Ultra-sensitive metaproteomics (uMetaP) redefines the dark field of metaproteome, enables single-bacterium resolution, and discovers hidden functions in the gut microbiome.

Feng Xian[1], Malena Brenek[1], Christoph Krisp[2], Ranjith Kumar Ravi Kumar[1], Manuela Schmidt[1] ✉, and David Gómez-Varela[1]✉

[1]Systems Biology of Pain, Division of Pharmacology & Toxicology, Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, Vienna, Austria. [2] Bruker Daltonics GmbH & Co. KG, Bremen, Germany.

✉Correspondence to: david.gomez.varela@univie.ac.at and manuela_schmidt@univie.ac.at

ABSTRACT

Metaproteomics uniquely characterizes host-microbiome interactions. However, most species detected by metagenomics remain hidden to metaproteomics due to sensitivity limits. We present a novel ultra-sensitive metaproteomic solution (uMetaP) that, for the first time, reaches full-length 16S rRNA taxonomic depth and can simultaneously decipher functional features. Querying the mouse gut microbiome, uMetaP achieved unprecedented performance in key metrics like protein groups (47925) alongside taxonomic (220 species) and functional annotations (223 KEGG pathways)- all within 30-min analysis time and with high reproducibility, sensitivity, and quantitative precision. uMetaP revealed previously unidentified proteins of unknown functions, small proteins, and potentially new natural antibiotics. Leveraging the extreme sensitivity of uMetaP and SILAC-labelled bacteria, we revealed the true limit of detection and quantification for the “dark” metaproteome of the mouse gut. Moreover, using a two-bacteria proteome mix, we demonstrated single-bacterium resolution (500 fg) with exceptional quantification precision and accuracy. From deciphering the interplay of billions of microorganisms with the host to exploring microbial heterogeneity, uMetaP represents a quantum leap in metaproteomics. Taken together, uMetaP will open new avenues for our understanding of the microbial world and its connection to health and disease.

INTRODUCTION

The interplay between members of microbial communities (microbiomes) is crucial for planetary ecosystem health[1, 2]. While genomic methods have greatly expanded the taxonomic characterization of microbiomes, complementary methods for directly quantifying biological functions are needed for deeper understanding.

Metaproteomics, analyzing microbial samples with liquid chromatography-mass spectrometry (LC-MS)-based proteomics, has emerged as a powerful tool for investigating host-microbiome interactions[3]. It offers a direct view of functional changes within hundreds of microbes and their host[4]. However, compared to single-organism LC-MS, metaproteomics faces significant hurdles. Microbial samples are immensely complex, harboring an estimated 100 million unique peptide species[5] with a vast dynamic range[6]. Consequently, current methods miss more than 80% of the species detectable by commonly used genomic-based methods (e.g., full-length 16S rRNA), limiting functional characterization[6]. A significant leap in sensitivity, depth, throughput, and accuracy is urgently needed.

We present a novel ultra-sensitive metaproteomic solution, termed uMetaP, leveraging cutting-edge technologies for unprecedented sensitivity. uMetaP combines an Ultra-High-Performance Liquid Chromatography (UHPLC) column for the minimization of peak broadening, an optimized ionization source for maximizing ion transfer[7], and the extreme sensitivity of the timsTOF Ultra mass spectrometer[8, 9].

Utilizing mouse feces as a model, uMetaP achieved dramatic improvements in key metrics. This includes quantifying unprecedented amounts of protein groups and associated functional pathways for both the host and its microbiome (47925 protein groups and 223 KEGG pathways). Remarkably, uMetaP quantified 220 microbial species, which is comparable to the average amount reached by full-length 16S rRNA, with exceptional reproducibility and quantitative precision in DIA-PASEF runs using low sample amounts (25 ng of peptides). Additionally, uMetaP revealed previously hidden functions, identifying 3252 proteins of unknown function[10] (PUFs), and taxonomic and functional annotating of 2,181 small proteins[11, 12] (sProt), and 553 proteins gathering predicted antimicrobial peptide sequences [13] (AMPs).

Leveraging uMetaP with a meticulous experimental design utilizing novel SILAC-labelled bacteria, we addressed two longstanding challenges in metaproteomics: 1) the true limit of detection for the dark metaproteome, and 2) achieving single-bacterium resolution. We confidently detected 10000 spiked-in bacteria among a theoretical amount of 10 billion, lowering the limit of detection 5000-fold[6]. Moreover, we confidently identified species-specific peptides from a 500-femtogram two-bacteria peptide mix with remarkable quantification precision and accuracy. These results position uMetaP as a powerful counterpart to genomic approaches for investigating microbial ecosystems, highlighting its potential for diverse biotechnological applications and uncovering unexplored biology.

MATERIAL AND METHODS

Reagents

All reagents were purchased from Sigma-Aldrich (St. Louis, Missouri) if not mentioned otherwise. Acetonitrile (ACN) and formic acid (FA) were purchased from Fisher Scientific (Hampton, New Hampshire; both FA and ACN were liquid chromatography-mass spectrometry (LC-MS) grade). LC-MS grade water from Sigma-Aldrich was used for all solutions. Protease inhibitor was purchased from Roche (Complete Ultra Tablets Mini, Roche, Basel, Switzerland).

Fecal sampling, protein extraction, SP3-assisted protein digestion, and peptide clean-up

In-house bred C57BL/6J mice were used. Housing and operation of mice were carried out with the approval of the University of Vienna animal care and use committee (license number 2021-0.138.925). All mice used in this study were group-housed in individually-ventilated cages in a 12 h light/dark cycle in the animal facility with water and food ad libitum. The procedures from fecal collection to final peptide sample were done as previously described[14]. A pooled peptide sample was made from samples previously prepared and was used subsequent analysis for data presented in Figure 1 and Figure 2.

Bacterial culture of *L.Murinus* and *S.Ruber*

Ligillacotobacillus murinus (DSM 20452, *L.Murinus*) and *Salinibacter ruber* (DSM 13855, *S.Ruber*) were purchased from DSMZ (Braunschweig, Germany). All culture mediums were autoclaved right after the preparation. *L.Murinus* was activated in 5 mL MRS medium (CARLROTH, Karlsruhe, Germany; prepared according to instruction) and incubated for 24 hours at 37 °C with 220 rpm agitation. At the end of this incubation period, 1 mL of the *L.Murinus* culture was taken and centrifuged at 3200 g for 5 minutes at 4 °C. The supernatant was carefully removed, and the bacterial pellet was gently resuspended in 5 mL SILAC-heavy medium (Glucose 10 g/L, KH_2PO_4 3 g/L, K_2HPO_4 3 g/L, sodium acetate 5 g/L, ammonium citrate dibasic 1 g/L, $MgSO_4 \cdot 7H_2O$ 0.2 g/L, $MnSO_4 \cdot 4H_2O$ 0.05 g/L, Tween-80 1 g/L, L-alanine 0.05 g/L, L-arginine-HCl ($13C_6$, $15N_4$; Fischer Scientific) 0.05 g/L, L-asparagine 0.1 g/L, L-aspartic acid 0.1 g/L, L-cysteine 0.2 g/L, L-glutamine 0.1 g/L, L-glutamic acid 0.1 g/L, glycine 0.05 g/L, L-histidine 0.05 g/L, L-isoleucine 0.05 g/L, L-leucine 0.05 g/L, L-lysine-2HCl ($13C_6$, $15N_2$; Fischer Scientific) 0.05 g/L, L-methionine 0.05 g/L, L-phenylalanine 0.05 g/L, L-proline 0.05 g/L, L-serine 0.05 g/L, L-threonine 0.05 g/L, L-tryptophan 0.05 g/L, L-tyrosine 0.05 g/L, L-valine 0.05 g/L, uracil 0.01 g/L, guanine 0.01 g/L, adenine 0.01 g/L, xanthine 0.01 g/L, biotin 0.01 g/, Vitamin Solution 2% (v/v)). The heavy-medium culture was incubated at 37 °C with 220 rpm agitation for 24 hours. Bacterial growth was monitored with spectrophotometric measurements (Eppendorf, Hamburg, Germany) at an optical density of 600 nm (OD600). An OD600 above 0.8 was aimed to ensure suitable growth conditions. For daily passage, 500 microliters of *L.Murinus* culture was taken and transferred to another 5 mL SILAC-heavy medium. The labelling efficiency was evaluated on timsTOF Pro after 10 passages in heavy-medium culture. *S.Ruber* was cultured in 5 mL DMSZ-936 medium according to the recommendation (https://mediadive.dsmz.de/medium/936) at 37 °C with 220 rpm agitation. The duration between passages for *S.Ruber* was around 7 days due to its slow growth. For enlarged culture, 1 mL/each of *L.Murinus* and *S.Ruber* cultures were transferred to 30 mL mediums, respectively.

At the end of cultivation, 2 mL bacteria aliquots were made and pelleted at 3200 g for 5 minutes at 4 °C, and one of aliquots was resuspended in 2 mL of either pre-chilled PBS (*L.Murinus*) or DSMZ-936 medium (*S.Ruber*). The resuspended bacteria were further serial diluted (2-50 times dilution) with either PBS (*L.Murinus*) or DSMZ-936 medium (*S.Ruber*) for bacteria counting using QUANTOM Tx Microbial Cell Counter (BioCat, Heidelberg, Germany) according to the procedures supplied with the device. The rest of the aliquots were snap-frozen in liquid nitrogen and stored at -80°C until further use.

High pH reversed-phase fractionation of pooled peptides

Peptide fractionation kit was purchased from Fisher Scientific (Cat. 84868). A total of 40 µg pooled peptides were processed according manufactures instruction. Eight peptide factions were dried using vacuum centrifugation and then re-suspended in 30 µL of MS-grade water. The peptide concentration was measured in duplicate using NanoPhotometer N60 (Implen, München, Deutschland) at 205 nm. Peptide samples were acidified with FA to a final concentration of 0.1% and stored at-20°C until LC-MS/MS analysis.

