OPEN

The ISME Journal (2017) 11, 309–314

www.nature.com/ismej

COMMENTARY

Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns

Emma Timmins-Schiffman, Damon H May, Molly Mikan, Michael Riffle, Chris Frazar, HR Harvey, William S Noble and Brook L Nunn

The ISME Journal (2017) **11,** 309–314; doi:10.1038/ismej.2016.132; published online 8 November 2016

Environmental meta-omics is rapidly expanding as sequencing capabilities improve, computing technologies become more accessible, and associated costs are reduced. The *in situ* snapshots of marine microbial life afforded by these data provide a growing knowledge of the functional roles of communities in ecosystem processes. Metaproteomics allows for the characterization of the dynamic proteome of a complex microbial community. It has the potential to reveal impacts of microbial metabolism on biogeochemical transport, storage and cycling (for example, Hawley *et al.*, 2014), while additionally clarifying which taxonomic groups perform these roles. Previous work illuminated many of the important functions and interactions within marine microbial communities (for example, Morris *et al.*, 2010), but a review of ocean metaproteomics literature revealed little standardization in bioinformatics pipelines for detecting peptides and inferring and annotating proteins. As prevalence of these data sets grows, there is a critical need to develop standardized approaches for mass spectrometry (MS) proteomic spectrum identification and annotation to maximize the scientific value of the data obtained. Here, we demonstrate that bioinformatics decisions made throughout the peptide identification process are as important for data interpretation as choices of sampling protocol and bacterial community manipulation experimental design. Our analysis offers a best practices guide for environmental metaproteomics.

MS-based metaproteomics is now practical due to advances in duty cycle and increased mass accuracy for both precursor and fragment masses. These improvements allow for the detection of over 10^4 tandem mass spectra from a single data-dependent acquisition MS analysis of a mixed microbial sample. These spectra must then be associated with peptides from thousands of proteins from diverse taxonomic groups. The most common approach is database searching: scoring observed tandem mass spectra against theoretical peptide spectra generated *in silico* from a protein or peptide database (Eng *et al.*, 1994). However, the approach to database selection, or construction, can vary dramatically. In an ocean metaproteomics experiment, the two main approaches for creating a protein identification database are to (1) leverage vast quantities of public sequence data or (2) sequence and assemble a metagenome. Further, when exploring and assembling possible public databases, a wide range of databases and sequence selection methods are used. As the field of environmental proteomics grows, the integrity of metaproteomics data sets and our ability to directly compare them across time and space depends on the adoption of a standardized procedure for peptide identification and annotation. Here, we reveal how highly influential the protein database selection is to the biological interpretations of a metaproteomics experiment.

We applied four database selection techniques in order to perform peptide detection, protein inference, and taxonomic and functional assignments from MS-based, oceanic, microbial community metaproteomics (Figure 1). The metaproteome in question represents a diverse and relatively undersequenced area of the ocean, the Pacific Arctic. Our results from this study offer a path forward as well as a caution for investigators that the biological conclusions drawn from metaproteomics data are highly database specific.

Our study followed traditional procedures currently employed in ocean metaproteomics (details in Supplementary Information 1). Water samples were collected and selectively filtered from the Bering Strait as described in May *et al.* (2016) and incubated shipboard over 10 days (T0 = day 0, T10 = day 10). Bacterial community proteomes from the incubations were analyzed on a Q-Exactive-HF (Thermo Fisher Scientific, Waltham, MA, USA) and resulting data were searched against four different peptide identification databases (Supplementary Information 2): (1) site/time-specific metagenome collected concurrently with the incubated water; (2) NCBI's env_NR database; (3) Arctic-bacterial database of NCBI protein sequences from known polar taxonomic groups (Supplementary Information 3) North Pacific database derived from a subset of the Ocean Microbiome sequencing project (Sunagawa *et al.*, 2015; Supplementary Information 4). Peptides were identified and proteins were inferred using Comet v. 2015.01 rev. 2 (Eng *et al.*, 2012, 2015), followed by

