003). Honor thy symbionts. *Proc. Natl. Acad. Sci.* 100, 10452–10459.

## BIBLIOGRAPHY

- Yadav, P., Ellinghaus, D., Rémy, G., Freitag-Wolf, S., Cesaro, A., Degenhardt, F., et al. (2017). Genetic Factors Interact With Tobacco Smoke to Modify Risk for Inflammatory Bowel Disease in Humans and Mice. *Gastroenterology* 153, 550–565.
- Yadavalli, S. S., and Yuan, J. (2022). Bacterial Small Membrane Proteins: the Swiss Army Knife of Regulators at the Lipid Bilayer. *J. Bacteriol.* 204, e0034421.
- Yan, D., Ikeda, T. P., Shauger, A. E., and Kustu, S. (1996). Glutamate is required to maintain the steady-state potassium pool in *Salmonella typhimurium*. *Proc. Natl. Acad. Sci. U. S. A.* 93, 6527–6531.
- Yang, P., Patrick, E., Tan, S. X., Fazakerley, D. J., Burchfield, J., Gribben, C., et al. (2014). Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. *Bioinformatics* 30, 808–814.
- Yang, Z., Shen, X., Chen, D., and Sun, L. (2020). Toward a Universal Sample Preparation Method for Denaturing Top-Down Proteomics of Complex Proteomes. *J. Proteome Res.* 19, 3315–3325.
- Yebra, M. J., Zúñiga, M., Beaufils, S., Pérez-Martínez, G., Deutscher, J., and Monedero, V. (2007). Identification of a Gene Cluster Enabling *Lactobacillus casei* BL23 To Utilize myo-Inositol. *Appl. Environ. Microbiol.* 73, 3850–3858.
- Yoshida, K. I., Yamaguchi, M., Morinaga, T., Kinehara, M., Ikeuchi, M., Ashida, H., et al. (2008). myo-inositol catabolism in *Bacillus subtilis*. *J. Biol. Chem.* 283, 10415–10424.
- Zallot, C., Quilliot, D., Chevaux, J. B., Peyrin-Biroulet, C., Guéant-Rodriguez, R. M., Freling, E., et al. (2013). Dietary beliefs and behavior among inflammatory bowel disease patients. *Inflamm. Bowel Dis.* 19, 66–72.
- Zhang, X., Fang, A., Riley, C. P., Wang, M., Regnier, F. E., and Buck, C. (2010). Multi-dimensional liquid chromatography in proteomics-A review. *Anal. Chim. Acta* 664, 101–113.
- Zhang, Z., and van der Donk, W. A. (2019). Nonribosomal Peptide Extension by a Peptide Amino-Acyl tRNA Ligase. *J. Am. Chem. Soc.* 141, 19625–19633.
- Zheng, Y., Wang, J., Bai, X., Chang, Y., Mou, J., Song, J., et al. (2018). Improving the acetic acid tolerance and fermentation of *Acetobacter pasteurianus* by nucleotide excision repair protein UvrA. *Appl. Microbiol. Biotechnol.* 102, 6493–6502.
- Zhuo, L., Hascall, V. C., and Kimata, K. (2004). Inter- $\alpha$ -trypsin Inhibitor, a Covalent Protein-Glycosaminoglycan-Protein Complex. *J. Biol. Chem.* 279, 38079–38082.
- Zocco, M. A., Ainora, M. E., Gasbarrini, G., and Gasbarrini, A. (2007). *Bacteroides thetaiotaomicron* in the gut: Molecular aspects of their interaction. *Dig. Liver Dis.* 39, 707–712.
- Zoetendal, E. G., Ben-Amor, K., Harmsen, H. J. M., Schut, F., Akkermans, A. D. L., and de Vos, W. M. (2002). Quantification of Uncultured *Ruminococcus obeum* -Like Bacteria in Human Fecal Samples by Fluorescent In Situ Hybridization and Flow Cytometry Using 16S rRNA-Targeted Probes. *Appl. Environ. Microbiol.* 68, 4225–4232.

## LIST OF ABBREVIATIONS

---

<b>ACN</b>	Acetonitrile
<b>ADP</b>	Adenosine diphosphate
<b>AMP</b>	Antimicrobial peptide
<b>ATP</b>	Adenosine triphosphate
<b>BCA</b>	Bicinchoninic acid
<b>BHI</b>	Brain-heart infusion
<b>BLASTp</b>	NCBI protein blast
<b>bRP</b>	High-pH (basic) reversed phase
<b>BUP</b>	Bottom-up proteomics
<b>CAA</b>	Chloroacetamide
<b>COG</b>	Clusters of orthologous genes
<b>CV</b>	Compensation voltage
<b>DDA</b>	Data dependent acquisition
<b>DDM</b>	Dodecyl- $\beta$ -D-maltoside
<b>DDT</b>	Dithiothreitol
<b>DIA</b>	Data-independent acquisition
<b>DltD</b>	D-alanyl-lipoteichoic acid biosynthesis protein
<b>EV</b>	Extracellular vesicle
<b>FAIMS</b>	High-field asymmetric waveform ion mobility spectrometry
<b>FASP</b>	Filter-aided sample preparation
<b>FDR</b>	False discovery rate
<b>GDP</b>	Guanosine diphosphate
<b>GELFrEE</b>	Gel-eluted liquid fraction entrapment electrophoresis
<b>GntR</b>	GntR family transcriptional regulator protein
<b>GRAVY</b>	Grand average of hydropathy
<b>GTP</b>	Guanosine triphosphate
<b>HGM</b>	Human gut microbiota
<b>HMWP</b>	High molecular-weight proteome
<b>HPr</b>	Histidine-containing phosphocarrier
<b>Hup</b>	DNA-binding protein HU
<b>IAA</b>	Iodoacetamide
<b>iBAQ</b>	Intensity-based absolute quantification
<b>iPtgxDB</b>	Integrated proteogenomics database
<b>ISD</b>	In-solution digestion
<b>KEGG</b>	Kyoto encyclopedia of genes and genomes
<b>LC-MS</b>	Liquid chromatography mass spectrometry
<b>LFQ</b>	Label free quantification
<b>LMWP</b>	Low molecular weight proteome
<b>LPS</b>	Lipopolysaccharide
<b>MALDI</b>	Matrix-assisted laser desorption/ionization
<b>NlpC</b>	NlpC/P60 domain-containing protein
<b>NME</b>	N-terminal methionine excision
<b>NTA</b>	Nanoparticle tracking analysis
<b>NFO</b>	Na <sup>+</sup> -translocating NADH:ferredoxin oxidoreductase
<b>NQR</b>	Na <sup>+</sup> -translocating NADH dehydrogenase
<b>OMV</b>	Outer membrane vesicle
<b>ORF</b>	Open reading frame

## LIST OF ABBREVIATIONS

<b>PCA</b>	Principal component analysis
<b>PD</b>	Proteome discoverer
<b>pH<sub>i</sub></b>	Intracellular pH
<b>pI</b>	Isoelectric point
<b>Pr</b>	Pearson correlation
<b>PrSM</b>	Proteoform spectrum match
<b>PSM</b>	Peptide spectrum match
<b>PTM</b>	Posttranslational modification
<b>SCFA</b>	Short chain fatty acid
<b>SCX</b>	Strong cation exchange
<b>SDC</b>	Sodium deoxycholate
<b>SDS</b>	Sodium dodecyl sulfate
<b>SEP</b>	Short-open reading encoded peptide
<b>sORF</b>	Short open reading frame
<b>SP3</b>	Single-pot solid-phase-enhanced
<b>SPE</b>	Solid-phase extraction
<b>SPEED</b>	Sample preparation by easy extraction and digestion
<b>STEP</b>	Sub-cellular topology and localisation of <i>E. coli</i> polypeptides
<b>TCEP</b>	Tris(2-carboxyethyl)phosphine
<b>TDP</b>	Top-down proteomics

<b>Ala (A)</b>	Alanine	<b>Leu (L)</b>	Leucine
<b>Arg (R)</b>	Arginine	<b>Lys (K)</b>	Lysine
<b>Asn (N)</b>	Asparagine	<b>Met (M)</b>	Methionine
<b>Asp (D)</b>	Aspartic acid	<b>Phe (F)</b>	Phenylalanine
<b>Cys (C)</b>	Cysteine	<b>Pro (P)</b>	Proline
<b>Gln (Q)</b>	Glutamine	<b>Ser (S)</b>	Serine
<b>Glu (E)</b>	Glutamic acid	<b>Thr (T)</b>	Threonine
<b>Gly (G)</b>	Glycine	<b>Trp (W)</b>	Tryptophan
<b>His (H)</b>	Histidine	<b>Tyr (Y)</b>	Tyrosine
<b>Ile (I)</b>	Isoleucine	<b>Val (V)</b>	Valine

## LIST OF FIGURES

FIGURE I-1   Sources of Proteome Complexity.....	2
FIGURE I-2   Protein Modifications in Bacteria.....	4
FIGURE I-3   The Orbitrap Fusion Lumos Tribrid Mass Spectrometer.....	6
FIGURE I-4   Fragment Ion Nomenclature. ....	7
FIGURE I-5   Generalized Bottom-up Proteomics Workflow. ....	8
FIGURE III-1   Experimental Design for the Comparison of Label-free and TMT-based Quantification.....	41
FIGURE III-2   Experimental Design for Top-Down Proteoform-Directed Analysis.....	42
FIGURE III-3   Overview of TMT Overlabeling.....	43
FIGURE III-4   Evaluation of Mixing and Two-dimensional Separation.....	44
FIGURE III-5   Protein and Peptide Identification Overview.....	45
FIGURE III-6   Comparison of Quantified Proteins for LFQ and TMT.....	46
FIGURE III-7   Analysis of the Proteomic Response between LFQ and TMT.....	47
FIGURE III-8   Overview of LFQ Quantitative Results. ....	48
FIGURE III-9   Sucrose Utilization Pathway in <i>B. thetaiotaomicron</i> . ....	49
FIGURE III-10   Top-down Methodology Comparison.....	50
FIGURE III-11   Neo-termini Analysis of Identified Proteoforms. ....	51
FIGURE III-12   Analysis of Methionine Cleavage. ....	52
FIGURE III-13   Analysis of Proteoform Neo-Termini.....	53
FIGURE III-14   Distribution of Precursor Delta Mass Shifts. ....	54
FIGURE III-15   Putative PTMs in <i>B. thetaiotaomicron</i> proteins.....	55
FIGURE III-16   Distribution of Precursor Mass Shifts on Acyl Carrier Protein (Q8A2E6). ....	56
FIGURE IV-1   Experimental Design to Analyze the pH-Induced Proteomic Response of HGM Bacteria. ....	66
FIGURE IV-2   Experimental Design for Top-down LFQ Analysis of the <i>B. producta</i> Proteome. ....	67
FIGURE IV-3   Evaluation of Data Processing of the <i>B. longum</i> Proteome. ....	68
FIGURE IV-4   <i>B. longum</i> Acidic and Alkaline Proteomic Response. ....	69
FIGURE IV-5   Comparison of the Acidic and Alkaline Response of the <i>B. longum</i> Proteome.....	70
FIGURE IV-6   Protein Identification Overview.....	71
FIGURE IV-7   Quantitative Proteome Analysis of Human Gut Bacteria. ....	72
FIGURE IV-8   Potential Role of Histidine Biosynthesis in <i>B. producta</i> pH <sub>i</sub> Maintenance. ....	73
FIGURE IV-9   Central Nitrogen Metabolism of <i>B. thetaiotaomicron</i> .....	75
FIGURE IV-10   Inositol Metabolism. ....	77
FIGURE IV-11   Identified Proteoforms using the Acidic and Basic HMWP Depletion Method.....	80

## LIST OF FIGURES

FIGURE IV-12   Comparison Top-down and Bottom-up Label-free-quantification. ....	82
FIGURE IV-13   Quantification of the C-terminal Region of the GntR family transcriptional regulator (A0A7G5MZW5). ....	83
FIGURE IV-14   Asp-Pro Cleavage in 30S Ribosomal Protein S16 (A0A4P6M2Y5). ....	84
FIGURE IV-15   Bottom-up Peptide Cleavage Analysis Highlights Asp-Pro Sequence Logos. ....	86
FIGURE IV-16   Peptide Abundance Distribution. ....	86
FIGURE IV-17   Phosphoserine Modification of HPr (A0A2S4GRU0). ....	87
FIGURE IV-18   Sequence Alignment of HPr Proteins Highlighting Proposed Phosphorylation Sites. ....	88
FIGURE IV-19   Quantification of Phosphorylation on HPr (A0A2S4GRU0). ....	88
FIGURE V-1   Experimental Design for Assessing the Impact of Culture Conditions on SEP Production. ....	107
FIGURE V-2   Flowchart of the Multistep SEP Verification Process. ....	110
FIGURE V-3   Validation of Non-canonical Peptides. ....	111
FIGURE V-4   Validation of Non-canonical Proteoforms. ....	112
FIGURE V-5   Metrics of Proteoform Identifications. ....	112
FIGURE V-6   Neo-termini Analysis of Identified SEP Proteoforms. ....	113
FIGURE V-7   Potential Alternative Initiation in SEP Proteoforms. ....	114
FIGURE V-8   AlphaFold Structure Prediction for BP12. ....	114
FIGURE V-9   Presence of SEP across Diverse Culture Conditions. ....	115
FIGURE V-10   Biochemical Predictions of Identified SEP. ....	116
FIGURE VI-1   Workflow to Evaluate OMV-associated proteins and Non-Ionic Detergent OMV Lysis. ....	126
FIGURE VI-2   SDS-PAGE Analysis of Protein Digestion using the Exobacteria Elution Buffer. ....	128
FIGURE VI-3   Nanoparticle Tracking Analysis of <i>E. coli</i> OMVs. ....	129
FIGURE VI-4   Loading Capacity Evaluation of the ExoBacteria OMV Isolation Kit. ....	130
FIGURE VI-5   Caco-2 Wound Healing Assay. ....	131
FIGURE VI-6   Subcellular Topological Distribution of OMV Proteins based on Total Protein Identifications. ....	132
FIGURE VI-7   Subcellular Topological Distribution of OMV Proteins based on Median Abundance. ....	133
FIGURE VI-8   iBAQ Distribution of Potential <i>E. coli</i> OMV Protein Markers. ....	134
FIGURE VI-9   Evaluation of Non-Ionic Detergents on OMV Lysis and Tryptic Digestion. ....	135
FIGURE VI-10   Peptide and Protein Identification Data of Different OMV Sample Preparations. ....	136
FIGURE VI-11   Variability of Different OMV Sample Preparations. ....	137
FIGURE A-12   Comparison of Top-Down Proteomic Sample Preparation Methods. ....	XXXI

