NovoLign

Metaproteomics by sequence alignment

Kleikamp, Hugo B.C.; van der Zwaan, Ramon; van Valderen, Ramon; van Ede, Jitske M.; Pronk, Mario; Schaasberg, Pim; Allaart, Maximilienne T.; van Loosdrecht, Mark C.M.; Pabst, Martin

DOI
10.1093/ismeco/ycae121

Publication date
2024

Document Version
Final published version

Published in
ISME Communications

ISME

ISME Communications, 2024, **4(1)**, ycae121

https://doi.org/10.1093/ismeco/ycae121

Advance access publication: 12 October 2024

Original Article

NovoLign: metaproteomics by sequence alignment

Hugo B.C. Kleikamp[1,2,*], Ramon van der Zwaan[1,‡], Ramon van Valderen[1,‡], Jitske M. van Ede[1], Mario Pronk[1], Pim Schaasberg[1], Maximilienne T. Allaart[1,3], Mark C.M. van Loosdrecht[1], Martin Pabst[1,*]

[1]Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, Delft 2629HZ, The Netherlands
[2]Present address: Department of Biology, University of Antwerp, Universiteitsplein 1, Antwerp, 2610 Wilrijk, Belgium
[3]Present address: Department of Geosciences, University of Tuebingen, Tuebingen, Germany

*Corresponding authors: Martin Pabst, Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, Delft 2629HZ, The Netherlands. Email: m.pabst@tudelft.nl and Hugo B.C. Kleikamp, Department of Biology, University of Antwerp, Universiteitsplein 1, Antwerp, Wilrijk 2610, Belgium. Email: hugo.kleikamp@uantwerpen.be

‡Ramon van der Zwaan and Ramon van Valderen contributed equally to this work.

Abstract

Tremendous advances in mass spectrometric and bioinformatic approaches have expanded proteomics into the field of microbial ecology. The commonly used spectral annotation method for metaproteomics data relies on database searching, which requires sample-specific databases obtained from whole metagenome sequencing experiments. However, creating these databases is complex, time-consuming, and prone to errors, potentially biasing experimental outcomes and conclusions. This asks for alternative approaches that can provide rapid and orthogonal insights into metaproteomics data. Here, we present NovoLign, a *de novo* metaproteomics pipeline that performs sequence alignment of *de novo* sequences from complete metaproteomics experiments. The pipeline enables rapid taxonomic profiling of complex communities and evaluates the taxonomic coverage of metaproteomics outcomes obtained from database searches. Furthermore, the NovoLign pipeline supports the creation of reference sequence databases for database searching to ensure comprehensive coverage. We assessed the NovoLign pipeline for taxonomic coverage and false positive annotations using a wide range of *in silico* and experimental data, including pure reference strains, laboratory enrichment cultures, synthetic communities, and environmental microbial communities. In summary, we present NovoLign, a *de novo* metaproteomics pipeline that employs large-scale sequence alignment to enable rapid taxonomic profiling, evaluation of database searching outcomes, and the creation of reference sequence databases. The NovoLign pipeline is publicly available via: https://github.com/hbckleikamp/NovoLign.

Keywords: metaproteomics, sequence alignment, de novo sequencing, microbial communities

Introduction

Microorganisms form complex multi-species communities in nature, inhabiting virtually every ecological niche on Earth. They are involved in global biogeochemical cycles and significantly affect human health and well-being [1–4]. Therefore, understanding these complex microbial ecosystems is required to accelerate the development of effective biotechnological solutions to address global challenges [5–8]. This requires advanced metaomics approaches to study them [9–12]. Among these techniques, metaproteomics is particularly powerful as it allows to determine expressed metabolic functions, the protein biomass composition, and species-to-species interactions [11–25]. However, compared to traditional single organism proteomics, metaproteomics faces the challenge of analyzing highly complex systems with potentially hundreds of species present at varying abundances [13, 22]. Consequently, metaproteomic experiments depend on reference sequence databases that accurately cover all organisms present in the microbial community [26–30]. Current strategies for constructing metaproteomic databases involve comprehensive generic reference databases, such as the NCBI nr, RefSeq, or UniRef. However, large databases require focusing of the sequence space to allow sensitive metaproteomics experiments. This can be achieved by using empirical knowledge about the microbial niche or by incorporating taxonomic information from other techniques such as 16S ribosomal ribonucleic acid (rRNA) sequencing [31–34]. More recent focusing approaches employ also multi-round searches and *de novo* sequence information. For example, iterative and two-step search approaches were employed, where the results of a primary search were used to create a smaller database for subsequent searches. This approach improved the handling of large databases and increased sensitivity [32, 33], albeit with the risk of introducing false positive matches [35]. A novel deep learning-based *de novo* tool KAIKO was employed, along with Diamond, to perform *de novo* sequencing and sequence matching against a large generic reference database [36]. However, currently, only exact sequence matches were considered for constructing a reference sequence database from genomes of matching species. MetaNovo generates short sequence tags which are then mapped to large generic databases [37]. The matching proteins were further used to construct a focused reference sequence database for subsequent DB searching, which peptide spectrum matches were analyzed for taxonomic uniqueness using Unipept [38]. These tools were primarily designed to focus databases for subsequent searches. Nevertheless, the ability to

Received: 9 June 2024. **Revised:** 3 September 2024. **Accepted:** 10 October 2024

annotate sequences from species not present in the database at different taxonomic ranks is commonly limited, and the capacity to match peptides with *de novo* sequencing errors has either not been thoroughly evaluated or is not feasible. Moreover, these tools do not evaluate database searching outcomes, including metagenomic database coverage, and spectral quality of metaproteomic experiments.

Generic reference sequence databases may also not encompass all organisms, strains, and sequence variants present in the community, which compromises accuracy and coverage. Therefore, for most applications the preferred method to create reference sequence databases is whole metagenome sequencing of the target community [13, 26, 39]. However, whole metagenome sequencing experiments are complex, expensive, and time-consuming. The coverage of the metagenomic database can be biased by the employed deoxyribonucleic acid (DNA) extraction method, sequencing errors, and variations between data processing pipelines [22, 27]. Consequently, metaproteomic experiments are highly influenced by the database construction procedure. Unfortunately, there are currently no dedicated pipelines available for evaluating the outcomes of metaproteomic experiments that use database searching and metagenomic reference sequence databases.

However, peptide *de novo* sequencing allows to annotate mass spectrometric fragmentation spectra with amino acid sequences in a database-independent manner [40–43]. Therefore, *de novo* sequencing has been already employed where reference sequence database are not available, such as for determining the amino acid sequence of antibodies [44, 45]. *De novo* sequencing has been also used to access the suitability of reference sequence files in proteomics experiments [46], and to string-search large public reference sequence databases to obtain taxonomic information from a microbial community [36, 47]. For example, the recently developed NovoBridge pipeline automates quality evaluation and matching of *de novo* sequences to a precomputed peptide database (Unipept) to provide rapid taxonomic information from complex samples [38, 47]. However, the quality of mass spectrometric sequencing data, *de novo* sequencing errors, and incomplete reference sequence databases limit the number of sequencing spectra that provide confident taxonomic information. Furthermore, using precomputed peptide databases require continuous update of the sequences and hamper the use of alternative proteolytic enzymes or the search for amino acid modifications.

To overcome the limitations posed by *de novo* sequencing errors and incomplete reference sequence databases, *de novo* sequence tags can also be matched to reference databases using sequence alignment. This has already been employed in tools such as PeptideSearch, CIDentify, MS-BLAST, MS-HOMOLOGY, FASTS, MS-Spider, and OpenSea, among others [48–55]. However, performing sequence alignment of large volumes of data against very large reference sequence databases is computationally highly demanding, and not suitable for processing complete metaproteomics experiments. In 2015, Buchfink *et al.* introduced the sequence aligner DIAMOND, which allows to align large volumes of data on standard desktops and eases the use of custom reference sequence databases [56]. This laid the foundation for creating a sequence alignment tool capable of handling large volumes of metaproteomic sequencing data with custom databases.

In this study, we introduce NovoLign, a *de novo* metaproteomics pipeline that classifies *de novo* sequences from complete metaproteomic experiments using sequence alignment against very large reference sequence databases. The pipeline aligns *de novo* sequences (including decoys) from complete metaproteomics experiments against generic reference sequence databases in short time frames on standard desktop personal computers (PCs). The sequence alignment pipeline overcomes challenges posed by *de novo* sequencing errors and incomplete reference sequence databases. The fraction of false positive classifications are estimated using randomized sequences. This metagenomics-independent approach allows to perform rapid taxonomic profiling with deep coverage, and to evaluate the sequence coverage of conventional metaproteomics experiments that employ database searching. Furthermore, the obtained taxonomic composition can be used to construct or complement reference sequence databases. We demonstrate the performance of NovoLign with a large spectrum of pure reference strains, enrichment cultures, synthetic, and complex natural communities.

