A Meta-proteogenomic Approach to Peptide Identification Incorporating Assembly Uncertainty and Genomic Variation

Authors

Sujun Li, Haixu Tang, and Yuzhen Ye

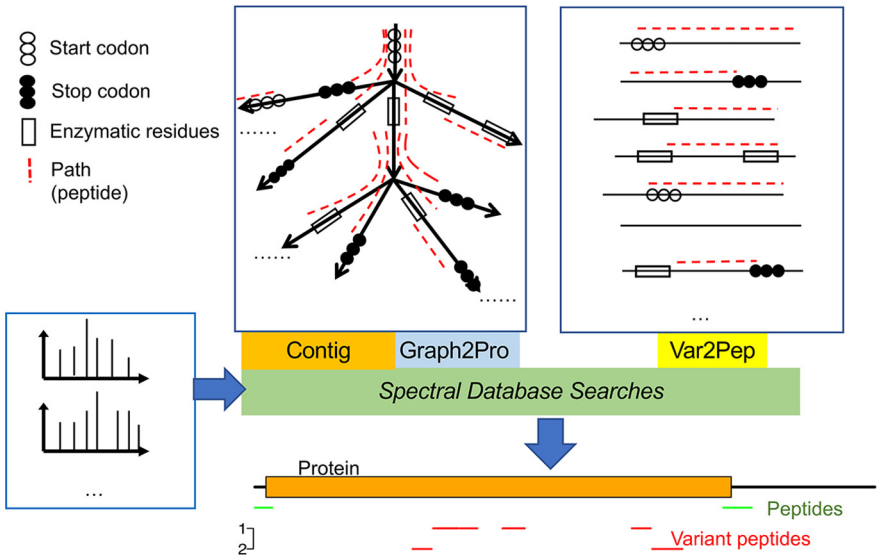
Correspondence

yye@indiana.edu

In Brief

A meta-proteogenomic approach analysis pipeline is developed for peptide identification from meta-proteomic MS/MS data, using matching metagenomic and/or metatranscriptomic sequencing data as the reference. The results reported reveal the importance of using metagenome assembly uncertainties and genomic variations remaining in the sequencing reads for meta-proteomic data characterization.

Graphical Abstract



Highlights

- A novel meta-proteogenomic analysis pipeline integrating Graph2Pro and Var2Pep approaches.
- Metaproteomic support of proteins with unknown functions.
- Improved functional profiling of microbiomes using variant peptides.

Li et al., 2019, Molecular & Cellular Proteomics *18*, S183–S192

https://doi.org/10.1074/mcp.TIR118.001233

A Meta-proteogenomic Approach to Peptide Identification Incorporating Assembly Uncertainty and Genomic Variation*[S]

Sujun Li, Haixu Tang, and Yuzhen Ye‡

Matching metagenomic and/or metatranscriptomic data, currently often under-used, can be useful reference for metaproteomic tandem mass spectra (MS/MS) data analysis. Here we developed a software pipeline for identification of peptides and proteins from metaproteomic MS/MS data using proteins derived from matching metagenomic (and metatranscriptomic) data as the search database, based on two novel approaches Graph2Pro (published) and Var2Pep (new). Graph2Pro retains and uses uncertainties of metagenome assembly for reference-based MS/MS data analysis. Var2Pep considers the variations found in metagenomic/metatranscriptomic sequencing reads that are not retained in the assemblies (contigs). The new software pipeline provides one stop application of both tools, and it supports the use of metagenome assembly from commonly used assemblers including MegaHit and metaSPAdes. When tested on two collections of multi-omic microbiome data sets, our pipeline significantly improved the identification rate of the metaproteomic MS/MS spectra by about two folds, comparing to conventional contig- or read-based approaches (the Var2Pep alone identified 5.6% to 24.1% more unique peptides, depending on the data set). We also showed that identified variant peptides are important for functional profiling of microbiomes. All results suggested that it is important to take into consideration of the assembly uncertainties and genomic variants to facilitate metaproteomic MS/MS data interpretation. ***Molecular & Cellular Proteomics 18: S183–S192, 2019. DOI: 10.1074/mcp.TIR118.001233.***

Microbiome research has been applied to various studies of microbial organisms associated with different ecosystems, habitats and hosts (1–8). Recent studies have further shown the broader impacts of microbiota on human health and diseases, including the influence of microbiota on the efficacy of cancer immunotherapy (9). Researchers started to investigate the therapeutic applications of microbiome, which are very promising as demonstrated in recent case studies: Zhu and colleagues edited gut microbiota on ameliorate colitis (10), and in another case, engineered commensal microbiomes were used for diet-mediated colorectal cancer chemoprevention (11). In microbiome research, metagenomic and metatranscriptomic (12, 13) data often tell about the taxonomic distribution and potential functions, whereas metaproteomic data provides more direct information about the functionality of microbial communities (14, 15, 16, 17). In a recent study of ocean microbiome (18), single-cell genomics and community metagenomics revealed that Nitrospinae are the most abundant and globally distributed nitrite-oxidizing bacteria in the ocean, whereas metaproteomic and metatranscriptomic analyses suggest that nitrite oxidation is the main pathway of energy production in Nitrospinae. And in another work, Kleiner and colleagues (19) showed that metaproteomics could be used to assess species biomass contributions in microbial communities, which is less prone to the biases in sequencing-based methods. Studies have shown the impact of microbiomes on human health and diseases. Metaproteomics provides a new opportunity of studying the functionality of microbial communities.

Although metaproteomics revealed complementary information to metagenomic and metatranscriptomic studies, it often applies conventional proteomics techniques using mass spectrometry (MS). Metaproteomic data analysis is challenging (20), which has motivated the development of recent algorithms and tools including MetaLab (21), cascaded search (22), Unipept (23), two-steps search (24), MetaProteomeAnalyzer (25) and ProteoStorm (26). Successful proteomic database search largely relies on the completeness and specificity of the target protein database. There are two broad categories of approaches for metaproteomic database search: approaches relying on a *generic* collection of reference proteins (for example, the human/mouse gut bacterial genes (27, 21)), and approaches that use *specific* proteins/peptides inferred from matching metagenomic and metatranscriptomic data sets of the same microbial community. Different from typical proteomics data analysis, peptide/protein identification from metaproteomics data generally lacks a good reference database of proteins for database searches,

From the School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN
Received November 21, 2018, and in revised form, April 25, 2019
Published, MCP Papers in Press, May 29, 2019, DOI 10.1074/mcp.TIR118.001233

except for well-studied microbiomes. On the other hand, although matching metagenomic or metatranscriptomic data sets can provide more adequate reference database, they are often under-used for metaproteomic data analysis. Microbiomes are normally composed of many species, many of those species are closely related, as demonstrated in studies of microbial communities empowered by long sequencing reads (28, 29). Despite of the recent advances in the development of assembly algorithms for metagenomes, it remains a challenge to assemble metagenomes from shotgun metagenomic sequences because of the nature of microbial communities: a large number of species co-exist in a sample; these species are of various abundances; and many species are closely related containing similar genome sequences with variations so that assemblers won't be able to untangle them. As a result, metagenome assemblies often provided limited references for peptide/protein identification from metaproteomic data sets.