FIGURE A-13   Identified Mass Shifts using a Discovery-Based Open Modification Search.	XXXII
FIGURE A-14   Normalization and Correlation of the Bottom-up Proteomic Data of <i>B. longum</i> .	XXXIII
FIGURE A-15   Normalization and Correlation Overview of HGM Proteomic Data.	XXXV
FIGURE A-16   Normalization Overview of the Top-Down Proteomic Data of <i>B. producta</i> .	XLIII
FIGURE A-17   Top-Down Proteomic Data Correlation of <i>B. producta</i> .	XLIII
FIGURE A-18   Structure Prediction of 30S Ribosomal Protein S16.	XLVII
FIGURE A-19   HPr Phosphocarrier Phosphoserine Modifications.	XLVIII
FIGURE A-20   Comparison of Peptides obtained by <i>in silico</i> Tryptic Digestion.	XLIX
FIGURE A-21   Proteoform Characterization of BP12.	XLIX
FIGURE A-22   <i>E. coli</i> Growth Curves.	L
FIGURE A-23   Summary of the Characteristics of STEPdb.	L
FIGURE A-24   Nanoparticle Tracking Analysis of Different Supernatant Volumes.	LII
FIGURE A-25   UpSet Plot Analysis of OMV Sample Preparation Methods.	LIII
FIGURE A-26   Correlation Heatmap of OMV Sample Preparation Methods.	LIII
FIGURE A-27   Impact of Various Sample Preparations on the Relative Abundance Distribution of <i>E. coli</i> OMV Protein Markers.	LIV

## LIST OF TABLES

TABLE II-1   Modified YCFA Medium .....	19
TABLE II-2   100X Trace Elements Solution.....	20
TABLE II-3   M9 Mineral Medium.....	20
TABLE II-4   Parameters for Nanoparticle Tracking Analysis .....	22
TABLE II-5   SCX Chromatographic Separation Scheme.....	26
TABLE II-6   SCX Fraction Collection Scheme .....	26
TABLE II-7   TMT-Labeling Scheme .....	28
TABLE II-8   Fraction Pooling Scheme .....	29
TABLE III-1   Effectiveness of TMT labeling .....	43
TABLE III-2   Top 10 Peptides Across the Seven Pools. ....	44
TABLE IV-1   Quantitative Results of Top-down Proteomics and Bottom-up Proteomics. ....	81
TABLE IV-2   Overlap of Differentially Abundant Proteins between Bottom-up and Top-down Proteome Analyses.....	83
TABLE IV-3   Sequence and Structure Predictions of 30S Ribosomal Protein S16. ....	85
TABLE VI-1   Overview of Protocol-specific Sample Processing Steps.....	127
TABLE VI-2   Summary of Nanoparticle Tracking Analysis.....	129
TABLE A-1   Significant Proteins of the Direction Analysis of <i>B. thetaiotaomicron</i> .....	XXX
TABLE A-2   Top 10 Identified N- and/or C-terminal Truncated Proteoforms.....	XXXI
TABLE A-3   Top-10 Significant Proteins of the Direction Analysis of <i>B. longum</i> . ....	XXXIV
TABLE A-4   Quantitative Data Summary of the Human Gut Bacteria. ....	XXXVI
TABLE A-5   <i>B. producta</i> 's Differentially Abundant Endospore-forming Proteins.....	XLII
TABLE A-6   Differentially Abundant Proteoforms of the Acidic and Basic HMWP Depletion. ....	XLIV
TABLE A-7   Comparison of Differentially Abundant Proteins of the Acidic and Basic HMWP Depletion.....	XLV
TABLE A-8   Quantitative Comparison of Top-down and Bottom-up Proteomic Data. ....	XLVI
TABLE A-9   Overview of Synthetic Peptides Potentially Acting as Antimicrobial Peptides. ....	XLVI
TABLE A-10   List of Proteins Suggested as Potential OMV Markers. ....	LI

<b>1</b>	<b>Proteomic Analysis of <i>B. thtaiotaomicron</i></b>	<b>XXX</b>
1.1	Directional Analysis of LFQ and TMT Quantitative Data	XXX
1.2	Top-down Proteomic Sample Preparation Methods	XXXI
1.3	Proteoform Neo-termini Analysis	XXXI
<b>2</b>	<b>Influence of pH on Bacterial Proteomes</b>	<b>XXXIII</b>
2.1	<i>B. longum</i> Bottom-up Proteomics Data Normalization	XXXIII
2.2	HGM Bottom-up Proteomics Data Normalization	XXXV
2.3	Differentially Abundant Proteins of the HGM Analysis	XXXVI
2.4	<i>B. producta</i> Top-down Proteomics Data Normalization	XLIII
2.5	Differentially Abundant Proteoforms in <i>B. producta</i>	XLIV
2.6	Comparison of LMWP-Top-Down and Full-Proteome Bottom-Up Analysis for <i>B. producta</i>	XLVI
2.7	Synthetic Peptides based on 30S ribosomal protein S16	XLVI
2.8	Phosphoserine Modification of HPr Proteins	XLVIII
<b>3</b>	<b>Proteogenomic Analysis of <i>B. producta</i></b>	<b>XLIX</b>
3.1	Potential Peptide Contamination Analysis	XLIX
3.2	Proteoform Characterization of BP12	XLIX
<b>4</b>	<b>Outer Membrane Vesicles Analysis</b>	<b>L</b>
4.1	<i>E. coli</i> Cultivation	L
4.2	Subcellular Topology and Localization of <i>E. coli</i> Proteins	L
4.3	Potential <i>E. coli</i> OMV Protein Markers	LI
4.4	Nanoparticle Tracking Analysis	LII
4.5	Comparison of OMV Sample Preparation Methods	LIII

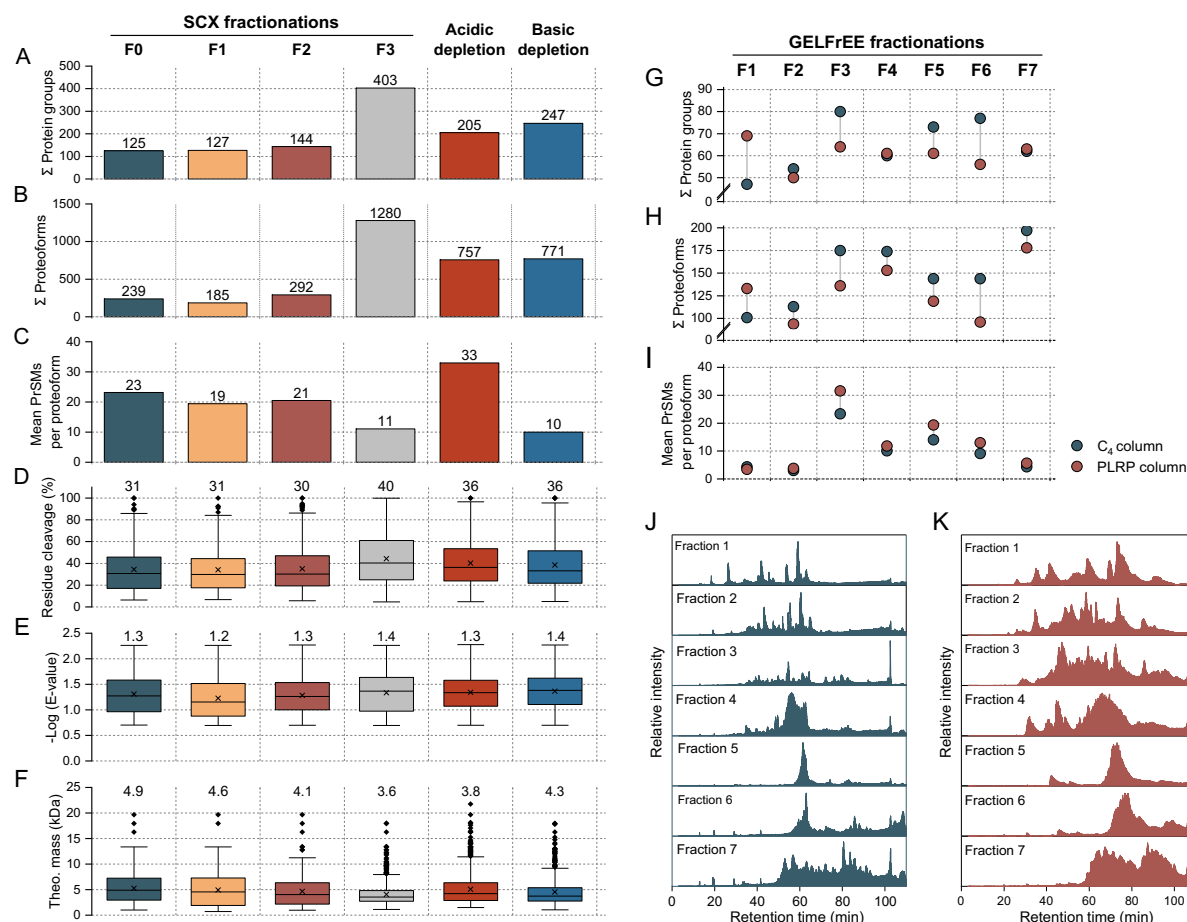
# 1 Proteomic Analysis of *B. thetaiotaomicron*

## 1.1 Directional Analysis of LFQ and TMT Quantitative Data

**TABLE A-1 | Significant Proteins of the Direction Analysis of *B. thetaiotaomicron*.** Higher abundant proteins in the presence of sucrose or glucose are ranked by their p-value.

	ACCESSION	COMMON NAME	P-VALUE
Sucrose	Q8A6W4	SusD homolog	5.0E-06
	Q8A6W6	Glycoside hydrolase family 32	7.2E-06
	Q8A6W9	Fructokinase	9.8E-06
	Q8A6W3	SusC homolog	1.3E-05
	Q8A6W5	DUF4960 domain-containing protein	1.6E-05
	Q8A6W1	Levanase (2,6-beta-D-fructofuranosidase)	2.4E-05
	Q8A6W7	Levanase (2,6-beta-D-fructofuranosidase)	2.9E-05
	Q8A178	SusC homolog	5.8E-05
	Q8A4H3	SusC homolog	7.2E-05
	Q8A193	Alpha-1,2-mannosidase, putative	9.7E-05
	Q8A186	Glycosidase, PH117-related	1.5E-04
	Q8A192	Alpha-mannosidase	1.7E-04
	Q8AA21	Pyruvate, phosphate dikinase	2.0E-04
	Q8A0R3	SusC homolog	2.2E-04
	Q8A182	Alpha-1,2-mannosidase	2.4E-04
	Q8A0R2	SusD homolog	2.6E-04
	Q89ZY1	Aminoglycoside phosphotransferase	2.6E-04
	Q8A185	Glycoside hydrolase family 125 protein	2.7E-04
	Q9RQ13	L-fucose isomerase	4.0E-04
	Q8A728	Pyruvate carboxylase subunit B	4.4E-04
	Q8A2Y6	Alpha-xylosidase	4.6E-04
	Q8A4H0	Alpha-1,2-mannosidase	4.8E-04
	Q8A7M2	Surface protein	5.4E-04
	G8JZS4	Glucan 1,4-alpha-glucosidase SusB	6.2E-04
	Q8A4H9	Alpha-glucosidase	7.9E-04
	Q8A578	SusC homolog	8.4E-04
	Q8AAU4	Galactokinase	8.4E-04
	Q89Z97	Cell surface protein	1.0E-03
	Q8A1Q5	SusC homolog	1.2E-03
	Q8A0Q9	Glycoside hydrolase family 92	1.4E-03
	Q8A373	2,6-beta-D-fructofuranosidase	1.7E-03
	Q8A2A7	Lipoprotein	2.2E-03
Glucose	Q8A2T3	DUF4848 domain-containing protein	5.0E-06
	Q89YT2	DUF4890 domain-containing protein	9.8E-06
	Q8A947	D-ribitol-5-phosphate phosphatase	1.3E-05
	Q8A7U6	ABM domain-containing protein	4.5E-05
	Q8A0B9	Sec-independent protein translocase	8.0E-05
	Q8A197	DUF4890 domain-containing protein	8.8E-05

