Introduction to Galaxy-P Project

- *Pratik Jagtap*
- *University of Minnesota*

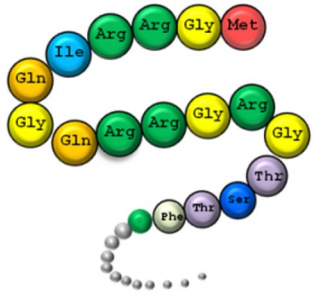
Google Scholar: z.umn.edu/pjgs
PubMed: z.umn.edu/pjagtapreferences

Email: pjagtap@umn.edu
Twitter: pratikomics
usegalaxyp

Learn more at galaxyp.org

z.umn.edu/itcrgalaxyvideo

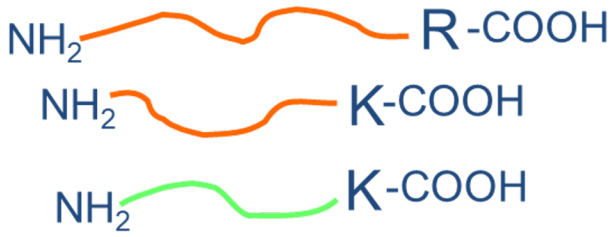
Shotgun proteomics: breaking proteins into peptides



1. Protein, MW = 10,000 +

digest into peptides (enzymatic, chemical)

2. Peptides,
MW < 4,000



Trypsin

N-THK.NCPHIVVGTPGR.IPD-C

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

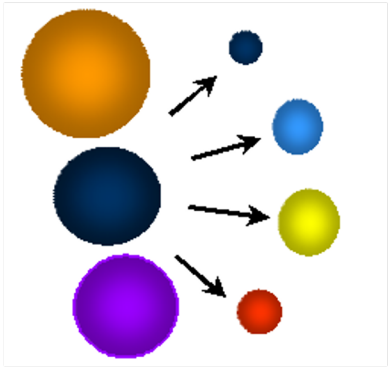
3. Peptide
fragments

GalaxyP
galaxyp.org

Mass Spectrometry-based proteomics: A primer

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

proteins



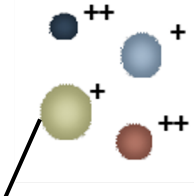
digestion

peptides

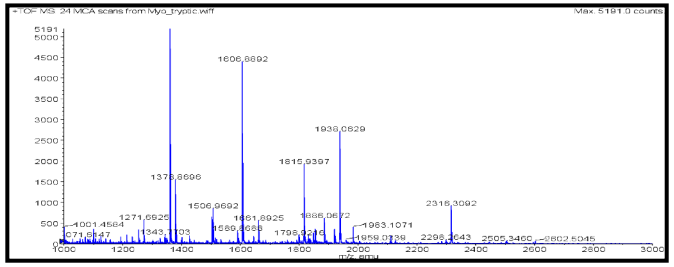
"Multidimensional" fractionation

"Molecular turn style"

µLC

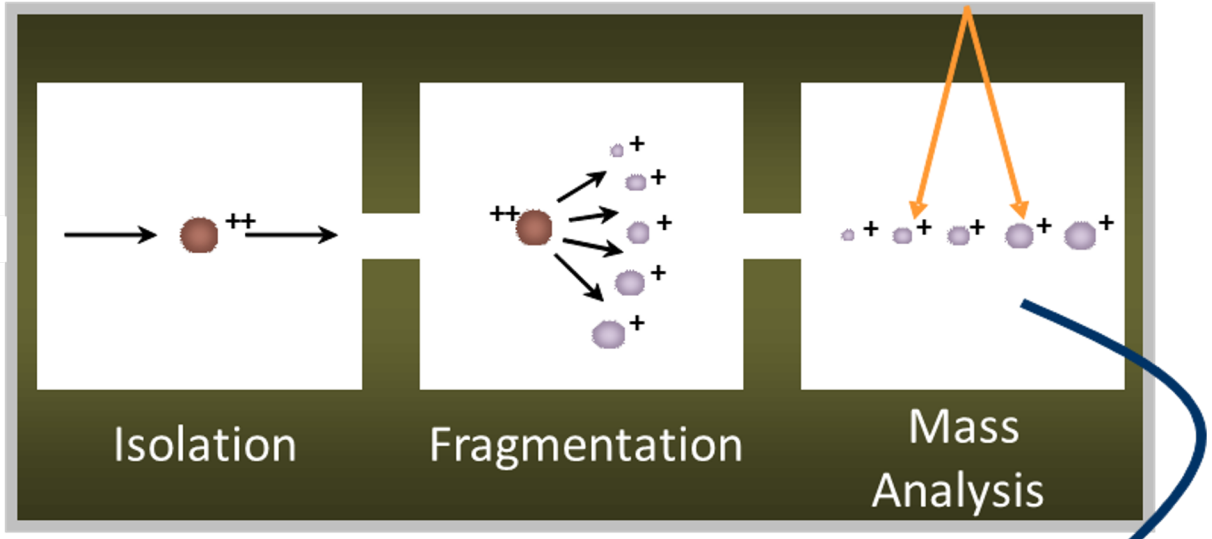


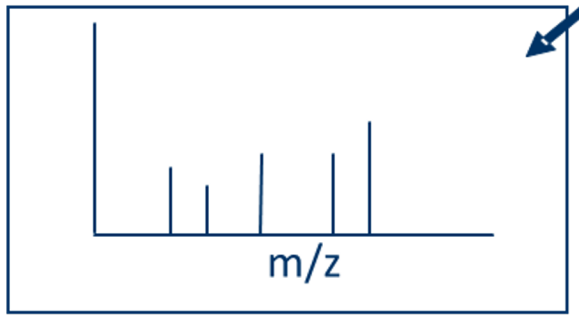
ESI



MS1

peptide fragments

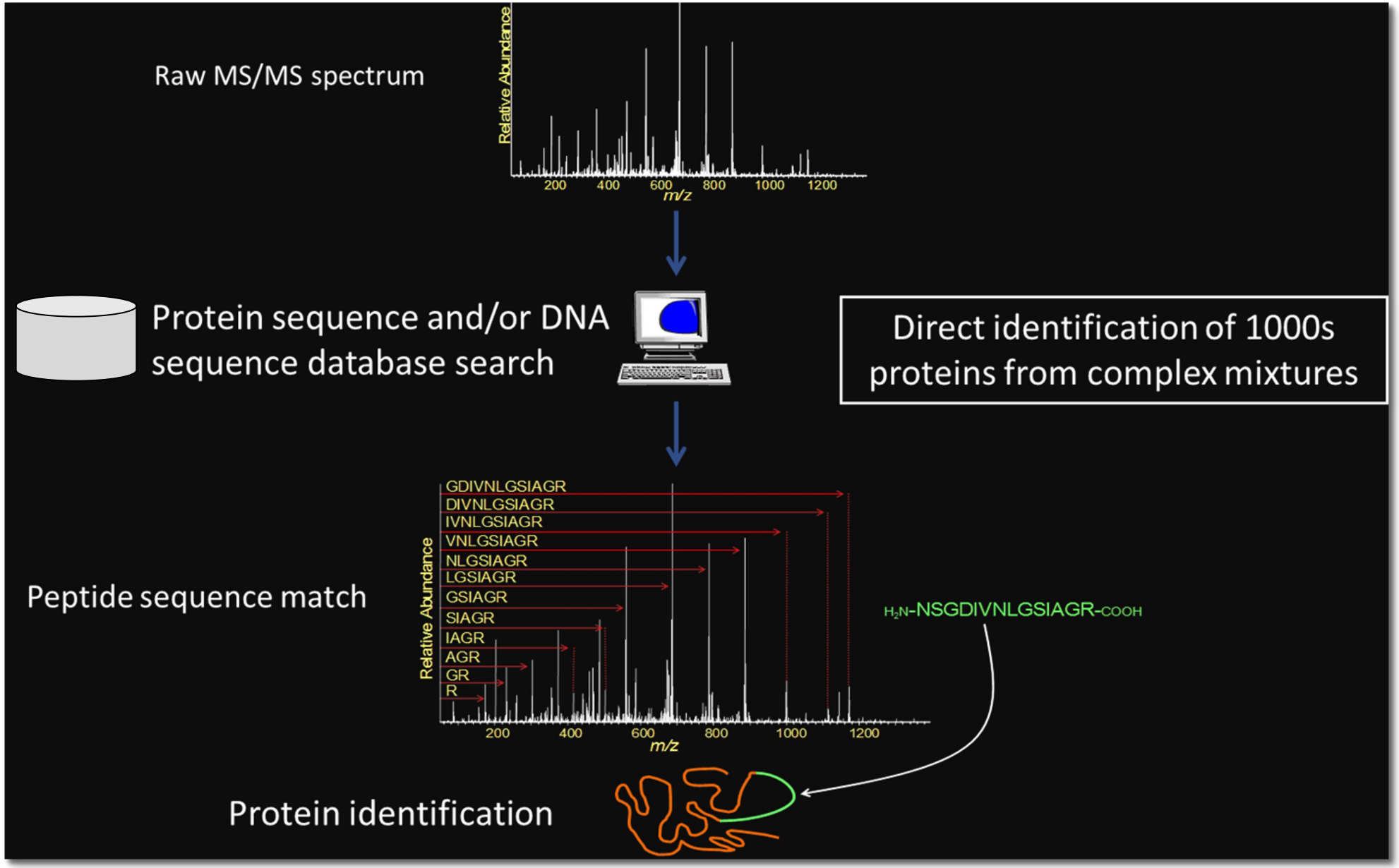




MS2

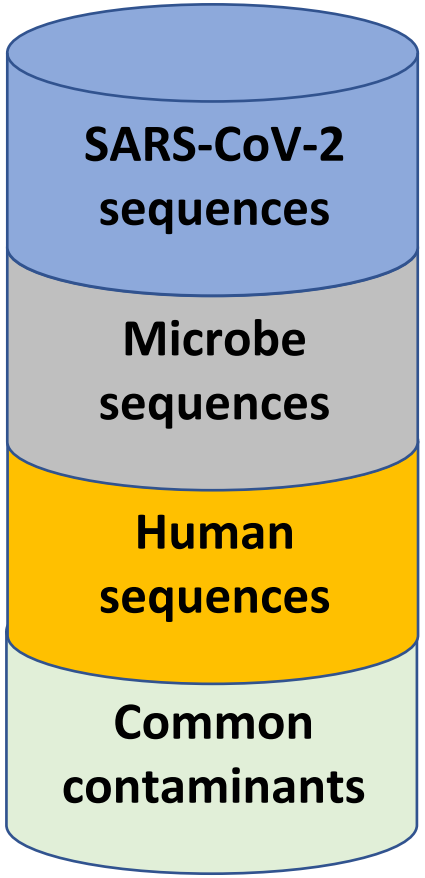
GalaxyP
galaxyp.org

Matching amino acid sequences to MS/MS data



GalaxyP
galaxyp.org

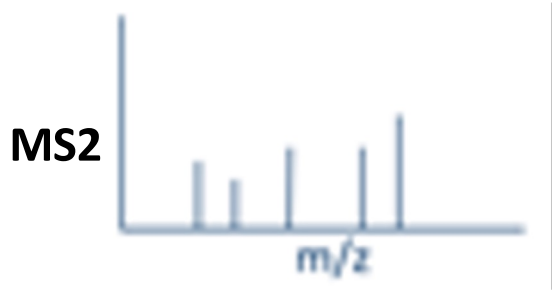
The importance of the protein sequence database