Materials and Methods

Application of publicly available data

Employed pure reference strains, enrichment cultures synthetic and natural communities, taxonomic lineages, and content of synthetic community samples are summarized in SI EXCEL Table 1. Briefly, the equal protein synthetic community proteomic raw data and reference database were obtained from ProteomXchange server project PXD006118 (Kleiner *et al.*, 2017, Nat Commun) [24], the simplified human gut microbiota model (SIHUMIx) proteomic raw data and reference database were obtained from ProteomXchange server project PXD023217 (Van den Bossche *et al.*, 2021, Nat Commun) [13]. *Acinetobacter baumannii* raw data and reference database were obtained from PXD011302 (Di Venanzio *et al.*, 2019, Nat Commun) [57], *Caldalkalibacillus thermarum* from PXD042369 (de Jong *et al.*, 2023, Front Microbiol) [58], *Nitrospira moscoviensis* from PXD019583 (Lawson *et al.*, 2021, mSystems) [59], *Chlamydomonas reinhardtii* from PXD010160 (Scholz *et al.*, 2019, Plant Journal) [60], *Lactobacillus sakei* from PXD011417 (Prechtl *et al.*, 2018, Front Microbiol) [61], *Paracoccus denitrificans* from PXD013274 (von Borzyskowski, 2019, Nature) [62], *Streptococcus mutans* from PXD006735 (Ahn *et al.*, 2017, Scientific Reports) [63], *Halanaeroarchaeum* sp. HSR-CO from PXD028241 (Sorokin *et al.*, 2022, The ISME Journal) [64], the *Clostridium kluyveri*-dominated enrichment, preparation and analysis was performed as described in Allaart *et al.*, (2021), Front. Bioeng. Biotechnol. [65], and Allaart *et al.*, (2023) Scientific Reports PXD040972 [66]. The raw data for the following samples are provided via PXD050548. More specifically, the *Saccharomyces cerevisiae* tryptic digest was purchased from Promega (cat no. V7461) and analyzed as described in Pabst *et al.*, The ISME Journal, 2020 [67], except using a shorter 1D 60-minute gradient for chromatographic separation. The preparation and analysis of *Aeromonas bestiarum* is described in Tugui *et al.*, 2024, bioRxiv [68]. The Ca. Kuenenia stuttgartiensis enrichment was prepared and analyzed as described in Lawson *et al.*, 2019, The ISME Journal [69]. The Ca. Accumulibacter phosphatis enrichment was prepared and analyzed as described in Kleikamp *et al.*, 2021, Cell Systems [47]. Aerobic granular sludge was sampled from the wastewater treatment plant in Utrecht, The Netherlands, and prepared and analyzed as described in Kleikamp *et al.*, Water Research, 2023 [22]. The MetaP reference sample measured with an Orbitrap Astral mass spectrometer was obtained from Dumas and co-workers (PXD045838) [70]. KO (KEGG Orthology) terms for the aligned and database-search-matched proteins of the Orbitrap Astral dataset were obtained using GhostKoala [71]. All analyzed proteomics and metaproteomics reference sample data are summarized in SI EXCEL Table 1.

De novo sequencing and database searching of mass spectrometric raw files

The mass spectrometric raw data were processed using PEAKS Studio X (Bioinformatics Solutions Inc., Canada) [72] for database search and *de novo* sequencing, or DeepNovo [73] for obtaining additional *de novo* sequencing output files for developing the pipeline. Both *de novo* sequencing and database searching were performed allowing 20 ppm parent ion and 0.02 Da fragment mass error. Carbamidomethylation was set as fixed and methionine oxidation as variable modifications. Database searching was performed with N/Q deamidation as additional variable modifications. Database searching further used decoy fusion for estimation of false discovery rates (FDRs) and subsequent filtering of peptide spectrum matches for 1% FDR. Only the top-ranked *de novo* sequence annotations were considered for processing. Proteome reference sequence databases for comparative database searching were obtained from the provided ProteomeXchange server projects or published supplementary data. The proteome reference sequence database for the MetaP reference dataset was constructed by combining UniProt genomes from the organisms listed in the SI data (40168_2024_1766_MOESM3_ESM.xlsx) of Dumas *et al.* [70]. The proteome reference database for the *C. kluyveri* dominated enrichment and the wastewater sludge was obtained from whole metagenome sequencing experiments. The aerobic granular sludge was fractionated by size and then homogenized, following the protocol described in Kleikamp *et al.*, 2023, Water Research [22]. For both samples (*C. kluyveri* enrichment and wastewater granule fractions), DNA was then extracted using a DNeasy UltraClean Microbial Kit (Qiagen, Germany), and the extracted DNA was quantified with a Qubit fluorometer. Whole metagenome sequencing was performed on an Illumina NovaSeq platform with paired-end reads (Novogene Co. Ltd, China). Raw reads were quality checked, trimmed and then assembled using MEGAHIT (v1.0.4-beta). Thereafter, open reading frames (ORFs) were predicted with MetaGeneMark (v3.05) using default parameters, for scaftigs ≥500 bp. Finally, redundancy in the predicted ORFs was eliminated using CD-HIT (v4.5.8). The individual metagenomic databases of the granule size fractions were merged before employing a two-round database search approach. All other samples, including the *C. kluyveri* enrichment, were analyzed using the above outlined database searching procedure.

Training sequences with *de novo* errors and mutations

The developed Python code and additional documentation to generate peptides with simulated *de novo* sequencing errors and mutations are freely available via GitHub: https://github.com/hbckleikamp/De-Novo-ErrorSIM/. *De novo* error files were created for "equal mass substitutions", "inversions of amino acids", and "Other" mutations [41]. Equal mass substitutions involve substituting a combination of one or more amino acids with another combination of amino acids that have the same mass. This, for example, includes the substitution of asparagine (N) with two glycine residues (GG). Equal mass combinations of up to 6 amino acids were created, where a sliding window detected substitutable regions within a peptide, which are then selected for substitution at a 25%, 50%, and 100% chance. For inversions, the peptide sequence was divided into fragments of two or three consecutive amino acids. Each fragment was iteratively tested for a 5% chance of randomization. If a randomization event occurred, the order of amino acids within the fragment was randomized. Finally, the fragments were reassembled to form the inverted sequence. The "Other" category included mutations where amino acid sequences were reverse-translated into trinucleotide codons. Each nucleotide had a 1% chance of being substituted by a different nucleobase. The mutated nucleotide sequences were then translated back into amino acid sequences to form the final mutated sequence. However, larger peptides are more likely to contain multiple *de novo* sequencing errors. To make sure that each error type generated similarly different peptides, the Levenshtein distance between the original and altered peptide was computed [74]. While inversions and mutations can occur in any position, equal mass substitutions can only occur if certain combinations of amino acids are present. Therefore, if similarly "different" peptides need to be generated, they need to occur at a higher error rate. Various error rates were tested for the different error types to create a similar distribution to 5% mutation rate. Final error rate was selected as 5% for mutation and inversion, and 25% for substitution errors. The individual *de novo* error sequences were combined using rates of occurrence for the individual error types as observed for common sequencing tools recently [41] (substitution of 1 by 1 or 2 AAs = 6.3%, substitution of 2 by 2 AAs = 13.7%, substitution of 3 by 3 AAs = 6.3%, substitution of 2 by 3 AAs = 3.7%, substitution of 4 by 4 AAs = 9.7%, substitution of 5 by 5 AAs = 7.6%, substitution of 6 by 6 AAs = 6.8%, inversion of 2 or 3 AAs = 16.1%, other = 29.7%).

Outline of NovoLign pipeline

The NovoLign pipeline is a single "tunable" python script in which parameters can be altered manually. The pipeline can be used with the commonly employed proteomic reference sequence databases including NCBI, UniprotKB, and GTDB. The pipeline, additional documentation and example data are freely available via GitHub: https://github.com/hbckleikamp/NovoLign. Setup of the pipeline and databases is described in the online documentation of the GitHub page. NovoLign was tested with output file formats from PEAKS Studio X [72] and DeepNovo [73]. Nevertheless, any tabular or .txt-like format can be supplied, provided it contains a column of peptide sequences with the header "Peptide". Supplying database-searched peptides (PSM file) and a reference sequence database (FASTA file with NCBI TaxIDs) will furthermore allow the evaluation of spectral and database coverage. The pipeline comprises the following modules:

1. DIAMOND alignment (write_to_fasta.py, diamond_alignment.py): firstly, peptide sequence lists are imported into the Python environment and filtered based on *de novo* score thresholds. Default filters include a minimum quality score of 70 ALC% for PEAKS data or −0.1 minimum score for DeepNovo. Optional (additional) filters include maximum ppm mass error, minimum peptide length, and peak intensity or area. Filtered sequences are written to a FASTA file. Decoy peptide sequences are created by scrambling the order of amino acids in front of the cleavage site (R or K) of every sequence. The resulting peptide file is aligned using the sequence aligner DIAMOND [56], using parameters optimized for *de novo* sequencing errors. Default parameters include a minimum percent identity of 85%, a minimum coverage of 80%, and a PAM70 matrix with gap opening penalty 2 and gap extension penalty of 4.
2. Lowest common ancestor (LCA) analysis (process_alignment.py): output files from sequence alignment are filtered for a minimum bitscore of 25. The obtained taxonomy IDs are then mapped to lineages for subsequent LCA

analysis. NovoLign includes 3 algorithms for LCA analysis, which contain different levels of stringency and false positives. The conventional LCA algorithm (CON) has the highest stringency and retains all aligned sequences. The weighted LCA algorithm (W) [75] counts the frequencies of all aligned taxonomies, which act as weights. For each peptide, the taxonomic weights of aligned sequences are sorted, cumulatively summed, and normalized to 1. The most frequent taxonomies are kept for LCA until the "weight_cutoff" (default 0.6) is reached. The bitscore LCA (BIT), operates similarly, but instead groups the aligned sequences by their taxonomic lineages for each ranks and sums the bitscores. Only taxa with a weight over the "weight_cutoff" (default 0.6) are retained [47, 76].