## 1.2 Top-down Proteomic Sample Preparation Methods



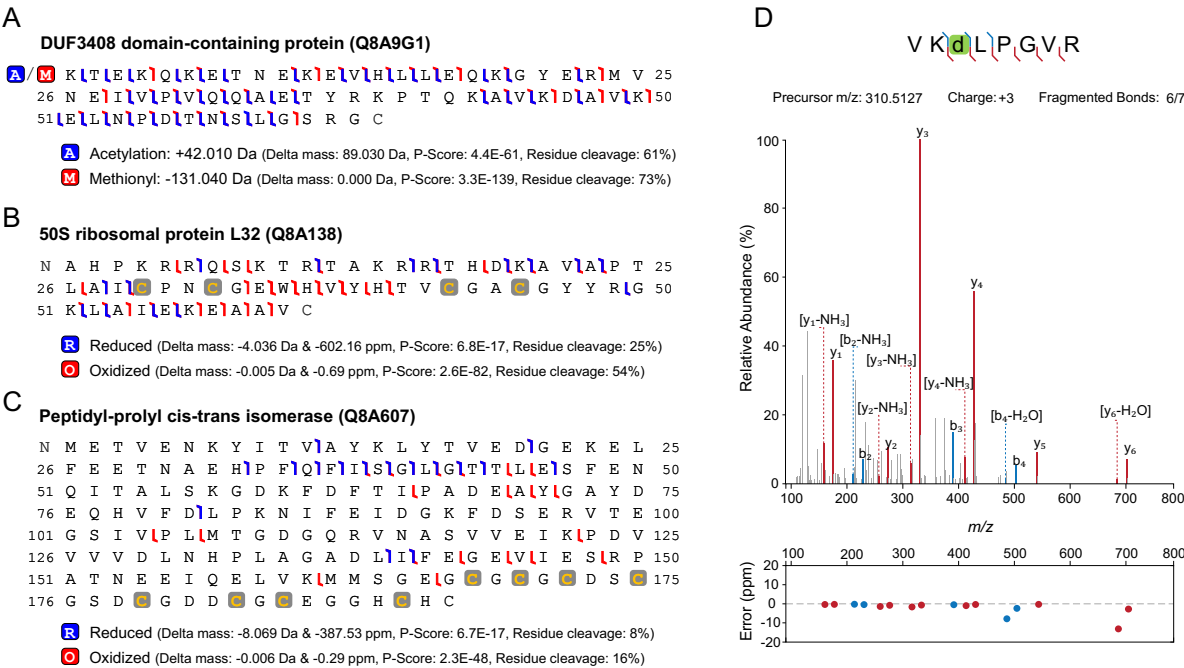
**FIGURE A-12 | Comparison of Top-Down Proteomic Sample Preparation Methods.** (A) Number of protein groups, (B) proteoforms, and (C) mean PrSMs per proteoform, along with the distribution of (D) residue cleavage, (E)  $-\log(E\text{-value})$ , and (F) theoretical proteoform mass for SCX fractions and acidic and basic HMWP depletions. (G) Number of protein groups, (H) proteoforms, and (I) mean PrSMs per proteoform for GELFrEE fractions separated using a C<sub>4</sub> or PLRP-S column. Total ion chromatograms of GELFrEE fractions separated using a (J) C<sub>4</sub> or (K) PLRP-S stationary phases.

## 1.3 Proteoform Neo-termini Analysis

**TABLE A-2 | Top 10 Identified N- and/or C-terminal Truncated Proteoforms.**

	ACCESSION	PROTEIN NAME	$\Sigma$ PROTEOFORMS
N-TERM	Q8ABX7	Uncharacterized protein	74
	Q9RQ15	30S ribosomal protein S16	54
	Q8A484	50S ribosomal protein L29	41
	Q8A758	Uncharacterized protein	40
	Q8AAA0	Heat shock protein	38
	Q8A733	50S ribosomal protein L31 type B	34
	Q8A627	Protein TonB	33
	Q8A468	50S ribosomal protein L7/L12	28
	Q8A487	50S ribosomal protein L24	26
	Q8A467	Ribosomal protein L10	25
C	Q8A758	Uncharacterized protein	37

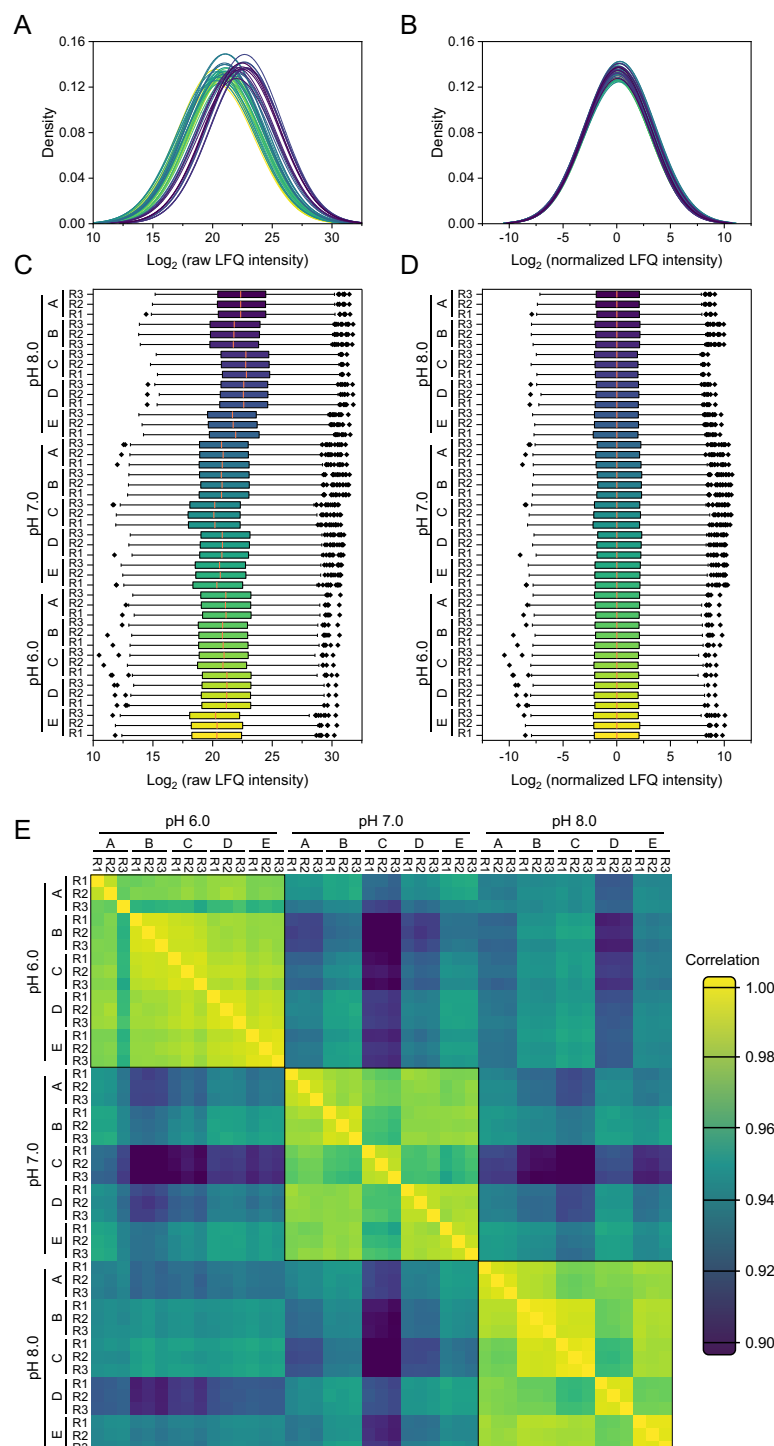
	ACCESSION	PROTEIN NAME	Σ PROTEOFORMS
N-TERM & C-TERM	Q8A6P8	60 kDa chaperonin (GroEL protein)	36
	Q8ABX7	Uncharacterized protein	35
	Q8A484	50S ribosomal protein L29	34
	Q8A494	50S ribosomal protein L30	33
	Q8A6P7	10 kDa chaperonin (GroES protein)	33
	Q8A627	Protein TonB	30
	Q8ABG3	Uncharacterized protein	21
	Q8A2E6	Acyl carrier protein (ACP)	21
	E7MCA7	30S ribosomal protein S21	20
	Q8ABX7	Uncharacterized protein	35
	Q8A627	Protein TonB	30
	Q8ABG3	Uncharacterized protein	21
	Q8A484	50S ribosomal protein L29	18
	Q8A862	Serine protease	16
	Q8A0Z4	30S ribosomal protein S2	14
	Q89ZV7	Large-conductance mechanosensitive channel	11
	Q8A2A1	Translation initiation factor IF-2	8
	Q8A493	30S ribosomal protein S5	7
	E7MCA7	30S ribosomal protein S21	6



**FIGURE A-13 | Identified Mass Shifts using a Discovery-Based Open Modification Search. (A)** Mass shift of +89.03 Da caused by the absence of initiator methionine and misassignments of acetylation (blue). Correct assignment of the missing start methionine (red) improved fragment spectra. Potential disulfide bonds for **(B)** 50S ribosomal protein L32 and **(C)** peptidyl-prolyl cis-trans isomerase, with fragment ions for the reduced (blue) and oxidized (red) proteoform. **(D)** β-Methylthio-aspartic acid on Ribosomal Protein S12. Bottom-up fragment ion spectra with the identified b- ions and y-ions in blue and red, respectively. The dot plot illustrates the mass error in ppm for each observed b and y ion, respectively.

## 2 Influence of pH on Bacterial Proteomes

### 2.1 *B. longum* Bottom-up Proteomics Data Normalization

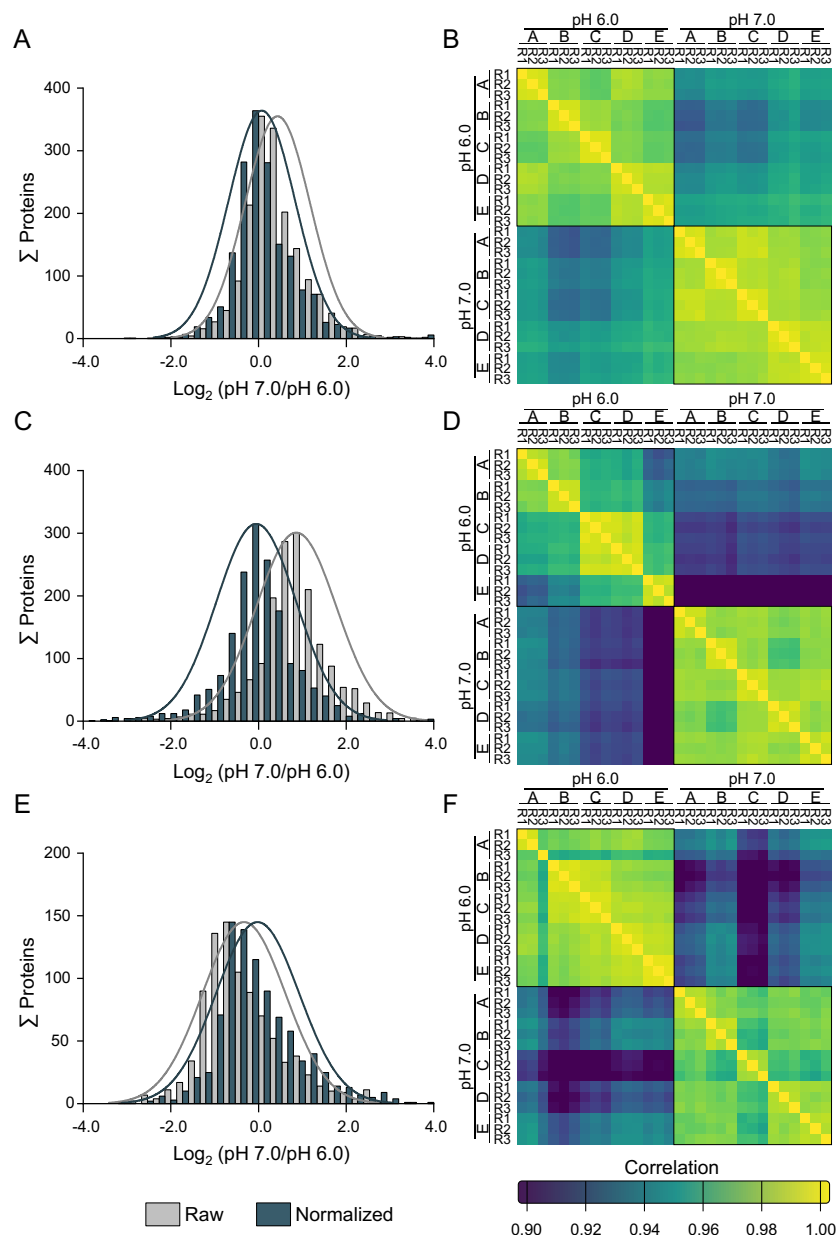


**FIGURE A-14 | Normalization and Correlation of the Bottom-up Proteomic Data of *B. longum*.** Overview of each pH condition which consisted of 5 biological replicates (A-E) and three technical replicates (R1-R3). Smoothed histograms of  $\text{Log}_2$  intensities, (A) before and (B) after median normalization. Box-and-whisker plots of  $\text{Log}_2$  intensities before (C) and (D) after normalization. (E) Pearson correlation coefficients based on protein normalized abundance.

