Enabling Proteo-Informatic Analysis via Galaxy-P Platform

Pratik Jagtap
University of Minnesota

Google Scholar: z.umn.edu/pjgs
PubMed: z.umn.edu/pjagtapreferences
Email: pjagtap@umn.edu
Twitter: pratikomics

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

• MASS SPECTROMETRY DATA ANALYSIS
• MULTIOMICS RESEARCH
• GALAXY BIOINFORMATICS PLATFORM
• METAPROTEOMICS RESEARCH
Mass Spectrometry and Proteomics

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

HPLC
Electro spray

Retention time
m/z
Isolation window

Intensity

m/z

Srivatsan P, et al. 2018
Enabling Proteo-Informatic Analysis via Galaxy-P Platform

- MASS SPECTROMETRY DATA ANALYSIS
- MULTIOMICS RESEARCH
- GALAXY BIOINFORMATICS PLATFORM
- METAPROTEOMICS RESEARCH
Enabling Proteo-Informatic Analysis via Galaxy-P Platform

- MULTIOMICS
- MASS SPECTROMETRY
- GALAXY BIOINFORMATICS PLATFORM
- METAPROTEOMICS
One solution: Galaxy

Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team.. *Genome Biol.* 2010, **11**: R86.

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:*
The Galaxy Interface

Main viewing window
(workflow development, results visualization etc)
Defining Workflows and Histories

**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)
Welcome to Galaxy Training!
Collection of tutorials developed and maintained by the worldwide Galaxy community

<table>
<thead>
<tr>
<th>Galaxy for Scientists</th>
<th>Tutorials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to Galaxy Analysis</td>
<td>10</td>
</tr>
<tr>
<td>Assembly</td>
<td>5</td>
</tr>
<tr>
<td>Climate</td>
<td>2</td>
</tr>
<tr>
<td>Computational chemistry</td>
<td>6</td>
</tr>
<tr>
<td>Ecology</td>
<td>5</td>
</tr>
<tr>
<td>Epigenetics</td>
<td>6</td>
</tr>
<tr>
<td>Genome Annotation</td>
<td>3</td>
</tr>
<tr>
<td>Imaging</td>
<td>2</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>4</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>6</td>
</tr>
<tr>
<td>Proteomics</td>
<td>15</td>
</tr>
<tr>
<td>Sequence analysis</td>
<td>2</td>
</tr>
<tr>
<td>Statistics and machine learning</td>
<td>8</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>23</td>
</tr>
<tr>
<td>Variant Analysis</td>
<td>8</td>
</tr>
<tr>
<td>Visualization</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Galaxy Tips &amp; Tricks</th>
<th>Tutorials</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Interface and Data Manipulation</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Galaxy for Developers and Admins</th>
<th>Tutorials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galaxy Server administration</td>
<td>35</td>
</tr>
<tr>
<td>Development in Galaxy</td>
<td>13</td>
</tr>
</tbody>
</table>

How to contribute?
First off, thanks for taking the time to contribute!

You can report mistakes or errors, create more contents, etc. Whatever is your background, there is probably a way to do it: via the GitHub website, via command-line. If you feel it is too much, you can even write it with any text editor and contact us: we will work together to integrate it.

To get you started, check our dedicated tutorials or our Frequently Asked Questions

<table>
<thead>
<tr>
<th>Galaxy for Contributors and Instructors</th>
<th>Tutorials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributing to the Galaxy Training Material</td>
<td>10</td>
</tr>
<tr>
<td>Teaching and Hosting Galaxy training</td>
<td>5</td>
</tr>
</tbody>
</table>

> 130 training materials

https://training.galaxyproject.org/training-material
GCC 2022
MINNEAPOLIS
COME TOGETHER  #UseGalaxy2022

https://gcc2022.sched.com/
Enabling Proteo-Informatic Analysis via Galaxy-P Platform

- MASS SPECTROMETRY DATA ANALYSIS
- MULTIOMICS RESEARCH
- GALAXY BIOINFORMATICS PLATFORM
- METAPROTEOMICS RESEARCH
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.

The genes in your microbiome outnumber the genes in our genome by about 150 to one.

1.3X
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 living in and on the human body.

https://worldmicrobiomeday.com/resources/

aPC
Microbiome Ireland
Interfacing Food & Medicine
The microbiome is more medically accessible and manipulable than the human genome.
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.