Liquid chromatography-mass spectrometry configurations

Nanoflow reversed-phase liquid chromatography (Nano-RPLC) was performed on NanoElute1 and NanoElute2 systems (Bruker Daltonik, Bremen, Germany) coupled with timsTOF Pro and timsTOF Ultra (Bruker Daltonik, Bremen, Germany) via CaptiveSpray ion source, respectively. Mobile solvent A consisted of 100% water, 0.1% FA and mobile phase B of 100%, 0.1% FA.

Data dependent acquisition (DDA-PASEF) of fractionated peptides on timsTOF Ultra and timsTOF Pro

Twenty-microgram of each peptide fraction was loaded on an Aurora™ ULTIMATE column (25 cm x 75 µm) packed with 1.6 µm C18 particles (IonOpticks, Fitzroy, Australia) with a total 66 minutes gradient. The mobile phase B was linearly increased from 5 to 23% in 56 minutes with a flowrate of 0.25 µL/min, followed by another linear increase to 35% within 4 minutes and a steep increase to 90% in 1 minute. The mobile phase B was maintained at 90% for the last 5 minutes with a flowrate increase from 0.25 µL/min to 0.35 µL/min. On both timsTOF Ultra and timsTOF Pro, the TIMS analyzer was operated in a 100% duty cycle with equal accumulation and ramp times of 166 ms each. Specifically, 5 PASEF scans were set per acquisition cycle (cycle time of 1.03 s) with ion mobility range from 0.7 to 1.3 (1/k0). The target intensity and intensity threshold were set to 14000 and 500 respectively. Dynamic exclusion was applied for 0.4 minutes. Ions with m/z between 100 and 1700 were recorded in the mass spectrum. Collision energies were dependent on ion mobility values with a linear increase in collision energy from 1/K0 = 0.6 Vs/cm² at 20 eV to 1/K0 = 1.6 Vs/cm² at 59 eV. In addition, 100 µg of each peptide fraction was analyzed in 90 minutes gradient on timsTOF Pro. In this setup, the mobile phase B was linearly increased from 0 to 20% in 65 minutes with a flowrate of 0.25 µL/min, followed by another linear increase to 35% within 15 minutes and a steep increase to 90% in 0.5 minute. The mobile phase B was maintained at 90% until the end of analysis with a flowrate of 0.4 µL/min. The TIMS analyzer was operated in 100% duty cycle with 100 ms defined for accumulation and ramp time. Ten PASEF scans were set per acquisition cycle (cycle time of 1.17

s) with ion mobility range from 0.65 to 1.45 (1/k0). The target intensity and intensity threshold were set to 10000 and 1750 respectively.

Data independent acquisition (DIA-PASEF) on timsTOF Ultra

Peptides were loaded onto an Aurora™ ULTIMATE column (25 cm x 75 µm) packed with 1.6 µm C18 particles (IonOpticks, Fitzroy, Australia) with either a total 30 minutes or 66 minutes gradient on a NanoElute2 system in triplicates. In the 30-min separation, the mobile phase B was linearly increased from 5 to 23% in 18 minutes with a flowrate of 0.25 µL/min, followed by another linear increase to 35% within 4 minutes and a steep increase to 90% in 2 minutes. The mobile phase B was maintained at 90% for the last 4 minutes with a flowrate increase from 0.25 µL/min to 0.35 µL/min. The composition of mobile phase B over the 66-min separation was the same as described above for the fractionated peptide samples. For the results presented in Figure 2, precursors with m/z between 400 and 1000 were defined in 8 scans (3 quadrupole switches per scan) containing 24 ion mobility steps in an ion mobility range of 0.64 – 1.45 (1/k0) with fixed isolation window of 25 Th in each step. The acquisition time of each DIA-PASEF scan was set to 100 ms, which led to a total cycle time of around 0.95 sec. For results presented in Figure 4 and 5, peptides were separated on the NanoElute 2 system with a 30-min gradient. Precursors with m/z between 350 and 800 were defined in 6 scans (3 quadrupole switches per scan) containing 18 ion mobility steps in an ion mobility range of 0.64 – 1.2 (1/k0) with fixed isolation window of 25 Th in each step. The acquisition time of each DIA-PASEF scan was set to 100 ms, which led to a total cycle time of around 0.74 sec.

L.Murinus and *S.Ruber* Spike-in experiment

Counted *L.Murinus* and *S.Ruber* stocks were resuspended and diluted in pre-chilled PBS to reach various numbers (ranging from 1 x e^{+4} to 1 x e^{+9}) in triplicates. Same number of *L.Murinus* and *S.Ruber* were mixed with 10 mg of mouse feces and subjected to protein extraction together (as previously described[14]). To keep consistency of spike-in background, the fecal sample used here was collected and pooled from the same mouse in two consecutive days at the same hour. The resulted peptide samples were analyzed on the timsTOF Ultra in a 30-min gradient as described above with 25 ng peptide / sample. The workflow is illustrated in Figure 4A.

DDA-PASEF data processing

Three sets of fractionated data generated using timsTOF Ultra and timsTOF Pro were separately submitted to MSfragger[15] (Version 3.7) integrated in FragPipe computational platform (Version 19.1) searching against a metagenome-translated protein database (PD1) containing 2.6 million protein sequences[16]. The decoy database was generated with reversed sequences. Trypsin was specified with a maximum of 2 missed cleavages allowed. The search included variable modifications of methionine oxidation and N-terminal acetylation and a fixed modification of carbamidomethyl on cysteine. The mass tolerances of 20 ppm for both precursor and fragment were used. Peptide length was set to 7 to 50 amino acids with a mass range from 500 to 5000 Da. The rest parameters were kept as default settings. During the validation, MSBooster (Version 1.1.10) was used for rescoring and Percolator[17] (version 3.5, default parameters) was checked for PSM validation. FDR level was set to 1% for PSM, peptide and protein. The identified

proteins from the search were combined to form a sample-specific protein database (PD2) containing 48970 sequences. For evaluating the labelling efficiency of *L.Murinus*, the data was searched against the standard proteome of *L.Murinus* downloaded from Uniprot (PD3, https://www.uniprot.org/proteomes/UP000051612, accessed on 2023-07-19) containing 1971 protein sequences. The rest parameters were kept the same in MSfragger as aforementioned.

DIA-PASEF data processing

DIA-NN[18] (version 1.8.1) was used to process DIA-PASEF data in library-free mode with PD2 and mus musculus reference proteome (PD4, https://www.uniprot.org/proteomes/UP000000589, accessed on 2023.04.07) to generate the predicted spectrum library. A deep learning-based method was used to predict theoretical peptide spectra along with their retention time and ion mobility. Trypsin/P was used for in silico digestion with an allowance of a maximum of 2 missed cleavages. Variable modifications on peptides were set to N-term methionine excision, methionine oxidation and N-term acetylation, while carbamidomethylation on cysteine was a fixed modification. The maximum number of variable modifications on a peptide was set to 2. Peptide length for the search ranged from 7 to 50 amino acids. Aligned with the DIA-PASEF acquisition method, m/z ranges were specified as 400 to 1000 for precursors and 100 to 1700 for fragment ions. Both MS1 and MS2 mass accuracy were set to automatic determination. Protein inference was set to “Protein names (from FASTA)” and the option of “Heuristic protein inference” was unchecked. Match-between-run (MBR) was enabled for cross-run analysis. RT-dependent cross-run normalization and Robust LC (high precision) options were selected for quantification. DIA-NN (version 1.8.2 Beta27) was used to process DIA-PASEF data presented in Figure 4 in library-free mode. The spectrum libraries were predicted based on PD2, PD3, PD4 and the reference proteome of *S.Ruber* (PD5, https://www.uniprot.org/proteomes/UP000008674, accessed on 2023-07-19, 2812 protein sequences). Trypsin/P was used for in silico digestion with an allowance of a maximum of 1 missed cleavage. Variable modifications on peptides were set to N-term methionine excision, methionine oxidation and N-term acetylation, while carbamidomethylation on cysteine was a fixed modification. The maximum number of variable modifications on a peptide was set to 2. Peptide length for the search ranged from 7 to 30 amino acids. Aligned with the DIA-PASEF acquisition method, m/z ranges were specified as 350 to 800 for precursors and 100 to 1700 for fragment ions. The same search parameters were used to process data presented in Figure 5, and the only difference was only PD3 and PD5 were used for the prediction of the spectrum library, as there was no complex fecal background used.

The DIA-NN search outputs were further processed with the R package, DIA-NN (https://github.com/vdemichev/diann-rpackage), to either extract the raw intensity or calculate the MaxLFQ[19] quantitative intensity for all identified peptides and protein groups with q-value < 0.01 as criteria at precursor and protein group levels.

DIA-PASEF spectrum visualization

Skyline[20] (version 23.1.0.380) was used to visualize the spectrum of identified peptides by DIA-NN. Briefly, the spectra library constructed by DIA-NN after database search was imported and used to build Skyline library containing precursor information of detected peptides. Then

199 precursors presented in the library and their associated fragment ions were extracted from the
200 raw DIA-PASEF data. During the extraction, the mass accuracy was set to 10 ppm for both
201 precursor and fragment, and only scans within 5 minutes of MS/MS identifications (retention
202 time included in the library) were extracted to avoid false matches. Results presented in Figure
203 4C-D, Figure 5B-C, and Supplementary Figure 4D-E were based on the Skyline process.