**TABLE A-3 | Top-10 Significant Proteins of the Direction Analysis of *B. longum*.** Proteins of higher abundance at pH 7.0 or higher abundance at pH 6.0 and pH 8.0 are ranked by their p-value. Abbreviations used: EPSP (phosphoenolpyruvate to 5-enolpyruvylshikimate-3-phosphate), DCP (domain-containing protein), TCS (two-component system).

	ACCESSION	COMMON NAME	P-VALUE
HIGHER ABUNDANT PROTEINS AT PH 7.0	Q8G7V3	YhgE/Pip DCP	2.3E-06
	Q8G6N5	Threonine/serine exporter	3.7E-05
	Q8G449	Teichoic acid transporter	9.7E-05
	Q8G7F0	Response regulator of TCS	1.1E-04
	Q8G762	Copper-transporting ATPase	2.3E-04
	Q8G7H2	50S ribosomal protein L32	2.3E-04
	Q8G6I8	MucBP DCP	3.3E-04
	Q8G410	50S ribosomal protein L29	4.8E-04
	Q8G5F5	Sulfur carrier protein ThiS	7.0E-04
	Q8G3P6	50S ribosomal protein L31	1.2E-03
HIGHER ABUNDANT PROTEINS AT PH 6.0 & PH 8.0	Q8G3T0	Probable glycosyltransferase	9.2E-06
	Q8G6W8	3-Oxoacyl-[acyl-carrier protein]	6.9E-05
	Q8G5N6	EPSP synthase	6.9E-05
	Q8G7F8	Inosine-uridine nucleoside hydrolase	1.1E-04
	Q8G3V2	DUF4037 DCP	1.3E-04
	Q8G7B9	Thioredoxin	2.8E-04
	Q8G5S9	T4 RNA ligase 1-like DCP	2.8E-04
	Q8G6D2	Probable aminotransferase Hi0286	4.8E-04
	Q8CY57	Beta-fructofuranosidase	5.9E-04
	Q8G7Q1	Histidine-aspartate DCP	7.4E-04

## 2.2 HGM Bottom-up Proteomics Data Normalization



**FIGURE A-15 | Normalization and Correlation Overview of HGM Proteomic Data. (A-B)** *B. thetaiotaomicron*, **(C-D)** *B. producta* and **(E-F)** *B. longum*. **(A, C and E)** Histograms show the data distribution and Gaussian distribution curve of proteins before and after normalization. **(B, D and F)** Heatmaps representing the distance matrix between the data sets. Pearson correlation coefficients based on protein normalized abundance between three technical replicates (R1-R3) and five biological replicates.

## 2.3 Differentially Abundant Proteins of the HGM Analysis

**TABLE A-4 | Quantitative Data Summary of the Human Gut Bacteria.** Quantitative data of *B. thetaiotaomicron* (BT), *B. producta* (BP), and *B. longum* (BL) were grouped based on COG, RAST, and KEGG annotations. Log<sub>2</sub> ratios (pH 7.0/pH 6.0) of significantly differentially abundant proteins (Two-sided Welch's t-test, Benjamini-Hochberg FDR corrected with a q-value of 0.05) are represented by different colors. A Log<sub>2</sub> ratio of 0.48 (■), 1.00 (■), 1.58 (■), 2.00 (■), 2.32 (■) and 2.58 (■) indicates higher abundance at pH 7.0, while a Log<sub>2</sub> ratio of -0.48 (■), -1.00 (■), -1.58 (■), -2.00 (■), -2.32 (■) - 2.58 (■) indicates higher abundance at pH 6.0.

	BT	BP	BL
<b>Amino acid metabolism</b>			
<i>Arginine biosynthesis</i>			
Acetylglutamate kinase (argB)			
Acetylornithine aminotransferase (argD)			
Ornithine carbamoyltransferase (argF)			
Arginine repressor (argR)			
<i>Aspartate and Asparagine metabolism</i>			
L-asparaginase (ansB)			
Asparagine synthetase B, glutamine-hydrolyzing (asnB)			
Aspartate ammonia-lyase (aspA)			
Aspartate aminotransferase (aspC)			
Aspartate decarboxylase AsdA (aspD)			
Aspartokinase (lysC)			
L-aspartate oxidase (EC 1.4.3.16) (nadB)			
<i>Cysteine and methionine metabolism</i>			
Adenosylhomocysteinase (ahcY)			
O-acetylhomoserine (Thiol)-lyase (cysD)			
Serine O-acetyltransferase (cysE)			
Cysteine synthase (cysK)			
Homoserine dehydrogenase (hom)			
Homoserine O-acetyltransferase (metA)			
Methionine synthase (metH)			
Methionine import ATP-binding protein MetN (metN)			
O-acetylhomoserine aminocarboxypropyltransferase (metY)			
Cysteine desulfurase (sufS)			
<i>Glycine, serine and threonine metabolism</i>			
Glycine dehydrogenase (gcvP)			
Homoserine dehydrogenase (hom)			
Glycine acetyltransferase (kbl)			
D-3-phosphoglycerate dehydrogenase (serA)			
Homoserine kinase (thrB)			
Threonine synthase (thrC)			
<i>Histidine biosynthesis</i>			
ProFAR isomerase (hisA)			
Histidine biosynthesis bifunctional protein HisB (hisB)			
Histidinol-phosphate aminotransferase (hisC)			
Histidinol dehydrogenase (hisD)			
Phosphoribosyl-ATP pyrophosphatase (hisE)			
Imidazole glycerol phosphate synthase subunit HisF (hisF)			
ATP phosphoribosyltransferase (hisG)			
Imidazole glycerol phosphate synthase subunit HisH (hisH)			
Histidine biosynthesis bifunctional protein HisI (hisI)			
ATP phosphoribosyltransferase regulatory subunit (hisZ)			

	BT	BP	BL
<i>Histidine catabolism</i>			
Histidine ammonia-lyase (hutH)			
<i>BCAA biosynthesis (leucine, isoleucine and valine)</i>			
Threonine ammonia-lyase (ilvA)			
Acetolactate synthase (ilvB)			
Dihydroxy-acid dehydratase (ilvD)			
Acetolactate synthase small subunit (ilvN)			
2-isopropylmalate synthase (leuA)			
3-isopropylmalate dehydratase small subunit (leuD)			
L-threonine dehydratase catabolic (TdcB)			
<i>Lysine biosynthesis</i>			
4-hydroxy-tetrahydrodipicolinate synthase (dapA)			
Dihydrodipicolinate reductase (dapB)			
N-succinyl-diaminopimelate aminotransferase (dapC)			
Tetrahydrodipicolinate N-succinyltransferase (dapD)			
Succinyl-diaminopimelate desuccinylase (dapE)			
Diaminopimelate epimerase (dapF)			
Acetyltransferase (dapH)			
LL-diaminopimelate aminotransferase (dapL)			
Diaminopimelate decarboxylase (LysA)			
<i>Phenylalanine, tyrosine and tryptophan biosynthesis</i>			
5-enolpyruvylshikimate-3-phosphate synthase (aroA)			
Bifunctional shikimate kinase/3-dehydroquinate synthase (aroB)			
Chorismate synthase (aroC)			
Shikimate dehydrogenase (aroE)			
3-deoxy-7-phosphoheptulonate synthase (aroG)			
3-dehydroquinate dehydratase (aroQ)			
Tryptophan synthase alpha chain (trpA)			
Anthranilate phosphoribosyltransferase (trpD)			
Anthranilate synthase component I (trpE)			
Prephenate dehydrogenase (tyrA)			
<i>Proline biosynthesis</i>			
Gamma-glutamyl phosphate reductase (proA)			
Glutamate 5-kinase (proB)			
Pyrroline-5-carboxylate reductase (proC)			
<i>GS-GOGAT &amp; GDH</i>			
Glutamate dehydrogenase (gdhA)			
Glutamine synthetase (glnA)			
Glutamate synthase large subunit (gltB)			
Glutamate synthase small subunit (gltD)			
Glutamate synthase small subunit (gltD2)			
<i>Glutamate metabolism</i>			
LL-diaminopimelate aminotransferase (dapL)			
Folypolyglutamate synthase (folC)			
UMAG ligase (murD)			
Glutamate racemase (murl)			
Glutamate 5-kinase (proB)			
<i>Glutamine metabolism</i>			
Asparagine synthase (asnB)			
Carbamoylphosphate synthase small subunit (carA)			
Carbamoylphosphate synthase large subunit (carB)			
Glucosamine 6-phosphate synthetase (glmS)			
L-aspartate oxidase (nadB)			

	<b>BT</b>	<b>BP</b>	<b>BL</b>
NAD <sup>+</sup> synthase (nadE)			
Amidophosphoribosyltransferase (purF)			
Phosphoribosylformylglycinamide synthase (purL)			
<b>GABA biosynthesis</b>			
Glutaminase (glsA)			
Glutamate decarboxylase (gadB)			
<b>Carbohydrate metabolism</b>			
<b>Glycolysis/Gluconeogenesis</b>			
Hexokinase			
Fructose-1,6-bisphosphatase (fbp)			
Phosphofructokinase (pfkA)			
Glyceraldehyde-3-phosphate dehydrogenase (gap)			
Phosphoglycerate kinase (pgk)			
Phosphoglyceromutase (gpmI)			
Pyruvate kinase (pyk)			
<b>Glycogen metabolism</b>			
Glycogen synthase (glgA)			
1,4- $\alpha$ -glucan branching enzyme GlgB (glgB)			
Glucose-1-phosphate adenylyltransferase (glgC)			
$\alpha$ -1,4 glucan phosphorylase (glgP)			
Glycogen debranching enzyme GlgX (glgX)			
4- $\alpha$ -glucanotransferase (Amylomaltase) (malQ)			
<b>Pentose phosphate pathway (oxidative branch)</b>			
Glucose-6-phosphate dehydrogenase (zwf)			
6-Phosphogluconate dehydrogenase (gnd)			
<b>Pentose phosphate pathway (non-oxidative branch)</b>			
Ribose-5-phosphate isomerase (rpiA)			
Ribulose-phosphate 3-epimerase (rpe)			
Transaldolase (tal)			
Transketolase (tkt)			
<b>SCFA production/TCA cycle</b>			
Aconitate hydratase (Aconitase) (acnA)			
Acetyl-coenzyme A synthetase (acsA)			
Aldehyde-alcohol dehydrogenase (adhE)			
Fumarate hydratase (fumB)			
CoA-acylating methylmalonate-semialdehyde dehydrogenase (mmsA)			
Methylmalonyl-CoA mutase small subunit (mutA)			
Methylmalonyl-CoA mutase large subunit (mutB)			
Pyruvate formate-lyase-activating enzyme (pflA)			
Formate acetyltransferase (pflB)			
Phosphoenolpyruvate carboxylase (ppc)			
Phosphate acetyltransferase (pta)			
Pyruvate carboxylase subunit B (pycB)			
Succinyl-CoA synthetase subunit beta (sucC)			
Succinyl-CoA synthetase subunit alpha (sucD)			
Phosphate butyryltransferase			
Dihydrolipoyl dehydrogenase (lpdA)			
Isocitrate dehydrogenase (icd)			
Pyruvate dehydrogenase (poxB)			
Citrate synthase (prpC)			
<b>Inositol biosynthesis</b>			
Inositol-phosphate synthase (ino1)			
Inositol-monophosphatase (suhB)			

	BT	BP	BL
<b>Inositol metabolism</b>			
Inositol 2-dehydrogenase (idhA)			
Scyllo-inositol 2-dehydrogenase (iolX)			
2-keto-myo-inositol dehydratase (iolE1)			
2-keto-myo-inositol dehydratase (iolE2)			
3D-(3,5/4)-trihydroxycyclohexane-1,2-dione acylhydrolase (iolD)			
5-deoxy-glucuronate isomerase (iolB)			
5-dehydro-2-deoxygluconokinase (iolC)			
6-phospho-5-dehydro-2-deoxy-D-gluconate aldolase (iolJ)			
Inosose isomerase (iolI)			
2-keto-myo-inositol dehydratase (iolE3)			
<b>Cellular respiration</b>			
<b>F-Type ATPase</b>			
Subunit alpha (atpA)			
Epsilon chain (atpC)			
Subunit beta (atpD)			
Subunit E (atpE)			
Gamma chain (atpG)			
Subunit delta (atpH)			
<b>V-Type ATPase</b>			
Alpha chain (ntpA)			
Subunit B (ntpB)			
Subunit F (ntpF)			
Subunit H (ntpH)			
<b>NADH ubiquinone oxidoreductase (NUO)</b>			
Subunit B (nuoB)			
Subunit C (nuoC)			
Subunit I (nuoI)			
<b>Na<sup>+</sup>-translocating NADH ubiquinone oxidoreductase (NQR)</b>			
Subunit F (nqrF)			
<b>Na<sup>+</sup>-translocating NADH:ferredoxin oxidoreductase (NFO)</b>			
Subunit B (nrfB)			
<b>Transcriptional and Translational Processes</b>			
<b>DNA-methyltransferases</b>			
DNA-methyltransferase (adenine-specific) (hsdM)			
6-O-methylguanine-DNA methyltransferase (ogt)			
<b>RNA-methyltransferases</b>			
Ribosomal RNA small subunit methyltransferase A (ksgA)			
tRNA (adenine(58)-N(1))-methyltransferase TrmI (trmI)			
tRNA N6-adenosine threonylcarbamoyltransferase (tsaD)			
RNA methyltransferase RlmN (rlmN)			
Ribosomal RNA small subunit methyltransferase E (rsmE)			
SAM-dependent methyltransferase (rsmF)			
Ribosomal RNA small subunit methyltransferase H (rsmH)			
Ribosomal RNA small subunit methyltransferase I (rsmI)			
tRNA (cytidine(34)-2'-O)-methyltransferase (spoU)			
<b>RNA polymerases</b>			
RNA polymerase subunit beta (rpoB)			
RNA polymerase subunit beta' (rpoC)			
RNA polymerase subunit omega (rpoZ)			
<b>Translation factors</b>			
50S ribosomal subunit assembly factor BipA (typA)			

