Analyzing bacterial community compositions by GroEL-taxoproteomics

Simon Klaes[1,2#], Shobhit Madan[1,3], Darja Deobald[1], Myriel Cooper[4], Lorenz Adrian[1,2]

[1]Department of Environmental Biotechnology, Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany

[2]Chair of Geobiotechnology, Technische Universität Berlin, Berlin, Germany

[3]Faculty of Engineering, Ansbach University of Applied Sciences, Ansbach, Germany

[4]Chair of Environmental Microbiology, Technische Universität Berlin, Berlin, Germany

#corresponding author: Simon Klaes (simon.klaes (AT) ufz.de)

Running head: GroEL-taxoproteomics

Word count, Abstract: 250

Word count (excluding Abstract, Materials and methods, Acknowledgments, Author contributions, Competing financial interests, References, Supplementary material, tables, and figure legends): 3797

Abstract

Bacterial communities are fundamental for many health-related, ecological, environmental, and biotechnological processes. With proteins constituting a large share of bacterial biomass and being responsible for metabolic activity, metaproteomics has become increasingly popular for characterizing bacterial communities. However, high sample complexity impairs the detection of low-abundant proteins, and a sample-specific database is necessary for accurate protein identification. Here, we present a targeted metaproteomic approach to quantify bacterial community composition at the family level using the taxonomic marker protein GroEL (synonyms: Cpn60 and Hsp60), which is conserved and abundant in almost all bacteria. We refer to our method as GroEL-taxoproteomics. GroEL-taxoproteomics is based on high-resolution mass spectrometry of extracted proteins and identification of GroEL-derived peptides in the data via a Galaxy workflow and a subsequent Python-based analysis script. Its advantage is that it can be performed with a pre-computed and extendable sample-independent database. GroEL-taxoproteomics was validated by applying it to raw metaproteome data from synthetic microbial communities and real human gut samples. In addition, we experimentally determined the relative detection limit of GroEL-taxoproteomics for detecting low-abundant bacterial taxa. To reduce sample complexity and improve GroEL identification while simultaneously decreasing the instrument time, proteins from crude extracts were pre-separated by SDS-PAGE. The gel bands corresponding to GroEL's molecular weight size of approximately 60 kDa were excised and analyzed by mass spectrometry followed by the GroEL-taxoproteomics workflow. Our data shows that GroEL-taxoproteomics enables fast and low-effort analysis of bacterial communities at good taxonomic resolution. Thus, GroEL-taxoproteomics allows protein-based community analyses at a large scale.

Importance

Metaproteomics has emerged as a powerful tool for investigating microbial dynamics in various environments, including global ecosystems and human microbiomes. However, the current limitations of traditional metaproteomics workflows hinder the widespread application due to the requirement to construct metagenomic databases for every sample site, which is time-consuming and expensive. This study presents a targeted metaproteomic approach that overcomes these limitations by using a sample-independent database. Hence, our approach presents a breakthrough by eliminating the necessity to construct metagenomic databases, significantly reducing costs and working time. In brief, our method allows researchers to rapidly assess bacterial community compositions at the protein level providing a fast, cost-effective, and scalable solution for applying metaproteomics in microbial research. By circumventing the resource-intensive steps of metagenomic database construction, our approach revolutionizes the field, enabling more efficient and accessible analysis of microbial communities.

1 Introduction

Bacterial communities govern our planet, impacting essential ecosystem services such as soil fertility (1), carbon dioxide fixation (2), bioremediation (3), and wastewater treatment (4). One aspect of understanding the functionality of microbial communities is the identification of their microbial composition. Commonly, microbial compositions are described by high-throughput sequencing of 16*S* small subunit ribosomal-RNA (16*S* rRNA) gene amplicons or sequencing of the total community DNA (metagenome sequencing). The 16*S* rRNA gene is ubiquitous in prokaryotic microorganisms and contains hypervariable regions enabling the identification of taxa down to the genus or species level with a gene identity threshold of about 94% and 98%, respectively (5, 6). Several 16*S* rRNA-gene-based fingerprinting techniques, such as denaturing gradient gel electrophoresis (DGGE) (7) and terminal restriction fragment length polymorphism (T-RFLP) analyses (8), have been developed to allow the monitoring of the microbial community dynamics at reasonable costs and speed. However, all these 16*S* rRNA gene-based methods are PCR-dependent and introduce a primer bias (9). Additionally, the variability and multiplicity of the 16*S* rRNA gene as well as the low taxonomic resolution of its short-reads, distort bacterial community analyses (10, 11). Analyzing several hypervariable regions of the 16*S* rRNA gene (12) or other taxonomic marker genes, such as *rpoB* (13) or *groEL* (14), can increase the taxonomic resolution.

Metagenome sequencing can reveal the composition of the microbial community and functional potentials within the community. However, this requires massive sequencing for complex communities with its associated costs and bioinformatics challenges when the study includes more than the few most abundant populations.

Another approach identifies active populations in mixed communities from the extracted total amount of proteins (metaproteomics) (15). Metaproteomics has become increasingly popular due to several advantages: *i)* it relies on abundant protein biomarkers avoiding an amplification step; *ii)* it assesses expressed functions of microbial community members indicating *in situ* activities (16); *iii)* it quantifies biomass distribution per individual species (17), *iv)* in combination with isotope labeling it can link microbial populations to physiological activities, *e.g.*, to the use of a specific carbon source (18), and *v)* it allows to monitor uptake, incorporation and interspecies transfer of isotopically labeled substrates by protein-based stable isotope probing (protein-SIP) (19). Metaproteomics has also become a promising approach for community analysis (16, 20–22). However, metaproteomics often fails to identify low-abundant proteins in complex communities (23–25) and classically relies on a large sample-specific metagenomic database (26). Generating such a sample-specific metagenomic database is a time-consuming and costly task that must be repeated for each new sample site. The challenge is not only that the metagenomic database might be incomplete and fails to include important sequences but also that large databases lead to false-positive identifications (27–31). However, it is possible to focus the community analysis on highly abundant and ubiquitous proteins, such as GroEL, elongation factors, and translation initiation factor 2. This quantitatively allowed the taxonomic differentiation of archaea, bacteria, and eukaryotes in soil samples using a universal database (21).

Our research aimed to develop and evaluate a workflow enabling the quantitative characterization of microbial communities with high taxonomic resolution utilizing protein mass spectrometric data focusing on GroEL as a bacterial biomarker. To achieve this objective, we assembled a sample-independent GroEL database for peptide identification and developed a Python script that facilitates protein and taxonomy inference. Our approach was evaluated by applying it to synthetic microbial communities and

metaproteome data obtained from the human gut, ensuring its robustness and applicability across different microbial ecosystems.

2 Results

2.1 Establishment of a GroEL database and redundancy analysis of protein and peptide sequences

Our download of bacterial GroEL sequences from NCBI in September 2021 resulted in 284,351 GroEL homologous sequences, of which 72,759 were non-redundant. We restricted the download to sequences of a length of 6–1500 amino acids resulting in a median length of 542 amino acids in the database. Prediction of trypsin digestion sites in the non-redundant protein sequences resulted in 9,099,605 peptides, of which 1,882,271 were non-redundant.

To evaluate the taxonomic relevance of GroEL-derived peptides, we calculated their taxon-specificity at different taxonomic levels. For this purpose, we first standardized taxon names in our database according to the "List of Prokaryotic names with Standing in Nomenclature" (LPSN) nomenclature. Taxon names matching the LPSN nomenclature were kept, whereas taxon names not matching the LPSN nomenclature were excluded. Then, peptides exclusively assigned to one taxon were counted as taxon-specific peptides. Our results show that the number of peptides with standardized taxon names and the number of taxon-specific peptides decreased with increasing taxonomic resolution (Table 1).

TABLE 1. Absolute and relative numbers of peptides with a standardized taxon name according to LPSN and taxon-specific peptides in the GroEL database at different taxonomic levels. Leucine and isoleucine were treated as indistinguishable.

Number of non-redundant GroEL peptides	**Taxonomic level**				
	Phylum	Class	Order	Family	Genus
Standardized taxon name[#]	1,805,476 (95.9%)	1,570,462 (83.4%)	1,416,704 (75.3%)	1,348,766 (71.7%)	1,320,069 (70.1%)
Taxon-specific*	1,739,630 (96.4%)	1,492,113 (95.0%)	1,239,269 (87.5%)	1,210,233 (89.8%)	1,127,184 (85.4%)

[#]: Relative numbers are based on the total number of non-redundant peptide sequences.
*: Relative numbers are based on the number of peptides affiliated with an organism with a standardized taxon name.