204 Taxonomy and function analysis

205 iMetaLab[21] (Version 2.3.0) was used for taxonomy and function annotation. The peptide
206 sequences and corresponding intensities data (microbiome and host) were imported into
207 iMetaLab, and the built-in taxonomy database was used for the mapping (ignore blanks below
208 rank: Super kingdom, unique peptide count >=3). Protein identifications with corresponding
209 MaxLFQ intensities were imported into iMetaLab for functional annotation (default
210 parameters). For the functional annotation of proteome databases (PD2, PD3, and PD5) related
211 to Figure 3C and Figure 4G, the fasta files were imported to EggNOG-mapper[22] (http://eggnog-
212 mapper.embl.de/) using default settings. Proteins with unknown functions (PUFs) were counted
213 only when they were not annotated in GO (Gene Ontology), KEGG (Kyoto Encyclopedia of
214 Genes and Genomes), and EC (Enzyme Commission), in addition, missing annotation in COG
215 (Clusters of Orthologous Genes) or associated with COG cluster S (unknown function). Small
216 proteins were considered with ≤ 100 amino acids (Figure 3C). Small proteins were subjected
217 for antimicrobial peptides (AMPs) analysis with Antimicrobial Peptide Scanner vr.2
218 (https://www.dveltri.com/ascan/v2/ascan.html).

219 Identification filters of *L.Murinus* and *S.Ruber* for spike-in experiment

220 Peptide and protein identifications generated from DIA-NN search (version 1.8.2 Beta27) of the
221 spike-in experiment were further filtered to have species-specific identifications: 1) only heavy-
222 labelled peptides were considered for *L.Murinus* to remove the interference from the
223 endogenous species, and heavy-labelled peptides assigned to *S.Ruber* were removed as they
224 are false-matches; 2) co-assigned peptides and protein groups (shared between *L.Murinus* and
225 *S.Ruber*) were removed; 3) peptides assigned to *L.Murinus* or *S.Ruber* which were also
226 identified in any of the non-spike controls (3 replicates) were removed. The filtered data is
227 presented in Figure 4 and Supplementary Table 10.

228 Pathway analysis of host proteome

229 The host protein identifications were analyzed in Metascape[23] (https://metascape.org/) and
230 searched against the KEGG database with default settings.

RESULTS

uMetap sets state-of-the-art in metaproteomic database generation.

Metaproteomic workflows often require as a first stage the creation of a sample-specific protein database in which a metagenome-translated protein database, containing millions of sequences, is reduced using Data-Dependent-Acquisition (DDA) mass spectrometry data[24]. The size and spectral quality of this reduced database is a crucial step that governs the depth and sensitivity of subsequent metaproteomic analyses[25].

To obtain a comprehensive and confident proteotyping of complex metaproteomic samples, we set a multi-faceted strategy combining uMetaP powered by the extra mobility dimension provided by DDA-Parallel Acquisition Serial Fragmentation[26] (DDA-PASEF), the use of pH-fractionated mouse fecal microbiome peptide samples, and a database creation pipeline including validation and rescoring at an FDR 1% level. As shown in Figure 1A, uMetaP fragmented up to 4 times more precursors across all 8 fractions compared to a similar approach using the previous generation timsTOF Pro mass spectrometer, even if the latter was used with a longer gradient and 5 times more injected peptides. As a result, we obtained a higher number of peptide-spectrum matches (PSMs; Supplementary Figure 1A; Supplementary Table 1), which resulted in 4 times more total microbial peptides identified (131,911; including 90.2% of the peptides detected with the timsTOF Pro; Supplementary Figure 1B; Supplementary Table 2). A detailed analysis of commonly identified peptides showed an enhanced detection signal along the whole range of intensities including the low-abundance species (Figure 1B), which strongly suggests a significant increase in sensitivity by uMetaP. We also observed similar gains in unique identifications across all peptide intensity levels (Supplementary Figure 1C). It's important to note that the two datasets (uMetaP and timsTOF Pro) were processed separately without any identification transfer (match-between-runs) in between. The peptide profiling improvement resulted in 3.3 times higher protein coverage (37724 vs. 11333 protein groups), including protein groups not detected with the timsTOF Pro (Figure 1C). Importantly, these proteotyping improvements primarily focused on medium and low-abundant proteins (Figure 1D).

The profiling results obtained with uMetaP were submitted through a careful database creation process that resulted in a mouse gut protein database containing 48970 unique protein sequences- a 4-fold versus the previous largest DDA-PASEF metaproteomic database[14]. In-silico digestion followed by taxonomic annotation (using a minimum of 3 taxa-specific peptides) indicated that the theoretical taxonomic profiling potential of this database is comprised of 82 phyla, 735 genera, and 812 species (Figure 1E; Supplementary Table 3). The improvements in depth and sensitivity provided by uMetaP are reflected in the breadth of the database and its ability to represent the classical dynamic range of the fecal microbiome: ten species represent 58.3% of the total peptide landscape (Figure 1E) while 511 species represented < 0.02% of the peptide content (Supplementary Table 3). The functional annotation of our database denoted the study of 221 KEGG pathways and 21 COG functional categories (Supplementary Table 4). A detailed analysis revealed the presence of vastly uncharacterized but crucially important proteins for host-microbiome dynamics: 3252 PUFs, 2181 sProt, and 553 AMPs (Supplementary Table 4).

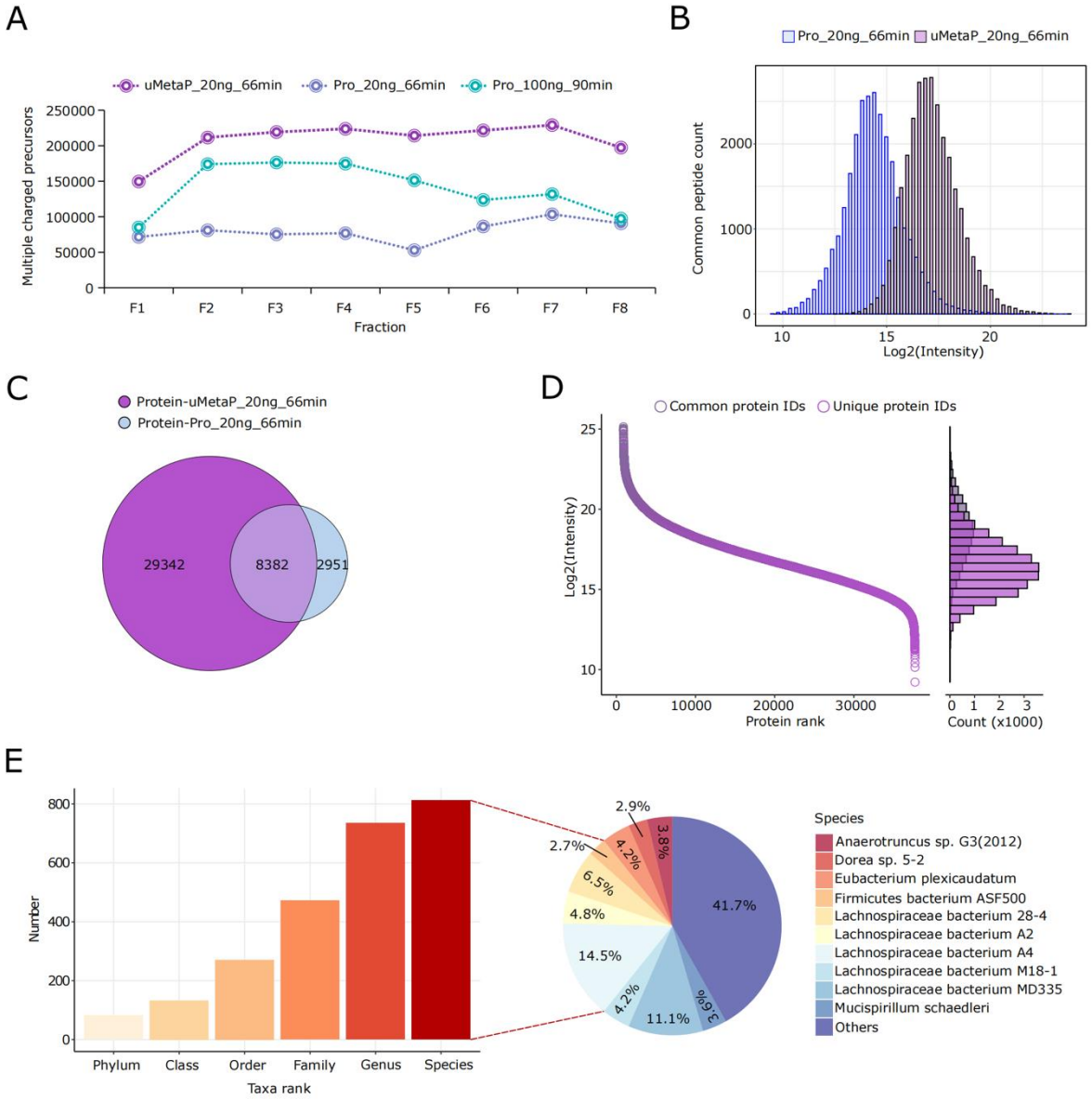


Figure 1. Benchmarking uMetaP's performance for metaproteomic database generation using DDA-PASEF. A) uMetaP DDA-PASEF-based fragmentation performance compared to timsTOF Pro in eight pH fractions of a mouse fecal sample B) Intensity distribution of commonly identified peptides between uMetaP and timsTOF Pro demonstrating increased sensitivity by uMetaP. C) Comparison of protein identifications between uMetaP and timsTOF Pro. D) Intensity ranks (left) and histogram (left) of common (uMetap and timsTOF Pro) and uMetaP unique identified protein identifications (IDs) showing that uMetaP primarily targets medium and low-abundant proteins. E) Taxonomic profiling potential of the in-silico digested uMetaP-based microbial protein database (PD2; left) and relative species abundance calculation using the number of annotated species-specific peptides divided by the total peptides count annotated to the species level (right). A filter of a minimum of 3 taxa-specific peptides was used.

