Our Experiments With Metaproteomics:
Ten Years of Galaxy-P

DECEMBER 2019
Our Experiments With Metaproteomics: Ten Years of Galaxy-P

Pratik Jagtap
University of Minnesota

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo
Enabling Proteo-Informatic Analysis via Galaxy-P Platform

• MICROBIOME RESEARCH
• MASS SPECTROMETRY DATA ANALYSIS
• GALAXY BIOINFORMATICS PLATFORM
• METAPROTEOMICS RESEARCH
Microbiome Research

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

5:1
Viruses:Bacteria in the gut microbiota

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

1.3X
The surface area of the gut tract is the same size as 2 tennis courts

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

>10,000
2kg
The gut microbiota can weigh up to 2kg

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.
Mass Spectrometry and Proteomics

1. Sample preparation
2. Separation
3. MS1
4. MS2
5. Data analysis

Retention time vs. m/z
Isolation window

Electro spray
HPLC

Metaproteomics Analytical Challenges

**Single-Organism Proteomics**

- Search Database:
  - Size: Small to medium size (10K to 100K sequences)
  - Complexity: Single + contaminants

**Metaproteomics**

- Search Database:
  - Size: Large (1 million and above)
  - Complexity: Multi-organism database with homologous proteins

- 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

Disparate tools and multiple processing steps.
Metaproteomics Workflow

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

**Database Search & Strategies**

**Functional Analysis**
- Known Function
  - Proteins

**Taxonomy Analysis**
- Unique
- Shared
- Unassigned
Software tools can be used in a sequential manner to generate analytical workflows that can be reused, shared and creatively modified.


The Galaxy Interface

Main viewing window
(workflow development, results visualization etc)

Tools

History
**Workflows:** Multiple software tools used in a sequential manner for an analysis

- Single software tool in Galaxy
- Integrated workflow using multiple, connected tools

**History:** recorded, complete analysis (workflow + all data and results)
Enabling Proteo-Informatic Analysis via Galaxy-P Platform

• MASS SPECTROMETRY DATA ANALYSIS
• MULTIOMIC RESEARCH
• GALAXY BIOINFORMATICS PLATFORM
• METAPROTEOMICS RESEARCH
**Metaproteomics Workflow**

**Database Generation**
- FASTQ
  - Protein / Peptide FASTA
    - Search Algorithm
      - Spectra

**Database Search & Strategies**
- PEPTIDES

**Quantitative Analysis**
- Spectral counts OR Intensity data

**Functional Analysis**
- Known Function
  - PROTEINS
    - Hypothetical Function
      - Unknown Function
    - Shared Taxonomy
      - Unassigned Taxonomy
    - Unique Peptides
      - Taxonomy Analysis
metaQuantome enables quantitative analysis of the taxonomic and functional state of a microbiome.

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

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

Serial dilution

Magnus Arntzen
NMBU, Norway
MetaQuantome Workflow
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.
Functions Associated With Cellulose Degradation in *Hungateiclostridium*

<table>
<thead>
<tr>
<th>Gene Ontology Terms</th>
<th>Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundance (LFQ intensities)</td>
<td></td>
</tr>
<tr>
<td>T4 (23-hrs)</td>
<td>0.0e+00</td>
</tr>
<tr>
<td>T6 (33-hrs)</td>
<td>5.0e+08</td>
</tr>
<tr>
<td>T7 (38-hrs)</td>
<td>1.0e+09</td>
</tr>
</tbody>
</table>

**Taxa associated with Glycosyl hydrolases and transferases**

**Hungateiclostridium**

**GLYCOSIDE HYDROLASE**

<table>
<thead>
<tr>
<th>Genus</th>
<th>Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundance (LFQ intensities)</td>
<td></td>
</tr>
<tr>
<td>T4 (23-hrs)</td>
<td>0e+00</td>
</tr>
<tr>
<td>T6 (33-hrs)</td>
<td>5e+07</td>
</tr>
<tr>
<td>T7 (38-hrs)</td>
<td>1e+08</td>
</tr>
</tbody>
</table>

**Taxonomic Contribution to a Function**

- **Hungateiclostridium**
- **metaQuantome Analysis**
METAPROTEOMICS: STRENGTHS & CHALLENGES

FUNCTIONAL COMPOSITION
HOST-MICROBIOME INTERACTION
QUANTIFYING BIOMASS CONTRIBUTIONS

PROTEIN EXTRACTION
VARIABILITY & ABUNDANCE
LARGE DATABASES FALSE POSITIVES
The Galaxy-P team seeks to continue integrating promising new software tools and workflows from leading laboratories into the Galaxy platform and encourages researchers to share their interests so that newer bioinformatics workflows can be made accessible via this resource.
EDUCATION PORTAL @ METAPROTEOMICS INITIATIVE

https://metaproteomics.org/education/

Want to stay up to date about the metaproteomics field and the Initiative? Become a member via www.metaproteomics.org!

