Journal of proteome research

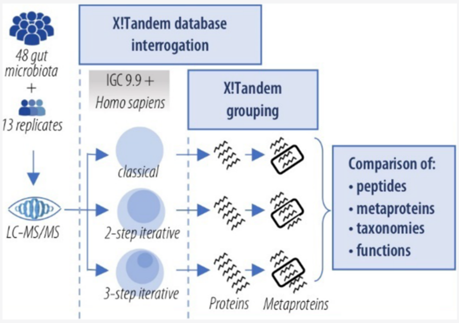
Benefits of Iterative Searches of Large Databases to Interpret Large Human Gut Metaproteomic Data Sets

Ariane Bassignani, Sandra Plancade,* Magali Berland, Melisande Blein-Nicolas, Alain Guillot, Didier Chevret, Chloé Moritz, Sylvie Huet, Salwa Rizkalla, Karine Clément, Joël Doré, Olivier Langella,* and Catherine Juste*

Cite This: https://dx.doi.org/10.1021/acs.jproteome.0c00669 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: The gut microbiota are increasingly considered as a main partner of human health. Metaproteomics enables us to move from the functional potential revealed by metagenomics to the functions actually operating in the microbiome. However, metaproteome deciphering remains challenging. In particular, confident interpretation of a myriad of MS/MS spectra can only be pursued with smart database searches. Here, we compare the interpretation of MS/MS data sets from 48 individual human gut microbiomes using three interrogation strategies of the dedicated Integrated nonredundant Gene Catalog (IGC 9.9 million genes from 1267 individual fecal samples) together with the *Homo sapiens* database: the classical single-step interrogation strategy and two iterative strategies (in either two or three steps) aimed at preselecting a reduced-sized, more targeted search space for the final peptide spectrum matching. Both iterative searches outperformed the single-step classical search in terms of the number of peptides and protein clusters identified and the depth of taxonomic and functional knowledge, and this was the most convincing with the three-step approach. However, iterative searches do not help in reducing variability of repeated analyses, which is inherent to the traditional data-dependent acquisition mode, but this variability did not affect the hierarchical relationship between replicates and all other samples.



KEYWORDS: *metaproteomics, gut microbiome, database search, X!Tandem*

■ INTRODUCTION

High-throughput sequencing of total fecal DNA of the gut microbiome provides an enormous reservoir for the discovery of unsuspected metagenomic signatures that can represent predictive biomarkers and new therapeutic targets for different human disease phenotypes or stages.[1,2] The accompanying challenge is a holistic metaproteomic approach to move beyond the genetic potential addressed by metagenomics and become closer to the real functionality of the gut microbiome by exploring the expression of metabolic and cellular pathways. Understanding how these pathways are altered in diseases can have a profound impact on patient diagnosis and treatment and eventually disease prevention. This may indeed open new avenues for reducing risks, including the discovery of new biomarkers, new targets, and new therapeutic molecules in common diseases. However, despite recent progress in mass spectrometry-based proteomics, which made possible the deep analysis of single-cell types like yeast[3] or human tissues and fluids,[4,5] proteomics of cell communities as diversified as those found in ecosystems (the gut microbiota forming the most densely populated community in the body) still remains a challenge at every stage.[6] This includes sample pretreatment, proteolytic digestion, peptide separation and analysis using liquid chromatography coupled to tandem mass spectrometry (LC−MS/MS), microbial protein identification and quantification, and lastly downstream taxonomic and functional annotations. This has been recently reviewed, leading to an early form of recommendations for performing and reporting label-free metaproteomics studies of human microbiomes in a more standardized way.[7]

Once the complex peptide mixture has been properly prepared and analyzed, a major difficulty remains: mass spectral interpretation, which is hampered by two main obstacles in metaproteomics. The first is the construction of a relevant multiorganism protein sequence database. Such a database can be generated from individual bacterial genomes translated into proteins. However, the risk is then to pick up a great number of nonrelevant sequences and miss many others from uncultured microorganisms.[8] Matching mass spectra against metagenomes specific to the samples studied would

Received: August 28, 2020

theoretically be the best solution.[9,10] However, such individual metagenomic data are not always available. In the absence of genomic information on the host and its symbiotic microbiome, an ingenious iterative search combining filtered generalist databases and RNA-seq-based protein database of the host has been developed, opening new avenues to put into perspective the taxonomic and functional compositions of a microbiome with the host proteome.[11] In the case of the human gut microbiome, an economic and more easily accessible solution is the use of the public metagenomic database IGC 9.9 generated by our consortium MetaHIT project.[12] It is directly derived from the whole genome sequencing of 1267 individual fecal samples from Europe, United States, and China plus a selection of sequenced gut bacterial genomes.[2]

The second obstacle is the optimization of the interrogation strategy. In addition to increasing the computation time, matching large mass spectral data sets to a large translated metagenomic database may also dramatically decrease the identification rates at a given false discovery rate (FDR) threshold calculated via classical target-decoy-based approaches.[13] To address this issue, Jagtap *et al.* proposed a two-step interrogation strategy known as "iterative database search".[14] Its principle is to first interrogate the target version of a large metaproteomic database with a relaxed FDR threshold, then to create a reduced database including all microbial sequences identified at this first stage, and re-interrogate the target-decoy version of this refined database with a stringent FDR threshold. Iterative database search has been used in metaproteomics studies[10,14,15] and is now implemented in automatic identification softwares.[16−18] This strategy has been successfully applied to interpret small numbers of human salivary metaproteomes[14,15] and gut metaproteomes of humans and mice.[10] When coupled with the interrogation of the human proteome, the search also allows identification of many human proteins that may be highly relevant in clinical contexts.[15,16] However, despite its increasing popularity, this strategy has been implemented on small data sets only (<10 samples), and its benefits have not been evaluated within the context of large-scale metaproteomic experiments nor was its possible impact on the reproducibility of the final peptide and protein identification list assessed.

An additional obstacle in metaproteomics is then to provide a reliable landscape of taxonomies and functions in a system where proteins are clustered into groups,[19−21] subgroups and groups,[22] or metaproteins,[23] using various grouping algorithms based on the shared peptide rule. Besides the fact that, even if designated by the same term, those protein assemblies do not necessarily correspond to the same thing due to various uses of the shared peptide rule, their taxonomic and functional annotation is not straightforward.[24] Many annotation strategies are proposed either at the peptide[25,26] or the protein[23] level by using different currently available knowledge databases and different taxonomic and functional annotation tools, which can provide significant variability in the annotation results.[27] In the absence of a consensus on which method to adopt, new approaches continue to emerge for taxonomic and functional annotations in order to view metaproteomes in a more physiologically meaningful way.

In this study, using X!Tandem and the grouping algorithm of X!TandemPipeline,[22] we aimed to optimize mass spectral interpretation and downstream knowledge of taxonomies and functions of the active bacterial community in the context of a large-scale metaproteomics study of the human gut microbiome including 48 individual samples. To this end, we compared three strategies of database interrogation: the classical (single-step) strategy and two iterative strategies. The first was the two-step strategy proposed by Jagtap *et al.*[14,15] In the second one, we included an intermediate step, consisting of a refined individual subdatabase interrogation before gathering the results into a concatenated final database that was interrogated in the last step. We applied the three strategies on the IGC 9.9 database concatenated with the human database. The comparisons are based on a series of well-defined qualitative and quantitative criteria such as the number of peptides and protein clusters identified throughout the entire data set or per sample, peptides and protein clusters that are consistently or specifically identified, reproducibility of MS/MS interpretations and final metaproteomic profilings of replicated samples, and taxonomic and functional knowledge brought by each strategy. Our results show that interrogating IGC 9.9 using the iterative search brought significant gain in terms of the number of peptides and protein clusters identified and taxonomic and functional knowledge, and this was the most convincing with the three-step approach. We also verified that iterative searches added only slight variability to the peptide and protein lists of replicated samples.

■ EXPERIMENTAL SECTION

Samples and MS/MS Data Sets

Stool samples were self-collected as previously detailed[28] by 48 overweight/obese subjects, which were recruited at the Human Research Center on Nutrition (CRNH), Pitié-Salpêtrière Hospital (75013 Paris), as part of the dietary intervention study MICRO-Obes project.[29,30] This study has been registered in ClinicalTrials.gov (NCT01314690) and approved by the Ethical Committee of Hôtel-Dieu Hospital in Paris, France, in 2008 (under the number 0811792). All participants provided written informed consent. Data collection occurred in 2009 and 2010. Fecal samples were transferred to a biobank at −80 °C within 2 h of collection in anaerobic containers. Then about 1 g stool aliquots were cut frozen, and the microbiota were separated from the fecal matrix by flotation in a preformed Nycodenz continuous gradient according to the method previously detailed,[28] except that we reduced the size of the gradient as described in Section S1 (all steps on ice or at 4 °C, under anaerobiosis). The extracted microbiota were lysed on ice with a probe sonicator in an antiprotease cocktail containing buffer without a chaotrope or detergent. Then the suspensions were centrifuged at 5000*g* for 30 min at 4 °C to remove unbroken cells and large cellular debris. The supernatants were finally ultracentrifuged at 220,000*g* for 30 min at 4 °C to pellet cell envelopes that were used for the present study. These cell-envelope-enriched fractions were resuspended in a surfactant containing buffer before acetone precipitation and in-solution digestion of proteins in the presence of a mass spectrometry-compatible surfactant. Finally, the peptide bulk was desalted on C18 cartridges and analyzed by liquid chromatography coupled with tandem mass spectrometry (LC−MS/MS, all steps detailed in Sections S1−S3), on the PAPPSO proteomic facility (http://pappso.inrae.fr/).

