Journal of proteome research

Article

pubs.acs.org/jpr

An Alignment-Free "Metapeptide" Strategy for Metaproteomic Characterization of Microbiome Samples Using Shotgun Metagenomic Sequencing

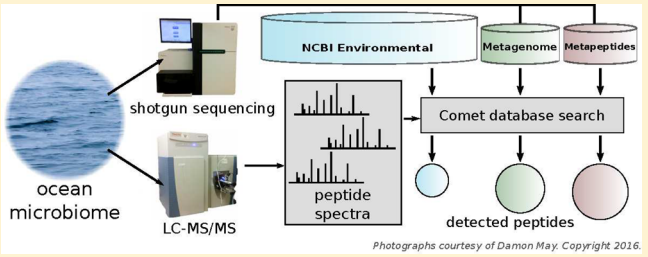
Damon H. May,[†] Emma Timmins-Schiffman,[†] Molly P. Mikan,[§] H. Rodger Harvey,[§] Elhanan Borenstein,[†,‡,∥] Brook L. Nunn,[†] and William S. Noble*[,†,‡]

[†]Department of Genome Sciences and [‡]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-5065, United States

[§]Department of Ocean, Earth & Atmospheric Sciences, Old Dominion University, Norfolk, Virginia 23529, United States

[∥]Santa Fe Institute, Santa Fe, New Mexico 87501, United States

ABSTRACT: In principle, tandem mass spectrometry can be used to detect and quantify the peptides present in a microbiome sample, enabling functional and taxonomic insight into microbiome metabolic activity. However, the phylogenetic diversity constituting a particular microbiome is often unknown, and many of the organisms present may not have assembled genomes. In ocean microbiome samples, with particularly diverse and uncultured bacterial communities, it is difficult to construct protein databases that contain the bulk of the peptides in the sample without losing detection sensitivity due to the overwhelming number of candidate peptides for each tandem mass spectrum. We describe a method for deriving "metapeptides" (short amino acid sequences that may be represented in multiple organisms) from shotgun metagenomic sequencing of microbiome samples. In two ocean microbiome samples, we constructed site-specific metapeptide databases 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. The increased peptide yield has the potential to enrich the taxonomic and functional characterization of sample metaproteomes.



KEYWORDS: *microbial ecology, metaproteomics, metagenomics, mass spectrometry, microbial communities*

1. INTRODUCTION

Because the ocean microbial community is the dominant driver of ocean biogeochemical processes such as the carbon cycle, a quantitative understanding of the ocean microbial taxa performing important functions is essential.[1−3] Due to culture technique limitations on mixed microbial communities, methods for examining whole microbiomes *in situ* are needed. Metaproteomic analysis of ocean samples has the power to detect peptides from thousands of proteins over a wide range of taxonomic groups within a single analysis.[4−7] Accordingly, metaproteomics has been used to investigate the functional roles of ocean microbes in a variety of ecological contexts.[5,7,8] However, the success of high-throughput proteomics on ocean samples has been limited by a lack of detection sensitivity.[9]

The majority of organisms active in the ocean microbiome do not have assembled genomes.[10] Public databases can provide partial metaproteome coverage, but without a precise guide to which organisms are present in the sample, those databases must be extremely large in order to accommodate as much sequence variation as possible. Searching against such very large databases severely and negatively impacts search sensitivity.[11−13] Because of the difficulty of constructing a protein database that accurately reflects an ocean bacterial microbiome, ocean metaproteomics experiments typically only detect a small proportion of the potentially detectable peptides in a sample.[5,14]

As sequencing technologies have become more accessible, "meta-omics" studies have integrated metagenomic, metatranscriptomic, and metaproteomic data. For example, databases for metaproteomic searches can be constructed using genes predicted from an assembled metagenome.[8] However, this approach can lead to low peptide detection sensitivity for two reasons. First, many gene fragments present in sequencing reads cannot be reliably assembled into longer contigs, so they will be missing from the gene prediction. Second, the process of optimal metagenome assembly requires expertise not necessarily shared by all researchers wishing to do metaproteomics analysis, and if not done optimally, then the metagenome may fail to contain much of the variation present in the sequencing data. For both of these reasons, even metaproteomic databases based on site-specific assembled metagenomes tend to provide substantially incomplete coverage of the sample metaproteome.[13,15,16]

Received: March 17, 2016
Published: July 11, 2016

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

An alternative approach takes advantage of the fact that most of the organisms present in many microbiome samples are prokaryotes, and therefore high proportions of their genomes are protein-coding. Tools such as MetaGeneAnnotator,[17,18] Orphelia,[16] and FragGeneScan[19] predict gene fragments directly from sequencing reads, without assembling the reads into contigs. These approaches can be used to construct metaproteomic databases suitable for database search. As Cantarel et al.[13] demonstrated, these databases enable a greater peptide yield via database search than other methods, with sensitivity greatly dependent on the specifics of the approach to database construction. However, the goal of these tools is sensitive gene prediction rather than peptide detection, so databases containing translations of their raw gene fragment output can be extremely large. This can lead to impractically long running times for database searches and, more importantly, reduced peptide detection sensitivity.

