Advanced MS analysis: proteogenomics

Tim Griffin

University of Minnesota

International Mass Spectrometry Conference

28 August, 2022

tgriffin@umn.edu

Learn more at galaxyp.org

z.umn.edu/itcrgalaxyvideo

Acknowledgements

Biochemistry,
Molecular Biology & Biophysics

Dr. Pratik Jagtap (Co-leader, Galaxy-P)
Praveen Kumar
Subina Mehta
Caleb Easterly
Ray Sajulga
Andrew Rajczewski
Dr. Shane Hubler
Mark Esler
Dr. Art Eschenlauer
Dr. Candace Guerrero
Matt Chambers

GalaxyP

Collaborators
David Largaespada
Frank Ondrey
Mo Heydarian/Karen Reddy
Brian Crooker/Wanda Weber
Bart Mesuere
Brook Nunn
Thilo Muth
Magnus Øverlie Arntzen

Minnesota
Supercomputing
Institute

James Johnson
Tom McGowan
Dr. Getiria Onsongo
Dr. Michael Milligan

COMMUNITY-BASED SOFTWARE DEVELOPMENT

Harald Barsnes and Marc Vaudel
University of Bergen, Bergen, Norway
Bjoern Gruening (Galaxy community…)
University of Freiburg, Freiburg, Germany
Lennart Martens
VIB Department of Medical Protein Research, UGent, Belgium
Lloyd Smith/Michael Shortreed
University of Wisconsin-Madison

ITCR groups
Rachel Karchin/Michael Ryan
Johns Hopkins University/In-Silico Solutions

Tom Doake/Jeremy Fischer
Indiana University

Jetstream

NIH award U24CA199347

NIH NATIONAL CANCER INSTITUTE
Informatics Technology for
Cancer Research

GalaxyP
galaxyp.org

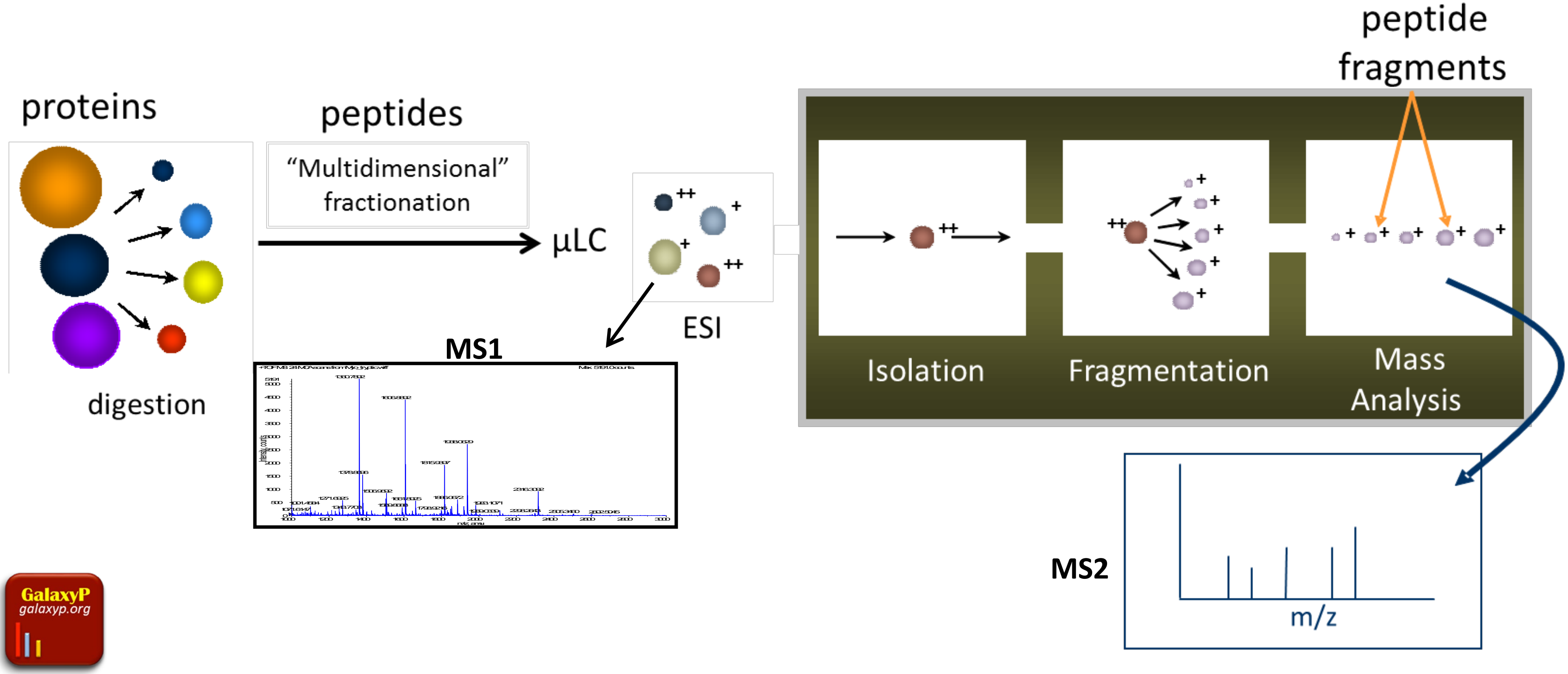
NSF

Outline: Proteogenomics and bioinformatics

- Background and informatics challenges
- Overview of components involved in proteogenomic workflows
- Hands-on workshop and tutorial

