Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Data Independent Acquisition Pipeline for Microbiome Samples (Microbe-DIA)

Samantha A. Obermiller
Pacific Northwest National Laboratory
Mary S. Lipton
Pacific Northwest National Laboratory
Paul D. Piehowski
Pacific Northwest National Laboratory
Victoria N. Prozapas
Pacific Northwest National Laboratory
Aivett Bilbao
Pacific Northwest National Laboratory
Lee Ann McCue
Pacific Northwest National Laboratory
Geremy C. Clair
Pacific Northwest National Laboratory
Isaac Kwame Attah
kwame.attah@pnnl.gov
Pacific Northwest National Laboratory

Method Article

Keywords: Data-independent analysis, DIA, microbial communities, metaproteomics, DIA-NN, microbiome.

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10396936/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The functional complexity inherent in microbiomes complicates analytical approaches aimed at defining phenotype. As proteins a re the functional effectors of microbiome phenotypes, improving the performance of mass spectrometry-based metaproteomics is critical to achieving the functional characterization of these systems. Data-independent acquisition (DIA) improves protein coverage and reduces data missingness when compared to data-dependent acquisition (DDA) in metaproteomics. However, the application of DIA to complex microbial systems remains constrained by analytical throughput and computational scalability. Here, we optimized LC–MS/MS acquisition parameters for both DDA and DIA using a model microbiome, demonstrating how DIA enables increased sample throughput without compromising quantitative performance. In addition, we demonstrated a computationally efficient, library-free DIA workflow that overcomes reliance on empirical spectral libraries. Our analytical and computational innovations establish a scalable and cost-effective pipeline for metaproteomics of complex microbial communities.

Introduction

Environmental microbiome samples possess a highly complex proteomic profile. This complexity arises from two sources: variation in the abundance of the microbial species as well as variation in the abundance of the proteins in each of the microbes. Further, each microbial genome typically contains thousands of protein-coding genes which are highly homologous across microbes, creating redundantly expressed proteins and peptides. Metaproteomics aims to detect and quantify the suite of proteins expressed in microbial communities, contributing to our understanding of critical functional modulations in these systems and enabling comprehensive multi-omics integration [1].

Liquid chromatography tandem mass spectrometry (LC-MS/MS) proteomic analysis focuses on the evaluation of tryptic peptides from samples where fragmentation of intact peptide precursors is necessary for peptide sequence identification. The traditional approach to fragmentation is data-dependent acquisition (DDA), where narrow mass-to-charge (m/z) windows are used to sequentially isolate the most abundant peptide precursors for MS/MS analysis. This strategy simplifies the analysis of the collected fragment spectra and increases the confidence in peptide identifications. However, because of the limited duty cycle of mass spectrometers, this strategy can acquire spectra for only a limited number of precursor ions, limiting peptide identifications and typically yielding low protein sequence coverage in complex samples, such as microbiomes. Additionally, accurate quantification requires the measurement of peptide intensities at multiple retention times to determine the area under the extracted ion chromatogram (XIC). Since DDA typically avoids fragmenting the same parent ion multiple times, this precludes the use of fragmentation spectra for quantification. In addition, traditional DDA struggles with high data missingness due to this stochastic sampling strategy [2]. Decreasing missing values in LC-MS/MS can provide enlightening functional information for microbiome samples by increasing the number of quantified proteins, especially for the low-abundance precursors [3].

In contrast to DDA, data-independent acquisition (DIA) fragments all precursors within wider m/z windows, enabling more comprehensive sampling of the peptide ion population [4]. In DIA, these windows are fragmented sequentially and cyclically to cover the desired mass range. By forgoing peak selection, DIA yields more reproducible data that includes fragments even from low-abundance precursors [5]. The m/z windows used for fragmentation can vary in width, with researchers weighing the complexity of overlapping peptide fragmentations against the increase in potential peptide identifications. The fragments themselves can also be used for quantification since the same precursors are fragmented multiple times, potentially improving quantification accuracy [6, 7, 8].

While data-independent acquisition reduces missing values and improves protein quantification, the resulting spectral complexity requires an additional layer of information for processing via a DDA-generated library or a library-free approach [2]. Generating both DDA and DIA data is resource-intensive, and library-free approaches are typically limited in feasibility due to the size of the search database prohibiting the use of large metaproteomes [9]. Previous studies have already compared various DIA software (such as DIA-NN, MaxDIA, and Spectronaut) and found that DIA-NN's peak selection algorithm is highly effective [10, 11, 12, 13, 14]. A recent study by Rajczewski *et al.* also evaluated the reproducibility of DDA and library-free DIA across multiple laboratories using modeled microbiome samples [9]. While they confirmed DIA's reproducibility and accuracy, their focus on inter-laboratory performance precluded systematic optimization of acquisition parameters. They also noted that further investigation was needed into the scalability of library-free processing to large metaproteomes.

Here, we address these challenges by developing a scalable, cost-effective pipeline for metaproteomic analysis of microbiomes (Fig. 1). Building on previous studies, we systematically optimized LC-MS/MS parameters for both DDA and DIA and evaluated multiple library processing strategies using a controlled model microbiome system. We demonstrated that DIA-NN's library-free mode [12] is currently the most efficient method of analysis while still accurately characterizing the model system. Critically, we present Metaproteomic Analysis with DIA Library-free (MADL), a two-stage approach that enables library-free DIA processing against a ~ 500,000 protein-coding metaproteome database, overcoming the primary computation limitation of this approach.

Methods

Designing the model sample and metaproteome. Each of the potential 46 strains of bacteria were cultured individually and underwent genome sequencing [15]. The resulting FASTA files were then used to assess the relatedness of the available species and ensure the model communities simulate microbiome diversity. The proteins for each strain underwent a simulated trypsin digestion, cleaving after arginine or lysine excluding instances followed by proline, allowing for up to 2 missed cleavages and a minimum string length of 10. The resulting fragments were combined for all strains within a species and deduplicated. Relatedness of a species was then calculated as the percent of its peptide fragments shared by any of the other species (Supplementary Fig. 1). Five, 24-strain mixtures were designed with species of varying relatedness and informed our comparison of MS1 and MS2 window

sizes for DIA and fractionation for DDA libraries (Fig. 2, Supplementary Fig. 2, Supplementary Table 1). For Fig. 1, only Mix1 of the 24-strain dataset was utilized. This decision was based on the consistency across mixtures observed in Supplementary Fig. 2. By focusing exclusively on Mix1, the figure provides informative comparisons including MS window size. Two 46-strain mixtures, Mix1 and Mix2, were designed using all available species and informed all further analyses (Figs. 3–6).

Sample preparation for proteomics analysis. The washed cell pellets were lysed on a bead mill homogenizer (Bead Ruptor Elite) under settings: Speed 6.3 for 45 seconds. Lysates were then centrifuged at 3220 x rpm in large centrifuge for 10min at 4C to remove debris. The supernatant were then transferred to fresh tubes with urea to achieve a final concentration of 8M. A BCA assay (Thermo Fisher Scientific) was then used to estimate protein concentration for downstream processing. Proteins were then reduced by incubating at 10 mM dithiothreitol (DTT) at 37 ºC for 1 hr. Lysates were then diluted 8-fold to reduce the urea concentration prior to digestion with trypsin at a 1:50 enzyme:substrate ratio, incubating at 37 ºC for 3 hrs. The digestion reaction was then quenched by plunge freezing in LN_2. Peptides were then desalted using strata C18 cartridges (Phenomenex). Eluents were dried in a speed vac and rehydrated in 50 µL of 18 MΩ water. The concentration of peptides was measured by BCA assay and diluted to 0.1 µg/µL for downstream sample creation (Supplementary Tables 1 and 2). The resulting peptide samples were then mixed as described in Supplementary Tables 1, 2, and 3 to create the model microbiomes for our studies.

LC-MS/MS Analysis. All peptide sample analysis were performed using a Vanquish Neo liquid chromatography system (Thermo Scientific). For longer reverse phase peptide analysis (~ 9 samples per day (SPD)), the Vanquish Neo LC system was set up with a single pump for both sample trapping and reverse-flow elution onto the analytical column, in the Trap and Elute mode. A 4-cm x 100 um i.d. fused silica column packed with 5µm Jupiter C2 in-house was used for sample trapping. For peptide separation, a fused silica column with an integrated emitter (30 cm × 75 µm i.d.) was packed in-house with 1.7 µm particle size of Waters Acquity BEH particles, Waters). Mobile phases used for sample analysis consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. The 9 SPD method, using a flow rate of 0.20µL/min, has an active gradient of 120 minutes and used the gradient profile; (min, %B): 0.0, 1.0; 11.0, 8.0; 86.0, 21.0; 106.0, 28.0; 116.0, 37.0; 121.0, 75.0; 124.0, 95.0; 129.0, 95.0, Column Wash; 130.0, 50.0; 131,0, 1.0; 132.0, 95.0; 133.0, 1.0; 149.0, 1.0.

For shorter reverse phase elution analysis (~ 50 SPD), the Vanquish Neo LC set up similarly, used a PepMap Neo Trap Cartridge (Thermo Scientific) for sample trapping, and a PepMap ES906 analytical column (Thermo Scientific). The 50 SPD method has a 23-minute active gradient, with 6 munities for sample loading and column equilibration. Below is the gradient profile used for this method; (min, Flow Rate, %B): 0.0, 3.50,1.0; 0 .001, 2.80, 1.0; 1.001, 1.30, 4.0; 2.000, 0.800, 8.0; 18.001, 0.800, 22.5; 22.001, 0.800, 35.0; 23.001, 1.10, 45.0; Column Wash; 23.501, 2.80, 99.0; 24.701, 99.0; 25.701, 3.50, 1.0.