APPENDIX

	<i>BT</i>	<i>BP</i>	<i>BL</i>
Elongation factor 4 (LepA)			
Elongation factor G (fusA2)			
Ribosome hibernation promoting factor (hpf)			
Translation initiation factor IF-1 (infA)			
Translation initiation factor IF-3 (infC)			
Ribosome-releasing factor (frr)			
Ribosome-binding factor A (rbfA)			
Ribosomal silencing factor RsfS (rsfS)			
<i>Ribosomal proteins</i>			
50S ribosomal protein L5 (rplE)			
50S ribosomal protein L9 (rplI)			
50S ribosomal protein L10 (rplJ)			
50S ribosomal protein L11 (rplK)			
50S ribosomal protein L7/L12 (rplL)			
50S ribosomal protein L13 (rplM)			
50S ribosomal protein L14 (rplN)			
50S ribosomal protein L15 (rplO)			
50S ribosomal protein L16 (rplP)			
50S ribosomal protein L17 (rplQ)			
50S ribosomal protein L18 (rplR)			
50S ribosomal protein L20 (rplT)			
50S ribosomal protein L21 (rplU)			
50S ribosomal protein L22 (rplV)			
50S ribosomal protein L23 (rplW)			
50S ribosomal protein L24 (rplX)			
50S ribosomal protein L27 (rpmA)			
50S ribosomal protein L28 (rpmB)			
50S ribosomal protein L29 (rpmC)			
50S ribosomal protein L30 (rpmD)			
50S ribosomal protein L31 (rpmE)			
50S ribosomal protein L32 (rpmF)			
50S ribosomal protein L33 (rpmG)			
50S ribosomal protein L35 (rpml)			
50S ribosomal protein L36 (rpmJ)			
30S ribosomal protein S5 (rpsE)			
30S ribosomal protein S6 (rpsF)			
30S ribosomal protein S7 (rpsG)			
30S ribosomal protein S8 (rpsH)			
30S ribosomal protein S9 (rpsI)			
30S ribosomal protein S10 (rpsJ)			
30S ribosomal protein S11 (rpsK)			
30S ribosomal protein S12 (rpsL)			
30S ribosomal protein S13 (rpsM)			
30S ribosomal protein S14 type Z (rpsN)			
30S ribosomal protein S15 (rpsO)			
30S ribosomal protein S16 (rpsP)			
30S ribosomal protein S17 (rpsQ)			
30S ribosomal protein S18 (rpsR)			
30S ribosomal protein S19 (rpsS)			
30S ribosomal protein S20 (rpsT)			
<i>Aminoacyl-tRNA synthesis</i>			
Alanyl-tRNA synthetase (alaS)			
Aspartyl-tRNA synthetase (aspS)			

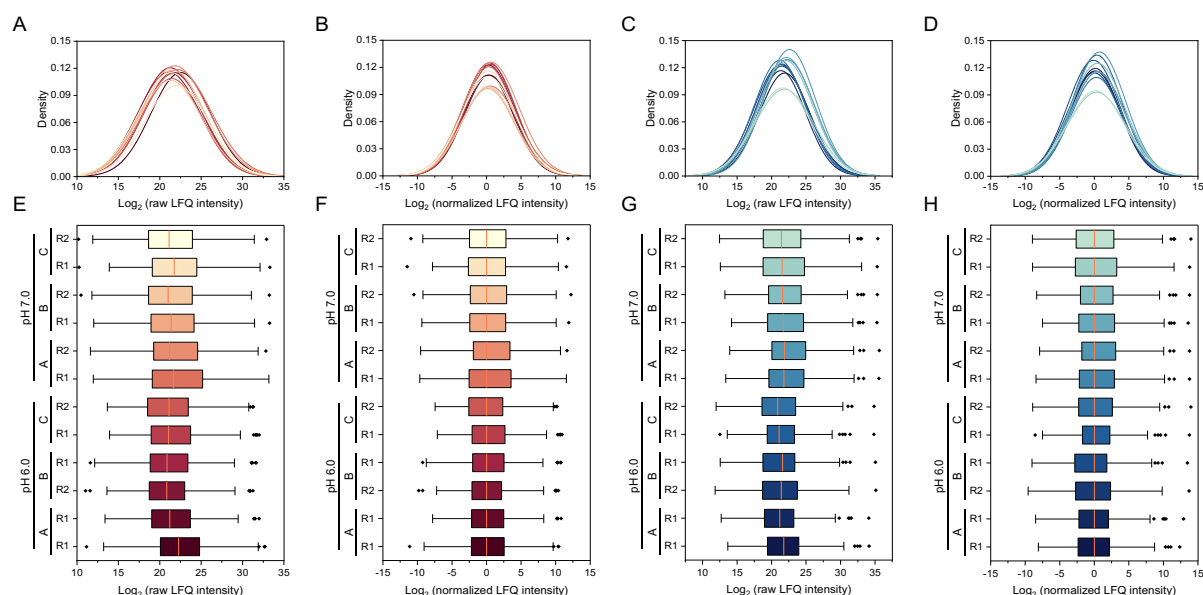
	BT	BP	BL
CysteinyI-tRNA synthetase (cysS)			
GlutaminyI-tRNA synthetases (glnS)			
Glycyl-tRNA synthetase (glyQS)			
Histidyl-tRNA synthetase (hisS)			
Isoleucyl-tRNA synthetase (ileS)			
Leucyl-tRNA synthetase (leuS)			
Methionyl-tRNA formyltransferase (fmt)			
Methionyl-tRNA synthetase (metG)			
Phenylalanyl-tRNA synthetase beta subunit (pheT)			
Prolyl-tRNA synthetase (proS)			
Threonyl-tRNA synthetase (thrS)			
Tyrosyl-tRNA synthetase (tyrS)			
Valyl-tRNA synthetase (valS)			
<b>Protection and repair of macromolecules</b>			
<i>DNA repair</i>			
ATP-dependent helicase/nuclease subunit A (addA)			
DNA polymerase IV (dinB)			
Replicative DNA helicase (dnaB)			
DNA polymerase III subunit alpha (dnaE)			
DNA primase (dnaG)			
Beta sliding clamp (dnaN)			
DNA gyrase subunit A (gyrA)			
DNA gyrase subunit B (gyrB)			
DNA polymerase III subunit delta (hoIA)			
DNA polymerase III subunit delta' (hoIB)			
DNA ligase (ligA)			
Transcription-repair-coupling factor (mfd)			
DNA mismatch repair protein MutS (mutS)			
DNA mismatch repair protein MutL (mutL)			
DNA topoisomerase (parC)			
DNA helicase (pcrA)			
DNA polymerase I (polA)			
Primosomal protein N' (priA)			
DNA repair protein RadA (radA)			
Recombinase A (recA)			
ATP-dependent exoDNAse (recB)			
ATP-dependent exoDNAse (recD)			
DNA replication and repair protein RecF (recF)			
Single-stranded-DNA-specific exonuclease RecJ (recJ)			
DNA helicase (recQ)			
Holliday junction ATP-dependent DNA helicase (RuvB)			
ATP-dependent exonuclease (sbcC)			
DNA topoisomerase III (topB)			
DNA primase TraC (traC)			
UvrABC system, subunit A (uvrA)			
UvrABC system, subunit B (uvrB)			
UvrABC system, subunit C (uvrC)			
UvrD-like helicase (uvrD)			
Exonuclease VII large subunit (xseA)			
Exonuclease VII small subunit (xseB)			
<i>Protein repair</i>			
Chaperone protein ClpB (clpB)			
Chaperone protein ClpC (clpC)			

	BT	BP	BL
Chaperone protein ClpX (clpX)			
Chaperone DnaJ (dnaJ)			
Chaperone protein DnaK (dnaK)			
Co-Chaperonin GroES (HSP10) (groS)			
Chaperone (small HSP)			
<b>Cell Wall Morphogenesis</b>			
<i>Cell cycle control &amp; cell division</i>			
ATP-dependent zinc metalloprotease FtsH (ftsH)			
Penicillin-binding transpeptidase (ftsI)			
Cell division protein FtsL (ftsL)			
Penicillin-binding protein 1A (ponA)			
<i>Cell morphogenesis</i>			
Capsule biosynthesis protein CapA (capA)			
Cell wall hydrolase			
Cell shape-determining protein MreB (mreB)			
Beta-lactamase (pbpA2)			
Phosphoglucosamine mutase (glmM)			
Glucosamine-6-phosphate synthase (glmS)			
Bifunctional protein GlmU (glmU)			
UDP-N-acetylglucosamine 1-carboxyvinyltransferase (murA)			
UDP-N-acetylmuramate dehydrogenase (murB)			
UDP-N-acetylmuramate--L-alanine ligase (murC)			
UDP-N-acetylmuramoylalanine--D-glutamate ligase (murD)			
UDP-N-acetylmuramyl-tripeptide synthetase (murE)			
UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase (murF)			
Alanine racemase (alr)			
D-alanine--D-alanine ligase (ddl)			
Isoprenyl transferase (uppS)			
Cyclopropane-fatty-acyl-phospholipid synthase (cfa)			
<i>Lipopolysaccharide biosynthesis (Gram-negative)</i>			
Lipopolysaccharide biosynthesis protein			
<i>Lipoteichoic acid biosynthesis (Gram-positive)</i>			
D-alanyl-lipoteichoic acid biosynthesis protein DltD (dltD)			
Teichoic acid biosynthesis protein			
Teichoic acid transporter			

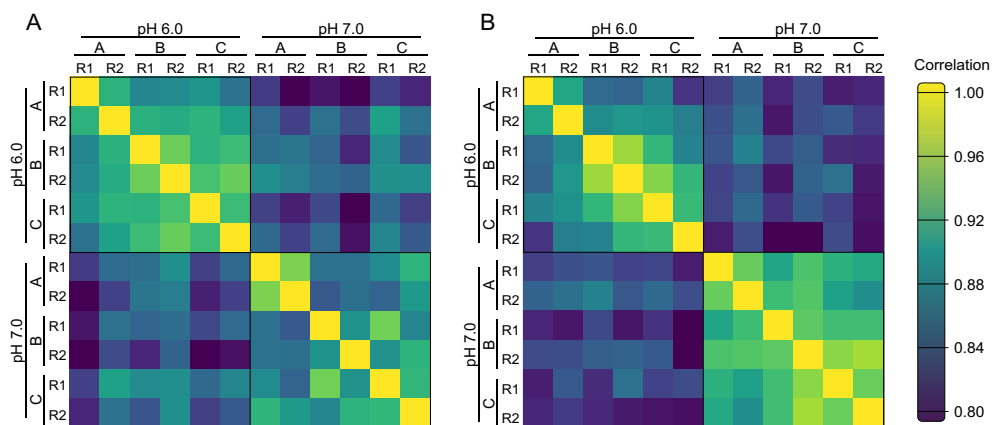
**TABLE A-5 | *B. producta*'s Differentially Abundant Endospore-forming Proteins.** Log<sub>2</sub> ratios (pH 7.0/pH 6.0) of significantly differentially abundant proteins (Two-sided Welch's t-test, Benjamini-Hochberg FDR corrected with a q-value ≤ 0.05) are represented by different colors. A Log<sub>2</sub> ratio of -0.48 (■), -1.58 (■), -2.32 (■) indicates higher abundance at pH 6.0.

GENE LOCUS	ACCESSION	PROTEIN NAME	Log <sub>2</sub> FC
E5259_04690	A0A2S4GGM0	Stage 0 sporulation protein A homolog	
E5259_18460	A0A4P6LZ09	Stage 0 sporulation protein A homolog	
E5259_09515	A0A4P6M3M0	Stage 0 sporulation protein A homolog	
E5259_07175	A0A4P6M2X9	Stage 0 sporulation protein A homolog	
E5259_07275	A0A4P6M749	Stage 0 sporulation protein A homolog	
E5259_19780	A0A4P6LX72	Stage 0 sporulation protein A homolog	
E5259_23915	A0A7G5N0J4	Stage 0 sporulation protein A homolog	
E5259_25430	A0A7G5N1B1	Stage III sporulation protein AB	

## 2.4 *B. producta* Top-down Proteomics Data Normalization



**FIGURE A-16 | Normalization Overview of the Top-Down Proteomic Data of *B. producta*.** Median data normalization of (A-D) the acidic and (E-H) basic HMWP depletion. Smoothed histograms of Log<sub>2</sub> intensities before (A-C) any preprocessing and (B-D) after median normalization. Box-and-whisker plots of Log<sub>2</sub> intensities in each sample before (E-F) and (G-H) after normalization. Boxes capture lower quartile and upper quartile with the median displayed as a horizontal orange line in the middle; whiskers represent minimum and maximum values that fall within 1.5 times the interquartile range.



