MetaDIA: A DDA-free Database Reduction Strategy for DIA Human Gut Metaproteomics

Haonan Duan (段浩楠)[1,2], Zhibin Ning (宁志斌)[2], Zhongzhi Sun (孙中智)[2], Tiannan Guo (郭天南)[3,4,5], Yingying Sun (孙莹莹)[3,4,5], Daniel Figeys[2,6,7,*]

[1] *Nanjing Women and Children's Healthcare Hospital, Women's Hospital of Nanjing Medical University, Nanjing 210004, China*

[2] *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa ON K1H 8M5, Canada*

[3] *Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310030, China*

[4] *School of Medicine, School of Life Sciences, Westlake University, Hangzhou 310030, China*

[5] *Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China*

[6] *Quadram Institute Bioscience, Norwich Research Park, Norwich NR4 7UQ, UK*

[7] *University of East Anglia, Norwich NR4 7TJ, UK*

[*] Corresponding author.

E-mail: Daniel.Figeys@quadram.ac.uk (Figeys D).

Running title: *Duan H et al / MetaDIA*: *Database Reduction*

The number of words: 9027

The number of references: 44

The number of figures: 8

The number of tables: 0

The number of supplementary files: 3

The number of supplementary figures: 20

The number of supplementary tables: 2

The number of letters in the article title: 68

The number of letters in the running title: 33

The number of keywords: 5

The number of words in abstract: 199

Abstract

Microbiomes, especially within the gut, are complex and may comprise hundreds of species. The identification of peptides in metaproteomics presents a substantial challenge, as it involves matching peptides to mass spectra within an enormous search space for complex and unknown samples. This poses difficulties for both the accuracy and the speed of identification. Specifically, analysis of data-independent acquisition (DIA) datasets has relied on libraries constructed from prior data-dependent acquisition (DDA) results. However, this method is resource-intensive, consumes samples, and limits identification to peptides previously identified. These limitations restrict the application of DIA in metaproteomics research. We introduced a novel strategy to reduce the search space by utilizing species abundance and functional abundance information from the microbiome to score each peptide and prioritize those most likely to be detected. Using this strategy, we have developed and optimized a workflow called MetaDIA for the analysis of microbiome data generated by DIA, which operates independently of DDA assistance. Our approach successfully created a smaller, yet sufficient database for DIA data search in metaproteomics. The results demonstrated strong consistency with the traditional DDA-based library approach at both protein and functional levels. MetaDIA is readily accessible as an open-source project hosted on GitHub (https://github.com/northomics/MetaDIA).

Key words: Metaproteomics; Human gut microbiome; Data independent acquisition; Data-dependent acquisition-free; diaPASEF

Introduction

The microbiome encompasses a diverse array of microorganisms residing in different organisms, ecosystems, and environmental settings such as the human body, animals, plants, soil, water bodies, and various ecological niches [1,2]. Metaproteomics serves as a tool for understanding the roles of proteins within these microbial communities [3]. Mass spectrometry-based proteomics aims to study all proteins in a sample. However, applying these techniques to the microbiome is challenged

by its complexity. Without prior knowledge of the microbes present in a sample, metaproteomics relies on searching mass spectra against a large database, making the task of matching peptides and spectra notably challenging. Employing an iterative search strategy significantly reduces the search complexity in which the final search is against a database generated from previous searching results [4,5]. The iterative strategy has been successfully used for the data acquired by data-dependent acquisition (DDA) mode [6,7]. Unfortunately, in DDA mode, only the most abundant precursor ions are selected for further inquiry, and lower abundant ones are overlooked [8].

In contrast, data-independent acquisition (DIA) uses a set of precursor isolation windows to collect all the fragment ions indiscriminately [9]. It has shown remarkable robustness, sensitivity, and reproducibility with fewer missing values [10]. DIA can be coupled with microLC enabling high-throughput analysis [11]. This makes it particularly suitable for conducting large-scale analyses. The diaPASEF [12] method integrates ion mobility separation with the DIA workflow, adding a fourth dimension of analyzing ion mobility to the traditional three-dimensional data set. This not only enriches the structural information of analytes but also enhances ion utilization efficiency, leveraging the linear relation between ion mobility and mass-to-charge ratio. Another improvement in mass spectrometer scanning speed enables the utilization of smaller isolation windows in DIA, termed as narrow-window DIA [13]. This approach achieves comprehensive peptide precursor coverage and high quantitative precision and accuracy. In bioinformatics, the development of prediction software for peptide properties (theoretically predicted spectrum [14,15], retention times [16–18]) enables the querying of DIA datasets without dependence on libraries generated by DDA. Those predicted libraries even showed better performance than the measured libraries [19]. Moreover, DIA-specific searching software such as DIA-NN [20,21], MaxDIA [22], and Spectronaut have shown reliable results for the identification and quantification of peptides. The above advantages make DIA increasingly popular in proteomics. However, it is noteworthy that the benefits conferred by these techniques have not yet been fully extended to the field of metaproteomics. The main reason is that the inherent complexity of DIA data requires a much more constrained searching space compared with DDA data. To date, only a few studies employed DIA for metaproteomics analysis, and the majority of them were compelled to use a spectral library derived from DDA data [23–25]. The DDA-derived method involves creating a spectral library from DDA runs for each sample, which is then used to interpret complex mass spectra from subsequent analyses. This approach requires multiple sample aliquots, extensive mass

spectrometry resources and is limited to detecting peptides previously identified by DDA. To surmount these limitations, several endeavors have been undertaken. Gladiator [26] uses DIA-Umpire [27] to assemble pseudo-DDA spectra from DIA data for microbiome samples. The method does not require a DDA-based spectral library for its operation; however, it still relies on spectrum-centric algorithms and does not fully exploit the potential advantages of DIA data. Dumas et al. [28] have refined the database by selecting the most abundant species from DDA data analysis. This approach, while reducing the size of database, preserves the entirety of species' proteins, thereby retaining peptides that may be overlooked by a DDA-based library. Similarly, Wu et al. [29] have derived species composition information of samples by analyzing DIA data, laying the foundation for subsequent database construction. This strategy, entirely independent of DDA data, reduces the demand on sample and instrumental resources. Nevertheless, the databases constructed via these two methodologies, encompassing the complete protein sequences of the selected species, remain vast, necessitating substantial computational resources.

Therefore, to leverage the benefits of DIA in metaproteomics, the searching space needs to be further reduced. In the previous iterative strategy [7,29,30], the high-abundant proteins were used for the first search to infer the species that existed in the sample then all the proteins belonging to those species were then used for the subsequent search. However, this database remains overly extensive when compared to the number of identified peptides. Since the abundance of species within the microbiome shows significant disparity [31], the species identified should not be considered equally. The same applies to proteins and peptides. Proteins with high abundance and peptides with high detectability [32] or shared among various species are more likely to be detected. Here we report on a DIA workflow for metaproteomics, called MetaDIA, that relies on an annotated peptide database. This database comprises peptides that are anticipated to be detected, leveraging information on species abundance and protein abundance to score each peptide. We conducted a proof-of-concept experiment on human gut microbiome data generated by diaPASEF mode [25]. The peptide identification number and quantitative results obtained through our peptide library are comparable to those from the DDA-based library. Moreover, the species and functional information obtained from both methods are highly consistent.

Method

Reference peptide sequence with detectability score for human gut microbiome

The Unified Human Gastrointestinal Protein (UHGP v2.0.2) catalog, encompassing 4744 assembled genomes from the human gut microbiome, served as the reference database for this study [33]. Within this catalog, each protein sequence is uniquely associated with a distinct genome and is accompanied by detailed taxonomic and functional annotations. The detectability of peptides derived from these protein sequences was predicted using DeepDetect [32], a deep learning algorithm specifically designed for this purpose. This process involved *in silico* digestion of the protein sequences and subsequent assignment of a detectability score to each resultant peptide. Consequently, the peptide sequence reference database was enhanced by annotating each peptide with three key pieces of information: the genome identifier, the protein identifier, and the detectability score of the peptide. Please note that the database is structured on an identifier-centric organization. This means that peptides with identical sequences may be present within the database; however, as long as they are not from same genome and protein, they are distinguished by unique identifiers.

Generation of FuncTax score

Firstly, the peptides identified by MetaPep [34] are mapped to the UHGP database to establish peptide-genome associations. Subsequently, a greedy algorithm is employed to identify the minimal set of genomes that encompasses all peptide sequences, effectively reducing the complexity of the dataset. Following this, the intensity of each peptide is summed to infer genome abundance. The relative genome abundance will be used as the taxonomic score. To address the assignment of shared peptides, a razor strategy is adopted, analogous to the MaxQuant approach for protein inference [35]. Specifically, when a peptide is found in multiple genomes, it is attributed to the genome with the greater number of associated peptides

For the functional score, we constructed a fixed table from the MetaPep project [34]. While building the database Metapep, the peptide identification was performed by the software MetaLab MAG [7], which provided quantifications of protein abundance. Those proteins are well annotated. Subsequently, the relative abundance of each Clusters of Orthologous Groups (COG) accession was computed. The mean of non-zero relative abundance of the COG accessions was determined across samples, establishing a metric referred to the functional score.

