Advanced MS Analysis: Metaproteomics

Pratik Jagtap
University of Minnesota

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo

International Mass Spectrometry Conference
28th August, 2022
Workshop acknowledgements

• **Instructors**
  • Pratik Jagtap

• **Other contributors**
  • Tim Griffin
  • Subina Mehta
  • James Johnson
  • Andrew Rajczewski
  • Reid Wagner
  • Katherine Do
  • usegalaxy.eu team
  • Galaxy community

---

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html
Mass Spectrometry and Proteomics
Microbiome
IN NUMBERS

100 Trillion
symbolic microbes live in and on every person and make up the human microbiota

95%
of our microbiota is located in the GI tract

150:1
The human body has more microbes than there are stars in the Milky Way

90%
It is thought that 90% of disease can be linked in some way back to the gut and health of the microbiome

5:1
Viruses:Bacteria in the gut microbiota

You have 1.3X more microbes than human cells

The surface area of the GI tract is the same size as 2 tennis courts

>10,000
Number of different microbial species that researchers have identified in and on the human body

2kg
The gut microbiota can weigh up to 2kg

Each individual has a unique gut microbiota, as personal as a fingerprint

https://worldmicrobiomeday.com/resources/
https://www.nature.com/articles/d41586-020-00193-3
Potential to unravel the mechanistic details of microbial interactions with host / environment by analyzing the functional dynamics of the microbiome.

- **METAGENOMICS**
  - TAXONOMY
  - function

- **METATRANSCRIPTOMICS**
  - TAXONOMY
  - function

- **METAPROTEOMICS**
  - TAXONOMY
  - FUNCTION

[Image: https://thedoctorweighsin.com/what-everyone-should-know-about-the-infant-microbiome/]

**MICROBIOME**
Bond and Wilmes 2004
“The large-scale characterization of the entire protein complement of environmental microbiota at a given point in time”

Bond and Wilmes 2015
“Through the application of metaproteomics to different microbial consortia over the past decade, we have learnt much about key functional traits in the various environmental settings where they occur.”
MICROBIAL TAXA VARY WHILE METABOLIC PATHWAYS REMAIN STABLE WITHIN A HEALTHY POPULATION

**Metaproteomics Analytical Challenges**

**Single-Organism Proteomics**

- Small to medium size (10 K to 100K sequences)
- Single + Contaminants

**Metaproteomics**

- Large (1 million and above)
- Multi-organism database with homologous proteins

**Search Database Size**

- **Single-Organism Proteomics**
  - Single
- **Metaproteomics**
  - Large (1 million and above)

**Search Complexity**

- Disparate tools and multiple processing steps.

- **Search Algorithms** being developed to address large and complex database searches
- **Protein Grouping** at multi-organism level
- **Identification Statistics** affected by large databases
- **Taxonomy** based on unique peptide identifications
- **Functional analysis** based on proteins identified
**Metaproteomics Workflow**

**Database Generation**
- FASTQ
- FASTA
- Mass Spectrometry Data

**Database Search & Strategies**
- Search Algorithm
- Spectra

**Functional Analysis**
- Known Function
- Proteins
- Unassigned
- Shared
- Unique

**Taxonomy Analysis**
- Peptides
- Known Function
- Proteins
- Unassigned
- Shared
- Unique
Please **Register** for creating an account with a valid email ID and Password at usegalaxy.eu.

Once Registered, click on TIAAS to join the IMSC 2022 Galaxy session. [https://usegalaxy.eu/join-training/imsc_galaxy_training](https://usegalaxy.eu/join-training/imsc_galaxy_training)

Go to Shared Data Published Histories

Go to Shared Data Published Workflows

Run the workflow on active history
Sample Collection
Water samples were collected from the Bering Strait and Chukchi Sea and oceanic marine bacteria retained on a 0.7 μM filter.

Metagenome: Illumina HiSeq
Mass Spectrometry: Q-Exactive-HF

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

Damon H. May,1 Emma Timmins-Schiffman,1 Molly P. Mikan,2 H. Rodger Harvey,2 Elhanan Borenstein,1,3 Brook L. Nunn,1 and William S. Noble1,3

1Department of Genome Sciences and 2Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-5605, United States
2Department of Ocean, Earth & Atmospheric Sciences, Old Dominion University, Norfolk, Virginia 23529, United States
3Santa Fe Institute, Santa Fe, New Mexico 87501, United States

http://noble.gs.washington.edu/proj/metapeptide/
SearchGUI

SearchGUI matches MS/MS spectra to peptide sequences

- SearchGUI allows for multiple search engines to run simultaneously
- Specific digestion conditions can be selected
- Mass spectrometer parameters can be selected to maximize the efficacy of spectral matches
- Post-Translational Modifications (PTMs) can be added to the search parameters
PeptideShaker

**PeptideShaker filters SearchGUI results.**
- Search GUI results are filtered by FDR to yield most confident peptide spectral matches (PSMs)
- Peptide Shaker generates outputs such as Protein Report, Peptide Report and mzIdentML files for subsequent analysis.