@MetaP_Init
info@metaproteomics.org
ANALYSIS OF FUNCTIONS EXPRESSED BY MICROBIOMES

Online Workshop
15-24 November 2021

Funded by American Society of Microbiology (ASM) and Indo-United States Science and Technology Forum (IUSSTF).

Hosted by CSIR-IMTech, Chandigarh and Galaxy-P Team, Minneapolis.

https://galaxyproject.org/events/2021-11-microbiomes/home/
ANALYSIS OF FUNCTIONS EXPRESSED BY MICROBIOMES

Online Workshop (15-24 November 2021) Hosted by CSIR-IMTech, Chandigarh and Galaxy-P Team, Minneapolis

- Reached out to students and researchers in India via twitter, emails, etc.

- Selected researchers based on their CV and statement on why they were interested in the workshop.

- 24 registered students from all across India along with 13 students from IMTech attended the workshop.

ONLINE WORKSHOP ON ANALYSIS OF FUNCTIONS EXPRESSED BY MICROBIOMES

15-24 November 2021

Application deadline 17th October 2021

SPEAKERS/INSTRUCTORS
- Robert Hettrich (Oak Ridge National Laboratory, Oak Ridge, TN)
- Dave Clements (Galaxy Project, Johns Hopkins University)
- Pratik Jagtap (Galaxy-P Team, University of Minnesota, MN)
- Saskia Hiltermann (Erasmus University Medical Center, Hague)
- Yogesh Shouche (National Centre for Cell Science, Pune)
- Dr. Gareth Price (QCFI Facility for Advanced Bioninformatics, St. Lucia)
- Berenice Batut (University of Freiburg, Freiburg)
- Manoj Kumar (CSIR-IMTECH, India)
- Subha Mehta (Galaxy-P Team, University of Minnesota, MN)
- Brook Nunn (University of Washington, Seattle, WA)
- Anshu Bhadwaj (CSIR-IMTECH, India)
- Paul Wilmes (University of Luxembourg, Luxembourg)
- Sharmila Mande (Tata Consultancy Services)

AGENDA: https://galaxypproject.org/events/2021-11-microbiomes/home/

WHO CAN APPLY: Research scholar, Early career scientists, Faculty from Universities in India

APPLICATION FORM: tinyurl.com/funmicro

Registration deadline for selected candidates 24th October
WEEK-LONG ONLINE WORKSHOP

Analytical of Functions Expressed by Microbiomes

Speakers from across the world and instructors from the Galaxy community will teach online courses on microbiome analysis through recorded videos and interactive resources.

Online Workshop at CSIR-IMTech - funded by American Society of Microbiology (ASM) and Indo-United States Science and Technology Forum (IUSSST)

Practical Information

When: 15th-24th of November 2021
Support: Slack channel (kde.foo), channel: event:functional-microbiome

Program

Day 1 | November 15

<table>
<thead>
<tr>
<th>Time</th>
<th>Description</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 - 8:45 AM</td>
<td>Welcome</td>
<td>Manoj Kumar (CSIR-IMTech, Chandigarh)</td>
</tr>
<tr>
<td>8:40 - 9:00 AM</td>
<td>Introduction to the Course</td>
<td>Pratik Jagtap (University of Minnesota)</td>
</tr>
<tr>
<td>9:00 - 10:30 AM</td>
<td>Introduction to Metagenomics</td>
<td>Robert Noftel (Oak Ridge National Laboratory)</td>
</tr>
<tr>
<td>10:00 - 11:00 AM</td>
<td>Introduction to Galaxy Video Tutorials</td>
<td>Dave Clements &amp; Pratik Jagtap</td>
</tr>
<tr>
<td>11:15 AM - 3:00 PM</td>
<td>Galaxy Video Tutorials</td>
<td>GTN Instructors on Slack</td>
</tr>
<tr>
<td>3:00 - 4:30 PM</td>
<td>Applications with Galaxy</td>
<td>Anusha Bhasawaj</td>
</tr>
<tr>
<td>5:00 - 6:00 PM</td>
<td>Ice-Breaker (interaction amongst attendees)</td>
<td>Attendees on GTN Slack</td>
</tr>
<tr>
<td>6:00 - 7:00 PM</td>
<td>Q&amp;A for Galaxy Tutorials</td>
<td>Subrata Mohanta &amp; Saswata Hikimish</td>
</tr>
</tbody>
</table>