The FuncTax score was obtained by multiplying two scores. In the case of peptides with the same sequence, their FuncTax scores were combined to give higher priority to shared peptides; the highest detectability score among them was utilized to ensure the inclusion of all possible peptides.

Taxonomic and functional analysis

The taxonomic analysis is similar to the generation of genomic abundance score. The identified peptides are mapped to a database to establish peptide-genome associations. The database contains only the top 50 genomes. In our workflow, the peptide database was filtered out from the top 50 genomes. So, all the identified peptides were from the top genomes and thus could be used for the taxonomic analysis (the peptides added from MetaPep might not be used). In DDA-based method, only the peptides mapping to the top 50 genomes were used for taxonomic analysis. Similarly, the razor strategy is used to process peptides shared by multiple genomes. Finally, the intensity of each peptide is summed to infer genome abundance.

For functional analysis, a protein abundance was firstly generated using the same strategy as taxonomic analysis. The proteins in the UHGP database have been extensively annotated thus the protein abundance can be further interpreted into functional abundance.

Deepdetect software configuration

Protein digestion was simulated using Trypsin with the following parameters: a maximum of two missed cleavages, and peptide lengths ranging from 7 to 50 amino acids. Default settings were applied for all other parameters. This step was executed on a system equipped with a single Intel® Xeon® Gold 5118 processor (2.30 GHz, 12 cores), resulting in a total elapsed runtime of 126.5 hours. The peptide database can be generated using Deepdetect by user or downloaded via the link on GitHub.

DIA software configuration

DIA-NN (version 1.8.1) was used to process all the DIA data in this study. Maximum mass accuracy tolerances were set to 10 ppm for both MS1 and MS2 spectra for data generated from tims-TOF. For data generated from Exploris 480, mass accuracy tolerances were set to 0 to let DIA-NN optimise them automatically. False discovery rate (FDR) threshold of 1% was applied to ensure the reliability of identification. The --relaxed-prot-inf option was used for library-free

searching (MetaPep and MetaDIA). The --no-norm option was used to disable the normalization for the quantification benchmark experiment. All other settings were left default. The precursor matrix containing the peptide information was used for taxonomic and functional analysis.

FragPipe (version 23.1) was used for quantification benchmark experiment and samples 1–10 for validation. For samples 1–10, DIA_SpecLib_Quant_diaPASEF workflow was used in which diaTracer was used for spectrum deconvolution. For the mixed samples, DIA_SpecLib_Quant workflow was used, the --no-norm option was used to disable the normalization. FDR threshold of 1% was applied to ensure the reliability of identification. All other settings were left default.

Metaproteomic datasets

The dataset used for optimizing workflow is sourced from a published study and downloaded from ProteomeXchange Consortium (http://www.proteomexchange.org) with Dataset identifier PXD051104 [25]. The dataset for evaluating accuracy is from in-house samples. *Blautia hydrogenotrophica* (DSM 101114; Leibniz Institute DSMZ-German collection of microorganisms and cell cultures) was cultured in LB broth. The human stool was collected from a healthy adult volunteer at the University of Ottawa, Ottawa, ON, CAN. The protocol was approved by Ottawa Health Science Network Research Ethics (Approval No. 20160585-01H). The protein extraction and digestion were performed as described previously [36]. Peptide concentrations were measured using Thermo Scientific Pierce Quantitative Colorimetric Peptide Assays according to the manufacturer's directions.

The in-house samples were then analysed using an UltiMate 3000 RSLCnano system (Thermo Fisher Scientific, USA) coupled to an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific, USA). Peptides were loaded onto an analytical column (75 μm inner diameter × 15 cm) packed with reverse phase beads (3 μm/120 Å ReproSil-Pur C18 resin, Dr. Maisch HPLC GmbH). The liquid chromatography system was directly connected to the analytical column, and no trap column was used in this analysis. A 60 min gradient of 5 to 35% (v/v) from buffer A (0.1% (v/v) formic acid) to B (0.1% (v/v) formic acid with 80% (v/v) acetonitrile) at a flow rate of 300 μL/min was used. The MS1 scan was conducted with a mass resolution of 60,000 (at 200 m/z) and covered a mass range of 380 to 985 m/z. The normalized automatic gain control (AGC) was set to 100%, with a maximum injection time of 100 ms. The MS2 scans were performed at a mass resolution of 15,000 (at 200 m/z), within a mass range of 380 to 980 m/z. Isolation windows of 10 m/z with 1

m/z overlap were used. The normalized AGC for MS2 was set to 200%, with a maximum injection time of 40 ms. The normalized collision energy was maintained at 38%.

Result

MetaDIA Workflow overview: taxonomy- and function-guided construction of peptide database for metaproteomics

Here we propose a new workflow for DIA based metaproteomics called MetaDIA. MetaDIA is a multistep workflow that systematically reduces the search space for DIA searching. At its basis, it relies on a combination of taxonomic abundance, functional abundance as a proxy of protein levels, and peptide detectability ultimately enabling DIA searching without the need for DDA results. Briefly, in the first step, we created a new database of peptides (UHGP peptide database), obtained by *in silico* digestion and detectability prediction of the UHGP database into 773,472,327 peptides [32,33]. Then each peptide is annotated with a FuncTax score (**Figure 1**). Both the FuncTax and the detectability scores are used to reduce the peptide database.

The FuncTax scores for each peptide in the UHGP peptide database are calculated using information from the MetaPep database [34]. MetaPep is a core peptide database compiling peptides previously identified in the published human gut metaproteomics studies. The information from MetaPep was used to create a static table of COG relative functional abundances and a sample-specific table of taxonomic relative abundances (Method). We noted that despite significant differences in the species composition of gut bacteria among different individuals, their functions are remarkably similar [37]. Therefore, functional abundance hierarchy information could act to estimate the likelihood of a protein being observed. We analyzed the search results used to construct the MetaPep database which contained 2134 raw files and 415 individuals [34] (Method). The functional ranking among various samples exhibits a strong correlation (Figure S1A and B). We observed a stable pattern in the functional hierarchy of human gut bacteria: abundant functions consistently remain high, while scarce functions persistently stay low across all samples (Figure S1C; Table S1). The sample-specific table of taxonomic relative abundances was generated by searching the DIA data against MetaPep [34]. The identified peptides and their quantitation were used to create the table (Method). However, this typically results in approximately 1000 genomes remaining, with many containing only a single peptide. The number substantially larger than that is found in a typical human gut microbiome which is around 200 [31]. So, we only choose

the most abundant species for subsequent analysis. The selection of species for consideration is further explored in the optimization section of the study. The FuncTax score for each peptide is calculated by multiplying the taxonomic score for its taxonomic annotation and the functional score of its functional annotation. For peptides with identical sequences, their FuncTax scores were summed thereby leading to higher rankings for shared peptides.

In the last step, sample-specific reduced peptide database is generated by filtering UHGP peptide database using the FuncTax score and the detectability score (Figure 1). The final search of the DIA data is done against the reduced peptide database. To validate the efficiency of our peptide ranking method, peptides were sorted by FuncTax score and partitioned into equal-sized subsets based on their percentile rank (*e.g.*, top 0–5%, 5%–10%, ..., 35%–40%). Each subset was subjected to database searching with uniform parameters. We observed a decline in the number of peptides identified as the percentile ranking of the subsets decreased (**Figure 2**). The decreasing trend suggested our ranking method effectively prioritizes peptides with a higher probability of detection.

Optimized MetaDIA parameters reduce the database size

We explored whether the number of microbes in the reduced peptide database, the threshold for FuncTax score and the threshold for detectability score influenced the identification of peptides. To examine the impact of these parameters, we utilized DIA data from 10 previously reported human gut microbiome samples [25] (File S1, sample information).

In particular, we first tested the effect of the number of microbes (genomes) ranging from 50 to 150 and FuncTax score ranging from top 1% to 40% (**Figure 3**). We keep the detectability threshold at the top 40% in this experiment which is suggested by the author of Deepdetect [32]. Interestingly, no matter how many genomes we chose, the size of the reduced peptide database had the strongest effect on the number of identified peptides. The identification number plateaued once the reduced peptide database size reached around 1.6 million entries (Figure 3A and B, Figures S2 and S3), corresponding to a FuncTax score threshold of 40% for 50 genomes, 20% for 100 genomes, and 15% for 150 genomes, respectively. We compared the three different reduced peptide databases, which led to consistent peptide identification results (Figure 3C and D, Figures S4 and S5). In our previous studies, we observed that low-abundance species were underrepresented [8]. In this context, we chose to focus on the top 50 genomes to prioritize high-abundance genomes. It

is important to note that this cut-off is a variable parameter that can be adjusted according to the specific objectives of different studies.

