Benchmarking and developing mass-spectrometry based omics approaches for complex microbiomes

Kleikamp, H.B.C.

DOI
10.4233/uuid:45f6ee1a-1ae2-4d7b-aac8-49809962aa83

Publication date
2022

Document Version
Final published version

Citation (APA)
Kleikamp, H. B. C. (2022). *Benchmarking and developing mass-spectrometry based omics approaches for complex microbiomes*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:45f6ee1a-1ae2-4d7b-aac8-49809962aa83

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Benchmarking and developing mass-spectrometry based omics approaches for complex microbiomes

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen, voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
Maandag 16 mei 2022 om 15:00 uur

door

Ir. Hugobert Bernhard Crijn KLEIKAMP

Master of Science in Life Science and Technology,
Technische Universiteit Delft, Nederland

Geboren in Den Haag, Nederland

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie bestaat uit:

Rector Magnificus,	voorzitter
Prof.dr.ir. M.C.M. van Loosdrecht,	Technische Universiteit Delft, promotor
Dr. M. Pabst,	Technische Universiteit Delft, copromotor
Dr. Y. Lin,	Technische Universiteit Delft, copromotor

Onafhankelijke leden:

Prof. dr. ir. M. J. T. Reinders,	Technische Universiteit Delft
Prof. dr. ir. T. J. G. Ettema,	Wageningen University & Research
Prof. dr. ir. M. Strous,	University of Calgary
Prof. dr. T. Curtis,	Newcastle University
Prof. dr. ir. J. M. Daran,	Technische Universiteit Delft, reserve member

Sleutelwoorden:

Mass spectrometry, Bioinformatics, Metaproteomics, De novo sequencing, Sialic acids

Drukkerij: Gildeprint
Boekomslag door: Enny Kleikamp-van Leeuwen (tekening, voorkant en achterkant)
Yuechun Lin (gemarmerd papier, voorkant)
ISBN: http://repository.tudelft.nl/

Table of contents

Samenvatting

De tomeloze ruimte aan microbiële diversiteit vormt een onuitputbare bron van enzymen en bioactieve stoffen. Genetische methodes hebben afgelopen decennia veel vooruitgang geboekt in hun analyse van microbiële gemeenschappen. Echter, louter kennis van een genoom bemoeilijkt de beschrijving van biologische aspecten, zoals tot expressie gebrachte eiwitten in het proteoom en complexe producten zoals bacteriële suiker modificaties, die tot uitdrukking komen in het fenotype. Massaspectrometrie levert zowel gerichte als algemene methodes voor de analyse van fenotypes, maar ondervindt hinder door een limiet in doorstroom en de gecompliceerde data-analyse van complexe monsters, waarvoor nieuwe strategieën noodzakelijk zijn. Het doel van deze thesis is daarom de vergelijking en verbetering van bestaande technieken, en het ontwikkelen van nieuwe methodes voor de analyse van microbiële mengsels met behulp van massaspectrometrie.

In **Hoofdstuk 2** wordt een vergelijking gemaakt tussen de 3 meest voorkomende analysetechnieken voor microbiële gemeenschappen: 16s rRNA sequencing, metagenomics en metaproteomics. Elke methode kwantificeert de contributie van individuele microben op een andere manier, en benadrukt hierdoor verschillende taxa. Een verdere complicatie komt door het gebruik van een scala aan taxonomische databases, die elk andere classificatiesystemen hanteren. Om de intrinsieke verschillen tussen de methodes beter te begrijpen, is een algemene vergelijking van databases en integratie gedaan aan de hand van de genome taxonomy database (GTDB), en wordt getracht om vertekeningen in contributie van specifieke organismen te koppelen aan zowel hun fysiologische eigenschappen, als aan de opzet van de experimenten en aanpak van de data-analyse. De geobserveerde verschillen in deze studie benadrukken het belang van een centrale taxonomische database en de combinatie van orthogonale kwantificatiemethoden voor bekrachtiging van observaties. Hierbij wordt, naast celgrootte, het patroon of de verdeling van eiwitexpressie vooropgesteld als een belangrijke bron van verschil tussen metagenoom en proteoom kwantificatie.

Een alternatieve manier voor het bepalen van peptidesequenties binnen massaspectrometrie is door middel van de novo sequencing, dat aminozuur-sequenties kan genereren zonder databases te gebruiken, hetgeen tijds-efficiënter (en goedkoper) is. In **Hoofdstuk 3** wordt de effectiviteit van deze methode holistisch vergeleken met conventionele spectrum annotatie, voor zowel pure cultures als voor microbiële mengsels. Daarbij wordt een strategie beschreven voor het verminderen van ruis en voor een vlotte taxonomische en functionele annotatie van de novo gemeten peptiden. Hiermee wordt aangetoond dat de novo sequencing een vergelijkbare taxonomische verdeling oplevert als de spectrum vergelijkende methodes, tot en met de familie rang.

Suiker modificaties zijn moeilijk te voorspellen vanuit genomische data, en vormen een minder begrepen aspect van microbiële fysiologie. De siaalzuren vormen een zeer complexe suikergroep, die vooral wordt geassocieerd met pathogenen, maar die ook hierbuiten beschreven is. In **Hoofdstuk 4** worden een nieuwe data acquisitie en dataverwerkingsmethode gecombineerd om doelgericht siaalzuren te ontdekken in ruwe celextracten met behulp van massaspectrometrie. Door middel van een grootschalige studie wordt bewezen dat siaalzuren, buiten ziekteverwekkers, door vele ecologische bacteriën geproduceerd worden. De methode toont bovendien in staat te zijn om tot nog toe onontdekte siaalzuurvarianten bloot te leggen.

Summary

The large unexplored space of microbial biodiversity forms a well of untapped potential of enzymes and bioactive compounds. Genomic methods have made significant improvements over the past decades in their analysis of microbial communities. However, several aspects of nature can't be described by a genome alone, such as expressed functions in the proteome and complex products such as bacterial glycosylation. Mass spectrometry can provide both general and targeted methods for the analysis of phenotypes, but is facing challenges in its limited throughput and the demanding data analysis of complex samples, which necessitates new methodologies. This thesis therefore aims to compare, improve and develop now methods for the analysis of microbial communities using mass spectrometry.

Within **Chapter 2** a comparison is made between the 3 most applied techniques for the quantification of microbial communities: 16S rRNA sequencing, metagenomics and metaproteomics. Each method quantifies in another manner and therefore emphasizes different taxa. A complicating factor is the use of different taxonomic databases, which employ varying methods of phylogenetic placement. To better understand the intrinsic differences between quantification methods, a general comparison and integration of databases is done according to the genome taxonomy database (GTDB), and an attempt is made to link biases in quantification to factors stemming from bacterial physiology, experimental setup, and data analysis. The observed quantification differences underline the need for centralized taxonomic databases and the use of orthogonal quantification methods for improved validation. Additionally, next to cell volume, protein expression patterns are proposed as a major source of quantification difference between metagenomics and metaproteomics.

De novo sequencing is an alternative strategy for obtaining peptide sequences without the need for a database, which results in a more efficient annotation. In **Chapter 3** the performance of de novo sequencing is benchmarked against the conventional spectrum searching approach, using pure cultures and microbial communities. Herewith a data analysis strategy is defined, to reduce noise, and for the rapid taxonomic and functional annotation of de novo peptides. With this annotation pipeline, it is shown that de novo sequencing can deliver similar taxonomic annotations to conventional database searching up to family rank.

Bacterial glycosylation is difficult to predict from genomic data alone, and forms a less understood aspect of microbial physiology. Sialic acids describe a complex class of sugars, which has been chiefly associated with pathogens, but has also been described in other contexts. In **Chapter 4** a new data acquisition strategy is combined with a novel data analysis method to selectively target sialic acids in crude cell extracts using mass-spectrometry. The widespread nature of sialic acids in environmental bacteria is proven by virtue of a large-scale screening. Additionally, this method is shown to be capable of uncovering new sialic acid variants.

1

General introduction

1.1 Why study microbial communities?

Over billions of years, microbes have shaped the world around us. Early microbes created the oxygen atmosphere and today play an essential role in biogeochemical cycles. From the early days, humans employed fermentation to increase food shelf-life, while the discovery of penicillin extended the average human live span by over 20 years. We live in a close symbiosis with microbes on our planet and our bodies, with a substantial impact on our culture, health, and well-being.

The discovery of life's source code: DNA, has enabled us to transmute microbes to do our bidding. This ability to directly alter and improve a microbe's functionality forms a key aspect of industrial microbiology. The field of environmental biotechnology, on the other hand, employs natural synergies between groups of microbes within non-sterile conditions. This enables the use of alternative feedstocks and the design of more sustainable processes. The main challenge, however, remains in understanding the underlying metabolic processes that take place in such microbial communities. Most microbes have not been sequenced to date and are unculturable under laboratory conditions. Nevertheless, natural communities have significant potential that, if harnessed, allows to apply the nearly unlimited number of enzymes and bioactive compounds that nature created during billions of years of evolution.

One of the major industrial applications of microbial communities is wastewater treatment, which is of paramount importance for our society and protection of the environment. Due to the scale at which these processes operate, thousands of tons of biomass are produced per plant annually. Nutrient removal from wastewater requires a synergistic effort of several groups of highly specialized microbes. Determining the key microbial converters and processes in wastewater treatment will assist in improving this widely used process and help to develop new applications for the otherwise discarded sludge. Given that community complexity easily transcends what is observed in a laboratory setting, studying full-scale plants is not an easy task. This requires the use of sophisticated analytical techniques that analyze pools of molecules, commonly grouped under the "Meta-Omics" umbrella.

Omics methodologies can characterize microbial communities from many angles, but most commonly target DNA (genomics) and proteins (proteomics). As these technologies are still evolving rapidly, there is an increased demand for deep, systematic, comparative studies to map out the potentials and limitations of the methods, but also for new strategies that can process the ever-increasing volumes of information that these technologies provide. This thesis therefore aims to benchmark, compare, improve, and develop new data processing approaches for mass spectrometry based "Meta-Omics" experiments. In particular, to study microbes in complex communities, such as those that exist within wastewater treatment plants.

1.2 A timeline of microbial taxonomy

Since the discovery of the microorganism by Antonie van Leeuwenhoek, Delft has been on the forefront of environmental microbiology (Fig 1.1). The Delft school of microbiology and its professors Kluyver, Beijerinck, Iterson, and Kuenen pushed the development of virology, enrichment cultures, and mathematical approaches to model the metabolism of microorganisms.

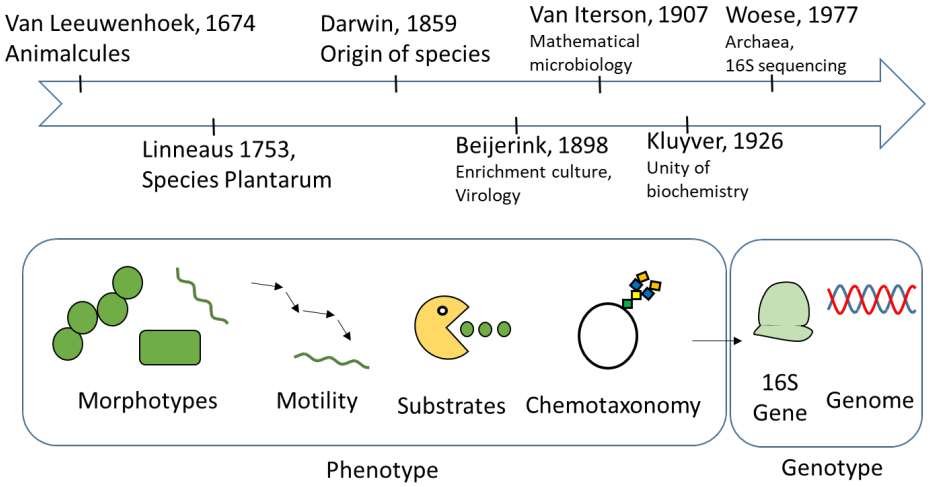


Fig. 1.1 Top: A timeline of microbial taxonomy, with a few major international highlights (Linneaus, Darwin, Woese), as well as achievements from Delft (Van Leeuwenhoek, Beijerinck, Van Iterson, & Kluyver). Bottom: the evolution of microbial classification methods, including the transition from phenotype to genotype-based classification.

While frameworks and tools have been developed to study and predict the behavior of microorganisms, the total expansive space of microbial diversity is still unexplored. The phrase Beijerink quipped a 100 years ago: “gelukkig zij, die nu beginnen” (Lucky are those who start now) seems ever more true, as recent advances in sequencing has caused an exponential surge in our knowledge of microbial diversity.

With the discovery of biodiversity, the need for a classification system arose. Traditionally, the taxonomy of an organism would be assigned according to its visible traits, which represent the phenotype. For macroorganisms, this would depend on morphological characteristics such as leaf shape, or number of legs, and behavioral characteristics such as social structure or mating patterns. For microorganisms, a phenotypical classification turned out to be more challenging (Fig. 1.2).

Fig. 1.2 The struggles of a microbial scientist. Microbes are hard to classify and identify due to (from left to right): their small size, their widespread presence across various environments, their complex dependencies, their high diversity.

First, by the microbe’s small size, which necessitates specialized instruments to observe their traits. Second, by their ability to persist in many conditions, which includes extreme environments that are not easily sampled (such as arctic or deep sea), but also conditions that are not easily mimicked in a laboratory setting. Third, by their complex symbiotic organization, which limits the study of species in isolation. Last, by their enormous diversity, which requires a much more elaborate range of characteristics to provide a systematic differentiation.

1.2.1 A shift in classification method

Historically, phenotypical classification of microorganisms was based simply on what could be seen with a microscope, focusing on morphological characteristics, such as cell shape, size, appearance, or motility (Fig. 1.1). Because these features could not explain fine-grained differences at higher taxonomic ranks, the inclusion of physiological typing such as substrate use proved helpful. However, since this required isolation and culturing in a laboratory setting, it could only be applied to a select fraction of bacteria (Staley 1985).

Later, additional biochemical characteristics were included to support the classification of microorganisms, which led to the term "chemotaxonomy", including "Gram-staining", DNA base ratios, or lipid and sugar compositions (Schleifer 2009, Pace 2012). However, this inclusion made the classification significantly more laborious and prone to human error. Therefore, the scientific field sought a universal evolutionary marker (or "semantide") that is found in all organisms and that evolves at low rates to enable the reconstruction of evolutionary lineages (Zuckerkandl 1965). Initially, protein-based semantides such as hemoglobin and cytochrome C were proposed but were subsequently replaced by the discovery of 16S rRNA as a more general evolutionary marker (Woese 1977). The 16S rRNA gene has the advantage of being present in all organisms because of its fundamental role in translation. This discovery also initiated the age of genotypic classification, which has expanded into considering full-genome sequences (Parks 2018).

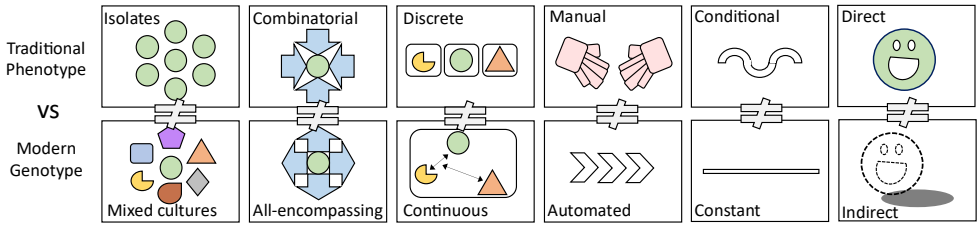


Fig. 1.3 Comparison of the main differences between traditional phenotypical classification (top row) and modern genotypical classification (bottom row) by their characteristics across various categories

1.2.2 Genotypical vs phenotypical classification

The transition from phenotypical to genotypical classification meant a radical shift in approach, which can be generalized to some key differences (Fig. 1.3):

Discrete vs Continuous: traditional phenotypical classification was discrete, which placed features into a predefined number of categories, like morphotype (bacillus/coccus) or nutrient utilization (chemotroph, phototroph), while genotypical placement occurs on a more gradual scale by use of similarity scores.
Qualitative vs Quantitative: Similarly, several of the traditional categories often gave qualitative information, and lacked in standardization, but genotypical information is purely quantitative.
Isolates vs Culture independent: Sequencing of 16S rRNA or whole genomes enables classifying organisms without the need for isolation or enrichment and can even measure samples directly.
Combinatorial vs All-encompassing: Phenotypical classification combined several features while genotypical classification bases this on a single feature.
Manual vs Automated: Out of the two classification methods, the multi-angular nature of phenotypical classification required significantly more manual labor.
Conditional vs Constant: Depending on culture conditions or growth stages, the phenotype can shift dramatically, while a genome would remain relatively constant within the same timeframe.
Direct vs Indirect: Compared to direct reflection of traits within the phenotype, the genotype forms an indirect representation of the organism's "behavior".

Since a phenotype is a direct product of gene expression, it will only include a fraction of what is present in the actual genotype. Therefore, the functional analysis of communities is limited, as there is no way to link activities to those of specific organisms. Genotypically-related groups might not fulfill the same ecological niches and utilize different substrates (Morissey 2016, Doolittle 2009, Jaspers 2004). In addition, varying expression levels can result in large activity differences in organisms possessing the same metabolic pathways (Amend 2016). Next to expression levels, there are several other factors that complicate the translation of a genotype into a phenotype.

These include transcriptional “alterations”, such as ribosomal frameshifting, codon reassignment, or stop codon read through, post transcriptional modifications, such as alternative splicing and mRNA-editing, but also the use of non-canonical amino acids, and small open reading frames (Orr 2020, Coruso 2017, Duval 2018, Meydan 2018, Ruiz-Orera 2019, Atkins 2010, Breinnicke 1999). This is reflected in the extended Dogma (Fig 1.4), which shows that between the genotype and phenotype there is a slew of steps which can change the final characteristics of a microbe.

1.2.3 Mass spectrometry as a bridge between phenotype and genotype

A molecular-level technique that enables work on the interface between genotypic and phenotypic characteristics, is mass spectrometry. This technique is widely applied within the field of analytical chemistry since it can analyze a large range of different molecules in complex mixtures at high sensitivity, including metabolites (metabolomics), lipids (lipidomics), carbohydrates (glycomics), and peptides or proteins (proteomics). The proteome represents the sum of the individual gene products and the active parts of a cell. Proteins convert, regulate, and maintain the cell and are, therefore, very close to the actual phenotype (Fig 1.4). Mass spectrometry-based methods also enable to study dynamic protein organization or substrate utilization of microbes using stable isotope incorporation experiments (Kleiner 2019).

Next to the proteome itself, mass spectrometry can study other aspects of microbial phenotypes (Fig. 1.5). One of the most abundant protein modifications is protein glycosylation, which is fundamental to all domains of life and plays an essential role in health by influencing virulence and pathogenicity (Varki 2008, Lewis 2009) but also in general protein localization, lifetime, and stability. Because this process decorates the cells (through lipids, proteins, and carbohydrate polymers), it represents the outermost interaction layer with its environment. Within biofilms, extensive glycosylation can support surface layer proteins (Boleij 2020) and extracellular polymers (Martinez 2020), which protect against degradation (Pinel 2020), whereas glycosylation may also be tuned to support substrate acquisition (Pabst 2021).

Sugar modifications present an aspect of the microbial phenotype that is hard to predict from the genotype alone, because of the typically large number of carbohydrate-active enzymes in a genome, and since sugar polymers are typically a collaborative effort of a cascade of enzymes. Because complex sugars are challenging to produce by chemical synthesis, they are a prime target for resource recovery of chemical building blocks to add value to otherwise disposed sludge. An example of sugars that have shown high abundance in granular sludge is sialic acids (de Graaff 2019), which are known for the complexity and diversity of their modifications. The flexibility of mass spectrometry enables the development of screening methods that specifically targets this class of sugars **(Chapter 4).**

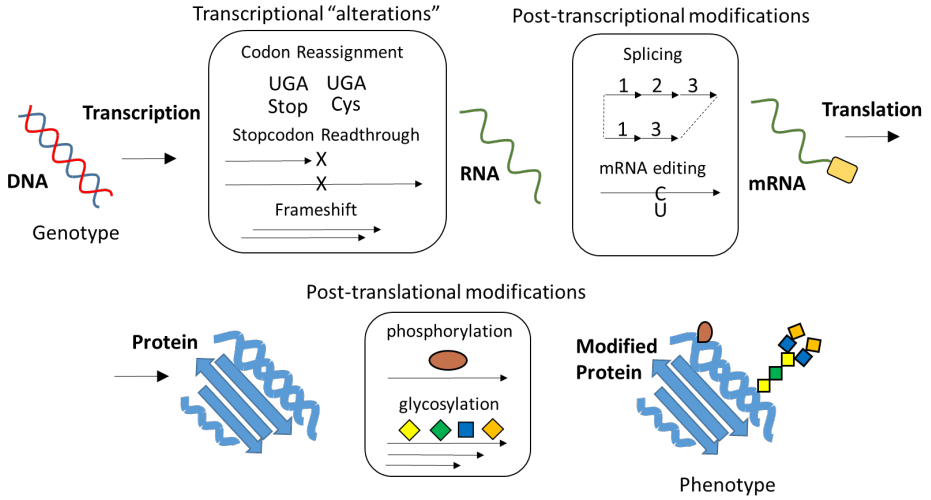


Fig. 1.4 An extended central biological dogma of transcription and translation, which highlights additional steps between genotype and phenotype (transcriptional "alterations", post-transcriptional and post-translational modifications, each of which affect the final Phenotype.

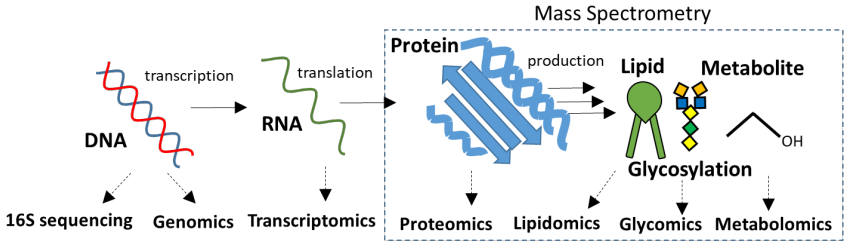


Fig. 1.5 An overview "-omics" methods (Top: target molecule, Bottom: the associated name of the "-omics" technique), and those that can be achieved by use of mass spectrometry: Proteomics, Lipidomics, Glycomics, Metabolomics.

1.3 Development of large-scale metaproteomics

In the early days of proteomics, sequencing of cDNA obtained from mRNA provided sequence tags (Adams 1993). Within a few years, proteins could be directly sequenced through Edman degradation (Erdjument-Bromage 1994). However, neither of these techniques could efficiently detect small amounts or complex mixtures of proteins (Wilm 1996). Because larger biomolecules were difficult to ionize with hard ionization techniques, the development of soft ionization techniques, such as electrospray ionization (ESI, Fenn 1989) and matrix-assisted laser desorption ionization (MALDI, Karas 1988), was a turning point in the field of biological mass spectrometry. The new ionization technique, in combination with tandem mass spectrometry (which allows to obtain sequence information from peptides or proteins) and high-performance separation techniques (Wilm 1996), ultimately enabled the field of "large-scale proteomics".

While the term "proteome" was introduced in the early 2000s, the first metaproteomics paper was published a few years later on activated sludge, which only identified a handful of proteins and employed laborious techniques, such as 2D-PAGE to analyze individual gel spots (Wilmes 2004). Shortly after, nearly 2000 proteins could be identified by employing a shotgun approach using LC-MS/MS and database searching with a metagenomics-constructed database (Ram 2005). Following further instrumental advancements (such as the development of the quadrupole Orbitrap mass spectrometer, Thermo) and optimizations of methods (e.g., data-independent fragmentation techniques, Gillet 2012), nearly 10000 proteins could be identified from one-dimensional analysis runs (Kleiner et al., 2017). The launch of mass spectrometers that employed additional advanced ion-mobility spectrometry and the further improved "next-generation" Orbitrap mass instruments further pushed the quality and number of spectra that could be obtained from proteomics experiments. Since proteomics advancements always come from a synergy between experimental, technological and data processing (Wilmes 2015, Steen 2004), each generation will also require new strategies in the field of data-processing pipelines.

1.4 Current trends and limitations

Recent advances in computational power and reduction of sequencing cost have caused an exponential influx of new sequences. While Moore's law projects a doubling of transistors per circuit every two years, the rate at which new sequence are submitted steadily outpaces the relative growth of computational power (Karl Rupp 2020, Yarza 2014, Nasko 2018, Schoch 2020). The increased availability of next generation sequencing by Illumina, and long read sequencing with nanopores, that enabled labs to sequence locally with MinIONs®, has led to a "space-race" of diversity where each scientific report aims to increase the number of reported organisms (Parks 2017, Passoli 2019, Nayfach 2021). Since the throughput of mass spectrometry-based proteomics is still lower than that of genome sequencing, it lags behind with the number of sequence identifications. Traditional metaproteomic spectrum matching suffers in performance from the increased protein sequence database size. De novo peptide sequencing approaches in metaproteomics provide a promising alternative to conventional spectrum searching algorithms, as they can provide faster annotation that scales more favorably with larger databases (**Chapter 3).**

Though field of meta-omics is experiencing rapid growth, it is still in its infancy. There is limited consensus on methodologies and the "ever-shifting" databases limit comparability of studies. While attempts are made to unify methods in data processing and sample preparation, we are still far from reaching an agreement (van den Bossche 2021, Sczyrba 2017). Many databases rely on archaic phylogenies and are not well curated for annotating microbial communities. Since most newly added genomes are incomplete, growing database discrepancies arise between 16S and metagenomics databases. The genome taxonomy database (GTDB) updates this taxonomic system (Parks 2018). Metaproteomics has recently developed itself as a method to quantify the taxonomic distribution in communities, next to DNA-based techniques (Kleiner 2017). However, there is a clear lack of comparative studies between the three techniques. While the quantitative biases between 16S and metagenomics have been thoroughly researched, the quantification differences between metagenomics and metaproteomics and the impact of different taxonomic databases are still a large unknown for complex communities **(Chapter 2)**.

1.5 Thesis Outline

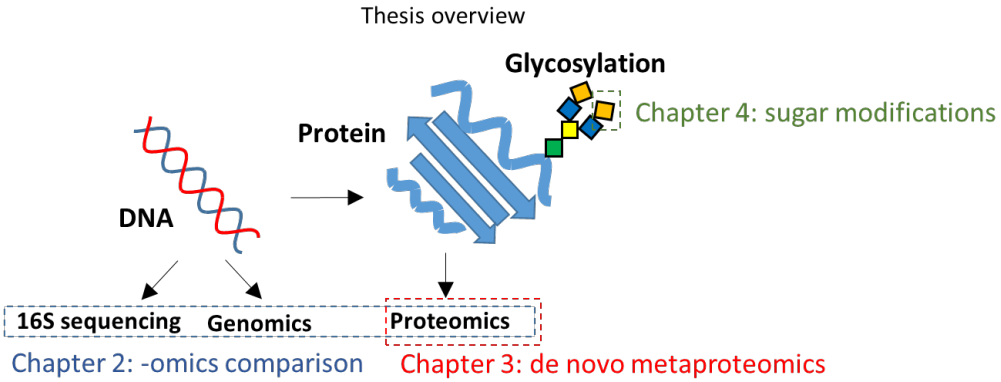


Fig. 1.6 An overview of research chapters within the thesis, and how they are positioned within biology. Chapter 2: compares 16S sequencing, genomics, and proteomics, while Chapter: 3 focuses on proteomics only, and Chapter 4 applies mass spectrometry to study sugar modifications.

Apart from chapter 1, there are 3 research chapters 2-4 (Fig. 1.6):

Chapter 2: Comparative metaproteomics demonstrates different views on the complex granular sludge microbiome

Several methods and databases exist to quantify the contribution of microorganisms within a community. Multiple recent studies benchmark the impact of parameters and approaches on a single omics technique across different labs. However, there is lack of comparative studies that analyze the influence of database choice, and the use different omics techniques for the same sample set. In **Chapter 2**, we provide an integrative study combining three popular methods of taxonomic quantification (metaproteomics, metagenomics and 16S rRNA sequencing) annotated with several different databases. We then compare the impact and origins of their quantitative biases, to further position metaproteomics as a technique in environmental biotechnology.

Chapter 3: Database-independent de novo metaproteomics of complex microbial communities

For metaproteomics to move forward and improve overall throughput, new approaches are required. A promising strategy that takes advantage of the increased mass accuracy and improved peptide fragmentation of new generation mass spectrometers is de novo sequencing. This method directly translates mass peaks into amino acid sequences, followed by string matching to a peptide database, which greatly speeds up the computational aspect of proteomics, while reducing database bias (Muth 2018). In **Chapter 3**, we validate de novo sequencing as a faster alternative to conventional spectrum search approach and provide an efficient annotation pipeline for taxonomy and function.

Chapter 4: Tackling the chemical diversity of microbial nonulosonic acids – a universal large-scale survey approach

Bacterial protein glycosylation presents a facet of phenotypes that is challenging to predict from genotypical information. Sialic acids are a particularly complex class of sugar modifications that are abundant within many environmental microbes. By leveraging the flexibility of mass-spectrometry, a method was designed to screen sialic acids in complex samples. By combining a data-independent acquisition method with a cheminformatics data-processing pipeline, an untargeted method was developed in **Chapter 4** for measuring sialic acids in complex microbial communities and discovering potential novel modifications.

1.6 References

- Adams, Mark D., et al. "3,400 new expressed sequence tags identify diversity of transcripts in human brain." Nature genetics 4.3 (1993): 256-267.
- Amend, Anthony S., et al. "Microbial response to simulated global change is phylogenetically conserved and linked with functional potential." The ISME journal 10.1 (2016): 109-118.
- Atkins, John F., and Pavel V. Baranov. "The distinction between recoding and codon reassignment." Genetics 185.4 (2010): 1535-1536.
- Brennicke, Axel, Anita Marchfelder, and Stefan Binder. "RNA editing." FEMS microbiology reviews 23.3 (1999): 297-316.
- Couso, Juan-Pablo, and Pedro Patraquim. "Classification and function of small open reading frames." Nature reviews Molecular cell biology 18.9 (2017): 575-589.
- Doolittle, W. Ford, and Olga Zhaxybayeva. "On the origin of prokaryotic species." Genome research 19.5 (2009): 744-756.
- Duval, Mélodie, and Pascale Cossart. "Small bacterial and phagic proteins: an updated view on a rapidly moving field." Current opinion in microbiology 39 (2017): 81-88.
- Eng, Jimmy K., Ashley L. McCormack, and John R. Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." Journal of the american society for mass spectrometry 5.11 (1994): 976-989.
- Erdjument-Bromage, Hediye, et al. "High-Sensitivity sequencing of large proteins: Partial structure of the rapamycin-fkbp12 target." Protein Science 3.12 (1994): 2435-2446.
- Fenn, John B., et al. "Electrospray ionization for mass spectrometry of large biomolecules." Science 246.4926 (1989): 64-71
- Gillet, Ludovic C., et al. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis." Molecular & Cellular Proteomics 11.6 (2012).
- de Graaff, Danny R., et al. "Sialic acids in the extracellular polymeric substances of seawater-adapted aerobic granular sludge." Water research 155 (2019): 343-351.
- Jaspers, Elke, and Jorg Overmann. "Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies." Applied and environmental microbiology 70.8 (2004): 4831-4839.
- Karas, Michael, and Franz Hillenkamp. "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons." Analytical chemistry 60.20 (1988): 2299-2301.
- Keller, Andrew, et al. "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." Analytical chemistry 74.20 (2002): 5383-5392.
- Lewis, Amanda L., et al. "Innovations in host and microbial sialic acid biosynthesis revealed by phylogenomic prediction of nonulosonic acid structure." Proceedings of the National Academy of Sciences 106.32 (2009): 13552-13557.
- Meydan, Sezen, Nora Vazquez-Laslop, and Alexander S. Mankin. "Genes within genes in bacterial genomes." Microbiology spectrum 6.4 (2018): 6-4.
- Muth, Thilo, et al. "A potential golden age to come—current tools, recent use cases, and future avenues for de novo sequencing in proteomics." Proteomics 18.18 (2018): 1700150.
- Nayfach, Stephen, et al. "A genomic catalog of Earth's microbiomes." Nature biotechnology 39.4 (2021): 499-509.
- Orr, Mona Wu, et al. "Alternative ORFs and small ORFs: shedding light on the dark proteome." Nucleic Acids Research 48.3 (2020): 1029-1042.
- Pabst, Martin, et al. "A general approach to explore prokaryotic protein glycosylation reveals the unique surface layer modulation of an anammox bacterium." The ISME Journal (2021): 1-12.

- Pace, Norman R., Jan Sapp, and Nigel Goldenfeld. "Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life." Proceedings of the National Academy of Sciences 109.4 (2012): 1011-1018.
- Parks, Donovan H., et al. "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life." Nature microbiology 2.11 (2017): 1533-1542.
- Pasolli, Edoardo, et al. "Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle." Cell 176.3 (2019): 649-662.
- Patterson, Scott D., and Ruedi H. Aebersold. "Proteomics: the first decade and beyond." Nature genetics 33.3 (2003): 311-323.
- Peng, Junmin, et al. "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC− MS/MS) for large-scale protein analysis: the yeast proteome." Journal of proteome research 2.1 (2003): 43-50.
- Pinel, Ingrid SM, et al. "Sialic acids: An important family of carbohydrates overlooked in environmental biofilms." Applied Sciences 10.21 (2020): 7694.
- Ram, Rachna J., et al. "Community proteomics of a natural microbial biofilm." Science 308.5730 (2005): 1915-1920.
- Ruiz-Orera, Jorge, and M. Mar Albà. "Translation of small open reading frames: roles in regulation and evolutionary innovation." Trends in Genetics 35.3 (2019): 186-198.
- Schleifer, Karl Heinz. "Classification of Bacteria and Archaea: past, present and future." Systematic and applied microbiology 32.8 (2009): 533-542.
- Sczyrba, Alexander, et al. "Critical assessment of metagenome interpretation—a benchmark of metagenomics software." Nature methods 14.11 (2017): 1063-1071.
- Staley, James T., and Allan Konopka. "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats." Annual review of microbiology 39.1 (1985): 321-346.
- Steen, Hanno, and Matthias Mann. "The ABC's (and XYZ's) of peptide sequencing." Nature reviews Molecular cell biology 5.9 (2004): 699-711.
- Sutcliffe, Iain C. "Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again." Frontiers in genetics 6 (2015): 218.
- Vandamme, Peter, et al. "Polyphasic taxonomy, a consensus approach to bacterial systematics." Microbiological reviews 60.2 (1996): 407-438.
- Van Den Bossche, Tim, et al. "Critical Assessment of Metaproteome Investigation (CAMPI): A Multi-Lab Comparison of Established Workflows." (2021).
- Varki, Ajit. "Sialic acids in human health and disease." Trends in molecular medicine 14.8 (2008): 351-360.
- Varki, Nissi M., and Ajit Varki. "Diversity in cell surface sialic acid presentations: implications for biology and disease." Laboratory investigation 87.9 (2007): 851-857.
- Washburn, Michael P., Dirk Wolters, and John R. Yates. "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nature biotechnology 19.3 (2001): 242-247.
- Wilmes, Paul, and Philip L. Bond. "The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms." Environmental Microbiology 6.9 (2004): 911-920.
- Wilmes, Paul, Anna Heintz-Buschart, and Philip L. Bond. "A decade of metaproteomics: where we stand and what the future holds." Proteomics 15.20 (2015): 3409-3417.
- Woese, Carl R., and George E. Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proceedings of the National Academy of Sciences 74.11 (1977): 5088-5090.
- Zuckerkandl, Emile, and Linus Pauling. "Molecules as documents of evolutionary history." Journal of theoretical biology 8.2 (1965): 357-366.

2

Comparative metaproteomics demonstrates different views on the complex granular sludge microbiome

A more updated version of this chapter is described in a preprint:

Kleikamp, H. B. C., Grouzdev, D., Schaasberg, P., van Valderen, R., van der Zwaan, R., van de Wijgaart, R., Lin, Y., Abbas, B., Pronk, M., van Loosdrecht, M.C. and Pabst, M., *bioRxiv* (**2022**)

2.1 Abstract

While sequencing developments have made DNA-based analyses the norm in field of microbial ecology, recent technological advances in metaproteomics showed that protein-based quantification offers an improved representation of biomass within a sample. However, there is a lack of comparative studies that delineate the observed quantification differences between the methods. Here, we systematically compare three microbiomes obtained from aerobic granular sludge wastewater treatment plant communities as derived from 16S rRNA gene sequencing, metagenomics and metaproteomics. To enable a quantitative comparison of the different approaches, we homogenized the Genome Taxonomy Database (GTDB) as uniform reference database for a contig-level metaproteomic database construction, and compared it to annotations done with conventionally used protein and 16S databases. By integrating multiple functional databases, we show that DNA-based methods tend to underrepresent key genera in the nutrient removal process, and we categorize sources of quantification bias into physiological, experimental, and computational sources, and compare their impacts. This systematic study underlines the complexity of quantification biases, the importance of orthogonal metaproteomic approaches (when characterizing complex microbiomes) and showcases the necessity of accurate reference sequence databases for enhancing comparability and accuracy in scientific reporting.

2.2 Introduction

In the past decades we have come to realize the instrumental role of microbial communities in health and biogeochemical cycles (Cho 2012, Falkowski 2008), and have started to apply their potential for conversions in industrial processes to produce energy and chemicals in a bio-based society (Angenent 2004, Rabaey 2005). A key industrial process that applies microbial communities is wastewater treatment, which will rise in priority as water-scarcity increases.

The challenge of designing processes around complex communities lies not only in the identification, but also in the quantification of its major constituents and their function. Next generation sequencing (NGS) has created studies of unprecedented scale that map the prevalence of specific taxa and led to established core microbiomes in wastewater treatment (Wu 2019, Saunders 2015, Dueholm 2021). Next to DNA-based methods such as 16S rRNA sequencing and shotgun metagenomics, recent technical advancements have enabled metaproteomics to analyze complex microbial communities (Kleiner 2019). However, significant quantification differences are observed between techniques. This was highlighted with mock communities, which first revealed cell volume as a main origin of quantification difference (Kleiner 2017). However, since metaproteomics was shown to have reduced performance as community complexity increases (Lohman 2020), quantification biases also need to be investigated in the context of highly complex communities.

Wastewater treatment is struggling with a lack of consensus on key role-players (Stokholm-Bjerregaard 2017). Traditional quantifications by 16S sequencing or qFISH revealed conflicting results (Albertsen 2012), while the addition of metagenomics and metaproteomics showed consistent discrepancies for the genus Accumulibacter, which is essential in phosphate removal (Barr 2016, Welles 2015, Azizan 2020). Recently, large-scale screening studies have led to the development of the wastewater specific MiDAS database, and the establishment of a core microbiome for activated sludge (McIlroy 2015 2017 Nierychlo 2020).

The aerobic granular sludge (AGS) technology is an emerging efficient wastewater technology with high space and energy efficiency (Pronk 2015a 2015b, Świątczak 2018, Ali 2019), but has even less consensus on its major role-players (Leventhal 2018, Ali 2020). Several roles in nutrient removal have been involved in the water purification process, such as phosphate accumulating organisms (PAO), glycogen accumulating organisms (GAO), nitrite-oxidizing bacteria (NOB), ammonia-oxidizing bacteria (AOB), and nitrate reducers (NR) (Weissbrodt 2013 2014, Szabó 2017). Next to their role in nutrient removal, bacteria maintain the structure of granular sludge by producing extracellular polymeric substances (EPS) (Adav 2009, Panchavinin 2019, Liang 2019). Therefore, knowing the composition of the core microbiome is a key step towards controlling the plants performance, and engineering more efficient processes.

Because of its low costs and ease-of-use, 16S rRNA gene sequencing has been the most frequently employed -omics approach for large sample sets and has been applied to both obtain a temporal (Zhou 2020, Ramos 2015, de Sousa Rollemberg 2019) and spatial resolution (Wu 2019, Zhang 2012, Ali 2019). When it comes to quantification accuracy, 16S sequencing is known to be biased due to varying 16S gene copy numbers (Stoddard 2015), which cannot be corrected for (Starke 2021, Louca 2018, Edgar 2017), and is further exacerbated by varying primer efficiencies (Brown 2015, Albertsen 2015, Brooks 2015). Additionally, functional annotation can only be done predictively (Morrissey 2016, Doolittle 2009, Jaspers 2004).

Whole genome shotgun sequencing (also referred to as metagenomics) is more costly and computationally demanding compared to 16S rRNA sequencing but provides a significantly higher taxonomic resolution and genome coverage. This enables differentiation between functionally distinct clades (e.g., Rubio-Rincón 2019), and provides insights into the metabolic potential of the individual community members. While considered to be less biased than 16S, both suffer from extraction biases that reduce recovery efficiency of DNA of certain taxa (Albertsen 2012, Pronk 2017).

Metaproteomics – the identification and quantification of proteins from complete microbial communities – is one of the most promising post-genomics approaches and has been applied on several occasions to activated sludge (Hagen 2013, Wilmes 2008 2015, Püttker 2015). Because the composition obtained by metaproteomics correlates to the mass of individual community members (Kleiner 2017), rather than to the number of cells as obtained for genomics approaches, metaproteomics can distinguish active and inactive fractions of biomass (Adav 2008), and can differentiate between distinct phenotypes (Nielsen 2019, Stokholm-Bjerregaard 2017, Welles 2015). Conventional metaproteomics approaches are hampered by a limited throughput and reduced sensitivity and taxonomic resolution compared to DNA-based techniques. Metaproteomics is therefore commonly paired with metagenomics to provide community-specific protein sequence databases (Narayanasamy 2015, Timmins-Schiffman 2017).

The first major hurdle towards an integrative comparison of analytical techniques is the use of different databases, which employ varying nomenclatures and phylogenies. These include 16S-specific databases, such as SILVA (Quast 2012) and the wastewater specific database MiDAS (Nierychlo 2020, Mcllroy 2017), and protein-specific databases such as UniprotKB, Uniref100,90,50 & Swiss-Prot (Boutet 2007), as well as general databases which include both 16S genes, genomes, and proteins, such as RefSeq (Pruitt 2005), and more recently the genome taxonomy database (GTDB) (Parks 2018). Since databases use varying methods of phylogenetic placement, there is significant divergence in their nomenclatures.

Modern phylogenetic placement tools include 16S gene % similarity (Yarza 2014), amino acid identity % (AAI) and average nucleotide identity % (ANI) (Konstantinidis 2005). The NCBI taxonomy, which is applied in RefSeq, and UniprotKB-derived databases uses a mixture of historical taxonomies and modern placement methods (Federhen 2012, Schoch 2020), and lacks rank normalization, which results in gapped lineages that lack one or more taxonomic ranks. Both SILVA and NCBI taxonomies contain dump taxa, which lump together groups of uncultured organisms (Hugenholz 2016).

The MiDAS database builds on the SILVA taxonomy, but adds several additional unique taxa and sequences, which get proprietarily named with the AutoTax system (Dueholm 2020). GTDB solves these issues by using a common placement method to normalize taxonomic ranks based on relative evolutionary divergence (Parks 2018).

Another challenge is the database completeness. Because of their divergent GC-content and multiple copies, 16S genes are lacking in many newly deposited metagenome-assembled genomes (MAGs) (Soo 2014, Rahman 2015). This requires an additional step of database homogenization to ensure that every taxon is represented equally. To prevent binning approaches from skewing taxonomic resolution towards specific taxa (Nelson 2020, Browne 2020, Chen 2019, Sieber 2019), a modified contig annotation method (von Meijenfeldt 2019) was employed to maximize the number of genus-level annotated open reading frames, while remaining closer to the diversity of 16S annotations.

In this study, we systematically compare 16S rRNA sequencing, shotgun metagenomics and metaproteomics for sludge samples from three full-scale AGS plants: Dinxperlo (DXP), Garmerwolde (GW) and Simpelveld (SP). We compare annotations done with homogenized GTDB to 16S databases SILVA and MiDAS, as well as protein databases derived from UniprotKB (Swiss-Prot+TrEBML, UniRef50,90,100) and RefSeq (non-redundant, redundant). Functional analysis is done by the integration of domain annotations by KEGG (Kanehisa 2016), COG terms (Tatusov 2000), PFAM (Mistry 2021) and TIGRFAM (Haft 2003), as well as UniprotKB genes (Boutet 2007). Lastly, we distinguish physiological, experimental, and computational sources of bias to identify the main variables that affect quantification differences between metaproteomics and metagenomics.

2.3 Results

2.3.1 Standardized contig-based reference database for metaproteomics

A standardized database with an accurate taxonomy is essential when describing the microbial diversity and when comparing scientific data between different techniques (Parks 2018, Godfray 2002). Here we demonstrate the application of the recently established genome taxonomy database (GTDB), that normalizes taxonomic ranks based on relative evolutionary divergence by a set of conserved proteins (Parks 2018 2020 2021, Chaumeil 2020), for a contig-based metaproteomic sequence database construction.

The Genome Taxonomy Database Toolkit (GTDB-Tk) was developed to classify bacterial and archaeal draft genome assemblies rather than contigs or reads (Chaumeil 2020, Sedlar 2017, Lin 2021). In metagenomics, the clustering and binning of contigs into individual genomes often leaves substantial unbinned fractions, which skews the taxonomic representation towards the more complete genomes present in the community. Therefore, to provide the most comprehensive sequence database for the metaproteomic analysis, and to standardize the taxonomic annotation between metaproteomics and metagenomics data, we performed taxonomic classification at the contig level.

Apart from rank normalization, another advantage of GTDB is that its 16S sequences share the same taxonomy. Silva and Midas, on the other hand, rely on 16S-derived phylogenies, while NCBI uses a combination of placements methods (Parks 2017). When comparing the three techniques, there are two key factors that influence the similarity: database homogenization and data processing. Since 16S genes are notoriously hard to include into assembled genomes due to their diverging GC-content and varying copy numbers (Soo 2014, Rahman 2015), there is a discrepancy between databases that causes a varying representation of taxa (SI Fig S2.3-S2.5). Of the representative GTDB genomes, 14% contains a fragmented 16S sequence shorter than 1200 base pairs, and 31% has no 16S sequence representative (SI Table S2.1).

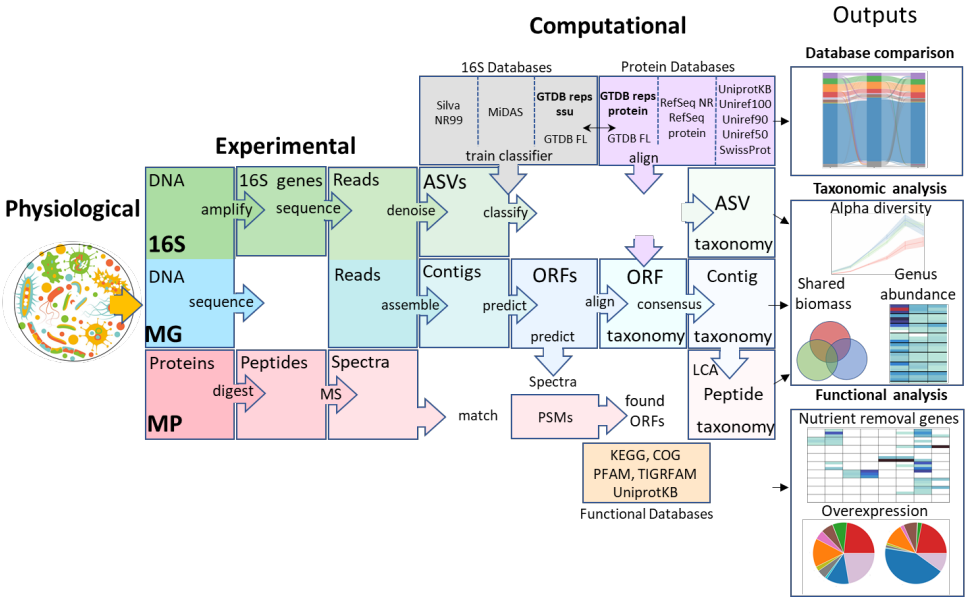


Fig. 2.1 Workflow diagram of techniques and databases used for the comparative study analyzing the microbiome of aerobic granular sludge. In addition to metaproteomics (red), orthogonal metagenomics (blue) and 16S rRNA amplicon sequencing (green) were performed. Metaproteomic spectra were matched against a metagenomic contig-based protein sequence reference database, to establish the peptide-spectrum matches (PSMs). In addition to multiple 16S databases (grey) used to annotate amplicon sequence variants (ASVs), several protein sequence databases (purple) were used for the taxonomic classification of the contig-based reference sequence database. To improve comparability, a homogenized GTDB database was constructed, that could be employed for the classification of all different techniques. The different outputs were finally compared for potential taxonomic and functional divergences between databases and Omics techniques.

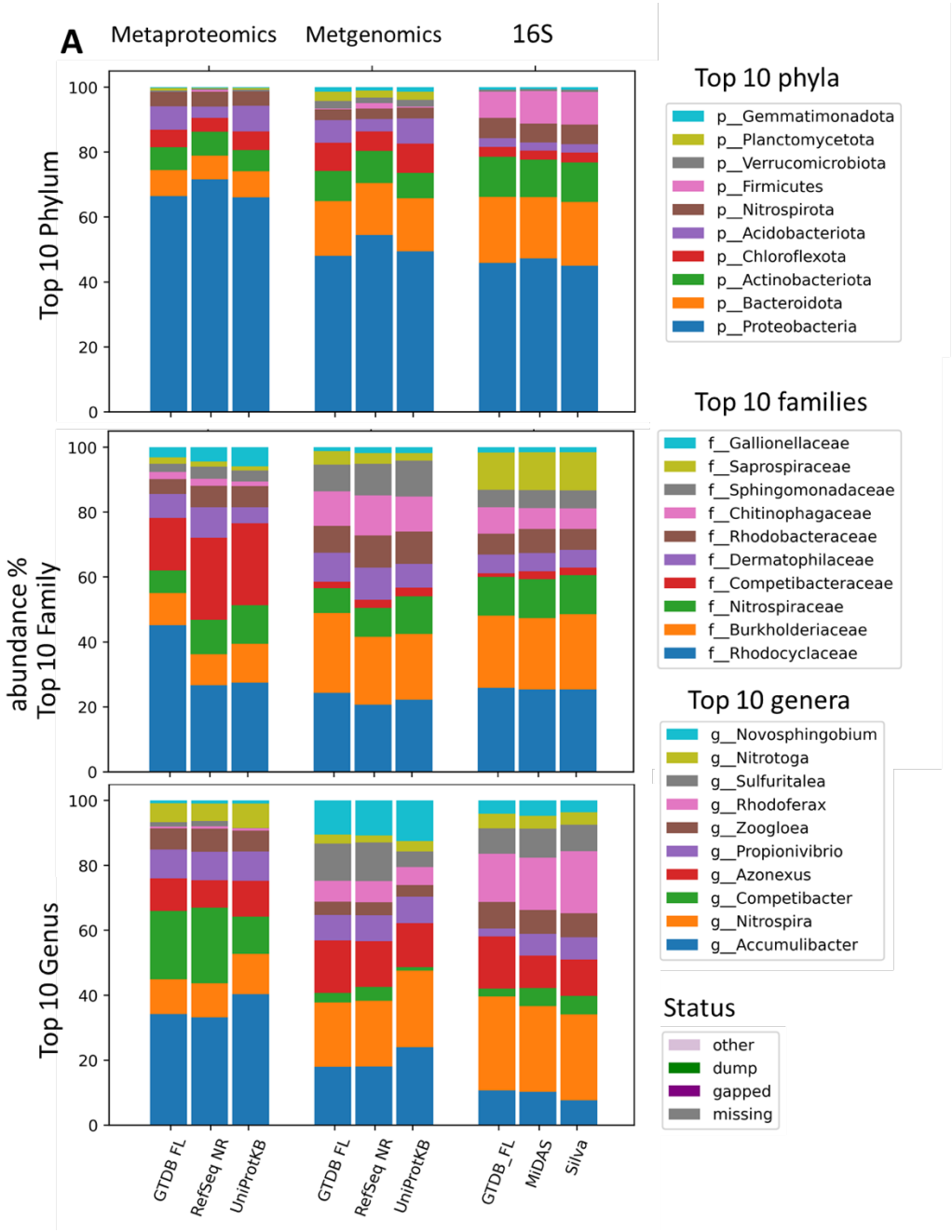
Therefore, GTDB was homogenized to select equally represented taxa with a full length 16S sequence. Comparatively, annotation with non-homogenized GTDB resulted in ~5% of total reads/peptide spectrum matches (PSMs) being matched to organisms not present in the 16S database (SI Fig. S2.7). Within this study, homogenized GTDB databases have been applied under the moniker GTDB FL (Full-Length). To allow an integrative comparison of this homogenized database, we established two tools: GTDB2DIAMOND and GTDB2QIIME (available from Github) that format GTDB sequence files for taxonomic annotation with the software tools DIAMOND and QIIME2, which are respectively used in the annotation of open reading frames (ORFs) and amplicon sequence variants (ASVs).

For contig consensus annotation, we provide two modifications to the CAT-algorithm, which constructs a consensus lineage for each contig from the lineages of its individual ORFs by selecting taxa based on their fraction of the total bit-score (von Meijenfeldt 2019). Since the high stringency of the default CAT algorithm resulted in reduced numbers of genus-level annotations, we adjusted the CAT algorithm to maximize genus level annotated contigs. To compensate, we added an additional filtering step to target contigs with only a few annotated ORFs, which can be less robust, and optimized parameters to obtain a similar level of diversity as 16S amplicon data (SI Fig. S2.1, S2.2, S2.6).

2.3.2 Database comparison

Annotations with the homogenized GTDB FL database were compared to Silva NR99 and the wastewater specific MiDAS database for 16S, and UniprotKB and RefSeq for metagenomics and metaproteomics. Since there are significant nomenclature differences between the databases, nomenclatures were standardized as much as possible. Some examples of nomenclature differences include "–a" versus "–ota" for phylum suffixes, the use of *Candidatus* prefixes, and GTDB unique sub-taxa annotation (Firmicutes_A, Firmicutes_B etc.), but can also include phylogenetic rearrangements, such as betaproteobacteria in GTDB. Without removing these small mismatches in nomenclature, an exact comparison results in drastic differences between databases, even though in most cases they effectively refer to the same taxa. To achieve this, the links between taxonomy were mapped using GTDB and Silva metadata, renaming taxa with more than 75% overlap to GTDB nomenclature.

When looking at the relative distribution of the top 10 most abundant taxa, limited effects are observed for database choice (Fig. 2.2A), but more so for the choice of technique. At phylum level, Proteobacteria are more dominant in metaproteomics than in metagenomics and 16S, while 16S has a reduced presence of Chloroflexeota and increased Firmicutes. On a family level, metaproteomics has a significantly higher abundance of Competibacteraceae, while genus level shows increased Competibacter, Accumulibacter and Nitrotoga for proteomics, with a decrease in Novospingobium, Rhodoferax and Sulfuritalea.


A
Metaproteomics
Metgenomics
16S
Top 10 Phylum
Top 10 Family
Top 10 Genus
abundance %
0
20
40
60
80
100
GTDB FL
RefSeq NR
UniProtKB
GTDB FL
RefSeq NR
UniProtKB
GTDB_FL
MiDAS
Silva
Top 10 phyla
p__Gemmatimonadota
p__Planctomycetota
p__Verrucomicrobiota
p__Firmicutes
p__Nitrospirota
p__Acidobacteriota
p__Chloroflexota
p__Actinobacteriota
p__Bacteroidota
p__Proteobacteria
Top 10 families
f__Gallionellaceae
f__Saprospiraceae
f__Sphingomonadaceae
f__Chitinophagaceae
f__Rhodobacteraceae
f__Dermatophilaceae
f__Competibacteraceae
f__Nitrospiraceae
f__Burkholderiaceae
f__Rhodocyclaceae
Top 10 genera
g__Novosphingobium
g__Nitrotoga
g__Sulfuritalea
g__Rhodoferax
g__Zoogloea
g__Propionivibrio
g__Azonexus
g__Competibacter
g__Nitrospira
g__Accumulibacter
Status
other
dump
gapped
missing

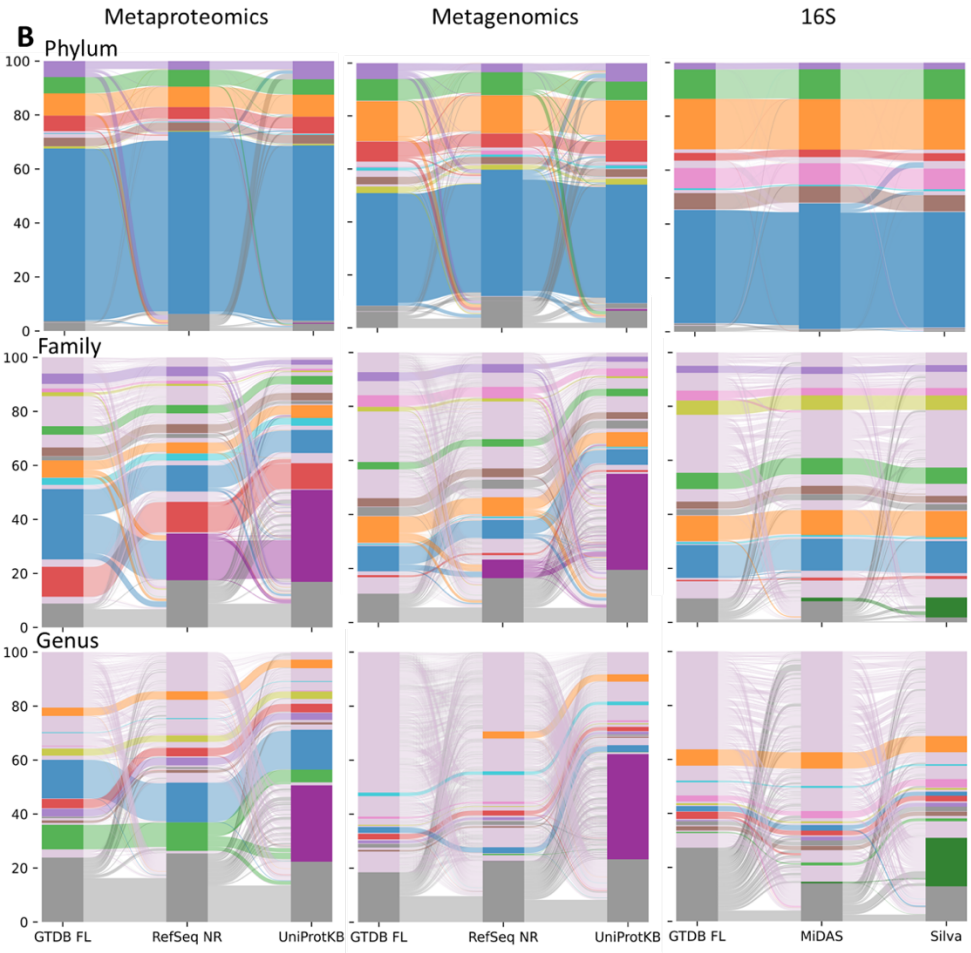


Fig. 2.2 A: Comparison of the top 10 most abundant taxa on phylum, family and genus level, averaged over the 3 combined samples, for the techniques metaproteomics (MP) metagenomics (MG) and 16S, against major protein databases (GTDB FL, RefSeq NR, UniprotKB) and 16S databases (GTDB FL, Midas, Silva), B: Sankey diagram of the relative abundance of all taxa, colored by top 10 most abundant taxa, as well as other taxa (thistle), unannotated taxa (grey), dump taxa (dark green) and gapped lineages (dark purple), for different protein and 16S databases.