The 9 SPD LCMS analysis were performed using the Orbitrap Exploris 480 mass spectrometer, with the analytical column interfaced using a nanoflex nano-electrospray ionization source. The ion transfer tube

was set at 300°C, and the electrospray voltage was 2.2 kV. Sample analysis on the Orbitrap Exploris 480 mass spectrometer was performed either in the data-dependent (DDA) or data-independent (DIA) mode, acquiring data for 120 minutes in both cases. For DDA analysis, full scans were acquired at a resolution of 60k in a scan range of 300–1800 m/z. The automatic gain control (AGC) target was set to standard, and maximum IIT set to auto. The Data Dependent Mode was set to Number of Scans, analyzing the top 20 peptide ion peaks per full scan. Peptide ions selected for MS2 analysis were fragmented using higher-energy collision dissociation (HCD) at a normalized collision energy of 30%. The isolation window width for isolating precursor ions was set to 0.7 m/z, with isolation offset set to "off". Resultant fragment ions were detected in the orbitrap, at a resolution of 30K, with first mass set to 110 m/z. The normalized AGC target was set at 250%, and the maximum injection time mode was set to “Auto”. The instrument was set to perform dependent scan on a single charge state precursor, after which the selected peptide is excluded for 45 s. For DIA analysis, the full scan spectra was acquired at a resolution of 120k, over the mass range of 300–900 m/z, using a normalized AGC target value of 300% and the and maximum IIT set to auto. DIA mode MS^2 was performed over the same mass range, using an isolation window size of 10 m/z, window overlap set to 0, and window placement optimization enabled. The cycle time was set to 3 seconds. Resultant fragment data was recorded at an orbitrap resolution of 30 k, over the mass range of 200–2000 m/z. The normalized AGC target was set at 800%, maximum IIT set to “Auto”. Similarly, HCD was used for fragmentation at a normalized collision energy of 30%.

All 50 SPD LCMS analysis were performed on the Orbitrap Astral mass spectrometer, with the ES906 analytical column interfaced to the mass spectrometer using an EASY-Spray ion source, and data acquired for 22 minutes. The ion source conditions were 2.2 kV and 300°C for the spray voltage and Ion transfer tube temperature respectively.

The full scan spectra acquired using the Orbitrap analyzer was performed over a scan range of 380 to 980 m/z at the resolution of 240 k. The normalized AGC target was set at 500%, with the maximum injection time set to 5 ms. MS^2 analysis was performed using the Astral analyzer, with the DIA window type was set to “Auto”, window placement optimization at “On”, window overlap at 0, and isolation window at 2 m/z. The normalized AGC target was set at 500%, with the maximum injection time set to 3 ms. Fragment analysis was performed using HCD at a normalized collision energy of 25%, and the Astral analyzer scan range set to 150–2000 m/z. Loop control mode was set to time, with a value of 0.6 seconds.

Peptide identification for DDA and DIA data. The raw mass spectrometry files were converted into mzML format using Msconvert with $peak\Pi ck \in g$ enabled [16]. The DDA runs were processed in Fragpipe (dx.doi.org/10.17504/protocols.io.bp2l6j6ddvqe/v1), with oxidation on methionine and n-terminal acetylation set as modifications and otherwise default settings [17, 18, 19, 20, 21, 22, 23, 24, 26, 26]. The same process is true for DDA library creation. The DIA runs were processed in either Fragpipe or DIA-NN with the same modifications as DDA [12]. In Fragpipe, library-free DIA was performed using DIA_DIA-Umpire_SpecLib_Quant (dx.doi.org/10.17504/protocols.io.x54v9q46ql3e/v1) and DIA_SpecLib_Quant workflows (dx.doi.org/10.17504/protocols.io.261gey7ydv47/v1), using settings

default to that workflow. In DIA-NN, a DDA library generated in Fragpipe (dx.doi.org/10.17504/protocols.io.36wgqxkwklk5/v1 & dx.doi.org/10.17504/protocols.io.j8nlkzpjxl5r/v1) was used as well as the library-free workflow (dx.doi.org/10.17504/protocols.io.yxmvm8jr5g3p/v1 & dx.doi.org/10.17504/protocols.io.5qpvoe4mdl4o/v1). In both cases, protein names were set as the inference method and MBR was enabled, otherwise default settings. The first dataset comparing mass spectrometry parameters utilized Fragpipe v22.0 and DIA-NN v1.9.1. The later datasets utilized Fragpipe v23.0 and DIA-NN v2.0.

Post-processing for search engine outputs. The combined_ion.tsv output for DDA was already filtered to 1% false discovery (FDR) by Fragpipe. The report.parquet output for DIA was filtered to 1% FDR using columns 'Lib.PG.Q.Value', 'Lib.Q.Value' and 'Q.Value' in R, along with the rest of the post-processing [27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. Peptides were then either 1) filtered to proteotypic peptides (aside from duplicate protein sequences within a species, which were collapsed) or 2) assigned according to parsimony assignment logic. Parsimony assignment logic involves assigning peptides to the protein with the highest number of mapped peptides, continuing in this loop until all peptides are assigned [37, 38, 39]. Peptides mapping to potential contaminants were removed. The charge states and modifications were then collapsed to the maximum value for each peptide, in accordance with DIA-NN's documentation. The peptide, mapped protein, species and strain information was then used to establish pmartR components [40]. Using pmartR's workflow, peptide intensities were log2 transformed and median normalized, then protein rollup was performed using the median peptide value for each protein. For the 46-strain dataset, the resulting pmartR objects were then filtered to proteins with a minimum of two non-missing values amongst the three replicates for each of the two mixtures. pmartR's imd_anova function was then used to perform a two-sample t-test, assuming equal variance between the two mixtures and using a Benjamin-Hochberg FDR adjustment.

Scaling DIA-NN library-free searches to metaproteomes. We obtained KEGG annotations from BLAST Koala for the proteins found in one of the 46-strain mixtures processed in DIA-NN's library-free mode with only proteotypic peptides [41, 42]. The pathways for those KO annotations were then obtained via the KEGG API [43]. Pathway enrichment analysis using a "holm" adjusted Fisher's exact test was then performed comparing the pathways associated with the 60 most abundant, strain unique proteins to those associated with all the proteins found in that dataset (Supplementary Fig. 3). The 20 most significant pathways were then used to filter the original database as well as three external FASTAs. The external FASTAs were chosen because they had varying sizes and represented varying microbiome communities (gut metaproteome originally sourced from the Human Microbiome Project in 2020 (1,162,845 proteins), a collection of 35 soil isolates generated by the Joint Genome Institute (171,005 proteins) and a soil metaproteome also generated by JGI (2,360,231 proteins) [44, 45, 46]. Each of these three external databases were then combined with the original and used to re-search the 46-strain DIA dataset in DIA-NN library-free mode. Organism attribution depended on what was provided by the original FASTAs.

Metaproteomic Analysis with DIA-NN Library Free (MADL). By generalizing the conserved pathways used for FASTA filtering and enforcing stricter confidence thresholds for organism identification, we built on the testing described in the previous section and developed a refined pipeline for using DIA-NN's library-free search in metaproteomic samples, described below and in Fig. 1. (1) KEGG annotations of the FASTA are performed locally using a fork of the kofam_scan repository available on Github [47]. Associated pathways are obtained via a static Json provided by KEGG [43, 48, 49]. Organism attributions are again dependent on what's provided in the FASTA. (2) The annotated FASTA is filtered to a generalized list of pathways determined by previous studies to be conserved across bacterial species [50, 51]. After the pathway down-selected FASTA has been run through DIA-NN's library-free mode, the results are filtered to only proteotypic peptides with an initial FDR threshold of 0.01. A stricter $PG.Q.Value$ (protein group false discovery rate) threshold is set by generating a histogram of PG.Q.Values for these peptides (100 bins) and finding the first bin with a large drop in frequency (< 0.01% of the largest bar). The new threshold is the max value of the last high-count bin [52–54]. (3) After filtering to peptides with PG.Q.Values less than or equal to new threshold, the number of proteins associated with each organism are counted and only organisms with at least 5 filter passing proteins are considered "detected". (4) The original FASTA is re-filtered using the list of detected organisms and (5) run through DIA-NN "library free" mode again to obtain phylogenetic and function results. The full pipeline is available at TBD.

Statistics and Reproducibility. For the primary model microbiome experiments, peptide samples were prepared from mixtures described in Supplementary Tables 1, 2, and 3. Biological replicates were defined as independently processed aliquots of each model microbiome mixture. Technical replicates were not included, as sample processing and LC-MS/MS acquisition were performed once per biological replicate. The 24-strain dataset involved five mixtures with no biological replicates and no downstream statistics beyond peptide and protein counts. The 46-strain dataset had two mixtures, Mix1 and Mix2, with 3 biological replicates each. Statistical analyses were detailed in the relevant Results and Methods subsections. Relevant versioning for R (version 4.4.2) packages includes tidyverse (version 2.0.0), here (version 1.0.1), pmartR (version 2.4.6), microseq (version 2.1.6), arrow (version 18.1.0), data.table (version 1.16.2), readxl (version 1.4.3), readr (version 2.1.5), scales (version 1.3.0), cowplot (version 1.1.3). Relevant versioning for Python (version 3.11.0) packages includes pandas (version 3.0.1), pyarrow (version 23.0.1), ruby (version 3.4.8), hmmer (version 3.4), parallel (version 20170422) and matplotlib (version 3.10.8). Figures in this paper were minimally adjusted using Adobe Illustrator 2026.