Sample preparation and LC−MS/MS analysis were carried out only once for 47 of the samples and repeated multiple times for one randomly selected sample for a study of reproducibility. More precisely, this sample was prepared in triplicate from microbiota extraction up to resolubilization of the peptide mixture in LC buffer. These preparations, called A, B, and C, were injected nine, three, and two times, respectively (Figure S1). In total, we thus performed 61 LC−MS/MS runs of which 47 corresponded to nonreplicated stool samples and 14 corresponded to the replicates, with a blank inserted between each sample injection. The analytical sequence of the samples is reported in Table S1. The mass spectrometry data have been deposited to the MassIVE repository[31] (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) with the data set identifier MSV000085993 and the ProteomeXchange identifier PXD021050.

Peptide and Protein Identifications

MS/MS data were searched using the X!Tandem software[32] version 2015.04.01.1. against three databases: (i) IGC 9.9 (available at http://meta.genomics.cn/meta/dataTools), (ii) the *Homo sapiens* protein catalog from UniProt (April 2018), including canonical and isoform proteins from Swiss-Prot and TrEMBL, and (iii) a contaminant database including 58 sequences of common contaminants of spectrometry experiments, such as keratins, BSA, and trypsin. For all identifications, four types of modifications were searched: carbamidomethylation of cysteines (fixed modification), oxidation of methionines, excision of the N-term methionine with or without acetylation, and cyclization of the N-term (potential modifications). The mass tolerance was set to 10 ppm for the precursor and 0.02 Da for the fragments, and one missed cleavage was allowed. Final results from X!Tandem searches were filtered based on statistical significance of individual peptide and protein identifications.[33]

Interrogation was either classical or according to two variants of the iterative search. The classical strategy consisted of a one-step target-decoy interrogation of the translated IGC 9.9 database, together with the *H. sapiens* database and the contaminant database. The *e*-value thresholds for peptides and proteins were set to 0.05. The iterative strategy was either in two steps[14,15] or in three steps. Briefly, the two-step strategy began with the search of mass spectra against the target version of the translated IGC-9.9 sequences, along with the *H. sapiens* and contaminants proteins, with a relaxed *e*-value threshold of 10 for peptides and proteins. In the second and last step, mass spectra were searched again against a refined target-decoy database constructed from all microbial proteins identified in the first step, concatenated with *H. sapiens* and contaminant proteins, with an *e*-value threshold of 0.05 for peptides and proteins. In the three-step variant of this strategy, we added an intermediate step, where each individual MS/MS data set was searched against the target version of its own specific subdatabase constructed from all bacterial proteins identified in the first step, concatenated with *H. sapiens* and contaminant proteins, with an *e*-value threshold of 0.05 for peptides and proteins. In the third and last step, mass spectra were searched again against the refined target-decoy database constructed from all microbial proteins remaining at the second step, along with *H. sapiens* and contaminant proteins, with an *e*-value threshold of 0.05 for peptides and proteins. The three methods used are illustrated in Figure S2.

Protein inference and clustering were performed using the grouping algorithm included in X!TandemPipeline[22] based on the principle of parsimony (Figure S3) in "combine" mode to generate a unique data table gathering all samples desired (Figure S3C). A minimum list of proteins present in the samples was thus generated based on the following rules. (1) A minimum of two peptides identified across all samples in the data set is set to validate a protein in order to exclude proteins with weak proof of presence. (2) The presence of a protein is attested if it contains at least one specific peptide, which is not seen in any other protein. (3) If a protein has no specific peptide (no proof of presence), it is eliminated. (4) Proteins identified with the same set of peptides are assembled into subgroups because one cannot distinguish which of these proteins are present. (5) At last, groups of proteins can be formed by gathering subgroups that share peptide(s). We based the present study on peptide and subgroup reports only (group entities as defined in X!TandemPipeline were not used, Figure S3), and we chose to designate "subgroups" of the X!TandemPipeline by "metaproteins", a term recently introduced in metaproteomics,[23,34] even if it may refer to different realities according to the grouping algorithm used.

Metaprotein Quantification

In this study, comparisons were essentially based upon inventories of peptides and metaproteins. We further quantified metaproteins by summing the spectral counts (SCs) of their specific peptides, i.e., excluding shared peptides, which bear information that is difficult to deconvolve.[35]

Taxonomic and Functional Annotations

We chose to deal with annotations at the protein level. For this purpose, all proteins embedded within each microbial metaprotein were taxonomically annotated with the sequence aligner DIAMOND (Double Index Alignment of Next-generation sequencing Data)[36] against the nonredundant NCBI database, with an *e*-value threshold of 10^{-4}. The complete taxonomic assignment (from superkingdoms to species) of the hit with the best bit-score was designated as the taxonomic assignment of the protein. Then only metaproteins whose all component proteins shared the same taxonomic annotation at the species level were functionally annotated using the KEGG (Kyoto Encyclopedia of Genes and Genomes) resource (release 89.0), with an *e*-value threshold of 10^{-5}, a bit-score threshold of 60, and using the sensitive mode of DIAMOND. The functional annotations with the better bit-scores were assigned to the protein. When multiple functions were assigned to the same protein, all of them were taken into account so that each species-specific metaprotein was functionally annotated with all KEGG Orthology (KO) entries assigned to all its component proteins. The functional and taxonomical annotations are available at https://doi.org/10.5281/zenodo.3997093.

Evaluation Criteria

The number of unique peptides and metaproteins and taxonomic and functional annotations were considered to compare the three search strategies at the individual or the whole cohort level. Differences were tested by ANOVA and with the post-hoc paired *t* test when the normality assumption was fulfilled (Shapiro test, $p > 0.05$) and by the Friedman test with the post-hoc paired Wilcoxon test otherwise. The *p*-values were adjusted by the Bonferroni method, and the significance threshold was set to 0.05. Interpolation of the total number of unique peptides and metaproteins identified with an increasing number of samples was performed with iNEXT.[37] The number of unique peptides and metaproteins specifically identified with a given search strategy was also considered. For that purpose,

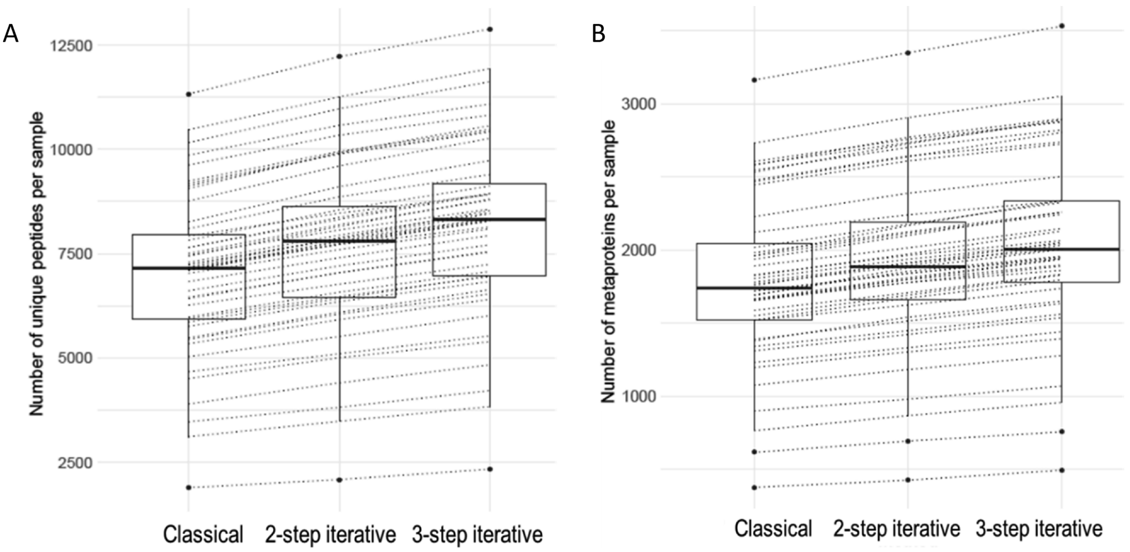


Figure 1. Number of unique peptides (A) and metaproteins (B) identified with each search strategy in every of the 48 individual enriched-envelope fractions of the human gut microbiome. Observations related to the same sample through the three methods are joined by dotted lines. All differences are significant (*p*.adj < 0.001 by ANOVA followed by the post hoc paired *t* test).

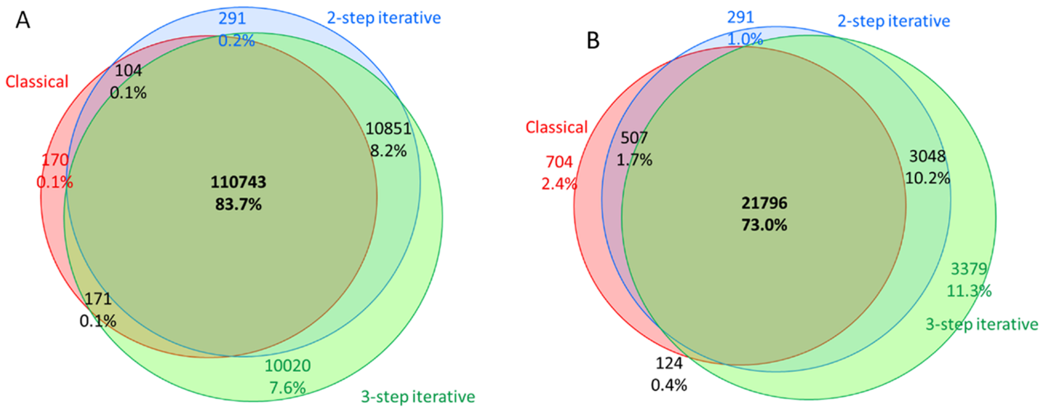


Figure 2. Venn diagrams of unique peptides (A) and metaproteins (B) identified with each search strategy across 48 individual enriched-envelope fractions of the human gut microbiome. The three-step strategy gave the highest number of peptides and metaproteins overall and specifically identified.

