Ultrafast metaproteomics for quantitative assessment of strain isolates and microbiomes

Elizaveta Kazakova[1], Mark Ivanov[1], Tomiris Kusainova[1], Julia Bubis[1], Valentina Polivtseva[2], Kirill Petrikov[2], Vladimir Gorshkov[3], Frank Kjeldsen[3], Mikhail Gorshkov[1], Yanina Delegan[2], Inna Solyanikova[2,4] and Irina Tarasova[1*]

[1]V. L. Talrose Institute for Energy Problems of Chemical Physics, N. N. Semenov Federal Research Center of Chemical Physics, Russian Academy of Sciences, 119334 Moscow, Russia;
[2]Skryabin Institute of Biochemistry and Physiology of Microorganisms, Pushchino Scientific Center for Biological Research, Russian Academy of Sciences, 142290 Pushchino, Russia;
[3]Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark;
[4]Belgorod National Research University, 308015 Belgorod, Russia

*Correspondence: iatarasova@yandex.ru; 38 Leninsky pr., bld 2, 119334 Moscow, Russia

ABSTRACT

BACKGROUND: Microbial communities play a crucial role in human health and environmental regulation, but present an especial challenge for the analytical science due to their diversity and dynamic range. Tandem mass spectrometry provides functional insights on microorganisms life cycle, but still lacks throughput and sensitivity. MALDI TOF is widely used for ultrafast identification of species, but does not assess their functional activity. Development of ultrafast mass spectrometry methods and bioinformatic approaches applicable for both accurate identification and functional assessment of microbial communities based on their protein content is of high interest.

RESULTS: We show for the first time that both identity and functional activity of microorganisms and their communities can be accurately determined in experiments as short as 7 minutes per sample, using the basic Orbitrap MS configuration without peptide fragmentation. The approach was validated using strain isolates, mock microbiomes composed of bacteria spiked at known concentrations and human fecal microbiomes. Our new bioinformatic algorithm identifies the bacterial species with an accuracy of 95 %, when no prior information on the sample is available. Microbiome composition was resolved at the genus level with the mean difference between the actual and identified components of 12 %. For mock microbiomes, Pearson coefficient of up to 0.97 was achieved in estimates of strain biomass change. By the example of *Rhodococcus* biodegradation of *n*-alkanes, phenols and its derivatives, we showed the accurate assessment of functional activity of strain isolates, compared with the standard label-free and label-based approaches.

SIGNIFICANCE: Our approach makes microbial proteomics fast, functional and insightful using the Orbitrap instruments even without employing peptide fragmentation technology. The approach can be applied to any microorganisms and can take a niche in routine functional assessment of microbial pathogens and consortiums in clinical diagnostics together with MALDI TOF MS and 16S rRNA gene sequencing.

KEYWORDS
Microbiological methods, bacteria, microbiome, metaproteomics, proteomics, metabolic pathways, biodegradation, LC-MS/MS, MALDI TOF

Introduction

Metaproteomics is a growing area of research[1] devoted to monitoring the physiology of biological communities, for example, microbial ones, at the level of their proteomes[2]. The metaproteomic studies are focused on determining the species diversity in a community, studying various stages of the life cycle of microorganisms, their response to stress, the functional characterization of microbiomes, etc[3]. Being a young field, metaproteomics possess challenges that must be overcome before using for routine research[2,4]. Currently, the research efforts are focused on accuracy, sensitivity, and reproducibility of the analysis, development of bioinformatic solutions, and standardization of workflows[5].

Proteomic analysis of microbiomes relies on liquid chromatography tandem mass spectrometry (LC-MS/MS) that allows the identification of thousands of protein molecules in a single experiment. A common approach is data-dependent acquisition (DDA) mass spectrometry, in which the success of peptide identification strongly depends on the choice of the most abundant ions. Reportedly, the performance of DDA decreases with increasing the sample complexity[6]. The time-consuming DDA analysis with insufficient peptide fragmentation and protein sequence coverage, and heavy bioinformatic processing of metaproteomic data still present a gap for more efficient solutions. Another acquisition method recently applied in metaproteomics is data independent acquisition (DIA)[7,8,9]. DIA allows collecting fragment spectra for all peptide ions without pre-selection of a fixed number of top abundant ones, and, thus, promises better sequence coverage and protein quantitation[10]. The third approach can be the use of short chromatographic gradients with acquisition of only MS1 spectra without fragmentation. It has been shown that interpretation of LC-MS1 data based on alignments of peptide feature intensities without peptide and protein identification allows large-scale screening of microbiomes and characterization of changes in MS1 profiles, depending on drug treatment[11]. Widely implemented in clinics, MALDI-TOF MS allows identification of the genus and species by matching mass spectra against spectral libraries of known microorganisms[12], but the strain resolution with this technique is problematic[13]. Identifying bacteria by matching the experimental LC-MS1 data against *in silico* generated taxon-specific tryptic peptide masses compiled based on UniProt, SwissProt, and TrEMBL databases, can resolve species and, sometimes, strain identity[14]. However, the functional characterization of microbiomes still cannot be assessed without the identification of proteins.

Identification of proteins relies on protein databases which represents one of the hardest issues in MS-based metaproteomics due to its huge size. The choice of the protein database affects the sensitivity of peptide identification because the larger the search space, the higher the chance of random and incorrect matches. The optimal database should only contain proteins that are present in the sample[15]. From this viewpoint, the protein database constructed using metagenome data or restricted using 16S rRNA gene analysis is the best. However, they are costly and require large sample amounts, as it is important to obtain

high-quality sequences of the genome. Attempts to fix the FDR problem typically include selecting top organisms or peptide sequences within a certain window of *m/z* values followed by repeating the search using the restricted database[16,17,18,19,20,21] . The MetaProteomeAnalyzer[22] implements a search with four different search engines and the integration of the results. Following the search, the identified proteins are grouped into so-called meta-proteins and annotated with protein-level information.

This study presents an adaptation of the novel ultrafast MS1-based method called DirectMS1[23,24] to the characterization of single microorganisms and microbiomes. Using DirectMS1 we demonstrate the blind identification (with no prior knowledge of the bacteria in the sample) of a single bacterial species and microbial compositions using two-step database search against the microbial part of SwissProt and TrEMBL (62Gb database size), the evaluation of fold changes in biomass of microorganisms between the metaproteomic samples, and the assessment of metabolic activity of *Rhodococcus* species responding to a change in growth conditions and cold shock.

Methods

Public datasets

Public dataset (https://zenodo.org/record/3573994) collected in DDA mode for 19 well-characterized bacterial strains[14] was used to validate the two-step search algorithm for blind identification of individual microorganisms against the microbial part of SwissProt and TrEMBL database (accessed on Jan 2020).

Public dataset (PRJEB42466) for human fecal microbiome, collected in DDA mode, was used for the testing of the blind search algorithm to identify composition of human fecal microbiome[5].

Strains, genomes and proteogenomic databases

Strains used in the study were isolated from different sources and characterized as previously described (*Rhodococcus erythropolis* X5[25], *Rhodococcus opacus* S8[26], *Rhodococcus qingshengii* 7B[27], *Rhodococcus qingshengii* VT6[28], *Priestia aryabhattai* 25[29], *Gordonia amicalis* 6-1[30], *Gordonia alkanivorans* 135[31,32], *Gordonia polyisoprenivorans* 135[33], *Rhodococcus opacus* 1CP[34], *Rhodococcus opacus* 3D[35]).

To construct protein databases using genome data, NCBI PGAP 6.6[36] was used. Annotations of proteins of interest were manually checked using UniProt[37].

Cell cultivation to characterize metabolic activity changes in response to stress

R. opacus 1CP was cultivated in the presence of four different carbon sources (glucose, benzoate, phenol, 4-chlorophenol) with six biological replicates per condition. The cultivation

was carried out in 750-mL Erlenmeyer flasks with 200 mL of the mineral medium of the following composition (g/L): Na_2HPO_4 - 0.7, K_2HPO_4 - 0.5, NH_4NO_3 - 0.75, $MgSO_4 \times 7H_2O$ - 0.2, $MnSO_4$ - 0.001, $FeSO_4$ - 0.02 with the addition of one of the following sources of carbon: phenol 500 mg/L, benzoate 500 mg/L, 4-chlorophenol 100 mg/L, glucose 10 g/L. The cultivation was performed for 48-72 hours in a shaking incubator (HZQ-111) at 24℃ and 220 rpm. During cultivation time, the carbon sources (phenol and 4-chlorophenol) were periodically injected into the medium after consumption of previously added. The disappearance of substrates was determined by characteristic absorption spectra in the 230-290 nm region. Culture growth was monitored with an optical absorption coefficient at 560 nm.

R. erythropolis X5 cells were grown in an orbital shaking incubator at 180 rpm at two different temperatures (28℃ and 6℃) for 3 and 11 days, respectively, with three biological replicates. The cultivation was carried out in 750 mL Erlenmeyer flasks with 100 mL of the modified Evans mineral salts medium supplemented with 2 mL of *n*-hexadecane. The composition of the cultivation medium was as follows (per liter): K_2HPO_4, 8.71 g; 5 M solution of NH_4Cl, 1 mL; 0.1 M solution of Na_2SO_4, 1 mL; 62 mM solution of $MgCl_2$, 1 mL; 1 mM solution of $CaCl_2$, 1 mL; 0.005 mM solution of $(NH_4)_6Mo_7O_{24}$, 1 mL; trace element solution, 1 mL. The value of pH was adjusted to 7.5 by concentrated HCl. Composition of the trace element solution in 1 % water solution of HCl (g/L): ZnO, 0.41 g; $FeCl_3$, 2.9 g; $MnCl_2$, 1.28 g; $CuCl_2$, 0.13 g; $CoCl_2$, 0.26 g; H_3BO_3, 0.06 g.

Cell cultivation for model microbiomes.

Cells were cultivated for 72 hours in a shaking incubator (HZQ-111) at 27°C and 250 rpm in Erlenmeyer flasks with 100 mL of Evans medium. The composition of the medium was as follows (g/L or mL/L:): K_2HPO_4 8.71 g, 5 M NH_4Cl 1 mL, 0.1 M Na_2SO_4 1 mL, 62 mM $MgCl_2$ 1 mL, 1 mM $CaCl_2$ 1 mL, 0.005 mM $(NH_4)_6Mo_7O_{24} \times 4H_2O$, micronutrients solution 1 mL (as follows in g/L): ZnO 0.41 g, $FeCl_3 \times 6H_2O$ 5.4 g, $MnCl_2 \times 4H_2O$ 2 g, $CuCl_2 \times 2H_2O$ 0.17 g, $CoCl_2 \times 6H_2O$ 0.48 g, H_3BO_3 0.06 g, (pH 7.0) - with the addition of 200 uL 10^8 CFU inoculum and glucose (1 % v/v for *R. opacus* 1CP and 2 % v/v for *R. erythropolis* X5, *G. alkanivorans* 135, *G. amicalis*, *P. aryabhattai* 25) as the sole source of carbon and energy.

