Enrichment and identification of small proteins in a simplified human gut microbiome

Hannes Petruschke[1], John Anders[2], Peter F. Stadler[2], Nico Jehmlich[1], Martin von Bergen[1,3]

[1]Helmholtz-Centre for Environmental Research – UFZ GmbH, Department of Molecular Systems Biology, Leipzig, Germany; hannes.petruschke@ufz.de, nico.jehmlich@ufz.de, martin.vonbergen@ufz.de
[2]Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics, Leipzig University, Germany
[3]Institute of Biochemistry, Faculty of Biosciences, Pharmacy and Psychology, University of Leipzig, Germany

Abstract

Small proteins (sProteins) with a size of 100 amino acids and less are involved in major biological processes and play an important role in different bacteria. Despite the increasing interest in them, the data on sProteins in bacterial communities, like the human gut microbiome is sparse. In this study, we are using the extended simplified human intestinal microbiota (SIHUMIx) as model system of the human gut microbiome to detect sProteins and to compare different sProtein enrichment methods. We observed that with our tested methods, the C8-cartridge enrichment resulted in the highest number of detected sProteins (n=295) with high reproducibility among replication analysis. However, in order to further increase the total number of sProteins, the combination of C8 cartridge enrichment with GelFree enrichment is favored because the latter complemented n=48 more sProteins compared to the C8 cartridge approach resulting in n=343 sProteins. Among all detected sProteins we were able to identify 79 so far uncharacterized sProteins, with no described protein evidence in the current released database. In total, 34 of those uncharacterized sProteins are localized in gene clusters conserved between different bacteria species allowing functional predictions. This study improves the assessment of sProtein detection and enables their functional characterization in future experiments.

Keywords

Small proteins (sProteins), human gut microbiome, proteomics

Introduction

The human gastrointestinal tract consists of thousands of different bacterial species (Almeida et al., 2019). It plays a pivotal role in protecting the host against pathogenic microbes, modulating immunity, regulating metabolic processes [1, 2] and has been associated with different diseases such as obesity, diabetes, inflammatory bowel disease and others [3]. It has also been shown that different

environmental influences such as nutrition and pharmaceuticals affect the shape of the gut community and thereby affect human health [4, 5]. However, due to the complexity of the gut microbiome it remains challenging to study host-microbiome interactions.

A suitable approach to address this challenge is the usage of bacterial model systems with a reduced number of species [6, 7]. Here, the extended simplified human intestinal microbiota (SIHUMIx) consisting of eight bacterial species was chosen. The reduced complexity makes this *in vitro* model system more suitable for -omics approaches such as metaproteomics. The bacterial consortium is continuously cultured in bioreactors and allows the analysis of the metabolic output and interdependence and interaction within SIHUMIx after a response to environmental stimuli under controlled conditions (e.g. diet, food additives, toxic compounds, etc.) (Schäpe et al., accepted).

In their natural habitats, bacteria usually live in complex multispecies communities. Since their complete proteome exhibits a wide range of functions that are essential for the physiology of the cells, we hypothesized that some proteins might be relevant for species interaction in a given community. These communities can comprise multicellular aggregates to billions or trillions of cells within the gastrointestinal tract [8]. It is well-known that bacteria`s behavior is different in communities compared to single strain cultivation [9]. Bacterial competition among available resources and space influences the structures and the functions of bacterial communities which can lead to a completely different set of gene translation that is not expressed in single strain cultivation [10].

Small proteins (sProteins) consist of up to 100 amino acids and are encoded in small open reading frames (sORFs). In prokaryotes, their importance has been demonstrated in several studies. The sProtein CydX (37 amino acids) for example regulates the activity of cytochrome oxidase in the electron transport chain during aerobic respiration and thereby influences ATP generation in *Escherichia coli* [11]. ArcZ (47 amino acids) binds to the AcrA-AcrB-TolC efflux pump and thereby affects susceptibility to certain antibiotics [12]. It was also shown that sProteins have an effect on the glucose uptake since SgrT (43 amino acids) inhibits the activity of the EIICBGlc glucose transporter [13]. Importantly, sProteins may also be used for e.g. communication between bacteria and bacteria and phages within niches such as the microbiota. A large metagenome study of the human microbiome revealed that it harbors more than 4,000 novel sProteins whereas more than 30% are predicted to be involved in cell-cell communication, highlighting the importance for the investigation of sProteins [14]. Notably, most known sProteins were discovered in single-strain approaches and it is well-known that bacteria`s behavior is different in communities compared to single strain cultivation [8].

Despite their relevance sProteins have not been fully explored for a long time. One reason is that in the past most genome annotation algorithms applied a 100 codon cutoff for annotation [15, 16]. This threshold was long time used to reduce the error rate of gene annotations although genome sequencing technologies enormously improved recently [17]. New methods such as ribosome profiling, in addition, allow the reanalysis of genomes and thus identify new ORFs containing sProteins [18].

Besides the annotation also the detection of sProteins using bottom-up proteomics is challenging since this technique relies on proteolytically cleaved proteins and subsequent analysis by LC-MS/MS [19]. Although mass spectrometers improved in sensitivity and scan speed, allowing fast in-depth proteome analysis, classic bottom-up proteomics protocols often result in low number of sProtein identifications [20-22]. This is caused by the fact that sProteins result in a low number of proteotypic peptides compared to large proteins and therefore to less identifiable peptides for LC-MS/MS analysis [23]. In fact, many sProteins are only identified by a single proteotypic peptide and not all possible peptides can be measured or identified through a combination of issues like peptide ionization characteristics, coelution with other peptides, peptide modifications and the protein inference problem [24-26]. An aspect that complicates the detection of sProteins further is that it may require specific proteomic methods or enrichment treatments, since sProteins can be lost in standard proteomic protocols, e.g. gel processing after SDS-PAGE if no strong fixation agent is used [27]. Hence, in order to increase sProtein identification, different enrichment strategies have been described, which can be divided into three different approaches related to the physicochemical properties of sProteins: i) based on protein size e.g. GelFree enrichment [28], Tricine SDS-Gel [29], Nanotrap particle enrichment [30], size exclusion chromatography [31] and molecular weight cutoff filtration (MWCO) [25]; ii) based on solubility, e.g. acetonitrile based precipitation and acidic precipitation [32]; or iii) hydrophobic interactions e.g. C8 cartridge enrichment [25]. Most of them focus on the enrichment of sProteins or the depletion of large proteins, before proteolytic cleavage and LC-MS/MS analysis. However, despite the achieved advancements none of these methods evolved as a gold standard yet. A comparison between different enrichment approaches is essential to enrich sProteins in order to find the best method for a given sample and matrix.

The aim of this study was to identify the most suitable method for the reliable detection of sProteins from a microbial community. We therefore compared different proteomic protocols and enrichment methods for sProtein identification in SIHUMIx. sProteins which may be important for the resilience and resistance of microbial communities after a response to environmental stimuli become present which

are probably hidden by applying only single strain experiments. The identified sProteins will be analyzed with respect to their genetic localization the conservation across bacterial species in order to obtain the first clue on their potential function.