Subsequently, we explored whether the detectability threshold impacted the number of peptides identified. While the recommended threshold by the author of Deepdetect was 40%, we explored thresholds ranging from 40% to 10%. We observed that a threshold of 25% was the point at which the number of identifications began to decrease significantly (**Figure 4**A). However, both the database size and the search time decreased substantially (Figure 4B). Comparing the identification results at thresholds of 25% and 40%, we found a substantial overlap (Figure 4C). Therefore, we selected a 25% threshold for detectability. Based on this analysis, we proceeded with peptides ranking in the top 40% by FuncTax score and the top 25% by detectability. Given that these two scores are entirely uncorrelated, applying both filters effectively reduces the database to one-tenth of its original size (25% times 40%, Figure S6). After applying these optimized parameters, approximately 1 million peptide sequences remain in the reduced database. Compared to previously reported database reduction strategies, our workflow substantially reduces the size of the database. Wu et al. [29] employed a high-abundance protein-guided method to construct their database, incorporating 256 or 366 complete proteomes. Rough estimates indicate that the database generated by our approach is 50 or 70 times smaller. A smaller database reduces computational resource consumption (Figure S7). Additionally, smaller searching space potentially facilitates peptide identification. Although different searching engines were used, almost double peptides were identified when we apply MetaDIA to the same data (File S2; Table S2).

MetaDIA maintains accuracy in DIA peptide identification

We next explored whether the enrichment of high abundant and highly detectable peptides in our reduced database impacted the accuracy of peptide identification when applying the false discovery rate strategy. To evaluate this, we conducted a benchmark experiment using three samples: a human gut microbiome sample (sample A), a *Blautia hydrogenotrophica* sample (sample C), and a 50:50 mixed sample of the two (sample B) (**Figure 5**A). *Blautia hydrogenotrophica* was selected due to its absence in the microbiome sample used here and its minimal peptide overlap with the microbiome sample. Each sample was subjected to triplicate DIA measurements. Samples A and B were analyzed using the reduced peptide database generated by our workflow with optimized parameters of top 40% FuncTax score and top 25% detectability

score, whereas samples B and C were searched against species-specific protein databases derived from NCBI (Genome assembly ASM15797v1). In the first search against the peptide database, 36,971 unique peptides were identified. Of these, 2072 peptides also present in the *Blautia hydrogenotrophica* database were excluded. Further, peptides unique to each sample were removed, leaving 32,210 peptides identified in both samples A and B. Ideally, the peptide abundance ratio between samples A and B should be approximately 2. In the second search, 16,173 unique peptides were identified. Among these, 12,109 peptides were unique to the *Blautia hydrogenotrophica* database and were found in both samples B and C. The expected ratio between samples B and C should be around 0.5. We found that, whether we used a protein database or a peptide database, the ratios of peptides identified in both searches closely aligned with the expected values (Figure 5B). To ensure that the performance is not specific to DIA-NN, we independently re-analyzed the three samples using FragPipe [38,39], a widely adopted alternative to DIA-NN with distinct algorithmic foundations. The results were consistent with those achieved using DIA-NN (Figure S8). This suggests that the employment of our reduced peptide database does not significantly affect the accuracy of peptide identification, thereby supporting its use in peptide identification workflows with a controlled FDR.

MetaDIA yields consistent peptide and protein identification results with DDA-based strategies.

We next evaluated whether MetaDIA performed similarly to a conventional DDA-based library for DIA data analysis. The DIA data and corresponding DDA-based library were obtained from a published study [25]. We found that the MetaDIA provided identification numbers comparable to those obtained through the DDA-based library (**Figure 6**A). Notably, in certain instances, such as with samples 8 and 9, the MetaDIA surpassed DDA library in the number of identifications. The initial step in our workflow involves searching the raw data against MetaPep. MetaPep was constructed by collecting searching result from an open search algorithm, Metalab-MAG [7]. Thus, it encompasses modified peptides not included in the original database. Subsequent integration of peptides identified by MetaPep into a refined peptide database resulted in a marked increase in identification rates (Figure 6A, Figure S9; File S3). Because of the augmentation in the number of peptide identifications attributable to MetaPep, we incorporate the identification results from MetaPep into the database in the following study.

Over 50% of peptides identified from the DDA-based library were also identified by MetaDIA+MetaPep (Figure 6B, Figure S10). The divergence in unique identifications between the two methods may be attributed to a variety of factors, including the size and composition of the data, as well as the inherent differences between DDA acquisition and DIA acquisition. For instance, we have found that approximately half of the unique peptides from the DDA-based library are indeed present within the database produced by MetaDIA+MetaPep; however, they remain unidentified in our workflow (Figure S11). Moreover, peptides provided by MetaDIA originate exclusively from a pre-selected group of 50 species, whereas the results from the DDA library encompass peptides from thousands of species. This, to some extent, elucidates the small overlap observed between the two methods. Upon examining the quantification results of those peptides found by both methods, we observed a significant consistency in the outcomes, with a Pearson coefficient above 0.9 (Figure 6D, Figure S12). It is worth noting that the fragment ions used for quantification in the DDA-based library correspond to actual DDA acquisitions. In contrast, our workflow leverages the library-free mode within DIA-NN which uses theoretical spectra that are predicted from peptide sequences. The high degree of agreement between the quantification results underscores the reliability of the MS-Simulator algorithm which is employed by DIA-NN for spectra prediction [14]. We further validated the MetaDIA using the FragPipe. In general, FragPipe identified less peptides (Figure S13A). This may potentially be attributed to its dependence on MS1 signals for assembling associated fragment ion intensities into pseudo MS2 spectra, whereas DIA-NN does not require such an approach. By using FragPipe, MetaDIA demonstrated a distinct advantage over DDA-based method in terms of the number of identifications. Similar number of common peptides (those detected by both the DDA-based library and MetaDIA) were identified by FragPipe and DIA-NN (Figure S13B). The DDA-based library and MetaDIA demonstrated greater consistency when using the FragPipe on peptide quantification (Figure S13C and D).

At the protein level, our findings revealed greater consistency in identification compared to the peptide level (Figure 6C). Around 70% of proteins found by the DDA-based library can be found by MetaDIA+MetaPep. The overlap on protein level reinforces the reliability of the identifications and indicates that a significant subset of proteins is consistently identified by both methods despite differences at the peptide level (Figure S14). Proteins like GYG000002545_00035 had greater sequence coverage and higher detection intensity with the DDA library, while others like

MGYG000002272_00452 showed higher coverage and intensity with MetaDIA+MetaPep. Given that the quantification of a protein was derived from different subsets of peptides in these two methods, we observed reduced consistency of quantification in the protein level between the methods, as reflected by Pearson correlation coefficients of approximately 0.7 (Figure 6E, Figure S15). However, it is important to note that in most proteomic studies, the primary interest lies in the differential abundance of the same protein across various samples. Therefore, it is crucial that we use the same fragment ions to quantify a protein. In this regard, the inconsistencies in protein quantification between the two methods do not undermine the utility of either approach. The substantial overlap in peptide and protein identification by both methods suggests a robust cross-validation of both methods. Then we annotated the proteins using COG accessions and calculated their relative abundances. Our analysis revealed that approximately 90% of the COG accessions identified by the DDA-based method were also covered by our method (Figure 6C). Furthermore, the Pearson correlation coefficient for the relative abundance of COG accessions exceeded 0.9, with a stronger correlation for those COG accessions that were highly abundant (Figure 6F, Figure S16).

MetaDIA provides taxonomic profiles highly similar to those obtained from searching DDA-libraries.

We verified whether both methods had a high degree of similarity in the taxonomic composition. We did comparative analysis of microbiome composition across different taxonomic levels using the results from both methods. In our workflow, taxonomic analysis was confined to the predefined database containing 50 genomes. In the DDA-based method, the peptide identified by the DDA library can be annotated to over 1000 genomes even after using the greedy algorithm (Method: generation of FuncTax score). However, we found that the top 50 genomes accounted for 79%–90% of peptides and 87%–92% of peptide intensity (Figure S17). To simplify the comparison between the two methods, we discarded the small number of peptides that could not be annotated to the top 50 genomes. Our findings indicate that there is a significant linear correlation between the compositions identified by both methods, with the degree of correlation strengthening at higher taxonomic levels (**Figure 7**A and B, Figure S18). The two methods showed remarkably consistent taxonomic composition at the genus level with a Pearson coefficient above 0.98 across all the samples tested. Even sample 9, which displayed the lowest correlation, demonstrated a substantial

degree of consistency between the two methods. To underscore the consistency, we have provided a detailed visualization of the taxonomic composition for sample 9 (Figure 7C and D, Figure S19) The species compositions observed by our method in these ten samples differ significantly as expected, indicating that our database and taxonomic analysis have the capability to identify a diverse range of microbiota (Figure S20; File S2). The most abundant species identified in the ten samples have been previously reported as high-abundance species in the human gut microbiome [40–44]. Except for *Phocaeicola dorei* which was identified as the top species in samples 2, 5, and 10, the other top species were all unique to each sample.