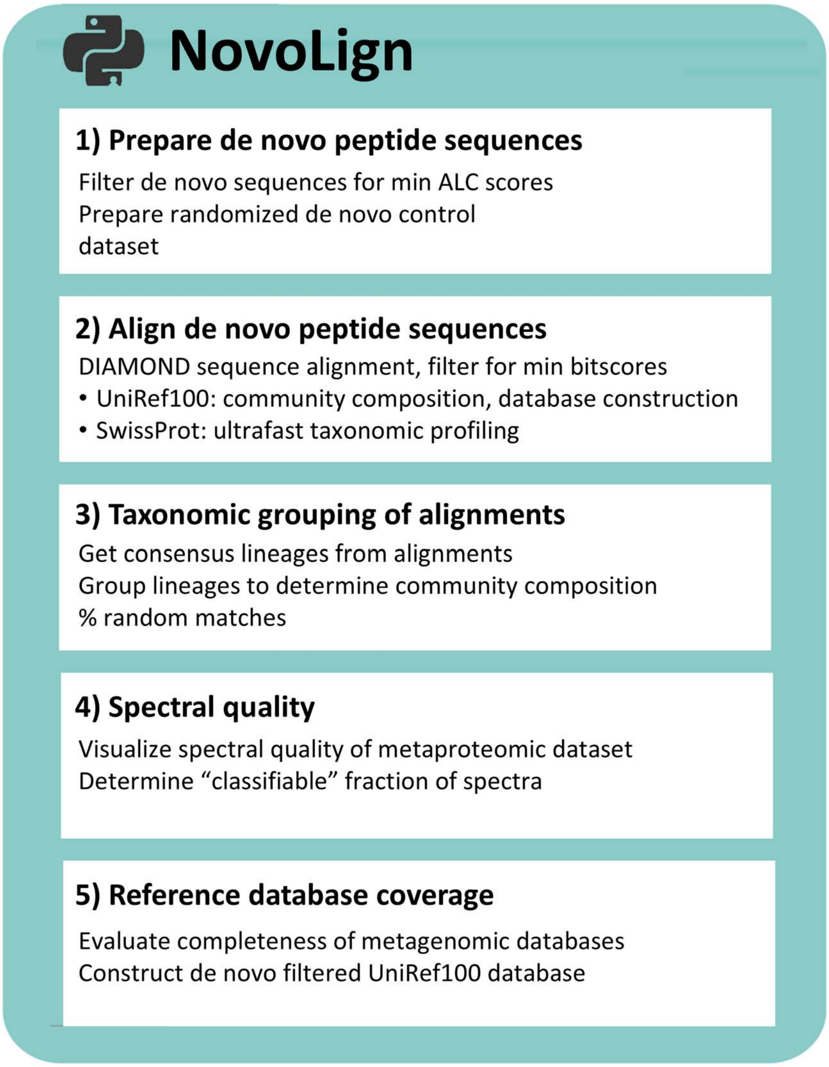
3. Taxonomic composition (bar_graphs.py): once the LCA has been determined, a frequency cutoff "freq_cut" (default 5) is applied, in order to remove taxonomies that occur at very low frequency. The remaining taxonomies are grouped together to provide a taxonomic composition of the analyzed microbial community. The results are then written to a table and displayed in bar graphs. Furthermore, the sequences from the target database (against which the query peptides were aligned) can be collected to create a specialized protein reference database. Optional arguments include an exclusion list, addition of decoy sequences, or addition of the query *de novo* peptides to the database. Alternatively, all sequences that belong to the identified taxonomies (at different levels, e.g. identified proteins, all species, genera, families) can be extracted and compiled into a database. Functional profiling can be performed using the compiled protein database, for example, by retrieving KO numbers for these proteins from GhostKoala [71].
4. Spectral quality analysis (experiment_qc.py): for spectral quality control, aligned *de novo* peptides are visualized with a scatter plot, histograms, and a stacked bar chart, based on *de novo* scores and annotation rates. Aligned *de novo* peptides are grouped into: "exact" (100% identity, 100% coverage), "exact tag" (100% identity, <100% coverage), "aligned" (<100% identity, 100% coverage), "aligned tag" (<100% identity, <100% coverage), and unmatched spectra. Optionally, if database searched peptides are also supplied, the stacked bar chart will show if the spectra of *de novo* sequenced peptides are detected in database searching ("matched" or "not detected").
5. Analysis of reference sequence database (database_qc.py): to access the coverage of the reference sequence database used for database searching, the aligned *de novo* sequences are compared to the database searching outputs. The annotated taxonomies will be compared for the ranks, order, family, and genus, and visualized with stacked bar charts. Taxonomic distributions will be shown for peptide spectrum matches detected exclusively with *de novo* sequencing (DN_only), all *de novo* peptides (DN_all), peptides detected exclusively in database searching (DB_only), and all database searching peptides (DB_all).

Results

Here, we present NovoLign, a *de novo* metaproteomics pipeline that performs large-scale sequence alignment of *de novo* sequences from complete metaproteomics experiments against large reference sequence databases (Fig. 1). To optimize the alignment and post-processing parameters, as well as to evaluate the accuracy of the taxonomic profiling, we employed a wide range of *in silico* and experimental data. These included *de novo* error datasets, pure reference strains, laboratory enrichments, and samples from synthetic and natural microbial communities.

Sequence alignment of *de novo* sequences

The sequence alignment of *de novo* sequences presents two major challenges. Firstly, bottom–up proteomics experiments result in short peptide sequences which are only poorly aligned using DIAMOND default settings. Secondly, in practice, many mass spectrometric fragmentation spectra show gaps in the fragment ion series and misleading fragments, resulting in sequencing errors such as amino acid inversions or substitutions [41, 43]. However, many of these errors occur in blocks within the amino acid sequence, and thus may still provide alignment to the correct taxonomy. To evaluate the correct alignment of such short sequences and sequences with errors, we created 13 *in silico* datasets, with 1000 sequences each (SI DOC Table 1), to reflect commonly observed sequencing errors like different types of amino acid inversions, substitutions, and mutations. Additionally, we constructed a dataset that combined all these errors, with error rates as commonly observed for the individual error types in proteomics experiments [41] (SI DOC Figs 1 and 2). This combined error dataset was then used to determine the most suitable DIAMOND parameters. First, we compared the alignment performance when using different scoring matrices, seed shapes, and DIAMOND seed search algorithms. Generally, for short sequences, PAM substitution matrices are more suitable than BLOSUM matrices. Shevchenko *et al.* (2001) used a PAM30 scoring matrix in their MS BLAST tool [50]. To improve the reporting of highly similar sequences, the authors made modifications to the matrix, such as adding scores for isobaric amino acids, trypsin cleavage sites, and substitutions. However, a modified matrix may interfere with the identification of homologous sequences from related organisms, which could be a disadvantage for metaproteomics experiments where the analyzed organisms may not be present in the database. Therefore, we accounted in our pipeline only for the isobaric amino acids isoleucine (I) and leucine (L) by replacing all "I" with an "L" in our reference sequence databases and sequencing data, but we did not modify scoring matrices. Additionally, the DIAMOND default sensitivity mode employs seeds with lengths of 12 and 15, which may discriminate shorter sequences. Furthermore, the different seed search algorithms (i.e. 0 = double-indexed, default algorithm, for large input files, less efficient for small query files; 1 = query-indexed, better performance for small query files; ctg = contiguous-seed mode, further improved performance for small query files) can affect the alignment of short sequences (with errors) and smaller datasets [77]. When aligning the combined error dataset against the UniRef100 and Swiss-Prot databases, we found that (as expected) the PAM 30 and 70 matrices generally performed better than BLOSUM62. Additionally, using shorter custom seeds and the "ctg" mode improved alignment for shorter peptides (<10 amino acids) significantly (SI DOC Figs 3 and 4). Furthermore, standard scoring matrices often use high gap opening and low gap extension penalties. This is suboptimal when aligning *de novo* peptide sequences with errors. Therefore, we evaluated different combinations of gap opening and gap extension penalties and found that a 2/4 (gap opening/extension penalties) combination performed best for our established *in silico de novo* sequences (SI DOC Figs 5 and 6). Finally, when comparing different reporting parameters for "query cover" and "percentage of sequence



NovoLign

1) Prepare de novo peptide sequences

Filter de novo sequences for min ALC scores
Prepare randomized de novo control dataset

2) Align de novo peptide sequences

DIAMOND sequence alignment, filter for min bitscores
- UniRef100: community composition, database construction
- SwissProt: ultrafast taxonomic profiling

3) Taxonomic grouping of alignments

Get consensus lineages from alignments
Group lineages to determine community composition
% random matches

4) Spectral quality

Visualize spectral quality of metaproteomic dataset
Determine "classifiable" fraction of spectra

5) Reference database coverage

Evaluate completeness of metagenomic databases
Construct de novo filtered UniRef100 database

Figure 1. The table outlines the main modules of the NovoLign pipeline, which conducts large-scale sequence alignment of complete metaproteomics experiments. This enables rapid (database searching independent) taxonomic profiling with deep coverage. It also offers an alternative approach for assessing the quality and coverage of metaproteomics outcomes obtained from database searching approaches. The pipeline contains five modules: (i) preparation of *de novo* sequences and generation of randomized decoy sequence dataset, (ii) alignment of the *de novo* and decoy sequences using DIAMOND, (iii) taxonomic grouping of the aligned sequences and determining the percentage of random matches, (iv) evaluating the spectral quality of the analyzed metaproteomics experiment, and (v) evaluating the coverage of the reference sequence database used for database searching.