2.5
The number of times your body's microbes would circle the earth if positioned end-to-end.

https://www.nature.com/articles/d41586-020-00193-3

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

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.
Metaproteomics Workflow

DATABASE GENERATION

- FASTQ
- FASTA
- Spectra
- Mass Spectrometry Data

DATABASE SEARCH & STRATEGIES

- Search Algorithm
- Peptides

FUNCTIONAL ANALYSIS

- Known Function
- Proteins

TAXONOMY ANALYSIS

- Unassigned
- Shared
- Unique

Search Algorithm
**Metaproteomics Workflow**

**Database Generation**
- FASTQ
- Protein / Peptide FASTA

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

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

Blank et al. Proteomes 2018, 6(1), 7; https://doi.org/10.3390/proteomes6010007


### Datasets and organisms detected

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

---


http://dx.doi.org/10.1021/acs.jproteome.0c00822
Protein assignment of detected and validated SARS-CoV-2 peptides

https://clinicalproteomicsjournal.biomedcentral.com/articles/10.1186/s12014-021-09321-1
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
Workshops

http://galaxyp.org/workshops/
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.

ONLINEWORKSHOP ON ANALYSIS OF FUNCTIONS EXPRESSED BY MICROBIOMES

15-24 November 2021

Application deadline 17th October 2021

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

AGENDA: https://galaxyproject.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

Talks By Eminent Researchers
Introduction to Training Material
Participants run workflows based on instructions
Q & A Session with Instructors
Quiz Sessions to test prior sessions

Analysis of Functions Expressed by Microbiomes

Speakers from across the world and instructors from Galaxy community will teach online courses on microbial 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 (USSSTF)

Practical Information
When: 15th-24th of November 2021
Support: Slack channel (register link), event functional microbiome
Contact: Prash Jagtap
Format: Online, Microsoft Teams Access Link

Program
Before you will find the full program for this event.

Some days will feature practical sessions, with video tutorials you can follow. Click the button below to see the tutorials.

Video Tutorials

Day 1 | November 15

<table>
<thead>
<tr>
<th>Time</th>
<th>Description</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 - 8:40 AM</td>
<td>Welcome</td>
<td>Minu Kaur (CSIR-IMTech, Chandigarh)</td>
</tr>
<tr>
<td>8:40 - 9:00 AM</td>
<td>Imaginovation</td>
<td>Dr. Sanjiv Khosla (Director CSIR-IMTech, Chandigarh)</td>
</tr>
<tr>
<td>9:00 - 9:30 AM</td>
<td>Introduction to the Course</td>
<td>Prash Jagtap (University of Minnesota)</td>
</tr>
<tr>
<td>9:30 - 10:30 AM</td>
<td>Introduction to Metagenomics Research</td>
<td>Robert Hettich (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; Prash Jagtap</td>
</tr>
<tr>
<td>11:15 AM - 12:00 PM</td>
<td>Galaxy Video Tutorials</td>
<td>GTN instructors on Slack</td>
</tr>
<tr>
<td>1:00 - 3:00 PM</td>
<td>Applications within Galaxy</td>
<td>Anshu Bhardwaj</td>
</tr>
<tr>
<td>3:00 - 3:30 PM</td>
<td>Ice-Breaker (interaction amongst attendees)</td>
<td>Attendees on GTN Slack</td>
</tr>
<tr>
<td>3:30 - 4:00 PM</td>
<td>Q&amp;A for Galaxy Tutorials</td>
<td>Subhash Mehta &amp; Sapsana Hithammar</td>
</tr>
</tbody>
</table>

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

My experience of using galaxy

Command line interphase

Galaxy

Learning

Execution

Difficult

Easy

Easy

Easy

The Galaxy

Galaxy of tools

Metabolomics

Metatranscriptomics

Metagenomics

Microbiome

Genomics

AMRs

Many more...

IF YOU ASK, YOU WILL NEVER KNOW.

BUT,

IF YOU KNOW, YOU NEED ONLY ASK...
The Feedback...

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

\textit{Room for Improvement:}
\begin{itemize}
  \item The workshop levels can be BASIC, INTERMEDIATE AND ADVANCED – with opportunity to analyze your own data.
  \item Follow up workshops with ‘hackathon’, ‘Meet the Speakers” as themes.
  \item Running and completing workflows was challenging for some (However, could complete it later).
\end{itemize}
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**
  - Instructions on how to use TPP to analyze MS data.
  - Answer questions from the participants
  - Michael Hoopmann - ISB, Seattle, WA

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

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

Galaxy Europe: https://proteomics.usegalaxy.eu/

Contact: http://galaxyp.org/contact/
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
What’s On The Horizon?

