Ocean Proteomics Data Sharing and Best Practices Workshop
Final Meeting Report

Convened May 3-5th 2017 at the Woods Hole Oceanographic Institution, Woods Hole MA
Report authored by Mak Saito and Danie Kinkade
msaito@whoi.edu and dkinkade@whoi.edu

Summary

Ocean metaproteomics is an exciting new datatype that has the potential to provide a myriad of valuable new insights into the biogeochemical functions of marine microbes throughout the oceans and their impact on ecological and chemical processes. A community workshop was organized to discuss and explore solutions to the challenges specific to data sharing of these ocean metaproteomic datasets. This workshop was held in May of 2017 with a diverse group of proteomic scientists, data scientists, and computer programmers, the latter groups associated with the Biological and Chemical Data Management Office and the development team of the EarthCube Ocean Protein Portal. The group identified areas that present challenges to data quality control and intercompatibility, including diverse data types and diversity and lack of standard approaches to informatic data processing. The group also recognized the important need for a metaproteomic intercalibration effort and demonstrated a willingness to organize and participate in a future intercalibration and in the development of intercalibration standards. The value of the future ocean protein portal, and the sustainability considerations in balancing capabilities with managing costs were also discussed. Finally, given that many participants had never met before, this workshop served as an important community-building effort for this nascent scientific community.

1. Introduction and purpose

As part of the EarthCube project “Laying the Groundwork for an Ocean Protein Portal”, a community workshop was organized and held in Woods Hole between May 3-5th 2017. For three days, proteomic domain scientists (from ocean, terrestrial and human metaproteomic research), data scientists, and computer programmers met to discuss the topic of challenges and best practices regarding the sharing of metaproteomic datasets from ocean and aquatic environments. Twenty two attendees participated in the conference from the US and Canada (see Figure 1 and attached Attendees list) and the agenda consisted of short talks, discussions and presentations of the design concept for the prototype EarthCube Ocean Protein Portal currently being designed and built at WHOI (see attached Agenda). The discussions centered on four topics: 1) relevant proteomic data types, 2) informatic challenges associated with processing, post-processing, and quality control, 3) specific details of sharing metaproteomic datasets, and 4) the role, sustainability, and data use policies for a future ocean protein portal and the community.

The measurement of many proteins within oceanic microbial communities, known as ocean metaproteomics, is a technique that is great interest to oceanographers and protein scientists. The potential ability to examine the functional attributes of these communities and their linkages to both ecology and biogeochemistry is particularly appealing as a means to better understand how these systems operate and respond to environmental change. However, there are numerous challenges facing the application of proteomic methods to environmental contexts. Primary among these is that by definition the ocean and other environmental contexts contain a multitude of organisms that are not easily separated, and hence are typically studied together in a ‘meta’ context. For example, in a typical ocean seawater sample, the microbial biological diversity includes prominent communities from each of the three major domains of life as well as from viruses. This natural biological diversity manifests itself in a tremendous chemical complexity for a proteomics analysis, where proteins from many organisms are typically lysed and digested into mass spectrometry peptides and analyzed together. The new generations of mass spectrometry instrumentation have combined blazing scanning speeds and high-resolution mass accuracy to allow deep interrogation of these complex samples as never before possible. With this combination of biological and chemical complexity, advances in instrumentation, and the resulting need for ‘big data’ analysis and interpretation, there is significant room for method development and identification of best practices throughout the data collection, analysis, and sharing process.

2. Summary of discussion/findings

Over the course of the workshop there was a vigorous discussion focused on topics pertaining to challenges in producing and verifying high data quality, and challenges facing effective data sharing for proteomics results, and the current Ocean Protein Portal design proposed by the Ocean Metaproteomics Portal team. These discussions culminated in a whiteboard diagram of challenges facing metaproteomics research, which was subsequently made into a graphic for a proposed best practices manuscript (Figure 2; see below).