Material and methods

Cultivation of SIHUMIx

The bacterial species *Anaerostipes caccae* (DSMZ 14662), *Bacteroides thetaiotaomicron* (DSMZ 2079), *Bifidobacterium longum* (NCC 2705), *Blautia producta* (DSMZ 2950), *Clostridium butyricum* (DSMZ 10702), *Clostridium ramosum* (DSMZ 1402), *Escherichia coli K-12* (MG1655) and *Lactobacillus plantarum* (DSMZ 20174) were cultivated as single strains in Brain-Heart-Infusion (BHI) medium under anaerobic conditions at 37°C and 175 rpm shaking for 72 h before inoculation in the bioreactor. A total of $8 \cdot x10^9$ bacteria cells ($1 \cdot x10^9$ cells/strain) were used for inoculation and the SIHUMIx community was continuously cultivated in modified complex intestinal medium in 250 mL culture vessels at 37°C, stirring at 150 rpm and constant pH of 6.5 as described (Schäpe et al., accepted). Samples were taken after 16 days of continuous cultivation. In total, 2 mL bacteria cell suspension were centrifuged (3,200 x g; 10 min; 4°C) and immediately frozen at -80°C for subsequent sample analysis.

Cell lysis

Bacteria cell pellets (approx. 7×10^9 cells) were lysed in 1 mL SDS-extraction buffer (20 mM Tris/HCl pH 7.5; 2% SDS) with 0.1% phenylmethylsulfonyl fluoride (PMSF). The lysed protein extracts were used for following proteomic protocols: In-Gel proteolytic cleavage, Tricine Gel fractionation and Nanotrap particles. For GelFree enrichment bacteria cell pellet was lysed in 1 mL UPX Lysis buffer (Expedeon, USA) with 0.1% PMSF as per the manufactures suggested. For reversed acetone precipitation bacteria cell pellet was lysed in 0.1% TFA with 0.1% PMSF. Afterward, bacterial cells were disrupted with a sonic probe (cycle 0.5; amplitude 60%, Branson Sonifier 250, Emerson, St. Louis, MO, USA), while the samples were kept on ice followed by heating on a shaker (37°C; 10 min; 1,400 rpm). Undissolved material was removed by centrifugation (10,000 × g; 10 min; 4°C). The supernatants containing the extracted protein were stored at -20°C.

For C8-Cartridge enrichment, In-Solution proteolytic cleavage, SP3, FASP, MWCO filtration the bacteria cell pellets were resuspended in 660 µL methanol. After adding 330 µL MilliQ water and 330 µL chloroform the samples were vortexed and incubated on ice for 10 min before being sonicated (see

above). To the lysate 330 µL MilliQ water and 330 µL chloroform were added and the samples were vortexed and centrifuged (1,700 x g; 10 min; 4°C). The interphase containing proteins was obtained and the remaining solvent evaporated in a vacuum concentrator. The pellet was resuspended in urea buffer: UT solution (8 M urea; 2 M thiourea) for In-Solution proteolytic cleavage, SP3, and C8-Cartridge enrichment. UA buffer (8 M urea on 0.1 M Tris/HCl; pH 8.5) for FASP and MWCO filtration.

Protein concentration was determined with bicinchoninic acid assay according to the user manual (Pierce™ BCA Protein Assay Kit, Thermo Fischer, USA) for samples in SDS or 0.1% TFA. For samples in UA or UT protein concentration was measured with Pierce™ 660 nm Protein Assay Reagent (Thermo Fischer, USA) as per the manufactures suggested. Protein concentration in UPX-Lysis buffer was determined with BradfordUltra (Expedeon, USA) according to kit instructions.

Small protein enrichment and proteolytic cleavage

In-Solution proteolytic cleavage

4 µg of protein in UT-solution was filled up to 20 µl with 20 mM Ammonium bicarbonate. For disulfide reduction 2 µl of 25 mM dithiothreitol- (DTT-) solution was added and the samples were heated on a thermoshaker (60°C; 1 h; 1,400 rpm). This was followed by alkylation by adding 14 µL of 20 mM ammonium bicarbonate and 4 µL of 100 mM 2-iodoacetamide and incubation at 37°C for 30 min at 1,400 rpm and overnight enzymatic digestion with trypsin (1:50) at 37°C.

In-Gel proteolytic cleavage

The protein lysate was precipitated overnight with acetone 1:5 (v/v) at -20°C and centrifuged at 14,000 x g for 10 min at 4°C. The protein pellet (100 µg) was resuspended in 25 µL loading buffer (4% SDS, 20% glycerol, 10% 2-mercaptoethanol, 0.004% bromphenol blue and 0.125 M Tris/HCl, pH 6.8) and incubated at 90°C for 5 min. The sample was loaded on sodium dodecyl sulfate polyacrylamide gel electrophorese (SDS–PAGE; 4% stacking gel and 12% separating gel). Electrophoresis was performed at 10 mA per gel. Proteins were stained with colloidal Coomassie Brilliant Blue G-250 (Roth, Kassel, Germany) and destained with MilliQ water. Protein bands were cut into gel pieces, washed twice with water for 10 min and once with NH_4HCO_3 (10 mM). InGel proteolytic cleavage was performed by adding modified porcine trypsin (150 ng, Sigma–Aldrich) and incubated overnight at 37°C. The reaction was stopped by adding formic acid (final concentration: 4%). Supernatant and gel elution solutions (first elution step: 40% (v/v) acetonitrile; second elution step: 80% (v/v)) were collected and the combined

mixtures were dried using vacuum centrifugation. Peptides were reconstituted in 0.1% formic acid. Extracted peptide lysates were dried using SpeedVac and resolved in 0.1% formic acid [33].

Tricine-Gel fractionation

50 µg protein were mixed with Tricine sample buffer (Biorad, USA) with 2% 2-mercaptoethanol 1:2 (v/v) and heated on a thermoshaker (70°C; 10 min; 1400 rpm). Samples were loaded on 12 % Tricine-SDS-Gel as described in (Haider, Reid, & Sharp, 2012). A prestained low molecular weight protein standard (5 µL of Precision Plus Protein™ Dual Xtra Prestained Protein Standards, BioRad, UK) was loaded as marker and electrophoresis were performed at 10 mA per gel. Proteins were fixed for 25 min in 5% (v/v) glutaraldehyde. Three gel fractions between 1-15 kDa were cut out, sliced into gel pieces and proteolytic cleavage was performed as described (see *In-Gel proteolytic cleavage*).

SP3

In total, 10 µg of protein lysate in UT-solution was used and 20 µL of 20 mM Ammonium bicarbonate was added. For disulfide reduction 2 µL of 25 mM Dithiothreitol- (DTT-) solution was added and the samples were heated on a thermoshaker (60°C; 1 h; 1400 rpm). This step was followed by adding 14 µL of 20mM ammonium bicarbonate and 4 µL of 100 mM 2-iodoacetamide and incubation at 37°C for 30 min at 1,400 rpm. SpeedBeads™ magnetic carboxylate modified particles (Sigma-Aldrich) were used for Single-pot, solid-phase-enhanced sample preparation (SP3) as previously described in [34]. Evaporated peptide samples were resolved in 20 µL 0.1% formic acid.

FASP/MWCO filtration

Vivacon 500 (Sartorius, Germany) with 2, 10 and 30 kDa MWCO membranes were equilibrated with 100 µL UA buffer and centrifuged (14,000 x g; 20 min; 20°C). All centrifugation steps were performed applying the same conditions. 100 µg protein in UA buffer was added and centrifuged. The flow-through was collected (2, 10 and 30 kDa MWCO) and taken for in-solution digest (see above). Proteins retained on the filter were incubated with 200 µL 10 mM DTT in UA buffer using a thermoshaker (37°C; 30 min; 1,400 rpm), centrifuged and further alkylated and proteolytically cleaved as previously described in [35].

C8 cartridges

Bond Elute C8 cartridges (100 mg, Agilent, USA) were equilibrated with 1 mL methanol followed by 2 mL 0.1% trifluoroacetic acid (TFA) before loading of 1 mL protein samples in UT-buffer. Afterwards, two washing steps with 1 mL 0.1% TFA were performed before eluting in two steps with 0.1%

TFA:acetonitrile (3:1 V/V) and (1:1 V/V). After evaporating eluted proteins using SpeedVac they were proteolytically cleaved as described above (*In-Solution proteolytic cleavage)*.

