

Cloud Resources for Proteomics Analysis

June 8, 2022

5:45 – 7:00 PM, Room 200 FG



Cloud computing and proteomics



<https://scouttg.com/blog/articles/what-is-cloud-computing/>

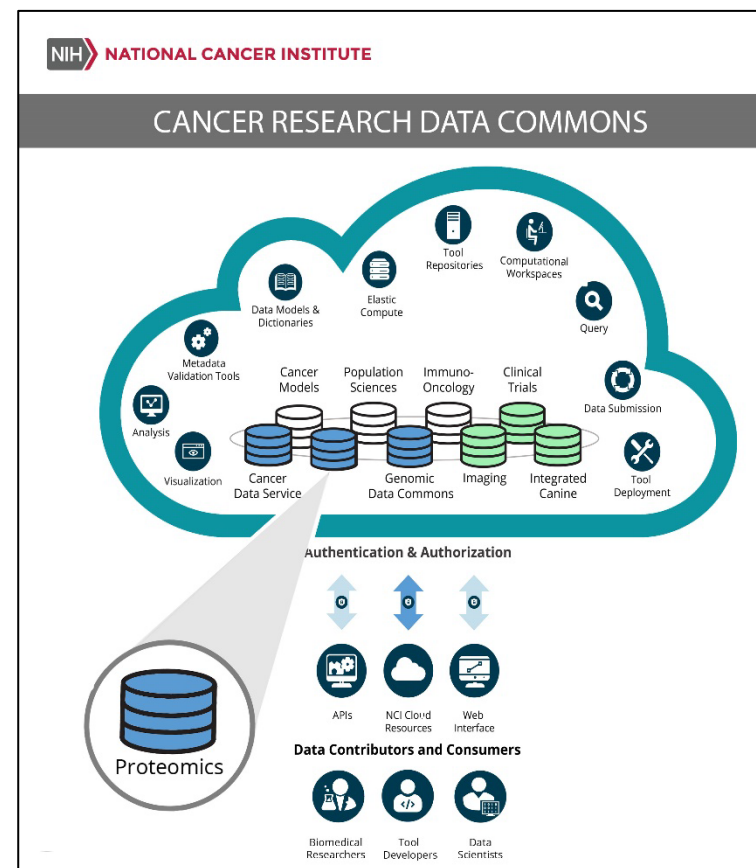


"The practice of using a network of remote servers hosted on the internet to store, manage, and process data, rather than a local server or a personal computer."

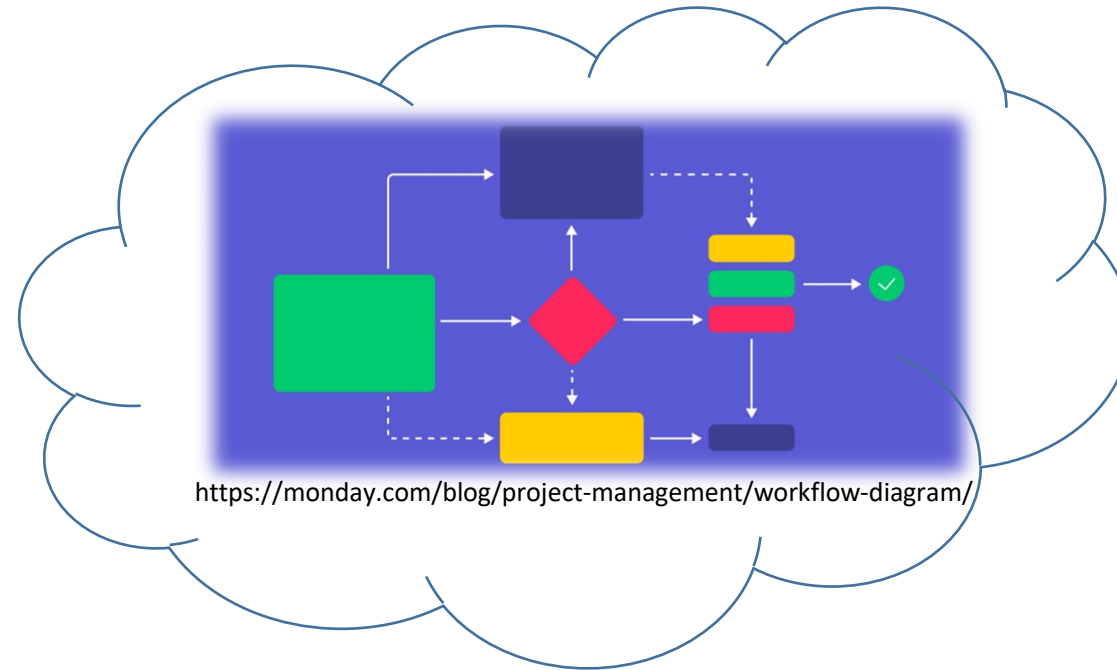
Cloud resources for proteomics: many and diverse



Mass Spectrometry
Interactive Virtual Environment



Our focus: cloud resources for data analysis tools and workflows



Talk 1: TPP resources for Proteomics Analysis (Michael Hoopmann): 10 minutes

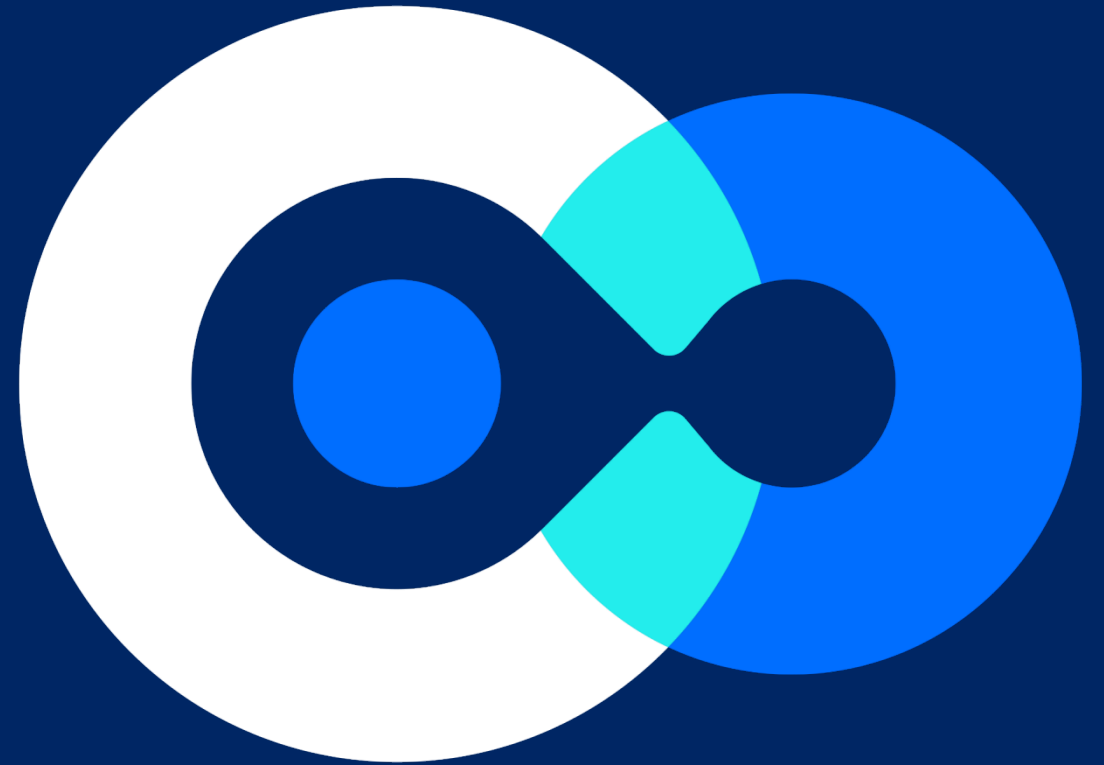
Talk 2: Galaxy resources for Proteomics Analysis (Pratik Jagtap): 10 minutes