**Unipept**

**FUNCTIONAL ANALYSIS**

- **Known Function**
- **Proteins**
- **Peptides**
- **Unassigned**
- **Shared**
- **Unique**

**TAXONOMY ANALYSIS**

- **Peptide**
- **Proteins** (Unipept)
- **Taxa** (NCBI Taxonomy)
- **Taxonomy** (Unipept)
- **LCA**

---

Unipept

FUNCTIONAL ANALYSIS

- Known Function
- Proteins
- Unassigned
- Shared
- Unique

TAXONOMY ANALYSIS

Peptides

Taxonomic and Functional analysis with Unipept

https://unipept.ugent.be/publications
https://usegalaxy.eu/u/pratikjagtap/h/metaproteomicsgtnimsc2022-completed
Who is there?

Get a taxonomy report from PSM report and Unipept pept2lca table

<table>
<thead>
<tr>
<th>peptide</th>
<th>superkingdom</th>
<th>genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADGHTMHFDVTGEK</td>
<td>Archaea</td>
<td>Nitrosopumilus</td>
</tr>
<tr>
<td>AALESFTGNVTSALK</td>
<td>Bacteria</td>
<td>Polaribacter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PSM#</th>
<th>Proteins</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGGEDMFVHKSDV...</td>
<td>EGGEDMFVHK</td>
</tr>
<tr>
<td>2</td>
<td>GKRVAAGVTVPE...</td>
<td>VAAAVGTPEQEWLK</td>
</tr>
</tbody>
</table>

SQLite Relational Data base

**SQL** query joins PSM and LCA to report number of PSMs and Peptides per genus

<table>
<thead>
<tr>
<th>genus</th>
<th>PSMs</th>
<th>DISTINCT PEPTIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planktomarina</td>
<td>161</td>
<td>20</td>
</tr>
<tr>
<td>Nitrosopumilus</td>
<td>122</td>
<td>27</td>
</tr>
</tbody>
</table>
Who is there? What are they doing?

How do we get taxonomy and function of a microbiome from a list of peptides?

<table>
<thead>
<tr>
<th>PSM#</th>
<th>Proteins</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGGEDMFVHKSDVDGFINEGDK</td>
<td>EGGEDMFVHK</td>
</tr>
<tr>
<td>2</td>
<td>GKRVAAAVGTVPQLEWLK, KVAAAVGTVPQLEWLK, RVAAAVGT...</td>
<td>VAAAVGTVPQLEWLK</td>
</tr>
</tbody>
</table>

**Unipept**
- Pept2lca: taxonomy lowest common ancestor for a peptide
- Pept2prot: Uniprot entries for a peptide with GO terms

### Peptide Taxonomy

<table>
<thead>
<tr>
<th>peptide</th>
<th>superkingdom</th>
<th>genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADGHTMHFDVITGEK</td>
<td>Archaea</td>
<td>Nitrospumilus</td>
</tr>
<tr>
<td>AALESFTGNVTSALK</td>
<td>Bacteria</td>
<td>Polaribacter</td>
</tr>
</tbody>
</table>

### Peptide Function

<table>
<thead>
<tr>
<th>peptide</th>
<th>uniprot_id</th>
<th>go_references</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAEKLSAAQAR</td>
<td>W5T6F9</td>
<td>GO:0016021</td>
</tr>
<tr>
<td>AAEKLSAAQAR</td>
<td>A0A0Q6ZKK0</td>
<td>GO:0005524 GO:0016887 GO:0015833</td>
</tr>
</tbody>
</table>
Questions

• How can I match metaproteomic mass spectrometry data to peptide sequences derived from shotgun metagenomic data?

• How can I perform taxonomy analysis and visualize metaproteomics data?

• How can I perform functional analysis on this metaproteomics data?

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html
Hands On Session

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html

https://usegalaxy.eu

- Download and start workflow
- Observe outputs
- Tool Basics
DATABASE SEARCH & STRATEGIES

DATABASE GENERATION

FASTQ

Protein / Peptide FASTA

Search Algorithm

Spectra

Peptides

FUNCTIONAL ANALYSIS

Known Function

Proteins

Hypothetical Function

Unknown Function

Shared Taxonomy

Unassigned Taxonomy

Unique Peptides

TAXONOMY ANALYSIS

Metaproteomics Workflow
**DATABASE GENERATION**
- FASTQ
- Protein / Peptide FASTA

**DATABASE SEARCH & STRATEGIES**
- **Search Algorithm**

**QUANTITATIVE ANALYSIS**
- Spectral counts OR Intensity data

**FUNCTIONAL ANALYSIS**
- Known Function
- Proteins
- Hypothetical Function
- Unknown Function
- Shared Taxonomy
- Unassigned Taxonomy

**TAXONOMY ANALYSIS**
- Unique Peptides

**METAPROTEOMICS WORKFLOW**
metaQuantome enables quantitative analysis of the taxonomic and functional state of a microbiome.

metaQuantome

**FUNCTION: VOLCANO PLOTS**

Fold-change: 33 hours versus 8 hours

**FUNCTION: HEATMAP**

**metaQuantome on Galaxy Training Network**

- [https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-data-creation/tutorial.html](https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-data-creation/tutorial.html)
- [https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-function/tutorial.html](https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-function/tutorial.html)
- [https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-taxonomy/tutorial.html](https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-taxonomy/tutorial.html)
Case Study: Cellulose Degradation in a BioGas Reactor