lists of peptides were compared based on their sequence and modifications, and lists of metaproteins were compared based on their identifiers directly provided in the output file of the grouping algorithm of X!TandemPipeline for comparisons of samples within the same combination (within-combination comparisons, Figure S3C) and lists of protein members embedded within each metaprotein for comparisons of samples from different combinations (between-combination comparisons). All computations were performed with RStudio version 1.1.383 and R version 3.3.3.[38]

RESULTS AND DISCUSSION

The Iterative Strategies Look More Deeply into Gut Metaproteomes

In order to compare the ability of each strategy to identify peptides and metaproteins, we randomly selected one of the 14 technical replicates for combination with all other nonreplicated 47 individual samples. Based on a cutoff of 0.05 for the *e*-value of individual peptide and protein identifications,[33] the a-posteriori peptide FDR returned by X!Tandem was always far below 1% (0.02, 0.03, and 0.05% for the classical, the two-step, and the three-step iterative approaches, respectively). Step-by-step reduction of the database and final percentages of interpreted spectra are summarized in Table S2. For the whole data set comprising the 48 nonreplicated biological samples, the three-step iterative method yielded 30,000 and 70,000 additional peptide spectrum matches (PSMs) compared to the two-step and the classical method, respectively. At the same time, distribution of the mass delta values remained unchanged whatever the method (Figure S4), indicating that PSMs were equally reliable for the three methods. Under these conditions, the overall number of identifications across the 48 individual samples increased when moving from the classical to the two-step or from the classical to the three-step strategy: +9.7 or +18.5% for unique peptides and +10.9 or +22.5% for metaproteins. This benefit was observed for every sample (Figure 1), resulting in very low adjusted *p*-values by ANOVA followed by the post hoc paired *t* test (all *p*.adj < 0.001).

The Venn diagram of Figure 2 further illustrates that only a few peptides and metaproteins were specifically identified by the classical or the two-step iterative search, while a substantial

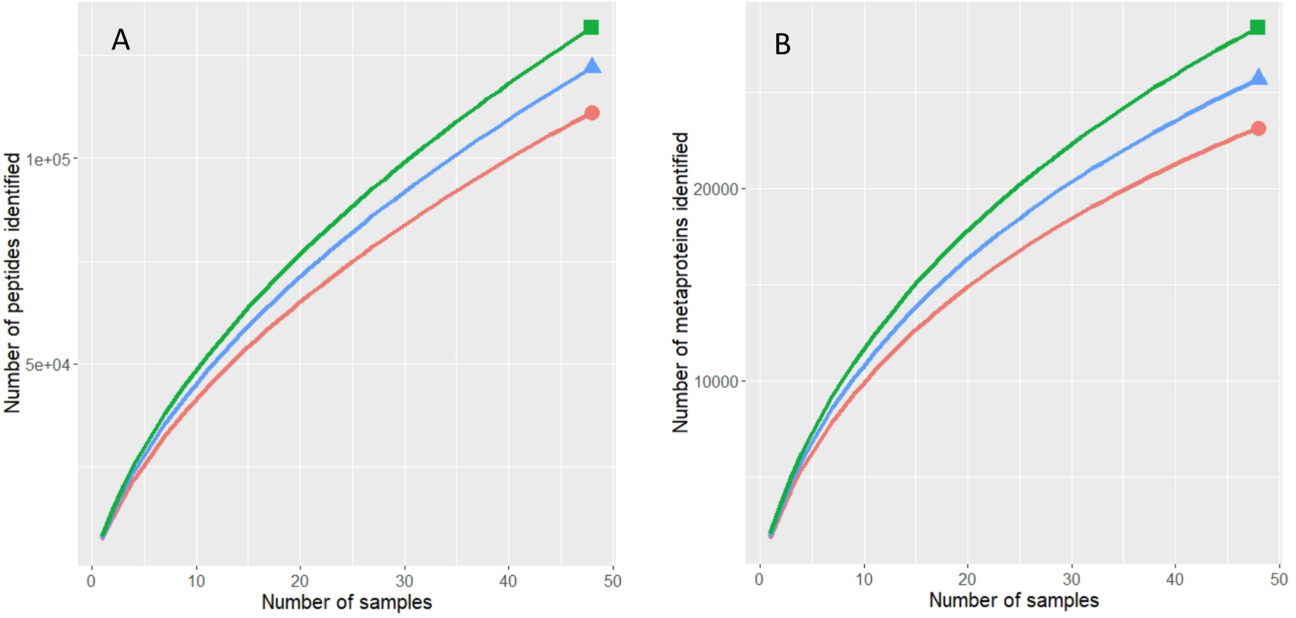


Figure 3. Diversity of unique peptides (A) and metaproteins (B) defined as the mean number identified with an increasing number of individual enriched-envelope fractions of the human gut microbiome. (red circle) Classical, (blue triangle) two-step iterative, (green square) three-step iterative. No plateau was observed even with the more refined search strategy.

number of peptides and metaproteins were identified by the three-step iterative strategy only.

The gain in peptide sequence matches provided by the two-step strategy compared to the traditional one-step search was first illustrated using the example of three human oral mucosa exudates.[15] Using the search algorithm ProteinPilot at a 0.05 local FDR threshold, the iterative two-step search increased by 26% the total number of unique peptides identified, and the gain was the most remarkable for the microbial fraction, whose contribution to the whole metaproteome was small in those oral samples (4.7% of total unique peptides with the traditional search) but almost doubled to 8.2% with the two-step search. The authors also provided evidence for the superiority of the two-step strategy when using other search algorithms, including X!Tandem, even if the magnitude of gain was program-dependent. In the present work including 48 individual human gut samples specially prepared to focus on microbial cell surface metaproteomes for clinical interest (beyond the scope of this paper), we demonstrate the interest of the two-step search on a whole new dimension as regards with sample size, database size, and the overwhelming predominance of microbial peptides. We further demonstrate that adding an intermediate step to refine each individual subdatabase and therefore consistently decrease its size (Table S2) before performing a last search of all MS/MS data sets against an assemblage of these subdatabases still increases peptide and metaprotein identifications by a substantial amount. Importantly, the three-step strategy also reduced the number of peptides that were matched in the classical search but missed in the iterative searches (341 and 274 unique peptides missed when using the two-step and the three-step searches, respectively, instead of the classical search; Figure 2A). Loss of a very small number of peptide matches with the iterative strategy was already observed by others, with the reason for this being not clear.[15] For the outcome of metaprotein identifications, loss was somewhat greater when moving from the classical to the three-step strategy (1211 metaproteins) than when moving from the classical to the two-step strategy (828 metaproteins lost) (Figure 2B). This is because increasing peptide matches by iterative search increased the likelihood of clustering several protein members within the same metaprotein based on the shared peptide rule. At the same time, the likelihood of distinguishing many more metaproteins based on the same shared peptide rule increased, so that the overall outcome weighted heavily in favor of the iterative strategies and even more of the three-step strategy. At last, we could verify that the gain brought when moving from the classical to the three-step strategy was as beneficial for a small-sized experiment (five metaproteomes randomly selected among the whole cohort) compared to several tens of samples as here.[39]

However, even if peptide and metaprotein discovery increased when moving from the classical to the two-step and then the three-step iterative search, none of the method enabled reaching a plateau with 48 individual samples (Figure 3), providing a clear illustration of the huge individual specificity of gut metaproteomic profiles, even within a relatively homogeneous group of overweighed patients recruited in the same region. Although this result could have been expected from what we know about specificity of individual microbiomes, either by 16s rDNA[40] or whole metagenome sequencing,[41] this is the first illustration of such a diversity at the metaproteomic level, even when using the most refined peptide mass matching strategy.

The Iterative Strategies Look More Deeply into the Taxonomies and Functions of Gut Metaproteomes

Our protein-centric taxonomic annotation allowed us to annotate all human proteins and more than 99% of microbial proteins (99.14, 99.18, and 99.16% for the classical, the two-step, and the three-step strategies, respectively). Then, we looked at the taxonomic consensus among protein members within each metaprotein. Only two metaproteins from the classical and the two-step iterative searches and one from the three-step iterative search contained a mix of human histones and various microbial proteins. The remainder was composed of 608, 676, and 716 "pure" human metaproteins and 22,521, 24,964, and 27,630 "pure" microbial metaproteins for the classical, the two-step, and the three-step approaches, respectively (Table S3). Therefore, in these pretreated samples where microbiota were first extracted from the fecal matrix by gradient and where spectra were searched against both IGC and human databases, the percentages of metaproteins from human origin were low: 2.6% for the classical and the two-step iterative approaches and 2.5% for the three-step iterative approach. Among them, the most abundant (based on the sum of SCs of their specific peptides) were pancreatic Elastase 3A, Chymotrypsin-C, IgGFc-binding protein, Submaxillary gland androgen-regulated protein 3B, Phospholipase A2, pancreatic triacylglycerol lipase, Polymeric immunoglobulin receptor, and Mucin-2. Bacterial surface-coating proteins from human origin can provide valuable information about the host−microbiome interactions. For instance, we found in three volunteers substantial amounts of the well-known neutrophil-derived proteins S100-A8 and S100-A9 (also called calprotectins), whose dosage in feces is used to identify an inflammatory bowel condition and determine the next course of action in diagnosis and treatment.[42] We also identified other immune-cell-derived proteins that are all related to host defense against bacterial infections and were over-represented in the same samples that already presented high levels of calprotectins, and this is despite the fact that these patients were not known to have symptoms of gastrointestinal disease. Interestingly, the abundance of those proteins slightly increased when moving from the classical to iterative strategies in the concerned samples but not in the other ones, making iterative strategies potentially useful to boost statistical power. In this respect, it should be mentioned that cell surface metaproteomes display more than twice the proportion of proteins from human origin than do whole metaproteomes (own scientific data), which might be of particular interest in clinical research and host−microbiome interaction studies.[43,44]

For all three search strategies, 1% of the "pure" microbial metaproteins had at least one component protein that could not be annotated, about 11% had components with more than one taxonomic annotation at the species level, and about 88% had components with a unique annotation at this level, with a slightly higher consensus when moving from the classical to the two-step and then the three-step search (Figure 4 and Table S3). Consensus substantially increased at the genus level and then at the order level, with nearly 94 and 98% of all microbial metaproteins, respectively, with a unique annotation at these taxonomic levels, for all three strategies (Figure 4).