- INTRODUCTION TO GALAXY
- METAGENOMICS
- METATRANSSCRIPTOMICS
- METAPROTEOMICS
- META-OMICS
- FEEDBACK
The Feedback...

❖ Positives:
  • Registered attendees learnt about Galaxy platform, metagenomics, metatranscriptomics and metaproteomics.
  • GTN as a Resource for learning about tools and workflows.
  • Great opportunity to listen to eminent scientists (‘learning from artisans’).
  • Quiz sessions were great and educational!

❖ Room for Improvement:
  • The workshop levels can be BASIC, INTERMEDIATE AND ADVANCED – with opportunity to analyze your own data.
  • Follow up workshops with ‘hackathon’, ‘Meet the Speakers” as themes.
  • Running and completing workflows was challenging for some (However, could complete it later).
What's On The Horizon?

Characterization of Host & Microbial Proteins from Clinical Samples

Quantitation

Multi-omics
CHARACTERIZATION OF HOST & MICROBIAL PROTEINS FROM CLINICAL SAMPLES

CFL/H = cystic fibrosis with low/high microbial diversity
DCL/H = disease control with low/high microbial diversity

Monica Kruk

Combined analytical and bioinformatic workflow
The vision: quantitative assay for host-microbial protein dynamics in clinical CF samples

<table>
<thead>
<tr>
<th>Source</th>
<th>Peptide sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>KLEDGYFGGAR</td>
</tr>
<tr>
<td></td>
<td>SNSGVRLDGYAR</td>
</tr>
<tr>
<td></td>
<td>FAVHVESVLRNRPQPEYR</td>
</tr>
<tr>
<td></td>
<td>RFSADEQFFSGQQASSAHSSK</td>
</tr>
<tr>
<td>Microbial</td>
<td>SSTPSSPTGTSSSDGSHHIGWGER</td>
</tr>
<tr>
<td></td>
<td>VYRVDPFGVWER</td>
</tr>
<tr>
<td></td>
<td>VVDSLAPSITNVLQGK</td>
</tr>
<tr>
<td></td>
<td>ATGHEFAVK</td>
</tr>
<tr>
<td></td>
<td>LLQVDPEAR</td>
</tr>
</tbody>
</table>

Targeted mass spectrometry-based assays for quantifying levels of host-microbe proteins of interest
ASMS 2021: DIA-MS offers benefits for quantitative metaproteomics

**DIA-MS has better reproducibility across replicates**

- **57.7% overlap**
- **96.7% overlap**

**DIA-MS offers deeper taxonomy coverage**

- DDA intensity: 9.4E+11
- DIA intensity: 4.5E+11

**DIA-MS offers deeper functional coverage**

**What level of accuracy does DIA-MS offer?**

- T4 Bacteriophage: 56.3%
- Bacillus subtilis: 54.5%
- Escherichia coli: 57.1%
- Salmonella typhimurium: 58.8%

---

**Susan T. Weintraub**

**Brook L. Nunn**
Metaproteomics Datasets

Pure cultures of 1 Archaeon, 1 Eukaryote, 5 bacteriophages, 3 gr+ Bacteria and 22 gr- Bacteria

Mix 3 community types with 28 to 32 of the species/strains

4 biological replicates for each community type

<table>
<thead>
<tr>
<th>Cell number</th>
<th>1x</th>
<th>31x</th>
<th>354x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein amount</td>
<td>30x</td>
<td>1x</td>
<td>229x</td>
</tr>
</tbody>
</table>