Results

Simulating microbiome complexity. The goal of this study was to develop and validate a robust workflow using DIA-based proteomics for the characterization of microbiomes. We used mixtures of peptides isolated from axenic cultured microbial species to model microbiome-like complexity, where microbial species composition and concentrations were predetermined. To design these mixtures and to ensure they reflected realistic complexity, we assessed the phylogenetic relatedness of the available species. In total, we used 46 strain isolates from 28 bacterial species originating from a soil microbiome

[15]. The full list of strains and their genome characteristics is provided in Supplementary Tables 1 & 2. We utilized previously cultured and sequenced organisms and performed in silico trypsin digestion on their individual FASTA files. In-silico peptide sequences from all strains of each species were merged into single-species level lists. The relatedness of the species was measured as the percent of tryptic peptide sequences shared with any of the other available species (Supplementary Fig. 1). This analysis indicated that there is a wide range of relatedness in our pool of organisms. For instance, multiple species in our pool belong to the *Bacillus* Genus; while some have a high degree of genetic overlap (i.e., *Bacillus aryabhattai*), others were less genetically related (i.e., *Bacillus pumilus*). This variation highlights the variability of taxonomic assignment and the importance of independently assessing relatedness. Using this information, we designed five, 24-strain mixtures using the same subset of low, medium, and highly related species (23 species in total, Supplementary Table 1). When multiple strains were available for a given species, one was chosen at random. We also designed two 46-strain mixtures using all available strains to compare DDA and DIA results (Supplementary Table 2). We varied the proportion of each strain between the two 46-strain mixtures, providing ground truth for quantification assessments (Supplementary Table 3).

DIA window schemes. Our first task was to determine the optimal instrument and search settings to be used for the evaluation of the microbiome model samples. To refine the generation of DIA mass spectrometry data, the 24-strain mixtures were analyzed using various MS1 and MS2 window sizes (Supplementary Table 1). The raw files were processed in DIA-NN using 24- and 48-fraction DDA libraries from Fragpipe as well as DIA-NN's library-free mode, as described in the methods section [12, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. In total, the 24-fraction DDA library generated 373,490 peptides while the 48 fraction DDA library generated 435,226 peptides. At this stage, we restricted the data analysis to proteotypic peptides, i.e., experimentally observable peptides that each map to only one protein [55], since the process for assigning redundant peptides varies across the field and can have a large impact on results. The 24-fraction DDA library had 259,787 proteotypic peptides and the 48 fraction DDA library had 308,221 proteotypic peptides.

The DIA datasets with a 10 Da MS2 isolation windows, analyzed with DDA libraries, yielded an average of 57,371 peptide identifications, whereas a 20 Da window yielded 47,627 (Fig. 2A). The relatively large difference of about 10,000 peptides indicates the smaller MS2 window size enabled more efficient identification, likely because it produced less complex MS2 spectra [56]. Further, there was relatively low variation in the number of peptides identified across MS1 m/z ranges. This is indicated by the small standard deviation of peptide counts within the 10 Da MS2 window size (4,470 peptides) and 20 Da MS2 window size (2,184 peptides). This result indicates that the MS1 m/z range had little effect on the number of peptides detected. The common *m/z* range of 400–900 is where a large portion of detectable peptides fall due to factors such as tryptic peptide length and ionization efficiency [4, 56]. Overall, these results indicate that the number of peptide and protein identifications is more dependent on MS2 window size than m/z range.

We observed a similar number of peptides identified between the two DDA libraries (i.e., 24- and 48-fraction libraries) suggesting that higher levels of fractionation are not beneficial for less complex microbial samples, as increasing the number of fractions did not result in a significant increase in the identified peptides and proteins. Next, we performed a “library-free” search using DIA-NN to evaluate the performance of this algorithm on complex metaproteomes. Using a 10 Da MS2 window, we identified an average of 71,494 peptides, which represents about a 25% increase in peptide identifications without relying on DDA data. Between these two strategies, only 13% of peptides and 9% of proteins were not observed using the library-free mode in place of a DDA library (Figs. 2B and 2C). The overlap in DIA identifications between library-free mode and the 24-fraction DDA library suggests that both approaches provide comparable information overall, lending confidence to the newer in silico methods.

Identification and quantitation. Having identified a suitable DIA instrument method for microbiome applications, the next step was to increase complexity by using the 46-strain mixtures to further investigate various processing methods and compare their results to DDA. DIA data were generated using a 300–900 Da MS1 m/z range, 10 Da MS2 window size, and a total of 6 runs (2 mixtures x 3 biological replicates). DDA data were generated using high pH reverse phase fractionation, resulting in 72 total runs (2 mixtures x 3 biological replicates x 12 fractions). Since each fraction was analyzed with a 2-hour gradient, the total analysis time for DDA was 144 hours, while the total analysis time for DIA was 12 hours. The 12 fraction DDA data was also used to generate a library, which had 379,015 total peptides and 271,829 proteotypic peptides. Considering the previous success of DIA-NN’s library-free mode, we evaluated two additional processing methods from Fragpipe that do not rely on DDA data: DIA_DIA-Umpire_SpecLib_Quant and DIA_SpecLib_Quant [16, 17, 18, 19, 20, 21, 22, 23, 24, 25].

The overall number of peptides identified does not appear to vary between the two 46-strain mixtures. DDA combined with fractionation yielded more average peptide (104,199) and protein (28,813) identifications than any of the DIA methods tested, including library-free (86,368 peptides and 24,176 proteins) (Fig. 3A). To calculate missingness, we defined the total peptide identification space as the set of all peptides identified across mixtures and replicates for a given acquisition mode (DDA vs DIA) and search method (library-based vs. library-free). We then calculated missingness as the percentage of those peptide values missing for each replicate within that acquisition mode and search method. DDA has the highest percentage of missing peptides in Mix1 (46.6%) and Mix2 (44.5%) when averaged across the three replicates, compared with DIA library-free, which has a 26.8% missingness for Mix1 and 26.3% for Mix2. The average number of peptide identifications from DIA in Fragpipe was 45.4% less than the average for DIA-NN. To simplify future computations, we decided to focus further comparisons on DIA results from DIA-NN.

One consideration in comparing DDA to DIA is the instrument time needed to fully analyze each sample. The DDA data was collected using two-dimensional liquid chromatography (2D-LC) reverse phase analysis to reflect the current approach applied in our lab, while DIA was one-dimensional liquid chromatography (1D-LC) reverse phase analysis. The total analysis time per mixture was 72 hours for DDA and 6 hours for DIA. When normalized for analysis time, the peptide identification yield in DIA mode

is nearly 900% higher than in DDA mode (Fig. 3B). Importantly, the DDA and DIA library-free data showed substantial overlap in peptide and protein identifications (Fig. 3C). Despite identifying fewer proteins than our DDA approach, DIA enables rapid microbiome characterization and can dramatically expand metaproteomics throughput, allowing larger experimental designs across space, time, and cohort size.

Up to this point, our comparisons have focused on qualitative assessment of peptide identifications. However, quantitative change in protein abundance can provide insight into the underlying functional phenotypes in microbial samples. To assess the quantitative performance of DIA and DDA, we tested for differential abundance (T-test with Benjamini-Hochberg FDR adjustment) of each protein. Directionality was determined from the log-transformed fold change between Mix1 and Mix2 in the 46-strain dataset. Since we altered the proportions of the peptides from each strain between Mix1 and Mix2, we expected the protein abundances to differ significantly between the two peptide mixtures. Previously, our assessment only included proteotypic peptides, because non-proteotypic peptides can introduce ambiguity. Here, to evaluate protein-level abundance, we compare a proteotypic-only approach with an approach that uses both proteotypic and non-proteotypic peptides assigned using the parsimony attribution approach [37, 38, 39]. This is a popular method that assigns peptides to the protein with the highest number of mapped peptides, as detailed in the methods. Importantly, all DIA processing strategies resulted in the detection of more significantly modulated proteins than DDA (i.e., over 14,776 proteins in DIA library-free vs. 7,817 in DDA), indicating improved quantification, likely due to fewer missing peptide values (Fig. 4A). Most statistically significant proteins also had the correct fold change direction (Fig. 4B). By utilizing the parsimony method, as described in the post processing section of the methods, to incorporate non-proteotypic peptides, we detected 6,824 additional proteins correctly quantified in DIA library-free and 3,954 in DDA. This suggests that strategies for peptide assignments, such as the parsimony method, can further improve quantification. We did not comprehensively evaluate the different peptide assignment methods as this was not the focus of our study. Since the proportions of the strains and species in each mixture are known, we tested if we were able to reconstitute the microbial composition by adding together the protein intensities of each species or strain and calculating it as a proportion of the total intensity for Mix1 (Fig. 4C & D). Across all the DIA processing strategies and the DDA data, Mix1 had proteins attributed to nearly all the same species (between 26 and 28 out of 28 species) and strains (between 37 and 44 out of 46 strains). The proportions are also consistent across processing and acquisition strategies, even when they deviate from the known proportions. This holds true with the use of the parsimony assignment method. This analysis demonstrates that DIA matches DDA's ability to accurately characterize highly complex mixtures, capturing not only protein abundance changes but also species composition.

The recently released Astral mass spectrometers (Thermo Scientific) can achieve DIA scan rates that exceed 200 Hz with its design incorporating an astral mass analyzer, whereas Orbitrap only instruments are typically limited to 40–60 hertz. Preliminary reports in the literature suggest the Astral will further improve identification efficiency and quantitative accuracy [57]. Therefore, we set out to benchmark this instrument against our model system and the processing approach described above. Each 46-strain mixture was analyzed using DIA mode on the Astral instrument (3 biological replicates, each run had an

analysis time of 21.88 minutes) for a total of 1.1 hours. Using DIA-NN's library-free mode and only proteotypic peptides, we report that the Astral 1-hour analysis time yielded 12,877 more peptide and 1,017 protein identifications, averaged across mixtures, than the Exploris 6-hour analysis time (Fig. 5A). The Astral also provides a similar number of correctly quantified proteins (10,506 proteins for the Astral vs. 10,624 for the Exploris), indicating that protein quantification is not impacted by the reduced run time on the Astral (Fig. 5C). Normalizing for instrument run time, this is a 468% improvement in hourly protein identification (Fig. 5B).