identity", we found that a combination of % identity 85 with a % query coverage in the range of 75–85 provided the highest accuracy (SI DOC Figs 7 and 8). After determining the most suitable DIAMOND parameters, we investigated the alignments when working with larger reference sequence databases, such as UniRef100. The main objective was to determine the fraction of sequences which provide (i) exact sequence matches (alignments to the correct sequence in the reference database), (ii) correct taxonomy matches (alignments to different peptides that correspond to the correct taxonomy), (iii) false-positive matches (alignments to peptides from taxa not present in the sample), (iv) random matches (alignments to the randomized decoy sequence database), and (v) unmatched sequences (peptides which did not return any alignments), as well as to investigate the length distribution of the aligned sequences. Remarkably, the combined error dataset provided for a large fraction of sequences exact or correct matches (SI DOC Figs 9–12). For example, at the lower ranks order, family, and genus, we obtained between 55.3%–62.8% alignments to the correct taxonomy for Swiss-Prot and in the range of 38.8%–59.7% for UniRef100. However, most importantly, the fraction of sequences that provided alignments to wrong taxonomies was between 3.4%–5.9% for Swiss-Prot and 5.0–8.0 for UniRef100. The remaining sequences did not return any alignments. Furthermore, the length distribution of the aligned sequences reflected the distribution of the input sequences, with only a minor discrepancy toward smaller peptides with <10 amino acids. The UniRef100 database provided more alignments for the random (decoy) sequences. These, however, generally show lower bitscores and can thus be eliminated by setting a minimum bitscore threshold. Overall, the individual *de novo* error datasets provided outcomes comparable to the combined dataset (SI DOC Figs 13–23). However, the randomly mutated sequences (datasets 12 and 14, as shown in SI DOC Table 1 and SI DOC Fig. 14) yielded significantly fewer alignments. Additionally, sequences that underwent substitutions involving three or more amino acids (datasets 31–42, detailed in SI DOC Table 1 and SI DOC Figs 20–23) exhibited increasing discrimination against shorter sequences. In summary, the optimized DIAMOND parameters allowed to align and retrieve correct taxonomies for a large fraction of *de novo* peptides with common sequencing errors, on average 53.1% with a span from 31%–70.5% and different lengths, from 10–50 amino acids. Most importantly, erroneous annotations were < 10% (on average 4.1%, from 0.9%–8%), which makes these alignments suitable for metaproteomic applications (exact numbers for individual error datasets are listed in SI EXCEL Table 2).

Rapid metaproteomic taxonomic profiling

As a first step, the sequences are filtered for high quality *de novo* sequences by applying a minimum *de novo* quality score threshold (e.g. ALC score for PEAKS [72]). Next, the sequences are randomized to provide an additional set of decoy sequences, as already introduced for the NovoBridge pipeline [47]. Both sequence datasets are then aligned, where the decoy alignments provide an estimate for the number of random alignments per dataset. The results are then filtered for confident alignments by applying a minimum bitscore threshold. Furthermore, because sequences often provide alignments to several sequences from different taxonomies, the pipeline includes algorithms to determine a consensus lineage. In order to maximize the annotations at lower taxonomic ranks, we investigated 3 different LCA approaches. The first "conventional" LCA approach ("CON") strictly determines the LCA from all lineages above the bitscore threshold. The second bitscore approach ("BIT" LCA) is based on the recently published BAT tool [76], which determines the consensus taxonomy stepwise. Thereby, for every taxonomic rank, the taxonomy which accounts for the majority of the total bitscore is chosen. The third approach is a "weighted" approach ("W" LCA), which was introduced for processing of metagenomics data, by Buchfink *et al.* in 2015 [75]. This approach first assigns weights to all taxonomies based on their frequency in the alignment results. Furthermore, the consensus lineage is determined from taxonomies which combined account for at least 80% of the sum of weights to which the query sequence provided alignments. Next, the pipeline performs grouping of the consensus lineages in order to estimate the microbial composition (the fraction of peptide sequence alignments to a specific taxonomy compared to the total number of sequences assigned to all taxonomies). The sequence counts assigned to each taxonomic group can be seen as an abundance estimate for these organisms. Nevertheless, in order to avoid reporting extensive lists of very low abundant taxonomies (with only few sequence counts), we implemented a minimum frequency threshold for the taxonomic reporting step, as also described earlier [47]. Sequence alignment of short sequences with *de novo* sequencing errors may result in some false positive annotations (SI DOC Figs 10–23). Although the fraction of false positives is generally low, they can inflate the number of identified taxa in the composition report. While this approach prevents false taxonomies from being reported, it may also result in the omission of low-abundance taxa (e.g. those <1%).

In order to verify whether the sequence alignment provides an accurate representation of the taxonomies present in the microbial community, we processed two synthetic communities with known content (SI EXCEL Table 1). For this purpose, raw data from the "Kleiner equal protein" and SIHUMIx synthetic communities were *de novo* sequenced and processed using NovoLign with various combinations of processing parameters. These included different ALC score, bitscore, and taxon reporting thresholds, as well as different LCA approaches. The evaluated parameter combinations are summarized in SI DOC Table 2. The NovoLign processing results are summarized in Fig. 2A (and SI DOC Fig. 24), where the complete parameter evaluation output is provided in SI EXCEL Table 2. This allowed to evaluate how these parameters impact on the number of (i) sequences with correct alignments, (ii) sequences with unexpected alignments (to other taxonomies), and the number of (iii) aligned decoy sequences. For example, the ALC and bitscore had the largest impact on % other matches and % decoy matches, while the LCA algorithm and the taxonomy reporting threshold mostly influenced the target coverage and the number of identified taxonomies (SI DOC Figs 25–29). The best NovoLign parameter combinations (namely those which provide a high target coverage by maintaining <10% "other" and "decoy matches") for both synthetic communities are listed in SI DOC Tables 3 and 4. These generally employed either weighted or bitscore-based LCA, with a minimum ALC of 70 and a minimum bitscore threshold of 25. Nevertheless, although all taxonomies were identified for both synthetic communities at the family level, the families *Bifidobacteriaceae* and *Lactobacillaceae*, present in the SIHUMIx sample, were only observed when employing lower taxonomic reporting thresholds of 5 instead of 15. However, these microbes were also very low abundant in the database searching results, at ∼0.15 and 0.05%, respectively. Taxonomies with ≥1% relative abundance could all be identified at the family or genus levels. The highly compatible taxonomic profiles from the NovoLign pipeline, compared to the expected composition for the Kleiner and SIHUMIx synthetic communities, are shown in SI DOC Figs 30 and 31 (for different LCAs and frequency thresholds). Additionally, a taxonomic bar graph comparing the NovoLign and database searching compositions at the peptide level is presented in SI DOC Fig. 32. A more detailed characterization of the employed SIHUMIx samples has been provided by Van Den Bossche *et al.* (2021) [13]. Moreover, the Kleiner synthetic community also included closely related strains and phages. While strain differentiation can be possible, this depends on strain abundance and the availability of complete genomes in the reference database. For example, the exact genomes of *Staphylococcus aureus* strains ATCC 13709 and ATCC 25923 could not be found in the UniRef100 database used in this study. Similarly, for *Rhizobium leguminosarum*, only the bv. viciae 3841 strain was available. The phages were also unresolved due to their limited protein targets or absence from the database.

Nevertheless, the overall microbial composition obtained by the employed parameters showed a strong correlation to the true (known) composition of both communities, as demonstrated by the strong Spearman's rank correlation shown in Fig. 2A (and SI DOC Figs 30 and 31). Good performance was achieved at both the family and genus levels. However, at the species level, performance heavily depends on species abundance and the genome coverage in the database. Species not present in the database will provide alignments only to related genera or family ranks.

Finally, by using the optimized NovoLign processing parameters, we aimed to demonstrate the application of the NovoLign pipeline using a broad spectrum of taxonomies and sample complexities. Therefore, we *de novo* sequenced and processed a range of pure reference strains, synthetic communities, microbial enrichment cultures, and complex microbial samples using NovoLign (SI EXCEL Table 1). This approach yielded taxonomic profiles for all samples that were very close to the expected (known) microbial composition, or matched those obtained from orthogonal database searching experiments (Fig. 2B, and SI DOC Figs 30 and 31). Stacked bar graphs based on spectral counts for the samples shown in Fig. 2B are provided in SI DOC Fig. 33.

