This is the preprint of the contribution published as:

Starke, R., Fiore-Donno, A.M., White III, R.A., Parente Fernandes, M.L., Martinović, T., Bastida, F., Delgado-Baquerizo, M., **Jehmlich, N.** (2022):
Biomarker metaproteomics for relative taxa abundances across soil organisms
Soil Biol. Biochem. **175** , art. 108861

The publisher's version is available at:

http://dx.doi.org/10.1016/j.soilbio.2022.108861

Title: Biomarker metaproteomics for relative taxa abundances across soil organisms

Running title: Biomarker metaproteomics

Robert Starke[1], Anna Maria Fiore-Donno[2,3], Richard Allen White III[4,5,6], Maysa Lima Parente Fernandes[1], Tijana Martinovic[1], Felipe Bastida[7], Manuel Delgado-Baquerizo[8,9], Nico Jehmlich[10]

[1]Institute of Microbiology of the Czech Academy of Sciences, Vídeňská 1083, 14220 Praha 4, Czech Republic
[2]Terrestrial Ecology Group, Institute of Zoology, University of Cologne, Cologne, Germany
[3]Cluster of Excellence on Plant Sciences (CEPLAS), Cologne, Germany
[4]Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, Charlotte, North Carolina, USA
[5]Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, Kannapolis, North Carolina, USA
[6]Australian Centre for Astrobiology, University of New South Wales Sydney, Sydney, Australia
[7]CEBAS-CSIC, Campus Universitario de Espinardo, Murcia, E-30100, Spain
[8]Laboratorio de Biodiversidad y Funcionamiento Ecosistemico, Instituto de Recursos Naturales y Agrobiología de Sevilla (IRNAS), CSIC, Av. Reina Mercedes 10, E-41012, Sevilla, Spain
[9]Unidad Asociada CSIC-UPO (BioFun), Universidad Pablo de Olavide, 41013 Sevilla, Spain.
[10]Department of Molecular Systems Biology, Helmholtz-Centre for Environmental Research, UFZ, Permoserstr. 15, 04318 Leipzig, Germany

*Corresponding author: robert.starke@biomed.cas.cz

Keywords: metaproteomics, biomarker, relative taxa abundances

The authors declare no conflict of interest.

Abstract

Soil organisms are often classified using methods targeting individual groups of taxa (e.g., bacteria, fungi and invertebrates), which hampers our ability to directly compare the relative abundance of different groups across environmental gradients. We posit that the use of protein biomarkers could help to provide a more real representation of the cross-kingdom soil microbial populations. Here, we tested if the abundant proteins ATP synthase F(0) complex (ATPS), elongation factors (EF), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), GroEL, pyruvate dehydrogenase (PyrDH), RNA polymerase beta chain (RNAP), and translation initiation factor 2 (TIF) could be used to describe the taxonomic composition of microbial communities. As positive control, we used a mock community with different relative abundances of algae, archaea, bacteria, and viruses. We tested this approach on a previously published soil metaproteomes from which we randomly selected samples from forests, grasslands, and shrublands (each n=10). Unfortunately, the biomarker approach is not feasible for viruses as these organisms do not share single genes. All biomarkers showed decent accuracy to determine the relative abundances of archaea, bacteria, and eukaryota in the mock community. However, false positive hits dominated on phylum level probably due to sequence homology. Archaeal proteins were only detected in the soil samples when EF was used as biomarker at an abundance of 0.7%. Bacteria dominated the EF-metaproteome and were most abundant in shrublands (64.4%) while eukaryotes were more abundant in forests (25.6%). In compliance with previously published results, the correlation analysis revealed the impact of mean annual temperature and pH on both bacteria and eukaryota. Our approach not only shows the potential to use biomarker metaproteomics to unveil the relative taxa abundances across soil organisms but also the need to create mock communities comprising members of all soil taxa.