Approaches have been proposed to address the above-mentioned limitation of using metagenome assemblies as reference for peptide/protein identification from MS/MS data. One successful effort is from May and colleagues (30), who developed a method to derive *metapeptides* (short amino acid sequences that may be encoded by multiple organisms) from shotgun metagenomic sequencing reads, and then use the metapeptides as the reference for peptide identification from metaproteomic MS/MS data. They showed that by constructing site-specific metapeptide databases, they were able to detect more than one and a half times as many peptides as by searching against predicted genes from an assembled metagenome and roughly three times as many peptides as by searching against the NCBI environmental proteome database. Also some benchmarking experiments have been optimized to explore better experimental and computational strategies for metaproteomics (31, 32) as well. We have also recognized the limitation of conventional approach of using assembled metagenomes (*e.g.* using only the contigs) for peptide/protein identification from MS/MS data and developed a graph-centric approach (Graph2Pro) (33) that uses the assembly graph to take into consideration of the assembly uncertainties. Our approach achieved significant improvement of peptide/protein identification from MS/MS data for microbiome research.

Comparing to (meta)genomic sequencing, metaproteomics (and proteomics in general) typically has relatively limited throughput. It is therefore important to develop methods that can extensively use the sequencing reads (in addition to assemblies) for peptide/protein identification from MS/MS data (sequencing reads from low abundant species are unlikely to be assembled into contigs). Further, microbial communities often contain closely related species and strains with genomic variations, many of which are of low abundances (28, 29). Short sequencing reads containing variations are often "ignored" by the assemblers, so as a result, our graph-based approach that uses assembly graphs for MS/MS identification is unable to use these reads for MS/MS identification. The open database search approach (34) in principle can be applied to detect variations, which however will be impractical for metaproteomic identification because of the unknown and extremely large search space. The goal of our new variant-aware approach is to incorporate the potential peptides encoded by the short sequencing reads that are not even assembled into the assembly graph, to improve the identification from MS/MS data. Instead of using variant callers (such as MIDAS (35) and metaSNV (36)) to detect potential variants from metagenomic and/or metatranscriptomic sequencing data (8), our variant-aware approach retains all potential variants containing peptides, and use them as reference for spectral database search. The rationale is that the variants are often of low abundance, so the typical variant detection tools that rely on sequencing coverage to distinguish variants from sequencing errors may not work. If a potential variant is supported by only few reads, but nevertheless is supported by the MS/MS data, it is a confident variant. We also note that, different from the previous method (30), which includes all potential short peptides from short sequencing reads, our variant-aware approach only includes the peptides that are similar to the proteins that have already been identified from MS/MS spectra. We showed that our approach effectively reduced the search space for MS/MS spectral database searching, and identified more peptides from metaproteomic MS/MS data.

EXPERIMENTAL PROCEDURES

Overview—We developed a pipeline for peptide/protein identification from metaproteomic mass spectrometry data, to optimize the use of matching genomic/transcriptomic data of the same microbial community. The pipeline is empowered by two novel approaches we developed, Graph2Pro (33) and Var2Pep proposed here. Graph2Pro uses the uncertainties of metagenome assembly captured in assembly graphs. Var2Pep uses genomic variations that are not retained in metagenome assembly. Specially, our pipeline uses two search databases for mass spectral identification: (1) the Graph2Pro database, which contains putative proteins and peptides inferred from the assembly graph (including the edges, *i.e.* the contigs, and the branching structures); and (2) the Var2Pep database, which contains putative peptides predicted from sequencing reads that potentially contain genomic variants as predicted by our new approach (see Fig. 1).

Rationale of the Variant-aware Approach (Var2Pep)—Our variant-aware approach Var2Pep aims to retain peptides that can be predicted from unassembled short sequencing reads in the search database to improve MS/MS spectral identification. Unassembled short reads containing sequence variations can be mapped to assembled contigs as long as the contigs contain similar sequences (so the variant-bearing reads can be "used" for downstream analysis based on genomic sequences). However, for MS/MS analysis, a peptide with only a single amino acid difference (caused by a single nucleotide variation) may be overlooked if the peptide was not included in the target database, because two peptides with a single amino acid difference may result in very different mass spectra.

In principle, keeping all putative peptides in short sequencing reads in the search database for MS/MS analysis will address the above-mentioned peptide variant problem. However, by doing so, we ex-

pand the search database tremendously, resulting in not only the increase of searching time, but also the decrease of identification performance because of the extremely large database containing unbalanced target and decoy entries. By contrast, Var2Pep only considers reads-derived putative peptides that share significant sequence similarity with the proteins/peptides in the Graph2Pro database.

Var2Pep and Its Integration with Graph2Pro—As illustrated in Fig. 1, our pipeline for protein/peptide identification from metaproteomic data integrates the Graph2Pro approach (33) and the new Var2Pep approach. The Var2Pep approach is composed of several steps, as shown in Fig. 1. The first step is to extract short sequencing reads that cannot be mapped to contigs, and the reads that can be mapped to contigs but with mismatches, based on bowtie2 (37) mapping results. In the second step, putative peptides are predicted from the reads with mismatches and unmapped reads using FragGeneScan (38). The third step is to prepare a new search database (called Var2Pep) that contains the putative peptides sharing high similarity (*e.g.* $\geq$ 70% identity by default, based on RAPSearch2 (39) search results) with proteins/peptides in the Graph2Pro database. The 70% sequence identity is chosen, considering that it allows up to 3 substitutions in identified variant peptides (the average length of identified peptide is 10). This identity cutoff works well in practice, but users can customize this identity threshold when they apply our pipeline.

The MS/MS database searching is applied to Var2Pep database, controlled by false discovery rate (FDR $\leq$ 1%). When combining the Graph2Pro and Var2Pep search results to derive a non-redundant collection of identified peptides, we chose the peptide identification with the higher score for a spectrum when the two searches result in different peptides for the same spectrum.