Talk 3: Nextflow for Proteomics Analysis. (Veit Schwämmle): 10 minutes

Panel Discussion

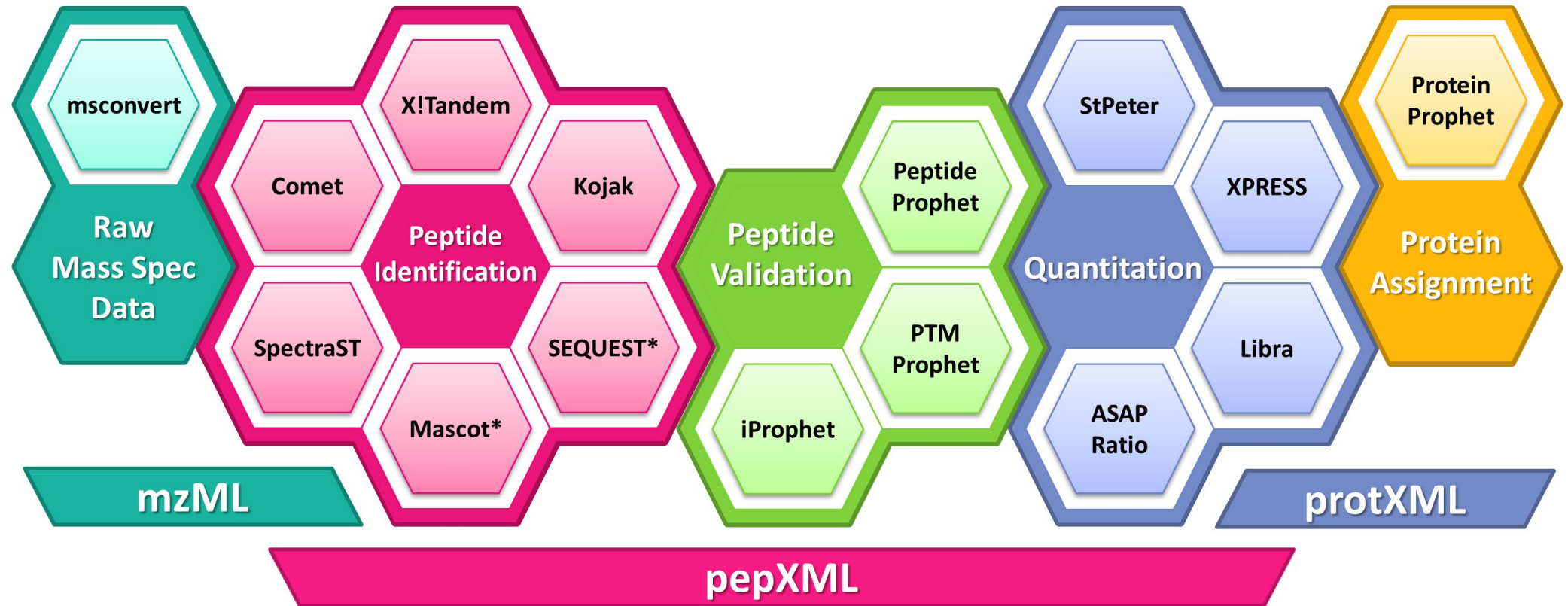


Trans-Proteomic Pipeline Resources for Proteomics Analysis



Michael Hoopmann, Institute for Systems Biology
ASMS 2022

Trans-Proteomic Pipeline (TPP) Overview

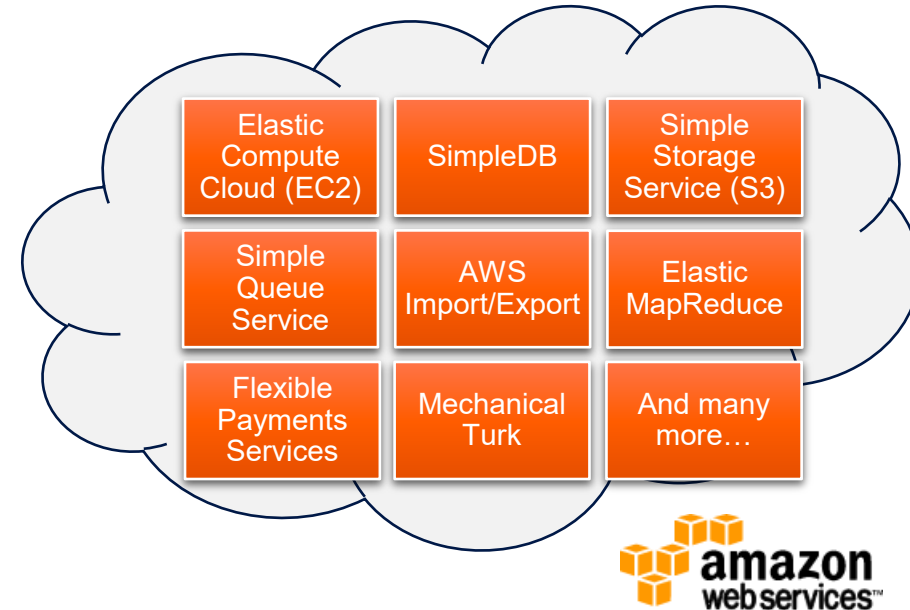
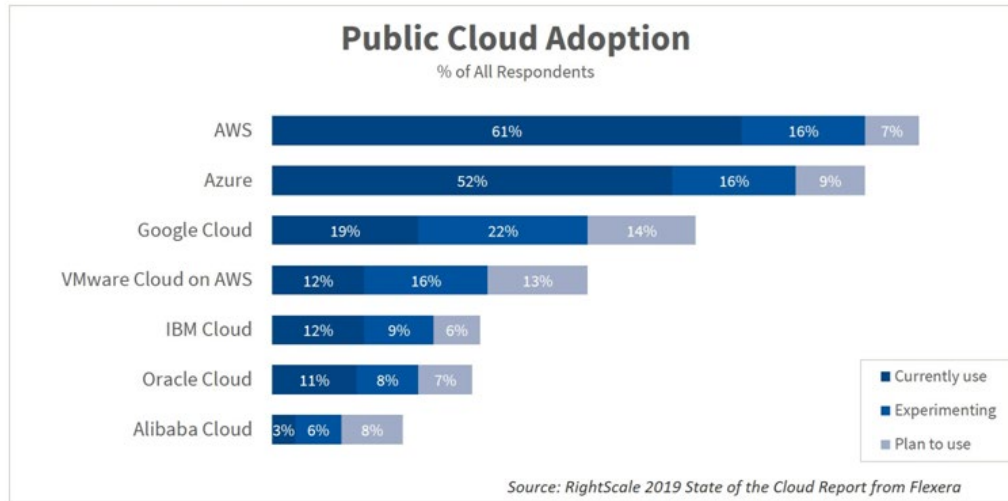


Free and open source suite of **software tools** and **file formats** that facilitates and standardizes proteomics analysis

Runs on Windows, Linux, and cloud platforms



Cloud Computing and Amazon Web Services



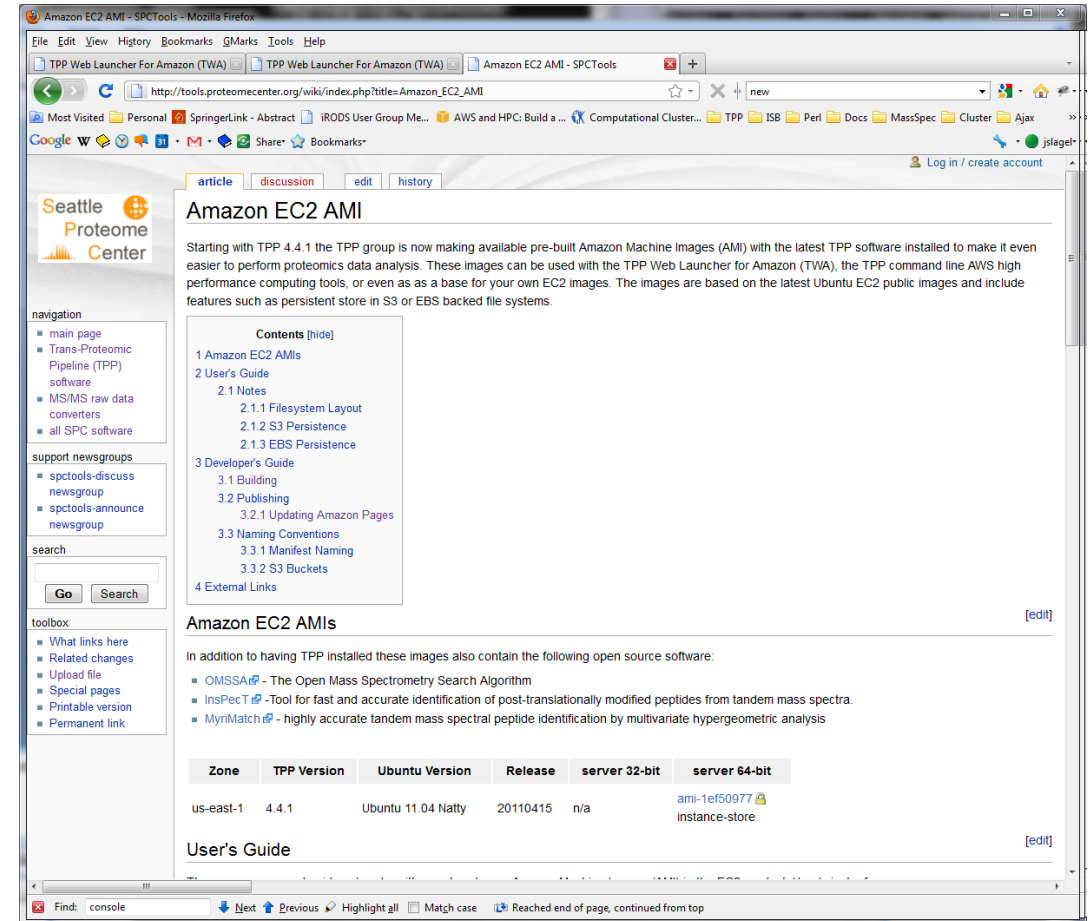
- Collection of web computing services offered by Amazon
- “Elastic” IT infrastructure – allocate computers, storage, and other services as needed
- Cost effective -- pay only for what you use
- Easy to use – simple API accessed over HTTP which supports almost every language
- Large number of tools available built for it



TPP Amazon Images

Publicly available Amazon Machine Instances (AMI) for the TPP

- **Based on official public releases of Ubuntu**
- **Contain additional open software (OMSSA, Myrimatch, etc.)**
- **Publicly available scripts for building, updating and publishing images**
- **Instructions on usage and details documented on wiki site**



The screenshot shows a Mozilla Firefox browser window displaying the 'Amazon EC2 AMI' page on the TPP wiki. The page title is 'Amazon EC2 AMI' and the URL is 'http://tools.proteomecenter.org/wiki/index.php?title=Amazon_EC2_AMI'. The page content includes a 'Contents' section with links to '1 Amazon EC2 AMIs', '2 User's Guide', '3 Developer's Guide', and '4 External Links'. Below the contents, there is a section titled 'Amazon EC2 AMIs' which lists the open source software included in the images: OMSSA, insPecT, and MyriMatch. A table provides details for the AMI, including the zone, TPP version, Ubuntu version, release, server architecture, and AMI ID.