Proteogenomics: A primer

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



proteins

digestion

peptides

“Multidimensional” fractionation

μLC

ESI

MS1

Isolation

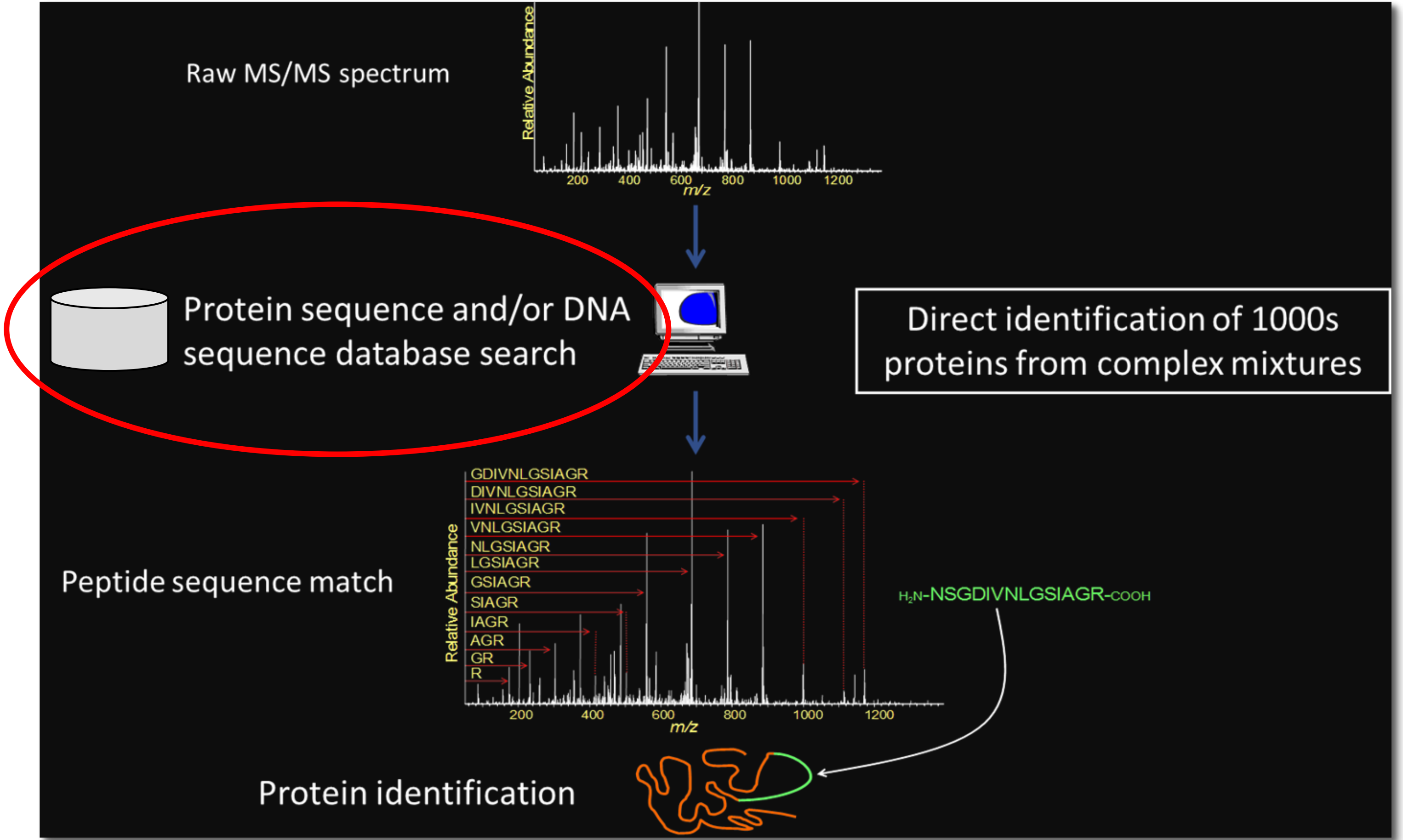
Fragmentation

Mass Analysis

peptide fragments

MS2

Matching amino acid sequences to MS/MS data



Raw MS/MS spectrum

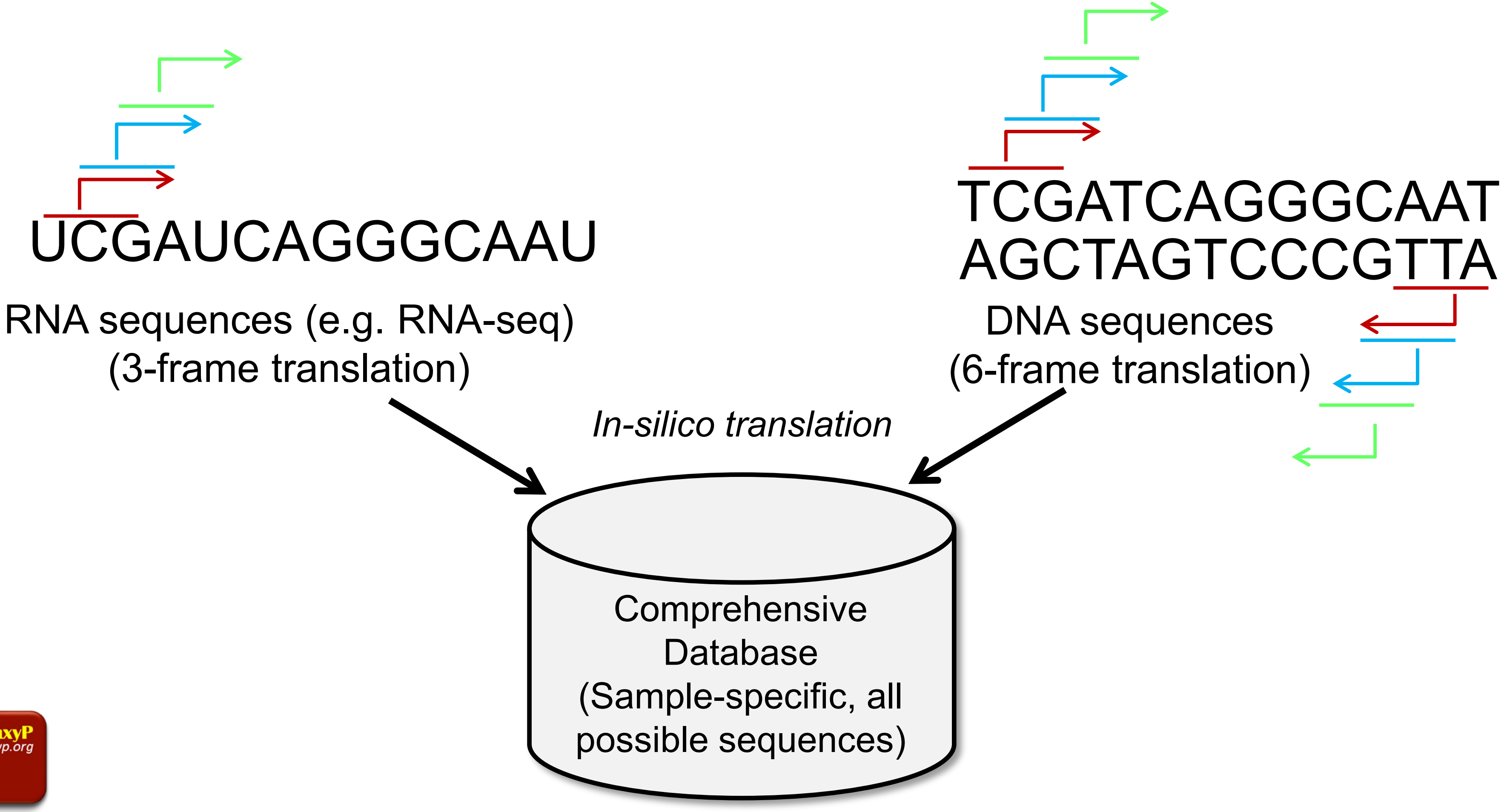
Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification

Detecting protein variants via proteogenomics



UCGAUCAGGGCAAU

RNA sequences (e.g. RNA-seq)
(3-frame translation)

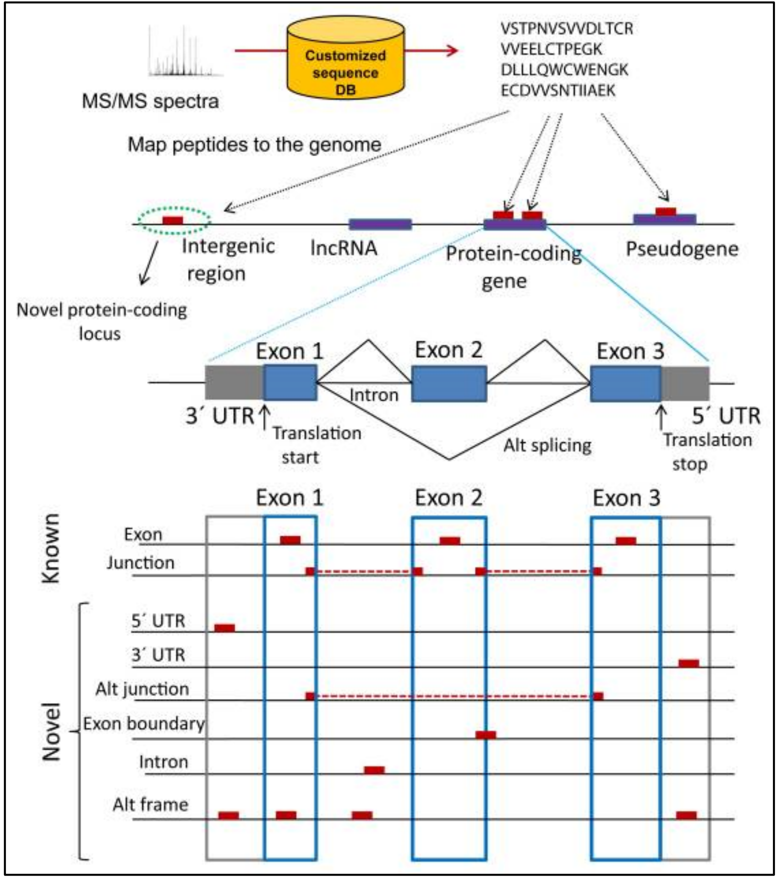
TCGATCAGGGCAAT
AGCTAGTCCCGTTA

DNA sequences
(6-frame translation)