**FIGURE A-17 | Top-Down Proteomic Data Correlation of *B. producta*.** Heatmap representing the distance matrix between (A) the acidic or (B) basic depletion data set. Pearson correlation coefficients based on proteoform normalized abundance between each pair of technical and biological replicates of each sample. Black squares indicate the biological replicates originating from the same conditions. For this experiment, each pH condition consists of 3 biological replicates (A-C) and two technical replicates (R1-R2).

## 2.5 Differentially Abundant Proteoforms in *B. producta*

**TABLE A-6 | Differentially Abundant Proteoforms of the Acidic and Basic HMWP Depletion.** Log<sub>2</sub> ratio (pH 7.0/pH 6.0). Protein abbreviations used: DCP (domain-containing protein), TSBP (transporter substrate-binding protein). PTM abbreviations used: C (cysteine dehydro modification, indicating disulfide bond), M (methionine formylation) and S (O-phosphopantetheine-L-serine)

	PROTEIN NAME	SEQUENCE	PTM	LOG <sub>2</sub> RATIO	LENTGH
ACIDIC HMWP DEPLETION	30S ribosomal protein S14 type Z	[M].AK...LA.[Y]	C	-3.38	47
	30S ribosomal protein S16	[M].AV...YD.[P]		-2.02	40
	30S ribosomal protein S16	[M].AV...QD.[P]		-1.54	44
	50S ribosomal protein L19	[-].MN...VK.[-]		-1.69	115
	50S ribosomal protein L29	[D].LK...AE.[-]		3.57	61
	50S ribosomal protein L29	[F].QN...AE.[-]		1.78	36
	50S ribosomal protein L30	[M].AD...EI.[-]		1.27	59
	Carbohydrate ABC TSBP	[V].KK...VK.[-]		-2.88	62
	D-alanyl-lipoteichoic acid synthesis	[I].NW...LS.[K]		-2.47	36
	DNA-binding protein HU	[-].MN...NE.[-]		-1.33	91
	DNA-binding protein HU	[-].MN...TG.[E]		-3.50	66
	DNA-binding protein HU	[-].MN...RE.[G]		-2.44	59
	DNA-binding protein HU	[-].MN...GF.[G]		-2.05	47
	RNA-polymerase omega subunit	[-].MI...EN.[-]		-1.96	86
	Flagellar export protein FliJ	[M].AR...NQ.[-]		-2.36	72
	GntR family transcriptional regulator	[N].MI...KK.[-]		2.27	44
	LPXTG cell wall anchor	[L].SP...EA.[-]		-2.03	51
	NlpC/P60 domain-containing protein	[A].AR...YL.[Y]		-6.36	94
	Recombination-promoting nuclease	[G].ME...KI.[-]		3.02	52
	Uncharacterized protein	[M].AE...KK.[-]	C	-2.76	60
	Uncharacterized protein	[L].IT...EQ.[I]		2.72	57
	Uncharacterized protein	[-].ME...RM.[-]		-2.68	103
	Uncharacterized protein	[M].PS...NM.[-]		-2.09	68
	Uncharacterized protein	[-].MK...GK.[-]		2.72	60
	UPF0297 protein C3R19_06560	[-].ME...RD.[-]		-2.45	90
	XRE family transcriptional regulator	[M].GL...KK.[-]		-1.66	140
BASIC HMWP DEPLETION	50S ribosomal protein L29	[-].MK...AE.[-]		1.55	70
	50S ribosomal protein L29	[N].AT...AE.[-]		2.24	34
	Acyl carrier protein	[-].ML...EE.[-]		3.35	77
	ATP synthase	[E].AK...SE.[-]	M, S	-2.43	86
	Carbohydrate ABC TSBP	[A].GY...AV.[T]		2.00	52
	Chaperonin GroEL	[G].YG...ID.[A]		3.75	63
	D-alanyl-lipoteichoic acid synthesis	[I].NW...LS.[K]		-1.87	36
	RNA-polymerase omega subunit	[L].AT...EN.[-]		3.21	53
	RNA-polymerase omega subunit	[A].TS...EN.[-]		3.19	52
	DUF1002 domain-containing protein	[L].ME...FS.[-]		-6.01	55
	DUF1858 domain-containing protein	[M].MT...DK.[-]	C	1.01	66
	DUF1858 domain-containing protein	[M].TI...DK.[-]	C	2.69	65
	DUF3502 domain-containing protein	[L].AP...VL.[N]		7.82	65
	DUF5067 domain-containing protein	[L].IM...IN.[F]		3.89	55

PROTEIN NAME	SEQUENCE	PTM	LOG <sub>2</sub> RATIO	LENTGH
Galactoside transporter permease	[K].VI...DE.[I]		1.56	46
GntR family transcriptional regulator	[L].VY...KK.[-]		3.93	60
GntR family transcriptional regulator	[N].MI...KK.[-]		2.29	44
GntR family transcriptional regulator	[D].NL...KK.[-]		1.85	41
NADH peroxidase	[D].AI...FG.[-]		0.98	29
NlpC/P60 domain-containing protein	[V].RA...DV.[S]		-6.75	70
NlpC/P60 domain-containing protein	[A].AR...ET.[K]		-2.85	69
NlpC/P60 domain-containing protein	[A].SN...VN.[-]		0.98	30
Recombinase RecT	[R].MG...QN.[A]	C	0.89	68
RNA-binding S4 DCP	[-].ME...VK.[-]	M	3.28	69
Sugar-binding protein	[I].GD...EL.[A]		-2.62	62
Uncharacterized protein	[-].MF...AG.[-]		4.93	62
Uncharacterized protein	[L].DD...PQ.[-]		1.79	53
Uncharacterized protein	[L].KQ...KL.[A]		-4.35	38
Uncharacterized protein	[G].ME...KI.[-]		2.38	52
Uncharacterized protein	[A].AR...EK.[-]		-4.35	37

**TABLE A-7 | Comparison of Differentially Abundant Proteins of the Acidic and Basic HMWP Depletion.** Colored fold changes (Log<sub>2</sub> ratio pH 7.0/pH 6.0) indicate differentially abundant proteoforms. PTM abbreviations used: F (methionine formylation) and A (methionine acetylation)

PROTEIN NAME	PROTEOFORM	FRAGMENT (#AA)	LOG <sub>2</sub> RATIO	
			ACIDIC	BASIC
NlpC/P60 domain-containing protein	[A].AR...KT.[K]	Intern (69)	<b>-6.36</b>	-
	[V].RA...DV.[S]	Intern (70)	-	<b>-6.75</b>
D-alanyl-lipoteichoic acid biosynthesis protein	[I].NW...LS.[K]	Intern (36)	<b>-2.47</b>	<b>-1.87</b>
	[T].AK...AI.[H]	Intern (59)	-0.26	-
GntR family transcriptional biosynthesis regulator	[N].MI...EK.[-]	N-term (44)	<b>2.27</b>	<b>2.29</b>
	[D].NL...KK.[-]	N-term (41)	2.60	<b>1.85</b>
	[L].VY...EK.[-]	N-term (60)	-	<b>3.93</b>
	[N].LK...KK.[-]	N-term (40)	-0.34	0.38
	[L].KT...KK.[-]	N-term (39)	-0.13	-0.89
	[N].MI...EE.[E]	Intern (42)	1.93	-
DNA-directed RNA polymerase subunit omega	[-].MI...EN.[-]	Complete (86)	<b>-1.96</b>	0.34
	[L].AT...EN.[-]	N-term (53)	-	<b>3.21</b>
	[A].TS...EN.[-]	N-term (52)	-	<b>3.19</b>
Ribosomal protein uL29	[-].mK...AE.[-]	Complete (70) <sup>F</sup>	<b>2.60</b>	-
	[-].mK...AE.[-]	Complete (70) <sup>A</sup>	-0.68	-
	[-].MK...AE.[-]	Complete (70)	-	<b>1.55</b>
	[N].AT...AE.[-]	N-term (34)	-	<b>2.24</b>
	[Y].VE...AE.[-]	N-term (64)	2.28	-
	[D].LK...AE.[-]	N-Term (61)	<b>3.57</b>	-
	[F].QN...AE.[-]	N-Term (36)	<b>1.78</b>	-
	[Q].LD...AE.[-]	N-Term (30)	1.65	-
Recombination-promoting nuclease RpnA	[G].ME...KI.[-]	N-term (52)	<b>3.02</b>	<b>2.38</b>

## 2.6 Comparison of LMWP-Top-Down and Full-Proteome Bottom-Up

### Analysis for *B. producta*

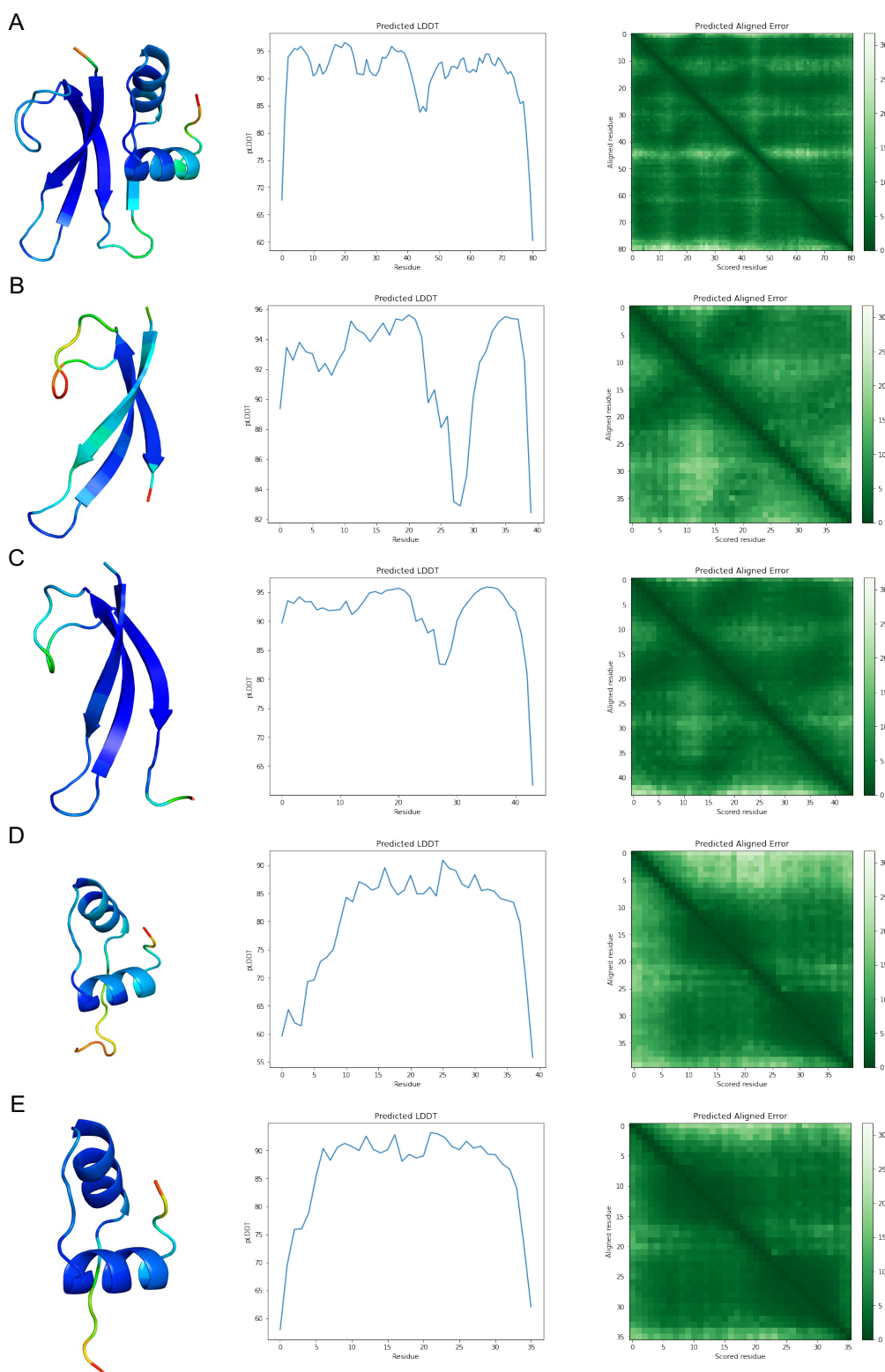
**TABLE A-8 | Quantitative Comparison of Top-down and Bottom-up Proteomic Data.** Comparison of proteoform- and protein-level changes between full proteome LFQ bottom-up and acidic and basic HMWP depletion top-down LFQ analysis.