Data Sets—We tested our approaches using two publicly available collections of meta-omics data sets. The first collection (called the *ocean* data set) contains meta-omics data sets derived from ocean water samples (30), which were collected from Bering Strait (BSt) chlorphyll maximum layer and from the more norther Chukchi Sea (CS) bottom water. This collection contains matching metagenomic and metaproteomic data. The LC-MS/MS spectra from triplicate acquisitions of peptides from the BSt and CS samples are designated as BSt 45–47, and CS 51–53. We downloaded the raw sequencing data and metaproteomic data from the following website: http://noble.gs.washington.edu/proj/metapeptide/. For the meta-proteogenomic analysis, the same metagenomic data set (BSt) was used for analyzing the three metaproteomics data sets (BSt45, BSt46 and BSt47), and similarly the same metagenomic data set (CS) was used for analyzing metaproteomic data sets CS51-53.

The second collection (called the *wastewater* data set) was derived from a multi-omic study of oleaginous mixed microbial communities (OMMC) sampled from an anoxic biological wastewater treatment tank (40). We used the metagenomic, metatranscriptomic and metaproteomic data acquired from the community at three sample dates, denoted as SD3 (January 25th, 2011), SD6 (October 12th, 2011) and SD7 (January 11th, 2012), respectively. We downloaded the metagenomic (MG) and metatranscriptomic (MT) data sets from the SRA website (SD3-MG: SRR1046369; SD3-MT: SRR1046681; SD6-MG: SRR1544596; SD6-MT:SRR1544599; SD7-MG: SRR1611146; and SD8-MT: SRR1611147). We downloaded the spectral data from PeptideAtlas (41, 42) : SD3 (ID: PASS00359), SD6 (PASS00577) and SD7 (PASS00578).

Metagenomic/Metatranscriptomic Assembly—Raw reads were preprocessed using Trimmomatic (version 0.32) (43) and only reads of at least 80 bps were used in downstream analyses. We used SOAPdenovo2 (44) and MegaHit (45) for the assembly of metagenomic and/or metatranscriptomic sequences. The default parameters of SOAPdenovo2, and the k-mer size of 31 were used. For MegaHit, we used -k-list 21,29,39,59,79,99 and default setting for other options; assembly graphs from the last k-mer size (*i.e.* 99) were used as inputs to our pipeline. Considering metaSPAdes (46) is becoming a popular assembler for metagenomes (a recent study showed that metaSPAdes produced longer scaffold, but had the longest run times (47)), we extended our Graph2Pro such that it can exploit assembly graphs produced by MetaSPAdes.

TABLE I
Summary of the read-based MS/MS search databases for the ocean data sets

	BSt	CS
Number of peptides in metapeptides database (obtained from 30)	15,911,893	19,194,693
Number of mismatched/unmatched reads	148,502,311	221,702,454
Number of peptides in Var2Pep database	2,702,655	4,891,690

Note: Bst, Bering strait; CS, Chukchi sea.

Peptide/Protein Identification from MS/MS Data Using Database Search—We used the MS-GF+ search engine (Version v10089) (48) in our pipeline. We used the following parameters for the MS-GF+ database searching: (1) instrument type was set as high-resolution LTQ; (2) precursor mass tolerance was set as 15 ppm; (3) Isotope error range was selected as −1,2; (4) modification was set to include at most 3 modification on the sequence, including variable oxidation on Methionine and fixed carboamidomethy on Cysteine; and (5) maximum charge was set as 7 and minimum charge was set as 1, (6) Number of tolerable termini is semi-tryptic. The false discovery rate (FDR) was estimated by using a target-decoy search approach (49), in which the decoy proteins were generated by reversing the target sequences. Our pipeline outputs two sets of identification results after imposing FDR at the spectrum level and at the peptide level, respectively.

RESULTS

We tested our integrated pipeline for peptide/protein identification using the ocean water and wastewater multi-omics data sets. We compared our approaches with *contig-based* (in which proteins predicted from contigs are used as the reference), *read-based* approach (using metapeptides derived from reads as the reference), and a *simple* combination of read- and contig-based approaches (in which peptides were derived using the read- and contig-based approaches separately, and then merged as the final results). Our results showed that by considering variations and uncertainties of assembly, we can significantly improve the MS/MS spectral identification rate, and therefore identify more peptides (and proteins) for downstream analyses including functional profiling. The Var2Pep approach alone identified 5.6% to 24.1% more unique peptides, and our pipeline integrating the Graph2Pro approach and the Var2Pep approach improved the overall identification rate of the metaproteomic MS/MS spectra by about two folds, as compared with conventional contig- or read-based approaches.

Performance of the Meta-proteogenomic Approach on the Seawater Microbiome Data Set—We first tested the pipeline for peptide and protein identification from metaproteomic MS/MS data using the *ocean* data set. Table I lists the size of

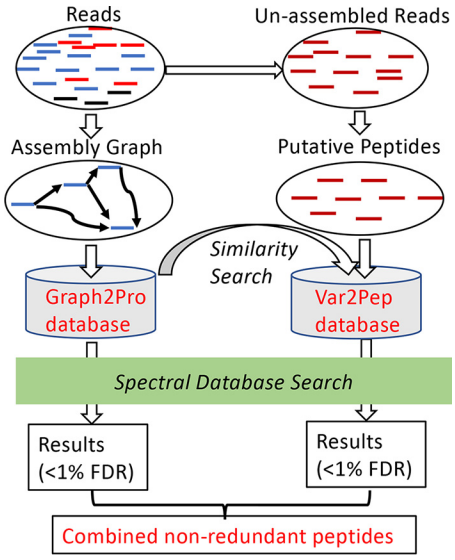


FIG. 1. **An overview of the pipeline for peptide/protein identification from metaproteomic MS/MS data.** The pipeline integrate two approaches: the Graph2Pro approach that uses assembly uncertainties and the new variant-aware approach Var2Pep.

the CS and BSt metapeptide databases, as well as the size of Var2Pep databases for comparison. Apparently, the Var2Pep database contains much fewer reads-derived peptides for MS/MS identification. For instance, for the BSt data set, there are about 15.9 million metapeptides, whereas the Var2Pep database only contains about 2.7 million peptides derived from 148 million mismatched or unmapped reads.