On the topics of data quality and sharing for metaproteomics, many topics were discussed. These included the challenges facing proteomics with regards to different data types and incomparability, usage of different genomic and metagenomic databases, the challenge of protein inference in metaproteomic settings, the constraints on peptide identification confidence, workflow reproducibility, necessary metadata for environmental and ocean metaproteomic datasets, and opportunities for standardization and intercalibration. In addition the pros and cons of different data usage policies were discussed in order to both encourage submission and usage of shared data.

One item of extended discussion was use of single peptides in metaproteomics. Single peptides from a protein have historically been discouraged for use in protein identifications in proteomics informatic workflow, yet for metaproteomics on environmental samples the available genomic and metagenomic may in many cases not be sufficiently deep to allow identification of multiple peptides from specific proteins (for example when those proteins are unknown) or there may be a population of protein diversity with co-existing related peptides. Hence the group agreed that, given the improvements in high resolution mass spectrometry and peptide identification and the complexities of protein inference in diverse metaproteomic samples, allowing the use of single peptides for protein identifications should be considered a useful tool for protein identifications and quantification in metaproteomics.

In a related discussion, the challenges of protein inference in an environmental population that contains a diversity of closely related sequences was discussed at length, and how connections to metagenomic resources influences this effort both by increasing proteome depth, but also in creating difficulties with peptide-to-spectrum matching algorithms.

On the second topic regarding feedback on the current design concept for the EarthCube Ocean Protein Portal, there was a significant discussion generated with workshop participants. Feedback received included creating connections to non-environmental mass spectrometry repositories (in particular ProteomeXchange), discussing the features and capabilities of the portal such as incorporating a spectra viewer and analysis capability, connecting and collaborating with workflow editors to facilitate data production such as Galaxy-P, and policies for data submission and use.

There was significant discussion about the quality of informatic pipelines to produce the peptide and protein inferences in complex metaproteomic samples. There was a lively debate about whether the scope of the portal should be expanded to allow users to examine individual spectra associated with peptides to directly assess peptide quality. This discussion weighed the benefits of visual inspection of the quality of peptide-to-spectrum match assignments versus the large logistical and sustainability challenges associated with expanding the portal scope to include to spectra analysis. The potential using external tools with raw files was also discussed as an alternative to this use case

Workshop participants expressed interest in the Metatryp software capability that is being updated from a previous version as part of the EarthCube Protein Portal project. Metatryp is a Python/SQL program that allows a user to determine the taxonomic group(s) a peptide of interest for targeted metaproteomics is found in. A new web version of Metatryp was demonstrated and is now able to ingest metagenomes in addition to the previous genome files, providing greater environmental relevance to the oceans. This feedback for community interest in a standalone Metatryp web capability was a welcome surprise, and the portal team has begun scoping and development plans for a product within the EarthCube project.

Finally, but certainly not least, this meeting served as an important community building event for the North American metaproteomics community, where basically all of the participants had not previously met some of the other participants at the meeting due to residing in different academic circles. It was hoped this effort could serve as the beginning for future meetings on the topic of mutual interest: measuring proteins in complex environments.

3. Significant outcomes

The workshop participants discussed and agreed on a number of characteristics that could constitute best practices in data sharing for ocean protein data including metadata types, required data files needed, availability and documentation of informatic pipeline workflows to generate these files, data use policies, and connections to other repositories. These will be described in greater detail in a workshop best practices document specific to ocean metaproteomics that would be submitted for peer-review publication. It was considered that the document would serve several functions including setting the quality control and standards expectations for ocean metaproteomic data sharing that an Ocean Protein Portal could utilize, In addition this document could provide community feedback on recommended policies for data sharing and use that would promotes both submission and fair use. Finally this document could provide a much needed reference document that could facilitate fair peer review of ocean metaproteomic data in the literature.

The participants also agreed that a future community effort and meeting to conduct intercalibration exercise as well as to develop best practices of informatic approaches would be beneficial for this young scientific community, and that we would look for opportunities to organize such an effort in the coming year(s).