Computational metagenomic approaches. Having confirmed that DIA library-free is a highly effective alternative to a DDA data-generated library, we next need to assess whether it can be used when searching against the large metagenomes typical of natural microbiomes, which range in microbial diversity from 100's (gut microbiome) to 1000's (soil microbiomes) of different species. We attempted running library-free on the 46-strain dataset using an external soil metaproteome (2,360,231 proteins) amended with the metaproteome from our 46 strains (248,658 proteins), but DIA-NN failed to complete the search and did not provide an error log file or any other output. We determined that a new approach was necessary for running DIA-NN library free on microbiome samples, both due to this run failure and because searching a metaproteome of that size has been known to produce false discovery issues even if completed [58].

The main goals of metaproteomics are (i) to identify which organisms are present and actively functioning in a community, and (ii) to determine what functions are performed by this community. To fulfill the goals of metaproteomics in a computationally efficient fashion, we developed a two-pass search approach which uses DIA-NN's library free mode: the first pass is used to identify the organisms potentially present in the samples, and the second pass is used for the functional characterization of the microbiome. We termed this approach Metaproteomic Analysis with DIA-NN Library-free (MADL). Previous work has demonstrated that multiple pass approaches can improve the performance of large database searches [58]. MADL builds on this concept by improving the stepwise logic and imposing strict FDR thresholds for organism identification. The full MADL pipeline was deposited on Github at TBD, allowing people to reproduce our approach.

In the first pass of MADL, we subset the search database to proteins from conserved pathways. Most microorganisms conserve essential processes, such as glycolysis, TCA cycle, ATP generation, ribosomal replication, or DNA replication. While core metabolic and other essential functions are shared across species, evolutionary pressure drives discrepancies in the genes and proteins involved. To test the viability of MADL's first pass, we needed to confirm that these essential processes exist for the organisms in our dataset and that they will reduce the search databases to a size reasonable for DIA-NN library free. To confirm that core processes are conserved in our 46-strain dataset, we performed KEGG pathway enrichment analysis comparing pathways for the 60 highest intensity proteins per strain, to all pathways identified in the dataset. As anticipated, the enriched pathways comprised many core functions, including "Glycolysis/Gluconeogenesis", "TCA cycle", "Microbial metabolism in diverse environments", "Ribosome", or "Nucleotide metabolism" (Supplementary Fig. 3A). We also confirmed

that these enriched pathways were specific enough to reduce both our FASTA and three external metaproteomes of varying size, as detailed in the methods (Supplementary Fig. 3B). Note that a large portion of the decrease in size came from whether we were able to attribute a Kegg Ontology (KO) to the protein sequences present in the FASTA.

Having confirmed the viability of our first pass strategy, we next needed to remove bias towards our test dataset and validate the full MADL pipeline. We therefore performed the full two-pass search, as depicted in Fig. 1, using our FASTA and the three external metaproteomes. (1) We annotated all three test FASTA databases (three external metaproteomes, each with our original database appended) and (2) filtered to proteins sequences associated with pathways previously reported as broadly conserved across bacterial species, as described in the methods section [50, 51]. (3) Each of these pathways filtered databases were then used to reprocess the 46-strain dataset in DIA-NN library free mode. To generate a list of 'detected' organisms for each set of results, identifications were first restricted to proteotypic peptides with an initial FDR of 0.01. Histograms of DIA-NN protein group confidence values (PG.Q.Value) were then generated after applying a -log10 transformation. To remove low confidence tails, each set of results received a new PG.Q.Value threshold defined as the upper boundary of the last histogram bin whose frequency was at least 0.01% of the maximum bin height. (4) Each test FASTA was then filtered to its list of detected organisms and (5) used to re-process the 46-strain dataset in DIA-NN, obtaining taxonomic and functional information on the sample.

The generalized set of pathways successfully reduced the test FASTAs to a size tractable for DIA-NN library-free processing (e.g. about 250,000 sequences) and successfully enabled detection of 25 of the 28 known organisms across all three test FASTAs (Fig. 6A, B & C). Based on the results of the two-pass pipeline, the number of falsely identified species varied by background: 1 in the soil genome composite, 8 in the gut genome composite, and 12 in the soil isolate composite (Fig. 6B & C). However, these false species accounted for at most 2.59% of the total signal intensity, comparable to the 1.94% of the total intensity attributed to external species by DDA when searching with the gut genome composite. In addition, the KO coverage of the 20 most abundant pathway groups in Mix1 was essentially unchanged, indicating that the functional information captured by these pathways is robust to both missing and spurious organism identifications (Fig. 6D). Together, these results demonstrate that this first pass search strategy supports accurate organism identification in complex mixtures and can be combined with downstream functional analysis, thereby enabling DIA-NN's library-free search on large metaproteomes.

Discussion

Prior work has shown that data-independent acquisition (DIA) offers clear advantages over data-dependent acquisition (DDA), including improved quantification and reduced missingness, even in the complex case of metaproteomics. However, most studies stop short of defining a complete, metaproteomics-specific DIA workflow that spans from LC-MS/MS data acquisition through to data analysis. A frequently cited obstacle to a fully DDA-independent pipeline is the computational burden of

library-free approaches when applied to very large metaproteomic databases, even though these methods have proven effective in more limited, well-defined systems.

Within this context, we systematically evaluated DIA acquisition parameters and found that MS2 window size is a major determinant of DIA identification performance, whereas refinement of the MS1 m/z range has a relatively minor impact on the number of proteins and peptides identified. A smaller MS2 window size appears to be preferable for complex metaproteomic samples, likely because it reduces peptide interferences in the already complex DIA spectra. As DIA scoring algorithms continue to improve, larger MS2 window sizes may enable higher throughput without sacrificing proteome depth. As LC-MS/MS instrumentation evolves, we anticipate incremental increase in throughput in metaproteome analyses, as demonstrated by the substantially higher identification yield on the newer Astral instruments compared to the older Exploris. The comparable number of correctly quantified proteins on the faster Astral instrument suggests that aggressive DIA duty cycles can be adopted without compromising quantitation accuracy. Collectively, these refinements in LC-MS/MS DIA can enable study designs with many more biological replicates, time points or environmental conditions. In the long term, this increase in throughput creates a powerful synergy for the age of AI; by generating massive datasets required for machine learning, our workflow enables the discovery of subtle, complex patterns in microbiome function that were previously beyond our analytical reach.

DIA's ability to recover information is comparable to deeply fractionated DDA data, but in an order-of-magnitude less instrument time, makes it the practical choice for experiments focused on broad functional profiling and comparative analyses across many samples. However, fractionated DDA data may remain the preferred method when fewer samples are being processed or when cleaner MS/MS for maximal sequence coverage is required, for instance, in novel protein or post-translational modification (PTM) discovery. Future work should focus on characterizing the physicochemical properties of the proteins or peptides uniquely identified by DDA to determine if they represent a specific biological class that is systematically under sampled by DIA. Aside from such targeted objectives, the marginal gain in identifications from extensively fractionated DDA is small relative to the cost in instrument time and sample consumption. Thus, for the majority of modern metaproteomic studies, focused on systems-level comparison, DIA represents the most judicious investment of analytical resources.

A major bottleneck in applying DIA-NN's library-free mode to real microbiomes is the computational cost of searching against large metaproteomic FASTAs. Without a focused spectral library, the algorithm must consider every possible peptide from the entire database, causing the search space to scale exponentially with the database size. Our MADL two-pass approach partly circumvents this limitation by effectively shrinking the search space while preserving organism-level and functional readouts. The low false organism attribution (< = 2.59%) across various sources of background noise, including soil isolates, a human gut metaproteome, and a soil metaproteome, suggests that our approach is robust and scalable.

Despite this success, we observed that the filtering step led to the omission of a few organisms. Specifically, *Streptomyces ambofaciens*, *Streptomyces albogriseolus*, and *Chryseobacterium gleum* were missed in the initial down selected FASTA, while *Bacillus aryabhattai* was missed across all three noise-added conditions. The exclusion of the *Streptomyces* and *Bacillus aryabhattai* may be explained by their high phylogenetic relatedness to other organisms in the community (Supplementary Fig. 1), making them difficult to discern from background noise or from other organisms present. This ambiguity was likely amplified by the presence of an unclassified species of *Streptomyces* in the gut metaproteome database. In contrast, the reason for the exclusion of *Chryseobacterium gleum*, which has a medium relatedness with the rest of the organism pool, remains unclear and requires further investigations.

With the computational hurdle of database size effectively removed by our approach, the quality and completeness of the input metagenome now become the primary determinant of analytical success. As such, the use of long-read sequencing to produce high-quality, contiguous metaproteome-assembled genomes (MAGs) are likely to improve the true and false-positive organism attribution by reducing genomic fragmentation and peptide attribution ambiguity. Equally critical are improvements in functional annotation, especially for less-studied organisms, as one of the main goals of metaproteome analyses is to determine the functional activity of bacterial communities. Investments in sequencing and functional sequence attributions will therefore be directly beneficial to both this pipeline and to metaproteomics analyses in general.

Our comparison between proteotypic-only and parsimony-based approaches demonstrates that peptide assignment strategies can substantially increase the number of correctly quantified proteins, especially in DIA. However, the additional complexity this introduces deserves deeper scrutiny, particularly for closely related organisms. Establishing standardized benchmarks to evaluate these peptide-to-protein attribution strategies for organisms with varying relatedness, as we've initiated here at the species and strain level, is crucial for ensuring that metaproteomics results are consistent and comparable across studies.