Model microbiome compositions

Model microbiomes were prepared by mixing tryptic peptide solutions to test the performance of DirectMS1 method. The composition of each sample is described in **Tables 1-3**.

Table 1. Model I: five low fold-change mixtures of peptides (ng) from 3 to 5 strains, total of 500 ng of the sample per injection.

Strain	M1.1	M1.2	M1.3	M1.4	M1.5
Rhodococcus opacus 1CP	167	250	167	100	71
Rhodococcus erythropolis X5	-	-	-	100	143
Priestia aryabhattai 25	167	125	250	100	71
Gordonia amicalis 6-1	-	-	-	100	143
Gordonia alkanivorans 135	167	125	83	100	71

Table 2. Model II: three mixtures of peptides (ng) from 8 strains, total of 1 ug of the sample per injection.

Strain	M2.1	M2.2	M2.3
Rhodococcus opacus 1CP	16	125	16
Rhodococcus opacus 3D	63	8	31
Rhodococcus opacus S8	251	63	4
Rhodococcus qingshengii 7B	31	502	251
Rhodococcus qingshengii VT6	125	251	502
Rhodococcus erythropolis X5	4	16	125
Gordonia alkanivorans 135	8	4	63
Gordonia polyisoprenivorans 135	502	31	8

Table 3. Model III: four ABRF-like mixtures of peptides (ng) from five strains, total of 1 ug of the sample per injection. Benchmark design was suggested earlier[38].

Strain	ABRF1	ABRF2	ABRF3	ABRF4
Rhodococcus opacus 1CP	417	96	146	3
Rhodococcus erythropolis X5	353	26	19	104
Priestia aryabhattai 25	96	3	631	88
Gordonia amicalis 6-1	71	1	97	803
Gordonia alkanivorans 135	64	873	107	1

Sample preparation for mass spectrometry-based proteomics

For mix models I and III (Tables 1 and 3), cell pellets were resuspended in 150 µL lysis buffer (50 mM ammonium bicarbonate (ABC), 10 % ACN, 0.1 % ProteaseMax (Promega, USA)). Cells incubated for 45 min at room temperature, after that they were boiled for 10 min at 95 °C, and then cooled down on ice. To lyse cells, the samples were sonicated for 10 min (1 s on 1 s off) on 60 % amplitude (QSonica Q125, Newtown, Connecticut, USA). Samples were centrifuged for 7 min on 10 000 g, then supernatant was taken, and protein concentration was measured by BCA kit (Thermo Scientific, Germany). 10 mM dithiothreitol (DTT) was added for 20 min at 56 °C, followed by incubation with iodoacetamide (IAA) for 30 min in the dark at room temperature. Lys-c was added in 1:100 m/m ratio and samples incubated for 2h at 37 °C then trypsin was added in 1:50 m/m ratio and samples incubated overnight at 37 °C. Digest was stopped by adding 1 % trifluoroacetic acid (TFA), then samples were desalted with OASIS HLB cartridges (Waters, USA) and dried in a vacuum concentrator.

For mix model II (Table 2) and characterization of metabolic activity changes, cell pellets were resuspended in a 150 µL lysis buffer (100 mM ABC, 4 % SDS, 10 mM DTT). Cells were boiled for 10 min at 95°C and then cooled down on ice. To lyse cells, the samples were sonicated for 10 min (1 s on 1 s off) on 60 % amplitude (QSonica Q125, Newtown, Connecticut, USA). Samples were centrifuged for 10 min on 10 000 g, then the supernatant was taken. 100 uL of supernatant was purified and concentrated with chloroform-methanol precipitation. Dry pellets were resuspended in 50 uL of 4 M urea in 50 mM ABC, urea concentration was further reduced to 1 M with 50 mM ABC. Protein concentration was measured by a BCA kit (Thermo Scientific, Germany). 10 mM DTT was added for 20 min at 56 °C, followed by incubation with IAA for 30 min in the dark at room temperature. Trypsin was added in a 1:50 m/m ratio and samples were incubated overnight at 37 °C. Digest was stopped by adding 1 % TFA, then samples were desalted with OASIS HLB cartridges (Waters, USA) and dried in a vacuum concentrator.

LC-MS1 data acquisition

The LC-MS experiments were performed using Orbitrap Q Exactive HF-X mass spectrometer (Thermo Scientific, San Jose, CA, USA) coupled to UltiMate 3000 LC system (Thermo Fisher Scientific, Germering, Germany) and Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific, San Jose, CA, USA) coupled to UltiMate 3000 LC system (Thermo Fisher Scientific, Germering, Germany) and equipped with FAIMS Pro interface. Trap column µ-Precolumn C18 PepMap100 (5 µm, 300 µm, i.d. 5 mm, 100 Å) (Thermo Fisher Scientific, USA) and self-packed analytical column (Reprosil-Pur 3 µm, 75 µm i.d., 5 cm length) were employed for separation. Mobile phases were as follows: (A) 0.1 % formic acid (FA) in water; (B) 80 % ACN, 0.1 % FA in water. The gradient was from 5 % to 35 % phase B in 4.8 min at 1500 nL/min. Total method time including column washing and equilibration was 7.5 min. Field asymmetric ion

mobility spectrometry (FAIMS) separations were performed with the following compensation voltages (CV) -50 V, -65 V, and -80 V in a stepwise mode during LC-MS analysis. Data acquisition was performed in MS1-only mode. Full MS scans were acquired in a range from *m/z* 375 to 1500 at a resolution of 120,000 at *m/z* 200 with AGC target of 4e5, 1 microscan, and 50 ms maximum injection time. Samples were resuspended in phase A and quantities of 1 μg were loaded per injection.

LC-MS2, data dependent acquisition

MS/MS-based analysis of samples was performed using Orbitrap Lumos Fusion mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) and Orbitrap Q Exactive HF mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to UltiMate 3000 LC system. Trap column µ-Precolumn C18 PepMap100 (5 µm, 300 µm, i.d. 5 mm, 100 Å) (Thermo Fisher Scientific, USA) and self-packed analytical column (Inertsil 2 µm, 75 µm i.d., 25 cm length) were employed for separation. Mobile phases were as follows: (A) 0.1 % FA in water; (B) 95 % ACN, 0.1 % FA in water.

For TMT-based quantitation of *R. erythropolis* X5, the gradient from 5 % to 30 % phase B for 114 min at a flow rate of 300 nL/min was used. Data was acquired in top20 mode. Full MS scans were acquired in a range from m/z 300 to 1400 at a resolution of 60000 at m/z 200 with AGC (Automatic Gain Control) target of 3e6, 1 microscan, and 50 ms maximum injection time. Precursor ions were isolated in a 1.4 m/z window and accumulated for a maximum of 100 ms or until the 1e5 AGC target was reached.

For label free quantitation of *R. erythropolis* X5, the gradient from 2 % to 40 % phase B for 60 min at a flow rate of 300 nL/min was used. Full MS scans were acquired in a range from m/z 300 to 1400 at a resolution of 60000 at m/z 200 with AGC target of 3e6, 1 microscan, and 50 ms maximum injection time. Precursor ions were isolated in a 1.4 m/z window and accumulated for a maximum of 50 ms or until the AGC target of 1e5 charges was reached. Precursors of charge states from 2+ to 6+ (inclusive) were scheduled for fragmentation. To save instrument time for label free DDA analysis, biological replicates were pooled and each pooled sample was measured in triplicate.

Data processing

Peptide feature detection was performed using the Biosaur2 software[39]. The proteomic search engine ms1searchpy[24] was used for protein identification. Searches were performed against the SwissProt + TrEMBL database (accessed on 06 Dec 2020) of bacterial proteins and strain-specific databases based on annotations of bacterial genomes. **Table 4** summarizes the origin of databases used for every type of sample. Mass tolerance for precursors in all data was ±10 ppm, fragment mass tolerance for DDA data was 0.01 Da. Carbamidomethylation of cysteine residues was fixed modification, no variable modifications and no missed cleavages

were allowed in MS1 search. For DDA data, the variable methionine oxidation and tmt10plex tags (where applicable) were allowed. The other search parameters were used by default.

To estimate fold changes of microorganism biomasses between samples of model microbiomes I and II, protein identification was made against the pooled proteogenomic databases (**Table 4**). Quantitation was performed with DirectMS1Quant[40].

To characterize changes in metabolic activity of *R. opacus* 1CP, the DirectMS1Quant was used for quantitation using the following parameters: differentially regulated proteins satisfy Benjamini Hochberg FDR < 0.05, fold change (FC) threshold was two standard deviations of $\log_2$FC distribution; intensity normalization by 1000 quantified peptides with maximal intensities was applied. Functional annotation and gene ontology analysis was performed using STRING db[41]. The annotation is publicly available at https://version-12-0.string-db.org/organism/STRG0A76PPK.

R. erythropolis X5 was quantified using Diffacto[42] and QRePS[43]. Differentially regulated proteins were selected using the following criteria: $|\log_2 FC| > 1.2$, $|\log_{10} FDR| > Q3+1.5*IQR$, where Q3 and IQR are 3rd quartile and interquartile range[43]. Functional annotation and gene ontology analysis were performed using STRING db. The annotation is freely available at https://version-12-0.string-db.org/organism/STRG0A44VNI. **Tables S1 and S2** containing quantitation results for *R. opacus* 1CP and *R. erythropolis* X5 are provided in supporting information.

Table 4. Summary on databases and target-decoy FDRs used to characterize different samples. *group-specific FDR, target-decoy approach.