4. Proposed next steps

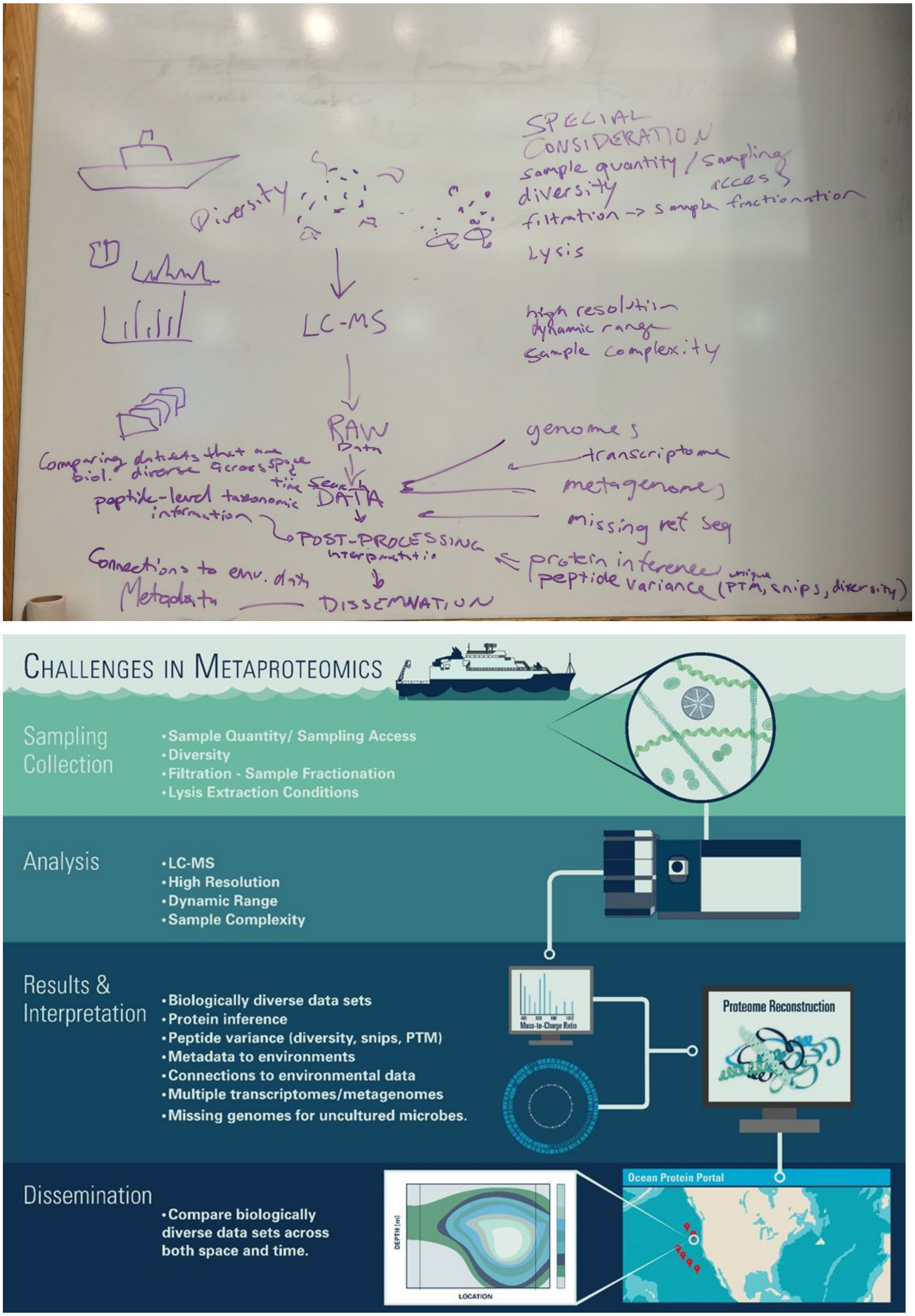
The group agreed that writing a publication on recommended best practices for data sharing in ocean proteomics would be a beneficial document for the community and committed to jointly authoring the document. An outline and assignments for this document were produced. The participants agreed that to try to author this short manuscript in a relatively short time frame in order to maintain momentum from the conference. Some of the potential journals to be considered for submission include Journal of Proteomics Research, Nature Microbiology, Frontiers in Marine Biogeochemistry, and Limnology and Oceanography Methods.