Introduction

Topsoil proteins catalyze a multitude of biological functions [1,2] that allow microbial communities to drive essential ecosystem services such as soil fertility, climate change regulation and waste decomposition [3–5]. DNA-based methods are known to be limited by their reduced capacity to account for activity and processes, restricting the capacity of metagenomics to efficiently predict ecosystem functions, and distinguish active from dormant microbial taxa. Metaproteomics is proposed for assessing functionality of soil microbial communities by identifying the actual catalyzers of soil processes, and therefore more active microbial taxa [6–9]. However, soil metaproteomics, still in its infancy, is far from assessing the full potential of this technology to better understand microbial community composition of the soil, and track the most dominant active microbial populations in soil [10]. Importantly, metaproteomics is biased by the preferential identification of highly abundant proteins from dominant populations. Due to the fact that different numbers of proteins are identified from each population in the metaproteomic data, it seems important to focus on ubiquitous proteins that are present in all organisms, and evaluate if they can provide a suitable taxonomical indicator. We posit that excluding lesser abundant proteins from high abundant species by targeting one high abundant protein will provide a better representation of the actual species distribution. This will allow for the estimation of true relative abundance abundances of prokaryotes (archaea and bacteria), eukaryotes (animals, fungi, plants, and protists), and viruses. For this, we screened for highly expressed genes as potential biomarkers in the proteome of *Escherichia coli* MC4100 [11] essential to the functioning of the cell and thus necessary to exist in all life forms. Logically, selecting abundant proteins from a bacterium will bias the results but our aim was to test the viability of a biomarker approach in metaproteomics rather than finding the most adequate biomarker for the whole tree of life. Future approaches should compare highly expressed genes in organisms from all domains, which will get easier once more eukaryotic genomes are sequenced. All available sequences from the chosen genes ATP synthase F(0) complex (ATPS), elongation factors (EF), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), GroEL, pyruvate dehydrogenase (PyrDH), RNA polymerase beta chain (RNAP), and translation initiation factor 2 (TIF) were used for protein identification. For the validation, a mock community of known abundances of one alga, one archaeon, 25 bacteria, and 5 viruses [12] was used. Then, the seven biomarkers were applied to a randomly chosen subset of a previously published metaproteomic dataset to describe the relative taxa abundances of soil organisms in forest, grassland, and shrubland samples (each n=10) as well as their correlation to edaphic and environmental factors. We argue that biomarker metaproteomics can yield a more adequate picture of the microbial community composition and thereby verify (or not) what has been published before. We hypothesized that (i) fungal proteins are more abundant in forests, positively correlate to mean annual temperature and negatively to pH [13] while archaea prefer shrublands [14].

Materials & Methods

Identifying ubiquitous proteins across the tree of life

The basis for the identification of highly expressed genes as potential biomarker in soil taxa from the tree of life was the proteome of *Escherichia coli* MC4100 [11]. Noteworthy, this choice biased towards Bacteria and the use of proteomes from organisms of different domains may yield a different result. We identified ATPS, EF, GAPDH, GroEL, PyrDH, RNAP, and TIF as potential biomarkers necessary for important cellular functions in all soil taxa from the tree of life and downloaded all available SwissProt and TrEMBL sequences for each of these from Uniprot [15] on 22/12/2021. The raw metaproteomic data for the first glimpse how well each biomarker represented both richness and composition of the soil community was obtained from a study used to separately describe the structure and function of archaea [14], bacteria [16], and fungi [13]. The data can be found on the PRIDE database [17,18] using the identifier

PXD018448. The validation of this approach was a positive control where we applied potential biomarkers to a previously published metaproteomic data set of a mock community with known abundances of algae, archaea, bacteria, and viruses [12]. The data can be found the PRIDE database [17,18] using the identifier PXD006118.

Applying ubiquitous proteins to understand the contribution of taxa from all domains of life across environmental gradients

The biomarkers were then applied to a randomly chosen subset of a previously published metaproteomic data set used to separately describe of archaea [14], bacteria [16], and fungi [13] in three ecosystems: forests, grasslands, and shrublands (each n=10). We unveiled community composition with the aim to investigate their relation to edaphic and environmental factors that included mean annual temperature (MAT), aridity index (AI), plant cover (PC), fine texture (FT), pH, electric conductivity (EC), soil phosphorus (SP), and soil carbon (SC).