Zone	TPP Version	Ubuntu Version	Release	server 32-bit	server 64-bit
us-east-1	4.4.1	Ubuntu 11.04 Natty	20110415	n/a	ami-1ef50977 instance-store

http://tools.proteomecenter.org/wiki/index.php?title=Amazon_EC2_AMI



Using TPP on the Cloud



TPP Web Application (TWA)

- Simple web based launcher to start petunia on a Amazon server
- Starts up an pre-configured TPP instance
- Doesn't require any software installation and is inexpensive to run
- Great tool for just trying out TPP
- Can be used when memory and better CPU is needed for an analysis



TPP Amazon Command Line Tools (amztp)

- Advanced command line toolset
- Launches parallel searches of files across multiple nodes
- Currently supports X!Tandem, OMSSA, MyriMatch, InsPect
- Manage all aspects of cloud computing including data transfer, scheduling, and instances
- Great for quickly and inexpensively processing large amounts of data



Direct Cloud support in TPP's User Interface, Petunia



TPP Web Launcher for Amazon (TWA)

1. Navigate to <http://tools.proteomecenter.org/twa>
2. Enter your Amazon Key ID and Secret
3. Click “Start Instance”
4. Welcome to Petunia
5. When you are done just click “Stop Instance”

TPP Web Launcher for Amazon v2.0.3 Beta Key ID: Secret: Tools AWS Shortcuts Start Instance

Welcome to the TPP Web Launcher for Amazon Web Services

The TPP Web-launcher for Proteomic Pipeline in the cloud is ready to use.

For more information on this software, please visit the following links:

- [TWA Documentation](#)
- [Trans-Proteomic Pipeline](#)
- [TWA Tutorial](#) - Simple
- [TPP Cloud](#) - Additional
- [Amazon Web Services](#) offered over the Internet

If you've reached this page, you are responsible for any charges that may occur, expected or otherwise, for the use of this software. It is strongly advised to use the AWS console manager to ensure that any services used are stopped and/or any storage deleted.

Amazon Web Services is not a free cloud service. Neither the developers of TPP nor the Institute of Systems Biology can be held responsible for any charges that may occur, expected or otherwise, for the use of this software. It is strongly advised to use the AWS console manager to ensure that any services used are stopped and/or any storage deleted.

ISB/SPC Trans Proteomic Pipeline - home

Home | Account | Pre-Process | mzXML UTILS | Analysis Pipeline (Sequest) | Decoy | UTILITIES | SpectraST Tools You are logged in as jslagel Log Out

Home ACCOUNT PRE-PROCESS mzXML UTILS ANALYSIS PIPELINE DECOY UTILITIES SPECTRAST TOOLS

Messages [Show / Hide]

- Welcome, jslagel.

Welcome

Welcome to the Trans-Proteomic Pipeline (TPP) web interface. These tools and interfaces were developed at the [Institute for Systems Biology \(ISB\)](#) under a grant from N-HLBI. Please visit www.proteomecenter.org and tools.proteomecenter.org for more information.

Please select analysis pipeline you want to use: Sequest

Analysis Pipeline

Follow these steps to convert, search, and analyze your data:

- 1. RAW to mzXML Conversion**
Convert original RAW files to the standard mzML input format used by the tools
- 2. Peptide Database Search and Identification**
This is a front-end to Sequest (runsearch)
- 3. Conversion to pepXML**
Convert original search results to the pepXML input format used by Xinteract
- 4. Data Curation and (optional) Peptide validation and Quantification**
Use Xinteract to filter, sort, group, and highlight data based on various criteria. You can also validate peptide identifications using PeptideProphet and/or use ASAPRatio or XPRESS to calculate the relative abundances of proteins and the corresponding confidence intervals from ICAT-type ESI-LCMS data.
- 5. Protein Assignment and Validation**
ProteinProphet provides a statistical model for validation of peptide identifications at the protein level

Please use the links on the top navigation bar to access these programs. Some of these interfaces contain inputs that only experienced users should modify.


We hope this tools suite and interface are useful. Please send feedback or post questions at our Google Groups [sptools-discuss](#) mailing list. Click [here](#) to find out how to join this list.

Resources and Links

- [SPC](#)
- [SPC Tools](#)
- [SPC Tools Wiki](#)
- [Sashimi](#)
- [SPCTools-Discuss at Google Groups](#)

Done



 Amazon Web Services is not a free cloud service. Neither the developers of TPP nor the Institute of Systems Biology can be held responsible for any charges that may occur, expected or otherwise, for the use of this software. It is strongly advised to use the AWS console manager to ensure that any services used are stopped and/or any storage deleted.

Costs of Cloud Computing

Canis lupus familiaris Data Set

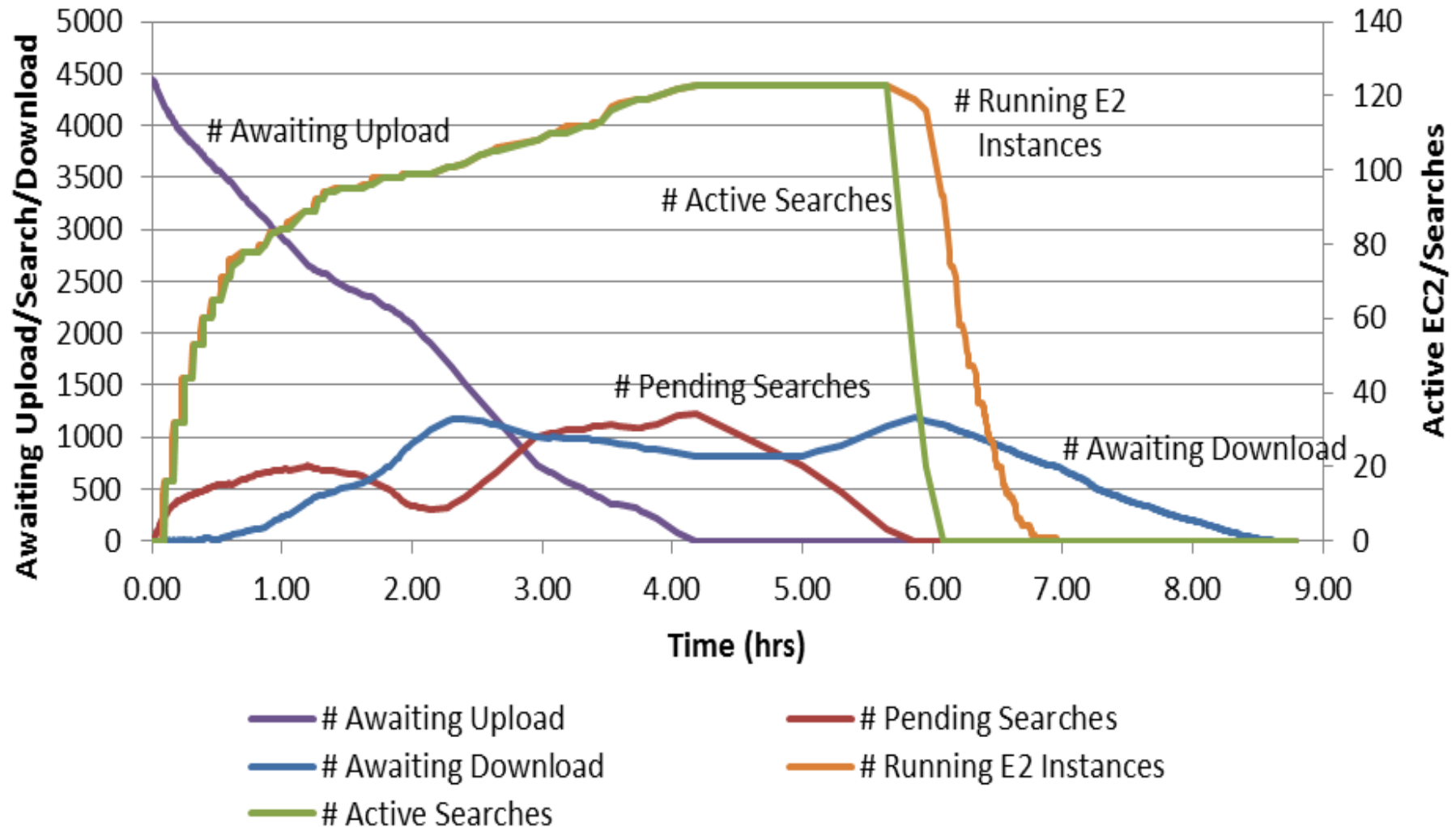
- Total 982 raw files organized in 35 folders
 - 598 raw files from LTQ Orbitrap
 - 288 raw files from LTQ
 - 96 raw files from Orbitrap
- Searched using MyriMatch and CLUSTAL
 - Total of 3,928 alignments
 - Total of 10,759 alignments
- Spot price (--ec2-spot-price) = \$0.2160
- Max # of EC2 instances (-m) = 200
- Max # of parallel upload/download processes (-P) = 10

• Total AWS cost of \$112.74
 • 82% was EC instances
 • Time to completion 5.95 hrs (+ download...)