2.2 Evaluating the quantification potential of GroEL-taxoproteomics and its protein filtering routine

To evaluate the suitability of GroEL peptide mass spectrometry for relative quantification of subpopulations, we experimented with a synthetic mixture of *T. aromatica* K172 and *P. putida* KT2440 in which the two strains were cultivated individually, proteins were extracted, and the crude protein extracts were mixed in defined ratios. The tryptic digests of the mixtures were analyzed by mass spectrometry, and the resulting data was evaluated against two databases: i) a database representing the two full proteomes of *T. aromatica* K172 and *P. putida* KT2440 and ii) our GroEL database. When matching against the database with the two full proteomes, the number of detected peptides assigned to each strain reflected the relative amount of the strains in the mixture (Table 2). In contrast, the number of detected GroEL peptides did not accurately reflect the input ratios. For each mixing ratio, more GroEL peptides were

detected for *T. aromatica* K172 than for *P. putida* KT2440. Only a tendency for less detected K172 peptides and more detected KT2440 peptides was observed with decreasing K172 share in the protein mix. When using the GroEL database instead of the whole proteome database, we detected slightly fewer GroEL peptides for both strains. Furthermore, approximately one-third of the detected GroEL peptides were exclusively linked to organisms not in the mixture, indicating false-positive identifications. The average precursor peak intensity for correctly and falsely detected peptides was similar. In summary, this initial experiment showed that differentiating two strains is possible, but measures are necessary to reduce the false discovery rate.

TABLE 2. Peptides identified from protein extracts of *Thauera aromatica* K172 and *Pseudomonas putida* KT2440 mixed in different ratios. Peptides were identified using the whole proteome database of *T. aromatica* K172 ("K172") containing 3,335 entries and *P. putida* KT2440 ("KT2440") with 5,450 entries, or the GroEL database with 72,759 non-redundant entries. If a peptide was present in both proteomes, it was counted for both organisms. Peptides that were not matching to K172 or KT2440 but other bacteria in the GroEL database were assigned to 'others'. Values represent means of biological triplicates ±SD.

rel. abundance of K172 proteins in the protein mixture (%)	**Database: whole proteomes**				**Database: GroEL**					
	Identified peptides		Identified GroEL peptides		Identified GroEL peptides			Mean intensity of GroEL peptides (x 10^6)		
	K172	KT2440	K172	KT2440	K172	KT2440	others	K172	KT2440	others
100	2441.0 ±201.5	48.7 ±1.2	19.7 ±1.2	0.7 ±0.5	14.7 ±0.5	0.0 ±0.0	8.3 ±4.8	540 ±120	0.0 ±0.0	310 ±130
75	188.7 ±90.7	815.0 ±98.1	18.0 ±0.8	8.3 ±0.9	16.0 ±0.0	6.0 ±0.8	13.0 ±1.4	420 ±51	130 ±21	110 ±50
50	1447.7 ±173.8	1438.7 ±116.3	16.9 ±0.7	11.1 ±0.9	15.0 ±1.2	7.6 ±1.4	12.9 ±4.3	290 ±50	210 ±62	130 ±36
25	772.7 ±171.8	1586.0 ±277.4	13.7 ±1.2	12.3 ±0.9	12.7 ±1.7	8.0 ±2.2	8.0 ±0.8	170 ±46	200 ±53	91 ±39
0	58.3 ±4.5	2961.7 ±152.9	0.0 ±0.0	11.7 ±1.7	0.0 ±0.0	8.3 ±2.1	5.3 ±2.4	0.0 ±0.0	500 ±20	190 ±33

The high number of false-positive detections made it necessary to introduce a filtering step in the data processing. As the average precursor peak intensity for correctly and falsely detected peptides was similar, filtering based on the precursor peak intensity was unsuitable (Table 2). Furthermore, peptides often can be assigned to multiple proteins, known as the protein inference problem. To remove false-positive detections and to infer proteins and taxonomy unambiguously, we developed a custom Python script that filters GroEL protein groups by the Top Rank Count, which only counts peptides once for the largest GroEL protein group (for the description of this filtering routine, see Materials & Methods). Then, we assessed the applicability of this filtering routine by mixing crude extracts of *T. aromatica* K172 and *P. putida* KT2440 in predetermined protein ratios with a total protein mass of 5 µg. Before nLC-MS/MS analysis, GroEL was separated from other proteins by SDS-PAGE, decreasing the heterogeneity of the sample, which resulted in a 1.8-fold increase in the number of detected GroEL peptides. Various Top Rank Count thresholds were tested to eliminate false positive identifications while retaining the ability to detect low-abundant organisms (Figure 1). Implementation of a Top Rank Count threshold ≥ 5 eliminated all false-positive identifications at the genus level. Notably, both organisms were consistently detected at the genus level in all biological

replicates at different protein ratios, demonstrating a relative detection limit of 1% of total protein content for low abundant organisms.

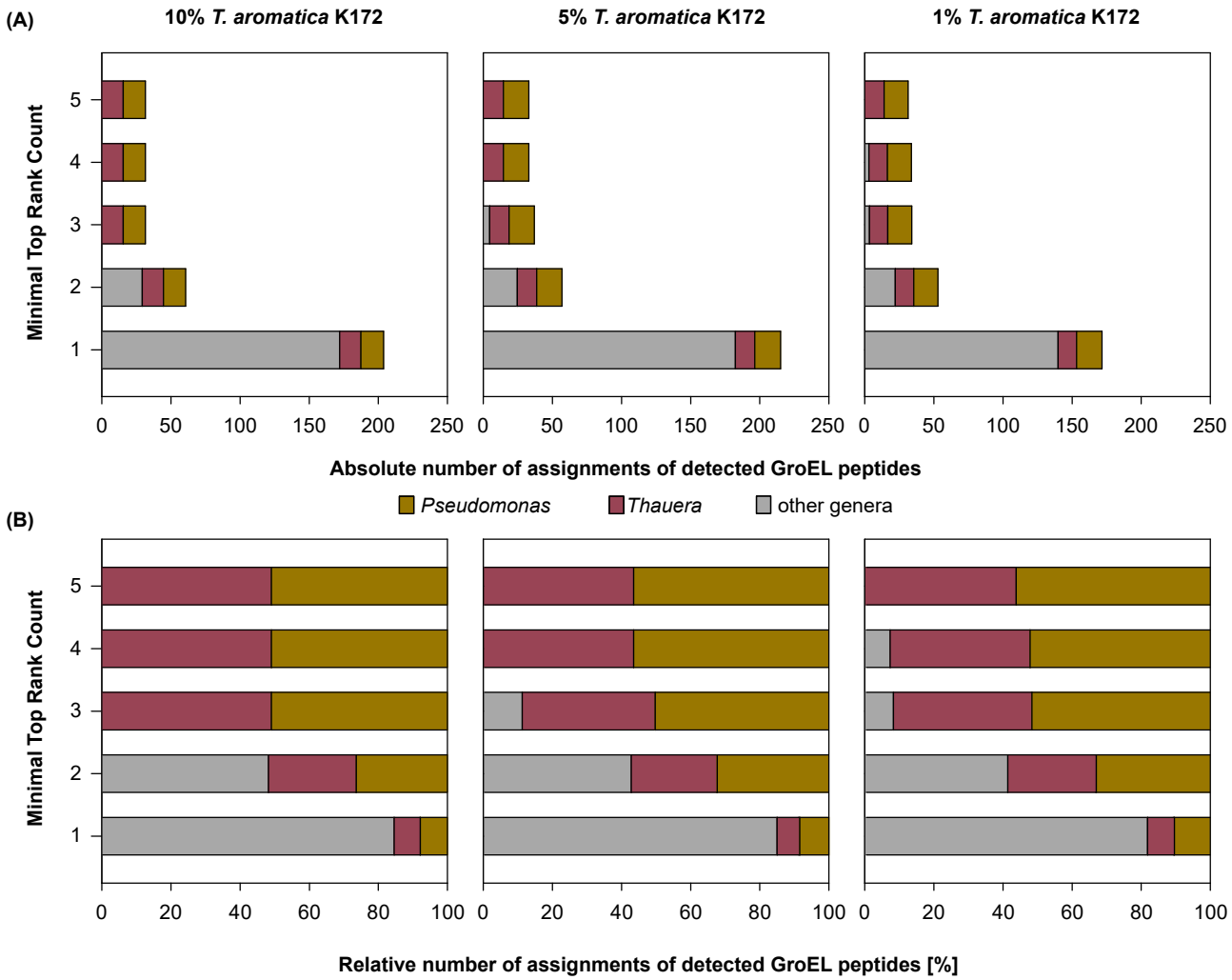


FIGURE 1. Evaluating sensitivity **(A)** and specificity **(B)** of GroEL-taxoproteomics in detecting organisms at the genus level in mixed protein extracts. Mixed protein extracts were composed of *Thauera aromatica* K172 and *Pseudomonas putida* KT2440 proteins in different ratios, and different filtering criteria were applied. Bars represent means of three biological replicates.