Metaproteomic search parameters and taxonomic identification

The raw files were searched with Proteome Discoverer (Thermo Fisher Scientific, v2.5.0.400) using SEQUEST HT and Percolator against the Uniprot database (as described above). The following search parameters were selected: enzyme: trypsin, precursor mass tolerance: 10 ppm, fragment mass tolerance: 0.05 Da, dynamic modification: oxidation / +15.995 Da (M) and static modification: carbamidomethyl / +57.021 Da (C). Only peptides with a false discovery rate (FDR) <1% calculated by Percolator were considered as identified. The strict search parameters may have caused the loss of identified protein groups as it was reported before [19]. The FDR concept was established for pure culture proteomics [20] using a defined threshold of 1% [21]. However, searches against large databases can decrease the number of identified protein groups due to FDR-overestimation [22], which can cause the loss of valuable protein identifications [23]. It becomes more and more common that higher FDRs up to 10% are used [24–27], particularly in soil metaproteomics [16,19]. Identified proteins were grouped by applying the strict parsimony principle, in which protein hits were reported. Since taxonomic precision depends on the grouping strategy [28], a comparison of our approach that quantifies protein groups based on unique and shared/razor peptides with the quantification using all peptides or only unique peptides as well as using unique and shared/razor peptides without parsimony grouping yielded similar results between all methods except when all peptides were used (Supplementary figure S1). Protein abundances were calculated using the minora feature detector implemented in Proteome Discoverer. The open-source software Prophane [29] was used to assign protein groups to their phylogenetic origin by diamond BLASTp. In the resulting files, only the biomarkers were selected in each sample. Visualization was performed in *R*.

Results

Validation of protein biomarkers

The EF-metaproteome was validated by a positive control using a mock community of known relative abundances from one archaeon, one alga, 5 bacteriophages, and 25 bacteria, which mixed in different amounts reflecting equal cell number, equal proteins number, and an uneven composition [12]. We used the metaproteomic raw data from this experiment to validate how accurately the biomarker metaproteomes cover abundances of soil taxa from different domains. On domain level, viral sequences were only found with GroEL as biomarker while both bacteria and eukaryota were identified in all biomarkers (**Figure 1**). On phylum level, a high number of false positive hits was found throughout all

domains, for example Euryarchaeota instead of Thaumarchaeota, which is why we decided to identify the community composition on domain level. Except ATPS and GAPDH, most biomarkers showed a high degree of accuracy to determine the relative abundances of archaea, bacteria, and eukaryota (**Figure 2**).

Applying biomarkers to estimate the relative abundance of soil taxa from different domains in soil samples across biomes

The different biomarkers yielded different soil community compositions in forests, grasslands and shrublands (**Figure 3**). Only EF was capable of identifying archaeal protein groups at a relative abundance of 0.72% in forests, 0.65% in grasslands, and 0.71% in shrublands while none of the other biomarkers showed any (**Figure 4a**). In addition, EF showed the expected decrease of eukaryota from forests (25.56%) to grasslands (16.77%) and shrublands (18.37%) while bacteria were most abundant in shrublands (64.37%) followed by grasslands (63.7%) and forests (55.97%). On average, 18% of the measured proteins could not be identified as archaeal, bacterial, or eukaryotic. Similar to the composition, the sequences in the EF-database showed a dominance of bacteria followed by eukaryotes and archaea (**Figure 4b**). These trends in relative abundances using EF-metaproteomics yielded differential significant Spearman correlation ($P < 0.05$) to edaphic and environmental parameters (**Figure 4c**). Archaea showed no correlation to both bacteria and eukaryota and any of the selected parameters while bacteria and eukaryota were negatively correlated, resulting in an inverse correlation to pH and mean annual temperature.

Number of sequences in the databases and identified proteins

The number of sequences ranged from 22,855 for ATPS to 408,998 for EF, which resulted in respectively the lowest and the highest protein richness from both mock and soil communities (**Table 1**). However, a drop down to 10% in richness of the mock community was found for the soil samples for many biomarkers.

Discussion

The aim of this study was to find a biomarker that could equally describe proteins across soil organisms from prokaryotes (archaea and bacteria), eukaryotes (animals, fungi, plants, and protists), and viruses. Unfortunately, this approach per se is not feasible for viruses as they do not share single genes and markers can only target specific viral groups, i.e. T7-like podoviruses with the DNA polymerase [30]. This was confirmed by the lack of viral proteins in both mock and soil communities regardless the biomarker. Therefore, viral abundances relative to other soil taxa have to be estimated with different approaches. Otherwise, the identification on domain level was reasonably accurate for archaea, bacteria, and eukaryotes for all used biomarkers except ATPS and GAPDH. However, on phylum level many false positive hits were identified. Particularly eukaryotes showed a multitude of different hits even though only the alga *Chlamydomonas reinhardtii* was present in the mock community. This might be caused by the focus of research on multicellular eukaryotes and their parasites as almost all eukaryotic genomes are from animals, fungi, or land plants [31] but they only represent 23% of environmental 18S DNA sequences [32]. Noteworthy, the mock community contained only one archaeon (*Nitrososphaera viennensis*) and the above-mentioned alga together with 25 bacteria, which makes the validity of the verification for archaea and eukaryotes questionable but it is also the only mock community for which proteomic data is available. In order to properly verify the quantitation of proteins across soil taxa, the measurement of mock communities with more non-bacterial strains is necessary in the future. Regardless, for now we hypothesize that the sequences of these ubiquitous biomarkers can only be differentiated on domain level but further sequence-based investigations are needed for verification. We then applied the biomarkers to describe the community composition of soil samples previously used to describe archaea [14], bacteria