Therefore, any of the three strategies using X!Tandem to match the mass spectra against the translated public metagenomic database IGC 9.9 (plus the *H. sapiens* protein catalog from UniProt), combined with the grouping algorithm of X!TandemPipeline[22] to cluster proteins into metaproteins, and taxonomic annotations at the protein level, allowed us to reliably predict the taxonomic lineage of about 88% of all microbial metaproteins up to the species level (a little bit less with the classical search and a little bit more with the three-step search). At the highest phylum and superkingdom levels, consensus reached almost 99.0%, with the three-step strategy giving the most consensual results. It should be remembered that, for each member of the protein list returned by X!Tandem, we chose to retain only the first alignment hit returned by the sequence aligner DIAMOND[36] against the nonredundant NCBI database. We then looked at the consensus of these first hits within each metaprotein and at the different taxonomic levels, which means finally complying with both principles of the lowest common ancestor[45] and taxonomic consensus annotation.[24] The proposed approach enables dealing with potential broad taxonomic ranges within metaproteomic assemblages. Remarkably, all the X!TandemPipeline assemblages resulted in a strong consensus even at the lowest species level and whatever search strategy was used. It would be very interesting to look at the results provided by other grouping algorithms, but this is beyond the scope of this paper.

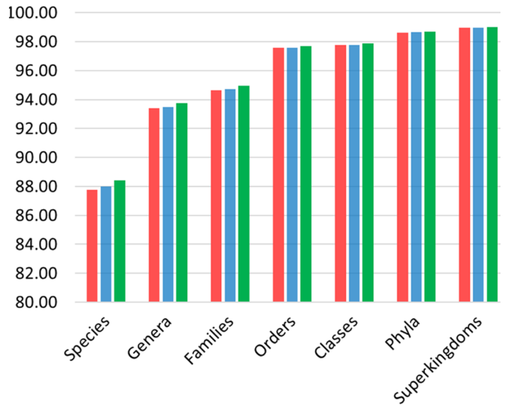


Figure 4. Percentage of microbial metaproteins with a unique annotation at the different taxonomic levels. (red square) Classical, (blue square) two-step iterative, (green square) three-step iterative. The remaining microbial metaproteins had either at least one component without taxonomic annotation or components with diverse annotations at the taxonomic level under consideration.

Overall, the three search strategies provided very close taxonomic landscapes in terms of percentages of metaproteins distributed among the different phyla, with a clear over-representation of metaproteins assigned to the phylum Firmicutes (Figure 5A). However, when considering the numbers of predicted species, iterative searches substantially increased taxonomic resolution at all levels and more especially at the lower species level (Figure 5B), with an additional 228 species predicted when moving from the classical to the two-step search and another additional 200 species when moving from the two-step to the three-step search. That said, a common core of 2111 species belonging to 29 phyla was predicted by all three strategies (Figure 5C). Specificities and intersections of all three strategies are illustrated in Figure 5C. Interestingly, looking in more detail at those 228 and 200 newly predicted species, we found a higher proportion of species belonging to the phylum Bacteroidetes and other less represented phyla and a lower proportion of species belonging to the predominant phylum Firmicutes, compared to what was observed within the intersection of the three searches or when using the classical search (Figure S5). For instance, the Bacteroidetes to Firmicutes species ratios in the additional pools brought by the two-step and the three-step iterative searches were 0.40 and 0.85, respectively, compared to 0.36 within the intersection of the three searches (Figure S5). In other words, boosting peptide mass matching by iterative searches added metaproteins to every phylum without reshaping the overall taxonomic landscape, but refined our taxonomic knowledge of the system particularly for less

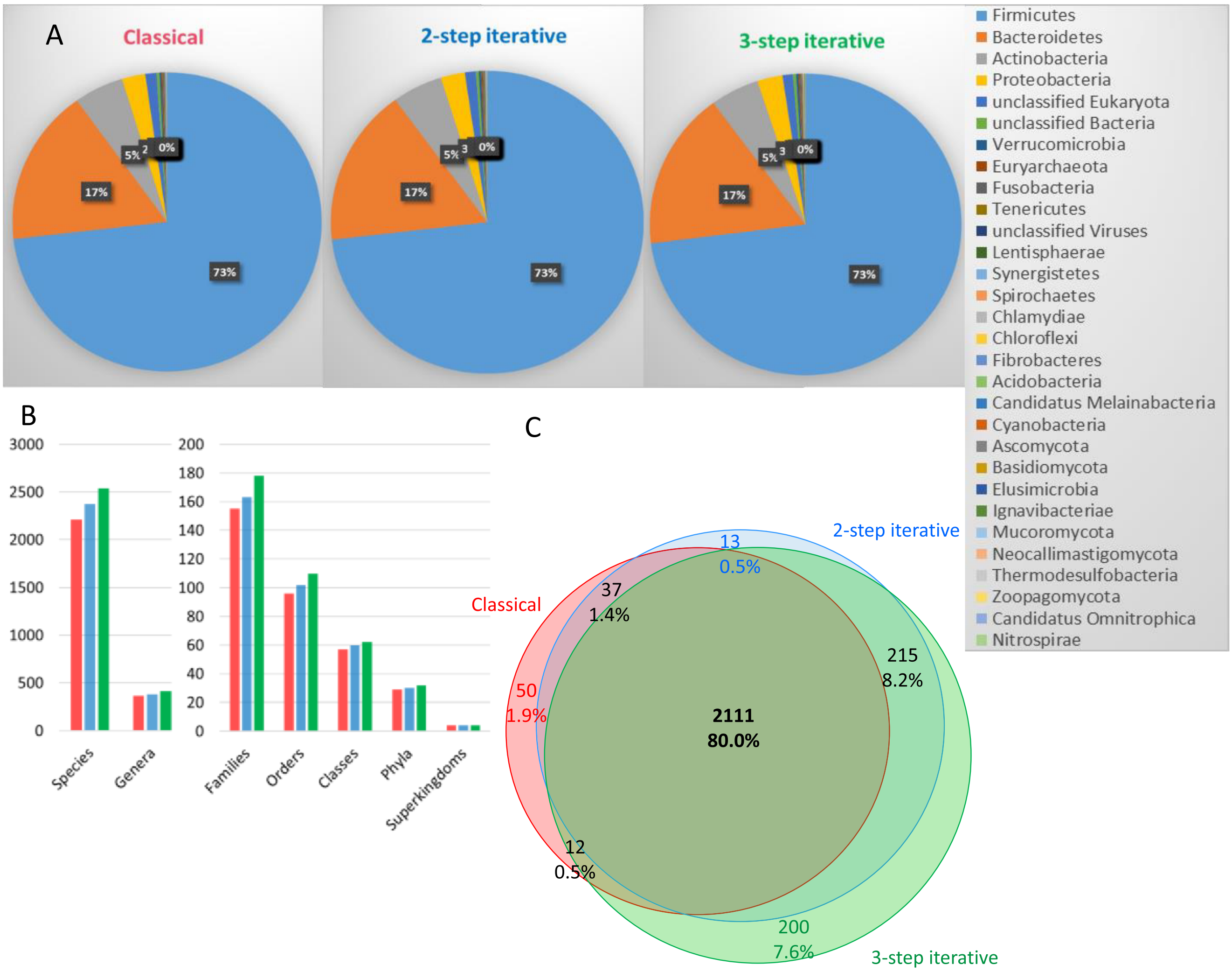


Figure 5. (A) Percentages of metaproteins assigned to the different phyla with the three search strategies. (red square) Classical, (blue square) two-step iterative, (green square) three-step iterative. Only metaproteins whose all component proteins were assigned to the same phylum were taken into account, i.e., 98.6 to 98.7% of all microbial metaproteins (see Figure 4). (B) Number of taxonomic entities predicted with the three search strategies. Only metaproteins whose all component proteins were consensually assigned at the taxonomic level on the *x*-axis were taken into account (see Figure 4). (C) Venn diagram of microbial species predicted with each search strategy. Only metaproteins whose all component proteins were assigned to the same species were taken into account, i.e., about 88% of all microbial metaproteins (see Figure 4).

represented taxa, and furthermore with a small increase in taxonomic consensus within metaproteins.

Using the KEGG resource, we then proceeded to the functional annotation of those 88% microbial metaproteins whose all component proteins were assigned to the same species. For each metaprotein, we took into account all KO entries of all its component proteins, so that each species-specific metaprotein was potentially assigned to several KO entries. Overall, 147 and 126 additional KO entries were predicted when moving from the classical to the two-step and from the two-step to the three-step strategy, respectively, while a core of 2198 KO entries was predicted by all three strategies (Figure 6). The functionalities predicted from the three data sets, as well as their differences and overlaps, are mapped in Figure S6. Examples of enhanced functional knowledge when moving from the classical to the two-step or the three-step iterative strategies (light blue lines in Figure S6) are amino acid, starch and sucrose, glycerophospholipid, and methane metabolisms, or fatty acid elongation, and examples when moving from the two-step to the three-step iterative strategy

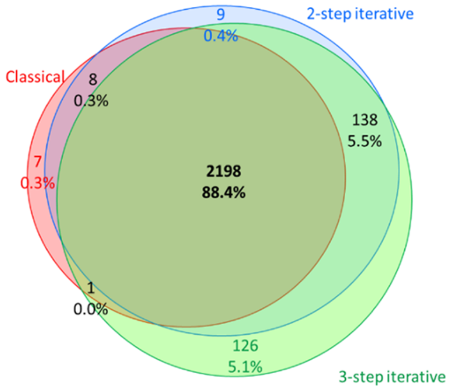


Figure 6. Venn diagram of KO entries predicted with each search strategy. Only microbial metaproteins whose all component proteins were assigned to the same species (i.e., about 88% of them, see Figure 4) were functionally assigned, considering all the KO entries of their protein members.