Performance validation of MetaDIA on large-scale diaPASEF datasets.

To further validate the versatility and applicability of our proposed metaproteomic workflow, we extended our analysis to a diverse set of 79 DIA datasets obtained from a published study [25]. This dataset encompasses samples from 62 individuals, featuring replicate injections, quality control samples, and pooled samples (File S1). We applied MetaDIA to this extensive dataset and compared the results with the conventional DDA-based approach. Remarkably, the number of peptides identified by both methods demonstrated a close equivalence, reinforcing the robustness and universal applicability of our metaproteomic workflow (**Figure 8**). Validating our method across diverse samples enhances confidence in its effectiveness and consistency, demonstrating its potential for widespread adoption in metaproteomics research.

Discussion

We propose a novel workflow for DIA data analysis from human gut microbiome called MetaDIA. The approach aims at prioritizing peptides with a higher likelihood of detection based on their detectability, as well as taxonomic and functional scores.

MetaDIA is entirely devoid of DDA, thereby circumventing the drawbacks of DDA-based methods. Not only does this approach save time and resources, but it also enables the creation of a tailored database for each sample. In contrast, DDA-based methods typically rely on a single pooled sample to generate a library. For instance, Gomez et al. [24] used a pooled sample to represent 12 individual mice, while Sun et al. [25] did so for a cohort of 62 individuals. However, such a pooled sample may not effectively represent every sample. In our study, the ten samples showed highly diverse taxonomic composition (Figure S20). To increase the sampling depth for

the pooled sample, Sun et al. [25] had to fractionate the pooled sample into 30 portions and Gomez et al. [24] repeatedly injected the pooled sample 10 times. Moreover, utilizing a static library to search various samples may potentially compromise the accuracy of peptide identification, as it includes peptides from the pooled samples that are absent in the specific sample under investigation.

In MetaDIA, we pre-defined the range of genomes for each microbiome sample (50 genomes in this study). This approach not only enabled us to narrow the search space but also to mitigate the issues associated with protein inference that arise from common peptides. When assigning peptides to proteins, we confined our consideration to the genomes within the predefined range rather than the entire dataset. This strategy significantly reduced the incidence of common peptides.

In the process of calculating the taxonomic composition of a sample, we employed a greedy algorithm to minimize the number of genomes required to account for all identified peptides. Practically, this was achieved through an iterative method, sequentially selecting the genome that encompasses the greatest number of peptides, until every peptide could be explained by the selected genomes. By using this, the matches between peptides and genomes are significantly simplified. However, this approach is prone to yielding locally optimal solutions that may not align with the global optimum, primarily due to the shared nature of peptides among different genomes. For example, a genome containing greatest number of identified peptides may not present in the sample, potentially compromising the precision of taxonomic analysis at the genomic level. Fortunately, genomes possessing a multitude of shared peptides typically align at higher taxonomic levels (*e.g.*, Genus, Family). Hence, the precision of taxonomic analysis is anticipated to be improved on higher taxonomic levels.

MetaDIA relies on MetaPep for the first search to acquire the taxonomic composition of the sample. Although MetaPep showed good performance in this study, it may exhibit bias due to the underrepresentation of peptides from rare species or functions, instrumentation-specific detection limits, and variations in sample processing methods, which compromise its comprehensiveness and reliability. Because MetaPep is for human gut microbiome, the current workflow can only be applied to the human gut microbiome. However, we foresee that its strategy could be extended to other types of microbiomes. In the bottom-up approach to metaproteomics, the abundance of a specific peptide is ascertained by its corresponding protein and species abundance, both of which are generally diminished during the processes of protein digestion and cellular lysis. Within this

strategy, efforts are made to digitally reconstruct these two crucial pieces of information using functional score and taxonomic score and employ them to prioritize the peptides with the highest probability of detection. In the future, a selected database containing proteins common to all taxonomy could be used to help us infer the taxonomic composition for other types of microbiome [28,29], and the functional score could be acquired by analyzing previously published data. Importantly, owing to the comprehensive nature of DIA data, researchers may also request the inclusion of specific functions or pathways of interest into the database.

Conclusion

In conclusion, we introduced a new strategy to prioritize peptides with a high probability of detection. This strategy simulates protein digestion procedures *in silico* and uses taxonomic and functional information to infer the peptide abundance. MetaDIA is a fully DDA-free workflow and provides a user interface to change the different parameters. We compared the performance of MetaDIA with the DDA-based library and observed a high degree of consistency. We further validated our method across a DIA-PASEF dataset with 79 samples, thereby confirming its wide applicability. We believe that our approach will help the application of DIA in metaproteomics.

Code availability

MetaDIA is publicly available at GitHub (https://github.com/northomics/MetaDIA). The code has also been submitted to BioCode at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB) (BioCode: BT007936), which is publicly accessible at https://ngdc.cncb.ac.cn/biocode/tool/BT007936.

Data availability

The datasets generated in this study were sourced from ProteomeXchange Consortium (http://www.proteomexchange.org) with dataset identifier PXD063632.

CRediT author statement

Haonan Duan: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft p reparation. **Zhibin Ning:** Conceptualization, Methodology, Writing – review & editing. **Zhongzhi Sun:** Writing – review & editing. **Tiannan Guo:** Data Curation, Writing – review &

editing. **Yingying Sun:** Data Curation, Writing – review & editing, **Daniel Figeys:** Supervision, Writing – review & editing, Funding acquisition. All authors read and approved the final manuscript.

Competing interests

Dainel Figeys is the founder of MedBiome Inc. a microbiome nutrition and therapeutic company. The other authors have declared that they have no competing interests.

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, Grant No. 2018-03905) discovery grant to Daniel Figeys. Haonan Duan was funded by a stipend from the NSERC CREATE in Technologies for Microbiome Science and Engineering (TECHNOMISE) Program (Grant No. 497995) and Jiangsu Funding Program for Excellent Postdoctoral Talent.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/xxxxxxx).

ORCID

0000-0002-2753-8594 (Haonan Duan)
0000-0003-2045-7596 (Zhibin Ning)
0000-0002-4862-8599 (Zhongzhi Sun)
0009-0008-6015-2109 (Yingying Sun)
0000-0003-3869-7651 (Tiannan Guo)
0000-0002-5373-7546 (Daniel Figeys)

Reference

[1] Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207–14.

[2] Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 2017;551:457–63.
[3] Wilmes P, Bond PL. Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 2006;14:92–7.
[4] Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, Seymour SL, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics 2013;13:1352–7.
[5] Zhang X, Ning Z, Mayne J, Moore JI, Li J, Butcher J, et al. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome 2016;4:31.
[6] Cheng K, Ning Z, Zhang X, Li L, Liao B, Mayne J, et al. MetaLab 2.0 enables accurate post-translational modifications profiling in metaproteomics. J Am Soc Mass Spectrom 2020;31:1473–82.
[7] Cheng K, Ning Z, Li L, Zhang X, Serrana JM, Mayne J, et al. MetaLab-MAG: a metaproteomic data analysis platform for genome-level characterization of microbiomes from the metagenome-assembled genomes database. J Proteome Res 2023;22:387–98.
[8] Duan H, Cheng K, Ning Z, Li L, Mayne J, Sun Z, et al. Assessing the dark field of metaproteome. Anal Chem 2022;94:15648–54.
[9] Gillet LC, Navarro P, Tate S, Rost 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. Mol Cell Proteomics 2012;11:O111.016717.
[10] Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nat Commun 2017;8:291.
[11] Vowinckel J, Zelezniak A, Bruderer R, Mulleder M, Reiter L, Ralser M. Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. Sci Rep 2018;8:4346.
[12] Meier F, Brunner AD, Frank M, Ha A, Bludau I, Voytik E, et al. diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. Nat Methods 2020;17:1229–36.

[13] Guzman UH, Martinez-Val A, Ye Z, Damoc E, Arrey TN, Pashkova A, et al. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. Nat Biotechnol 2024;42:1855–66.
[14] Sun S, Yang F, Yang Q, Zhang H, Wang Y, Bu D, et al. MS-Simulator: predicting y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions. J Proteome Res 2012;11:4509–16.
[15] Zeng WF, Zhou XX, Zhou WJ, Chi H, Zhan J, He SM. MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer tearning. Anal Chem 2019;91:9724–31.
[16] Bouwmeester R, Gabriels R, Hulstaert N, Martens L, Degroeve S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods 2021;18:1363–9.
[17] Ma C, Ren Y, Yang J, Ren Z, Yang H, Liu S. Improved peptide retention time prediction in liquid chromatography through deep learning. Anal Chem 2018;90:10881–8.
[18] Al Musaimi O, Valenzo OMM, Williams DR. Prediction of peptides retention behavior in reversed-phase liquid chromatography based on their hydrophobicity. J Sep Sci 2023;46:e2200743.
[19] Cox J. Prediction of peptide mass spectral libraries with machine learning. Nat Biotechnol 2023;41:33–43.
[20] 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;17:41–4.
[21] Demichev V, Szyrwiel L, Yu F, Teo GC, Rosenberger G, Niewienda A, et al. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. Nat Commun 2022;13:3944.
[22] 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;39:1563–73.
[23] Aakko J, Pietila S, Suomi T, Mahmoudian M, Toivonen R, Kouvonen P, et al. Data-independent acquisition mass spectrometry in metaproteomics of gut microbiota-implementation and computational analysis. J Proteome Res 2020;19:432–6.