[16], and fungi [13]. Even though almost all biomarkers showed high accuracy to determine domain level relative abundances in the mock community, there was much higher variation in the soil samples. GroEL, RNAP, and TIF identified almost no eukaryotic proteins while GAPDH, as it was similar in the mock community, had a much higher abundance of eukaryotes. From the other three biomarkers, only ATPS and EF unveiled the expected decrease of eukaryotes from forests to shrublands but of those two, only EF was able to identify archaea. Combined with the accuracy of identifying the relative abundances in the mock community, we believe that EF is the best choice among the tested biomarkers to identify domain level abundances across soil taxa. Indeed, the dominance of bacterial sequences (57.0%) in the EF-database compared to eukaryotes (37.0%) and archaea (5.3%) aligns well with previously reported metatranscriptomic results [33,34]. Consistent with fungal proteins [13], eukaryotes decreased in abundance from forests to shrublands, were positively correlated to mean annual temperature and negatively to pH, again highlighting the focus of genomic research on multicellular eukaryotes like fungi [31]. Perhaps unicellular eukaryotes like protists have different correlative patterns. Interestingly, our approach revealed the inverse relationship between bacteria and eukaryotes, probably resulting from their different niches as it was reported for bacteria and fungi before [19]. The trend of increasing archaeal protein abundance in shrublands and the correlation to aridity [14] could not be seen in the EF-data. However, archaeal sequences are present in the EF-database which is why we are confident that this approach is better to identify true relative abundances compared to using only a subset of organisms as done before. Possibly, the taxonomic specificity of the archaeal peptides is too low to warrant more identifications. We therefore hypothesize that the previously reported trends for archaea were artifacts introduced by the database. In fact, the used database impacted the number of identified proteins. Generally, more sequences in the database resulted in a higher number of identified proteins, which makes the comparison of protein richness impossible unless the sequences are aligned to similar numbers as commonly done in sequencing approaches (rarefaction). However, the richness results also showed the decrease in protein identification rate in soil compared to the mock community, which means that finding new and better strategies in the metaproteomic workflow is inevitable for future research.

Taken together, we investigated the potential of using biomarkers in metaproteomics to equally describe archaea, bacteria, and eukaryotes. Among seven chosen biomarkers, EF showed the highest potential for accurate quantification but a different choice of biomarkers focusing on archaea or eukaryotes can yield different results. A higher number of genomes of non-bacterial organisms will not only make the EF-approach better but will also allow for the search of other biomarkers with the potential to unveil relative abundances from all soil taxa. Importantly, cellular protein concentration has been shown to depend on experimental conditions causing system-wide proteome allocation, expression regulation, and post-translational adaptations [35], which in turn questions the validity of the biomarker approach if they are not ubiquitously abundant in all cells across soil taxa.

Acknowledgements

M.D-B. is supported by a project from the Spanish Ministry of Science and Innovation (PID2020-115813RA-I00), and a project of the Fondo Europeo de Desarrollo Regional (FEDER) and the Consejería de Transformación Económica, Industria, Conocimiento y Universidades of the Junta de Andalucía (FEDER Andalucía 2014-2020 Objetivo temático “01 - Refuerzo de la investigación, el desarrollo tecnológico y la innovación”) associated with the research project P20_00879 (ANDABIOMA). RS thanks Alexandra Elbakyan for accessing literature. We thank Kay Schallert for his continuous support with the bioinformatic pipeline. The authors thank Petra Marschner and the anonymous reviewer for their insightful comments that immensely improved our manuscript.

Compliance with ethical standards

The authors declare no conflict of interest.