		BOTTOM-UP		
		DIFFERENTIALLY ABUNDANT	NOT DIFFERENTIALLY ABUNDANT	NOT QUANTIFIED
ACIDIC HMWP DEPLETION	DIFFERENTIALLY ABUNDANT	8 proteins 8 proteoforms	12 proteins 17 proteoforms	0 proteins 0 proteins
	NOT DIFFERENTIALLY ABUNDANT	15 proteins 23 proteoforms	53 proteins 110 proteoforms	12 proteins 15 proteins
	NOT QUANTIFIED	481 proteins N/A	1,103 proteins N/A	
BASIC HMWP DEPLETION	DIFFERENTIALLY ABUNDANT	8 proteins 10 proteoforms	17 proteins 27 proteoforms	3 proteins 3 proteoforms
	NOT DIFFERENTIALLY ABUNDANT	13 proteins 17 proteoforms	34 proteins 51 proteoforms	15 proteins 27 proteoforms
	NOT QUANTIFIED	480 proteins N/A	1,122 proteins N/A	

## 2.7 Synthetic Peptides based on 30S ribosomal protein S16

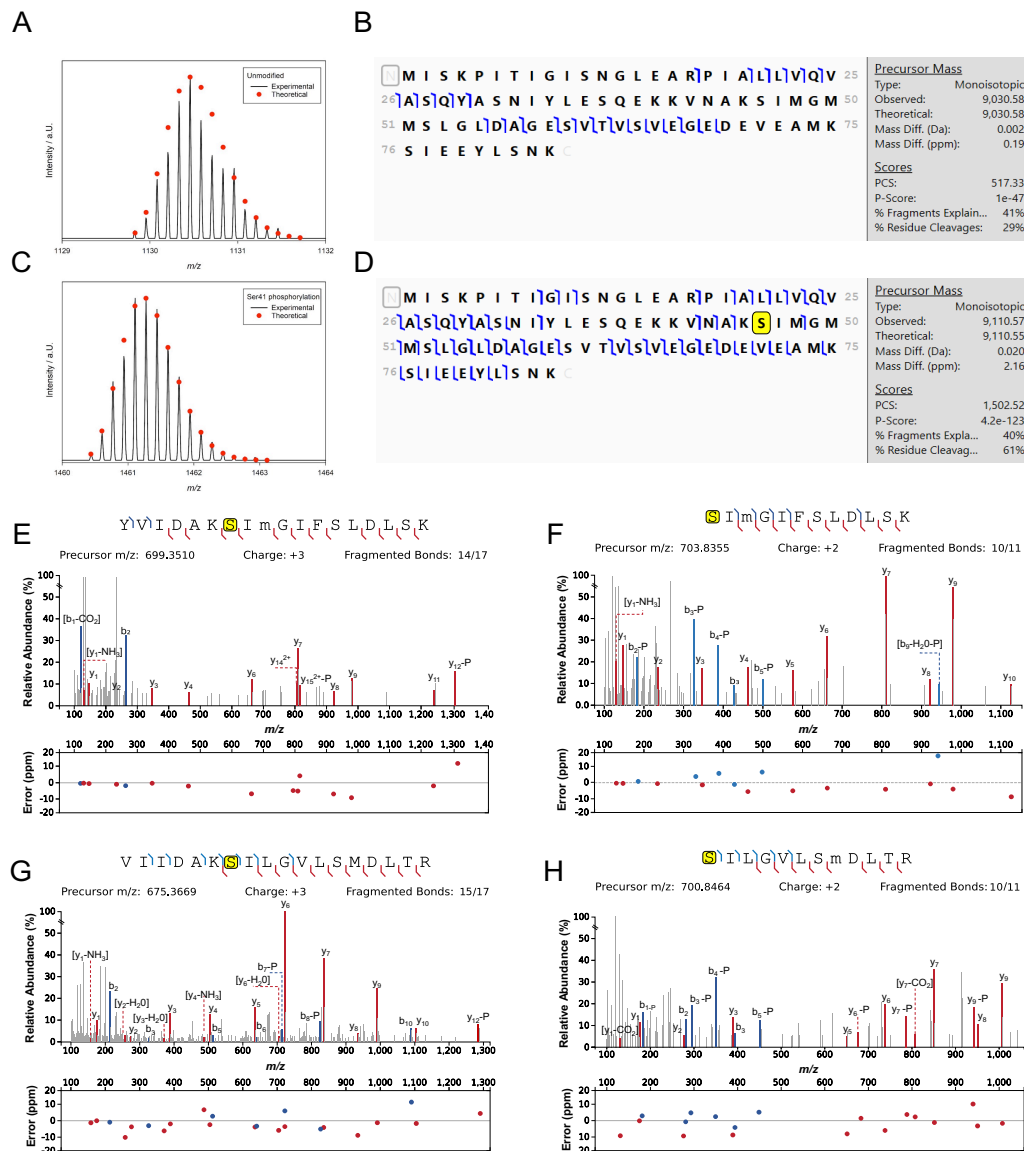
**TABLE A-9 | Overview of Synthetic Peptides Potentially Acting as Antimicrobial Peptides.**

PEPTIDE	SEQUENCE	MW (DA)
1	NH <sub>2</sub> .FIEEIGTYDPNQDP SVYKVDEE.OH	2587.72
2	NH <sub>2</sub> .DSRSPRDGKFIEEIGTYDPNQD.OH	2539.63
3	NH <sub>2</sub> .IIVADSRSPRDGKFIEEIGTYD.OH	2481.72
4	NH <sub>2</sub> .PNQDP SVYKVDEEAAKKWLANG.OH	2459.68
5	NH <sub>2</sub> .P SVYKVDEEAAKKWLANGAQPT.OH	2402.67



**FIGURE A-18 | Structure Prediction of 30S Ribosomal Protein S16.** Possible fragments after Asp-Pro cleavage of the 30S ribosomal protein S16 (A0A4P6M2Y5) from *B. producta*. **(A)** Full canonical form, **(B)** N-terminal fragment after cleavage of the first Asp-Pro bond, **(C)** N-terminal fragment after cleavage of the second Asp-Pro bond, **(D)** C-terminal fragment after cleavage of the first Asp-Pro bond, and **(E)** C-terminal fragment after cleavage of the second Asp-Pro bond. Structures were predicted using AlphaFold and visualized using PyMol, with the per-residue confidence score (pLDDT) indicated by color. Plots showing the predicted LDDT for each residue and heatmaps showing the predicted positional error (in Ångströms) for each position.

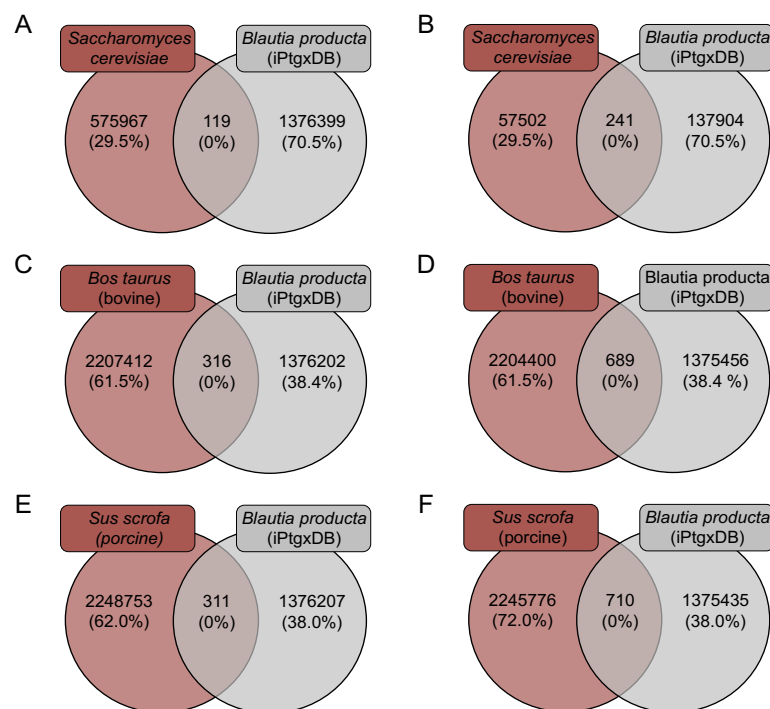
## 2.8 Phosphoserine Modification of HPr Proteins



**FIGURE A-19 | HPr Phosphocarrier Phosphoserine Modifications.** (A-D) Identified Proteoforms of A0A4V0Z7D2, with the distribution of experimentally observed and theoretical peaks for (A) unmodified and (B) phosphorylated HPr proteoforms and proteoform sequence and fragment ion spectrum of (C) unmodified and (D) phosphorylated HPr proteoforms. (E-F) Identified phosphopeptide sequence and fragment ion spectrum of A0A2S4GRU0 and (G-H) A0A7G5MYS7 phosphopeptides. The identified b- and y-ions are highlighted in blue and red, respectively. Dot plots illustrate the mass error in ppm for each observed b and y ion, respectively.

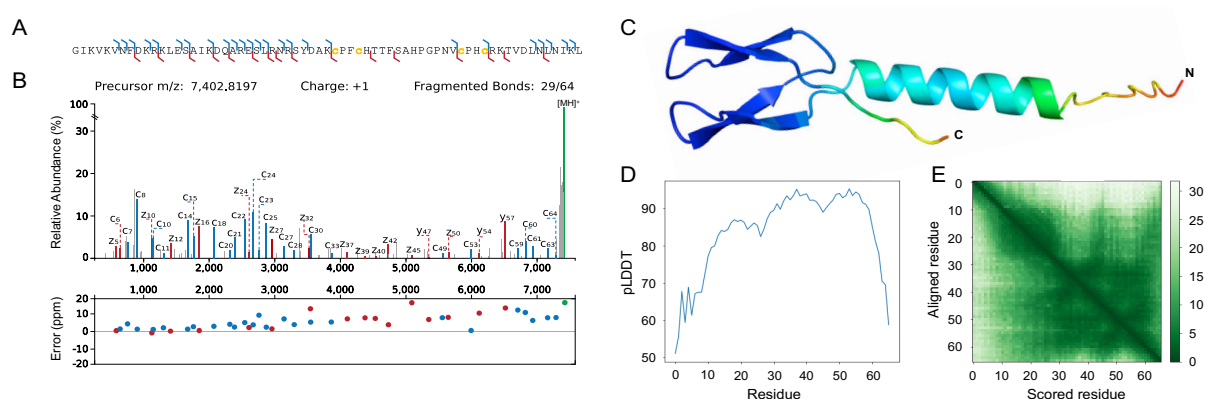
### 3 Proteogenomic Analysis of *B. producta*

#### 3.1 Potential Peptide Contamination Analysis



**FIGURE A-20 | Comparison of Peptides obtained by *in silico* Tryptic Digestion.** Identified *in silico* peptides from protein-containing media were compared with *in silico* peptides from *B. producta* iPtgxDB. **(A-B)** Peptide overlap for *Saccharomyces cerevisiae*, **(C-D)** *Sus scrofa* and **(E-F)** *Bos Taurus*. The venn diagrams in **(A, C, E)** are based on 20 amino acids, while for **(B, D, F)** leucine and isoleucine were treated as a single amino acid.

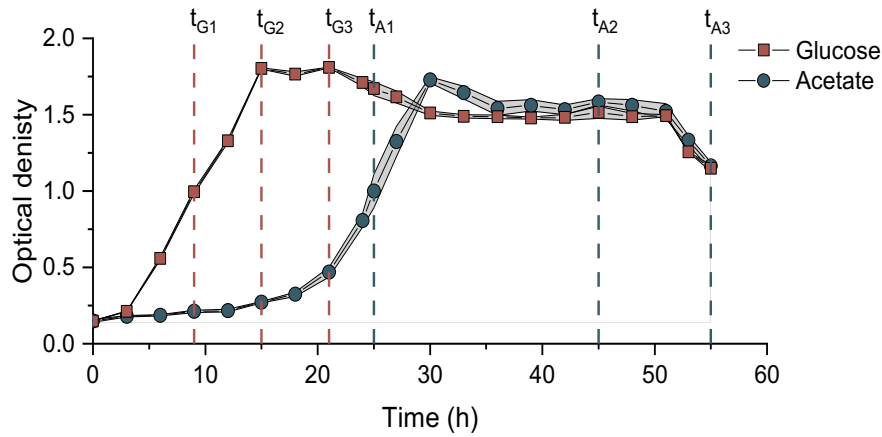
#### 3.2 Proteoform Characterization of BP12



**FIGURE A-21 | Proteoform Characterization of BP12.** **(A)** Identified N-terminal methionine excised proteoform. **(B)** MS<sup>2</sup> spectra indicates the presence of two disulfide bonds. The proteoform sequence is annotated with the identified c- ions and y-, z-ions in blue and red, respectively. The dot plot illustrates the mass error in ppm, confirming the correct identification for each observed ion. **(C)** AlphaFold structure prediction with color-coded per-residue confidence scores (pLDDT). **(D)** predicted LDDT and **(E)** predicted positional error (in Ångströms) for each position.