DIA-PASEF-powered uMetaP maximizes proteotyping at unprecedented sensitivity, quantification precision and reliability.

Our previous work demonstrated the advantages of combining Data-Independent-Acquisition (DIA)-PASEF together with deep neural network-based data analysis for the study of complex metaproteomic samples[14]. As shown in Figure 2A, using uMetaP powered by DIA-PASEF increased the microbial and host peptide and protein identifications as more peptides were injected during a short gradient of 30 minutes. The identification rate continued to increase linearly until plateauing at a total of 101987 peptides (93119 microbial and 8868 mouse peptides; an average of three injection replicates) at 25 ng of injected peptides. Extending the LC gradient to 66 minutes increased the number of identified peptides and protein groups, reaching saturation at 50 ng of injected peptides (total of 126611 peptides and 47925 protein groups; Figure 2A; Supplementary Table 5). In line with an increase in sensitivity, uMetaP detected an average of 145 microbial peptides and 416 host peptides, corresponding to 63 microbial and 85 host protein groups, at an unprecedented low sample amount of 10 picograms (Supplementary Table 5). We identified peptides spanning over four orders of magnitude using 10 ng of injected peptides with a 30-min gradient. More than 83% of all peptides exhibited a coefficient of variation (CV) lower than 0.2 (Figure 2B) across different peptide quantities and gradient lengths (examples in Supplementary Figure 2 A-B). This high quantitative precision improved as the peptide amount increased, from a CV of 18.11% (median value) with 10pg to 8.31% with 50 ng (Figure 2C). The high correlations obtained when 10 ng of peptides were analyzed under two different gradient lengths, demonstrated the extraordinary quantification reliability both for the host (r-value = 0.992) and the microbial proteins (r-value = 0.972; Figure 2D).

Reaching unprecedented taxonomic depth while quantifying previously hidden functions.

We set to discover the potential of uMetaP for understanding the taxonomic and functional profiles of complex microbial systems. We observed an increasing taxonomic and functional coverage as more peptides were injected during a short gradient of 30 minutes, with a minor improvement offered by doubling the gradient length (Figure 3A; Supplementary Table 6). uMetaP confidentially identified 203 and 220 microbial species (cutoff: 3 peptide-specific species) when as low as 25 ng sample was loaded onto a 30- or a 60-min gradient (Figure 3A). Our results encompassed organisms from all four major kingdoms present in the mouse gut microbiome (examples shown in Supplementary Figure 3A). uMetaP reached notable improvements across all taxonomic levels (a 3-fold at the species level) in comparison to our previously published dataset obtained in a timsTOF Pro mass spectrometer[14] in which we used 10 times injected peptides and 2.3 times longer gradient (Supplementary Figure 3B). This taxonomic coverage sets a new state-of-the-art in metaproteomics by reaching the average performance offered by full-length 16S rRNA in a human gut microbiome[27]. Notably, we identified 3 bacterial species (*Anaerotruncus* sp. G3, *Lachnospiraceae bacterium A4*, and *Lachnospiraceae bacterium MD335*) with 10 pg, indicating an unprecedented taxonomic sensitivity.

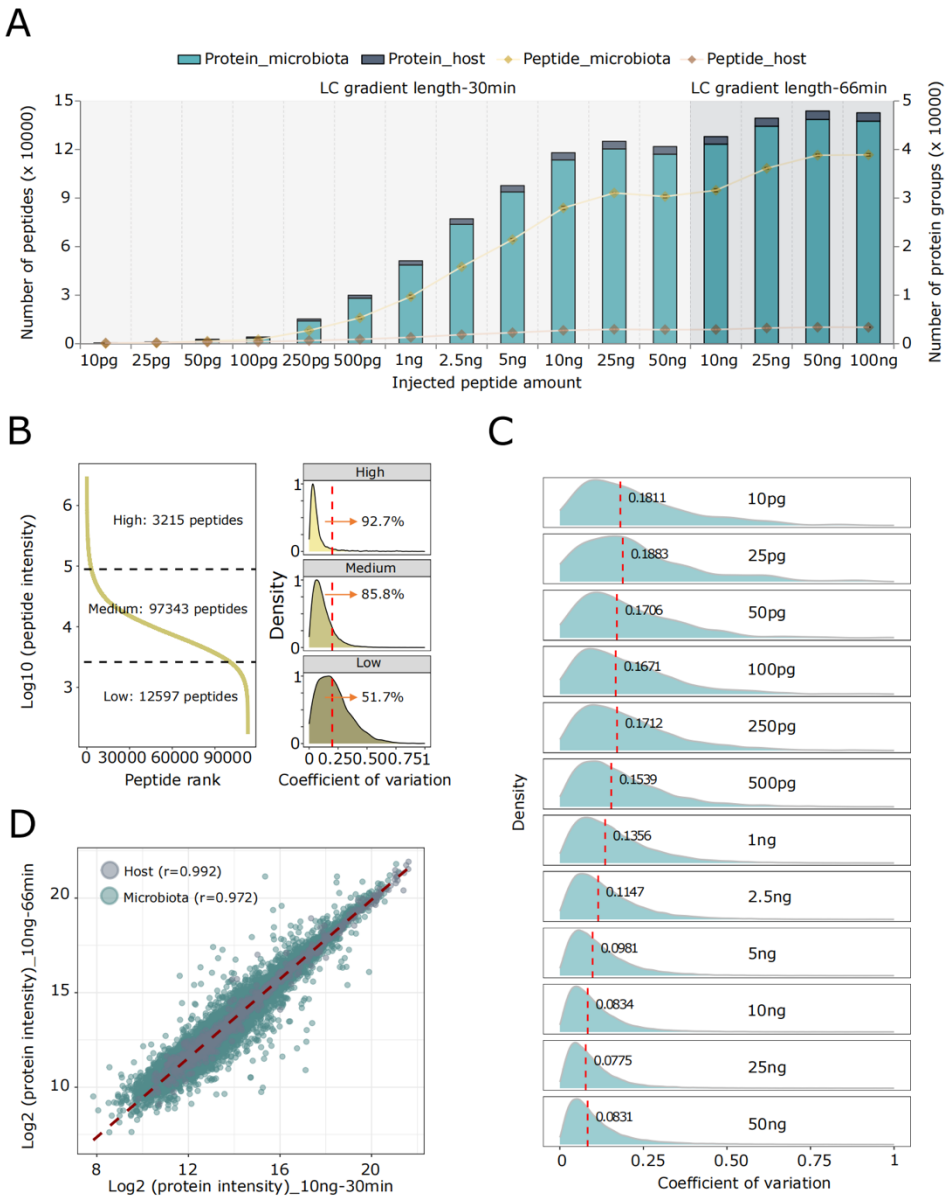


Figure 2. Proteotyping characterization of uMetaP powered by DIA-PASEF. A) Peptide and protein identifications obtained with uMetaP along the peptide dilution series and LC gradient lengths (average of injection triplicates). Error bars show the standard deviations. B) Intensity rank of peptides identified using 10 ng in a 30-min gradient (left) and the coefficient of variations (CV) across different abundance regions (right) delimited by the dotted lines in the intensity rank graph. The red-dotted lines indicate CV = 0.2. C) CVs of identified peptides (triplicates) indicate the high quantitative precision along the dilution series. The red-dotted lines indicate the median CV values. D) Quantitative reliability by calculating the correlations of identified protein groups (host and microbiota) using 10 ng with two different LC gradients.

310 Focusing on functional profiling, the detected microbial proteins enabled the annotation of 161
311 and 44 KEGG terms using 25 ng and 10pg of peptides, respectively, in a 30-min gradient (Figure
312 3A; Supplementary Table 6). We revealed the representation of 24 COG functional categories
313 for the microbiome (Figure 3B; Supplementary Table 7). At the same time, identified host
314 proteins belonged to 62 and 8 KEGG pathways using 25 ng and 10 pg, respectively (Figure 3A).
315 Remarkably, we identified 13 COG functional categories for the host with 10 pg of peptides
316 (Supplementary Table 6). Detailed analysis identified a total of 583 PUFs and 1405 sProt, among
317 which 384 contain predictive AMPs using 25 ng and 30-min gradient (Figure 3C; Supplementary
318 Table 7). Taxonomic and functional annotation of these sProt showed that they are mostly
319 enriched in 6 taxa (Clostridia, Bacteroidetes, Bacilli, Firmicutes, Deferribacteres, and
320 Deltaproteobacteria) and involved in pathways including ribosome, carbon metabolism,
321 microbial metabolism in diverse environments, biosynthesis of secondary metabolites, and
322 carbon fixation pathways in prokaryotes (Supplementary Table 8).

