Mass Spectrometry (MS)-based multi-omics analysis using the Galaxy-P bioinformatics platform: A case study in COVID-19 data analysis

Great Lakes Bioinformatics Conference
May 10, 2021

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

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo
Overview and objectives

- Introduce the Galaxy framework as a solution for data analysis across ‘omics’ domains, with a focus on MS-based proteomics
- Provide **hands-on** experience to attendees in using Galaxy
- Demonstrate use of Galaxy for MS-based proteomics and multi-omic analysis using COVID-19 studies as representative data
- Lay the foundation for attendees to implement Galaxy in their work, and/or contribute tools to the Galaxy community
Workshop instructors and acknowledgements

• **Instructors**
  • Pratik Jagtap
  • Subina Mehta
  • Andrew Rajczewski
  • An Nguyen

• **Other contributors**
  • James Johnson
  • Reid Wagner
  • usegalaxy.eu team
  • Galaxy community

• **GLBIO Conference organizers**
Topic(s) of workshop

- Mass spectrometry-based proteomic informatics
- COVID-19 research
- Metaproteomics (multi-omics)
- Protein sequence analysis
Shotgun proteomics: breaking proteins into peptides

1. Protein, MW = 10,000 + 

   digest into peptides (enzymatic, chemical)

2. Peptides, MW < 4,000

   Trypsin

   cleaves c-terminal side of arginine (R) and lysine (K)

3. Peptide fragments
Mass Spectrometry-based proteomics: A primer

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)

proteins → digestion → peptides → "Multidimensional" fractionation → \( \mu \text{LC} \) → ESI → "Molecular turn style"

\( \text{MS1} \) → Isolation → Fragmentation → Mass Analysis → MS2

peptide fragments

\[ \text{Intensity, counts} \]

1000 1200 1400 1600 1800 2000 2200 2400 2600 2800


1360.7892 1606.8892 1938.0629 1815.9397 1378.8696 2316.3092 1506.9692 1886.0672 1271.6925 1661.8925

+TOF MS: 24 MCA scans from Myo_trypsin.wiff. Max. 5191.0 counts.
Matching amino acid sequences to MS/MS data

- Raw MS/MS spectrum
- Protein sequence and/or DNA sequence database search
  - Direct identification of 1000s proteins from complex mixtures
- Peptide sequence match
- Protein identification
Inferring protein identity for peptide sequences

Cytochrome C

\( \text{NH}_2 \text{GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLENPKKYIPGTKMIFAGIKKTEREDLIAYLKATNE} \text{COOH} \)
Driven by bioinformatics

- Many step workflow for analyzing large and diverse datasets
One solution: Galaxy


- A web-based, community developed bioinformatics workbench for integrating disparate software -- flexible
- 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)

**Working philosophy:**
Galaxy: an integrative workbench well-suited for multi-omics

Integrate datasets, analysis tools, visualizations, and computing resources for large-scale biomedical data science.

**Interfaces**
- Web UI
- Programmatic API

**Datasets**

**Computing Resources**
- slurm
- Condor
- Google Cloud Platform
- Amazon Web Services

**Analysis Tools and Visualizations**
- Filter by columns
- Differential expression
- Drug response predictor

*Courtesy Jeremy Goecks, OHSU*
## Accessing Galaxy via a public gateway

1. **Login/Register:** usegalaxy.eu

2. **Go to TlaaS link:** https://usegalaxy.eu/join-training/glbiotraining

3. **Return to usegalaxy.eu site**
The Galaxy Interface

Main viewing window
(workflow development, results visualization etc)

Tools

History
Built for the community: Tool sharing and access
**Workflows**: Multiple software tools used in a sequential manner for an analysis

**History**: recorded, complete analysis (workflow + all data and results)
Built for the community: Sharing workflows and histories

**HISTORY:** https://galaxyp.msi.umn.edu/u/pjagtap/h/itraq-search-yang-2-xtandem-scaffold

**WORKFLOW:** https://galaxyp.msi.umn.edu/u/pjagtap/w/workflow-for-4-plex-itraq-xtandem-search-scaffold-processing

Galaxy-Workflow-Workflow_for_4-plex_iTRAQ_X_tandem_Search_Scaffold_Processing.ga
Accessing and using Galaxy
Agenda and schedule

1 PM - 1:30 PM: Introduction to Galaxy and Galaxy-P

1:30 PM - 2:00 PM: Mass Spectrometry, Multi-omics and Galaxy Workflows for COVID-19 Research