2.3 Tracing bacterial community composition in a synthetic biculture

To test the applicability of GroEL-taxoproteomics in tracing bacterial community composition in mixed cultures, we performed a time course analysis of a co-cultured biculture comprising *T. aromatica* K172 and *P. putida* KT2440. In pure cultures, both organisms grew on benzoate as the sole carbon source. However, the lag phase of *T. aromatica* was approximately 15 h longer than that of *P. putida*. Thus, for the co-cultivation of both organisms, the medium was initially inoculated with approximately 10^6 cells mL^{-1} of *T. aromatica* and incubated for 15 h before injecting approximately 10^5 cells mL^{-1} of *P. putida* into the medium and continuing the cultivation. From the biculture, cells were harvested at specified time intervals, and protein crude extracts were prepared. GroEL proteins were enriched in relation to other proteins by SDS-PAGE. The tryptic digests were analyzed by mass spectrometry. Mass spectra were matched with

our GroEL database. Obtained peptides were associated with GroEL protein groups, and resulting GroEL protein groups were filtered using a Top Rank Count ≥ 5. Then, taxonomic classification was performed at the genus level. During the initial 25 h of incubation, only the genus *Thauera* was detected, with *Pseudomonas* not being detected until a total incubation time of 39 h. The relative abundance of *Pseudomonas* was consistently lower than that of *Thauera* (Figure 2). At an incubation time of 48 h, *Pseudomonas* was only detected in two out of three biological replicates, indicating a low abundance close to the detection limit. Additionally, our data revealed discrepancies in quantifying the relative abundance of *Pseudomonas*. The relative abundance of this family was consistently higher when determined by the absolute number of detected GroEL peptides, as compared to the sum of the precursor intensities of these peptides. Overall, this experiment indicated that the biculture was dominated by *Thauera* and that GroEL-taxoproteomics can monitor community compositions over time.

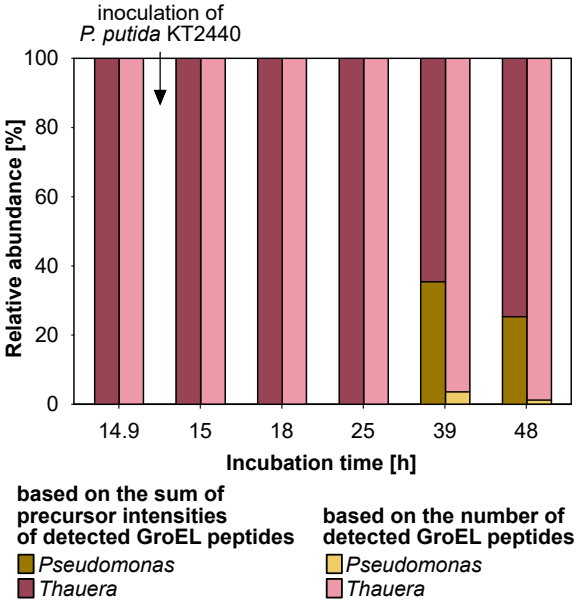


FIGURE 2. Analysis of the composition of a co-cultivated biculture of *Thauera aromatica* K172 and *Pseudomonas putida* KT2440 at various time points by GroEL-taxoproteomics. Cultivation medium DSM 586 was inoculated with *T. aromatica* K172 to initiate the experiment, followed by the inoculation of *P. putida* KT2440 after 14.9 h of incubation. Bacterial composition was quantified based on the sum of the precursor intensities or the number of detected GroEL peptides. Bars represent means of three biological replicates.

2.4 GroEL-taxoproteomics in action I: Reanalyzing raw data of synthetic microbial communities

In the previous experiments, communities of only two bacteria were investigated. To examine if the filtering and quantification techniques are applicable for characterizing more complex bacterial consortia, we reanalyzed proteomic raw data files of three synthetic microbial communities originally assembled by Kleiner et al. (17). The synthetic communities comprised crude protein extracts of one archaeon, one eukaryote, five bacteriophages, three strains of Gram-positive bacteria, and 18 (Mock A and B) or 22 (Mock C) strains of Gram-negative bacteria mixed in different ratios (Supplementary Tables 1–3 of Kleiner et al. (17)). The bacterial population contained 12 (Mock A and B) to 15 (Mock C) bacterial families.

Since our method is specialized in bacterial community characterization, we focused only on these bacterial families.

We again investigated different thresholds of the Top Rank Count to eliminate false-positive identifications while still detecting low-abundance microorganisms (Figure 3). Filtering by a Top Rank Count threshold ≥ 5 (as described above) resulted in some false identifications at the genus level. However, at the family level, this filter effectively eliminated all incorrect identifications while maintaining an acceptable sensitivity for detecting low-abundance organisms. Therefore, we used this threshold for subsequent community analysis at the family level.

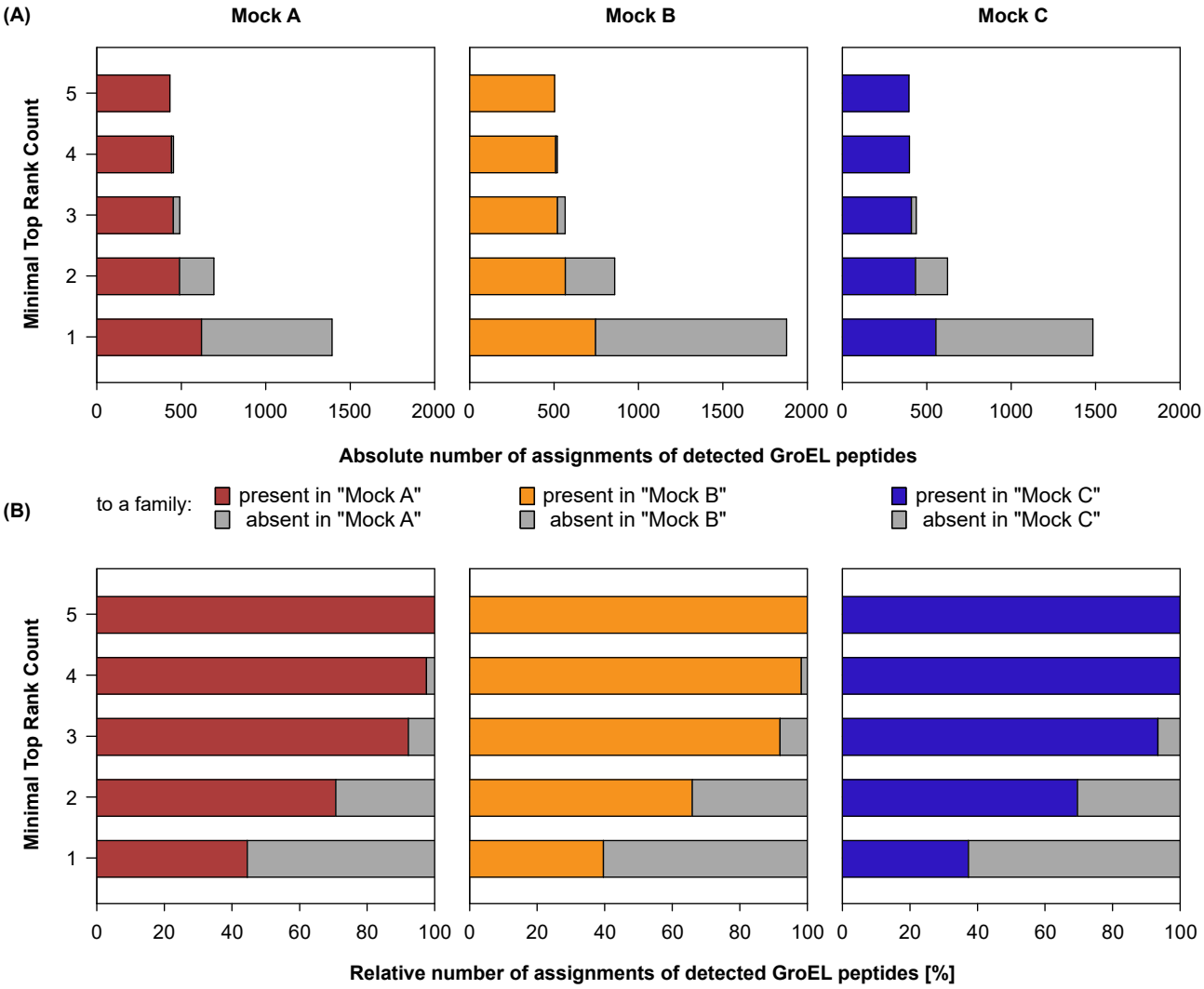


FIGURE 3. Evaluating sensitivity **(A)** and specificity **(B)** of GroEL-taxoproteomics using different Top Rank Count filtering thresholds in detecting peptides assigned to bacterial families that were present in three synthetic microbial communities (Mock A–C) assembled by Kleiner et al. (17). The detected peptides were assigned to specific GroEL proteins, and GroEL proteins with the same set of detected peptides were merged into GroEL groups, which were then filtered by the Top Rank Count ≥ 5. Bars represent the means of four biological replicates.