Reversed acetone precipitation

In total, 1 mL protein lysate was used and 5 mL acetone was added and precipitated at -20°C overnight. The samples were centrifuged (7,000 x g; 15 min; 4°C) and the supernatant was collected, protein lysates were dried using a SpeedVac, and the proteolytic cleavage of proteins was done as described above (*In-Solution proteolytic cleavage)*.

GelFree

GelFree® 8100 12% Cartridge Kit was prepared as described in the manufacturer's instruction. 500 µg protein was mixed with 6× GelFree loading buffer and 10 mM DTT and heated for 10 min at 50°C before being loaded to the chamber into the GeLFree 8100 fractionator (Expedeon, USA) according to manufacturer's instructions. The samples were separated into 12 fractions and collected as suggested by the manufacturer. The first five fractions containing small proteins were further purified to exclude SDS as described [36] and evaporated using SpeedVac. Protein pellets were proteolytic cleaved following the *In-Solution* protocol (as described above).

Nanotrap particles

Nanotrap particles (Ceres Nanosciences, USA) [containing white, blue and purple particles] were prepared as described in the manufacturer's instruction before being incubated with 100 µg protein sample for 30 min at 20°C. All washing and elution steps were performed as described in the kit manual. After evaporating eluted proteins using SpeedVac they were proteolytic cleaved following the *In-Solution* protocol (as described above).

Peptide desalting

Extracted peptides were purified by SOLAµ (Thermo Scientific, USA) as per the manufactures suggested. After evaporation peptides were resuspended in 20 µL 0.1% formic acid.

Mass spectrometric analysis

In total, 5 µL of peptide lysates were injected into nanoHPLC (UltiMate 3000 RSLCnano, Dionex, Thermo Fisher Scientific) followed by separation on a C18-reverse phase trapping column (C18 PepMap100, 300 µm x 5 mm, particle size 3 µm, Thermo Fischer Scientific), followed by separation on a C18-reverse phase analytical column (Acclaim PepMap® 100, 75 µm x 25 cm, particle size 3 µm, nanoViper, Thermo Fischer Scientific). Mass spectrometric analysis of eluted peptides was performed on a Q Exactive HF

mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) coupled with a TriVersa NanoMate (Advion, Ltd., Harlow, UK) source in LC chip coupling mode. LC Gradient, ionization mode and mass spectrometry mode was performed as described [37].

Raw data repository

Database construction and small protein identification

MS data processing was performed using Proteome Discoverer (v.2.2, Thermo Fischer Scientific, Waltham, MA, USA). Search settings for Sequest HT search engine were set to trypsin (Full), max. missed cleavage sites: 2, precursor mass tolerance: 10 ppm, fragment mass tolerance: 0.05 Da. Carbamidomethylation of cysteines was specified as a fixed modification and the oxidation of methionine and N-Acetylation of the protein N-terminus as a variable modification. As database the protein coded sequences of the eight SIHUMIx strains were downloaded from UniProt (http://www.uniprot.org/) and combined in one protein-coding sequence database (*.fasta). False discovery rates (FDR) were determined using Percolator [38]. Proteins were considered as identified when at least one unique peptide passed the FDR of 0.01 and a Sequest HT Score of ≥2. Proteins ≤ 100 amino acids were defined as sProteins. Sequence similarity search was performed using *Diamond* v.0.9.21.0 with default settings; NCBI NR (2019-04-03) as reference database, sensitivity mode: sensitive. Figures were created using *Prism* v8.1.2 (GraphPad Software, La Jolla, CA, USA) and R. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016298 [39].

Protein statistics

The isoelectric point (p*I*) was calculated using Proteome Discoverer (v.2.2, Thermo Fischer Scientific, Waltham, MA, USA). *The* grand average of hydropathy score (GRAVY score) was calculated using the web tool http://www.gravy-calculator.de. The Frequency distribution of p*I* and GRAVY score was calculated and illustrated using R.

Transmembrane helices prediction in sProteins was performed using TMHMM v.2 (http://www.cbs.dtu.dk/services/TMHMM/). The signal peptide analysis was performed using SignalP-5.0 (http://www.cbs.dtu.dk/services/SignalP/).

The normalized spectral abundance factor (NSAF) was calculated as previously described [40].

KEGG SSDB

To predict potential functions of the uncharacterized sProteins, we used the KEGG SSDB (Sequence Similarity DataBase), containing information about amino acid sequence similarity's [41]. The KEGG SSDB offers a search for conserved gene clusters for a query gene, based on conserved gene synteny. We used the KEGG definition and the information of the KEGG BRITE (functional hierarchies) database of those genes in the cluster to get information about the function of the gene cluster. For mapping between different database identifier, we used the API of bioDBnet [42].

Results

Comparison of different proteomic methods and enrichment methods

First, we selected 14 different standard proteomic protocols and enrichment methods to define the most suitable approach for total protein and sProtein identification (for complete method collection please refer to Supplement Information 1). As a criterion for the most useful method, we selected the number of sProtein identifications and required experimental time. As a main result, SP3, FASP, In-Solution, and In-Gel proteolytic cleavage performed best according to the total number of protein identifications ranging from 2,673 (In-Gel) to 2,964 (SP3) identified proteins (**Supplement Figure 1**). More specific, using the SP3 and the In-Solution cleavage protocol 121 sProteins and 189 sProteins were identified, respectively. Two enrichment protocols, namely the C8 cartridge- and the GelFree enrichment approach identified 329 and 323 sProteins respectively, which is about 2-fold higher compared to the tested standard proteomic protocols. Notably, Reversed acetone precipitation, a method used to precipitate large proteins and to keep sProtein in the supernatant resulted in only six identified sProteins. A low number of identified proteins and sProteins was also observed for In-Gel fractionation on a Tricine-SDS-Gel, molecular weight cut off filtration (MWCO) on 2, 10 and 30 kDa cut off filters and Nanotrap particle enrichment. Based on these first findings, we selected five promising methods with a high number of sProtein identifications and short experimental time for further replication and detailed comparison: C8 cartridge enrichment, GelFree enrichment, FASP on 10 kDa cut off filters, In-Solution cleavage and SP3 (**Supplement, Figure 1**).

Performing the main experiment (**Figure 1**), we were able to confirm the results of the pre-experiment. First, we compared the total number of identified proteins. For the SP3 protocol, we observed the highest number with 3,200 ±159 (mean SD) proteins followed by FASP (3,034 ±17) and In-Solution cleavage (2,888 ±261). Compared to GelFree enrichment (1,717 ±395), the other four methods including

the C8 cartridge enrichment (2,630 ±94) showed a significantly higher number of total proteins (**Figure 2A**). But more importantly, when comparing the number of sProteins, the C8 cartridge– (339 ±4) and the GelFree enrichment (325 ±32) yielded significant higher numbers compared to the standard protocols (**Figure 2B**). Two methods, namely In-Solution cleavage (189 ±1 sProteins), and FASP (202 ±5) performed similarly in case of total protein and sProtein identifications. In contrast to the total protein identifications, the SP3 protocol (114 ±8) showed significant fewer identified sProteins. In order to determine the sProtein enrichment efficiency we analyzed the relative amount of identified sProteins (**Figure 2C**). Interestingly GelFree enrichment resulted in the highest relative number of sProteins (19%), followed by C8 cartridge enrichment (13%). In contrast, the standard proteomic protocols FASP, In-Solution (7%) and SP3 (4%) showed a much lower relative amount of sProteins. Additionally, the technical reproducibility of the appropriate protocol for sProtein identifications was analyzed. Therefore, we compared the process replicates of the preparation and observed that two enrichment methods showed high reproducibility with 87% of sProteins for C8 cartridge and 83% for GelFree enrichment (**Figure 2B**). In general, the protocols of C8 cartridges, GelFree and FASP showed higher reproducibility of sProteins identifications, whereas for In-Solution and SP3 more total proteins than sProteins were recovered. When comparing the identified sProteins detected in triplicates, we observed that C8 cartridge- and GelFree enrichment strategy showed an overlap of 222 sProteins while 56 sProteins remained unique to C8 cartridge- and 41 to GelFree enrichment (**Figure 3**). In contrast, the standard proteomic protocols only increased the number of identified sProteins by 9. Four unique sProteins were identified with FASP, and only two with In-Solution proteolytic cleavage and SP3.