[24] Gomez-Varela D, Xian F, Grundtner S, Sondermann JR, Carta G, Schmidt M. Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. Front Microbiol 2023;14:1258703.
[25] Sun Y, Xing Z, Liang S, Miao Z, Zhuo LB, Jiang W, et al. metaExpertPro: a computational workflow for metaproteomics spectral library construction and data-independent acquisition mass spectrometry data analysis. Mol Cell Proteomics 2024;23:100840.
[26] Pietilä S, Suomi T, Elo LL. Introducing untargeted data-independent acquisition for metaproteomics of complex microbial samples. ISME Commun 2022;2:51.
[27] 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;12:258–64.
[28] 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;12:46.
[29] Wu E, Yang Y, Zhao J, Zheng J, Wang X, Shen C, et al. High-abundance protein-guided hybrid spectral library for data-independent acquisition metaproteomics. Anal Chem 2024;96:1029–37.
[30] 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;9:80.
[31] Yang J, Pu J, Lu S, Bai X, Wu Y, Jin D, et al. Species-level analysis of human gut microbiota with metataxonomics. Front Microbiol 2020;11:2029.
[32] Yang J, Cheng Z, Gong F, Fu Y. DeepDetect: deep learning of peptide detectability enhanced by peptide digestibility and its application to DIA library reduction. Anal Chem 2023;95:6235–43.
[33] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 2021;39:105–14.
[34] Sun Z, Ning Z, Cheng K, Duan H, Wu Q, Mayne J, et al. MetaPep: A core peptide database for faster human gut metaproteomics database searches. Comput Struct Biotechnol J 2023;21:4228–37.
[35] Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 2016;11:2301–19.

[36] Zhang X, Li L, Mayne J, Ning Z, Stintzi A, Figeys D. Assessing the impact of protein extraction methods for human gut metaproteomics. J Proteomics 2018;180:120–7.
[37] Li L, Wang T, Ning Z, Zhang X, Butcher J, Serrana JM, et al. Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. Nat Commun 2023;14:3428.
[38] Yu F, Teo GC, Kong AT, Frohlich K, Li GX, Demichev V, et al. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat Commun 2023;14:4154.
[39] Li K, Teo GC, Yang KL, Yu F, Nesvizhskii AI. diaTracer enables spectrum-centric analysis of diaPASEF proteomics data. Nat Commun 2025;16:95.
[40] Davis-Richardson AG, Ardissone AN, Dias R, Simell V, Leonard MT, Kemppainen KM, et al. Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. Front Microbiol 2014;5:678.
[41] Hosomi K, Saito M, Park J, Murakami H, Shibata N, Ando M, et al. Oral administration of Blautia wexlerae ameliorates obesity and type 2 diabetes via metabolic remodeling of the gut microbiota. Nat Commun 2022;13:4477.
[42] Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. Elife 2013;2:e01202.
[43] Lan PTN, Sakamoto M, Sakata S, Benno Y. Bacteroides barnesiae sp. nov., Bacteroides salanitronis sp. nov. and Bacteroides gallinarum sp. nov., isolated from chicken caecum. Int J Syst Evol Microbiol 2006;56:2853–9.
[44] Ferreira-Halder CV, Faria AVS, Andrade SS. Action and function of Faecalibacterium prausnitzii in health and disease. Best Pract Res Clin Gastroenterol 2017;31:643–8.

Figure legend

Figure 1 The flowchart for the MetaDIA

All proteins in Unified Human Gastrointestinal Protein (UHGP) database were firstly *in silico* digested into peptides. The detectability of each peptide was predicated by DeepDetect algorithm. Following this prediction, each peptide was assigned a functional score and a taxonomic score, derived from a predetermined functional relative abundance table and a sample-specific taxonomic relative abundance table, respectively (Method). The FuncTax score was calculated by multiplying

the two scores. For peptides with identical sequence, their FuncTax scores were summed to prioritize shared peptides; the maximum of their detectability scores was used to ensure the inclusion of all potential peptides. The detectability and FuncTac scores are both used for filtering peptides. The reduced peptide database was used for a second search.

Figure 2 The number of peptides identified from each subset

Ten samples were tested in the experiment. For constructing the peptide database, the top 100 genomes were considered; the detectability threshold was set at 40%. Each subset contains around 400,000 peptides. Peptide identification was performed by DIA-NN under same conditions. The maximum identification in the ten samples from the last subset was heighted in the figure.

Figure 3 Optimization for genome number and FuncTax score (Sample 1)

Peptides from n (50, 100, and 150) genomes were ranked by the FuncTax score and top x% (1–40 for 50 and 100 genomes; 1–35 for 150 genomes) peptides was used as database. **A.** Number of identified peptides against database percentage. **B.** Number of identified peptides against database size. The inflection point has been highlighted with a red box. The overlap of the reduced peptide database (**C**) and identified peptide (**D**) when taking top 40% peptides for 50 genomes, top 20% for 100 genomes, and top 15% for 150 genomes as database. Peptide identification was performed by DIA-NN under same conditions.

Figure 4 Optimization for detectability threshold

The number of peptides identified (**A**) and the searching time (**B**) under detectability threshold from 10% to 40%. Only the searching time is included, time consumed by DeepDetect is not accounted for. **C**. The overlap of peptides identified by top 25% and top 40% of the database. Peptide identification was performed by DIA-NN under same conditions.

Figure 5 Benchmark experiment for peptide identification

A. The experimental design. Each sample was subjected to triple-run measurements. **B.** Log-transformed ratios are plotted as a function of peptide intensity for microbiome peptides (n = 32,210, in green) and *Blautia hydrogenotrophica* peptides (n = 12,109, in purple). The point density for ratio was plotted at right. Dashed lines indicate the expected ratio. Peptide identification

was performed by DIA-NN under same conditions. The intensities for x axis were derived from pure samples (Samples A and C).

Figure 6 Comparison between the DDA-based method and DDA-free method

A. The peptide identified by each method. **B.** The overlap of peptide identified in sample 1 by each method. **C.** Coverage of peptides, proteins and cog accessions identified by DDA-based method with those found using DDA-free method. The intensity correlation of the overlapped peptides (**D**), proteins (**E**), and COG accessions (**F**) in sample 1. The dashed line indicates y = x. For DDA-based method, the peptides identified as derived from human proteins are removed.

Figure 7 Comparison of the taxonomic composition between the DDA-based method and DDA-free method

Pearson correlation (**A**) and Bray-Curtis distance (**B**) analysis between DDA-based method and DDA-free method on different taxonomic levels from Phylum to Species. The sunburst chart showing taxonomic composition (Phylum to Family) of sample 9 derived from DDA-based method (**C**) and DDA-free method (**D**). The relative taxonomic abundance was used for the analysis. In the correlation analysis, taxonomic categories that were unique to one method were imputed with a value of zero. In sunburst chart, some species names are omitted due to spatial constraints. Each color corresponds to a distinct species, and consistent coloring is applied for the same species across panels.

Figure 8 Number of peptides identified by both methods

Mean values from 79 diaPASEF samples were shown. For DDA-based method, the peptides identified as derived from human proteins are removed. The specific values can be found in File S3.