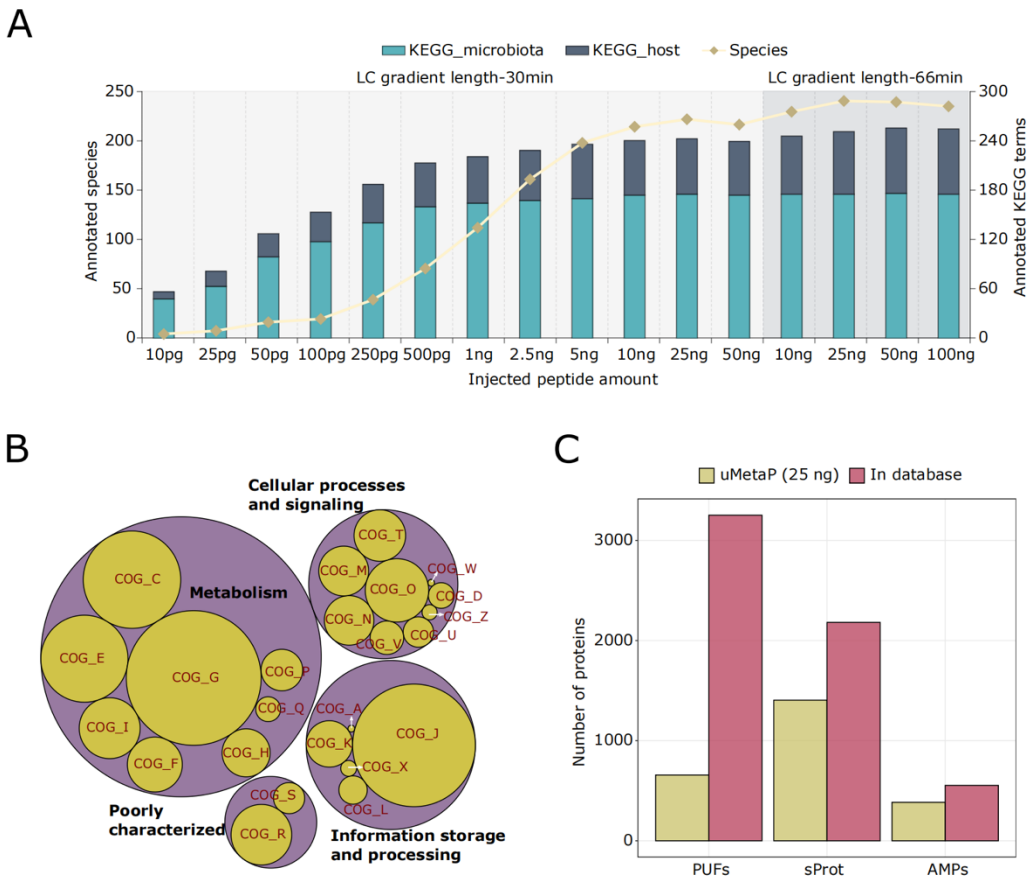


Figure 3. Taxonomic and functional profiling with uMetaP powered by DIA-PASEF. A) Taxonomic and functional coverage with varying peptide quantities and LC gradient lengths (average of injection triplicates). B) Representation of 24 COG functional categories annotated for microbial proteins identified using 25 ng with a 30-min gradient. C) Amount of PUFs, sProt, and AMPs experimentally detected using 25 ng with a 30-min gradient and present in the reduced protein database (PD2).

Redefining the dark field of metaproteome by Improving 5000 times the detection limit.

The human gut microbiome harbors an average of 200 bacterial species[27, 28] comprising a core of abundant species present in most individuals[29, 30] and a second pool formed by low-abundant species (more than 50% of the total) differing among people[6]. Current metaproteomic approaches do not achieve sufficient sensitivity to study these low-abundant species[6], resulting in a vastly uncharacterized dark metaproteome.

We established a novel strategy to calculate, for the first time in metaproteomics, the real Lowest Limit of Detection (LLoD) and Quantification (LLoQ): the lowest number of bacteria that can be reliably identified and quantified by metaproteomics in a complex microbiome sample. To reduce identification uncertainties, we set up a stable isotope labelling in cell culture[31] (SILAC) for *Ligillacotobacillus murinus* (*L.Murinus*), which is endogenous to the mouse gut microbiome. DDA-PASEF analysis of the SILAC culture indicated an average of 97.42% incorporation efficiency for *L.Murinus* (Supplementary Figure 4A; Supplementary Table 9). Additionally, we also used the exogenous *Salinibacter Ruber* bacterium (*S.Ruber*). We found that the number of detected peptides and protein groups decreased as we lowered the number of SILAC-labeled *L.Murinus* and non-labelled *S.Ruber* bacteria spiked into 10 mg of mouse feces (Figure 4A-B; Supplementary Figure 4B). After applying strict filtering criteria to taxonomic identifications for both bacteria (including a Non-spiked control; see Methods), we identified on average 17 and 29 peptides for *L.Murinus* and *S.Ruber* when 10000 bacteria were spiked (Figure 4B; Supplementary Table 10). We confirmed the identifications by inspecting identified spectra (examples in Supplementary Figure 4D-E). In LC-MS-based proteomics, LLoD often differs from the Lowest Limit of Quantification (LLoQ). By extracting precursors and their corresponding ion fragments from complex DIA-PASEF spectra, we found a LLoQ of 1 and 5 million bacteria for *L.Murinus* (example in Figure 4D) and *S.Ruber* (example in Figure 4E), respectively, to maintain a reliable quantification. We hypothesized that the differences in bacteria size (Supplementary Figure 4C), therefore in the total protein content of each bacteria species, underlined this difference in LLoQ.

The relative abundance of the species in a microbial community is a key parameter in the study of microbiomes. Using the sum of intensities of species-specific peptides[32], we showed that taxa abundance assessed by uMetaP is dominated by a few species: 8 species (excluding spiked-in *L.Murinus* and *S.Ruber*) contribute to 55% biomass of the microbiota (Figure 4E; Supplementary Table 11). Calculation based on identified peptide intensity indicated that the 10000 spiked *L.Murinus* and *S.Ruber* detected by uMetaP correspond to 0.001% and 0.006% of the total biomass, respectively (Figure 4F; Supplementary Table 12). When LLoQ was considered, we confidently quantified spiked bacteria representing 0.011% and 0.025% for *L.Murinus* and *S.Ruber*, respectively (Figure 4F). Based on genomic calculations[33] (predicting 1 e^{+12} bacteria in 1 gr of mouse feces), we reached the ground-breaking LLoD of $1e^{-4}$ % (or 1 in 1 million) for both bacteria and an LLoQ of $1e^{-3}$ % and $5e^{-3}$ % for *L.Murinus* and *S.Ruber*, respectively. Our data positively compared to some of the deepest profiling efforts using full-length 16S rRNA[27] and improved up to 5,000 times previous state-of-the-art in metaproteomics[6].

363 Furthermore, we investigated the maintenance of the functional profile along the spike-in
364 dilution series. Annotation of identified protein groups to KEGG pathways decreased notably
365 after 100 million and plateaued after 1 million spiked bacteria (Figure 4G; Supplementary Table
366 13). Notably, we quantified 47 and 27 functional pathways using as low as 10000 *L.Murinus* and
367 *S.Ruber*, including diverse metabolic and biosynthesis pathways (Supplementary Table 13).

