Data Science Using Galaxy-P Platform

Pratik Jagtap
pjagtap@umn.edu

Galaxy-P Team; University of Minnesota
September 5, 2020
1st International Conference on Data Science in Biology
Outline

• Galaxy Platform and Galaxy-P (Galaxy for Proteomics)
• Mass spectrometry-based research
• Proteogenomics
• Metaproteomics
• COVID-19 Mass spectrometry datasets – reanalysis & new insights
• Accessibility and future directions
A web-based, community developed bioinformatics workbench for integrating disparate software

Geared towards use by bench scientists; many training resources available

Already home to genomic/transcriptomic tools

Provenance tracking, sharing and reproducibility

Amenable to other ‘omic tools (e.g. Galaxy for proteomics project, Galaxy-P)
Galaxy-P Platform
Boekel et al Nature Biotechnology 2015 33:137

Galaxy-P workflows

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

https://galaxyproject.org/
Proteogenomics

- Confirms translation of variants
- Direct evidence of potential functional variants
- Applications in neoantigen discovery (immuno-oncology)

Galaxy Training Network: https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/proteogenomics-dbcreation/tutorial.html
Quantitative ProteoTranscriptomics

Multiple studies have shown correlation of microbial composition with physiological conditions.
Potential to unravel the mechanistic details of microbial interactions with host / environment.
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
13h
18h
23h T4
28h
33h T6
38h
43h

Mass Spectrometer

Magnus Arntzen
NMBU, Norway
# Metaproteomics Workflow

## Microbiome

https://www.biospectrumasia.com

---

## Metaproteomics Workflow

<table>
<thead>
<tr>
<th></th>
<th>T4 (23hr)</th>
<th>T6 (33hr)</th>
<th>T7 (38hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># PSM</strong></td>
<td>~450,000</td>
<td>~330,000</td>
<td>~360,000</td>
</tr>
<tr>
<td><strong># Peptides</strong></td>
<td>80,619</td>
<td>79,253</td>
<td>89,761</td>
</tr>
<tr>
<td><strong># GO terms (Molecular Function)</strong></td>
<td>2,150</td>
<td>1,874</td>
<td>1,945</td>
</tr>
<tr>
<td><strong># Genera</strong></td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
- Enables quantitative analysis of the taxonomic and functional state of a microbiome.
- Unravels the complex and hierarchical data structure of taxonomic and functional ontologies.
- Enables data exploration, tests hypotheses, and generates high-quality visualizations.
- Deciphers the contribution of taxa to a functional process and vice versa.
FUNCTION: Volcano Plots

79 GO terms were found to be differentially expressed in both timepoints T6 & T7 as compared to T4.
Hungateiclostridium and Coprothermobacter do not seem to change significantly during the time points under study.

Functional abundance values cluster time points T4 (23 hr), T6 (28 hr) and T7 (38 hr), while taxonomy values do not.
Functions Associated With Cellulose Degradation in Hungateiclostridium

Cellulose degradation (*Hungateiclostridium*)

- Abundance (LFQ intensities)
  - cellulase activity
  - cellulose 1,4-beta-cellobiosidase activity
  - cellulose binding

TAXONOMY-FUNCTION INTERACTION

Taxa Associated with Glycosyl Hydrolases and Transferases

**GLYCOSIDE HYDROLASE**

- Genus
  - Coprothermobacter
  - Hungateiclostridium
  - Thermoctosrlidium

**GLYCOSYL TRANSFERASES**

- Genus
  - Coprothermobacter
  - Hungateiclostridium
  - Thermoctosrlidium
  - Methanothermobacter
COVID-19 Mass spectrometry datasets – reanalysis & new insights

- Clinical Datasets.
- In-depth proteomes of SARS-CoV-2 cell line models.
- Proteomic response of cells to SARS-CoV-2 infection.
- Synthetic peptides, spectral libraries, and targeted assays for SARS-CoV-2 proteins.

Zecha et al., 2020, Mol Cell Proteomics 19(9), 1503–1522
September 2020 https://doi.org/10.1074/mcp.RA120.002164
Deep proteomic analysis of CoViD-19 virus infected Vero cells. (PXD018804; Armengaud Lab)

Time course analysis of CoViD-19 infection. (PXD018594; Armengaud Lab)

Proteogenomics analysis of CoViD-19 infection (PXD018241; Matthews Lab)

Gargling samples from CoViD-19 infected patients (PXD019423; Sinz Lab)

Nasopharyngeal samples from CoViD-19 infected patients (PXD019119; Carvalho Lab)
Gargling samples from Covid-19 infected patients (PXD019423; Sinz Lab)

Respiratory tract samples from Covid-19 infected patients (PXD019119; Carvalho Lab)

Deep proteomic analysis of Covid-19 virus infected Vero cells. (PXD018804; Armengaud Lab)

Time course analysis of Covid-19 infection. (PXD018594; Armengaud Lab)

Proteogenomics analysis of Covid-19 infection (PXD018241; Matthews Lab)

COVID-19 validated peptides

139
99
550

COVID-19 Peptidomics

9
71

https://doi.org/10.1074/mcp.RA120.002164
Gargling samples from Covid-19 infected patients (PXD019423; Sinz Lab)

Respiratory tract samples from Covid-19 infected patients (PXD019119; Carvalho Lab)

Deep proteomic analysis of Covid-19 virus infected Vero cells. (PXD018804; Armengaud Lab)

Time course analysis of Covid-19 infection. (PXD018594; Armengaud Lab)

Proteogenomics analysis of Covid-19 infection (PXD018241; Matthews Lab)

COVID-19 validated peptides

Deep proteomic analysis of Covid-19 virus infected Vero cells. (PXD018804; Armengaud Lab) 139 (9)

Time course analysis of Covid-19 infection. (PXD018594; Armengaud Lab) 99 (4)

Proteogenomics analysis of Covid-19 infection (PXD018241; Matthews Lab) 550 (17)

COVID-19 Peptidomics

Gargling samples from Covid-19 infected patients (PXD019423; Sinz Lab) 9 (0)

Respiratory tract samples from Covid-19 infected patients (PXD019119; Carvalho Lab) 71 (0)
Gargling samples from Covid-19 infected patients (PXD019423; Sinz Lab)

Respiratory tract samples from Covid-19 infected patients (PXD019119; Carvalho Lab)

COVID-19 co-infection studies

- Stenotrophomonas maltophilia
- Streptococcus pneumoniae

PepQuery
Unipept
COMPIL 2.0

Anaerococcus
Dietzia
Prevotella buccalis
Johnsonella ignava
Accessing the Multi-omic Workflows

ALSO AVAILABLE ON:
GitHub: https://github.com/galaxyproteomics

Galaxy Toolshed: https://toolshed.g2.bx.psu.edu/

Docker: https://jraysajulga.github.io/cravatp-galaxy-docker/

Training Workflows also available on: https://training.galaxyproject.org
We can be Reached at:

Published Manuscripts: z.umn.edu/galaxypreferences

Galaxy-P Presentations: http://galaxyp.org/conference-presentations

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

Twitter: twitter.com/usegalaxyp