Sample type	NCBI:txid	Database	# proteins in fasta	FDR
19 single strains	all	SwissProt+TrEMBL		0.05
Fecal human microbiome	all	SwissProt+TrEMBL		0.05*
Model I (*R. opacus* 1CP	37919	SwissProt+TrEMBL	7769	0.05
R. erythropolis X5	-	Proteogenomic	6407	
G.alkanivorans 135	-	Proteogenomic	5855	
G. amicalis 6-1	1220574	SwissProt+TrEMBL	4528	
P. aryabhattai 25)	1358420	SwissProt+TrEMBL	6393	
Model II (*R. opacus* 1CP	37919	TrEMBL	7769	0.05
R. opacus 3D	-	Proteogenomic	8870	
R. opacus S8	-		8102	
R. erythropolis X5	-		6407	
R. qingshengii 7B	-		6277	
R. qingshengii VT6	-		6789	
G. alkanivorans 135	-		5855	
G. polyisoprenivorans 135)	-		6248	
Model III (*R. opacus* 1CP	37919	SwissProt+TrEMBL	7769	0.05
R. erythropolis X5	-	Proteogenomic	6407	

P. aryabhattai 25 *G. amicalis* 6-1 *G. alkanivorans* 135)	-	Proteogenomic	5855	
	1220574	SwissProt+TrEMBL	4528	
	1358420	SwissProt+TrEMBL	6393	
Single strain response 1 *R. opacus* 1CP	37919	SwissProt+TrEMBL	7769	0.01
Single strain response 2 *R. erythropolis* X5	-	Proteogenomic	6407	0.01

Results

1. *Algorithm for two-stage blind database search identifies genus and species of a bacteria from LC-MS1 data*

1.1 Algorithm for two-stage search against SwissProt+TrEMBL database. The standard DirectMS1 analysis using ms1searchpy search engine cannot handle protein databases with more than few hundred thousands of proteins. To solve that issue, we first used a preliminary search to identify the most probable strains presented in the sample using the whole bacterial database SwissProt+TrEMBL as input candidates. The algorithm for this search is based on a simple matching of theoretical peptide *m/z* values with experimental MS1 spectra. Based on the identifications from the preliminary search, the shortened database is composed of the most probable species. This shortened database is further used for accurate protein identification using the standard ms1searchpy search. Details on the algorithm are provided in the Supporting information (Appendix A, **Figure S1**). As an example, the runtime for full analysis on a computer with Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz (4 cores) takes 4 hours: 3 hours for SwissProt+TrEMBL parsing and calculation of theoretical *m/z* (performed once), 6 minutes for preliminary search and 1 hour for standard search against the shortened database (a sum for 5 sample replicates).

1.2 Two-stage blind search against SwissProt+TrEMBL demonstrates high accuracy in the identification of isolated bacterial species. To estimate the efficiency of the proposed two-stage identification approach, public LC-MS/MS data for 19 strain isolates was used[14]. The results of blind identification are presented in **Table 5**. The top identified microorganism matched the analyzed strain at the level of genus+species in 95 %, and 32 % of those cases matched the correct strain ("Level of match" column in **Table 5**). Analysis of top-3 identified microorganisms reveals that 11 of 19 cases showed the undoubted leader supported by at least 10 times higher numbers of identified proteins than the top-second and top-third taxons. The remaining eight cases demonstrated less than 2-fold difference in numbers of protein identifications distributed between the top-first and top-second organisms. This observation corresponds to the identification of taxons of the same species or species group and means a high similarity of proteomes. *E. coli* was the only strain identified at the level of the genus due to the classification within the ete3 toolkit[44]. The *E. coli* strain K-12 (NCBI:txid83333, 4518 proteins in

SwissProt) is not reported with NCBITaxa().get_descendant_taxa() function for *E.coli* species (NCBI:txid562) which leads to it being excluded from the search space. At the next step, *E.coli* species database is also excluded in blind search due to its excessive size (1501705 proteins). Reportedly, differentiation between *Escherichia coli* and *Shigella*[45] or *Bacillus cereus* and *Bacillus anthracis*[46] is also problematic for MALDI TOF that requires development of specific solutions. The proposed blind search algorithm can be successfully applied for identification of individual species and strains using fast proteome profiling data.

Table 5. Summary of blind identification of bacteria against SwissProt+TrEMBL database shows correct match between the analyzed and top identified strains at the level of species groups, species and strains. Level of match was determined using phylogenetic trees available at https://phylot.biobyte.de/. Percentage is defined as (#proteins identified per species / total identified proteins)*100 %. Strains taken in multiple replicates were analyzed calculating the mean number of identifications across replicates.

Strain analyzed	NCBI:txid	1st taxon by #proteins, %	2nd taxon by #proteins, %	3rd taxon by #proteins, %	Level of match
Acinetobacter baumannii DSM 30007	470 575584	*Acinetobacter baumannii* *(45 %)*	*Acinetobacter* sp. FDAARGOS_559 (41 %)	*Acinetobacter* sp. 25977_8 (6 %)	strain
Bacillus cereus DSM 31	1396 226900	*Bacillus anthracis* *(51 %)*	*Bacillus* sp. S66 *(28 %)*	*Bacillus* sp. BI3 *(5 %)*	Species group
Bacillus velezensis DSM 23117/FZB42	492670 326423	*Bacillus velezensis* FZB42 (32 %)	*Bacillus velezensis* *(19 %)*	*Bacillus* sp. VMFN-A1 (18 %)	strain
Burkholderia cepacia ATCC 25416	292 983594	*Burkholderia reimsis* (22 %)	*Burkholderia* sp. LS-044 (16 %)	*Burkholderia lata* (14 %)	Species group
Burkholderia thailandensis DSM 13276	57975 271848	*Burkholderia thailandensis* *(65-68 %)*	*Burkholderia oklahomensis* (6-8 %)	*Burkholderia* sp. MSMB1552 (6-8 %)	strain
Burkholderia thailandensis E125	57975	*Burkholderia thailandensis* *(67-70 %)*	*Burkholderia* sp. MSMB1552 (8-12 %)	*Burkholderia oklahomensis* (6-8 %)	species
Burkholderia thailandensis E131	57975	*Burkholderia thailandensis* *(67-68 %)*	*Burkholderia* sp. MSMB1552 (7 %)	*Burkholderia oklahomensis* (1-9 %)	species
Burkholderia thailandensis E153	57975	*Burkholderia thailandensis* *(65-68 %)*	*Burkholderia* sp. MSMB1552 (7-13 %)	*Burkholderia oklahomensis* (8 %)	species
Burkholderia thailandensis LMG 20219	57975	*Burkholderia thailandensis* *(67-71 %)*	*Burkholderia* sp. MSMB1552 (8-11 %)	*Burkholderia oklahomensis* (1-8 %)	species
Burkholderia oklahomensis DSM 21774	342113	*Burkholderia oklahomensis* *(78-83 %)*	*Burkholderia lata* (4 %)	*Burkholderia* sp. ABCPW 14 (3 %)	strain

Citrobacter freundii DSM 30039	546 1006003	uncultured *Citrobacter* sp. (57 %)	*Citrobacter freundii complex* sp. CFNIH9 *(14 %)*	*Citrobacter* sp. LUTT5 (8 %)	Species group
Enterococcus faecalis DSM 20371	1351	*Enterococcus faecalis (92 %)*	*Enterococcus faecium* (2 %)	*Enterococcus haemoperoxidus* ATCC BAA-382 (1 %)	strain
Mycobacteroides abscessus DSM 44196	36809 561007	*Mycobacteroides abscessus* ATCC 19977 *(57-58 %)*	*Mycobacteroides abscessus* subsp. Abscessus (18-19 %)	*Mycobacteroides franklinii* (11 %)	strain
Escherichia coli DSM 3871 (K12 W3110 derivative)	562 83333	*Escherichia fergusonii* ATCC 35469 (20 %)	*Escherichia* sp. KTE114 (18 %)	*Shigella boydii* (12 %)	genus
Pseudomonas aeruginosa ATCC 27853	287	*Pseudomonas aeruginosa* (93-94 %)	*Pseudomonas* sp. ATCC 13867 (2 %)	*Pseudomonas nitroreducens* (1 %)	species
Staphylococcus epidermidis DSM 1798	1282	*Staphylococcus epidermidis (85 %)*	*Staphylococcus* sp. HMSC34G04 (4 %)	*Staphylococcus* sp. RIT622 (2 %)	species
Staphylococcus aureus DSM 20231/NCTC 8532	1280 1241616	*Staphylococcus aureus* (95 %)	*Staphylococcus pasteuri* (1 %)	*Staphylococcus argenteus* (1 %)	species
Vibrio cholerae NIH 41	666	*Vibrio cholerae* (77 %)	*Vibrio mimicus* (8 %)	*Vibrio metoecus* (6 %)	species
Yersinia pseudotuberculosis DSM 8992	633	*Yersinia wautersia (41 %)*	*Yersinia pseudotuberculosis (28 %)*	*Yersinia similis* (5 %)	species

2. *Two-stage blind database search identifies relative composition of microbial community at level of genus*

2.1 Model mixtures of soil bacteria.

To study the performance of a two-stage blind search to identify microbiome composition, the soil bacteria mixed at known concentrations (**Tables 1-3**) were used. The main motivation was to investigate how blind searches resolve the metaproteomic samples containing different species of the same genus. The results of blind identification of sample compositions are summarized in **Table 6**. The actual and estimated compositions of each model microbiome were compared at two levels: genus and species. In blind identification, the mean difference between the actual and identified components of microbiome composition was 12 % at the level of genus. The difference between the actual and estimated compositions depended on the microorganisms under study. In our example, the content of *Rhodococcus* was easily identified and often overestimated, while *Gordonia* was identified in all samples (even though its content was below 15 %), but always underestimated. *Priestia* was identified with high accuracy within a few percent, but only if its actual fraction in the sample was higher than 30

%. Below 30 %, *Priestia* was not identified at all. At the level of species, the mean difference between the estimated and actual content was 15 %. However, the number of missing identifications of microorganisms increased from 20 % at the level of genus to 40 % at the level of species. Thus, the current realization of blind search algorithm for identification of microbiome composition provides the most complete information at the level of genus.

Table 6. Summary of blind identification of microbial compositions against SwissProt+TrEMBL database. Sample labels and actual compositions correspond to Tables 1-3. Percentage in actual compositions is defined as (peptide mass per genus (species) / total peptide mass)*100 %. Percentage in estimated compositions is defined as (#proteins per genus (species) / total #proteins)*100 %. Empty cells mean missing identification of microorganisms from a given genus or species. The mean difference between actual and estimated composition was calculated as ∑|Actual % - Estimated %|/N, where N=33 is the number of rows in the column "actual composition, % by mass". Missings were imputed by zero.