Table II summarizes the peptide identification for selected data sets (BSt45 and CS51; see results for all data sets in the supplementary Data File S1) using different approaches. Fig. 2 shows the comparison results for all ocean water data sets. Briefly, our approach approximately doubled the number of identified peptides as compared with the contig-based and read-based approaches, when the same FDR ($\leq$ 0.01, at spectrum level) estimated using the target-decoy search approach was applied. When FDR $\leq$ 0.01 at peptide level was applied, all methods unsurprisingly identified fewer peptide, however, we observed similar trends of improvements by Graph2Pro and Var2Pep: supplemental Fig. S1 shows the comparison of the different approaches on all ocean water data sets, using FDR $\leq$ 0.01 at peptide level, and supplemental Fig. S2 shows the side-by-side comparison of the results by using FDR at spectrum or peptide level for the CS51 data set, respectively. It was shown in reference (30) that read-based approach using short amino acid sequences predicted from shotgun sequencing reads as the search database identified significantly more peptides as compared with the contig-based approach using genes predicted from contigs assembled by SOAPdenovo version 1.06 (30)). Our results confirmed the observation that when SOAPdenovo2 was used as the assembler, contig-based approach resulted in worse MS/MS identification (see Table II) than the read-based approach. However, when MegaHit was used as the assembler, the trend reversed and the contig-based approach identified more peptides than the read-based approach (see Fig. 2), whereas in both cases adding Graph2Pro and Var2Pep steps further improved the peptide identification.

Our pipeline significantly outperformed the *simple* combination of contig-based and read-based approaches (red bars *versus* blue bars in Fig. 2). For instance, for the Bst45 data set, reads-based method identified 7795 (8.6% of total) spectra, and the contig-based (using MegaHit) approach identified 8631 (9.5%) of total spectra. Combining contig-based search and reads-based search resulted in the characterization of a total of 16,426 spectra/3813 peptides. By comparison, our pipeline reported the identification of comparable number of spectra (15,913), but many more unique peptides (6240).

Comparison of Our Approach with Grouped and Cascaded Searches—Our pipeline uses separate database searches on the Graph2Pro and Var2Pep databases, and then combines the search results (we called it *separate* search for comparison purpose). We compared this approach with the other two approaches: a single spectral search on a grouped database combining both databases (called *grouped* approach), and cascaded search using only rejected spectra from the spectral search against the Graph2Pro database for the subsequent search against Var2Pep database (*i.e.* the *cascaded* approach, similar to the approach developed by Kertesz-Farkas and colleagues (22)). Table III summarizes the comparison of separate and cascaded searches at the same FDR level. For example, for CS51, our pipeline based on separate search against the Var2Pep database identified additional 2073 spectra (917 unique peptides); by comparison, the cascaded search against the Var2Pep database identified slightly fewer 1598 spectra (629 unique peptides). Notably, this result appears different from the previous report that *cascaded search* identified more spectra and unique peptides (22) than the separate search. One possible reason is that removing some high qualify spectra in the cascaded searches may alter the overall score distribution, and thus the estimation of false discovery rate for different searching results. Our approach also outperformed (marginally) the search using combined database. For example, for the CS51 data set, combined database search only identified a total of 20,046 spectra and 7662 unique peptides; by contrast, our approach based on separate searches identified 31,145 spectra and 12,380 peptides, and the cascaded search identified 30,670 spectra and 12,092 peptides from the same spectra data set. Based on our tests, we recommend using either separate or cascaded search, but not using the combined database for spectrum search. Our pipeline supports both separate search (default) and cascaded-search, and the results reported below were based on the separate searches.

Performance of the Meta-proteogenomic Approach on the Wastewater Microbiome Data Set—Next we tested our approach using the *wastewater* data sets. Fig. 3 illustrates the

TABLE II
Summary of peptide identification from MS/MS spectra for selected ocean datasets

	BSt45 (90,072 spectra)		CS51 (100,588 spectra)	
	PSMs (%)	Unique peptide	PSMs (%)	Unique peptide
Reads based	7795 (8.6%)	2958	15,167 (15.1%)	5652
Contigs based (S*)	1892 (2.10%)	817	6526 (6.49%)	2442
Graph2Pro (S*)	12,728 (14.1%)	4542	26,576 (26.4%)	9932
Contigs based (M*)	8631 (9.6%)	3367	17,427 (17.3%)	6388
Graph2Pro (M*)	15,172 (16.8%)	5857	29,072 (28.9%)	11,463
Graph2Pro (M*) + Var2Pep	15,913 (17.7%)	6240	31,145 (31.0%)	12,380

Note: S* represents SOAPdenovo2, M* represents MegaHit, PSM stands for peptide spectrum match. Bst stands for Bering strait, and CS stands for Chukchi sea. Graph2Pro (S*) and Graph2Pro (M*) represent using assembly graph from SOAPdenovo2 (S*) and MegaHit (M*) as the reference in Graph2Pro, respectively. In all cases, FDR (false discovery rate) was estimated using a target-decoy search approach, and a cutoff of 1% at spectrum level was applied. This table only shows the results for two datasets. See Fig. 2 and supplementary Data File S1 for results of all data sets.

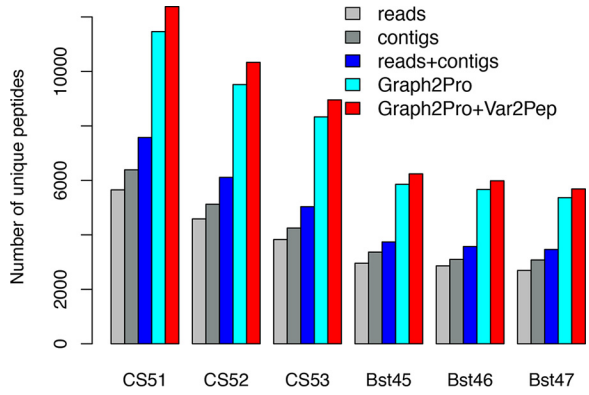


FIG. 2. **Comparison of peptide identification by the different approaches on the ocean data sets.** The barplot shows the total number of unique peptides identified from six ocean metaproteomic MS/MS data sets, by different approaches.