Conclusions

In conclusion, we establish the Metaproteomic Analysis with DIA Library-free (MADL) pipeline as a practical and scalable approach for high-throughput metaproteomic analyses. By combining DIA acquisition with a two-pass library-free search, MADL empowers researchers to replace extensive DDA library generation in most studies, dramatically increasing the throughput while maintaining high organismal and functional resolution. While DDA-based strategies will remain essential for specialized applications like de novo protein discovery or the establishment of the role of post-translational modifications in communities, MADL is poised to become a central method for large-scale, quantitative metaproteomics studies that are critical for advancing our understanding of complex microbial systems.

Declarations

Funding

This research was performed on a project award (10.46936/intm.proj.2023.60905/60008964) from the Environmental Molecular Sciences Laboratory, a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830. Portions of this work were supported by the National Microbiome Data Collaborative with funding from the Genomic Science Program in the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract numbers DE-AC02-05CH11231 (LBNL) and DE-AC05-76RL01830 (PNNL). PNNL is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

Author Contribution

IKA, GCC, MSL, and PDP conceived the design and implementation of the research. VNP performed sample preparation. IKA acquired the instrument data. SAO, GCC performed data and statistical analysis, and SAO performed the computations. GCC lead the development effort of MADL. PDP, MSL, GCC and AB verified the data and statistical analysis methods. LAM was part of a project that was used to inform decisions and experiments in this research. SAO lead the manuscript writing. All authors contributed by providing critical feedback that helped shape the research, analysis and contributed to the final manuscript.

Acknowledgement

Thank you to the Pacific Northwest National Laboratory's Persistent Control Research team and Soil SFA microbial communities' team, especially to Dr. Rob Egbert, Dr. Ryan McClure, and Andrew Wilson for providing the cultured microbes.

Data Availability

We will make all data and code available.

References

1. Van Den Bossche T, Armengaud J, Benndorf D, Blakeley-Ruiz JA, Brauer M, Cheng K, et al. The microbiologist's guide to metaproteomics. iMeta. 2025;4(3):e70031. doi:10.1002/imt2.70031
2. Dowell JA, Wright LJ, Armstrong EA, Denu JM. Benchmarking Quantitative Performance in Label-Free Proteomics. ACS Omega. 2021;6(4):2494–504. doi:10.1021/acsomega.0c04030
3. Xian F, Brenek M, Krisp C, Urbauer E, Ravi Kumar RK, Aguanno D, et al. Ultra-sensitive metaproteomics redefines the dark metaproteome, uncovering host-microbiome interactions and

drug targets in intestinal diseases. Nat Commun. 2025;16:6644. doi:10.1038/s41467-025-61977-7 PubMed PMID: 40681571; PubMed Central PMCID: PMC12274446.

4. Zhang Y, Bilbao A, Bruderer T, Luban J, Strambio-De-Castillia C, Lisacek F, et al. The Use of Variable Q1 Isolation Windows Improves Selectivity in LC–SWATH–MS Acquisition. J Proteome Res. 2015;14(10):4359–71. doi:10.1021/acs.jproteome.5b00543
5. Barkovits K, Pacharra S, Pfeiffer K, Steinbach S, Eisenacher M, Marcus K, et al. Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-based Data-independent Acquisition*. Molecular & Cellular Proteomics. 2020;19(1):181–97. doi:10.1074/mcp.RA119.001714
6. Bruderer R, Bernhardt OM, Gandhi T, Miladinović SM, Cheng LY, Messner S, et al. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues*[S]. Molecular & Cellular Proteomics. 2015;14(5):1400–10. doi:10.1074/mcp.M114.044305
7. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, 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. 2012;11(6):O111.016717. doi:10.1074/mcp.O111.016717
8. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol. 2018;14(8):e8126. doi:10.15252/msb.20178126 PubMed PMID: 30104418; PubMed Central PMCID: PMC6088389.
9. Rajczewski AT, Blakeley-Ruiz JA, Meyer A, Vintila S, McIlvin MR, Van Den Bossche T, et al. Data-Independent Acquisition Mass Spectrometry as a Tool for Metaproteomics: Interlaboratory Comparison Using a Model Microbiome. PROTEOMICS. 2025;25(9–10):e202400187. doi:10.1002/pmic.202400187
10. Lou R, Cao Y, Li S, Lang X, Li Y, Zhang Y, et al. Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. Nat Commun. 2023;14(1):94. doi:10.1038/s41467-022-35740-1
11. Fröhlich K, Brombacher E, Fahrner M, Vogele D, Kook L, Pinter N, et al. Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. Nat Commun. 2022;13(1):2622. doi:10.1038/s41467-022-30094-0
12. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods. 2020;17(1):41–4. doi:10.1038/s41592-019-0638-x
13. Sinitcyn P, Hamzeiy H, Salinas Soto F, Itzhak D, McCarthy F, Wichmann C, et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. Nat Biotechnol. 2021;39(12):1563–73. doi:10.1038/s41587-021-00968-7
14. Martinez-Val A, Bekker-Jensen DB, Hogrebe A, Olsen JV. Data Processing and Analysis for DIA-Based Phosphoproteomics Using Spectronaut. Methods Mol Biol. 2021;2361:95–107. doi:10.1007/978-1-0716-1641-3_6 PubMed PMID: 34236657.

15. Fonseca-García C, Pettinga D, Wilson A, Elmore JR, McClure R, Atim J, et al. Defined synthetic microbial communities colonize and benefit field-grown sorghum. ISME J. 2024;18(1):wrae126. doi:10.1093/ismejo/wrae126 PubMed PMID: 38984785; PubMed Central PMCID: PMC11410050.
16. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008;24(21):2534–6. doi:10.1093/bioinformatics/btn323
17. Teo GC, Polasky DA, Yu F, Nesvizhskii AI. A fast deisotoping algorithm and its implementation in the MSFragger search engine. J Proteome Res. 2021;20(1):498–505. doi:10.1021/acs.jproteome.0c00544 PubMed PMID: 33332123; PubMed Central PMCID: PMC8864561.
18. da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang HY, Shanmugam AK, Mellacheruvu D, et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods. 2020;17(9):869–70. doi:10.1038/s41592-020-0912-y PubMed PMID: 32669682; PubMed Central PMCID: PMC7509848.
19. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in shotgun proteomics. Nat Methods. 2017;14(5):513–20. doi:10.1038/nmeth.4256 PubMed PMID: 28394336; PubMed Central PMCID: PMC5409104.
20. Tsou CC, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras AC, et al. DIA-Umpire: comprehensive computational framework for data independent acquisition proteomics. Nat Methods. 2015;12(3):258–64. doi:10.1038/nmeth.3255 PubMed PMID: 25599550; PubMed Central PMCID: PMC4399776.
21. Yang KL, Yu F, Teo GC, Li K, Demichev V, Ralser M, et al. MSBooster: improving peptide identification rates using deep learning-based features. Nat Commun. 2023;14:4539. doi:10.1038/s41467-023-40129-9 PubMed PMID: 37500632; PubMed Central PMCID: PMC10374903.
22. Yu F, Teo GC, Kong AT, Haynes SE, Avtonomov DM, Geiszler DJ, et al. Identification of modified peptides using localization-aware open search. Nat Commun. 2020;11:4065. doi:10.1038/s41467-020-17921-y PubMed PMID: 32792501; PubMed Central PMCID: PMC7426425.
23. Yu F, Haynes SE, Teo GC, Avtonomov DM, Polasky DA, Nesvizhskii AI. Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. Mol Cell Proteomics. 2020;19(9):1575–85. doi:10.1074/mcp.TIR120.002048 PubMed PMID: 32616513; PubMed Central PMCID: PMC7996969.
24. Yu F, Haynes SE, Nesvizhskii AI. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. Mol Cell Proteomics. 2021;20:100077. doi:10.1016/j.mcpro.2021.100077 PubMed PMID: 33813065; PubMed Central PMCID: PMC8131922.
25. Yu F, Teo GC, Kong AT, Fröhlich K, Li GX, Demichev V, et al. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat Commun. 2023;14(1):4154. doi:10.1038/s41467-023-39869-5

26. Yu F, Deng Y, Nesvizhskii AI. MSFragger-DDA+ Enhances Peptide Identification Sensitivity with Full Isolation Window Search. bioRxiv. 2024;2024.10.12.618041. doi:10.1101/2024.10.12.618041 PubMed PMID: 39463976; PubMed Central PMCID: PMC11507693.
27. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, et al. data.table: Extension of “data.frame” [Internet]. 2026 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/data.table/index.html
28. Snipen L, Liland KH. microseq: Basic Biological Sequence Handling [Internet]. 2025 [cited 2026 Jul 7]. Available from: https://cran.r-project.org/web/packages/microseq/index.html
29. Iannone R, Cheng J, Schloerke B, Haughton S, Hughes E, Lauer A, et al. gt: Easily Create Presentation-Ready Display Tables [Internet]. 2026 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/gt/index.html
30. Müller K, Bryan J. here: A Simpler Way to Find Your Files [Internet]. 2025 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/here/index.html
31. Richardson N, Cook I, Crane N, Dunnington D, François R, Keane J, et al. arrow: Integration to “Apache” “Arrow” [Internet]. 2026 [cited 2026 Jul 7]. Available from: https://cran.r-project.org/web/packages/arrow/index.html
32. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686. doi:10.21105/joss.01686
33. Wickham H, Bryan J, Posit, attribution) P (Copyright holder of all R code and all C code without explicit copyright, code) MK (Author of included R, code) KV (Author of included libxls, et al. readxl: Read Excel Files [Internet]. 2025 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/readxl/index.html
34. Wickham H, Hester J, Francois R, Bryan J, Bearrows S, Software P, et al. readr: Read Rectangular Text Data [Internet]. 2026 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/readr/index.html
35. Wickham H, Pedersen TL, Seidel D, Software P, PBC [cph, fnd. scales: Scale Functions for Visualization [Internet]. 2025 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/scales/index.html
36. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2” [Internet]. 2025 [cited 2026 Apr 11]. Available from: https://cran.r-project.org/web/packages/cowplot/index.html
37. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, et al. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. J Proteome Res. 2009;8(8):3872–81. doi:10.1021/pr900360j PubMed PMID: 19522537; PubMed Central PMCID: PMC2810655.
38. Dou Y, Liu Y, Yi X, Olsen LK, Zhu H, Gao Q, et al. SEPepQuant enhances the detection of possible isoform regulations in shotgun proteomics. Nat Commun. 2023;14(1):5809. doi:10.1038/s41467-023-41558-2

39. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res. 2007;6(9):3549–57. doi:10.1021/pr070230d PubMed PMID: 17676885; PubMed Central PMCID: PMC2810678.
40. Stratton KG, Webb-Robertson BJM, McCue LA, Stanfill B, Claborne D, Godinez I, et al. pmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data. J Proteome Res. 2019;18(3):1418–25. doi:10.1021/acs.jproteome.8b00760
41. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. Journal of Molecular Biology. 2016;Computation Resources for Molecular Biology428(4):726–31. doi:10.1016/j.jmb.2015.11.006
42. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587–92. doi:10.1093/nar/gkac963
43. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. doi:10.1093/nar/28.1.27 PubMed PMID: 10592173; PubMed Central PMCID: PMC102409.
44. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. Genome Res. 2009;19(12):2317–23. doi:10.1101/gr.096651.109 PubMed PMID: 19819907; PubMed Central PMCID: PMC2792171.
45. Zhalnina K, Louie KB, Hao Z, Mansoori N, da Rocha UN, Shi S, et al. Dynamic root exudate chemistry and microbial substrate preferences drive patterns in rhizosphere microbial community assembly. Nat Microbiol. 2018;3(4):470–80. doi:10.1038/s41564-018-0129-3 PubMed PMID: 29556109.
46. Foley MM, Blazewicz SJ, McFarlane KJ, Greenlon A, Hayer M, Kimbrel JA, et al. Active populations and growth of soil microorganisms are framed by mean annual precipitation in three California annual grasslands. Soil Biology and Biochemistry. 2023;177:108886. doi:10.1016/j.soilbio.2022.108886
47. Aramaki T. takaram/kofam_scan [Ruby] [Internet]. 2026 [cited 2026 Mar 27]. Available from: https://github.com/takaram/kofam_scan
48. Abulfaraj AA, Shami AY, Alotaibi NM, Alomran MM, Aloufi AS, Al-Andal A, et al. Exploration of genes encoding KEGG pathway enzymes in rhizospheric microbiome of the wild plant Abutilon fruticosum. AMB Express. 2024;14:27. doi:10.1186/s13568-024-01678-4 PubMed PMID: 38381255; PubMed Central PMCID: PMC10881953.
49. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182-185. doi:10.1093/nar/gkm321 PubMed PMID: 17526522; PubMed Central PMCID: PMC1933193.
50. Lee SH, Lee H, You HS, Sung H joong, Hyun SH. Metabolic pathway prediction of core microbiome based on enterotype and orotype. Front Cell Infect Microbiol. 2023;13. doi:10.3389/fcimb.2023.1173085

51. Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. Nat Commun. 2023;14(1):7318. doi:10.1038/s41467-023-43000-z
52. McKinney W. Data Structures for Statistical Computing in Python. In. 2010. p. 56–61. doi:10.25080/Majora-92bf1922-00a
53. Hunter J. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007;9:90–5. doi:10.1109/MCSE.2007.55
54. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62. doi:10.1038/s41586-020-2649-2
55. Kuster B, Schirle M, Mallick P, Aebersold R. Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol. 2005;6(7):577–83. doi:10.1038/nrm1683
56. Zhang H, Bensaddek D. Narrow Precursor Mass Range for DIA–MS Enhances Protein Identification and Quantification in Arabidopsis. Life (Basel). 2021;11(9):982. doi:10.3390/life11090982 PubMed PMID: 34575131; PubMed Central PMCID: PMC8469718.
57. Dumas T, Martinez Pinna R, Lozano C, Radau S, Pible O, Grenga L, 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(1):46. doi:10.1186/s40168-024-01766-4
58. Muth T, Kolmeder CA, Salojärvi J, Keskitalo S, Varjosalo M, Verdam FJ, et al. Navigating through metaproteomics data: A logbook of database searching. PROTEOMICS. 2015;15(20):3439–53. doi:10.1002/pmic.201400560

Figures

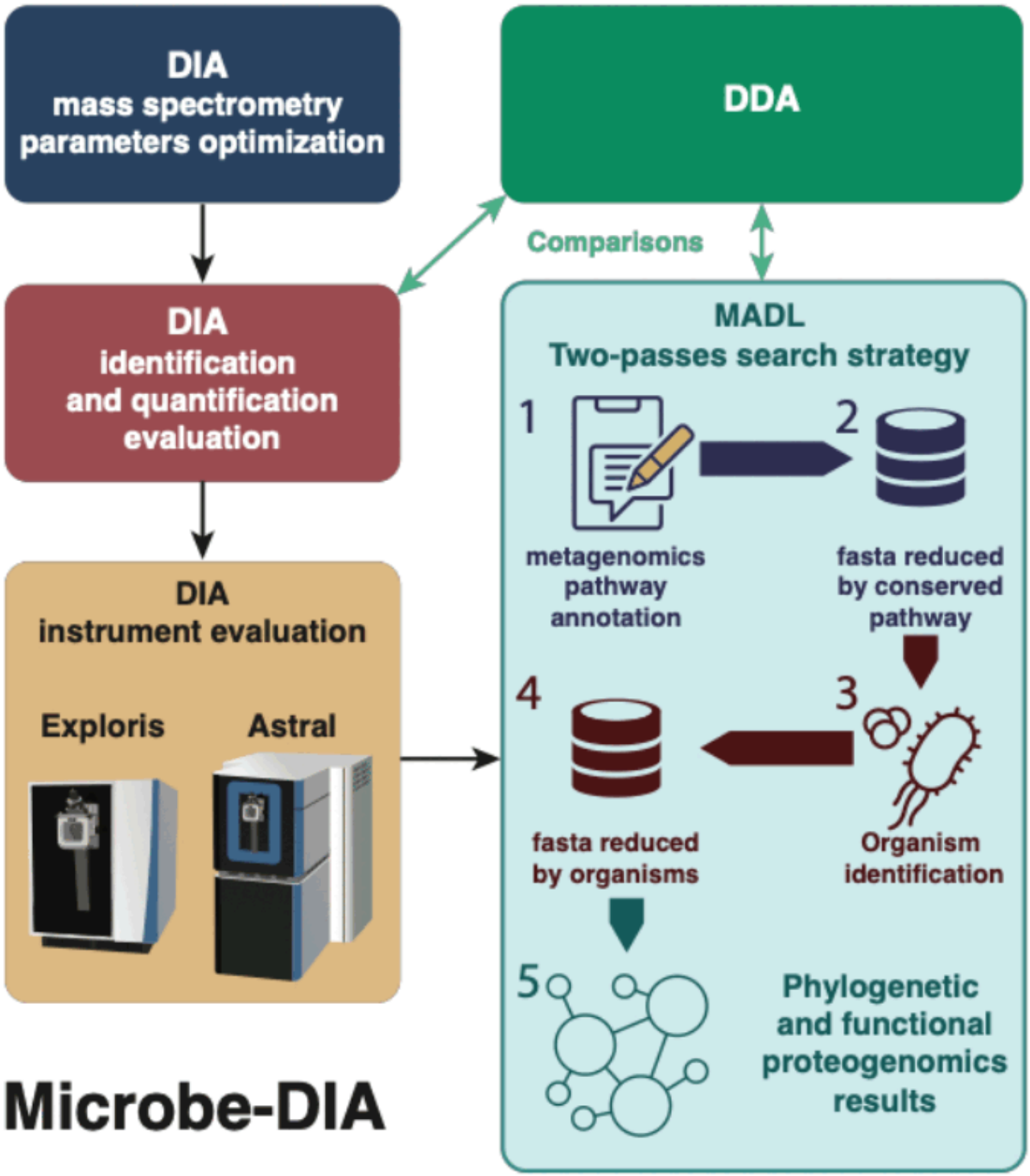


Figure 1

Graphical abstract of the experimentation described in this paper, including the optimization of LC-MS/MS parameters, instrumentation, and analysis. These results were used to develop and validate the Metaproteomic Analysis with DIA-NN Library free (MADL) pipeline.