In the approach described here, we begin with either the gene fragments predicted by MetaGeneAnnotator or six-frame translations of raw reads. We trim and filter these sequences to build a database of "metapeptides": short amino acid sequences derived from open reading frame fragments found in individual reads that are more likely to be identifiable via LC-MS/MS (Figure 1A). This approach exploits more of the metagenomic

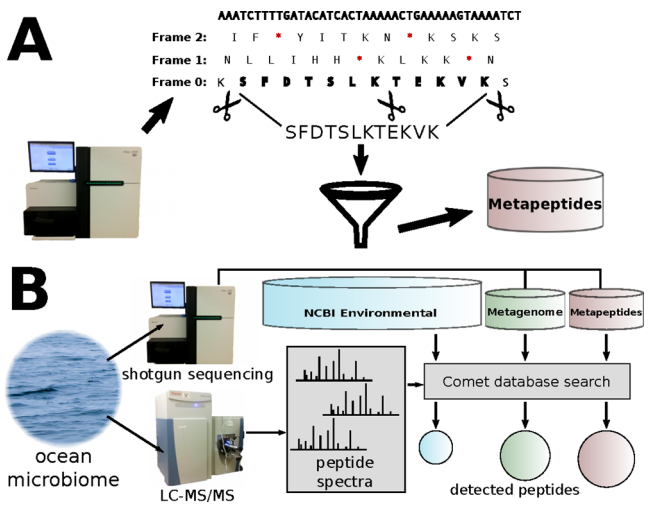


Figure 1. Multiple approaches for metaproteomics of microbiome samples. (A) After high-throughput sequencing, metapeptide database construction begins with six-frame translations of raw sequencing reads, or with gene fragments predicted from reads. Amino acid sequences are trimmed to their outermost tryptic sites to yield metapeptide sequences. Candidate metapeptides are filtered on per-base quality scores, sequence length and other features. Passing candidates are added to the database. (B) Alternative proteomics workflows. Microbiome samples are subjected to shotgun metagenomic sequencing and LC-MS/MS analysis. MS/MS spectra are searched with Comet against the NCBI environmental database, against predicted genes from an assembled metagenome, or against a metapeptide database, resulting in peptide yields of different size. Photographs copyright 2016 Damon May.

data than an approach based on an assembled metagenome, incorporating reads that fail to be integrated into a contig as well as all of the sample variation for each gene sequence while avoiding a loss of sensitivity due to overinclusivity. It is both more complete and more focused on the sample at hand than a strategy based on public databases, potentially including sequences never before observed in any organism and excluding sequences from species not present in the sample.

To evaluate the utility of our metapeptide approach, we compared the sets of peptide sequences detected in two Arctic Ocean microbiome samples at a 1% false discovery rate (FDR) via database search against three different databases (Figure 1B): the NCBI nonredundant database of environmental protein sequences (env_nr), which is commonly used to interrogate ocean and soil microbiome samples,[4,20] a database derived from a metagenome assembled from shotgun metagenomic sequencing of the two Arctic Ocean samples, and metapeptide databases constructed from the same sequencing reads.

Two microbiome samples were collected from the Arctic Ocean: one sample from the surface chlorophyll maximum layer in the Bering Strait (BSt) and one from bottom waters in the Chuckchi Sea (CS). A total of 2050 peptides were detected in the BSt sample by searching the environmental database. A metagenome-derived database search yielded 2.12 times as many peptides, and a metapeptide search yielded 3.37 times as many peptides. Results were similar in the CS sample, though with many fewer peptides detected in each search. Integrating the results from all three databases further increased peptide yield.

This substantial advantage in peptide yield contributes greatly to the taxonomic and potential functional classification of the sample proteomes. We used Unipept[21,22] to infer the lowest common ancestor taxon for peptides detected in each search, as well as the list of Gene Ontology (GO) "biological process" categories associated with proteins containing each peptide. Comparison of the results revealed a much richer taxonomic characterization of the proteins present in the samples from the metapeptide search than from either of the other methods, and a much higher number of detected peptides with the potential for functional annotation. Thus, in addition to dramatically increasing the number of peptides detected in a given ocean sample, the metapeptide-based approach can significantly expand our understanding of the organisms producing the biochemically active molecules in a microbiome. This understanding is crucial to developing a functional model of the microbiome.

2. METHODS

The data described in the following sections may be downloaded at http://noble.gs.washington.edu/proj/metapeptide.

2.1. Experimental Methods

2.1.1. Sample Collection. Water samples were collected in August of 2013 from the Bering Strait (BSt) chlorophyll maximum layer (7 m depth, 65°43.44″ N, 168°57.42″ W) and from the more northern Chukchi Sea (CS) bottom waters (55.5 m depth, 72°47.624″ N, 16°53.89″ W) using a 24-bottle CTD (conductivity, temperature, and depth) rosette (10 L General Oceanics Niskin X). The measurement of integrated water column chlorophyll was 226.88 mg/m^2 at station BSt and 2.64 mg/m^2 at station CS. As our previous work has shown, to examine bacterial contributions, it is essential to remove the very high background contribution from algal inhabitants.[23] Also, oceanic marine bacteria are typically smaller than bacteria in gut biomes or freshwater systems, with the majority passing a 1.0 μm filter.[24,25] Accordingly, a 15 L water sample was prefiltered through two high-volume cartridges (10 μm and

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

then 1 μm) to remove larger eukaryotes, and the filtrate comprising the bacterial microbiome was then collected on a glass fiber filter (GF/F) with nominal pore size of 0.7 μm. Filters were flash frozen and stored at −80 °C until extraction.

2.1.2. Metagenome DNA Extraction, Library Preparation, and Sequencing. Filters were sliced, and DNA extraction was accomplished using the protocol developed for planktonic biomass on Sterivex filters, as described in Wright et al.[26] Briefly, DNA was extracted from the collected cells using phenol/chloroform and chloroform extractions. DNA was then purified using a cesium chloride density gradient. Extracted DNA was sheared to <1 kb, and excess salts were cleaned up using Agencourt AMPure XP purification (Beckman Coulter, Brea, CA). Library preparation was done with the Kapa Hyper Kit, following the manufacturer's instructions (Kapa Biosystems, Wilmington, MA), and library quality was confirmed using the Bioanalyzer (Agilent, Santa Clara, CA). Libraries were sequenced in one lane on an Illumina HiSeq. The resulting 100 bp, paired-end sequencing reads were trimmed and filtered using SolexaQA,[27] with a minimum Phred quality score[28] of 20 on any base.