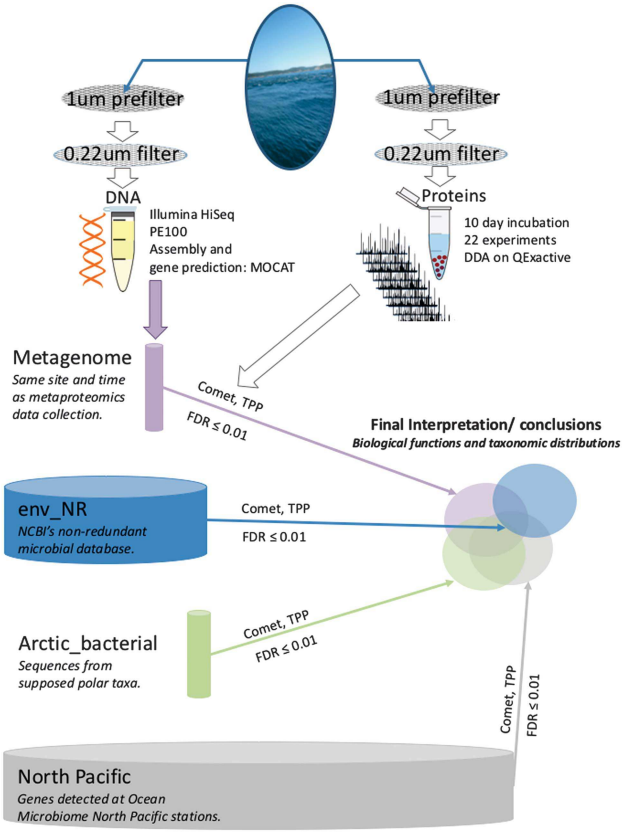

Figure 1 Schematic of the workflow for the database searches of the metaproteomics samples. The width of the cylinders depicting each database are scaled to the number of unique tryptic peptides in each database (Supplementary Information 4).

peptide and protein match scoring (Pedrioli, 2010; Deutsch *et al.*, 2015) at a false discovery rate threshold of 0.01 (Supplementary Information 5). Proteins from all databases were annotated using BLASTp (Altschul *et al.*, 1990; Camacho *et al.*, 2009) against the UniProtKB TrEMBL database (downloaded April 28, 2015) with an e-value cutoff of 1E-10 (Supplementary Information 6). Shifts in community biological functions over the 10-day incubation were quantified using a Gene Ontology (GO) analysis where peptide spectrum matches were associated with GO terms. Additionally, database-driven peptide score sensitivity as a function of database size was investigated by searching the site/time-specific metagenome database with increasing numbers of decoy peptides.

The number of peptide experimental spectra that yielded spectrum matches was very different among databases. The highest number of confidently scored unique peptide matches and protein inferences resulted from the search against the site/time-specific metagenome database. This number of peptide matches was augmented 1.5 times by searching the same data against unassembled reads. This 'metapeptide' approach (May *et al.*, 2016) avoids sequence loss and potential noise introduced by read assembly (for example Cantarel *et al.*, 2011). The peptides identified by the four assembled databases overlapped relatively little, suggesting that the different databases cover different parts of the acquired metaproteome (May *et al.*, 2016). In a direct comparison of the unassembled metagenome peptides and env_NR, the metagenome contained more peptides from the metaproteome (May *et al.*, 2016). Additionally, database size, especially in the cases of env_NR and North Pacific, had a substantial impact on search sensitivity, making statistically confident detection of peptides difficult (Supplementary Information 7; May *et al.*, 2016). In agreement with others, we found large database searches suffer from a loss of statistical power from multiple hypothesis testing against the vast number of sequences unrepresented in the expressed metaproteome (Nesvizhiskii, 2010; Jagtap *et al.*, 2013; Tanca *et al.*, 2013). This paradox of too many sequences resulting in too few identifications will become increasingly problematic with the availability of more sequence data. Our results point to the success obtained by searching a metaproteome-specific database that excludes non-specific sequences, while balancing the need to retain a sufficient amount of sequence variation.

Taxonomic and functional interpretations resulting from the different searches of the same metaproteome against different databases were divergent, suggesting that each database would yield a different biological conclusion. The four resulting community taxonomy profiles diverged even at the phylum level, and these differences were amplified at finer taxonomic levels (Figure 2). The metagenome also yielded a greater variety of taxa at ranks more specific than class compared to env_NR (May *et al.*, 2016). In addition to taxonomic discrepancies, functional response to the 10-day incubation differed depending on database used, differences that have been noted by others (Rooijers *et al.*, 2011; Tanca *et al.*, 2013). In our arctic microbiome, there was little agreement among database searches in the ten GO terms that changed the most between the beginning and end of the incubation experiment (Table 1, Supplementary Information 8). These GO terms would be considered the most significant contributors to changes in community function in the particular experiment, and would lead to substantially different interpretations depending on the database selected. The importance of these differences in functional assignments among search results can direct downstream analyses and interpretations. For example, they are of critical importance when inferring and reporting community function. Our results and others (for example, Rooijers *et al.*, 2011) stress the importance of database choice for metaproteomics functional assignments and community biological process, especially in the case of a previously uncharacterized, complex community.