The group also expressed strong interest in future efforts in intercalibration of measurements and development of reference standards, including involvement with scientists at NIST. Based on this interest avenues for supporting a marine proteomics intercalibration effort will be explored, perhaps in concert with those of other ‘omics communities such as metabolomics. Connections will be made with the scientists operating ProteomeXchange, which is a portal that connects various proteomics repositories to make their data discoverable to a broader community. Results of this workshop will be presented at the EarthCube All Hands Meeting, Chemical Oceanography Gordon Conference in the summer of 2017, and the American Society for Mass Spectrometry in 2018, which is the major international proteomics meeting.

5. Acknowledgements

The Workshop and Workshop report were supported by an EarthCube Grant 1639714 to M. Saito and D. Kinkade. We also gratefully acknowledge meeting support from the Woods Hole Oceanographic Institution and an anonymous donor.

Figure 1. Participant photo for the Ocean Protein Data Sharing Workshop in Woods Hole May 3-5, 2017.



Sampling Collection

- Sample Quantity/ Sampling Access
- Diversity
- Filtration - Sample Fractionation
- Lysis Extraction Conditions

Analysis

- LC-MS
- High Resolution
- Dynamic Range
- Sample Complexity

Results & Interpretation

- Biologically diverse data sets
- Protein inference
- Peptide variance (diversity, snips, PTM)
- Metadata to environments
- Connections to environmental data
- Multiple transcriptomes/metagenomes
- Missing genomes for uncultured microbes.

Dissemination

- Compare biologically diverse data sets across both space and time.

Figure 2. Collaborative whiteboard sketch of overview figure (top), and final product (bottom) for use in a Workshop Best Practices Manuscript.

6. Appendices

Participants:
Ocean Protein Portal Project Team (WHOI): Mak Saito, Danie Kinkade, Noelle Held, Matt Mcilvin, Dawn Moran, Nick Symmonds, Adam Shepherd, David Gaylord

External Advisors for Workshop Agenda and Planning: Michael Janech, Dasha Leary and Judson Hervey

Full Participant List:
Erin Bertrand
Assistant Professor, Tier II CRC Chair in Marine Microbial Proteomics
Department of Biology
Dalhousie University
1355 Oxford St.
P.O. Box 15000
Halifax, Nova Scotia, Canada
B3H 4R2 LSC 5076B
Phone: 902 494-1853
Erin.Bertrand@Dal.Ca

Megan Duffy
Graduate Student
School of Oceanography
University of Washington
Box 355351
Seattle, WA 98195
Phone: 802-279-8715
duffyme@uw.edu

David Gaylord
Information Systems Associate II
Woods Hole Oceanographic Institution
266 Woods Hole Rd., MS# 46
Woods Hole, MA 02543-1050
Phone: 508-289-3271
dgaylord@whoi.edu

Noelle Held
Graduate Student
Woods Hole Oceanographic Institution
266 Woods Hole Rd., MS# 51
Woods Hole, MA 02543-1050

Phone: 508-289-3994
nheld@whoi.edu

Judson Hervey
Research Biologist
Laboratory for Bio/Nano Science & Technology
(Code 6910)
Naval Research Laboratory
4555 Overlook Avenue - SW
Washington, DC 20375
Phone: 202-767-0599 - Office
judson.hervey@nrl.navy.mil

Robert(Bob) Hettich
Research Scientist
Organic and Biological Mass Spectrometry Group
Chemical Sciences Division
Oak Ridge National Laboratory and
Microbiology Department, University of Tennessee
PO Box 2008
Oak Ridge, TN 37831
Phone: 865.574.4968
hettichrl@ornl.gov

Pratik Jagtap
Research Assistant Professor
Department of Biochemistry, Molecular Biology and Biophysics
University of Minnesota
7-166 MCB, 420 Washington Ave SE,
Minneapolis, MN 55455
Phone: 612-624-0381
pjagtap@umn.edu