(including variant-specific sequences)

GalaxyP
galaxyp.org

Inferring protein identity for peptide sequences



CAQCHTVEK

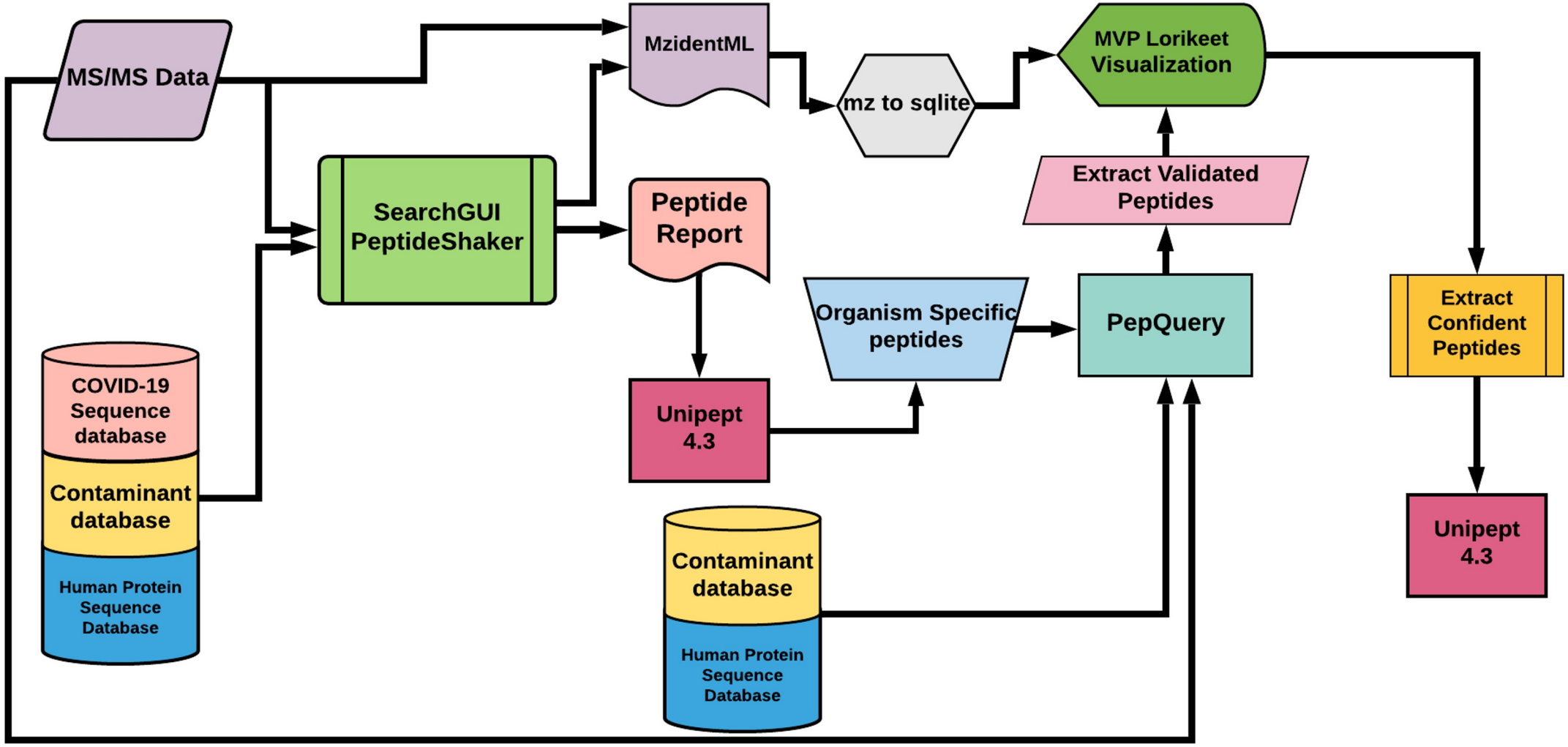
Cytochrome C

NH_2GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGIT
WKEETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE$COOH$

GalaxyP
galaxyp.org

Driven by bioinformatics

- Many step workflow for analyzing large and diverse datasets



GalaxyP
galaxyp.org

https://galaxyproject.org

Data Intensive *analysis* for everyone

Introduction to Galaxy PROJECT

- **Web-based** platform for computational biomedical research
- Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic
- Community driven

- **Open source** under Academic Free License
- More than 10,000 citations

- More than 125 **public Galaxy servers**
- Usegalaxy.* instances:
  - Usegalaxy.org, usegalaxy.org.au, usegalaxy.eu

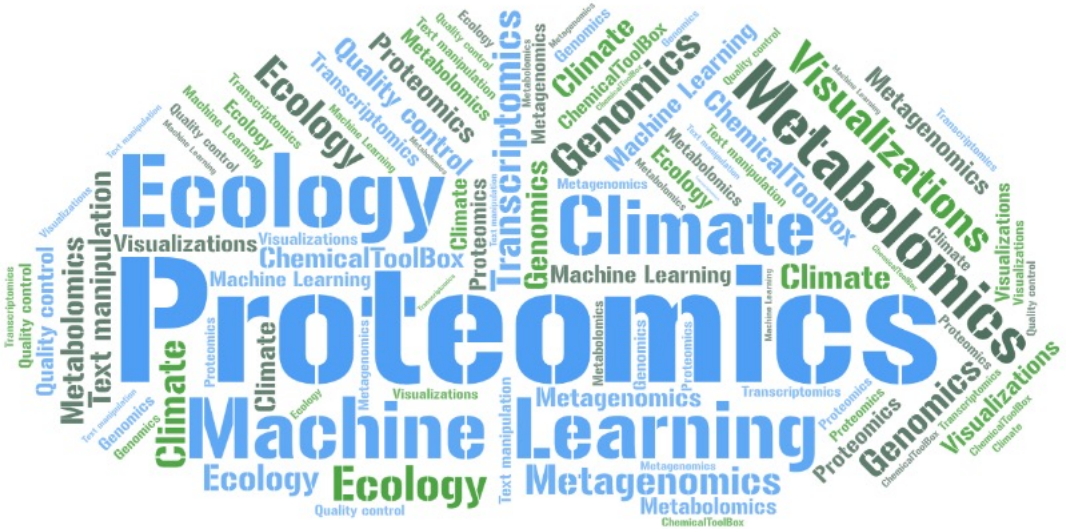
Core values

Galaxy PROJECT

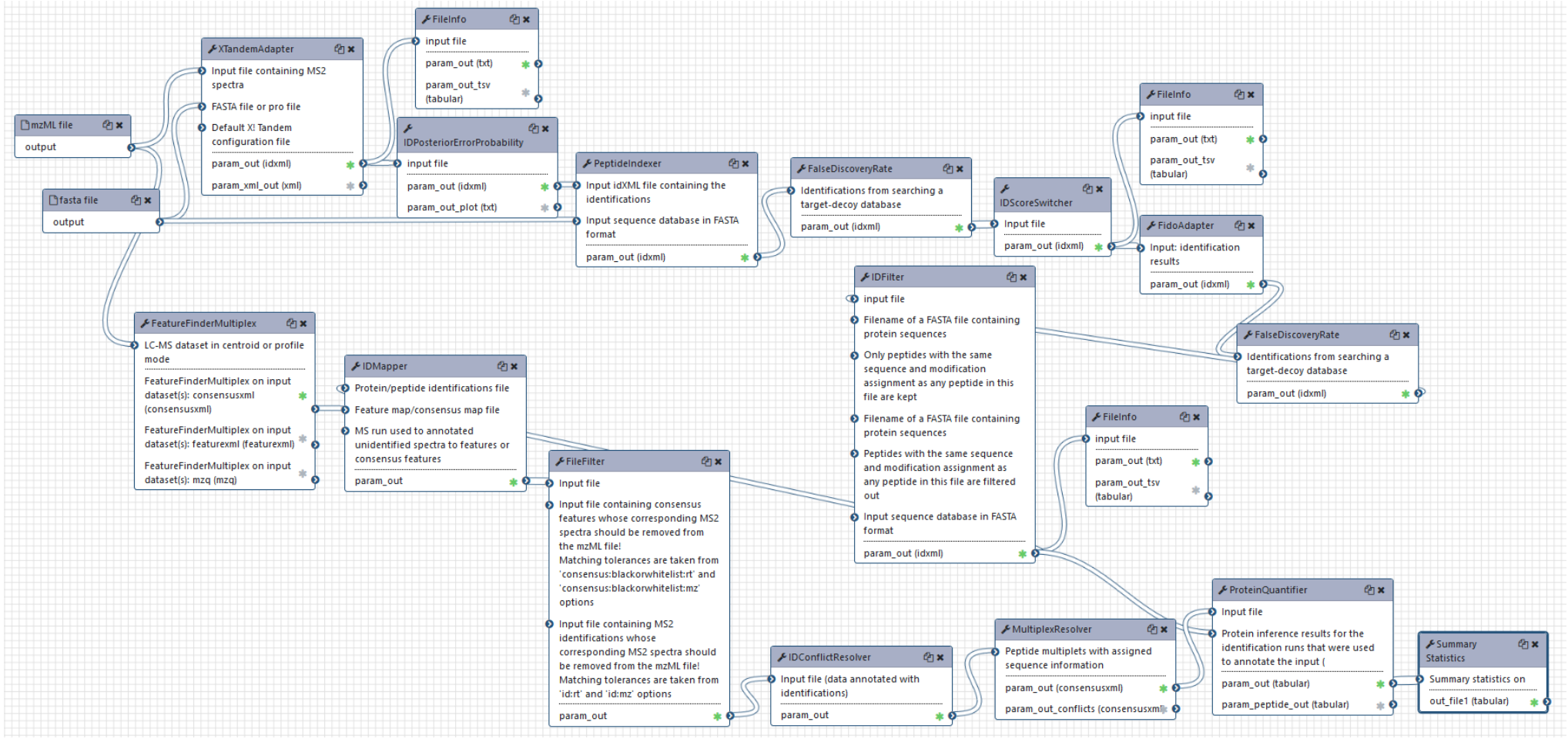
- **Accessibility**
  - Graphical user interface
  - Pre-installed tools
- **Reproducibility**
  - All information is captured in histories
  - Controlled tool versions
- **Transparency**
  - Sharing of histories and workflows

Available tools

- Galaxy tool shed: more than 7,000 tools
- Any open source tool can be integrated
- Areas:
  - Genomics
  - Transcriptomics
  - Proteomics
  - Metabolomics
  - ChemicalToolBox
  - Statistics
  - Ecology
  - Text manipulation
  - …



Tools can be connected into workflows



Hands-on training material

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	10
Assembly	5
Climate	2
Computational chemistry	6
Ecology	5
Epigenetics	6
Genome Annotation	3
Imaging	3
Metabolomics	4
Metagenomics	6
Proteomics	15
Sequence analysis	2
Statistics and machine learning	8
Transcriptomics	23
Variant Analysis	8
Visualisation	2

Galaxy Tips & Tricks

Topic	Tutorials
User Interface and Data Manipulation	16

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	35
Development in Galaxy	13

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

Galaxy for Contributors and Instructors

Topic	Tutorials
Contributing to the Galaxy Training Material	10
Teaching and Hosting Galaxy training	5