Sample	GENUS		SPECIES	
	Actual composition, % by mass	Estimated composition, % by #proteins	Actual composition, % by mass	Estimated composition, % by #proteins
M1.1	*Rhodococcus* (33 %)	*Rhodococcus* (38.4 %)	*R. opacus* (33 %)	*R. opacus (5.3 %)*
	Priestia (33 %)	*Priestia (39.7 %)*	*P. aryabhattai* (33 %)	*P. aryabhattai (39.7 %)*
	Gordonia (33 %)	*Gordonia (*15.4 %)	*G. alkanivorans* (33 %)	
M1.2	*Rhodococcus* (50 %)	*Rhodococcus* (80.5 %)	*R. opacus* (50 %)	*R. opacus* (6.1 %)
	Priestia (25 %)		*P. aryabhattai* (25 %)	
	Gordonia (25 %)	*Gordonia* (9.9 %)	*G. alkanivorans* (25 %)	
M1.3	*Rhodococcus* (33 %)	*Rhodococcus* (37.2 %)	*R. opacus* (33 %)	*R. opacus* (5.9 %)
	Priestia (50 %)	*Priestia* (50.3 %)	*P. aryabhattai* (50 %)	*P. aryabhattai* (48.4 %)
	Gordonia (17 %)	*Gordonia* (6.6 %)	*G. alkanivorans* (17 %)	
M1.4	*Rhodococcus* (40 %)	*Rhodococcus* (64.3 %)	*R. opacus* (20 %)	*R. opacus* B4 (3.5 %)
			R. erythropolis (20 %)	*R. erythropolis* PR4 (12.7 %)
	Priestia (20 %)		*P. aryabhattai* (20 %)	
	Gordonia (40 %*)*	*Gordonia* (26.2 %)	*G. amicalis* (20 %)	*G. amicalis* NBRC 100051 = JCM 11271 (3.8 %)
			G. alkanivorans (20 %)	*G. alkanivorans* NBRC 16433 (13.7 %)
M1.5	*Rhodococcus* (43 %)	*Rhodococcus* (64.3 %)	*R. opacus* (14 %)	*R. opacus* B4 (3.5 %)
			R. erythropolis (29 %)	*R. erythropolis* (25 %) *R. erythropolis* PR4 (5.9 %)
	Priestia (14 %)		*P. aryabhattai* (14 %)	
	Gordonia (43 %)	*Gordonia* (29.7. %)	*G. amicalis* (29 %)	
			G. alkanivorans (14 %)	*G. alkanivorans* NBRC 16433 (22.0 %)
M2.1	*Rhodococcus* (49 %)	*Rhodococcus* (64.8 %)	*R. opacus* (33 %)	*R. opacus* (31.8 %) *R. opacus* B4 (2.8 %)
			R. qingshengii (16 %)	
			R. erythropolis (0.4 %)	*R. erythropolis* PR4 (10.8 %)
	Gordonia (51 %)	*Gordonia* (30.0 %)	*G. alkanivorans* (0.8 %)	
			G. polyisoprenivorans (50 %)	*G. polyisoprenivorans* NBRC 16320 = JCM 10675 (23.8 %)
M2.2	*Rhodococcus* (96.5 %)	*Rhodococcus* (95.6 %)	*R. opacus* (20 %)	*R. opacus* (19.5 %) *R. opacus* B4 (2.4 %)
			R. qingshengii (75 %)	*R. qingshengii* (21.4 %)

			R. erythropolis (1.6 %)	*R. erythropolis* (31.1 %) *R. erythropolis* PR4 (7.5 %)
	Gordonia (3.5 %)	*Gordonia* (0.9 %)	*G. alkanivorans* (0.4 %)	
			G. polyisoprenivorans (3 %)	
M2.3	*Rhodococcus* (92.9 %)	*Rhodococcus* (95.2 %)	*R. opacus* (5.1 %)	*R. opacus* B4 (1.1 %)
			R. qingshengii (75.1 %)	*R. qingshengii* (28.5 %)
			R. erythropolis (13 %)	*R. erythropolis* (43.7 %) *R.erythropolis* PR4 (11.0 %)
	Gordonia (7.1 %)	*Gordonia* (1.1 %)	*G. alkanivorans* (6 %)	
			G. polyisoprenivorans (0.8 %)	
ABRF1	*Rhodococcus* (77 %)	*Rhodococcus* (89.2 %)	*R. opacus* (42 %)	*R. opacus* B4 (4.8 %)
			R. erythropolis (35 %)	*R. erythropolis* (32.3 %) *R. erythropolis* PR4 (8.9 %)
	Priestia (10 %)		*P. aryabhattai* (10 %)	
	Gordonia (13 %)	*Gordonia* (3.1 %)	*G. amicalis* (7 %)	
			G. alkanivorans (6 %)	
ABRF2	*Rhodococcus* (12.6 %)	*Rhodococcus* (18.7 %)	*R. opacus* (9.6 %)	*R. opacus* B4 (3.1 %)
			R. erythropolis (3 %)	
	Priestia (0.3 %)		*P. aryabhattai* (0.3 %)	
	Gordonia (87.1 %)	*Gordonia* (73.9 %)	*G. amicalis* (0.1 %)	*G. amicalis* NBRC 100051 = JCM 11271 (4.8 %)
			G. alkanivorans (87 %)	*G. alkanivorans* NBRC 16433 (46.9 %)
ABRF3	*Rhodococcus* (17 %)	*Rhodococcus* (21.8 %)	*R. opacus* (15 %)	*R. opacus* B4 (5.6 %)
			R. erythropolis (2 %)	*R. erythropolis* PR4 (1.9 %)
	Priestia (63 %)	*Priestia* (65.0 %)	*P. aryabhattai* (63 %)	*P. aryabhattai (65.0 %)*
	Gordonia (20 %)	*Gordonia* (6.0 %)	*G. amicalis* (10 %)	
			G. alkanivorans (10 %)	
ABRF4	*Rhodococcus* (10.7 %)	*Rhodococcus* (32.4 %)	*R. opacus* (0.3 %)	*R. opacus* (0.7 %)
			R. erythropolis (10.4 %)	*R. erythropolis PR4* (11.1 %)
	Priestia (8.8 %)		*P. aryabhattai* (8.8 %)	
	Gordonia (80.5 %)	*Gordonia* (63.7 %)	*G. amicalis* (80.4 %)	*G. amicalis* NBRC 100051 = JCM 11271 (12.6 %)
			G. alkanivorans (0.1 %)	

2.2 Fecal microbiome. To test the algorithm in identifying the composition of the real microbiome, the public dataset for human fecal microbiome was used[14]. In the referenced publication , the sample was analyzed using a multi-omics approach involving metagenomics, metatranscriptomics, and metaproteomics. Taxonomic profiling of the results obtained using these methods has shown that the families *Ruminococcaceae*, *Lachnospiraceae, Eubacteriaceae* and *Clostridiaceae* were the most abundant. Among the organisms identified using the blind search (**Figure 1**), the following bacteria were found: (1) *Ruminococcaceae* family: *Faecalibacterium prausnitzii*, *Ruminococcus* sp. AM54-14NS, *Faecalibacterium* sp. AF28-13AC, *Ruminococcus* sp. AM29-26, *Ruminococcus* sp. AF31-14BH; (2) *Eubacteriaceae* family: *Eubacterium rectale* CAG:36; (3) *Lachnospiraceae* family: *E. rectale* ATCC 33656; and (4) *Clostridiaceae* family*: Clostridia* sp. AF34-10BH. Despite the low number of identified proteins for the sample, it was possible to identify four families composing approximately 90 % of the

sample (according to the data from[14]) and suggest a genus composition for them. Thus, the proposed algorithm can be applicable for the identification of both individual strains and complex samples such as the real human gut microbiome.

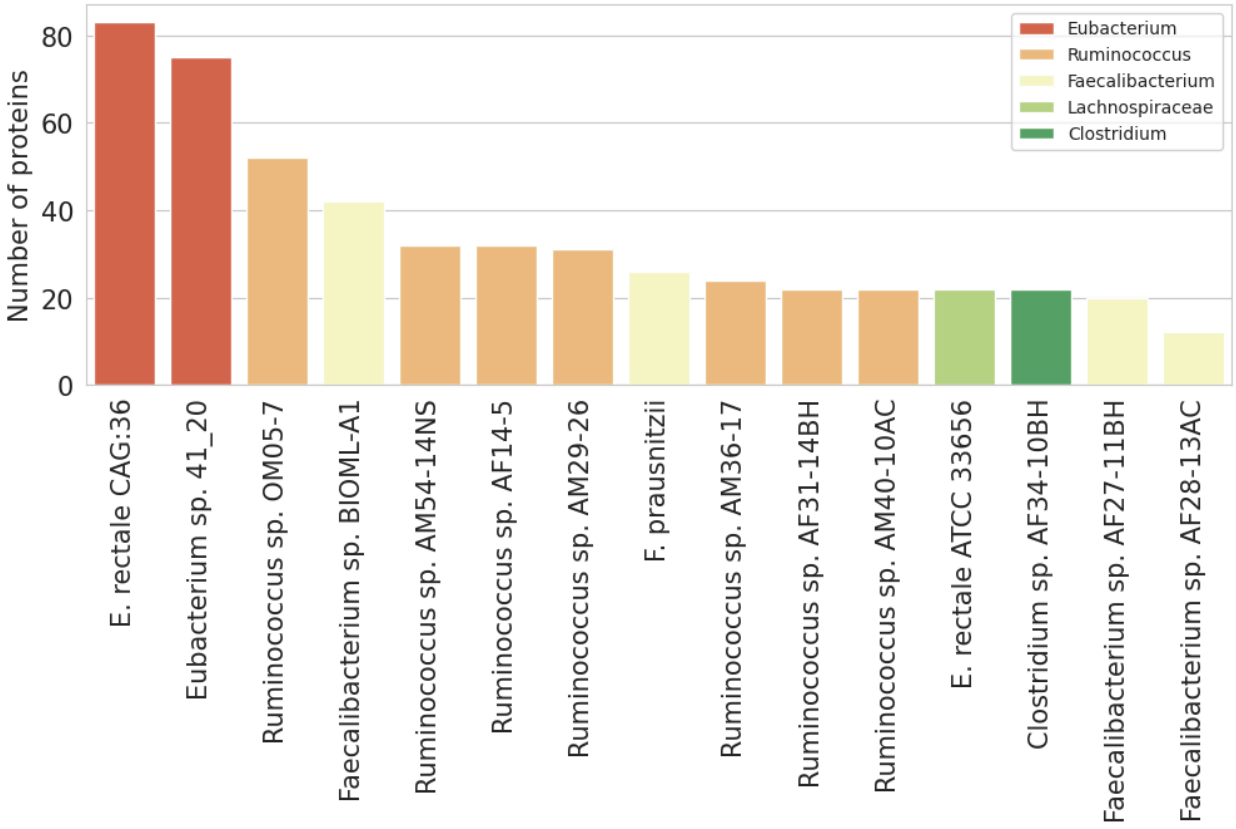


Figure 1. Blind identification of fecal microbiome composition using DirectMS1. Organism-specific FDR is 0.05.

3. *DirectMS1 allows correct determination of the fold change of species biomass between microbiome samples*

Model microbiomes I, II and III (**Tables 1-3**) were used to analyze strain abundance variation between different samples with the following procedure. After quantitation at peptide level, the fold changes of all quantified peptides without decoys were used to plot the baseline distribution **(Figure 2a**). Then, fold change distribution was built for all peptides from each strain of the model microbiome ("strain" distribution). The baseline was subtracted from the "strain" distribution and the corrected distribution was analyzed **(Figures 2 b,c,d**). A weighted average of fold changes for the corrected distribution provides an approximation of the biomass fold change for the selected strain, where weights are non-negative normalized densities.