2:00 PM - 3:30 PM: Detecting SARS-CoV-2 and microbial peptides from patient samples

3:30 PM - 4:00 PM: Summary and Questions and Answers

Please ask questions as they come up!
Galaxy Workflows for analysis of COVID-19 Mass Spectrometry datasets

Pratik Jagtap
University of Minnesota
Galaxy-P Team
**COVID-19 DETECTION METHODS**

**Molecular Tests (Nucleic Acid Detection)**
*Diagnose active SARS-CoV-2 infections*

1. Obtain Specimen: Swab.
2. Extract RNA from specimen and convert to DNA.
3. Amplify by PCR with SARS-CoV-2 specific primers.
4. Interpret results: presence of viral RNA indicates active SARS-CoV-2 infection.

**Antibody Tests (Serology)**
*Detect immune response to SARS-CoV-2 exposure*

1. Obtain Specimen: Blood Sample.
2. Expose specimen to SARS-CoV-2 specific antigens.
3. Interpret results: color change indicates previous exposure to SARS-CoV-2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ProteomeXchange ID</th>
<th>Pubmed ID</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gargling Solution</td>
<td>PXD019423</td>
<td>PMID: 32568543</td>
<td>Sinz Lab (Halle, Germany)</td>
</tr>
<tr>
<td>Nasopharyngeal swabs</td>
<td>PXD020394</td>
<td>PMID: 32835036</td>
<td>Lima Lab (Montevideo, Uruguay)</td>
</tr>
<tr>
<td>Respiratory tract samples</td>
<td>PXD021328</td>
<td>PMID: 33273458</td>
<td>Carvalho Lab (São Paulo, Brazil)</td>
</tr>
<tr>
<td>Broncho-alveolar lavage fluid (BALF)</td>
<td>PXD022085</td>
<td>PMID: 33098359</td>
<td>Cheng Lab (Wuhan, China)</td>
</tr>
<tr>
<td>Lung Samples</td>
<td>PXD018094</td>
<td>PMID: 33060566</td>
<td>Zhong Lab (Beijing, China)</td>
</tr>
<tr>
<td>Gut Microbiome</td>
<td>PXD023099</td>
<td>Unpublished</td>
<td>Yan Lab (Guangzhou, China)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ProteomeXchange ID</th>
<th>Pubmed ID</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time series</td>
<td>PXD018594</td>
<td>PMID: 32619390</td>
<td>Armengaud Lab (Bagnols-sur-Cèze, France)</td>
</tr>
<tr>
<td>8 hours time point</td>
<td>PXD018804</td>
<td>PMID: 32462744</td>
<td>Armengaud Lab (Bagnols-sur-Cèze, France)</td>
</tr>
<tr>
<td>Proteo-transcriptomics analysis</td>
<td>PXD018241</td>
<td>PMID: 32723359</td>
<td>Matthews Lab (Bristol, UK)</td>
</tr>
<tr>
<td>Host-viral protein interaction</td>
<td>PXD018117</td>
<td>PMID: 32353859</td>
<td>Krogan Lab (San Francisco, CA)</td>
</tr>
</tbody>
</table>
COVID-19 DETECTION MASS SPECTROMETRY METHODS

In silico approach toward the identification of unique peptides from viral protein infection: Application to COVID-19.
Orsburn et al doi: https://doi.org/10.1101/2020.03.08.980383 April 2020

In Vivo Datasets
Saline gargling (PXD019423)
Oro/nasopharyngeal swabs (PXD021328)

In Vitro Datasets
Gouveia et al. (PXD018804)
Grenga et al. (PXD018594)
Davidson et al. (PXD018241)


Multiple datasets were used in the creation of a peptide panel and the validation of their utility in diagnosing SARS-CoV-2.
Protein assignment of detected and validated SARS-CoV-2 peptides

Metaproteomics analysis of SARS-CoV-2 infected patient samples

Gargling solution (PXD019423)

Oro- and Nasopharyngeal tract (PXD020394)

Respiratory tract (PXD021328)

http://dx.doi.org/10.1021/acs.jproteome.0c00822
Co-Infection in COVID-19 Patients

- Co-infection has an effect on the diagnosis, symptoms, treatment and mortality.
- Patient could be infected prior to COVID-19 infection or during hospitalization.
- Nosocomial infections can affect antibiotics treatment plans due to antibiotic resistance.
- Culture-based detection methods prolong diagnosis of the disease.