In addition to differences in peptide search results, the true complexity in annotating detected proteins

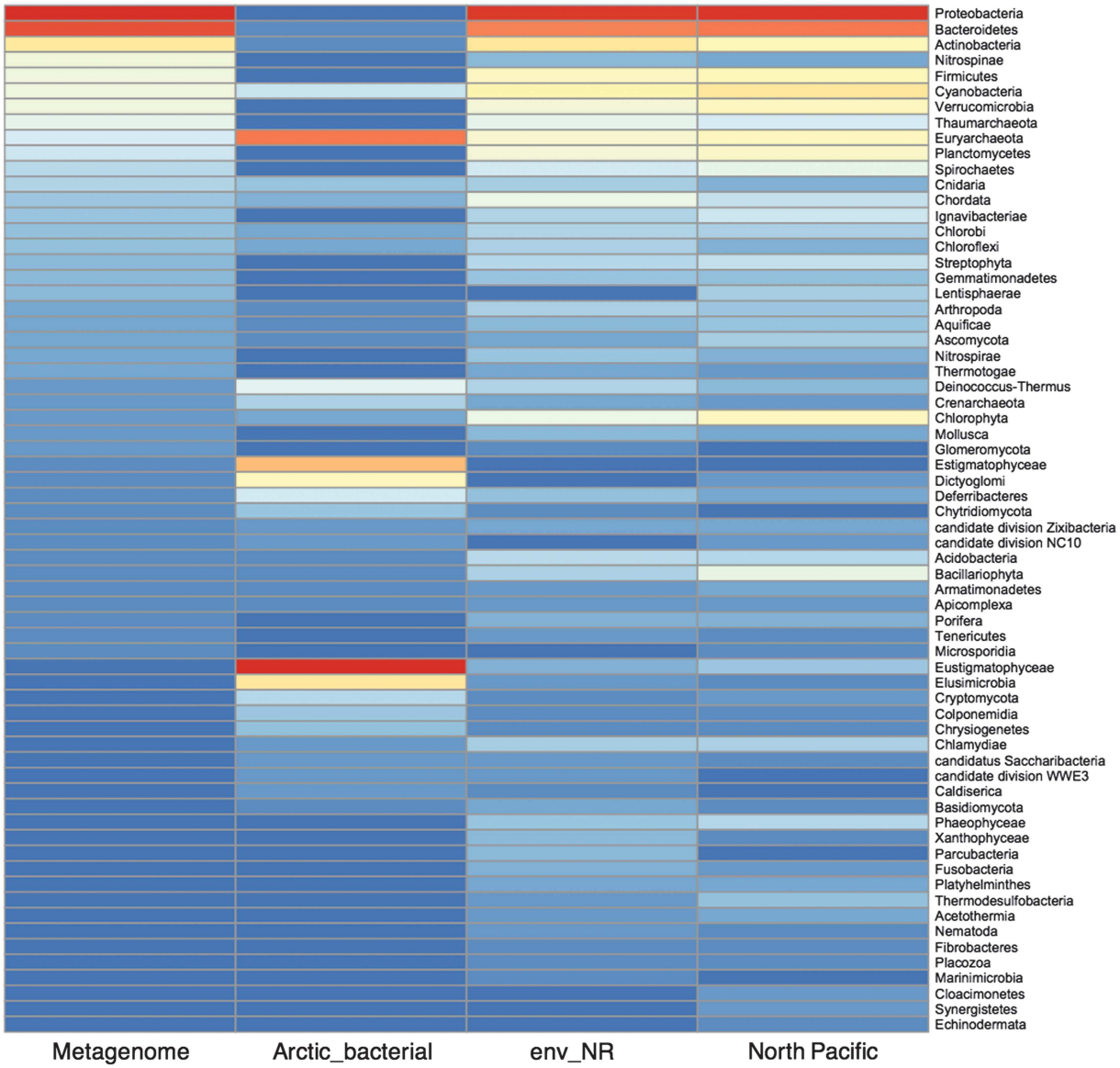

Figure 2 A heat map depicting the amount of agreement of taxonomic assignments at the phylum level derived from inferred proteins across searched databases. For each phylum, a colored box represents the number of proteins (log(x+1)-transformed) associated with that phylum for each set of search results (red = highly abundant phylum; blue = low/non-existent phylum). The results are ordered by phylum abundance in the site/time-specific metagenome search results.