The Sankey diagrams of Fig. 2.2B. show the overall shifts between taxa when using different databases. These diagrams show how unannotated sequences (grey) become annotated when using different databases, but also how they shift to gapped lineages (dark purple). Gapped lineages result from databases that lack rank normalization and are caused by incomplete lineages which lack annotation at one or more key ranks (phylum, class, order, family, genus), or by dump taxa, which consist of uncultured and poorly characterized microbes. An example of an incomplete lineage can be seen in the NCBI taxonomy of the genus Accumulibacter, which lacks order and family ranks: Proteobacteria (phylum); Betaproteobacteria (class); Betaproteobacteria incertae sedis (no rank); Candidatus Accumulibacter (genus). Dump taxa often lack even more taxonomic ranks, for example: Bacteroidetes (phylum); unclassified Bacteroidetes (no rank); Bacteroidetes bacterium (species). When constructing the Sankey diagrams, dump taxa are flagged based on a list of keywords, such as “unclassified” or “bacterium”.

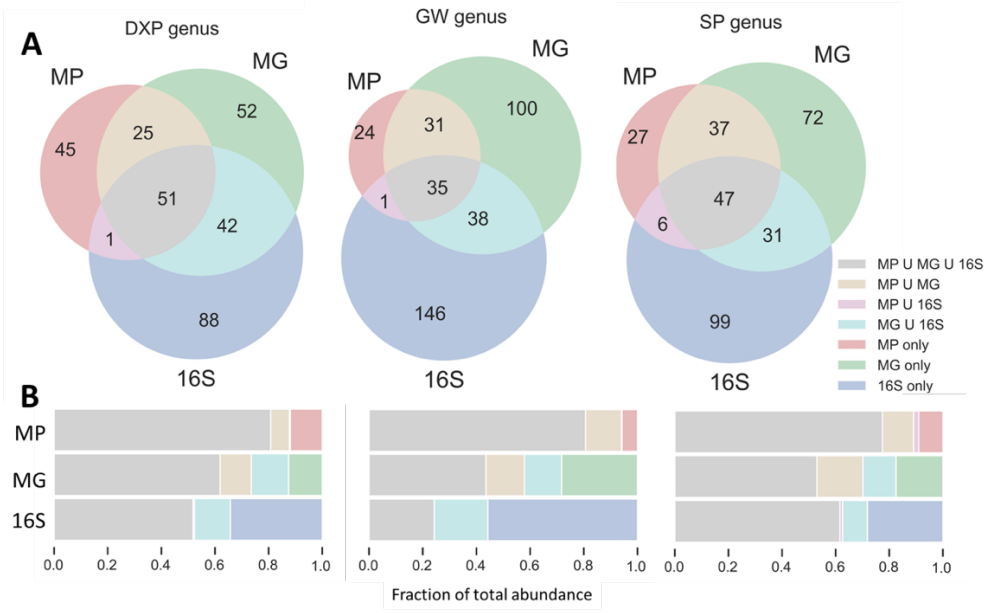
Both UniprotKB and RefSeq contain incomplete lineages, as they employ NCBI taxonomy, while UniprotKB also contains dump taxa. In metaproteomics and metagenomics annotations with UniprotKB, the total fraction of gapped lineages can be significant, and account for half of the total annotated contigs (Fig. 2.2A). This halves the number of taxonomic annotations at certain ranks, which highlights the importance of rank normalization, as is done in GTDB. For UniprotKB, dump taxa are more clearly visible at species level, while for 16S annotations with SILVA, they occur also at earlier ranks (SI Sup Fig S2.11-2.12). On the other hand, gapped lineages are not observed in Midas and SILVA. Depending on the rank, up to 30-70% of mismatched annotations can be attributed to name shifts (SI Sup Tables S2.2-2.4). This underlines the importance of a standardized nomenclature when comparing different databases. A larger comparison including other taxonomic ranks, non-homogenized GTDB databases as well as expanded and clustered databases like Uniref100,90,50, Swissprot & RefSeq nonredundant protein sequences are shown in the supporting information (SI Fig S2.8-S2.13).

Another major difference between techniques can be observed when comparing the abundance of the top 10 taxa in respect to the total sample abundance (Fig 2.2B). On phylum level the top 10 most abundant taxa cover most of the total abundance, but as ranks increase, the total sample coverage by the top 10 most abundant taxa decreases. For metaproteomics, the top 10 taxa still cover ~50% at genus level, while in DNA-based techniques they only cover a small fraction. This clearly shows the different levels of diversity obtained with the techniques and is also reflected in the reduced evenness of metaproteomics (Fig. 2.3C).

A notable absence in GTDB annotations is that of the genus Tetrasphaera, a key genus in biological phosphate removal systems (Stockholm-Bjerregaard 2017, Ali 2019), which is identified in MIDAS and Silva annotations, but not in GTDB annotations. Though its parent family Dermatophilacae is found amongst top 10 most abundant families in GTDB, these annotations are not carried over to genus level annotations due to the lowest common ancestor algorithm. A closer inspection with blast+ alignment of 16S ASVs annotated with Tetrasphaera in SILVA and Midas, to GTDB SSUs shows that *Tetrasphaera japonica* is matched to each of the ASVs, but it is not the taxon with the highest percent identity. Instead, higher identity matches are typically achieved by other genera belonging to the family Dermatophilacae (SI Fig S2.14). This might be a limitation of V3-V4 primer resolution and stem from a difference in phylogenetic placement, as GTDB reassigns many Tetraspheara taxa from NCBI to other genera.

2.3.3 Diversity comparison

Since different analysis methods rely on different denoising and confidence metrics, a fixed cutoff of 0.1% relative abundance was used to standardize data treatment. The total number of unique and shared taxa above this cutoff were visualized with Venn-diagrams (Fig 2.3.A), and their contribution to total abundance was shown in Fig 2.3B. In both DXP and SP samples, the taxa shared between the three techniques cover the major fraction of total abundance (Beige area, Fig 2.3B), while GW shows a larger number of genera unique to 16S and metagenomics (dark blue circle, Fig. 2.3A). Overall, 16S has the largest number of unique genus-level taxa, which is also reflected in its alpha-diversity metrics (Fig. 2.3C), where it has the highest richness, as well as long tailing (Fig. 2.3D) Metaproteomics has lower richness, but also lower evenness (Fig. 2.3C), indicating that it focusses on highly abundant organisms, which is also reflected in its shorter tail (Fig. 2.3D).



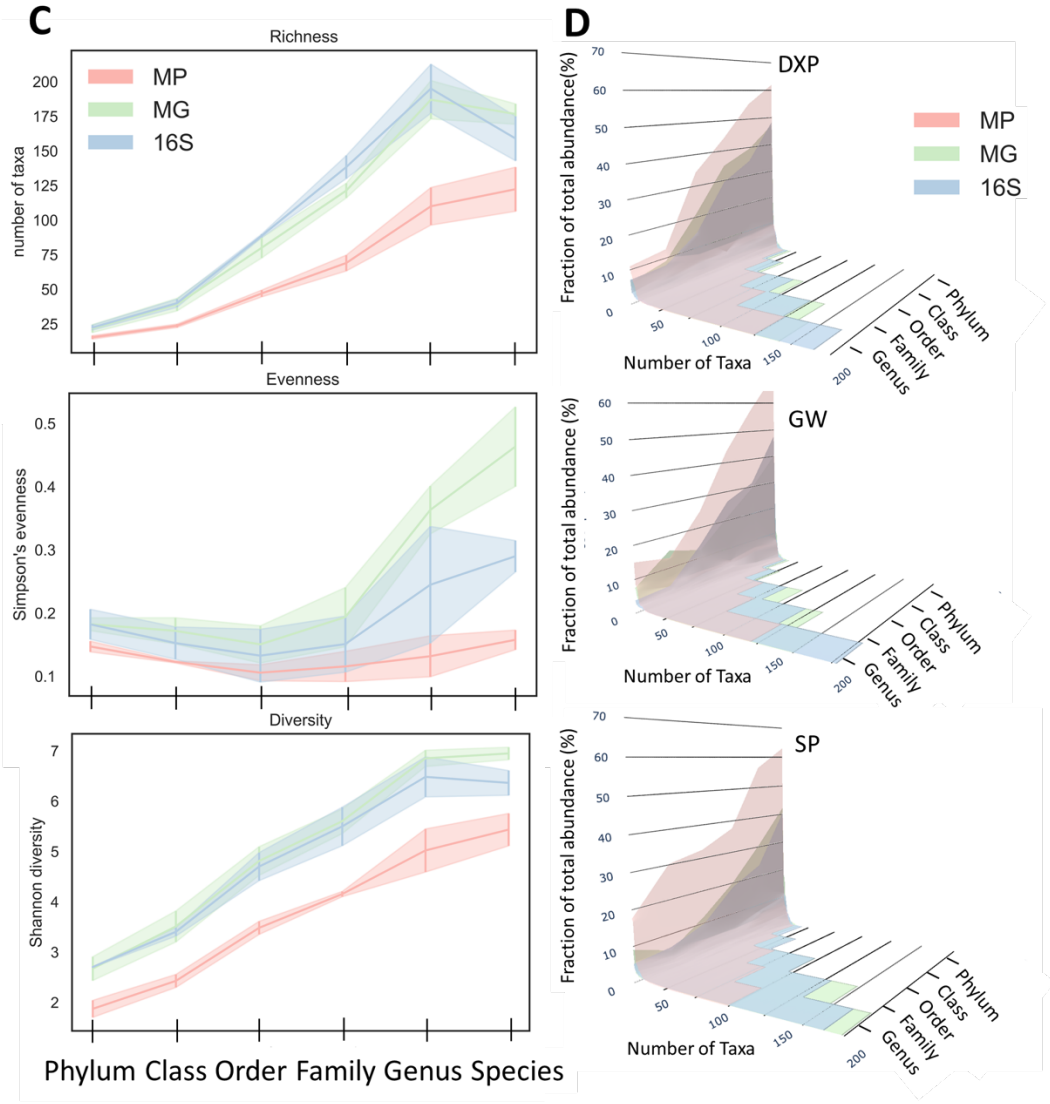


Fig. 2.3 Comparison of technique diversity and tailing A: Venn diagrams of shared genus level taxonomies between metaproteomics (MP), metagenomics (MG) and 16S from the samples Dinxperlo (DXP), Garmerwolde (GW), and Simpelveld (SP), B: The contribution of shared and unique taxa to the total biomass, C: Alpha diversity metrics for combined samples: Richness, Simpson's evenness and Shannon diversity, D: Tailing diagrams of taxa sorted by abundance for each separate plant.

The richness of the analytical technique automatically affects distribution of abundances. If a technique detects a larger number of taxa, each taxon will have a lower fraction of the total abundance compared to a technique that detects fewer taxa. Therefore, to quantitatively compare structural biases between techniques, it is essential that we look at a shared fraction of taxa, and renormalize their abundances, instead of comparing to the fraction of total abundance. Normalizing to all taxa would be disadvantageous to techniques that have higher richness (metagenomics and 16S). The taxa that were selected for further analysis (Fig. 2.4A) were required to have been measured by two techniques and either are found in more than 3% of total abundance or express key nutrient removing genes (Fig 2.4B).

2.3. 4 Functional comparison

Nutrient removal in wastewater treatment comprises the removal of nitrogen and phosphorous. Key role players in the removal of phosphorous are polyphosphate removing organisms (PAO), which remove phosphate from water by producing polyphosphate with the gene *ppk* while *ppa* hydrolyzes pyrophosphates. Glycogen accumulating organisms (GAO) synthesize glycogen with *glg*, while *blg* (beta-glucosidase) breaks down glycosides. Nitrogen removal is done in subsequent nitrification and denitrification steps, which are performed by *Hao* and *Amo* genes of ammonia-oxidizing bacteria (AOB) and *Nxr* of nitrite-oxidizing bacteria (NOB). The gene cluster *Nar, Nap* (nitrate reductase) is the first step in denitrification with subsequent *NirK, NirS* to reduce nitrite, *Nor* to convert nitric oxide to nitrous oxide and *Nos to* convert nitrous oxide to dinitrogen gas, but also the first step in DNRA/ammonification, which uses *Nrf* and *Nir* to convert nitrite into ammonia. Anammox metabolic genes *hzs* and *hdh* were observed in plant GW, which is known to be dosed biomass from a neighboring SHARON process. Lastly, cytochrome C is generally present as electron acceptor, but has been implicated to play a role in various stages of the nitrogen cycle (Black 2018, Mundinger 2019, Camejo 2019). To assess genera involved in the conversion of these components a combined annotation was done integrating KEGG, COG terms; PFAM, TIGRFAM domains; and UniprotKB genes. This method is particularly effective at annotating nitrogen metabolism, which is otherwise fragmented across different databases.

The overall relative abundance of key genera and their expressed nutrient removal genes are shown across the three plants (DXP top row, GW middle row, SP bottom row) in Fig. 2.4A,B. The order of genera is determined by their averaged quantification difference between metaproteomics-metagenomics and metaproteomics-16S (Fig. 2.5A,B). Genera quantified more strongly in metaproteomics are positioned on top and genera quantified more in DNA-based techniques are positioned at the bottom. Key roles are generally performed by the same genera in each plant. Accumulibacter is highly abundant and expresses both PAO, GAO, and denitrification pathways in each plant, while Competibacter expresses GAO metabolism, and Nitrosomonas is an AOB in each plant. The role of NOB is shared over Nitrospira and Nitrotoga, where DXP only has Nitrospira, GW has both and SP predominantly uses Nitrotoga as NOB. Denitrification is much more dispersed, but is consistently expressed in several genera, including Accumulibacter, Azoxenus and Propionivibrio, while the DXP plant has strong expression by Zoogloea.

Comparing the quantification differences between 16S-metaproteomics and metagenomics-metaproteomics (Fig. 2.5A,B) shows that several nutrient removing organisms are underrepresented in DNA-based techniques compared to metaproteomics. This is especially true for two key genera in the phosphate removal process: Accumulibacter and Competibacter (Rubio-Rincón 2017). Both of which are highly abundant in metaproteomic analysis, but less so in DNA-based techniques. For Accumulibacter, this aligns with previous observations where the genus was underrepresented in DNA-based techniques (Bar 2016, Welles 2017, Azizan 2020), however for Competibacter this observed difference has not been reported yet. Apart from these two, several other key nutrient removing genera are underrepresented in DNA-based methods, including Propionivibrio (PAO/GAO/DN), Nitrosomonas (AOB), Nitrotoga (NOB/DN) and Zoogloea (DN). However, not all genera that express nutrient removal genes are underrepresented in this manner. Azonexus (PAO/DN) and Nitrospira (NOB/DN) are quantified similarly by metagenomics and metaproteomics, and the genus Sulfuritalea (PAO/DN) is underrepresented in metaproteomics.

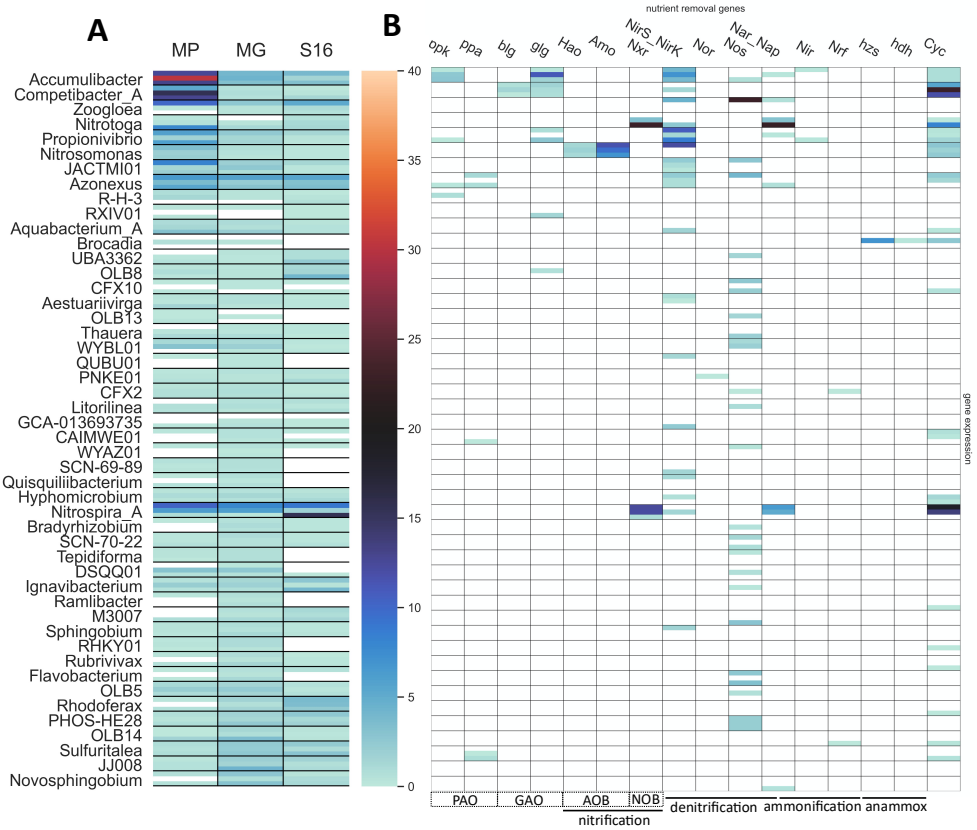


Fig. 2.4 A: The genus level abundance for a subset of genera in DXP, GW, SP (top, middle, bottom) across metaproteomics (MP), metagenomics (MG) and 16S. B: nutrient removal genes found in metaproteomics for the three samples. Genes are grouped according to associated roles: phosphate accumulating organisms (PAO), glycogen accumulating organisms (GAO), ammonium oxidizing bacteria (AOB) and nitrite oxidizing bacteria (NOB). Nitrogen metabolism is further grouped into nitrification, denitrification, ammonification, and annamox pathways. Genera are sorted according to the absolute abundance difference between metagenomics and metaproteomics (Fig. 2.6A).

Therefore, the distinction between “active and inactive” biomass does not relate directly to observed quantification differences. To see if the quantification difference instead could be linked to specific protein groups, a more general functional annotation was performed (Fig. 2.6&7). Here, emphasis was put on the relative allocation of reads in metagenomics versus peptide spectrum matches (PSMs) in metaproteomics, to see if the overexpression of certain proteins groups related to quantification differences.

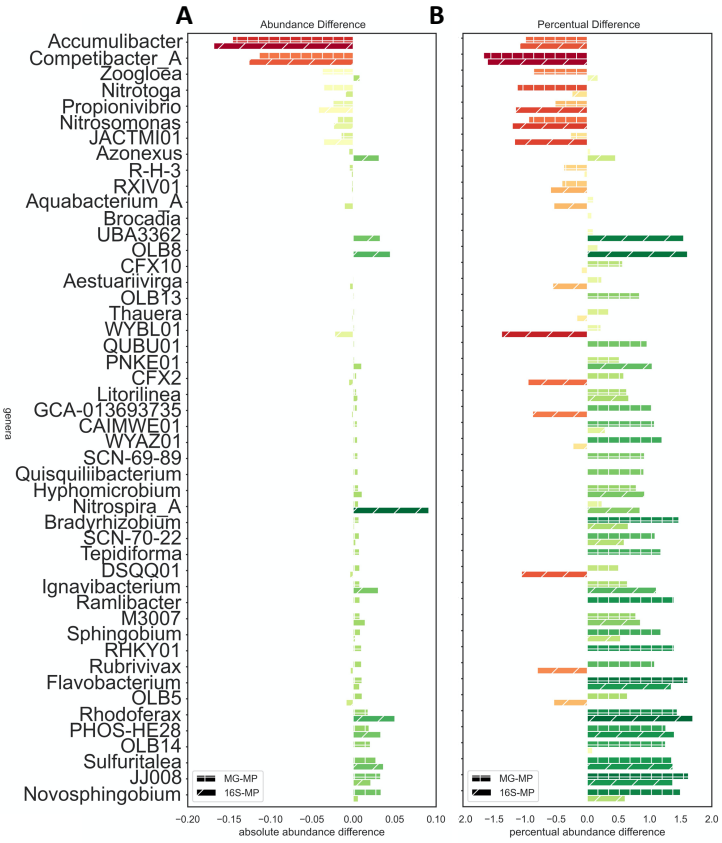


Fig. 2.5 The average abundance difference for each genus between quantification methods for metagenomics versus metaproteomics (MG-MP) (top bar) and 16S versus metaproteomics 16S-MP (bottom bar), for absolute (A) and percentual difference (B).

2.3.5 Protein expression comparison

Apart from annotating nutrient removal genes, the COG classification system was used as basis to further categorize the phenotype and strongly expressed protein groups, subdividing into “Metabolism and transport”, “Membrane associated”, “Cell cycle” and “Other” categories (Fig. 2.6). By integrating KEGG, PFAM and TIGRFAM domain annotations, 83% percent of found peptides could be grouped. Comparing the distribution of peptide spectrum matches (PSMs) and reads allocated to the proteins found in metaproteomics yields a metric of overexpression to reveal strongly expressed protein groups (Fig. 2.6).

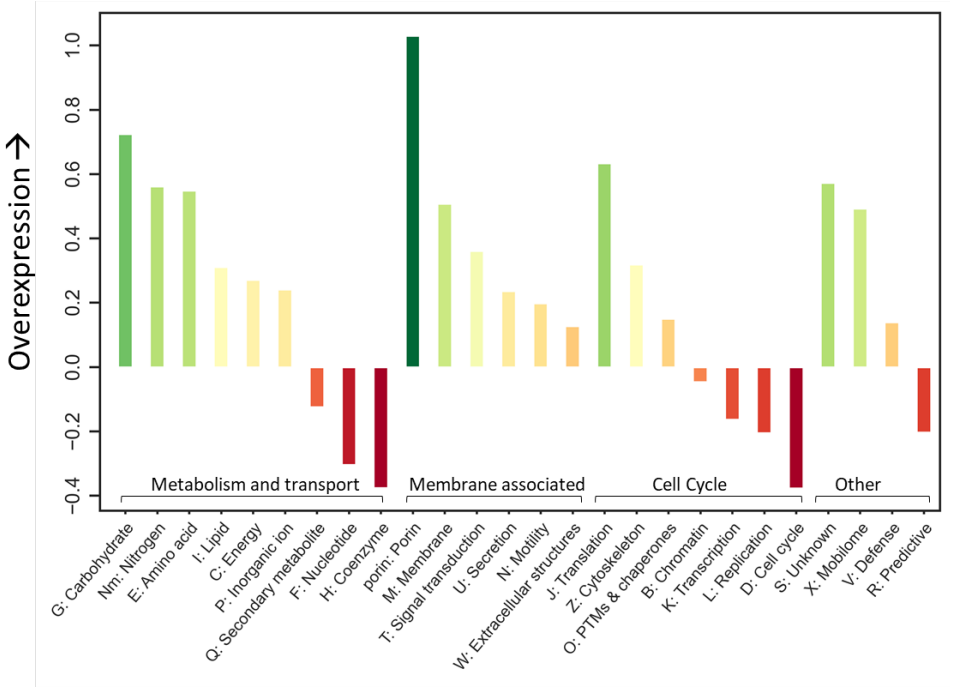


Fig. 2.6 Overexpression analysis and annotation with COG terms for combined samples, described by the percentual difference between median fold changes for peptide spectrum matches and read counts for all proteins identified in metaproteomics.

Next to protein groups associated with nutrient removal (Carbohydrate, Nitrogen, amino acid metabolism) and growth (Translation), membrane associated proteins are heavily expressed. A striking observation is the prevalence of porin proteins, a beta-barrel forming class of transporters, whose targets range from fatty acids (fadL) to small inorganic molecules (cirA, fepA, OVP1), and coenzyme transport (BtuB). Since shotgun metaproteomics generally selects for the most intense signals, overexpression of certain protein groups could be source of quantification biases observed between protein and DNA-based techniques (Fig. 2.5). The expression patterns of COG categories were further investigated for highly abundant genera with >3% relative abundance (Fig. 2.7). Each genus shows two bars: the left for the relative distribution of reads, and right for the relative distribution of PSMs for proteins found by metaproteomics. it could be observed that some genera showed significant difference between the relative allocation of reads and peptide spectrum matches of identified proteins.

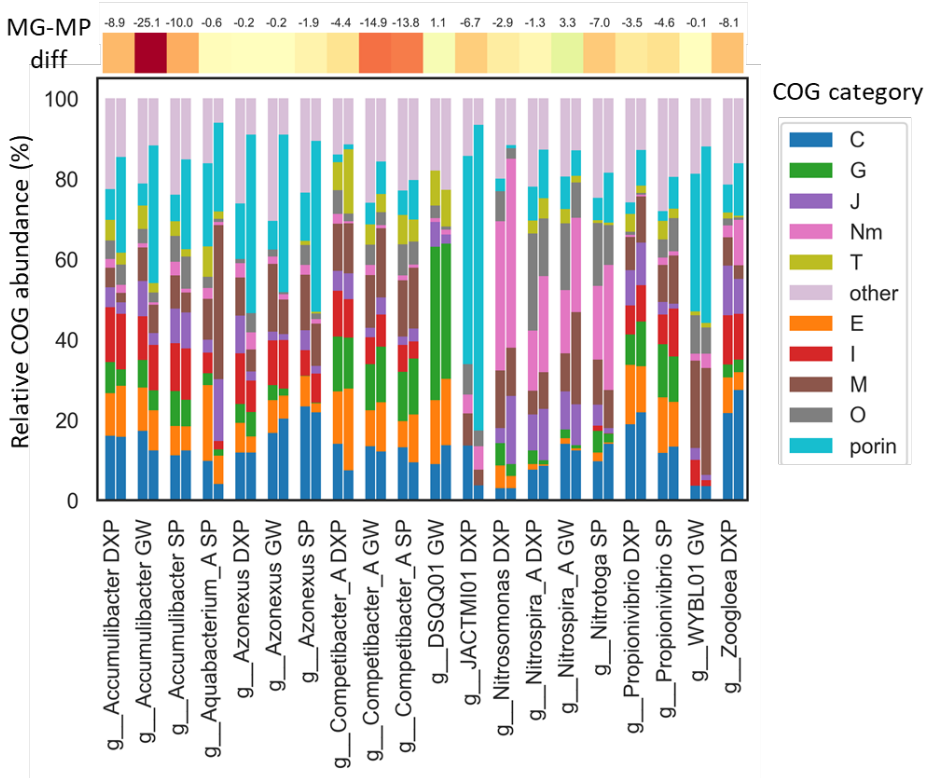


Fig. 2.7 A comparison of relative read distribution (left bar) and PSM distribution (right bar) of proteins found in highly abundant metaproteomics genera. The absolute quantification bias between metagenomics and metaproteomics is shown by the bar on top, which was calculated by subtracting relative abundance of each genus in metagenomics (MG) with the relative abundance in metaproteomics (MP). Abbreviations of COG categories are explained in Fig. 2.6.

Extreme examples include Accumulibacter and Azonexus, which strongly express proteins of the porin class, while Nitrotoga and Nitrospira highly express proteins related to nitrogen metabolism. The presence of strongly expressed membrane protein groups however does not directly equate to a quantification bias (Fig. 2.7), as Competibacter and Propionivibrio show little difference in the allocation of PSMs and reads but are under-quantified in DNA-based techniques, while Aquabacterium_A, Azonexus all are quantified similarly by DNA-based techniques and metaproteomics but show significant differences between allocation of reads and PSMs. These factors indicate that quantification bias cannot be linked directly to expression profiles.

2.3.6 Further bias investigation

Since no general trend was observed between quantification bias and the expressed proteins, an in-depth look was performed on different sources of bias. For 16S analysis, this bias is well described and attributed to variations in 16S gene copy number and primer choice. On phylum rank, these effects are clearly visible (Fig. 2.8A). Increased 16S gene copy number resulted in an overestimation of Firmicutes, while a reduced copy number led to lower Acidobacteriota and Verrucomicrobiota. The V3-V4 primers used in this study did not properly amplify Chloroflexeota, which was underestimated, and Planktomycetota, which was not detected by 16S. A cumulative sum of the total abundance with increasing 16S gene copy number is shown in Fig 2.8B. This shows that genera with a copy number of less than 5 constitute 95% of all abundance in metagenomics and metaproteomics, but only 80% in 16S.

Because quantification biases can come from multiple angles, their influence on measured percentual differences should not be taken separately, but as a combinatorial "bias phenotype". Here, we can distinguish bias that stems from cell physiology (P), experimental bias that is inherent to decisions made in experimental design (E), and computational bias that relates to decisions made in data processing (C). By using multiple linear regression, weights can be assigned to separate standardized metrics to assess their contribution to the overall observed quantification differences. By multiplying the standardized metrics with the weights calculated from multiple regression, a weighted clustergram was constructed (Fig 2.8C, 2.9). The metrics used for comparison of 16S and metagenomics in this way include: 16S gene copy number (P1), primer efficiency (E1) and BLAST vs Bayesian classification (C1). Copy number (P1) is inherent to cell physiology, while primer efficiency (E1) is classified as an experimental bias. As computational bias, the effect of Bayesian classifiers vs alignment was tested by aligning 16S sequences with BLAST+ (C1) (SI Fig S2.15). The largest impact was seen for primer efficiency, which affected the detection of Brocadia, and the Chloroflexota genera OLB13,14 & CFX2,10. Parameters P1 & C1 had a lower effect and correlated poorly with observed differences. Only an R2 of 0.07 was reached in multiple regression, indicating poor correlation.

This matches prior reports that indicate that even if the source is known, 16S bias cannot be corrected for (Starke 2021, Edgar 2017).

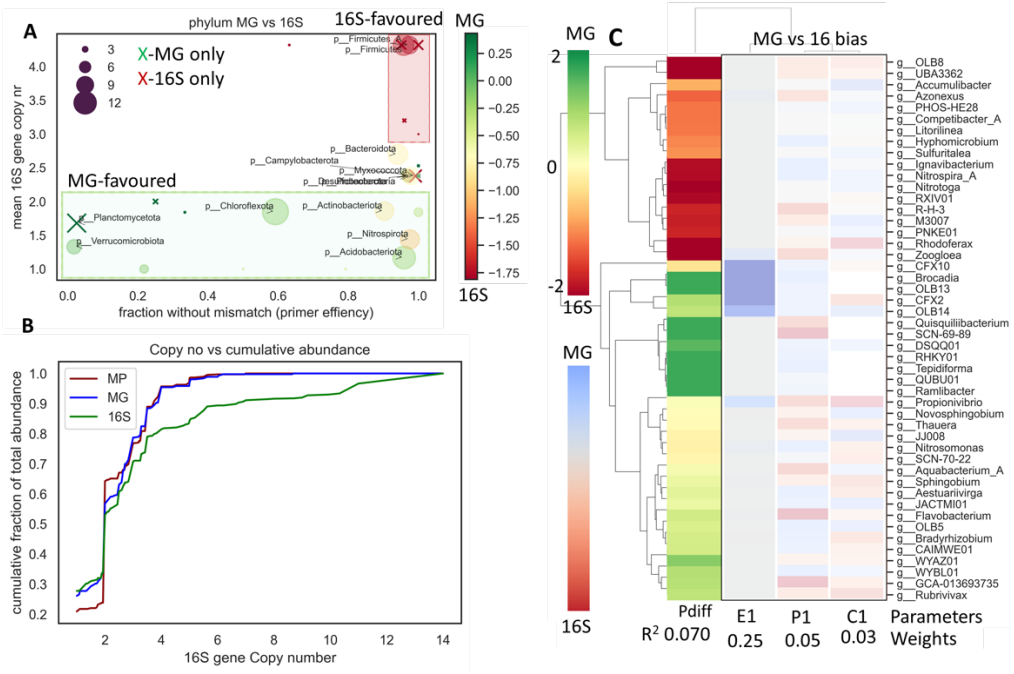


Fig. 2.8 An investigation of quantification bias between 16S analysis and metagenomics. A: Phylum level analysis for combined samples of 16S gene copy number and primer efficiency versus the percentual difference between 16S and metagenomics. B: Cumulative abundance of taxa versus 16S gene copy number for combined plants, and C: Integrated bias clustergram of 16S versus metagenomics, with metrics: E1=primer efficiency, P1=16S gene copy number and C1=Bayesian vs BLAST. Multiple regression of standardized metrics to the percentual difference (Pdiff) was used to assign weights in clustering.

For metaproteomics apart from cell volume (Strous 2017) (which is a physiological bias), limited research has been done to explain observed quantification differences. This metric is included in bias comparison (Fig. 2.9), which uses cell volumes reported from literature (SI Table S5) (P2). As seen in Fig. 2.7, the presence highly expressed proteins could be another source of physiological bias.

As a metric to describe this bias from protein expression patterns, the equation PSMs/K was used, where K equals the number of proteins required to reach 30% of total abundance measured for that genus, and PSMs equals relative abundance of the genus (P1) (SI Table S2.6). Potential sources of experimental bias stem from the ionization and fragmentation of peptides. The ionization efficiency depends on peptide hydrophobicity, here approximated with GRAVY-score (E1) (Bagog 2013, Kite 1982), but also on the amino acid composition and length of a peptide (E2) (Liigand 2019). The charge state and its fragmentation are influenced by the presence of basic amino acid residues (E3). (Biniossek 2012). Metaproteomics and metagenomics largely use the same data processing, so little computational bias is expected. However, since homology and lowest common ancestor analysis are used to annotate open reading frames, there can be difference for strongly expressed proteins, which are often more conserved (C1). Also, an additional lowest common ancestor (LCA) selection is done if detected peptides are detected within multiple organisms (C2).

When comparing metaproteomics to metagenomics (Fig. 2.9) the tested metrics were: the effects of protein expression patterns (P1), literature reported cell volume (P2), peptide hydrophobicity (E1), ionization efficiency (E2), number of basic residues (E3). For the computational effects of conserved proteins on LCA (C1) and LCA of shared peptides (C2) were considered. Physiological bias originating from expression level (P1) is highly impactful. Since metaproteomics is employed in a data-dependent manner, highly abundant peptides will take precedent over less abundant peptides. This causes organisms with high expression of specific proteins to be detected more frequently. These include Accumulibacter and Azonexus, with porin proteins, but also Nitrotoga and Nitrospira with nitrate reductase (nxr) (Lin 2021). Cell volume (P2) is the next most correlating parameter, with the largest cell volumes observed for competibacteriacae. The combination of these two physiological parameters might largely explain quantification bias. Some genera with smaller cells such as Azonexus, Nitrospira and Nitrotoga are detected more in metaproteomics because of their high expression of specific proteins, while Competibacter and RXV01 have a more gradual expression pattern but larger cell size.

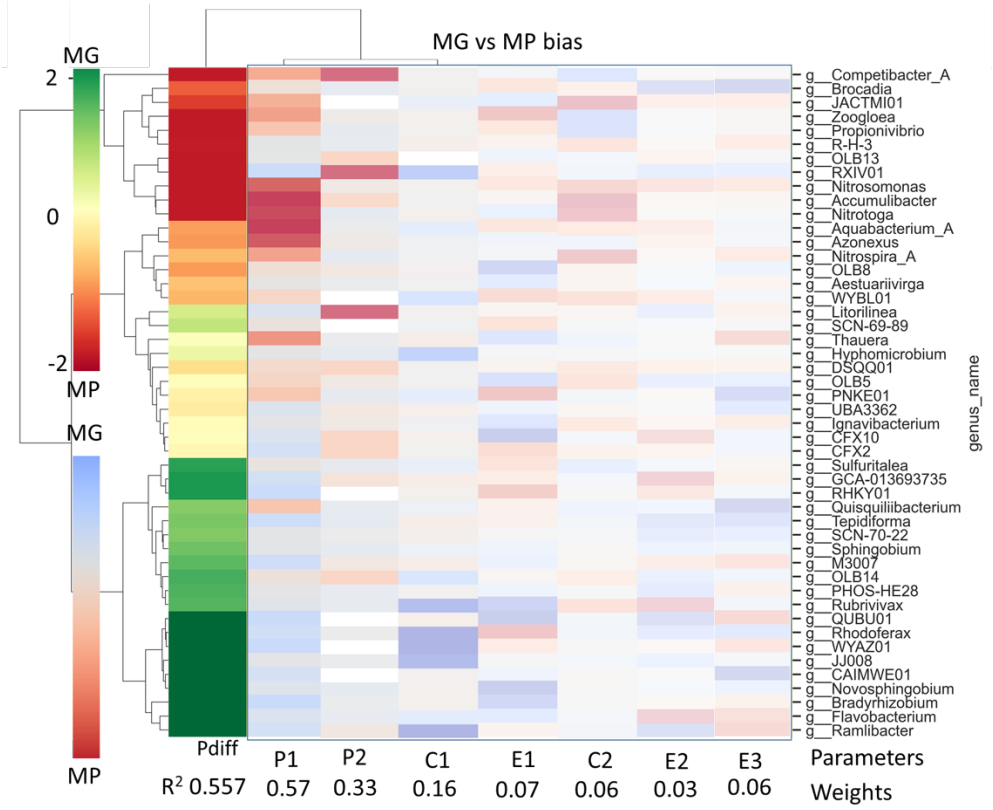


Fig. 2.9 An integrated bias clustergram of metaproteomics versus metagenomics, P1=Shotgun metaproteomics bias, P2=reported cell volumes, C1=homology bias, E1=peptide hydrophobicity, C2=peptide LCA, E2=Ionization efficiency, E3=basic residues. Multiple regression of standardized parameters to the percentual difference (Pdiff) was used to assign weights in clustering.

Bias from the contig LCA method (C1) helps explain underrepresentation of certain genera in metaproteomics (SI Fig S2.16). In case of organisms with no exact match in the database, assigning a consensus contig taxonomy lowers taxonomic resolution. This effect is more pronounced for conserved proteins, which are more likely to be matched to multiple taxa. Strongly affected genera include Rhodoferax and JJ008. Other metrics such as peptide hydrophobicity, ionization efficiency, basic amino acids and peptide LCA (E1, E2,E3,C2) showed little impact.

2.4 Discussion

2.4.1 16S sequencing vs Metagenomics

Overall, the community structure of the aerobic granular sludge evaluated by 16S and genomics analysis was similar, but differences were observed due to primer efficiency and 16S gene copy number (Stoddard 2015, Brown 2015, Albertsen 2015). The effect on the samples was shown in Fig. 2.8. Major underrepresented phyla in 16S include Acidobacteriota with a low copy number, Chloroflexota with reduced primer efficiency, and Planktomycetota and Verrucomirobiota, which were negatively impacted by both. The Planktomycetota phylum contains Anammox bacteria, for which presence was observed in the Garmerwolde plant (Fig 2.5B), while Chloroflexota have been associated with sludge bulking (Speirs 2019), and Acidobacteriota has been linked to PAO and GAO activity in activated sludge (Kristensen 2021).

The 16S gene copy number likely has an influence on the high community evenness of 16S (Fig. 2.3C). On genus level (Fig. 2.8B), high copy number organisms accounted for a significant fraction of total abundance. As granular sludge has a longer solid retention time, it is enriched in slow growing organisms (de Kreuk 2004, Ali 2019) that generally have a lower 16S gene copy number (Roller 2016), which could lead to inflated diversity metrics.

The choice of database showed to be impactful as well. While Midas and Silva had similar rates of genus-level annotation, Silva showed an increasing content of dump taxa at higher taxonomic ranks (Fig. 2.2A). Though Midas had the highest annotation rate, GTDB r202 still outperformed Silva, annotating around 70% and 40% of AVSs at genus and species level respectively, which is a significant improvement over previously reported V4 annotation rates with GTDB r89 (Dueholm 2021). Tetrasphaera, a key organism in biological phosphate removal systems (Nielsen 2019), was not annotated by GTDB after family level. BLASTn alignment of Midas FL-ASVs to GTDB SSUs however revealed that highest identity scores (>95%) were attributed to several other genera within Dermatophilacae (SI Fig S 14). This underlines the limitation 16S derived phylogenetic placement, and the need for longer amplicons than V3-V4 to assess the identity of these taxa.

2.4.2 Metaproteomics vs Metagenomics

As in a previous study (Kleiner 2017), cell volumes derived from literature correlated to the observed quantification differences. Several of the genera underrepresented in metaproteomics have lower reported cell volumes, including Hyphomicrobium, Spingobium and Sulturitalea, while Competibacter, RXIV01, Accumulibacter, have larger cell volumes. However, some genera adhered poorly to this cell volume relation, like Nitrotoga and Propionivibrio, which are small, but still were more abundant in metaproteomics analysis. For Nitrospira, Nitrotoga, and Azonexus, a few strongly expressed proteins were detected. Quantification bias stemming from highly expressed repetitive protein complexes likely affects the abundance of Azonexus and Accumulibacter (porins), and Nitrospira and Nitrotoga (Nxr) (Fig 2.5B), which can form tubular filaments (Chicano 2021). This data-dependent acquisition likely also causes the reduced diversity of metaproteomics (Fig. 2.3C).

While this focus of metaproteomics on the most dominant signals can be seen as a detriment towards accurate quantification, it can also help distinguish the actual key-players from the background of potential converters. The presence of certain enzymes on the genome can be a poor descriptor of functionality. Some ubiquitous genes include (Singleton 2021) *ppk* for PAO, *glg* for GAO and cytochrome c, which is a general protein, but has been implicated with various parts of the nitrogen cycle (Black 2018, Mundinger 2019, Camejo 2019). In particular, *ppk* proteins were mostly detected in known PAO organisms, such as Accumulibacter and Azonexus, but also for Propionivibrio (Albertsen 2016, Nielsen 2019). Similarly, *glg* was expressed in organisms with reported GAO activity: Accumulibacter, Competibacter, and Propionivibrio, while cytochrome C was found heavily expressed in Competibacter (Denitrificans) and Nitrospira, which otherwise show limited expression of denitrifying enzymes (Fig. 2.5B). Metaproteomics can effectively highlight the division of labor within a sample. While PAO, GAO and AOB was consistently performed by the same genera, NOB activity was alternated by Nitrospira and Nitrotoga and denitrification was split over several genera. Denitrification genes are commonly found in many organisms (Singleton 2021), however in metaproteomics *Nor* is barely detected, *Nirs/NirK* are found in abundant PAO/GAO organisms and *Nos* seems the most dispersed. Additionally, based on proteome analysis, truncated nitrogen metabolism was observed within AGS, similarly to activated sludge (Singleton 2021).

2.4.3 Sources of bias

While the sources of quantification bias for 16S are well described, they still showed poor correlation with observed bias (Fig. 2.8). On the other hand, the differences between metagenomics and metaproteomics are less well covered in literature. In this study we therefore categorized potential sources of quantification bias into sources of physiological, experimental and computational bias. Still, many aspects of bias remain unexplored.

It is a major challenge to assess physiological traits within mixed cultures. Though cell volume can be monitored with microscopy, it is not directly representative of total protein biomass in a cell, since a system grown under feast-famine regime will commonly produce storage polymers such as polyphosphate glycogen and polyhydroxyalkanoates that affect the cell volume. Determining the average protein content per cell would require cell sorting combined with cell staining. There is also a limited understanding of the biases of shotgun metaproteomics, which would require modelling approaches with in-silico communities (Herbst 2016). Additionally, next to genome size, DNA-content per cell can differ based on the organism's growth rate and physiology, as exponential growth has been associated with duplicate chromosomes (Haugan 2018), and slow growing organisms have more time to accumulate plasmids (Suzuki 2019).

Experimental bias on the other hand can be investigated more easily and should become a focal point. Especially the role of extraction bias was not explored within this study. To extract DNA or protein, physical homogenization is typically employed when working with activated sludge. This is known to impact the quantification, either by shearing DNA or by insufficient cell lysis (Pronk 2017, Guo 2013). On the other hand, proteome extraction can apply more drastic physical lysis, but more so relies on detergents to properly solubilize membrane-bound and structural proteins (Boleij 2019, Danielsen 2017), both of which could require optimization for aerobic granular sludge. Besides extraction, the bias generated through shotgun metaproteomics might be mitigated using data independent acquisition, which also fragments less abundant peptides (Elo 2020, Gillet 2012).

2.4.4 Use cases and synergy

Though each of the three techniques showed different levels of information and varying quantification of the community, none is without merit. Since each comes with its own advantages, use will depend on the research question. The most common applications of DNA-based techniques in wastewater are the detection of pathogens, viruses, and antimicrobial resistance genes (Garner 2021), for which metaproteomics will be less effective due to its lower sensitivity and coverage. However, large-scale studies have also used DNA-based techniques to infer functional roles from metabolic potential and have established a core-microbiome based on prevalence (Wu 2019, Saunders 2015, Singleton 2021). This is where metaproteomics should be the preferred method, as it is more effective at identifying the main metabolic organisms within a microbial community and their roles within the system. The use of orthogonal techniques such as FISH RAMAN can confirm functions predicted from genomes (McIlroy 2018, Huang 2007), but does so less efficiently. Considering the quantification differences observed for several key metabolic organisms, the inclusion of orthogonal metaproteomic analysis on selected samples should become a standard within large-scale metagenomic studies moving forward, as metabolic insights can lead to different conclusions on what is “core”.

Metaproteomics is still limited by its lack of throughput and parallelization compared to metagenomic sequencing and will require further improvements to compete in the field of ecology. It also still typically relies on DNA-based techniques for database construction (Tanca 2015). As techniques evolve further, they should do so collaboratively. While the current focus lies on improving techniques independently (Scyrba 2017, van den Bossche 2021), there also lies a risk of divergence. Current attention towards ultra-small endosymbionts (Brown 2015, He 2021) will result in more extreme differences in cell-counting versus biomass-based quantification, while the increasing read length in metagenomics will exacerbate extraction biases, as it more susceptible to shearing (Maghini 2019). Lastly, the lack of 16S genes in most published MAGs results in increasing database discrepancy between genomic and 16S databases (Soo 2014, Rahman 2015). A collaborative evolution between analysis techniques should be desired for both the fields of ecology and environmental biotechnology. As a start the use of standardized databases and taxonomies of GTDB will be key, to achieve a convergent evolution that maximizes synergy.

2.5 References

- Abdul Rahman, Nurdyana, et al. "A phylogenomic analysis of the bacterial phylum Fibrobacteres." Frontiers in microbiology 6 (2016): 1469.
- Adav, Sunil S., Duu-Jong Lee, and Joo-Hwa Tay. "Extracellular polymeric substances and structural stability of aerobic granule." Water Research 42.6-7 (2008): 1644-1650.
- Adav, Sunil S., Duu-Jong Lee, and Juin-Yih Lai. "Proteolytic activity in stored aerobic granular sludge and structural integrity." Bioresource technology 100.1 (2009): 68-73.
- Albertsen, Mads, et al. "A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal." The ISME journal 6.6 (2012): 1094-1106.
- Albertsen, Mads, et al. "Back to basics–the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities." PloS one 10.7 (2015): e0132783.
- Ali, Muhammad, et al. "Importance of species sorting and immigration on the bacterial assembly of different-sized aggregates in a full-scale aerobic granular sludge plant." Environmental science & technology 53.14 (2019): 8291-8301.
- Angenent, Largus T., et al. "Production of bioenergy and biochemicals from industrial and agricultural wastewater." TRENDS in Biotechnology 22.9 (2004): 477-485.
- Avila-Calderón, Eric Daniel, et al. "Outer Membrane Vesicles of Gram-Negative Bacteria: An Outlook on Biogenesis." Frontiers in Microbiology 12 (2021): 345.
- Azizan, Asma, et al. "Using proteomics for an insight into the performance of activated sludge in a lab-scale WWTP." International Biodeterioration & Biodegradation 149 (2020): 104934.
- Barr, Jeremy J., et al. "Metagenomic and metaproteomic analyses of Accumulibacter phosphatis-enriched floccular and granular biofilm." Environmental microbiology 18.1 (2016): 273-287.
- Biniossek, Martin L., and Oliver Schilling. "Enhanced identification of peptides lacking basic residues by LC-ESI-MS/MS analysis of singly charged peptides." Proteomics 12.9 (2012): 1303-1309.
- Black, Ellen M., and Craig L. Just. "The genomic potentials of NOB and comammox Nitrospira in river sediment are impacted by native freshwater mussels." Frontiers in microbiology 9 (2018): 2061.
- Boleij, Marissa, et al. "Solubilization and characterization of extracellular proteins from anammox granular sludge." Water research 164 (2019): 114952.
- Bolyen, Evan, et al. "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2." Nature biotechnology 37.8 (2019): 852-857.
- Bonk, Fabian, et al. "PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls." Journal of microbiological methods 153 (2018): 139-147.
- Boutet, Emmanuel, et al. "Uniprotkb/swiss-prot." Plant bioinformatics. Humana Press, 2007. 89-112.
- Brooks, J. Paul, et al. "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies." BMC microbiology 15.1 (2015): 1-14.
- Brown, Christopher T., et al. "Unusual biology across a group comprising more than 15% of domain Bacteria." Nature 523.7559 (2015): 208-211.
- Browne, Patrick Denis, et al. "GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms." GigaScience 9.2 (2020): giaa008.
- Callahan, Benjamin J., Paul J. McMurdie, and Susan P. Holmes. "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis." The ISME journal 11.12 (2017): 2639-2643.
- Camejo, Pamela Y., et al. "Integrated omic analyses provide evidence that a “Candidatus Accumulibacter phosphatis” strain performs denitrification under microaerobic
- Chaumeil, Pierre-Alain, et al. "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." (2020): 1925-1927.
- Cheng, M., et al. "Evidence of compositional differences between the extracellular and intracellular DNA of a granular sludge biofilm." Letters in applied microbiology 53.1 (2011): 1-7.
- Chicano, Tadeo Moreno, et al. "Structural and functional characterization of the intracellular filament-forming nitrite oxidoreductase multiprotein complex." Nature microbiology 6.9 (2021): 1129-1139.
- Cho, Ilseung., Blaser, Martin J. "The human microbiome: at the interface of health and disease." Nature reviews genetics 13 (2012): 260–270

- Danielsen, Heidi N., et al. "Direct identification of functional amyloid proteins by label-free quantitative mass spectrometry." Biomolecules 7.3 (2017): 58.
- Dueholm, Morten Simonsen, et al. "Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax)." MBio 11.5 (2020): e01557-20.
- Dueholm, Morten Simonsen, et al. "MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants." bioRxiv (2021).
- Edgar, Robert C. "UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success." BioRxiv (2017): 124149.
- Edgar, Robert C. "UNCROSS: filtering of high-frequency cross-talk in 16S amplicon reads." Biorxiv (2016): 088666.
- Falkowski, Paul G., Tom Fenchel, and Edward F. Delong. "The microbial engines that drive Earth's biogeochemical cycles." science 320.5879 (2008): 1034-1039.
- Garner, Emily, et al. "Next generation sequencing approaches to evaluate water and wastewater quality." Water Research (2021): 116907.
- Gillet, Ludovic C., et al. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis." Molecular & Cellular Proteomics 11.6 (2012).
- Guo, Feng, and Tong Zhang. "Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing." Applied microbiology and biotechnology 97.10 (2013): 4607-4616.
- Haft, Daniel H., Jeremy D. Selengut, and Owen White. "The TIGRFAMs database of protein families." Nucleic acids research 31.1 (2003): 371-373.
- Hagen, Live H., et al. "Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester." Applied and environmental microbiology 83.2 (2017): e01955-16.
- Hansen, Susan Hove, et al. "Metaproteomics: evaluation of protein extraction from activated sludge." Proteomics 14.21-22 (2014): 2535-2539.
- Haugan, Maria Schei, et al. "Chromosome replication as a measure of bacterial growth rate during Escherichia coli infection in the mouse peritonitis model." Scientific reports 8.1 (2018): 1-13.
- He, Christine, et al. "Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems." Nature microbiology 6.3 (2021): 354-365.
- Herbst, Florian-Alexander, et al. "Enhancing metaproteomics—the value of models and defined environmental microbial systems." Proteomics 16.5 (2016): 783-798.
- Huang, Wei E., et al. "Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function." Environmental microbiology 9.8 (2007): 1878-1889.
- Hugenholtz, Philip, Adam Skarshewski, and Donovan H. Parks. "Genome-based microbial taxonomy coming of age." Cold Spring Harbor perspectives in biology 8.6 (2016): a018085.
- Karthik, L., et al. "Protease inhibitors from marine actinobacteria as a potential source for antimalarial compound." PloS one 9.3 (2014): e90972.
- Kleikamp, Hugo BC, et al. "Database-independent de novo metaproteomics of complex microbial communities." Cell Systems 12.5 (2021): 375-383.
- Kleiner, Manuel, et al. "Assessing species biomass contributions in microbial communities via metaproteomics." Nature communications 8.1 (2017): 1-14.
- de Kreuk, MK van, and MCM van Van Loosdrecht. "Selection of slow growing organisms as a means for improving aerobic granular sludge stability." Water Science and Technology 49.11-12 (2004): 9-17.
- Kolakovic, Srdana, et al. "Accumulibacter diversity at the sub-clade level impacts enhanced biological phosphorus removal performance." Water Research 199 (2021): 117210.
- Konstantinidis, Konstantinos T., and James M. Tiedje. "Towards a genome-based taxonomy for prokaryotes." Journal of bacteriology 187.18 (2005): 6258-6264.
- Koskinen, Kaisa, et al. "First insights into the diverse human archaeome: specific detection of archaea in the gastrointestinal tract, lung, and nose and on skin." MBio 8.6 (2017): e00824-17.

- Kyte, Jack, and Russell F. Doolittle. "A simple method for displaying the hydropathic character of a protein." Journal of molecular biology 157.1 (1982): 105-132.
- Leventhal, Gabriel E., et al. "Strain-level diversity drives alternative community types in millimetre-scale granular biofilms." Nature microbiology 3.11 (2018): 1295-1303.
- Liang, Zixuan, et al. "Formation, extracellular polymeric substances, and structural stability of aerobic granules enhanced by granular activated carbon." Environmental Science and Pollution Research 26.6 (2019): 6123-6132.
- Liigand, Piia, Karl Kaupmees, and Anneli Kruve. "Influence of the amino acid composition on the ionization efficiencies of small peptides." Journal of Mass Spectrometry 54.6 (2019): 481-487.
- Lin Y, Wang L, Xu K, Li K, Ren H. Revealing taxon-specific heavy metal-resistance mechanisms in denitrifying phosphorus removal sludge using genome-centric metaproteomics. Microbiome. 2021;9:1-17.
- Lin, Yuan, et al. "Revealing taxon-specific heavy metal-resistance mechanisms in denitrifying phosphorus removal sludge using genome-centric metaproteomics." Microbiome 9.1 (2021): 1-17.
- Louca, Stilianos, Michael Doebeli, and Laura Wegener Parfrey. "Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem." Microbiome 6.1 (2018): 1-12.
- McIlroy, Simon J., et al. "Genomic and in situ analyses reveal the Micropruina spp. as abundant fermentative glycogen accumulating organisms in enhanced biological phosphorus removal systems." Frontiers in microbiology 9 (2018): 1004.
- McIlroy, Simon J., et al. "'Candidatus Competibacter'-lineage genomes retrieved from metagenomes reveal functional metabolic diversity." The ISME journal 8.3 (2014): 613-624.
- McIlroy, Simon Jon, et al. "MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups." Database 2017 (2017).
- McIlroy, Simon Jon, et al. "MiDAS: the field guide to the microbes of activated sludge." Database 2015 (2015).
- McLaren, Michael R., Amy D. Willis, and Benjamin J. Callahan. "Consistent and correctable bias in metagenomic sequencing experiments." Elife 8 (2019): e46923.
- von Meijenfeldt, FA Bastiaan, et al. "Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT." Genome biology 20.1 (2019): 1-14.
- Metzl-Raz, Eyal, et al. "Principles of cellular resource allocation revealed by condition-dependent proteome profiling." Elife 6 (2017): e28034.
- Mirdita, Milot, et al. "Fast and sensitive taxonomic assignment to metagenomic contigs." Bioinformatics 37.18 (2021): 3029-3031.
- Mistry, Jaina, et al. "Pfam: The protein families database in 2021." Nucleic Acids Research 49.D1 (2021): D412-D419.
- Morrissey, Ember M., et al. "Phylogenetic organization of bacterial activity." The ISME journal 10.9 (2016): 2336-2340.
- Mundinger, Aniela B., et al. "Cultivation and transcriptional analysis of a canonical Nitrospira under stable growth conditions." Frontiers in microbiology 10 (2019): 1325.
- Narayanasamy, Shaman, et al. "Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities." Microbial biotechnology 8.3 (2015): 363-368.
- Nayfach, Stephen, et al. "A genomic catalog of Earth's microbiomes." Nature biotechnology 39.4 (2021): 499-509.
- Nelson, William C., Benjamin J. Tully, and Jennifer M. Mobberley. "Biases in genome reconstruction from metagenomic data." PeerJ 8 (2020): e10119.
- Nielsen, Per Halkjær, et al. "Re-evaluating the microbiology of the enhanced biological phosphorus removal process." Current opinion in biotechnology 57 (2019): 111-118.
- Nierychlo, Marta, et al. "MiDAS 3: An ecosystem-specific reference database, taxonomy and knowledge platform for activated sludge and anaerobic digesters reveals species-level microbiome composition of activated sludge." Water Research 182 (2020): 115955.
- Nikaido, Hiroshi. "Multidrug efflux pumps of gram-negative bacteria." Journal of bacteriology 178.20 (1996): 5853-5859.