Figure 1

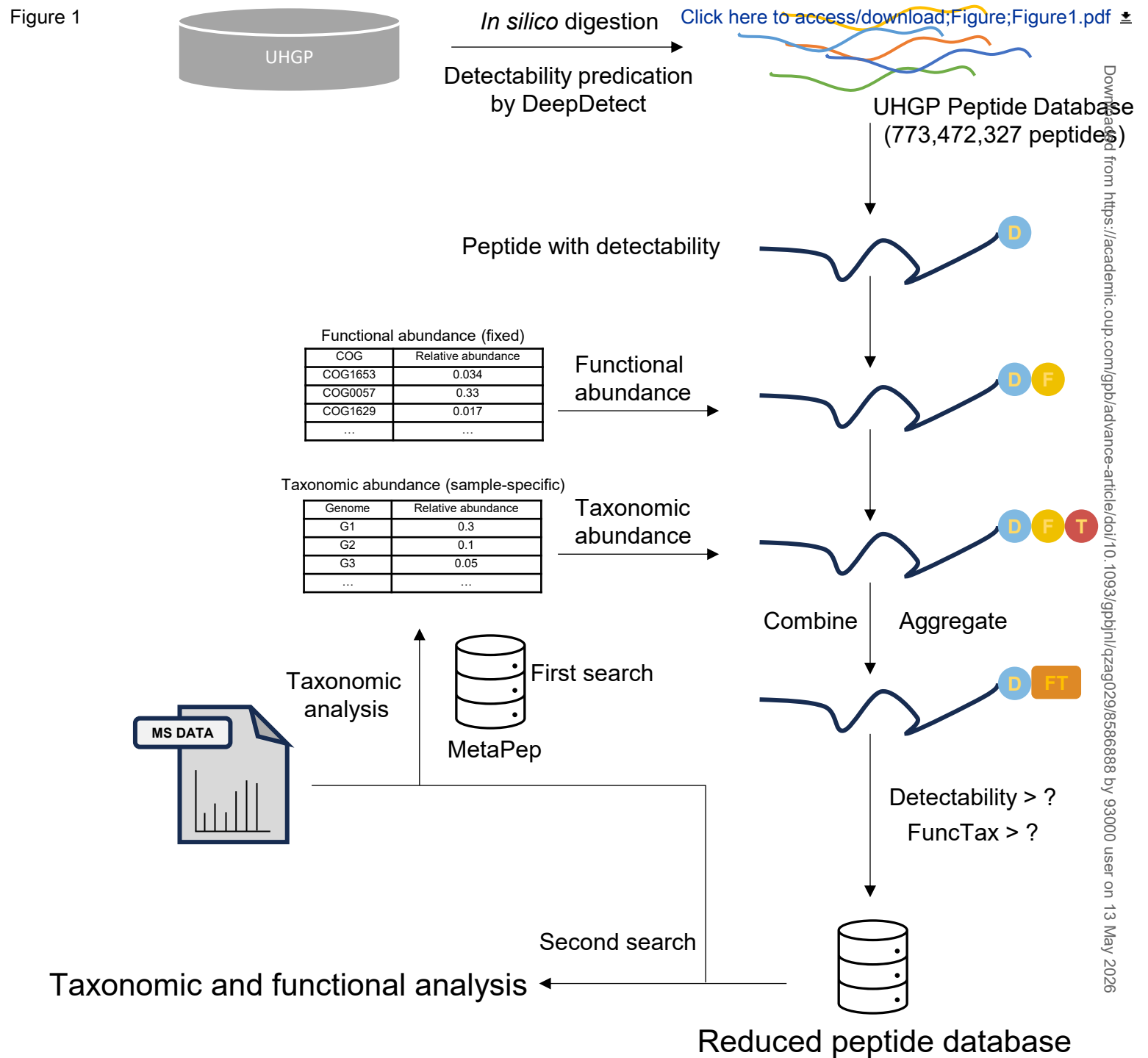


Figure 2

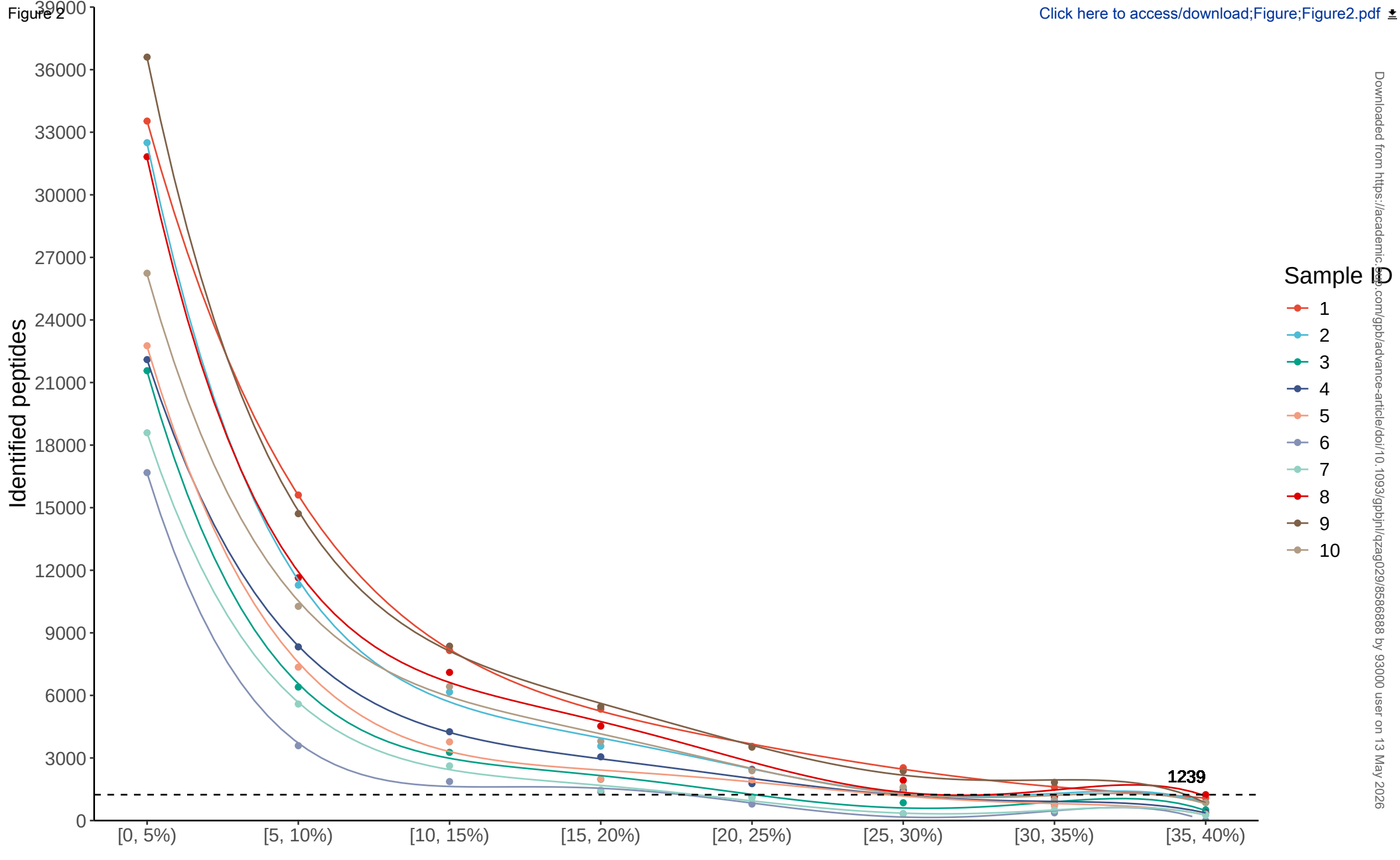
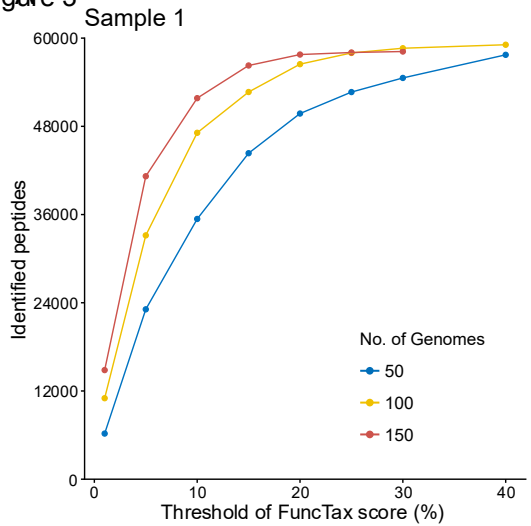
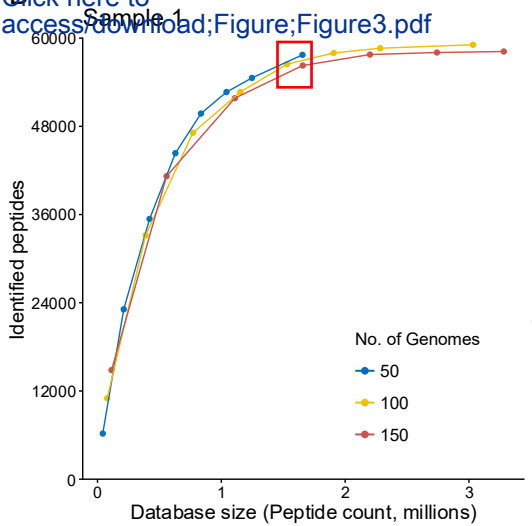
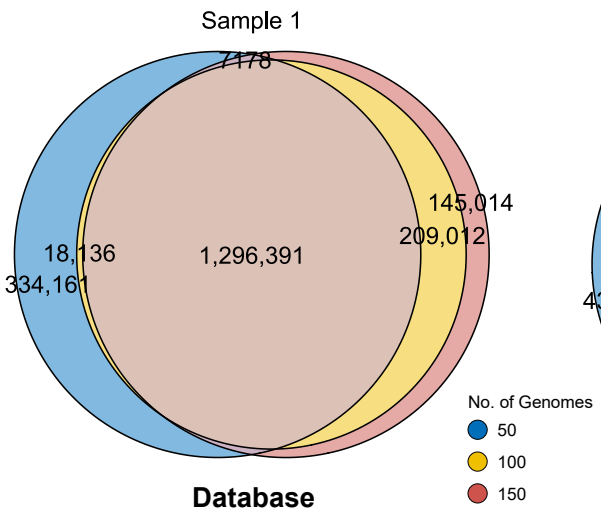