QUANTITATION

TARGETED METAPROTEOMICS

MS DATA

DDA  DIA

PEPTIDE QUANTITATION

metaQuantome

TABULAR OUTPUT WITH PEPTIDE m/z, RT INFORMATION

TARGETED METAPROTEOMICS ASSAYS WITHIN SKYLINE

CHARACTERIZATION OF PROTEINS WITH UNKNOWN FUNCTION

CHARACTERIZATION OF PROTEINS

PROTEINS WITH UNKNOWN FUNCTION

METAGENOME METADATA

METAGENOME ANNOTATION TABULAR OUTPUTS

CONSERVED DOMAINS

PEPTIDE / PROTEIN QUANT

HOMOLOGY WITH PROTEINS OF KNOWN STRUCTURE / FUNCTION

output

metaQuantome

mzTab outputs

FOR FURTHER DETAILS, SEE EXPLANATION...
Combined analytical and bioinformatic workflow
CFL/H = cystic fibrosis with low/high microbial diversity
DCL/H = disease control with low/high microbial diversity
Results: 87 stringently filtered microbial peptides

<table>
<thead>
<tr>
<th>Genus</th>
<th>CF/DC Ratio (MS Intensity)</th>
<th>Peptides Detected</th>
<th>Function Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudomonas</td>
<td>109.7</td>
<td>4</td>
<td>Type II secretion system</td>
</tr>
<tr>
<td>Stenotrophomonas</td>
<td>49.0</td>
<td>2</td>
<td>SGL protein</td>
</tr>
<tr>
<td>Staphylococcus</td>
<td>21.1</td>
<td>3</td>
<td>Eap/MAP</td>
</tr>
<tr>
<td>Prevotella</td>
<td>16.0</td>
<td>2</td>
<td>ISS family transposase</td>
</tr>
<tr>
<td>Fusobacterium</td>
<td>8.8</td>
<td>1</td>
<td>Penicillin-binding prot.</td>
</tr>
<tr>
<td>Streptococcus</td>
<td>5.8</td>
<td>2</td>
<td>LKG protein</td>
</tr>
<tr>
<td>Sphingomonas</td>
<td>4.5</td>
<td>1</td>
<td>Uncharacterized protein</td>
</tr>
<tr>
<td>Moraxella</td>
<td>0.001</td>
<td>3</td>
<td>Outer membrane prot.</td>
</tr>
</tbody>
</table>

Microbe MS Intensity in CF vs. DC

Monica Kruk
What about the human proteins?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CF/DC Ratio (MS Intensity)</td>
<td>196.54</td>
<td>13.99</td>
<td>6.62</td>
<td>5.69</td>
<td>4.56</td>
<td>0.25</td>
<td>0.16</td>
<td>0.12</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Protein Examples</td>
<td>Deoxyribonuclease-1</td>
<td>Dynein light chain Tctex-type 1</td>
<td>Protein S100-A8</td>
<td>Isoform 3 of Phosphorylase b kinase regulatory subunit beta</td>
<td>Neutrophil elastase</td>
<td>Guanylate-binding protein 2</td>
<td>Calcium-activated chloride channel regulator 1</td>
<td>Apolipoprotein E</td>
<td>Rootletin</td>
<td>Ankyrin-3</td>
</tr>
<tr>
<td>Pendrin</td>
<td>Intraflagellar transport protein 56</td>
<td>Neutrophil collagenase</td>
<td>Serine/threonine-protein phosphatase PP1-beta catalytic subunit</td>
<td>Azurocidin</td>
<td>Ubiquitin/ISG15-conjugating enzyme E2 L6</td>
<td>Chloride intracellular channel protein 3</td>
<td>Serum paraoxonase/arylesterase 1</td>
<td>Ciliary rootlet coiled-coil protein 2</td>
<td>Cingulin</td>
<td></td>
</tr>
</tbody>
</table>
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>FAVHVESVLNRVPQPEYR</td>
</tr>
<tr>
<td></td>
<td>RFSADEQFFSGQAASSAHSSK</td>
</tr>
<tr>
<td>Microbial</td>
<td>SSTPSSPTGTSSDSGGHHIGWER</td>
</tr>
<tr>
<td></td>
<td>VYRVPDFGVWER</td>
</tr>
<tr>
<td></td>
<td>VVDSLAPSITNVLVOQGK</td>
</tr>
<tr>
<td></td>
<td>ATGHEFAVKK</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
TEN YEARS OF GALAXY-P!

PubMed

z.umn.edu/galaxypreferences

https://github.com/galaxyproteomics

GalaxyP
galaxyp.org

https://training.galaxyproject.org

http://galaxyp.org/workshops/

Galaxy Training Network
TEN YEARS OF GALAXY-P!

Mass spectrometry data → Galaxy-P workflows → Peptide spectral matches → Proteogenomics

Sequence database search → database generation

Determine novel peptide sequence variants

Metaproteomics

Functional & taxonomy analysis