- Oren, Aharon, George M. Garrity, and Aidan C. Parte. "Why are so many effectively published names of prokaryotic taxa never validated?." International journal of systematic and evolutionary microbiology 68.7 (2018): 2125-2129.
- Panchavinin, Smarch, et al. "Candidates of quorum sensing bacteria in activated sludge associated with N-acyl homoserine lactones." Chemosphere 236 (2019): 124292.
- Parks, Donovan H., et al. "A complete domain-to-species taxonomy for Bacteria and Archaea." Nature biotechnology 38.9 (2020): 1079-1086.
- Parks, Donovan H., et al. "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy." Nucleic Acids Res. (2021): gkab776.
- Pietila, Sami, Tomi Suomi, and Laura L. Elo. "Metaproteomics boosted up by untargeted data-independent acquisition data analysis framework." bioRxiv (2020).
- Pronk, Mario., et al. "Effect and behaviour of different substrates in relation to the formation of aerobic granular sludge." Applied microbiology and biotechnology 99.12 (2015a): 5257-5268.
- Pronk, Mario, et al. "The acid soluble extracellular polymeric substance of aerobic granular sludge dominated by Defluviicoccus sp." Water Research 122 (2017): 148-158.
- Pronk, Mario., et al. "Full scale performance of the aerobic granular sludge process for sewage treatment." Water Research 84 (2015b): 207-217.
- Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic acids research 33.suppl_1 (2005): D501-D504.
- Püttker, Sebastian, et al. "Metaproteomics of activated sludge from a wastewater treatment plant–A pilot study." Proteomics 15.20 (2015): 3596-3601.
- Quast, Christian, et al. "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." Nucleic acids research 41.D1 (2012): D590-D596.
- Rabaey, Korneel, and Willy Verstraete. "Microbial fuel cells: novel biotechnology for energy generation." TRENDS in Biotechnology 23.6 (2005): 291-298.
- Ramos, Carlos, María Eugenia Suárez-Ojeda, and Julián Carrera. "Long-term impact of salinity on the performance and microbial population of an aerobic granular reactor treating a high-strength aromatic wastewater." Bioresource technology 198 (2015): 844-851.
- Roller, Benjamin RK, Steven F. Stoddard, and Thomas M. Schmidt. "Exploiting rRNA operon copy number to investigate bacterial reproductive strategies." Nature microbiology 1.11 (2016): 1-7.
- Rubio-Rincón, F. J., et al. "Cooperation between Candidatus Competibacter and Candidatus Accumulibacter clade I, in denitrification and phosphate removal processes." *Water Research* 120 (2017): 156-164.
- Rubio-Rincón, F. J., et al. "“Candidatus Accumulibacter delftensis”: A clade IC novel polyphosphate-accumulating organism without denitrifying activity on nitrate." *Water research* 161 (2019): 136-151
- Saunders, Aaron M., et al. "The activated sludge ecosystem contains a core community of abundant organisms." The ISME journal 10.1 (2016): 11-20.
- Schoch, Conrad L., et al. "NCBI Taxonomy: a comprehensive update on curation, resources and tools." Database 2020 (2020).
- Schwechheimer, Carmen, and Meta J. Kuehn. "Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions." Nature reviews microbiology 13.10 (2015): 605-619.
- Sedlar, Karel, Kristyna Kupkova, and Ivo Provaznik. "Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics." Computational and Structural Biotechnology Journal 15 (2017): 48-55.
- Sieber, Christian MK, et al. "Unusual metabolism and hypervariation in the genome of a gracilibacterium (Bd1-5) from an oil-degrading community." MBio 10.6 (2019): e02128-19.
- Soo, Rochelle M., et al. "An expanded genomic representation of the phylum Cyanobacteria." Genome biology and evolution 6.5 (2014): 1031-1045.
- de Sousa Rollemberg, Silvio Luiz, et al. "Comparison of the dynamics, biokinetics and microbial diversity between activated sludge flocs and aerobic granular sludge." Bioresource technology 294 (2019): 122106.

- Starke, Robert, Victor Satler Pylro, and Daniel Kumazawa Morais. "16S rRNA gene copy number normalization does not provide more reliable conclusions in metataxonomic surveys." Microbial ecology 81.2 (2021): 535-539.
- Stoddard, Steven F., et al. "rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development." Nucleic acids research 43.D1 (2015): D593-D598.
- Stokholm-Bjerregaard, Mikkel, et al. "A critical assessment of the microorganisms proposed to be important to enhanced biological phosphorus removal in full-scale wastewater treatment systems." Frontiers in microbiology 8 (2017): 718.
- Suzuki, Yoshihiko, et al. "Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut." Microbiome 7.1 (2019): 1-16.
- Świątczak, Piotr, and Agnieszka Cydzik-Kwiatkowska. "Performance and microbial characteristics of biomass in a full-scale aerobic granular sludge wastewater treatment plant." Environmental Science and Pollution Research 25.2 (2018): 1655-1669.
- Szabó, Enikö, et al. "Comparison of the bacterial community composition in the granular and the suspended phase of sequencing batch reactors." Amb Express 7.1 (2017): 1-12.
- Tanca, Alessandro, et al. "The impact of sequence database choice on metaproteomic results in gut microbiota studies." Microbiome 4.1 (2016): 1-13.
- Tatusov, Roman L., et al. "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic acids research 28.1 (2000): 33-36.
- Timmins-Schiffman, Emma, et al. "Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns." The ISME journal 11.2 (2017): 309-314.
- Weissbrodt, David G., Noam Shani, and Christof Holliger. "Linking bacterial population dynamics and nutrient removal in the granular sludge biofilm ecosystem engineered for wastewater treatment." FEMS microbiology ecology 88.3 (2014): 579-595.
- Weissbrodt, David G., et al. "Assessment of bacterial and structural dynamics in aerobic granular biofilms." Frontiers in microbiology 4 (2013): 175.
- Welles, L., et al. "Accumulibacter clades Type I and II performing kinetically different glycogen-accumulating organisms metabolisms for anaerobic substrate uptake." Water research 83 (2015): 354-366.
- Wilmes, Paul, Anna Heintz-Buschart, and Philip L. Bond. "A decade of metaproteomics: where we stand and what the future holds." Proteomics 15.20 (2015): 3409-3417.
- Wilmes, Paul, Margaret Wexler, and Philip L. Bond. "Metaproteomics provides functional insight into activated sludge wastewater treatment." PLoS One 3.3 (2008): e1778.
- Winkler, MK H., et al. "Microbial diversity differences within aerobic granular sludge and activated sludge flocs." Applied microbiology and biotechnology 97.16 (2013): 7447-7458.
- Wu, Linwei, et al. "Global diversity and biogeography of bacterial communities in wastewater treatment plants." Nature microbiology 4.7 (2019): 1183-1195.
- Wu, Sitao, et al. "WebMGA: a customizable web server for fast metagenomic sequence analysis." BMC genomics 12.1 (2011): 1-9.
- Xia, Juntao, et al. "Microbial community structure and function in aerobic granular sludge." Applied microbiology and biotechnology 102.9 (2018): 3967-3979.
- Yarza, Pablo, et al. "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences." Nature Reviews Microbiology 12.9 (2014): 635-645.
- Zhang, Tong, Ming-Fei Shao, and Lin Ye. "454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants." The ISME journal 6.6 (2012): 1137-1147.
- Zhou, Jun, and Qianyu Sun. "Performance and microbial characterization of aerobic granular sludge in a sequencing batch reactor performing simultaneous nitrification, denitrification and phosphorus removal with varying C/N ratios." Bioprocess and biosystems engineering 43.4 (2020): 663-672.
- Zukancic, Alem, et al. "Staphylococcal Protein A (spa) locus is a hot spot for recombination and horizontal gene transfer in Staphylococcus pseudintermedius." Msphere 5.5 (2020): e00666-20.

2.6 Supplemental information

2. 6.1 Material & Methods

Sampling of aerobic granular sludge.

Aerobic granular sludge (AGS) was collected from 3 different full scale AGS wastewater treatment plants, from the Netherlands: Simpelveld (SP), Dinxperlo (DX) and Garmerwolde (GW). Each plant performed stable operation with simultaneous denitrification and phosphorous removal. AGS granules were sieved to select a size fraction of a diameter of 2.0 mm, which furthermore were stored at -80˚until further processed.

DNA extraction and sequencing.

Extraction of DNA for both 16S rRNA gene sequencing and shotgun metagenomics was done using a DNeasy UltraClean Microbial Kit (Qiagen, Germany), and was quantified with Qubit. 16S rRNA gene amplification was done by Novogene (Novogene Co., Ltd., China), by amplifying V3-V4 regions with 341F, 806R primers. Sequencing of 16S rRNA genes and shotgun metagenomics was done with paired-end reads on the Illumina Novaseq platform.

Protein extraction and proteolytic digestion.

The collected 2 mm granules were freeze-dried and grinded with a mortar and pestle. To approximately 5 mg starting material, 200 mg of glass beads (150 – 212 µm) and 350 µL of both TEAB and B-PER were added. Beads beating was performed for 20 seconds, three times with a 30 second break in-between. Samples were shortly centrifuged, and 3 freeze/thaw were performed by placing the sample into a -80°C freezer, and for thawing into a 95°C hot water bath. The sample was centrifuged, and the supernatant was collected. Further, protein precipitation was performed by adding TCA at a ratio of 1 to 4 to the supernatant. The samples are then incubated at 4°C for 10 minutes and centrifuged at 14,000 rpm for five minutes. The pellet was washed with 200 µL ice cold acetone. To every tube, 250 µL of 6 M urea was added to reconstitute the protein pellet for further digestion. Furthermore, the protein extract was reduced using 10 mM Dithiothreitol (DTT) for 60 minutes at 37°C. Next the samples were alkylated using 20 mM Iodoacetamide (IAA) in the dark and at room temperature for 30 minutes. Thereafter, the samples were diluted with 200 mM Ammonium bicarbonate (AmBiC), to <1 M Urea. Finally, sequencing grade trypsin was added (Promega) at an approx. enzyme to protein ratio of 1:50 and incubated at 37°C for overnight digestion. The obtained peptides were further solid-phase extraction purified by using Oasis HLB well plates (Waters), according to the manufactures protocol. Purified peptide

fractions were then dried in a Speedvac concentrator, and samples were finally reconstituted in aqueous 0.1% TFA solution. Finally, peptides were fractionated using a Pierce high pH reversed-phase peptide fractionation kit (Thermo Scientific), according to the instructions of the manufacturer, and dried in a Speedvac concentrator. Peptide samples were finally dissolved in H2O, containing 3% acetonitrile and 0.1% formic acid, protein digest was determined using a NanoDrop micro volume spectrophotometer.

Shotgun metaproteomic analysis.

The speed-vac tried peptide fraction was resuspended in H2O containing 3% acetonitrile and 0.1% formic acid. An aliquot corresponding to approx. 300 ng protein digest was analyzed in duplicates using an one dimensional shotgun proteomics approach. Briefly, the samples were analyzed using a nano-liquid-chromatography system consisting of an EASY nano-LC 1200, equipped with an Acclaim PepMap RSLC RP C18 separation column (50 µm x 150 mm, 2 µm and 100 Å), and an QE plus Orbitrap mass spectrometer (Thermo Scientific, Germany). The flow rate was maintained at 300 nL/min over a linear gradient to 30% solvent B over 60 or 90 minutes, and finally to 75% B over additional 30 minutes. Solvent A was H2O containing 0.1% formic acid, and solvent B consisted of 80% acetonitrile in H2O and 0.1% formic acid. The Orbitrap was operated in data dependent acquisition mode where the top 10 signals were isolated at a window of 2.0 m/z and fragmented using a NCE of 28. The AGC target was set to 1e5, at a max IT of 54 ms and 17.5 K resolution.

Processing of 16S rRNA raw sequencing data.

Standard read preparation was done by Novogene including de-multiplexing, trimming and assembly. Cleaned reads were used to pick amplicon sequence variants (ASVs) with Usearchv11 command -unoise3 (Edgar 2016). To improve ASV-picking, the dataset was padded with additional sample sets containing different size fractions from each plant: flocs, >0,2, >0.7, >1.0 mm. (data not shown). Taxonomic annotation was performed using QIIME2 with trained V3-V4 classifiers. As a comparison 16S rRNA sequences were annotated with GTDB representative small subunit ribosomal RNA (SSU rRNA) sequences, Midas 3.7 flASVs, and a Silva NR99 v138 pre-trained V3-V4 classifier (Robeson 2021, (Bokulich 2018). To compare the effects of database homogenization, GTDB r202 complete 16S (ssu all) and representative (ssu reps), were analyzed with all sequences, and full-length sequences of >1200 base pairs.

Processing of metagenomics raw sequencing data.
Reads were assembled for all samples using metaSPAdes v3.14.0 (Nurk et al. 2017) at the default settings. Prodigal v2.6.3 (Hyatt et al. 2010) was used as a gene caller to identify open reading frames (ORFs). DIAMOND v2.11 (Buchfink 2015) was used align ORFs with parameters -fast -top10 -e 0.001 to protein databases of GTDB r202, from Uniprot release 2021 03: UniProtKB, Swiss-Prot UniRef100, UniRef90, UniRef50, and from NCBI RefSeq protein and RefSeq protein NR release 205. Additionally, homogenized databases were annotated by constructing databases from organisms represented in GTDB ssu reps to create GTDB HOM, and organisms with full length 16S sequences to create GTDB FLHOM. Contig-level taxonomic annotation was done by first determining the taxonomy of each ORF by lowest ancestor analysis (LCA) of the aligned proteins (protein LCA), followed by constructing a consensus lineage for each contig from the annotated ORFs (contig LCA). (Supplemental 2.6.2)

Processing of metaproteomics raw data.
The mass spectrometric raw data were analysed using PEAKS Studio X by either database search using metagenomic databases constructed from predicted ORFs, or by de novo sequencing (Supplemental). Taxonomic annotation of database matched peptides was done by performing an LCA on associated contigs. Metabolic annotation with KEGG orthologies was performed on identified proteins using BlastKOALA (Kanehisa 2016), while WEBMGA (Wu 2011) was used to annotate COG, PFAM and TIGRFAM terms, while DIAMOND v2.11 (Buchfink 2015) was used to annotate ORFs with UniprotKB genes.

Comparative analysis and visualization of large-scale omics data.
Sankey diagrams were constructed on an annotation level. By using GTDB metadata, taxa with >75% overlap between databases were homogenized to GTDB nomenclature. Quantification of taxa was performed on total ASV count in 16S, total depth of associated contigs in metagenomics, and the total number of peptide spectrum matches in metaproteomics. To homogenize data treatment across the three techniques, an abundance cutoff of 0.1% was applied. This taxa-set was used to compute alpha-diversity metrics and shared biomass content. For further analysis abundance was renormalized to a subset of shared taxa that either contained nutrient removal genes, or had >3% abundance in one technique. Between techniques, the absolute abundance difference: (x-y) and percentual abundance difference: (x-y)/(((x+y)⁄2) was computed for each genus. Functional analysis was done by integrating KEGG,COG,PFAM,TIGRFAM and UniprotKB genes. Two manually curated subclassifications were added to the COG system

"nitrogen metabolism" based on KEGG pathways and "porin", which includes beta-barrel proteins. As a measure of overexpression a percentual difference between the fold change respective to the median of reads and PSMs for each class as computed. Sources of bias were grouped into experimental (E), physiological (P) and computational (C). Comparison of 16S and metagenomics was done on primer efficiency (E1), 16S gene copy number (P1), and BLAST vs Bayesian annotation (C1). Primer efficiency was approximated by the percentage of genomes belonging to that taxa that match a primer pair without mismatches, 16S gene copy numbers were obtained from rrnDB (Stoddard 2015) and BLAST+ v2.12.0 was used to align ASVs to full length representative GTDB ssu sequences Comparison of metaproteomics and metagenomics was done on experimental parameters: GRAVY-score (E1) (Kite 1982) ionization efficiency (E2) (Liigand 2019) and number of basic amino acids (E3) of each peptide, physiological parameters: expression profile (P1) (Table S2.6, Fig. S2.18-S2.20), literature reported cell volume (P2) (Table S5) and computational parameters contig LCA (C1) (Fig. S2.16) and peptide LCA (C2) (Fig. S2.17). To assess the influence of each parameter, values were standardized and fitted to observed percentual differences using multiple regression. Standardized values were the fitted coefficients before clustering on Euclidian distance.

2.6.2 optimization of contig annotation

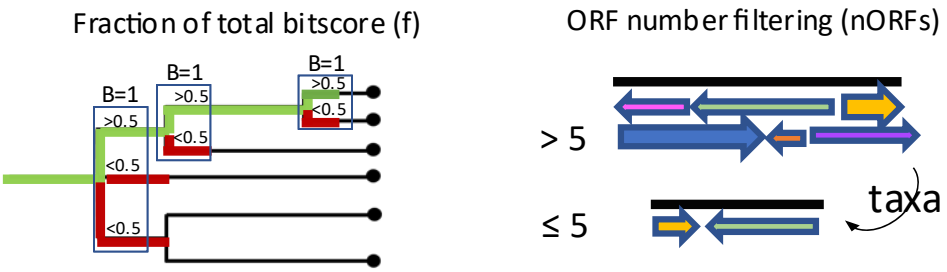


Fig. S2.1 Diagram highlighting the modifications applied to the CAT algorithm (von Meijenfeldt 2019) The tolerated fraction of total bitscore (f) is recomputed at each rank, and after annotation, contigs with less than a certain number of matched ORFs are filtered based on larger contigs, allowing only taxonomies observed in contigs with more than the specified number of ORFs (nORFs).

Since an optimized binning approach did not cover several of the expected organisms (Table S2.7), a general contig annotation was explored. Contig-level taxonomic annotation was done by first determining the taxonomy of each ORF by lowest ancestor analysis (LCA) of the aligned proteins with a protein-level LCA, followed by constructing a consensus lineage for each contig from the annotated

ORFs as a contig-level LCA. An existing tool that uses this method is CAT (von Meijenfeldt 2019). Since CAT is designed to annotate novel organisms, it's stringency results in a reduced number of genus-level matches. We therefore modified the CAT algorithm to construct a stepwise lineage by recalculating the total bitscore for each rank from the taxa above the cutoff. Another disadvantage of CAT is that it weighs short contigs equally strong as long contigs, which could more easily result in false positives, if only a single ORF can be matched to a database within a contig. Therefore, a filtering step was included to filter taxa that were only found in shorter contigs. This filtering step helped to maintain a similar diversity as 16S sequences. For this Jaccard similarity as used as a metric, as this compares which taxa are detected, but not their abundances. Both protein LCA and contig LCA approaches were optimized to maximize genus-level Jaccard similarity to 16S annotations, and to maximize the fraction of genus-level annotations, which is essential for functional annotation. As baseline comparison, the CAT algorithm with parameters r=10 and f=0.5 was used, in which r denotes the range of top bitscore % used in protein LCA, and f the minimum fraction classification support (von Meijenfeldt 2019). For optimizing protein LCA, 5 approaches were compared 1. A standard approach (LCA, r=10), 2. LCA on top bitscore (BLCA, r=0), 3. Top hit only (Top hit), 4. FtLCA, fraction of taxa (FtLCA, r=10, f=0.5), 5. FbLCA, fraction of bitscore (FbLC, r=10, f=0.5). FtLCA and FbLCA are newly proposed, modified versions of the CAT algorithm. In both, a lineage is built up sequentially from the most frequent taxa, provided it appears above a threshold (Fig 1). This threshold is either based on the fraction of taxa belonging to the taxon selected at the previous rank (FtLCA), or the fraction of the bitscore of these taxa (FbLCA). For optimizing contig LCA, the FbLCA algorithm was selected with an added filtering step, in which only taxa were allowed from contigs with more than x ORFs (nORF). A parameter sweep of f (0.1-1) and nORF (2-10,20) was used to optimize the algorithm on the GTDB FLHOM annotated dataset (Fig 2.). As objective score a multiplication was used of the genus coverage and the Jaccard similarity to 16S taxa. As a comparison the CAT algorithm was combined with the different protein LCA methods. The baseline CAT annotation would be CAT with standard LCA. The final selected parameters used for annotation within the paper are BLCA for protein LCA and FbLCA (nORFs=5, f=0.5) for contig LCA. Using the optimized parameters, genus level annotation improved from 0.27 to 0.82 while Jaccard similarity to 16S from 0.19 to 0.25. Performance of BLA and standard LCA are shown in (Fig. S2.2).

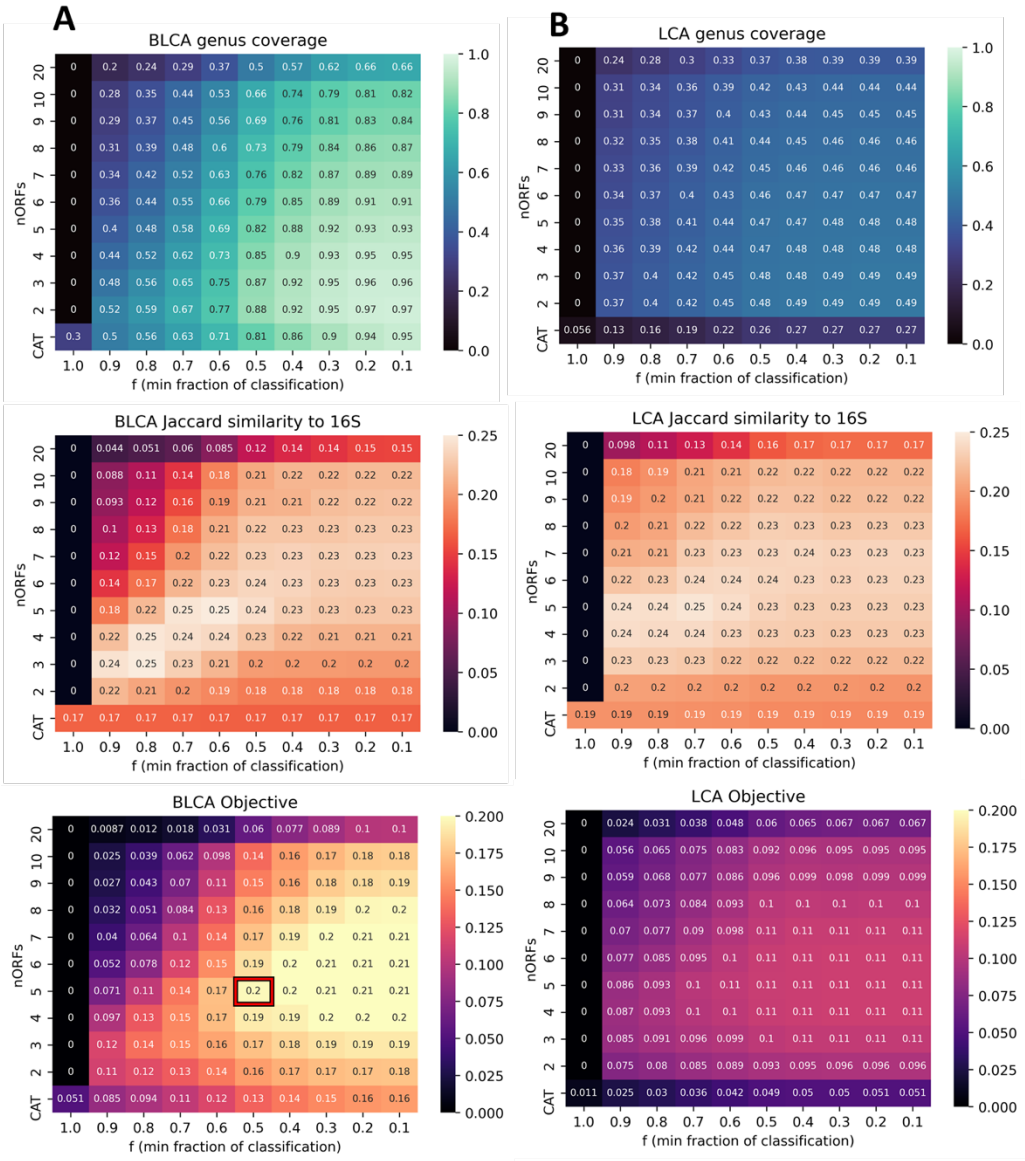


Fig. S2.2 Parameter sweep of fraction of bitscore (f) and mimunum number of ORFS (nORFS) using the FbLCA algorithm for contig LCA, and different protein LCAs (Top hit, BLCA, FtLCA, FbLCA, LCA), and the resulting heatmaps: fraction of annotated sequences with genus level annotation (genus coverage), Jaccard similarity to 16S annotations, and a multiplication of the two (Objective). The final selected parameter set is highlighted in a red square.

2.6.3 16S database completeness and homogenization

Since 16S genes are notoriously hard to include into assembled genomes due to their diverging GC-content and varying copy numbers, there is a discrepancy between databases. If we look at different databases, we also see different length distributions (Fig. S2.3) For Silva ssu parc, which is less curated than silva NR99 we see many shorter fragmented sequences, while for MiDAS we only see full-length sequences. GTDB offers two different ssu (short subunit) 16S databases, reps, which is representative, and all which is all. However only representative databases are available for download from the ftp website for proteins. Therefore, we can compare the taxonomic representation of GTDB protein reps to 16S reps and 16S all of GTDB. When comparing GTDB protein reps to GTDB 16S reps we see that around 30% of genomes in GTDB protein reps do not have a corresponding 16S sequence (MG only), and around 15% of sequences is shorter than 1100 base pairs, which was selected as cutoff for full-length sequences in this study (Table S2.1). Comparing GTDB 16S all to GTDB protein reps shows that here several genomes are present that are not represented in GTDB protein reps (16S only), which would also lead to database discrepancies.

A closer look at how database discrepancy occurs over the taxonomic tree (Fig S5.) shows that on the surface this include Bacteroidetes and Firmicutes_A, which are phyla that typically higher 16S gene copy numbers, which limits how easily they are assembled within MAGs. Still only around 75% of all phyla have one representative in the 16S database (Table S2.1). Looking at the fraction of representation of each individual phylum (Fig. S2.4-5), we see that while several candidate phyla with provisionary names do not have a single 16S representative, but common phyla such as Proteobacteria still have a higher representation.

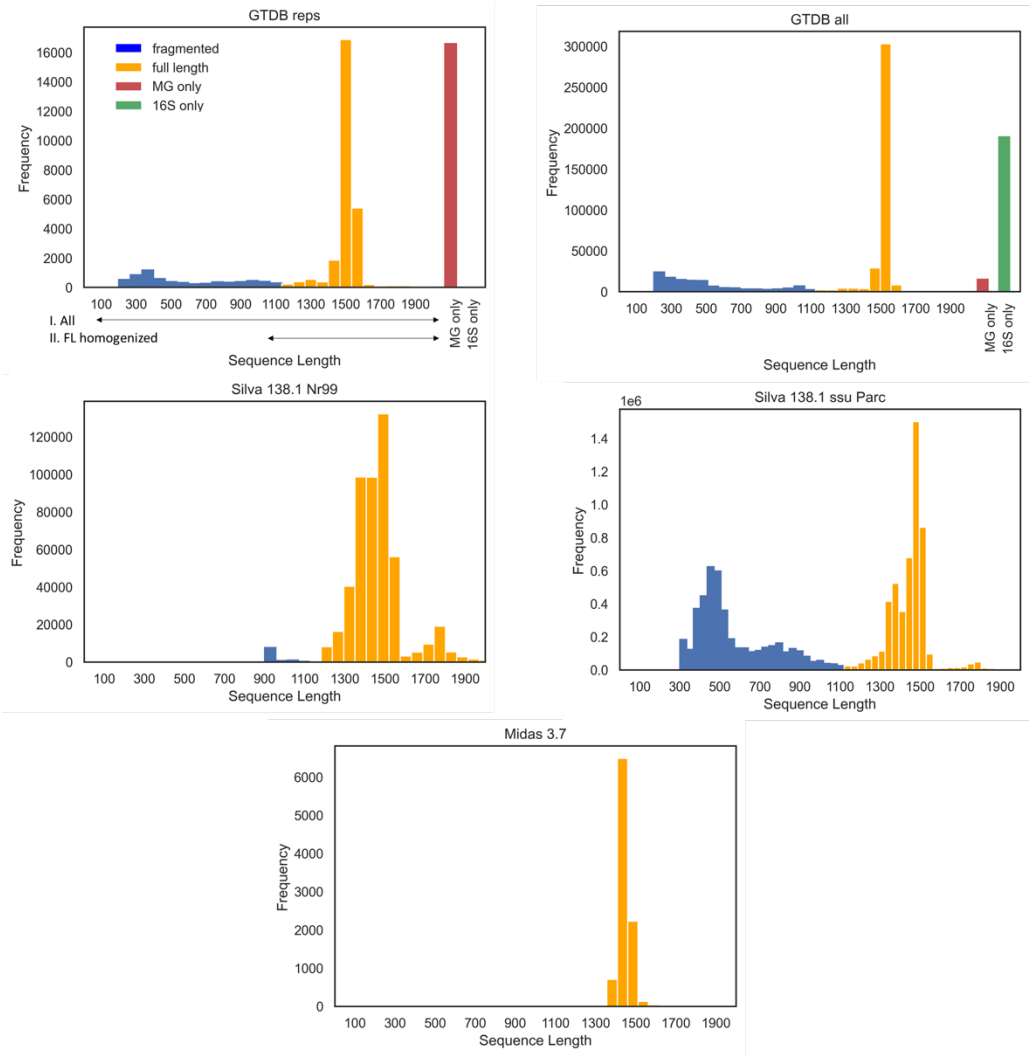


Fig. S2.3 16S length distributions of entries for different databases. For GTDB 16S reps and 16S all, representation was compared GTDB protein reps, which showed entries only present in GTDB protein reps (MG only) and entries present only in the 16S database (16S only).

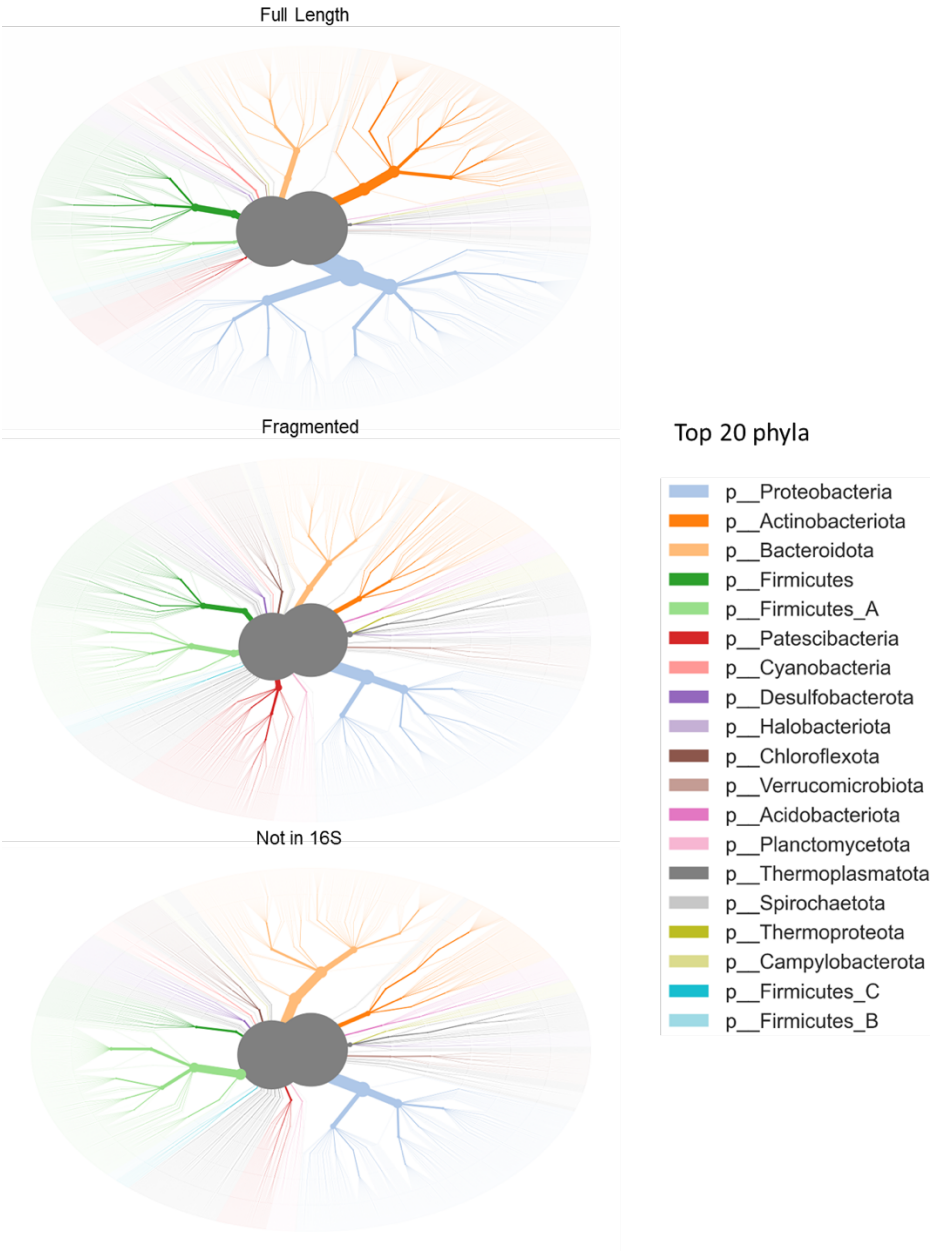
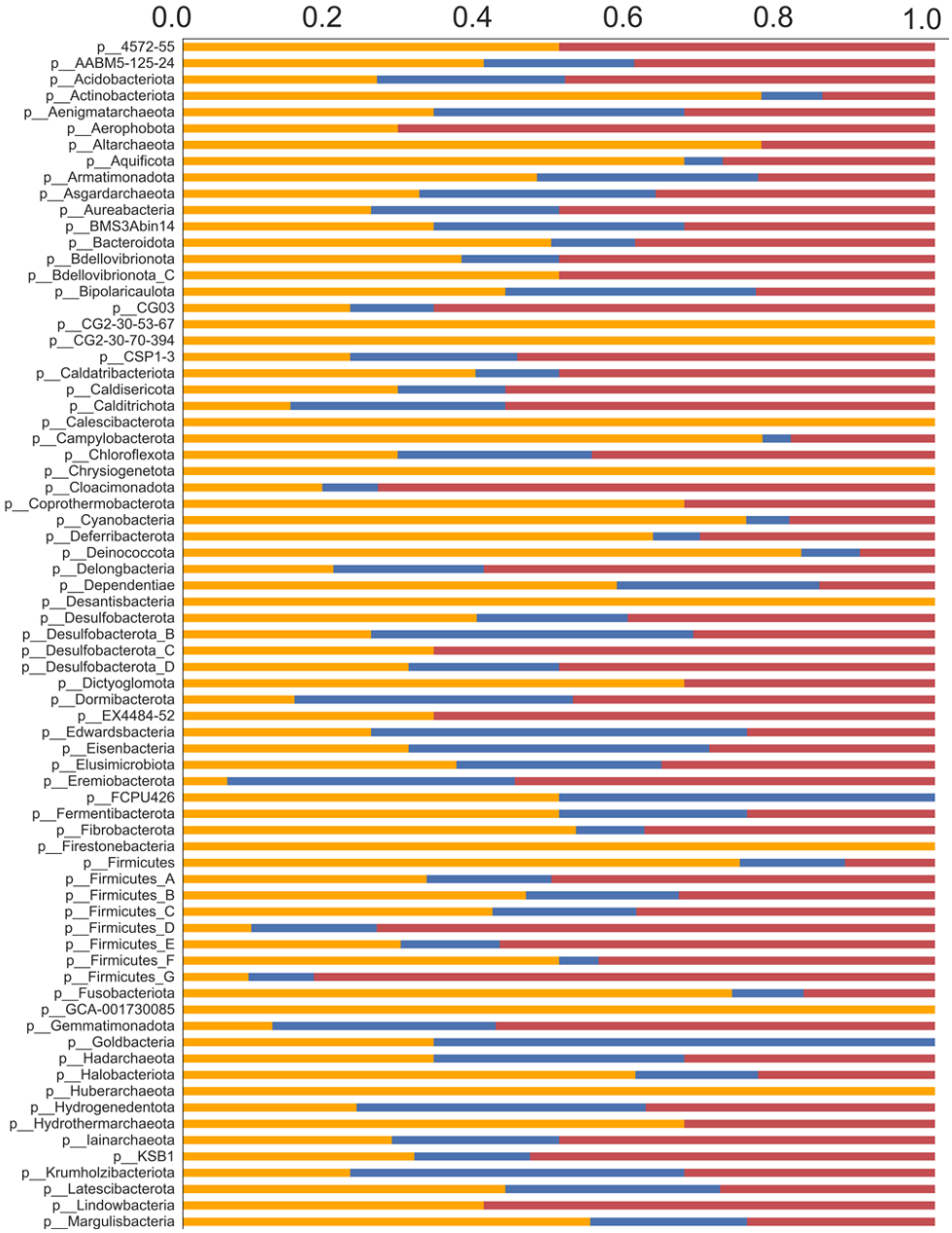


Fig. S2.5 Representation of taxas over taxonomic ranks (superkingdom, phylum, class, order, family, genus & species) within a taxonomic tree for GTDB protein reps genomes, that either contain a full length 16S representative(>1200 base pairs), a fragmented represntative (<1200 base pairs) or no representative (not in 16S).


0.0
0.2
0.4
0.6
0.8
1.0
p__4572-55
p__AABM5-125-24
p__Acidobacteriota
p__Actinobacteriota
p__Aenigmatarchaeota
p__Aerophobota
p__Altarchaeota
p__Aquificota
p__Armatimonadota
p__Asgardarchaeota
p__Aureabacteria
p__BMS3Abin14
p__Bacteroidota
p__Bdellovibrionota
p__Bdellovibrionota_C
p__Bipolaricaulota
p__CG03
p__CG2-30-53-67
p__CG2-30-70-394
p__CSP1-3
p__Caldatribacteriota
p__Caldisericota
p__Calditrichota
p__Calescibacterota
p__Campylobacterota
p__Chloroflexota
p__Chrysiogenetota
p__Cloacimonadota
p__Coprothermobacterota
p__Cyanobacteria
p__Deferribacterota
p__Deinococcota
p__Delongbacteria
p__Dependentiae
p__Desantisbacteria
p__Desulfobacterota
p__Desulfobacterota_B
p__Desulfobacterota_C
p__Desulfobacterota_D
p__Dictyoglomota
p__Dormibacterota
p__EX4484-52
p__Edwardsbacteria
p__Eisenbacteria
p__Elusimicrobiota
p__Eremiobacterota
p__FCPU426
p__Fermentibacterota
p__Fibrobacterota
p__Firestonebacteria
p__Firmicutes
p__Firmicutes_A
p__Firmicutes_B
p__Firmicutes_C
p__Firmicutes_D
p__Firmicutes_E
p__Firmicutes_F
p__Firmicutes_G
p__Fusobacteriota
p__GCA-001730085
p__Gemmatimonadota
p__Goldbacteria
p__Hadarchaeota
p__Halobacteriota
p__Huberarchaeota
p__Hydrogenedentota
p__Hydrothermarchaeota
p__Iainarchaeota
p__KSB1
p__Krumholzibacteriota
p__Latescibacterota
p__Lindowbacteria
p__Margulisbacteria

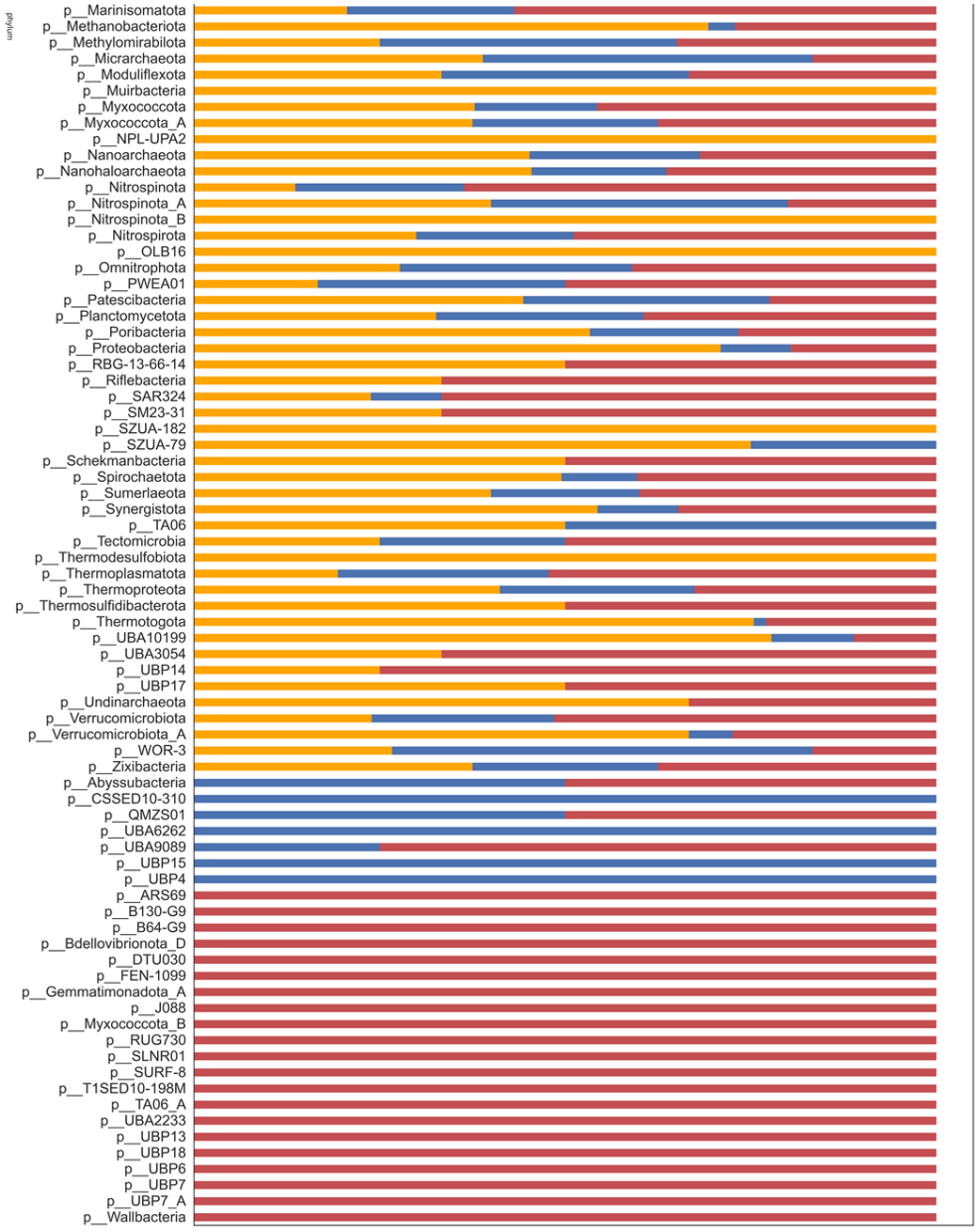


Fig. S2.4 Fraction of coverage of each phylum in GTDB protein reps compared to GTDB 16S reps. Coverage is based on the percentage of species within the phylum that are represented in GTDB 16S reps. Fragmented sequences (blue) are shorter than 1200 base pairs, while full length sequences (orange) are longer, and red is not represented.

Table S 2.1 Comparison of GTDB protein reps and GTDB 16S reps representation for different taxonomic ranks. Coverage indicates which percentage of taxa at each taxonomic rank has at least one representative in the 16S reps database. Fragmented sequences are shorter than 1200 base pairs, while full length sequences are longer.

target	full length	fragmented	MG only	coverage (%)
genomes	26031	6853	15010	54.35
superkingdom	2	0	0	100.0
phylum	121	7	24	79.61
class	290	29	71	74.36
order	721	139	239	65.61
family	1578	474	761	56.1
genus	5185	1741	2893	52.81
species	17753	4202	8618	58.07

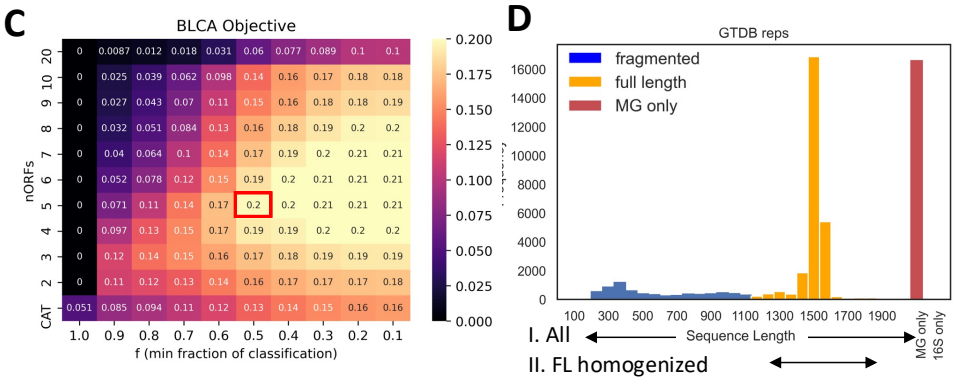


Fig. S2.6 A: Selected SSU GTDB sequences and B: selected parameters for contig annotation.

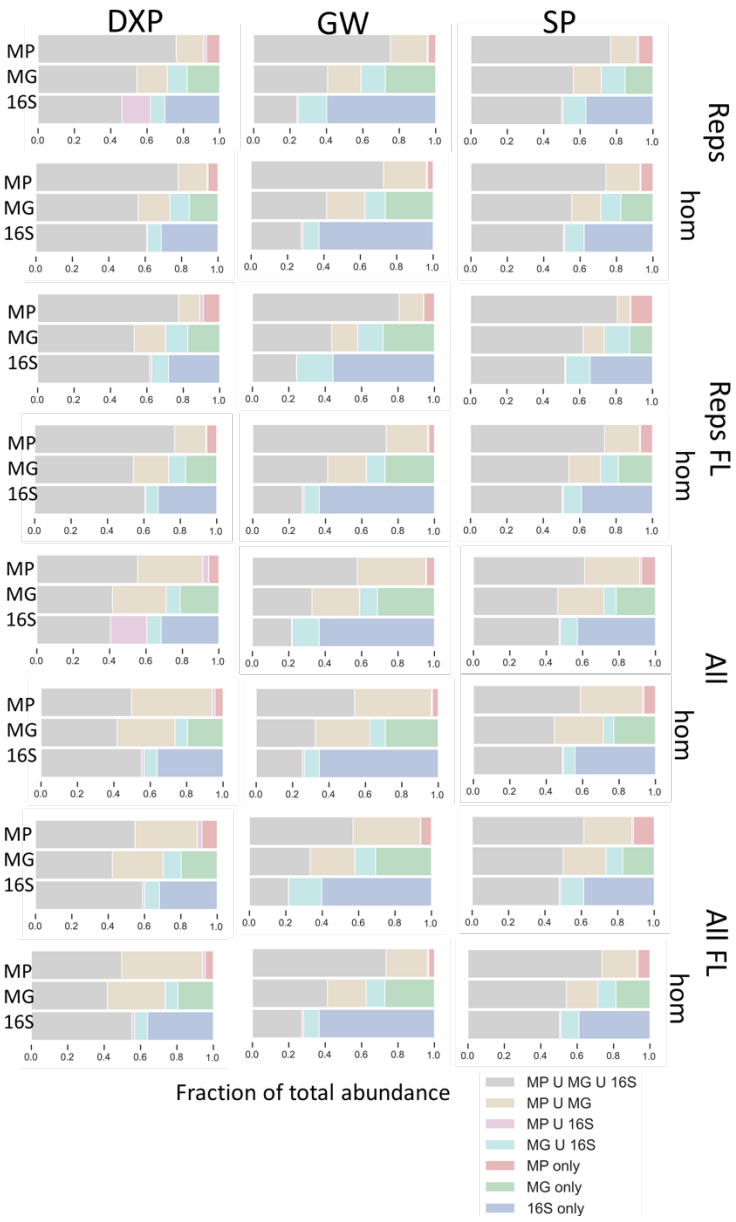


Fig. S2.7 Impact of database homogenization the shared biomass between techniques 16S, metagenomics (MG) and metaproteomics (MP) on different plants: Dinxperlo (DXP), Garmerwolde (GW) and Simpelveld (SP). The techniques were annotated using either homogenized or non homogenized protein databases, and either GTDB 16S reps, GTDB reps 16S full length sequences (FL) or GTDB 16S all or GTDB 16S all full-length sequences. The database used in further analysis is reps FL hom.

2.6.4 Mismatch categorization and extended Sankeys Diagrams

The effect of database homogenization is further compared for individual annotations using Sankey diagrams. Additionally, the full length homogenized GTDB database is compared to other conventionally used databases for 16S: Silva, Midas and proteins: UniprotKB, Uniref100,90,50 Swiss-Prot, and bacterial Refseq and Refseq non-redundant databases. To effectively compare annotations of different databases, 3 issues need to be solved: 1. Small nomenclature shifts, 2. Gapped sequences, and 3. Dump taxa. Between databases there are several small shifts in nomenclature. These can include Candidatus- prefixes, standardized suffixes such as -ota for phyla, or the additional specifications GTDB employs such as Firmicutes_A. Lastly some phylogenetic rearrangements can occur such as shifting annotations from Betaproteobacteria to Gammaproteobacteria (Parks 2018). The distribution of the types of matches and mismatches between annotated contigs, and ASVs are shown in tables S2-5, where GTDB FL homogenized is compared to RefSeq NR or UniprotKB for metagenomics and metaproteomics, and to Midas and Silva for 6S. Categorization goes as follows: Same includes identical annotation in both databases, which can be either matched for both and unmatched (no taxa assigned) for both. Different, indicating different annotation for the 2 compared databases includes name shifts, which are effectively the same organisms but differ only due to naming conventions employed by the database. Dump taxa concern taxa for uncultured or unspecified organisms, which are flagged based on certain keywords (“Uncultured”, “Unclassified” etc.). Other concerns different annotations that cannot be categorized in another way. Missing concerns mismatches that are either annotated only in the GTDB FL annotation (missing-GTDB) or in the database it is compared to (missing-db). Gapped lineages can be caused either by incomplete lineages, or (gapped-lineage) due to lack of rank normalization, or specifically due to dump taxa, which often lack multiple taxonomic ranks (gapped-dump).

Table S2.2 Categorization of the types of matches and mismatches observed for metaproteomic annotation between GTDB FL either RefSeq NR or UniprotKB databases over different taxonomic ranks.

db	rank	Same			Different							
		total	matched	not matched	total	name shift	dump taxa	other	missing		gapped	
									GTDB	db	lineage	dump
Refseq NR	phylum	0.65	0.64	0.01	0.35	0.24	0.00	0.05	0.01	0.04	0.00	0.00
	class	0.31	0.29	0.02	0.69	0.43	0.00	0.16	0.02	0.08	0.01	0.00
	order	0.31	0.27	0.04	0.69	0.21	0.00	0.10	0.02	0.10	0.27	0.00
	family	0.32	0.26	0.06	0.68	0.26	0.00	0.12	0.02	0.11	0.17	0.00
	genus	0.31	0.16	0.16	0.69	0.21	0.00	0.31	0.08	0.09	0.00	0.00
	species	0.27	0.06	0.21	0.73	0.13	0.00	0.27	0.10	0.23	0.00	0.00
Uniprot KB	phylum	0.64	0.63	0.01	0.36	0.29	0.00	0.03	0.02	0.02	0.00	0.00
	class	0.31	0.30	0.02	0.69	0.42	0.00	0.12	0.02	0.06	0.00	0.07
	order	0.27	0.24	0.03	0.73	0.17	0.00	0.08	0.03	0.10	0.25	0.13
	family	0.27	0.22	0.05	0.73	0.19	0.00	0.06	0.04	0.12	0.16	0.18
	genus	0.24	0.11	0.14	0.76	0.21	0.00	0.13	0.10	0.09	0.00	0.28
	species	0.21	0.03	0.19	0.79	0.09	0.28	0.17	0.12	0.19	0.00	0.00

Table S2.3 Categorization of the types of matches and mismatches observed for metagenomics annotation between GTDB FL either RefSeq NR or UniprotKB databases over different taxonomic ranks.

db	rank	Same			Different							
		total	matched	not matched	total	name shift	dump taxa	other	missing		gapped	
									GTDB	db	lineage	dump
Refseq NR	phylum	0.43	0.42	0.01	0.57	0.35	0.00	0.12	0.03	0.07	0.00	0.00
	class	0.29	0.27	0.01	0.71	0.35	0.00	0.23	0.03	0.09	0.02	0.00
	order	0.44	0.41	0.02	0.56	0.16	0.00	0.21	0.04	0.10	0.05	0.00
	family	0.34	0.31	0.03	0.66	0.19	0.00	0.24	0.05	0.11	0.07	0.00
	genus	0.34	0.27	0.07	0.66	0.06	0.00	0.38	0.09	0.13	0.00	0.00
	species	0.30	0.12	0.18	0.70	0.05	0.00	0.24	0.09	0.33	0.00	0.00
Uniprot KB	phylum	0.42	0.41	0.01	0.58	0.41	0.00	0.08	0.04	0.05	0.00	0.01
	class	0.28	0.26	0.02	0.72	0.29	0.00	0.16	0.04	0.10	0.00	0.14
	order	0.37	0.34	0.03	0.63	0.12	0.00	0.10	0.05	0.12	0.03	0.23
	family	0.26	0.23	0.03	0.74	0.11	0.00	0.09	0.06	0.15	0.05	0.31
	genus	0.22	0.15	0.07	0.78	0.05	0.00	0.14	0.11	0.16	0.00	0.39
	species	0.21	0.06	0.15	0.79	0.03	0.39	0.07	0.13	0.27	0.00	0.00

Table S2.4 Categorization of the types of matches and mismatches observed for 16Sannotation between GTDB FL either Midas or Silva databases over different taxonomic ranks.

db	rank	Same			Different							
		total	matched	not matched	total	name shift	dump taxa	other	missing		gapped	
									GTDB	db	lineage	dump
Midas	phylum	0.45	0.45	0.00	0.55	0.41	0.00	0.09	0.04	0.00	0.00	0.00
	class	0.72	0.72	0.00	0.28	0.02	0.00	0.20	0.05	0.01	0.00	0.00
	order	0.42	0.42	0.01	0.58	0.08	0.00	0.39	0.06	0.04	0.00	0.00
	family	0.57	0.56	0.01	0.43	0.05	0.01	0.20	0.10	0.07	0.00	0.00
	genus	0.19	0.15	0.04	0.81	0.05	0.01	0.41	0.25	0.10	0.00	0.00
	species	0.13	0.00	0.13	0.87	0.00	0.00	0.32	0.47	0.09	0.00	0.00
Silva	phylum	0.86	0.86	0.00	0.14	0.03	0.00	0.06	0.04	0.00	0.00	0.00
	class	0.85	0.85	0.00	0.15	0.00	0.00	0.10	0.05	0.00	0.00	0.00
	order	0.71	0.70	0.01	0.29	0.11	0.01	0.11	0.06	0.01	0.00	0.00
	family	0.55	0.54	0.01	0.45	0.13	0.07	0.15	0.10	0.01	0.00	0.00
	genus	0.29	0.23	0.06	0.71	0.06	0.18	0.24	0.22	0.06	0.00	0.00
	species	0.44	0.00	0.44	0.56	0.00	0.21	0.05	0.15	0.23	0.00	0.00

Using GTDB metadata, both Silva and NCBI taxonomy can be mapped to GTDB taxa. By comparing the frequency of mapping with a cutoff of 75% taxa's can be normalized to the GTDB nomenclature system. Another important aspect is annotation gaps. Since many databases do not employ rank normalization, there are several bacteria that lack certain taxonomic ranks. An extreme example of these taxa are dump taxa, which often lack all ranks between class and species, and have unspecific species identifiers such as unclassified Betaproteobacterium. Gapped taxa also occur on non-dump taxa, such as Accumulibacter and Competibacter, both of which lack the order rank, and are essential wastewater treatment organisms. Gapped sequences present problems in the fact that they would allow organisms to be identified only at specific ranks, but this also affects lowest common ancestor algorithms. To be prevent loss of information, gapped sequences can be filled based and flagged based on their nearest child, while dump taxa can be flagged by targeting keywords such as unclassified and uncultured. A first comparison of 16S databases shows high similarity between annotations (Fig. S2.13). The Midas database is partly based on Silva annotations but has many manual additions and nomenclature shifts. Since the exact links are not accessible, the nomenclature cannot be directly homogenized in all cases to GTDB standards, which does result in some shifts like Nitrospiria -Nitrospira on class level. Other than that, Silva contains dump taxa, and commonly employ the

name “metagenome” for uncultured samples. This is apparent at species level, for which Silva annotations are dominated by dump taxa. Still up to genus level annotations seem highly similar between databases, except for the lack of Tetrasphaera annotations in GTDB. Tetraspheara is considered a key phosphate accumulating organism that is detected in both activated sludge and large granules (Stockholm-Bjerregaard 2017, Ali 2019). Upon closer inspection with Blast+ only Tetrapsheara Japonica is matched in alignment, but other Dermatophilacae have better identity scores in GTDB, which results in an LCA annotation only at family level. Aligned genera with better scores include genera not represented in Midas, such as Terrabacter and Phycococcus (Sup excel sheet x). Therefore, lack of Tetrasphaera could be a result of database incompleteness on either side, phylogenetic rearrangement in GTDB, and also lack of taxonomic resolution of the V3-V4 16S primer used in this study, however it is also important to mention that Tetrasphaera is not significantly detected in metagenomics and metaproteomic annotations with any protein database. For protein database comparison, also a similar distribution between top taxa is observed at genus level (Fig. S2.11-12), except for SwissProt, a very small, curated database included as negative control. However, while the ratios of top taxa seem similar there is a clear difference in the number of annotated reads and peptides. This is due the presence of dump taxa and gapped ranks, which are observed for both RefSeq and UniprotKB derived databases, both of which employ NCBI taxonomy. The effect is still more pronounced for UniprotKB derived databases. Uniref50 shows the worst performance of commonly used metaproteomic databases and annotates around one third of the sequences annotated with GTDB on family level. Some differences observed for top taxa is that UniprotKB annotates less Competibacter sequences for metaproteomics than RefSeq and GTDB, while GTDB annotates less rhodobacter sequences. We can also use Sankey diagrams to visualize changes in annotation upon using different GTDB 16S databases and length trimming (Fig. S2.8-10). When comparing different GTDB databases: reps, reps full length, and all, all full length, some difference observed at species level where the absence of certain taxa leads to contig LCAs being skewed in a different direction. For metagenomic and meproteomic database homogenization overall has limited effect, mostly visible at species level, but does lead to ~5% of genus level annotations changing (Fig. S2.7). An example is the top metaproteomic genus uba_7236, a provisionary taxon that does not contain a full-length 16S representative and is not present in the full-length homogenized annotation.