Characteristics of identified SIHUMIx sProteins

For sProteins different properties and localization in the cell have been reported (Storz, Wolf, & Ramamurthi, 2014). One relevant biochemical aspect of proteins is the isoelectric point (p*I*) since it determines the solubility dependent on the pH of the matrix it occurs in. We observed that the identified proteins larger than 100 AA showed a highly skewed bimodal distribution with an increased relative number of acidic proteins. The minimum around the physicological pH within bacterial cells is expected since non-charged proteins are less soluble. In contrast, the bimodal p*I* distribution of sProteins shows two equal peaks and a higher relative number of basic proteins (**Figure 4A**). We further analyzed the distribution of the relative protein number of identified sProteins and total proteins between different localizations in the cell according to gene ontology annotation (**Figure 4B**). We observed an increase of ribosomal proteins (27% compared to 8%), whereas the relative number of proteins localized in membranes, cell wall, flagellum and extracellular was reduced. The relative number

of proteins localized in the cytosol was equal and many proteins were not assigned in both groups. Interestingly, analyzing the p*I* of sProteins in between the different locations in the cell we observed a high p*I* for ribosomal and membrane sProteins, while most cytosolic sProteins showed an acidic p*I* (**Figure 4C**). The grand average of hydropathy (GRAVY) is a score describing the potential hydrophobicity of a protein. A positive GRAVY score corresponds to hydrophobic proteins, whereas a negative GRAVY score corresponds to hydrophilic proteins. The comparison of the GRAVY score frequency distribution between the relative number of sProteins and proteins larger than 100 AA showed a wider curve for sProteins but no significant difference between both groups (**Figure 4D**).

Identification of hypothetical and uncharacterized sProteins

sProteins are getting more attention and several sProteins can be found in public protein databases, however, most of them are uncharacterized and have no protein evidence so far. Overall, 134 so far uncharacterized sProteins were identified and 105 of them were identified in at least three samples independent of the proteomic method. To further validate that these sProteins have no annotated function, their amino acid sequences were subjected to a sequence similarity search against the NCBI non-redundant protein database using *Diamond*. Still, 79 sProteins with no however predicted function were identified (**Supplement Table 1**; representative examples in **Table 1**). In order to provide unambiguous protein identification, we applied the following stringent criteria: Sequest HT score >2, False Discovery rate <1%, one unique identified peptide, and identification in triplicates. High-quality MS/MS spectra were observed for many sProtein peptides (examples shown in **Figure 5**). Importantly this also holds true for sProteins identified by a single peptide as for example *P56614*, a 57-amino acid long uncharacterized sProtein in *E.coli*, which was identified with high-quality spectra for an 11-amino acid long peptide shown in **Figure 5**. To further analyze the characteristics of the uncharacterized sProteins we compared the p*I* and GRAVY score between characterized and uncharacterized sProteins (**Figure 6**). Interestingly, the uncharacterized sProteins showed a skewed bimodal distribution shifting to a higher number of acidic sProteins, compared to the characterized with equal relative number of basic proteins and acidic sProteins. In contrast to the differences in the p*I* the distribution of the GRAVY score did not show significant changes between characterized and uncharacterized sProteins. Analysis of the secondary structure of detected uncharacterized sProteins using *TMHMM* showed that 9 sProteins have predicted transmembrane helices and 2 sProteins contain a Lipoprotein signal peptide (Sec/SPII) predicted by *SignalP-5.0* (**Supplement Table 2**).

Although many sProteins are not associated with any function in bacteria, the localization in a conserved gene cluster can give information about its potential function. We therefore used KEGG SSDB, containing gene clusters conserved by gene synteny between different bacteria species and analyzed the localization of the detected uncharacterized sProteins. Interestingly, 34 of the 79 uncharacterized sProteins were identified to be in a gene cluster according to KEGG SS data base. A list of the gene clusters can be found in **Supplement Table 3 and 4**. The *Bacteroides thetaiotaomicron* sProtein Q8A9G1 for example, lies in a gene cluster that contains 12 proteins, including members of a two-component system, as well as ABC transporter proteins, a permease and the outer membrane channel protein TolC. Another example is Q8A8B1 which was found in *B. thetaiotaomicron* in a conserved gene cluster containing 13 protein-coding sequences. Among those proteins are a multidrug resistance protein, an outer membrane efflux protein, a Multiple Antibiotic Resistance Regulator (MarR) transcriptional regulator, choloylglycine hydrolase and an acetyltransferase.

Distribution of sProtein identifications within the SIHUMIx culture

SIHUMIx community consists of eight different bacteria species with different genome sizes. The abundance of all SIHUMIx members in the community is not equal during cultivation (as previously described by Schäpe et al., accepted). Therefore the relative number of detected proteins and the protein abundance between all SIHUMIx species were compared (**Figure 7**). The relative protein number of detected sProteins was 9% compared to detected proteins > 100 amino acids (**Figure 7A**). Interestingly, when comparing the relative abundance using the normalized spectral abundance factor (NSAF) we observed a much stronger presence of sProteins with 27.7% (**Figure 7C).** When comparing the relative number of detected proteins between species, most of the total proteins were identified for *B. thetaiotaomicron* (42%) followed by *Escherichia coli* (20%), *Blautia producta* (15%), *Clostridium ramosum* (13%) and *Anaerostipes caccae* (9%). *Clostridium butyricum*, *Bifidobacterium longum* and *Lactobacillus plantarum* only show a relative abundance below 1% (**Figure 7B**). sProteins of *Escherichia coli* (28%) and *B. thetaiotaomicron* (23%) were less represented than compared to the total protein number. In contrast the comparison of the relative protein number of uncharacterized sProteins showed the highest number of identified sProteins for *B. thetaiotaomicron* (38%) followed by *C. ramosum* (24%) and *Escherichia coli* (18%). The comparison of species abundance of SIHUMIx using the NSAF resulted in higher relative abundance of *B. thetaiotaomicron* in total proteins (55 %), sProteins (42 %) and uncharacterized sProteins (68 %) compared to the relative protein number (**Figure 7D**).

Discussion

Comparison of different sProtein detection methods

The detection of sProteins with bottom-up proteomics is still quite challenging. In the past, different methods were described to improve sProtein detection [21, 28]. This study compares different proteomic protocols and enrichment strategies to detect sProteins in a model system of the human gut microbiome (SIHUMIx). In a pre-experiment, we compared 14 different standard proteomic and sProtein enrichment methods. The standard proteomic methods namely FASP, In-Gel, SP3, and In-solution proteolytic cleavage are well established, need a short experimental time and resulted in a high number of total protein identifications. Nevertheless, the number of identified sProteins was low in all these methods. One reason for that is the loss of sProteins during experimental workflow e.g. no retaining at filter membranes during FASP or diffusion out of Gel pores in SDS-Gels [27]. Another reason is the fact, that with no enrichment of sProteins or exclusion of large proteins, the small amount of sProtein derived peptides are missed in the large background of total peptides and the low abundance of sProteins [32].