was obscured by the standard approach that uses only the top BLAST hit as the defined protein annotation. The BLAST algorithm returns a list of possible hits with associated Expect values (e-values) when a sequence is searched; to better understand the downstream effects of this approach, we included up to 500 BLAST results per protein. On average, 403 protein matches per metagenome sequence were returned that passed the e-value cutoff of 1E-10. Disagreements in functional and taxonomic assignment among the BLAST hits for a single protein are very common, even when the results all have very low e-values (Supplementary Information 9). This casts doubt on the 'top' BLAST hit as the correct annotation for the protein of interest, even though this is common practice in 'omics' literature. Inaccuracy or lack of precision of protein annotation via BLAST methodology would further obscure an accurate interpretation of metaproteomics data when combined with an uninformed database choice.

The selection of a protein database for peptide identifications is one of the most critical bioinformatics decisions for accurate biological and ecological interpretation of *in situ* community functions. Although more time and money are required to complete a site/time-specific metagenome, we have demonstrated that these investments lay the groundwork for more complete metaproteome interpretation (Tanca *et al.*, 2013; May *et al.*, 2016). Whether or not a metagenome is assembled, data interpretation must proceed with care. Based on current and

Table 1 Ten GO terms with the biological aspect 'biological process' with the greatest log fold change from each database search; five that changed the most to have higher abundance at T10 (light gray) and five that have higher abundance at T0 (dark gray)

Rank	*Higher in*	*Site/time-specific metagenome*	*Arctic-bacterial*	*env_NR*	*North Pacific*
1	T10	Regulation of nitrogen utilization (3.8)	Sodium ion transport (3.6)	**Ammonium transport (3.5)**	Acyl-CoA biosynthetic process (3.3)
2	T10	**Ammonium transport (3.2)**	Alpha-amino acid catabolic process (3.2)	Tetrahydrofolate metabolic process (3.4)	Thioester biosynthetic process (3.3)
3	T10	Ammonium transmembrane transport (3.1)	**Ammonium transport (3.2)**	Tetrahydrofolate interconversion (3.4)	Acetyl-CoA biosynthetic process (3.1)
4	T10	Monocarboxylic acid catabolic process (3.1)	Organonitrogen compound catabolic process (3.1)	Acyl-CoA biosynthetic process (3.3)	Taurine catabolic process (2.9)
5	T10	Pyrimidine-containing compound biosynthetic process (2.8)	Glutamate metabolic process (3.0)	Alpha-amino acid catabolic process (3.3)	Taurine metabolic process (2.9)
1	T0	Cobalamin biosynthetic process (2.6)	Microtubule-based process (3.9)	Carbon fixation (5.8)	Photosynthesis (4.1)
2	T0	Cobalamin metabolic process (2.6)	Positive regulation of biological process (3.6)	**Photosynthetic electron transport chain (4.0)**	Microtubule-based process (3.4)
3	T0	Cellular ketone metabolic process (2.6)	Positive regulation of metabolic process (3.6)	Photosynthesis, light reaction (3.5)	Transcription-coupled nucleotide-excision repair, DNA damage recognition (3.2)
4	T0	Cytoplasmic transport (2.0)	Positive regulation of gene expression (3.6)	Microtubule-based process (3.5)	Nucleotide-excision repair, DNA damage recognition (3.2)
5	T0	**Photosynthetic electron transport chain (2.0)**	Positive regulation of macromolecule metabolic process (3.6)	Photosynthetic electron transport in photosystem II (3.3)	**Photosynthetic electron transport chain (2.7)**

Abbreviation: GO, gene ontology.
Bold terms are matches with the top 10 terms from the metagenome. Fold changes for each term are in parentheses next to the term name.