In-silico translation

Comprehensive
Database
(Sample-specific, all
possible sequences)

Proteogenomic outcomes and applications

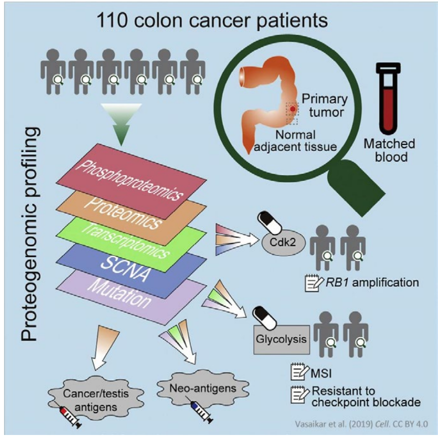


VOL.11 |NO.11 | NOVEMBER 2014 | NATURE METHODS

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

CPTAC

National Cancer Institute
at the National Institutes of Health

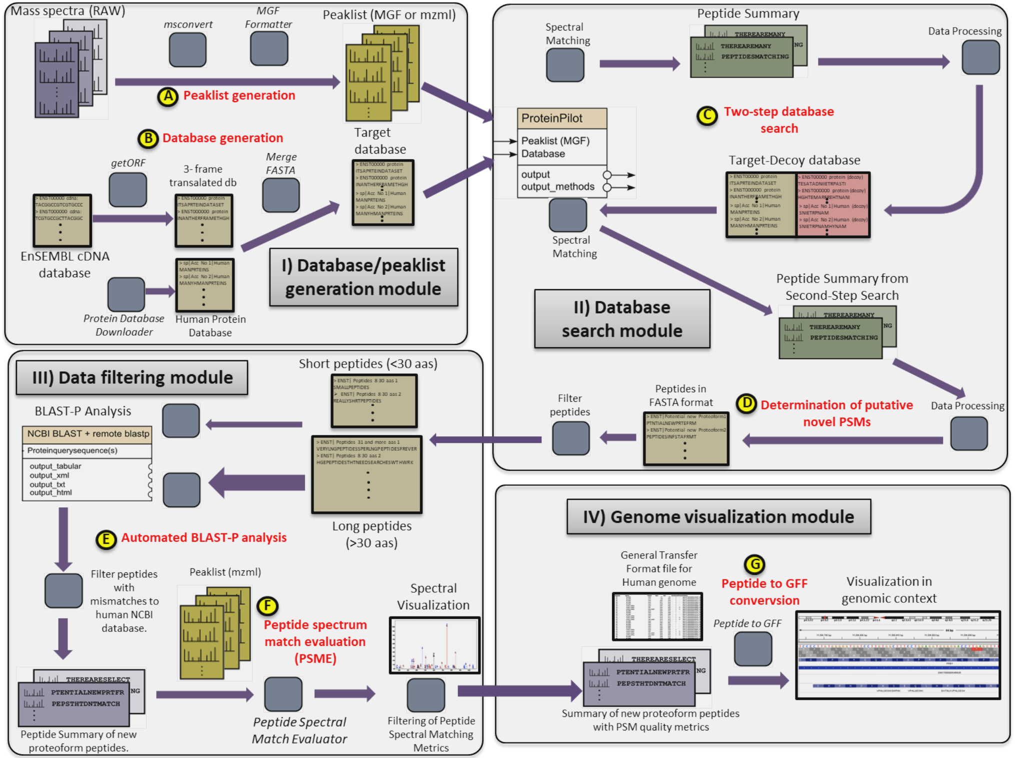


Vasaikar et al. (2019) Cell. CC BY 4.0

GalaxyP
galaxyp.org

Bringing proteogenomics to the masses: informatics challenges

- Many software tools, integration, automation….



Mass spectra (RAW)
msconvert
MGF Formatter
Peaklist (MGF or mzml)
A Peaklist generation
B Database generation
getORF
3-frame translated db
Merge FASTA
Target database
EnSEMBL cDNA database
Protein Database Downloader
Human Protein Database
I) Database/peaklist generation module
Spectral Matching
Peptide Summary
Data Processing
ProteinPilot
Peaklist (MGF)
Database
output
output_methods
C Two-step database search
Target-Decoy database
Spectral Matching
Peptide Summary from Second-Step Search
II) Database search module
Filter peptides
Peptides in FASTA format
D Determination of putative novel PSMs
Data Processing
III) Data filtering module
Short peptides (<30 aas)
BLAST-P Analysis
NCBI BLAST + remote blastp
Proteinquerysequence(s)
output_tabular
output_xml
output_txt
output_html
Long peptides (>30 aas)
E Automated BLAST-P analysis
Filter peptides with mismatches to human NCBI database.
Peaklist (mzml)
F Peptide spectrum match evaluation (PSME)
Spectral Visualization
Peptide Summary of new proteoform peptides.
Peptide Spectral Match Evaluator
Filtering of Peptide Spectral Matching Metrics
IV) Genome visualization module
General Transfer Format file for Human genome
G Peptide to GFF convervsion
Peptide to GFF
Visualization in genomic context
Summary of new proteoform peptides with PSM quality metrics

J. Proteome Res., 2014, 13, pp 5898–5908

GalaxyP
galaxyp.org

Proteogenomic informatics challenges

- *Assembly and variant calling from DNA/RNA sequencing data*
- *Customized protein sequence database generation*
- *Matching sequences to MS/MS data: best practices?*
- *Filtering, QC and verification of putative variant sequences*
- *Interpretation! Beyond a list....*
- *Access and usability by the research community*

A solution: The Galaxy Ecosystem

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)

GalaxyP
galaxyp.org

Working philosophy:

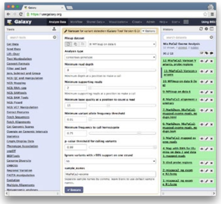
GalaxyP
galaxyp.org

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

Galaxy Training Network

Interfaces

Web UI



Programmatic API

```
>_
```

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

Datasets

Galaxy

DB

API

Computing Resources

slurm workload manager

Condor High Throughput Computing

Google Cloud Platform

amazon web services

Analysis Tools and Visualizations

MS/MS sequence matching

Protein sequence database generation

ID of novel sequence variants

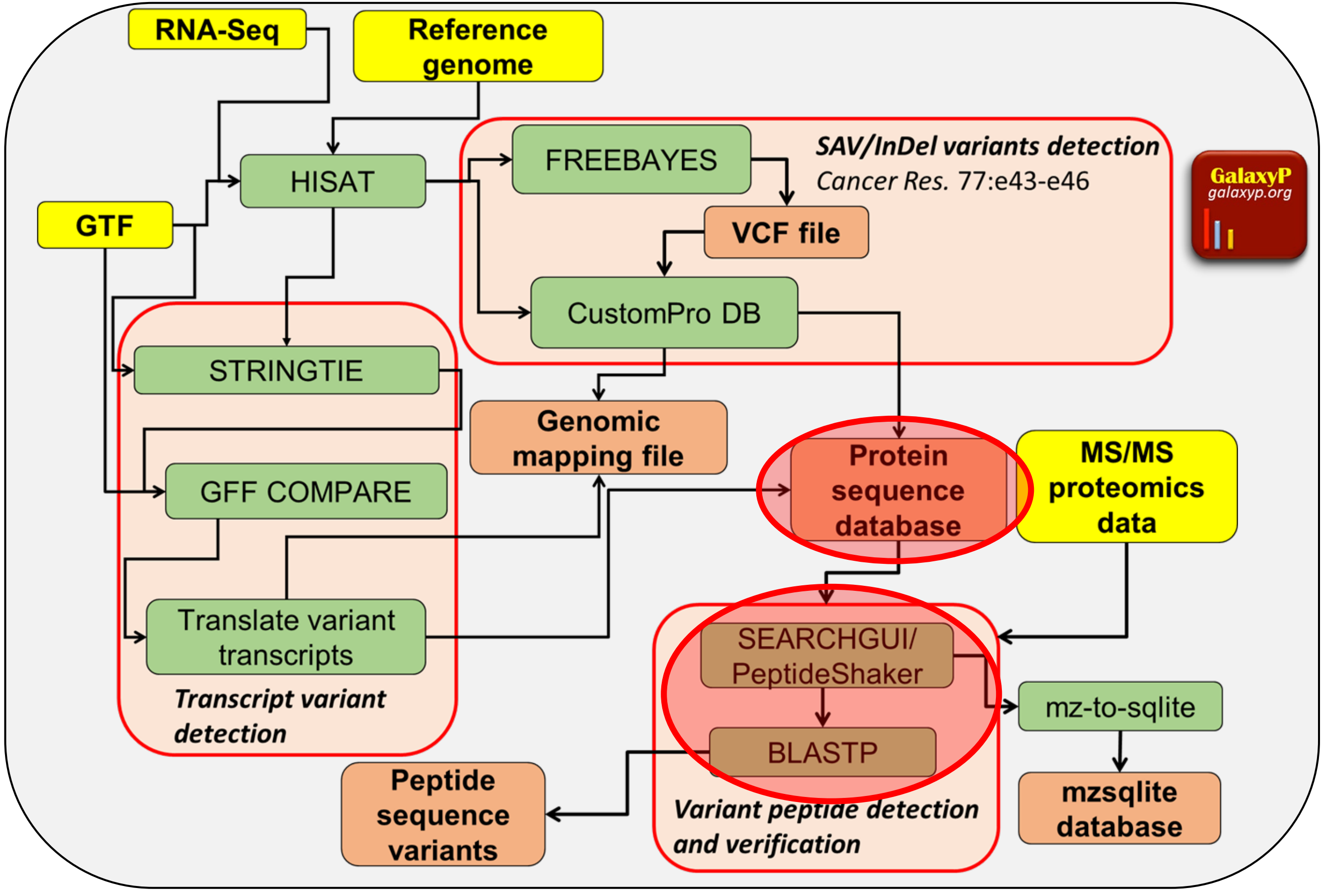
GalaxyP galaxyp.org

Courtesy Jeremy Goecks, OHSU

Integrative data processing: RNA-Seq + proteomics

1. *Generate protein sequence database from RNA-Seq data*

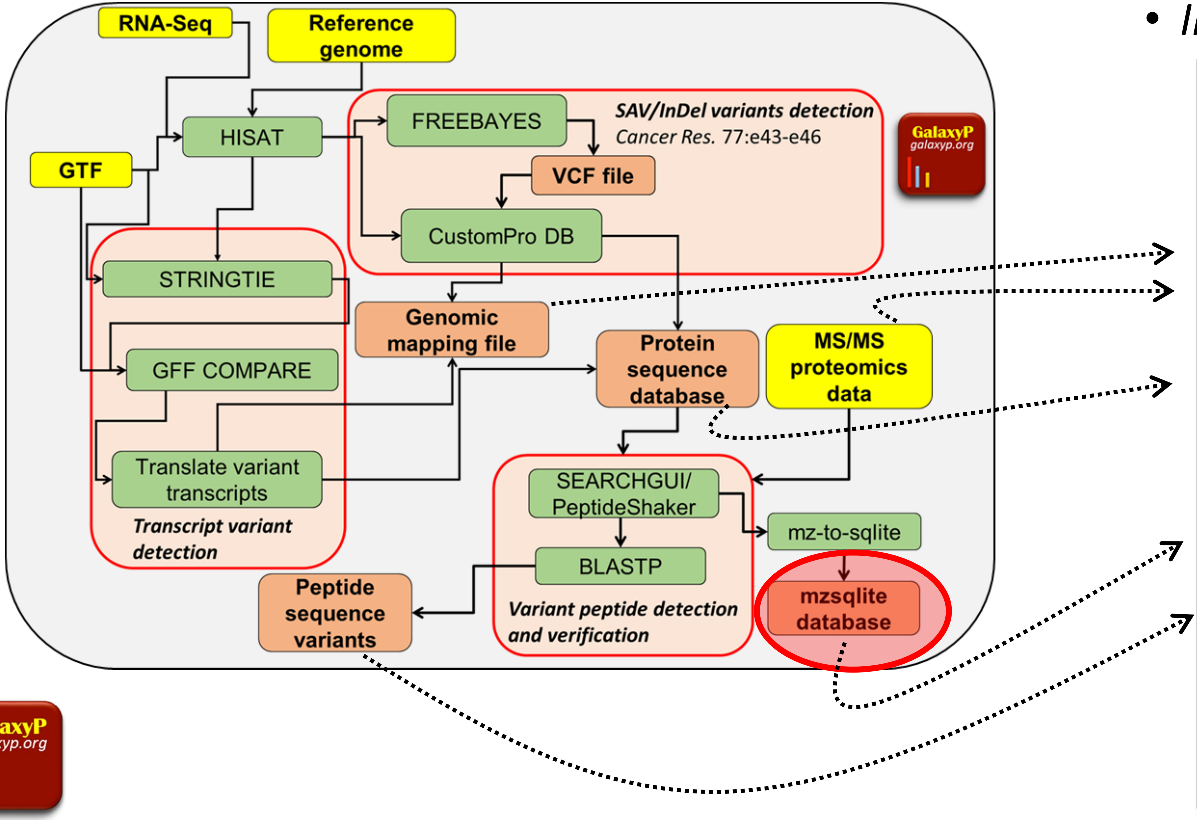
2. *Match empirical MS/MS data to protein sequences*

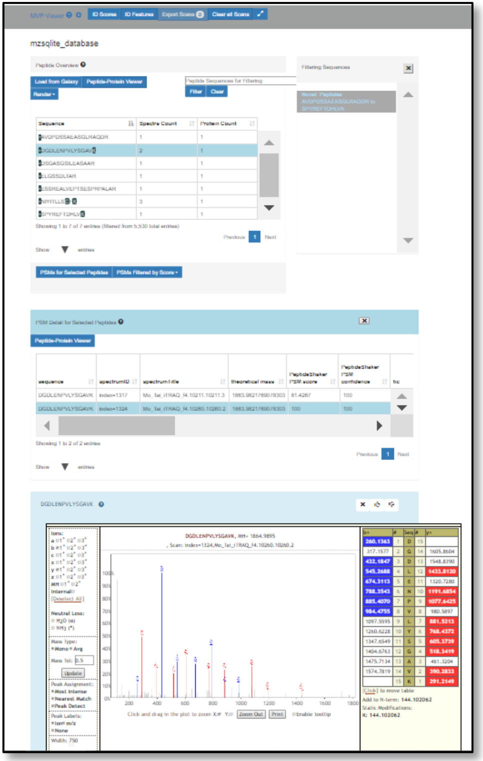


What’s next? Beyond a big list….