Proteomics, GTDB-only comparison

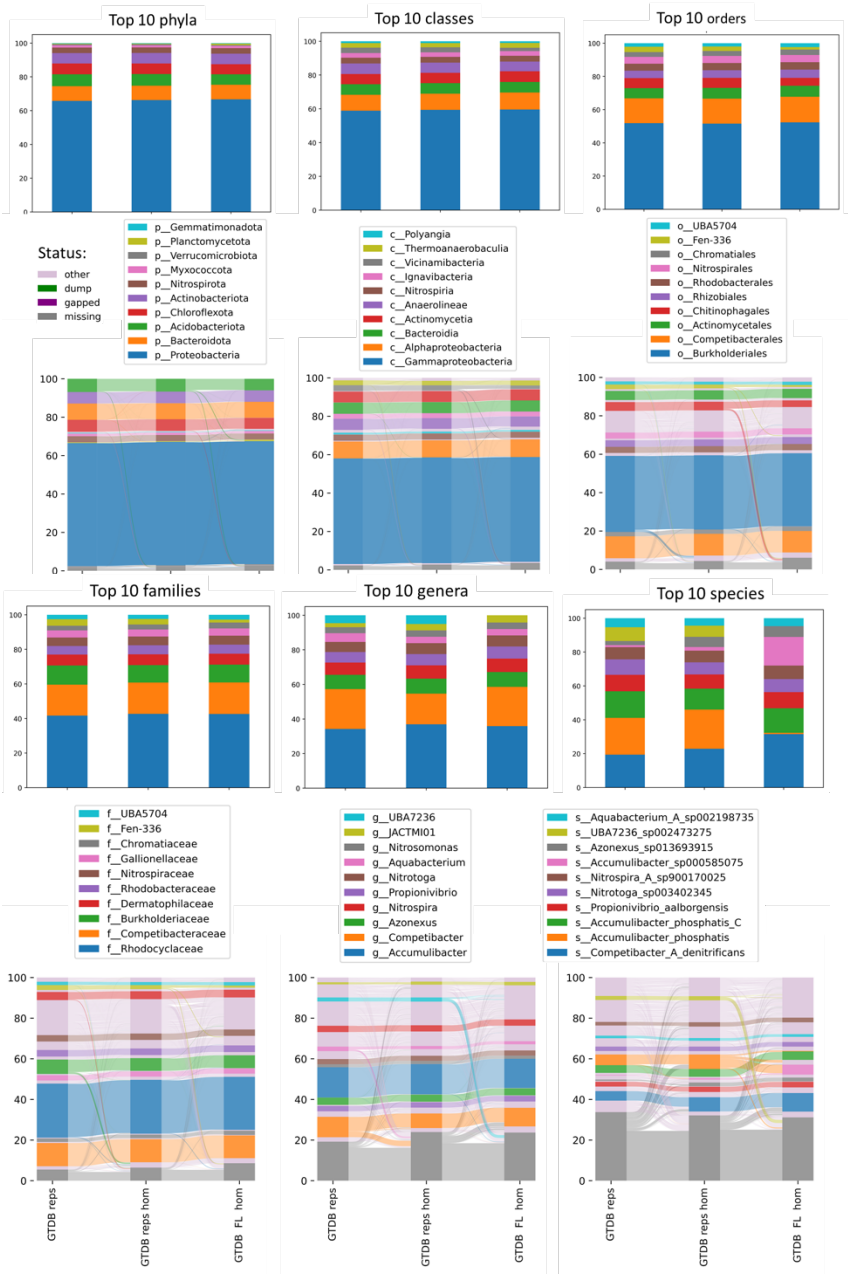


Fig. S2.8 Sankey diagram and top 10 taxa for metaproteomic annotation of combined with 3 different GTDB derived protein databases (GTDB non-homogenized, GTDB reps homogenized and GTDB reps FL homogenized).

Genomics, GTDB-only comparison

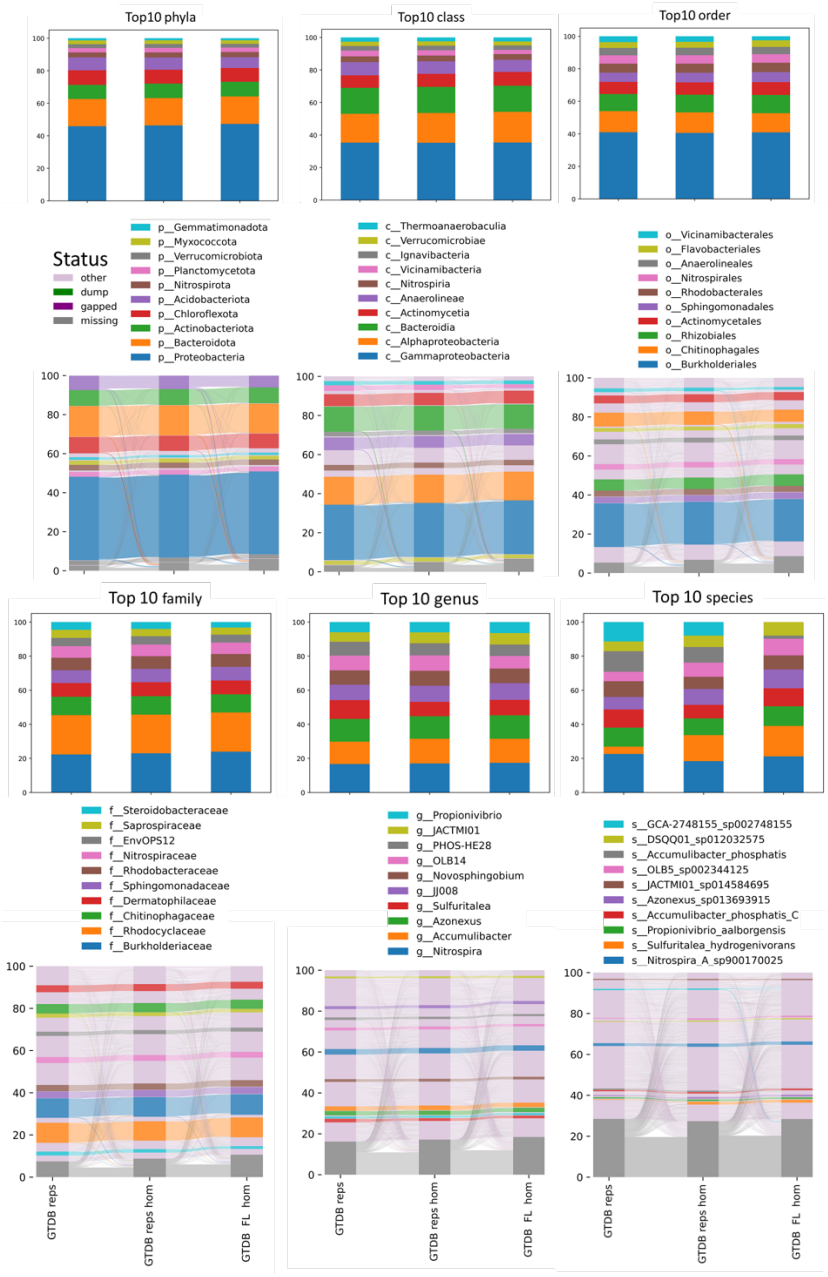


Fig. S2.9 Sankey diagram and top 10 taxa for metagenomic annotation of combined with 3 different GTDB derived protein databases (GTDB non-homogenized, GTDB reps homogenized and GTDB reps FL homogenized).

16S, GTDB-only comparison

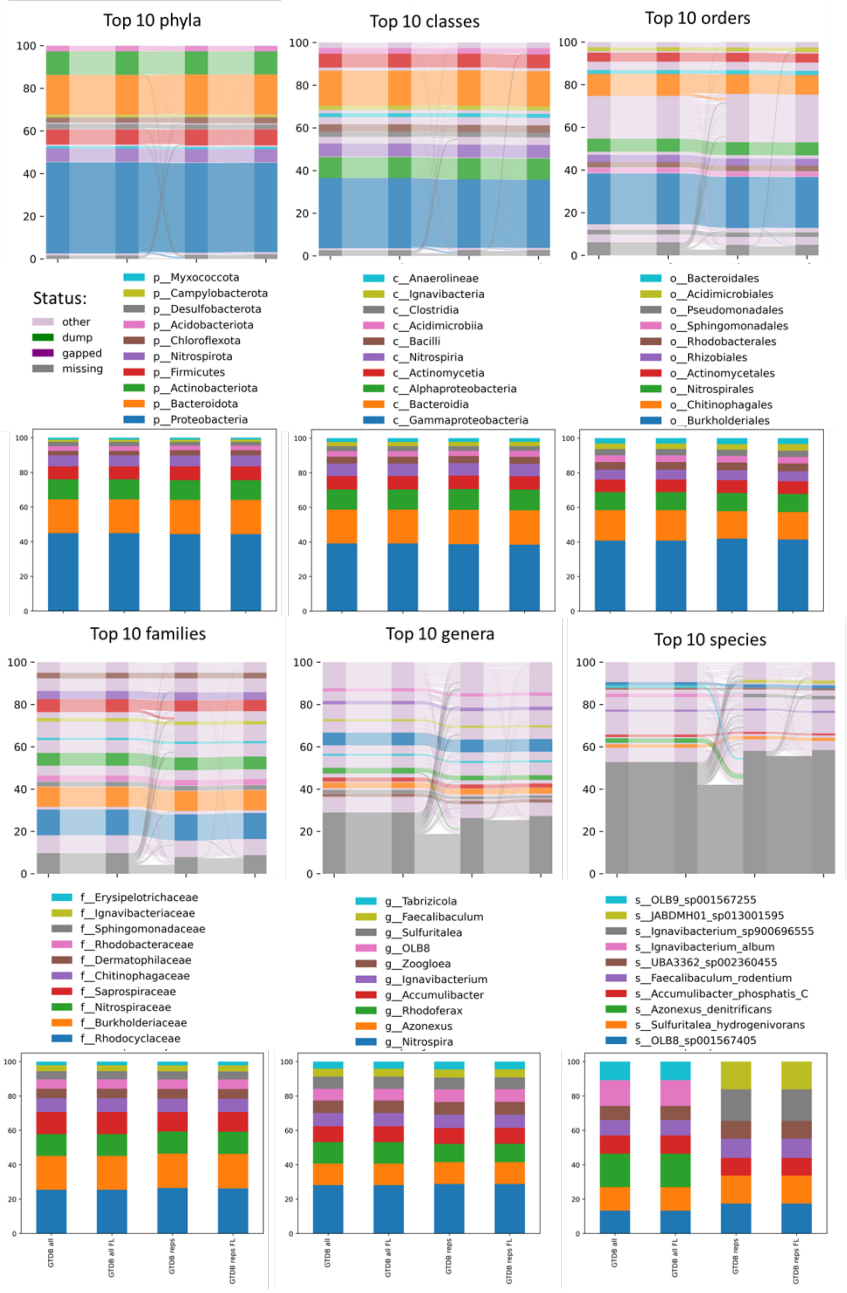


Fig. S2.10 Sankey diagram and top 10 taxa for 16S annotation of combined with 4 different GTDB derived ssu databases (Reps, Reps FL ,All, All FL).

Proteomics, extended db and ranks comparison

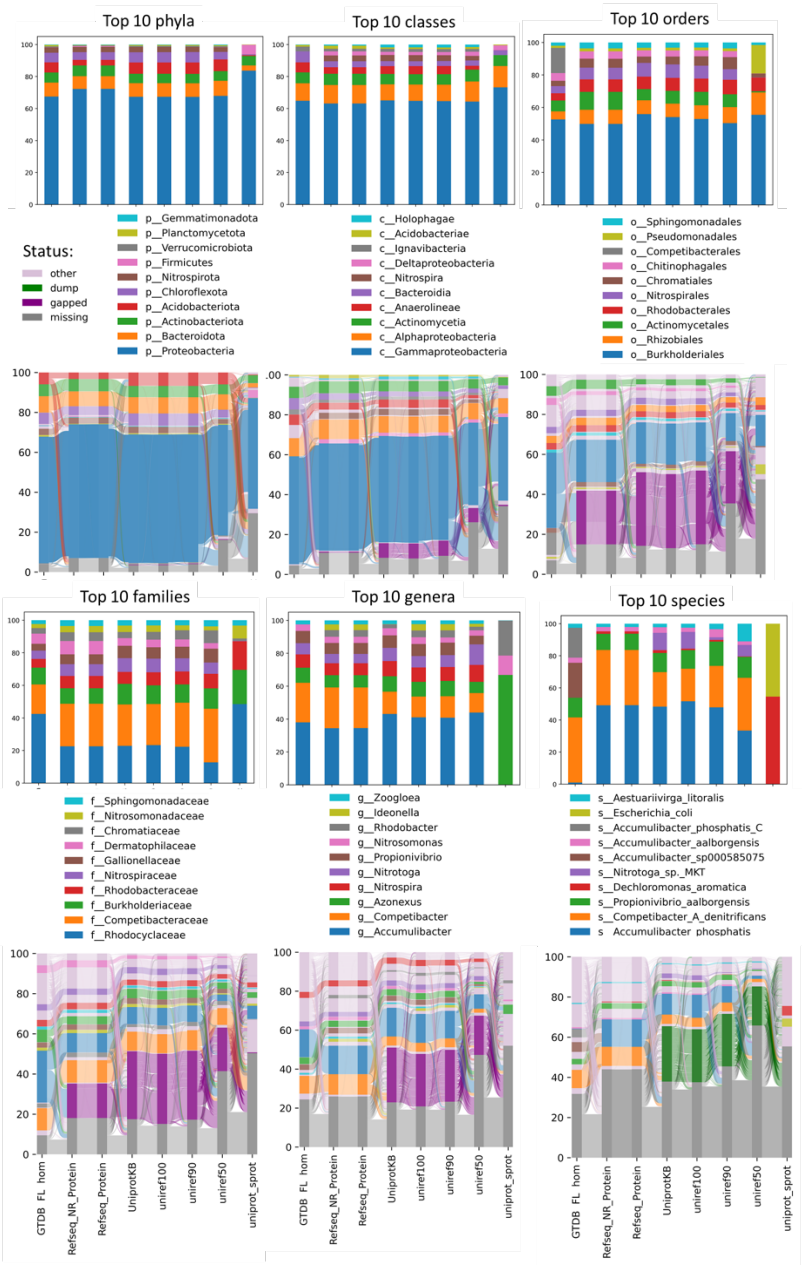


Fig. S2.11 Sankey diagram and top 10 taxa for metaproteomics annotated to 8 different protein databases (full length GTDB reps, UniprotKB, Uniref100,90,50 SwissProt, and bacterial redundant and non-redudant RefSeq)

Genomics, extended db and ranks comparison

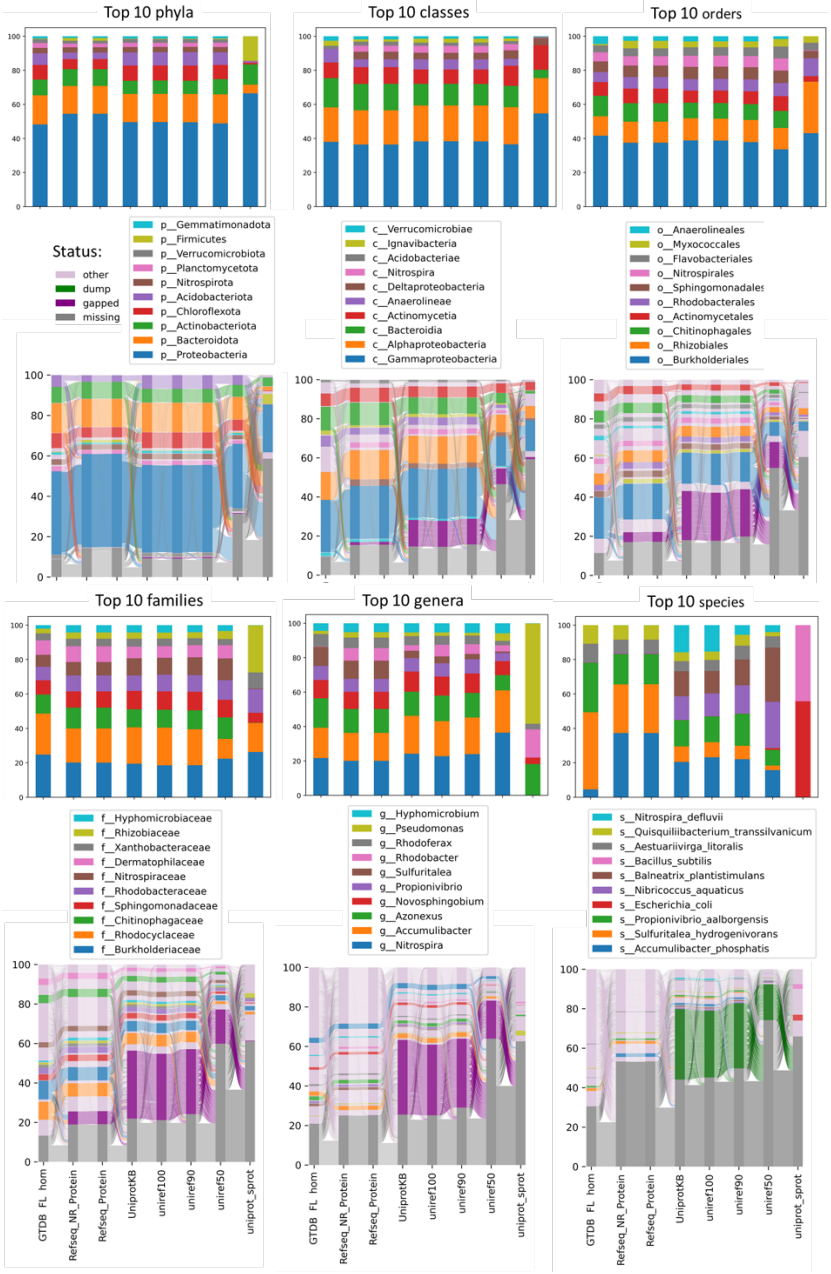


Fig. S2.12 Sankey diagram and top 10 taxa for metaproteomics annotated to 8 different protein databases (full length GTDB reps, UniprotKB, Uniref100,90,50 SwissProt, and bacterial redundant and non-redudant RefSeq)

16S extended ranks comparison

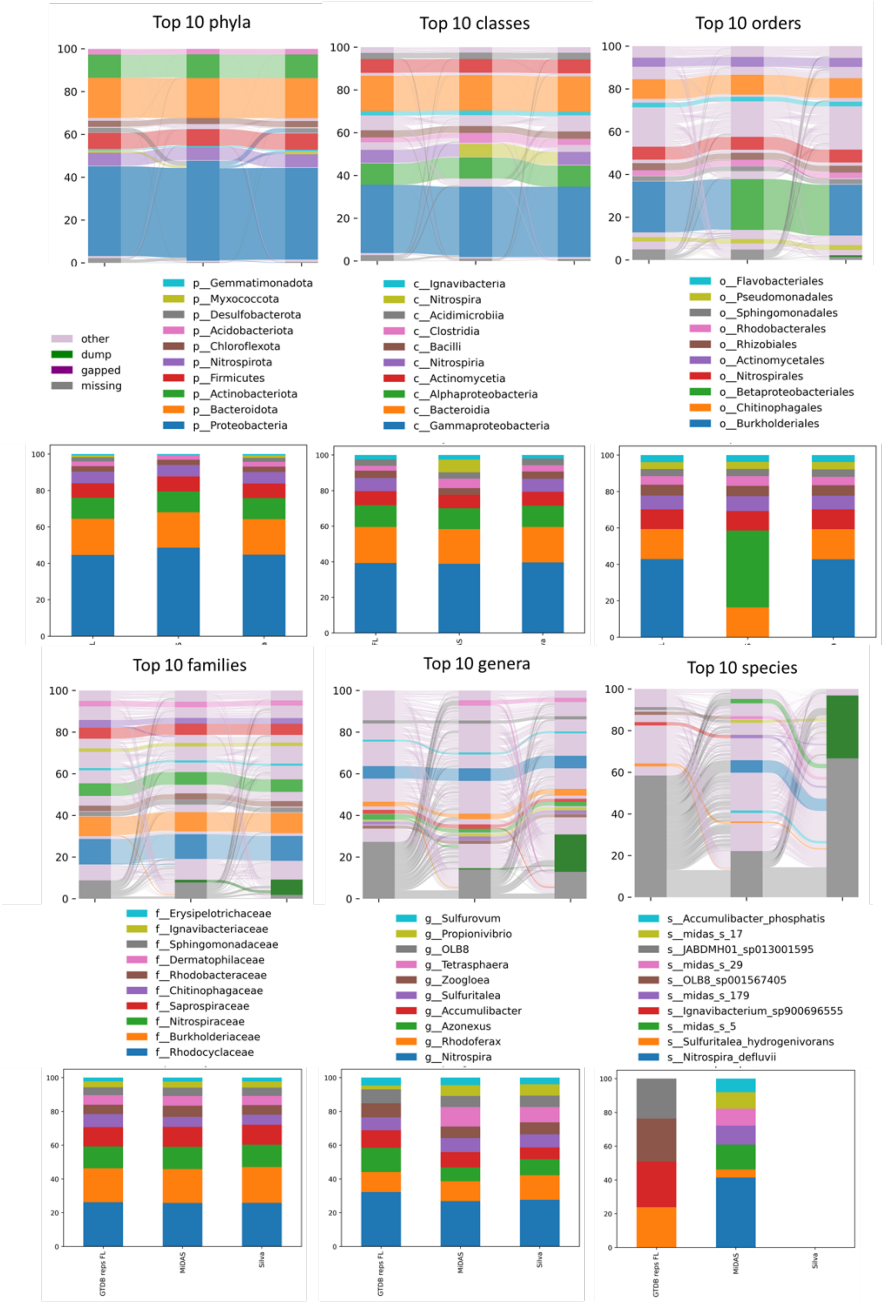


Fig. S2.13 Sankey diagram and top 10 taxa for metaproteomics annotated to 3 different ssu databases (full length GTDB reps, Midas and Silva)

2.6.5 Tetrasphaera annotation with BLAST+

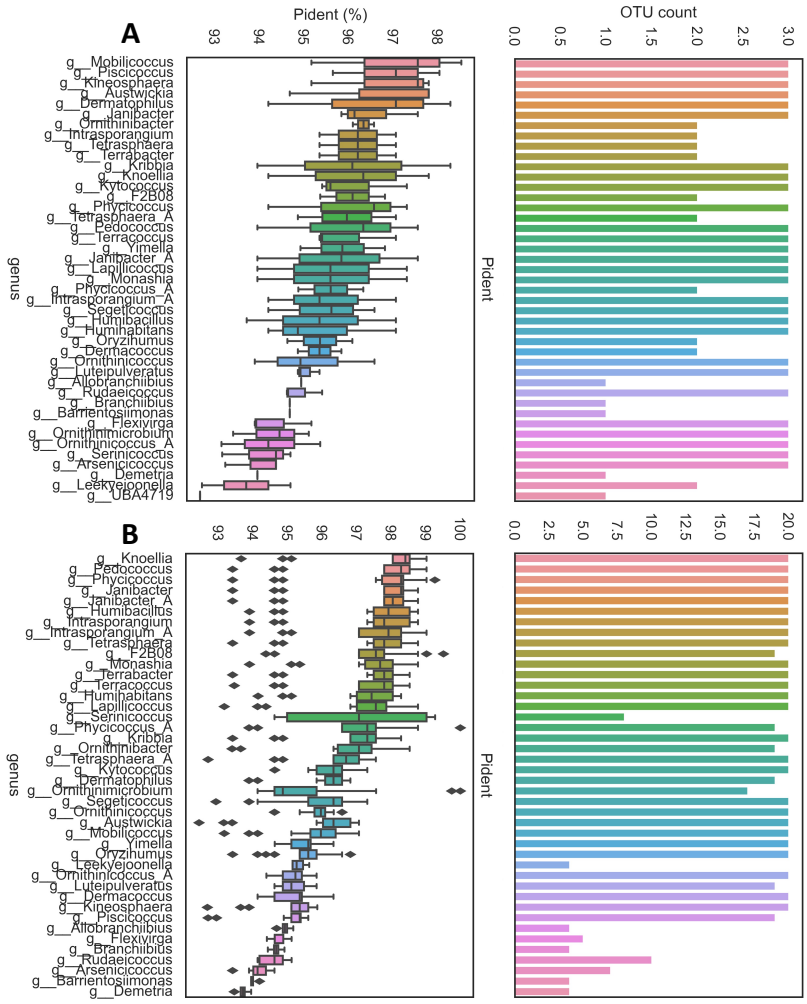


Fig. S2.14 Tetrasphaera ssu annotations in GTDB. A: distribution of GTDB genera aligned with blastn for Silva ASVs with family Dermatophilaceae, A: distribution of GTDB genera aligned with blastn for Midas ASVs with family Intrasporangiaceae.

A key difference between GTDB ssu anntoations and Silva, Midas was the absence of Tetrasphaera. Using BlASTN ASVs were aligned to GTDB FL ssus. For Silva 3 ASVs were annotated with Tetrasphaera (Fig. S2.14A), and for Midas 20 (Fig. S2.14B). The alignment shows that several genera within GTDB show better identity percentage than genera annotated as Tetrasphaera within GTDB.

2.6.6 Computational bias

Contig LCA influence

An important driver of bias is the effect of conserved proteins on contig level LCA. Sweeping the fraction of bitscore parameter (f) applied in contig LCA shows that some genera are more sensitive towards increased stringency of "f". By normalizing the relative abundance of organisms to their abundance at f=0.1 their shift in abundance can be expressed. Depending how well they match sequences in the database, some taxa will more easily lose annotations at higher ranks because of dispersed alignments with matched candidates belonging to different taxa. Since expressed proteins are often more conserved this effect is more pronounced in metaproteomics (Fig. S2.15A) than in metagenomics (Fig. S2.15B) which creates a quantification bias. Normalized abundance values to f=0.1 can be subtracted between techniques (Fig. S2.15C), which is highlighted more clearly at genus level (Fig. S2.15D). This helps explain why certain genera are underrepresented in metaproteomic annotations. Sequences aligned to Expressed proteins of Rhodoferax, JJ008 (Chitinofagacae), WYBL01 (Vicinamibacteria), OLB14 (Anaerolineales), Aquabacterium_A, both decrease steeply fraction as stringency of the f parameter increases. For metagenomics, this also includes Competibacter_A. This indicates that aligned sequences of these organisms include a wider variety of matched taxa within its contigs. If the match has a high evalue, this infers that the predicted ORF is unspecific, if the evalue is low, this could indicate database incompleteness.

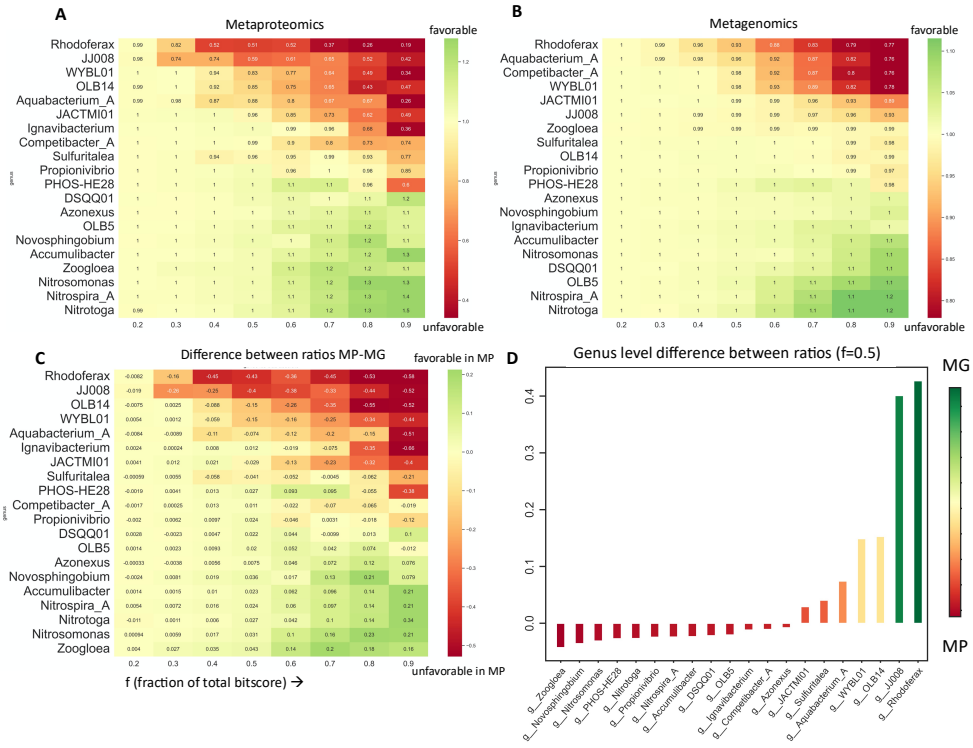


Fig. **S2.15** Effect of sweeping the allowed fraction of bitscore parameter (f) in contig LCA applied on combined samples (DXP, GW, SP), on metaproteomics (A), metagenomics (B), and the difference between the shifts (C), zoomed in at genus level (D) with f=0.5, which was the parameter setting applied in this study, with genera favored in metagenomics (MG) colored green, and genera favored in metaproteomics (MP) colored red.

Bayesian vs BLAST

Standard annotation of 16S sequences is done with Bayesian classifiers, which create a statistical estimator for different taxa. Since BLAST weighs each residue equally, the different taxonomic annotation methods employed between 16S (Bayesian) and metagenomics (BLAST) could be a source of bias. To investigate this, the percentual difference was computed between normal 16S annotation vs BLAST metagenomics (Fig. S2.16B) and 16S annotated with nucleotide blast vs BLAST metagenomics (Fig. S2.16A). The two percentual differences can be subtracted from one another to see which taxa are favored by either annotation. This showed that genera Rhodoferax and Propionivibrio are annotated less with Bayesian annotation than BLAST, which can help explain quantification differences between the methods.

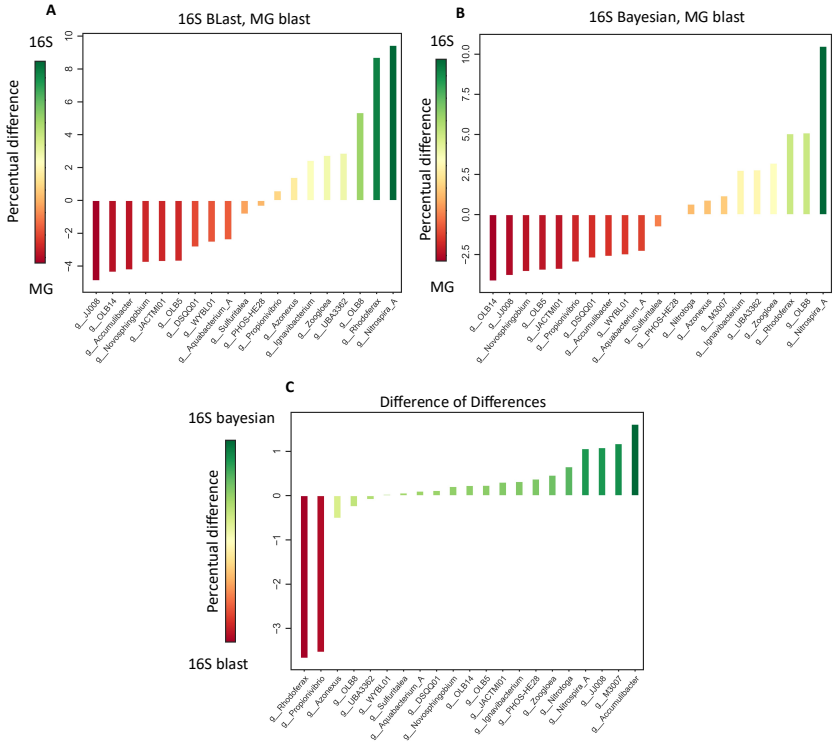


Fig. S2.16 Effect of using Bayesian classifiers in 16S vs BLAST applied on combined samples (DXP, GW, SP), both compared to metagenomics (MG) annotations with BLAST. With 16S BLAST vs MG BLAST (A), 16S Bayesian vs MG BLAST (B), and both percentual differences subtracted (C).

Peptide LCA

Another source of computational bias is the additional peptide LCA step which is required when a detected peptide is found in multiple organisms within the database. The resulting annotation takes the lowest common ancestor of the shared peptide. If an organisms has many peptides in the database that overlap with other taxa, it will be disadvantaged. This is observed strongly for Zoogloea, which has significantly more annotations in metaproteomics when no peptide LCA is applied.

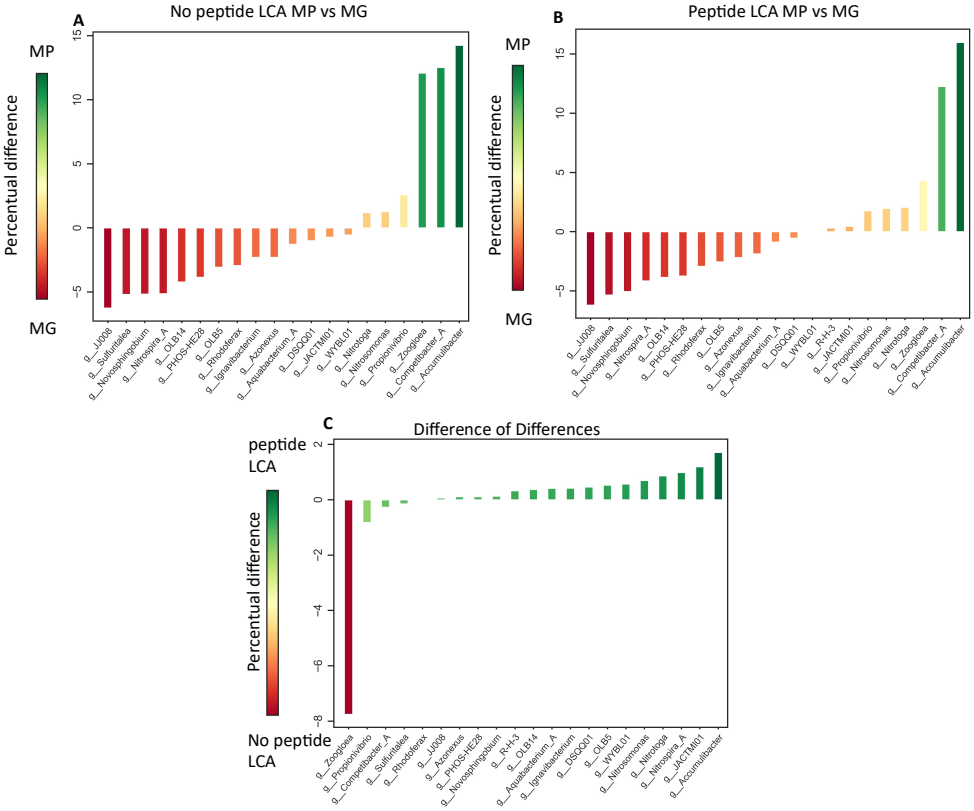


Fig. S2.17 Effect of using peptide LCA applied on combined samples (DXP, GW, SP), on metaproteomic (MP) quantification both compared to metagenomics (MG) annotations. With no peptide LCA (A), peptide LCA (B), and both percentual differences subtracted (C).

2.6.7 Physiological bias

Since obtaining accurate cell volume measurements for separate genome is a massive challenge, literature reported cell volumes were used instead to correlate to observed quantification bias. Not all genera of the subset could be linked to existing cell volume data, as this was not available for some of the rarer taxa.

Table S2.5 Literature reported cell volumes

genus_name	Literature volume (μ^3)	Area (μ^2)	Width	Length	Reference
g__Accumulibacter	4.21	14.43	1.75	1.75	Kim 2010
g__Aestuariivirga	0.70	4.52	0.80	1.40	Li 2019
g__Aquabacterium_A	1.05	5.94	0.90	1.65	Chen 2016
g__Azonexus	1.66	8.22	1.00	2.12	Reinhold-Hurek 2006
g__Bradyrhizobium	0.81	5.39	0.70	2.10	Somasegaran 2012
g__Brocadia	0.40	3.02	0.80	0.80	Van Niftrik 2008
g__CAIMWE01					
g__CFX10	4.91	78.64	0.25	100.00	Yamada 2006
g__CFX2	4.91	78.64	0.25	100.00	Yamada 2006
g__Competibacter_A	12.57	31.42	2.00	4.00	Mcllroy 2015
g__DSQQ01	4.91	78.64	0.25	100.00	Yamada 2006
g__Flavobacterium	0.39	3.53	0.50	2.00	Inzana 1999
g__GCA-013693735	3.03	12.73	1.13	3.02	Marin 2014
g__Hyphomicrobium	0.04	0.69	0.25	0.75	Oren 2014
g__Ignavibacterium	1.90	30.53	0.25	38.75	Liu 2012
g__JACTMI01					
g__JJ008	0.87	5.72	0.70	2.25	Siddiqi 2017

g__Litorilinea	19.63	157.47	0.50	100.00	Barberán 2017
g__M3007	2.20	15.98	0.57	8.64	Mcllroy 2014
g__Nitrosomonas	1.78	8.33	1.12	1.82	Soliman 2018
g__Nitrospira_A	0.11	1.60	0.30	1.55	Ehrich 1995
g__Nitrotoga	0.20	1.96	0.50	1.00	Kitzinger 2018
g__Novosphingobium	0.19	2.12	0.41	1.46	Choi 2015
g__OLB13	4.91	78.64	0.25	100.00	Yamada 2006
g__OLB14	4.91	78.64	0.25	100.00	Yamada 2006
g__OLB5	1.90	30.53	0.25	38.75	Liu 2012
g__OLB8	2.20	15.98	0.57	8.64	Mcllroy 2014
g__PHOS-HE28	0.39	3.53	0.50	2.00	Inzana 1999
g__PNKE01	0.35	2.81	0.65	1.05	Barberán 2017
g__Propionivibrio	0.34	3.03	0.51	1.63	Oren 2014
g__QUBU01					
g__Quisquiliibacterium	0.37	2.91	0.65	1.10	Felföldi 2017
g__R-H-3	0.48	3.56	0.68	1.32	Tamura 2013
g__RHKY01					
g__RXIV01	12.57	31.42	2.00	4.00	Mcllroy 2015
g__Ramlibacter	2.14	8.04	0.80		De Luca 2011
g__Rhodoferax	0.99	6.19	0.75	2.25	Hiraishi 1991
g__Rubrivivax	0.48	3.93	0.55	2.00	Nagashima 2102
g__SCN-69-89					
g__SCN-70-22	1.10	7.04	0.70	2.85	Hanada 2014
g__Sphingobium	0.11	1.38	0.40	0.90	Qin 2020

g__Sulfuritalea	0.28	3.08	0.40	2.25	Oren 2014
g__Tepidiforma	0.18	2.01	0.40	1.40	Kochetkova 2020
g__Thauera	0.97	5.96	0.77	2.08	Oren 2014
g__UBA3362	2.20	15.98	0.57	8.64	Mcllroy 2014
g__WYAZ01					
g__WYBL01					
g__Zoogloea	1.42	7.46	0.94	2.06	Oren 2014

Next to the cell volume, one of the strongest factors of the quantification difference between metaproteomics and metagenomics comes from strongly expressed proteins. To express this into numbers, proteins were counted until a fraction 0.3 of the total abundance was reached for a signal (expressed in K). Some K values are shown in Table S3 and Fig S18-S20, however K values generally depend also on the abundance of the organism, as low abundant organisms will have lower coverage, they will also have a lower K value, as there are less proteins identified for those taxa. Therefore if an organism is highly abundant and still has a low K value, this means that it is strongly overexpressing certain proteins, which will affect shotgun metaproteomics.

Table S2.6 Showcase of some of the more abundant genera and their K values, and relative abundance (%) based on peptide spectrum matches (PSMs)

Index	genus_name	K	PSMs	PSMs/K
1	GW g__Accumulibacter	8	26.26	3.28
2	GW g__Competibacter_A	12	13.76	1.15
3	SP g__Accumulibacter	13	11.16	0.86
4	DXP g__Accumulibacter	11	10.11	0.92
5	DXP g__Zoogloea	8	9.02	1.13
6	SP g__Competibacter_A	10	7.98	0.80
7	DXP g__Nitrospira_A	8	7.26	0.91
8	GW g__Nitrospira_A	6	6.32	1.05
9	DXP g__Azonexus	5	5.51	1.10
10	SP g__Propionivibrio	10	5.18	0.52
11	SP g__Nitrotoga	3	5.17	1.72
12	DXP g__Propionivibrio	5	5.12	1.02

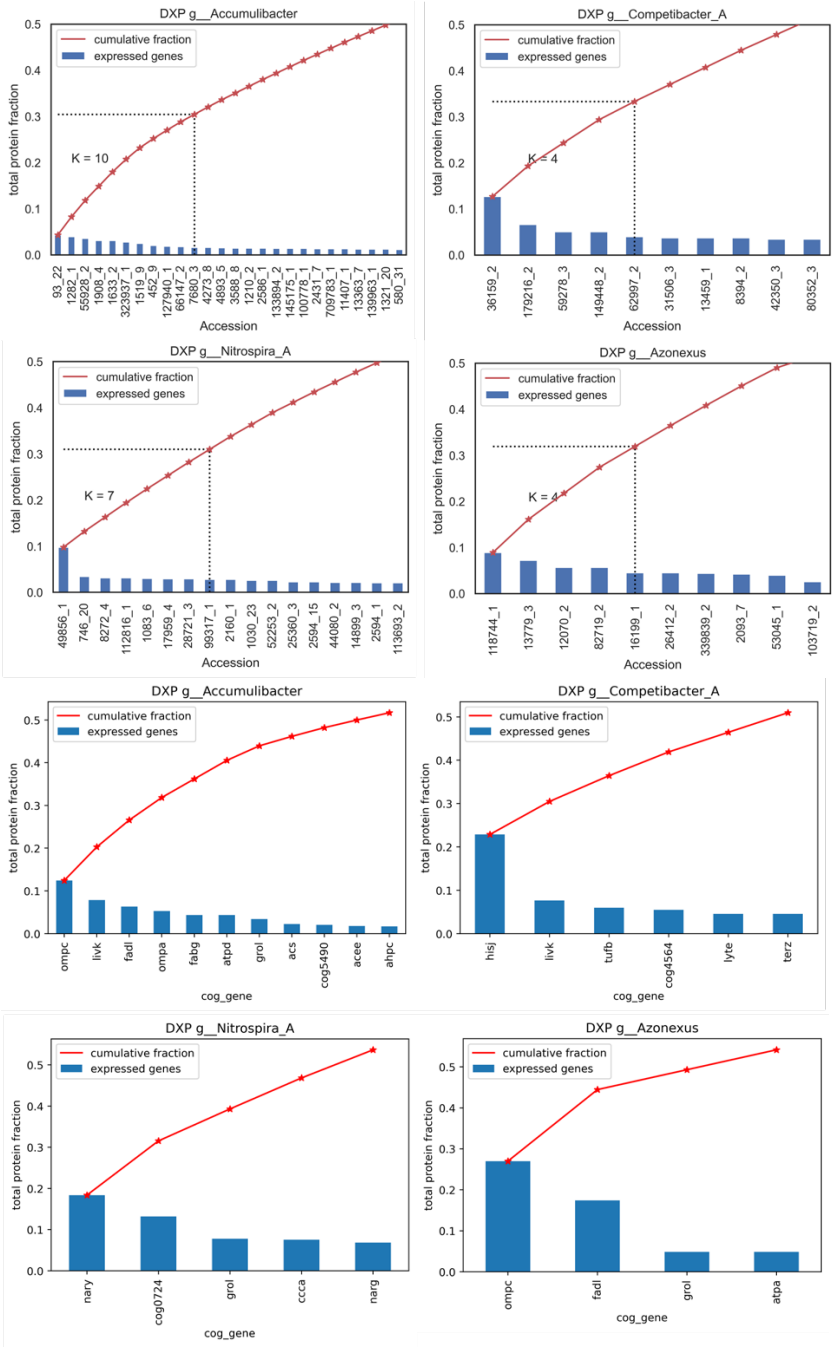


Fig. S2.18 Cumulative fraction of proteins, and COG annotated genes, for various genera in the sample DXP and their K-values.

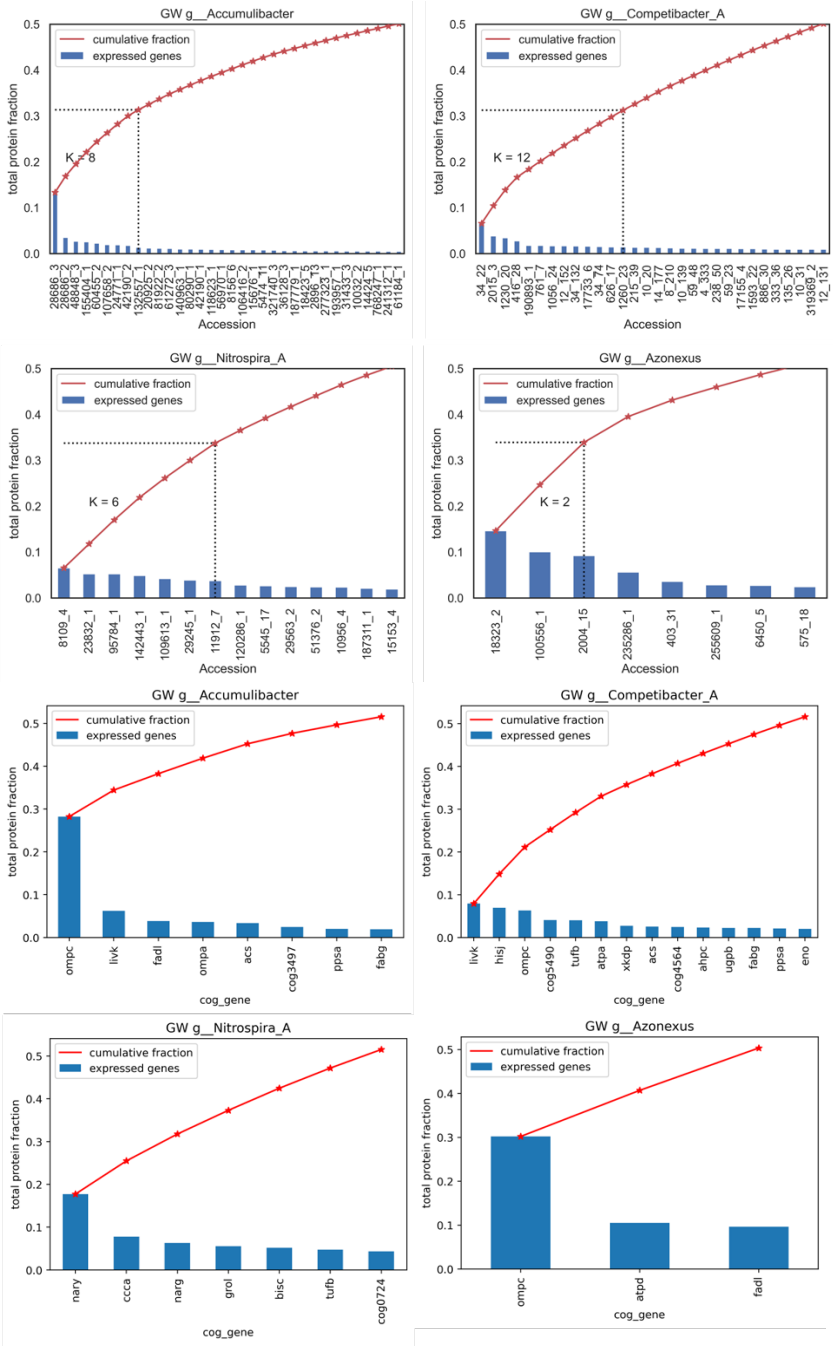


Fig. S2.19 Cumulative fraction of proteins, and COG annotated genes, for various genera in the sample GW and their K-values.

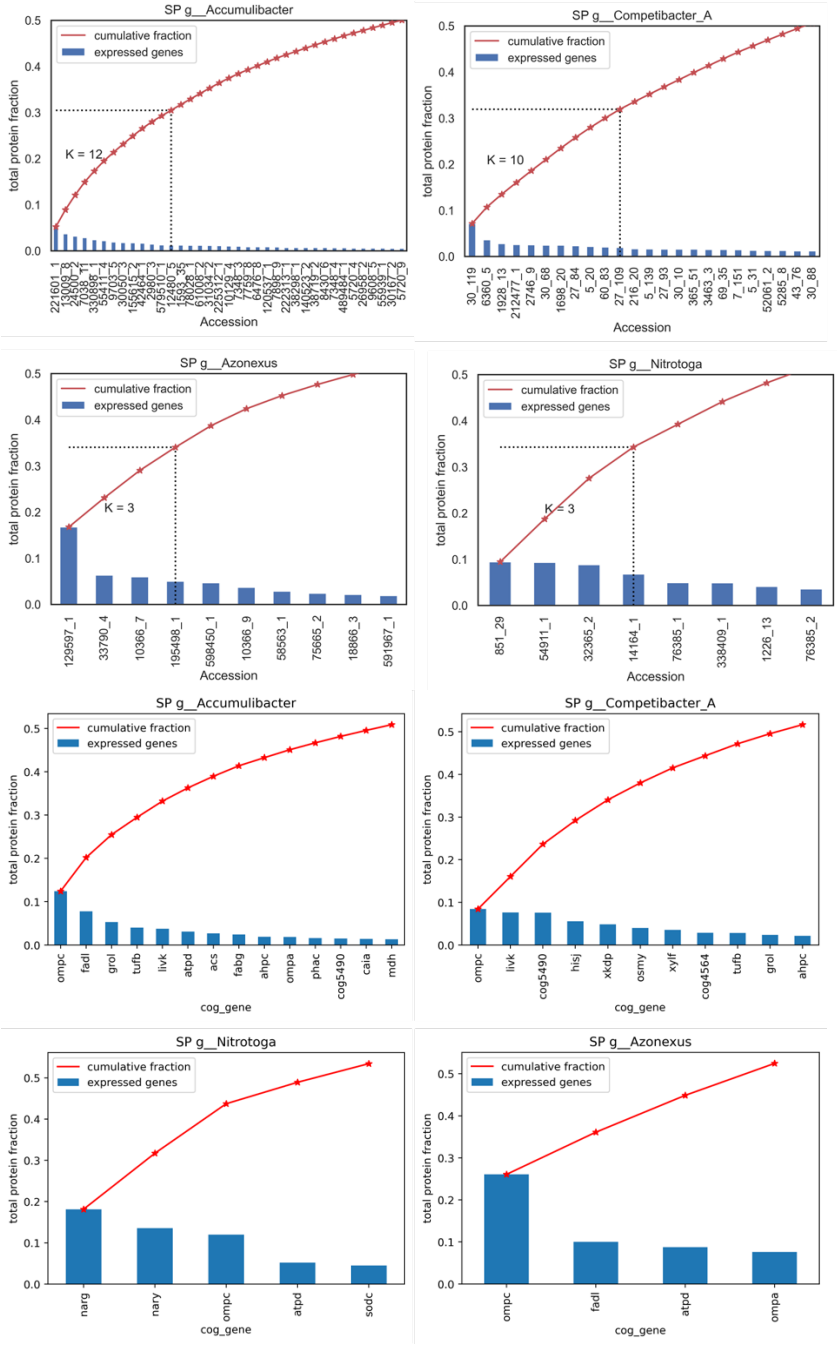


Fig. S2.20 Cumulative fraction of proteins, and COG annotated genes, for various genera in the sample SP and their K-values.

2.6.8 Porin associated bias

Since porins are so strongly detected by proteomics in the samples, it was investigated if there was anything giving them an advantage in detection, other than their strong expression. Protein conservedness would not improve detection per se, but influences at which taxonomic rank proteins are observed, which could bias annotation. The conservedness of nutrient removal genes and cog categories was analyzed by making box-plots of the highest taxonomic ranks they are annotated at (Fig S21). This revealed that porin proteins are less conserved than most other COG categories. Since they are membrane bound, porin proteins could be advantaged in ionization and fragmentation due to the increased presence of hydrophobic residues. Peptides of porin proteins showed limited difference in their hydrophobicity and ionization efficiency (Fig S22).

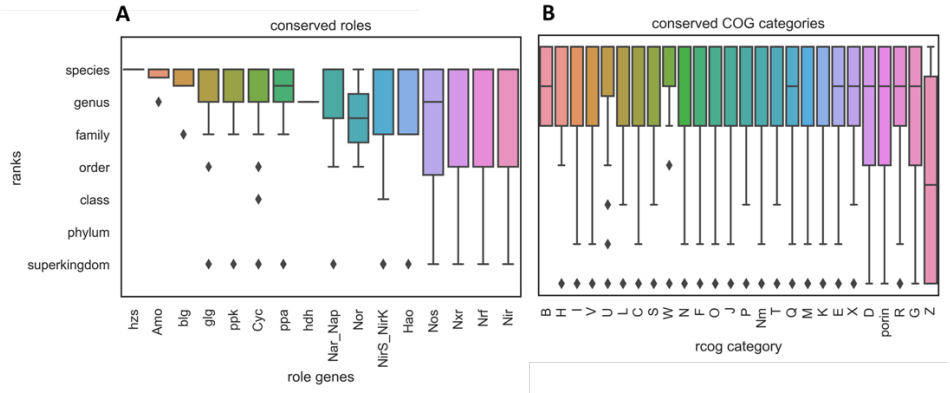


Fig. S 1 Conservedness of nutrient removal genes (A) and COG categories (B).

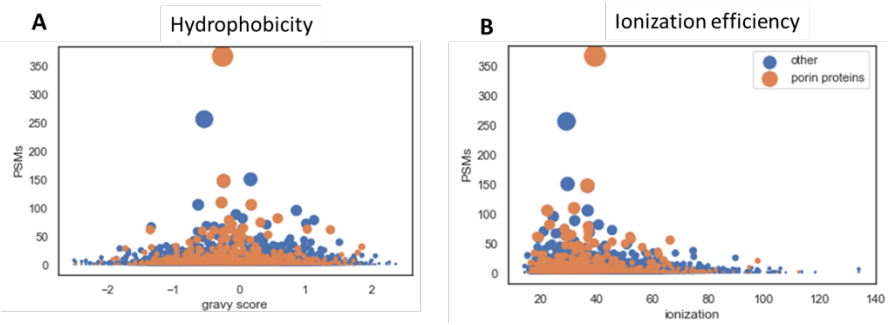


Fig. S2.2 Peptide hydrophobicity (A) and ionization efficiency (B) of porin proteins compared to other proteins, versus the number of peptide spectrum matches (PSMs). Spots are scaled to PSMs.

2.6.9 Simpelveld Binning

Initial binning of the Simpelveld metagenome was done by external collaborator Dennis Grouzdev, using the following methods:

Reads were assembled for all samples using metaSPAdes v3.14.0 (Nurk et al. 2017) at the default settings. Metagenome binning was performed using three different binning algorithms: CONCOCT v1.1 (Alneberg et al. 2014), MaxBin 2.0 v2.2.4 (Wu, Simmons and Singer 2016) and MetaBAT 2 v1.7 (Kang et al. 2019). The three bin sets were supplied to DAS Tool v1.0 (Sieber et al. 2018) for consensus binning to obtain the final optimized bins. Metagenome assembled genomes (MAGs) were assessed for completeness and contamination using lineage-specific marker genes and default parameters in CheckM v1.0.1271 (Parks et al. 2015). RefineM v. 0.0.2450 was used to remove contamination based on taxonomic assignments. The quality metrics were assessed using the QUAST v5.0.1 (Gurevich et al. 2013). The GTDB-Tk v1.4.0 (Chaumeil et al. 2020) 'classify_wf' command was used to get the taxonomic assignment of reconstructed MAGs using the GTDB r95 database (Parks et al. 2018). Prodigal v2.6.3 (Hyatt et al. 2010) was used as a gene caller to identify open reading frames, whereas genes were functionally annotated using GhostKOALA (Kanehisa et al. 2016)

The selected top genome bins did not contain several of the well-studied aerobic granular sludge organisms, such as Accumulibacter and Nitrosomonas, which prompted the investigation of a more comprehensive contig-based annotation.

Table S2.7 Das Tool optimized bins of Simpelveld plant.