Galaxy Training Network

> 130 training materials

https://training.galaxyproject.org/training-material

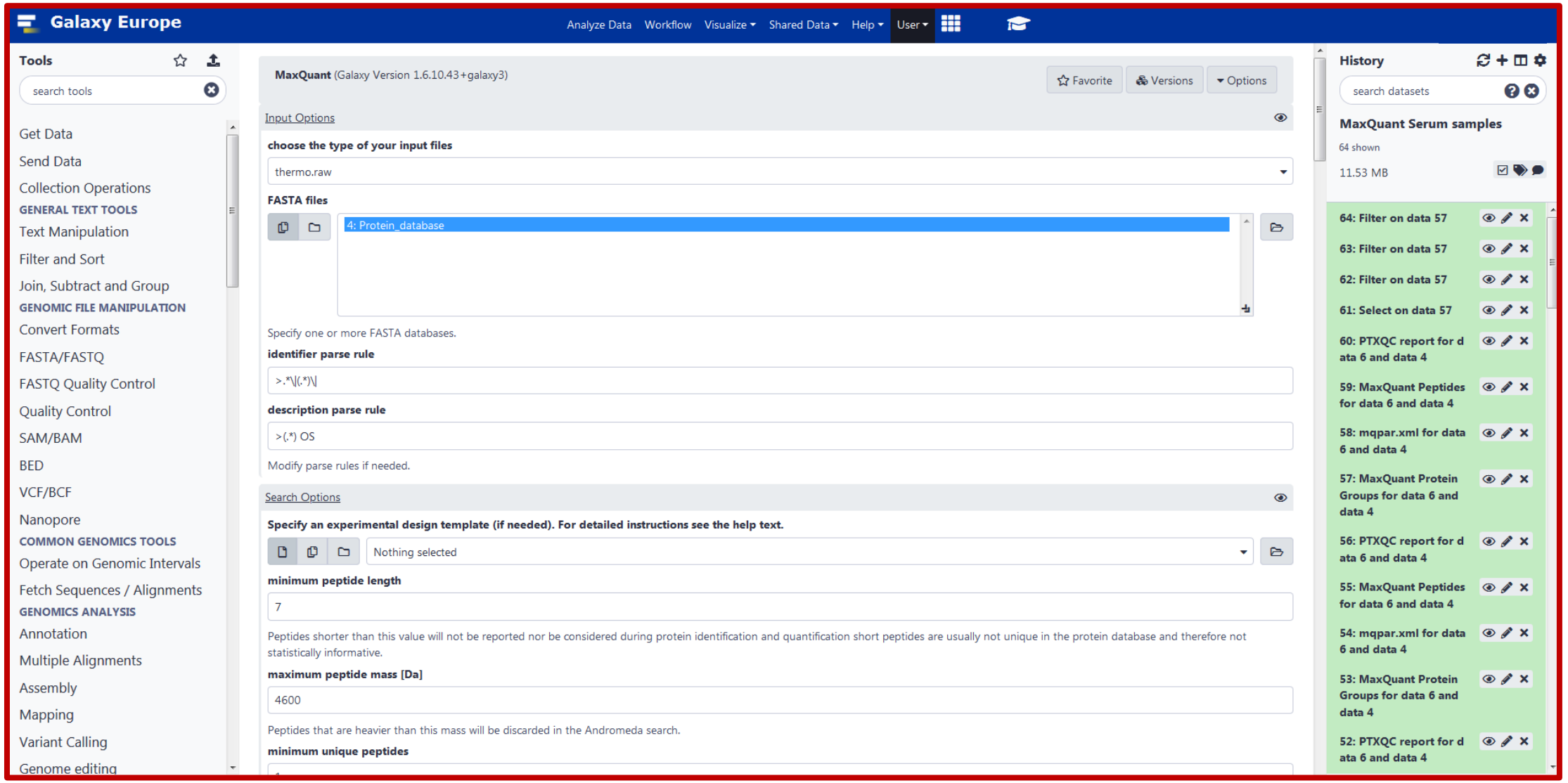
Galaxy community

GCC 2019 Freiburg

- > 200,000 user world wide
- > 250 contributor to Galaxy framework
- > 200 tool developer
- > 150 Galaxy training material contributors
- Annual international community conferences (GCC)

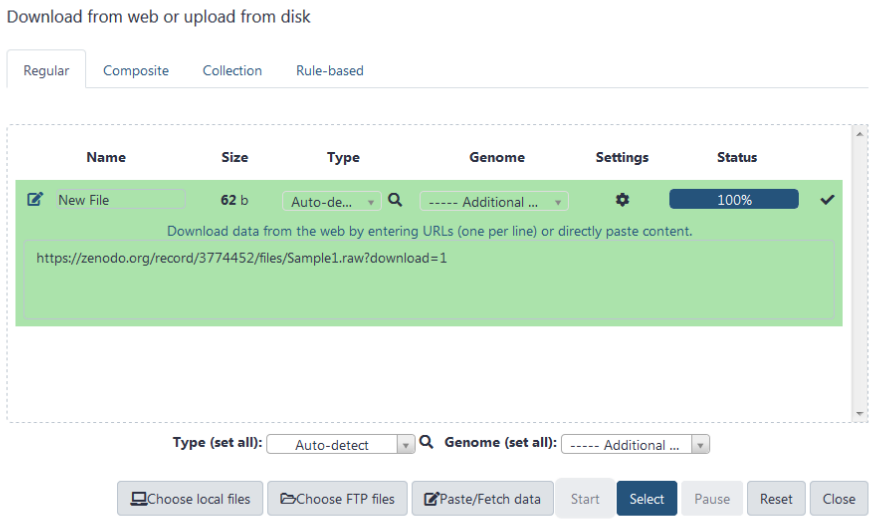
- Galaxy help forum: https://help.galaxyproject.org