In order to improve the number of identified sProteins different enrichment methods based on the physicochemical properties of sProteins were compared. GelFree enrichment, a method based on protein size separation showed an increase of sProtein identification. This method showed good results in *Archea* as previously described in [28]. An advantage of this method is the fractionation by size on a gel matrix allowing deep proteome analysis without the risk of losing sProteins by diffusion as it might be the case in normal SDS-PAGE approaches. Though, it is very time consuming, needs an additional SDS-removal step before proteolytic cleavage of proteins and is not able to retain sProteins below 3 kDa based on the GelFree membrane size in the protein collection chamber. In contrast, other methods based on protein size separation did not improve sProtein identification. Also sProteins separation based on solubility by precipitating large proteins with acetone did not result in increased sProtein detection. This may be due to the fact that the concentration of acetone was too high resulting in total protein precipitation. Recently, Cassidy et al. used acetonitrile as precipitation agent for depleting large proteins and achieved good results [32]. The third approach of enriching sProteins was by hydrophobic interactions using the unpolar surface of a C8 cartridge. C8 cartridges were first established to obtain small peptide hormones from blood plasma [43, 44] and afterward for polypeptide enrichment in eukaryotes [25]. We modified the protocol and observed the highest amount of sProtein identifications. Additionally also a relatively high number of total proteins were identified. We hypothesize that the

unpolar C8 material retains large and hydrophobic proteins, while sProteins and polar proteins are already eluted with 25-50% organic solvent. This method together with GelFree enrichment showed high reproducibility for sProteins with less working steps and experimental time. Interestingly, when comparing identified sProteins between all methods a high number remained unique for GelFree or C8 cartridge enrichment, indicating that enrichment based on size retains different sProteins, compared to enrichment based on hydrophobicity. We therefore recommend applying both methods for increased coverage of sProtein identification.

p*I* shift and sProtein characteristics

For the so far investigated sProteins several cellular localizations have been reported. Often sProteins are bound to membranes or large protein complexes [45]. We therefore analyzed characteristics of the identified sProteins. In contrast to the identified large proteins, we observed a shift to a more basic p*I* for sProteins and in case of cellular localizations, more ribosomal proteins and fewer membrane proteins. Further analysis revealed that sProteins with basic p*I* are mostly ribosomal proteins and membrane proteins. We hypothesize that the relatively high amount of basic amino acids in those sProteins may be essential for binding the acidic DNA/RNA, as it has been shown in some ribosomal proteins [46, 47]. Analyzing the differences between characterized and uncharacterized sProteins, we observed a higher relative number of acidic and hydrophilic uncharacterized sProteins. Only 9 uncharacterized sProteins showed transmembrane helices, so we hypothesize that several potential new sProteins are hydrophilic, localized in the cytoplasm and have an acidic p*I*.

Potential functions of identified uncharacterized sProteins

To predict functions of hypothetical and uncharacterized proteins we analyzed their localization in conserved bacterial gene clusters using KEGG SSDB. We were able to detect 34 uncharacterized sProteins in gene clusters. One example is the *B. thetaiotaomicron* sProtein Q8A9G1. It may play a role in membrane transport to certain stress as its gene cluster contains 11 additional proteins which are part of a two component system, ABC transporter proteins, a permease and the outer membrane channel protein TolC. The two-component system has been described as a recognition system to bacterial stressors and TolC as well as ABC transporters can work as drug efflux pumps in bacteria [48-50]. Interestingly, ArcZ an *E.coli* sProtein was already described to bind to the AcrA-AcrB-TolC efflux pump affecting susceptibility to antibiotics [12]. Q8A9G1, therefore, might be an important sProtein for survival after uptake of toxic compounds which can be beneficial in competing in bacterial communities. Another example is Q8A8B1, a 93 amino acid long sProtein identified in *B. thetaiotaomicron*. Several

proteins in its gene cluster are associated with drug resistance and extracellular export. Interestingly choloylglycine hydrolase, catalyzing the hydrolysis of bile acids can also be found in the gene cluster [51]. Primary bile acids facilitate the uptake of fatty acids but also have an antibacterial capacity that might be used to regulate the microbiota in the intestinal tract [52, 53]. Q8A8B1 might, therefore, be associated with detoxification of primary bile acids and drug export and as a consequence influence host-host and host-microbe interactions.

The study of sProteins in SIHUMIx is well feasible because all 8 bacteria strains are known and have a sequenced and annotated genome. The total amount of different proteins is much lower compared to the natural microbiome with up to hundreds various species expressing different proteins. In natural microbiomes this may lead to a dilution of each protein species and hampers the identification of low abundant proteins. sProteins might be affected if they are expressed in low amounts and due to the fact that they contain a small number of amino acids resulting in fewer peptides after proteolytic cleavage [23]. Based on that, the sProtein enrichment strategies may lead to much higher sProtein identifications for microbiomes compared to standard proteomic protocols. Nevertheless, with increasing number of bacteria species, the sProtein identification per species will be reduced. Another point is the bioinformatics challenge. The high number of different protein species increases the possibility of shared peptides between different proteins and thereby reduces the number of truly identified proteins [54]. Furthermore natural microbiome samples contain species that are not fully sequenced and annotated which could lead to a lot unmatched spectra or false positive identifications [55].

Conclusion

Although sProteins are difficult to detect by bottom-up proteomics our proteomic method comparison showed that C8 cartridge enrichment performed well in case of enrichment and detection of sProteins. To further increase sProtein identification we recommend using a second enrichment strategy based on different physicochemical properties as for example GelFree (based on protein size). In case of SIHUMIx this enrichment method could complement additional 48 sProteins not detected with C8 cartridge enrichment. We were further able to detect 79 uncharacterized sProteins with no evidence on protein level so far. In addition, 34 of them were identified to be localized in a potential gene cluster, which gives information about their potential functions and localization.

Author Contributions

Conceptualization, HP, NJ and MvB; Data curation, HP; Formal Analysis, HP and JA; Funding Acquisition, MvB and PS; Methodology, HP; Supervision, NJ and MvB; Validation, HP; Visualization, HP and JA; Writing-original draft, HP; Writing-review&editing, all authors.

Acknowledgment

We thank Prof. Dr. Michael Blaut (German Institute of Human Nutrition, Potsdam-Rehbruecke) for providing the SIHUMIx bacteria. We are thankful for technical assistance from Stephanie S. Schäpe and Jannike Lea Krause for SIHUMIx cultivation.