2.1.3. Protein Sample Preparation and Tandem Mass Spectrometry (LC-MS/MS). GF/F filters with the bacterial fraction were placed in 1.5 mL tubes with 100 μL of 0.5 mm glass beads, 100 μL of 6 M urea, and 500 μL of nanopure water. Filters were shaken on a bead beater for 1 min and then placed in ice for 5 min. This process was repeated 10 times to ensure cell lysis and filter breakup. A needle was then heated by flame and used to create a <0.5 mm hole at the bottom of the 1.5 mL sample tube. The sample tubes were then placed atop an open 1.5 mL tube and centrifuged (3000*g*, 10 min). This process was completed to isolate protein lysate from extracted particles and glass beads. Protein concentrations were determined using BCA colormetric assay; 100 μg of total protein was used for digestion. Each 100 μg protein sample received 300 ng of purified human ApoA1 to monitor protein digesion. Samples were reduced, alkylated, enzymatically digested with trypsin, and desalted following Nunn et al.[29] Prior to MS injections, 50 fmol of the Pierce Peptide Retention Time Standard (ThermoFisher Scientific) was added to each autosample vial at 50 fmol per 2 μg of total protein. Peptides were separated using an inline NanoAquity HPLC with a 4 cm precolumn (5 μm; 200A; Magic C18) and 30 cm Reprosil-Pur Basic 3 μm C18 analytical column (Dr. Maisch GmbH, Germany). Peptides were eluted using a 2−30% ACN, 0.1% formic acid nonlinear gradient in 120 min at 300 nL/min. LC-MS/MS was performed with a Q-Exactive-HF (ThermoScientific) on technical triplicates for each sample. The instrument was operated in Top 20 data-dependent acquisition mode, collecting data on 400−1600 m/z range with a 5 s dynamic exclusion.

2.2. Computational Methods

All computation was performed on a Univa Grid Engine cluster with 1.90 GHz AMD Opteron processors.

2.2.1. Gene Prediction from Shotgun Sequencing with Existing Methods. The MOCAT pipeline[30] was used to assemble a metagenome and predict genes as follows. Trimmed and filtered reads from both BSt and CS samples were aligned to the human hg19 reference using SOAPaligner v2.21, and aligned reads were removed. The remaining reads were assembled into contigs and scaftigs with SOAPdenovo v1.06. The assembly was revised, correcting for indels and chimeric regions, with SOAPdenovo v1.06 and BWA v0.7.5a-r16. Genes were predicted using Prodigal v2.60.

We used three well-established gene fragment prediction tools to predict gene fragments directly from shotgun metagenomic sequencing reads from each sample: MetaGeneAnnotator (in multiple species mode), FragGeneScan version 1.2.0 (illumina_10 model parameters), and Orphelia (with Net300 prediction model).

2.2.2. Metapeptide Databases. Separate metapeptide databases were constructed from the BSt and CS sequencing runs, from either predicted gene fragments or raw read sequences. When starting from raw read sequences, each read was translated in all six reading frames, and reading frames containing a stop codon were discarded. The results described in section 3 were obtained by starting with predicted gene fragments from MetaGeneAnnotator.

Whether starting from gene fragments or raw read sequences, amino acid sequences from each nucleotide sequence were trimmed to the first and last tryptic cleavage site (or discarded if fewer than two sites), and the remaining ends were discarded (Figure 1A). This was done in order to remove partial tryptic peptide sequences that are unlikely to be detected by LC-MS/MS of a trypsinized metaproteome. The resulting candidate sequences were discarded if they were less than 10 amino acids long, if they contained no tryptic peptides with seven or more amino acids, or if the minimum Phred quality score over the length of the sequence was less than 30. Finally, metapeptide candidates meeting all the above criteria were discarded if they were represented by fewer than two reads. A FASTA database was constructed from the remaining metapeptides.

For purposes of comparison, we also made use of a metagenome-derived database of translated genes from the metagenome described above and the NCBI nonredundant database of protein sequences from large environmental sequencing projects ('env_nr', downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/env_nr.gz on December 1, 2015).

2.2.3. Database Search. All database searches were performed using Comet[31] version 2015.01 rev. 2, using a concatenated decoy database in which peptide sequences were reversed but C-terminal amino acids were left in place. Search parameters included a static modification for cysteine carbamidomethylation (57.021464) and a variable modification for methionine oxidation (15.9949). Enzyme specificity was trypsin, with one missed cleavage allowed. Parent ion mass tolerance was set to 10 ppm around five isotopic peaks, and fragment ion binning was 0.02, with offset 0.0. Peptide-spectrum matches (PSMs) from all technical replicates were combined into a single data set. As described previously,[32] after each unique peptide was associated with its top-scoring spectrum, irrespective of charge state, we used the widely used target−decoy search strategy of estimating the false discovery rate (FDR) associated with a given set of accepted peptides.[33] In this context, the FDR is defined as the proportion of the accepted peptides that are not responsible for generating observed spectra. We then empirically examined the trade-off between FDR and the number of accepted peptides, since in practice the mass spectrometrist is typically interested in accepting as many peptides as possible while maintaining an acceptable FDR. Note that this trade-off is similar to the distinction between precision (1 − FDR) and recall or sensitivity.

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

Results of searches of individual samples against multiple databases were integrated as follows. PSMs from searches against all databases were combined into a single tab-delimited file of features for input to Percolator.[34] For each database, a new binary feature was added to the combined feature file indicating whether the PSM was derived from a search against that database. Percolator was then used to analyze the combined set, thereby computing a discriminant score for each PSM. For each scan with multiple PSMs (from multiple databases), all but the highest-scoring PSM were removed. Peptide-level FDR was then calculated as described above.

2.2.4. Taxonomic and Functional Inference. Detected peptides were given taxonomic assignments by Unipept version 1.1.0. For all tryptic peptides with no missed cleavages present in UniProtKB, Unipept assigns a lowest common ancestor (LCA) taxon from the NCBI Taxonomy Database, the most-granular taxon common to all organisms containing the peptide. For peptides with missed tryptic cleavages, Unipept calculates an LCA based on the LCAs associated with all completely cleaved peptide sequences contained in the peptide.