Graphical user interface

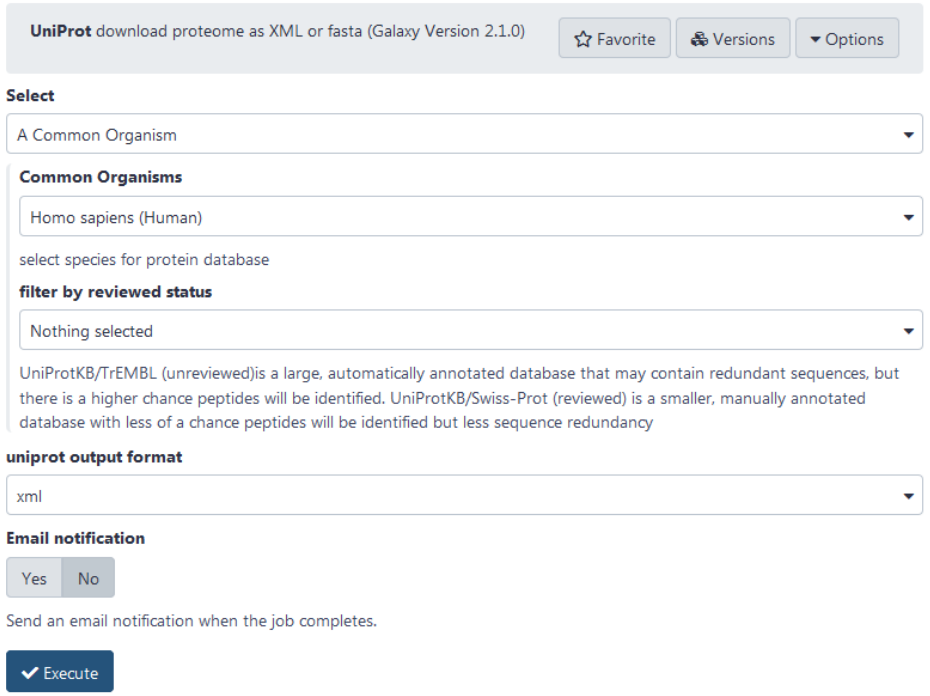


Data Input

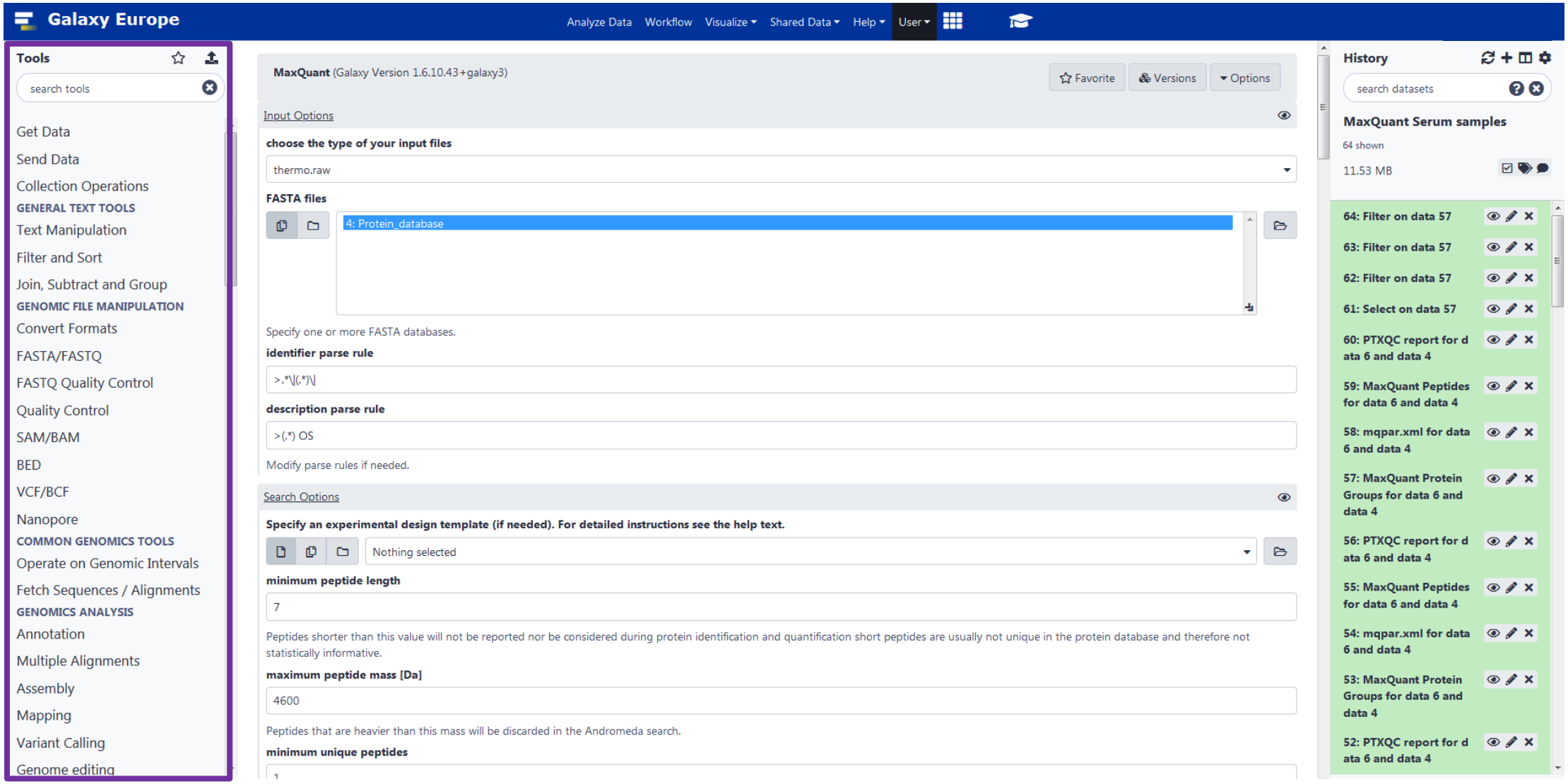
Upload from local, paste, ftp



Directly from repositories such as Uniprot



Tool panel


Galaxy Europe
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
search tools
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Sequences / Alignments
GENOMICS ANALYSIS
Annotation
Multiple Alignments
Assembly
Mapping
Variant Calling
Genome editing
MaxQuant (Galaxy Version 1.6.10.43+galaxy3)
Favorite
Versions
Options
Input Options
choose the type of your input files
thermo.raw
FASTA files
4: Protein_database
Specify one or more FASTA databases.
identifier parse rule
>.*\|(.*)\|
description parse rule
>(.*) OS
Modify parse rules if needed.
Search Options
Specify an experimental design template (if needed). For detailed instructions see the help text.
Nothing selected
minimum peptide length
7
Peptides shorter than this value will not be reported nor be considered during protein identification and quantification short peptides are usually not unique in the protein database and therefore not statistically informative.
maximum peptide mass [Da]
4600
Peptides that are heavier than this mass will be discarded in the Andromeda search.
minimum unique peptides
History
search datasets
MaxQuant Serum samples
64 shown
11.53 MB
64: Filter on data 57
63: Filter on data 57
62: Filter on data 57
61: Select on data 57
60: PTXQC report for data 6 and data 4
59: MaxQuant Peptides for data 6 and data 4
58: mqpar.xml for data 6 and data 4
57: MaxQuant Protein Groups for data 6 and data 4
56: PTXQC report for data 6 and data 4
55: MaxQuant Peptides for data 6 and data 4
54: mqpar.xml for data 6 and data 4
53: MaxQuant Protein Groups for data 6 and data 4
52: PTXQC report for data 6 and data 4

The Galaxy Interface

Main viewing window
(workflow development, results visualization etc)

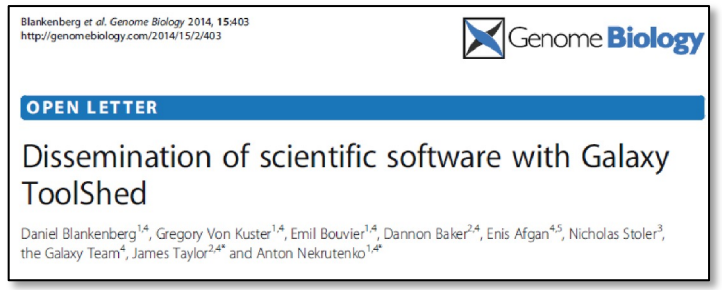
Tool
s

History

GalaxyP
galaxyp.org

Built for the community: Tool sharing and access

Galaxy Tool Shed



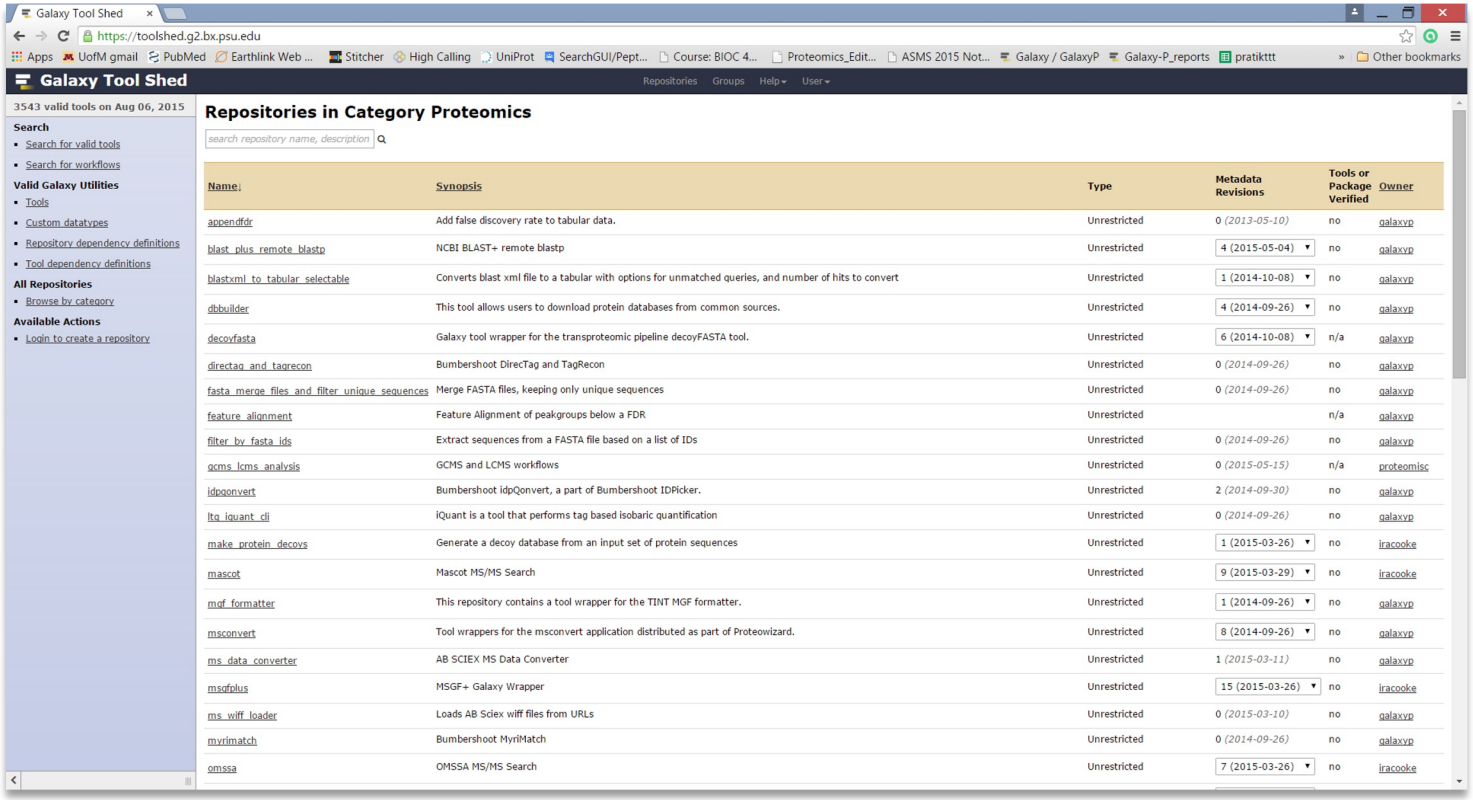
Blankenberg *et al. Genome Biology* 2014, **15**:403
http://genomebiology.com/2014/15/2/403

Genome Biology

OPEN LETTER

Dissemination of scientific software with Galaxy ToolShed

Daniel Blankenberg[1,4], Gregory Von Kuster[1,4], Emil Bouvier[1,4], Dannon Baker[2,4], Enis Afgan[4,5], Nicholas Stoler[3], the Galaxy Team[4], James Taylor[2,4*] and Anton Nekrutenko[1,4*]



Galaxy Tool Shed

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

Apps | UofM gmail | PubMed | Earthlink Web ... | Stitcher | High Calling | UniProt | SearchGUI/Pept... | Course: BIOC 4... | Proteomics_Edit... | ASMS 2015 Not... | Galaxy / GalaxyP | Galaxy-P_reports | pratikttt | Other bookmarks

Galaxy Tool Shed — Repositories | Groups | Help | User

3543 valid tools on Aug 06, 2015

Search
- Search for valid tools
- Search for workflows

Valid Galaxy Utilities
- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories
- Browse by category

Available Actions
- Login to create a repository

Repositories in Category Proteomics

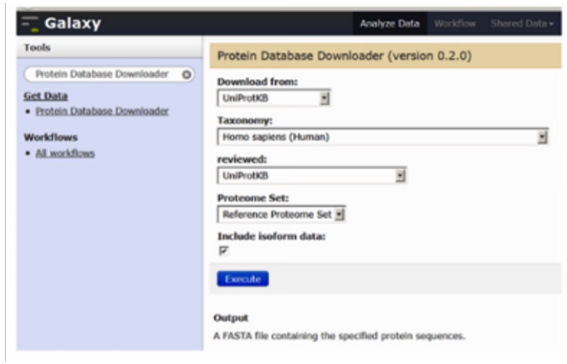
search repository name, description

Name	Synopsis	Type	Metadata Revisions	Tools or Package Verified	Owner
appendfdr	Add false discovery rate to tabular data.	Unrestricted	0 *(2013-05-10)*	no	galaxyp
blast_plus_remote_blastp	NCBI BLAST+ remote blastp	Unrestricted	4 (2015-05-04)	no	galaxyp
blastxml_to_tabular_selectable	Converts blast xml file to a tabular with options for unmatched queries, and number of hits to convert	Unrestricted	1 (2014-10-08)	no	galaxyp
dbbuilder	This tool allows users to download protein databases from common sources.	Unrestricted	4 (2014-09-26)	no	galaxyp
decoyfasta	Galaxy tool wrapper for the transproteomic pipeline decoyFASTA tool.	Unrestricted	6 (2014-10-08)	n/a	galaxyp
directag_and_tagrecon	Bumbershoot DirecTag and TagRecon	Unrestricted	0 *(2014-09-26)*	no	galaxyp
fasta_merge_files_and_filter_unique_sequences	Merge FASTA files, keeping only unique sequences	Unrestricted	0 *(2014-09-26)*	no	galaxyp
feature_alignment	Feature Alignment of peakgroups below a FDR	Unrestricted		n/a	galaxyp
filter_by_fasta_ids	Extract sequences from a FASTA file based on a list of IDs	Unrestricted	0 *(2014-09-26)*	no	galaxyp
gcms_lcms_analysis	GCMS and LCMS workflows	Unrestricted	0 *(2015-05-15)*	n/a	proteomisc
idpqonvert	Bumbershoot idpQonvert, a part of Bumbershoot IDPicker.	Unrestricted	2 *(2014-09-30)*	no	galaxyp
ltq_iquant_cli	iQuant is a tool that performs tag based isobaric quantification	Unrestricted	0 *(2014-09-26)*	no	galaxyp
make_protein_decoys	Generate a decoy database from an input set of protein sequences	Unrestricted	1 (2015-03-26)	no	iracooke
mascot	Mascot MS/MS Search	Unrestricted	9 (2015-03-29)	no	iracooke
mgf_formatter	This repository contains a tool wrapper for the TINT MGF formatter.	Unrestricted	1 (2014-09-26)	no	galaxyp
msconvert	Tool wrappers for the msconvert application distributed as part of Proteowizard.	Unrestricted	8 (2014-09-26)	no	galaxyp
ms_data_converter	AB SCIEX MS Data Converter	Unrestricted	1 *(2015-03-11)*	no	galaxyp
msgfplus	MSGF+ Galaxy Wrapper	Unrestricted	15 (2015-03-26)	no	iracooke
ms_wiff_loader	Loads AB Sciex wiff files from URLs	Unrestricted	0 *(2015-03-10)*	no	galaxyp
myrimatch	Bumbershoot MyriMatch	Unrestricted	0 *(2014-09-26)*	no	galaxyp
omssa	OMSSA MS/MS Search	Unrestricted	7 (2015-03-26)	no	iracooke