Bin	Classification GTDB	Coverage	Completeness
SP005	o__Bryobacterales;f__Bryobacteraceae;g__UBA690;s__	15.5	95.6
SP029	o__Pyrinomonadales;f__Pyrinomonadaceae;g__OLB17;s__	7.4	85.3
SP039	o__Holophagales;f__Holophagaceae;g__;s__	15.4	98.4
SP021	o__UBA5704;f__;g__;s__	51.5	92.5
SP015	o__Vicinamibacterales;f__UBA2999;g__;s__	51.2	90.3
SP022	o__Vicinamibacterales;f__UBA2999;g__;s__	17.6	80.7
SP032	o__Microtrichales;f__Microtrichaceae;g__UBA11034;s__	13.6	93.6
SP006	o__Actinomycetales;f__Dermatophilaceae;g__GCA-2748155;s__	58.8	81.2
SP030	o__Actinomycetales;f__Dermatophilaceae;g__GCA-2748155;s__	101.6	88.1
SP014	o__AKYH767-A;f__2013-40CM-41-45;g__;s__	9.5	76.2
SP024	o__Bacteroidales;f__vadinHA17;g__LD21;s__	8	86.3
SP017	o__Chitinophagales;f__BACL12;g__UBA7236;s__	100.6	92.2
SP011	o__Chitinophagales;f__Chitinophagaceae;g__Ferruginibacter;s__	30.8	96.0
SP016	o__Chitinophagales;f__Chitinophagaceae;g__Ferruginibacter;s__Ferruginibacter sp002400445	22.5	96.3
SP001	o__Chitinophagales;f__Chitinophagaceae;g__JJ008;s__	30.3	89.5
SP012	o__Chitinophagales;f__Chitinophagaceae;g__JJ008;s__	19	49.2
SP002	o__Chitinophagales;f__Saprospiraceae;g__;s__	85.8	89.5
SP031	o__Chitinophagales;f__Saprospiraceae;g__;s__	9	42.5
SP013	o__Chitinophagales;f__Saprospiraceae;g__UBA3362;s__	16.3	94.9
SP020	o__Flavobacteriales;f__PHOS-HE28;g__PHOS-HE28;s__	15.5	65.3
SP028	o__Flavobacteriales;f__PHOS-HE28;g__PHOS-HE28;s__	19.7	86.3
SP038	o__Flavobacteriales;f__PHOS-HE28;g__PHOS-HE28;s__	20.6	87.7
SP041	o__Flavobacteriales;f__PHOS-HE28;g__PHOS-HE28;s__	7.7	46.3
SP009	o__Ignavibacteriales;f__Ignavibacteriaceae;g__Ignavibacterium;s__	31	76.4
SP008	o__SJA-28;f__OLB5;g__OLB5;s__	30.5	71.6
SP010	o__Anaerolineales;f__envOPS12;g__;s__	14.1	71.6
SP018	o__Obscuribacterales;f__Obscuribacteraceae;g__Obscuribacter;s__	5.9	61.9
SP035	o__F11;f__FEN-1173;g__;s__	11.1	95.5
SP033	o__Haliangiales;f__Haliangiaceae;g__;s__	15.3	99.0
SP025	o__Nitrospirales;f__Nitrospiraceae;g__Nitrospira_A;s__	13.9	95.3
SP042	o__Rhizobiales;f__Anderseniellaceae;g__PALSA-927;s__	18.4	97.5
SP007	o__Sphingomonadales;f__Sphingomonadaceae;g__Ga0077559;s_	31.6	97.7
SP040	o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingobium;s__	13.4	81.9
SP026	o__UBA1301;f__UBA1301;g__UBA6038;s__	14.5	85.1
SP037	o__Burkholderiales;f__;g__;s__	14.4	20.6
SP019	o__Burkholderiales;f__Gallionellaceae;g__UBA7399;s__	39.5	86.8
SP023	o__Burkholderiales;f__Rhodocyclaceae;g__Fen-999;s__	47.1	81.0
SP027	o__Burkholderiales;f__Rhodocyclaceae;g__Propionivibrio;s__	58.5	96.1
SP036	o__Competibacterales;f__Competibacteraceae;g__Competibacter;s__	17.2	89.8
SP004	o__Competibacterales;f__Competibacteraceae;g__Competibacter;s__Competibacter denitrificans	25.6	93.3
SP003	o__Pseudomonadales;f__UBA5518;g__UBA5518;s__	10.9	80.1
SP034	o__Pedosphaerales;f__Pedosphaeraceae;g__;s__	10.6	88.2

2.6.9 Supplemental References

- Alneberg, Johannes, et al. "Binning metagenomic contigs by coverage and composition." *Nature methods* 11.11 (2014): 1144-1146.
- Barberán, Albert, et al. "Hiding in plain sight: mining bacterial species records for phenotypic trait information." MSphere 2.4 (2017): e00237-17.
- Bokulich, Nicholas A., et al. "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin." Microbiome 6.1 (2018): 1-17.
- Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. "Fast and sensitive protein alignment using DIAMOND." Nature methods 12.1 (2015): 59-60.
- Chaumeil, Pierre-Alain, et al. "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." (2020): 1925-1927.
- Chen, Wen-Ming, et al. "Aquabacterium limnoticum sp. nov., isolated from a freshwater spring." International journal of systematic and evolutionary microbiology 62.Pt_3 (2012): 698-704.
- Choi, Dong Hee, et al. "Complete genome sequence of Novosphingobium pentaromativorans US6-1 T." Standards in genomic sciences 10.1 (2015): 1-8.
- De Luca, Gilles, et al. "The cyst-dividing bacterium Ramlibacter tataouinensis TTB310 genome reveals a well-stocked toolbox for adaptation to a desert environment." PLoS One 6.9 (2011): e23784.
- Ehrich, Silke, et al. "A new obligately chemolithoautotrophic, nitrite-oxidizing bacterium, Nitrospira moscoviensis sp. nov. and its phylogenetic relationship." Archives of Microbiology 164.1 (1995): 16-23.
- Felföldi, Tamás, et al. "Quisquiliibacterium transsilvanicum gen. nov., sp. nov., a novel betaproteobacterium isolated from a waste-treating bioreactor." International journal of systematic and evolutionary microbiology 67.11 (2017): 4742-4746.
- Gurevich, Alexey, et al. "QUAST: quality assessment tool for genome assemblies." *Bioinformatics* 29.8 (2013): 1072-1075.
- Hanada, Satoshi, and Yuji Sekiguchi. "The phylum gemmatimonadetes." The prokaryotes 11 (2014): 677-681.
- Hiraishi, Akira, Yasuo Hoshino, and Toshio Satoh. "Rhodoferax fermentans gen. nov., sp. nov., a phototrophic purple nonsulfur bacterium previously referred to as the "Rhodocyclus gelatinosus-like" group." Archives of microbiology 155.4 (1991): 330-336.
- Hyatt, Doug, et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC bioinformatics 11.1 (2010): 1-11.
- Inzana, Thomas J. "Miscellaneous Glucose-Nonfermenting Gram-Negative Bacteria." Diagnostic Procedure in Veterinary Bacteriology and Mycology. Academic Press, 1990. 165-176.
- Kanehisa, Minoru, et al. "KEGG as a reference resource for gene and protein annotation." Nucleic acids research 44.D1 (2016): D457-D462.
- Kang, Dongwan D., et al. "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies." *PeerJ* 7 (2019): e7359.
- Kim, Jeong Myeong, et al. "Analysis of the fine-scale population structure of "Candidatus Accumulibacter phosphatis" in enhanced biological phosphorus removal sludge, using fluorescence in situ hybridization and flow cytometric sorting." Applied and environmental microbiology 76.12 (2010): 3825-3835.
- Kitzinger, Katharina, et al. "Characterization of the first "Candidatus Nitrotoga" isolate reveals metabolic versatility and separate evolution of widespread nitrite-oxidizing bacteria." MBio 9.4 (2018): e01186-18.
- Kochetkova, Tatiana V., et al. "Tepidiforma bonchosmolovskayae gen. nov., sp. nov., a moderately thermophilic Chloroflexi bacterium from a Chukotka hot spring (Arctic, Russia), representing a novel class, Tepidiformia, which includes the previously uncultivated lineage OLB14." International journal of systematic and evolutionary microbiology 70.2 (2020): 1192-1202.
- Li, Xin, et al. "Aestuariivirga litoralis gen. nov., sp. nov., a proteobacterium isolated from a water sample, and proposal of Aestuariivirgaceae fam. nov." International journal of systematic and evolutionary microbiology 69.2 (2019): 299-306.

- Liu, Zhenfeng, et al. "Complete genome of Ignavibacterium album, a metabolically versatile, flagellated, facultative anaerobe from the phylum Chlorobi." Frontiers in microbiology 3 (2012): 185.
- Liu, Zhenfeng, et al. "Complete genome of Ignavibacterium album, a metabolically versatile, flagellated, facultative anaerobe from the phylum Chlorobi." Frontiers in microbiology 3 (2012): 185.
- Marin, I. A. D. R., and D. Ruiz Arahal. "The family beijerinckiaceae." The prokaryotes 115 (2014): 33.
- McIlroy, Simon J., et al. "Re-appraisal of the phylogeny and fluorescence in situ hybridization probes for the analysis of the C ompetibacteraceae in wastewater treatment systems." Environmental microbiology reports 7.2 (2015): 166-174.
- McIlroy, Simon Jon, and Per Halkjær Nielsen. "The family saprospiraceae." The prokaryotes: Other major lineages of bacteria and the Archaea. Springer Science+ Business Media, 2014. 863-889.
- von Meijenfeldt, FA Bastiaan, et al. "Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT." Genome biology 20.1 (2019): 1-14.
- Nagashima, Sakiko, et al. "Complete genome sequence of phototrophic betaproteobacterium Rubrivivax gelatinosus IL144." (2012): 3541-3542.
- Nurk, Sergey, et al. "metaSPAdes: a new versatile metagenomic assembler." Genome research 27.5 (2017): 824-834.
- Oren, Aharon. "The family rhodocyclaceae." The prokaryotes 975 (2014).
- Qin, Dan, et al. "Sphingobium estronivorans sp. nov. and Sphingobium bisphenolivorans sp. nov., isolated from a wastewater treatment plant." International journal of systematic and evolutionary microbiology 70.3 (2020): 1822-1829.
- Parks, Donovan H., et al. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." *Genome research* 25.7 (2015): 1043-105
- Parks, Donovan H., et al. "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life." *Nature biotechnology* 36.10 (2018): 996-1004.
- Reinhold-Hurek, B. A. R. B. A. R. A., and T. H. O. M. A. S. Hurek. "The genera Azoarcus, Azovibrio, Azospira and Azonexus." Prokaryotes 5 (2006): 873-891.
- Robeson, Michael S., et al. "RESCRIPt: Reproducible sequence taxonomy reference database management." PLoS computational biology 17.11 (2021): e1009581.
- Siddiqi, Muhammad Zubair, Zubair Aslam, and Wan-Taek Im. "Arachidicoccus ginsenosidivorans sp. nov., with ginsenoside-converting activity isolated from ginseng cultivating soil." International journal of systematic and evolutionary microbiology 67.4 (2017): 1005-1010.
- Soliman, Moomen, and Ahmed Eldyasti. "Ammonia-Oxidizing Bacteria (AOB): opportunities and applications—a review." Reviews in Environmental Science and Bio/Technology 17.2 (2018): 285-321.
- Somasegaran, Padma, and Heinz J. Hoben. Handbook for rhizobia: methods in legume-Rhizobium technology. Springer Science & Business Media, 2012.
- Stoddard, Steven F., et al. "rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development." Nucleic acids research 43.D1 (2015): D593-D598.
- Tamura Tomohiko, Suzuki Ken-Ichiro. (2014) "The Suborder Kineosporiineae." The prokaryotes 975 (2014).
- van Niftrik, Laura, et al. "Linking ultrastructure and function in four genera of anaerobic ammonium-oxidizing bacteria: cell plan, glycogen storage, and localization of cytochrome c proteins." Journal of bacteriology 190.2 (2008): 708-717.
- Wu, Sitao, et al. "WebMGA: a customizable web server for fast metagenomic sequence analysis." BMC genomics 12.1 (2011): 1-9.
- Wu, Yu-Wei, Blake A. Simmons, and Steven W. Singer. "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets." *Bioinformatics* 32.4 (2016): 605-607.
- Yamada, Takeshi, et al. "Anaerolinea thermolimosa sp. nov., Levilinea saccharolytica gen. nov., sp. nov. and Leptolinea tardivitalis gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes Anaerolineae classis nov. and Caldilineae classis nov. in the bacterial phylum Chloroflexi." International Journal of Systematic and Evolutionary Microbiology 56.6 (2006): 1331-1340.

3

Database-independent de novo metaproteomics of complex microbial communities

This chapter is based on:

Kleikamp, H. B. C., Pronk, M., Tugui, C., Guedes, L., Lin, Y. M., Loosdrecht, M. C. M. Van, & Pabst, M. (**2021**). Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems*, (12), 375–383.

3.1 Introduction

State-of-the-art approaches for analyzing the composition of microbial communities are based on *in situ* staining, 16S ribosomal RNA sequencing, or whole-genome shotgun-based approaches. Moreover, metatranscriptomics provides additional gene activity information, but unfortunately, mRNA levels often only poorly correlate with actual protein abundances (Maier et al., 2009). Therefore, those approaches do not directly assess the actual phenotype of a community, and the actively expressed pathways responsible for metabolic conversions remain elusive (Martin and Uroz, 2016). On the other hand, metaproteomics targets the functional parts—the proteins—of a community directly, and therefore provides insights into the community phenotype. Furthermore, because proteins make up the bulk mass of a cell, metaproteomics also estimates the contribution of individual community members to the community biomass (Kleiner et al., 2017).

In recent years, metaproteomics has gained substantial momentum with the development of high-resolution proteomics workstations and the establishment of next-generation sequencing (NGS) technologies, which provide affordable high-quality (protein) sequence databases from complete communities (Wilmes and Bond, 2006). Classical metaproteomics approaches employ peptide-spectrum-matching algorithms used for subsequent protein and species identification. The quality and completeness of the employed databases are therefore of utmost importance (Timmins-Schiffman et al., 2017; Xiao et al., 2018). A complete database covers the genetic potential of all community members and may contain hundreds of thousands of sequences. Alternatively, comprehensive (and even larger) public sequence databases such as NCBI, UniProtKB/Swiss-Prot, or GenBank may be accessed (in addition) (Xiao et al., 2018), which, however, require advanced focusing/filtering strategies to manage computational efforts (Heyer et al., 2017; Muth et al., 2015, 2016; Potgieter et al., 2019). Very large protein sequence databases challenge the common ‘‘peptide-spectrum-matching’’ algorithms and associated statistical parameters, which have been historically developed for single-species proteomics.

Consequently, conventional metaproteomics experiments can be compromised in regard to sensitivity, accuracy, and throughput (Heyer et al., 2017; Muth et al., 2015; Timmins-Schiffman et al., 2017). Moreover, database-matching inherently biases the outcome of a metaproteomics measurement toward the (constructed) protein sequence database.

A database-independent approach, such as *de novo* peptide sequencing, that directly annotates mass spectrometric fragmentation spectra with amino acid sequences overcomes the above-mentioned database-related limitations. Ultimately, the generation of the peptide sequence lists from the mass spectrometric raw data can be regarded as inherently unbiased (Muth et al., 2016). Following a successful *de novo* sequencing, the sequence lists only require retrieving taxonomic and functional annotations from comprehensive taxonomic databases using efficient ‘‘text-search’’ tools. Thereby, *de novo* sequencing also avoids the loss of taxonomic and functional information from community members not covered by the database. Those signals (not covered by the target database) can be further matched to related species through sequence homology searching approaches (Ma and Johnson, 2012). Homology searching further increases proteome coverage, by annotating also ‘‘partially correct’’ sequences (sequence tags), which are common ‘‘by-products’’ of the *de novo* sequencing process (Ma and Johnson, 2012).

Moreover, *de novo* sequencing may serve as a direct measure of the proportion of unsequenced members in a community. In a similar manner, the usefulness of *de novo* sequencing for evaluating the target sequence database completeness, or ‘‘suitability,’’ has been demonstrated only recently (Johnson et al., 2020). On the other hand, *de novo* peptide sequencing strongly depends on high-quality mass spectrometric data and efficient sequence annotation tools. Therefore, *de novo* sequencing commonly provides fewer spectral identifications when compared with database search approaches (Medzihradszky and Chalkley, 2015). Nevertheless, whether *de novo* sequencing provides sufficient qualitative and quantitative information for (quantitative) metaproteomic applications has not been effectively established to date.

Over the past years, several high-performance *de novo* sequencing algorithms have been introduced (Tran et al., 2019; Ma et al., 2003; Behsaz et al., 2020), and some have also been proposed for taxonomic profiling applications (Lee et al., 2018; Mooradian et al., 2019). In addition, a number of advanced web-based services that support taxonomic and functional analyses of metaproteomic protein and peptide sequences have been introduced only recently (Mesuere et al., 2015; Boekel et al., 2015; Zhang et al., 2016; Singh et al., 2019; Riffle et al., 2017). In this study, we introduce and evaluate a newly established *de novo* metaproteomics workflow for its quantitative performance and taxonomic resolution using synthetic and natural environmental community data. Furthermore, we introduce a new validation strategy and demonstrate how to establish the actual content of individual community members within community proteomics data. The new pipeline (NovoBridge) efficiently bridges individual components from *de novo* sequencing, automated annotation of sequences with taxonomies, a new validation procedure, and the provision of an output summary.

3.2 Results

The presented metaproteomics pipeline employs conventional high-resolution shotgun proteomics data in which fragmentation spectra are subsequently translated into peptide sequence lists by *de novo* sequencing. The lists are then submitted by programmed access to the (public) Unipept database to retrieve taxonomic and metabolic information (Singh et al., 2019). Annotations are then processed by the established pipeline, which includes grouping into taxonomic branches and translation of enzyme commission numbers into KEGG pathways. We investigated fundamental aspects and evaluated the performance of the established workflow using synthetic and natural microbial communities.

Taxonomic resolution

The first question concerns the taxonomic resolution that can be achieved when matching de novo peptide sequences against particularly large taxonomy databases to retrieve taxonomic and functional annotations. A large number of peptide sequences is common to several taxa and can therefore only be unique to a certain taxonomic ranking.

Hence, the number of unique peptide sequences decreases from higher to lower taxonomic rankings. For example, because of the relatedness between taxa, there will be many more peptide sequences unique only to the phylum level compared with the more distinguished genus or species levels.

For our study, we aimed to retrieve taxonomic information from the Unipept database, which contains processed peptide sequences pre-allocated with taxonomic and functional annotations derived from the Uniprot database, using NCBI taxonomy (Mesuere et al., 2012, 2016). The Unipept ranking uses the hierarchical structure of the NCBI taxonomy for which consensus taxa have been determined using the lowest common ancestor approach (Mesuere et al., 2012). To test the Unipept database for the achievable taxonomic resolution, we generated *in silico* peptide sequences from >1,000 species retrieved from the NCBI reference sequences database (www.ncbi.nlm.nih.gov/refseq/). This provided for approximately 90% of all peptide sequences taxonomic annotations, but as expected, showed a steady decrease in the number of assigned peptides from higher to lower taxonomic rankings (= ‘‘drop-off rate’’), with a particularly large drop between genus and species levels (Figure 1C). It is worth noting that deviations from this ‘‘drop-off rate’’ can be observed for species from highly sampled taxa and species with inconsistent taxonomic classifications. This impacts not only the quantitative performance but may also limit the taxonomic resolution, because a certain number of peptides is required for the identification of a respective taxon.

Furthermore, because there is no complete taxonomy database available, there is always a high likelihood of ‘‘unsequenced’’ community members—those that are not in the taxonomy database—being present in the community. Those retrieve annotations through related species mostly at higher taxonomic rankings and will therefore provide only a comparatively low taxonomic resolution. A quantitative analysis should therefore aim to investigate the ‘‘drop-off rates’’ for individual taxonomic branches, in order to flag poorly quantitative traits. For this, *in silico* peptidomes may serve as highly useful comparators to establish the actual content of a member within the community proteomics data.

A validation procedure

De novo sequencing commonly generates a fraction of only partially correct peptide sequences. This raises the question of whether those incomplete sequences lead to false-positive assignments, which bias the taxonomic representation of the community.

As a measure of confidence for *de novo*-established peptide sequences, the software platform PEAKS provides the average local confidence (ALC) score, and DeepNovo, the p score (Ma et al., 2003; Tran et al., 2019, 2017). Although these parameters are useful for ranking *de novo* sequences based on their quality, an estimate on the actual number of incorrect sequences is not provided.

Consequently, additional measures are required to give confidence in the taxonomic representation achieved by *de novo* generated sequences. A recently proposed solution employs a taxonomic database containing sequences not only in correct but also in reverse order. This strategy enables to make use of the widely employed target/decoy approach (Mooradian et al., 2019). However, database volumes are thereby duplicated, and considering only single taxonomic points does not allow performance of a quantitative investigation of the taxonomic profiles.

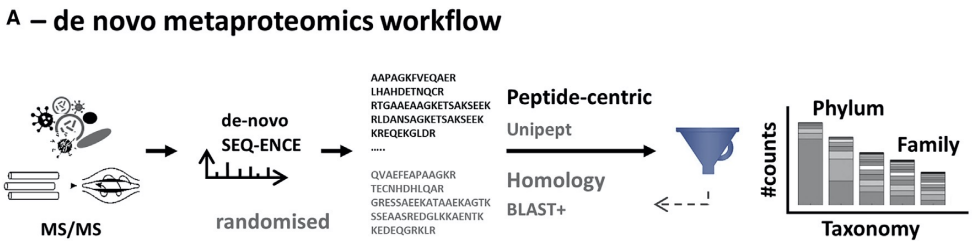
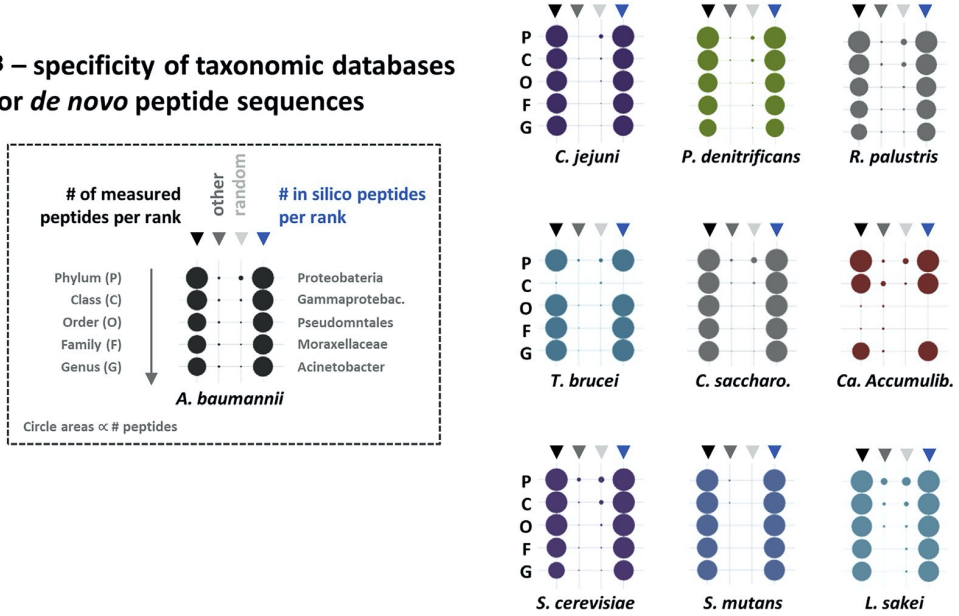
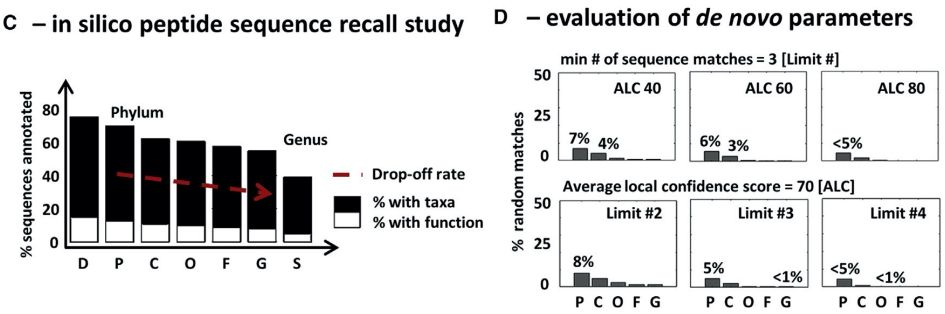


Figure 1. Overview of the de novo metaproteomics workflow and an evaluation of fundamental characteristics (A) Shotgun metaproteomics workflow. Shotgun metaproteomic raw data from microbial communities are *de novo* sequenced and processed through the established pipeline as ''correct'' and randomized sequences. The peptide-centric approach accesses Unipept (Mesuere et al., 2016) to obtain taxonomic and functional annotations. Further processing includes grouping into taxonomic branches and translation of functional annotations into KEGG pathways. High quality unmatched sequences are further made accessible for homology search approaches such as BLAST+.

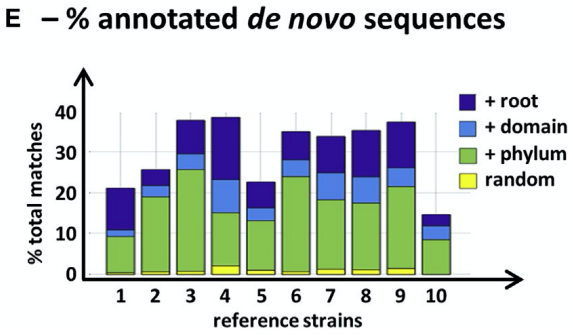


(B) Specificity of taxonomy databases for *de novo* peptide sequence lists. Shotgun proteomic data from pure reference strains were *de novo* sequenced and processed through the established *de novo* metaproteomics pipeline to retrieve taxonomic annotations. The annotated sequences were then grouped into taxonomic lineages (phylum, class, order, family, and genus) and represented as circle graphs. The circle areas correlate to the normalized sequence counts of the respective taxonomic rank. Every reference strain is represented by four circle lanes: black triangle arrow, ‘‘# of measured peptides per rank,’’ which counts the number of peptide sequences annotated to the lineage of the target strain, e.g., *A. baumannii*; gray triangle arrow, ‘‘other,’’ which counts the number of peptide sequences annotated to other taxonomic lineages than the target strain; light gray triangle arrow, ‘‘random,’’ which counts the number of randomized peptide sequences which received a taxonomic annotation; blue triangle arrow, ‘‘# of *in silico* peptides per rank,’’ which counts the number of *in silico* target strain sequences for every rank. The experiment confirms that erroneous or only partially correct *de novo* sequences only insignificantly interfere with the taxonomic representation of the metaproteomic sample. Furthermore, the low number of ‘‘other’’ strain assignments confirmed the purity of the selected reference strain samples. Except for the *in silico* experiments, the averages of duplicate analyses are shown.



(C) *In silico* proteome recall study. The bar graph shows the average number of *in silico* peptide sequences which retrieved taxonomic or ''enzyme commission number'' annotations. The *in silico* peptide sequences were generated from a large number of proteomes (>1,000, retrieved from the NCBI reference proteome database). The individual taxonomic rankings domain (D), phylum (P), class (C), order (O), family (F), genus (G), and species (S) are shown as separate bars. Approximately 90% of the peptides obtained taxonomic annotations (black bars), and 10%–20% retrieved additional functional annotations (enzyme commission numbers, white bars). The number of sequence annotations per taxon showed a steady decrease from the phylum to the genus level (''drop-off'' rate, red arrow).

(D) Evaluation of *de novo* sequence quality parameters. The bar graph shows the average number of random sequences which obtained a taxonomic annotation, when considering different quality parameter thresholds. The randomized sequences were generated from the ''correct'' reference strains *de novo* sequence lists (excluding *T. brucei* and *Ca*. Accumulibacter). The quality parameter thresholds evaluated were the average local confidence score (ALC, PEAKS platform) and frequency limits (# of peptide sequences observed for an individual taxonomic identifier). ALCs below 60 and frequency limits <3 increased the percentage of random sequence annotations to >5%. Therefore, an ALC of 70 and a minimum of 3 sequence annotations per taxon were set as default thresholds for the experiments in this study.



(E) The percentage of annotated *de novo* sequences. The bar graph outlines the percentage of *de novo* sequences submitted to Unipept, which retrieved taxonomic annotations. The bars (1–10) represent the strains shown in Figure 1B (*A*. *baumannii*, top of the image; *L*. *sakei*, bottom right of the image). The blue bars represent all annotations including ''root'' level, which are sequences common to all domains of life; the light blue bars represent annotations assigned to domain level and lower; and the green bars show annotations assigned to phylum level and lower. The yellow bars indicate the average number of random annotations at the lower taxonomic rankings. The observed differences in the degree of sequence annotations are supposedly a consequence of differences in employed sample preparation protocols and instrumental setups. Therefore, although the percentage of assigned sequences are difficult to compare between different laboratories, those parameters are likely to provide a useful quality parameter when operations are standardized within one laboratory.

Therefore, we aimed not to randomize the target database sequences but to randomize the peptide query sequences instead. To qualify this approach, we processed proteomics data from pure reference species, once in correct order, and once after peptide sequence randomization. The randomized sequences retrieved a surprisingly large number of taxonomic annotations at the root (>20%) and super kingdom levels (>10%) but were consistently low for the lower taxonomic rankings (Figures 1B and 1D). Only small proportions of other taxa were observed, mostly related to culturing and sample preparation conditions, or the samples themselves (such as *virus L-A related proteins* for the yeast *S*. *cerevisiae*). Several of those unexpected matches were only identified at certain taxonomic levels, which underlines the importance of measuring the taxonomic profiles across several taxonomic rankings (e.g., from phylum, family, or genus level) rather than single taxonomic points (e.g., only genus level) (Figure 1B; Table S5). Next, we constructed the theoretical drop-off rates using the reference proteomes of the test strains to investigate for ''hidden'' side populations, not covered by the taxonomic database. This, however, showed that the theoretical and the observed drop-off rates were very comparable, which confirmed the purity of the selected reference strains.

In summary, using the pure reference strain samples and the sequence randomization strategy, we could demonstrate that *de novo* sequence lists provide only small numbers of erroneous assignments at lower taxonomic rankings (phylum and genus).

Quantitative community profiling

Finally, we investigated the quantitative aspect when measuring more complex communities. Kleiner et al. only recently demonstrated the usefulness of metaproteomics for estimating species biomass contributions (Kleiner et al. 2017). Thereby, the authors generated highly useful metaproteomic reference data from synthetic communities consisting of species with ‘‘equal protein’’ and ‘‘equal cell’’ content. We *de novo* sequenced the publicly available raw data from both synthetic communities and subjected the obtained sequence lists to our data-processing pipeline. By employing the abovementioned multi-point taxonomic evaluation, we achieved a particularly good quantitative representation of the community as shown for the ‘‘equal protein’’ community (phylum and family) in Figure 2A. The 17 genus-level identifiers provided a comparably good correlation, although 3 strains did not provide sufficient unique peptides at this lower level. The same good species abundance correlation was achieved when analyzing another dataset of the same ‘‘equal cell’’ community, thereby also comparing 2 different *de novo* sequencing platforms, PEAKS and DeepNovo (Figure 3). Verification of parameters such as ALC scores and mass error, including species abundance correlations, obtained for the ‘‘equal cell’’ synthetic community are shown in Figures S1–S4.

Furthermore, we aimed to apply the *de novo* pipeline to datasets from two natural communities. Thereby, we first processed a publicly available metaproteomic dataset published by Mikan et al., representing microbiomes sampled from the Bering Sea (Mikan et al., 2020). We generated peptide sequences once using *de novo* sequencing and once using peptide-spectrum-matching employing the metagenomics constructed database published by the authors. Thereby, the taxonomic profiles between both approaches were highly comparable (Figure S5A), where only some of the very-low-abundance members were not resolved by the *de novo* approach. However, the metaproteomics approaches indicate a stronger contribution of the Alphaproteobacteria, compared with the 16S rRNA sequencing data published by the authors (Figure S5A).

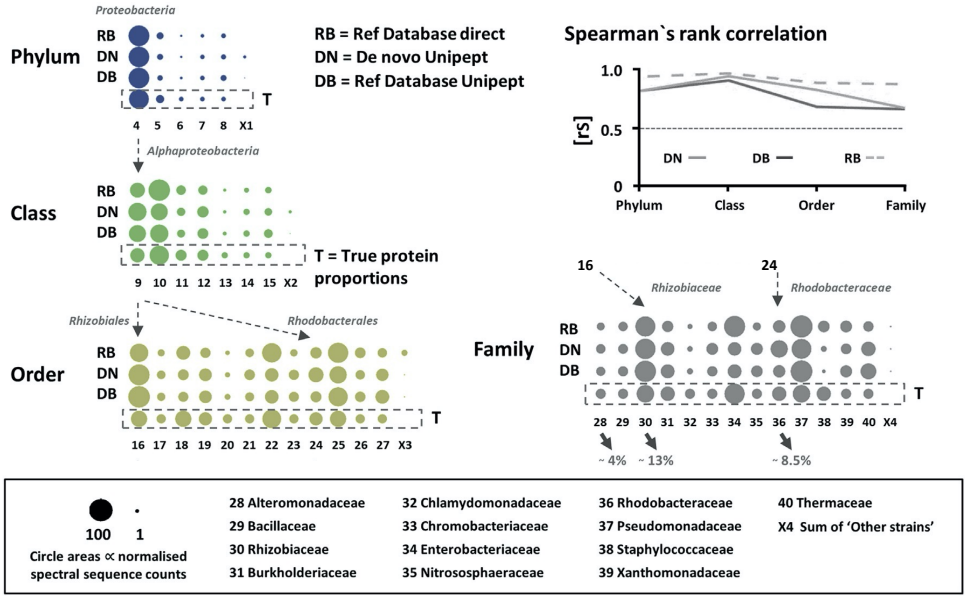
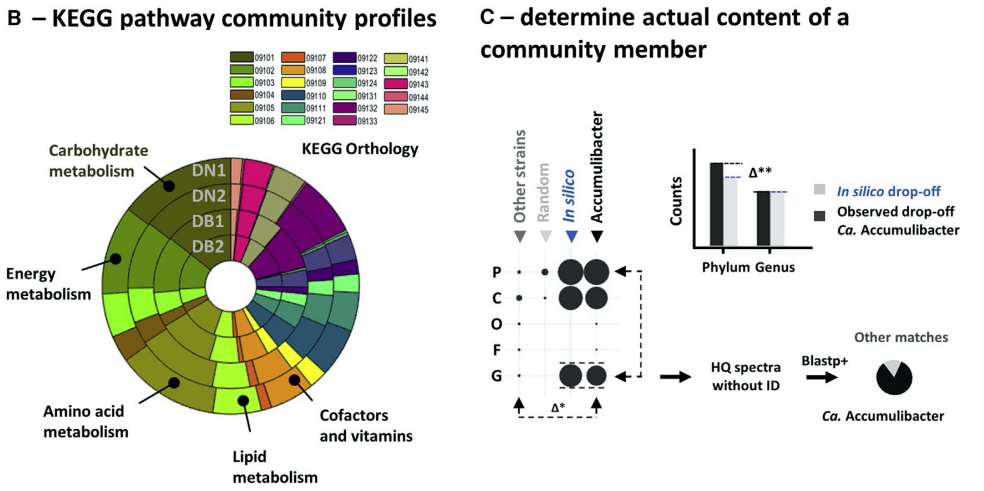


Figure 2. Quantitative taxonomic profiling of microbial communities

(A) Analyzing the community composition by *de novo* metaproteomics. Proteomics data from a synthetic community, as established by Kleiner et al. (Kleiner et al., 2017), were used to evaluate the quantitativeness of the established *de novo* metaproteomics workflow. For this, the raw data were once *de novo* sequenced and once analyzed using the constructed target database published by the authors. The taxonomic rankings from phylum and family are represented as circle graphs. Thereby, rows annotated with: ''DN'' show the protein abundances of each taxon using the *de novo* sequences; ''DB'' show the protein abundances obtained for each taxon using the sequences established by database matching; ''RB'' show the protein abundances obtained after grouping the taxon annotated database matched peptides directly; ''T'' shows the theoretical (true) protein abundances for each taxon. The circle areas correlate to the normalized spectral sequence counts of the respective taxon. All community members show abundance profiles, which strongly correlate to the expected/true (T) species protein abundances. The taxonomic lineages of *Rhizobiaceae* and *Rhodobacteriaceae* are outlined with arrows for exemplification purposes. Those account for approximately 13% and 8.5% of the total community protein content, respectively. Shown is the average of duplicate analyses. The taxonomic identifiers with the numbers 1–27 represent: (1) Bacteria, (2) Eukaryotes, (3) Archaea, (4) Proteobacteria, (5) Firmicutes, (6) Chlorophyta, (7) Thaumarchaeota, (8) Deinococcus-Thermus, (9) Alphaproteobacteria, (10) Gammaproteobacteria, (11) Bacilli, (12) Betaproteobacteria, (13) Chlorophyceae, (14) Nitrososphaeria, (15) Deinococci, (16) Rhizobiales, (17) Alteromonadales, (18) Bacillales, (19) Burkholderiales, (20) Chlamydomonadales,

(21) Neisseriales, (22) Enterobacterales, (23) Nitrososphaerales, (24) Rhodobacterales, (25) Pseudomonadales, (26) Xanthomonadales, and (27) Thermales. The lower graph shows the Spearman's rank correlation between the peptide sequence lists (obtained by DB matching, ''DB,'' and DN sequencing, ''DN'') and the expected protein abundance ratios (''T''). Overall, the correlation to the expected protein abundances was strong for both sequence list approaches (e.g., >0.82 for the DN sequence lists from phylum to order, and 0.67 only at the family level, considering all taxonomic identifiers, including ''x''). The very comparable correlation between the *de novo* and the database-matching generated sequence lists confirms the high quality of the *de novo* established peptide sequences. The profiles obtained after directly grouping the database spectrum-matched peptides show, as expected, a slightly better correlation. Therefore, the difference between these profiles and the profiles obtained by the sequence lists shows the impact of the database, such as sequence coverage and volumes. The database used for the database-matching experiments consisted of the reference proteomes of the strains present in the synthetic community, and therefore represented a comparatively focused, complete, and non-redundant database. Moreover, the very large and generic Unipept peptide sequence database, used to annotate the peptide sequence lists, contained only closely related taxa for some strains (e.g., for *Roseobacter sp*. AK199).



(B) KEGG pathway community profiles. The graphs compare profiles for the major KEGG categories ''metabolism'' and ''genetic information processing,'' obtained by sequence lists from *de novo* (outer circles) or peptide-spectrum matching approaches (inner circles) of the ''equal protein'' community. Both *de novo* (DN) and the database (DB) sequences provide very comparable profiles. Nevertheless, since peptide sequence lists are compared against a large genomic space, sequences can be matched to several enzymes or different pathways, which may inflate functional annotations. See also Figure S6. (C) Establishing the actual contribution of community members. The *de novo* metaproteomic analysis of a *Ca*. Accumulibacter enrichment culture suggests a very high enrichment (>95%, ''other'' versus ''Accumulibacter,''

Δ*). Furthermore, comparing the experimental with the *in silico* ‘‘drop-off’’ rates, shows only a discrepancy of approximately 17% (small bar graphs, Δ**). To investigate for potential ‘‘hidden’’ members not covered by the taxonomic database, the high- quality (HQ) unmatched sequences (top 20% fraction based on ALC scores) were analyzed using BLAST+ for homolog sequences. Thereby, more than 80% of the newly retrieved annotations were again assigned to *Ca*. Accumulibacter (small pie chart), confirming the content estimated after drop-off correction. The individual circle graph columns represent: black triangle arrow, ‘‘# of measured peptides per rank,’’ which counts the peptide sequences annotated to the lineage of *Ca*. Accumulibacter; blue triangle arrow, ‘‘# of *in silico* peptides per rank,’’ which represents the number of *Ca*. Accumulibacter in silico sequences per taxon; light gray triangle arrow, ‘‘random,’’ which counts the number of randomized peptide sequences which received a taxonomic annotation; gray triangle arrow, ‘‘other,’’ which counts the number of measured peptide sequences annotated to other taxonomic lineages than *Ca*. Accumulibacter. The circle areas correspond to spectral sequence (peptide) counts for the respective taxonomic ranking.

The second dataset was derived from the metaproteomic analysis of a wastewater treatment plant community, published by Hansen et al. (Hansen et al., 2014). The authors investigated different protein extraction procedures to maximize extraction reducibility and community coverage. We therefore analyzed the mass spectrometric raw data obtained from the most efficient protocol through the established metaproteomics pipeline. Furthermore, the observed community profiles appeared very comparable between the *de novo*-generated peptide sequences and the (metagenomics) database search peptide sequence matches. Again, differences were only observed in the very low-abundance community members. Moreover, the *de novo* phylum-level profile of the *de novo* dataset was found highly comparable to qFISH data established from the same community at an earlier time point (Albertsen et al., 2012) (Figure S6).

Database incompleteness and spectral volume dependency

To evaluate the impact of incomplete databases, we simulated scenarios where the taxonomies present in the microbiomes are not covered by the taxonomic database (e.g., Unipept). As a consequence, measured peptide sequences from those taxonomic identifiers would only match to related taxa (potentially) present in the same database. Interestingly, when all species or genera (present in the synthetic ‘‘Kleiner community’’ or Bering Sea microbiome) were removed, the obtained community profiles at the higher taxonomic rankings (e.g.,

family/phylum) changed only marginally compared with the unfiltered database output (Figures S2B and S5B). However, an incomplete taxonomic database unavoidably limits the achievable taxonomic resolution. Nevertheless, this is expected to become an increasingly less impactful parameter over time. Proteome/genome databases have been rapidly expanding over the past years, and this is likely to further accelerate due to the continuous advancements in sequencing technologies. For example, the RefSeq database expanded by approximately 25,000 entries for bacteria alone over 5 years (November 2015–November 2020), which corresponds to an increase of >100 million protein sequences in that period of time (www.ncbi.nlm.nih.gov/refseq/statistics/).

Furthermore, to evaluate the dependency of achieving a comprehensive taxonomic coverage on the volume of peptide sequences, we performed a random downsampling of the peptide sequences from the synthetic ‘‘Kleiner community’’ and the Bering sea metaproteomics datasets. To evaluate the impact of the downsampling procedure, we plotted the number of the remaining sequences against the obtained number of taxonomic identifiers. This showed a plateau for the number of obtained taxonomic identifiers at a certain percentage of the original number of peptide sequences (approximately 40%–60% for the ‘‘Kleiner community’’ and approximately 80%–90% for the Bering Sea microbiome) for both metaproteomic experiments. This means that (nearly) no new taxonomic identifiers were obtained after this fraction of peptide sequences and that the acquired datasets therefore indeed comprehensively cover the microbiome biomass.

Establishing the actual content of a community member

Finally, we aimed to investigate the usefulness of *in silico* drop-off curves (the decrease in the number of peptides, assigned to different taxonomic ranks from the higher to the lower taxonomic ranks using the (lowest common ancestor [LCA] approach) and BLAST+ homology search, for investigating the actual content of an enrichment culture. Evaluating the drop-off rates of a lineage enables one to evaluate whether the observed numbers of peptides at the higher taxonomic levels (e.g., phylum level = Proteobacteria) are aligned with the number of peptides observed at lower taxonomic levels (e.g., *Ca*. Accumulibacter, genus

level). This approach allows one to evaluate whether the proportion of proteobacteria is likely derived from *Ca*. Accumulibacter or whether there are other lineages present that are not covered by the database. *Ca*. Accumulibacter has been described frequently as showing strong discrepancies in the proposed community contribution when comparing between FISH and 16S RNA sequencing-based techniques (Stokholm-Bjerregaard et al., 2017). Therefore, we analyzed an Accumulibacter enrichment culture metaproteomic dataset through the described pipeline, which indicated a particularly high enrichment (Figure 2C, approximately 98% at the genus level [D*], in contrast to 16S RNA data for the same reactor at an earlier time point of approximately 34% [Da Silva et al., 2018]). When comparing the experimental drop-off rate for the lineage of *Ca*. Accumulibacter with the *in silico* constructed drop-off curve, we observed a discrepancy of only approximately 17% (D**), meaning that nearly all sequences assigned to proteobacteria translate to the *Ca*. Accumulibacter genus-level annotations.

Nevertheless, to fully exclude significant quantities of potential other populations—e.g., from other phyla, not captured by the (Unipept) database—the high-quality unmatched sequences (top 20% based on ALC scores) were analyzed using BLAST+ against the non-redundant NCBI protein sequence database (for the sake of speed using a local installation). Thereby, approximately 83% of newly retrieved (genus level) sequences could be attributed again to *Ca*. Accumulibacter (Table S7; Figure 2D), reflecting the estimated content obtained after drop-off correction. Moreover, the high degree of enrichment indicated by our metaproteomics experiments is in good agreement with the observed phosphate accumulation activity, observed for this culture during lab experiments (data not shown). Determining the fraction of unmatched (high-quality) spectra has already been proposed as an indicator for the presence of community members not captured by the database (Kleiner et al., 2017; Johnson et al., 2020). The fraction of unmatched high-quality spectra, however, may considerably depend on the applied analytical procedures. The same was observed for the reference strains used in this study, in which raw data were acquired from different laboratories and thus showed large variations in their fraction of peptides that obtained taxonomic annotations (Figure 1E). Although this approach appears very promising, it may provide misleading conclusions if not corrected for individual analytical procedures.

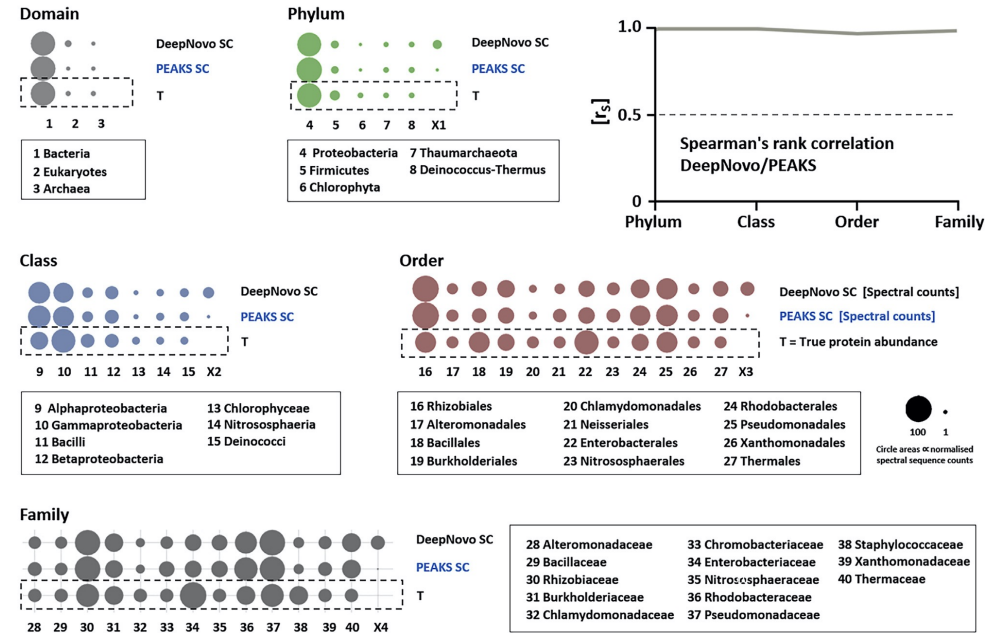


Figure 3. Comparison of microbiome profiles established by PEAKS and DeepNovo (A) Community profiles of the ''equal protein'' community established by PEAKS and DeepNovo. The circle graphs show the taxonomic profiles obtained from the ''equal protein'' community (Kleiner et al., 2017) established by PEAKS or DeepNovo. *De novo* sequence lists from both platforms were processed by the established *de novo* metaproteomics pipeline using the same parameters. ''T'' represents the true abundance of the respective community members (dashed box). ''PEAKS SC'' represents the established profiles obtained from the PEAKS *de novo* sequences using spectral sequence counting. ''DeepNovo SC'' represents profiles obtained from the DeepNovo *de novo* sequences using spectral sequence counting. The unexpected, ''other'' taxonomic annotations were summed and are shown as circles labeled with ''X.'' The experiment demonstrates that both tools provide very comparable taxonomic profiles and only differ in the proportions of the unexpected ''other'' matches. The circles represent the average of 2 analyses, where the circle areas correlate to the normalized spectral sequence counts. The left upper graph shows the Spearman's rank correlation of the taxonomic profiles between PEAKS and DeepNovo. The very strong correlation (rS between 1.0–0.97, from phylum–family, considering the expected taxonomic identifiers, 1–40) confirms that both tools provide highly comparable peptide sequence lists.

3.3 Discussion

Metaproteomics has emerged as one of the most promising post-genomics approaches to study microbial dynamics in nature or in the context of human health, such as the microbial dynamics of the gut microbiome (Behsaz et al., 2020; Timmins-Schiffman et al., 2017). However, common metaproteomics workflows require the laborious construction of high-quality protein sequence databases. Thus, spectrum-matching algorithms are challenged by very large databases or unsequenced community members not covered by the database. Furthermore, the quantitative aspect is often only poorly supported, despite being utmost important when investigating community dynamics.

Here, we introduce a newly established *de novo* metaproteomics workflow, which enables quantitative profiling of microbial communities within a very short analysis time. We provide a systematic evaluation of the taxonomic resolution and quantitative performance using reference strains and natural communities. Thereby, we introduce a validation procedure and demonstrate how to establish the actual content of community members within community proteomics data. The established pipeline automates data filtering, taxonomic annotation, additional validation procedures, grouping, and reporting of taxonomic and functional outputs with only minutes of processing time for a typical shotgun metaproteomics dataset. In comparison, metagenomics including database construction, or the analysis of the mass spectrometric data against very large generic databases, typically requires (several) days of processing time.

Notably, because our approach is database independent, it generates peptide sequences also from ‘‘not-in-the-database’’ community members, making them accessible for further interpretation. The achievable resolution in *de novo* metaproteomics, however, depends not only on the taxonomic database but also on the abundance of the individual community members. Moreover, a completely metagenomics-independent evaluation of a community, containing only unsequenced community members, will likely provide only a comparatively low taxonomic resolution or provide assignments only to the closest taxa present in the database.

The evaluation we performed demonstrates that the highest accuracy is achieved up to the family level, which could therefore be suggested as the default level of operation. However, an improved resolution and quantification (number of peptide matches) for the lower taxonomic rankings—such as genus or even species level—could currently be achieved by performing a *de novo*/database-matching hybrid approach. *De novo*-established taxonomies thereby guide the construction of a focused database from large generic databases, which subsequently can be used for comparatively efficient peptide-spectrum-matching experiments.

Nevertheless, the current vast technical advancements in the field of mass spectrometry and sequencing algorithms are likely to continue improving the quality of the sequencing spectra and thus the number of correct *de novo* sequence annotations in the near future. Ultimately, this will strengthen and expand the scope of *de novo* metaproteomics as either a hybrid, orthogonal, or stand-alone approach.

3.4 References

- Albertsen, M., Hansen, L.B.S., Saunders, A.M., Nielsen, P.H., and Nielsen, K.L. (2012). A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. ISME J. *6*, 1094–1106.
  Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P.C., Mylne, J.S., and Pevzner, P.A. (2020). De novo peptide sequencing reveals many cyclopeptides in the human gut and other environments. Cell Syst. *10*, 99–108.e5.
- Boekel, J., Chilton, J.M., Cooke, I.R., Horvatovich, P.L., Jagtap, P.D., Ka€ll, L., Lehtio¨ , J., Lukasse, P., Moerland, P.D., and Griffin, T.J. (2015). Multi-omic data analysis using Galaxy. Nat. Biotechnol. *33*, 137–139.
  Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.
- Camacho, C., Madden, T., Tao, T., Agarwala, R., and Morgulis, A. (2008). BLASTÒ Command Line Applications User Manual (National Center for Biotechnology Information).
  Da Silva, L.G., Gamez, K.O., Gomes, J.C., Akkermans, K., Welles, L., Abbas, B., Van Loosdrecht, M.C., and Wahl, S.A. (2018). Revealing metabolic flexibility of Candidatus Accumulibacter phosphatis through redox cofactor analysis and metabolic network modeling.
- Hansen, S.H., Stensballe, A., Nielsen, P.H., and Herbst, F.A. (2014). Metaproteomics: evaluation of protein extraction from activated sludge. Proteomics *14*, 2535–2539.
  Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. J. Biotechnol. *261*, 24–36.
- Johnson, R.S., Searle, B.C., Nunn, B.L., Gilmore, J.M., Phillips, M., Amemiya, C.T., Heck, M., and MacCoss, M.J. (2020). Assessing protein sequence database suitability using de novo sequencing. Mol. Cell. Proteomics *19*, 198–208.
- Junqueira, M., Spirin, V., Balbuena, T.S., Thomas, H., Adzhubei, I., Sunyaev, S., and Shevchenko, A. (2008). Protein identification pipeline for the homology-driven proteomics. J. Proteomics *71*, 346–356.
- Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.
- Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., Liu, D., Li, C., and Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. Nat. Commun. *8*, 1558.
- Lee, J.-Y., Mitchell, H.D., Burnet, M.C., Jenson, S.C., Merkley, E.D., Shukla, A.K., Nakayasu, E.S., and Payne, S. (2018). Proteomics of natural bacterial isolates powered by deep learning-based de novo identification. bioRxiv https:// www.biorxiv.org/content/10.1101/428334v1.
- Ma, B., and Johnson, R. (2012). De novo sequencing and homology searching. Mol. Cell. Proteomics *11*, O111.014902.
- Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *17*, 2337–2342.
- Madden, T. (2013). Chapter 16: The BLAST sequence analysis tool. The NCBI handbook, Second Edition (National Center for Biotechnology Information)), pp. 1–15.
  Maier, T., Gu€ell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett. *583*, 3966–3973.
- Martin, F., and Uroz, S. (2016). Microbial environmental genomics (MEG) (Springer).
- Medzihradszky, K.F., and Chalkley, R.J. (2015). Lessons in de novo peptide sequencing by tandem mass spectrometry. Mass Spectrom. Rev. *34*, 43–63.
- Mesuere, B., Debyser, G., Aerts, M., Devreese, B., Vandamme, P., and Dawyndt, P. (2015). The Unipept metaproteomics analysis pipeline. Proteomics *15*, 1437–1442.

- Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., and Dawyndt, P. (2012). Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J. Proteome Res. *11*, 5773–5780.
- Mesuere, B., Willems, T., Van Der Jeugt, F., Devreese, B., Vandamme, P., and Dawyndt, P. (2016). Unipept web services for metaproteomics analysis. Bioinformatics *32*, 1746–1748.
- Mikan, M.P., Harvey, H.R., Timmins-Schiffman, E., Riffle, M., May, D.H., Salter, I., Noble, W.S., and Nunn, B.L. (2020). Metaproteomics reveal that rapid perturbations in organic matter prioritize functional restructuring over taxonomy in western Arctic Ocean microbiomes. ISME J. *14*, 39–52.
- Mooradian, A.D., Van Der Post, S., Naegle, K.M., and Held, J.M. (2019). ProteoClade: a taxonomic toolkit for multi-species and metaproteomic analysis. bioRxiv https://www.biorxiv.org/content/10.1101/793455v1.
- Muth, T., Kolmeder, C.A., Saloja€rvi, J., Keskitalo, S., Varjosalo, M., Verdam, F.J., Rensen, S.S., Reichl, U., De Vos, W.M., Rapp, E., and Martens, L. (2015). Navigating through metaproteomics data: a logbook of database searching. Proteomics *15*, 3439–3453.
- Muth, T., Renard, B.Y., and Martens, L. (2016). Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. Expert Rev. Proteomics *13*, 757–769.
- Potgieter, M.G., Nel, A.J., Tabb, D.L., Fortuin, S., Garnett, S., Blackburn, J., and Mulder, N.J. (2019). MetaNovo: a probabilistic approach to peptide and polymorphism discovery in complex mass spectrometry datasets. bioRxiv https://www.biorxiv.org/content/10.1101/605550v6.
- Riffle, M., May, D.H., Timmins-Schiffman, E., Mikan, M.P., Jaschob, D., Noble, W.S., and Nunn, B.L. (2017). MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. Proteomes *6*, 2.
- Sayers, E. (2009). The E-utilities in-depth: parameters, syntax and more. In Entrez Programming Utilities Help (National Center for Biotechnology Information), pp. 1–156.
- Singh, R.G., Tanca, A., Palomba, A., Van Der Jeugt, F., Verschaffelt, P., Uzzau, S., Martens, L., Dawyndt, P., and Mesuere, B. (2019). Unipept 4.0: functional analysis of metaproteome data. J. Proteome Res. *18*, 606–615.
- Stokholm-Bjerregaard, M., Mcilroy, S.J., Nierychlo, M., Karst, S.M., Albertsen, M., and Nielsen, P.H. (2017). A critical assessment of the microorganisms proposed to be important to enhanced biological phosphorus removal in full-scale wastewater treatment systems. Front. Microbiol. *8*, 718.
- Timmins-Schiffman, E., May, D.H., Mikan, M., Riffle, M., Frazar, C., Harvey, H.R., Noble, W.S., and Nunn, B.L. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. ISME J. *11*, 309–314.
- Tran, N.H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nat. Methods *16*, 63–66.
- Tran, N.H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017). De novo peptide sequencing by deep learning. Proc. Natl. Acad. Sci. USA *114*, 8247–8252.
- Wilmes, P., and Bond, P.L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol. *14*, 92–97.
- Xiao, J., Tanca, A., Jia, B., Yang, R., Wang, B., Zhang, Y., and Li, J. (2018). Metagenomic taxonomy-guided database-searching strategy for improving metaproteomic analysis. J. Proteome Res. *17*, 1596–1605.
- Zhang, X., Ning, Z., Mayne, J., Moore, J.I., Li, J., Butcher, J., Deeke, S.A., Chen, R., Chiang, C.-K., Wen, M., et al. (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome *4*, 31.

3.5 methods

Application of publicly available data

The synthetic community proteomic raw data were downloaded from ProteomXchange server project PXD006118, established by M. Kleiner and M. Strous labs (Kleiner et al., 2017). Protein content and taxonomic lineages of the synthetic community samples used have been further outlined in the Tables S2 and S3. Due to incomplete coverage of viral strains in the Unipept database, viruses were not further considered in the quantitative analysis. Shotgun proteomic raw data from *Rhodopseudomonas palustris* were retrieved from the project PXD013729 generated by E. Nakayasu, Pacific Northwest National Laboratory and C.S. Harwood, University of Washington, *Campylobacter jejuni* raw data were retrieved from PXD005306 generated by M. Monroe, and J. Adkins, Pacific North-west National Laboratory, Paracoccus denitrificans raw data were downloaded from project PXD013274 generated by T. J. Erb and M. Glatter, MPI Marburg, respectively. *Lactobacillus sakei* PXD011417 from C. Ludwig, Bavarian Center for Biomolecular Mass Spectrometry (BayBioMS), Technical University Munich. Acinetobacter baumannii PXD011302 from M. Feldmann Washington University School of Medicine and J. Scott, University of Melbourne; Streptococcus mutans PXD006735 from J. Koh and K.C. Rice, University of Florida; Trypanosoma brucei PXD009073 from J.W. Dupuy form Centre de Ge´ nomique Fonctionnelle Bordeaux, France and M. Boshart, from Ludwig-Maximilians-University Munich Martinsried, Germany. Additional environmental community reference dataset shown in Figure S5, was obtained from PXD008780, as published by B. L. Nunn and E. Timmins-Schiffman of the University of Washington (Mikan et al., 2020). The waste water treatment plant community data shown in Figure S6, were obtained from processing PXD000862, which were published by S. A. Hansen and F. A. Herbst, from Aalborg University.(Hansen et al., 2014) Comparative database-search peptide sequences were retrieved from published supplemental information, which were filtered for sequences with PEP<0.01 before processing through the pipeline. qFISH abundances were obtained from the paper published by Albertsen et al. (2012),(Albertsen et al., 2012) using the ‘GetData Graph Digitizer’ tool.

Whole cell lysate proteolytic digestion

Y-PER reagent (Thermo ScientificÔ, for yeast cells), respectively. The supernatant was collected by centrifugation at 14.000xg. The protein content was precipitated using TCA (1 vol TCA 100 w/v % to 4 vol sample) followed by washing with ice cold acetone. The protein pellet was resuspended in 200 mM ammonium bicarbonate containing 6M Urea, reduced in a 10 mM DTT solution at 40C for 1 hour, and alkylated using 20 mM IAA in the dark, at room temperature, for 30 minutes. The solution was diluted to below 1 M Urea and digested using sequencing grade Trypsin at a protease to protein ratio of approximately 1:50. Peptides were desalted using Oasis HLB solid phase extraction cartridges (Waters Corporation) according to the protocol provided by the manufacturer, speed-vac dried and resuspended in 3% acetonitrile in H2O, containing 0.1% formic acid.