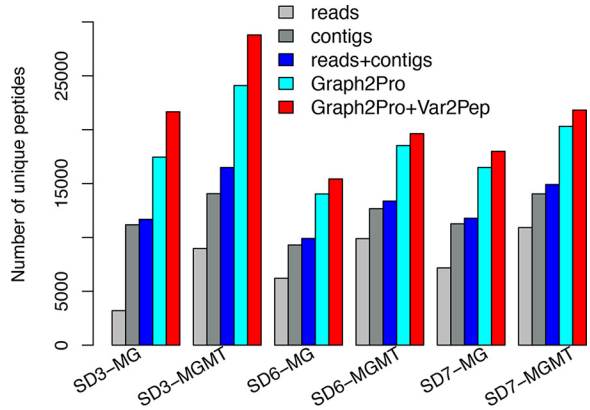


FIG. 3. **Comparison of peptide identification results by the different approaches on the wastewater data sets.** The barplot shows the total number of unique peptides identified from three wastewater samples (SD3, SD6, and SD7), using either matching metagenomic data alone (MG), or both metagenomic and metaproteomic data (MGMT) as the reference.

TABLE III
Comparison of additional spectra and unique peptides identified by searching against the Var2Pep database using the separate search approach (our approach) versus cascaded search

	Separate (our approach)		Cascaded	
	PSMs	Peptides	PSMs	Peptides
CS51	2073	917	1598	629
CS52	2282	816	1692	533
CS53	1969	625	1466	392
Bst45	741	383	495	221
Bst46	651	318	390	173
Bst47	640	322	418	192

Note: PSM stands for peptide spectrum match. Bst, Bering strait; CS, Chukchi sea.

results of peptide identification using different approaches on this data set using FDR $\leq$ 0.01 at spectrum level (see supplementary Data File S2 for details; also see supplemental Fig. S3 for the comparison of the different approaches using FDR $\leq$ 0.01 at peptide level, and supplemental Fig. S4 for side-by-side comparison of the results by using FDR at spectrum *versus* peptide level for the SD3-MG data set). For comparison purposes, for reads-based approach, we used the Sixgill software (https://github.com/dhmay/sixgill) (30) to predict putative peptides (*i.e.* metapeptides) from reads for the reads-based approach. Here are the numbers of metapeptides predicted from the different data sets: 266,430 (SD3-MG), 1,640,859 (SD3-MGMT), 2,614,012 (SD6-MG), 4,782,170 (SD6-MGMT), 1,800,892 (SD7-MG), and 3,977,520 (SD7-MGMT) (sequences of the metapeptides can be found at our supplementary website at http://omics.soic.indiana.edu/MP). In all cases, the contig-based approach achieved significantly better performance than the read-based approach, and our approach integrating Graph2Pro and Var2pep significantly outperformed other approaches.

As shown in Fig. 3, significant improvement of peptide identification was achieved by using both metatranscriptomic and metagenomic data sets as the reference as compared with using only metagenomic sequencing data (see supplementary Data File S2 for details). This result is expected, because metatranscriptomic data may provide sequence information for otherwise low-abundant species that may have been missed by metagenomic sequencing. We note that among the wastewater data sets, the Var2Pep approach resulted in the most significant improvements of the peptide

TABLE IV
Comparison of the performance using metagenome assembly from MegaHit and MetaSpades as the reference on SD3-MG dataset

	MegaHit	MetaSpades
Number of contigs	111,739	446,773
Total bases (MB)	78	166
Contig only (PSM/peptide)	43650 (11173)	40161 (10337)
Graph2Pro (PSM/peptide)	63178 (17452)	57760 (17651)
Graph2Pro + Var2pep (PSM/peptide)	74317 (21662)	71366 (22043)

identification for the SD3 data set, with an additional 24.1% and 19.5% more identified unique peptides when using matched metagenomic data set alone (SD3-MG), and both the metagenomic and metatranscriptomic data sets (SD3-MGMT).

Considering the increasing popularity of metagenome assembler MetaSPAdes, we also used the wastewater data sets to test if more contigs assembled from metagenome resulted in better identification of metaproteomics data. Table IV shows the comparison of the results for SD3-MG data set using assemblies from MegaHit and MetaSPAdes as the reference in our pipeline. Strikingly, MetaSPAdes produced about three times more contigs (more than doubled the total bases in the contigs) as compared with MegaHit. However, the increasing of contigs did not lead to more peptide identifications.

Abundant Known (and Unknown) Proteins Revealed by Metaproteomics—We examined proteins that were identified using our pipeline. For example, for CS51 data set, its first most abundant protein was supported by 270 spectra (0.8% of the total 31,145 identified spectra for this data set). A total of 663 proteins or protein fragments (of minimal length of 60 aa) were each supported by at least 10 spectra (see supplementary Data File S3 for details). Functional annotation of these proteins by myRAST (http://blog.theseed.org/downloads/myRAST-Intel.dmg) (50) revealed 90 ribosomal proteins, 61 heat shock proteins (HSP60), 59 elongation factors (GTP_EFTU), 38 chaperone proteins (DnaK), 30 glutamate aspartate periplasmic binding protein precursors, 26 branched-chain amino acid ABC transporters, 10 TonB-dependent receptors, among other functions (see supplementary Data File S3 for details of the annotations). Our result is consistent with the fact that many of these functions (*e.g.* ribosomal proteins) are typically the most abundant and highly conserved in prokaryotic genomes.

Among the CS51's 663 proteins each with support of at least 10 spectra, a significant fraction (93, 14%) have no functional annotation based on myRAST annotation, including a protein (ID: NODE_17410_length_722_cov_188.0000_ID_34819_1_456_+) that has the third most spectra support with 257 spectra and many others with strong spectral supports. See supplementary Data File S3 for the complete list of these highly expressed proteins that lack any functional annotations (and their sequences can be found at our supplementary website). We argue that these proteins are *most wanted* proteins that wait to be further studied experimentally for their functions. Fig. 4 illustrates two examples of proteins supported by a significant number of variant peptides (details of the variant peptides are available at our supplementary website). Protein366 was predicted from the assembly graph by Graph2Pro and was shown to be supported by a total of 27 peptides, including six variant peptides. The myRAST annotation did not reveal any significant hits for this protein. We also did sequence similarity search of this protein using ncbi blast (http://www.ncbi.nlm.nih.gov/blast) against the nr database (as of Nov 18th, 2018), which returned no hits either. Although no significant sequential similarity was detected between Protein366 and proteins in the public databases, a structural comparison (by FATCAT (51)) of this protein's predicted structure (using QUARK (52)) revealed that it contains a Cystatin-like fold composed of helical segments packed against coiled antiparallel beta-sheet. Protein2811 is another example, which was annotated as a putative peptide/opine/nickel uptake transporter. This protein has a total of 17 supporting peptides, among which nine are variant peptides.