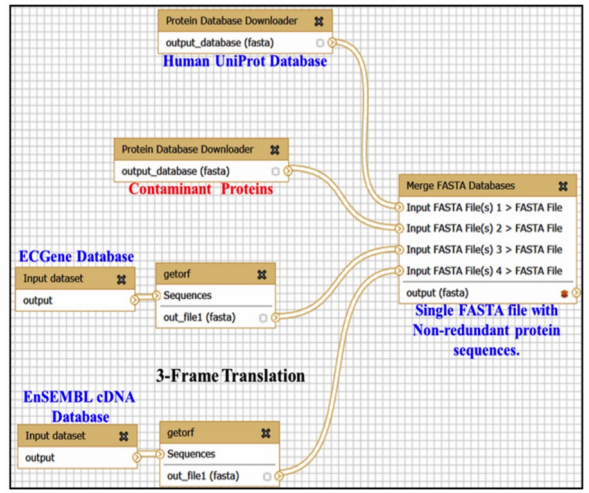
GalaxyP
galaxyp.org

Defining Workflows and Histories

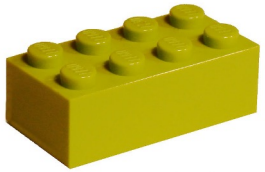
Workflows: Multiple software tools used in a sequential manner for an analysis

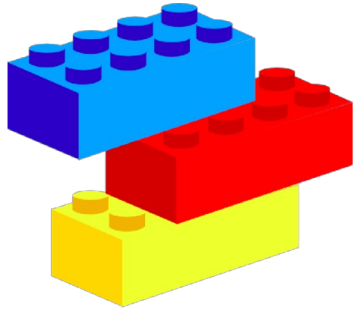


Single software tool in Galaxy

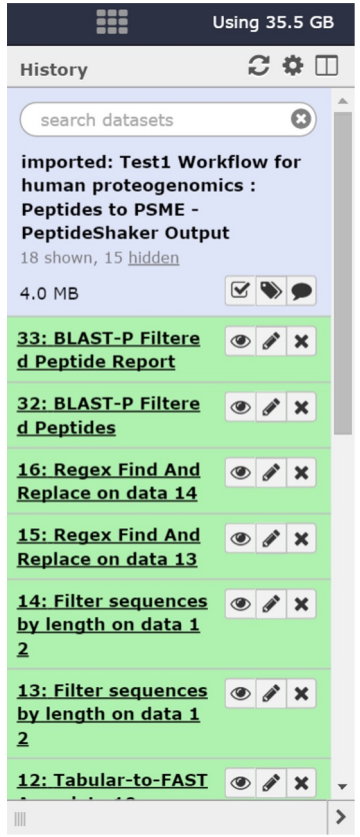


Integrated workflow using multiple, connected tools





History: recorded, complete analysis (workflow + all data and results)



GalaxyP
galaxyp.org

Galaxy EUROPE

2500 ready to use tools

(https://galaxyproject.eu/tools)

&

~ 6000 cores, 50 TB RAM, 2 PB storage

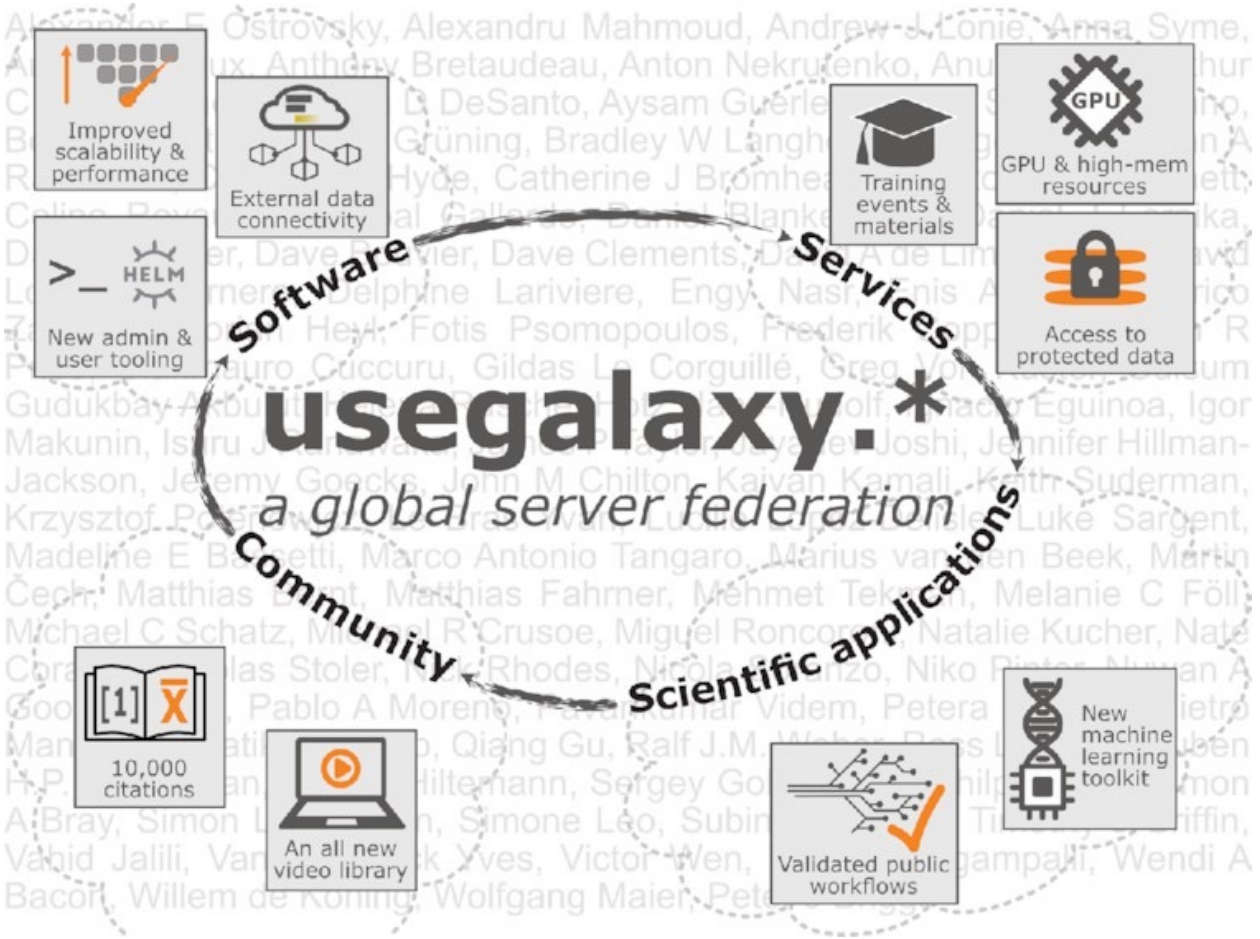
de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

DFG

Bundesministerium für Bildung und Forschung

elixir

Graphical Abstract The Galaxy Project is represented by a combination of software, managed services, application of the software ...

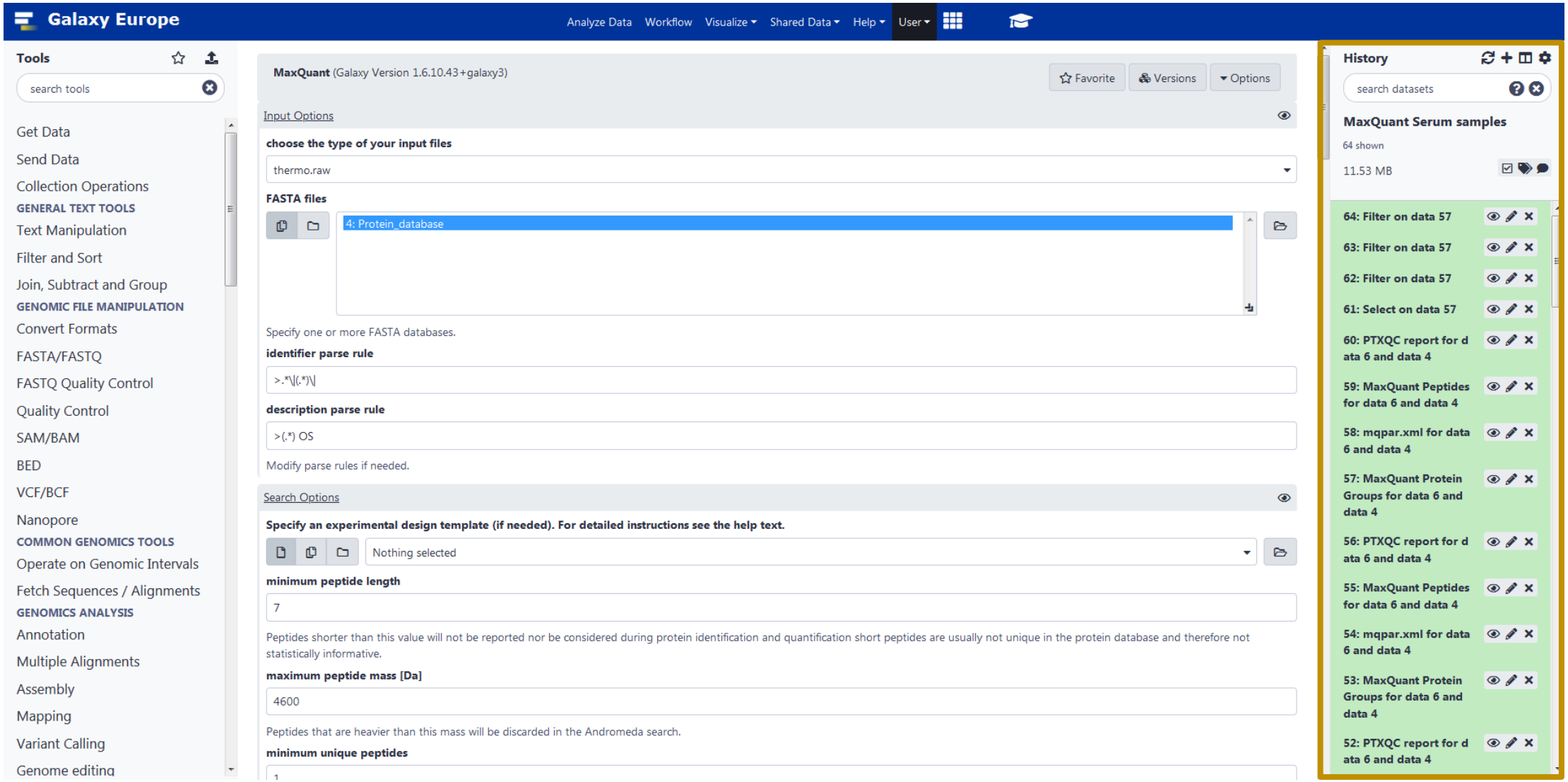


Nucleic Acids Res, gkac247, https://doi.org/10.1093/nar/gkac247

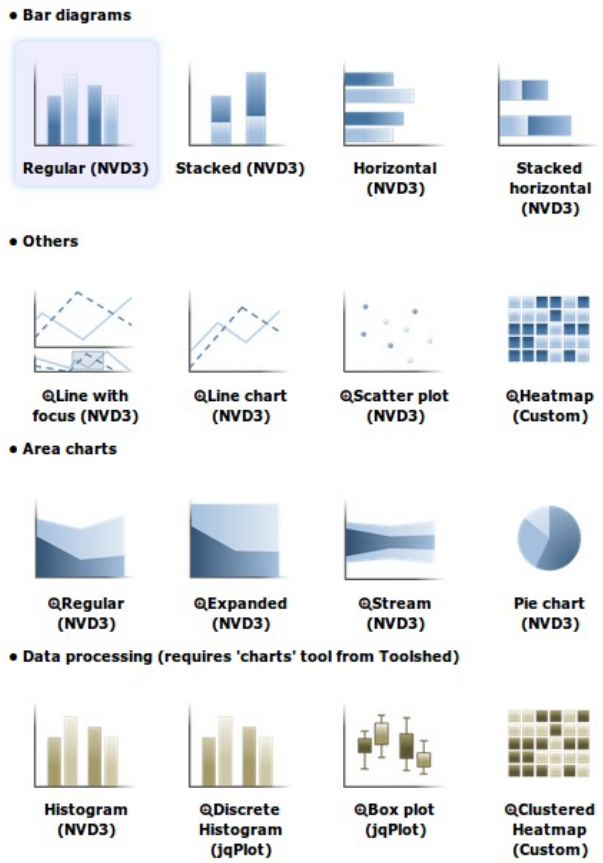
The content of this slide may be subject to copyright: please see the slide notes for details.