Characterizing the nature of detected variants

- *HTML-based Galaxy plugin*
- *Interactive reading of mzsqlite dB*





mzsqlite_database

Novel Peptides AVDPDSSAEASGLRAQDR to SPYREFTDHLVK

Sequence	Spectra Count	Protein Count
AVDPDSSAEASGLRAQDR	1	1
DGDLENPVLYSGAVK	2	1
DSGASGSILEASAAR	1	1
ELGSSDLTAR	1	1
ESSREALVEPTSESPRPALAR	1	1
NIYITLLS[illegible]	3	1
SPYREFTDHLVK	1	1

Showing 1 to 7 of 7 entries (filtered from 5,530 total entries)

Previous 1 Next

Show entries

PSM Detail for Selected Peptides

sequence	spectrumID	spectrumTitle	theoretical mass	PeptideShaker PSM score	PeptideShaker PSM confidence	bc
DGDLENPVLYSGAVK	index=1317	Mo_Tai_iTRAQ_f4.10211.10211.3	1863.9821769078303	81.4287	100	
DGDLENPVLYSGAVK	index=1324	Mo_Tai_iTRAQ_f4.10260.10260.2	1863.9821769078303	100	100	

Showing 1 to 2 of 2 entries

Previous 1 Next

Show entries

DGDLENPVLYSGAVK

Ions:
a □1+ □2+ □3+
b ☑1+ □2+ □3+
c □1+ □2+ □3+
x □1+ □2+ □3+
y ☑1+ □2+ □3+
z □1+ □2+ □3+
MH □1+ □2+
Internal □
[Deselect All]

Neutral Loss:
□ H2O (o)
□ NH3 (*)

Mass Type:
● Mono ○ Avg

Mass Tol: 0.5

Peak Assignment:
● Most Intense
○ Nearest Match
☑ Peak Detect

Peak Labels:
● Ion ○ m/z
○ None

Width: 750

DGDLENPVLYSGAVK, MH+ 1864.9895
, Scan: index=1324,Mo_Tai_iTRAQ_f4.10260.10260.2

Click and drag in the plot to zoom X:☑ Y:□ □ Enable tooltip

b+	#	Seq	#	y+
260.1363	1	D	15	
317.1577	2	G	14	1605.8604
432.1847	3	D	13	1548.8390
545.2688	4	L	12	1433.8120
674.3113	5	E	11	1320.7280
788.3543	6	N	10	1191.6854
885.4070	7	P	9	1077.6425
984.4755	8	V	8	980.5897
1097.5595	9	L	7	881.5213
1260.6228	10	Y	6	768.4372
1347.6549	11	S	5	605.3739
1404.6763	12	G	4	518.3419
1475.7134	13	A	3	461.3204
1574.7819	14	V	2	390.2833
	15	K	1	291.2149

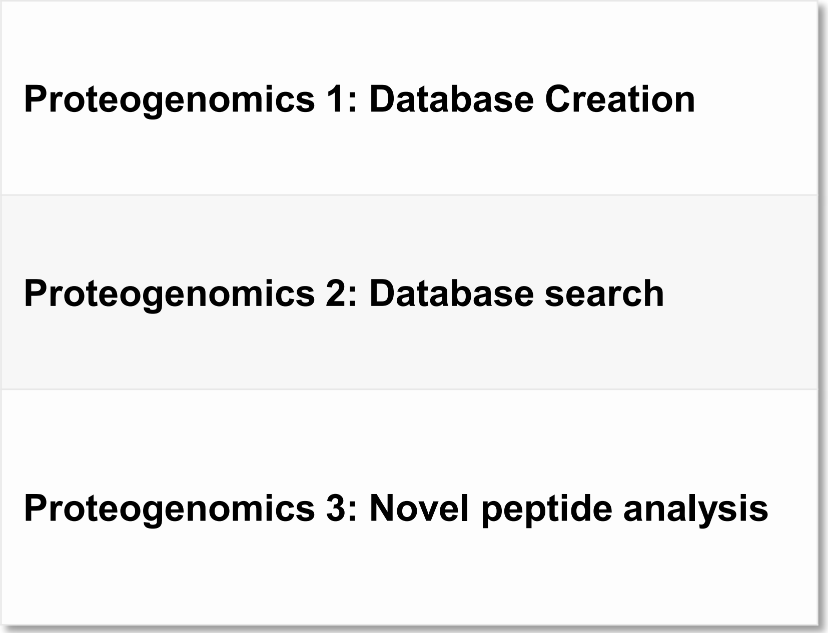
[Click] to move table

Add to N-term: 144.102062

Static Modifications:

K: 144.102062

Training tutorials for main components of proteogenomics



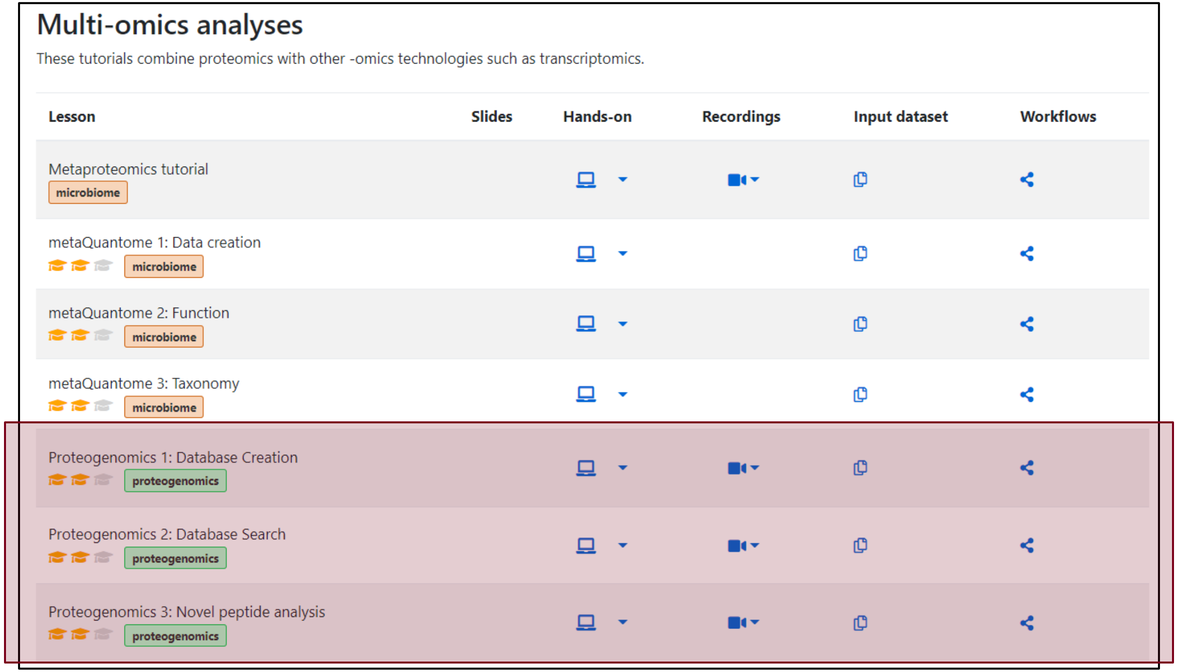
Proteogenomics 1: Database Creation

Proteogenomics 2: Database search

Proteogenomics 3: Novel peptide analysis

Please ask questions as they come up!

Training tutorials for main components of proteogenomics



Multi-omics analyses

These tutorials combine proteomics with other -omics technologies such as transcriptomics.

Lesson	Slides	Hands-on	Recordings	Input dataset	Workflows
Metaproteomics tutorial microbiome		💻 ▾	🎥 ▾	📋	🔗
metaQuantome 1: Data creation 🎓🎓 microbiome		💻 ▾		📋	🔗
metaQuantome 2: Function 🎓🎓 microbiome		💻 ▾		📋	🔗
metaQuantome 3: Taxonomy 🎓🎓 microbiome		💻 ▾		📋	🔗
Proteogenomics 1: Database Creation 🎓🎓 proteogenomics		💻 ▾	🎥 ▾	📋	🔗
Proteogenomics 2: Database Search 🎓🎓 proteogenomics		💻 ▾	🎥 ▾	📋	🔗
Proteogenomics 3: Novel peptide analysis 🎓🎓 proteogenomics		💻 ▾	🎥 ▾	📋	🔗

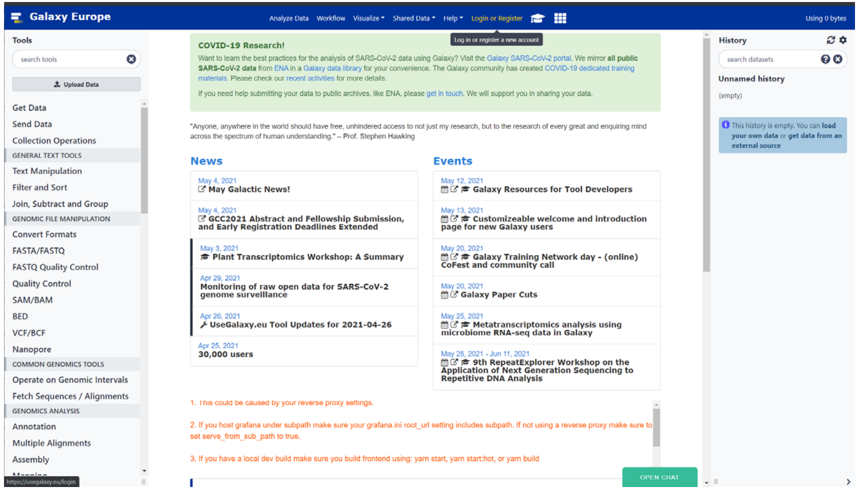
https://training.galaxyproject.org/training-material/topics/proteomics/

Accessing Galaxy via a public gateway

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

2 **Go to TIaaS link:** https://usegalaxy.eu/join-training/imsc_galaxy_training

3 **Return to usegalaxy.eu site**



COVID-19 Research!