Figure 3







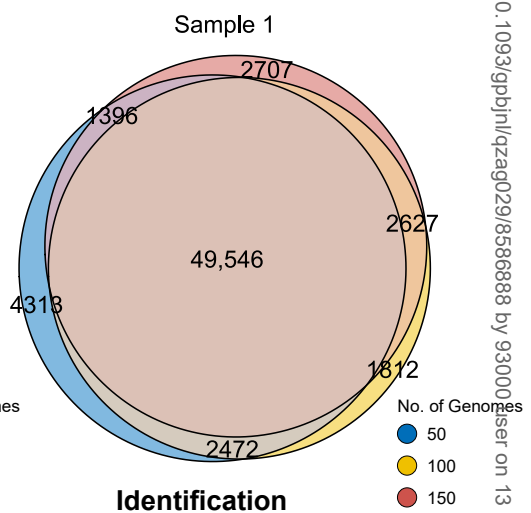
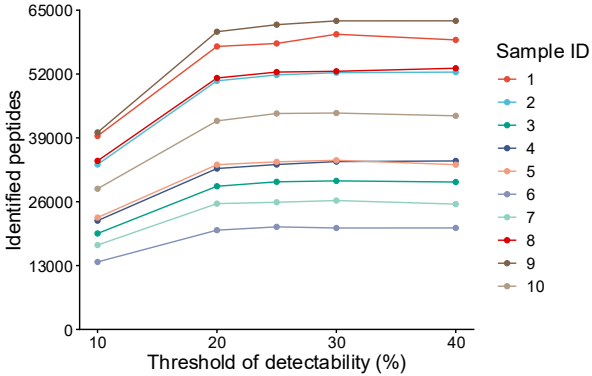
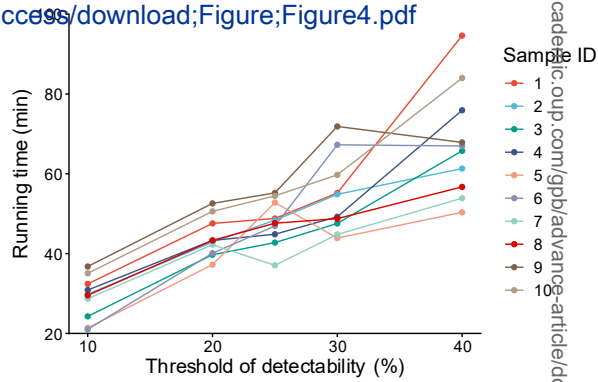