Biogas-plant (60°C) Fredrikstad, Norway

Lab-scale reactor (55°C)

Anaerobic bottles (65°C)

Food waste Manure

Food waste Manure

Cellulose

Serial dilution

0h
8h T1
13h
18h
23h T4
28h
33h T6
38h
43h

Magnus Arntzen NMBU, Norway
Functional abundance values separate time point T1 (8 hr) from other time points thus highlighting the importance of understanding functional state of the microbiome.
Gene Ontology terms were found to be differentially expressed in both timepoints T6 and T7 as compared to T4.
metaQuantome Analysis

**Functions Expressed by a Taxon**

*Hungateiclostridium*

- Cellulose binding
- Cellulase activity
- Cellulose 1,4-beta-cellobiosidase activity

**Gene Ontology Terms**

- Cellulose binding
- Cellulase activity
- Cellulose 1,4-beta-cellobiosidase activity

*Functions Associated With Cellulose Degradation in Hungateiclostridium*

**Taxonomic Contribution to a Function**

**GLYCOSIDE HYDROLASE**

- Genus: Coprothermobacter, Hungateiclostridium, Thermoclostridium

*Taxa associated with Glycosyl hydrolases and transferases*
Metaproteomics Publications

**PROTEOMICS**


What’s On The Horizon?

CHARACTERIZATION OF HOST & MICROBIAL PROTEINS FROM CLINICAL SAMPLES

QUANTITATION

MULTIOMICS
A METAPROTEOMICS BIOINFORMATICS WORKFLOW TO STUDY HOST-MICROBE DYNAMICS IN CLINICAL SAMPLES

Pratik Jagtap

Galaxy for Proteomics (Galaxy-P) team
University of Minnesota

Co-authors:
Monica E. Kruk, Subina Mehta, Katherine Do, James E. Johnson, Reid Wagner, Chris H. Wendt, John B. O’Connor, Theresa Laguna and Timothy J. Griffin
**DIA-MS HAS BETTER REPRODUCIBILITY ACROSS REPLICATES**

- **57.7% OVERLAP**
- **96.7% OVERLAP**

**DIA-MS PEPTIDES**

**DDA-MS PEPTIDES**

**DIA-MS OFFERS DEEPER TAXONOMY COVERAGE**

- **TAXONOMY-SPECIFIC INTENSITY**
  - **DDA**
  - **DIA INTENSITY**

**DIA-MS OFFERS DEEPER FUNCTIONAL COVERAGE**

- **HIGH ABUNDANCE PROTEINS**
  - DdRP
  - CypA
  - GP21

- **MEDIUM ABUNDANCE PROTEINS**
  - MGEAS
  - OTC
  - MgtA

- **LOW ABUNDANCE PROTEINS**
  - ArcC
  - ICL
  - DNase I

**WHAT LEVEL OF ACCURACY DOES DIA-MS OFFER?**

- **T4 Bacteriophage**
- **Bacillus subtilis**
- **Escherichia coli**
- **Salmonella typhimurium**

- **B. subtilis**
  - DDA: 42%
  - DIA: 86%

- **T4 Phage**
  - DDA: 94%
  - DIA: 86%

- **S. typhi**
  - DDA: 54.5%
  - DIA: 58.8%

- **E. coli**
  - DDA: 29.6%
  - DIA: 29.5%
META-OMICS APPROACH BY NMBU TEAM

METAGENOMICS

METATRANSCRIPTOMICS

METAPROTEOMICS

Francesco Delogu
Magnus Arntzen

https://galaxyproject.eu/posts/2020/04/14/integrative-meta-omics/
Cloud Computing Workshop (2022)

The iPRG will conduct a series of online video tutorials about the use of cloud computing resources for MS-based proteomics, focusing on Nextflow, the Trans-Proteomic Pipeline (TPP) and Galaxy Platform.

September 2022
- Michael Hoopmann - ISB, Seattle, WA
  - Instructions on how to use TPP to analyze MS data.
  - Answer questions from the participants

October 2022
- Melanie Foell - Freiburg University, Germany
  - Instructions on how to use Galaxy to analyze MS data.
  - Answer questions from the participants

November 2022
- Yasset Perez-Riverol - EBI, Hinxton, UK
  - Instructions on how to use Nextflow to analyze MS data.
  - Answer questions from the participants
Accessing tools and Workflows

METAGENOMICS:
Toolshed: z.umn.edu/metagenomics_toolshed
Galaxy Training Network: https://training.galaxyproject.org/training-material/topics/metagenomics/

METATRANSCRIPTOMICS:
Workflow: http://z.umn.edu/MTWF2020

METAPROTEOMICS:
Workflow: z.umn.edu/MPWF2020
Galaxy Training Network: http://z.umn.edu/gtn-metaproteomics

Also available on: https://proteomics.usegalaxy.eu/ and Metaproteomics Gateway: z.umn.edu/metaproteomicsgateway

galaxyp.org/contact