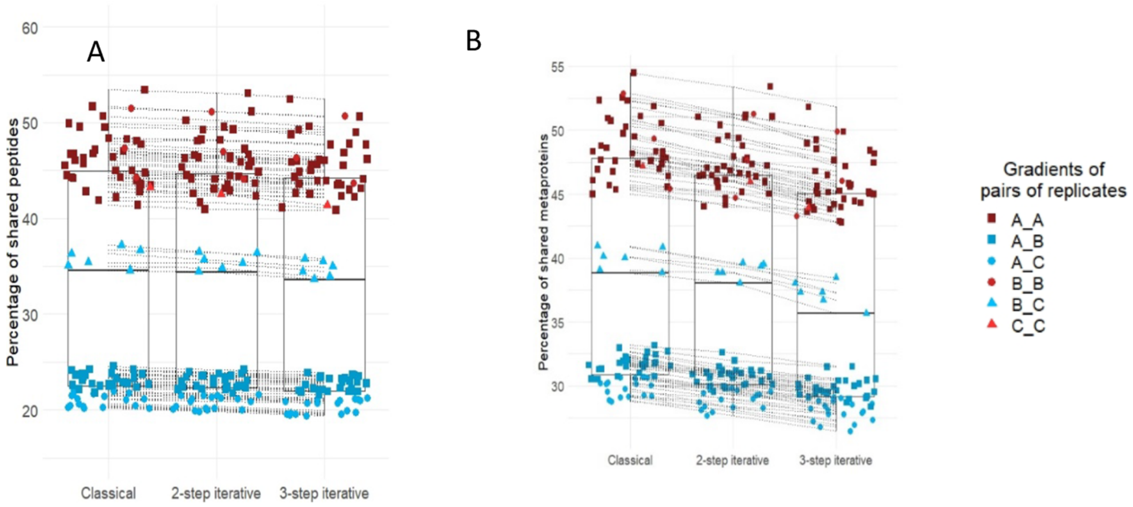


Figure 7. Peptide (A) and metaprotein (B) overlapping in all pairs of replicates, with each interrogation strategy. All differences are significant (p.adj < 0.001 by the Friedman test with the post hoc paired Wilcoxon test).

(green lines in Figure S6) are amino acid biosynthesis and degradation, lipopolysaccharide biosynthesis, vitamin B6 and B9 biosynthesis, or terpenoid backbone biosynthesis.

We found that most "species-consensual" metaproteins (82.4, 82.6, and 82.9% for the classical, the two-step, and the three-step strategies, respectively) were assigned to a single KO entry (Figure S7). Importantly, the functional diversity (the count of different KO entries per metaprotein) of the remaining 17.0−17.6% was not related to the number of their protein members but to the own functional diversity of their individual members, even if reduced to one (Figure S8). As a corollary, the functional diversity within each metaprotein scarcely ever exceeded that of its functionally best-endowed protein member (except in 0.08% of the metaproteins from the classical or two-step searches and in 0.07% of the metaproteins from the three-step search), which means that, when multiple proteins were embedded within the same metaprotein, they were assigned to redundant KEGG annotations.

The end result is that, using any one of the three search strategies, we were able to say "who does what in the system" for almost 90% of the microbial metaproteins, with an increasing taxonomic consensus when moving from the classical to the two-step and then the three-step search strategy. At last, although this was not the focus of the present study, we wanted to have a glimpse of the distribution of microbial genes among phyla as we did for metaproteins in Figure 5A. The gene count table of the 48 microbiomes, obtained by read mapping on the IGC 9.9 database, was extracted and annotated from previous data obtained by our consortium,[29,30,46] thus allowing us to compare the profiles of gene potential and expression from a taxonomic point of view (Figure S9). Both images proved to be close. However, a lower proportion of genes was attributed to Firmicutes and Actinobacteria, and a higher proportion was attributed to Bacteroidetes and Proteobacteria, compared to the distribution reported for metaproteins. It is uncertain whether this is a biological reality or simply the reflection of a more efficient lysis of Gram-positive bacteria in the metaproteomic preparative workflow or any other technological bias in either approach.

The Iterative Strategies Add Minor Variability to Technical Replicates

Protein inference and clustering for a given sample can vary somewhat depending on the other samples included in the experiment. Thus, to properly evaluate the reproducibility of metaproteomic profiling of technical replicates, we combined each of the 14 replicates together with the 47 nonreplicated samples. The 14 independent peptide and protein identification tables served to measure reproducibility in the actual context of large-scale experiments. Peptide overlapping in pairs of replicates (n = 91 pairs) slightly but significantly decreased when moving from the classical to the two-step and then to the three-step iterative strategy (all p.adj < 0.001, the Friedman test with the post hoc paired Wilcoxon test, Figure 7A). As metaprotein identifiers differ between groupings, assessment of their overlapping between pairs of replicates required prior alignment of metaproteins based on the "Description" of their protein members. The end result was that metaprotein overlapping followed the same pattern as peptide overlapping (all p.adj < 0.001, the Friedman test with the post hoc paired Wilcoxon test, Figure 7B).

In order to assess whether this variability affected the positioning of each replicated sample relative to the other 47 samples, we computed all pairwise Spearman correlations between metaproteomic profiles (expressed as the sum of SCs of specific peptides per metaprotein) of the replicate and that of every nonreplicated sample within each combination (within-combination correlations), and this was repeated for each search strategy. Figure 8 illustrates that all pairwise correlations (14 × 47 for each interrogation strategy) slightly but significantly decreased when moving from the classical to the two-step and the three-step iterative search (all p.adj < 0.001, the Friedman test with the post hoc paired Wilcoxon test). However, this did not affect the positioning of the replicate relative to all other samples as exemplified by dendrograms of Figure S10.

In order to assess both technical and biological variabilities within each of the three search strategies, we computed all pairwise Spearman correlations between metaproteomic profiles of all samples from the 14 combinations taken in pairs (between-combination correlations), and this was repeated for each search strategy (91 pairwise comparisons

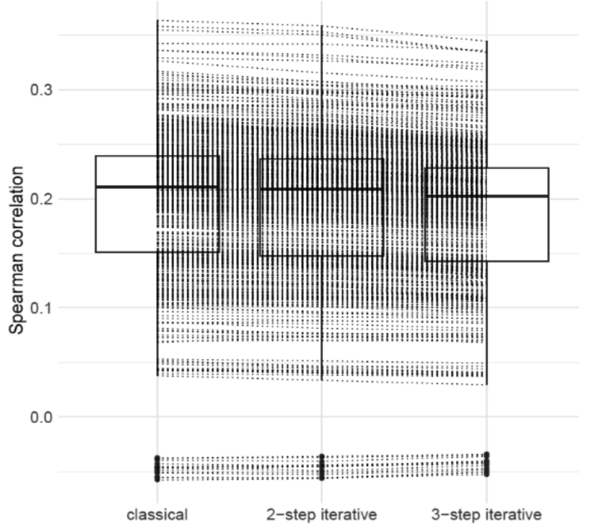


Figure 8. Within-combination Spearman correlations of each replicate with every nonreplicated sample. Each box contains 14 × 47 pairwise Spearman correlation values based on metaproteomic profiling (sum of SCs of specific peptides per metaprotein). Pairwise correlations of the same pair of samples through the three methods are joined by dotted lines. All differences are significant (p.adj < 0.001 by the Friedman test with the post hoc paired Wilcoxon test).

for each strategy). Between-combination pairwise correlations were invariably around 0.99 for the 47 identical nonreplicated samples (upper lines of dots in Figure 9A). As expected, between-combination pairwise correlations for replicated samples, which is a measurement of the technical variability, were substantially lower (intermediate values in Figure 9A) but fortunately still much higher than any pairwise correlation between biologically different samples, which is a measurement of the biological variability (lower clouds of dots in Figure 9A). Focusing on replicated samples only, correlation values fell into the ranges of 0.6−0.4 when the gradient and the MS run differed and 0.7−0.6 when only the MS run differed (Figure 9B). For all types of paired samples (identical, or replicated, or biologically different), we observed a moderate but significant decrease of correlation values when moving from the classical to the two-step and the three-step strategy (all p.adj < 0.001, the Friedman test with the post hoc paired Wilcoxon test). Although these p-values are only indicative, given that the correlations are not independent (since the same samples are used several times), this means that both technical and biological variabilities slightly increased when the iterative searches were used instead of the classical one.