previous work, we propose a general best practices guide (Figure 3) to identifying peptides and inferring biological function and taxonomic distributions of natural microbial assemblages: (1) For previously uncharacterized communities, construct as accurate and efficient a database as possible by (a) using the metapeptide approach (May *et al.*, 2016), (b) sequencing the metagenome and utilizing gene prediction software (for example, Hyatt *et al.*, 2012) or (c) constructing the most accurate database possible to avoid loss of sensitivity due to large search space when metagenome sequencing is not possible; (2) when annotating proteins, go beyond the top BLAST hit to base the annotation for taxonomy and function on an agreement among BLAST hits above a specific e-value threshold (Supplementary Information 9); (3) to increase peptide identifications, leverage publicly available sequences via the more statistically robust multistep or iterative searches (for example, Jagtap *et al.*, 2013; Kertesz-Farkas *et al.*, 2015). As researchers begin to explore these different search methods with a variety of metaproteomics data sets, this approach will provide the most robust search methods and most reliable taxonomic and functional inference for environmental metaproteomics.

Supplementary Information is available at ISME Journal's website.

1: Detailed methods for metagenome sequence, metaproteomics MS, database searching and biological interpretation of data.

2: Minimum, maximum and mean protein lengths for each protein identification database used in this study.

3: Taxonomic groups used to create the Arctic-bacterial database. The first and second columns list the group name and taxonomic level for the protein sequences that were downloaded from NCBI, followed by the complete taxonomic tree. For each taxonomic group, citations are given from peer-reviewed literature that were used to infer this group's presence near our study site. The second tab in the workbook has the full citations listed.

4: Summary of total unique protein and peptide sequences in each database. Three different scenarios are given for peptide sequence generation to construct a database: (1) 3 missed cleavages and oxidized methionine; (2) 0 missed cleavages, no oxidation; (3) 3 missed cleavages, no oxidation.

5: Comet parameter file used to run all database searches.

6: Query protein, top UniProt BLAST hit and corresponding e-value are provided for all proteins detected with high confidence. Protein lists for the different database searches can be found in the different sheets of the Excel workbook.

7: Adding large numbers of random decoy peptides to an 11-million-peptide metagenome-derived database depressed peptide detection sensitivity. Horizontal axis is the number of peptides in each search database (11 million metagenome peptides,

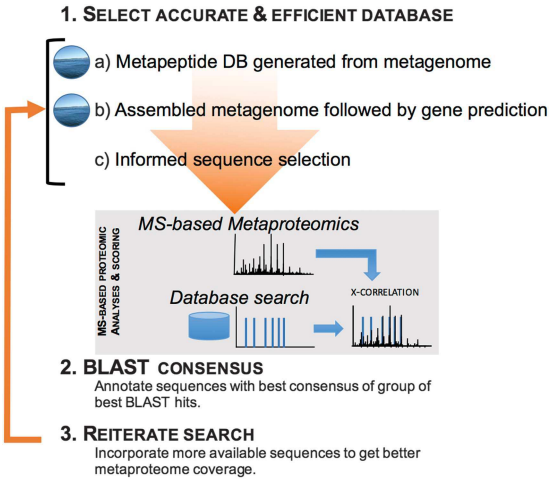

Figure 3 Depiction of the recommended 'best practices in metaproteomics' workflow. The ocean circles represent data derived from the same sample. (1) Selection of an accurate and efficient database is followed by (2) finding the consensus BLAST hit among the group of best hits, and (3) re-searching the data against more sequences to achieve greater metaproteome coverage using a robust multi-step or iterative algorithm.

with increasing numbers of random decoy peptides). The vertical axis is the number of metagenome peptides detected at a false discovery rate of 0.01 as determined by forward–reverse database search for five different sample files. False discovery rate was calculated from Trans Proteomic Pipeline probabilities.

8: Direction of $\log_2$ fold change for GO terms detected at total PSM count > 50 in T0 vs T10, T0' vs T10', T0 vs T0', and T10 vs T10' (' ' 'denotes a technical replicate). A $\log_2$ fold change >1 is 'positive', <-1 is 'negative', between -1 and 1 is 'none' and if a GO term was not detected at above 50 PSM in a database there is an 'X'. Results for each database (site/time-specific metagenome, env_NR, Arctic-bacterial and North Pacific) are listed in separate columns for each comparison.