Shotgun metaproteomic analysis

An aliquot of each sample was analysed using a nano-liquid-chromatography system consisting of an EASY nano LC 1200 equipped with an Acclaim PepMap RSLC RP C18 reverse phase column (75mm x 150mm, 2mm) coupled to a QE plus Orbitrap mass spectrometer (Thermo, Germany). Solvent A was H2O containing 0.1% formic acid, and solvent B consisted of 80% acetonitrile in H2O, containing 0.1% formic acid. The flow rate was maintained at 300 nL/min. The Orbitrap was operated in top 10 data dependent acquisition (DDA) mode, acquiring peptide signals form 350-1400 m/z, at 70K resolution in MS1 with an AGC target of 3e6 and max IT of 100ms. For yeast, approx. 250ng protein digest were analysed using a short linear gradient from 4 to 30% B over 32.5 minutes, and further to 70% B over 12.5 minutes. MS2 acquisition was performed at 17.5K resolution, with an AGC target of 2e5, and a max IT of 54ms, using a NCE of 28. Unassigned, singly charged as well as 7, 8 and >8 charged mass peaks were excluded. For bacterial samples, approx. 100ng protein digest were analysed using a linear gradient from 5-30% B over 85 minutes and further to 75% B over 25 minutes. MS2 acquisition was performed at 17.5K resolution, with an AGC target of 1e5, and a max IT of 54ms, at a NCE of 30. Unassigned, singly charged, 8 and >8 times charged mass peaks were excluded.

PEAKS and DeepNovo raw data processing

published by Albertsen et al. (2012),(Albertsen et al., 2012) using the 'GetData Graph Digitizer' tool. Peptide sequencing procedures: Mass spectrometric raw data were processed using PEAKS Studio X (Bioinformatics Solutions Inc., Canada)(Ma et al., 2003) for database search and *de novo* sequencing, or DeepNovo(Tran et al., 2019) for comparative *de novo* sequencing studies. Both, *de novo* sequencing and database search was performed allowing 15ppm parent ion and 0.015Da fragment mass error (depending on the acquisition, slightly more tolerant parameters such as 20ppm/0.02Da were applied). Carbamidomethylation was set as fixed and methionine oxidation as variable modifications. Database search allowed in addition N/Q deamidation as variable modifications. The same settings were applied to DeepNovo where applicable, otherwise software default settings were used. Database search further used decoy fusion for estimation of false discovery rates (FDR) and subsequent filtering of peptide spectrum matches for 1% FDR. Only the top ranked *de novo* sequence annotations were considered for processing. Both, sequence lists were further processed through the same metaproteomics pipeline. Except for the comparative study, shown in Figure 3, PEAKS was used to generate sequence lists.

NovoBridge data processing pipeline

A Matlab 'main script' was constructed that links together functions for pre-filtering, sequence randomisation, automated submissions to Unipept to obtain taxonomic and functional information, threshold filtering, taxonomic grouping and visualisation of output data. The pipeline was established and tested with peptide sequence lists generated by *de novo* sequencing using PEAKS or Deep-Novo, from high-resolution QE Orbitrap shotgun proteomics raw data. The script was constructed using Matlab 2017b and 2019 respectively.

Function 1, pre-filtering, sequence randomisation and Unipept submission

The first part of the script involves importing peptide sequence lists (obtained from PEAKS/DeepNovo) into the Matlab environment and to perform pre-filtering based on the sequence annotation quality parameters. The default pre-filtering thresholds were set to ALC scores >40, less than 20ppm mass error and a minimum peptide length of 7 amino acids. Sequence lists were 'cleaned' from peptide modification annotations and mass errors were corrected for mass drifts.

The Matlab 'rand' function was further used to generate additional randomised sequences from imported *de novo* lists. Thereby, the order of amino acids in front of the cleavage site (R or K) of every sequence was randomised, keeping original sequence parameters attached. Automated sequence submission to Unipept was done using Unipept's inbuilt API (https://unipept.ugent.be/apidocs) option.(Mesuere et al., 2015) For

retrieving taxonomic information, '*pep2lca*' including the options *'&equate_il=true'*, to equate leucine and isoleucine, were used. Further, '*&extra=true &names=true*' are specified to get the complete taxonomic lineage and the names of every taxonomic rank. The script automatically filters for the main categories super kingdom, phylum, class, order family, genus and species. The '*pept2- funct*' combined with the option *'&equate_il=true'* was used to retrieve additional EC number information.(Mesuere et al., 2015) Thereby, a single peptide sequence can generate multiple EC numbers or pathways which cause functional inference and inflation, particularly when searching against a large sequence database space. For this study, only the top scoring peptide sequence per scan was considered.

Function 2, compositional analysis

The compositional analysis considered the major taxonomic categories super kingdom, phylum, class, order, family, genus and species. Depending on data quality/abundance, lower ranks (such as species or genus) were excluded from quantitative analysis/representation due to low numbers or insufficient annotations. In a first step, tables were filtered for sequences with ALCs >70 (or less than -0.1 for DeepNovo), and a mass error of less than 15 ppm. To exclude random matches from erroneous *de novo* sequences or low-abundance signals, a taxonomic identifier of a branch was only considered when occurring at least 3 times. Frequency and ALC cut-offs/thresholds were established using randomised sequences of the pure reference strains. Remaining taxonomic branches are further grouped and visualised using the 'bar(x..,stacked)' function in Matlab for both, absolute and normalized peptide sequence counts (or areas/intensities, respectively). Visualising the relative abundances of the individual community members were performed using circle graphs using the 'surf' function in Matlab. Circle areas represent thereby the number of normalised spectral sequence

counts and show the average of 2 separate analyses (except stated otherwise). True/expected abundances of individual community members of the synthetic communities were retrieved from the supplemental information materials, as published by Kleiner et al. (2017).(Kleiner et al., 2017)

Function 3, functional analysis

KEGG pathways, from global classifications to individual conversions within a pathway, correspond to the KEGG orthology (KO) codes.(Kanehisa and Goto, 2000) Therefore, we established a script, which translates the retrieved enzyme commission numbers (EC) into KO codes. This was done by integrating the KEGG annotation database, downloaded from https://www.genome.jp/kegg-bin/get_htext?ko00001 (10/19), into the Matlab environment. The analysis of the global community metabolic functions, considered thereby only branches which were also used for compositional analysis. Sequences assigned to root and super kingdom levels were excluded. EC assignments matched more than twice (based on unique spectral sequence counts) were further translated into KO codes, normalised to the total number of spectral sequence counts and grouped into pathways. Obtained functional community profiles were visualised using heat maps or circle graphs based on KEGG pathways/category levels 2 (global) and 3 (carbohydrate and energy metabolism). Further information regarding 'KEGG pathway categories' are outlined below.(Kanehisa and Goto, 2000)* Heat maps were generated using the 'heatmap' function, and circle graphs were created using Matlab's 'donut.m' function as available through www.mathworks.com 'file exchange' website.

* Second category codes: 09101 Carbohydrate metabolism, 09102 Energy metabolism, 09103 Lipid metabolism, 09104 Nucleotide metabolism, 09105 Amino acid metabolism, 09106 Metabolism of other amino acids, 09107 Glycan biosynthesis and metabolism, 09108 Metabolism of cofactors and vitamins, 09109 Metabolism of terpenoids and polyketides, 09110 Biosynthesis of other secondary metabolites, 09111 Xenobiotics biodegradation and metabolism, 09121 Transcription 09122 Translation, 09123 Folding, sorting and degradation, 09124 Replication and repair, 09131 Membrane transport, 09132 Signal transduction, 09133 Signalling molecules and interaction, 09141 Transport and catabolism,

09143 Cell growth and death, 09144 Cellular community – eukaryotes, 09145 Cellular community – prokaryotes, 09142 Cell motility.

* Third category codes: 00010 Glycolysis/Gluconeogenesis, 00020 Citrate cycle (TCA cycle), 00030 Pentose phosphate pathway, 00040 Pentose and glucuronate interconversions, 00051 Fructose and mannose metabolism, 00052 Galactose metabolism, 00053 Ascorbate and aldarate metabolism, 00500 Starch and sucrose metabolism, 00520 Amino sugar and nucleotide sugar metabolism, 00620 Pyruvate metabolism, 00630 Glyoxylate and dicarboxylate metabolism, 00640 Propanoate metabolism, 00650 Butanoate metabolism, 00660 C5-Branched dibasic acid metabolism, 00562 Inositol phosphate metabolism, 00190 Oxidative phosphorylation, 00195 Photosynthesis, 00196 Photosynthesis - antenna proteins, 00710 Carbon fixation in photosynthetic organisms, 00720 Carbon fixation pathways in prokaryotes, 00680 Methane metabolism, 00910 Nitrogen metabolism, 00920 Sulfur metabolism. *www.genome.jp/kegg/pathway.html

Function 4. Peptide sequence outputs

To interface with other tools, a peptide sequence table output is provided in form of '.xls' or '.mat' files. Thereby either all sequences, only identified or non-identified sequences can be selected. The later can be filtered for high quality spectra, such as selecting for the top 20% (based on ALC score), which was exemplified using the BLAST+ homology search module, to investigate for potential un-sequenced community members.

Alternative BLAST+ search of unidentified spectra

Alternatively, high quality unidentified *de novo* sequences were subjected to BLAST+ homology search(Madden, 2013; Camacho et al., 2008). Even though there are homology search web services available(Junqueira et al., 2008), we used a local installation to maintain sufficient throughput and integrity with the established *de novo* metaproteomics pipeline. For this ncbi-blast-2.9.0+ and the non-redundant protein sequence database 'nr.gz' (segmented for more efficient use, due to size) were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/blast, updated 12/19) and installed on a local windows 10 workstation. BLAST searches were operated using the Matlab 'system' command function. All BLAST searches used the PAM30 scoring matrix. The top 5 assignments per query sequence (based on bit-scores) were combined and

filtered for best e values and scores, respectively. Taxon ID and name databases were downloaded from the NCBI server. Full taxonomic lineages were retrieved form NCBI using E-utilities calls 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=taxonomy&id=' and 'taxurl_right='&retmode=xml'.(Sayers, 2009)

Taxonomic annotation of metagenomic sequence database

The metagenomics protein (assembly) sequence database from Mikan et al.(Mikan et al., 2020) was annotated with taxonomies using DIAMOND v2.0.6 and the non-redundant bacterial NCBI RefSeq database (Reference Sequence, release 203) and default parameters.(Buchfink et al., 2015) Furthermore, for the top 20% of sequence alignments (based on bit score), complete lineages were determined using the Unipept taxonomy API. The LCA was established using the LCA approach, and the taxonomy ID was retrieved from the prot.accession2taxid database downloaded from the NCBI repository (ftp.ncbi.nih.gov).

Quantification and statistical analysis

In silico evaluation of 'drop-off curves'

Large-scale reference proteomes in silico study: A large number of reference proteomes (>1500) covering all 3 domains of life were retrieved from the NCBI reference database (www.ncbi.nlm.nih.gov/refseq/). *In silico* trypsin cleavage, random selection of 1K sequences (each) and programmed submission to Unipept was done and determination of drop-off curves were performed using Matlab2017b (The MathWorks, Inc., US). Reference proteome *in silico* drop off analysis: A random selection of 3.5K unique trypsin cleaved *in silico* peptides (7-15 amino acids length, to approximate real samples) for the pure strains analysed in this study, as listed in the Table S4, was performed using MATLAB's bioinformatics toolbox. The *in silico* peptidomes were processed through the same NovoBridge pipeline, as described above.

Simulation of peptide sequence database lacking specific taxonomies

Peptide sequences were submitted (as usual), using the above-mentioned NovoBridge pipeline, to Unipept to retrieve taxonomic lineages based on the lowest common ancestor (LCA) approach using 'pept2lca'. Unlike in the default processing pipeline, peptide sequences that obtained a class or lower taxonomic annotation were further annotated with taxIDs using the 'pept2prot' and 'taxonomy' API to obtain all underlying taxonomic lineages. This moreover enabled to selectively remove lineages for peptides from taxonomic entries at defined taxonomic rankings, e.g. species, genus, or family. Following the selective filtering, the remaining taxonomic lineages (of the respective peptide sequences) were regrouped using the LCA approach and further processed through the conventional NovoBridge pipeline with default parameters. The evaluation of the obtained taxonomic profiles was compared to the true (synthetic Kleiner community) or the initially determined taxonomic profiles (Bering Sea community) by determining the Spearman rank-order correlation coefficient (rs) using the MATLAB 'corr' function and the 'Spearman' option.

Simulation of metaproteomics data containing different volumes of peptide sequences

To evaluate the taxonomic profiles obtained from metaproteomics containing different amounts of spectral information, we performed a (random) down-sampling of the peptide sequences. For this, the metaproteomics data from the synthetic 'Kleiner community' or the natural Bering Sea microbiome were down-sampled stepwise to finally contain only 90, 80, 70, 60, 50, 40, 30, 20, 10, 8, 6, 4, 2, or 1% of the original number of peptide sequences. The remaining sequences were further processed through the NovoBridge pipeline using default parameters. The change in the obtained number of taxonomic identifiers (at different taxonomic ranks) was compared using line plots created with the MATLAB 'plot' function and hill equation curve fitting.

Spearman rank correlation

Generally, the evaluation of the obtained taxonomic profiles were compared to the true or otherwise comparatively determined taxonomic profiles by determining the Spearman rank-order correlation coefficient (rs) using the MATLAB 'corr' function and the 'Spearman' option.

3.6 Supporting information

1. Additional materials and methods

Large-scale *in silico* study. Reference proteomes (>1500) covering all 3 domains of life were retrieved from the NCBI reference database (www.ncbi.nlm.nih.gov/refseq/). *In silico* trypsin cleavage, random selection of 1K sequences (each) and programmed submission to Unipept was done using Matlab2017b (The MathWorks, Inc., US).

Whole cell lysate proteolytic digestion. Approximately 25-50mg biomass (wet weight) of each cell pellet/material were homogenised by beads beating in TEAB/B-PER reagent (Thermo Scientific™, for bacterial cells such as *Ca.* Accumulibactor phosphatis enrichment and Clostridium sacch.) or Y-PER reagent (Thermo Scientific™, for yeast cells), respectively. The supernatant was collected by centrifugation at 14.000xg. The protein content was precipitated using TCA (1 vol TCA 100 w/v % to 4 vol sample) followed by washing with ice cold acetone. The protein pellet was resuspended in 200 mM ammonium bicarbonate containing 6M Urea, reduced in a 10 mM DTT solution at 40C for 1 hour, and alkylated using 20 mM IAA in the dark, at room temperature, for 30 minutes. The solution was diluted to below 1 M Urea and digested using sequencing grade Trypsin at a protease to protein ratio of approximately 1:50. Peptides were desalted using Oasis HLB solid phase extraction cartridges (Waters corporation) according to the protocol provided by the manufacturer, speed-vac dried and resuspended in 3% acetonitrile in H_2O, containing 0.1% formic acid.

Shotgun (meta)proteomics. An aliquot of each sample was analysed using a nano-liquid-chromatography system consisting of an EASY nano LC 1200 equipped with an Acclaim PepMap RSLC RP C18 reverse phase column (50µm x 150mm, 2µm) coupled to a QE plus Orbitrap mass spectrometer (Thermo, Germany). Solvent A was H_2O containing 0.1% formic acid, and solvent B consisted of 80% acetonitrile in H_2O, containing 0.1% formic acid. The flow rate was maintained at 300 nL/min. The Orbitrap was operated in top 10 data depended acquisition mode, acquiring peptide signals form 350-1400 m/z, at 70K resolution in MS1 with an AGC target of 3e6 and max IT of 100ms. For yeast, approx. 250ng protein digest were analysed using a short linear gradient from 4 to 30% B over 32.5 minutes, and further to 70% B over 12.5 minutes. MS2 acquisition was performed at 17.5K

resolution, with an AGC target of 2e5, and a max IT of 54ms, using a NCE of 28. Unassigned, singly charged as well as 7, 8 and >8 charged mass peaks were excluded. For bacterial samples, approx. 100ng protein digest were analysed using a linear gradient from 5-30% B over 85 minutes and further to 75% B over 25 minutes. MS2 acquisition was performed at 17.5K resolution, with an AGC target of 1e5, and a max IT of 54ms, at a NCE of 30. Unassigned, singly charged, 8 and >8 times charged mass peaks were excluded. Shotgun proteomic raw data have been made available via ProteomeXchange server project PXD016992.

***De novo* metaproteomics pipeline outline (NovoBridge).** A Matlab 'main script' was constructed that links together functions for pre-filtering, sequence randomisation, automated submissions to Unipept to obtain taxonomic and functional information, threshold filtering, taxonomic grouping and visualisation of output data. The pipeline was established and tested with peptide sequence lists generated by *de novo* sequencing using PEAKS or DeepNovo, from high-resolution QE Orbitrap shotgun proteomics raw data. The script was constructed using Matlab 2017b and 2019 respectively. **Function 1, pre-filtering, sequence randomisation and Unipept submission:** The first part of the script involves importing peptide sequence lists (obtained from PEAKS/DeepNovo) into the Matlab environment and to perform pre-filtering based on the sequence annotation quality parameters. The default pre-filtering thresholds were set to ALC scores >40, less than 20ppm mass error and a minimum peptide length of 7 amino acids. Sequence lists were 'cleaned' from peptide modification annotations and mass errors were corrected for mass drifts. The Matlab 'rand' function was further used to generate additional randomised sequences from imported *de novo* lists. Thereby, the order of amino acids in front of the cleavage site (R or K) of every sequence was randomised, keeping original sequence parameters attached. Automated sequence submission to Unipept was done using Unipept`s inbuilt API (https://unipept.ugent.be/apidocs) option.[1] For retrieving taxonomic information, '*pep2lca'* including the options *'&equate_il=true',* to equate leucine and isoleucine, were used. Further, '*&extra=true &names=true*' are specified to get the complete taxonomic lineage and the names of every taxonomic rank. The script automatically filters for the main categories super kingdom, phylum, class, order family, genus and species. The '*pept2funct'* combined with the option *'&equate_il=true'* was used to retrieve additional EC number information.[1]

Thereby, a single peptide sequence can generate multiple EC numbers or pathways which cause functional inference and inflation, particularly when searching against a large sequence database space. For this study, only the top scoring peptide sequence per scan was considered. **Function 2, compositional analysis:** The compositional analysis considered the major taxonomic categories super kingdom, phylum, class, order, family, genus and species. Depending on data quality/abundance, lower ranks (such as species or genus) were excluded from quantitative analysis/representation due to low numbers or insufficient annotations. In a first step, tables were filtered for sequences with ALCs >70 (or less than -0.1 for DeepNovo), and a mass error of less than 15 ppm. To exclude random matches from erroneous *de novo* sequences or low abundant signals, a taxonomic identifier of a branch was only considered when occurring at least 3 times. Frequency and ALC cut-offs/thresholds were established using randomised sequences of the pure reference strains. Remaining taxonomic branches are further grouped and visualised using the 'bar(x..,stacked)' function in Matlab for both, absolute and normalized peptide sequence counts (or areas/intensities, respectively). Visualising the relative abundances of the individual community members were performed using circle graphs using the 'surf' function in Matlab. Circle areas represent thereby the number of normalised spectral sequence counts and show the average of 2 separate analyses (except stated otherwise). True/expected abundances of individual community members of the synthetic communities were retrieved from the supplemental information materials, as published by Kleiner et al., 2017.[2] **Function 3, functional analysis:** KEGG pathways, from global classifications to individual conversions within a pathway, correspond to the KEGG orthology (KO) codes.[3] Therefore, we established a script, which translates the retrieved enzyme commission numbers (EC) into KO codes. This was done by integrating the KEGG annotation database, downloaded from https://www.genome.jp/kegg-bin/get_htext?ko00001 (10/19), into the Matlab environment. The analysis of the global community metabolic functions, considered thereby only branches which were also used for compositional analysis. Sequences assigned to root and super kingdom levels were excluded. EC assignments matched more than twice (based on unique spectral sequence counts) were further translated into KO codes, normalised to the total number of spectral sequence counts and grouped into pathways. Obtained functional

community profiles were visualised using heat maps or circle graphs based on KEGG pathways/category levels 2 (global) and 3 (carbohydrate and energy metabolism). Further information regarding 'KEGG pathway categories' are outlined below.[3]* Heat maps were generated using the 'heatmap' function, and circle graphs were created using Matlab`s 'donut.m' function as available through www.mathworks.com 'file exchange' website.

*Second category codes: 09101 Carbohydrate metabolism, 09102 Energy metabolism, 09103 Lipid metabolism, 09104 Nucleotide metabolism, 09105 Amino acid metabolism, 09106 Metabolism of other amino acids, 09107 Glycan biosynthesis and metabolism, 09108 Metabolism of cofactors and vitamins, 09109 Metabolism of terpenoids and polyketides, 09110 Biosynthesis of other secondary metabolites, 09111 Xenobiotics biodegradation and metabolism, 09121 Transcription 09122 Translation, 09123 Folding, sorting and degradation, 09124 Replication and repair, 09131 Membrane transport, 09132 Signal transduction, 09133 Signalling molecules and interaction, 09141 Transport and catabolism, 09143 Cell growth and death, 09144 Cellular community – eukaryotes, 09145 Cellular community – prokaryotes, 09142 Cell motility.

*Third category codes: 00010 Glycolysis/Gluconeogenesis, 00020 Citrate cycle (TCA cycle), 00030 Pentose phosphate pathway, 00040 Pentose and glucuronate interconversions, 00051 Fructose and mannose metabolism, 00052 Galactose metabolism, 00053 Ascorbate and aldarate metabolism, 00500 Starch and sucrose metabolism, 00520 Amino sugar and nucleotide sugar metabolism, 00620 Pyruvate metabolism, 00630 Glyoxylate and dicarboxylate metabolism, 00640 Propanoate metabolism, 00650 Butanoate metabolism, 00660 C5-Branched dibasic acid metabolism, 00562 Inositol phosphate metabolism, 00190 Oxidative phosphorylation, 00195 Photosynthesis, 00196 Photosynthesis antenna proteins, 00710 Carbon fixation in photosynthetic organisms, 00720 Carbon fixation pathways in prokaryotes, 00680 Methane metabolism, 00910 Nitrogen metabolism, 00920 Sulfur metabolism. *www.genome.jp/kegg/pathway.html

Function 4. Peptide sequence outputs. To interface with other tools, a peptide sequence table output is provided in form of '.xls' or '.mat' files. Thereby either all sequences, only identified or non-identified sequences can be selected. The later can be filtered for high quality spectra, such as selecting for the top 20% (based on ALC score), which was exemplified using the BLASTp+ homology search module, to investigate for potential un-sequenced community members.

***De novo* sequence homology search.** Alternatively, high quality unidentified *de novo* sequences were subjected to BLASTp+ homology search[4, 5]. Even though there are homology search web services available[6], we used a local installation to maintain sufficient throughput and integrity with the established *de novo* metaproteomics pipeline. For this ncbi-blast-2.9.0+ and the non-redundant protein sequence database 'nr.gz' (segmented for more efficient use, due to size) were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/blast, updated 12/19) and installed on a local windows 10 workstation. BLAST searches were operated using the Matlab 'system' command function. All BLAST searches used the PAM30 scoring matrix. Top search results (based on bit-scores) for every sequence were combined and filtered for best e values and scores, respectively. Taxon ID and name databases were downloaded from the NCBI server. Full taxonomic lineages were retrieved form NCBI using E-utilities calls 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=taxonomy&id=' and 'taxurl_right='&retmode=xml'.[7]

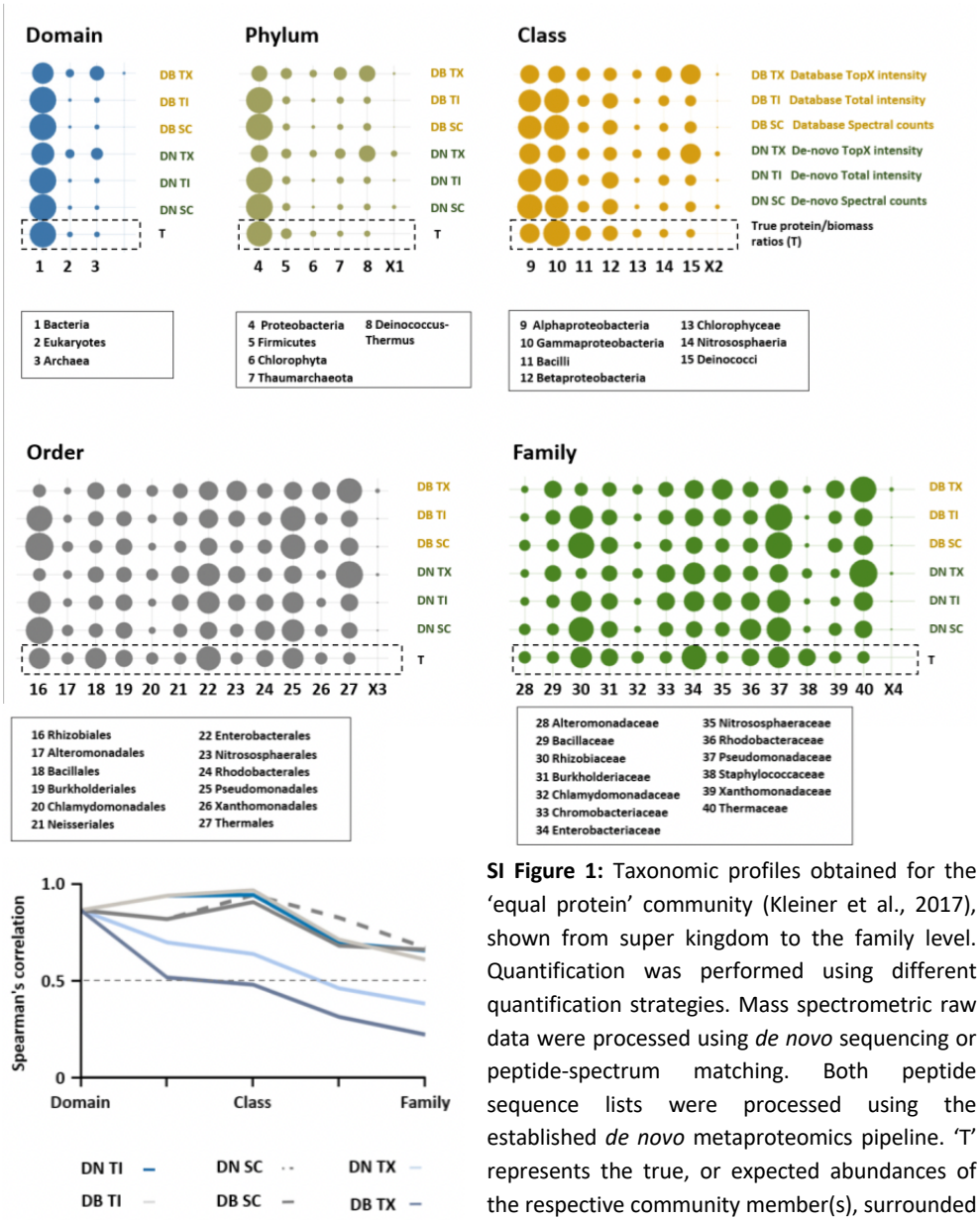
Simulation of peptide sequence databases lacking specific taxonomic entries. Peptide sequences were submitted (as usual), using the above-mentioned NovoBridge pipeline, to Unipept to retrieve taxonomic lineages based on the lowest common ancestor (LCA) approach using 'pept2lca'. Unlike in the default processing pipeline, peptide sequences that obtained a class or lower taxonomic annotation were further annotated with taxIDs using the 'pept2prot' and 'taxonomy' API to obtain all underlying taxonomic lineages. This moreover enabled to selectively remove lineages for peptides from taxonomic entries at defined taxonomic rankings, e.g. species, genus, or family. Following the selective filtering, the remaining taxonomic lineages (of the respective peptide sequences) were regrouped using the LCA approach and further processed through the conventional NovoBridge pipeline with default parameters. The evaluation of the

obtained taxonomic profiles was compared to the true (synthetic Kleiner community) or the initially determined taxonomic profiles (Bering Sea community) by determining the Spearman rank-order correlation coefficient (r_s) using the MATLAB 'corr' function and the 'Spearman' option.

Simulation metaproteomics data containing different volumes of peptide sequences. To evaluate the taxonomic profiles obtained from metaproteomics containing different amounts of spectral information, we performed a (random) down-sampling of the peptide sequences. For this, the metaproteomics data from the synthetic 'Kleiner community' or the natural Bering Sea microbiome were down-sampled stepwise to finally contain only 90, 80, 70, 60, 50, 40, 30, 20, 10, 8, 6, 4, 2, or 1% of the original number of peptide sequences. The remaining sequences were further processed through the NovoBridge pipeline using default parameters. The change in the obtained number of taxonomic identifiers (at different taxonomic ranks) was compared using line plots created with the MATLAB 'plot' function and hill equation curve fitting.

Taxonomic annotation of metagenomics protein sequence database. The metagenomics protein (assembly) sequence database from Mikan et al.[8] was annotated with taxonomies using DIAMOND v2.0.6 and the non-redundant bacterial NCBI RefSeq database (Reference Sequence, release 203) and default parameters.[9] Furthermore, for the top 20% of sequence alignments (based on bit score), complete lineages were determined using the Unipept taxonomy API. The LCA was established using the LCA approach, and the taxonomy ID was retrieved from the prot.accession2taxid database downloaded from the NCBI repository (ftp.ncbi.nih.gov).

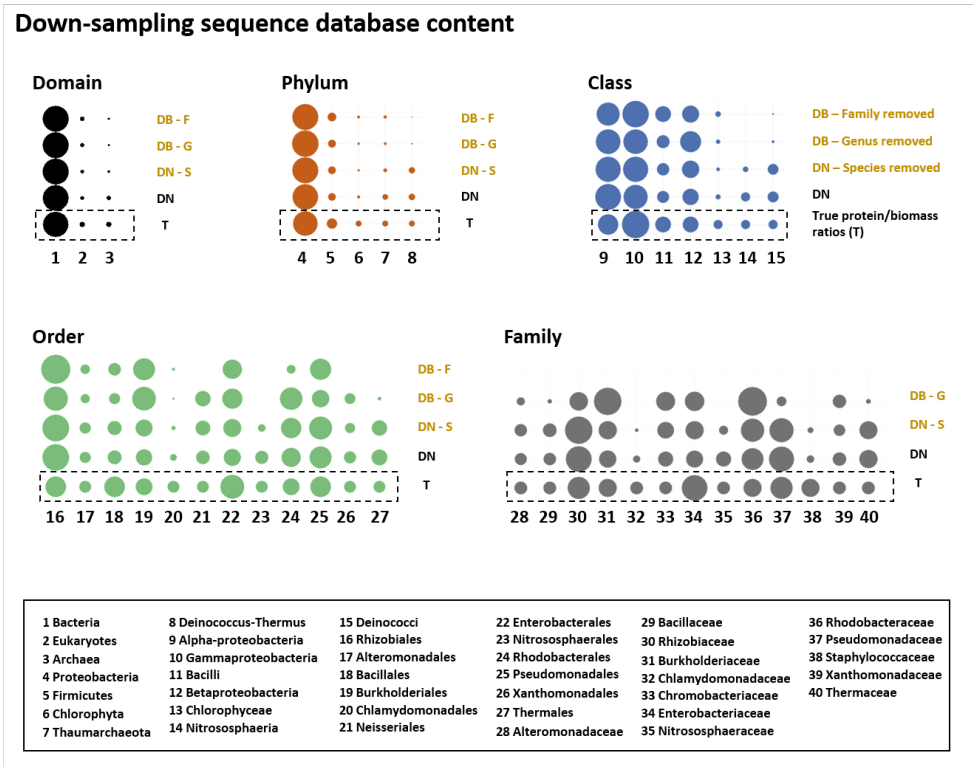
2. Evaluation of different protein quantification strategies

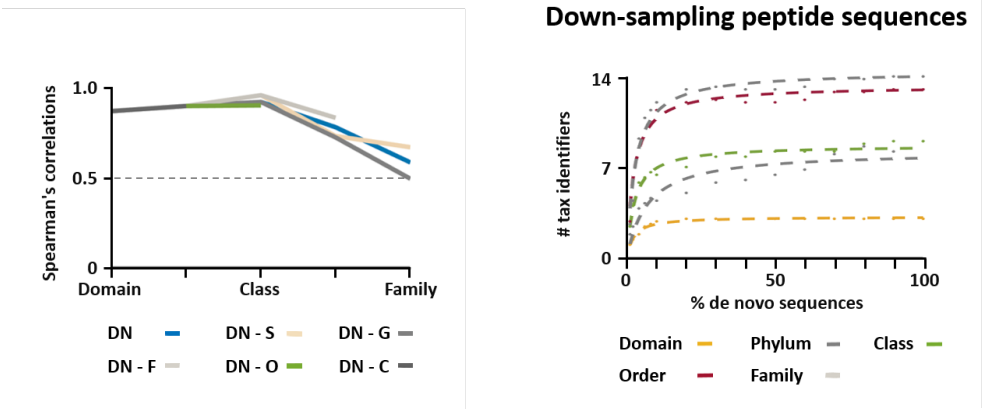


SI Figure 1: Taxonomic profiles obtained for the 'equal protein' community (Kleiner et al., 2017), shown from super kingdom to the family level. Quantification was performed using different quantification strategies. Mass spectrometric raw data were processed using *de novo* sequencing or peptide-spectrum matching. Both peptide sequence lists were processed using the established *de novo* metaproteomics pipeline. 'T' represents the true, or expected abundances of the respective community member(s), surrounded by a dashed box. 'DN SC' shows abundances obtained from the *de novo* sequence lists using spectral sequence counting. 'DN TI' shows abundances obtained from the *de novo* sequence lists using the total spectral (peptide) intensities. 'DN TX' shows abundances obtained from the *de novo* sequences using the intensity of the top 5 most intense spectra. 'DB SC' shows abundances obtained from the

database search sequences using spectral sequence counting. 'DB TI' shows abundances obtained from the database search sequences using the total sequence (peptide) intensities. 'DB TX' shows abundances obtained from the database search sequences using the intensity of the top 5 most intense spectra. The sum of 'other' taxonomic annotations is shown as circles annotated by 'X'. Best correlation was observed for the spectral sequence counting strategy, whereas the 'top 5' approach showed the poorest correlation. The bottom left graph shows the Spearman's rank correlation coefficients plotted over the investigated taxonomic profiles (domain to family level). For both approaches (de novo sequencing and database-matching) the spectral sequence counting (and also the total intensity approach) provided significantly better correlations compared to the 'TX' approach, for which only the counts or area of the top 'X (=5)' strongest signals are considered. Ultimately, the spectral sequence counting approach, is the most straightforward and (likely) most robust approach when applied to different operational conditions. The correlation was determined considering all taxonomic identifiers (1–40, including the non-target 'other' matches labelled with 'X').

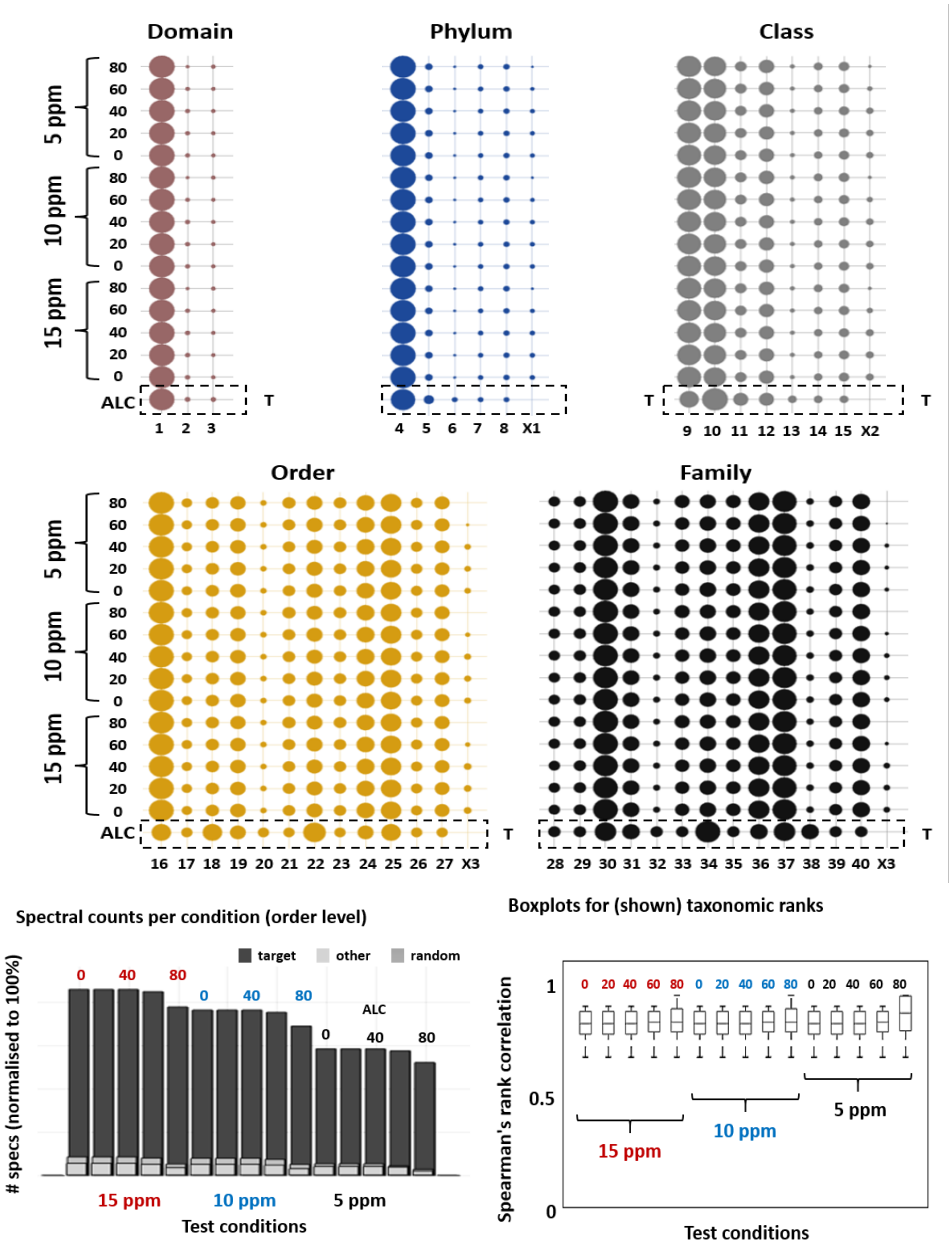
3. Influence of DB completeness and peptide sequence quantities on taxonomic profiles





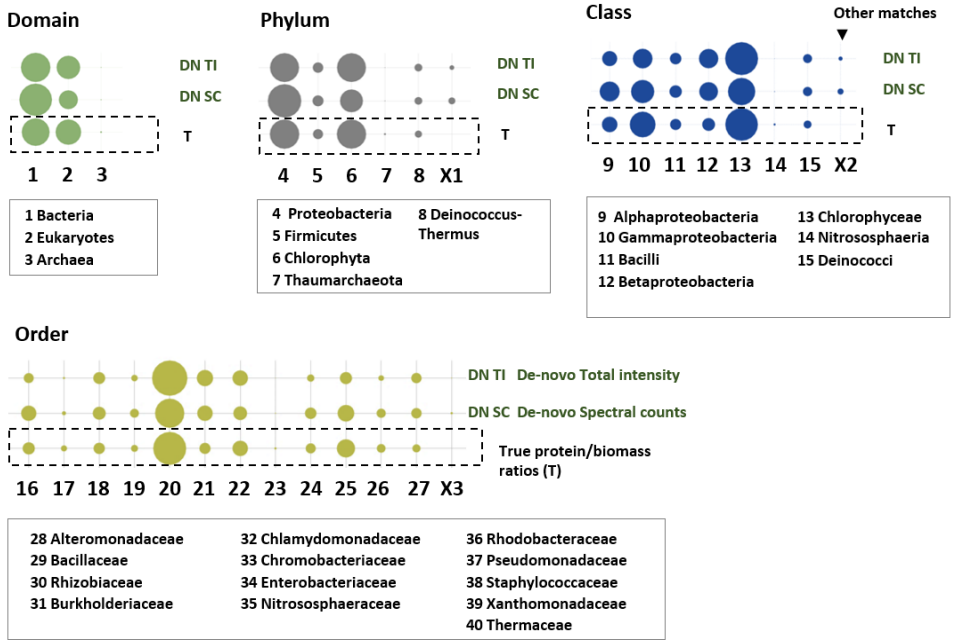
SI Figure 2: The circle graphs show the community profiles obtained for the Kleiner equal protein community (Kleiner et al., 2017)[2] after simulating a Unipept peptide sequence database lacking specific taxonomic entries. The reduced database content was simulated by selectively removing species, genus or family member sequence branches known to be present in the synthetic community. The circle graphs labelled with 'T' show the expected (or true) protein abundance ratio. Circle graphs labelled with 'DN' show the community profiles obtained by using the complete (or unfiltered) Unipept sequence database. 'DN-S', 'DN-G' and 'DN-F' show the community profiles obtained after selectively removing sequence branches corresponding to species, genus or family level of the synthetic community, respectively. Overall, the impact of removing branches/sequences corresponding to the species and genus level had only a relatively moderate impact on the community profiles. Removing the complete set of sequences representing a taxonomic rank, however, abolished (as expected) the annotations to this level. Such a scenario, even if not impacting significantly on the protein abundance ratios at the higher taxonomic levels, would limit the taxonomic resolution. These observations are further demonstrated by the Spearman's rank correlations shown in the bottom left, where the protein abundance profiles induce only moderate changes in the overall Spearman's rank correlation coefficients. The correlations were determined considering the above shown taxonomic identifiers (1–40). The lower right plot shows the impact of different numbers of de novo peptide sequences on the number of identified taxonomies. For the here-investigated equal protein 'Kleiner community', at approx. 40–60% of the original number of peptide sequences, all taxonomic levels reach a plateau. Hence, no new taxonomic identifiers were obtained after further increasing the number of peptide sequences, demonstrating the robustness of the dataset in regard to providing a comprehensive coverage of the community biomass.

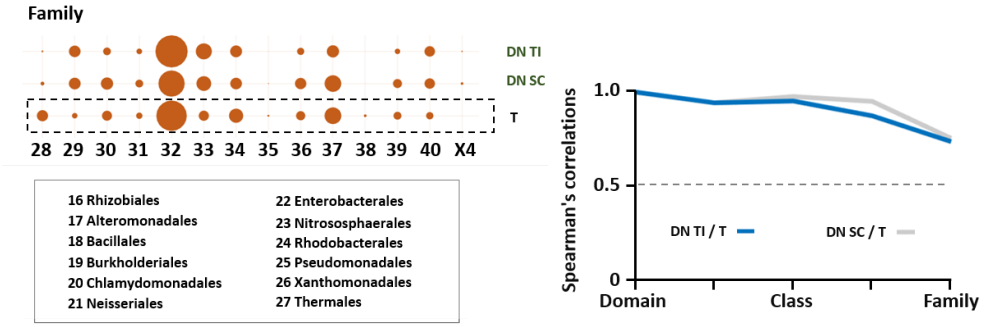
4. Influence of de novo sequence quality parameters on taxonomic profiles



SI Figure 3: The influence of the *de novo* sequence quality parameters 'average local confidence score' (ALC) and mass error (Δ ppm, of the proposed sequence) on quantitative taxonomic profiles was investigated using the 'equal protein' community (Kleiner et al., 2017). Mass spectrometric raw data were *de novo* sequenced and processed by the established *de novo* metaproteomics pipeline. The taxonomic profiles were visualised as circle graphs, where the circle areas correlate to the normalised spectral sequence counts. Shown are the average of 2 separate analyses. The sum of unexpected 'other' taxonomic annotations is shown by circles labelled with 'X'. The experiment demonstrates that the obtained taxonomic profiles were comparatively constant and close to the true abundances ('T') throughout the investigated parameter settings. However, the quantities of 'unexpected' taxonomic annotations decreased when considering only de novo sequences with an ALC score >60.The bottom left bar graph shows the proportion of sequence spectral counts that obtained correct taxonomic annotations, random annotations or assignment to 'other' taxonomic identifiers. With increasingly stringent filtering parameters, the proportion of random sequence matches and annotations to unexpected taxa is reduced. The bottom right boxplot shows the Spearman's rank correlation comprising the investigated taxonomic ranks (1–40, inclusive 'other' matches labelled with 'X') across the testing conditions. The highly comparable correlation plots, moreover, confirm that the taxonomic profiles are relatively stable over the investigated parameter space.

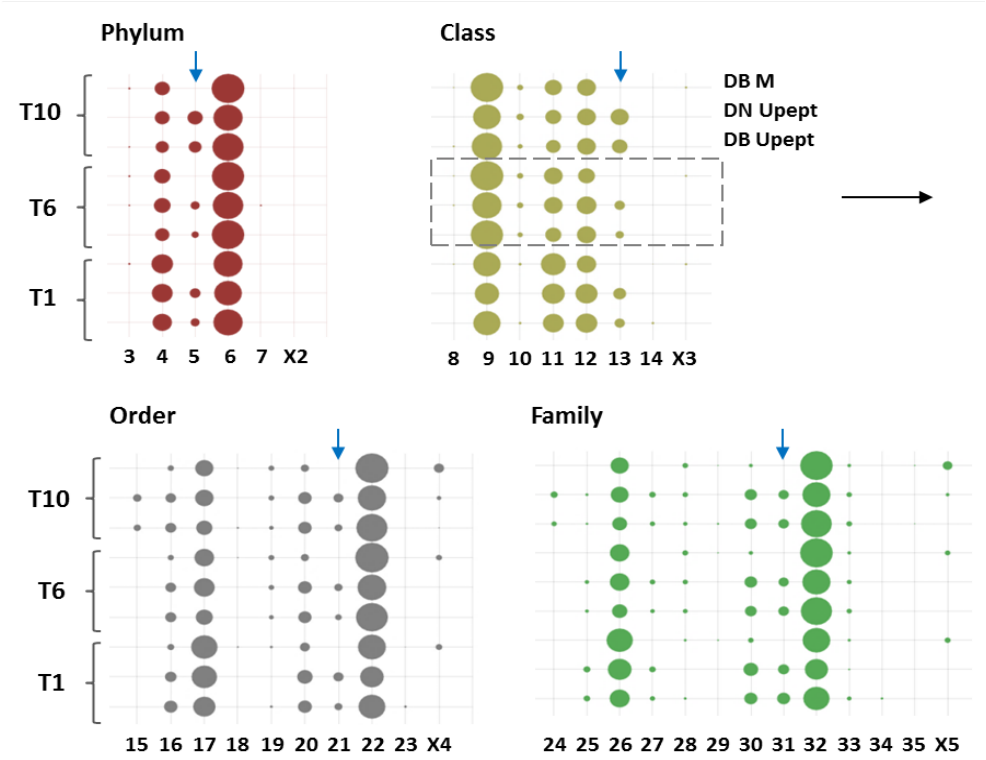
5. Community profiles of 'equal cell' community

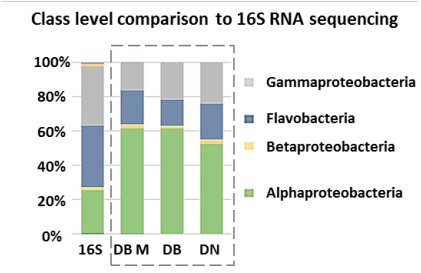


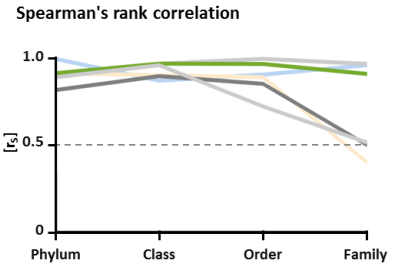


SI Figure 4: The above circle graphs show the quantitative profiles of the 'equal cell' community (Kleiner et al., 2017) with the aim to investigate a community with large protein abundance differences. Mass spectrometric raw data were processed by the described *de novo* metaproteomics pipeline, as described in the materials and methods section. The quantification was performed by either summing up spectral sequence counts or by summing up spectral intensities. 'T' represents the true, or expected abundances of the respective community member(s). 'DN SC' shows abundances obtained from the *de novo* sequences using spectral counting (sum of all sequence annotations). 'DN TI' shows abundances obtained from the *de novo* sequences using the total peptide intensity (sum of all sequence intensities). The sum of 'other' taxonomic annotations are shown by circle graphs labelled by 'X'. Overall, the observed community profiles were highly comparable to the true/expected profiles ('T'). Except *Staphylococcaceae* (#38, family level), all other taxonomic identifiers (from phylum - family) could be observed and were close to the true/expected abundance profiles. The circles areas represent the normalised spectral sequence counts of the respective taxonomic identifiers. Shown is the average of 2 separate analyses. The bottom left plot shows the Spearman's rank correlation of the taxonomic profiles (1–40, including unexpected 'other' matches labelled with 'X') obtained from DB and DN sequence lists compared to the expected taxonomic profiles. Overall, both approaches showed a very high correlation over the investigated taxonomic ranks and only dropped to approx. 0.75—which is still considered a strong correlation—at the family level.

6. Bering Sea microbiome







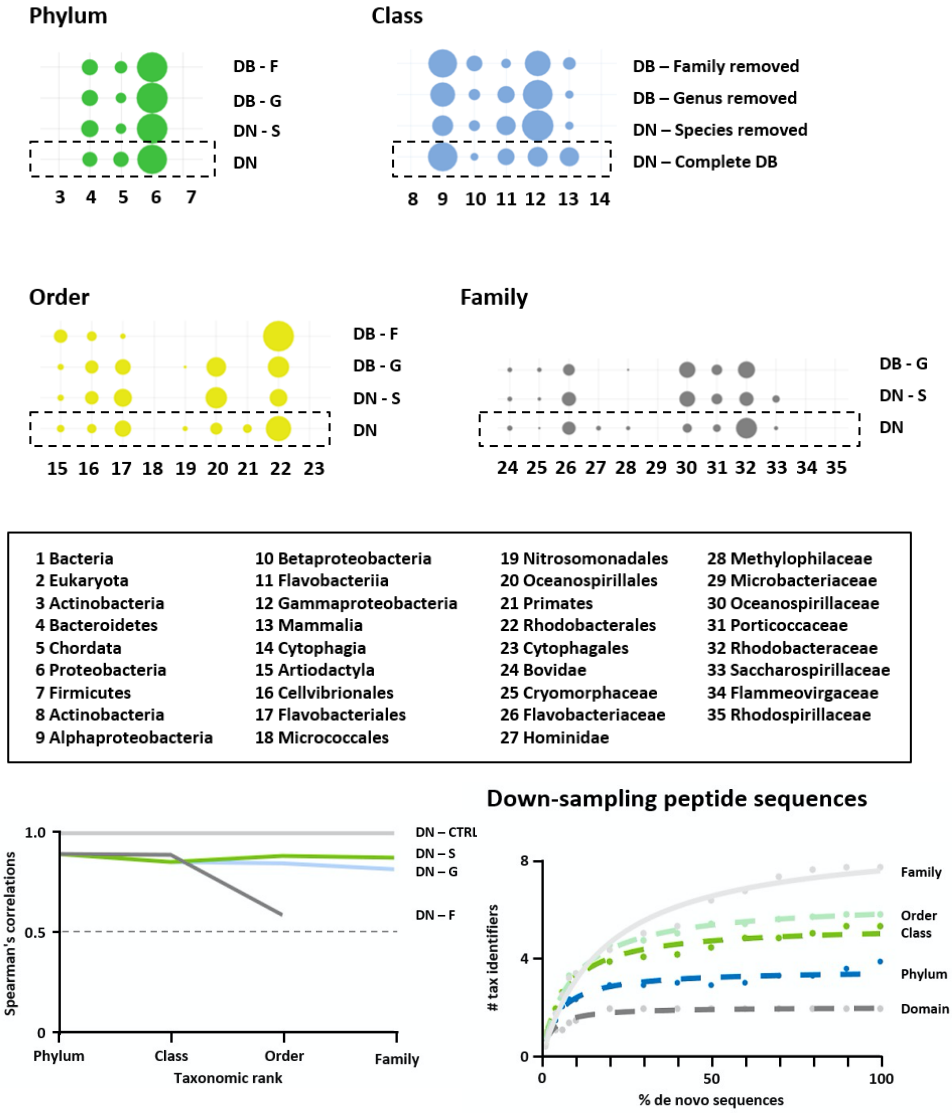
DB M: database matching using the taxonomy annotated metagenomics database

Eukaryotic origin

DN to DB T1
DN to DB M T1
DN to DB T6
DN to DB M T6
DN to DB T10
DN to DB M T10

SI Figure 5a: The circle graphs show the obtained community profiles obtained for the different marine community samples BS-T1, BS-T6 and BS-T10, which were sampled from the Bering sea, as published by Mikan et al (2019).[8] Metaproteomic raw data were retrieved from the proteome exchange server, *de novo* sequenced and analyzed by the established *de novo* metaproteomics pipeline. The circle graphs labelled with 'DB M' show the community profiles achieved from grouping peptide-spectrum matches obtained from using the taxonomy-annotated metagenomics-constructed sequence database. 'DN Upept' show the community profiles obtained from the sequence lists generated by de novo sequencing. 'DB Upept' illustrates the community profiles obtained from the sequence lists generated from the comparative peptide-spectrum matching experiment, using the protein sequence database published by the authors. All approaches used spectral sequence counting for visualizing the community profiles. The overall taxonomic profiles obtained from the different approaches are highly comparable. Differences are only found for the very low abundant community members, at the lower taxonomic rankings. Other differences were a result of the metagenomics sequence database annotation procedure, which did not consider eukaryotic taxonomies (DB M, blue arrows). The circle areas correlate to the normalized spectral sequence counts of the respective taxonomic identifiers. 'Other' taxonomic annotations (not found by the DB Upept submissions) were summed and are shown as circles annotated by 'X'. The circle areas shows the average of 2 separate analyses. The bottom left bar graph provides a comparison of the class-level community composition established by 16S RNA sequencing (Mikan et al., 2019)[8] and by different proteomics approaches. Although the overall composition between the substantially different approaches was highly comparable, the fraction of Alphaproteobacteria was consistently more strongly pronounced in the metaproteomics experiments.

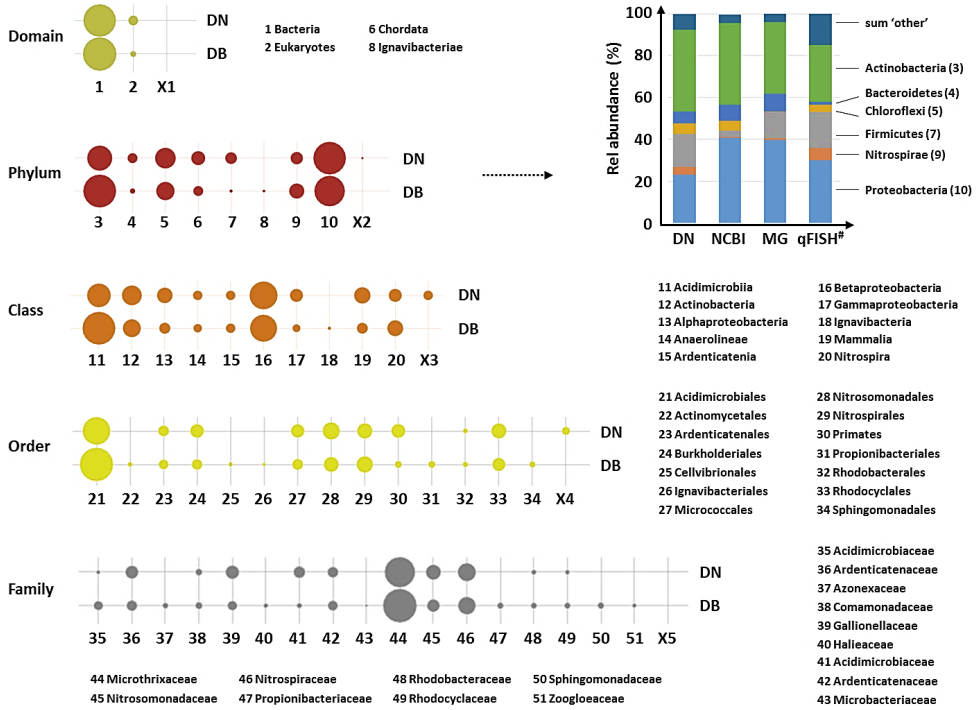
The bottom right plot shows the Spearman's rank correlation between the obtained taxonomic profiles (including taxonomies 3–35 including 'X', but excluding eukaryotic annotations because a purely bacterial DB was used to annotated the metagenomic database) of the different approaches. Peptide sequences obtained from *de novo* sequencing and database-matching provided highly comparable profiles, shown by their high correlation across the investigated taxonomic levels (DN to DB T1/6/10; r_S at family level = 0,96; 0,97 and 0,91 respectively). Comparison of the *de novo* sequence lists with the directly grouped peptide-spectrum matches of the annotated metagenomics database showed only at the lower taxonomic rankings a decrease of the otherwise strong correlation (DN to DB M T1/6/10, r_S at family level = 0,41; 0,51 and 0,52 respectively). The observed (overall comparatively low) deviation is likely related to the database completeness (or differences in content) of the Unipept database compared to the NCBI reference database used to annotate the metagenomics sequence database.



SI Figure 5b: The circle graphs show the community profiles obtained for the marine community BS-T10 (Mikan et al., 2019)8, after simulating a Unipept peptide sequence database lacking specific taxonomic entries. The reduced database content was simulated by selectively removing identified species, genus or family members, previously identified for the same community, using the constructed metagenomics database. The circle graphs labelled with 'DN' show the community profiles obtained by using the complete/unfiltered Unipept sequence database. 'DN-S', 'DN-G' and 'DN-F' show the community profiles obtained after selectively removing all sequences corresponding to previously identified species, genus or family members, respectively.

Overall, the selective removal of branches/sequences at the species and genus level showed only a moderate impact on the community profiles. The relatively low impact is underlined by the Spearman's rank correlation curves shown in the bottom left graph (for the taxonomic identifiers 3–35). 'DN-CTRL' thereby indicates the maximum correlation (=1) for the case of obtaining identical profiles compared to the unfiltered database. 'DN-S', 'DN-G' and 'DN-F' show the Spearman's rank correlation of the species-, genus- or family-filtered community profiles, respectively. Only after removing all sequence branches belonging to the family level was a more significant impact on the community profiles observed (decrease in correlation to approx. 0.76 and 0.75 for DN-S and DN-G (at family level), and drop to 0.4 (at order level) for DN-F). The lower right graph shows the impact of the number of de novo sequences available for representing the community profiles. All taxonomic levels reach approx. 80–90% of the original number of peptide sequences—a plateau. Therefore, after this, no new taxonomic identifiers are obtained and, at the same time, the robustness of the acquired dataset in regard to providing a comprehensive coverage of the community biomass is demonstrated.