No such standard methods currently exist for assigning functional annotations to detected peptide sequences, so we estimated the maximum number of peptides that could be assigned functional annotations. We used Unipept to retrieve all of the proteins containing each detected peptide, along with their GO category annotations. GO annotations are divided into three namespaces: "biological process", "molecular function", and "cellular compartment". We declared a peptide to be potentially functionally informative if at least one protein containing it was annotated with at least one GO category in the "biological process" namespace.

3. RESULTS

3.1. Gene Fragment Predictions from Deep Shotgun Metagenomics Sequencing Are Not Directly Usable for Proteomics Database Search

Shotgun sequencing of the BSt and CS samples generated 171 million and 245 million reads, respectively. We evaluated three different gene prediction tools: FragGeneScan, Orphelia, and MetaGeneAnnotator. None of these tools were originally developed for this high depth of coverage, nor have they been updated to accommodate high-depth sequencing. On the BSt reads, MetaGeneAnnotator ran to completion in 3.5 h, producing 133 million fragments. Orphelia quickly exceeded 100 GB of memory usage; as its output on smaller inputs was 33 times the size of the output of MetaGene, with no scoring mechanism to use for filtering, we decided not to pursue it further. After 5 days of running time, FragGeneScan had not yet completed, and its output on smaller inputs was 32 times the size of the output of MetaGene, so we decided not to pursue it further.

The 133 million fragments produced by MetaGeneAnnotator contained 222 million unique peptides and required more than 4 days to search one replicate against. However, 177 million of the peptides in the database represented ragged ends of peptides terminating at the beginning or end of a metagenomic sequencing read and did not represent a detectable tryptic peptide.

3.2. Environmental and Assembled Metagenome Databases Provide Incomplete Coverage of Peptides in Ocean Samples

Next, we quantified the extent to which a public database and a metagenome-derived database could be used to detect the peptides present in the two ocean microbiome samples. The environmental and metagenome databases contained 119 million and 11 million peptides, respectively. All three replicates of each sample were searched against both databases, and the set of peptides detected with FDR < 0.01 in searches against each database was determined with Percolator, as described above (Figure 2).

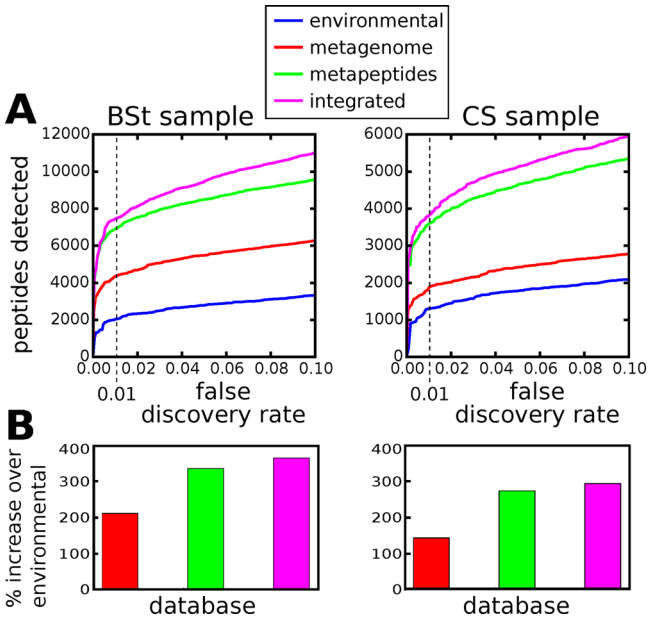


Figure 2. Peptides detected in different searches. (A) Line plots of peptide false discovery rate (FDR, horizontal axis) vs number of peptide sequences detected at that FDR in the Bering Strait (BSt) and Chukchi Sea (CS) samples (vertical axis), when searched four different ways. Dashed line indicates peptide yield at FDR > 0.01 from searches against the NCBI environmental database (2050 in BSt and 1317 in CS), against the metagenome-derived database (4344 and 1877), against metapeptides derived from the sample being searched (6918 and 3606), and integrated results from all three databases (7508 and 3871). (B) Detected peptide counts at FDR < 0.01 in the metagenome, metapeptide, and integrated database searches as a percentage of the counts detected in environmental search.

For the BSt sample, of the 4344 peptides present in the metagenome database and detected in the metagenome search, 61.2% did not occur in the environmental database; similarly, 46.4% of the 2050 peptides present in the environmental database and detected in the environmental search were absent from the metapeptide database. This high complementarity indicates that large numbers of peptides present in the sample are absent from each database. Furthermore, of the 1708 peptides present in both databases (and therefore potentially detectable by either search) and detected in one or both searches, only 1.3% were detected in the environmental database search but not in the metapeptide database search. By contrast, 35.7% were detected in the metapeptide search but not in the environmental search. Those peptides were present in the environmental database, so we conclude that the failure

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

to detect them is due to a loss of statistical power stemming from the much larger size of the environmental database.

3.3. Searching Metapeptide Databases Increases Peptide Yield and Enriches Taxonomic and Functional Characterization

Next, we evaluated the ability of our metapeptide databases to increase peptide detection sensitivity relative to the environmental and metagenome databases. The metapeptide databases constructed from the shotgun metagenomic sequencing reads from the BSt and CS samples contained 12 million and 14 million peptides, respectively. The BSt metapeptide database (12 million peptides) contained 2 million peptides in common with the environmental database (129 million peptides) and 4 million in common with the metagenome database (11 million peptides). All MS/MS replicates of the BSt and CS ocean microbiome samples were searched against the metapeptide database constructed from the sample being searched, and the set of peptides detected with FDR < 0.01 was derived with Percolator. In the BSt and CS samples, the numbers of peptides detected were 1.59 and 1.92 times the number detected by searching against the metagenome-derived database and 3.37 times and 2.74 times the number detected by searching against the environmental database, respectively (Figure 2).