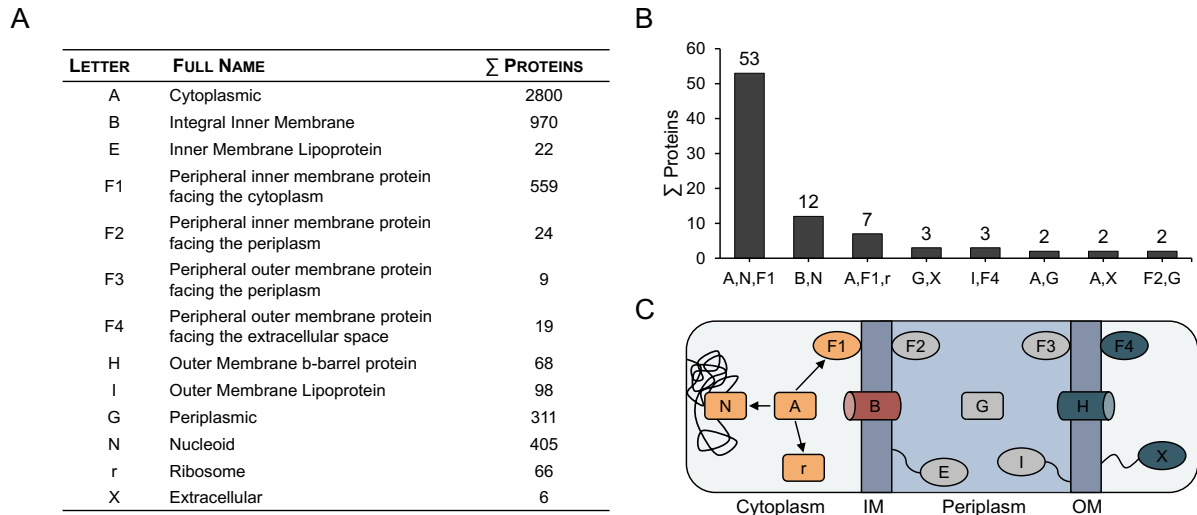
## 4 Outer Membrane Vesicles Analysis

### 4.1 *E. coli* Cultivation



**FIGURE A-22 | *E. coli* Growth Curves.** Times points indicate sample collection at mid-logarithmic, prestationary, stationary or death phases of *E. coli* cultures from M9 media containing either 15 mM glucose or 45 mM acetate. Data points are the mean of two biological replicates with minimum and maximum values represented in grey.

### 4.2 Subcellular Topology and Localization of *E. coli* Proteins



**FIGURE A-23 | Summary of the Characteristics of STEPdb. (A)** STEPdb's 13 subcellular topology classes. **(B)** Proteins with multiple subcellular locations. **(C)** Subcellular topological classifications categorized into four groups for cytoplasm (F1, A, R, and N), inner membrane (B), periplasm (I, G, F2, F3, and E), and outer membrane/extracellular group (H, X, and F4).

### 4.3 Potential *E. coli* OMV Protein Markers

**TABLE A-10 | List of Proteins Suggested as Potential OMV Markers.** Based on (Daleke-Schermerhorn et al., 2014; Hong et al., 2019).

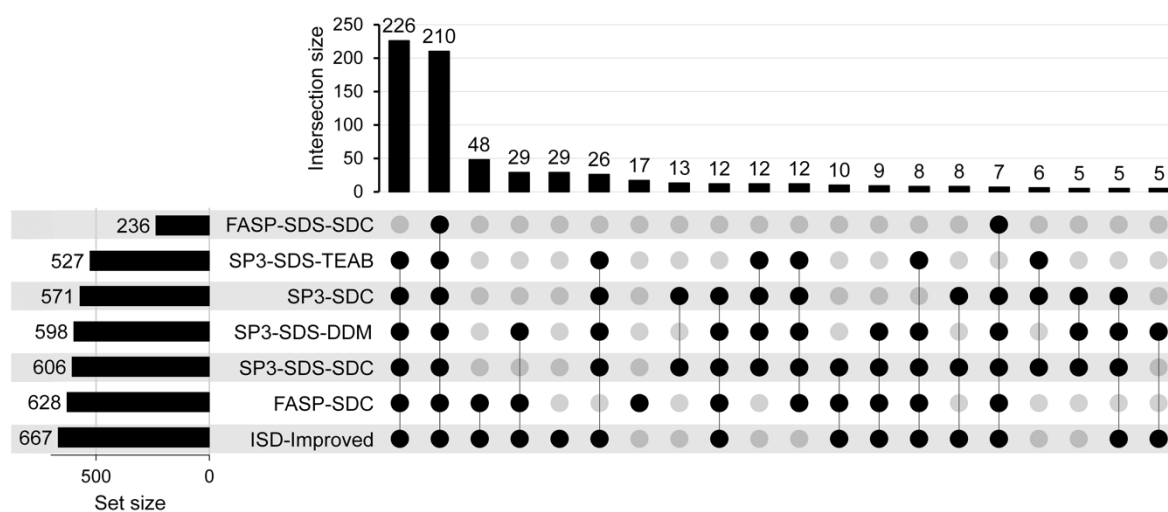
	GENE	UNIPROT	PROTEIN NAME
PORINS	OmpA	P0A910	Outer membrane protein A
	OmpC	P06996	Outer membrane porin C
	OmpF	P02931	Outer membrane porin F
	OmpT	P09169	Protease 7
	OmpW	P0A915	Outer membrane protein W
	OmpX	P0A917	Outer membrane protein X
	Tsx	P0A927	Nucleoside-specific channel-forming protein Tsx
LIPOPROTEINS	LolB	P61320	Outer-membrane lipoprotein LolB
	Lpp	P69776	Major outer membrane lipoprotein Lpp
	RcsF	P69411	Outer membrane lipoprotein RcsF
	RlpA	P10100	Endolytic peptidoglycan transglycosylase RlpA
	SlyB	P0A905	Outer membrane lipoprotein SlyB
	Slp	P37194	Outer membrane protein Slp
	LolB	P61320	Outer-membrane lipoprotein LolB
ASSEMBLY PROTEINS	BamA	P0A940	Outer membrane protein assembly factor BamA
	BamB	P77774	Outer membrane protein assembly factor BamB
	BamC	P0A903	Outer membrane protein assembly factor BamC
	BamD	P0AC02	Outer membrane protein assembly factor BamD
	BamE	P0A937	Outer membrane protein assembly factor BamE
	LpoA	P45464	Penicillin-binding protein activator LpoA

## 4.4 Nanoparticle Tracking Analysis

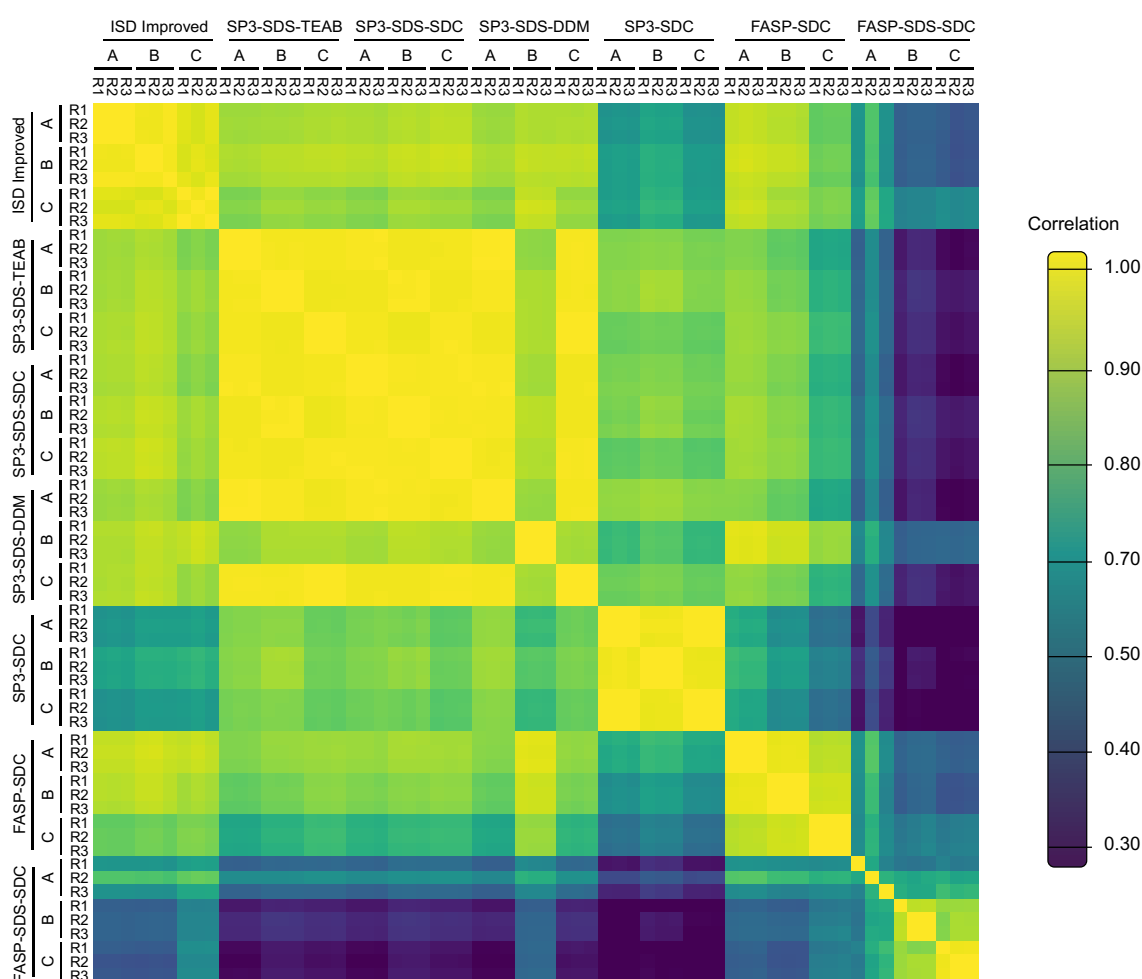


**FIGURE A-24 | Nanoparticle Tracking Analysis of Different Supernatant Volumes.** Analysis of **(A)** 20 ml, **(B)** 40 ml and **(C)** 60 ml supernatant from three biological replicas (A-C) processed using the OMV isolation kit. Shown are the particle size distribution, mode diameter and particle concentration of six 60 s videos of the three biological replicates and the particle size distribution after averaging the six replicates. Mode particle diameter and mean particle concentration ( $\pm$  standard error).

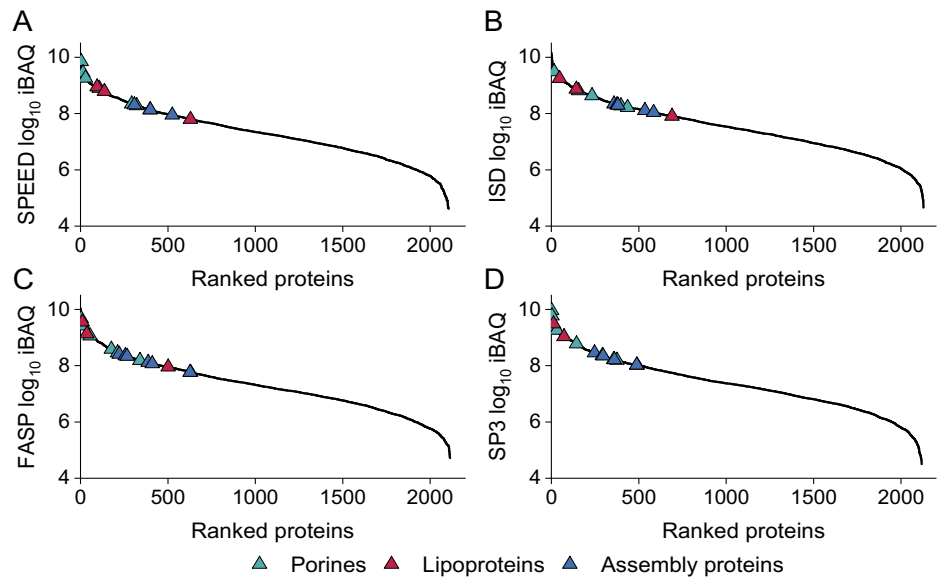
#### 4.5 Comparison of OMV Sample Preparation Methods



**FIGURE A-25 | UpSet Plot Analysis of OMV Sample Preparation Methods.** The set size represents the total number of identified proteins by each method, while the intersection size represents the number of shared proteins identified by different methods.



**FIGURE A-26 | Correlation Heatmap of OMV Sample Preparation Methods.** Pearson correlation coefficients based on normalized protein abundance between technical and biological replicates of each OMV-based sample preparation protocol (3 biological replicates A-C, 3 technical replicates R1-R3).



**FIGURE A-27 | Impact of Various Sample Preparations on the Relative Abundance Distribution of *E. coli* OMV Protein Markers.** Distribution of protein iBAQ values in (A) SPEED, (B) In-solution digestion, (C) filter-aided sample preparation and (D) single-pot, solid-phase-enhanced samples. Based on supplementary material from (Doellinger et al., 2020).

# Erklärung

Hiermit erkläre ich, Jerome Genth, dass ich die vorliegende Arbeit selbstständig und nur mithilfe der von mir angegebenen Quellen und Hilfsmittel unter Anleitung meiner akademischen Betreuer verfasst habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet. Die vorliegende Arbeit entstand zwischen April 2020 und Dezember 2023 in der Arbeitsgruppe von Prof. Dr. Andreas Tholey am Institut für Experimentelle Medizin unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft und wurde keiner anderen Universität als der Christian-Albrechts-Universität zu Kiel zur Begutachtung vorgelegt. Mir wurde zuvor kein akademischer Grad entzogen.

Jerome Genth

Kiel, den 15 August 2024