7. Wastewater treatment plant microbiome



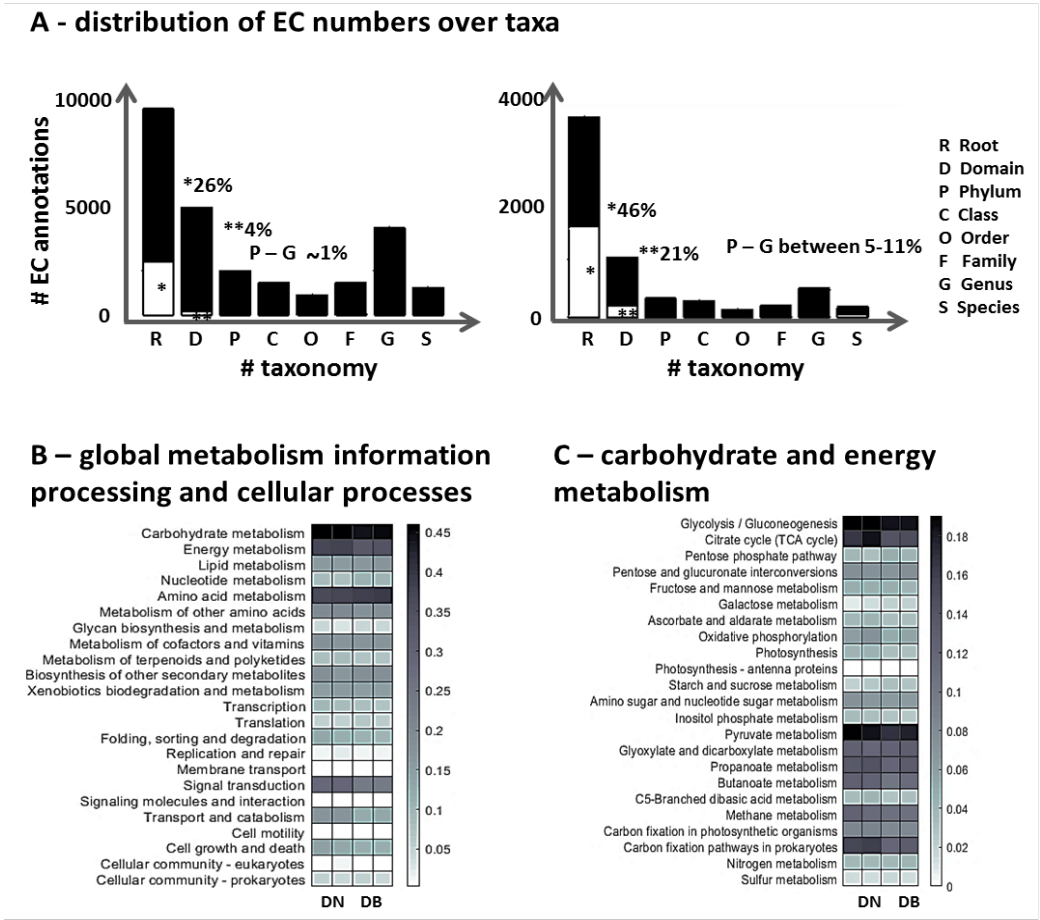
SI Figure 6: The circle graphs show the community profiles of a waste water treatment plant community as published by Hansen et al.[10] Metaproteomic raw data (from the best performing extraction protocol using B-BER buffer and beads beating, FAH_SludgeExtr_B_BB_1 and

FAH_SludgeExtr_B_BB_2), were retrieved from the proteome exchange server project PXD000862, de novo sequenced and analysed by the established de novo metaproteomics pipeline. The circle graphs labelled with 'DN' show the community profiles obtained from the de novo sequence lists using spectral sequence counting. 'DB' shows the comparative community profiles obtained from the sequence lists (as published by the authors) obtained from peptide-spectrum matching using the metagenomics constructed protein sequence database. The taxonomic profiles obtained from both approaches are very comparable and differ only in very low abundant community members. Noteworthy, the phylum-level profiles further appeared highly comparable to qFISH data acquired for the same community at an earlier time point (stacked bar graph). The bar labelled with 'DN' shows thereby the *de novo* sequence list generated phylum profile; the bar labelled with 'NCBI' shows the phylum profile from sequence lists established by using a global NCBI database; and the bar labelled with 'MG' shows the phylum profiles from sequence lists established by using the metagenomics constructed database, as published by Hansen et al.[10] The bar labelled with 'qFISH' shows the profiles obtained by FISH staining from the same community at an earlier time point, published by Albertsen et al.[11] The circle areas correlate to the normalised spectral sequence counts of the respective taxonomic identifiers. Taxonomic annotations only observed for the *de novo* sequences were summed and are shown as circles annotated by 'X'. The graph shows the average of 2 separate analyses.

8. Global community functions visualised as KEGG pathway

Since GO terms have been reported challenging for enrichment analyses due to unclear hierarchies and dependencies,[12, 13] our workflow translates the obtained EC assignments into KO terms and KEGG pathways. Although the peptide-centric approach compares sequences to a large genomic space (sequences may be matched to several enzymes and different pathways), the overall metabolic profiles appear very comparable between peptide-spectrum matching and *de novo* generated sequence lists. (Figure 2B, and SI Figure 5) Nevertheless, the large database space may limit taxonomic resolution and inflate functional annotations. An investigation with deeper taxonomic resolution requires to maximise proteome coverage and to increase the numbers of unique assignments at lower taxonomic rankings. This is supported through spectrum-matching approaches using tailored databases or potentially also by extensive homology search on high quality sequences. EC numbers were translated into KEGG Orthology (KO) pathways, where only sequences which were kept for taxonomic analysis, were considered for evaluation of the functional profiles. Higher taxonomic rankings, such as annotations to the root or super-kingdom levels were not considered. 2

unique annotations per EC number were required as minimum. Thresholds were generally investigated/tested by using randomised peptide sequences.



SI Figure 7. A) The bar graphs visualise the distribution of enzyme commission number (EC) assignments for database search (A, left bar graph) and *de novo* sequencing (B, right bar graph) generated sequence lists. Annotations from 'correct' sequences are shown by black bars and EC annotations retrieved from the randomised sequences are represented by grey bars. Random EC assignments were frequently retrieved for 'root' and 'super kingdom' levels. Therefore, EC annotations from those taxonomic levels were excluded from functional analysis. The total number of assignments (phylum to species level) were lower for the *de novo* sequence lists, which is in agreement with the lower number of spectral assignments for *de novo* sequencing, compared to database search approaches. The profiles have been established from the 'equal protein'

community as provided by Kleiner et al., 2017.[2] **B/C)** Global functional community profiles of the 'equal protein' community, established by sequence lists from database search (DB) and *de novo* (DN) sequencing. EC numbers retrieved from Unipept were translated into KO codes, grouped into KEGG pathways[#] and visualised using heat maps. **A** shows the global profiles for the KEGG 'categories' (1) metabolism, (2) genetic information processing, (3) environmental information processing and (4) cellular processes. **B** further details the metabolic profiles for the (1.1) carbohydrate metabolism and the (1.2) energy metabolism. This experiment shows very comparable profiles between the different peptide sequence annotation approaches. Clustering into KEGG pathways provides a well-structured way to visualise overall community profiles, but may also lead to inflating pathways, because sequences may retrieve more than one EC number or pathway annotation, respectively. The large sequence space of generic databases may further lead to multiple annotations for a single sequence, thereby lowering the accuracy of metabolic profiles. The above shown community profiles were generated from the 'equal protein' community raw data as established by Kleiner et al., 2017.[2] [#]https://www.kegg.jp/kegg/pathway.html#metabolism

9. Additional references

- Mesuere, B. et al. The Unipept metaproteomics analysis pipeline. Proteomics 15, 1437-1442 (2015).
- Kleiner, M. et al. Assessing species biomass contributions in microbial communities via metaproteomics. Nature communications 8, 1558 (2017).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28, 27-30 (2000).
- Madden, T. in The NCBI Handbook [Internet]. 2nd edition (National Center for Biotechnology Information (US), 2013).
- Christiam Camacho, T.M., Tao Tao, Richa Agarwala, Aleksandr Morgulis BLAST® Command Line Applications User Manual. Bookshelf NCBI (2008).
- Junqueira, M. et al. Protein identification pipeline for the homology-driven proteomics. Journal of proteomics 71, 346-356 (2008).
- Sayers, E. The E-utilities in-depth: parameters, syntax and more. Entrez Programming Utilities Help [Internet] (2009).
- Mikan, M.P. et al. Metaproteomics reveal that rapid perturbations in organic matter prioritize functional restructuring over taxonomy in western Arctic Ocean microbiomes. The ISME journal 14, 39-52 (2020).
- Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. Nature methods 12, 59-60 (2015).
- Hansen, S.H., Stensballe, A., Nielsen, P.H. & Herbst, F.A. Metaproteomics: evaluation of protein extraction from activated sludge. Proteomics 14, 2535-2539 (2014).
- Albertsen, M., Hansen, L.B.S., Saunders, A.M., Nielsen, P.H. & Nielsen, K.L. A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. The ISME journal 6, 1094-1106 (2012).
- Gaudet, P. & Dessimoz, C. in The Gene Ontology Handbook 189-205 (Humana Press, New York, NY, 2017).
- Simopoulos, C.M. et al. pepFunk, a tool for peptide-centric functional analysis in metaproteomic human gut microbiome studies. bioRxiv, 854976 (2019).

4

Tackling the chemical diversity of microbial nonulosonic acids – a universal large-scale survey approach

This chapter is based on:

Kleikamp, H.B., Lin, Y.M., McMillan, D.G., Geelhoed, J.S., Naus-Wiezer, S.N., Van Baarlen, P., Saha, C., Louwen, R., Sorokin, D.Y., Van Loosdrecht, M.C. and Pabst, M., (**2020**). Tackling the chemical diversity of microbial nonulosonic acids–a universal large-scale survey approach. *Chemical science*, *11*(11), 3074-3080.

4.1 Introduction

Nonulosonic acids (NulOs, including animal-type sialic acids§) are a class of 9-carbon α-keto acid sugars essential to many cellular processes throughout all domains of life.[2–6] The first nonulosonic acid (*N*-acetyl-neuraminic acid or Neu5Ac) was discovered during the 1940s in salivary mucins (from Greek 'sialon'), establishing the nowadays commonly used abbreviation 'sialic acids'.[7,8] Since then, the broader group of discovered NulO derivatives expanded to some 100 members.[9] Neuraminic acid (Neu) and the unmodified variant keto-deoxy-neuraminic acid (Kdn) are commonly associated with animal tissues. Pseudaminic acid (Pse), its stereoisomers legionaminic (Leg) and acinetaminic acid (Aci) are variants exclusive to prokaryotes.[10] Neus are involved in processes such as cell–cell interaction, signalling, adhesion, regulation of protein half-life and mediating an immune response. Similarly, derivatives of Pse/Leg have been found as part of bacterial cell surface structures such as lipopolysaccharides, peptidoglycans or glycoproteins of cell surface layers, adhesins, pili and flagella.[11–14] Most uniquely, every type of NulO can undergo diversification at multiple positions, which enables an (potential) enormous chemical diversity.[9] Where diversification on animal-type NulOs is mostly limited towards acetylation, glycolylation, methylation, and more rarely, phosphorylation or sulphation, diversifications in prokaryotes have been found to be remarkably innovative. This includes additions of formyl, glyceryl, hydroxybutyryl, lactoyl and glutamyl groups, to just name a few out of many more.[9] Moreover, new types of modifications are being discovered continuously. Currently there is no good estimate on the natural boundaries of this process in prokaryotes. Therefore, the driving forces behind the evolution of a certain configuration remain elusive.

When only considering a number of some 15 possible modifications—occurring at several diversification points—the number of modified NulO derivatives exceeds several thousands, thereby not considering stereochemistry or linkage isomers (Fig. 1A and B). Furthermore, any changes in the chemical structure are likely to affect physicochemical properties and therefore may interfere with highly important molecular recognition processes.

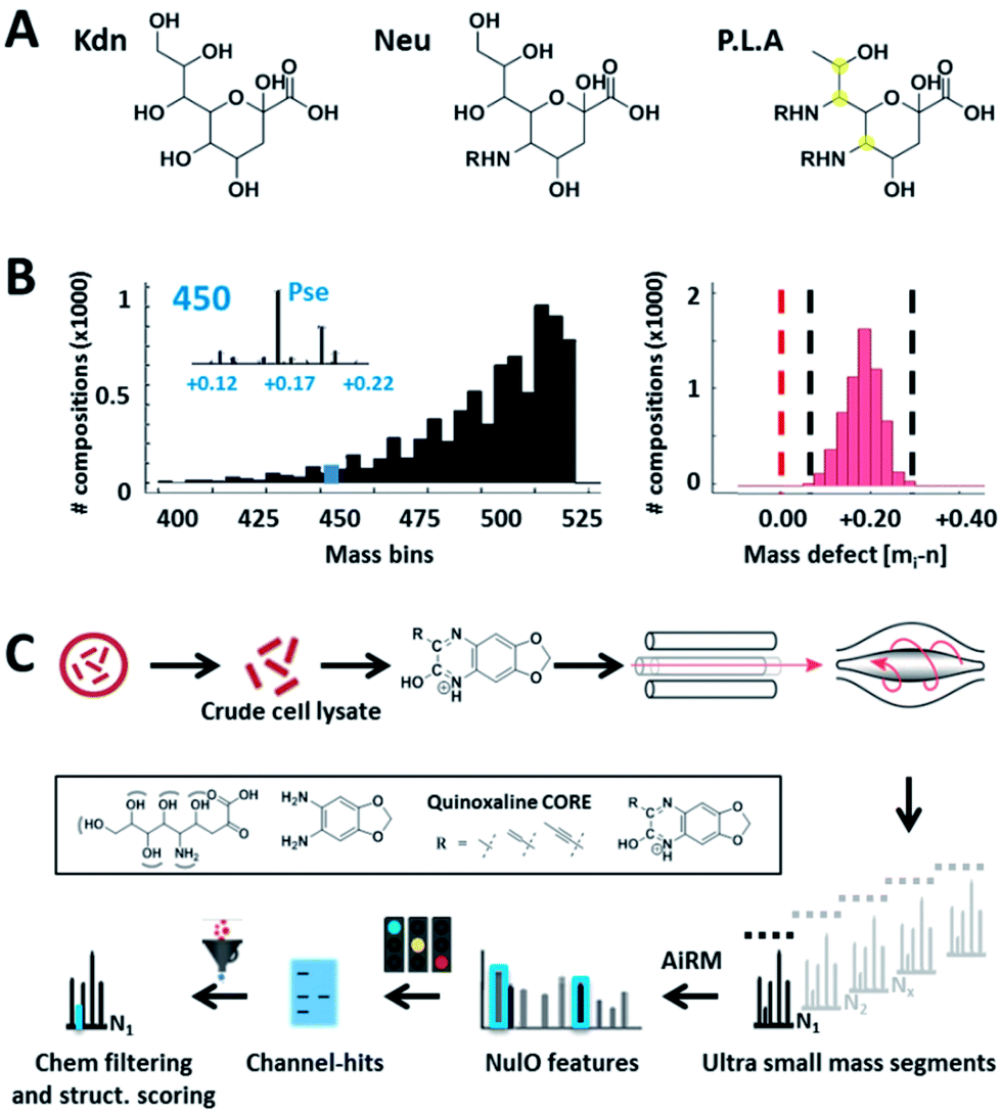

A
Kdn
Neu
P.L.A
B
450
Pse
+0.12
+0.17
+0.22
compositions (x1000)
Mass bins
Mass defect [m_i-n]
C
Crude cell lysate
Quinoxaline CORE
AiRM
Chem filtering
and struct. scoring
Channel-hits
NuIO features
Ultra small mass segments

Fig. 1 Universal sialic acid large-scale survey approach. (A) outlines the three main NulO core structures known to date: Kdn (deaminoneuraminic acid), Neu (neuraminic acid) and P.L.A = Pse (pseudaminic acid), Leg (legionaminic acid) and Aci (acinetaminic acid). (B) demonstrates the potential diversity generated through (combinations) of (natural) diversifications of the three main NulO core compositions Kdn, Neu and P.L.A (Pse/Leg/Aci). The left graph shows a binned mass histogram (5 Da) of the theoretically possible chemical compositions between 380–520 Da. Already more than 25 different compositions are found within ±0.05 Da of Pse5Ac7Ac. The image on the right shows the mass defect binned for every calculated structure, which all fall within a very narrow window between 0.05–0.25 Da. The mass spectrometric mass defect is defined as the delta of monoisotopic mass and nominal mass. For combinatorial space and abbreviations see ESI-Table, sheet 8.‡ This filters for possible realistic structures during parent NulO annotation. (C) outlines the established large-scale survey pipeline starting from crude cell lysates, α-keto acid specific chemical labelling, small mass channel scanning (2.75 *m/z* windows, from 380–520 Da), and the automated structural filtering and evidence scoring pipeline to annotate parent NulOs. In summary, channel hits and compositions are discovered based on highly conserved quinoxaline-core fragments together with the diversification-independent carbon chain length and structural features.

Bacterial NulOs have been commonly linked to virulence and pathogenicity, presumed to mimic the host's glycosylation for evading an immune response.[15] However, a recent genome level study by Lewis *et al.* across bacteria and archaea revealed an unexpectedly wide distribution of homologous neuraminic acid biosynthesis (NAB) pathway genes.[1] Unfortunately, genome-level studies are only predictive and do not allow for conclusions on active gene products, and the possibility that yet-to-be-discovered NulO pathways have evolved completely independently cannot be ruled out completely. Most importantly however, the above mentioned diversifications processes, which make sialic acids and other NulOs so unique, remain unnoticed because they are not accessible through genome level analysis.[16] Furthermore, fully untargeted large-scale molecular level studies on the diversity of natural NulOs is by current state-of-the-art approaches not achievable.

Current detection of sialic acids and other NulOs is achieved by imaging, staining approaches, or by selective fluorescent labelling of the alpha-keto acid group followed by liquid chromatography with or without additional mass-spectrometric detection.[6,17,18] Complex lysates, however, produce a large background derived from lower carbon-chain ulosonic acids or common bulk metabolites. Therefore, the classical approach engages pre-fractionation of target conjugates or applies only to samples with low complexity.[17,19] Unfortunately, this procedure is

laborious and biased, particularly when dealing with uncharacterised (or low abundance) compounds. Alternatively, gas chromatography combined with mass spectrometry has been successfully employed, but this approach also requires pre-fractionation and involves manual data curation of low resolution spectra.[20]

Over the past decade, new high-resolution mass spectrometers have paved the way to a new era in metabolite analysis. Rapid identification of thousands of compounds in a single experimental measurement is achieved, mostly by using shotgun-type sampling. Due to the stochastic nature of this approach, sub-stoichiometric compounds are overlooked, and frequent co-isolation of closely related metabolite peaks generates hybrid-type and low-informative spectra.

The common route for identifying a given metabolite peak is by matching mass and fragmentation spectra to database library entries.[21] More recent advancements include isotope pattern analysis, consider spectral relatedness, and make use of common fragmentation patterns or ultra-high resolution mass spectrometers.[21–23] Even so, the virtually large (and yet unmapped) chemical diversification of prokaryotic NulOs, their sub-stoichiometric occurrence, poorly recorded species distribution and difficult to measure nature, hampers the application of shotgun as well as database-matching approaches.

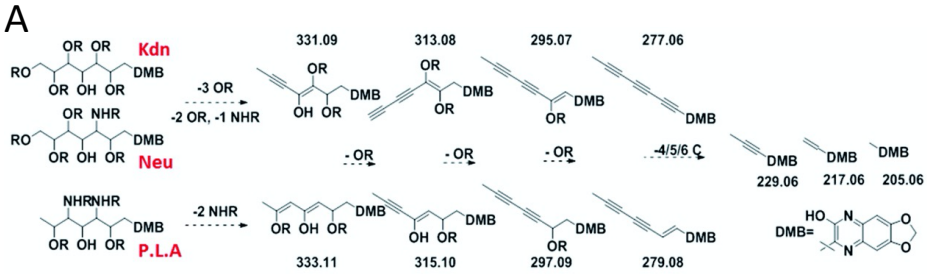
In this study, we describe a nonulosonic acid (NulO) universal screening approach which tackles the (yet unmapped) chemical diversity of prokaryotic nonulosonic acids (NulOs). By doing so, we take advantage of chemical labelling combined with continuous small mass channel mass spectrometric scanning and a systematic matching for NulO unique core features. We applied our approach to a large number of yet unexplored environmental microbes revealing a yet undescribed diversity.

4.2 Results and discussions

Given the limitations of current methods, we sought to establish a universal ‘sialo-omics’ approach for untargeted large-scale molecular studies starting from crude cell lysates (Fig. 1C). However, the search for sub-stoichiometric compounds with an unknown chemical composition in complex mixtures compares to the search for a needle in a haystack.

To tackle this challenge, we searched for mass spectrometric features which are unique for ulosonic acids but are independent of any diversification. Thereby, we took advantage of the alpha-keto-acid specific labelling reagent 1,2-diamino-4,5-methylene dioxybenzene, which is otherwise utilized for fluorescence detection. However, for our purpose, the probe-of-choice was selected since it significantly shifts the double bond equivalents (+7.5), the mass defect, and since it introduces highly conserved ulosonic acid core fragment features which are different from the bulk cell lysate background. Since lower carbon ulosonic acid molecules are highly common in nature, we aimed to establish additional features which contain structural information.

Therefore, we first screened the experimental data acquired in this study for commonly co-occurring fragment compositions using an automated script which confirmed the theoretical fragmentation space (Fig. 2). By doing so, we uncovered universal features which contain the carbon chain length information and which were not influenced by any diversification process (Fig. 2). While the conserved core features are important for identifying ulosonic acids, the carbon chain length features enable differentiation from structurally related eight (KDO) and lower carbon ulosonic acids. However, most importantly, the carbon chain length features provide also information on the degree of saturation and oxidation, and therefore allow differentiation between types of NulOs, such as Neu/Kdn-type and Pse/Leg-type ulosonic acids (Fig. 2A–C). Where this serves as tool to identify derivatives related to known NulOs, the systematic extrapolation of those features makes it universal for the discovery of completely novel types of NulOs. In addition, to maintain high sensitivity for sub-stoichiometric signals, we established a high-resolution mass spectrometric fragmentation of highly small mass segments covering the mass range established by our initial theoretical calculations. The very small mass segments generate hybrid fragmentation spectra of only medium complexity. However, in ‘all-ion-type’ fragmentation techniques, a large number of parent ions contribute to a single spectrum, and the link between parent ions and their product fragment peaks is very difficult to reconstruct.



B

Carbon class

	C-7			C-8			C-9			C-10		
O6	307.09	309.11	311.12	**319.09**	321.11	323.12	331.09	333.11	335.12	343.09	345.11	347.12
O5	289.08	291.10	293.11	**301.08**	303.10	305.11	313.08	315.10	317.11	325.08	327.10	329.11
O4	271.07	273.09	275.10	**283.07**	285.09	287.10	295.07	297.09	299.10	307.07	309.09	311.10
O3	253.06	255.08	257.09	265.06	267.08	269.09	277.06	279.08	281.09	289.06	291.08	293.09

Oxidation

Saturation

Kdo

Ulosonic acid unique CORE			
205.06	217.06	229.06	231.06

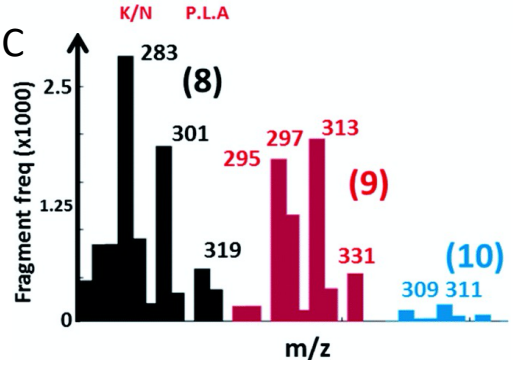


Fig. 2 Conserved and diversification independent feature space for nonulosonic acids (NulOs). (A and B) outline the empirical fragmentation space systematically spanning the oxidation and saturation space of ulosonic acids with different carbon chain lengths. More specifically, (A) outlines (from left to right) one universal fragmentation route for the three main sialic acid compositions (Kdn, Neu and Pse/Leg/Aci). (B) The alpha-keto acid specific chemical labelling introduced a significantly altered mass defect and fragment features such as the (quinoxaline-based) ulosonic acid unique core fragments and carbon chain length features. Deviations in the degree of saturation and oxidation allow further differentiation between types of NulOs and the identification of completely new structures.(C) shows the frequency (>50 counts, see ESI-Table sheet 4‡) of binned fragments from ulosonic acid related compounds observed in the large-scale study. This confirmed the theoretically established fragment feature space for different ulosonic acid classes. The high frequency for 283 and 301 (8 C, black bars) and 295 and 297 (9 C, red bars) correspond to fragments from KDO, Neu/Kdn and P.L.A respectively. Furthermore, this provided mass spectrometric evidence for potential higher carbon chain ulosonic acids (blue).

Nevertheless, identification of the above described conserved ulosonic acid features in mass segments provides very sensitive global maps (channel hits) of the NulO landscape (Fig. 3A and B, ESI-Table, sheet 3‡). The channel hit generation has been automated by a Matlab script, which loops through high-resolution spectra of complete runs within only a few minutes. Nevertheless, molecular studies typically require the exact chemical identity and nature of the NulO. Therefore, we established a chemical filtering and structural evidence scoring pipeline to identify the parent NulO from positively assigned mass segments. In most cases, the small mass segments showed sufficiently low complexity to unambiguously filter for a NulO candidate. To enable higher-throughput studies, the complete process from channel-hit generation to chemical filtering and structural evidence scoring and reporting was automated in a pipeline (ESI‡). Where this provides a very efficient solution to screen classes of NulOs based on the backbone, the closer investigation of fragmentation spectra provides hints on the nature of the modification(s) present.

A molecular level survey across eukaryotes and prokaryotes

Using our newly established pipeline, we performed a first molecular level survey on prokaryotic NulOs where we included a range of well-characterised commercial standards, animal cells, model plants and algae (Fig. 3A and B, ESI-Table, sheet 3‡). To qualify our approach, we first analysed the well-characterised animal-derived cell materials such as CHO and HeLa cells (cultivated on 2 growth media).Apart from low quantities of Kdn, Neu5Ac was identified in all samples as the predominant species, which appeared further diversified through glycolylation, methylation and higher degrees of acetylation. HeLa cells cultivated on foetal calf serum showed a large increase of the *N*-glycolylated variant compared to serum-free conditions, which presumably results from uptake and internal release in lysosomes, followed by metabolic incorporation.

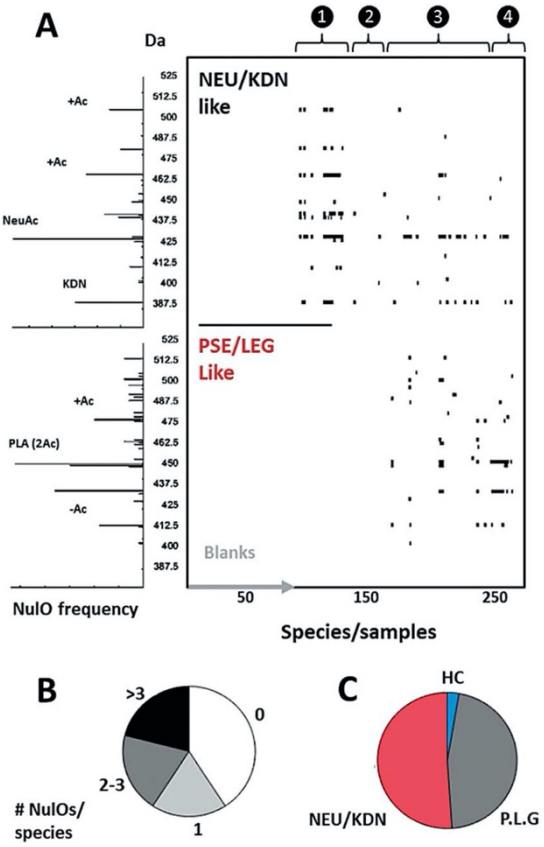


Fig. 3 NulO large-scale survey represented as channel hit-map and chemical diversity graphs. (A) shows a hit-map for channels with detected nonulosonic acid (NulO) candidates in samples (duplicates shown separate for samples) individual injections are plotted along the *x*-axis. The *y*-axis is binned into 2.5 Da mass channel units. The upper map is filtered for Kdn and Neu-like species and the lower map is filtered for Pse/Leg-type hits. The same data are visualised as binned (0.1 *m*/*z*) spectra on the left. The numbers on the top of the hit-map represent: (1) references, spiked *E. coli* and animal cell samples; (2) plants, fungi and algae; (3) *Archaea*, alpha-, beta-, gamma-, delta- and epsilon proteobacteria, actinobacteria, firmicutes, cyanobacteria; (4) enrichment/communities. (B) Sialic acids were found at high frequency throughout environmental samples and almost half of all samples showed two or more types per species. (C) Kdn/Neu-type sugars were observed with comparable frequency to bacterial-type NulOs. While the former were found in eukaryotes and prokaryotes, the bacterial types were exclusive to prokaryotes. Potential higher carbon chain ulosonic acids (blue) were only detected in rare cases. (see also ESI-Table, sheets 1–3‡).

The well-established model plants *Arabidopsis thaliana* and *Nicotiana benthamiana*, as well as baker's yeast, contained lower carbon chain ulosonic acids (not shown), but as expected, did not show any observable nonulosonic acid analogues (9-carbon). Noteworthy, in our experiments, one plant showed misleading artefacts that were difficult to distinguish from Kdn by mass only ('pseudo-Kdn', ESI, Section I‡). Such or related other misleading candidates may have fuelled historical debates about the presence of (active) NulO biosynthesis pathways in plants. Furthermore, we included selected algae, which are increasingly considered as production systems in biotechnology.

In contrast to a recent report which speculated about the presence of sialylated glycans in *Chlamydomonas reinhardtii*, nonulosonic acids were not identified in either of the strains analysed. On the other hand, we could confirm the existence of (genuine) Kdn and associated derivatives in some microalgae. In animal tissues, sialic acids provide a barrier which protects from predators. However, some pathogens, including viruses, have adapted to this strategy and developed specific proteins that bind to sialic acids, to use them as an entry point for successful colonisation of the host. Some other pathogens appeared to have adapted to cleave terminal sialic acids and utilise those as a source of energy. Furthermore, many pathogenic or symbiotic bacteria display sialic acids or other NulOs on their own surface, which is regarded as molecular mimicry of the host sialylation. This strategy has been found effective in delaying, or abolishing the host's immune response.[24] Hence, we included in our survey a well-studied pathogenic *Campylobacter jejuni* strain that is able to utilise bacterial-type and animal-type NulOs. As expected, we observed several heavily expressed hits for both animal- and bacterial-type NulOs.

The majority of Neu variants are expected to derive from medium uptake (and potential incorporation), but the strain further produced at least 3 distinct bacterial-type NulO variants (ESI-Table, sheet 3‡). Moreover, it is suggested that bacteria apply cell surface diversification strategies (such as through NulOs) to evade recognition by bacteriophages.[25] Bacteriophages are viruses that infect and replicate within prokaryotes and attach specifically through bacterial cell surface carbohydrates. For that reason, we selected a range of environmental, non-pathogenic species, spanning most phyla described to date. Surprisingly, the NulOs uncovered from those microbes were rich in diversity and frequency – nonulosonic acid peaks were detected in more than half of the measured species: two-thirds showed more than one hit, and one-half showed three or more NulO hits. The relative sialic acid abundances compared to the mammalian cell lines (normalised to dry weight biomass starting material) was in many cases only slightly lower compared to quantities observed in animal cells. The more rigid bacterial cell walls may, however, make a fair comparison not particularly meaningful (ESI-Table, sheets 1–3‡).

The authors also want to emphasize that incubation with acetic acid at higher temperature is required to ensure sufficient release of different types of NulOs from various biological matrices. On the other hand, the release and labelling procedures may introduce additional degradation variants, which impacts on quantitation and comparability. Furthermore, we aimed to ensure excess of labelling reagent over the biological matrix to ensure efficient NulO labelling. An additional reagent mass channel however might be highly useful to monitor the labelling process when dealing with unknown matrices. Nevertheless, as expected, there was a strong trend towards Pse/Leg-type NulOs, but also Kdn and Neu-type acids were found with high frequency. When animal derived media supplements were used (such as casein tryptone) the detected NeuAc may originate from uptake or scavenging rather from neo-biosynthesis. In some cases, additional Neu derivatives were observed indicating that these did result from further cellular processing. Several environmental microbes, isolated from distinctive niches, displayed a remarkable diversity of NulOs, including *Streptomyces coelicolor*, which is seemingly specialized in abundant Kdn and modified variants thereof, as well as *Paracoccus denitrificans* harbouring at least 4–5 different bacterial-type sialic acid derivatives (ESI-Table, sheets 1–3‡). The authors also would like to point out, that the investigated microbes were analysed from one particular culturing condition, and the extend of variability between conditions and stimuli remains to be explored.

In support of the observed frequency, we analysed the proteomes of selected strains for homologues of known NAB pathways. This largely confirmed observed trends from the molecular identifications, but also underlined the notion that gene and protein sequence databases are still incomplete und selection of comparator sequences is crucial. For example, the homologous genes for Kdn pathways in algae were not identified, although their existence had been confirmed previously in an independent study[26] (Fig. 4).

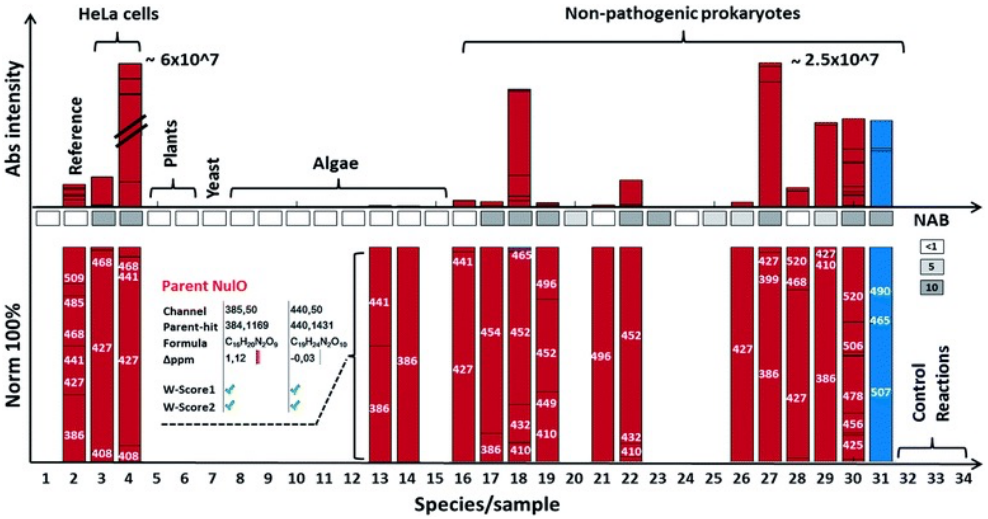


Fig. 4 NulO channel-hits and compositions for selected species. The bar graph highlights results for selected species from the NulO survey. The lower bar graphs are normalised to 100%. Individual bars represent a channel hit (annotated by the channel centre mass). The bar graph size correlates to the MS1 intensity of the aligned parent mass peaks. The following samples are shown from left to right: (1) DMB blank incubation; (2) reference sialic acids spiked into *E coli*, (3) free grown HeLa serum free grown, (4) HeLa grown on serum, (5) *A. thaliana*, (6) *N. benthamiana*, (7) *Saccharomyces cerevisiae*; (8) *Chlamydomonas reinhardtii*-s1; (9) *Chlamydomonas reinhardtii*-s2; (10) *Cryptomonas*; (11) *Asterionella formosa*, (12) *Diatoma*, (13) *Alexandrium ostenfeldii*, (14) *Cricophaera carterae*, (15) *Galdiera sulpharia*, (16) *Haloferax volcanii*, (17) *Geitlerinema* n., (18) Methanogenic archaeon, (19) *Scalindua*, (20) *Nitrospira moscoviensis*; (21) *Clostridium* S. (22) *Thiothrix* enr. (23) *Pseudomonas putida* (24) *E. coli* BW25113 (25) *Myxococcus xanthus*, (26) *Acidovorax avenae*, (27) *Streptomyces coelicolor*, (28) *Mycobacterium smegmatis*, (29) *Arthrospira platensis*, (30) *Paracoccus denitrificans*, (31) *Magnetospirillum gryphiswaldense* (MG) acid hydrolysed + labelled, (32) MG non-hydrolysed + labelled, (33) MG acid hydrolysed + unlabelled, and (34) MG non-hydrolysed + unlabelled. The insert for *Alexandrium ostenfeldii* (13) exemplifies the parent NulO deconvolution from the channel-hit using the chemical and structural evidence scoring and filtering. A comparative gene level NAB pathway search is illustrated in the box between bars (white box = no significant, light grey = <5 and dark grey >5 homologue hits). All survey channel-hits and NulO parent annotations are summarised in the ESI-Table, sheets 1–3,‡ and the NAB pathway search results are summarised in ESI-Table sheet 7.‡

Carbon chain length diversification: from octulosonic, nonulsonic and higher carbon ulosonic acids

All described nonulosonic acids (9 C) are products of evolutionarily related synthase families. A condensation of a hexose derivative with phosphoenolpyruvate (PEP) takes place by a mechanism comparable to the reaction of the Neu5Ac synthase in humans. Alternative (reversible) routes exist through aldolase activities, which utilize pyruvate instead of PEP. A highly comparable condensation reaction produces other lower carbon ulosonic acids such as heptulosonic (7 C) or octulosonic (8 C) acids. Heptulosonic acids are intermediates of the shikimic-acid pathway (deoxy-d-arabino-heptulosonate phosphate, DHAP), an essential part of the neo-biosynthesis of amino acids. Octulosonic acids (keto-deoxy-octulosonate, KDO) are essential building blocks of cell walls of many prokaryotes and plants. Indeed, the large majority of ulosonic acids (UAs) observed in our survey were of modified 7-, 8-, or 9-carbon nature. However, surprisingly, some rare cases evidenced the existence of higher carbon ulosonic acids (potential 10 C, or other by means of MS indistinguishable variants), such as in *Magnetospirillum gryphiswaldense*, a Gram-negative magnetotactic alpha-Proteobacterium, capable of orienting and navigating along geomagnetic field lines[27] (Fig. 2B). Additional experiments on the unlabelled and fractionated NulO further support the presence of a triply acetylated higher carbon ulosonic acid (ESI, Section H‡). The same species was detected in lower abundance in *Methanococcus*. In support of this finding, Hsu *et al.*, reported on directed evolution of sialic acid aldolases with low specificity, that accept a variety of (different) monosaccharide substrates for condensation with pyruvate to uncommon NulOs.[28] Furthermore, Jacobs *et al.* demonstrated the efficient incorporation of mannosamine analogues with alternative *N*-acyl groups.[29] The promiscuousness of biosynthetic routes together with substrate availability may be frequent mediators for species/niche diversification processes of sialic acids and other NulOs.

4.3 Conclusions

In summary, we developed the first universal large-scale survey approach that tackles the enormous chemical diversity of prokaryotic NulOs by using chemical labelling, segmented mass spectrometric scanning and structural evidence filtering and scoring. Comparative large-scale studies will advance the understanding of important molecular level processes relevant to medical applications. Furthermore, we provide a first molecular-level comparative study on selected species representing a large number of the prokaryote phyla described to date, with a focus on the largely unexplored environmental niches. In this study, we observed NulOs at a high frequency and diversity, which challenges the current model of evolution and utilisation of sialic acids and other NulOs being predominantly driven by mechanistic advantages during host–pathogen interactions. This supports a broader utilisation of these compounds, such as for diversification of cell surface attachment points and protection from bacteriophages, which are a major driver of bacterial evolution in the laboratory and in nature.[30] Furthermore, our large-scale data evidence potential higher carbon ulosonic acids, which would further expand the boarders on the chemical diversity of natural NulOs. At the same time, we emphasise, although the approach presented here serves as an ideal tool to survey for known and completely new UAs, the latter require orthogonal analysis, such as by NMR, to completely unravel and confirm their chemical structures.

Nevertheless, chemical synthesis of sialic acid and other NulO derivatives is highly challenging and often only achieved at low yields.[31] The exploration of non-pathogenic microbes for novel biosynthetic routes gives access to new NulO derivatives that have thus far been difficult to produce by chemical synthesis.

.

4.4 References

1. A. L. Lewis , N. Desa , E. E. Hansen , Y. A. Knirel , J. I. Gordon , P. Gagneux , V. Nizet and A. Varki , *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106** , 13552 —13557 CrossRef CAS PubMed .
2. L. Warren *Comp. Biochem. Physiol.*, 1963, **10** , 153 —171 CrossRef CAS .
3. T. Angata and A. Varki , *Chem. Rev.*, 2002, **102** , 439 —469 CrossRef CAS PubMed .
4. A. Varki , R. L. Schnaar and R. Schauer , *Essentials of Glycobiology* , A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar and P. H. Seeberger, Cold Spring Harbor, NY, 2015, pp. 179–195 Search PubMed .
5. M. Cohen and A. Varki , *OMICS*, 2010, **14** , 455 —464 CrossRef CAS PubMed .
6. D. D. Park , G. Xu , M. Wong , C. Phoomak , M. Liu , N. E. Haigh , S. Wongkham , P. Yang , E. Maverakis and C. B. Lebrilla , *Chem. Sci.*, 2018, **9** , 6271 —6285 RSC .
7. G. Blix , L. Svennerholm and I. Werner , *Acta Chem. Scand.*, 1952, **6** , 358 —362 CrossRef CAS .
8. G. Blix *Hoppe-Seyler's Z. Physiol. Chem.*, 1936, **240** , 43 —54 CrossRef CAS .
9. R. Schauer and J. P. Kamerling , *Adv. Carbohydr. Chem. Biochem.*, 2018, **75** , 1 —213 CrossRef PubMed .
10. M. J. Morrison and B. Imperiali , *Biochemistry*, 2014, **53** , 624 —638 CrossRef CAS PubMed .
11. Y. A. Knirel , A. S. Shashkov , Y. E. Tsvetkov , P. E. Jansson and U. Zahringer , *Adv. Carbohydr. Chem. Biochem.*, 2003, **58** , 371 —417 CrossRef CAS PubMed .
12. J. J. Kenyon , A. M. Marzaioli , C. De Castro and R. M. Hall , *Glycobiology*, 2015, **25** , 644 —654 CrossRef CAS PubMed .
13. E. Morello , A. Mallet , Y. Konto-Ghiorghi , T. Chaze , M. Y. Mistou , G. Oliva , L. Oliveira , A. M. Di Guilmi , P. Trieu-Cuot and S. Dramsi , *PLoS One*, 2015, **10** , e0138103 CrossRef PubMed .
14. A. I. M. Salah Ud-Din and A. Roujeinikova , *Cell. Mol. Life Sci.*, 2018, **75** , 1163 —1178 CrossRef CAS .
15. A. F. Carlin , S. Uchiyama , Y. C. Chang , A. L. Lewis , V. Nizet and A. Varki , *Blood*, 2009, **113** , 3333 —3336 CrossRef CAS PubMed .
16. T. Angata and A. Varki , *Chem. Rev.*, 2002, **102** , 439 —470 CrossRef CAS PubMed .
17. A. Klein , S. Diaz , I. Ferreira , G. Lamblin , P. Roussel and A. E. Manzi , *Glycobiology*, 1997, **7** , 421 —432 CrossRef CAS PubMed .
18. S. Hara , M. Yamaguchi , Y. Takemori , M. Nakamura and Y. Ohkura , *J. Chromatogr.*, 1986, **377** , 111 —119 CrossRef CAS PubMed .
19. R. Zeleny , D. Kolarich , R. Strasser and F. Altmann , *Planta*, 2006, **224** , 222 —227 CrossRef CAS PubMed .
20. J.-P. Zanetta , A. Pons , M. Iwersen , C. Mariller , Y. Leroy , P. Timmerman and R. Schauer , *Glycobiology*, 2001, **11** , 663 —676 CrossRef CAS PubMed .
21. A. C. Schrimpe-Rutledge , S. G. Codreanu , S. D. Sherrod and J. A. McLean , *J. Am. Soc. Mass Spectrom.*, 2016, **27** , 1897 —1905 CrossRef CAS PubMed .
22. J. J. J. van Der Hooft , J. Wandy , M. P. Barrett , K. E. V. Burgess and S. Rogers , *Proc. Natl. Acad. Sci.*, 2016, **113** , 13738 —13743 CrossRef PubMed .

23. D. C. P. Lozano , R. Gavard , J. P. Arenas-Diaz , M. J. Thomas , D. D. Stranz , E. Mejía-Ospino , A. Guzman , S. E. Spencer , D. Rossell and M. P. Barrow , *Chem. Sci.*, 2019, **10** , 6966 —6978 RSC .
24. G. Sekot , G. Posch , P. Messner , M. Matejka , X. Rausch-Fan , O. Andrukhov and C. Schäffer , *J. Dent. Res.*, 2011, **90** , 109 —114 CrossRef CAS PubMed .
25. R. Louwen , D. Horst-Kreft , A. De Boer , L. Van Der Graaf , G. de Knegt , M. Hamersma , A. Heikema , A. Timms , B. Jacobs and J. Wagenaar , *Eur. J. Clin. Microbiol. Infect. Dis.*, 2013, **32** , 207 — 226 CrossRef CAS PubMed .
26. B. A. Wagstaff , M. Rejzek and R. A. Field , *J. Biol. Chem.*, 2018, **293** , 16277 — 16290 CrossRef CAS PubMed .
27. K. H. Schleifer , D. Schüler , S. Spring , M. Weizenegger , R. Amann , W. Ludwig and M. Köhler , *Syst. Appl. Microbiol.*, 1991, **14** , 379 —385 CrossRef .
28. C.-C. Hsu , Z. Hong , M. Wada , D. Franke and C.-H. Wong , *Proc. Natl. Acad. Sci.*, 2005, **102** , 9122 — 9126 CrossRef CAS PubMed .
29. C. L. Jacobs , S. Goon , K. J. Yarema , S. Hinderlich , H. C. Hang , D. H. Chai and C. R. Bertozzi , *Biochemistry*, 2001, **40** , 12864 —12874 CrossRef CAS PubMed .
30. B. Koskella and M. A. Brockhurst , *FEMS Microbiol. Rev.*, 2014, **38** , 916 —931 CrossRef CAS PubMed .
31. S. Matthies , P. Stallforth and P. H. Seeberger , *J. Am. Chem. Soc.*, 2015, **137** , 2848 — 2851 CrossRef CAS PubMed .

4.5 Supporting information

Experimental Procedures

A. Cell lysis, acid hydrolysis and alpha-keto acid specific labelling. Liquid cultures were pelleted at 4200 rpm for 10 minutes, and supernatants were removed and washed with PBS. Further, cell pellets and all other solid materials were freeze-dried, followed by physical disruption and homogenisation. Then, 1 mL of 2 M acetic acid solution was added to 2.5 mg of freeze-dried biomass. The resolubilised material was hydrolysed at 80 °C for 2 hours, after which samples were centrifuged for 5 min at 14K rpm. From every sample hydrolysate, 10 µL were dried at 45 °C under reduced pressure using a speed vac concentrator. Labelling was performed by adding 20 µL labelling solution at 50 °C for 2.5 hours. The labelling solution consisted of 1.4 M acetic acid, 0.75 M 2-mercaptoethanol (beta-mercaptoethanol), 18 mM of sodium dithionate and 7 mM of DMB. Reference sialic acid standards were labelled directly and were prepared to reach a final concentration of 2.5 pmol per µL injection solution. MS-grade water was added to the reference panel to reach a concentration of 10 pmol/µl. The labelling procedure follows the original work published by Hara et al.,[1] where 1,2-diamino-4,5- methylene dioxybenzene was introduced as the selective label for alpha-keto acids, forming a fluorescence-active quinoxaline derivative. Here, we decided to select the well-established DMB label (otherwise used for fluorescence detection) because it has a high degree of saturation and therefore the mass defect of labelled ulosonic acids distinguishes significantly from non-labelled metabolites or background signals. Furthermore, the large quinoxaline core guided highly unique ulosonic fragmentation pattern. The reaction is outlined in Figure S2.

B. Reverse-phase-Orbitrap-MS and segmented very small mass window scanning. An M-Class HSS T3 300µm x150mm C18 was mounted to an Acquity M-Class UPLC (Waters) using 97% H2O plus 3% acetonitrile as solvent A and 97% acetonitrile plus 3% H2O as solvent B (both 0.1% formic acid). A gradient from 12% B to 25% B was maintained at a flow-rate of 9 µL/min over 20 minutes, followed by a washing-step. Samples were injected in duplicates followed by blanks. Continuous fragmentation of very small mass segments was performed using a QE plus bench top Orbitrap mass spectrometer, operated in ES+ mode, in 2.5 Da steps from 380–520 Da (2.75 Da width). Fragmentation was performed at a NCE of 28. Alternating MS1 and MS2 scans at a loop count of 51 and a resolution of 70K with an AGC target of 5e5 for MS1, and 17K with an AGC target of 5e4 for MS2, were acquired in centroid mode. Confirmatory high-resolution and targeted experiments were performed at 140K resolution in HRMS, or PRM mode at 1 Da isolation, respectively. Calibration was carried out within every 24 hours.

C. Channel hit detection and NulO parent assignment by chemical filtering and structural scoring. Nonulosonic acid parent candidate identification was performed using Matlab (R2019b). Briefly, raw files were converted to mzxml ('.mzXML') format using the msConvert software tool. Peak lists were imported into Matlab using the 'mzxmlread' function, followed by 'mzxml2peaks' to extract MS1 and MS2 mass lists. MS1 data were deisotoped (within 5 ppm) by removing up to three 13C isotopes, provided that intensity ratios were close to natural abundance ratios (113C=1, 213C<0.25, 313C<0.05, 413C<0.01). Furthermore, MS2 peaks with less than 103 counts were excluded. For each fragmentation channel (2.75 Da window) a cell array was created, containing scan index, mass channel range, retention time, fragment peaks and peak intensities. MS2 scans containing core, class or reporter fragments were extracted (allowing a max mass error of 15ppm, see SI-table, sheet 9, script box 1), and only MS2 scans containing both, class and core fragments, were considered for further processing. For each scan, the class of the ulosonic acid was determined by the carbon length of the largest class fragment (see SI-table, sheet 9, script box 2). Neighbouring MS1 scans were then analysed for potential precursors (see SI-table, sheet 9, script box 3), where only precursor masses with a mass defect between 0.075–0.225 m/z were considered.

For this, an empirical (NulO) chemical composition space was constructed considering min/max element counts (C15-30H16-35N2-5O7-15P0-1S0-1) and constrained ratios between elements and degree of unsaturation (RDBE >7.75, C/H ratio = 0.65–0.9, C/O ratio = 1.25–2.5, which was further corrected for amidation, phosphorylation and sulphation) (see SI-table, sheet 9, script box 4-5).[3] P/O and S/O ratios were set to <0.09, which elements also required the presence of in-source loss fragments of -79.96 Da or -79.95 Da, respectively. All precursors matching the constructed (NulO) chemical composition space, thereby allowing no more than 5 ppm mass error, received a chemical composition and were considered for further processing (see SI-table, sheet 9, script box 6). Precursor candidates occurring in >75% of all scans were regarded as static background and not further considered.

The passed MS1 parent mass candidates were finally subjected to a structural evidence scoring as described in the following (=decision tree). Parent mass candidates were scored based on the occurrence of water loss peaks, either observed as a neutral loss peak in the MS2 scan or as an in-source fragment in the MS1 scan (=Water score, +1 score for MS1/MS2 scans with water loss, [M+H+]+-H2O, - 18.0105). Further, scores were added according to the max. number of observed ulosonic acid (UA) fragments (2. # Fragments, +X scores equal to X = #core + #class-fragments (one per oxidation state)), and for the uniqueness of a parent mass candidate within a MS1 fragmentation window (=Window score, +1 score if parent mass candidate is the only 'realistic candidate'). Additional scores were granted for the fraction of co-elution of a parent mass candidate with the MS2 scans containing UA fragments (=Overlap, expressed as fraction= aligned/total). Parent mass candidates belonging to the same class and chemical composition were combined (intensity, scores, retention times and fragment signatures). In order to provide a comparative measure (to compare to random matches, as described below) a total score was established (=sum of Water score, Window score and # Fragments, multiplied by the fraction of Overlap).

To verify the significance over purely random matches, a 'total score-cut-off, was defined for every sample. For this, every dataset was subjected to semi-randomisation of the MS scan mass peaks and processed through the same pipeline as described above. Randomisation was performed using Matlabs 'rand' function, with the constrain, that randomised masses fall within the window of the lowest and highest masses of the original data (see SI-table, sheet 9, script box 7). The largest total score observed for the randomised data set defined the 'total score-cut-off'. Only parent mass candidates (from the correct dataset) with a total score greater than the 'score-cut-off' were considered for further analysis.
Finally, the minimum threshold for a valid ulosonic parent mass candidate (to be reported in SI-table, sheet 1) required a minimum MS1 intensity of 10^4 counts, a MS1 water loss in-source peak (a water loss peak (MS1 or MS2, during filtering), being matched twice in 2 consecutive runs, a minimum of 3 fragments (at least one class), and in addition either an MS2 water score or a window score. Hits showing strong mono-linker marker peaks as well as degradation makers were rejected. The hits were finally exported using the 'writetable' function. Peaks from the LC gradient wash region were excluded. The complete output of channel-hits and assigned NulO parents for all species/references as analysed above is presented in the **SI-table, sheets 1-3.**

D. Genome-level analysis of ulosonic acid biosynthesis (UAB) pathways. Analysis for homologues of oct- and non-ulosonic acid biosynthetic routes in genomes of species highlighted in Figure 2B was established as described in Lewis et al., 2009.[2] The following protein sequences were used for protein homology search using the NCBI Blast tool: Kdn Q8A710 (NAB1) and Q8A711 (NAB2), Neu P13266/Q8NFW8 (NAB1) and A0A0H3MPX1/Q9NR45 (NAB2); Pse A0A3K5CFB7 (NAB1) and A0A3X8VGM4 (NAB2), Leg Q0P8S7 (NAB1) and Q0P8T1 (NAB2).The Matlab function '[RID1, ROTE]=blastncbi(seq, 'blastp',' Entrez', species, 'expect', 1e-3)' was used to obtain the homology search request ID ('RID1') and estimated search time ('ROTE') for every fasta template (seq) and selected species ('species'). The homology search report was collected using 'report1= getblast(RID1,'WaitTime',ROTE,'ToFile','1CIV_report.xml')'. Results are summarised in **SI-table, sheet 7**.

E. Graphical representation of Channel-hit maps, mass binning graphs and fragment statistics. The theoretical chemical space for NulOs was established by considering the three currently known core compositions (**Figure 1, A**) and by performing a combinatorial addition of up to 4 ('functional') modifications out of 10 frequently found modifications, plus one small mass-shift such as oxidation, dehydration and reduction. To further expand the chemical space and mass defect for yet undescribed diversifications we allowed also for one additional single amino acid and one additional sugar modification (see SI-table, sheet 8, script boxes 1-3). The formula weight for every composition was calculated and masses were binned in 5 Da channels using the 'histogram' function of Matlab. The difference to the nominal mass unit of every composition was binned and visualised in a mass defect histogram (**Figure 1, B**). The established combinatorial space described above is outlined in detail in the **SI-table, sheet 8.**

Initial mining for realistic ulosonic acid species in the large scale data was performed using a Matlab script. Thereby, all acquired spectra were searched for the highly conserved core ulosonic acid DMB fragment features 205.06, 217.06, 229.06 and 231.06. Spectra containing those features were extracted and fragments were binned according to their chemical compositions. Continuous carbon chain fragments with chain lengths of 7, 8, 9 and potential higher, allowing different degrees of oxidation and saturation (=realistic backbone compositions), were extracted and translated into structures (considering continuous and steady fragmentation trees) and visualised using the 'histogram' function of Matlab (SI-table, sheet 4, including script boxes 1-4, **Figure 2 A-C**).

The channel-hit map (**Figure 3**) showing all references/species analysed was prepared using the 'pcolor' function in Matlab. Limits were set so that every valid channel hit (containing a potential parent NulO) was represented by a black/red channel. In-between sample run blanks are shown first, followed by the analysed samples/references. The order of samples shown in the map (from left to right) follows the order listed in the **SI-table, sheet 3**, except that duplicates were not combined, but represented separately. The channel hit-map was once filtered for Neu/Kdn-like signals and once for bacterial-type NulOs across samples analysed. The binning of identified NulOs from every channel was performed using the 'histogram' function of Matlab. Thereby, masses calculated from chemical compositions (of individual runs) were combined within 0.1 Da bins. Selected

species from the survey (**Figure 4**) are presented as a combined graph using Matlabs 'bar' function (stacked). For this, the intensity of the most intense parent ion was used to present the proportion of every channel hit. Bars are annotated by the channel mass (rounded to unit mass). Boxes representing the NAB hits were included manually between the relative and absolute proportions. For sample 13, NulOs are annotated with their chemical composition, mass error and class (Kdn derivatives) and scores.

Additions to results and discussions

F. NulO core compositions. The to-date three distinct core NulO compositions are known, which are outlined in Figure S1. The only composition without any amino functionality is keto-deoxy-neuraminic acid (Kdn, Figure S1, middle). The second (Figure S1, right) refers to N-acetyl-neuraminic acid (Neu5Ac, 5-acetamido-2-keto-3,5-dideoxy-D-glycero-D-galactononononic acid) commonly found in the D- glycero-D-galacto configuration. The third class is a 5,7-diamino variant with an additional deoxy on the 9 position. This composition is commonly found for pseudaminic acid (Pse, 5,7-Diamino-3,5,7,9-tetradeoxy-L-glycero-L-manno-non-2-ulopyranosonic) found in L- glycero-L-manno configuration or its stereoisomers legionaminic acid (Leg, D-glycero-D-galacto configuration, including a 3/8-epilegionaminic acid referred to as 3/8eLeg) as well as acinetaminic (Aci, D-glycero-L-altro configuration 8 epimer, D-glycero-L-altro and a 8 epimer referred as to 8eAci). Diversification can be found on any of the amino and hydroxyl positions. A recent comprehensive summary of a large collection of discovered sialic acids and other NulOs has been collected by Schauer et al.[3]

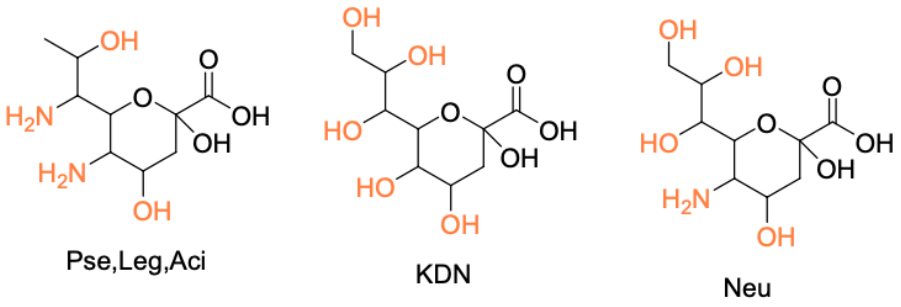


Figure S1. Three most commonly found classes of non-ulosonic acids, including groups undergoing frequent diversification (orange). Common isomerisation points for Pse, Leg and Aci (also referred to as to P.L.A) are not further differentiated in this study.