For the wastewater data sets, an interesting example is a protein (Protein264) identified from SD6 data set. This protein has the most spectra support (2266 spectra), and the domain prediction shows that it contains three SLH (S-layer homology) domains and another larger domain that has unknown function (DUF2815). The presence of SLH domains in this protein suggest that it is likely a cell-surface protein, as studies have shown the utilization of SLH domains as vehicles for surface location of function proteins (53, 54); however, the precise function of this protein remains to be determined. Fig. 4 (bottom) shows the locations of the domains and the distribution of identified peptides and variant peptides supporting this protein.

Low Abundant Proteins with Supports Augmented by Variant Peptides—As shown above, metaproteomics is likely limited to revealing highly abundant proteins expressed from highly abundant bacterial species, so metaproteomics-based functional profiling of microbiomes will only provide a shallow glimpse into the functionality of the underlining microbial community. A hope to alleviate this limitation is to augment the identification of proteins using variant peptides. Here we use the wastewater data set (SD3-MG) as an example to illustrate this approach. We used BlastKOALA (https://www.kegg.jp/blastkoala/) (55) to annotate the functions of identified proteins from this data set. A total of 1401 KEGG families (*i.e.* the KO families) were identified using proteins with at least one peptide support; however, this number increased to 1938 if variant peptides were also considered (a 38% improvement). We acknowledge that the number of families is still small, because of the shallow functional profiling of microbiomes based on metaproteomics.

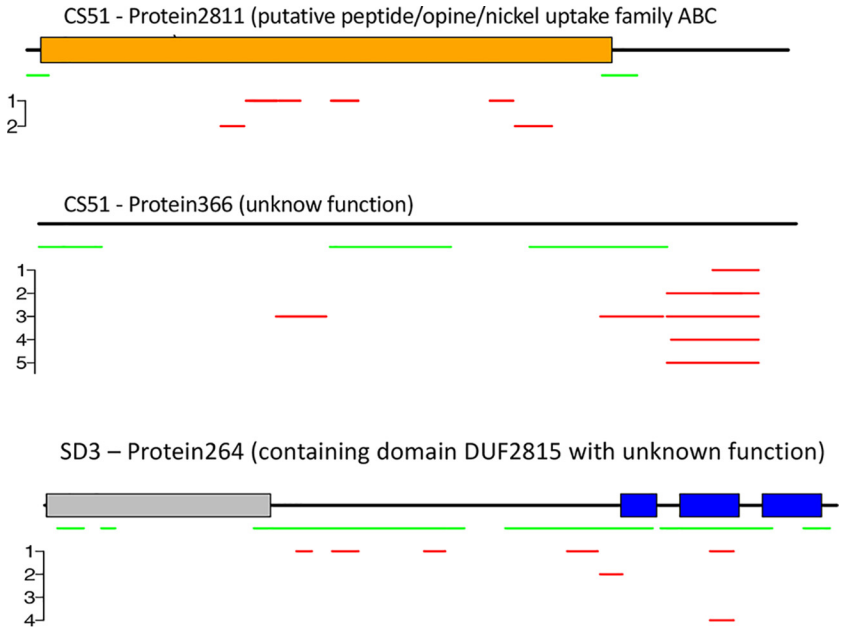


FIG. 4. **Selected examples of identified proteins and variants supported by metaproteomic MS/MS data (ocean water sample CS51 and wastewater sample SD3).** The plots depict the regions in the proteins that are supported by identified peptides. The black lines on the top represent the proteins, with boxes in different colors showing predicted PFAM domains in these proteins (orange box represents the SBP_bac_5 domain, gray box represents DUF2815, and the blue boxes represent the SLH domains. The green lines below the protein lines represent MS/MS supported peptides from each protein (no mismatches), and the red lines represent peptide variants that share similarity with the protein, with the number of mismatches indicated by the bar on the left.

We note that the variant peptides identified by our pipeline have both nucleotide level (from metagenomic and/or metatranscriptomic sequences) and mass spectral level supports. Here, we examined the types of substitutions found in the variant peptides. Not surprisingly, we found that most variations are the substitutions likely to be found in homologous proteins, with positive BLOSUM scores (BLOSUM matrices contain log-odd scores of all possible substitution pairs of amino acids, where positive scores represent favorable substitutions observed in homologous proteins such as Asp to Asn substitution and negative scores represent unfavorable substitutions such as Arg to Asp) (56). As shown in Fig. 5, the average BLOSUM (BLOSUM62) score between two amino acids is −2 (see "all" box), whereas the average BLOSUM score of substitutions in variant peptides is positive for Protein366. This result indicates that Var2Pep likely detected biologically meaningful variations.

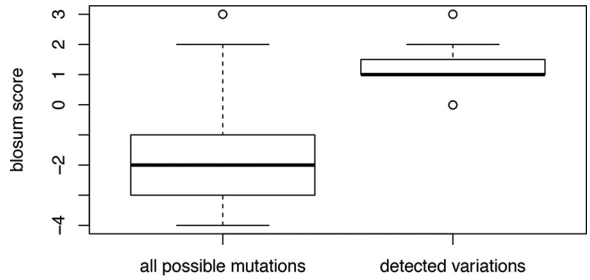


FIG. 5. **Var2Pep detected variations that are preferred substitutions among homologous proteins.** The *y* axis shows the BLOSUM score of pairs of amino acids. The "all" box is for all pairs of possible amino acids (excluding pairs of identical residues), whereas the other box is for variations found by Var2Pep in Protein366.

DISCUSSION AND CONCLUSIONS

We developed the Var2Pep algorithm, which combined with Graph2pro achieved drastic improvement of peptide identification from metaproteomics data. Var2Pep makes use of shotgun sequencing reads for MS/MS identification, and it reduces the search space imposed by the large number of sequencing reads by only keeping peptides that share similarity with proteins already identified by Graph2Pro. As shown in Table I, there were 15.9 million putative peptides in BSt data set from the mismatched/unmatched reads, and Var2Pep only contained 2.7 million peptides for MS/MS database search. This process significantly reduced the number of possible decoys in the database to permit the database searches, which significantly contribute to increasing the identification rate of metaproteomic MS/MS data. Similar ideas have been applied to improve metaproteomics data identification: for example MetaLab (21) uses a reduced gut reference gene catalogue to improve MS/MS identification). We note Var2Pep may still be used when assembly graph is not available (*i.e.* only the contigs are available). In such a case, peptides encoded in short sequencing reads that share similarity with MS/MS supported peptides identified from the contigs can be included in the database searching for MS/MS analysis.