Maximum fold differences between individual bacterial community members

- **8%**
- **16%**
- **32%**
- **64%**

- A. tumefaciens
- A. macroloidii
- B. subtilis
- B. xenovorans
- Chl. reichardtii
- C. violaceum
- C. metallidurans
- D. vulgaris
- E. coli
- N. europaea
- N. ureae
- N. viennensis
- N. multiformis
- Par. denitrificans
- Phage ES18
- Phage F0
- Phage F2
- Phage M13
- Phage P22
- Pseud. denitrificans
- P. fluorescents
- P. pseudoalcaligenes
- R. leg. bv. viciae 3841
- R. leg. bv. viciae VF39
- Roseobacter sp. AK199
- S. typhimurium (3 strains)
- S. aureus ATCC 13709
- S. aureus ATCC 25923
- S. maltophilia
- T. thermophilus
Protein digestion of microbial-mix sample (EQUAL CELL & EQUAL PROTEIN & UNEVEN)

NC STATE UNIVERSITY

4X

4X

4X

Peptide Digests to Participating Labs

WHOI WORKFLOW

UMN WORKFLOW

Fusion

Scaffold

Scaffold-DIA

DDA-MS

DIA-MS

DDA-MS

DIA-MS

Q-Exactive

EncyclopeDIA

FragPipe

Matthew McIlvin

Andrew Rajczewski
PEPTIDES DETECTION

EQUAL PROTEIN | DDA | UMN | PEPTIDES

4673 (34.7%)

EQUAL PROTEIN | DIA | WHOI | PEPTIDES

10859 (92.53%)

EQUAL PROTEIN | DDA | WHOI | PEPTIDES

9203 (50.66%)

EQUAL PROTEIN | DIA | WHOI | PEPTIDES

22692 (90.42%)
• Most of the taxonomic members are detected by both UMN and WHOI workflows.

• DIA detected previously undetected viral proteins (such as phage ES18).
**Coefficient of Variation**

- **Cupriavus metallireducens** (High abundance taxon)
- **Bacillus subtilis** (Medium abundance taxon)
- **Staphylococcus aureus** (Low abundance taxon)

The graph shows the coefficient of variation for different taxa across three categories: DDA Spectral Counts, DDA Intensity, and DIA Intensity.
Metatranscriptomics on Galaxy Training Network (ASaiM-MT)

https://training.galaxyproject.org/training-material//topics/metagenomics/tutorials/metatranscriptomics/tutorial.html

ASaiM
Batut et al Gigascience. 2018 7(6)
doi: 10.1093/gigascience/giy057
TEN YEARS OF GALAXY-P!

z.umn.edu/galaxypreferences
https://github.com/galaxyproteomics

https://training.galaxyproject.org
http://galaxyp.org/workshops/
TEN YEARS OF GALAXY-P!

Galaxy-P workflows

- Mass spectrometry data
- Database generation
- Sequence database search
- Peptide spectral matches
- Determine novel peptide sequence variants
- Proteogenomics
- Metaproteomics
- Functional & taxonomy analysis

z.umn.edu/galaxypreferences
Accessing Galaxy Tools and Workflows

Tools and Workflows also available on:
https://proteomics.usegalaxy.eu/

Galaxy Training Network:
https://training.galaxyproject.org/training-material/topics/proteomics

Contact: http://galaxyp.org/contact/
ACKNOWLEDGMENTS

Minnesota Supercomputing Institute
James Johnson
Reid Wagner

University of Minnesota
Timothy Griffin
Subina Mehta
Monica Kruk
Katherine Do
Danielle Weise
Andrew Rajczewski

Biologists / collaborators
Chris Wendt
Teresa Laguna
Maneesh Bhargava
Ryan Hunter
David Largaespada

Brook Nunn
U of Washington

Peter Thuy-Boun
Dennis Wolan
Scripps Institute

Lloyd Smith (Co-I)
Michael Shortreed
UW-Madison

Jeremy Fisher
Indiana University

Brian Searle
Ohio State University

Andrew McCardle
Imperial College,
London, UK

Saskia Hiltemann

Lennart Martens
Bart Mesuer
UGhent, Belgium

Michigan State University

Valdemir Carvalho
Fleury Group, Brazil

Matt Chambers
Nashville, TN

Harald Barsnes
Marc Vaudel
University of Bergen, Norway

Phil Pope
Magnus Överlie Arntzen
NMBU, Oslo, Norway

Thilo Muth
Robert Koch Institut, Germany

Sanjeeva Srivastava
IIT, Mumbai

Anamika Krishanpal
Persistent Systems Limited

India

Bjoern Gruening
Bérénice Batut
University of Freiburg,
Freiburg, Germany

Maria Doyle
Melbourne, Australia

Twitter.com/usegalaxy

http://galaxyp.org/contact/

galaxyp.org