Such an extensive reproducibility study has never been tried before in metaproteomics or even in single organism proteomics, with the exception of a broad interlaboratory study,[47] which compared a wide range of conditions (multiple laboratories and instruments, lab-specific protocols, or common SOPs) on peptide and protein inventories of diverse samples from defined protein mixtures to complex proteomes such as yeast extract. When a peptide mixture from a single digestion was repeatedly analyzed on the same instrument and HPLC column, the overlapping fraction of peptides in pairs of replicates was around 45%, whatever the complexity of the peptide mixture.[47] This figure coincides with ours, related to pairs of LC−MS/MS injections from the same gradients, where peptide overlapping always exceeded 40% and ranged up to almost 55% whatever the search strategy (red dots of Figure 7A). Yet, overlapping at the metaprotein level was much lower in our samples (43−54%, red dots of Figure 7B, all search strategies considered) than that reported before for yeast extract proteins (60−70%), whose overlapping was already lower than that in less complex defined protein mixtures.[47] This indicates that variable peptide lists from repeated analyses may be sufficient to infer reproducible protein identifications in low-complexity protein mixtures or at most in simple organisms but not to infer reproducible metaprotein assemblages, even with the most sophisticated search strategy. However, omitting metaproteins with uncertainty of data acquisition, i.e., metaproteins with only one specific peptide of very low abundance (one SC) as sole proof of evidence in one of the two replicates only, strikingly increased metaprotein overlapping between replicates, over 82% and up to more than 90% for same-gradient paired samples, and over 60% for the least overlaps (Figure S11).

This re-emphasizes the difficulty in obtaining highly reproducible identifications from repeated MS analyses, even from the same wet preparation, a statement already pointed out by those who dared the comparison, even on much less complex samples. This is of course inherent to the traditional data-dependent acquisition (DDA), which is plagued by the stochastic nature of precursor selection and low sampling efficiency at the lower end of the dynamic range, due to the limited speed of mass spectrometers.[48] This is still more challenging in metaproteomics where a myriad of low-abundance precursors are embedded within each injection. This might have been less intense with a more recent instrumentation, but clearly, multiple-step database search slightly increases both biological and technical variabilities, with the first being beneficial but at the cost of the second. However, we show that this is clearly not a major drawback for providing a reproducible global view of similarities and dissimilarities between tens of samples in a cohort because biological variability between individual microbiomes largely exceeds technical variability, even when keeping metaproteins with low proof of presence. Indeed, the common core throughout repeated identifications exactly preserved the position of any replicated sample with respect to other biological samples in the cohort. This means that questioning the position of a metaproteome among many others is robust, even if it is analyzed only once, and this is true for all three search strategies. This is an important finding as replicate analyses, feasible on a reduced number of experimental samples, cannot be routinely extended to large cohorts. Nonetheless, it may be advantageous to discard metaproteins with uncertainty of data acquisition in a reasoned calibrated strategy adapted to each study when one seeks to highlight robust significant discriminating traits between clinical groups or any other environmental phenotypes.

Finally, our data also documented between-combination correlations, which were invariably close to 0.99 for all identical samples and all search strategies. This means that the grouping algorithm of X!TandemPipeline is highly stable even if the input peptide list differs due to one nonidentical sample. Moreover, between-combination correlations for replicates well reflected the bimodal distribution already observed for overlapping metaproteins, with the highest values corresponding to the same sample preparation repeatedly analyzed and the lowest ones corresponding to different sample preparations.

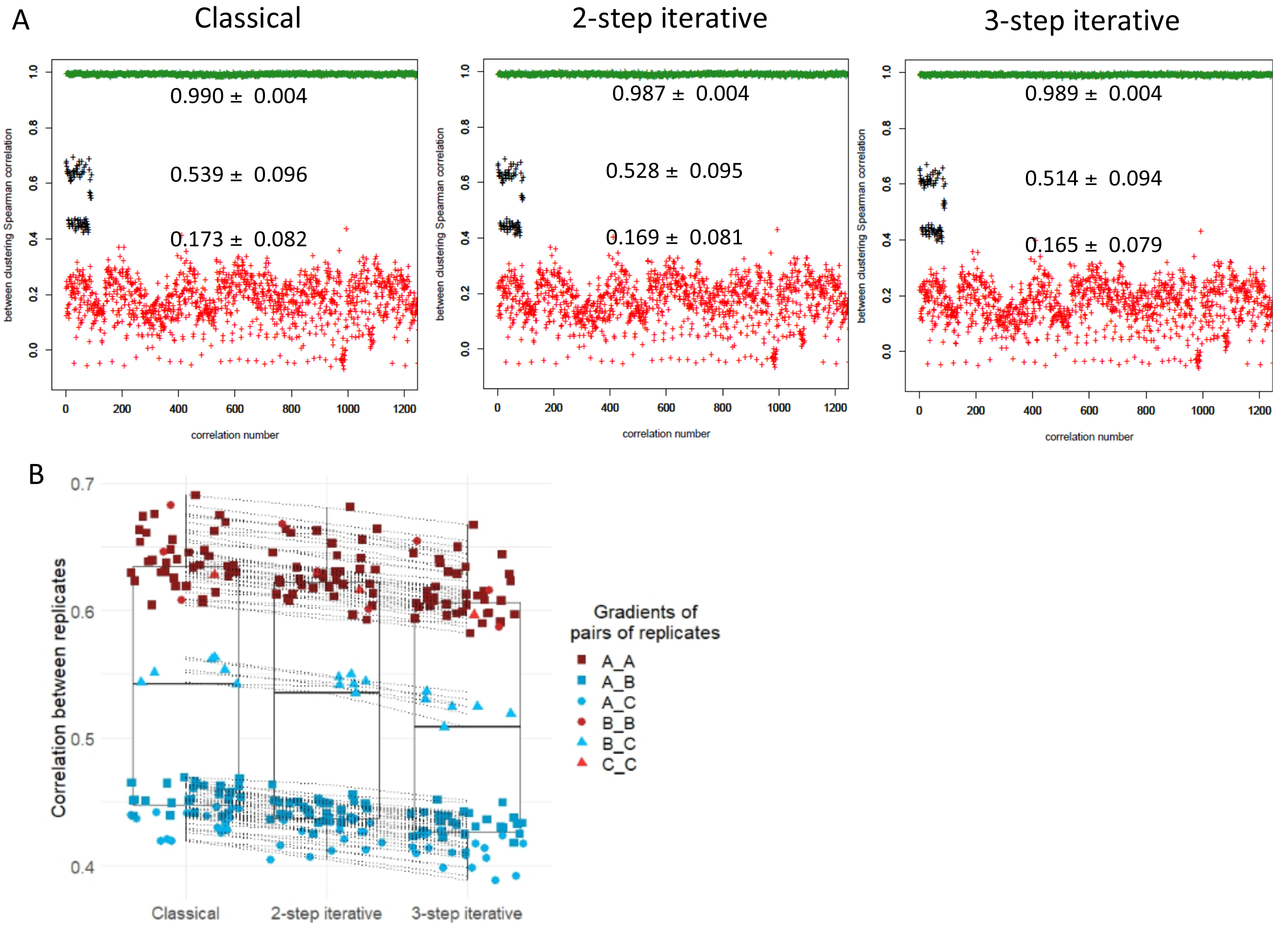


Figure 9. (A) All between-combination Spearman correlations for each of the three interrogation strategies. Identical nonreplicated samples (green markers) are closely correlated. Biologically different samples (red markers) are poorly correlated. Replicated samples (black markers) have intermediate correlation values. For each sample type, mean ± sd is superimposed onto the figure (all between-method means are significant, *p*.adj < 0.001, the Friedman test with the post hoc paired Wilcoxon test). (B) Detailed between-grouping Spearman correlations of replicates only (all *p*.adj < 0.001, the Friedman test with the post hoc paired Wilcoxon test).

■ CONCLUSIONS

Although MS acquisition was on an instrumentation that has now been exceeded, this study clearly proves the advantage of iterative searches of huge databases to which one refers to interpret several tens of thousands of mass spectra acquired from multiorganism communities, typically gut microbiomes. Although the iterative search strategy has already been proven to be very efficient on a limited number of oral or gut samples and is now implemented in automatic identification software, the present study was on a whole new dimension regarding sample size, database size, the overwhelming predominance of microbial peptides, interpretation of mass data not only in terms of the numbers of peptides and metaproteins identified but also of added knowledge of taxonomies and functionalities, and extended analysis of reproducibility of repeated analyses. This should encourage the newly emerging metaproteomic scientific community to systematically integrate this multistep search strategy as part of the challenging interpretation workflow of huge metaproteomic data sets. We further demonstrate an interest in adding to the two-step reference method an intermediate compilation stage consisting of a refined individual subdatabase interrogation before gathering the results into a concatenated final database. This three-step iterative search gave the highest numbers of peptides and metaproteins identified or taxa and functions predicted, providing the best understanding and knowledge of the system from the available mass spectral data sets, even if somewhat decreasing reproducibility of peptide and metaprotein identifications but, importantly, without modifying the relative positioning of replicated samples within a cohort of almost 50 individual microbiomes. Finally, the presence of those additional peptides that are specifically identified in such complex environments with the three-step method could be confirmed by a targeted spectrometry method, like parallel reaction monitoring (PRM).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00669.