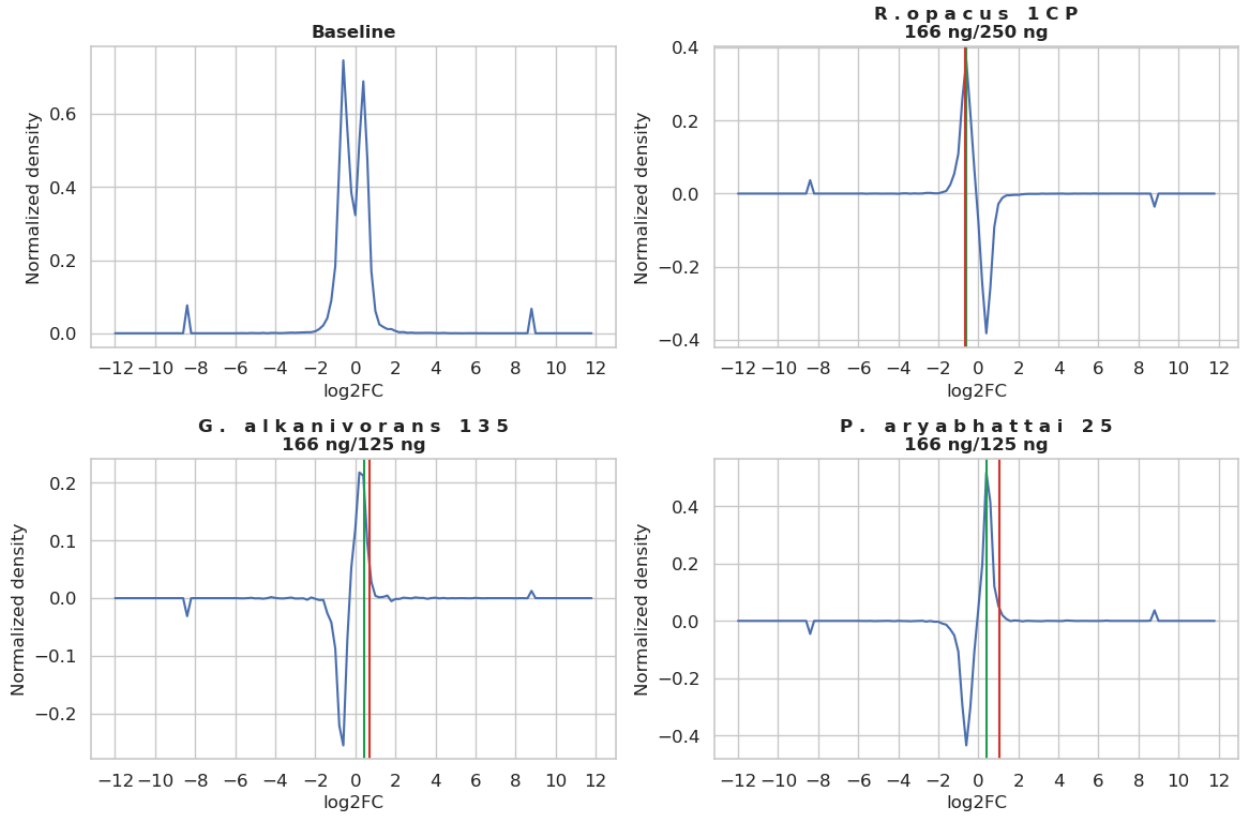


Figure 2. The normalized densities of $\log_2$FC distribution for M1.1/M1.2 comparison. The green line stands for the actual ratio of peptide masses (given in subtitles) and the red line stands for the calculated ratio, log2-scaled. FC is the fold change in strain biomass between model microbiomes.

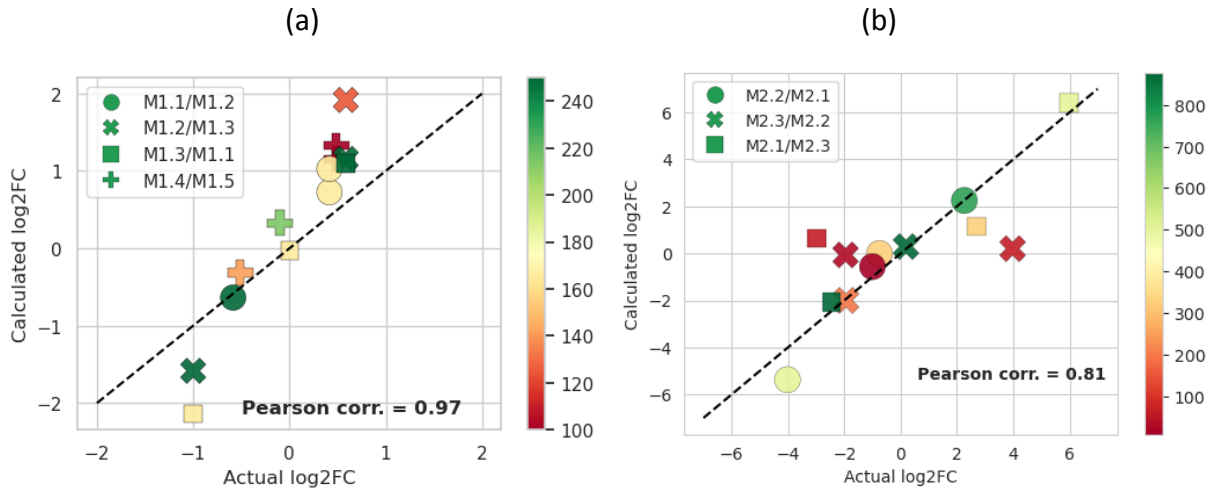
Figure 3 summarizes estimates of fold changes in strain biomass between samples and compares it to actual mass ratios. For all samples from the Model I, the experimentally measured fold changes coincide well with the actual values (Pearson correlation coefficient of 0.97, **Figure 3a**). It means that species mixed at ratios 1:1, 1:2, 1:3 and composing at least from 15 to 20 % of a microbiome are typically well resolved with our method. For Model II and III, a decrease in fold change predictions was observed (**Figure 3b,c**). Model II represented the extreme case of uniform composition of the same species (*R. opacus* 1CP, *R. opacus* 3D, *R. opacus* S8; *R. erythropolis* X5, *R. qingshengii* 7B, *R. qingshengii* VT6, *G. alkanivorans* 135, *G. polyisoprenivorans* 135) mixed in a range of concentrations differing by up to two orders (**Table 2** in M&M). This complex case resulted in a decrease of Pearson correlation to 0.81 due to *Gordonia* species composing just 3.5 % (M2.2) and 7.1 % (M2.3) of the model microbiome (**Figure 3b**).

Model III had compositions differing at the level of species (*R. opacus* 1CP, *R. erythropolis* X5, *P. aryabhattai* 25, *G. amicalis* 6-1, *G. alkanivorans* 135) (**Table 3** in M&M) and covering the similar range of concentrations as samples from Model II. Model III is less complex

compared with Model II that results in more accurate fold change estimations (R = 0.93, **Figure 3c**).

For the low amounts (species or genera content within a few percent) in one of the samples, the missing values can occur and subsequent imputation of peptide abundance results in a decreased accuracy. **Figure S2** illustrates such a case with the bacterium *P. aryabhattai* 25 from Model III ABRF2/ABRF3, for which the peptide content was 3 ng and 631 ng, respectively. If the total biomass of microorganisms is low in both samples, the method could not quantify enough peptides to construct the distribution. This case is illustrated with the bacterium *R. erythropolis* X5 in **Figure S2**, for which there was no fold change estimation provided.

Another case we explored is the fold change predictions for mixtures with high overlap of the identified peptides coming from different species. For example, the fold change distributions for *G. alkanivorans* 135 and *G. amicalis* 6-1 from Model III ABRF2/ABRF3 coincide (**Figure S2**) and the fold change estimate matches the sum of the peptide masses of both strains. This trend was observed for all mixtures analyzed. Next example is the Model II M2.2/M2.3 comparison demonstrating an absence (or undetectable concentrations) of unique peptides required for distinguishing the *R. opacus* strains, as well as *R. erythropolis* and *R. qingshengii* species (**Figure S3**). From a practical viewpoint, it means that the measured fold change between the components of complex microbiomes matches the sum of the biomasses of all strains identified by the shared peptides.



(c)

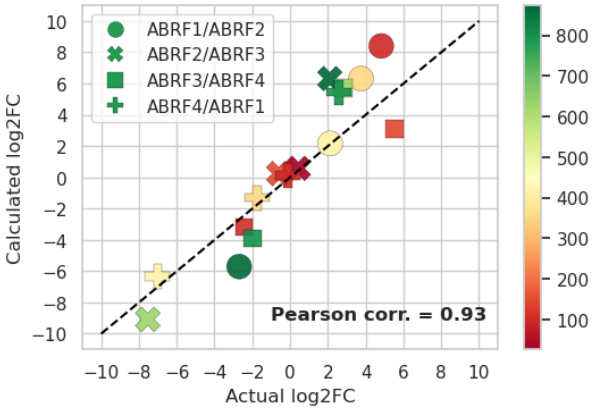


Figure 3. Correlation of actual and experimentally measured fold changes in strain biomass for Model I (a), Model II (b), Model III (c). Y: fold change estimate from experimental data, X: actual ratios, in log2-scale. Legends correspond to labels in Tables 3-5. The color corresponds to the maximum of two compared species biomasses, in nanograms.

4. *DirectMS1 allows the characterization of changes in metabolic pathways and biodegradation ability of strain isolates*

The performance of DirectMS1 to characterize metabolic response of bacteria to the stimulus was tested using two pollutant-degrading actinobacteria: (1) *Rhodococcus opacus* 1CP grown in presence of glucose, benzoate, phenol, and 4-chlorophenol; and (2) *Rhodococcus erythropolis* X5 grown at 28 °C and 6 °C in presence of *n*-hexadecane. **Figure 4-5** summarizes the results of quantitation analysis.

Figure 4 summarizes the key differences in the proteomes of *Rhodococcus opacus* 1CP grown in the presence of phenol, 4-chlorophenol and benzoate, compared with glucose. *R. opacus* 1CP is the actinobacterium able to degrade phenol and their derivatives[34]. In our study, the presence of phenol resulted in upregulation of enzymes involved in phenol degradation: three different phenol hydroxylases ((**pheA1(1**), **pheA1(2**), and **pheA1(3**)), and 2-oxopent-4-enoate hydratase (**R1CP_00930**). These enzymes are also induced in the presence of benzoate and 4-chlorophenol, but with smaller magnitude. The data obtained are in full agreement with previously obtained real-time PCR results[47]. Using specific primers for the small subunit of all three phenol hydroxylases of strain 1 CP growing in the presence of phenol, the gene activation has been shown; an increase for phe A1(3) was approximately 2000 times.