E C 2	Operation	Spot Price	Hours	Cost
	m1.xlarge	\$ 0.216	95	\$ 20.52
	m1.xlarge	\$ 0.22	328	\$ 72.16
	Subtotal		423	\$ 92.68
S 3	Operation	Price	Usage	Cost
	PublicIP-In	\$ 0.12/GB	0.0062	\$ 0.00
		\$ 0.12/GB	0.0105	\$ 0.00
		\$ 0.12/GB	0.0211	\$ 0.00
		\$ 0.12/GB	0.0005	\$ 0.00
			\$ 0.00	\$ 92.68
S Q S	Operation	Price	Usage	Cost
	Requests	\$0.01/1,000	11,909	\$ 0.12
		\$0.01/10,000	17,433	\$ 0.02
	S3 Total			\$ 20.00
	Data Transfer In	\$ -	118.08	\$ -
	Data Transfer Out	\$ 0.12/GB	165.56	\$ 19.87
	SQS Total			\$ 0.06



Amazon Web Services Cost Management



Learn More About AWS & TPP

Technological Innovation and Resources

© 2015 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at <http://www.mcponline.org>

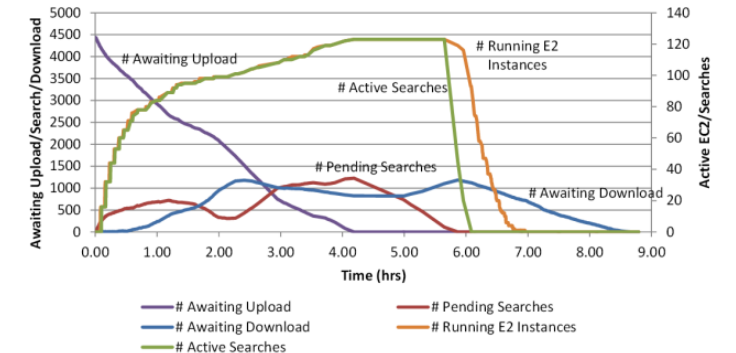
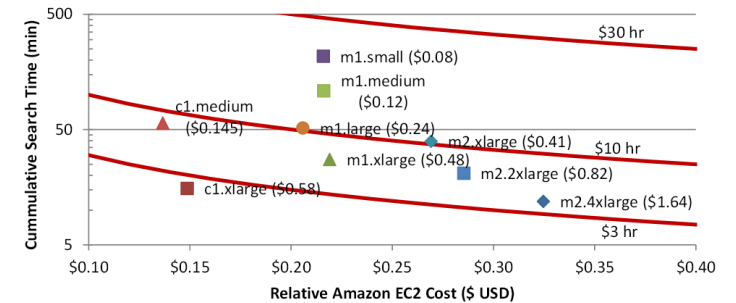
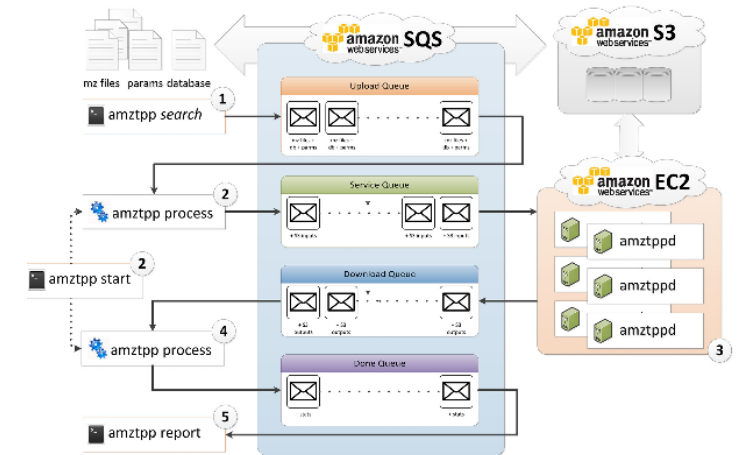
Processing Shotgun Proteomics Data on the Amazon Cloud with the Trans-Proteomic Pipeline*

Joseph Slagel†, Luis Mendoza‡, David Shteynberg‡, Eric W. Deutsch†§, and Robert L. Moritz‡

Cloud computing, where scalable, on-demand compute cycles and storage are available as a service, has the potential to accelerate mass spectrometry-based proteomics research by providing simple, expandable, and affordable large-scale computing to all laboratories regardless of location or information technology expertise. We present new cloud computing functionality for the Trans-Proteomic Pipeline, a free and open-source suite of tools for the processing and analysis of tandem mass spectrometry datasets. Enabled with Amazon Web Services cloud computing, the Trans-Proteomic Pipeline now accesses large scale computing resources, limited only by the available Amazon Web Services infrastructure, for all users. The Trans-Proteomic Pipeline runs in an environment fully hosted on Amazon Web Services, where all software and data reside on cloud resources to tackle large search studies. In addition, it can also be run on a local computer with computationally intensive tasks launched onto the Amazon Elastic Compute Cloud service to greatly decrease analysis times. We describe the new Trans-Proteomic Pipeline cloud service components, compare the relative performance and costs of various Elastic Compute Cloud service instance types, and present on-line tutorials that enable users to learn how to deploy cloud computing technology rapidly with the Trans-Proteomic Pipeline. We provide tools for estimating the necessary computing resources and costs given the scale of a job and demonstrate the use of cloud enabled Trans-Proteomic Pipeline by performing over 1100 tandem mass spectrometry files through four proteomic

an important proteomics technique that has enabled researchers to identify and quantify proteins in complex biological samples in a high throughput manner. Mass spectrometers continue their incremental increases in sensitivity, mass accuracy, and speed of data collection, thereby generating comprehensive highly accurate data on smaller and smaller sample sizes. Software tools have likewise become more sophisticated and have enabled improved interpretation of the mass spectra that are generated all at the cost of greater computational resources (1). Simply applying cutoffs to native search scores has been replaced with algorithms that model the output scores and other attributes of the peptide-spectrum matches (PSMs) to yield improved probabilistic metrics for peptide and protein identifications (2–4).

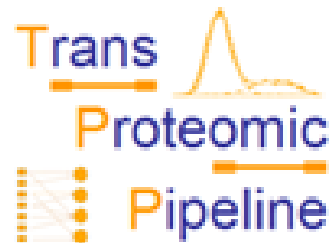
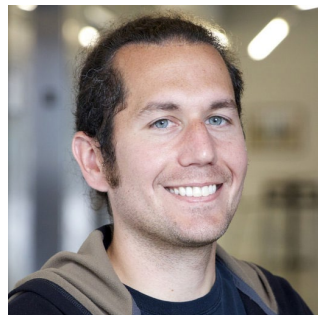
The typical bioinformatics workflow for analyzing such shotgun data (5) relies on an algorithm that matches the set of spectra generated by the instrument against a set of candidate matches. These candidates can be either theoretical spectra generated from a set of plausible candidate peptides selected from a set of protein sequences, termed sequence searching, or a set of previously identified mass spectra, termed spectral library searching. There are a large number of both commercial and open-source sequence search engines available for use (see (5) for references to many of these). They perform comparably over a wide range of data sets, although the output scores and formats vary significantly, thereby making comparison and integration of results challenging. How-



Cloud Computing Workshop (2022)

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.

12-16 September 2022



Michael Hoopmann - *ISB, Seattle, WA*

- Instructions on how to use TPP to analyze MS data.
- Answer questions from the participants

3-7 October 2022



Melanie Foell - *Freiburg University, Germany*

- Instructions on how to use Galaxy to analyze MS data.
- Answer questions from the participants

14-18 November 2022



Yasset Perez-Riverol - *EBI, Hinxton, UK*

- Instructions on how to use Nextflow to analyze MS data.
- Answer questions from the participants

REGISTRATION LINK:

<https://abrf.memberclicks.net/cloudcomputingworkshop>



Proteomic Data analysis in Galaxy



Galaxy PROJECT

<https://galaxyproject.org>

Galaxy for
SCIENTISTS



Galaxy for
TRAINERS



Galaxy for
TOOL AUTHORS



Galaxy for
DEVELOPERS



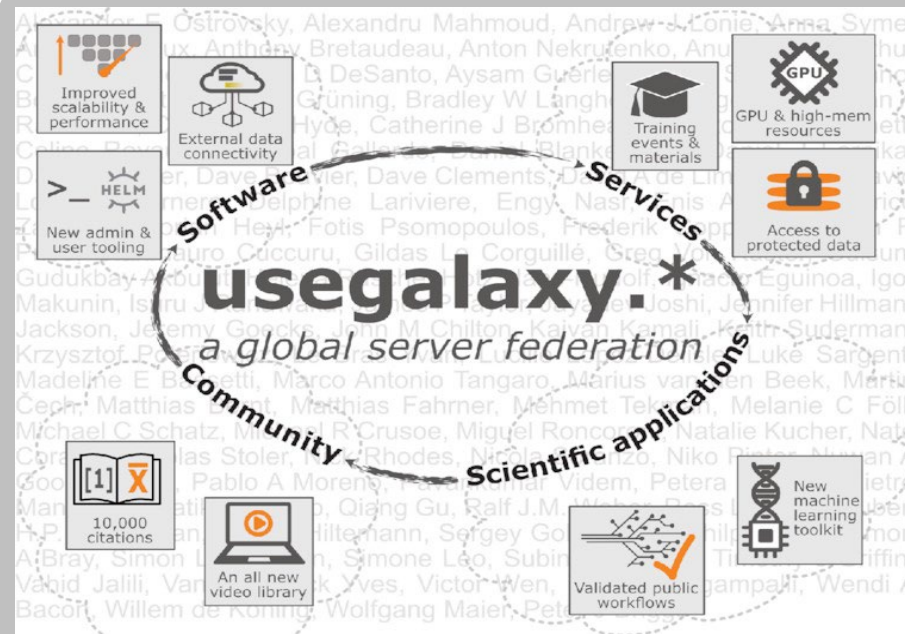
Galaxy for
ADMINS



Data Intensive *analysis* for everyone

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



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



Analysis history

The screenshot displays the Galaxy Europe web interface. The top navigation bar includes 'Galaxy Europe' and menu items: 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and a grid icon. A left sidebar lists various tool categories: 'Tools' (with a search bar), '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), and 'GENOMICS ANALYSIS' (Annotation, Multiple Alignments, Assembly, Mapping, Variant Calling, Genome editing).

The main workspace shows the 'MaxQuant' tool configuration (Galaxy Version 1.6.10.43+galaxy3). The 'Input Options' section includes a dropdown for 'choose the type of your input files' set to 'thermo.raw'. The 'FASTA files' section shows a file browser with '4: Protein_database' selected. Below this, there are fields for 'identifier parse rule' (>.*\((.*)\|), 'description parse rule' (>.* OS), and a note to 'Specify one or more FASTA databases.' The 'Search Options' section includes a dropdown for 'Specify an experimental design template (if needed)' set to 'Nothing selected'. It also has input fields for 'minimum peptide length' (7) and 'maximum peptide mass [Da]' (4600), with explanatory text for each. The 'minimum unique peptides' field is set to 1.