In the raw data files of the three synthetic communities assembled by Kleiner et al. (17), we detected all bacterial families that were present at a relative bacterial protein content above 2.8%, while no absent

family was falsely detected (Figure 4A). However, the detection of scarce families was inconsistent between the assessments of the different synthetic communities. Specifically, we detected all bacterial families in all four biological replicates of Mock B, which has the most balanced actual protein distribution of 4.8–19.1% of the total protein content for each family (Figure 4A, yellow). In Mock A, all families except the lowest-abundant family *Staphylococcaceae*, accounting for 0.7% of the total bacterial protein content, were detected (Figure 5A, red). On the other hand, in Mock C, which had the highest number of low-abundant bacterial families (relative bacterial protein content ≤ 2.8%), and the most uneven actual protein distribution, with families representing 0.2–41.8% of the total bacterial protein content, we detected all highly abundant bacterial families (relative bacterial protein content > 2.8%) (Figure 4A, blue). However, the detection of low-abundant families with relative bacterial protein content ≤ 2.8% was inconsistent: we detected *Chromobacteriaceae* (1.3%), *Desulfovibrionaceae* (1.0%), *Rhodobacteraceae* (1.0%), *Roseobacteraceae* (1.7%), and *Thermaceae* (1.8%), but not *Staphylococcaceae* (2.8%), *Alteromondaceae (1.0%)*, *Bacillaceae (0.8%)*, *Nitrosomonadaceae (0.7%)*, and *Nitrospiraceae (0.2%)* (Figure 4A, blue).

In addition, we compared two quantification methods: *i)* analyzing the number of detected GroEL peptides and *ii)* analyzing the sum of the precursor intensities of detected GroEL peptides (Figure 4B). Both methods performed similarly for Mock A and Mock B, with the medians centering around zero, indicating high agreement of our analysis and input (Figure 4B, red and yellow). However, differences between the two quantification methods became apparent for Mock C. The method based on peptide counts overestimated most families (*e.g.*, *Thermaceae*) while underestimating others (*e.g.*, *Enterobacteriaceae*) (Figure 4B, blue). The deviation of the measured abundance from the actual input was smaller for the method based on precursor intensities than the method based on peptide counts. Overall, we show that the characterization of complex synthetic communities by GroEL-taxoproteomics is robust at the family level and more consistent when based on the sum precursor intensities of detected GroEL peptides than when based on peptide counts.

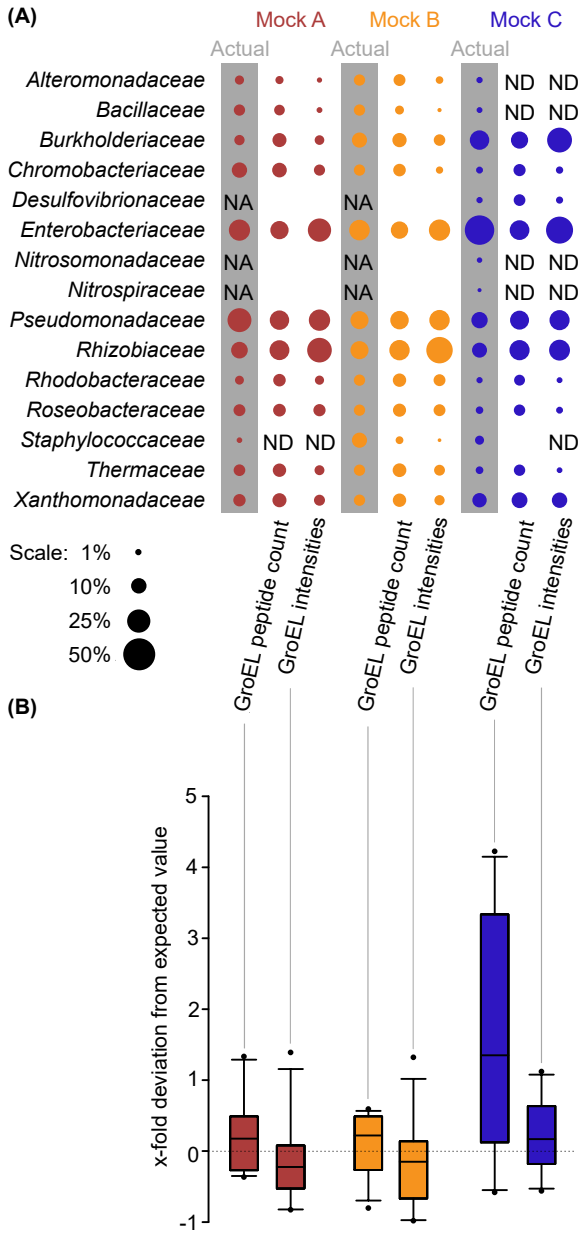


FIGURE 4. Profiling the taxonomic composition of bacterial subpopulations in synthetic microbial communities using GroEL-taxoproteomics. The metaproteome dataset of synthetic communities was obtained from Kleiner et al. (17). We used the same graphic layout to facilitate comparison. **(A)** Comparison of actual community shares (shaded values) with GroEL-based quantification: 'Peptide number' - based on the number of non-redundant GroEL peptides assigned to that family; 'Intensities' - based on the sum of the precursor intensities of the detected GroEL peptides. Quantities are represented as bubble areas (41). The figure includes information on families that were not added (NA) or not detected with our method (ND). Results are means of four replicates. **(B)** Comparison of two quantification methods for GroEL peptides to the actual protein abundance of each family. $x\text{-}fold\ deviation\ from\ the\ expected\ value = \frac{measured\ value}{expected\ value} - 1$. Boxes show the 1st and 3rd quartile, the median, and the whiskers indicate the 10th and 90th percentile, with filled circles representing outliers. ND were removed before plotting. A value of 0 is depicted as a dotted line, indicating equal measurement and input. Negative values indicate underestimation, while positive values indicate overestimation.

2.5 GroEL-taxoproteomics in action II: Reanalyzing raw data of human gut microbiomes

To test the applicability of GroEL-taxoproteomics for characterizing complex bacterial communities, we reanalyzed a gut metaproteomic dataset previously published by García-Durán et al. (22) pertained to six healthy individuals. We focused on the human gut microbiome because it harbors a diverse bacterial community of up to 1,150 species (42). The original study used a protein database based on the human gut microbial gene catalog (9,878,647 sequences) (43) and human proteins from UniProt (74,451 sequences) to detect an average of 11,712 peptides per sample, 56% of which were assigned at the family level. They identified 33 different bacterial families, of which 11 showed a relative family abundance of at least 1% (Supplementary Table 3 of García-Durán et al. (22)).

In our reanalysis, we used our GroEL database to detect peptides and assign them to GroEL protein groups. We filtered GroEL protein groups by the Top Rank Count with a threshold ≥ 5 and evaluated the taxonomy at the family level to achieve high specificity. Our reanalysis closely mirrored the identification of abundant families reported by García-Durán et al. (22) (Figure 5). We detected all 11 families with a relative family abundance of at least 1% (in the original publication). Among the 22 families exhibiting a relative family abundance below 1%, our analysis successfully identified three families: *Streptococcaceae*, *Enterobacteriaceae*, and *Coprobacillaceae* (identified as *Erysipelotrichaceae* by (22)) (Table S1). We did not detect the remaining 19 families, identified with a relative abundance below 1% in the original publication. We observed only minor differences in the abundance of highly abundant families (at least 1% in the original publication), depending on the quantification parameter used (peptide count, peptide spectrum matches (PSMs), or sum precursor intensities). However, for a few families, we noted a substantial discrepancy between our quantification based on GroEL and the quantification based on the human gut microbial gene catalog applied in the original publication. Specifically, GroEL-taxoproteomics led to a higher abundance of *Lachnospiraceae* and *Clostridiaceae* and a lower abundance of *Bacteriodaceae*, *Eubacteriaceae*, and *Prevotellaceae* than in the original analysis (Figure 5).

In the second case study, we demonstrate that GroEL-taxoproteomics can effectively be applied to real and complex samples, as it yields results comparable to traditional metaproteomic approaches in detecting highly abundant bacteria. However, we also identify certain limitations associated with our method, specifically the unreliable detection of low-abundant families.