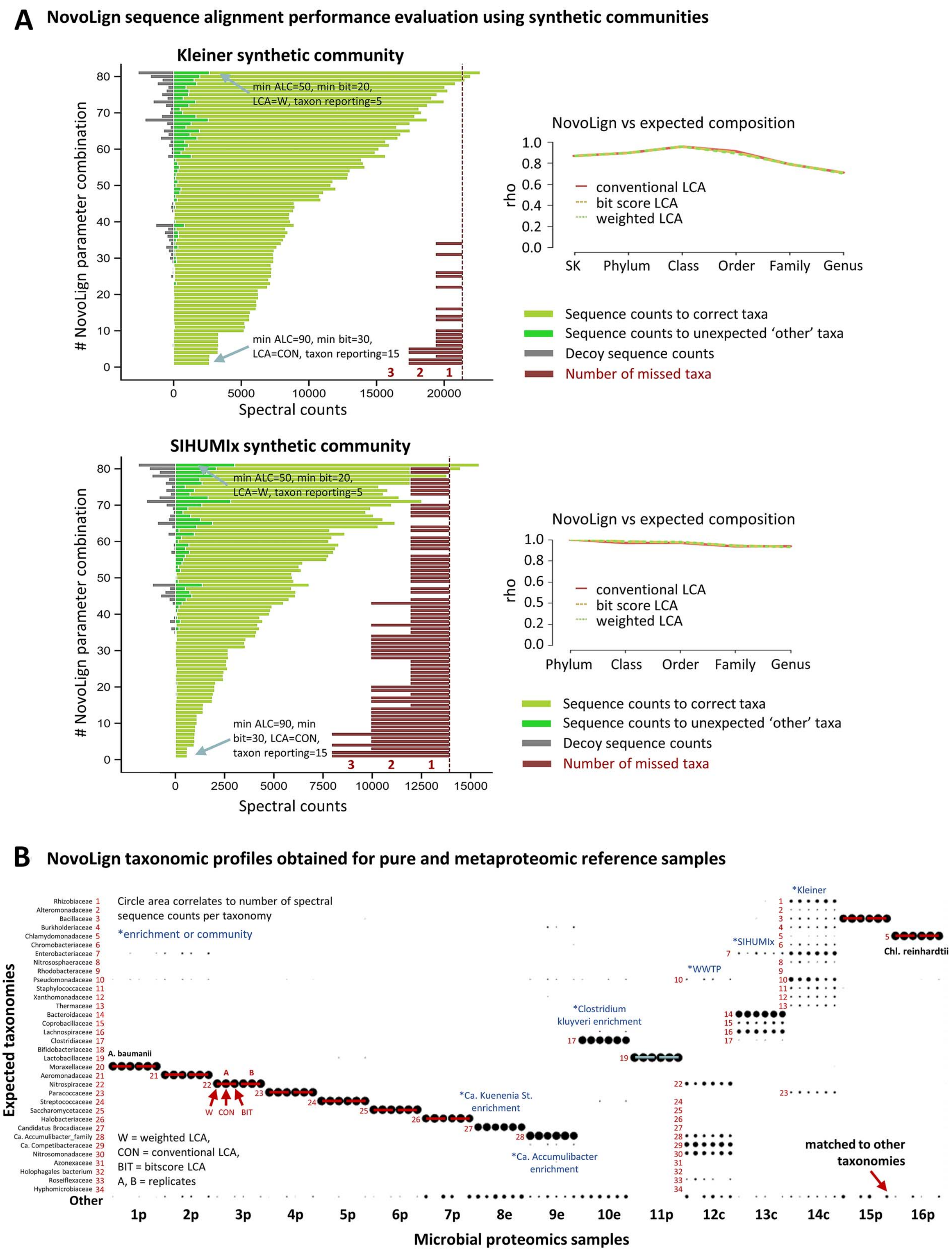


Figure 2. (A) The left bar graphs depict the NovoLign performance across various parameter combinations (detailed in SI DOC Table 2 and SI EXCEL Table 2) for the synthetic Kleiner equal protein (21 species/strains) and SIHUMIx communities (8 species). The influence of different post-processing parameter combinations, including ALC score, bitscore, LCA, and taxon reporting thresholds, was assessed. The bar graphs show for each parameter combination the number of (i) alignments to expected taxonomies, (ii) unexpected taxonomies, and the number of (iii) decoy sequence matches (see figure legend for color code). The bars located on the right of the bar graphs depict the number of expected taxonomies that were not identified ("number of missed taxa"). Generally, the proportion of decoy (random) and other taxonomies were very low for all parameter combinations. Furthermore, missed taxonomies for the Kleiner community were observed only when using very high minimum ALC, bitscore, and taxonomic reporting thresholds. For the SIHUMIx sample, only taxonomies with an abundance <1% were missed. The best parameter combinations (namely those which provided the highest target coverage by maintaining <10% "other" and decoy matches) are listed in SI DOC Tables 3 and 4. The line graphs on the right illustrate the

Finally, we aimed to test the NovoLign pipeline using a complex dataset obtained from an Orbitrap Astral mass spectrometer. The information density of the Orbitrap Astral raw data is extremely high due to its rapid MS/MS scan speed and high sensitivity. We selected a shotgun proteomics experiment recently published by Dumas *et al.* (2024), which analyzed a human fecal sample spiked with two microbes at low abundance (MetaP reference sample) [70]. Two 30-minute DDA runs were *de novo* sequenced and processed using the default NovoLign parameters. The taxonomic profiles generated by the NovoLign pipeline were highly comparable to those obtained from database searching, by Dumas *et al.* (SI DOC Fig. 34). Additionally, the spiked species *Deinococcus proteolyticus* (2%) and *B. vulgaris* (1%) were both clearly detected. Furthermore, we wanted to investigate how well the global functional profiles obtained by NovoLign and database searches compare. Therefore, we visualized the KO terms of the sequence-aligned proteins, alongside the KO terms of the proteins matched by database searching. The results showed that the profiles were highly reproducible between replicates and very comparable between both methods (SI DOC Figs 35 and 36).

Compared to the previously developed *de novo* pipeline, which employed exact sequence matches (NovoBridge), sequence alignment also significantly increased the number of annotated sequences. For instance, for *S. cerevisiae*, *Aeromonas*, and *Nitrospira*, as well as for the Kleiner and the complex Wastewater microbiome samples, the bitscore based LCA provided on average a 3.6-fold increase, and the weighted LCA on average a ∼4.8-fold increase in annotated sequences at the taxonomic levels, order, family, and genus (SI DOC Fig. 37).

Finally, processing the *S. cerevisiae* dataset (58 368 *de novo* sequences, including decoys) was very time efficient. For example, NovoLign processing using the complete Swiss-Prot database (containing 567 413 sequences, 280 MB) takes only 0.7 minutes, and NovoLign processing with the extensive UniRef100 database (containing 352 965 587 sequences, ∼180 GB) only 32.4 minutes, respectively (on a desktop with an Intel$^{(R)}$ Core$^{(\text{™})}$ i7-7700K and 32 GB RAM). A more detailed breakdown of the processing times for the individual NovoLign modules is presented in SI DOC Fig. 38.

Metaproteomic quality plots, database completeness, and complementation

A crucial step in metaproteomics is constructing the reference sequence database. To date, there is no consensus on the best method for creating such databases. However, generic reference sequence databases have a very large search space, which reduces sensitivity and increases the likelihood of false discoveries. Additionally, these databases can be regarded as incomplete, potentially missing crucial organisms or proteins [22, 27, 35]. Therefore, constructing sample-specific databases through whole metagenome sequencing is currently regarded as the most effective approach for many metaproteomic applications. However, because metagenomics experiments are time-consuming and expensive, generic databases are often employed to enable rapid quality monitoring in laboratory experiments. Furthermore, several factors can affect the coverage and accuracy of these databases, including DNA extraction, sequencing, genome assembly, identification of open reading frames, and taxonomic classification. Moreover, differences in the timing of sampling and the storage conditions of biomass between metagenomics and metaproteomics experiments can lead to significant discrepancies due to possible degradation and alterations in the microbial composition [22, 26, 27, 30, 78]. As a result, constructing an accurate protein sequence database from metagenomic data remains a delicate task.

While the large number of comparable datasets in single-species proteomics allows for setting expectations for peptide spectrum matches, the increased complexity of metaproteomics samples makes it more difficult to monitor the quality of these experiments. Consequently, poorly constructed reference sequence databases that miss members of the community may go unnoticed without the application of additional approaches.

One advantage of the alternative *de novo* sequencing in proteomics is that it provides next to a likely amino acid sequence also a quality metric for each spectrum. Spectra with good fragment ion coverage will obtain a high quality score, and are also expected to give a strong peptide spectrum match if the sequence is present in the reference database.

This offers the opportunity for NovoLign to identify missed taxonomies or proteins through processing *de novo* sequences from unmatched high-quality spectra. In order to demonstrate this procedure we determined the fraction of high-quality spectra (ALC $\geq$ 90) that were matched during database searching for our employed reference strains, enrichment cultures, and synthetic and natural communities (Fig. 3A). In order to investigate for missed taxonomies we also processed the *de novo* sequences from unmatched high-quality spectra through NovoLign and compared the taxonomic profile with those obtained from database searching (Fig. 3B).

Interestingly, for most samples >80% of the high-quality spectra were matched during database searching, with a strong agreement between the taxonomic profiles obtained by NovoLign and database searching. The remaining fraction of unmatched high-quality spectra did also not reveal new organisms for most samples and may therefore originate from modified peptides that were not considered during the database search process.

However, in one sample ("*Clostridium kluyveri* enrichment"), the fraction of matched high-quality spectra was <25% (Fig. 3A, sample 10e), and the taxonomic profile obtained from NovoLign significantly differed from the one obtained through database searching (Fig. 3B and C). For instance, while the NovoLign composition indicates that the family *Clostridiaceae* is dominant (∼80%), the outcomes from database searching shows that *Desulfovibrionaceae* is dominant, with *Clostridiaceae* being only a minor component. This reveals that the metagenomic reference database used

Spearman's correlation coefficients for both synthetic communities across the taxonomic ranks: superkingdom (SK), phylum, class, order, and family. Both synthetic communities show an excellent correlation between the abundances derived from NovoLign and the expected abundances. For the SIHUMIx sample, the composition determined by database searching served as the reference. Graphs detailing replicate experiments can be found in SI DOC Figure 24. (A) The figure illustrates the family-level microbial profiles derived from the proteomics and metaproteomics reference samples processed with NovoLign. The size of the circles corresponds to the sequence counts for each taxonomic identifier. Each sample is depicted through six distinct outputs, specifically three different LCA approaches—weighted LCA (W), conventional LCA (CON), and bitscore LCA (BIT)—for duplicate proteomics samples. The reference samples displayed the expected microbial profiles, with only minor portions of unexpected taxonomies, denoted as "other" on the y-axis. The labels below the graph (x-axis) provide the microbial sample identifier, with 1p corresponding to *A. baumannii*, 2p to *Aeromonas b.*, 3p to *N. moscoviensis*, 4p to *P. denitrificans*, 5p to *S. mutans*, 6p to *S. cerevisiae*, 7p to *Halanaeroarchaeum sp.*, 8e to Ca. Kuenenia stuttgartiensis enrichment, and 9e to Ca. Accumulibacter phosphatis enrichment, 10e to *C. kluyveri* enrichment, 11p to *Lactobacillus sakei*, 12c to aerobic granular sludge community, 13c to SIHUMIx synthetic community, 14c to Kleiner equal protein synthetic community, 15p to *C. thermarum*, and 16p to *C. reinhardtii*. The letters next to each number stand for: p = pure reference strain, e = enrichment, and c = community.