On the right, a 'History' panel titled 'MaxQuant Serum samples' shows a list of 64 datasets. The top entry is '64: Filter on data 57'. Below it, a green-shaded section contains entries 63 through 52, including '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', and '52: PTXQC report for data 6 and data 4'. Each entry has icons for viewing, editing, and deleting.



Accessibility

Graphical user interface
Pre-installed tools



Reproducibility

All information is captured in histories
Controlled tool versions

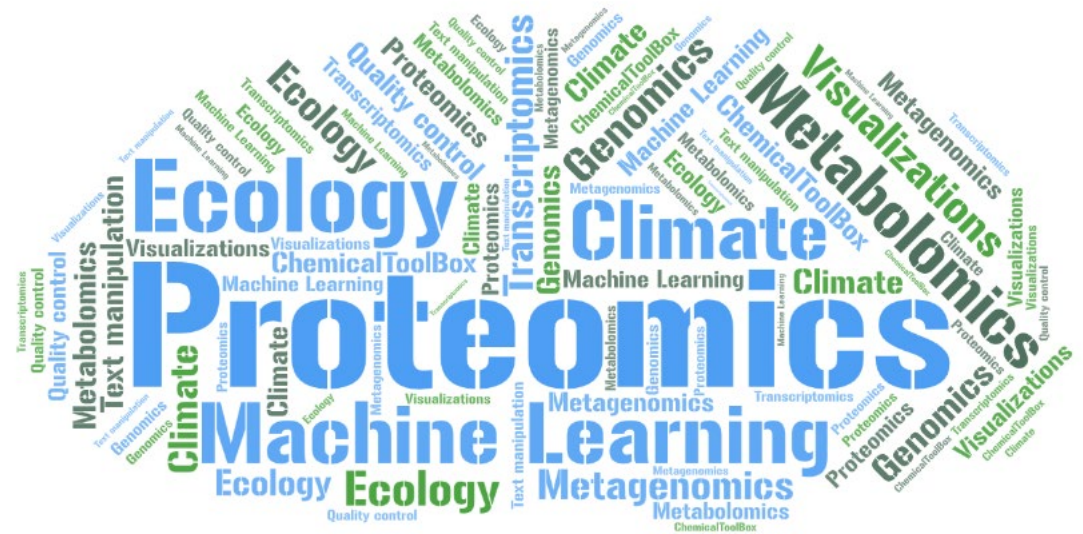


Transparency

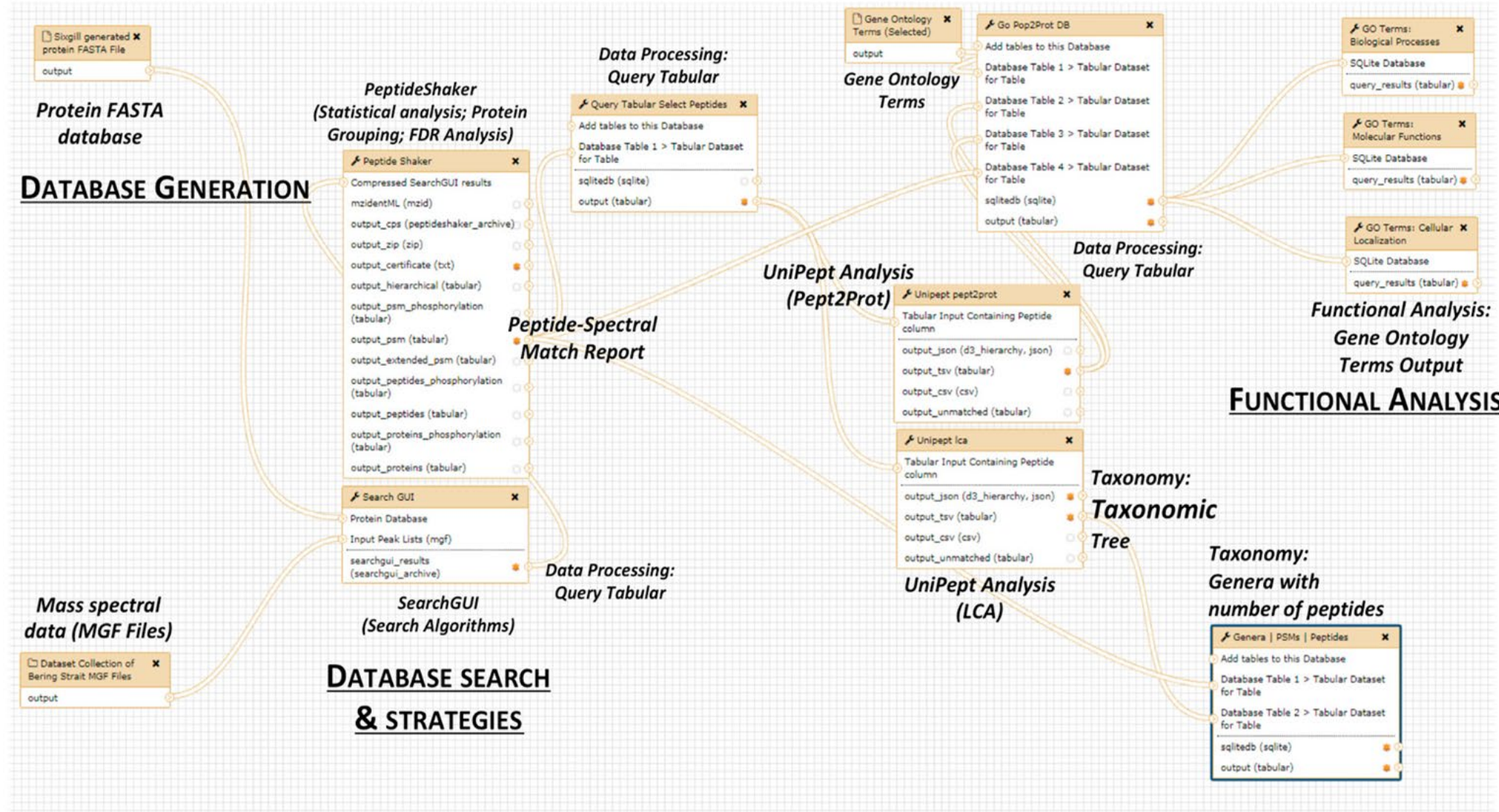
Sharing of histories and workflows



more than 7,000 tools



SOLUTION: GALAXY BIOINFORMATICS PLATFORM



FUNCTIONAL ANALYSIS

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

GTN: HANDS-ON TRAINING MATERIAL WITH INSTRUCTIONAL VIDEOS

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

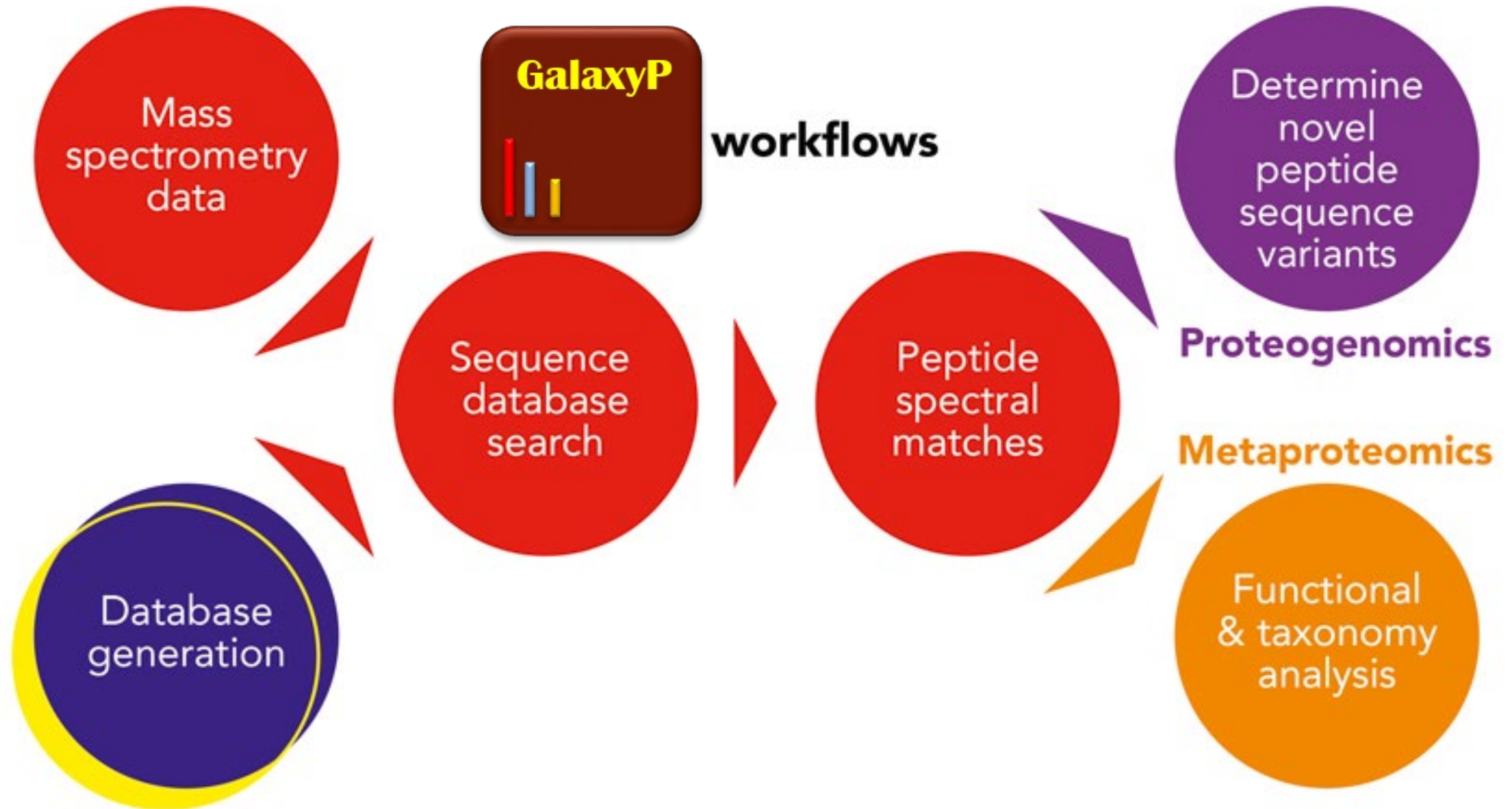
Topic	Tutorials
Contributing to the Galaxy Training Material	10
Teaching and Hosting Galaxy training	5