Funding

This research project was funded by the DFG grant within the Priority Programme entitled “Small Proteins in Prokaryotes, an Unexplored World” (SPP 2002)

Figures

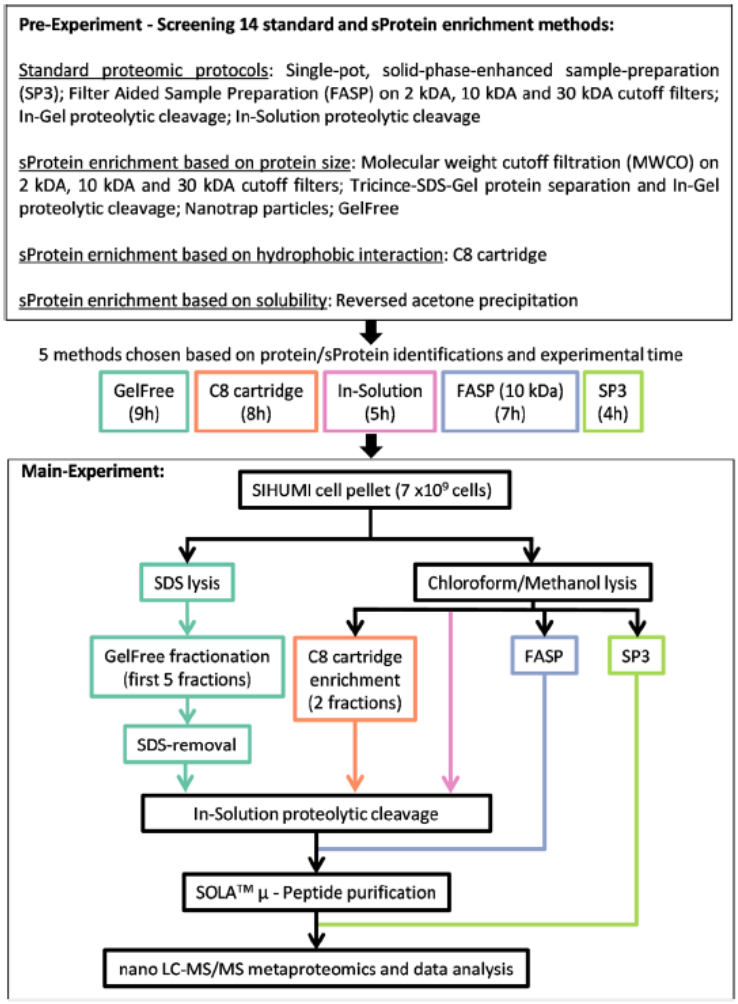

Figure 1: Experimental workflow of tested proteomic and sProtein enrichment methods: Shows the 14 different standard proteomic and sProtein enrichment methods tested in a pre-experiment. GelFree, C8 cartridge, In-Solution, FASP (10 kDa) and SP3 were chosen as most suitable methods and listed with the approximately experimental time from cell lysis to LC-MS/MS measurement (excluding overnight trypsin incubation). Starting from the SIHUMIx cell pellet, the cells were lysed in UPX-Buffer containing SDS or with chloroform/methanol lysis. SDS cell lysates were fractionated using GelFree fractionation, keeping the first 5 fractions containing sProteins. After SDS removal the proteins were proteolytic cleaved, purified and measured using nano LC-MS/MS. Cells lysed in chloroform/methanol were resuspended in Urea buffer and prepared for nano LC-MS/MS analysis using SP3, FASP, In-Solution proteolytic cleavage or loaded on C8 cartridges for sProtein enrichment. Two fractions were collected followed by In-Solution proteolytic cleavage and peptide purification before LC-MS/MS measurement.

Figure 2: Number of detected total proteins **(A)** and sProteins (≤100 amino acids) **(B)** with the percentage of proteins detected in triplicates for standard proteomic protocols: SP3, FASP on 10 kDa MWCO filter and In-Solution cleavage, and two sProtein enrichment strategies: C8 cartridge and GelFree fractionation. **C** shows the relative amount of detected sProteins; n=3. Data are mean ± s.e.m., *P* values calculated by ANOVA with Tukey's multiple comparisons test. A letter on the bar graph indicates the level of significance. Bars denoted by the same letter are not statistically significant (p>0.005).

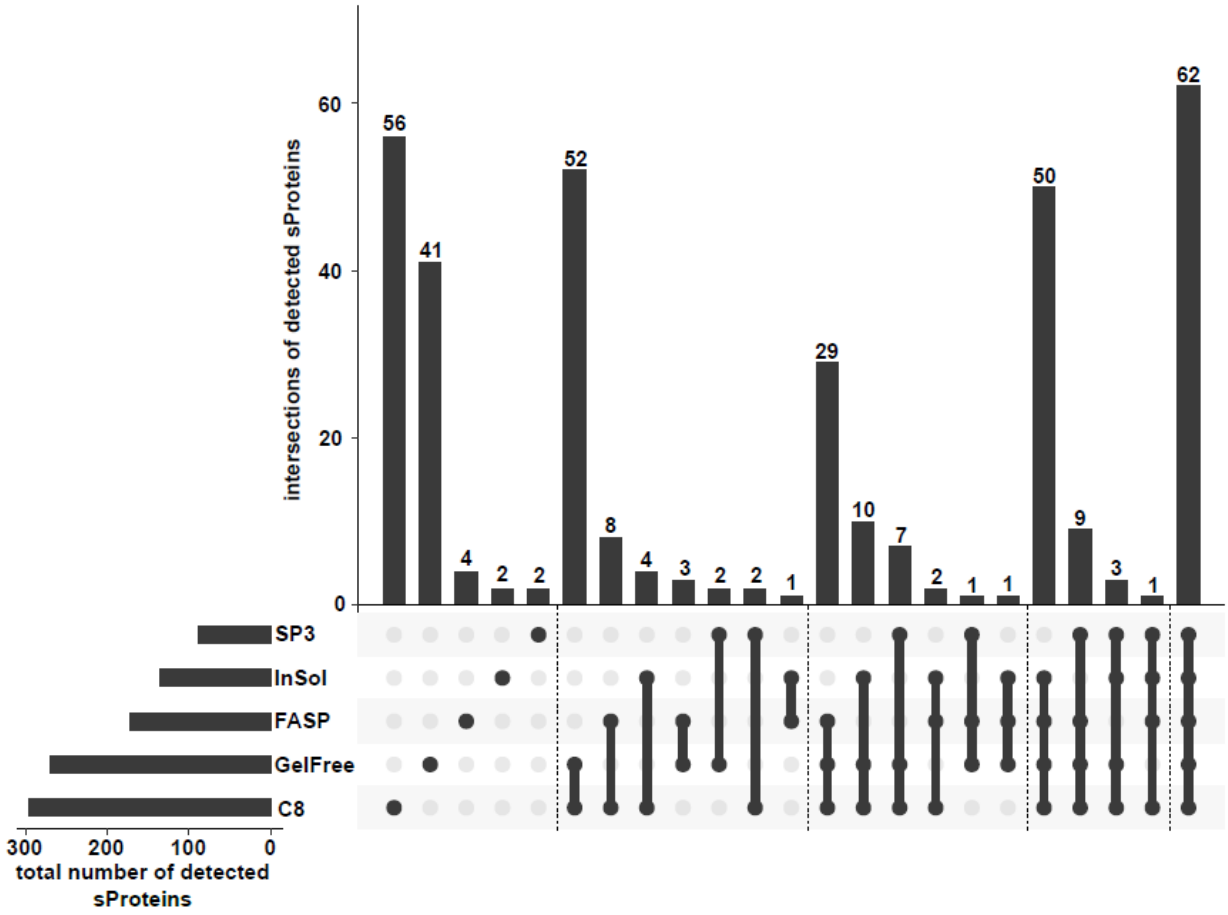

Figure 3 Comparison of number of identified sProteins in triplicates between standard proteomic protocols: SP3, FASP on 10 kDa MWCO filter, In-solution cleavage and two sProtein enrichment strategies: C8 cartridge and GelFree enrichment; n=3.

Figure 4 Comparison of the isoelectric point (pI) (**A**) and grand average of hydropathy (GRAVY) (**D**) between the relative sProtein (≤100 AA) and protein (>100 AA) number. Furthermore the p*I* of relative sProtein number (**C**) and distribution of sProtein and total protein (**B**) is compared between different cell localization based on gene ontology annotation.