Enzyme cluster consisting of upregulated 3-methyl-2-oxobutanoate hydroxymethyltransferase (**panB)**, catechol 1,2-dioxygenases (**catA1/catA2)**, 4-nitrophenol 2-monooxygenase (**nphA1)**, metapyrocatechase (**pheB)**, and Fe-ADH domain-containing protein were specific to phenol. For degradation of 4-chlorophenol, we observed no specific features, the upregulated enzymes were the same as for phenol conditions. For benzoate conditions, we

found specific pattern of enzymes degrading benzoate: benzene 1,2-dioxygenase (**A8787_2660)**, benzoate 1,2-dioxygenase subunit alpha (**benA)**, benzoate 1,2-dioxygenase small subunit (**benB)**, 1,6-dihydroxycyclohexa-2, 4-diene-1-carboxylatedehydrogenase (**xylL)**, and ferrienterobactin-binding periplasmic protein (**fepB)**.

In total, the results of ultrafast proteomics profiling were consistent with the expected changes in metabolic activity and can serve as a valuable support in characterization of pollutant degradation activity of different microorganisms.

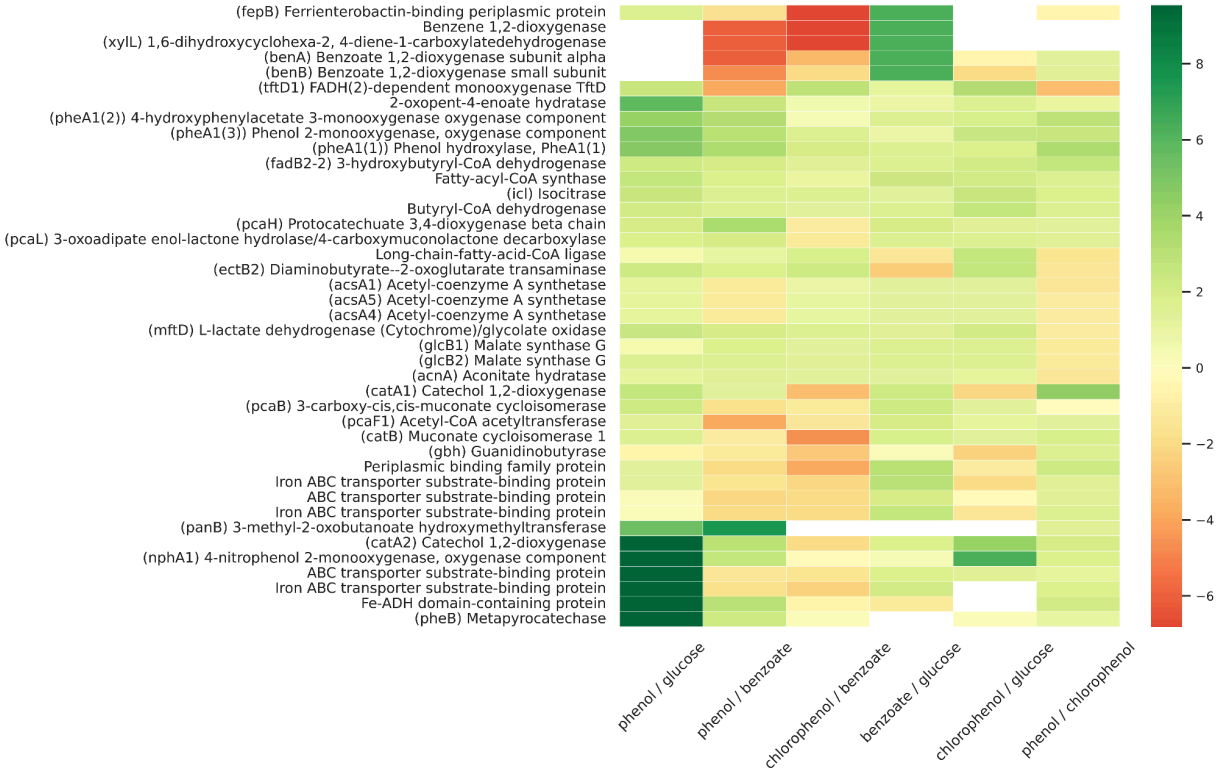


Figure 4. Proteomic response of actinobacteria *Rhodococcus opacus* 1CP to the presence of aromatic compounds: benzoate, phenol, and 4-chlorophenol, compared with glucose. Quantitation: DirectMS1Quant[40] [10.1021/acs.analchem.2c02255]. Protein selection corresponds to the most enriched biological processes (GO score ≥ 6).

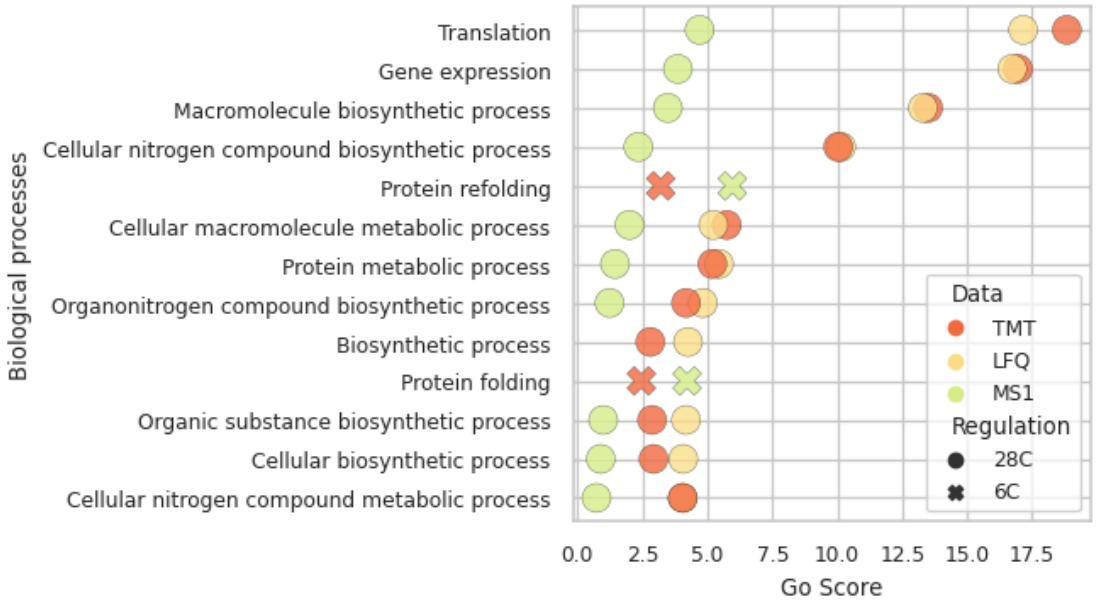
Figure 5 shows response of the bacterium *Rhodococcus erythropolis* X5 grown on *n*-hexadecane to temperature change from 6 °C to 28 °C. *R. erythropolis* X5 is the psychrotrophic bacterium that possesses a wide range of catabolic activities within 4°C to 28°C temperature range. Specifically, this strain can degrade *n*-alkanes which constitute the main fraction of oil pollution[25]. In our study, the proteome of X5 strain grown on *n*-hexadecane at 6 °C was compared with its proteome at 28 °C. At 6 °C, we observed proteins related to all steps of *n*-alkane degradation including oxidation to alcohols (QEX08599, QEX09356, etc), oxidation to aldehydes (QEX08400, QEX09399, etc), oxidation to fatty acids and fatty acid metabolism

(QEX08594, QEX12285, QEX08597, etc), exopolysaccharide production (QEX09055, QEX10519, QEX12651, QEX11399, etc), and iron transport (QEX13740, QEX13741, QEX09379, etc) (**Table S2, Appendix B**). Besides enzymes involved in *n*-alkane biodegradation, the chaperons/chaperonins (QEX09859, QEX13151, QEX09986, QEX09823) and stress/protection proteins (QEX09757, QEX11651, QEX11169) were upregulated at 6 °C. Also, a significant loss of 30S/50S ribosomal proteins was observed at cold conditions. Numerous studies report that loss of ribosomal proteins in bacteria is the reaction to stress and is necessary for survival under nutrient- or energy-limited conditions[48,49,50,51]. GO analysis revealed that translation, gene expression, and protein folding were significantly affected at cold conditions (**Figure 5a**). Proteins involved in DNA and RNA processing, regulation of transcription and translation are shown in **Figure 5b** and **Appendix B.**

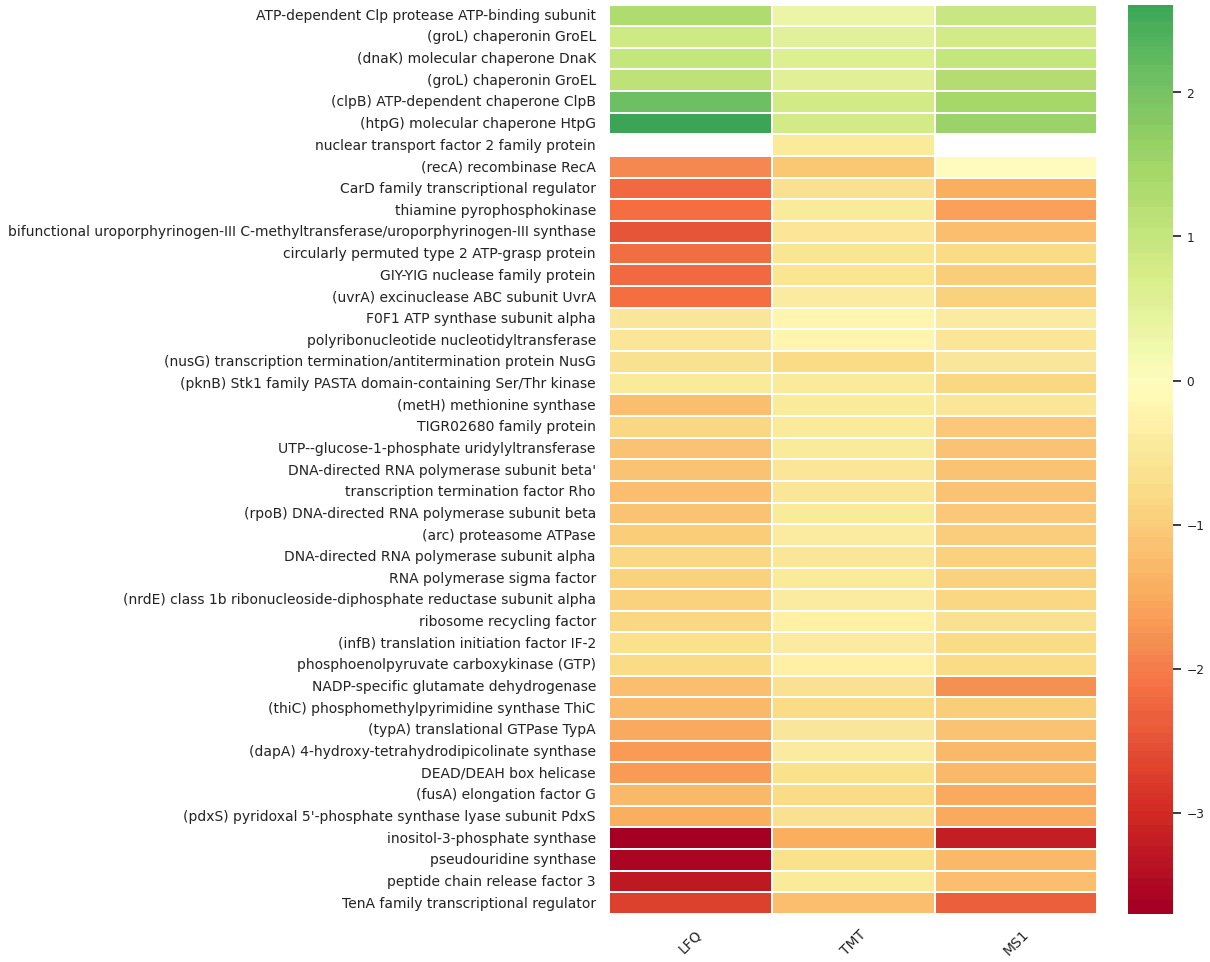
At 28 °C, we observed the abundant protein group corresponding to degradation of *n*-alkanes, but expressed from other parts of genome: *n*-alkanes to alcohols (QEX08601, QEX08447), alcohols to aldehydes (QEX10266, QEX10225, QEX10036), aldehydes to fatty acids / fatty acid metabolism (QEX12404, QEX10265, QEX10256, and many others), exopolysaccharide production (QEX12160, QEX13966, QEX14217, etc), and iron transport (QEX11646, QEX12370, QEX11669, etc) (**Table S2, Appendix B**). This observation suggests that different parts of the *R. erythropolis* X5 genome were transcribed to translate the enzymes needed for *n*-alkane degradation under optimal and cold conditions. This transcriptional behavior can be a conserved mechanism of coping with stress.