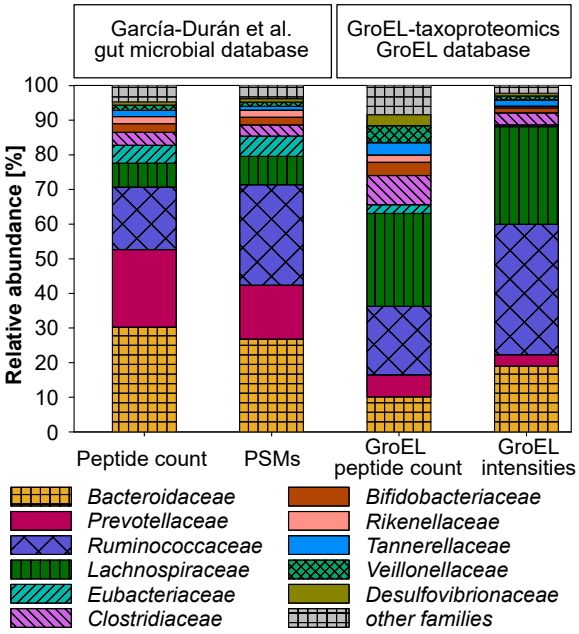
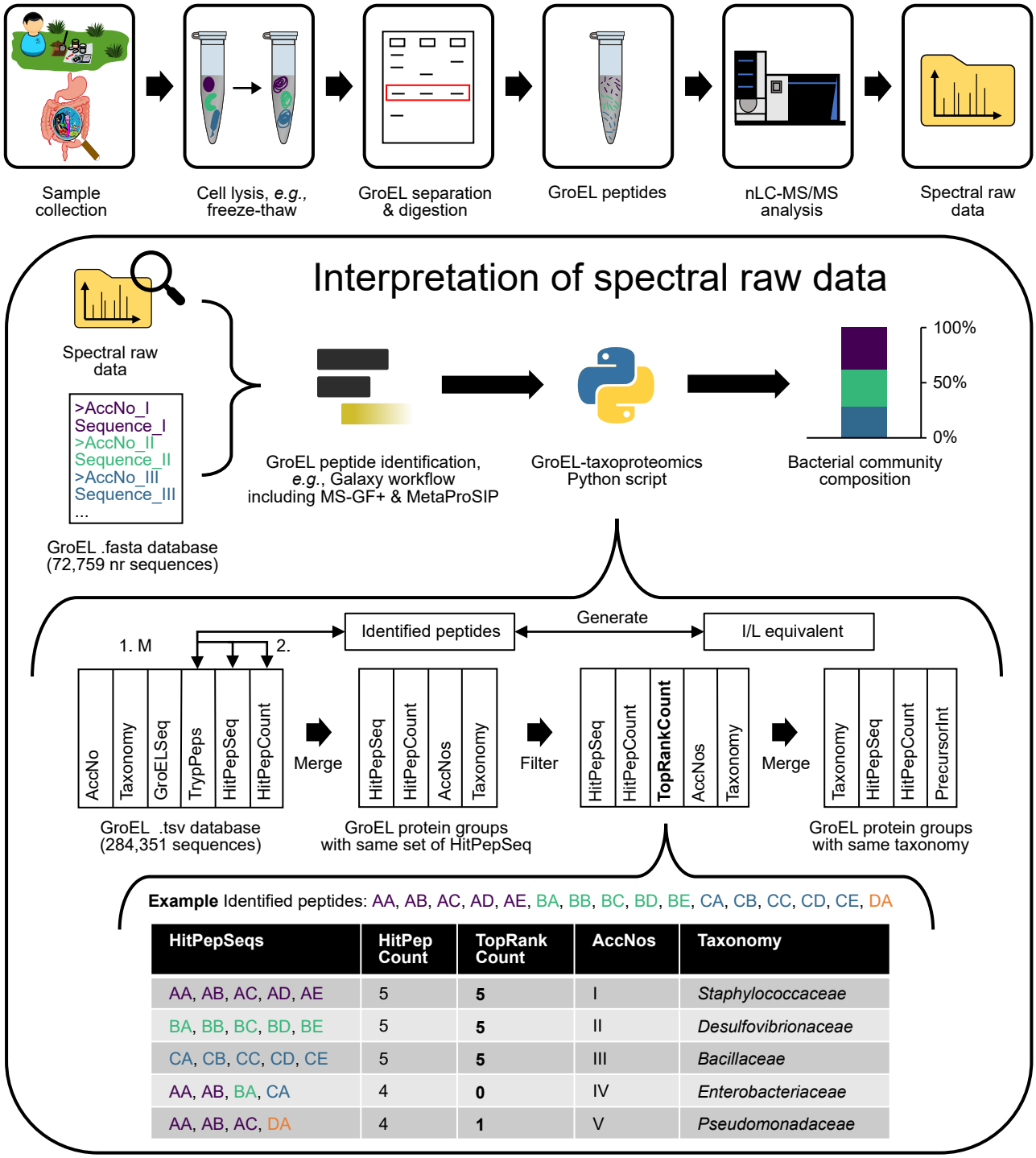


FIGURE 5. The average composition of bacterial communities at the family level, as described from the gut metaproteomic dataset of six healthy individuals by García-Durán et al. (22), compared with our reanalysis of the same data set applying GroEL-taxoproteomics. For comparison, peptides that could not be assigned at the family level were excluded. In the original analysis, the relative abundance of a family was determined by the number of peptide spectrum matches (PSMs) or unique peptides assigned to that family as a proportion of the total number of PSMs or unique peptides assigned at the family level, respectively. In our analysis, the relative abundance of a family was calculated based on the sum of the precursor intensities or the number of non-redundant GroEL peptides assigned to that family, relative to the total sum of the precursor intensities or the number of all filtered GroEL peptides. Families with an abundance below 1% in the original analysis were merged and represented as 'other families'.

3 Discussion

Our findings show that compositions of bacterial communities can be analyzed by using GroEL as a taxonomic marker protein and a sample-independent database. This aspect holds substantial advantages, particularly in situations where rapid or continuous monitoring of community compositions is required. In our work, highly abundant families were reliably identified in all tested samples. We propose a flexible workflow (Figure 6) that can be adapted to a variety of sample preparations, nLC-MS/MS setups, peptide search engines, and quantification strategies. This workflow facilitates the analysis and can be automatized, *e.g.*, as a Galaxy workflow. It provides quantitative identifications with both, sample-independent or sample-specific databases. While the sample-independent GroEL database gave satisfactory results on the family level, a sample-specific GroEL database can improve identifications, as shown with our defined bicultures. We propose introducing the term 'taxoproteomics' as a distinct subcategory of proteomics, which would specify the identification and quantification of taxa in complex microbial communities based on proteins. Adopting the term 'GroEL-taxoproteomics' would then differentiate this particular method from other taxoproteomics approaches.



HitPepSeqs	HitPep Count	TopRank Count	AccNos	Taxonomy
AA, AB, AC, AD, AE	5	**5**	I	*Staphylococcaceae*
BA, BB, BC, BD, BE	5	**5**	II	*Desulfovibrionaceae*
CA, CB, CC, CD, CE	5	**5**	III	*Bacillaceae*
AA, AB, BA, CA	4	**0**	IV	*Enterobacteriaceae*
AA, AB, AC, DA	4	**1**	V	*Pseudomonadaceae*

FIGURE 6. Proposed workflow for analyzing mixed bacterial communities using GroEL-taxoproteomics. The workflow is adaptable to various sample preparations, nano-liquid chromatography-tandem mass spectrometry (nLC-MS/MS) configurations, peptide search engines, and quantification strategies. While SDS-PAGE pre-separation of GroEL proteins can enhance the detection limit, it is not compulsory. The lower part explains the TopRankCount approach used to filter hits. See the text for further explanations.

The analysis of our database shows that most tryptic GroEL peptides are highly taxon-specific similar to the nucleotide sequence of the *groEL* gene that has already been established as a barcode for bacteria (14). Likewise, our workflow identified bacteria at the genus level in low-complexity samples, such as bicultures composed of *T. aromatica* K172 and *P. putida* KT2440. In more complex samples, the identification was reliable and robust at the family level. The classical metaproteomic approach allowed for characterizing the same sample down to the species level (17). However, this better resolution comes with the cost of requiring a sample-specific metagenomic database. Therefore, while GroEL-taxoproteomics currently has a lower taxonomic resolution, it is much faster and cheaper than metaproteomics.

Previous analyses based on GroEL or other taxonomic marker proteins without additional filtering procedures could differentiate populations only at kingdom level (21). This demonstrates the importance of effective peptide filtering and protein/taxonomy inference. Here, we achieved a higher taxonomic resolution by employing a filtering strategy for GroEL protein groups based on the number of peptides counted for top-scored proteins, which we refer to as 'Top Rank Count' (≥ 5). While this approach is not entirely novel, as similar filtering techniques have been implemented by Proteome Discoverer (Thermo Scientific) and MassSieve (44), its integration into our GroEL-taxoproteomics workflow enabled us to attain superior taxonomic resolution compared to previous GroEL-based approaches.

Taxa quantification was more robust based on the sum of the precursor intensities compared to the peptide count and a good estimate for the actual protein input in the complex synthetic communities. Surprisingly, none of the quantification approaches accurately reflected the relative protein amount of the strains in the synthetic biculture. One reason could be that the relative GroEL expression differs between microbes, growth phases, and environmental conditions and that the difference became more apparent for the biculture setup than in the complex communities (45–47).

Our data indicate that the detection of low-abundant taxa strongly depends on the sample complexity and protein distribution across taxa. In a complex synthetic community with an uneven protein distribution, we only reliably detected bacterial families with a relative protein mass of more than 2.8%. However, pre-separation of GroEL by SDS-PAGE increased the identification of GroEL peptides 1.8-fold, resulting in a relative detection limit of 1% in low-complex bicultures. Thus, we hypothesize that using a separable marker protein allows for reducing sample complexity without losing taxonomic information, consequently enhancing the detection of low-abundant taxa. Larger protein input, pooling of gel bands, and longer LC gradients during mass spectrometry may further improve the detection of low-abundant taxa.