> 130 training materials

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

- proteomics tutorials in Galaxy:
<https://training.galaxyproject.org/training-material/topics/proteomics>
- Global online Galaxy course in March (much more than proteomics)
<https://gallantries.github.io/posts/2021/12/14/smorgasbord2-tapas/>



PUBLICATIONS: z.umn.edu/galaxypreferences

ACCESSING MULTIOMIC GALAXY 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/>



3-7 October 2022

MaxQuant and MSstats in Galaxy Enable Reproducible Cloud-Based Analysis of Quantitative Proteomics Experiments for Everyone

Niko Pinter, Damian Glätzer, Matthias Fahrner, Klemens Fröhlich, James Johnson, Björn Andreas Grüning, Bettina Warscheid, Friedel Drepper, Oliver Schilling, and Melanie Christine Föll*

✔ **Cite this:** *J. Proteome Res.* 2022, 21, 6, 1558–1565

Publication Date: May 3, 2022 ▾

<https://doi.org/10.1021/acs.jproteome.2c00051>

Copyright © 2022 American Chemical Society

Article Views

288

Altmetric

18

Citations

-

[LEARN ABOUT THESE METRICS](#)

Share



Add to



Export



Melanie Foell, Freiburg University, Freiburg (Germany)

Hands-on

Tutorial: MaxQuant and Msstats for the analysis of label-free data

<https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/maxquant-label-free/tutorial.html>

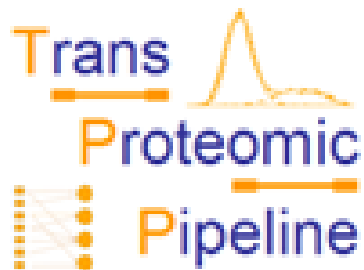
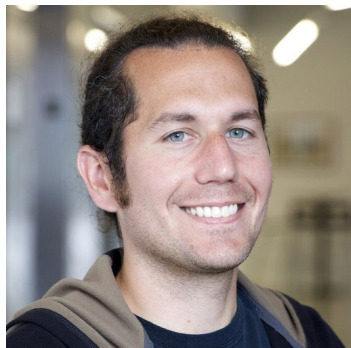
Video with demonstration of tutorial in youtube: <https://www.youtube.com/watch?v=IXdLAt2PAT4>



Cloud Computing Workshop (2022)

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.

12-16 September 2022



Michael Hoopmann - *ISB, Seattle, WA*

- Instructions on how to use TPP to analyze MS data.
- Answer questions from the participants

3-7 October 2022



Melanie Foell - *Freiburg University, Germany*

- Instructions on how to use Galaxy to analyze MS data.
- Answer questions from the participants

14-18 November 2022



Yasset Perez-Riverol - *EBI, Hinxton, UK*

- Instructions on how to use Nextflow to analyze MS data.
- Answer questions from the participants

REGISTRATION LINK: <https://abrf.memberclicks.net/cloudcomputingworkshop>

Proteomics Workflows in NextFlow with Focus on Benchmarking

— Veit Schwämmle —

University of Southern Denmark

Motivation

Data and software heterogeneity in proteomics

Many different methods and algorithms to analyze proteomics data

Different workflows exist but mostly not portable and not very scalable

Composition of new workflows built on information from already implemented ones

Different data analysis approaches can lead to vast differences

Parameter settings



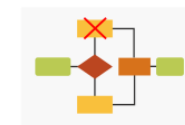
Algorithms



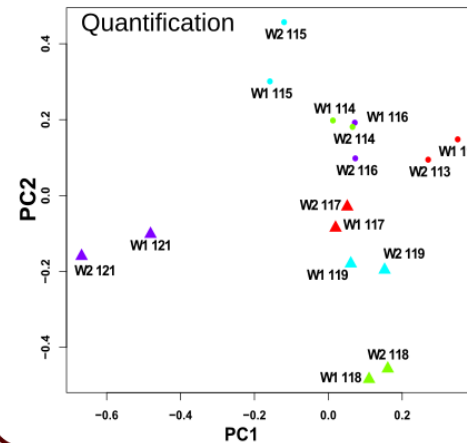
Operating environment



Omitted components



Different workflows perform like different replicates



Limited overlap in identification

Data: iTRAQ quantification of samples for muscle-invasive vs non-invasive bladder cancer

Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis
Latosinska *et al*, PLOS One 2015

Workflow 2
2515 proteins
TPP + Libra for protein
quantification



Workflow 1
2544 proteins
SearchGUI/PeptideShaker +
isobar (R package)

M Palmblad, A-L Lamprecht, J Ison & VS, Bioinformatics, (2019)

Now what?

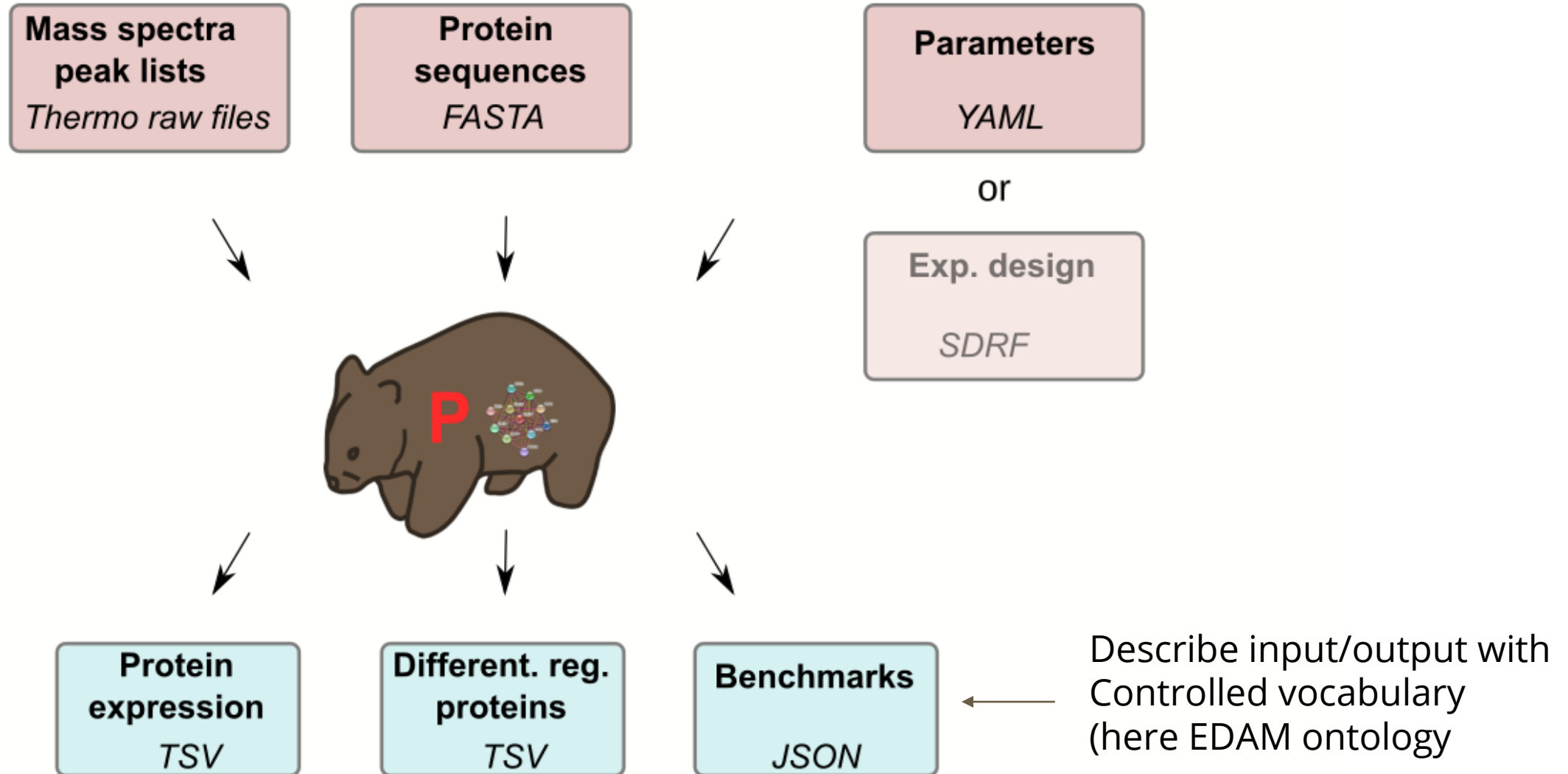


Everything: proteomics x

Search bio.tools

1851 tools

Full analysis of data from label-free bottom-up MS



Implementations in

nextflow

5 different workflows with dockerized software

- a) Compomics / FlashLFQ / MSqRob
- b) MaxQuant / Normalyzer
- c) OpenMS / ProteomicsLFQ
- d) Trans-Proteomic Pipeline / ROTS
- e) SearchGUI / Proline / PolySTest



WOMBAT: Workflow Metrics, Benchmarking and AnalyTics in Proteomics

github.com/wombat-p

Fundamentals of Nextflow Pipelines

nextflow script

Write code
in any language.



Define orchestration with
dataflow programming.