To confirm the technical reproducibility of the results, a comparison was made between label-free DirectMS1, DDA LFQ and TMT-based quantitation. **Figure 5a** shows the enriched biological processes obtained using the three quantitation methods. The results of the analysis show that the set of enriched biological processes for all methods has a high overlap. The protein content behind the top GO enrichments, measured using different quantitation workflows, is shown in **Figure 5b**. Legend in **Figure 5b** compares the duration of the experimental methods for the used quantitation methods. DirectMS1 requires less experimental time providing comparable performance.

(a)


Translation
Gene expression
Macromolecule biosynthetic process
Cellular nitrogen compound biosynthetic process
Protein refolding
Cellular macromolecule metabolic process
Protein metabolic process
Organonitrogen compound biosynthetic process
Biosynthetic process
Protein folding
Organic substance biosynthetic process
Cellular biosynthetic process
Cellular nitrogen compound metabolic process
Biological processes
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Go Score
Data
TMT
LFQ
MS1
Regulation
28C
6C

(b)



Method	Time per run	Total time
DirectMS1, LC-MS1 (MS1)	7.5 min	45 min
Label Free Quantitation, DDA (LFQ)	87 min	522 min
Tandem mass tag labeling, DDA (TMT)	137 min	137 min

Figure 5. Response of actinobacteria *Rhodococcus erythropolis* X5 grown on *n*-hexadecane to temperature change from 6 °C to 28 °C. Quantitation: Diffacto[42] (LFQ, MS1-only acquisition), MSFragger+Scavager+NSAF (LFQ, DDA), and MSFragger+Scavager+Diffacto (DDA, TMT labeling). GO Score = E * log10 fdr, where E is the enrichment of biological processes, fdr is the statistical significance of the GO enrichment corrected for multiple comparisons using Benjamini-Hochberg method[52].

Discussion

We demonstrate that the DirectMS1 method for fast proteomic profiling can be effectively used for functional characterization of strain isolates and quantitative assessment of microbial communities. Our algorithm for two-stage blind database search discriminated

bacteria to species level from LC-MS1 data with high accuracy of 95 %, and 32 % of those cases were matched to the correct strains. The algorithm identifies the composition of microbial communities at the level of genus and provides quantitative estimates in strain biomass. DirectMS1 for metaproteomics takes a position between MALDI TOF MS and LC-MS/MS-based proteomics, combining both fast identification and quantitative assessment of strain isolates and microbiomes at reasonable costs.

Besides MALDI-TOF MS, 16S rRNA gene sequencing and whole-genome sequencing are widely used for microorganism identification[53]. The accuracy of microorganism identification depends on the number of strain-specific features measured: the larger the number of features, the higher the level of discrimination. For example, 16S rRNA gene sequencing can possess limits in discriminating to the species level[54,55]. MALDI-TOF MS is often positioned as a more accurate technique capable of demonstrating high discrimination to the species and even strain levels[56]. For example, *Citrobacter* strains are identified to the species level by the MALDI-TOF MS, while 16S rRNA gene sequencing could not discriminate them clearly[54]. Another example, *Yersinia pestis* and *Yersinia pseudotuberculosis* are difficult to separate[57]. These species have similar genomes and identical 16S rRNA genes[58], and their differentiation by MALDI-TOF MS is also challenging[46]. *Yersinia pseudotuberculosis* and *Yersinia wautersii* are also genetically close[59]. *Bacillus cereus* and *Bacillus anthracis*[46], *Escherichia coli, Escherichia fergusonii* and *Shigella*[60,61,45] are also represent challenges in discrimination. These earlier findings agree well with the top-ranking identifications reported during the blind search of strain isolates (**Table 5**). From the literature, we see that some species largely sharing genomes can generate identical 16S rRNA and very similar proteome profiles, and then can be resolved by only a combination of methods including whole genome sequencing.

The 5-minute LC-MS1-based proteome profiling of strain isolates responding to environmental conditions demonstrated high performance in quantitation (**Figures 4-5**). We detected the key enzymes degrading aromatic compounds, such as phenol, benzoate and 4-chlorophenol, that are in full agreement with previous studies[62,63]. We measured the key enzymes involved in the oxidation of *n*-alkanes to alcohols (i.e. alkane monooxygenases and cytochrome P450 family proteins); alcohols oxidation to aldehydes (alcohol dehydrogenases); aldehydes oxidation to fatty acids (aldehyde dehydrogenases); fatty acid metabolism (acyl-CoA dehydrogenases, fatty acid-CoA synthases, fatty acid synthesis proteins); exopolysaccharide production (glycosyltransferases, glucose dehydrogenases, fructose synthases, mannose dehydrogenases, etc.); and iron transport (ferredoxins, ferredoxin reductases, ABC transporter proteins)[64]. We conclude that DirectMS1 was proved as a valuable approach to functional annotation of strain isolates responding to the environment.

Acknowledgements

I.T. and M.G. thank Prof. Victor Zgoda from "Human Proteome" Core Facility at the Institute of Biomedical Chemistry for help with implementation of DirectMS1 method on the Orbitrap FTMS system.

Data availability statement

LC-MS1 and LC-MS/MS data are available at proteomexchange.org (PXD050587, PXD050761, PXD050807, PXD050887). Code for blind search, identification of microbial sample composition and assessment of species biomass change between samples is available at GitHub (https://github.com/kazakova/Metaproteomics-DirectMS1).

Supporting information

Supporting information is available (Appendix A-B, Figures S1, S2, S3, Tables S1 and S2).

Funding

The study (developing DirectMS1 and bioinformatic solutions for metaproteome profiling, study design, sample preparation, data processing, etc) was supported by the Russian Science Foundation grant No. 20-14-00229.

F.K. and V.G. acknowledge generous grants to the VILLUM Center for Bioanalytical Sciences (VILLUM Foundation grant no. 7292), PRO-MS: Danish National Mass Spectrometry Platform for Functional Proteomics (grant no. 5072-00007B), and the Novo Nordisk Foundation (INTEGRA, NNF20OC0061575) for support of instrumentation infrastructure at the University of Southern Denmark.

Author contributions

Conceptualization, I.T.; Methodology, E.K.,J.B., M.I., Y.D., I.S., and I.T.; Software, M.I. and E.K.; Validation, E.K., J.B., and Z.Z.; Formal Analysis, M.I. and E.K.; Investigation, E.K., T.K., J.B., V.P., K.P, V.G., F.K., Y.D., I.S.; Data Curation, T.K. and E.K.; Resources, E.K., T.K., J.B., Y.D., I.S.; Writing – Original Draft Preparation, I.T.; Writing – Review & Editing, I.A., M.I., V.G., F.K, M.G.; Visualization, E.K.; Supervision, I.T.; Project Administration, I.T., M.G.

References

1. Ascandari, A., Aminu, S., Safdi, N. E. H., El Allali, A. & Daoud, R. A bibliometric analysis of the global impact of metaproteomics research. *Front. Microbiol.* **14**, (2023).
2. Kleiner, M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems* **4**, e00115-19 (2019).
3. Van Den Bossche, T. *et al.* The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome* **9**, 243 (2021).
4. Miura, N. & Okuda, S. Current progress and critical challenges to overcome in the bioinformatics