In our approach, the detection of organisms depends on the presence of its GroEL sequence in the database. For example, *Roseobacteraceae* was only detected in complex synthetic communities after adding the GroEL sequence of *Roseobacter* sp. AK199 to the database, demonstrating that incomplete databases bias the identification and quantification of taxa. Thus, applying GroEL-taxoproteomics to environmental samples containing many non-sequenced organisms is still challenging (48). However, we successfully applied GroEL-taxoproteomics to human gut proteome data, roughly mirroring the original study's results. Furthermore, we are confident that the rapid growth of sequence databases will massively increase database coverage. In addition, universal primers can amplify a 549–567 bp region of the *groEL* gene, allowing a targeted, fast, and sample-specific extension of the database (49).

At present, our taxoproteomics approach is limited to bacteria because GroEL is highly abundant in bacteria (except very few intracellular *Mycoplasma* and *Ureaplasma* strains (50)), while it is only present in some archaea that most likely acquired it through horizontal gene transfer (51). Consequently, the

current scope of our approach does not encompass the detection of eukaryotes and archaea. To expand the applicability of our method to also detect eukaryotes and archaea, the incorporation of additional putative marker proteins such as ribosomal proteins, chaperonin TCP-1, or thermosome proteins should be considered.

In summary, our study introduces GroEL-taxoproteomics as a rapid and cost-effective method for characterizing bacterial communities at the family level. Our approach offers several advantages over traditional methods. By utilizing a sample-independent database, the need for constructing a sample-specific database is eliminated, saving time and reducing costs associated with database generation. Although the implementation of our method requires access to a protein mass spectrometer, the actual analysis process, with a 60 min LC gradient, is fast, which enables the characterization of up to 20 complex samples per day, making our approach highly efficient for high-throughput analyses. Furthermore, our automatable bioinformatics workflow enables the comprehensive analysis of microbial communities within 48 hours. This can particularly be advantageous for monitoring defined or highly enriched mixed communities over time. Moreover, as the field progresses very fast, the applicability of GroEL-taxoproteomics will expand with increasing GroEL protein sequence databases, the development of automated workflows, and the optimization of mass spectrometric techniques such as ion mobility devices.

4 Materials and methods

4.1 Cultivation, harvesting, and protein extraction

All cultivations were done in 100 mL cotton plugged Erlenmeyer flasks with 50 mL of nitrate-free DSMZ 586 medium at 30 °C in the dark on a rotary shaker under aerobic conditions. *Thauera aromatica* K172 and *Pseudomonas putida* KT2440 were separately grown and harvested during the early stationary phase by centrifuging at 16,000 x g for 10 min. Cell pellets were washed with 50 mM ammonium bicarbonate (AMBIC) buffer, pH 7.9. For the co-cultivation of *P. putida* and *T. aromatica*, we first inoculated the medium with approximately 10^6 cells mL^{-1} of *T. aromatica*. After 15 h of incubation, we injected approximately 10^5 cells mL^{-1} of *P. putida*. Samples were taken directly before adding *P. putida* and after 0; 18; 25; 39; and 48 h incubation time. Cells were harvested from 1 mL samples by centrifugation at 16,000 x g for 10 min, and the pellet was washed with 50 mM AMBIC buffer. Cell pellets were stored at -80 °C until further analysis. For protein extraction, cell pellets were resuspended in 50 mM AMBIC buffer and lysed by three freeze and thaw cycles, utilizing liquid nitrogen and a thermal shaker operating at 40 °C. Additionally, a 30 s treatment in a sonication bath was applied to enhance lysis efficiency. Cell debris and insoluble proteins were removed by centrifuging at 16,000 x g for 10 min. Protein concentrations were determined with the enhanced protocol of the bicinchoninic acid (BCA) assay kit (Pierce, Thermo Scientific). For synthetic bicultures, crude extracts of *T. aromatica* and *P. putida* were mixed in varying protein ratios, resulting a total protein content of 5 µg in 30 µL AMBIC buffer.

4.2 Sample preparation for protein mass spectrometry

Crude extract samples containing 5 µg protein in 30 µL AMBIC buffer were first amended with 40 ng bovine serum albumin (BSA) as a quality control measure. Then, the samples were sequentially treated with 62.5 mM dithiothreitol and 128 mM 2-iodoacetamide to reduce and alkylate cysteine residues. Subsequently, proteins were digested overnight with 0.63 µg trypsin (Promega). Tryptic digestion was stopped by adding formic acid to a final concentration of 1.8% (v/v). Undigested and precipitated proteins

were removed by centrifugation at 16,000 x g for 10 min, and samples were dried by vacuum centrifugation. Subsequently, the peptides were resuspended in 0.1% (v/v) formic acid and desalted using C_{18} Zip tips (Pierce, Thermo Scientific). The peptides were again dried by vacuum centrifugation before being reconstituted in 50 µL of 0.1% (v/v) formic acid for nano-liquid chromatography coupled to tandem mass spectrometry (nLC-MS/MS) analysis.

In separate experiments, we decreased the heterogeneity of the crude extracts before nLC-MS/MS analysis by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). For SDS-PAGE, the crude extract containing 5 µg protein in 30 µL AMBIC buffer was mixed with 10 µL SDS reducing buffer and incubated at 95 °C for 10 min to fully denature proteins. After cooling to room temperature, samples were subjected to SDS-PAGE at 110 V for 60–90 min and Coomassie stained. Protein bands with a molecular weight of approximately 60 kDa, corresponding to the size of GroEL, were excised. The gel slices were destained with acetonitrile. Then, proteins captured within the gel slice were reduced and alkylated by sequential incubation in 50 µL of 10 mM dithiothreitol and 50 µL of 100 mM 2-iodoacetamide for 60 min each. Subsequently, 40 ng of reduced and alkylated BSA was added as a quality control measure. Proteins were then digested with 0.1 µg trypsin (Promega) at 37 °C overnight. After digestion, peptides were extracted from gel pieces with 50% (v/v) acetonitrile and 5% (v/v) formic acid and dried by vacuum centrifugation. Next, peptides were resuspended in 20 µL of 0.1% (v/v) formic acid, desalted using C_{18} Zip tips (Pierce, Thermo Scientific), and repeatedly dried by vacuum centrifugation. Finally, the desalted peptides were resuspended in 50 µL of 0.1% (v/v) formic acid for nLC-MS/MS analysis.

4.3 Mass spectrometry

Desalted peptides were separated on an UltiMate 3000 RSLCnano high-performance nano-UPLC system (Thermo Scientific) coupled to an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific) via a TriVersa NanoMate nano-electrospray ionization (nano-ESI) ion source (Advion). For in-solution digested samples, 3 µL of the peptide solution was injected, whereas 5 µL was injected for in-gel digested samples. Peptides were first loaded for 3 min onto the Acclaim PepMap 100 C_{18} trap column (75 µm x 15 cm, 3 µm material, Thermo Scientific) with 5 µL min^{-1} of 3.2% (v/v) acetonitrile in water at 0.1% formic acid. Then, the trap column was switched into line with the Acclaim PepMap 100 C_{18} analytical column (75 µm x 25 cm, 3 µm material, Thermo Scientific) heated up to 35 °C to separate peptides at a flow rate of 0.3 µL min^{-1} using a gradient of 145 min (for in-solution digested samples) or 60 min (for in-gel digested samples) from 3.2% to 72% (v/v) acetonitrile in water at 0.1% (v/v) formic acid. The ion source operated in positive mode at a spray voltage of 2.4 kV and a source temperature of 275 °C. The mass spectrometer run in data-dependent mode with a cycle time of 3 s. Internal mass calibration was performed using a lock mass of 445.12003 *m/z*. Precursor ions were scanned in the Orbitrap mass analyzer over a range of 350–2000 *m/z* with a resolution of 120,000, automatic gain control (AGC) target of 4 x 10^5 ions, and a maximum injection time of 50 ms. Precursor ions with a minimum intensity of 5 x 10^4 and charge state of +2 and +3 (for in-solution digested samples) or +2 to +7 (for in-gel digested samples) were selected and isolated by the quadrupole in a window of 1.6 *m/z* accumulating to an AGC target of 5 x 10^4 ions with a maximum injection time of 54 ms (for in-solution digested samples) and 120 ms (for in-gel digested samples). The isolated precursor ions were fragmented using higher energy collisional dissociation (HCD) at 30% relative collision energy. Fragment ions were scanned with the Orbitrap mass analyzer at a resolution of 30,000 (for in-solution digested samples) or 60,000 (for in-gel digested samples), respectively. Precursor ions with the same mass (±10 ppm) were excluded for further precursor selection for 30 s.

4.4 Databases for mass spectrometric analysis

We assembled a comprehensive dataset of bacterial GroEL protein sequences by downloading all available amino acid sequences from the National Center for Biotechnology Information (NCBI) in GenBank format on September 7, 2021 (32). The dataset was used to generate a file in fasta format as input for peptide identification search engines (see below) and to generate a second file with tab-separated values (tsv) for protein and taxonomy inferences. The tsv-database included in separated columns accession number, amino acid sequence, expected tryptic peptides, and taxonomic classification at various levels (kingdom, phylum, class, order, family, and genus). Tryptic peptides were calculated with pyOpenMS (33), accepting two missed cleavages and a peptide length of 6–144 amino acids. The taxonomic classification was updated by cross-referencing the "List of Prokaryotic names with Standing in Nomenclature" (LPSN) on October 18, 2022 (34). We also included GroEL sequences from the separately published genomes of *Roseobacter* sp. AK199 and *Chromobacterium violaceum* CV026 (17). The common Repository of Adventitious Proteins (cRAP) was appended to the database to identify common contaminants. Additionally, we downloaded the whole proteome databases of *T. aromatica* K172 (CP028339.1) and *P. putida* KT2440 (NC_002947.4) from NCBI.