References

[1] R.L. Hettich, C. Pan, K. Chourey, R.J. Giannone, Metaproteomics: Harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities, Anal Chem. (2013). https://doi.org/10.1021/ac303053e.

[2] R. Starke, N. Jehmlich, F. Bastida, Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics, J Proteomics. (2018). https://doi.org/10.1016/j.jprot.2018.11.011.

[3] R.D. Bardgett, W.H. van der Putten, Belowground biodiversity and ecosystem functioning, Nature. (2014). https://doi.org/10.1038/nature13855.

[4] T.W. Crowther, J. van den Hoogen, J. Wan, M.A. Mayes, A.D. Keiser, L. Mo, C. Averill, D.S. Maynard, The global soil community and its influence on biogeochemistry, Science (1979). (2019). https://doi.org/10.1126/science.aav0550.

[5] M. Delgado-Baquerizo, P.B. Reich, C. Trivedi, D.J. Eldridge, S. Abades, F.D. Alfaro, F. Bastida, A.A. Berhe, N.A. Cutler, A. Gallardo, L. García-Velázquez, S.C. Hart, P.E. Hayes, J.Z. He, Z.Y. Hseu, H.W. Hu, M. Kirchmair, S. Neuhauser, C.A. Pérez, S.C. Reed, F. Santos, B.W. Sullivan, P. Trivedi, J.T. Wang, L. Weber-Grullon, M.A. Williams, B.K. Singh, Multiple elements of soil biodiversity drive ecosystem functions across biomes, Nat Ecol Evol. 4 (2020). https://doi.org/10.1038/s41559-019-1084-y.

[6] F. Bastida, I.F. Torres, J.L. Moreno, P. Baldrian, S. Ondoño, A. Ruiz-Navarro, T. Hernández, H.H. Richnow, R. Starke, C. García, N. Jehmlich, The active microbial diversity drives ecosystem multifunctionality and is physiologically related to carbon availability in Mediterranean semi-arid soils, Mol Ecol. (2016). https://doi.org/10.1111/mec.13783.

[7] J. Hultman, M.P. Waldrop, R. Mackelprang, M.M. David, J. McFarland, S.J. Blazewicz, J. Harden, M.R. Turetsky, A.D. McGuire, M.B. Shah, N.C. VerBerkmoes, L.H. Lee, K. Mavrommatis, J.K. Jansson, Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes, Nature. (2015). https://doi.org/10.1038/nature14238.

[8] D. Liu, K.M. Keiblinger, S. Leitner, U. Wegner, M. Zimmermann, S. Fuchs, C. Lassek, K. Riedel, S. Zechmeister-Boltenstern, Response of microbial communities and their metabolic functions to drying–rewetting stress in a temperate forest soil, Microorganisms. 7 (2019). https://doi.org/10.3390/microorganisms7050129.

[9] R. Starke, F. Bastida, J. Abadía, C. García, E. Nicolás, N. Jehmlich, Ecological and functional adaptations to water management in a semiarid agroecosystem: A soil metaproteomics approach, Sci Rep. (2017). https://doi.org/10.1038/s41598-017-09973-w.

[10] J.A. Blakeley-Ruiz, M. Kleiner, Considerations for Constructing a Protein Sequence Database for Metaproteomics, Comput Struct Biotechnol J. (2022).

[11] Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F.U. Harlt, M.J. Kerner, D. Frishman, Protein abundance profiling of the Escherichia coli cytosol, BMC Genomics. 9 (2008). https://doi.org/10.1186/1471-2164-9-102.

[12] M. Kleiner, E. Thorson, C.E. Sharp, X. Dong, D. Liu, C. Li, M. Strous, Assessing species biomass contributions in microbial communities via metaproteomics, Nat Commun. 8 (2017). https://doi.org/10.1038/s41467-017-01544-x.

[13] M.L.P. Fernandes, F. Bastida, N. Jehmlich, T. Martinović, T. Větrovský, P. Baldrian, M. Delgado-Baquerizo, R. Starke, Functional soil mycobiome across ecosystems, J Proteomics. 252 (2022). https://doi.org/10.1016/j.jprot.2021.104428.

[14] R. Starke, J.A. Siles, M.L.P. Fernandes, K. Schallert, D. Benndorf, C. Plaza, N. Jehmlich, M. Delgado-Baquerizo, F. Bastida, The structure and function of soil archaea across biomes, J Proteomics. (2021). https://doi.org/10.1016/j.jprot.2021.104147.