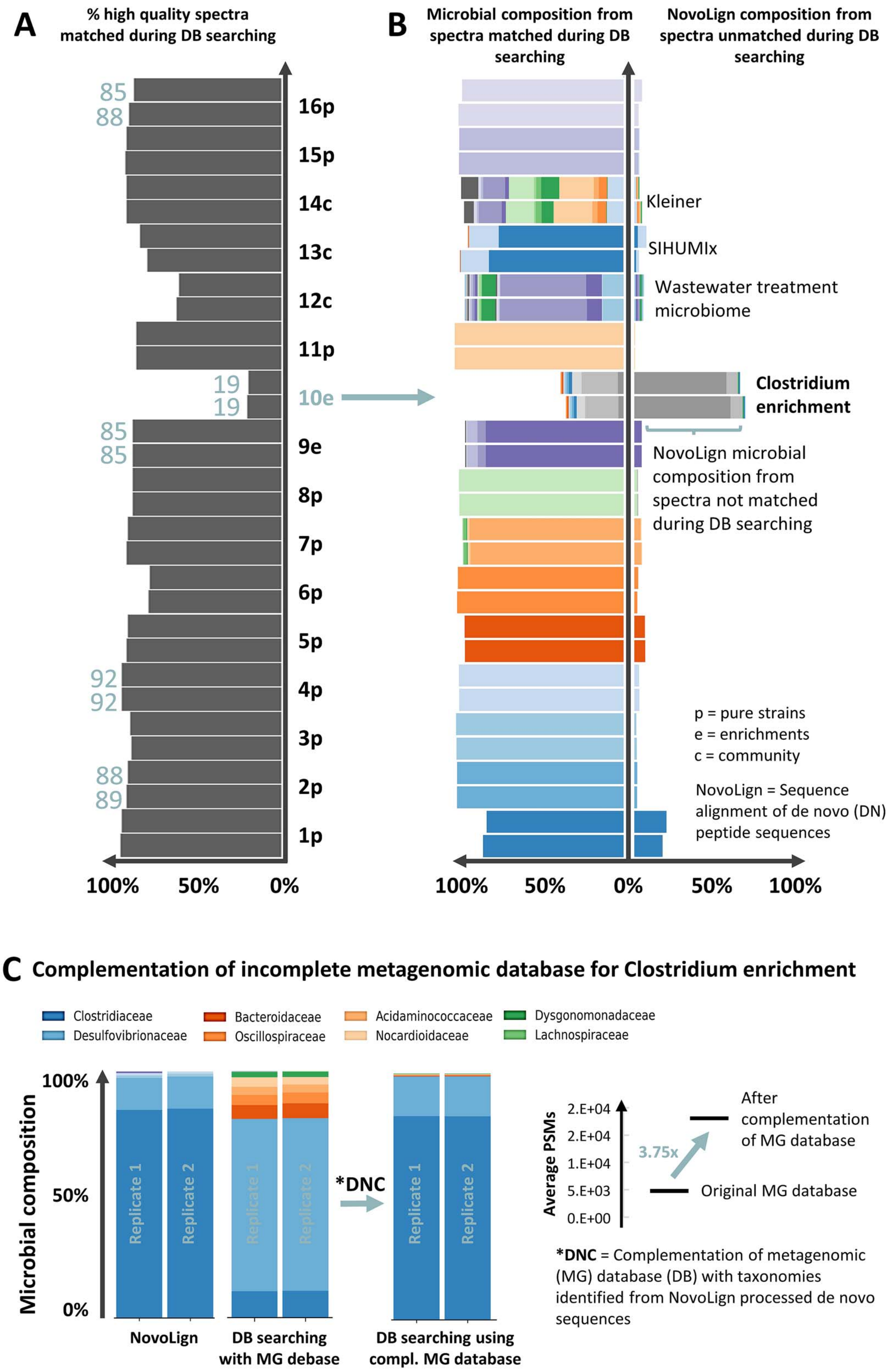


Figure 3. (A) The graph depicts the percentage of high quality spectra (ALC >90) that were matched during the database searching for the individual (meta)proteomics samples. Except for one sample (*C. kluyveri* enrichment, 10e), the fraction of spectra matched during the database search was very high. The labels next to the bars indicate the microbial sample, with 1p corresponding to *A. baumannii*, 2p to *Aeromonas b.*, 3p to *N. moscoviensis*, 4p to *P. denitrificans*, 5p to *S. mutans*, 6p to *S. cerevisiae*, 7p to *Halanaeroarchaeum sp.*, 8e to Ca. Kuenenia stuttgartiensis enrichment, and 9e to Ca. Accumulibacter phosphatis enrichment, 10e to *C. kluyveri* enrichment, 11p to *L. sakei*, 12c to aerobic granular sludge microbial community, 13c to SIHUMIx synthetic community, 14c to Kleiner equal protein synthetic community, 15p to *C. thermarum*, and 16p to *C. reinhardtii*. The letters next to each number stand for: p = pure reference strain, e = enrichment, and c = community. (B) The graph compares the taxonomic profiles at the family level obtained by database searching (left bars) with those obtained by processing spectra that were not matched during database searching using NovoLign (right bars). For the majority of the samples, NovoLign provided only a few additional taxonomic annotations, and the taxonomic profiles were very similar compared to

for database searching poorly represents the sample analyzed in the proteomics experiment. Although there could be several reasons for this (as described above), the observed discrepancy is most likely due to differences in storage times and sample processing between metagenomics and metaproteomics experiments.

Advantageously, NovoLign allows also to extract sequences of the identified taxonomies from the UniRef100 database. These can then be used to complement the existing database with the missing organisms. Therefore, we complemented the original *C. kluyveri* enrichment database with all *de novo* identified taxonomies (at the family level) and repeated the database searching with the complemented database. This resulted in a database that was significantly larger (~3.5 million sequences, compared to the original 50 K sequences in the metagenomic database), which strongly increased the number of peptide-spectrum matches (3.75× compared to the original metagenomic database) and aligned the taxonomic profiles with the one obtained by NovoLign (Fig. 3C). The microbial profile obtained after database complementation is also consistent with the 16S rRNA amplicon sequencing data and performed reactor experiments for the same enrichment previously [66].

Discussion

Metaproteomics relies on reference sequence databases that comprehensively cover all organisms present in a microbial community. Current methods for constructing these databases vary, but most applications require creating reference sequence databases from whole metagenome sequencing experiments. However, these experiments are time-consuming and complex, potentially biasing the content of reference sequence databases toward certain taxonomies. For instance, in the Clostridium enrichment analyzed in this study, the metagenomic sequence database covered the taxonomies present in the sample only partially. Although sampling for proteomics and metagenomics was performed simultaneously, suboptimal storage or extraction biases led to significant mismatches between both experiments. Unfortunately, unlike in single-species proteomics, biases in metaproteomics often remain unidentified due to the lack of benchmarks for highly complex samples.

Nevertheless, *de novo* sequencing allows to obtain the amino acid sequence directly from mass spectrometric fragmentation spectra. While other tools already employed *de novo* sequencing, they only used exact sequence matches or precomputed peptide databases. However, the sequence alignment employed by NovoLign overcomes these limitations, which improves coverage and broadens the range of applications.

Nevertheless, the generation of comprehensive *de novo* peptide sequence lists depends on robust validation strategies and high-quality peptide sequencing spectra. Therefore, the pipeline was established using shotgun data obtained from high-resolution Orbitrap mass spectrometers and validated with a broad spectrum of *in silico*, randomized, and experimental data. Furthermore, the *de novo* sequence lists in this study were predominantly generated using PEAKS. However, any sequence list (or sequence tags) in a tabular or text-like format that includes a column of peptide sequences with the header “Peptide” can be processed.

Rapid taxonomic and functional profiling of complex microbial samples—without requiring parallel metagenomics experiments or extensive databases—can be highly beneficial. It enables quick monitoring of compositional changes over time or quality checks before undertaking more extensive experiments, thus saving time and costs. However, beyond rapid taxonomic profiling and validation of conventional metaproteomics experiments that employ database searching, we foresee additional applications. For example, *de novo* sequence alignment will also be useful for identifying fragmentation spectra of modified peptides, which often remain unmatched in complex metaproteomics experiments [67, 79, 80]. Advanced *de novo* pipelines will also become increasingly important in other life sciences applications, where reference sequence databases are challenging to obtain. For instance, *de novo* sequence alignment can help to assemble antibody sequences when sequencing of the coding mRNA is not feasible [44, 45], or in identifying HLA-associated peptides that derive from mutated sequences, spliced peptides, or from non-coding regions [81]. Additionally, viruses, particularly bacteriophages that selectively target and kill bacteria, are increasingly studied. However, many viruses mutate with high frequency, which makes their proteomic analysis using conventional database searching a challenge [82].

In summary, we introduce a novel *de novo* metaproteomics pipeline based on sequence alignment, called NovoLign. This pipeline enables rapid taxonomic profiling with deep coverage and allows for the evaluation of the quality of metaproteomics experiments and reference sequence databases. Additionally, NovoLign facilitates the complementation of reference databases with sequences from organisms not present in the existing database.