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal. We mirror **all public SARS-CoV-2 data** from ENA in a Galaxy data library for your convenience. The Galaxy community has created COVID-19 dedicated training materials. Please check our recent activities for more details.

If you need help submitting your data to public archives, like ENA, please get in touch. We will support you in sharing your data.

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

May 4, 2021
May Galactic News!

May 4, 2021
GCC2021 Abstract and Fellowship Submission, and Early Registration Deadlines Extended

May 3, 2021
Plant Transcriptomics Workshop: A Summary

Apr 29, 2021
Monitoring of raw open data for SARS-CoV-2 genome surveillance

Apr 26, 2021
UseGalaxy.eu Tool Updates for 2021-04-26

Apr 25, 2021
30,000 users

Events

May 12, 2021
Galaxy Resources for Tool Developers

May 13, 2021
Customizeable welcome and introduction page for new Galaxy users

May 20, 2021
Galaxy Training Network day - (online) CoFest and community call

May 20, 2021
Galaxy Paper Cuts

May 25, 2021
Metatranscriptomics analysis using microbiome RNA-seq data in Galaxy

May 25, 2021 - Jun 11, 2021
9th RepeatExplorer Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis

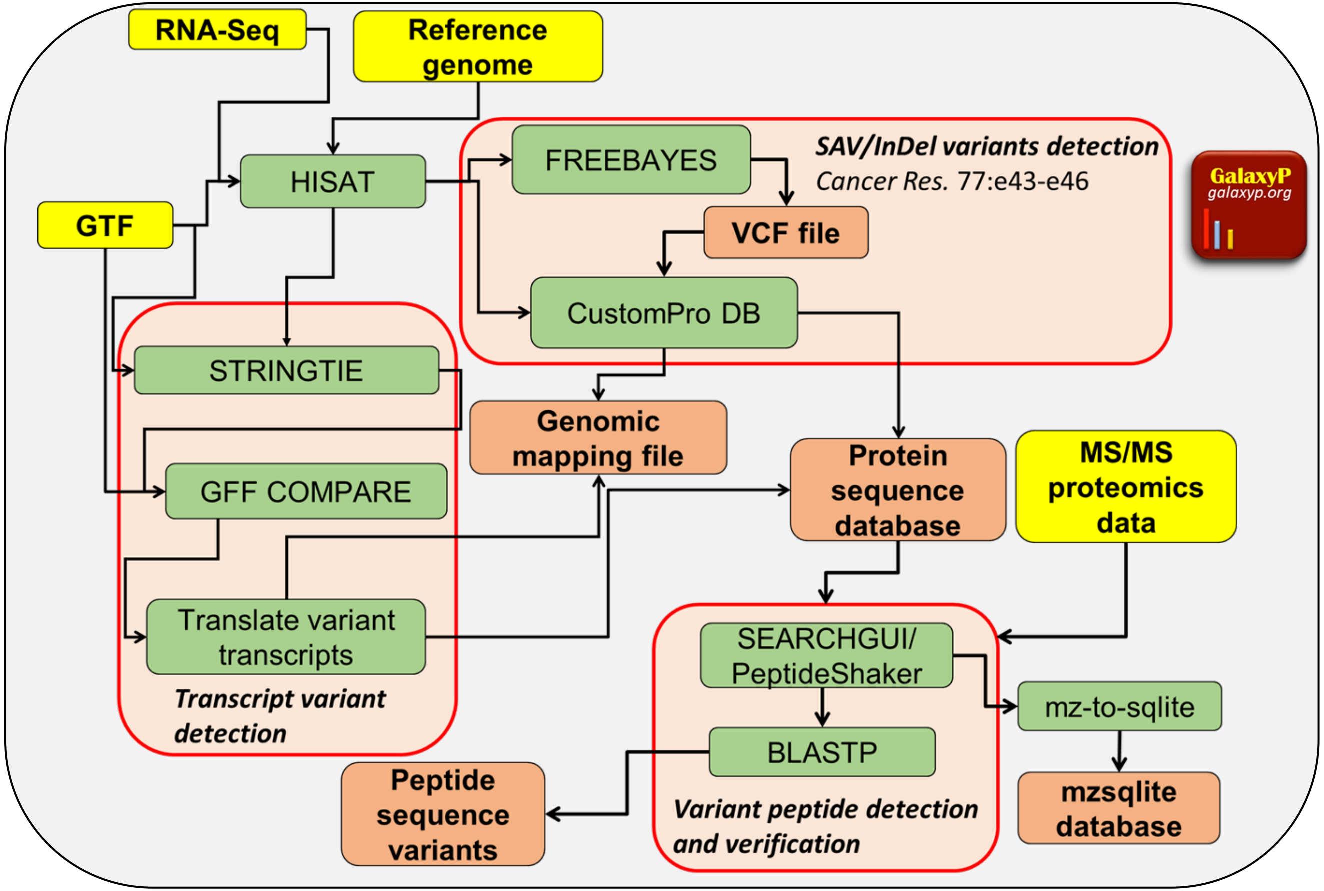
1. This could be caused by your reverse proxy settings.

2. If you host grafana under subpath make sure your grafana.ini root_url setting includes subpath. If not using a reverse proxy make sure to set serve_from_sub_path to true.

3. If you have a local dev build make sure you build frontend using: yarn start, yarn start:hot, or yarn build

This history is empty. You can **load your own data** or **get data from an external source**

Database generation from RNA-Seq data: SAV/Indels + unexpected novel translation events



GalaxyP
galaxyp.org

Working with Galaxy

1. Start with a new and blank History in usegalaxy.edu. Give it a name of your choosing. Go to the link below, which will open a Data Library in usegalaxy.eu. Select each of the three datasets ending in the text shown below (a-c) and hit “Export to History”. Export each “As dataset” and select your newly named History as the destination.

https://z.umn.edu/imscdata

a) FASTQ_ProB_22LIST.fastqsanger
b) Mus_musculus.GRCm38.86.gtf
c) Trimmed_ref_5000_uniprot_cRAP.fasta (On page 2 of the data library)

2. In your active History, hit the “pen” icon and rename each History item as shown for a-c above, taking out the URL information to simplify the names

3. Listen to lecture on the details of the upcoming analysis
   (https://youtu.be/b_kZf8mXHdo)

Running a workflow: Database creation

1. Go to Shared Data → Workflows and search for “IMSC Proteogenomics 1”
2. On the dropdown menu select “Run”
3. In the dialogue box select the correct three input files from your History, matching names to the input item in your History
4. Hit “Run Workflow”

Part 2: Sequence database searching

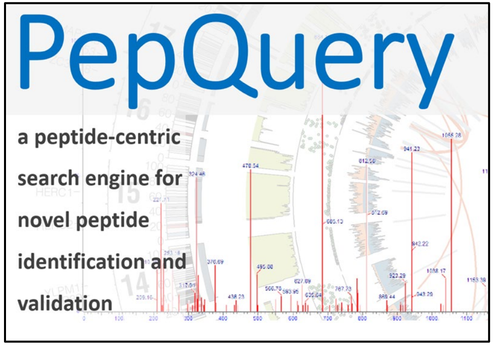
After listening to the introductory lecture information, import the completed History for this module of the workshop:

1. Go to Shared Data → Histories and search for “IMSC2022”

2. Select “IMSC2022 Proteogenomics 2: Database Search” and hit the “+Import” button in the upper right corner of the page

3. Listen to the remaining lecture material and then we will explore some of the inputs, tools and outputs.
   (https://youtu.be/q1OjmTcbvBA)

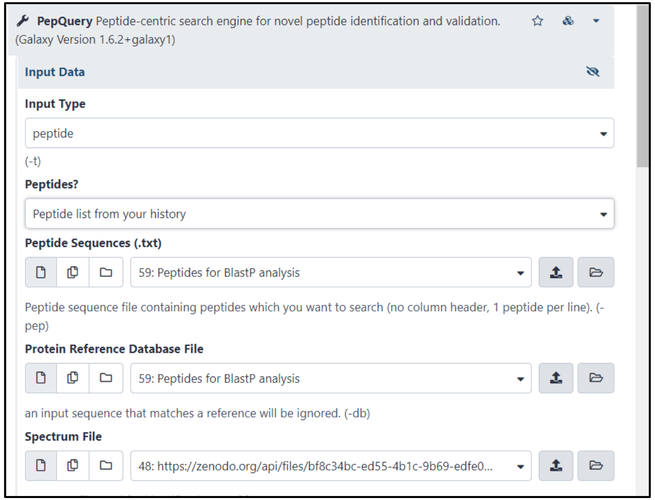
Ensuring confidence in MS/MS matches to novel peptide sequences: important!



PepQuery

a peptide-centric search engine for novel peptide identification and validation

Genome research 29.3 (2019): 485-493.



PepQuery Peptide-centric search engine for novel peptide identification and validation. (Galaxy Version 1.6.2+galaxy1)

Input Data

Input Type

peptide

(-t)

Peptides?

Peptide list from your history

Peptide Sequences (.txt)

59: Peptides for BlastP analysis

Peptide sequence file containing peptides which you want to search (no column header, 1 peptide per line). (-pep)

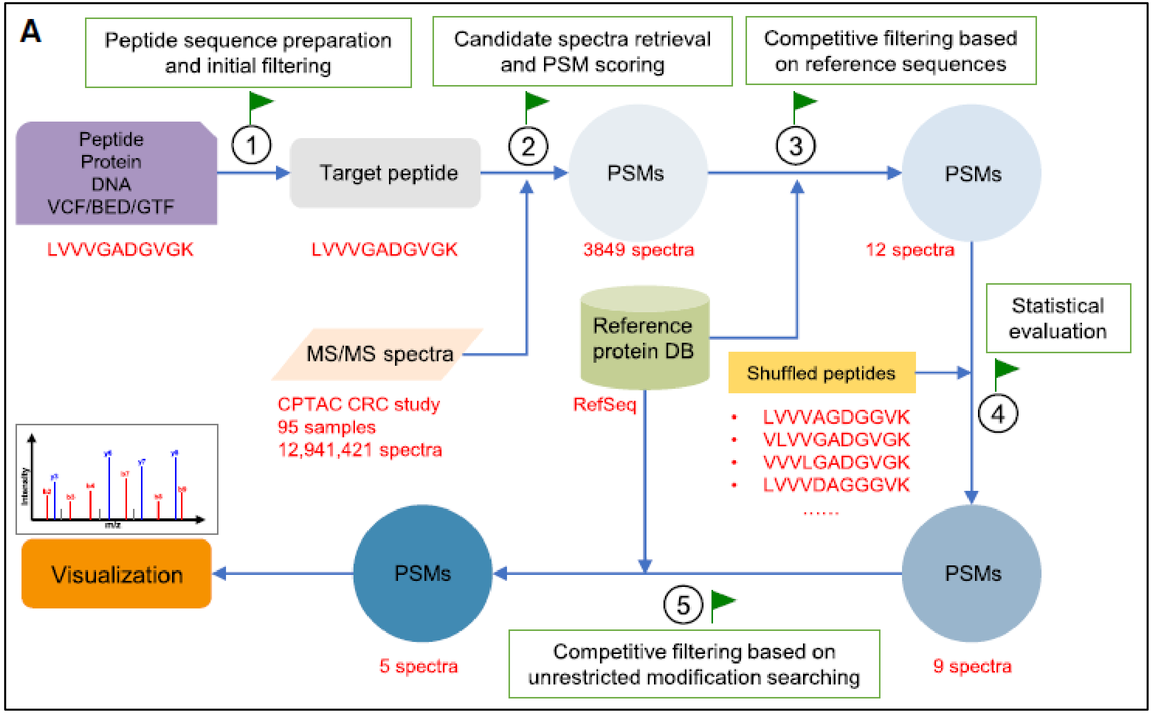
Protein Reference Database File

59: Peptides for BlastP analysis

an input sequence that matches a reference will be ignored. (-db)

Spectrum File

48: https://zenodo.org/api/files/bf8c34bc-ed55-4b1c-9b69-edfe0...



A

Peptide sequence preparation and initial filtering

Candidate spectra retrieval and PSM scoring

Competitive filtering based on reference sequences

Peptide
Protein
DNA
VCF/BED/GTF

LVVVGADGVGK

1

Target peptide

LVVVGADGVGK

2

PSMs

3849 spectra

3

PSMs

12 spectra

MS/MS spectra

CPTAC CRC study
95 samples
12,941,421 spectra

Reference protein DB

RefSeq

Shuffled peptides

- LVVVAGDGGVK
- VLVVGADGVGK
- VVVLGADGVGK
- LVVVDAGGGVK

……

Statistical evaluation

4

PSMs

9 spectra

5

Competitive filtering based on unrestricted modification searching

PSMs

5 spectra

Visualization

Part 3: Novel peptide analysis

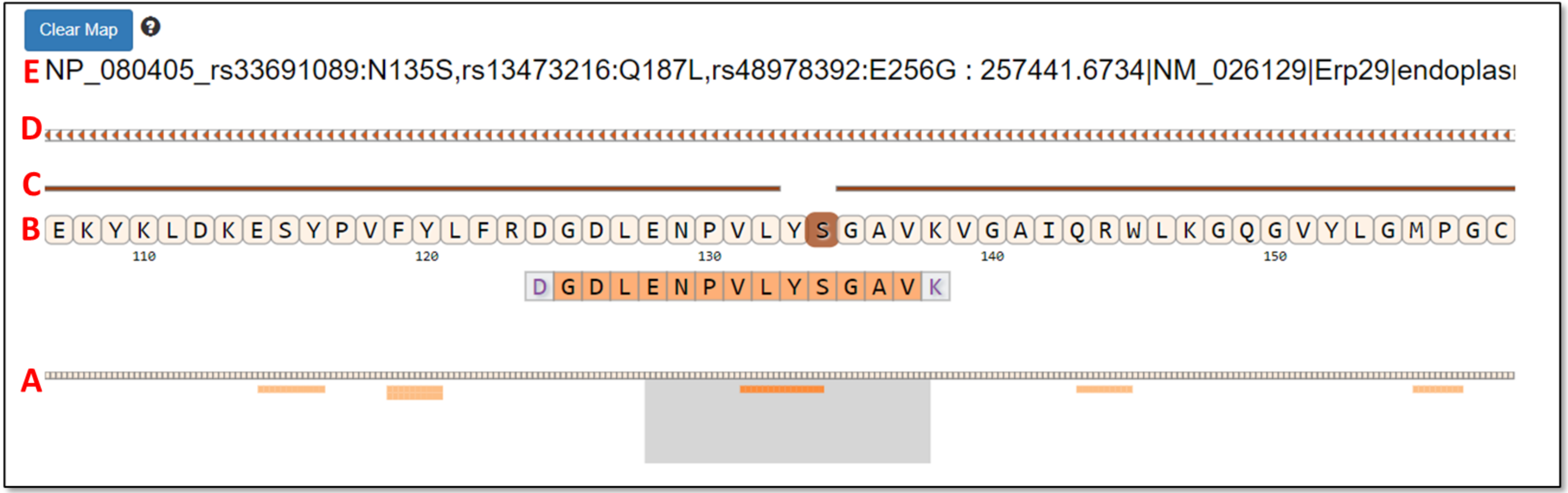
After listening to the introductory lecture information, import the completed History for this module of the workshop:

1. Go to Shared Data → Histories and search for “IMSC2022”

2. Select “IMSC2022 Proteogenomics 3: Novel Peptide Analysis” and hit the “+Import” button in the upper right corner of the page

3. Listen to the remaining lecture material and guidance on exploring results (https://youtu.be/Ku274KwFh1Y)

Viewing and characterizing novel peptide sequences

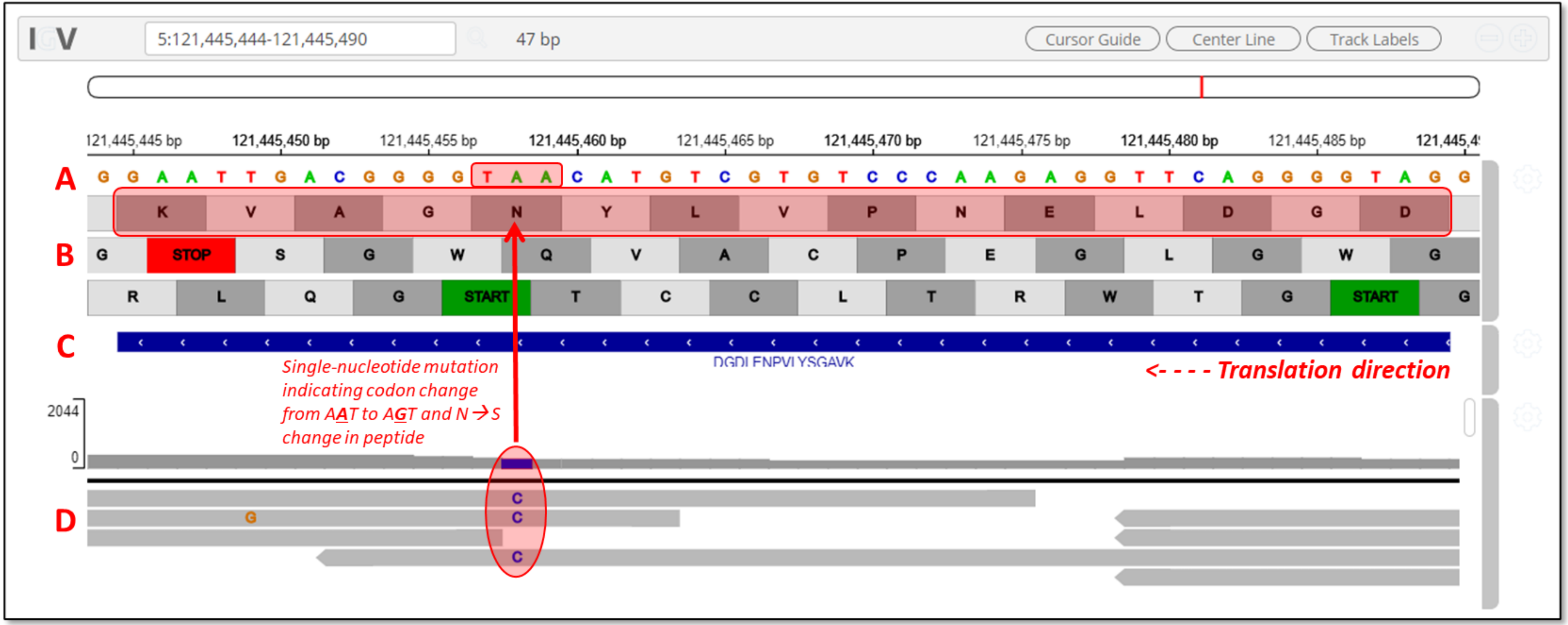


E NP_080405_rs33691089:N135S,rs13473216:Q187L,rs48978392:E256G : 257441.6734|NM_026129|Erp29|endoplas

Clear Map
D
C
B
E K Y K L D K E S Y P V F Y L F R D G D L E N P V L Y S G A V K V G A I Q R W L K G Q G V Y L G M P G C
110
120
130
140
150
D G D L E N P V L Y S G A V K
A

Gigascience (2020), **9**:giaa025

IGV viewer for visualizing peptide, RNA and DNA sequences

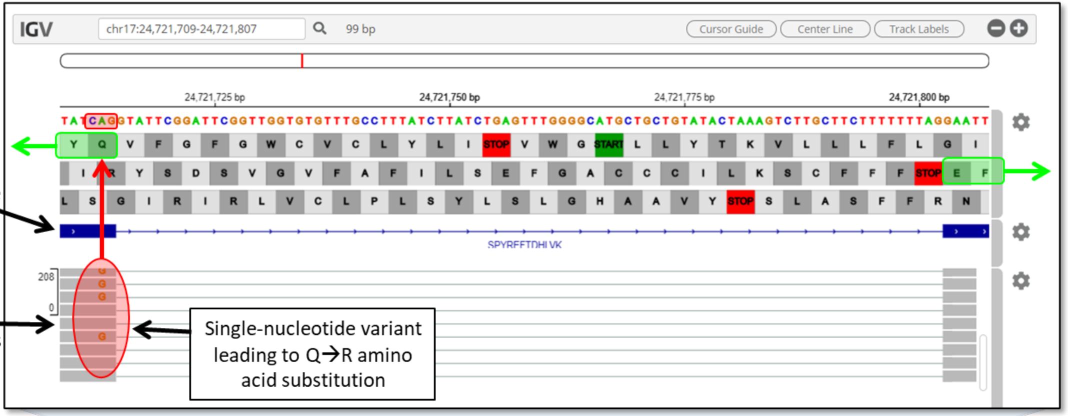


Single-nucleotide mutation indicating codon change from AAT to AGT and N → S change in peptide

Splice isoform example

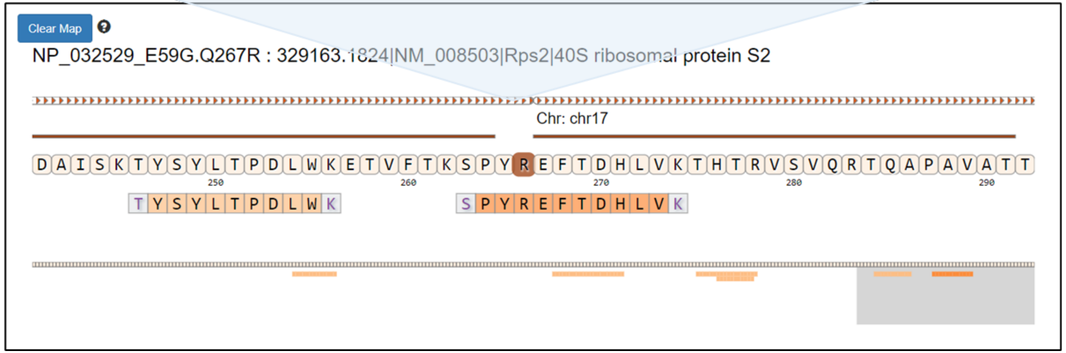
Translated peptide spanning junction

Assembled RNA-Seq reads



chr17:24,721,709-24,721,807 99 bp

Single-nucleotide variant leading to Q→R amino acid substitution



NP_032529_E59G.Q267R : 329163.1824|NM_008503|Rps2|40S ribosomal protein S2

Chr: chr17