[15] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. da Silva, M. de Giorgi, T. Dogan, F. Fazzini, L.G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, J. Zhang, UniProt: The universal protein knowledgebase, Nucleic Acids Res. (2017). https://doi.org/10.1093/nar/gkw1099.

[16] F. Bastida, N. Jehmlich, R. Starke, K. Schallert, D. Benndorf, R. López-Mondéjar, C. Plaza, Z. Freixino, C. Ramírez-Ortuño, A. Ruiz-Navarro, M. Díaz-López, A. Vera, J.L. Moreno, D.J. Eldridge, C. García, M. Delgado-Baquerizo, Structure and function of bacterial metaproteomes across biomes, Soil Biol Biochem. 160 (2021).

[17] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, Ş. Yılmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A.F. Jarnuczak, T. Ternent, A. Brazma, J.A. Vizcaíno, The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucleic Acids Res. 47 (2018) D442–D450. https://doi.org/10.1093/nar/gky1106.

[18] T. Ternent, A. Csordas, D. Qi, G. Gómez-Baena, R.J. Beynon, A.R. Jones, H. Hermjakob, J.A. Vizcaíno, How to submit MS proteomics data to ProteomeXchange via the PRIDE database, Proteomics. (2014). https://doi.org/10.1002/pmic.201400120.

[19] R. Starke, R.L. Mondéjar, Z.R. Human, D. Navrátilová, M. Štursová, T. Větrovský, H.M. Olson, D.J. Orton, S.J. Callister, M.S. Lipton, A. Howe, L.A. McCue, C. Pennacchio, I. Grigoriev, P. Baldrian, Niche differentiation of bacteria and fungi in carbon and nitrogen cycling of different habitats in a temperate coniferous forest: A metaproteomic approach, Soil Biol Biochem. 155 (2021). https://doi.org/10.1016/j.soilbio.2021.108170.

[20] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat Methods. (2007). https://doi.org/10.1038/nmeth1019.

[21] K. Barnouin, Guidelines for experimental design and data analysis of proteomic mass spectrometry-based experiments, Amino Acids. (2011). https://doi.org/10.1007/s00726-010-0750-9.

[22] R. Heyer, K. Schallert, R. Zoun, B. Becher, G. Saake, D. Benndorf, Challenges and perspectives of metaproteomic data analysis, J Biotechnol. (2017). https://doi.org/10.1016/j.jbiotec.2017.06.1201.

[23] T. Muth, C.A. Kolmeder, J. Salojärvi, S. Keskitalo, M. Varjosalo, F.J. Verdam, S.S. Rensen, U. Reichl, W.M. de Vos, E. Rapp, L. Martens, Navigating through metaproteomics data: A logbook of database searching, Proteomics. (2015). https://doi.org/10.1002/pmic.201400560.

[24] T. Schneider, K.M. Keiblinger, E. Schmid, K. Sterflinger-Gleixner, G. Ellersdorfer, B. Roschitzki, A. Richter, L. Eberl, S. Zechmeister-Boltenstern, K. Riedel, Who is who in litter decomposition Metaproteomics reveals major microbial players and their biogeochemical functions, ISME Journal. (2012). https://doi.org/10.1038/ismej.2012.11.

[25] S.M. Figarska, S. Gustafsson, J. Sundström, J. Ärnlöv, A. Mälarstig, S. Elmståhl, T. Fall, L. Lind, E. Ingelsson, Associations of circulating protein levels with lipid fractions in the general population, Arterioscler Thromb Vasc Biol. (2018). https://doi.org/10.1161/ATVBAHA.118.311440.

[26] K.M. Keiblinger, I.C. Wilhartitz, T. Schneider, B. Roschitzki, E. Schmid, L. Eberl, K. Riedel, S. Zechmeister-Boltenstern, Soil metaproteomics - Comparative evaluation of protein extraction protocols, Soil Biol Biochem. (2012). https://doi.org/10.1016/j.soilbio.2012.05.014.

[27] F. Bastida, T. Hernández, C. García, Metaproteomics of soils from semiarid environment: Functional and phylogenetic information obtained with different protein extraction methods, J Proteomics. (2014). https://doi.org/10.1016/j.jprot.2014.02.006.