Michael G. Janech
Associate Professor
Director - Nephrology Proteomics Laboratory
Medical University of South Carolina
96 Jonathan Lucas St.
829 CSB - Nephrology
MSC 623
Charleston, SC 29425
Office address:
Strom Thurmond Building - Room 625

114 Doughty St.
Charleston, SC 29403-5729
Office: 843 792 1083
janechmg@musc.edu

Danie Kinkade
Information Systems Associate III
Biological and Chemical Oceanography Data Management Office
Woods Hole Oceanographic Institution
266 Woods Hole Rd., MS# 36
Woods Hole, MA 02543-1050
Phone: 508 289 2291
dkinkade@whoi.edu

Dasha Leary
Research Biologist
Naval Research Laboratory
Center for Bio/Molecular Science & Engineering Bldg. 30 / Code 6920
4555 Overlook Avenue - SW
Washington D.C. 20375
Tel: 202-404-6055
dasha.leary@nrl.navy.mil

Matt McIlvin
Research Associate III
Woods Hole Oceanographic Institution
266 Woods Hole Rd.
MS# 51
Woods Hole, MA 02543-1050
Phone: 508 289 2884
mmcilvin@whoi.edu

Eli Moore
Post Doctoral Researcher
Rutgers University
71 Dudley Rd., Room 303A
New Brunswick, NJ 08901
Phone: 848-932-3450
moore@marine.rutgers.edu

Robert Morris
University of Washington
School of Oceanography

Box 357940
Seattle, WA 98195-7940
206-221-7228
morrisrm@uw.edu

Benjamin Neely
Research Chemist
Marine Biochemical Sciences Group (CSD/MML)
National Institute of Standards and Technology
Hollings Marine Laboratory
Charleston, South Carolina, USA
Phone: (843) 762-8999
Ben.neely@noaa.gov

Brook Nunn
Dept. of Genome Sciences
University of Washington
Foege Building S113
3720 15th Ave NE
Seattle, WA 98195
www.environmentalproteomics.org
Phone: 206-200-5871
brookh@uw.edu

Mak Saito
Associate Scientist with Tenure
Woods Hole Oceanographic Institution
266 Woods Hole Rd., MS# 51
Woods Hole, MA 02543-1050
Phone: 508-289-2393
msaito@whoi.edu

Jaclyn K. Saunders
Post Doc
Center for Environmental Genomics
University of Washington, School of Oceanography
School of Oceanography
Seattle, WA 98195-7940
Phone: 610-704-8649
jaclynk@uw.edu

Adam Shepherd
Information Systems Associate III

Woods Hole Oceanographic Institution
266 Woods Hole Rd., MS# 36
Woods Hole, MA 02543-1050
Phone: 508-289-2772
ashepherd@whoi.edu

Nick Symmonds
Manager of Applications
Information Services
Woods Hole Oceanographic Institution
266 Woods Hole Rd., MS# 46
Woods Hole, MA 02543-1050
Phone: 508-289-3114
nsymmonds@whoi.edu

David Walsh
CRC Microbial Ecology and Genomics
Associate Professor
Biology Department
Concordia University
7141 Sherbrooke St. West
Montreal, QC
H4B 1R6
514-848-2424 ext 3477
david.walsh@concordia.ca

Meeting Agenda

Agenda for Ocean Proteomics Data Sharing Meeting - May 3-5th

Tuesday May 2nd
Arrive in Falmouth MA, Inn on the Square Hotel
Dinner on your own - informal drinks/social for those in town (Liam's McGuire's)