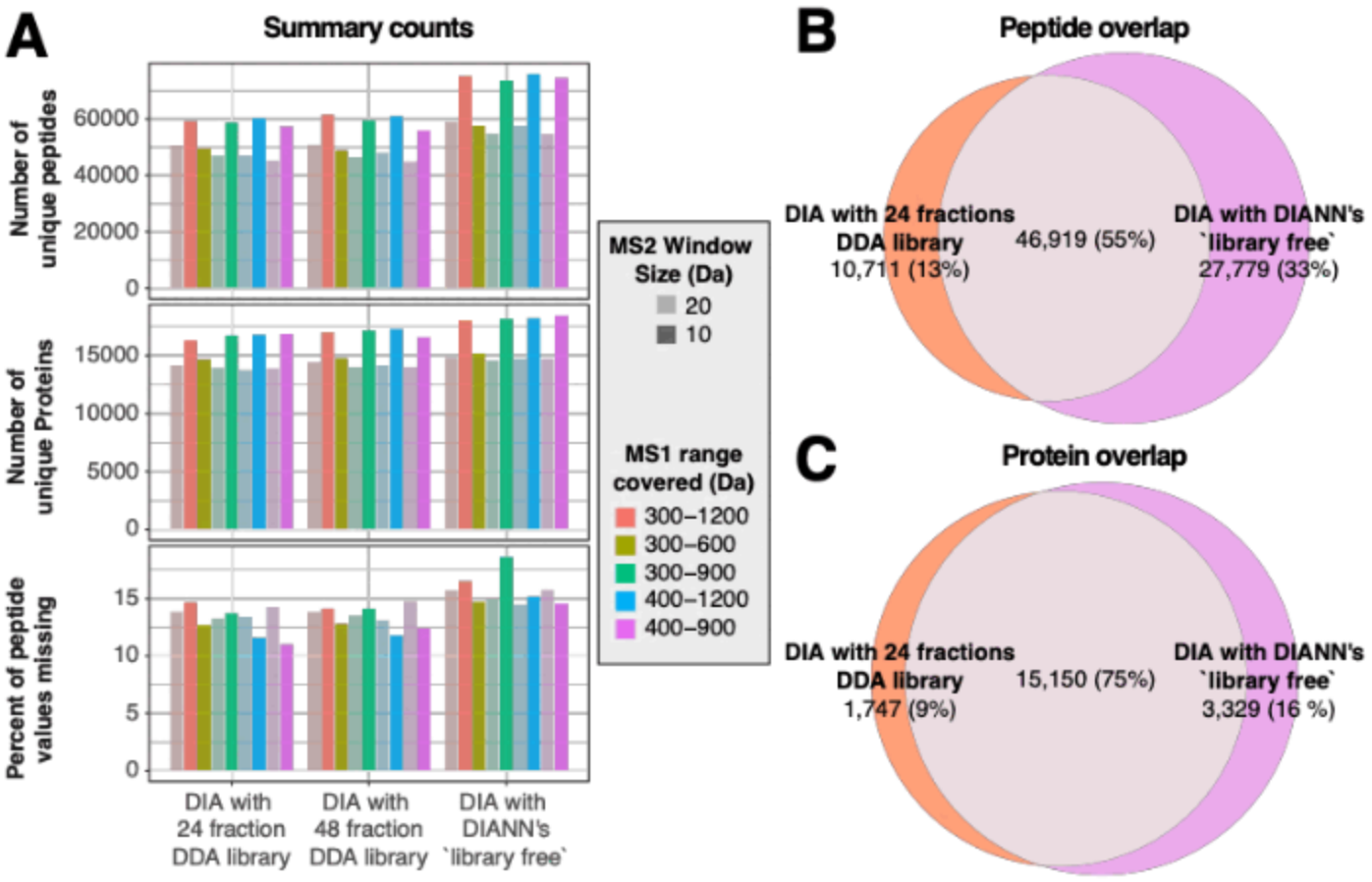


Figure 2

Comparison of DIA window schemes. **(A)** Mix1 of the 24-strain dataset was analyzed by DIA using different MS1 m/z ranges (indicated by bar color), MS2 window sizes (transparent and solid bars), and library types. Results indicate the number of peptide and protein identifications mostly depends on the MS2 window size and using a DDA library versus the in-silico library-free method. Smaller MS2 window sizes and library-free searches yield more detections. Venn diagrams show the overlap in peptide **(B)** and protein identifications **(C)** when DIA data is processed with a DDA library (orange) versus a library-free approach (purple).

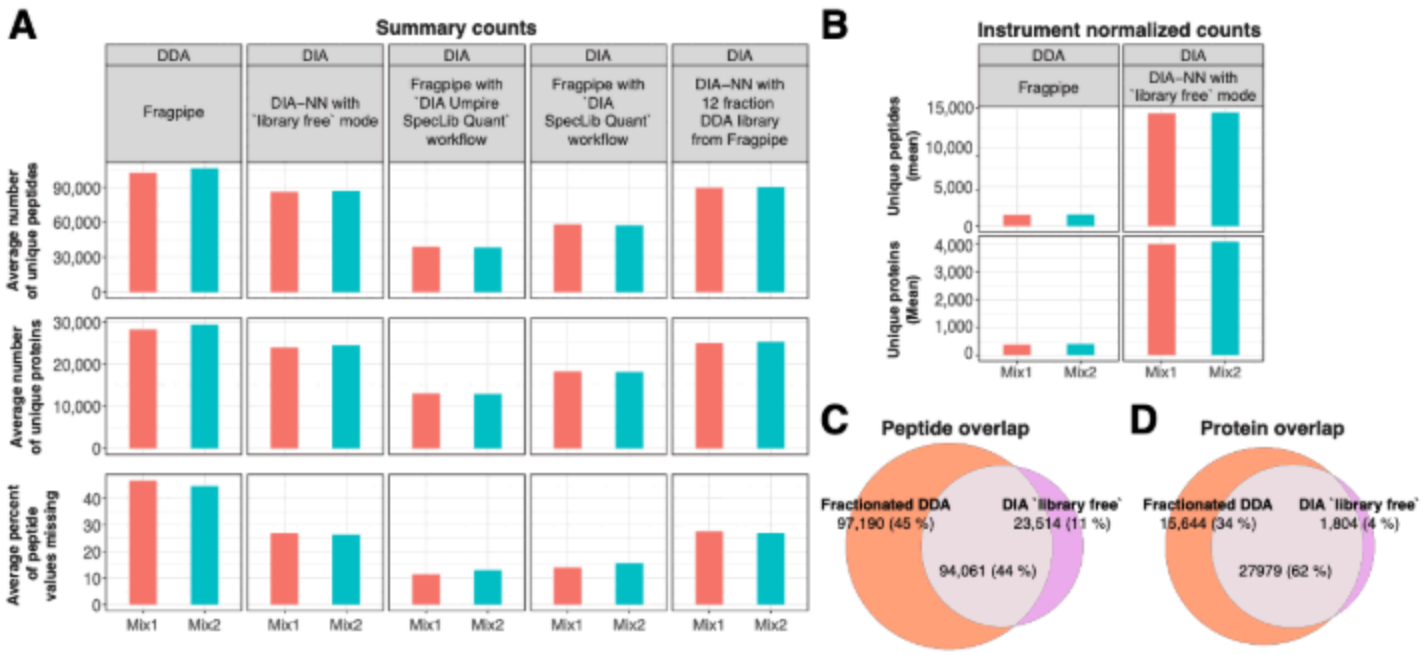


Figure 3

Comparison of DDA and DIA identifications in Mix1 (red) and Mix2 (blue) of the 46-strain dataset using various software tools. **(A)** The number of peptide and protein identifications is lower for DIA than DDA, but DIA has lower percentages of missing peptide values. Amongst the DIA methods, processing DIA with the 12 fraction DDA library provides nearly the same number of identifications as library-free. **(B)** Comparing DDA to the best-performing DIA method, library-free has nearly double the efficiency of DDA when normalized for instrument run time (6 hours per mix vs. 72 hours per mix). **(C)** There is also general agreement amongst the Mix1 peptide and protein identifications made by DDA (orange) and the DIA library-free approach (purple).

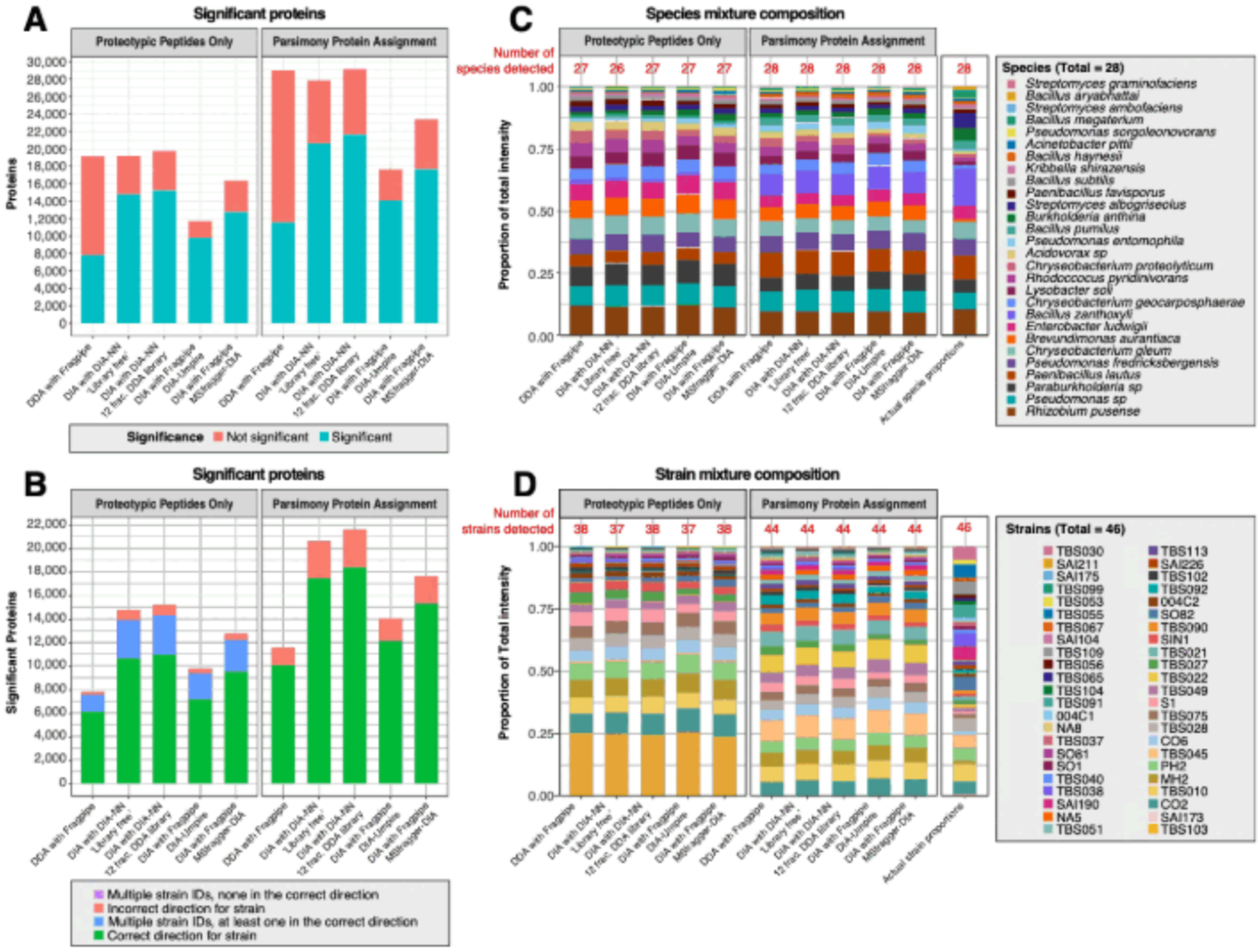


Figure 4

Comparison of DDA and DIA quantification. Quantification of the 46-strain mixtures indicates that**(A)** DIA provides a higher proportion of statistically significant proteins than DDA and **(B)** of those significant proteins, more are correctly quantified in DIA than DDA. Mixture compositions for Mix1 at both the species (**C**) and strain (**D**) level are consistent across DDA and the various DIA processing strategies. These trends hold true both when using only proteotypic peptides for quantification, this study's default, and when using parsimony protein assignment logic, a common approach in metaproteomics.