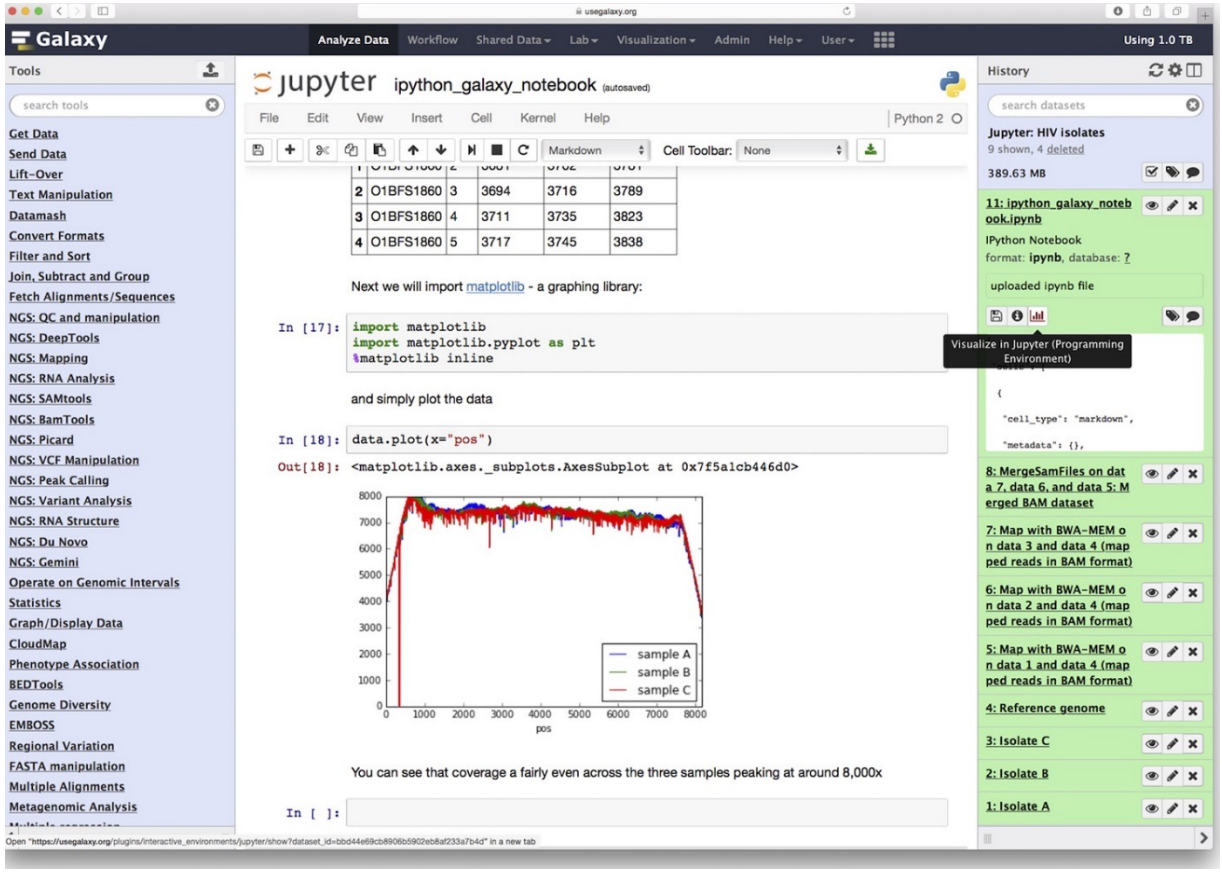
OXFORD UNIVERSITY PRESS

Analysis history



Visualizations & Interactive Environments





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

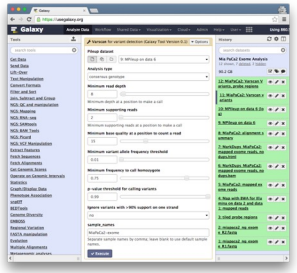
Amenable to other 'omic tools (e.g. Galaxy for proteomics project, Galaxy-P)

Working philosophy:

Galaxy: an integrative workbench well-suited for multi-omics

Interfaces

Web UI



Programmatic API

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

Datasets

Computing Resources

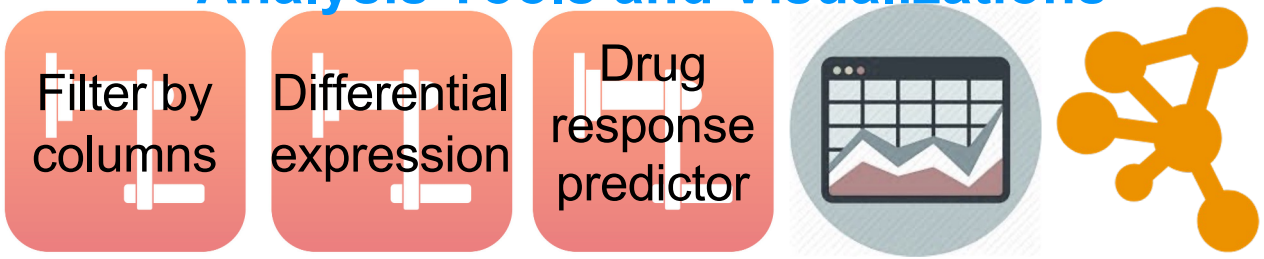
slurm workload manager

Condor High Throughput Computing

Google Cloud Platform

amazon web services

Analysis Tools and Visualizations



GalaxyP galaxyp.org

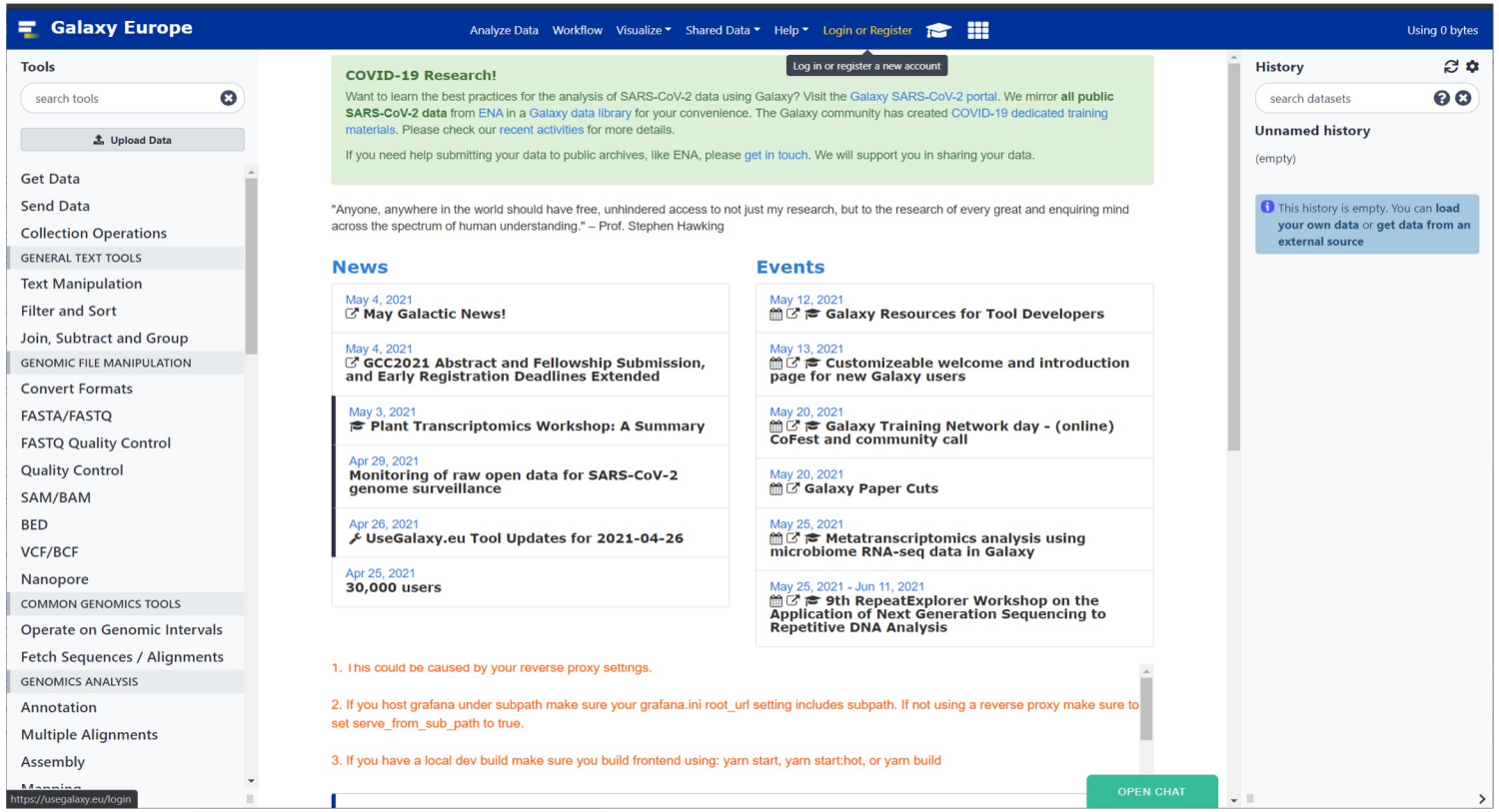
Courtesy Jeremy Goecks, OHSU

Accessing Galaxy via a public gateway

1. Login/Register: usegalaxy.eu

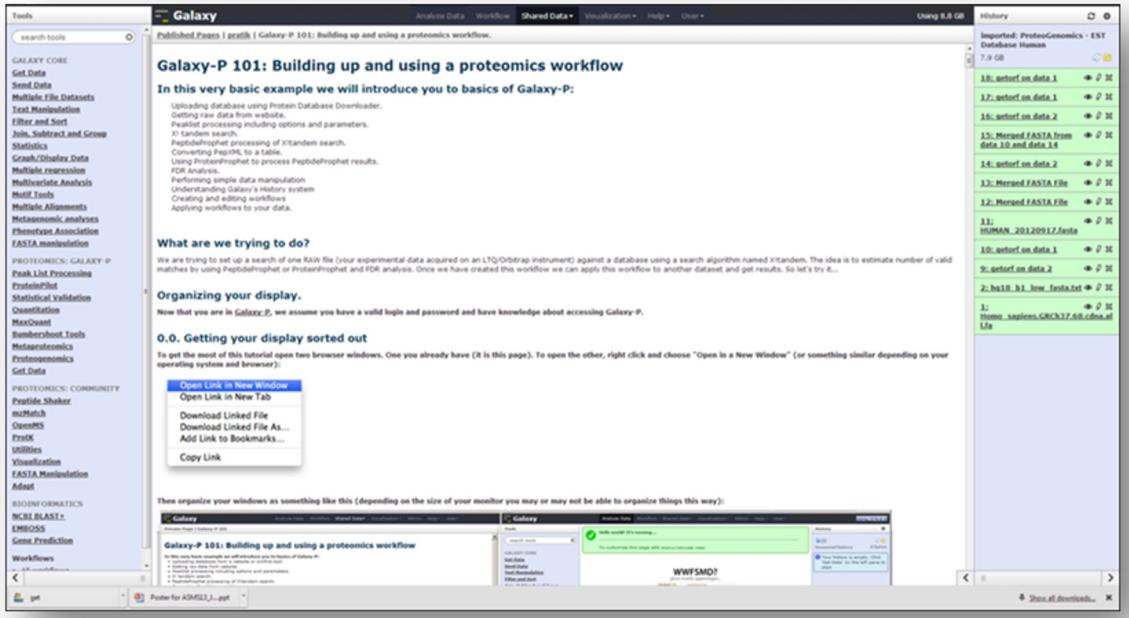
2. Go to TIaaS link: https://usegalaxy.eu/join-training/quant-meta

3. Return to usegalaxy.eu site



GalaxyP
galaxyp.org

Built for the community: Sharing workflows and histories



URL export

File export (.ga)

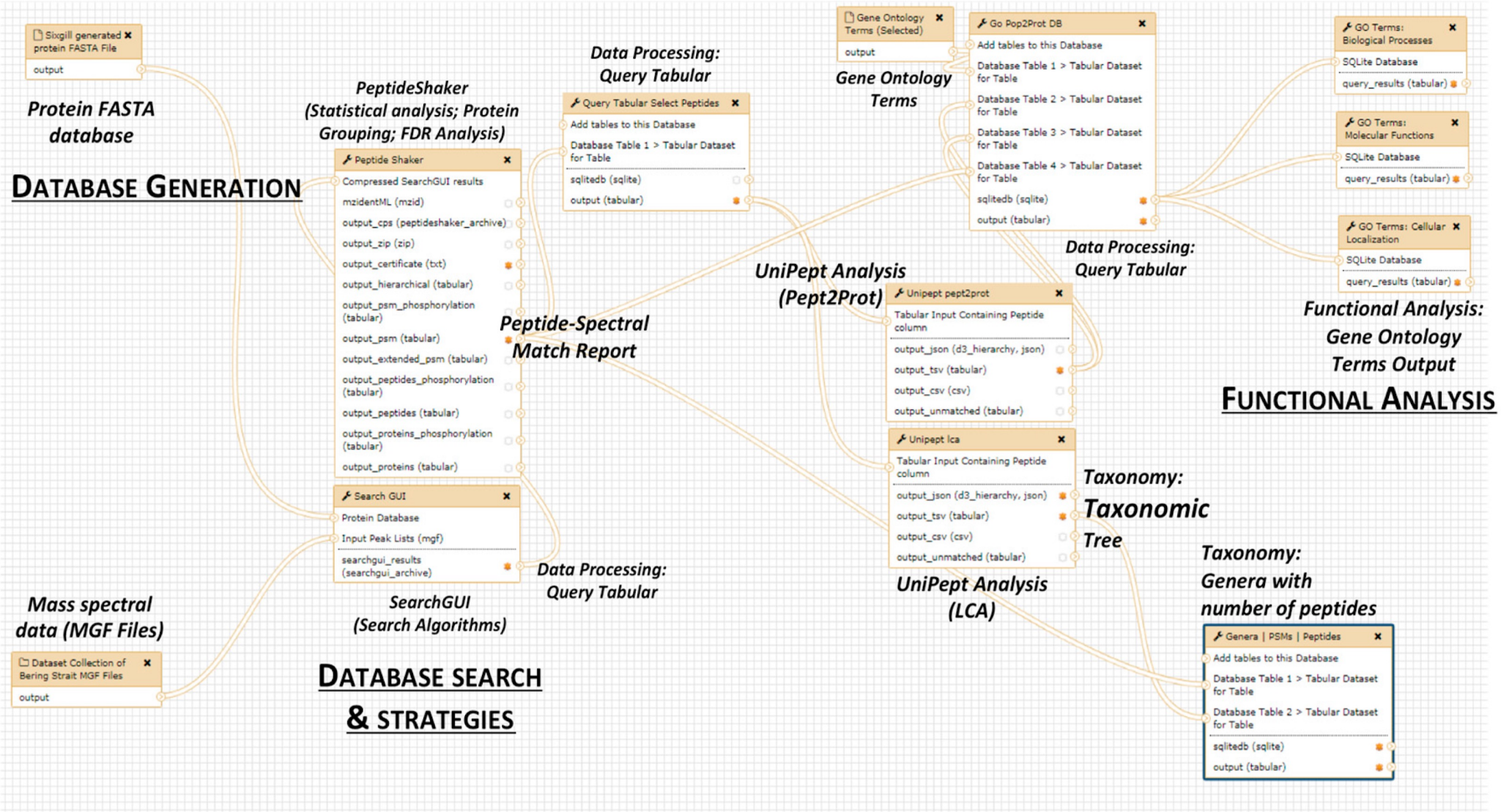
HISTORY: https://galaxyp.msi.umn.edu/u/pjagtap/h/itraq-search-yang-2-xtandem-scaffold

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**

GalaxyP
galaxyp.org

Solution: Galaxy Bioinformatics Platform



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

Metaproteomics Publications

PROTEOMICS

Jagtap P *et al* (2012) Deep metaproteomic analysis of **human salivary supernatant**. Proteomics. 12(7):992-1001. doi: 10.1002/ pmic.201100503.

Microbiome

Rudney JD *et al* (2015) Protein relative abundance patterns associated with **sucrose-induced dysbiosis** are conserved across taxonomically diverse oral microcosm **biofilm models of dental caries.** Microbiome. 3:69. doi: 10.1186/s40168-015-0136-z.

SCIENTIFIC REPORTS nature research

Afiuni-Zadeh S *et al* (2018) Evaluating the potential of residual Pap test fluid as a resource for the metaproteomic analysis of the **cervical-vaginal microbiome**. Sci Rep. ;8(1):10868. doi: 10.1038/s41598-018-29092-4.

ATS Journals

AJRCMB

Jagtap PD et al (2018) **BAL Fluid Metaproteome** in Acute Respiratory Failure. Am J Respir Cell Mol Biol. 59(5):648-652. doi: 10.1165/rcmb.2018-0068LE.

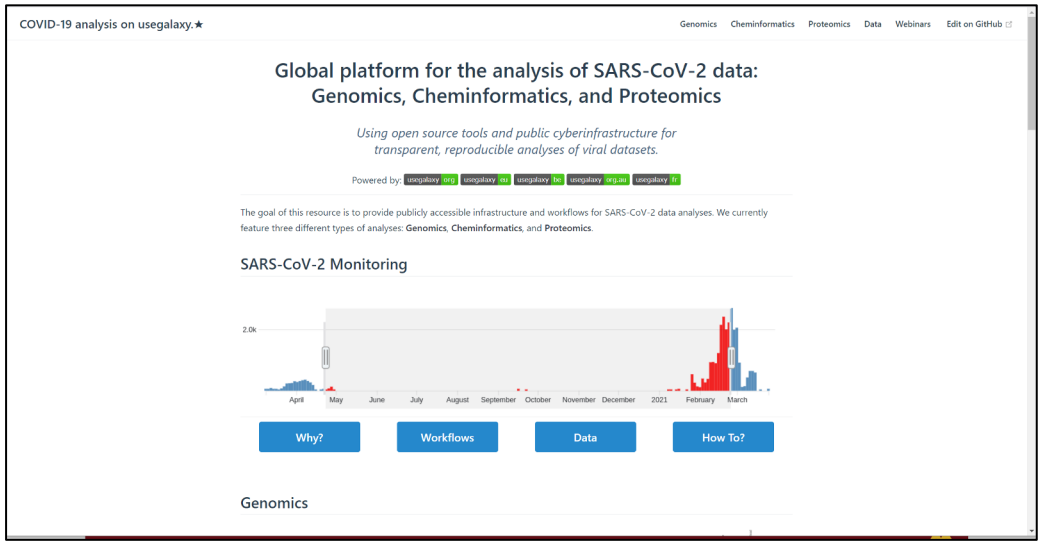
ACS Publications Most Trusted. Most Cited. Most Read.

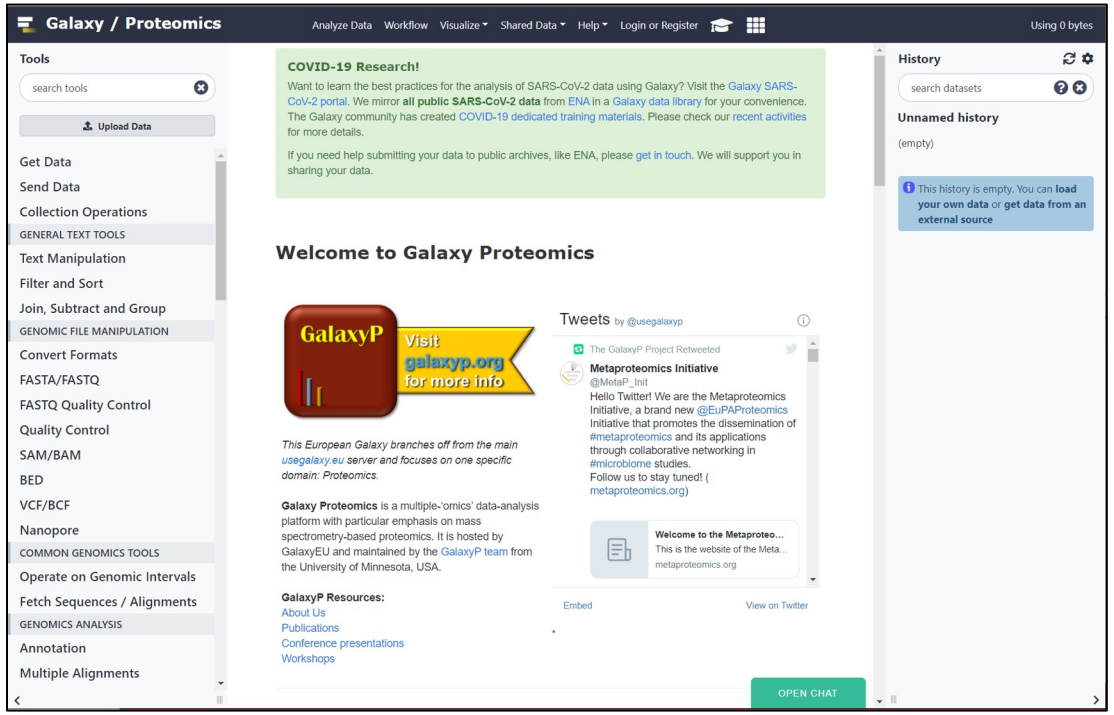
Journal of proteome research

Thuy-Boun PS *et al* (2021). Metaproteomics Analysis of **SARS-CoV-2-Infected Patient Samples** Reveals Presence of Potential Coinfecting Microorganisms. J Proteome Res. 20(2):1451-1454.doi: 10.1021/acs.jproteome.0c00822

Accessing and using Galaxy

useGalaxy.*

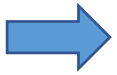


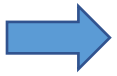


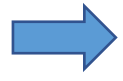
Galaxy Tool Shed

HANDS-ON SESSION

Instructions







Please **Register** for creating an account with a valid email ID and Password at usegalaxy.eu.