Wednesday May 3rd
8:20 Pick up at the Inn on the Square by Mak, Danie, Adam, Noelle
8:30-9:00 Breakfast - Clark 5th floor
9:00-9:05 Around the Room Introductions
9:05-9:25 Welcome, Logistics, and Meeting Objectives - Mak Saito
9:25-9:40 Introduction to EarthCube and BCO-DMO - Danie Kinkade
9:40-10:00 Homologous proteins in metagenomic searches - Bob Morris
10:00-10:20 Metaproteomic Workflows in Galaxy-P - Pratik Jagtap
10:30-10:50 Coffee Break
10:50-11:20 Summary of Current Ocean Protein Portal Workflow and Design
11:20-12:00 Discussion on Portal Introduction
12:00-1:15 Lunch
1:15 -1:30 Group Photo Clark Balcony
1:30-1:50 Databases in Metaproteomics - Brook Nunn
1:50-3:00 Discussion #1 - Moderator – Mike Janech

1. What are the data types that should/could be shared?
2. What is the role(s) of an ocean protein portal relative to NIH/EBI supported repositories?
3. How will data types evolve as proteomics evolves?
4. How can mis-interpretation of data by non-expert users be avoided?

3:00-3:30 Coffee Break
3:30-5:00 Discussion #2 - Moderator Dasha Leary

1. What are metaproteomic challenges in protein inference?
2. What kinds of proteomics quality control are possible?
3. How can data sharing accommodate future improvements in methodologies?

5:00-5:30 Individual writing contributions to product report discussions
5:30 Depart for Hotel
7:00 Dinner: Walk to La Cucina on Main Street Falmouth

Thursday May 4th
8:20 Pick up at the Inn on the Square
8:30-9:00 Breakfast- Clark 5th floor
9:00-9:20 Reflections on prior day's discussions, recalibrations for meeting products
9:20-9:40 NIST in Standardizing Measurement Science - Ben Neely
9:40-10:00 Challenges in Interoperability in data sharing - Adam Shepherd

10:00-10:15 Biogeotraces peptide nonmenclature submission experience – Mak Saito
10:15-10:30 Short Talk Discussing HUPO Standards
10:30-10:45 Coffee Break
10:45 - 12:15 Discussion #3 – Moderator Danie Kinkade

1. What are the metadata needs/requirements for documenting protein datasets?
2. What are useful/appropriate naming schemes for biomarkers, proteins, peptides?
3. Could of intercalibrations and certified standards could be created for marine proteomics?
4. What challenges confront and solutions toward producing sharable datasets

12:15-1:00 Lunch
1:00-1:20 Advances in metagenomics and connections to proteomics - Erin Bertrand
1:20-1:40 Revisiting the Ocean Protein Portal Design and Metatryp 2.0 - Mak Saito and David Gaylord
1:40-2:00 Short Coffee break
1:55-3:15 Discussion #4 – Moderator Noelle Held

1. How can data submission be encouraged and facilitated?
2. What guidelines should be made on acknowledging/attributing shared data?
3. How can an ocean proteomics repository connect with genomics and non-marine mass spectrometry data centers
4. Can connections be imagined for future methods (e.g. metabolomics)?
5. How can environmental based ‘omics portals be designed to be sustainable?

3:15-3:30 Coffee Break
3:30-3:50 Individual writing, offline discussions or continued conversation
4:00-5:00 Tour of the Woods Hole village AUV Clio and Dock
5:15 Drinks at Landfall Restaurant, Woods Hole
7:00 Dinner Landfall Restaurant, Woods Hole

Friday May 5th
8:20 Pick up at the Inn on the Square
8:30-9:00 Breakfast
9:00-9:30 Discuss progress towards meeting goals, goals for future meeting(s).
9:30-10:45 Wrap up thoughts and discussion, writing assignments
9:45-10:30 Meeting outputs, report writing and discussions
10:30-10:45 Coffee Break
10:45-12:00 Meeting outputs, report writing and discussions
12:00 Bag lunches and meeting end