The main goal of our approaches is to optimize the use of matching metagenomic and/or metatranscriptomic data for

metaproteomics data analysis. Our integrated pipeline results in two databases for MS/MS identification, one from Graph2Pro (Graph2Pro DB), and the other from Var2Pep (Var2Pep DB). Our pipeline provides one way of using these databases as shown in Fig. 1. Users can make use of these databases in different ways. In order to get high confident peptide identification, as well as variant peptides, it is highly recommended that users apply strict false discovery rate control, either in spectral level or peptide level. A reasonable concern on our approach of combining the results from two separated searches each at 1% FDR is that the actually FDR would be higher. We re-calculated the FDRs for combined results, based on the unique sets of identified spectra for all the data sets we tested. The results showed that the FDRs estimated using this approach were slightly higher, but still reasonably low (at about 1.5%).

It is possible to combine our approaches with those that are recently developed, including the utilization of spectral clustering to speed up the search as shown in (57). We used the MS-GF+ search engine (48), which is one of the fastest MS/MS search engines. More recently developed tools for the fast database searching of metaproteomic MS/MS data such as ProteoStorm (26) can also be incorporated into the pipeline. Our significantly improved results indicate there is still a huge space for algorithmic improvement in the field of metaproteomics and proteomics in general.

The goal of metaproteomics is not only to identify proteins expressed in the microbial community, but also to estimate their abundances (*i.e.* their expression levels) under different conditions. Nevertheless, a protein can be quantified only if it can be identified by using the metaproteomic data. Therefore, the methods presented here that increase the coverage of protein identification will also help the subsequent steps for protein quantification. Furthermore, identification of peptide variants may contribute to accurate quantification of the proteins from which these peptides are derived. Tools such as MetaGOmics (58) can be applied to infer functional and taxonomic contents of the microbial communities based on identified proteins/peptides. Better identification of peptides and/or proteins from metaproteomic MS/MS data can improve the functional and pathway annotation of the corresponding microbial communities, as we have shown in (33).

We showed that metaproteomics (and matching metagenomics/metatranscriptomics) data can be used to identify hypothetical proteins that sometimes are very abundant in the microbial communities. These MS/MS supported hypothetical proteins provide interesting candidates for further studies of their functions using computational and experimental approaches. However, our analyses of the identified proteins also suggested that unless we significantly increase the depth of the metaproteomic experiments, what we identify are limited to those highly abundant (often well studied) proteins.

In conclusion, we developed a pipeline for metaproteomic MS/MS data analysis using matching metagenomic and/or metatranscriptomic sequencing data as the reference. Tests of our pipeline using publicly available meta-omics data sets showed that it is important to consider the assembly uncertainties (captured in assembly graph) and genomic variants to maximize the utilization of metagenomes and/or metatranscriptomes as the reference for metaproteomic data interpretation.

DATA AVAILABILITY

Our pipeline for peptide/protein identification from metaproteomic data using matching metagenomic and/or metatranscriptomic data as the reference can be downloaded from https://github.com/COL-IU/graph2pro-var. We also made available the results (including the reference databases and search results) from analyzing the two collections of microbiome data sets in a supplementary website at http://omics.soic.indiana.edu/MP/and Zenodo at http://doi.org/10.5281/zenodo.2691363.

* The NIH grants 1R01AI108888 and 1R01AI143254, and the Indiana University (IU) Precision Health Initiative (PHI).

S This article contains supplemental Figures.

‡ To whom correspondence should be addressed. E-mail: yye@indiana.edu.

REFERENCES

1. Crump, B. C., Armbrust, E. V., and Baross, J. A. (1999) Phylogenetic analysis of particle-attached and free-living bacterial communities in the columbia river, its estuary, and the adjacent coastal ocean. *Appl. Environ. Microbiol.* **65,** 3192–3204
2. Santelli, C. M., Orcutt, B. N., Banning, E., Bach, W., Moyer, C. L., Sogin, M. L., Staudigel, H., and Edwards, K. J. (2008) Abundance and diversity of microbial life in ocean crust. *Nature* **453,** 653–656
3. Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., and Knight, R. (2012) Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* **6,** 1007–1017
4. Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., and Caporaso, J. G. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. of the Natl. Acad. Sci. USA* **109,** 21390–21395
5. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., and Yamada, T. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65
6. Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359
7. Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023
8. Wilmes, P., Andersson, A. F., Lefsrud, M. G., Wexler, M., Shah, M., Zhang, B., Hettich, R. L., Bond, P. L., VerBerkmoes, N. C., and Banfield, J. F. (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J.* **2,** 853
9. Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P. M., Alou, M. T., Daillere, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragon, L., Jacquelot, N., Qu, B., Ferrere, G., Clemenson, C., Mezquita, L., Masip, J. R., Naltet, C., Brosseau, S., Kaderbhai, C., Richard, C., Rizvi, H., Levenez, F., Galleron, N., Quinquis, B., Pons, N., Ryffel, B., Minard-Colin, V., Gonin, P., Soria, J. C., Deutsch, E., Loriot, Y.,