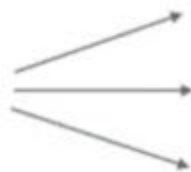
Define software
dependencies
with containers.



Version
control.

nextflow runtime

Orchestration of tasks to
deploy anywhere with ease.



```
/*  
 * STEP 1.2 - Convert mzDB files to MGF  
 */  
process convert_mzdb_to_mgf {  
  label 'process_low'  
  label 'process_single_thread'  
  
  publishDir "${params.outdir}/MGFs", mode:'copy'  
  
  input:  
  file mzdbfile from mzdb_convert_to_mgf  
  
  output:  
  file "${mzdbfile.baseName}.mgf" into mgfs_searchgui  
  
  script:  
  """  
  mzdb2mgf "${mzdbfile}"  
  mv "${mzdbfile}.mgf" "${mzdbfile.baseName}.mgf"  
  """  
}
```



Compute platforms

Automatically provision, manage and scale compute environments in the cloud, or tap existing on-premises or cloud HPC and Kubernetes clusters for maximum flexibility.



AWS BATCH



Azure Batch



Google Life Sciences



Altair | PBS Works™



IBM Spectrum LSF



Amazon EKS



Google GKE



kubernetes

Benchmarking

Set of performance and data quality metrics

Category	Aspect	Subgroup	Name	Definition	Value
Functionality	Traceability	Spectra	Tracable spectra	Results tracable to original spectra	Y/N
Functionality	Traceability	Spectra	Universal spectrum identifiers	Workflow generates USIs (Universal Spectrum Identifier)	Y/N
Functionality	Traceability	Spectra	Peptide to spectra	Corresponding spectrum numbers/ids available from peptide level	Y/N
Functionality	Traceability	Spectra	Protein to spectra	Corresponding spectrum numbers/ids available from protein level	Y/N
Functionality	Traceability	File names	Results to raw files	Raw input file names preserved in tables on PSM/peptide/protein level	Y/N
Functionality	Traceability	File names	Public raw files	Raw files publicly available	Y/N
Functionality	Traceability	Parameters	Settings	Are all processing settings available with result files	Y/N
Functionality	Traceability	Parameters	Experimental design	Biological and technical replicates can be identified in results	Y/N
Functionality	Reproducibility	Files	Identity	Can exact result be reproduced	Y/N
Functionality	Performance	Identification	PSM number	Number of identified PSMs passing preset FDR	Integer
Functionality	Performance	Identification	Peptide number	Number of uniquely identified peptide identifications passing preset FDR	Integer
Functionality	Performance	Identification	Protein number	Number of uniquely identified protein identifications passing preset FDR	Integer
Functionality	Performance	Identification	Protein group number	Number of different protein groups passing preset FDR	Integer
Functionality	Performance	Identification	Peptide coverage	Percentage of peptides identified in all samples	Double
Functionality	Performance	Identification	Protein coverage	Percentage of proteins identified in all samples	Double
Functionality	Performance	Identification	Peptides per protein	Distribution of peptides per protein group	Set of Integer

```

    ▼ Quantification:
      ▼ CVPeptides:
        0: 0.0793
      ▼ CVProteins:
        0: 0.0566
      ▼ CorrelationPeptides:
        0: 0.9773
      ▼ CorrelationProteins:
        0: 0.9989
      ▼ NumberOfPeptides:
        0: 6458
      ▼ DynamicPeptideRange:
        0: 43.36
      ▼ NumberOfProteinGroups:
        0: 984
      ▼ DynamicProteinRange:
        0: 190.2964
      ▼ Statistics:
        ▼ DifferentialRegulatedPeptides5Perc:
          0: 37.6944
          1: 86.6389
          2: 1.6389
          3: 54.6111
          4: 48.3611
          5: 0.5278
          6: 43.9444
          7: 33.6944
          8: 82.6389
          9: 28.6944
  
```

```

    ▼ Performance:
      ▼ Identification:
        ▼ PSMNumber:
          0: null
        ▼ PeptideNumber:
          0: 6959
        ▼ ProteinNumber:
          0: 955
        ▼ ProteinGroupNumber:
          0: 998
        ▼ PeptideCoverage:
          0: 5014
        ▼ ProteinCoverage:
          0: 944
        ▼ PeptidesPerProtein:
          ▼ 0:
            Var1: "1"
            Freq: 5646
          ▼ 1:
            Var1: "2"
            Freq: 3541
          ▼ 2:
            Var1: "3"
            Freq: 2383
          ▶ 3: {...}
          ▶ 4: {...}
          ▶ 5: {...}
          ▶ 6: {...}
          ▶ 7: {...}
          ▶ 8: {...}
          ▶ 9: {...}
        ▼ Quantification:
          ▼ CVPeptides:
            0: 0.0793
          ▼ CVProteins:
            0: 0.0566
  
```

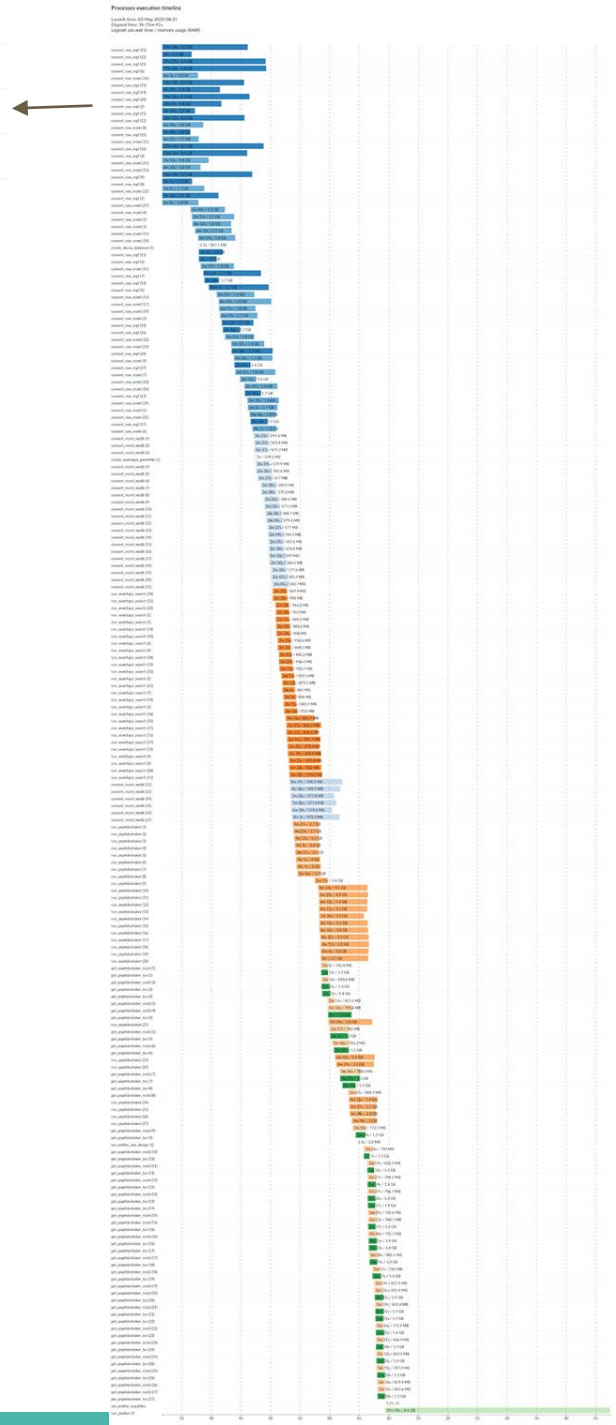
Harmonized workflow output on peptide and protein level

Statistical testing as workflow component adds valuable information

Execution

convert_raw_mgf (17)
convert_raw_mzml (6)
convert_mzml_mzdb (1)
convert_mzml_mzdb (2)

convert_raw_mgf (17)	4m 1s / 1.8 GB
convert_raw_mzml (6)	2m 33s / 591.6 MB
convert_mzml_mzdb (1)	2m 27s / 593.8 MB
convert_mzml_mzdb (2)	



Nextflow allows extensive scaling

Further tasks for better performance and usage:

- Simple web interface for execution on the cloud. Alternative (commercial) solution: nf.tower
- Create missing bioconda packages
- Automatic runs using new PRIDE metadata standard (SDRF format)
- Generalized input parameter set for analysis
- Done: automatic calculation of generalized benchmarks on standardized output

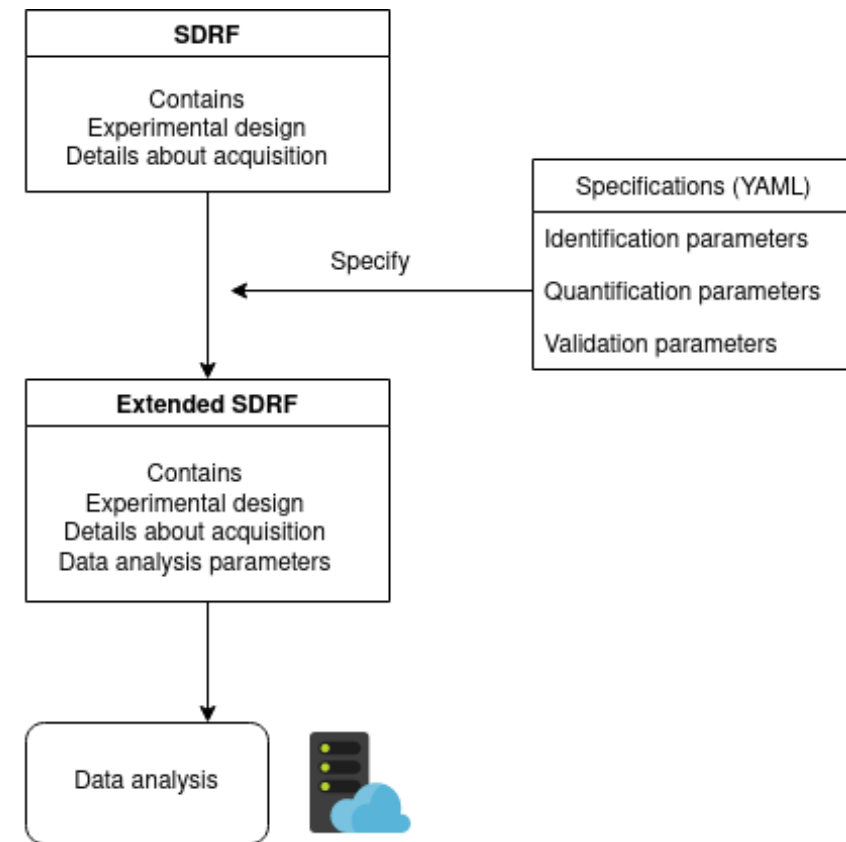
SDRF format as starting point

Standardize SDRF for data analysis: Extend SDRF to include parameter settings for analysis