Outline for Best Practices Manuscript

1. Challenges unique to metaproteomics - Mak/Dasha/Noelle
    a. Diversity and number protein varies between samples
    b. Role of proteomics in the realm of Big Data
    c. Lack of biological replicates in environment, put forward concept of environmental / oceanographic consistency, Noelle
    d. Mapping to multiple genomes/metagenomics, what is appropriate database - Judson, Brook
    e. Inability to standardize, due to diversity (what is the standard?)
    f. Challenges of normalization in complex samples
    g. Challenges of sample extraction in complex env samples - Eli
    h. Challenge of characterizing natural diversity of protein familiies (Homologous protein), Bob/Brook/Erin
    i. Challenges of acquiring accurate annotations in genomes/metagenomes- Jaci, David W., Pratik
        i. No means to accumulate manual curations
        ii. Different nonmenclature
        iii. Metagenomic resources constantly changing

    https://img.jgi.doe.gov/cgi-bin/m/main.cgi
    https://metacyc.org/
2. Data: Spectral Counts, Precursors Intensities, Targeted; Mak
3. Recommendations for best practices in data analysis/acquisition Erin/Brook/Megan/Mak
    a. Not prescriptive, a best practice
    b. Recommendation for high resolution instruments
    c. Single peptides
    d. Express the room to evolve methods
    e. Don't want to be too restrictive
    f. Encourage documentation and sharing of database construction/resources
        i. Be wary of challenges in protein inference, databases that are not representative
        ii. Cross references to NCBI page to obtain sequence data
4. Metadata needed for data sharing - required/no, unit, datatype(str, integer, identifier): -Noelle/Danie
    a. Sampling
        i. Geospatial information (required)
        ii. Connections to environmental, connections to external repositories (string)
        iii. Basic hydrography as part of metadata submission (T, S, chl, O2)
        iv. Habitat type (water column, sediment, wetlands, lakes, host-associated microbiomes) (IMG ontology)
        v. Sampling type: filter type, sediment trap, dissolved
        vi. Expeditions, lab,

vii. Other analytes analyzed co-located

b. Data acquisition
   i. Sample prep - adopt standards?
      1. reducing/alkylating
      2. digestion enzyme
   ii. Standards
      1. External standards
      2. Targeted standards
   iii. Acquisition
      1. Instrument
      2. Mass accuracy MS1 and MS2
      3. Activation method
      4. Chromatography details
      5. Experiment type (DDA, DIA, SRM/MRM)

c. Data analysis - Pratik
   i. Document workflow
   ii. Database type (metagenome, genome, metatranscriptome, custom)
      1. Link to fasta - can be included in ProteomeXchange (unique identifiers for file)
   iii. Search engine
   iv. Recommend moving forward testing with contaminant database to false (which ones? GPM-CRAP or marine specific?)
   v. Recommend deposition of raw and search database files into established repository. Describe what a good repository is.

http://www.proteomexchange.org/

5. Encouraging proper use
   a. Statistics on peptide level - Brook/Pratik/Dasha/Matt M.
      i. Single hits with mass accuracy, multiple sample observation
   b. Peptide quality metric Brook/Pratik/Dasha/Matt M.
      i. [Percent b and y ions]
      ii. Other metrics - Confident, Doubtful
      iii. Best scoring PSM?
   c. Potential challenges Mak/Bob H./Ben
      i. Comparing relative quantitative units across different datasets (presence OK)
         1. Standards for comparability
            a. internal/external
            b. general/specific
         2. Spectral ct thresholds?
   d. Data use policy - Jaci/David/Danie/Noelle
      i. Warning message of comparing relative datasets
      ii. May want to contact data generator(s)
      iii. Automatic Citation descriptors

iv. Open licences options at submission time due to institutional
v. Document the need and value of reanalyzed metaproteomic datasets (e.g. as genomic resources expand)
vi. Examples of good (use cases for non-proteomic scientists) and bad data use (vignettes, pitfalls)
1. Normalized spectral count example
2. Overloading of trypsin in normalized spectral counts

e. Need/Niche for a environmental/ocean portal/repository

6. Recommendations for improvements in data quality - Ben, Mike, Mak
a. Development of internal and external standards
b. Reference datasets
c. Intercalibration efforts
d. Improvements in metagenomic resources/standardization
e. Workflow standardization/reproducibility’
f. Development of metaproteomic capable metrics and benchmarks