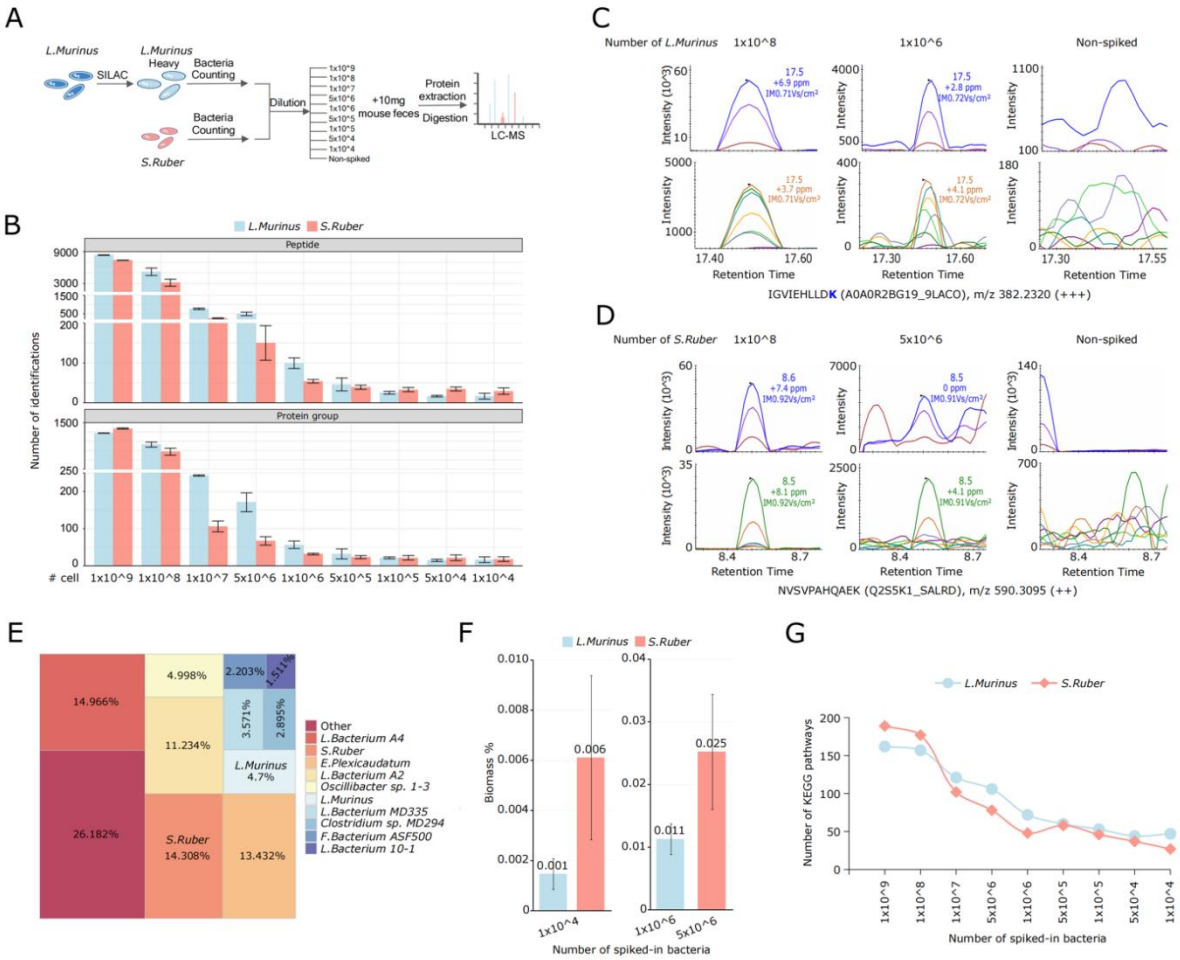


Figure 4. Unveiling the real limits of detection and quantification in the dark metaproteome. A) Experimental workflow of uMetaP analysis with stable isotope labelling (SILAC) of *L.Murinus* and *S.Ruber* to assess the sensitivity limits in the dark metaproteome. Spike-in triplicates were prepared for each spike-in bacterial amount. B) Identification summary of identified peptides and protein groups cross varying numbers of spiked-in *L.Murinus* and *S.Ruber* (see Methods for detailed filtering steps). Bars show the average identifications of triplicates, and error bars indicate the standard deviations across the triplicates. C-D) Examples of extracted precursors (upper panels) and ion fragments (lower panels) for spiked-in *L.Murinus* (C) and *S.Ruber* (D) at different spike-in bacterial amounts and a non-spiked control. Blue-labelled K (lysine) represents the heavy isotope labelling. Inserted info indicates the retention time, mass shift, and ion mobility (from the top to bottom). E) Example of relative biomass abundance assessment by uMetaP with 1 x 10^8 spiked-in bacteria. F) Calculation of relative biomass abundance using identified species-specific peptides considering LLoD (1 x 10^4 bacteria; left) and LLoQ (1 x 10^6 bacteria for *L.Murinus* and 5 x 10^6 bacteria for *S.Ruber*; right) G) Investigation of functional profile (KEGG pathways) maintenance across spiked-in dilutions.

Pushing the sensitivity limits to quantify biology at single-bacterium resolution.

While proteotyping at single-eukaryotic cell level inputs (250 picograms[34]) is possible, the challenge of detecting single bacterium peptide amounts (sub-picogram levels) hinders our ability to explore bacterial diversity and its connection to health and disease[34].

Encourage by the incredible sensitivity of uMetaP in complex microbiomes, we set to calculate the sensitivity limit of uMetaP using a dilution series of a two-proteome mix (SILAC-labelled *L.Murinus* and non-labelled *S.Ruber*). To ensure an unambiguous identification, we applied a filter to exclude heavy-labeled peptides assigned to *S.Ruber* (therefore, false identifications; see Methods). Furthermore, we did not consider non-species-specific peptides (the same peptides assigned to both species). As shown in Figure 5A, uMetaP confidently identified peptides and proteins in a whole dilution series ranging down to 500 fg of injected peptides (343 peptides and 93 protein groups for *L.Murinus*, 288 peptides and 91 protein groups for *S.Ruber*; Supplementary Table 14). Notably, as little as 50 pg enabled the identification of 56.9% and 58.5% of the complete proteome for *L.Murinus* and *S.Ruber*, respectively (1122 and 1644 proteins). These metrics compare favorably with the coverage reached by the analysis of 50 pg yeast samples in the Orbitrap Astral[35] (1782 protein groups, corresponding to 26.5% of the whole proteome). A comparative analysis of the spectra demonstrates the quality of these identifications at sub-picogram levels (Figure 5B-C).

From the quantitative point of view, we observed a decrease in the quantitative precision from 79.1% to 48.5% (*L.Murinus* peptides with a CV<0.2 at 250 pg and 500 fg, respectively; Figure 5D; Supplementary Table 14). Quantitative accuracy was calculated using the ratio between the median ratios of all common peptides along the dilution series (peptide intensities of 250 pg were set as reference) and the theoretical ratios between the dilutions. We observed a minimal accuracy decrease until 25 pg (1.0 to 1.2 from 250pg to 25 pg) which augmented with further peptide dilutions (Figure 5E; Supplementary Table 15).

These results demonstrate that uMetaP can reliably identify species-specific peptides with favourable quantification precision and accuracy at unprecedented ultra-low sample amounts, paving the way towards single-bacterium proteomics (SBP).

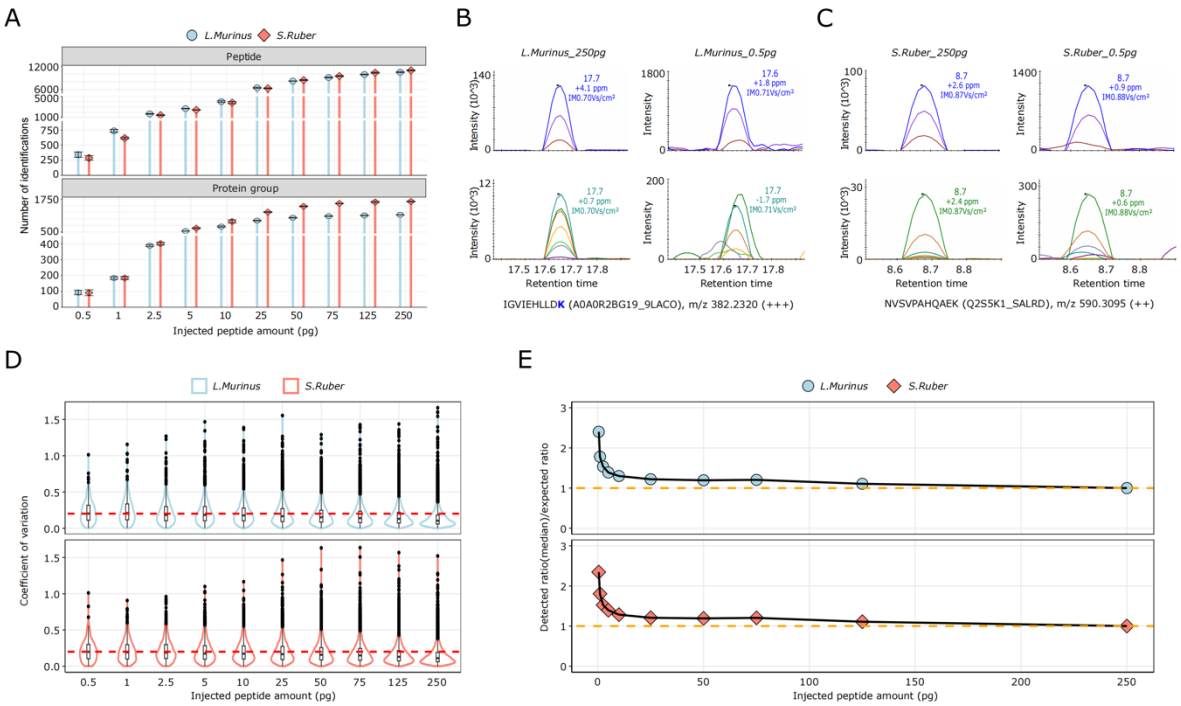


Figure 5. Pushing the sensitivity limits to quantify biology at single-bacterium resolution. A) Number of identified peptides and protein groups in a dilution series (ranging from 0.5 to 250 pg) of a two-proteome mix (SILAC-labeled *L.Murinus* and non-labeled *S.Ruber*). Symbols (circle and rhomboid) indicated the average identifications and the error bars show the standard deviations across injection triplicates. B-C) Representative extracted precursors (upper panels) and ion fragments (lower panels; same peptides as shown in Figure 4) obtained from uMetaP's analysis of 500 fg in comparison with 250 pg injected peptide for *L.Murinus* (B) and *S.Ruber* (C). Blue-labelled K (lysine) represents the heavy isotope labelling. D) Quantitative precision across the dilution series measured by the coefficient of variation (CV) of all peptides detected among triplicates. The red-dotted lines indicate the CV = 0.2. E) Quantitative accuracy across the dilution series, measured as the median ratio of commonly identified peptides across the whole dilution series (Detected ratio; reference intensity at 250 pg) to the theoretical dilution factor (expected ratio). The yellow-dotted lines indicate the perfect accuracy of 1.

DISCUSSION

Current metaproteomic methods fall short of achieving the proteotyping performance of single-organism proteomics and of capturing the taxonomic depth offered by the widely used 16S rRNA approaches. The existing methods struggle to simultaneously achieve high sensitivity, deep protein profiling, quantitative accuracy, precision, and high sample throughput. uMetaP offers a solution by addressing these limitations.

We demonstrated an impressive performance using DDA-PASEF on pH-fractionated mouse fecal samples. Over 1.5 million charged precursors were sent for fragmentation, resulting in 450,000 identified PSMs in all eight fractions. We expect DeNovo algorithms to further enhance this by deciphering unassigned spectral data, leading to more comprehensive metaproteomic databases[36].

uMetaP, powered by DIA-PASEF, represents a groundbreaking advancement in metaproteomics for protein identification. It surpasses current standards[6], the most optimistic performance forecast in the field [37], and a preliminary evaluation of the Orbitrap Astral mass spectrometer[38]. uMetaP offers additional advantages: fast analysis times (30-min) and low variability. These features make it highly promising for large-scale metaproteomic studies. Beyond solely proteotyping, uMetaP demonstrates exceptional precision and reliability in quantification, along with ultra-high sensitivity. This unique combination caters to various project needs. For samples with ample quantity, it offers comprehensive protein coverage. For limited samples (e.g., ocean metaproteomics[39]), it can provide meaningful taxonomic and functional profiling. On the taxonomic front, uMetaP achieves significantly deeper analysis than previously possible. It surpasses the typical coverage of 16S rRNA experiments[40] and even rivals the species detection capabilities of full-length 16S rRNA used in human gut microbiome studies[27]. The use of a SILAC-labelled bacterium and an exogenous bacterium (accounting for sample preparation losses), establishes the first reliable LLoD and LLoQ in a complex metaproteome. Remarkably, uMetaP can detect a single bacterium in a theoretical pool of 1 million, representing a 5000-fold improvement in investigating the hidden dark metaproteome[6]. This exceptional sensitivity opens doors for early diagnosis by detecting changes in the abundance of disease-causing bacteria. Importantly, a closer look at the MS2 spectra indicated a clear shift in the reliable quantification limit for *L. Murinus* and *S. Ruber*. This finding likely applies across metaproteomes, highlighting the importance of rigorous spectral quality control for accurate peptide-based quantification.