To determine the reasons for this larger peptide yield, we compared the sets of peptides detected by searching the BSt spectra against the metapeptide and environmental databases (Figure 3). Of the 6918 peptides present in the metapeptide database and detected in the metapeptide search, 67.8% did not occur in the environmental database; by contrast, only 27.5% of the 2050 peptides present in the environmental database and detected in the environmental search were absent from the metapeptide database. This discrepancy suggests that the metapeptide database contains more of the peptides present in the sample. Furthermore, of the 2261 peptides present in both databases and detected in one or both searches, only 1.5% were detected in the environmental database search but not in the metapeptide database search. By contrast, 34.2% were detected in the metapeptide search but not in the environmental search. Those peptides were present in the environmental database, so we conclude that the failure to detect them is due to a loss of statistical power stemming from the much larger size of the environmental database. We also compared the sets of peptides detected by searching the BSt spectra against the metapeptide and metagenome databases, which are of much more similar size. Of the 6918 peptides present in the metapeptide database and detected in the metapeptide search, 41.9% did not occur in the environmental database; by contrast, only 7.9% of the 4344 peptides present in the metagenome database and detected in the metagenome search were absent from the metapeptide database, suggesting more complete sample coverage in the metapeptide database. Furthermore, of the 4250 peptides present in both databases and detected in one or both searches, 5.3% were detected in the metagenome search but not in the metapeptide database search, whereas 5.8% were detected in the metapeptide search but not in the metagenome search. This suggests that the metapeptide database advantage is essentially due to greater coverage and that the searches of the two similarly sized databases are roughly equally sensitive.

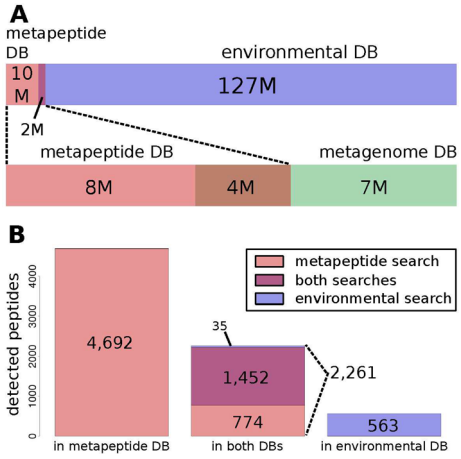


Figure 3. Database and detected peptide comparisons. (A) The BSt metapeptide database contains roughly 12 million tryptic peptides. The environmental database contains 129 million, with an intersection between the two databases of 2 million peptides. The metagenome database contains 11 million peptides, with 4 million peptides in common with the BSt metapeptide database. (B) Searching against the BSt metapeptide database detects 6918 unique peptides at FDR < 0.01, vs 2050 when searching against the environmental database, with 1452 in common. Of the 2261 peptides detected in either search that were present in both databases, 1452 were detected in both searches, 774 were only detected in the BSt metapeptide database search, and 35 were only detected in the environmental database search.

By themselves, detected peptide sequences provide limited information about a sample. However, the peptides can be used to provide important insight into the sample's community composition. Accordingly, we assessed the extent to which the additional peptides detected using the metapeptide database can enrich the taxonomic and functional classification of the metaproteome.

For taxonomic inference, we used the Unipept tool to assign a least common ancestor (LCA) taxon to all possible peptides detected in a search of the BSt sample replicates against a given database. The metapeptide search detected 1.28 times as many peptides that were assigned LCAs as the metagenome-derived database search, and 1.76 times as many as the environmental database search. At every taxonomic rank more granular than class, the highest number of taxa were detected by the integrated search, followed by the metapeptide, metagenome, and then environmental searches (Figure 4). The same order was observed when examining the number of peptides with an LCA at each taxonomic rank. As a side note, the number of peptides detected by a given database search with an LCA at a given rank decreases monotonically with the granularity of the rank, which is a reflection of the LCA ranks of detectable peptides as a whole, as determined by Unipept based on UniProt annotations. The number of taxa detected increases with rank granularity until the rank of species, which shows a modest decline from the rank of genus. This relationship is also a function of the LCA ranks of detectible peptides and does not reflect any particular characteristics of the various searches.

Because many metapeptides are likely from unsequenced microbes not present in public protein databases (and therefore uninformative to Unipept), an important question is whether the detected peptides that were present in the metapeptide database but absent from the environmental database conferred any taxonomic information via this method. Considering the peptides detected in the metapeptide database search, the percentage of peptides that are assignable to an LCA by Unipept is much greater for the subset of those peptides that

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

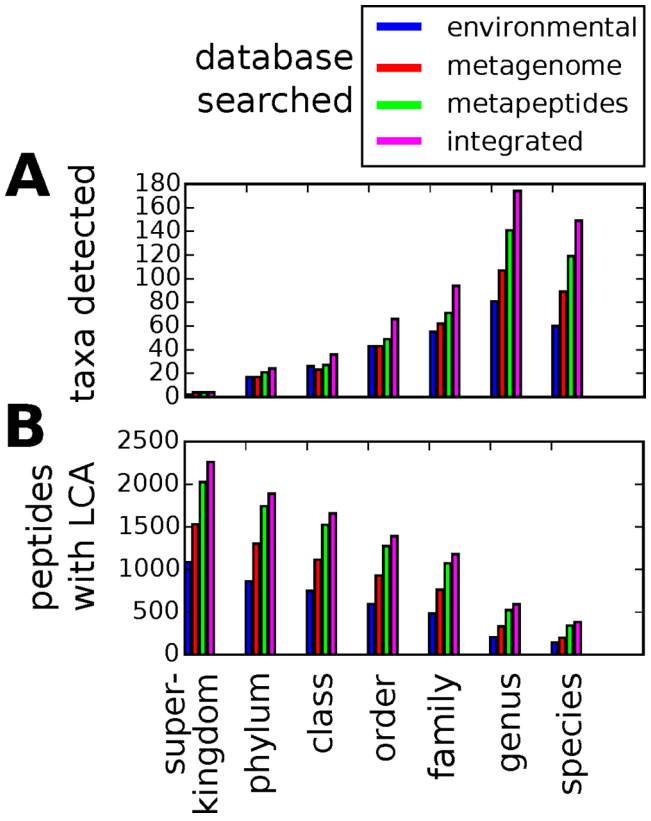


Figure 4. Taxonomic inference summary. Bar charts comparing the taxonomic information derived from four different searches of the BSt sample: against the NCBI environmental database, against the metagenome-derived database, against site-specific metapeptides, and integrating results from all three databases. (A) Counts of taxa detected, by rank from superkingdom to species. (B) Counts of peptides associated with an LCA at each rank.