Figure 5 MS/MS spectra of peptide GGSGNFAEDRE derived from the *E.coli* sProtein *P56614* (**A**), AAFSYAGLEEATEKK derived from *A.caccae* sProtein B0MAN7 (**B**), EDAKDICYEAK derived from *C.ramosum* sProtein BON696 (**C**), LLKLPSETKPSTR derived from *B. thetaiotaomicron* sProtein Q8A9G7 (**D**) and HDSIAEENIEPNGRPAK derived from *B. thetaiotaomicron* sProtein Q8A758 (**E**) (Table 1, Supplement information).

Figure 6 Comparison of the isoelectric point (p*I*) (**A**) and grand average of hydropathy (GRAVY) (**B**) between the relative sProtein number of characterized and uncharacterized sProteins.

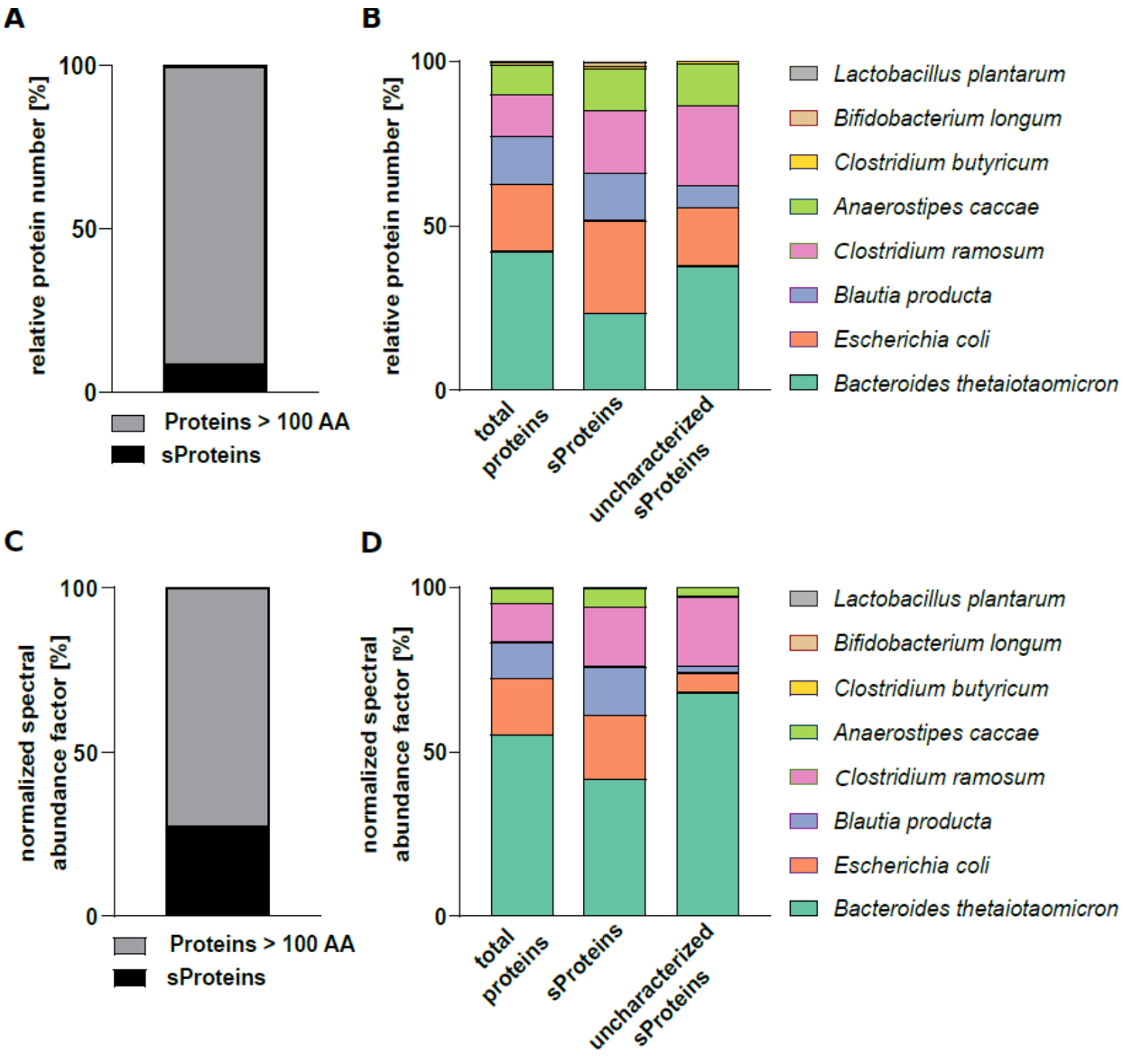

Figure 7 Comparison of relative protein number in % (**A**) and the normalized spectral abundance factor (NSAF) in % (**C**) of detected total proteins and sProteins. The SIHUMIx species distribution based on the

relative protein number in % (**B**) and the normalized spectral abundance factor (NSAF) in % (**D**) of total proteins, sProteins and uncharacterized sProteins is shown.

Supplement Figure 1: Number of detected total proteins **(A)** and sProteins (≤100 amino acids) **(B)** for standard proteomic protocols: SP3, FASP on 2,10 and 30 kDa MWCO filter and In-Solution cleavage, and sProtein enrichment strategies: C8 cartridge, Nanotrap particles, In-Gel fractionation, reverse acetone precipitation, molecular weight cutoff filtration on 2, 10 and 30 kDa MWCO filter and GelFree fractionation.

Table 1 Representative list of identified uncharacterized sProteins including the length in amino acid (AA), identified unique peptides, Sequest HT score, species, and identification methods.

Supplement Table 1: List of identified uncharacterized sProteins including the length in amino acid (AA), identified unique peptides, Sequest HT score, species, and identification methods.

Supplement Table 2: List of identified uncharacterized sProteins including the number of transmembrane helices predicted by TMHMM and signal peptides predicted by SignalP-5.0.

Supplement Table 3: List of uncharacterized sProteins with all proteins predicted to be in a gene cluster, according to KEGG SSDB

Supplement Table 4: List of uncharacterized sProteins with KEGG-Definitions of each gene in the gene clusters

References

1. MacDonald, T.T. and G. Monteleone, *Immunity, Inflammation, and Allergy in the Gut.* Science, 2005. **307**(5717): p. 1920.
2. Tremaroli, V. and F. Bäckhed, *Functional interactions between the gut microbiota and host metabolism.* Nature, 2012. **489**: p. 242.
3. Kinross, J.M., A.W. Darzi, and J.K. Nicholson, *Gut microbiome-host interactions in health and disease.* Genome Medicine, 2011. **3**(3): p. 14.
4. Flint, H.J., et al., *The role of the gut microbiota in nutrition and health.* Nature Reviews Gastroenterology &Amp; Hepatology, 2012. **9**: p. 577.
5. Maurice, Corinne F., Henry J. Haiser, and Peter J. Turnbaugh, *Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome.* Cell, 2013. **152**(1): p. 39-50.
6. Guzman-Rodriguez, M., et al., *Using bioreactors to study the effects of drugs on the human microbiota.* Methods, 2018. **149**: p. 31-41.
7. Kostic, A.D., M.R. Howitt, and W.S. Garrett, *Exploring host–microbiota interactions in animal models and humans.* Genes & Development, 2013. **27**(7): p. 701-718.