Figure 4





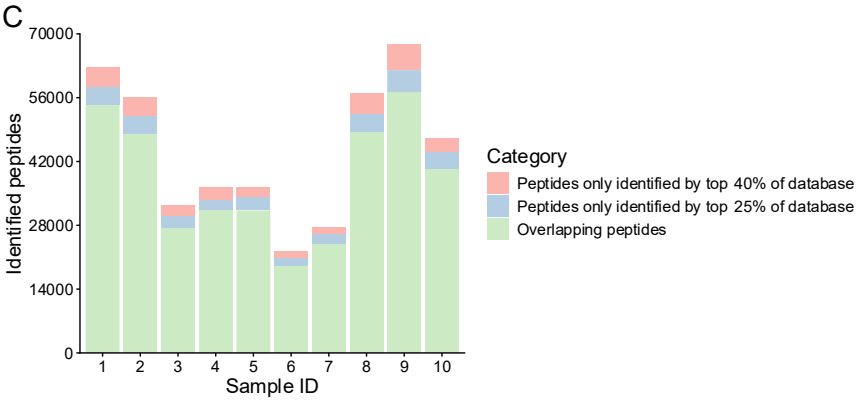


Figure 5

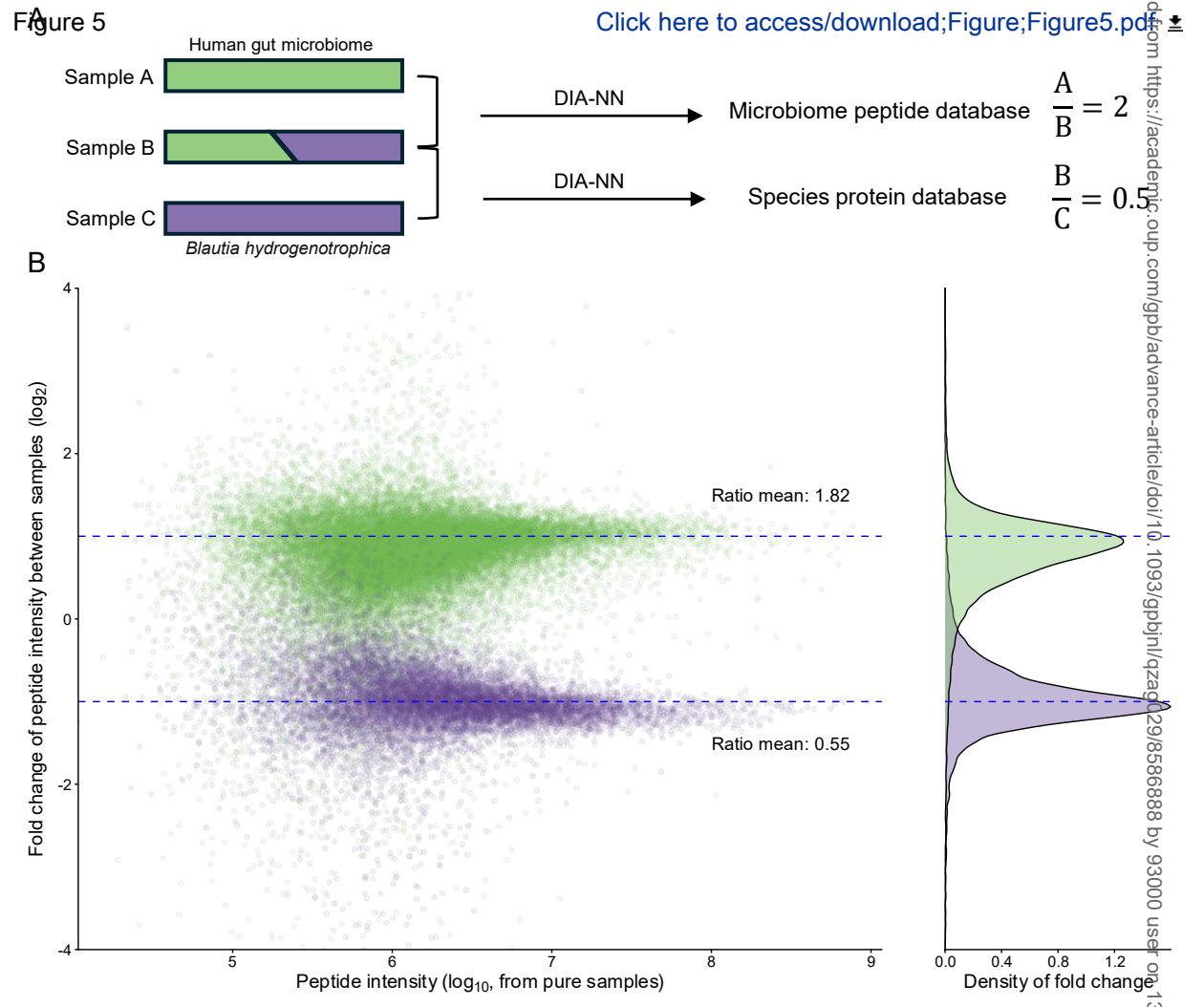


Figure 6

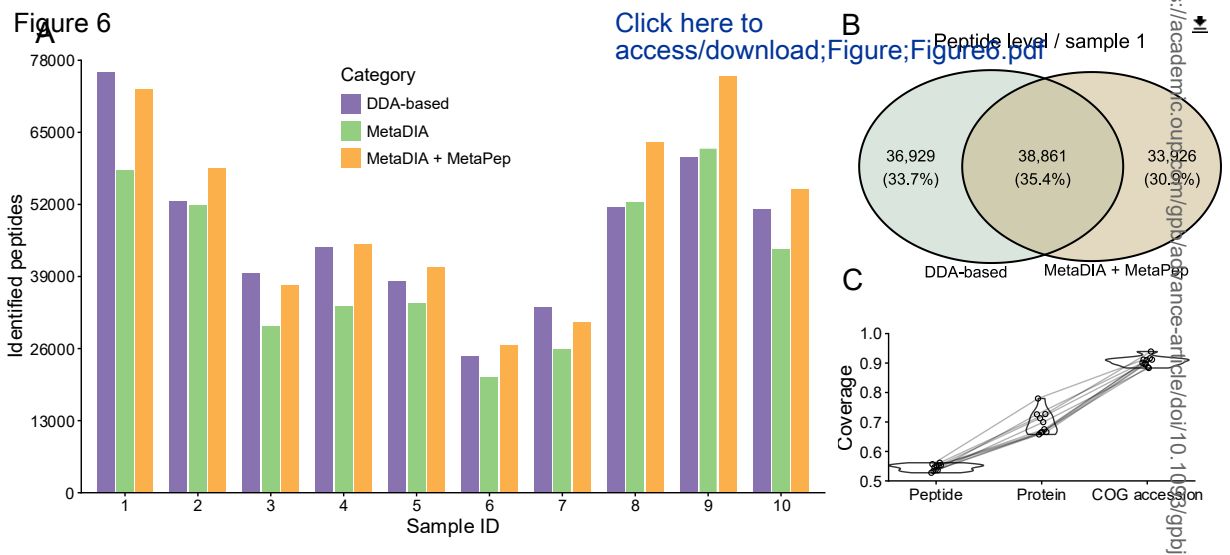


Figure 7

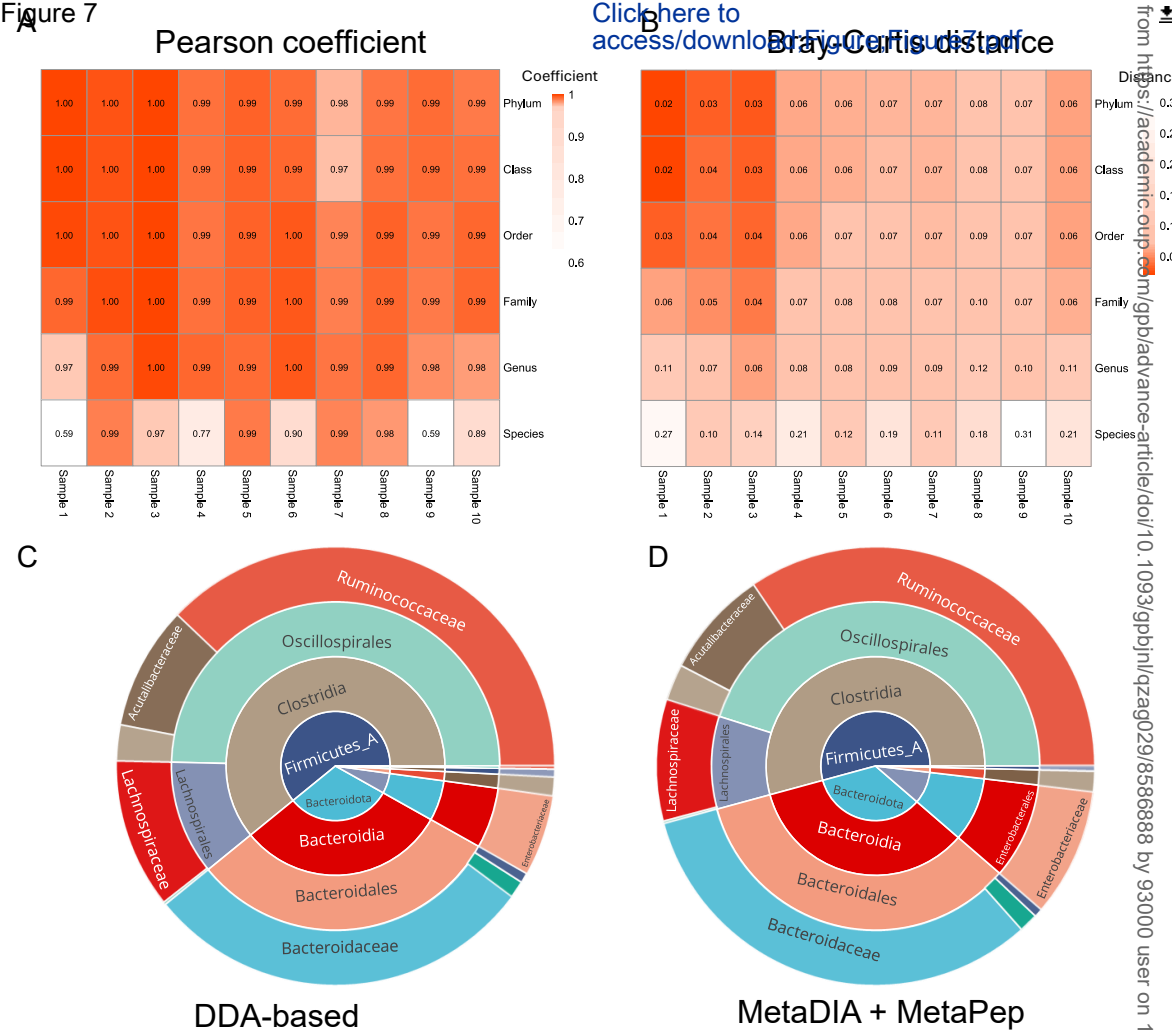
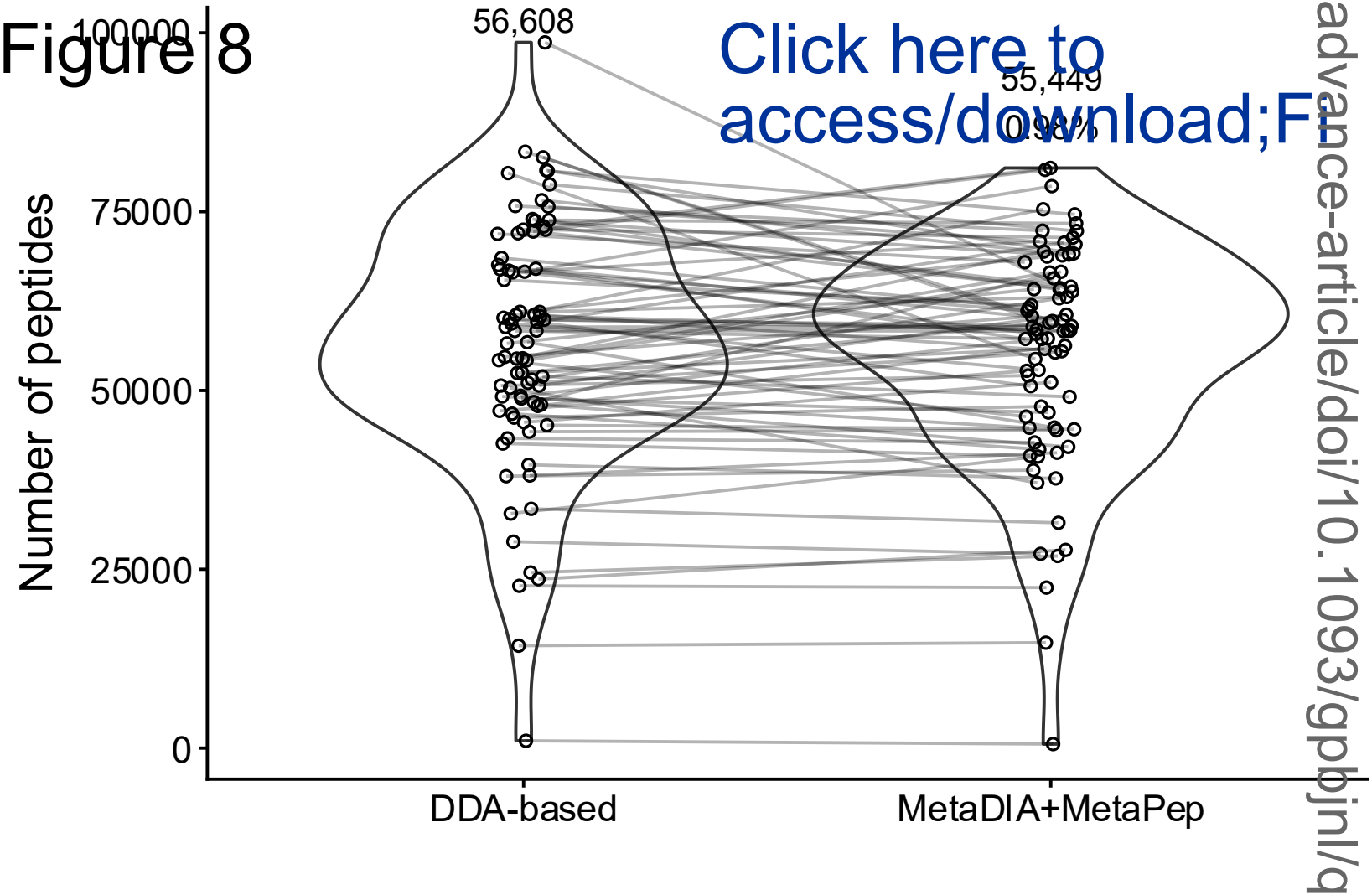


Figure 8



Graphical Abstract