From a functional point of view, uMetaP's focus on identifying medium and low-abundance proteins unlocks a new world of biological information. Our curated metaproteomic database allows researchers to study thousands of previously undetectable PUFs. By applying advanced machine learning techniques[41], we can now begin to understand their potential functions, a significant gap in our current knowledge of the metaproteome. Furthermore, uMetaP's enables the discovery of hundreds of sProts and potential AMPs without the need for additional enrichment steps[11]. We expect these resources to increase our understanding of the role of these proteins in regulating the balance within microbial ecosystems.

To the best of our knowledge, we offer the first assessment of detection, quantitative accuracy, and precision at the single-bacterium resolution. Detailed analysis revealed a decrease in quantitative accuracy for peptide injections below 25 pg. The combination of two factors might explain this decrease: errors introduced while creating the dilute samples, and/or technical limitations in detecting such low signal levels. We believe the first factor to parallel the expected losses during future SBP sample preparation protocols, highlighting the need for further improvements in current methods[34]. Beyond its role in paving the way for SBP, uMetaP excels at analyzing extremely small sample amounts as shown by the higher whole proteome coverage for *L.Murinus* and *S.Ruber* compared to the one obtained using the same amount of injected yeast peptides (50 pg) in an Orbitrap Astral mass spectrometer[35]. This unprecedented sensitivity has the potential to shed light on the biochemical resources hidden in the dark field of ultra-low abundant microbiomes.

This study introduces a significant breakthrough in microbiome research with uMetaP. The leap forward in taxonomic and functional capabilities paves the way for a new understanding of the host-microbiome interactions and their role in health and disease.

AUTHOR CONTRIBUTIONS

Conceptualization: D.G.V.; experimental design: D.G.V. and F.X.; biochemistry and mass spectrometry: F.X. and C.K.; sample collection and preparation: F.X., M.B. and R.K.; data analysis: F.X. and D.G.V.; writing initial draft: D.G.V. and F.X. All authors edited and approved the final manuscript; supervision: D.G.V.; project administration: D.G.V. and M.S.; Funding acquisition: D.G.V. and M.S.

NOTES

D.G.V, F.X., M.B, R. K., and M.S. declare no competing financial interest. C.K is an employee of Bruker Daltonics GmbH & Co. KG.

ETHICS

All animal experiments were carried out in strict accordance with institutional IACUC guidelines, international ARRIVE guidelines, and the principles of the 3Rs of animal research in accordance with the BMBWF (Federal Ministry of Education, Science and Research, Austria). All experimental procedures were approved by the BMBWF (GZ 2020-0.592.919).

ACKNOWLEDGMENTS

We thank Elisabeth Haberleitner (the Division of Pharmacology & Toxicology, University of Vienna, Austria) for assistance during sample preparation. We thank Astrid Horn Ph.D. (Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, Austria) for suggestions during bacterial counting. This work is funded by a Research grant from the FWF (P35856-B to MS) and the University of Vienna.

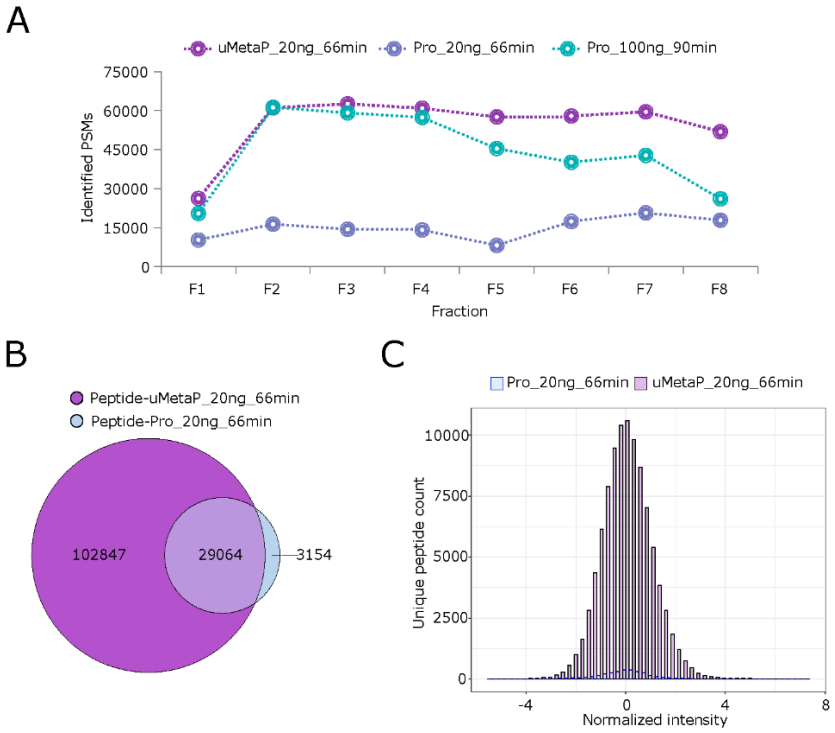
REFERENCES

1. Prescott, S.L., Wegienka, G., Logan, A.C. & Katz, D.L. Dysbiotic drift and biopsychosocial medicine: how the microbiome links personal, public and planetary health. *Biopsychosoc Med* **12**, 7 (2018).
2. Clemente, J.C., Ursell, L.K., Parfrey, L.W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**, 1258-1270 (2012).
3. Van Den Bossche, T. et al. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat Commun* **12**, 7305 (2021).
4. Stamboulian, M., Canderan, J. & Ye, Y. Metaproteomics as a tool for studying the protein landscape of human-gut bacterial species. *PLoS Comput Biol* **18**, e1009397 (2022).
5. Armengaud, J. Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. *Environ Microbiol* **25**, 115-125 (2023).
6. Duan, H. et al. Assessing the Dark Field of Metaproteome. *Anal Chem* **94**, 15648-15654 (2022).
7. Chapman, J.D., Edgar, J.S., Goodlett, D.R. & Goo, Y.A. Use of captive spray ionization to increase throughput of the data-independent acquisition technique PAcIFIC. *Rapid Commun Mass Spectrom* **30**, 1101-1107 (2016).
8. Brunner, A.D. et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol Syst Biol* **18**, e10798 (2022).
9. Ctortecka, C. et al. Automated single-cell proteomics providing sufficient proteome depth to study complex biology beyond cell type classifications. *bioRxiv* (2024).
10. Paik, Y.K. et al. Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. *J Proteome Res* **17**, 4042-4050 (2018).
11. Petruschke, H., Anders, J., Stadler, P.F., Jehmlich, N. & von Bergen, M. Enrichment and identification of small proteins in a simplified human gut microbiome. *J Proteomics* **213**, 103604 (2020).
12. Sberro, H. et al. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* **178**, 1245-1259 e1214 (2019).
13. Ma, Y. et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol* **40**, 921-931 (2022).
14. Gomez-Varela, D. et al. Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. *Front Microbiol* **14**, 1258703 (2023).
15. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. & Nesvizhskii, A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520 (2017).
16. Xiao, L. et al. A catalog of the mouse gut metagenome. *Nat Biotechnol* **33**, 1103-1108 (2015).
17. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923-925 (2007).

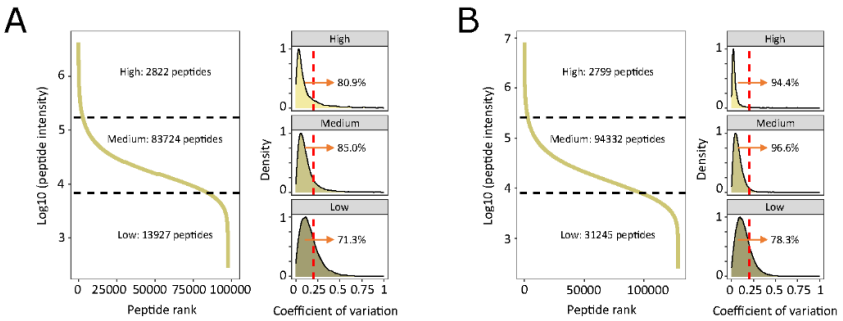
18. Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41-44 (2020).
19. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526 (2014).
20. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010).
21. Cheng, K. et al. MetaLab 2.0 Enables Accurate Post-Translational Modifications Profiling in Metaproteomics. *J Am Soc Mass Spectrom* **31**, 1473-1482 (2020).
22. Cantalapiedra, C.P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829 (2021).
23. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
24. Jagtap, P. et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**, 1352-1357 (2013).
25. Blakeley-Ruiz, J.A. & Kleiner, M. Considerations for constructing a protein sequence database for metaproteomics. *Comput Struct Biotechnol J* **20**, 937-952 (2022).
26. Gómez-Varela, D. et al. Expanding the Characterization of Microbial Ecosystems using DIA-PASEF Metaproteomics. *bioRxiv*, 2023.2003.2016.532922 (2023).
27. Yang, J. et al. Species-Level Analysis of Human Gut Microbiota With Metataxonomics. *Front Microbiol* **11**, 2029 (2020).
28. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).
29. Martinez, I., Muller, C.E. & Walter, J. Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. *PLoS One* **8**, e69621 (2013).
30. Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).
31. Soufi, B. & Macek, B. Stable isotope labeling by amino acids applied to bacterial cell culture. *Methods Mol Biol* **1188**, 9-22 (2014).
32. Kleiner, M. et al. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat Commun* **8**, 1558 (2017).
33. Taibi, A. et al. Data on cecal and fecal microbiota and predicted metagenomes profiles of female mice receiving whole flaxseed or its oil and secoisolariciresinol diglucoside components. *Data Brief* **38**, 107409 (2021).
34. Gatto, L. et al. Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments. *Nature Methods* **20**, 375-386 (2023).
35. Bubis, J.A. et al. Challenging the Astral™ mass analyzer- up to 5300 proteins per single-cell at unseen quantitative accuracy to study cellular heterogeneity. *bioRxiv*, 2024.2002.2001.578358 (2024).
36. Kleikamp, H.B.C. et al. Database-independent de novo metaproteomics of complex microbial communities. *Cell Syst* **12**, 375-383 e375 (2021).