4.5 Analysis of mass-spectrometric raw data files

Thermo raw files were first converted into mzML format using msConvert (ProteoWizard) (35). Next, we analyzed the mzML files using a customized MetaProSIP workflow (36) on the Galaxy platform (37). Briefly, peptides were identified using MS-GF+ (38) and MetaProSIP (39). We accepted two missed trypsin cleavages, peptide lengths of 6–40 amino acids, and precursor *m/z* deviations of 5 ppm. Oxidation of methionine and carbamidomethylation of cysteine were set as dynamic and static modifications, respectively. The false discovery rate of identified peptide sequences was kept below 0.01, based on the SpecEValue calculated by MS-GF+. The peptide-centric output of MetaProSIP was used as input for a custom GroEL-taxoproteomics Python script. The isobaric amino acids leucine (L) and isoleucine (I) were treated as indistinguishable. Peptides listed in the cRAP database were excluded from the analysis with the customized GroEL-taxoproteomics Python script. The Python script assigns detected peptides to each possible GroEL protein in the tsv-database based on matching sequences. GroEL proteins with the same set of detected peptides are merged into protein groups and sorted by the number of detected peptides in descending order. For each GroEL protein group, only the detected peptides not present in another GroEL protein group containing more peptides are counted, resulting in a Top Rank Count. GroEL protein groups are then filtered by the Top Rank Count with a threshold (≥ 5), and the taxonomy of the remaining GroEL protein groups is read from the tsv-file. For each taxonomic level, the most frequent taxonomic description and its frequency within the GroEL protein group are calculated. The most frequent taxonomic description is used to merge GroEL protein groups into taxonomic groups at the level of interest, which are then quantified by the sum of the precursor intensities (based on peak area) or the number of non-redundant peptides within the taxonomic group. In the current version of the Python script (version 1.0.0), non-redundant peptides are only counted once for the same taxonomic group and multiple times for different taxonomic groups. Furthermore, the precursor intensity of peptides assigned to multiple taxonomic groups is distributed proportionately to the total number of peptides assigned to each group. To evaluate our workflow, we used published metaproteome data from synthetic microbial communities (PRIDE repository PXD006118). Specifically, we used the raw data acquired with an LC gradient of 260 min (run 4 and 5) as this provided sufficient data for community analysis (17). The identified peptides of technical replicates

were pooled before running our GroEL-taxoproteomics Python script. Furthermore, we applied our analysis workflow to a metaproteome dataset of human gut microbiota (PRIDE repository PXD020786) (22). Identified peptides of biological replicates were pooled before running the GroEL-taxoproteomics Python script to allow comparison with the original study (Supplementary Table 3 of García-Durán et al. (22)).

4.6 Data availability

The mass spectrometric raw data and protein sequence databases have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (40) with the data set identifier PXD043106 (Username: reviewer_pxd043106@ebi.ac.uk

Password: KzeQ5W8r). The GroEL-taxoproteomics Python script is publicly available via GitLab (https://git.ufz.de/klaes/groel-taxoproteomics/). Any other relevant data will be made available upon request.

5 Acknowledgments

This project was funded by the German Research Foundation (DFG) as part of the Research Training Group "Urban Water Interfaces (UWI)" (GRK 2032/2, Project H8: "Microbial transformation of mobile halogenated aromatics in redox gradients of urban hyporheic zones"). Open access was supported by the UFZ library.

We acknowledge Benjamin Scheer for assistance with mass spectrometric measurements. Furthermore, we thank Matthias Bernt and Bryson Gibbons for rapid troubleshooting regarding the Galaxy server and MS-GF+.

6 Author contributions

Simon Klaes: conceptualization (lead), data curation (lead), formal analysis, investigation (equal), methodology (lead), software (lead), visualization, writing – original draft preparation, writing – review & editing (equal) **Shobhit Madan:** data curation (supporting), investigation (equal) **Darja Deobald:** conceptualization (supporting), methodology (supporting), writing – review & editing (equal) **Myriel Cooper:** conceptualization (supporting), funding acquisition (equal), writing – review & editing (equal) **Lorenz Adrian:** conceptualization (lead), funding acquisition (equal), methodology (supporting), software (supporting), writing – review & editing (equal)

7 Competing financial interests

The authors declare that they have no competing financial interests.

8 References

1. Siciliano SD, Palmer AS, Winsley T, Lamb E, Bissett A, Brown M V., van Dorst J, Ji M, Ferrari BC, Grogan P, Chu H, Snape I. 2014. Soil fertility is associated with fungal and bacterial richness, whereas pH is associated with community composition in polar soil microbial communities. Soil Biol Biochem 78:10–20.
2. Berg IA. 2011. Ecological aspects of the distribution of different autotrophic CO_2 fixation pathways. Appl Environ Microbiol 77:1925–1936.
3. Lopes PRM, Cruz VH, de Menezes AB, Gadanhoto BP, Moreira BR de A, Mendes CR, Mazzeo DEC, Dilarri G, Montagnolli RN. 2022. Microbial bioremediation of pesticides in agricultural soils: an integrative review on natural attenuation, bioaugmentation and biostimulation. Rev Environ Sci

Biotechnol 21:851–876.

4. Al Ali AA, Naddeo V, Hasan SW, Yousef AF. 2020. Correlation between bacterial community structure and performance efficiency of a full-scale wastewater treatment plant. J Water Process Eng 37:101472.

5. Kim M, Oh HS, Park SC, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16*S* rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351.

6. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16*S* rRNA gene sequences. Nat Rev Microbiol 12:635–645.

7. Muyzer G, De Waal EC, Uitterlinden AG. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16*S* rRNA. Appl Environ Microbiol 59:695.

8. Liu WT, Marsh TL, Cheng H, Forney LJ. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16*S* rRNA. Appl Environ Microbiol 63:4516.

9. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA. 2015. The truth about metagenomics: quantifying and counteracting bias in 16*S* rRNA studies. BMC Microbiol 15:66.

10. Větrovský T, Baldrian P. 2013. The variability of the 16*S* rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One 8:e57923.

11. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029.

12. Liu Z, Desantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 36:e120.

13. Ogier JC, Pagès S, Galan M, Barret M, Gaudriault S. 2019. *rpoB*, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. BMC Microbiol 19:1–16.

14. Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE. 2012. The chaperonin-60 universal target is a barcode for bacteria that enables *de novo* assembly of metagenomic sequence data. PLoS One 7:e49755.

15. Bastida F, Torres IF, Moreno JL, Baldrian P, Ondoño S, Ruiz-Navarro A, Hernández T, Richnow HH, Starke R, García C, Jehmlich N. 2016. The active microbial diversity drives ecosystem multifunctionality and is physiologically related to carbon availability in Mediterranean semi-arid soils. Mol Ecol 25:4660–4673.

16. Heyer R, Schallert K, Siewert C, Kohrs F, Greve J, Maus I, Klang J, Klocke M, Heiermann M, Hoffmann M, Püttker S, Calusinska M, Zoun R, Saake G, Benndorf D, Reichl U. 2019. Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. Microbiome 7:69.

17. Kleiner M, Thorson E, Sharp CE, Dong X, Liu D, Li C, Strous M. 2017. Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun 8:1588.

18. Kleiner M, Dong X, Hinzke T, Wippler J, Thorson E, Mayer B, Strous M. 2018. Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. Proc Natl Acad Sci U S A 115:E5576–E5584.

19. Taubert M, Vogt C, Wubet T, Kleinsteuber S, Tarkka MT, Harms H, Buscot F, Richnow HH, Von Bergen M, Seifert J. 2012. Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium. ISME J 6:2291–2301.

20. Ouyang WY, Su JQ, Richnow HH, Adrian L. 2019. Identification of dominant sulfamethoxazole-degraders in pig farm-impacted soil by DNA and protein stable isotope probing. Environ Int 126:118–126.

21. Starke R, Fiore-Donno AM, White RA, Parente Fernandes ML, Martinović T, Bastida F, Delgado-Baquerizo M, Jehmlich N. 2022. Biomarker metaproteomics for relative taxa abundances across

soil organisms. Soil Biol Biochem 175:108861.

22. García-Durán C, Martínez-López R, Zapico I, Pérez E, Romeu E, Arroyo J, Hernáez ML, Pitarch A, Monteoliva L, Gil C. 2021. Distinct human gut microbial taxonomic signatures uncovered with different sample processing and microbial cell disruption methods for metaproteomic analysis. Front Microbiol 12:1849.

23. Weston LA, Bauer KM, Hummon AB. 2013. Comparison of bottom-up proteomic approaches for LC-MS analysis of complex proteomes. Anal Methods 5:4615–4621.