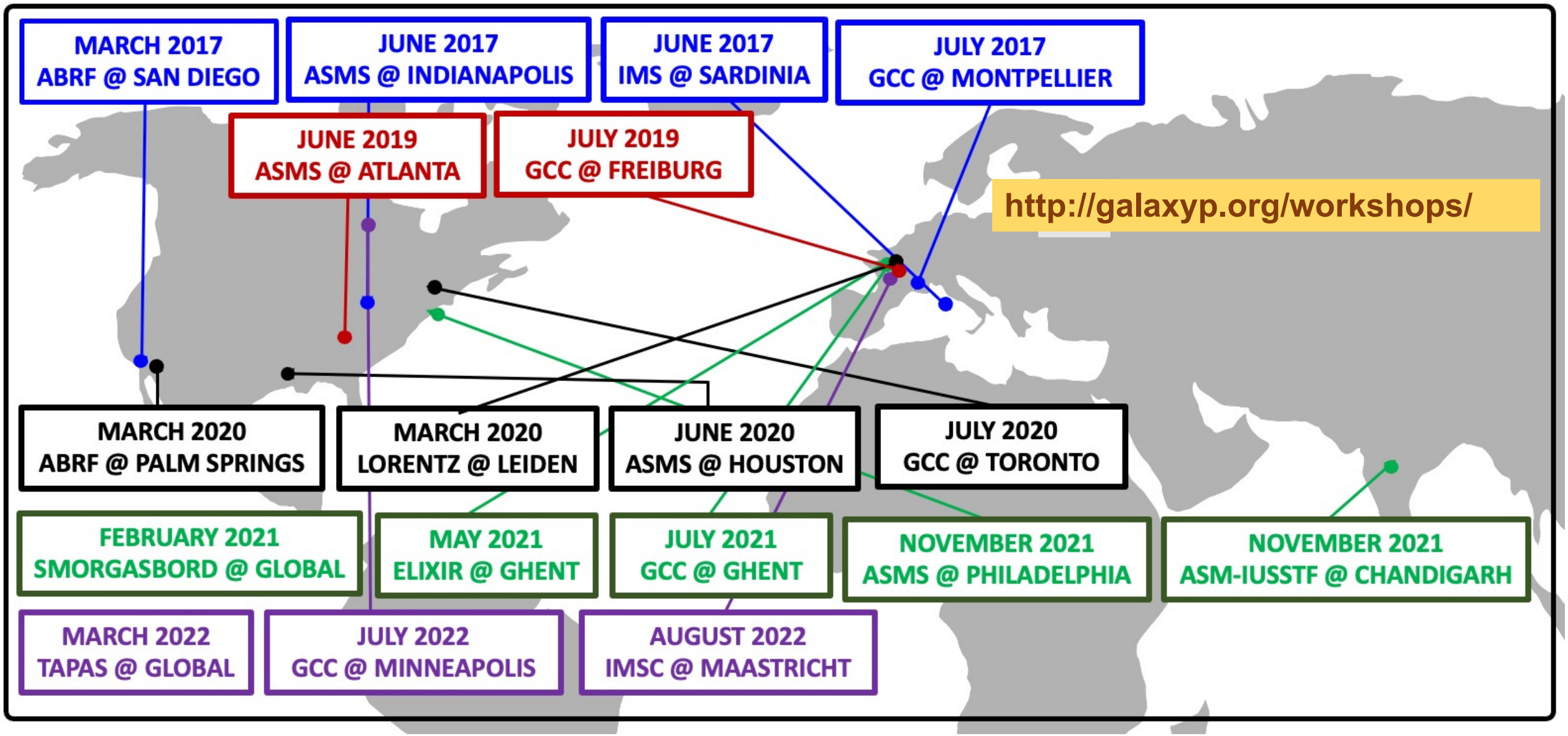
Once Registered, click on TIAAS to join the GCC 2022 Galaxy session. https://usegalaxy.eu/join-training/quant-meta

Go to Shared Data

Go to Shared Data

Run the workflow on active history

Workshops


MARCH 2017
ABRF @ SAN DIEGO
JUNE 2017
ASMS @ INDIANAPOLIS
JUNE 2017
IMS @ SARDINIA
JULY 2017
GCC @ MONTPELLIER
JUNE 2019
ASMS @ ATLANTA
JULY 2019
GCC @ FREIBURG
http://galaxyp.org/workshops/
MARCH 2020
ABRF @ PALM SPRINGS
MARCH 2020
LORENTZ @ LEIDEN
JUNE 2020
ASMS @ HOUSTON
JULY 2020
GCC @ TORONTO
FEBRUARY 2021
SMORGASBORD @ GLOBAL
MAY 2021
ELIXIR @ GHENT
JULY 2021
GCC @ GHENT
NOVEMBER 2021
ASMS @ PHILADELPHIA
NOVEMBER 2021
ASM-IUSSTF @ CHANDIGARH
MARCH 2022
TAPAS @ GLOBAL
JULY 2022
GCC @ MINNEAPOLIS
AUGUST 2022
IMSC @ MAASTRICHT

Galaxy Training Network

https://galaxyproject.org/events/gcc2021/training/

https://training.galaxyproject.org/training-material/topics/proteomics/

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.

TEN YEARS OF GALAXY-P !

PubMed

z.umn.edu/galaxypreferences

https://github.com/galaxyproteomics

GalaxyP

galaxyp.org

https://training.galaxyproject.org

Galaxy Training Network

http://galaxyp.org/workshops/

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/

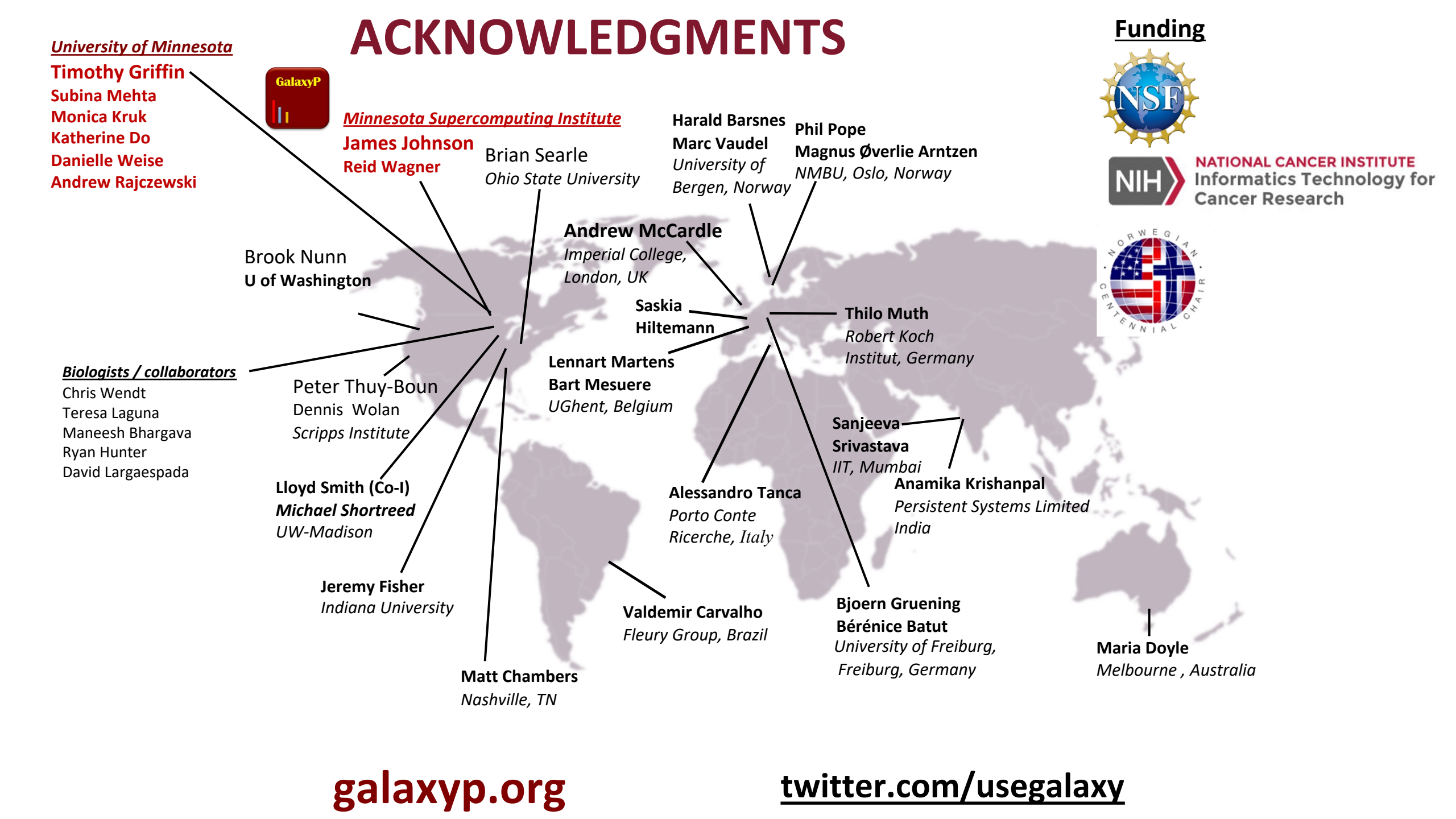
ACKNOWLEDGMENTS

Funding

NSF

NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

NORWEGIAN CENTENNIAL CHAIR



University of Minnesota

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

Minnesota Supercomputing Institute

James Johnson
Reid Wagner

Brian Searle
Ohio State University

Harald Barsnes
Marc Vaudel
University of Bergen, Norway

Phil Pope
Magnus Øverlie Arntzen
NMBU, Oslo, Norway

Andrew McCardle
Imperial College, London, UK

Brook Nunn
U of Washington

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

Peter Thuy-Boun
Dennis Wolan
Scripps Institute

Saskia Hiltemann

Thilo Muth
Robert Koch Institut, Germany

Lennart Martens
Bart Mesuere
UGhent, Belgium

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

Sanjeeva Srivastava
IIT, Mumbai

Anamika Krishanpal
Persistent Systems Limited India

Alessandro Tanca
Porto Conte Ricerche, Italy

Jeremy Fisher
Indiana University

Valdemir Carvalho
Fleury Group, Brazil

Bjoern Gruening
Bérénice Batut
University of Freiburg, Freiburg, Germany

Maria Doyle
Melbourne , Australia

Matt Chambers
Nashville, TN

galaxyp.org

twitter.com/usegalaxy