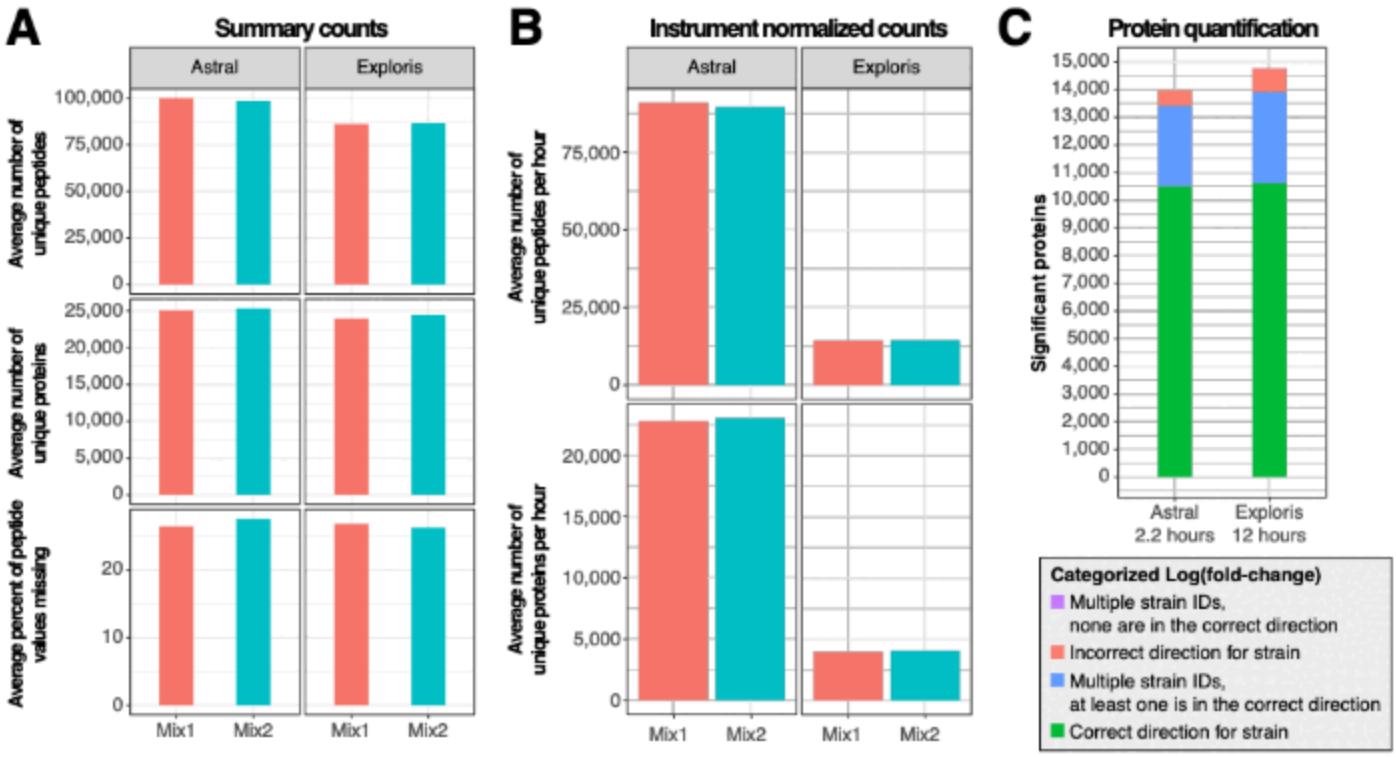


Figure 5

Comparison of the Astral DIA and the Exploris DIA results on Mix1 (red) and Mix2 (blue) from the 46-strain dataset. **(A)** The Astral only detects slightly more peptides and proteins than the Exploris, and **(C)** both instruments correctly quantify a similar number of proteins, but when **(B)** normalized for instrument time (1 hour per mix for the Astral vs. 6 hours per mix for the Exploris), the Astral provides about 9-times more peptide and protein identifications per hour than the Exploris.

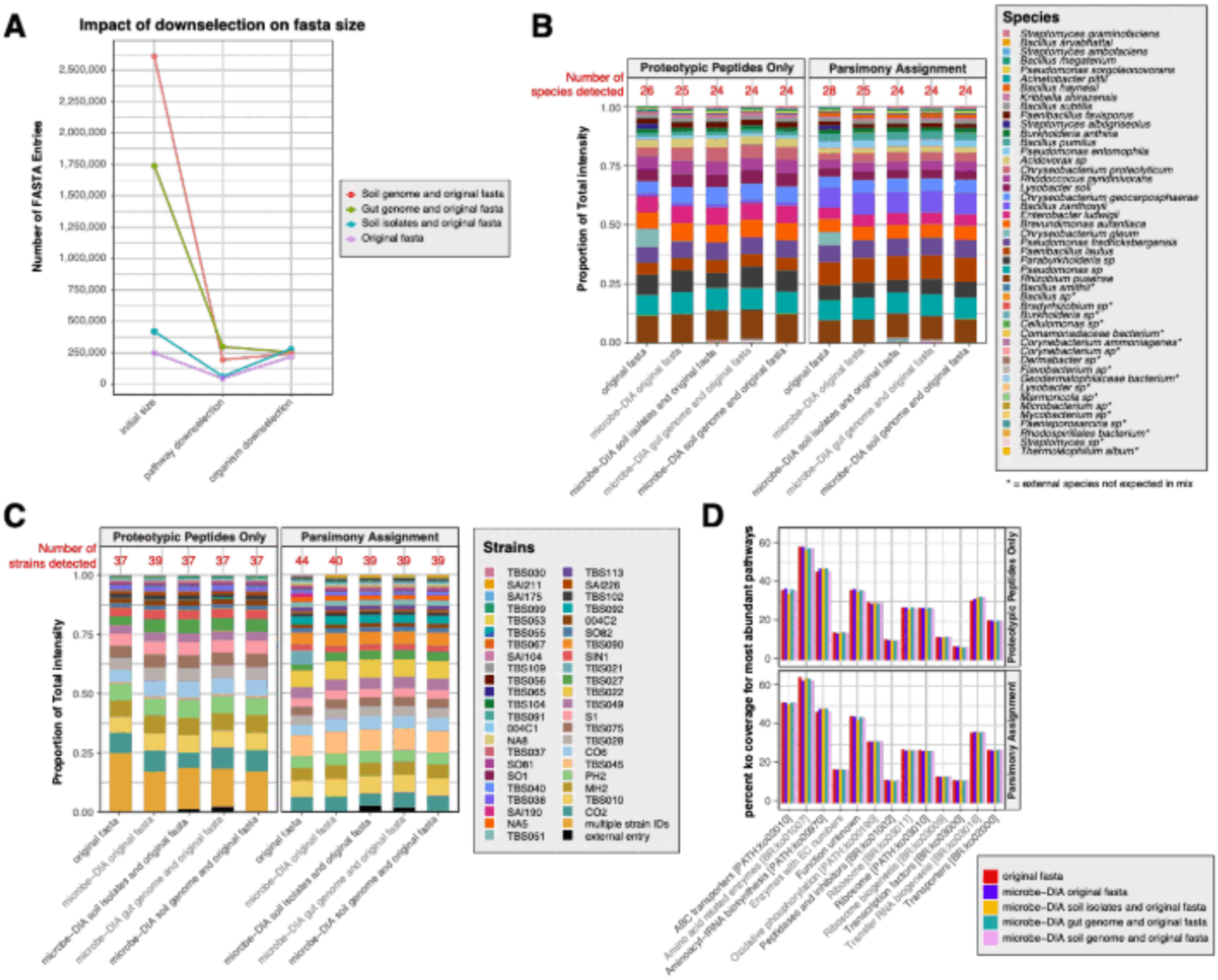


Figure 6

Validation of the two-pass library-free strategy using commonly conserved pathways on the 46-strain DIA Exploris dataset. **(A)** Commonly conserved pathways effectively downselected the test genomes to a manageable search space. After the first and second passes of DIA-NN library-free, the 46-strain dataset's Mix1 was accurately reconstituted at both the species **(B)** and strain levels **(C),** with a maximum species-level false discovery rate of 2.59% of the total abundance. **(D)** The limited false discovery and missed species do not substantially alter the KO coverage for the most abundant pathway groups, indicating that the functional readout is preserved.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DIApaperDraftmicrobiomeSupplementaryInformation.docx