9: A heatmap representing the granularity of taxa returned from a BLAST search (e-value ⩽ 1E-10) as a function of percent identity threshold. Each colored bin represents the number of protein hits at a given least common taxonomic unit level for up to 500 protein hits. Horizontal axis: minimum percent sequence identity between query protein and BLAST hits. Vertical axis: rank of the lowest common taxonomic unit representing all BLAST hits above the threshold. Color indicates the natural log of the number of query proteins that fall into each bin, according to the scale at right. 'None' indicates hits that were assigned to multiple superkingdoms.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported and funded by a grant from the National Science Foundation (NSF-OCE 1233014) for ETS, BLN, DHM and MPM as well as a Training Grant from the National Institutes of Health for ETS (T32 HG00035). DHM and WSN were supported by the National Institute of General Medical Sciences of the NIH under award number P41 GM103533. Microbial community sampling was supported through the Bureau of Ocean Energy Management (BOEM—Hanna Shoal Ecosystem Study) to HRH. This work is supported in part by the University of Washington's Proteomics Resource (UWPR95794). We thank Jimmy Eng for aiding with database searching and bioinformatics; Jody Wright for advice on DNA extraction; Marcos Perez and Marsha Wheeler for assistance with metagenome sequencing; Ohad Manor for his help with annotations; Brian Searle for writing the code to download sequences from NCBI; Jarrett Egertson and the UW Genome Sciences Information Technology team for their assistance with data analysis; and Luis Pedro Coehlo for his advice and help with subsetting the Ocean Microbiome data set. BLN and ETS would like to thank TAN and IJE for their ongoing inspiration.

E Timmins-chiffman, DH May, C Frazar, WS Noble and BL Nunn are at University of Washington, Department of Genome Sciences, Seattle, WA, USA
HR Harvey and M Mikan and at Old Dominion University, Department of Ocean, Earth, and Atmospheric Sciences, Norfolk, VA, USA
M Riffle is at University of Washington, Department of Biochemistry, Seattle, WA, USA
WS Noble is at University of Washington, Department of Computer Science and Engineering, Seattle, WA, USA
E-mail: emmats@uw.edu or brookh@uw.edu

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421–430.

Cantarel BL, Erickson AR, VerBerkmoes NC, Erickson BK, Carey PA, Pan C *et al.* (2011). Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One* **6**: e27173.

Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. (2015). Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl* **9**: 745–754.

Eng JK, Hoopmann MR, Jahan TA, Egertson JD, Noble WS, MacCoss MJ. (2015). A deeper look into Comet–implementation and features. *J Am Soc Mass Spectrom* **26**: 1865–1874.

Eng JK, Jahan TA, Hoopmann MR. (2012). Comet: an open source tandem mass spectrometry sequence database search tool. *Proteomics* **13**: 22–24.

Eng JK, Mccormack AL, Yates JR. (1994). An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J Am Soc Mass Spectrom* **5**: 976–989.

Hawley AK, Brewer HM, Norbeck AD, Pasa-Tolic L, Hallam SJ. (2014). Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *PNAS* **111**: 11395–11400.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. (2012). Gene and translation initiation site prediction in metagenome sequences. *Bioinformatics* **28**: 2223–2230.

Jagtap P, Goslinga J, Kooren JA, McGowen T, Wroblewski MS, Seymour SL *et al.* (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomic and proteogenomic studies. *Proteomics* **13**: 1352–1357.

Kertesz-Farkas A, Keich U, Noble WS. (2015). Tandem mass spectrum identification via cascaded search. *J Proteome Res* **14**: 3027–3038.

May DH, Timmins-Schiffman E, Mikan MP, Harvey HR, Borenstein E, Nunn BL, Noble WS. (2016). Metaproteomics characterization of microbiome samples by translating shotgun metagenomic reads. *J Proteome Res* **15**: 2697–2705.

Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4**: 673–685.

Nesvizhskii AI. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**: 2092–2193.

Pedrioli PGA (2010). Trans-Proteomic Pipeline: A pipeline for proteomic analysis. *Proteome Bioinformatics* **604**: 213–238.

Rooijers KK, Kolmeder C, Juste C, Doré J, de Been M, Boeren S *et al.* (2011). An iterative workflow for mining the human intestinal metaproteome. *BMC Genomics* **12**: 6.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al.* (2015). Structure and function of the global ocean microbiome. *Science* **348**: 1261359.

Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G *et al.* (2013). Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* **8**: e82981.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)