[28] K. Schallert, P. Verschaffelt, B. Mesuere, D. Benndorf, L. Martens, T. van den Bossche, Pout2Prot: An Efficient Tool to Create Protein (Sub)groups from Percolator Output Files, J Proteome Res. 21 (2022). https://doi.org/10.1021/acs.jproteome.1c00685.

[29] H. Schiebenhoefer, K. Schallert, B.Y. Renard, K. Trappe, E. Schmid, D. Benndorf, K. Riedel, T. Muth, S. Fuchs, A complete and flexible workflow for metaproteomics data analysis based on

MetaProteomeAnalyzer and Prophane, Nat Protoc. 15 (2020). https://doi.org/10.1038/s41596-020-0368-7.

[30] M.B. Sullivan, Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities, J Virol. 89 (2015). https://doi.org/10.1128/jvi.03289-14.

[31] J. Pawlowski, S. Audic, S. Adl, D. Bass, L. Belbahri, C. Berney, S.S. Bowser, I. Cepicka, J. Decelle, M. Dunthorn, A.M. Fiore-Donno, G.H. Gile, M. Holzmann, R. Jahn, M. Jirků, P.J. Keeling, M. Kostka, A. Kudryavtsev, E. Lara, J. Lukeš, D.G. Mann, E.A.D. Mitchell, F. Nitsche, M. Romeralo, G.W. Saunders, A.G.B. Simpson, A. v. Smirnov, J.L. Spouge, R.F. Stern, T. Stoeck, J. Zimmermann, D. Schindel, C. de Vargas, CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms, PLoS Biol. 10 (2012). https://doi.org/10.1371/journal.pbio.1001419.

[32] J. del Campo, M.E. Sieracki, R. Molestina, P. Keeling, R. Massana, I. Ruiz-Trillo, The others: Our biased perspective of eukaryotic genomes, Trends Ecol Evol. 29 (2014). https://doi.org/10.1016/j.tree.2014.03.006.

[33] T. Urich, A. Lanzén, J. Qi, D.H. Huson, C. Schleper, S.C. Schuster, Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome, PLoS One. 3 (2008). https://doi.org/10.1371/journal.pone.0002527.

[34] A.T. Tveit, T. Urich, P. Frenzel, M.M. Svenning, Metabolic and trophic interactions modulate methane production by Arctic peat microbiota in response to warming, Proc Natl Acad Sci U S A. (2015). https://doi.org/10.1073/pnas.1420797112.

[35] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, M. Heinemann, The quantitative and condition-dependent Escherichia coli proteome, Nat Biotechnol. 34 (2016). https://doi.org/10.1038/nbt.3418.

Figures & figure legends

Figure 1: Relative abundances on domain and phylum level of a mock community (Mock) comprising one archaeon, 25 bacteria, one eukaryote, and five viruses in different mixtures (Mock-1 = equal cells, Mock-2 = equal proteins, and Mock-3 = uneven) estimated by biomarker metaproteomics (each n=4).

Figure 2: Accuracy of biomarker metaproteomics to determine the relative abundances of archaea, bacteria, and eukaryota of a mock community comprising one archaeon, 25 bacteria, one eukaryote, and five viruses.

Figure 3: Relative abundances of soil microbial communities across three global biomes estimated by metaproteomics using seven biomarkers (a).

Figure 4: Soil community composition using EF-metaproteomics (a), the composition of sequences in the EF-database (b) as well as significant ($P < 0.05$) Spearman correlation of the relative abundances of archaea, bacteria, and eukaryota using EF-metaproteomics to edaphic and environmental variables (b). MAT stands for mean annual temperature, AI for aridity index, PC for plant cover, FT for fine texture, EC for electric conductivity, SC for soil carbon, and SP for soil phosphorus.

Tables & table legends

Table 1: Number of sequences in the FASTA-files as well as the average number and standard deviation of identified protein groups across three mock (each n=4) and three soil biome samples (each n=10) for the seven biomarkers.

Gene (abbreviation)	Sequences	Mock proteins	Soil proteins
ATP synthase F(0) complex (ATPS)	22,855	101±4	38±6
Elongation factor (EF)	408,998	901±187	247±49
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	83,257	195±27	22±6
GroEL	112,733	476±81	312±63
Pyruvate dehydrogenase (PyrDH)	312,478	301±30	23±6
RNA polymerase beta chain (RNAP)	266,861	395±66	89±19
Translation initiation factor 2 (TIF)	135,452	69±8	20±4