8. Stubbendieck, R.M. and P.D. Straight, *Multifaceted Interfaces of Bacterial Competition.* Journal of Bacteriology, 2016. **198**(16): p. 2145.
9. Stubbendieck, R.M., C. Vargas-Bautista, and P.D. Straight, *Bacterial Communities: Interactions to Scale.* Frontiers in Microbiology, 2016. **7**(1234).
10. Mukherjee, S. and B.L. Bassler, *Bacterial quorum sensing in complex and dynamically changing environments.* Nature Reviews Microbiology, 2019. **17**(6): p. 371-382.
11. VanOrsdel, C.E., et al., *The Escherichia coli CydX Protein Is a Member of the CydAB Cytochrome bd Oxidase Complex and Is Required for Cytochrome bd Oxidase Activity.* Journal of Bacteriology, 2013. **195**(16): p. 3640.
12. Hobbs, E.C., et al., *Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance.* Proceedings of the National Academy of Sciences, 2012. **109**(41): p. 16696.
13. Kosfeld, A. and K. Jahreis, *Characterization of the Interaction Between the Small Regulatory Peptide SgrT and the EIICBGlc of the Glucose-Phosphotransferase System of E. coli K-12.* Metabolites, 2012. **2**(4): p. 756-774.
14. Sberro, H., et al., *Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes.* Cell, 2019. **178**(5): p. 1245-1259.e14.
15. Dujon, B., et al., *Complete DNA sequence of yeast chromosome XI.* Nature, 1994. **369**(6479): p. 371-378.
16. Su, M., et al., *Small proteins: untapped area of potential biological importance.* Frontiers in Genetics, 2013. **4**: p. 286.
17. Metzker, M.L., *Sequencing technologies — the next generation.* Nature Reviews Genetics, 2009. **11**: p. 31.
18. Olexiouk, V., W. Van Criekinge, and G. Menschaert, *An update on sORFs.org: a repository of small ORFs identified by ribosome profiling.* Nucleic Acids Research, 2017. **46**(D1): p. D497-D502.
19. Zhang, Y., et al., *Protein Analysis by Shotgun/Bottom-up Proteomics.* Chemical Reviews, 2013. **113**(4): p. 2343-2394.
20. Ma, J., et al., *Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides.* Anal Chem, 2016. **88**(7): p. 3967-75.
21. Müller, S.A., et al., *Optimization of parameters for coverage of low molecular weight proteins.* Analytical and Bioanalytical Chemistry, 2010. **398**(7): p. 2867-2881.
22. Shishkova, E., A.S. Hebert, and J.J. Coon, *Now, More Than Ever, Proteomics Needs Better Chromatography.* Cell systems, 2016. **3**(4): p. 321-324.
23. Yang, X., et al., *Discovery and annotation of small proteins using genomics, proteomics, and computational approaches.* Genome Res, 2011. **21**(4): p. 634-41.
24. Duncan, M.W., R. Aebersold, and R.M. Caprioli, *The pros and cons of peptide-centric proteomics.* Nature Biotechnology, 2010. **28**: p. 659.
25. Ma, J., et al., *Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides.* Analytical Chemistry, 2016. **88**(7): p. 3967-3975.
26. Picotti, P., R. Aebersold, and B. Domon, *The Implications of Proteolytic Background for Shotgun Proteomics.* Molecular & Cellular Proteomics, 2007. **6**(9): p. 1589.
27. Sasse, J. and S.R. Gallagher, *Staining Proteins in Gels.* Current Protocols in Molecular Biology, 2003. **63**(1): p. 10.6.1-10.6.25.
28. Cassidy, L., et al., *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. Journal of Proteome Research, 2016. **15**(10): p. 3773-3783.
29. Schägger, H., *Tricine–SDS-PAGE.* Nature Protocols, 2006. **1**: p. 16.
30. Shafagati, N., et al., *The Use of Nanotrap Particles in the Enhanced Detection of Rift Valley Fever Virus Nucleoprotein.* PLOS ONE, 2015. **10**(5): p. e0128215.
31. Müller, S.A., et al., *Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics.* J Proteome, 2013. **86**.
32. Cassidy, L., P.T. Kaulich, and A. Tholey, *Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes.* Journal of Proteome Research, 2019. **18**(4): p. 1725-1734.
33. Bastida, F., et al., *Differential sensitivity of total and active soil microbial communities to drought and forest management.* Global Change Biology, 2017. **23**(10): p. 4185-4203.
34. Hughes, C.S., et al., *Single-pot, solid-phase-enhanced sample preparation for proteomics experiments.* Nature Protocols, 2019. **14**(1): p. 68-85.
35. Wiśniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nature Methods, 2009. **6**: p. 359.
36. Wessel, D. and U.I. Flügge, *A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids.* Analytical Biochemistry, 1984. **138**(1): p. 141-143.
37. Haange, S.-B., et al., *Disease Development Is Accompanied by Changes in Bacterial Protein Abundance and Functions in a Refined Model of Dextran Sulfate Sodium (DSS)-Induced Colitis.* Journal of Proteome Research, 2019. **18**(4): p. 1774-1786.
38. Käll, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nature Methods, 2007. **4**: p. 923.
39. Perez-Riverol, Y., et al., *The PRIDE database and related tools and resources in 2019: improving support for quantification data.* Nucleic Acids Research, 2018. **47**(D1): p. D442-D450.
40. Zybailov, B., et al., *Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces cerevisiae.* Journal of Proteome Research, 2006. **5**(9): p. 2339-2347.
41. Sato, Y., et al., *SSDB: Sequence Similarity Database in KEGG.* Genome Informatics, 2001. **12**: p. 230-231.
42. Mudunuri, U., et al., *bioDBnet: the biological database network.* Bioinformatics, 2009. **25**(4): p. 555-556.
43. Vale, W., et al., *[38] Assay of corticotropin releasing factor*, in *Methods in Enzymology*. 1983, Academic Press. p. 565-577.
44. Vale, W., et al., *[28] Assay of growth hormone-releasing factor*, in *Methods in Enzymology*. 1986, Academic Press. p. 389-402.
45. Storz, G., Y.I. Wolf, and K.S. Ramamurthi, *Small proteins can no longer be ignored.* Annu Rev Biochem, 2014. **83**: p. 753-77.
46. Davis, R.L., et al., *The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation.* Cell, 1990. **60**(5): p. 733-746.
47. Wang, W., et al., *RNA binding by the novel helical domain of the influenza virus NS1 protein requires its dimer structure and a small number of specific basic amino acids.* RNA, 1999. **5**(2): p. 195-205.
48. Ann M. Stock, a. Victoria L. Robinson, and P.N. Goudreau, *Two-Component Signal Transduction.* Annual Review of Biochemistry, 2000. **69**(1): p. 183-215.
49. Postle, K. and H. Vakharia, *TolC, a macromolecular periplasmic 'chunnel'.* Nature Structural Biology, 2000. **7**(7): p. 527-530.
50. Davidson, A.L. and J. Chen, *ATP-Binding Cassette Transporters in Bacteria.* Annual Review of Biochemistry, 2004. **73**(1): p. 241-268.

51. Stellwag, E.J. and P.B. Hylemon, *Purification and characterization of bile salt hydrolase from Bacteroides fragilis subsp. fragilis.* Biochimica et Biophysica Acta (BBA) - Enzymology, 1976. **452**(1): p. 165-176.
52. Urdaneta, V. and J. Casadesús, *Interactions between Bacteria and Bile Salts in the Gastrointestinal and Hepatobiliary Tracts.* Frontiers in Medicine, 2017. **4**(163).
53. Schubert, K., et al., *Interactions between bile salts, gut microbiota, and hepatic innate immunity.* Immunological Reviews, 2017. **279**(1): p. 23-35.
54. Schiebenhoefer, H., et al., *Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis.* Expert Review of Proteomics, 2019. **16**(5): p. 375-390.
55. Xiong, W., et al., *Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota.* Proteomics, 2015. **15**(20): p. 3424-3438.