are present in the environmental database (70.8%) than for the subset that are absent (16.3%). However, because 67.8% of the peptides detected in the metapeptide database search are absent from the environmental database, in absolute terms 32.6% of the peptides assignable to LCAs come from that latter group. Thus, both the greater peptide detection sensitivity and the greater peptide coverage afforded by the metapeptide database contribute to its increased potential for metaproteome taxonomic classification.

To estimate the number of peptides with potential functional annotations, we used Unipept to locate all of the proteins containing each detected peptide, along with their associated GO categories. The metapeptide search detected 1.50 times as many peptides associated with one or more GO categories in the "biological process" namespace as the metagenome-derived database search and 2.12 as many as the environmental database search.

3.4. Combining Results from Multiple Databases Further Increases Peptide Coverage

Although the metapeptide databases are the most valuable individual databases for searching these samples, a higher overall peptide yield can be obtained by combining results from multiple databases. PSMs from searches against the environmental, metagenome, and metapeptide databases were integrated as described above. In the BSt and CS samples, respectively, 1.09 and 1.07 times as many peptides were detected by this method as by searching against the individual metapeptide databases (Figure 2). For context, the largest set of peptides that could possibly be detected at FDR 0.01 by any method combining these three searches is the union of all peptides detected at FDR 0.01 in each of the three separate database searches of each sample. This number was 1.16 and 1.18 times the number of peptides detected via metapeptide searches of the two samples, respectively. We determined that this modest increase was statistically significant via an analysis of the three technical replicates for each sample. In the metapeptide database searches, technical replicates of the BSt and CS samples detected a mean of 4691.3 and 2991.6 peptides, respectively, with a mean pairwise intersection of 3075.0 and 2377.7 peptides between replicates, respectively. The per-replicate means for the BSt and CS samples in the integrated searches were 5090.0 and 3193.0, respectively, and for each sample, the increase was statistically significant by both one-tailed and two-tailed paired t tests at $p < 0.05$. In terms of taxonomic inference, the integrated searches of the BSt and CS samples detected 1.13 and 1.12 times as many peptides assignable to LCAs compared with a metapeptide search (Figure 4 for BSt comparison), with more taxa observed at every taxonomic rank lower than superkingdom.

This method of integrating database search results yielded 14.2% more peptides at FDR < 0.01 as searching a concatenated database combining the environmental, metagenome and metapeptide databases. Because of the reduced statistical power of a search against a larger database, searching the concatenated database yielded 5.0% fewer peptides than searching the metapeptide database alone. The extra Percolator features representing the database against which each PSM was made were of modest benefit, increasing peptide yield by 4.3% versus an integrated search with those features removed.

3.5. Metapeptide Databases from Two Microbiome Samples Can Be Used to Interrogate Each Other

Constructing a metapeptide database is a relatively expensive endeavor, requiring library preparation, short read sequencing, and computational time, so it would be convenient to use a single database to interrogate the metaproteome from multiple samples. Our two samples are from two different locations and from two very different positions in the water column (chlorophyll maximum layer and bottom water). In each case, overall peptide yield from a database search against the metapeptide database derived from the other sample was a large improvement over the yield from a search against the environmental database (2.17 and 1.95 times as many peptides, respectively). In each case, however, searching a sample against its site-specific metapeptide database detected many more peptides than searching against the database derived from the other sample. Notably, the BSt sample appeared to benefit greatly from a search against the BSt database rather than against the CS database (1.55 times as many peptides), whereas the effect in the opposite direction was not as pronounced (1.40 times as many). A potential explanation for this difference lies in the depth from which the two samples were taken: the BSt sample, from the upper water column, is expected to contain more biodiversity than the CS sample taken from the bottom layer, which has no light.

3.6. Filtering Protocols Are Critical to Resulting Metapeptide Database Size

Prior to filtering, the trimmed MetaGene output contained 51 million tryptic peptides. To investigate the effects of the filtering criteria, we systematically varied each parameter while

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

leaving the remaining parameters set to the values described in section 2.2. The results (Figure 5) demonstrate that filtering

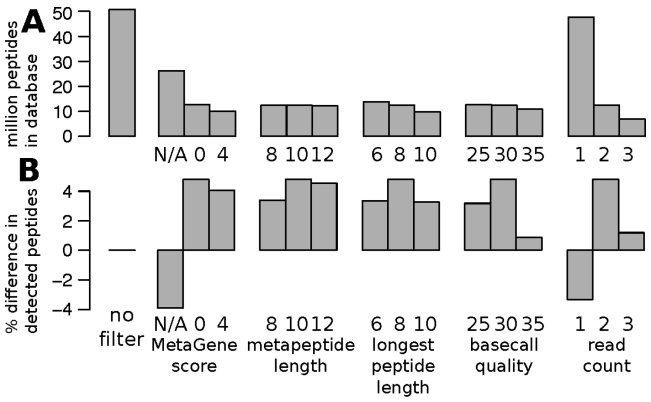


Figure 5. Metapeptide parameter comparison. Comparisons between metapeptide databases constructed with different filtering parameters. The bars on the far left represent a database of MetaGene fragments trimmed to outermost tryptic ends but otherwise are unfiltered. Each group of three bars represents three different values for a single parameter, with all other parameters set as described in section 2.2. In each group, the middle value represents the value used to generate the results described in Figures 2−5. The N/A value for MetaGene score represents the use of raw six-frame translations of reads instead of MetaGene output. (A) Millions of tryptic peptides in each database. (B) Percent difference in counts of peptides detected in a search of 24 000 scans from the three BSt sample replicates at FDR < 0.01 against each database, as compared with search against the unfiltered, trimmed MetaGene database.

metapeptides based on the support of two or more reads and the use of MetaGeneAnnotator fragments rather than six-frame translations of raw reads had particularly large effects on database size, reducing the number of unique tryptic peptides by 74.0 and 51.8%, respectively, and decreasing search time by similar proportions.