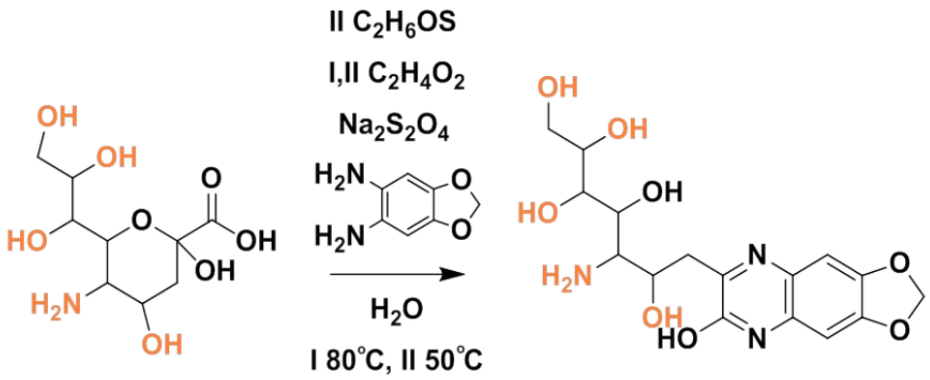


Figure S2. DMB labelling reaction scheme, first described by Hara et al.[1] I) first step includes incubation with 2M acetic acid at 80°C for 2 hours to release glycosidically as well as nucleotide activated sialic acids or other NulOs. II) Alpha keto acid specific labelling using 1,2-diamino-4,5-methylene dioxybenzene at 50°C for 2.5 hours.

G. Sialic acid reference standard spiking experiments. Initial spiking experiments were performed using a mix of 6 commercial sialic acid standards Neu5Ac, NeuGc, Neu(AC)2, Neu(Ac)3, NeuGcAc (Ludger, CatNo CM-SRP-01) and Kdn (Sigma Aldrich, CatNo 60714). All 6 sialic acid standards could be successfully recovered when spiking into E-coli K12 lysate at concentrations of 1:1, 1:5 and 1:1). It was also possible to distinguish from *E. coli*'s octulosonic acid derivatives.

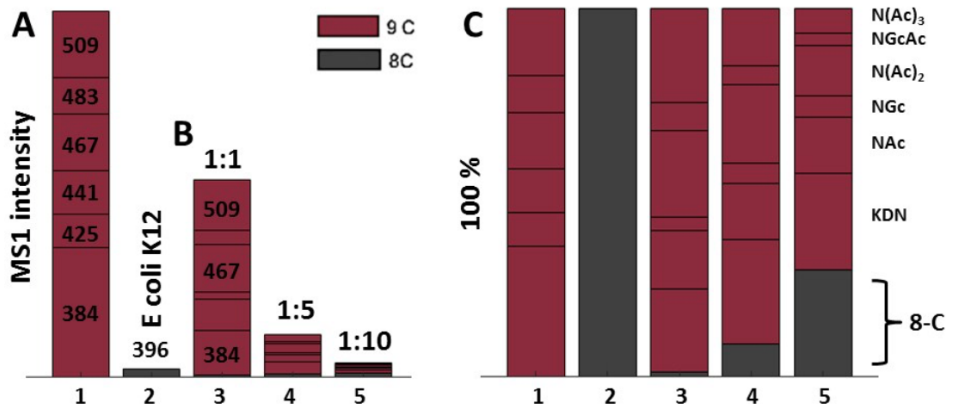


Figure S3. Stacked bar graphs show the observed peak intensities/proportions for reference sialic acids Kdn, NeuAc, NeuGc, NeuAc2, NeuGcAc and NeuAc3 and an 8-Carbon ulosonic acid naturally present in E-coli (presumably Kdo-Ac, the mass analysis did not include the native Kdo mass). A) Bars 1 to 2 represent the (absolute) summed MS1 peak intensities of the 6 sialic acid reference standards and the *E. coli* octulosonic acid derivate measured separately. B) Bars 3-4 show the (absolute) summed MS1 peak intensities of the 6 reference standards, after spiking in ratios 1:1 (3) 1:5 (4) and 1:10 (5) into the *E. coli* extract. The extract was prepared with protocols and biomass quantities, exactly as used for all other samples in this survey. B) shows the same order of samples normalised to 100%.

H. Towards establishing universal fragmentation features. We systematically established a core fragmentation framework for ulosonic acids with different carbon chain lengths ranging from C-8 to theoretical C-10 sugars (C-9+), deoxygenation states and degrees of saturation. Ulosonic acids are identified for which class fragments determine the carbon length, core fragments show attachment to the DMB-label and reporter fragments show side products from DMB-labelling. After an initial water loss peak (-H2O), the neutral losses will be all –N and –O modified side groups. This leaves behind the backbone carbon chain, with many unmodified hydroxyl groups, varying from 0 to 3 -OHs that remain attached to the backbone. Depending on which –O positions are unmodified, fragmentation of the backbone can occur before all hydroxyl groups are fragmented, such as in Kdn. The fragments with the largest complete intact carbon backbone can be used to allocate the carbon length and corresponding class of the ulosonic acid (C-8–C-10). Due to the higher degree of modification, C-9 backbone with 1 –OHs Kdn/Neu core (295.07 m/z) can be differentiated from Pses with a less-saturated backbone (297.09 m/z). Extrapolating the fragmentation route can also be used to include octulosonic acids Kdo, Kdo8N (283.07 m/z), as well as more theoretical chemical space of deoxy Pse/Leg 299.10, deoxy Kdo (285.09 m/z) and larger C-10 sugars (311.09). Almost exclusively, the methylenedioxybenzene proximate C1-hydroxyl remained stabilised and contributed to the 3-carbon core fragments as described further below. Fractionation of the product alkene and alkyne chains led then to losses of 12.00 Da and 27.99 Da corresponding to the loss of C or CO, respectively. The smallest, but unique ulosonic fragments observed (core fragments), were a C-3 (205.061 m/z), C-4 (217.061) and C-5 (229.061 m/z) respectively, and an alternative C-5 fragment for 4-deoxy ulosonic acids (231.076 m/z).

Finally, reporter fragments were determined for detection of mono-labelled species (223.071, 343.114, 283.093, 297.108, 313.103) and DMB label degradation products (201.066, 189.066), which shows chemical instability at the 4,5-methylenedioxy end. Lower mass fragments at 177.07 m/z displayed additional DMB label-related fragmentation events.

Table S1. Empirical conserved fragmentation trees for different kinds of ulosonic acid classes considering varying degrees of oxidation, saturation and carbon backbone length. For every backbone class, we considered a range of 12 indicative fragments. Following the same rationale, potential higher carbon ulosonic acid derivatives e.g. (10-carbon) where considered using 289.06, 291.08, 293.09, 307.07, 309.09, 311.10, 325.08, 327.10, 329.11, 343.09, 345.11 and 347.12. Out of these derivatives, 293 and 311 were most prominently observed in our study. This mass table was used for the screening study to discriminate between different carbon chain length ulosonic acids.

Core	205.06	217.06	229.06	231.08

	C-6			C-7			C-8			C-9			Higher carbon (C-10)		
O6	295.09	297.11	299.12	307.09	309.11	311.12	319.09	321.11	323.12	331.09	333.11	335.12	343.09	345.11	347.12**
O5	277.08	279.10	281.11	289.08	291.10	293.11	301.08	303.10	305.11	313.08	315.10	317.11	325.08	327.10	329.11*
O4	259.07	261.09	263.10	271.07	273.09	275.10	283.07	285.09	287.10	295.07	297.09	299.10	307.07	309.09	311.10
O3	241.06	243.08	245.09	253.06	255.08	257.09	265.06	267.08	269.09	277.06	279.08	281.09	289.06	291.08	293.09

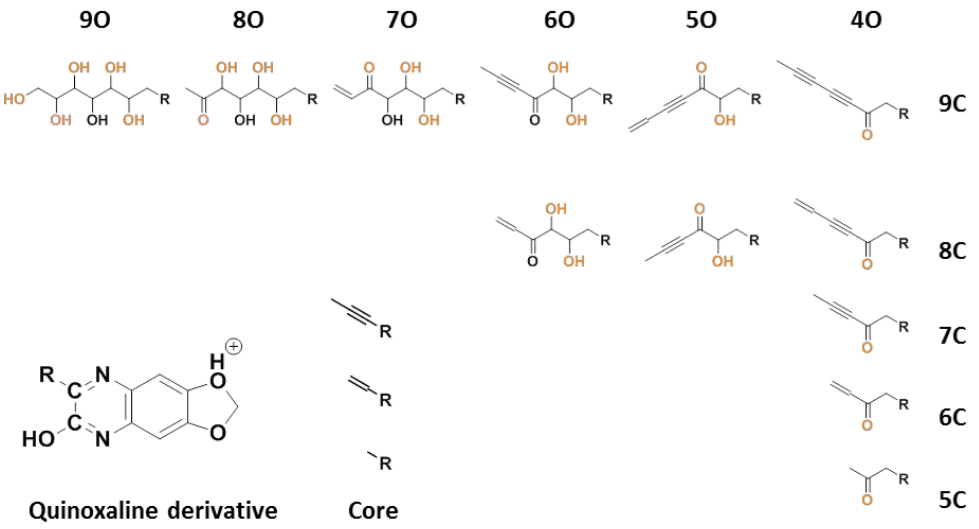


Figure S4. CID fragmentation tree for Kdn including quinoxaline (label) core.

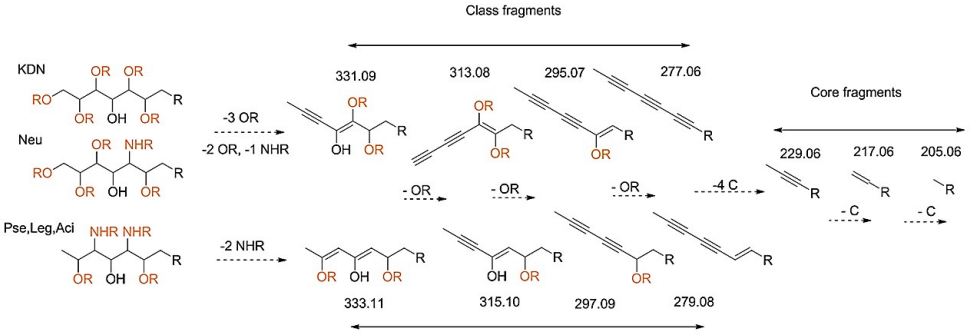


Figure S5. Comparative fragmentation trees of Kdn/Neu and Pse/Leg/Aci, where R represents the quinoxaline label core. The here shown fragmentation tree could also be extended towards other C-9 derivatives, showing highly comparable fragmentation behaviour. The main difference between Pse-type and Neu/Kdn is the 2 Da difference, which is not caused by the number of substitutions reducing the saturation. This difference is seen for the 'Class fragments' because the Pse-type C-9 sugars have an exposed primary carbon. However, this difference did not appear to be influenced by the number of amine versus hydroxyl groups; amine groups were not further considered during fragment mining in automatic data processing. Apart from core fragment and class fragments, reporter fragments were included to assess false positives due to side reactions of the DMB-label, or artefacts generated upon fragmentation of the DMB-labelled compounds.

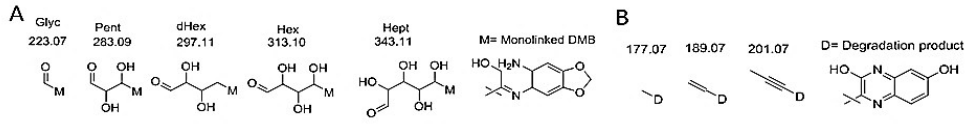


Figure S6. Proposed fragmentation for 'mono-linked' species (A), and proposed instability of the DMB label itself (B).

I. Verification of the empirical fragmentation features. The empirical fragmentation tree was verified with known ulosonic acid standards for Kdo (E-Coli), Kdn (commercial), Neu5Ac (commercial) and Pseudaminic acid (Pse) as obtained from Campylobacter.

Table S2. Kdo C-8 fragmentation markers were as predicted 283.07, 301.08 and 319.09; Neu/Kdn C-9 markers were 295.07, 313.08 and 331.09. For Pse, the C-9 markers were 297.09, 315.10 and 333.11. Pse showed compared to Neu an additional lower oxidation peak of 279.08. The identified peaks for the ulosonic acid form Gryphiswaldense (GW) were 331.10 and 328.13 (amine instead of H2O*). GW showed in addition (similar to Pse) the lower oxidised species 293.09. Due to high levels of acetylation, the further oxygen was only added after addition of 2 acetylations (**).

	C-6			C-7			C-8			C-9			Higher carbon (C-10)		
O6	295.09	297.11	299.12	307.09	309.11	311.12	319.09	321.11	323.12	331.09	333.11	335.12	343.09	345.11	347.12**
O5	277.08	279.10	281.11	289.08	291.10	293.11	301.08	303.10	305.11	313.08	315.10	317.11	325.08	327.10	329.11*
O4	259.07	261.09	263.10	271.07	273.09	275.10	283.07	285.09	287.10	295.07	297.09	299.10	307.07	309.09	311.10
O3	241.06	243.08	245.09	253.06	255.08	257.09	265.06	267.08	269.09	277.06	279.08	281.09	289.06	291.08	293.09

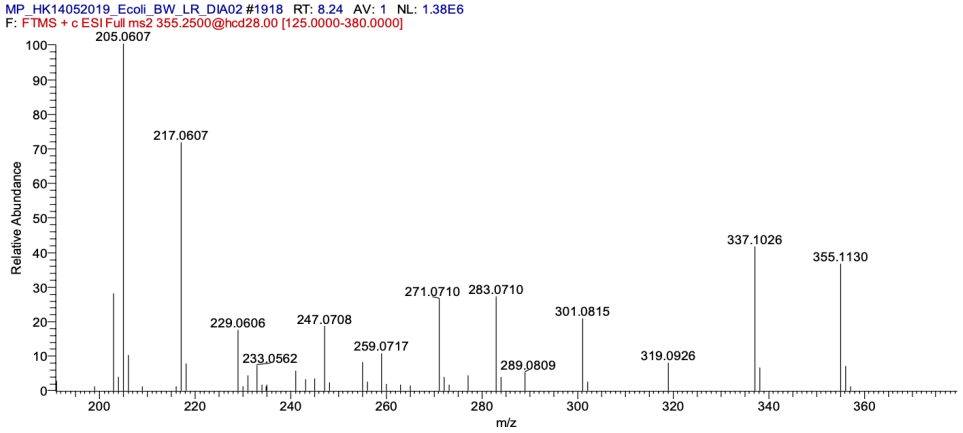


Figure S7a. DMB-Kdo fragmentation profile.

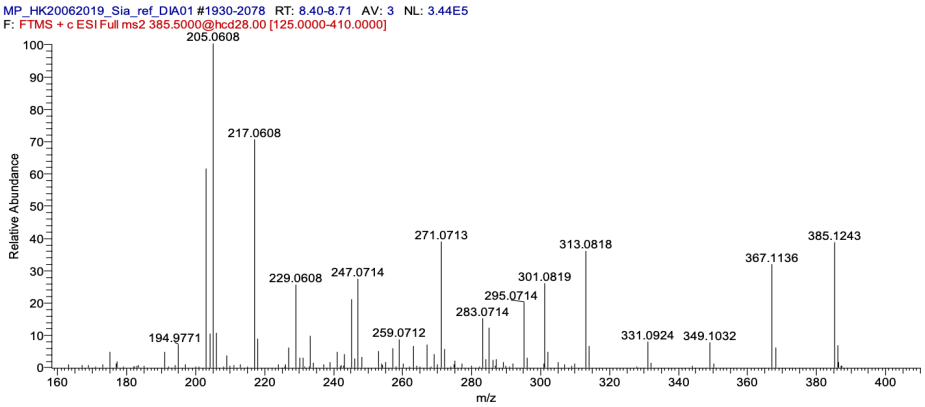


Figure S7b. DMB-Kdn fragmentation profile.

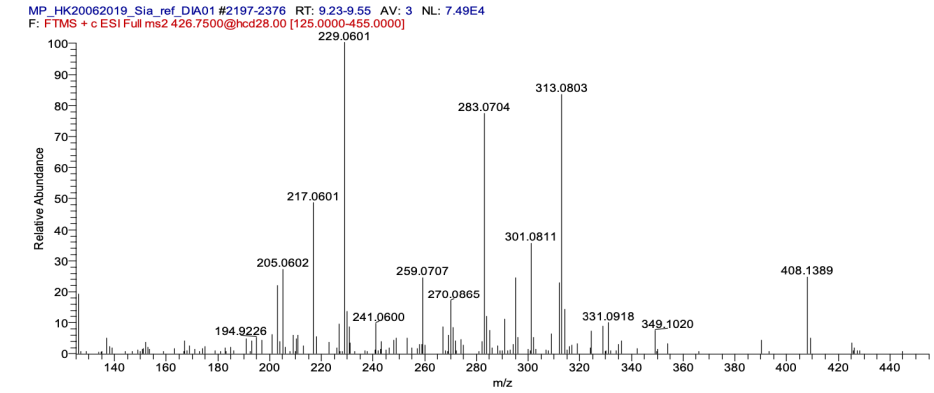


Figure S7c. DMB-NeuAc fragmentation profile.

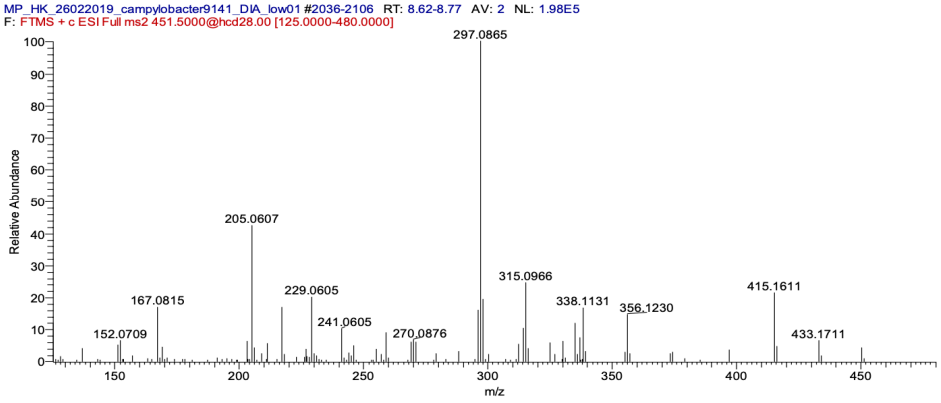


Figure S7d. DMB-Pse fragmentation profile.

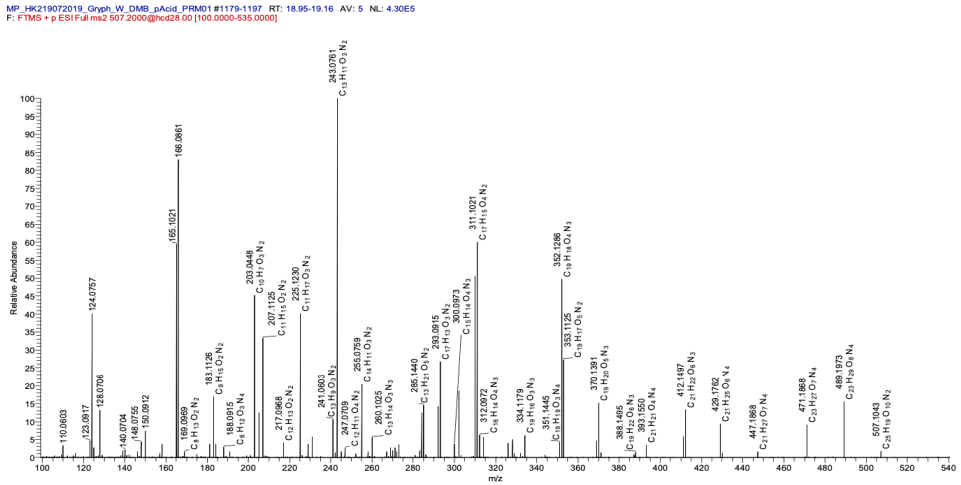


Figure S7e. DMB-UA GW fragmentation spectrum (of 507 m/z) annotated with sum formulae as determined by the Xcalibur chemical composition annotation tool (max mass deviation 10 ppm, DMB core set as min elemental composition)

J. Potential higher carbon ulosonic acids. The ulosonic acid peaks observed for *M. Gryphiswaldense* matched a fragmentation tree of a potential 10 carbon backbone (or any other structure indistinguishable by means of MS). More precisely, three ulosonic acid peaks were observed that matched the UA backbone peaks 293.09 and 311.10 [M+H]+, which correspond to 3 and 4 oxygen-containing species (2 oxygens derive from the DMB label, and one from a hydroxyl group which is part of the unique core quinoxaline fragment). The mass peak with the sum formula C23H30N4O9 (mi = 506.2013, confirmed by high-resolution mass spectrometry) revealed following fragmentation the loss of three acetyl groups. Other losses derived from the loss of water, ammonia and -CH2 groups. No other type of backbone modification was observed that could have increased the carbon count compared to the known ulosonic acid fragmentation features. The other mass peaks observed derived from a water loss of 506.2013 mass peak and from a species with only 2 acetyl groups instead of 3. The latter (twofold acetylated) species was further also observed in the methanogenic bacterium sample (**SI-table, sheet 1**- **3**). The same sample (*M. Gryphiswaldense*) was also measured without prior DMB labelling (plus/minus acid hydrolysis) which did not show any of the abovementioned ulosonic acid peaks. Further, the same sample

was measured without acid hydrolysis, but with DMB labelling, which resulted in only trace quantities of the above mentioned ulosonic acid peaks, only observable after manual investigations of the mass traces.

To confirm the said peaks as ulosonic acids and to obtain additional fragmentation spectra, we analysed the acid hydrolysed and non- derivatised (non DMB labelled) material for the theoretical free ulosonic acid mass peak (predicted sum formula from the AiRM experiment: C16H26N2O9 after subtraction of the label). Thereby, we indeed observed an abundant mass peak corresponding to the predicted accurate mass of the free ulosonic acid and fragmentation profiles similar to those observed for the nonulosonic acid standards. To verify the carbon backbone length, we investigated the fragmentation tree closer and compared it with Neu5Ac (commercial standard) and Pse (Campylobacter jejuni). Since the fragmentation tree of the triply acetylated ulosonic acid from M. Gryphiswaldense (GW) appeared somewhat more complex, we undertook a chemical deacetylation using 20 mM sodium hydroxide. To do so, a microscale quantity of the peak (giving the parent ion 391 [M+H+]+) was fractionated manually from the capillary HPLC, speed- vacuum dried and redissolved in 20 mM aqueous sodium hydroxide. The mixture was incubated at 30°C for 12 hours and (re)injected to the LC-MS system. We performed a targeted analysis for masses of species with 3 (starting material), 2, 1 and no acetylation(s). The fractionated ctrl, as expected, showed only a peak for the threefold-acetylated species, where the (mild) base treated sample showed only a peak for a twofold-acetylated species (loss of O-Ac, but not N-Ac). In the following, the fragmentation tree for the twofold (N) acetylated species was compared to Neu5Ac (commercial, single-NAc, 9 carbon) and Pse (Campylobacter, di-NAc, 9 carbon).

Neu5Ac shows major fragments for a threefold water loss followed by the fragmentation of the N-acetyl group (-42.01 Da -C2H2O or - 59.03Da -C2H5ON). From the later peak (197.04 m/z), the backbone fragmentation occurred from either loss of carboxylic acid (-46.005 Da, -H2CO2) or the cleavage of the C-9 carbon methanol group (-30.01, -CH2O). The fractionation from the resulting peak at 176.03 Da was followed by the carboxylic acid loss (-46.005 Da, -H2CO2) to give a peak at 121.02 m/z.

Pse shows the same major fragments for a threefold water loss followed by the fragmentation of one acetyl group (major -42.01 Da, -C2H2O) and a second N-

acetyl group (major - 59.03Da, -C2H5ON). The resulting C-9 backbone fragment (180.06 m/z) showed further the loss of carboxylic acid (-46.005 Da, -H2CO2) or the cleavage of the C-9 carbon methanol group (-30.01, -CH2O). The fractionation from the resulting peak at 176.03 Da was followed by the carboxylic acid loss (-46.005 Da, -H2CO2) to give a peak at 121.02 m/z.

The (doubly N-acetylated) ulosonic acid from GW showed a comparable (simple) fragmentation tree compared to Neu5Ac and Pse. First, 2 water loss peaks were observed followed by twice a loss of N-Acetyl groups (major -59.03 Da -C2H5ON) to 254.1 m/z and to the 10- carbon fragment 195.06 m/z (C10H11O4). From here, we observed the backbone fragmentation by the loss of the carboxylic acid (-46.005 Da, -H2CO2) to give a 9-carbon peak at 149.05 m/z. A second lower abundant route branching from 2 water loss peaks to the loss of one N-Acetyl group (major -59.03 Da, -C2H5ON) to the peak 254.1 m/z. From there, an (early) CO2 loss to 210.1 m/z (including a minor carboxylic acid loss to 208.09 m/z) was found taking place. This loss was followed by the loss of acetyl to the C-9 168.1 m/z (-42.01 Da, - C2H2O), where the loss of -59.03 Da (-C2H5ON) was much weaker. The (early) loss of CO2 (-43.98 Da) in parallel to the carboxylic acid loss was also observed in Neu5Ac and Pse but was significantly less pronounced.

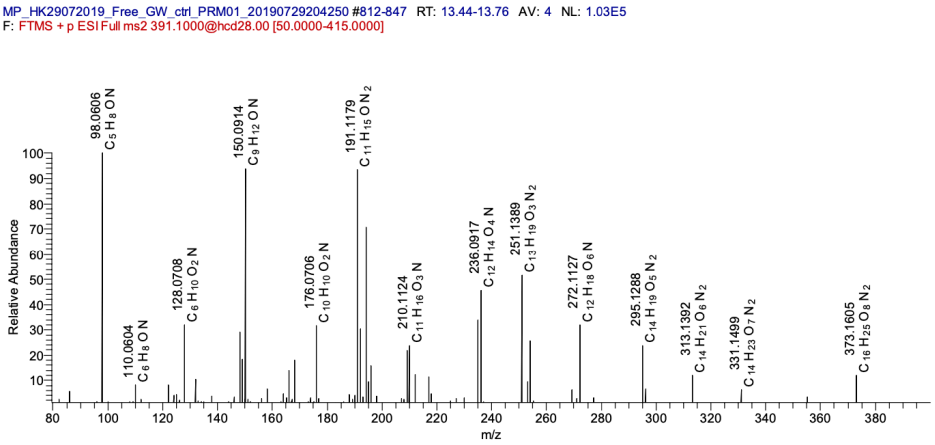


Figure S8a. HCD fragmentation profile of unlabelled ulosonic acid from GW with the proposed sum formula C16H26N2O9 and the theoretical [M+H]+ of 391.1711 Da.

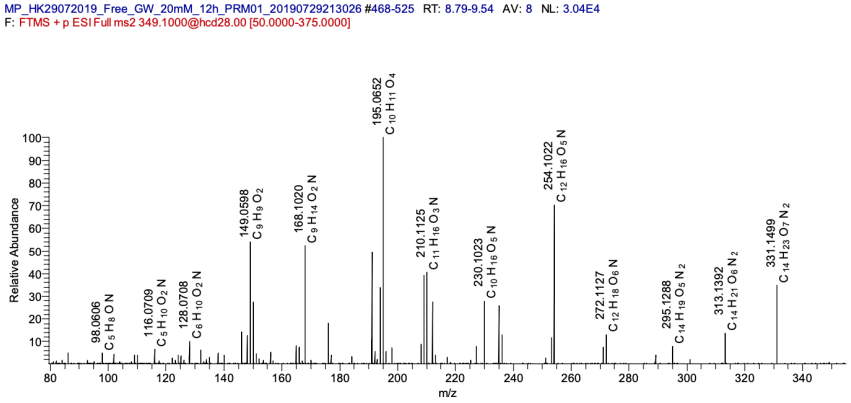


Figure S8b. HCD fragmentation profile of unlabelled and de O-acetylated ulosonic acid from GW.

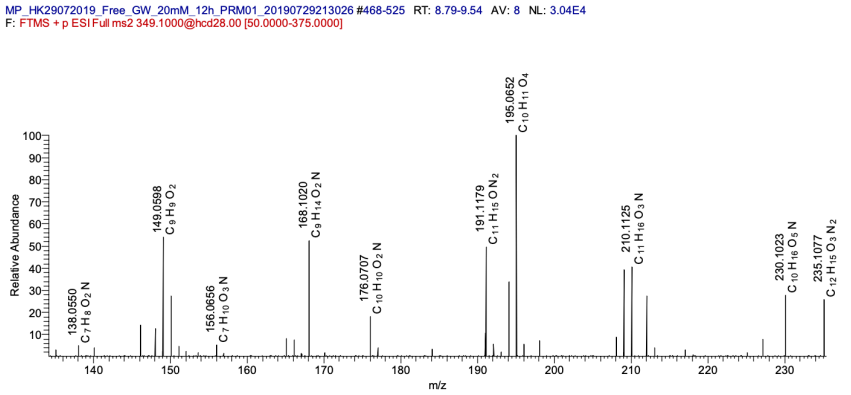


Figure S8c. HCD fragmentation profile of unlabelled and de O-acetylated ulosonic acid from GW, with zoom to the C-10 peak which fragments to a C-9 following the loss of carboxylic acid.

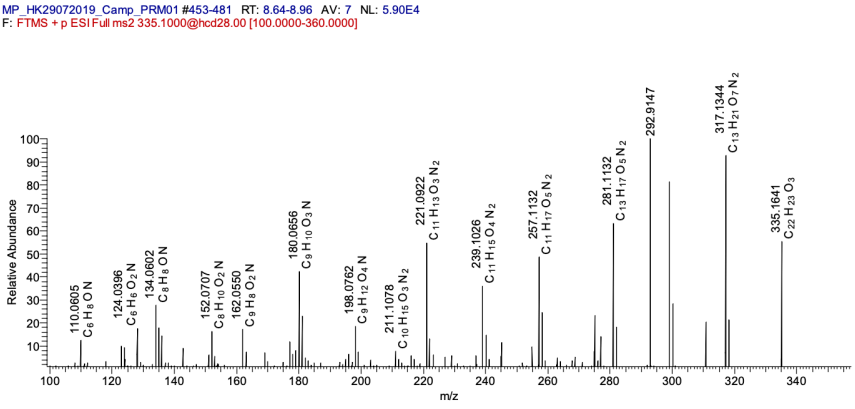


Figure S8d. HCD fragmentation profile of unlabelled Pse from Campylobacter jejuni.

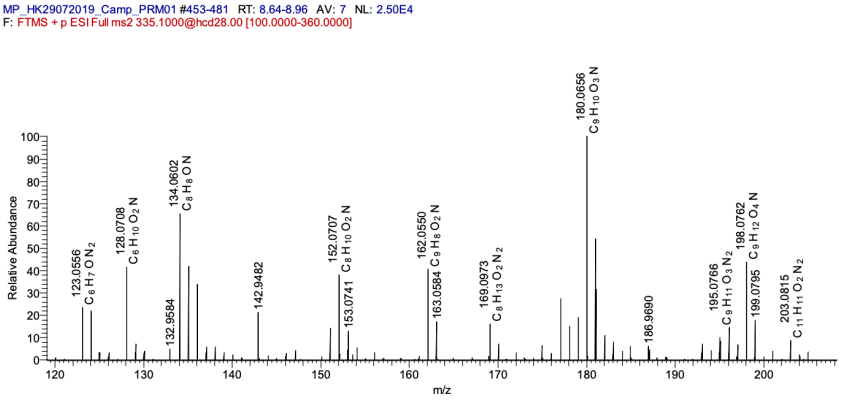


Figure S8e. HCD fragmentation profile of Pse from Campylobacter, with zoom to the C-9 peak which fragments to a C-8 by carboxylic acid loss.

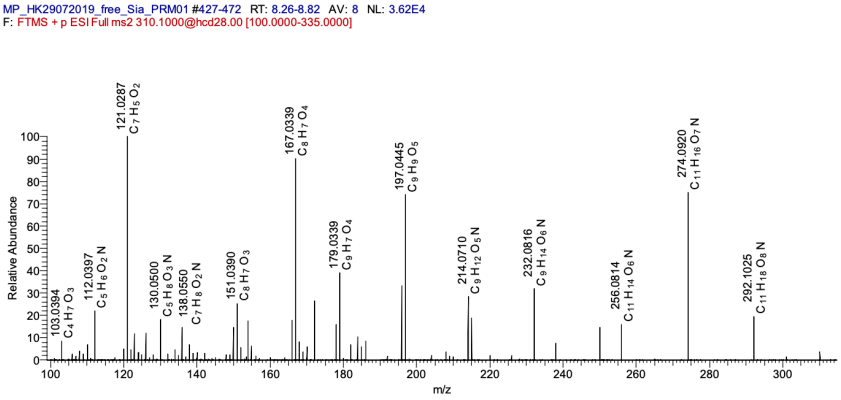


Figure S8f. HCD fragmentation profile of the commercial, unlabelled Neu5Ac standard.

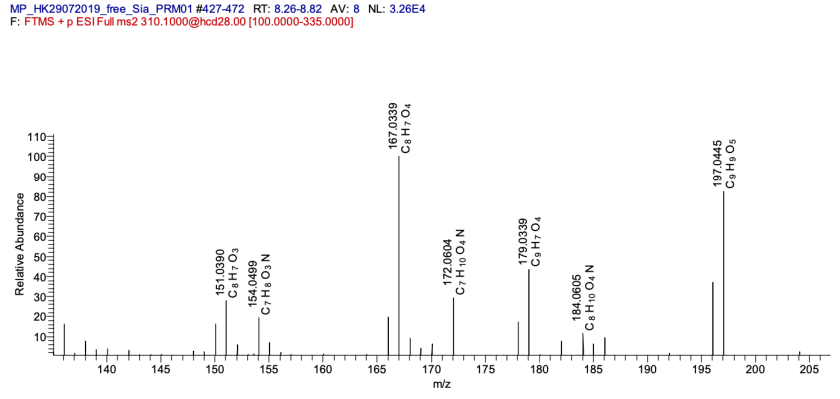


Figure S8g. HCD fragmentation profile of Neu5Ac, with zoom to the C-9 peak which fragments to a C-8 by carboxylic acid loss.

K. Degradation and non-specific labelling side products. Most lysates showed additional artefacts deriving from nonspecific labelling side reactions with other sugars and that unfortunately showed upon fragmentation the ulosonic acid core fragments 205.061, 217.061 and 229.061 [M+H+]+. However, to distinguish from actual hept-, oct- and nonulosonic acids, we included the 'Reporter ions 1' that were established through initial validation experiments using common bulk monosaccharides. In addition, we included in-source fragments of common bulk monosaccharides such as for 162.053, 146.058 and 132.042. Spectra showing reporter ions and in-source fragments at high frequency were rejected, or marked as possible labelling artefacts. Furthermore, chemical instability of the DMB derivative may produce low abundant artefacts peaks. A presumable hydrolysis of the methylenedioxy group followed by elimination would result in a mass decrease by 28 Da. Unfortunately, higher fragments may mimic the intact quinoxaline core fragments. For example degraded DMB- Neu5Ac would appear with the mass of potential 'N-methyl-neuraminic acid' and could only be distinguished by fragmentation tree analysis. Therefore, we included in our fragmentation trees the 'Reporter ions 2' 177.066, 189.066 and 201.066 [M+H+]+. Spectra were rejected, or marked as degradation products when all 3 reporter ions were present at high frequency.

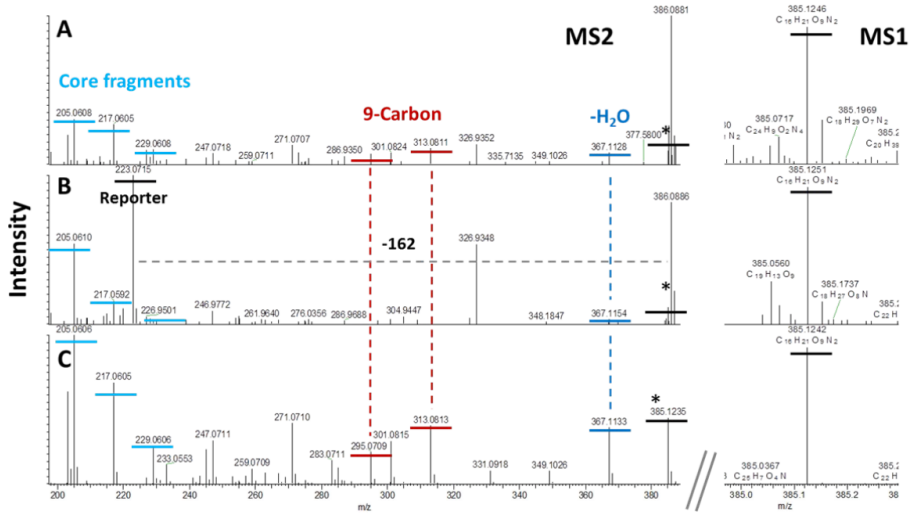


Figure S9. Non-specific derivatisation (artefacts) mimic ulosonic acid mass peaks, including core ulosonic acid fragment ions, as shown for a "pseudo-Kdn" peak C16H21O9N2. Spectra could be differentiated from genuine peaks by the carbon number fragments and reporter ions. A) Genuine Kdn from Cricosphaera carterae, B) 'pseudo-Kdn' peak observed in Nicotiana benthamiana and C) commercial Kdn standard.

References

[1] S. Hara, M. Yamaguchi, Y. Takemori, M. Nakamura, Y. Ohkura, *J Chromatogr* **1986**, *377*, 111-119.

[2] A. L. Lewis, N. Desa, E. E. Hansen, Y. A. Knirel, J. I. Gordon, P. Gagneux, V. Nizet, A. Varki, *Proc Natl Acad Sci U S A* **2009**, *106*, 13552-13557.

[3] R. Schauer, J. P. Kamerling, *Adv Carbohydr Chem Biochem* **2018**, *75*, 1-213.

5

Outlook

5.1 The snake that eats its own tail

It is often claimed that the number of stars in the universe equals the number of microbial species on earth (Dykhuizen 1998). Likewise, our knowledge of microbial diversity can be seen as a universe that is continuously increasing in entropy until its eventual death. For the universe, a state of death is reached when all energy is dispersed, while for ecological knowledge, a point of "death" would be reached when the vastness of total sequence space completely halts our ability to annotate taxa. In this sense the current explosion of sequence data is like a snake eating its own tail, similar to the classical symbol of the ouroboros, which can signify both a self-destructive and cyclic behavior. The more our knowledge increases, the more difficult it will be to place new findings, which would make the accelerated discovery of new species its own worst enemy. As knowledge increases, the surrounding framework will have to be iteratively reborn to categorize these new insights, as is happening at currently with the creation of new taxonomic databases. Another association of the ouroboros is "εν το παν" (The all is one), which can be interpreted as the more knowledge is gained, the more connections are formed between entities, which will end up dissipating boundaries between bodies of knowledge. The 3 main challenges that I expect to threaten future ecological developments are 1. Computational burden, 2. Reduced sequence uniqueness, and 3. Phylogenetic placement, each of which will require us to reconsider how we treat and classify our information.

5.1.1 Computational burden

Since sequence growth is significantly outpacing Moore's law at the moment, we will continuously have to change our approaches to accommodate such large databases, like data compression, representative databases or the use of "de novo" database independent methods. To exemplify: during the time of my thesis, the UniprotKB protein database has tripled in size. We will have to think about minimizing the data that will be compared to the ever-growing reference. In metagenomics, taxonomic annotation often was done by assembly with read-mapping sequences to reference genomes (Nasko 2018) but is currently outpaced by de novo assembly. Similarly, de novo metaproteomics methods can circumvent using large databases for peptide spectrum matching, which is computationally

taxing. Another strategy involves data reduction before using database annotation to minimize the total amount of data that is matched. In metagenomics annotation, binning is often performed before taxonomic annotation. Data reduction is also standard in 16S analysis. Either by clustering sequences into operational taxonomic units (OTUs) or denoising them into amplicon sequence variants (ASVs). Data reduction strategies are not commonly applied with metaproteomics, though spectral clustering pipelines have been developed (Bandeira 2007, Frank 2008).

5.1.2 Sequence uniqueness

An additional side effect of increasing reference database size is that sequence uniqueness decreases. This has significant effects on lowest common ancestor algorithm and requires us to shift towards longer sequences to improve the specificity of our sequences. We see this in 16S sequencing where currently the emphasis is put on increasing amplicon length, shifting to full length 16S sequences to enable species level annotations, and even adding ITS and 23S regions to extend the analyzed ribosomal region. In metagenomics, long-read sequencing is also gaining traction as it has several other benefits and improves the generation of fully circular metagenome assembled genomes (MAGs). For metaproteomics a potential step would be to increase peptide length. Especially for de novo sequencing this will be a major step, as here peptides are matched against an entire reference database. Shifting from trypsin to separately digested Lys-K and Arg-C could be a relatively easy way to accomplish longer overall peptide length (Wu 2018). Though not included in this thesis, the influence of different proteases on peptide uniqueness was explored by bachelor student Jelle Keijzer. However, moving towards longer sequences also generally results in more errors per read, which will limit methods that annotate by exact sequence matches. To this extent Ramon van Valderen researched error-tolerant alignment for de novo sequencing in his master thesis, which was also not included in this thesis.

5.1.3 Phylogenetic placement

While a species definition in macroorganisms is based on the ability to reproduce, a species of microorganisms was based on arbitrary genotypical cutoffs, which shifted from 70% DNA:DNA hybridization (Brenner 1973) to 97% 16S rRNA (Stackebrandt 1994) and more recently to a 95% average nucleotide identity (Konstantidinis 2005, Hugenholtz 2021). This means that distance metrics form a continuous scale, which enables normalization of evolutionary distance. However, as the tree of life ends up getting filled, continuous species cutoffs will become meaningless as the boundaries between groups will result in overlapping species groups, from which no tree can be constructed. Still, many of the recent expansion is instead added to existing genera (Nasko 2018). The normalization of GTDB started introducing suffixes to existing ranks (e.g. g__Nitrospira A,C,F), splitting them into subclusters (Parks 2017), showing that more intricate grouping is required to separate taxa.

Additionally, it should be reconsidered whether an evolutionary system of classification (phylogenetic) is desired over a functional or trait-based system (phenetic/cladistic). Events such as horizontal gene transfer and microdiversity stemming from mutations or genomic rearrangements can cause massive changes in functionality (Nelson 2015), while strain-to-strain variations can determine key traits such as pathogenicity, each of which is poorly reflected within evolutionary distance. Before the advent of genotypical phylogeny, phenetic mathematical taxonomy with automated systems was on the rise (Sokal 1966, Sneath 1962). While mass spectrometry has the potential to enable mathematical taxonomy by quantifying expressed functions, the fact remains that functionality remains conditional. Like the pupa and the butterfly, that share the same phylogeny (Sneath 1995), microorganisms can exhibit growth cycles, dramatic examples being stigmatella and myxococcus, but can also have various states of being and modes of operation. Should the different states of being then be separate entities in a phenetic system? As the presence and absence of pathways does not equate functionality (ppk and PAO), it is also a question if this classification should be discrete, which minimizes conditional effects, or continuous by relying on similarity between expression profiles.

5.2 A coming of age

What is a worrying trend in the field in -omics is that while most people agree that it is not an established field, the rate at which people are generating data makes it seem otherwise. There is yet to be consensus within single methodologies as evidenced by CAMI and CAMPI studies (van den Bossche 2021, Sczyrba 2017), and on quality requirements for database inclusion (Bowers 2017, Chen 2020), which leads to a flood of imperfect data, resulting in database discrepancies between 16S and metagenomics data. The high time-investment required for mastering an –omics technique leads to people getting blindsided to the potential bias of their employed techniques. A recent discovery that highlighted this is that of candidate phyla radiation (Brown 2015) and DPANN (Rinke 2013), two phyla of ultra-small symbiotic microorganisms that are barely found by 16S sequencing. However similar results concerning large sulphur bacteria have largely flown under the radar (Salman 2012). This reeks of survivorship bias, where a method is continuously applied and built upon, instead of continuously challenged and teared down. As potentially a lot of organisms remain unsequenced because of extraction bias in DNA-based methods (McIlroy 2008), and differential binning bias (Miller 2016), the same is also true for proteomics, where posttranslational modifications and use of non-canonical amino acids can affect detection (Den Ridder 2020). While control samples exist in the form of idealized mock-communities, they poorly approximate true community complexity. As each methodology has varying sensitivity and noise, a tradeoff remains between measuring low-abundant signals and artifacts from erroneous sequencing. Since each technique has its own “dark matter” we will need to be wary of unidentified signals. A key factor in the maturation of the field will remain the implementation of control experiments, quality control metrics and in-silico models, to assess the fraction of missing signals.

5.3 The future of metaproteomics

Compared to metagenomic methods, metaproteomics is lagging behind. It still mainly employs methodologies designed for regular pure culture proteomics, and has remained stagnant for over a decade in its number of identified proteins per sample. The main limitation of metaproteomics at the moment is its throughput. The complex separation of peptides in mixed communities leads to co-eluting peptides, which greatly limits the detection of low abundant organisms in complex samples (Mccain 2019). While this is likely to be an inherent aspect of chromatographic separation and shotgun methods, there are several angles that should be explored to improve resolution, on an experimental level, an acquisition level, and a data processing level.

As discussed prior in the outlook, an experimental improvement would be the increase of peptide length, to create middle-down proteomics (Cristobal 2017), either with incomplete digestion or the use of different proteases. Which would result more unique peptides with less mass and sequence overlap. On the other end of the spectrum is nanopore sequencing of proteins, which has the potential to sequence full length proteins at very high rates, and also measure low abundant organisms. However, this is still far from being realized, as there is limited ability to distinguish amino acids and their modifications (Restropo-Perez 2018). On the acquisition level, data-independent methods (Gillet 2012, Tsou 2015, Pietilä 2020) are promising, as they can potentially sequence more peptides, and circumvent the bias of shotgun fragmentation. Spectral deconvolution will remain a challenge, which is currently done with graph-based methods, but seems a prime target for artificial intelligence. As an alternative, small tweaks with exclusion lists to reduce the impact of shotgun bias should be explored. With data processing there are several angles for improvement, including automated detection of glycosylation and other modifications (Pabst 2021), de novo sequencing (Kleikamp 2021) combined with homology searching (Schevchenko 2006) or spectral clustering (Frank 2008).

Within the field, the “metaproteomic iceberg” is often discussed, as the shotgun nature of metaproteomics often tends to focus on the most abundant proteins, and struggles with increasing community complexity (Lohmann 2020). The question remains which fraction of the total diversity is represented. This presents a challenge on its own, since currently only simple mock communities exists to validate communities. Creating a mock community that approximates actual microbial complexity (>1000 species) is unfeasible from a practical point of view. Though not included within this thesis, master student Ramon van der Zwaan created a model to mimic metaproteomics experiments, with the aim to uncover the limits of community complexity in metaproteomics. However, to develop a generally applicable model, further research would be required into the prediction of retention times, ionization and fragmentation.

In the long run, metaproteomics could be integrated with orthogonal experiments to enable de novo functional annotation without the use of metagenomics. Once peptide length and coverage have been improved, de novo assembly of peptide sequences into proteins could be a possibility. This could in turn enable true database independent annotation when combined with structure-based functional annotation tools such as Alphafold or Rosetta (Jumper 2021, Rohl 2004). Added value could be gained with integrative approaches, such as metabolomics, glycomics and stable isotope probing. Metabolomics combined with molecular docking could help understand more about substrate utilization and enable high-throughput enzyme discovery platforms.

5.4 Concluding remarks

While genomic methods have booked great progress, there is a need for orthogonal analysis methods that help investigate phenotypes of microorganisms. Mass spectrometry is not only very powerful but also very flexible will carry us to distant lands. However, we have our work cut out for us as we are just beginning on this journey.

5.5 References

- Bandeira, Nuno, et al. "Protein identification by spectral networks analysis." Proceedings of the National Academy of Sciences 104.15 (2007): 6140-6145.
- Bowers, Robert M., et al. "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea." Nature biotechnology 35.8 (2017): 725-731.
- Brenner, Don J. "Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria." International Journal of Systematic and Evolutionary Microbiology 23.4 (1973): 298-307.
- Chen, Lin-Xing, et al. "Accurate and complete genomes from metagenomes." Genome research 30.3 (2020): 315-333.
- Cristobal A, Marino F, Post H, van den Toorn HW, Mohammed S, Heck AJ. Toward an Optimized Workflow for Middle-Down Proteomics. Anal Chem. 2017;89(6):3318-3325. doi:10.1021/acs.analchem.6b03756
- Dykhuizen, Daniel E. "Santa Rosalia revisited: why are there so many species of bacteria?." Antonie van Leeuwenhoek 73.1 (1998): 25-33.
- Frank, Ari M., et al. "Clustering millions of tandem mass spectra." Journal of proteome research 7.01 (2008): 113-122.
- Gillet, Ludovic C., et al. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis." Molecular & Cellular Proteomics 11.6 (2012).
- Hugenholtz, Philip, et al. "Prokaryotic taxonomy and nomenclature in the age of big sequence data." The ISME Journal (2021): 1-14.
- Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.
- Kleikamp, Hugo BC, et al. "Database-independent de novo metaproteomics of complex microbial communities." Cell Systems 12.5 (2021): 375-383.
- Konstantinidis, Konstantinos T., and James M. Tiedje. "Genomic insights that advance the species definition for prokaryotes." Proceedings of the National Academy of Sciences 102.7 (2005): 2567-2572.
- Lohmann, Patrick, et al. "Function is what counts: how microbial community complexity affects species, proteome and pathway coverage in metaproteomics." Expert review of proteomics 17.2 (2020): 163-173.
- McCain, J. Scott P., and Erin M. Bertrand. "Prediction and consequences of cofragmentation in metaproteomics." Journal of proteome research 18.10 (2019): 3555-3566.
- Miller, Ian J., et al. "Single sample resolution of rare microbial dark matter in a marine invertebrate metagenome." Scientific reports 6.1 (2016): 1-10.
- Nasko, Daniel J., et al. "RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification." Genome biology 19.1 (2018): 1-10.
- Nelson, William C., et al. "Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia." Applied and environmental microbiology 82.1 (2016): 255-267.
- Pabst, Martin, et al. "A general approach to explore prokaryotic protein glycosylation reveals the unique surface layer modulation of an anammox bacterium." The ISME Journal (2021): 1-12.
- Restrepo-Pérez, Laura, Chirlmin Joo, and Cees Dekker. "Paving the way to single-molecule protein sequencing." Nature nanotechnology 13.9 (2018): 786-796.
- den Ridder, Maxime, Pascale Daran-Lapujade, and Martin Pabst. "Shot-gun proteomics: Why thousands of unidentified signals matter." FEMS yeast research 20.1 (2020): foz088.

- Rohl, Carol A., et al. "Protein structure prediction using Rosetta." Methods in enzymology 383 (2004): 66-93.
- Sczyrba, Alexander, et al. "Critical assessment of metagenome interpretation—a benchmark of metagenomics software." Nature methods 14.11 (2017): 1063-1071.
- Shevchenko, Andrej, et al. "Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching." Analytical chemistry 73.9 (2001): 1917-1926.
- Sneath, Peter HA, and Robert R. Sokal. "Numerical taxonomy." Nature 193.4818 (1962): 855-860.
- Sneath, Peter HA. "Thirty years of numerical taxonomy." Systematic Biology 44.3 (1995): 281-298.
- Sokal RR. Numerical taxonomy. Sci Am. 1966;215:106–17.
- Stackebrandt, Erko, and Brett M. GOEBEL. "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology." International journal of systematic and evolutionary microbiology 44.4 (1994): 846-849.
- Tsou, Chih-Chiang, et al. "DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics." Nature methods 12.3 (2015): 258-264.
- Wu, Zhen, et al. "Lys-C/Arg-C, a more specific and efficient digestion approach for proteomics studies." Analytical chemistry 90.16 (2018): 9700-9707.

Acknowledgements

Firstly, I'd like to thank the external committee **Jean-Marc, Marc**, **Marcel, Thijs** and **Tom**, for reading this thesis, and making it past the 2nd chapter.

Dear supervisors, thank you for supporting me throughout this journey. **Martin**, you started your position at TU Delft some months after I started, but I am happy to have joined you. It was clear that you were very eager, both from your ideas and your dedication. However, while aiming high with your science, you also aimed to be a good supervisor, one that was patient and considerate. I am proud to be your first PhD student in Delft and that we could grow together. **Yuemei**, working with you in my master thesis is what motivated me to go into science, and I'm grateful that you gave me the opportunity to continue at EBT. Your creative approach to science is inspiring, and sorely needed in this explorative area of research. **Mark**, thank you for creating a healthy and free-thinking environment. You always prioritize personal needs over protocols, and I will take your advices to heart.

Dear students, since my supervised thesis topics were explorative, most projects were challenging and frustrating, so thank you for hanging in there. **Leanne** as my first victim, **Pim** with his contagious enthusiasm, **Melati** with her attention to detail, **Suzanne** and **Jelle** with their perseverance during early covid, and "**The Ramones**", very different but equally talented, I hope you continue in science.

EBT, thank you for being a special place. A hippie group with communal feeling is easily taken for granted. When I started it had a great atmosphere, but it got even more fun with the arrival of the **Spanish** (also and those that speak Spanish, Portuguese, Greek, Italian, or French). thank you for bringing such "southern" vibrancy to EBT. From BBQs with amazing food that showcased true synergy to chaotic Friday drinks, I am sure that the next generation will carry on these traditions. My only wish is for you to have more lunches outside.

Lastly, I'd like to thank my parents **Enny** and **Ben** for being my safe haven and support during the weekends, and my dear paranymphs **Luuk** and **Stefan** for doing such a great job reading my propositions.

Table 1. Some quick shoutouts in alphabetical order:

Name:	Reason:
Anna-Maria	for trying to pronounce my name
Chris	for being a big bully
Danny	for his inferior synths
Duncan	for his inferior staff
Felipe	for joining me outside
Florence	for being nice
Jules	for his inferior beers
Luuk	for buying me fries
Maria-Paula	for enjoying my beer
Mariana	for ojeeguaposito
Martin	for his many bottles of champagne
Maxime	for the coffee breaks
Phillip	for the mexicaners
Stefan	for letting me win in chess
Sue and Mau	for books beers and movies
Team Team Team	for the explosions, bad smells and 2e pek
Wijb	for the great meals

Curriculum Vitae

Hugobert Bernhard Crijn Kleikamp was born in Den Haag, The Netherlands on July the 3rd 1993. Initially focusing on arts and literature during high school, his scientific curiosity diverted him towards natural sciences. He completed his Bachelor's in Life Science and Technology in Leiden University and Delft Technical University in 2014 and became the first to follow this up with a scientific double degree MSc in both chemical and biotechnological Life Science and Technology programs. In 2017 Hugo started his PhD thesis at TU Delft, which centered around the development of computational strategies for mass spectrometry on microbial communities. Currently, Hugo is continuing this research in Aalborg, where he is employed as Postdoc.

Publications within thesis

Kleikamp, H. B., Lin, Y. M., McMillan, D. G., Geelhoed, J. S., Naus-Wiezer, S. N., Van Baarlen, P., ... & Pabst, M. (2020). Tackling the chemical diversity of microbial nonulosonic acids–a universal large-scale survey approach. *Chemical science*, *11*(11), 3074-3080.

Kleikamp, H. B., Pronk, M., Tugui, C., da Silva, L. G., Abbas, B., Lin, Y. M., ... & Pabst, M. (2021). Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems*, *12*(5), 375-383

Kleikamp, H. B., Grouzdev D., Schaasberg P., Van Valderen R., Van der Zwaan R., … & Pabst, M. (2022). Kleikamp, Hugo BC, et al. "Comparative metaproteomics reveals the divergent views on the complex granular sludge microbiome", *BioRxiv*

Scientific contributions

Felz, S., **Kleikamp, H.**, Zlopasa, J., van Loosdrecht, M. C., & Lin, Y. (2020). Impact of metal ions on structural EPS hydrogels from aerobic granular sludge. *Biofilm*, *2*, 100011.

Boleij, M., **Kleikamp, H.**, Pabst, M., Neu, T. R., van Loosdrecht, M. C., & Lin, Y. (2020). Decorating the anammox house: sialic acids and sulfated glycosaminoglycans in the extracellular polymeric substances of anammox granular sludge. *Environmental science & technology*, *54*(8), 5218-5226.

Pinel, I. S., **Kleikamp, H. B.**, Pabst, M., Vrouwenvelder, J. S., van Loosdrecht, M., & Lin, Y. (2020). Sialic acids: An important family of carbohydrates overlooked in environmental biofilms. Applied Sciences, 10(21), 7694.

Tomás-Martínez, S., **Kleikamp, H. B.**, Neu, T. R., Pabst, M., Weissbrodt, D. G., van Loosdrecht, M., & Lin, Y. (2021). Production of nonulosonic acids in the extracellular polymeric substances of "Candidatus Accumulibacter phosphatis". *Applied microbiology and biotechnology*, *105*(8), 3327-3338.

Pabst, M., Grouzdev, D. S., Lawson, C. E., **Kleikamp, H. B.**, de Ram, C., Louwen, R., ... & Laureni, M. (2022). A general approach to explore prokaryotic protein glycosylation reveals the unique surface layer modulation of an anammox bacterium. *The ISME Journal*, *16*(2), 346-357.

Patents

Van Loosdrecht MC, Lin Y, Pabst M, **Kleikamp HB**, inventors; Technische Universiteit Delft, assignee. Production of biomedical compounds by enrichment cultures of microorganisms. United States patent application US 17/288,463. 2021 Dec 16.