Supporting Experimental Section S1. Stool sample collection and processing. Supporting Experimental Section S2. Protein digestion and peptide desalting. Supporting Experimental Section S3. LC−MS/MS analysis. Table S1. Analytical sequence of the samples. Table S2. Summary of the three search strategies. Table S3. Taxonomic diversity allowed by each of the three search strategies. Figure S1. Experimental design. Figure

S2. Overview of the three interrogation methods used in this paper. Figure S3. Clustering of proteins using the X! Tandem Grouping Algorithm of X!TandemPipeline. Figure S4. Plots of the mass delta distributions for the three methods. Figure S5. Species distribution among phyla in the intersection of the three searches (A) or in the additional pool brought by the two-step (B) and the three-step strategies (C). Figure S6. iPath projection of KO entries highlighted with the three approaches. Figure S7 Distribution of metaproteins as a function of KO entries embedded. Figure S8. If present, the functional diversity of metaproteins is related to the own functional diversity of their component proteins, not to their number. Figure S9. Percentages of genes (A) and metaproteins (B, recall of Figure 4A) assigned to the different phyla within the 48 microbiomes. Figure S10. Example of reproducibility of the positioning of replicates relative to other samples in the cohort, based on relative abundances of all metaproteins. Figure S11. Metaproteins overlapping in all pairs of replicates when singletons are omitted (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Sandra Plancade − *MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France; INRAE, UR875 MIAT, F-31326 Castanet-Tolosan, France*; Email: sandra.plancade@inrae.fr

Olivier Langella − *Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE − Le Moulon, 91190 Gif-sur-Yvette, France*; Email: olivier.langella@universite-paris-saclay.fr

Catherine Juste − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France*; Email: catherine.juste@inrae.fr

Authors

Ariane Bassignani − *Université Paris-Saclay, INRAE, MGP, 78350 Jouy-en-Josas, France; Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France; Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE − Le Moulon, 91190 Gif-sur-Yvette, France; MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France*; orcid.org/0000-0001-6267-4519

Magali Berland − *Université Paris-Saclay, INRAE, MGP, 78350 Jouy-en-Josas, France*

Melisande Blein-Nicolas − *Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE − Le Moulon, 91190 Gif-sur-Yvette, France*

Alain Guillot − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France*

Didier Chevret − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France*

Chloé Moritz − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France*

Sylvie Huet − *MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France*

Salwa Rizkalla − *Sorbonne Université, Inserm, UMRS Nutrition et Obésités; approches systémiques, Paris 75006, France; Assistance Publique Hôpitaux de Paris, Service de Nutrition, CRNH Ile-de-France, Pitié-Salpêtrière Hospital, Paris 75013, France*

Karine Clément − *Sorbonne Université, Inserm, UMRS Nutrition et Obésités; approches systémiques, Paris 75006, France; Assistance Publique Hôpitaux de Paris, Service de Nutrition, CRNH Ile-de-France, Pitié-Salpêtrière Hospital, Paris 75013, France*

Joël Doré − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00669

Author Contributions

S.P., O.L., M.B., and C.J. jointly conceived, planned, and directed this work. S.P. and S.H. inspired a rigorous approach in the choice of quantitative and qualitative criteria used in this study. S.P. designed all statistical analyses. A.B. performed all bioinformatics computations and statistical analyses. A.B. and C.J. designed the figures, interpreted the results, and drafted the core of the manuscript. S.P., M.B., M.B.-N., and O.L. helped shape the final manuscript. O.L. developed and upgraded the X!TandemPipeline to guarantee the management of large amounts of metaproteomic data. S.P., M.B., M.B.-N., O.L., and C.J. jointly supervised the findings and discussed and interpreted the results with A.B. C.J. and C.M. prepared all samples. A.G. optimized the mass tandem analysis of the microbiota. A.G. and D.C. performed the mass tandem analyses. K.C. and J.D., as leaders of the MICRO-Obes ANR project, made this community project possible. K.C. and S.R. planned and supervised the recruitment of patients and collection of samples. All authors have given approval to the final version of the manuscript and may have contributed to the very last version of the manuscript.

Notes

The authors declare no competing financial interest.

The mass spectrometry (MS) data are publicly accessible at https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp via the MassIVE repository with the data set identifier MSV000085993 and at http://www.proteomexchange.org/ via the ProteomeXchange repository with the data identifier PXD021050.

■ ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche (ANR) as part of the MICRO-Obes (ANR-07-GMGE-002.1-01) and the ProteoCardis (ANR-15-CE14-0013) projects and by the Métaprogramme INRA as part of the ObOmics project. We are grateful to Dr. Nicolas Pons, senior research scientist at Université Paris-Saclay, INRAE, MGP, for critical reading of the manuscript. The contribution of Emmanuelle Le Chatelier, senior research scientist at Université Paris-Saclay, INRAE, MGP, in providing the gene count table of the 48 microbiomes for comparison of gene potential and expression from a taxonomic point of view, is gratefully acknowledged.

■ ABBREVIATIONS

MS, mass spectrometry; IGC, integrated nonredundant gene catalog; LC−MS/MS, liquid chromatography coupled to tandem mass spectrometry; FDR, false discovery rate; SC, spectral counts; KEGG, Kyoto Encyclopedia of Genes and Genomes; KO, KEGG Orthology; DDA, data-dependent acquisition

REFERENCES

(1) Qin, J.; Li, Y.; Cai, Z.; Li, S.; Zhu, J.; Zhang, F.; Liang, S.; Zhang, W.; Guan, Y.; Shen, D.; Peng, Y.; Zhang, D.; Jie, Z.; Wu, W.; Qin, Y.; Xue, W.; Li, J.; Han, L.; Lu, D.; Wu, P.; Dai, Y.; Sun, X.; Li, Z.; Tang, A.; Zhong, S.; Li, X.; Chen, W.; Xu, R.; Wang, M.; Feng, Q.; Gong, M.; Yu, J.; Zhang, Y.; Zhang, M.; Hansen, T.; Sanchez, G.; Raes, J.; Falony, G.; Okuda, S.; Almeida, M.; LeChatelier, E.; Renault, P.; Pons, N.; Batto, J.-M.; Zhang, Z.; Chen, H.; Yang, R.; Zheng, W.; Li, S.; Yang, H.; Wang, J.; Ehrlich, S. D.; Nielsen, R.; Pedersen, O.; Kristiansen, K.; Wang, J. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **2012**, *490*, 55−60.

(2) Wen, C.; Zheng, Z.; Shao, T.; Liu, L.; Xie, Z.; Le Chatelier, E.; He, Z.; Zhong, W.; Fan, Y.; Zhang, L.; Li, H.; Wu, C.; Hu, C.; Xu, Q.; Zhou, J.; Cai, S.; Wang, D.; Huang, Y.; Breban, M.; Qin, N.; Ehrlich, S. D. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **2017**, *18*, 142.

(3) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **2014**, *13*, 339−347.

(4) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T.-C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509*, 575−581.

(5) Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582−587.

(6) Blackburn, J. M.; Martens, L. The challenge of metaproteomic analysis in human samples. *Expert. Rev. Proteomics* **2016**, *13*, 135−138.

(7) Zhang, X.; Figeys, D. Perspective and Guidelines for Metaproteomics in Microbiome Studies. *J. Proteome Res.* **2019**, *18*, 2370−2380.

(8) Rudney, J. D.; Xie, H.; Rhodus, N. L.; Ondrey, F. G.; Griffin, T. J. A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry. *Mol. Oral Microbiol.* **2010**, *25*, 38−49.

(9) Tanca, A.; Palomba, A.; Pisanu, S.; Deligios, M.; Fraumene, C.; Manghina, V.; Pagnozzi, D.; Addis, M. F.; Uzzau, S. A straightforward and efficient analytical pipeline for metaproteome characterization. *Microbiome* **2014**, *2*, 49.

(10) Zhang, X.; Ning, Z.; Mayne, J.; Moore, J. I.; Li, J.; Butcher, J.; Deeke, S. A.; Chen, R.; Chiang, C.-K.; Wen, M.; Mack, D.; Stintzi, A.; Figeys, D. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **2016**, *4*, 31.

(11) Gouveia, D.; Pible, O.; Culotta, K.; Jouffret, V.; Geffard, O.; Chaumot, A.; Degli-Esposti, D.; Armengaud, J. Combining proteogenomics and metaproteomics for deep taxonomic and functional characterization of microbiomes from a non-sequenced host. *npj Biofilms Microbiomes* **2020**, *6*, 23.

(12) Li, J.; Jia, H.; Cai, X.; Zhong, H.; Feng, Q.; Sunagawa, S.; Arumugam, M.; Kultima, J. R.; Prifti, E.; Nielsen, T.; Juncker, A. S.; Manichanh, C.; Chen, B.; Zhang, W.; Levenez, F.; Wang, J.; Xu, X.; Xiao, L.; Liang, S.; Zhang, D.; Zhang, Z.; Chen, W.; Zhao, H.; Al-Aama, J. Y.; Edris, S.; Yang, H.; Wang, J.; Hansen, T.; Nielsen, H. B.; Brunak, S.; Kristiansen, K.; Guarner, F.; Pedersen, O.; Doré, J.; Ehrlich, S. D.; MetaHIT Consortium; Bork, P.; Wang, J. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **2014**, *32*, 834−841.

(13) Muth, T.; Kolmeder, C. A.; Salojärvi, J.; Keskitalo, S.; Varjosalo, M.; Verdam, F. J.; Rensen, S. S.; Reichl, U.; de Vos, W. M.; Rapp, E.; Martens, L. Navigating through metaproteomics data: A logbook of database searching. *Proteomics* **2015**, *15*, 3439−3453.

(14) Jagtap, P.; McGowan, T.; Bandhakavi, S.; Tu, Z. J.; Seymour, S.; Griffin, T. J.; Rudney, J. D. Deep metaproteomic analysis of human salivary supernatant. *Proteomics* **2012**, *12*, 992−1001.

(15) Jagtap, P.; Goslinga, J.; Kooren, J. A.; McGowan, T.; Wroblewski, M. S.; Seymour, S. L.; Griffin, T. J. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **2013**, *13*, 1352−1357.

(16) Jagtap, P. D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J. E.; Rhodus, N. L.; Rudney, J.; Griffin, T. J. Metaproteomic analysis using the Galaxy framework. *Proteomics* **2015**, *15*, 3553−3565.

(17) Cheng, K.; Ning, Z.; Zhang, X.; Li, L.; Liao, B.; Mayne, J.; Stintzi, A.; Figeys, D. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* **2017**, *5*, 157.

(18) Beyter, D.; Lin, M. S.; Yu, Y.; Pieper, R.; Bafna, V. ProteoStorm: An Ultrafast Metaproteomics Database Search Framework. *Cell Syst.* **2018**, *7*, 463−467.e6.