To investigate the effect of database filtering parameters on peptide detection sensitivity, we generated a small sample set of 24 000 MS/MS spectra from the BSt sample (8000 random spectra from each replicate run) to compare the number of peptides detected at FDR < 0.01 by searching each database. Beginning with MetaGeneAnnotator fragments rather than with a six-frame translation of raw reads increased detected peptides by 9.0%, demonstrating that the MetaGeneAnnotator is valuable but not crucial to the metapeptide strategy. The MetaGeneAnnotator score was not useful as a filtering criterion: higher score thresholds resulted in monotonically lower peptide yield. Requiring two or more reads increased detected peptides by 8.4%. Higher read count thresholds monotonically reduced yield. Sufficiently restrictive values for each parameter reduce peptide yield much more severely (data not shown). However, in general, within the range of values shown in Figure 5 the reduction in yield was minor, suggesting a relative robustness of the parameter settings.

4. DISCUSSION

In this work, we have demonstrated the value of interrogating microbial metaproteomes by constructing metapeptide databases from site-specific shotgun metagenomic sequencing reads. These databases afford much greater peptide detection sensitivity than the NCBI environmental database or a database of genes predicted from an assembled metagenome. Furthermore, we have shown that a database derived from one sample can be used to interrogate another sample from a different location and position in the water column. By combining metapeptide databases from a variety of samples, sequencing efforts could potentially be centralized to an extent, and metapeptide databases integrated into existing metaproteomics workflows such as the MetaProteomeAnalyzer.[35] In principle, these methods should be applicable to other microbiomes, such as riverine and soil-derived microbial communities, in which prokaryotes dominate the microbiome and the great majority of organisms are unsequenced. These methods may also provide additional sensitivity in a better-understood environment such as the human gut microbiome.

From a practical standpoint, the larger environmental database required more computational time for database search than the metapeptide databases. On our hardware, the three BSt sample replicates, with an average of 104 000 MS/MS scans, took us an average of 1.15 h to search against the BSt database and an average of 14 h to search against the environmental database. The much larger raw database of MetaGeneAnnotator fragments took more than 4 days to search. The strategy of trimming reads to the outermost tryptic sites and the filtering criteria applied are responsible for the much smaller metapeptide database size, making the metapeptide database easier to integrate into a proteomics pipeline.

De novo sequencing is another strategy for increasing peptide detection sensitivity. Although this method can yield many confident partial peptide sequences, it is less effective at confidently detecting full-length peptides. Furthermore, as a result of codon degeneracy, *de novo* sequencing also cannot easily link detected peptides with their corresponding nucleotide sequences for taxonomic annotation. As others have noted, the space of peptides likely to be present in a metaproteomics sample should remain tractable to a database search approach if search databases are constructed with an eye toward detection sensitivity.[36] However, *de novo* sequencing remains a viable approach for assignment of spectra that cannot be assigned with database search.

Some of the peptide sequences detected by metapeptide database search are present in organisms with publicly available genomes, enabling putative taxonomic assignment using existing peptide-based tools and enriching taxonomic characterization. However, a large proportion of peptides detected by searching against metapeptide databases have never been reported in an assembled genome. In future work, we will place those peptides within a taxonomic hierarchy using sequence homology. This may be accomplished using all of the nucleotide sequences of the reads that contributed to the inclusion of each metapeptide in the database.

Sequence homology could also be used to infer the putative function of proteins containing these detected peptides. With both taxonomic and functional assignments, a large number of detected peptides could be used in comparisons of the activity of various microbes between samples. This research will quantify the protein functions responsible for chemical transformations at meaningful taxonomic levels, thereby exposing microbial ecosystems at the molecular level to improve our understanding of their interactions and biological roles. Applying this approach in conjunction with recent advances in quantitative proteomics can bring about a fundamental change in how we view, analyze, and model microbial ecosystems.

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

An important area for future research lies in the development of improved methods for combining search results from multiple databases. The approach we have adopted here relies upon the machine-learning algorithm Percolator to calibrate scores between the different database searches. A more powerful approach might be to adopt a strategy similar to cascade search,[37] searching against, in order, the metapeptide, metagenome, and environmental databases. In future work, we plan to develop and validate a statistical method for combining cascade search with a machine-learning postprocessing step such as Percolator.

The software tools described here have been implemented in Python 2.7. The software (including source code) and data described in this manuscript may be downloaded at http://noble.gs.washington.edu/proj/metapeptide.

AUTHOR INFORMATION

Corresponding Author

*E-mail: william-noble@uw.edu. Tel.: (206) 221-4973.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number P41 GM103533, the National Science Foundation Directorate for Geosciences under award numbers OCE-1233014 and OCE-1233589, and the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

(1) Allison, S. D.; Martiny, J. B. H. Colloquium paper: resistance, resilience, and redundancy in microbial communities. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (Supplement 1), 11512.

(2) Pinhassi, J.; Azam, F.; Hemphälä, J.; Long, R. A.; Martinez, J.; Zweifel, U. L.; Hagström, Å. Coupling between bacterioplankton species composition, population dynamics, and organic matter degradation. *Aquat. Microb. Ecol.* **1999**, *17* (1), 13−26.