-> Unified format to describe experimental design and “optimal” parameters

-> Re-run by specifying extended SDRF only

-> Need for more annotations of experimental design



Other NextFlow implementations

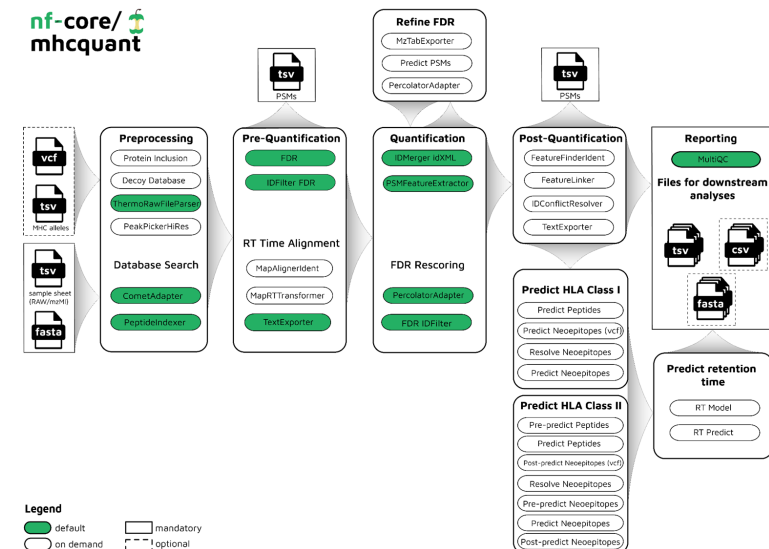
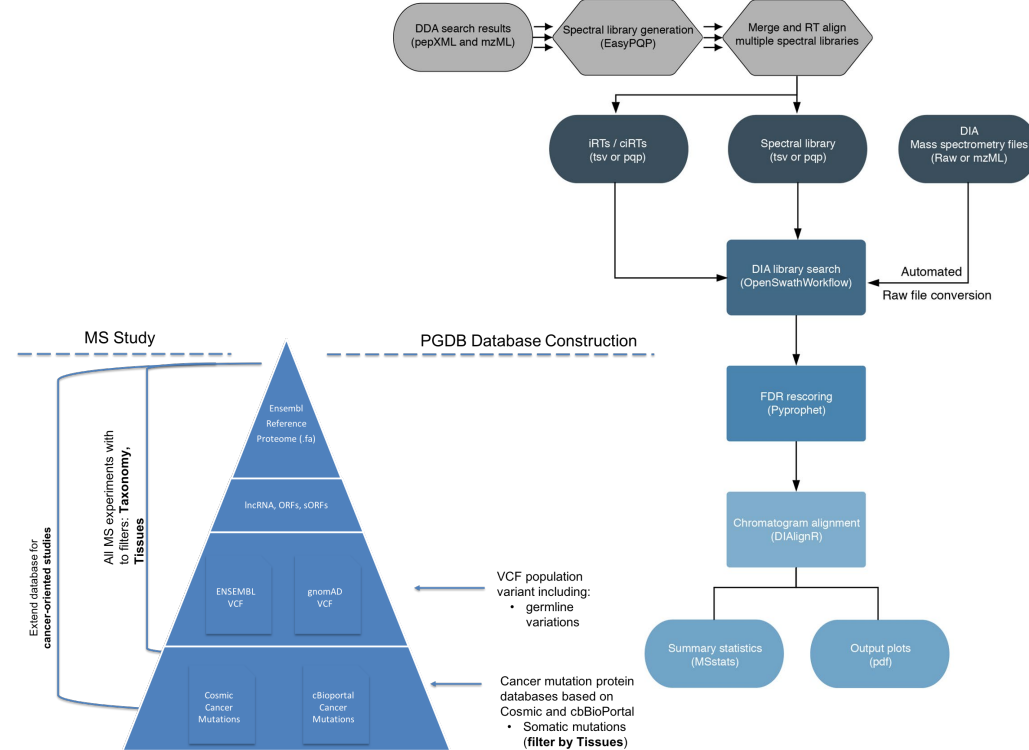
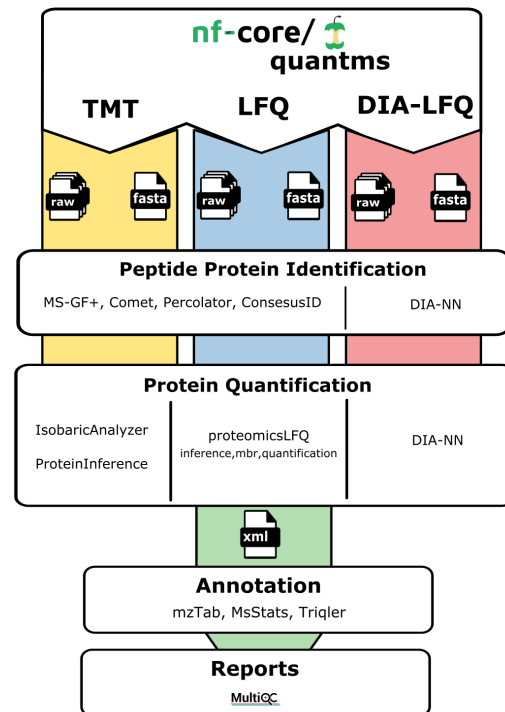
nf-core to make workflows "production-ready"

nf-co.re/pgdb

nf-co.re/diaproteomics

nf-core/quantms

nf-core/mhcquant



Acknowledgements

Implementation Study:

Comparison, benchmarking and dissemination of proteomics data analysis pipelines



Node	Name of PI	Role (lead or member)
Denmark	Veit Schwämmle, Jon Ison	Lead
EMBL-EBI	Juan Antonio Vizcaíno	Member
Netherlands	Magnus Palmblad, Anna-Lena Lamprecht, Peter Horvatovich	Member
Spain	Salvador Capella-Gutierrez, Josep Ll. Gelpi	Member
Spain	Eduard Sabidó (Fernando Corrales, ProteoRed)	Member
France	Yves Vandenbrouck, David Bouyssié, Wolfgang Raffelsberger	Member
Sweden	Fredrik Levander (Lund University)	Member
Sweden	Ola Spjuth (Uppsala University)	Member
Italy	Gianluigi Zanetti (CRS4)	Member
Czech Republic	Martin Hubalek	Member
Germany	Martin Eisenacher, Julian Uszkoreit	Member
Germany	Oliver Kohlbacher, Timo Sachsenberg	Member
EMBL-EBI	Steven Newhouse	Member

Wolfgang Raffelsberger, University of Strasbourg

EuBIC-MS Developers Meeting 2023

SAVE THE DATE 15-20 January 2023, ETH Congressi Stefano Franscini, Monte Verità, Switzerland

SUBMIT YOUR HACKATHON PROPOSAL Deadline 30 September 2022

Confirmed keynote speakers



Karsten Borgwardt



Alexey Nesvizhskii



Maximilian Strauss

Keynotes by experts in the field of
bioinformatics and proteomics

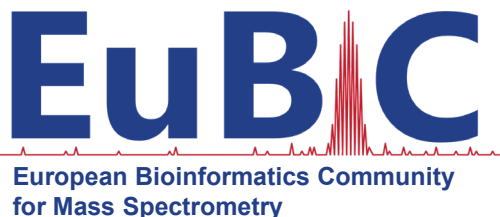
Hackathons from selected abstracts

Meet and team up with developers

Poster session

... and much more!

MORE INFO www.eubic-ms.org @EuBIC_ms



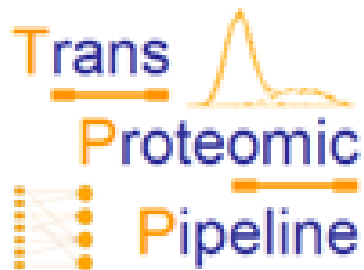
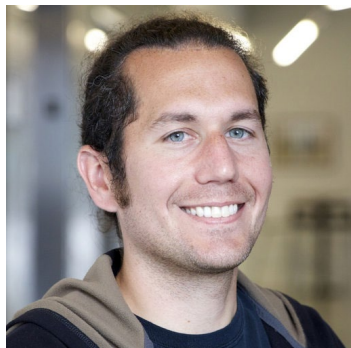
ETH zürich



Cloud Computing Workshop (2022)

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.

12-16 September 2022



Michael Hoopmann - *ISB, Seattle, WA*

- Instructions on how to use TPP to analyze MS data.
- Answer questions from the participants

3-7 October 2022



Melanie Foell - *Freiburg University, Germany*

- Instructions on how to use Galaxy to analyze MS data.
- Answer questions from the participants

14-18 November 2022



Yasset Perez-Riverol - *EBI, Hinxton, UK*

- Instructions on how to use Nextflow to analyze MS data.
- Answer questions from the participants

REGISTRATION LINK: <https://abrf.memberclicks.net/cloudcomputingworkshop>