We optimized the DIAMOND alignment for short sequences and *de novo* sequencing errors using a wide spectrum of *in silico* peptide data. Moreover, we validated NovoLign with a broad spectrum of data from pure reference strains, synthetic communities, enrichment cultures, and environmental microbial communities. We also evaluated post-processing parameters, including taxonomic grouping, to enhance taxonomic coverage and minimize false positive annotations.

Finally, the taxonomic profiles obtained for the reference strains and synthetic communities closely matched the expected taxonomic composition, with <10% unexpected taxonomies or decoy sequence matches. Complete proteomics datasets (~50 K *de novo* sequences) can be aligned to Swiss-Prot in <1 minute and to

those derived from database searching. However, in the case of the *C. kluyveri* enrichment (10e), NovoLign processing of unmatched spectra showed a substantial number of additional annotations, and a taxonomic profile very different from that obtained through database searching. (C) The graphs illustrate the complementation of the incomplete metagenomic sequence database for the *C. kluyveri* enrichment (10e). The bars, from left to right, display the microbial composition identified by NovoLign, database searching using the metagenomic database, and database searching using the *de novo* complemented metagenomic database. The taxonomic profile obtained from processing the metaproteomics dataset against the UniRef100 database with NovoLign was significantly different from that obtained through database searching using the metagenomic reference sequence database. Therefore, family level taxonomies identified by NovoLign were extracted from the UniRef100 database and integrated with the metagenomic database (DNC = *de novo* complementation). The complemented database was then utilized for database searching, providing a profile that closely aligned with the one obtained by NovoLign. Furthermore, the graph on the right highlights the increase in peptide spectrum matches (~3.5 times more compared to using the original metagenomic database) after the database complementation.

UniRef100 in ∼30 minutes, using a conventional desktop PC. The large set of validation data and the description of the NovoLign pipeline are publicly available and can be used to further optimize the pipeline for individual applications.

NovoLign serves as the urgently needed orthogonal approach to the widely used database searching in metaproteomics. Thereby, NovoLign allows for assessing the quality of conventional metaproteomics experiments and furthermore supports complementation of reference sequence databases to improve sequence coverage.

Acknowledgements

The authors acknowledge Dita Heikens for her support with sample preparation and all other colleagues from the Department of Biotechnology for valuable discussions.

Author contributions

HK and MPa conceived and designed the project. HK, MPa, RZ, RV, MPr, PS, JvE, and MA performed the experiments. HK, MPa, RZ, RV, and MvL provided critical input to experiments and/or protocols. HK and MP conducted the data analysis. HK and MPa generated the figures and wrote the original draft. All coauthors contributed in reviewing and editing the manuscript.

Supplementary material

Supplementary material is available at *ISME Communications* online.

Conflicts of interest

The authors declare that they have no conflict of interest.

Funding

This work was supported by a TU Delft Startup fund and the SIAM Gravitation Grant 024.002.00, the Netherlands Organization for Scientific Research (NWO).

Data availability

All mass spectrometric proteomics raw data are available via the ProteomeXchange consortium database. The raw data acquired within this project are available through the identifier PXD050548. ProteomeXchange identifiers for all other datasets are as following: the equal protein synthetic community proteomic raw data were obtained from PXD006118; the SIHUMIx were obtained from PXD023217; *A. baumannii* raw data were obtained from PXD011302; *C. thermarum* from PXD042369; *N. moscoviensis* from PXD019583; *C. reinhardtii* from PXD010160; *L. sakei* from PXD011417; *P. denitrificans* from PXD013274; *S. mutans* from PXD006735; Halanaeroarchaeum sp. HSR-CO from PXD028241; and the *C. kluyveri*-dominated was obtained from PXD040972. The MetaP reference sample measured with the Orbitrap Astral mass spectrometer was obtained from PXD045838.

The NovoLign pipeline is publicly available via: https://github.com/hbckleikamp/NovoLign.

Consent for publication

Not applicable.

References

1. Madsen EL. Microorganisms and their roles in fundamental biogeochemical cycles. *Curr Opin Biotechnol* 2011;**22**:456–64. https://doi.org/10.1016/j.copbio.2011.01.008
2. Pflughoeft KJ, Versalovic J. Human microbiome in health and disease. *Annu Rev Pathol Mech Dis* 2012;**7**:99–122.
3. Rousk J, Bengtson P. Microbial regulation of global biogeochemical cycles. *Front Microbiol* 2014;**5**:103.
4. Wierzchos J, de los Ríos A, Ascaso C. Microorganisms in desert rocks: the edge of life on Earth. *Int Microbiol* 2012;**15**:173–83.
5. González-Cabaleiro R, Martinez-Rabert E, Argiz L *et al.* A framework based on fundamental biochemical principles to engineer microbial community dynamics. *Curr Opin Biotechnol* 2021;**67**:111–8. https://doi.org/10.1016/j.copbio.2021.01.001
6. Kleerebezem R, van Loosdrecht MC. Mixed culture biotechnology for bioenergy production. *Curr Opin Biotechnol* 2007;**18**:207–12.
7. Li P, Roos S, Luo H *et al.* Metabolic engineering in human gut microbiome: recent developments and future perspectives. *Metab Eng* 2023;**79**:1–13. https://doi.org/10.1016/j.ymben.2023.06.006
8. Zaramela LS, Moyne O, Kumar M *et al.* The sum is greater than the parts: exploiting microbial communities to achieve complex functions. *Curr Opin Biotechnol* 2021;**67**:149–57. https://doi.org/10.1016/j.copbio.2021.01.013
9. Quince C, Walker AW, Simpson JT *et al.* Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;**35**:833–44. https://doi.org/10.1038/nbt.3935
10. Su C, Lei L, Duan Y *et al.* Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. *Appl Microbiol Biotechnol* 2012;**93**:993–1003. https://doi.org/10.1007/s00253-011-3800-7
11. Wilmes P, Bond PL. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* 2006;**14**:92–7.
12. Armengaud J. Metaproteomics to understand how microbiota function: the crystal ball predicts a promising future. *Environ Microbiol* 2023;**25**:115–25. https://doi.org/10.1111/1462-2920.16238
13. Van Den Bossche T, Kunath BJ, Schallert K *et al.* Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat Commun* 2021;**12**:7305.
14. Kleiner M. Metaproteomics: much more than measuring gene expression in microbial communities. *mSystems* 2019;**4**:10.1128/msystems.00115-19. https://doi.org/10.1128/msystems.00115-19
15. Van Den Bossche T, Arntzen MØ, Becher D *et al.* The metaproteomics initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome* 2021;**9**:243.
16. Schiebenhoefer H, Schallert K, Renard BY *et al.* A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and prophane. *Nat Protoc* 2020;**15**:3212–39. https://doi.org/10.1038/s41596-020-0368-7
17. Salvato F, Hettich RL, Kleiner M. Five key aspects of metaproteomics as a tool to understand functional interactions in host associated microbiomes. In: *Advances in Clinical Immunology, Medical Microbiology, COVID-19, and Big Data*. 1st edition. Jenny Stanford Publishing, 2021, 647–60.
18. Sun Z, Ning Z, Figeys D. The landscape and perspectives of the human gut metaproteomics. *Mol Cell Proteomics* 2024;**23**:100763.