(19) AbSciex *Understanding the Pro Group^{TM} Algorithm*; https://sciex.com/Documents/manuals/proteinPilot-ProGroup-Algorithm.pdf, Accessed on 2019-06-12.

(20) Schneider, T.; Schmid, E.; de Castro, J. V., Jr.; Cardinale, M.; Eberl, L.; Grube, M.; Berg, G.; Riedel, K. Structure and function of the symbiosis partners of the lung lichen (Lobaria pulmonariaL. Hoffm.) analyzed by metaproteomics. *Proteomics* **2011**, *11*, 2752−2756.

(21) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehtevä, M.; Reichl, U.; Martens, L.; Rapp, E. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.* **2015**, *14*, 1557−1565.

(22) Langella, O.; Valot, B.; Balliau, T.; Blein-Nicolas, M.; Bonhomme, L.; Zivy, M. X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *J. Proteome Res.* **2017**, *16*, 494−503.

(23) Heyer, R.; Schallert, K.; Büdel, A.; Zoun, R.; Dorl, S.; Behne, A.; Kohrs, F.; Püttker, S.; Siewert, C.; Muth, T.; Saake, G.; Reichl, U.; Benndorf, D. A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 h Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer. *Front. Microbiol.* **2019**, *10*, 1883.

(24) Werner, J.; Géron, A.; Kerssemakers, J.; Matallana-Surget, S. mPies: a novel metaproteomics tool for the creation of relevant protein databases and automatized protein annotation. *Biol. Direct* **2019**, *14*, 21.

(25) Mesuere, B.; Willems, T.; Van der Jeugt, F.; Devreese, B.; Vandamme, P.; Dawyndt, P. Unipept web services for metaproteomics analysis. *Bioinformatics* **2016**, *32*, 1746−1748.

(26) Riffle, M.; May, D. H.; Timmins-Schiffman, E.; Mikan, M. P.; Jaschob, D.; Noble, W. S.; Nunn, B. L. MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. *Proteomes* **2018**, *6*, 2.

(27) Sajulga, R.; Easterly, C.; Riffle, M.; Mesuere, B.; Muth, T.; Mehta, S.; Kumar, P.; Johnson, J.; Gruening, B. A.; Schiebenhoefer, H.; Kolmeder, C. A.; Fuchs, S.; Nunn, B. L.; Rudney, J.; Griffin, T. J.; Jagtap, P. D. Survey of metaproteomics software tools for functional microbiome analysis. *PLoS One* **2020**, *15*, No. e0241503.

(28) Juste, C.; Kreil, D. P.; Beauvallet, C.; Guillot, A.; Vaca, S.; Carapito, C.; Mondot, S.; Sykacek, P.; Sokol, H.; Blon, F.; Lepercq, P.; Levenez, F.; Valot, B.; Carré, W.; Loux, V.; Pons, N.; David, O.;

https://dx.doi.org/10.1021/acs.jproteome.0c00669

Schaeffer, B.; Lepage, P.; Martin, P.; Monnet, V.; Seksik, P.; Beaugerie, L.; Ehrlich, S. D.; Gibrat, J. F.; Van Dorsselaer, A.; Doré, J. Bacterial protein signals are associated with Crohn's disease. *Gut* **2014**, *63*, 1566−1577.

(29) Cotillard, A.; Kennedy, S. P.; Kong, L. C.; Prifti, E.; Pons, N.; Le Chatelier, E.; Almeida, M.; Quinquis, B.; Levenez, F.; Galleron, N.; Gougis, S.; Rizkalla, S.; Batto, J.-M.; Renault, P.; ANR MicroObes consortium; Doré, J.; Zucker, J.-D.; Clément, K.; Ehrlich, S. D. Dietary intervention impact on gut microbial gene richness. *Nature* **2013**, *500*, 585−588.

(30) Kong, L. C.; Wuillemin, P.-H.; Bastard, J.-P.; Sokolovska, N.; Gougis, S.; Fellahi, S.; Darakhshan, F.; Bonnefont-Rousselot, D.; Bittar, R.; Doré, J.; Zucker, J.-D.; Clément, K.; Rizkalla, S. Insulin resistance and inflammation predict kinetic body weight changes in response to dietary weight loss and maintenance in overweight and obese subjects by using a Bayesian network approach. *Am. J. Clin. Nutr.* **2013**, *98*, 1385−1394.

(31) Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; Moritz, R. L.; Carver, J. J.; Wang, M.; Ishihama, Y.; Bandeira, N.; Hermjakob, H.; Vizcaíno, J. A. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **2017**, *45*, D1100−D1106.

(32) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.

(33) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. Target-decoy approach and false discovery rate; when things may go wrong. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111−1120.

(34) Saito, M. A.; Bertrand, E. M.; Duffy, M. E.; Gaylord, D. A.; Held, N. A.; Hervey, W. J., IV; Hettich, R. L.; Jagtap, P. D.; Janech, M. G.; Kinkade, D. B.; Leary, D. H.; McIlvin, M. R.; Moore, E. K.; Morris, R. M.; Neely, B. A.; Nunn, B. L.; Saunders, J. K.; Shepherd, A. I.; Symmonds, N. I.; Walsh, D. A. Progress and Challenges in Ocean Metaproteomics and Proposed Best Practices for Data Sharing. *J. Proteome Res.* **2019**, *18*, 1461−1476.

(35) Blein-Nicolas, M.; Xu, H.; de Vienne, D.; Giraud, C.; Huet, S.; Zivy, M. Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. *Proteomics* **2012**, *12*, 2797−2801.

(36) Buchfink, B.; Xie, C.; Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59−60.

(37) Hsieh, T. C.; Ma, K. H.; Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **2016**, *7*, 1451−1456.

(38) R Core Team *R: A Language and Environment forStatistical Computing*; https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf.

(39) Bassignani, A. Metaproteomics bioanalysis to study functionnalities of gut microbiota in large cohorts. PHD thesis, Sorbonne University, Paris **2919**, chapter 2, pp. 52. https://tel.archives-ouvertes.fr/tel-02871891.

(40) Benson, A. K.; Kelly, S. A.; Legge, R.; Ma, F.; Low, S. J.; Kim, J.; Zhang, M.; Oh, P. L.; Nehrenberg, D.; Hua, K.; Kachman, S. D.; Moriyama, E. N.; Walter, J.; Peterson, D. A.; Pomp, D. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18933−18938.

(41) Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; Mende, D. R.; Li, J.; Xu, J.; Li, S.; Li, D.; Cao, J.; Wang, B.; Liang, H.; Zheng, H.; Xie, Y.; Tap, J.; Lepage, P.; Bertalan, M.; Batto, J.-M.; Hansen, T.; Le Paslier, D.; Linneberg, A.; Nielsen, H. B.; Pelletier, E.; Renault, P.; Sicheritz-Ponten, T.; Turner, K.; Zhu, H.; Yu, C.; Li, S.; Jian, M.; Zhou, Y.; Li, Y.; Zhang, X.; Li, S.; Qin, N.; Yang, H.; Wang, J.; Brunak, S.; Doré, J.; Guarner, F.; Kristiansen, K.; Pedersen, O.; Parkhill, J.; Weissenbach, J.; MetaHIT Consortium; Bork, P.; Ehrlich, S. D.; Wang, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010**, *464*, 59−65.

(42) Manceau, H.; Chicha-Cattoir, V.; Puy, H.; Peoc'h, K. Fecal calprotectin in inflammatory bowel diseases: update and perspectives. *Clin. Chem. Lab. Med.* **2017**, *55*, 474−483.

(43) Palm, N. W.; de Zoete, M. R.; Cullen, T. W.; Barry, N. A.; Stefanowski, J.; Hao, L.; Degnan, P. H.; Hu, J.; Peter, I.; Zhang, W.; Ruggiero, E.; Cho, J. H.; Goodman, A. L.; Flavell, R. A. Immunoglobulin A Coating Identifies Colitogenic Bacteria in Inflammatory Bowel Disease. *Cell* **2014**, *158*, 1000−1010.

(44) Fadlallah, J.; Sterlin, D.; Fieschi, C.; Parizot, C.; Dorgham, K.; El Kafsi, H.; Autaa, G.; Ghillani-Dalbin, P.; Juste, C.; Lepage, P.; Malphettes, M.; Galicier, L.; Boutboul, D.; Clément, K.; André, S.; Marquet, F.; Tresallet, C.; Mathian, A.; Miyara, M.; Oksenhendler, E.; Amoura, Z.; Yssel, H.; Larsen, M.; Gorochov, G. Synergistic convergence of microbiota-specific systemic IgG and secretory IgA. *J Allergy Clin Immunol.* **2019**, *143*, 1575−1585.e4.

(45) Huson, D. H.; Beier, S.; Flade, I.; Górska, A.; El-Hadidi, M.; Mitra, S.; Ruscheweyh, H.-J.; Tappu, R. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **2016**, *12*, No. e1004957.

(46) Shoaie, S.; Ghaffari, P.; Kovatcheva-Datchary, P.; Mardinoglu, A.; Sen, P.; Pujos-Guillot, E.; de Wouters, T.; Juste, C.; Rizkalla, S.; Chilloux, J.; Hoyles, L.; Nicholson, J. K.; MICRO-Obes Consortium; Dore, J.; Dumas, M. E.; Clement, K.; Bäckhed, F.; Nielsen, J. Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. *Cell Metab.* **2015**, *22*, 320−331.

(47) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2010**, *9*, 761−776.

(48) Saba, J. *ASMS 2016 - Day 2 - DDA : Still the Gold Standard in Label-Free Quantitation*; https://analyteguru.com/asms-2016-day-2-dda-still-the-gold-standard-in-label-free-quantitation/.