(3) Azam, F.; Fenchel, T.; Field, J. G.; Gray, J. C.; Meyer-Reil, L. A.; Thingstad, F. The ecological role of water-column microbes in the sea. *Mar. Ecol.: Prog. Ser.* **1983**, *10* (3), 257−264.

(4) Morris, R. M.; Nunn, B. L.; Frazar, C.; Goodlett, D. R.; Ting, Y. S.; Rocap, G. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J.* **2010**, *4* (5), 673−685.

(5) Georges, A. A.; El-Swais, H.; Craig, S. E.; Li, W. K. W.; Walsh, D. A Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. *ISME J.* **2014**, *8* (6), 1301−1313.

(6) Yoshida, M.; Yamamoto, K.; Suzuki, S. Metaproteomic characterization of dissolved organic matter in coastal seawater. *J. Oceanogr.* **2014**, *70* (1), 105−113.

(7) Hawley, A. K.; Brewer, H. M.; Norbeck, A. D.; Paša-Tolic, L.; Hallam, S. J. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (31), 11395−11400.

(8) Teeling, H.; Fuchs, B. M.; Becher, D.; Klockow, C.; Gardebrecht, A.; Bennke, C. M; Kassabgy, M.; Huang, S.; Mann, A. J; Waldmann, J.; et al. Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science* **2012**, *336*, 608−611.

(9) 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* (20), 3439−3453.

(10) Rappé, M. S.; Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **2003**, *57*, 369−394.

(11) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092−2123.

(12) Noble, W. S. Mass spectrometrists should search only for peptides they care about. *Nat. Methods* **2015**, *12* (7), 605−608.

(13) 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.; Hettich, R. L. Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One* **2011**, *6*, e27173.

(14) Keiblinger, K. M.; Wilhartitz, I. C.; Schneider, T.; Roschitzki, B.; Schmid, E.; Eberl, L.; Riedel, K.; Zechmeister-Boltenstern, S. Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biol. Biochem.* **2012**, *54*, 14−24.

(15) Xiong, W.; Abraham, P. E.; Li, Z.; Pan, C.; Hettich, R. L. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics* **2015**, *15* (20), 3424−3438.

(16) Hoff, K. J.; Lingner, T.; Meinicke, P.; Tech, M. Orphelia: Predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* **2009**, *37*, W101−W105.

(17) Noguchi, H.; Park, J.; Takagi, T. MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **2006**, *34* (19), 5623−5630.

(18) Noguchi, H.; Taniguchi, T.; Itoh, T. MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* **2008**, *15* (6), 387−396.

(19) Rho, M.; Tang, H.; Ye, Y. FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res.* **2010**, *38* (20), e191.

(20) Moore, E. K.; Nunn, B. L.; Faux, J. F.; Goodlett, D. R.; Harvey, H. R. Evaluation of electrophoretic protein extraction and database-driven protein identification from marine sediments. *Limnol. Oceanogr.: Methods* **2012**, *10*, 353−366.

(21) Mesuere, B.; Devreese, B.; Debyser, G.; Aerts, M.; Vandamme, P.; Dawyndt, P. Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* **2012**, *11* (12), 5773−5780.

(22) Mesuere, B.; Debyser, G.; Aerts, M.; Devreese, B.; Vandamme, P.; Dawyndt, P. The Unipept metaproteomics analysis pipeline. *Proteomics* **2015**, *15* (8), 1437−1442.

(23) Moore, E. K.; Nunn, B. L.; Goodlett, D. R.; Harvey, R. H. Identifying and tracking proteins through the marine water column: Insights into the inputs and preservation mechanisms of protein in sediments. *Geochim. Cosmochim. Acta* **2012**, *83*, 324−359.

(24) Kirchman, D. *Microbial ecology of the oceans*; John Wiley & Sons: Hoboken, NJ, 2010.

(25) Needham, D. M.; Fuhrman, J. A. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology* **2016**, *1* (4), 16005.

(26) Wright, J. J.; Lee, S.; Zaikova, E.; Walsh, D. A.; Hallam, S. J. DNA extraction from 0.22 microM Sterivex filters and cesium chloride density gradient centrifugation. *J. Visualized Exp.* **2009**, *31*, 3−6.

(27) Cox, M. P.; Peterson, D. A.; Biggs, P. J. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinf.* **2010**, *11* (1), 485.

(28) Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* **1998**, *8*, 186.

(29) Nunn, B. L.; Slattery, K.; Cameron, K. A.; Timmins-Schiffman, E.; Junge, K. Proteomics of Colwellia psychrerythraea at subzero temperatures - a life with limited movement, flexible membranes and vital DNA repair. *Environ. Microbiol.* **2015**, *17*, 2319−2335.

(30) Kultima, J. R.; Sunagawa, S.; Li, J.; Chen, W.; Chen, H.; Mende, D. R.; Arumugam, M.; Pan, Q.; Liu, B.; Qin, J.; Wang, J.; Bork, P.

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705

MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS One* **2012**, *7* (10), e47656.
(31) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open source tandem mass spectrometry sequence database search tool. *Proteomics* **2013**, *13* (1), 22−24.
(32) Granholm, V.; Navarro, J. F.; Noble, W. S.; Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *J. Proteomics* **2013**, *80* (27), 123−131.
(33) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.
(34) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.
(35) 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.
(36) Saito, M. A.; Dorsk, A.; Post, A. F.; Mcilvin, M. R.; Rappé, M. S.; Ditullio, G. R.; Moran, D. M. Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **2015**, *15* (20), 3521−3531.
(37) Kertesz-Farkas, A.; Keich, U.; Noble, W. S. Tandem mass spectrum identification via cascaded search. *J. Proteome Res.* **2015**, *14* (8), 3027−3038.

DOI: 10.1021/acs.jproteome.6b00239
J. Proteome Res. 2016, 15, 2697−2705