19. Chirania P, Holwerda EK, Giannone RJ *et al*. Metaproteomics reveals enzymatic strategies deployed by anaerobic microbiomes to maintain lignocellulose deconstruction at high solids. *Nat Commun* 2022;**13**:3870.
20. Karaduta O, Dvanajscak Z, Zybailov B. Metaproteomics—an advantageous option in studies of host-microbiota interaction. *Microorganisms* 2021;**9**:980.
21. Li L, Wang T, Ning Z *et al*. Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. *Nat Commun* 2023;**14**:3428.
22. Kleikamp HB, Grouzdev D, Schaasberg P *et al*. Metaproteomics, metagenomics and 16S rRNA sequencing provide different perspectives on the aerobic granular sludge microbiome. *Water Res* 2023;**246**:120700.
23. Yates JR III. Proteomics of communities: metaproteomics. *J Proteome Res* 2019;**18**:2359.
24. Kleiner M, Thorson E, Sharp CE *et al*. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat Commun* 2017;**8**:1558.
25. Wilmes P, Andersson AF, Lefsrud MG *et al*. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* 2008;**2**:853–64. https://doi.org/10.1038/ismej.2008.38
26. Blakeley-Ruiz JA, Kleiner M. Considerations for constructing a protein sequence database for metaproteomics. *Comput Struct Biotechnol J* 2022;**20**:937–52. https://doi.org/10.1016/j.csbj.2022.01.018
27. Wu E, Mallawaarachchi V, Zhao J *et al*. Contigs directed gene annotation (ConDiGA) for accurate protein sequence database construction in metaproteomics. *bioRxiv* 2023.; 2023.2004.2019.537311
28. Miura N, Okuda S. Current progress and critical challenges to overcome in the bioinformatics of mass spectrometry-based metaproteomics. *Comput Struct Biotechnol J* 2023;**21**:1140–50. https://doi.org/10.1016/j.csbj.2023.01.015
29. Heyer R, Schallert K, Zoun R *et al*. Challenges and perspectives of metaproteomic data analysis. *J Biotechnol* 2017;**261**:24–36. https://doi.org/10.1016/j.jbiotec.2017.06.1201
30. Tanca A, Palomba A, Deligios M *et al*. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* 2013;**8**:e82981.
31. Stamboulian M, Li S, Ye Y. Using high-abundance proteins as guides for fast and effective peptide/protein identification from human gut metaproteomic data. *Microbiome* 2021;**9**:1–17.
32. Bassignani A, Plancade S, Berland M *et al*. Benefits of iterative searches of large databases to interpret large human gut metaproteomic data sets. *J Proteome Res* 2021;**20**:1522–34. https://doi.org/10.1021/acs.jproteome.0c00669
33. Jagtap P, Goslinga J, Kooren JA *et al*. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* 2013;**13**:1352–7. https://doi.org/10.1002/pmic.201200352
34. Xiao J, Tanca A, Jia B *et al*. Metagenomic taxonomy-guided database-searching strategy for improving metaproteomic analysis. *J Proteome Res* 2018;**17**:1596–605. https://doi.org/10.1021/acs.jproteome.7b00894
35. Nalpas N, Hoyles L, Anselm V *et al*. An integrated workflow for enhanced taxonomic and functional coverage of the mouse fecal metaproteome. *Gut Microbes* 2021;**13**:1994836.
36. Lee J-Y, Mitchell HD, Burnet MC *et al*. Uncovering hidden members and functions of the soil microbiome using de novo metaproteomics. *J Proteome Res* 2022;**21**:2023–35. https://doi.org/10.1021/acs.jproteome.2c00334
37. Potgieter MG, Nel AJ, Fortuin S *et al*. MetaNovo: an open-source pipeline for probabilistic peptide discovery in complex metaproteomic datasets. *PLoS Comput Biol* 2023;**19**:e1011163.
38. Mesuere B, Debyser G, Aerts M *et al*. The Unipept metaproteomics analysis pipeline. *Proteomics* 2015;**15**:1437–42. https://doi.org/10.1002/pmic.201400361
39. Cantarel BL, Erickson AR, VerBerkmoes NC *et al*. Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One* 2011;**6**:e27173.
40. Vitorino R, Guedes S, Trindade F *et al*. De novo sequencing of proteins by mass spectrometry. *Expert Rev Proteomics* 2020;**17**:595–607. https://doi.org/10.1080/14789450.2020.1831387
41. Muth T, Renard BY. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* 2018;**19**:954–70. https://doi.org/10.1093/bib/bbx033
42. Devabhaktuni A, Elias JE. Application of de novo sequencing to large-scale complex proteomics data sets. *J Proteome Res.* 2016;**15**:732–42. https://doi.org/10.1021/acs.jproteome.5b00861
43. Muth T, Hartkopf F, Vaudel M *et al*. A potential golden age to come—current tools, recent use cases, and future avenues for de novo sequencing in proteomics. *Proteomics* 2018;**18**:1700150.
44. Beslic D, Tscheuschner G, Renard BY *et al*. Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Brief Bioinform* 2023;**24**:bbac542.
45. Peng W, Pronker MF, Snijder J. Mass spectrometry-based de novo sequencing of monoclonal antibodies using multiple proteases and a dual fragmentation scheme. *J Proteome Res* 2021;**20**:3559–66. https://doi.org/10.1021/acs.jproteome.1c00169
46. Johnson RS, Searle BC, Nunn BL *et al*. Assessing protein sequence database suitability using de novo sequencing. *Mol Cell Proteomics* 2020;**19**:198–208. https://doi.org/10.1074/mcp.TIR119.001752
47. Kleikamp HB, Pronk M, Tugui C *et al*. Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems* 2021;**12**:375–383. e375.
48. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1997;**11**:1067–75.
49. Taylor JA, Johnson RS. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* 2001;**73**:2594–604.
50. Shevchenko A, Sunyaev S, Loboda A *et al*. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem* 2001;**73**:1917–26.
51. Chalkley RJ, Baker PR, Huang L *et al*. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol Cell Proteomics* 2005;**4**:1194–204.
52. Mackey AJ, Haystead TA, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol Cell Proteomics* 2002;**1**:139–47.
53. Searle BC, Dasari S, Turner M *et al*. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal Chem* 2004;**76**:2220–30.
54. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. In:

Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. IEEE; 2004. p. 206–15. https://doi.org/10.1109/CSB.2004.1332434

55. Leprevost FV, Valente RH, Lima DB *et al.* PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. *Mol Cell Proteomics* 2014;**13**:2480–9. https://doi.org/10.1074/mcp.M113.037002
56. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60. https://doi.org/10.1038/nmeth.3176
57. Di Venanzio G, Flores-Mireles AL, Calix JJ *et al.* Urinary tract colonization is enhanced by a plasmid that regulates uropathogenic Acinetobacter baumannii chromosomal genes. *Nat Commun* 2019;**10**:2763.
58. De Jong SI, Sorokin DY, van Loosdrecht MC *et al.* Membrane proteome of the thermoalkaliphile Caldalkalibacillus thermarum TA2. A1. *Front Microbiol* 2023;**14**:1228266.
59. Lawson CE, Mundinger AB, Koch H *et al.* Investigating the chemolithoautotrophic and formate metabolism of Nitrospira moscoviensis by constraint-based metabolic modeling and 13C-tracer analysis. *Msystems* 2021;**6**:e00173–21.
60. Scholz M, Gäbelein P, Xue H *et al.* Light-dependent N-terminal phosphorylation of LHCSR3 and LHCB4 are interlinked in Chlamydomonas reinhardtii. *Plant J* 2019;**99**:877–94. https://doi.org/10.1111/tpj.14368
61. Prechtl RM, Janßen D, Behr J *et al.* Sucrose-induced proteomic response and carbohydrate utilization of lactobacillus sakei TMW 1.411 during dextran formation. *Front Microbiol* 2018;**9**:2796.
62. Schada von Borzyskowski L, Severi F, Krüger K *et al.* Marine proteobacteria metabolize glycolate via the *β*-hydroxyaspartate cycle. *Nature* 2019;**575**:500–4. https://doi.org/10.1038/s41586-019-1748-4
63. Ahn S-J, Gu T, Koh J *et al.* Remodeling of the Streptococcus mutans proteome in response to LrgAB and external stresses. *Sci Rep* 2017;**7**:14063.
64. Sorokin DY, Merkel AY, Messina E *et al.* Anaerobic carboxydotrophy in sulfur-respiring haloarchaea from hypersaline lakes. *ISME J* 2022;**16**:1534–46. https://doi.org/10.1038/s41396-022-01206-x
65. Allaart MT, Stouten GR, Sousa DZ *et al.* Product inhibition and pH affect stoichiometry and kinetics of chain elongating microbial communities in sequencing batch bioreactors. *Front Bioeng Biotechnol* 2021;**9**:693030.
66. Allaart MT, Fox BB, Nettersheim IH *et al.* Physiological and stoichiometric characterization of ethanol-based chain elongation in the absence of short-chain carboxylic acids. *Sci Rep* 2023;**13**:17370.
67. Pabst M, Grouzdev DS, Lawson CE *et al.* A general approach to explore prokaryotic protein glycosylation reveals the unique surface layer modulation of an anammox bacterium. *ISME J* 2022;**16**:346–57. https://doi.org/10.1038/s41396-021-01073-y
68. Tugui C, Sorokin D, Hijnen W *et al.* Exploring the metabolic potential of Aeromonas to utilise the carbohydrate polymer chitin. *bioRxiv* 2024.; 2024.2002. 2007.579344
69. Lawson CE, Nuijten GH, de Graaf RM *et al.* Autotrophic and mixotrophic metabolism of an anammox bacterium revealed by in vivo 13C and 2H metabolic network mapping. *ISME J* 2021;**15**:673–87. https://doi.org/10.1038/s41396-020-00805-w
70. Dumas T, Martinez Pinna R, Lozano C *et al.* The astounding exhaustiveness and speed of the astral mass analyzer for highly complex samples is a quantum leap in the functional analysis of microbiomes. *Microbiome* 2024;**12**:46.
71. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* 2016;**428**:726–31. https://doi.org/10.1016/j.jmb.2015.11.006
72. Ma B, Zhang K, Hendrie C *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;**17**:2337–42.
73. Tran NH, Zhang X, Xin L *et al.* De novo peptide sequencing by deep learning. *Proc Natl Acad Sci* 2017;**114**:8247–52.
74. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*
75. Buchfink B, Huson D, Xie C. Metascope—fast and accurate identification of microbes in metagenomic sequencing data. arXiv. *arXiv preprint arXiv:151108753.*. 2015
76. Von Meijenfeldt FB, Arkhipova K, Cambuy DD *et al.* Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* 2019;**20**:1–14.
77. Vroland C, Salson M, Bini S *et al.* Approximate search of short patterns with high error rates using the 01* 0 lossless seeds. *J Discrete Algorithms* 2016;**37**:3–16.
78. Timmins-Schiffman E, May DH, Mikan M *et al.* Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J* 2017;**11**:309–14. https://doi.org/10.1038/ismej.2016.132
79. Devabhaktuni A, Lin S, Zhang L *et al.* TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat Biotechnol* 2019;**37**:469–79. https://doi.org/10.1038/s41587-019-0067-5
80. Cheng K, Ning Z, Zhang X *et al.* MetaLab 2.0 enables accurate post-translational modifications profiling in metaproteomics. *J Am Soc Mass Spectrom* 2020;**31**:1473–82. https://doi.org/10.1021/jasms.0c00083
81. Rye-Weller A, Dhanik A, Cygan K *et al.* Proteogenomics and de novo sequencing based approach for neoantigen discovery from the immunopeptidomes of patient CRC liver metastases using mass spectrometry. *J Immunol* 2020;**204**:217.216.
82. Holstein T, Kistner F, Martens L *et al.* PepGM: a probabilistic graphical model for taxonomic inference of viral proteome samples with associated confidence scores. *Bioinformatics* 2023;**39**:btad289.