37. Lohmann, P. et al. Function is what counts: how microbial community complexity affects species, proteome and pathway coverage in metaproteomics. *Expert Rev Proteomics* **17**, 163-173 (2020).
38. Dumas, T. et al. The astounding exhaustiveness and speed of the Astral mass analyzer for highly complex samples is a quantum leap in the functional analysis of microbiomes. *Microbiome* **12**, 46 (2024).
39. Sabine, M.-S., Pratik, D.J., Timothy, J.G., Mélanie, B. & Ruddy, W. in Metagenomics. (ed. N. Muniyandi) 327-363 (Academic Press, 2018).
40. Wang, J. et al. Core Gut Bacteria Analysis of Healthy Mice. *Front Microbiol* **10**, 887 (2019).
41. Tavis, S. & Hettich, R.L. Multi-Omics integration can be used to rescue metabolic information for some of the dark region of the Pseudomonas putida proteome. *BMC Genomics* **25**, 267 (2024).

SUPPLEMENTARY FIGURES

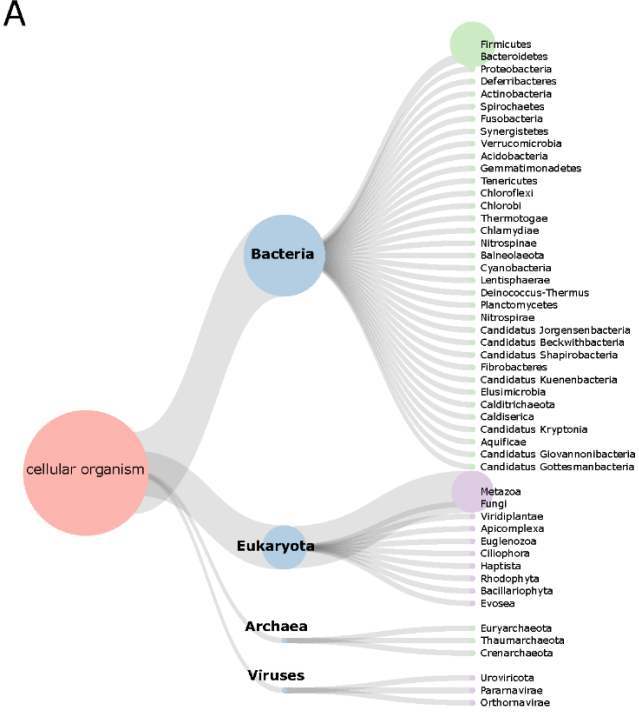


Supplementary Figure 1. Comparative performance for metaproteomic database generation between uMetaP and timsTOF Pro using DDA-PASEF. A) Peptide-spectrum matches (PSMs) obtained by uMetaP compared to timsTOF Pro in a pH-fractionated mouse fecal sample. B) Comparison of total microbial peptides identified by uMetaP and timsTOF Pro. C) Intensity distribution of uniquely identified peptides using uMetaP and timsTOF Pro.

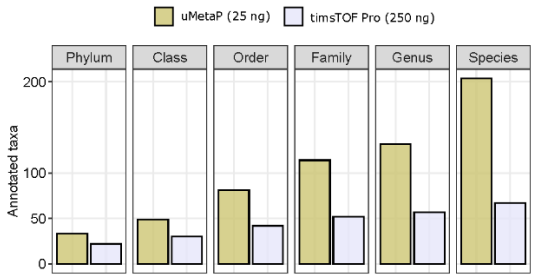


Supplementary Figure 2. Assessment of quantitative precision and reliability. Left: Intensity rank of peptides identified when 50ng of peptides were analyzed in a 30 min gradient (A) or a 66 min gradient (B). Right: coefficient of variations across different abundance regions delimited by the dotted lines in the intensity rank graph. The red-dotted lines indicate CV = 0.2.

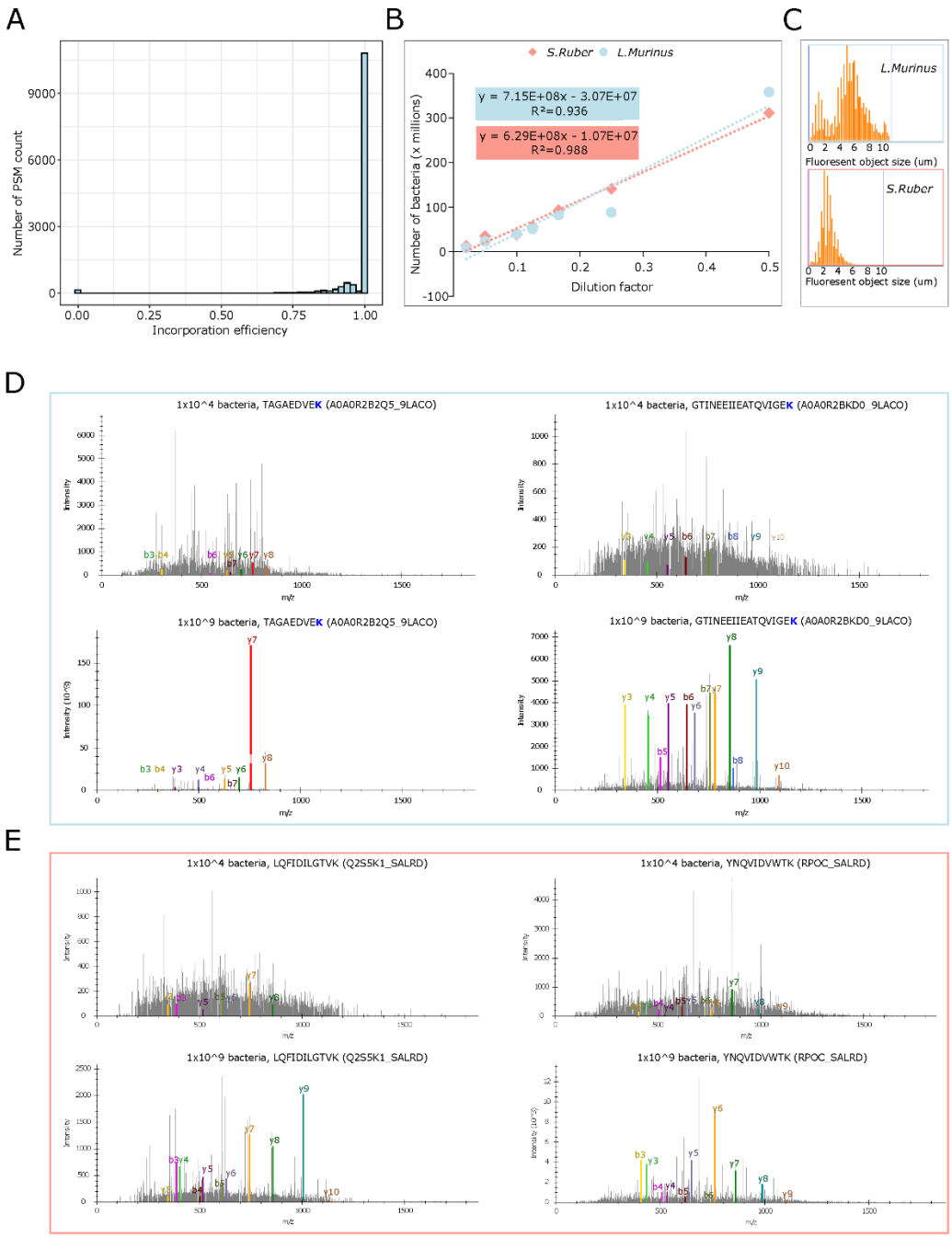
A



B



Supplementary Figure 3. Enhanced taxonomic and functional annotation with uMetaP. A) Representation of annotated taxa covering 4 major kingdoms in the mouse gut microbiome (an example using taxa annotation from 25 ng in a 30 min gradient). B) Taxonomic profiling of uMetaP compared with the previous DIA-PASEF published state-of-the-art dataset (Gomez-Varela et. al.).



Supplementary Figure 4. Peptide detection at single-bacterium resolution. A) Assessment of SILAC incorporation efficiency for *L.Murinus* using DDA-PASEF analysis (see Methods for details). B) Accuracy of bacterial counting using QUANTOM Tx Microbial Cell Counter and regression analysis of different dilutions of *L.Murinus* and *S.Ruber*. C) Example of the bacteria size distribution detected with QUANTOM Tx Microbial Cell Counter. Blue lines mark the size threshold used during automatic counting. D-E) Examples of the same peptide spectrum at LLoD (1 x 10^4 bacteria) and 1 x 10^9 bacteria for *L.Murinus* (D) and *S.Ruber* (E). Blue-labelled K (lysine) represents the heavy isotope labeling.