24. Hu A, Noble WS, Wolf-Yadlin A. 2016. Technical advances in proteomics: new developments in data-independent acquisition. F1000Res 5:419.

25. Van Den Bossche T, Kunath BJ, Schallert K, Schäpe SS, Abraham PE, Armengaud J, Arntzen M, Bassignani A, Benndorf D, Fuchs S, Giannone RJ, Griffin TJ, Hagen LH, Halder R, Henry C, Hettich RL, Heyer R, Jagtap P, Jehmlich N, Jensen M, Juste C, Kleiner M, Langella O, Lehmann T, Leith E, May P, Mesuere B, Miotello G, Peters SL, Pible O, Queiros PT, Reichl U, Renard BY, Schiebenhoefer H, Sczyrba A, Tanca A, Trappe K, Trezzi JP, Uzzau S, Verschaffelt P, von Bergen M, Wilmes P, Wolf M, Martens L, Muth T. 2021. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. Nat Commun 12:7305.

26. Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, Muth T, Rapp E, Martens L, Addis MF, Uzzau S. 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome 4:51.

27. Rechenberger J, Samaras P, Jarzab A, Behr J, Frejno M, Djukovic A, Sanz J, González-Barberá EM, Salavert M, López-Hontangas JL, Xavier KB, Debrauwer L, Rolain JM, Sanz M, Garcia-Garcera M, Wilhelm M, Ubeda C, Kuster B. 2019. Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. Proteomes 7:2.

28. Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, Pagnozzi D, Addis MF, Uzzau S. 2013. Evaluating the impact of different sequence databases on metaproteome analysis: Insights from a lab-assembled microbial mixture. PLoS One 8:e82981.

29. Sticker A, Martens L, Clement L. 2017. Mass spectrometrists should search for all peptides, but assess only the ones they care about. Nat Methods 14:643–644.

30. Colaert N, Degroeve S, Helsens K, Martens L. 2011. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 10:5555–5561.

31. Muth T, Kolmeder CA, Salojärvi J, Keskitalo S, Varjosalo M, Verdam FJ, Rensen SS, Reichl U, de Vos WM, Rapp E, Martens L. 2015. Navigating through metaproteomics data: a logbook of database searching. Proteomics 15:3439–3453.

32. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2020. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 48:D9–D16.

33. Röst HL, Schmitt U, Aebersold R, Malmström L. 2014. pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. Proteomics 14:74–77.

34. Parte AC, Carbasse JS, Meier-Kolthoff JP, Reimer LC, Göker M. 2020. List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ. Int J Syst Evol Microbiol 70:5607–5612.

35. Kessner D, Chambers M, Burke R, Agus D, Mallick P. 2008. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24:2534–2536.

36. Jehmlich N, von Bergen M. 2020. Protein-based stable isotope probing (Protein-SIP): applications for studying aromatic hydrocarbon degradation in microbial communities, p. 277–284. *In* Anaerobic utilization of hydrocarbons, oils, and lipids. Springer International Publishing, Cham.

37. Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Ech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537–W544.

38. Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5:5277.

39. Sachsenberg T, Herbst FA, Taubert M, Kermer R, Jehmlich N, Von Bergen M, Seifert J, Kohlbacher O. 2015. MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. J Proteome Res 14:619–627.

40. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, Vizcaíno JA. 2022. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res 50:D543–D552.

41. Zaikova E, Walsh DA, Stilwell CP, Mohn WW, Tortell PD, Hallam SJ. 2010. Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. Environ Microbiol 12:172–191.

42. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J, Antolin M, Artiguenave F, Blottiere H, Borruel N, Bruls T, Casellas F, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Forte M, Friss C, Van De Guchte M, Guedon E, Haimet F, Jamet A, Juste C, Kaci G, Kleerebezem M, Knol J, Kristensen M, Layec S, Le Roux K, Leclerc M, Maguin E, Melo Minardi R, Oozeer R, Rescigno M, Sanchez N, Tims S, Torrejon T, Varela E, De Vos W, Winogradsky Y, Zoetendal E. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65.

43. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, Bork P, Wang J, Pons N, Le Chatelier E, Batto JM, Kennedy S, Haimet F, Winogradski Y, Pelletier E, Lepaslier D, Artiguenave F, Bruls T, Weissenbach J, Turner K, Parkhill J, Antolin M, Casellas F, Borruel N, Varela E, Torrejon A, Denariaz G, Derrien M, van Hylckama Vlieg JET, Viega P, Oozeer R, Knoll J, Rescigno M, Brechot C, M'rini C, Mérieux A, Yamada T, Tims S, Zoetendal EG, Kleerebezem M, de Vos WM, Cultrone A, Leclerc M, Juste C, Guedon E, Delorme C, Layec S, Khaci G, van de Guchte M, Vandemeulebrouck G, Jamet A, Dervyn R, Sanchez N, Blottière H, Maguin E, Renault P, Tap J, Mende DR. 2014. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 32:834–841.

44. Slotta DJ, McFarland MA, Markey SP. 2010. MassSieve: panning MS/MS peptide data for proteins. Proteomics 10:3035–3039.

45. Klančnik A, Botteldoorn N, Herman L, Možina SS. 2006. Survival and stress induced expression of *groEL* and *rpoD* of *Campylobacter jejuni* from different growth phases. Int J Food Microbiol 112:200–207.

46. Kupper M, Gupta SK, Feldhaar H, Gross R. 2014. Versatile roles of the chaperonin GroEL in microorganism-insect interactions. FEMS Microbiol Lett 353:1–10.

47. Gifford SM, Sharma S, Booth M, Moran MA. 2013. Expression patterns reveal niche diversification in a marine microbial assemblage. ISME J 7:281–298.

48. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.

49. Hill JE, Town JR, Hemmingsen SM. 2006. Improved template representation in *cpn60* polymerase chain reaction (PCR) product libraries generated from complex templates by application of a specific mixture of PCR primers. Environ Microbiol 8:741–746.

50. Musatovova O, Dhandayuthapani S, Baseman JB. 2006. Transcriptional heat shock response in the smallest known self-replicating cell, *Mycoplasma genitalium*. J Bacteriol 188:2845–2855.

51. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz R, Martínez-Arias R, Henne A, Wiezer A, Bäumer S, Jacobi C, Brüggemann H, Lienard T, Christmann A, Bömeke M, Steckel S, Bhattacharyya A, Lykidis A, Overbeek R, Klenk H, Gunsalus R, Fritz H, Gottschalk G. 2002. The

genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol 4:453–461.

9 Supplemental material

Table S1: Composition of bacterial families with a relative family abundance below 1% in the original study of the human gut microbiome dataset published by García-Durán et al. (22).

	García-Durán et al.				Our analysis based on GroEL			
Family	Number of unique peptides	#PSMs	relative abundance based on unique peptides	relative abundance based on PSMs	Number of unique peptides	peptide sum intensity	relative abundance based on unique peptides	relative abundance based on intensity
Acidaminococcaceae	135	385	0.86%	0.70%			0.00%	0.00%
Sutterellaceae	129	344	0.82%	0.63%			0.00%	0.00%
Oscillospiraceae	124	322	0.79%	0.59%			0.00%	0.00%
Erysipelotrichaceae	71	180	0.45%	0.33%			0.00%	0.00%
Streptococcaceae	71	180	0.45%	0.33%	18	3.48×10^9	3.80%	0.92%
Coriobacteriaceae	58	178	0.37%	0.33%			0.00%	0.00%
Odoribacteraceae	34	78	0.22%	0.14%			0.00%	0.00%
Eggerthellaceae	20	64	0.13%	0.12%			0.00%	0.00%
Enterobacteriaceae	17	32	0.11%	0.06%	8	1.08×10^9	1.69%	0.28%
Akkermansiaceae	15	32	0.10%	0.06%			0.00%	0.00%
Acetobacteraceae	10	26	0.06%	0.05%			0.00%	0.00%
Methanobacteriaceae	9	15	0.06%	0.03%			0.00%	0.00%
Chlamydiaceae	9	20	0.06%	0.04%			0.00%	0.00%
Lactobacillaceae	7	57	0.04%	0.10%			0.00%	0.00%
Porphyromonadaceae	7	35	0.04%	0.06%			0.00%	0.00%
Barnesiellaceae	6	12	0.04%	0.02%			0.00%	0.00%
Rhodospirillaceae	4	10	0.03%	0.02%			0.00%	0.00%
Selenomonadaceae	4	14	0.03%	0.03%			0.00%	0.00%
Flavobacteriaceae	4	16	0.03%	0.03%			0.00%	0.00%
Rhodobacteraceae	3	21	0.02%	0.04%			0.00%	0.00%
Succinivibrionaceae	3	8	0.02%	0.01%			0.00%	0.00%
Spirochaetaceae	3	24	0.02%	0.04%			0.00%	0.00%
Peptoniphilaceae	3	10	0.02%	0.02%			0.00%	0.00%
Coprobacillaceae	0	0	0.00%	0.00%	14	3.82×10^9	2.95%	1.01%