Ghiringhelli, F., Zalcman, G., Goldwasser, F., Escudier, B., Hellmann, M. D., Eggermont, A., Raoult, D., Albiges, L., Kroemer, G., and Zitvogel, L. (2018) Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors, *Science* **359,** 91–97
10. Zhu, W., Winter, M. G., Byndloss, M. X., Spiga, L., Duerkop, B. A., Hughes, E. R., Büttner, L., de Lima Romão, E., Behrendt, C. L., Lopez, C. A., Sifuentes-Dominguez, L., Huff-Hardy, K., Wilson, R. P., Gillis, C. C., Tükel, Ç., Koh, A. Y., Burstein, E., Hooper, L. V., Bäumler, A. J., and Winter, S. E. (2018) Precision editing of the gut microbiota ameliorates colitis. *Nature* **553,** 208
11. Ho, C. L., Tan, H. Q., Chua, K. J., Kang, A., Lim, K. H., Ling, K. L., Yew, W. S., Lee, Y. S., Thiery, J. P., and Chang, M. W. (2018) Engineered commensal microbes for diet-mediated colorectal-cancer chemoprevention, *Nat. Biomed. Eng.* **2,** 27–37
12. Shi, Y., and Tyson, G. W. (2009) DeLong, E. F. Metatranscriptomics reveals unique microbial small rnas in the ocean's water column. *Nature* **459,** 266–269
13. Stewart, F. J., Ulloa, O., and DeLong, E. F. (2012) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* **14,** 23–40
14. Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., Lefsrud, M. G., Apajalahti, J., Tysk, C., and Hettich, R. L. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME.J.* **3,** 179–189
15. Wilmes, P., and Bond, P. L. (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **14,** 92–97
16. Maron, P.-A., Ranjard, L., Mougel, C., and Lemanceau, P. (2007) Metaproteomics: a new approach for studying functional microbial ecology. *Microbial Ecol.* **53,** 486–493
17. Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N. C., Fraser, C. M., Hettich, R. L., and Jansson, J. K. (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PloS One* **7,** 49138
18. Pachiadaki, M. G., Sintes, E., Bergauer, K., Brown, J. M., Record, N. R., Swan, B. K., Mathyer, M. E., Hallam, S. J., Lopez-Garcia, P., Takaki, Y., Nunoura, T., Woyke, T., Herndl, G. J., and Stepanauskas, R. (2017) Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. *Science* **358,** 1046–1051
19. Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., and Strous, M. (2017) Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **8,** 1558
20. Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017) Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261,** 24–36
21. Cheng, K., Ning, Z., Zhang, X., Li, L., Liao, B., Mayne, J., Stintzi, A., and Figeys, D. (2017) MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* **5,** 157
22. Kertesz-Farkas, A., Keich, U., and Noble, W. S. (2015) Tandem mass spectrum identification via cascaded search. *J. Proteome Res.* **14,** 3027–3038
23. Mesuere, B., Debyser, G., Aerts, M., Devreese, B., Vandamme, P., and Dawyndt, P. (2015) The unipept metaproteomics analysis pipeline. *Proteomics* **15,** 1437–1442
24. Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T., Wroblewski, M. S., Seymour, S. L., and Griffin, T. J. (2013) A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13,** 1352–1357
25. Muth, T., Behne, A., Heyer, R., Kohrs, F., Benndorf, D., Hoffmann, M., Lehtevä, M., Reichl, U., Martens, L., and Rapp, E. (2015) The metaproteomeanalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.* **14,** 1557–1565
26. Beyter, D., Lin, M. S., Yu, Y., Pieper, R., and Bafna, V. (2018) Proteostorm: An ultrafast metaproteomics database search framework. *Cell Syst.* **7,** 463–467
27. 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. F., and Igeys, D. (2016) MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4,** 31
28. Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., and Snyder, M. (2016) Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* **34,** 64–69
29. Sharon, I., Kertesz, M., Hug, L. A., Pushkarev, D., Blauwkamp, T. A., Castelle, C. J., Amirebrahimi, M., Thomas, B. C., Burstein, D., Tringe, S. G., Williams, K. H., and Banfield, J. F. (2015) Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res.* **25,** 534–543
30. May, D. H., Timmins-Schiffman, E., Mikan, M. P., Harvey, H. R., Borenstein, E., Nunn, B. L., and Noble, W. S. (2016) An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *J. Proteome Res.* **15,** 2697–2705
31. Cantarel, B. L., Erickson, A. R., VerBerkmoes, N. C., Erickson, B. K., Carey, P. A., Pan, C., Shah, M., Mongodin, E. F., Jansson, J. K., Fraser-Liggett, C. M., and Hettich, R. L. (2011) Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PloS One* **6,** 27173
32. Rooijers, K., Kolmeder, C., Juste, C., Doré, J., De Been, M., Boeren, S., Galan, P., Beauvallet, C., de Vos, W. M., and Schaap, P. J. (2011) An iterative workflow for mining the human intestinal metaproteome. *BMC Genomics* **12,** 6
33. Tang, H., Li S., and Ye, Y. (2016) A graph-centric approach for metagenome-guided peptide and protein identification in metaproteomics, *PLoS Computational Biol.* **12,** 1005224
34. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14,** 513
35. Nayfach, S., Rodriguez-Mueller, B., Garud, N., and Pollard, K. S. (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26,** 1612–1625
36. Costea, P. I., Munch, R., Coelho, L. P., Paoli, L., Sunagawa, S., and Bork, P. (2017) metaSNV: A tool for metagenomic strain level analysis, *PLoS ONE* **12,** 0182392
37. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.* **10,** 25
38. Rho, M., Tang, H., and Ye, Y. (2010)Fraggenescan: predicting genes in short and error-prone reads, *Nucleic Acids Res.* **747,** e191
39. Zhao, Y., Tang, H., and Ye, Y. (2012) Rapsearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28,** 125–126
40. Muller, E. E., Pinel, N., Laczny, C. C., Hoopmann, M. R., Narayanasamy, S., Lebrun, L. A., Roume, H., Lin, J., May, P., Hicks, N. D., Heintz-Buschart, A., Wampach, L., Liu, C. M., Price, L. B., Gillece, J. D., Guignard, C., Schupp, J. M., Vlassis, N., Baliga, N. S., Moritz, R. L., Keim, P. S., and Wilmes, P. (2014) Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Commun.* **5,** 5603
41. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) Peptideatlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Reports* **9,** 429–434
42. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The peptideatlas project. *Nucleic Acids Res.* **34**, 655–658
43. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30,** 2114–2120
44. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S. M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T. W., and Wang, J. (2012) Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18
45. Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015) Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph, *Bioinformatics* **31,** 1674
46. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017) metaspades: a new versatile metagenomic assembler. *Genome Res.* **27,** 824–834
47. Vollmers, J., Wiegand, S., and Kaster A. K. (2017) Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLoS ONE* **12,** 0169662

48. Kim, S., and Pevzner, P. A. (2014) Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nat. Communications* **5,** 5277
49. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214
50. Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42,** 206–214,
51. Ye, Y., and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **19,** 246–255
52. Xu, D., and Zhang, Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80,** 1715–1735
53. Misra, C. S., Basu, B., and Apte, S. K. (2015) Surface (S)-layer proteins of Deinococcus radiodurans and their utility as vehicles for surface localization of functional proteins. *Biochim. Biophys. Acta* **1848,** 3181–3187
54. Sychantha, D., Chapman, R. N., Bamford, N. C., Boons, G. J., Howell, P. L., and Clarke, A. J. (2018) Molecular basis for the attachment of S-layer proteins to the cell wall of Bacillus anthracis. *Biochemistry* **57,** 1949–1953
55. Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428,** 726–731
56. Henikoff, S., and Henikoff, J. G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17,** 49–61
57. Griss, J., Foster, J. M., Hermjakob, H., and Vizcaino, J. A. (2013) PRIDE Cluster: building a consensus of proteomics data. *Nat. Methods* **10,** 95–96
58. Riffle, M., May, D. H., Timmins-Schiffman, E., Mikan, M. P., Jaschob, D., Noble, W. S., and Nunn, B. L. (2017) A web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes* **6,** 2