of mass spectrometry-based metaproteomics. *Comput. Struct. Biotechnol. J.* **21**, 1140–1150 (2023).
5. Van Den Bossche, T. *et al.* Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat. Commun.* **12**, 7305 (2021).
6. Baquero, F. & Nombela, C. The microbiome as a human organ. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* **18 Suppl 4**, 2–4 (2012).
7. Long, S. *et al.* Metaproteomics characterizes human gut microbiome function in colorectal cancer. *NPJ Biofilms Microbiomes* **6**, 14 (2020).
8. Zhao, J. *et al.* Data-independent acquisition boosts quantitative metaproteomics for deep characterization of gut microbiota. *NPJ Biofilms Microbiomes* **9**, 4 (2023).
9. Pietilä, S., Suomi, T. & Elo, L. L. Introducing untargeted data-independent acquisition for metaproteomics of complex microbial samples. *ISME Commun.* **2**, 51 (2022).
10. Lou, R. & Shui, W. Acquisition and Analysis of DIA-Based Proteomic Data: A Comprehensive Survey in 2023. *Mol. Cell. Proteomics MCP* **23**, 100712 (2024).
11. Simopoulos, C. M. A. *et al.* MetaProClust-MS1: an MS1 Profiling Approach for Large-Scale Microbiome Screening. *mSystems* **7**, e0038122 (2022).
12. Schubert, S. & Kostrzewa, M. MALDI-TOF MS in the Microbiology Laboratory: Current Trends. *Curr. Issues Mol. Biol.* **23**, 17–20 (2017).
13. Sandrin, T. R., Goldstein, J. E. & Schumaker, S. MALDI TOF MS profiling of bacteria at the strain level: a review. *Mass Spectrom. Rev.* **32**, 188–217 (2013).
14. Lasch, P., Schneider, A., Blumenscheit, C. & Doellinger, J. Identification of Microorganisms by Liquid Chromatography-Mass Spectrometry (LC-MS1) and in Silico Peptide Mass Libraries. *Mol. Cell. Proteomics MCP* **19**, 2125–2139 (2020).
15. Heyer, R. *et al.* Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36 (2017).
16. 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).
17. Bassignani, A. *et al.* Benefits of Iterative Searches of Large Databases to Interpret Large Human Gut Metaproteomic Data Sets. *J. Proteome Res.* **20**, 1522–1534 (2021).
18. Kumar, P. *et al.* A Sectioning and Database Enrichment Approach for Improved Peptide Spectrum Matching in Large, Genome-Guided Protein Sequence Databases. *J. Proteome Res.* **19**, 2772–2785 (2020).
19. Zhang, X. *et al.* MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**, 31 (2016).
20. Xiao, J. *et al.* Metagenomic Taxonomy-Guided Database-Searching Strategy for Improving Metaproteomic Analysis. *J. Proteome Res.* **17**, 1596–1605 (2018).
21. Beyter, D., Lin, M. S., Yu, Y., Pieper, R. & Bafna, V. ProteoStorm: An Ultrafast Metaproteomics Database Search Framework. *Cell Syst.* **7**, 463-467.e6 (2018).
22. Muth, T. *et al.* The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.* **14**, 1557–1565 (2015).
23. Ivanov, M. V. *et al.* DirectMS1: MS/MS-Free Identification of 1000 Proteins of Cellular Proteomes in 5 Minutes. *Anal. Chem.* **92**, 4326–4333 (2020).
24. Ivanov, M. V. *et al.* Boosting MS1-only Proteomics with Machine Learning Allows 2000 Protein

Identifications in Single-Shot Human Proteome Analysis Using 5 min HPLC Gradient. *J. Proteome Res.* **20**, 1864–1873 (2021).

25. Delegan, Y. *et al.* Complete Genome Sequence of Rhodococcus erythropolis X5, a Psychrotrophic Hydrocarbon-Degrading Biosurfactant-Producing Bacterium. *Microbiol. Resour. Announc.* **8**, e01234-19 (2019).
26. Delegan, Y. *et al.* Complete Genome Analysis of Rhodococcus opacus S8 Capable of Degrading Alkanes and Producing Biosurfactant Reveals Its Genetic Adaptation for Crude Oil Decomposition. *Microorganisms* **10**, 1172 (2022).
27. Iminova, L. *et al.* Physiological and biochemical characterization and genome analysis of Rhodococcus qingshengii strain 7B capable of crude oil degradation and plant stimulation. *Biotechnol. Rep. Amst. Neth.* **35**, e00741 (2022).
28. Delegan, Y. *et al.* Complete Genome Sequence of Rhodococcus qingshengii VT6, a Promising Degrader of Persistent Pollutants and Putative Biosurfactant-Producing Strain. *Microbiol. Resour. Announc.* **11**, e0117921 (2022).
29. Esikova, T. Z., Anokhina, T. O., Abashina, T. N., Suzina, N. E. & Solyanikova, I. P. Characterization of Soil Bacteria with Potential to Degrade Benzoate and Antagonistic to Fungal and Bacterial Phytopathogens. *Microorganisms* **9**, 755 (2021).
30. Frantsuzova, E., Delegan, Y., Bogun, A., Sokolova, D. & Nazina, T. Comparative Genomic Analysis of the Hydrocarbon-Oxidizing Dibenzothiophene-Desulfurizing Gordonia Strains. *Microorganisms* **11**, 4 (2022).
31. Delegan, Y., Valentovich, L., Vetrova, A., Frantsuzova, E. & Kocharovskaya, Y. Complete Genome Sequence of Gordonia sp. 135, a Promising Dibenzothiophene- and Hydrocarbon-Degrading Strain. *Microbiol. Resour. Announc.* **9**, e01450-19 (2020).
32. Delegan, Y., Kocharovskaya, Y., Frantsuzova, E., Streletskii, R. & Vetrova, A. Characterization and genomic analysis of Gordonia alkanivorans 135, a promising dibenzothiophene-degrading strain. *Biotechnol. Rep. Amst. Neth.* **29**, e00591 (2021).
33. Frantsuzova, E. *et al.* Complete Genome Sequence of Gordonia polyisoprenivorans 135, a Promising Degrader of Aromatic Compounds. *Microbiol. Resour. Announc.* **12**, e0005823 (2023).
34. Emelyanova, E. V. & Solyanikova, I. P. Evaluation of phenol-degradation activity of Rhodococcus opacus 1CP using immobilized and intact cells. *Int. J. Environ. Sci. Technol.* **17**, 2279–2294 (2020).
35. Anokhina, T. O. *et al.* Alternative Naphthalene Metabolic Pathway Includes Formation of ortho-Phthalic Acid and Cinnamic Acid Derivatives in the Rhodococcus opacus Strain 3D. *Biochem. Biokhimiia* **85**, 355–368 (2020).
36. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
37. Coudert, E. *et al.* Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinforma. Oxf. Engl.* **39**, btac793 (2023).
38. Choi, M. *et al.* ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. *J. Proteome Res.* **16**, 945–957 (2017).
39. Abdrakhimov, D. A. *et al.* Biosaur: An open-source Python software for liquid chromatography-mass spectrometry peptide feature detection with ion mobility support. *Rapid*

Commun. Mass Spectrom. RCM **20**, e9045 (2021).
40. Ivanov, M. V. *et al.* DirectMS1Quant: Ultrafast Quantitative Proteomics with MS/MS-Free Mass Spectrometry. *Anal. Chem.* **94**, 13068–13075 (2022).
41. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447-452 (2015).
42. Zhang, B., Pirmoradian, M., Zubarev, R. & Käll, L. Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences. *Mol. Cell. Proteomics MCP* **16**, 936–948 (2017).
43. Kazakova, E. M. *et al.* Proteomics-based scoring of cellular response to stimuli for improved characterization of signaling pathway activity. *Proteomics* **23**, e2200275 (2023).
44. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
45. Paauw, A. *et al.* Rapid and reliable discrimination between Shigella species and Escherichia coli using MALDI-TOF mass spectrometry. *Int. J. Med. Microbiol. IJMM* **305**, 446–452 (2015).
46. Lasch, P. *et al.* Identification of Highly Pathogenic Microorganisms by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry: Results of an Interlaboratory Ring Trial. *J. Clin. Microbiol.* **53**, 2632–2640 (2015).
47. Egozarian, N. S. *et al.* Removal of Phenol by Rhodococcus opacus 1CP after Dormancy: Insight into Enzymes' Induction, Specificity, and Cells Viability. *Microorganisms* **12**, 597 (2024).
48. Cheng-Guang, H. & Gualerzi, C. O. The Ribosome as a Switchboard for Bacterial Stress Response. *Front. Microbiol.* **11**, 619038 (2020).
49. Njenga, R., Boele, J., Öztürk, Y. & Koch, H.-G. Coping with stress: How bacteria fine-tune protein synthesis and protein transport. *J. Biol. Chem.* **299**, 105163 (2023).
50. Zhang, S., Scott, J. M. & Haldenwang, W. G. Loss of ribosomal protein L11 blocks stress activation of the Bacillus subtilis transcription factor sigma(B). *J. Bacteriol.* **183**, 2316–2321 (2001).
51. El-Sharoud, W. M. Ribosome inactivation for preservation: concepts and reservations. *Sci. Prog.* **87**, 137–152 (2004).
52. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
53. Franco-Duarte, R. *et al.* Advances in Chemical and Biological Methods to Identify Microorganisms-From Past to Present. *Microorganisms* **7**, 130 (2019).
54. Kwak, H.-L. *et al.* Development of a Rapid and Accurate Identification Method for Citrobacter Species Isolated from Pork Products Using a Matrix-Assisted Laser-Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS). *J. Microbiol. Biotechnol.* **25**, 1537–1541 (2015).
55. Caudill, M. T. & Brayton, K. A. The Use and Limitations of the 16S rRNA Sequence for Species Classification of Anaplasma Samples. *Microorganisms* **10**, 605 (2022).
56. Vargha, M., Takáts, Z., Konopka, A. & Nakatsu, C. H. Optimization of MALDI-TOF MS for strain level differentiation of Arthrobacter isolates. *J. Microbiol. Methods* **66**, 399–409 (2006).
57. Feng, B. *et al.* Effective discrimination of Yersinia pestis and Yersinia pseudotuberculosis by MALDI-TOF MS using multivariate analysis. *Talanta* **234**, 122640 (2021).
58. Trebesius, K., Harmsen, D., Rakin, A., Schmelz, J. & Heesemann, J. Development of rRNA-targeted PCR and in situ hybridization with fluorescently labelled oligonucleotides for detection of Yersinia species. *J. Clin. Microbiol.* **36**, 2557–2564 (1998).

59. Savin, C. *et al.* Genus-wide Yersinia core-genome multilocus sequence typing for species identification and strain characterization. *Microb. Genomics* **5**, e000301 (2019).
60. Dahal, R. H., Choi, Y.-J., Kim, S. & Kim, J. Differentiation of Escherichia fergusonii and Escherichia coli Isolated from Patients with Inflammatory Bowel Disease/Ischemic Colitis and Their Antimicrobial Susceptibility Patterns. *Antibiot. Basel Switz.* **12**, 154 (2023).
61. Liu, R. *et al.* Genomic Characterization of Two Escherichia fergusonii Isolates Harboring mcr-1 Gene From Farm Environment. *Front. Cell. Infect. Microbiol.* **12**, 774494 (2022).
62. Nešvera, J., Rucká, L. & Pátek, M. Catabolism of Phenol and Its Derivatives in Bacteria: Genes, Their Regulation, and Use in the Biodegradation of Toxic Pollutants. *Adv. Appl. Microbiol.* **93**, 107–160 (2015).
63. Rucká, L., Nešvera, J. & Pátek, M. Biodegradation of phenol and its derivatives by engineered bacteria: current knowledge and perspectives. *World J. Microbiol. Biotechnol.* **33**, 174 (2017).
64. Laczi, K. *et al.* Metabolic responses of Rhodococcus erythropolis PR4 grown on diesel oil and various hydrocarbons. *Appl. Microbiol. Biotechnol.* **99**, 9745–9759 (2015).