### Datasets and Organisms Detected

<table>
<thead>
<tr>
<th>Organism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Streptococcus pneumoniae</strong></td>
<td>Causes pneumonia (respiratory-tract infection)</td>
</tr>
<tr>
<td><strong>Lactobacillus rhamnosus</strong></td>
<td>Probiotic</td>
</tr>
<tr>
<td><strong>Pseudomonas sp. BcH</strong></td>
<td>Unclassified <em>Pseudomonas</em></td>
</tr>
<tr>
<td><strong>Pseudomonas monteilii</strong></td>
<td>Meningoencephalitis</td>
</tr>
<tr>
<td><strong>Acinetobacter ursingii</strong></td>
<td>Bacteremia</td>
</tr>
</tbody>
</table>

---

**Thuy-Boun et al (2021) Journal of Proteome Research**

http://dx.doi.org/10.1021/acs.jproteome.0c00822
Galaxy workflows are available for analysis of COVID-19 MS datasets.

We could detect peptides that spanned the SARS-CoV-2 proteome.

Metaproteomics analysis revealed presence of potential co-infecting microorganisms in COVID-19 patient samples.

Reanalysis of PXD018804
Reanalysis of PXD018682
Reanalysis of PXD018117
Reanalysis of PXD018241
Reanalysis of PXD018594
Reanalysis of PXD020394
Reanalysis of PXD021328
Metaproteomics of mPXD019423
Metaproteomics of mPXD021328
Metaproteomics of mPXD020394

A rigorous evaluation of optimal peptide targets for MS-based clinical diagnostics of Coronavirus Disease 2019 (COVID-19).
Rajczewski et al (Preprint in MedRxiv)
https://www.medrxiv.org/content/10.1101/2021.02.09.21251427v1
ACCEPTED FOR PUBLICATION IN CLINICAL PROTEOMICS
Please **Register** for creating an account with a valid email ID and Password at [usegalaxy.eu](https://usegalaxy.eu). Once Registered, click on [https://usegalaxy.eu/join-training/glbiotraining](https://usegalaxy.eu/join-training/glbiotraining) to join the GLBIO 2021 Galaxy session.

Go to Shared Data Histories, GLBIO inputs Import • Go to Shared Data Workflows, GLBIO workflow Import • Run the workflow on active history
Determining the presence of COVID-19 peptides in Galaxy
Dataset Information

Oro-Nasopharyngeal

Patient data - Positive and Negative

Coronavirus present?

Any other species – co-infection?

Oro- and Nasopharyngeal tract (PXD020394)
Analytical Workflow
Instructions

• Go to Shared Data ➔ Histories ➔ GLBIO workshop Sample 1 ➔ Import

OR

• Go to Shared Data ➔ Histories ➔ GLBIO workshop Sample 2 ➔ Import
Unipept

Match taxa

select organisms with at least 2 peptides
Peptide Validation

PepQuery

Spectral Quality
PepQuery Workflow

1. Peptide sequence preparation and initial filtering
   - Peptide: LVVVGADGVGK
   - Protein: DNA
   - VCF/BED/GTF

2. Candidate spectra retrieval and PSM scoring
   - Target peptide: LVVVGADGVGK
   - PSMs:
     - 3849 spectra

3. Competitive filtering based on reference sequences
   - PSMs:
     - 12 spectra

4. Statistical evaluation
   - CPTAC CRC study
     - 95 samples
     - 12,941,421 spectra
   - RefSeq
   - Shuffled peptides:
     - LVVVGADGGVK
     - VLVVGADGVGK
     - VVVGADGVGK
     - LVVDAGGGVK
     - .......

5. Visualization
   - PSMs:
     - 5 spectra
   - Competitive filtering based on unrestricted modification searching
     - 9 spectra

PepQuery

Features
PSM rank

Eliminate peptides in reference database

P-value

confidence
What do you think is the COVID and co-infection status of this sample?
Spectral Validation (Lorikeet)
Thank you!

Questions?

Useful links:

• [galaxyp.org](galaxyp.org)
• [proteomics.usegalaxy.eu](proteomics.usegalaxy.eu)
• [Galaxy Workflows for COVID-19 Research](Galaxy Workflows for COVID-19 Research)
• [Galaxy Training Network](Galaxy Training Network)
• [Co-infection Metaproteomics](Co-infection Metaproteomics)
• [Peptides for clinical diagnosis of COVID-19](Peptides for clinical diagnosis of COVID-19)