Proteogenomic workflows

Tim Griffin
University of Minnesota

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

NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

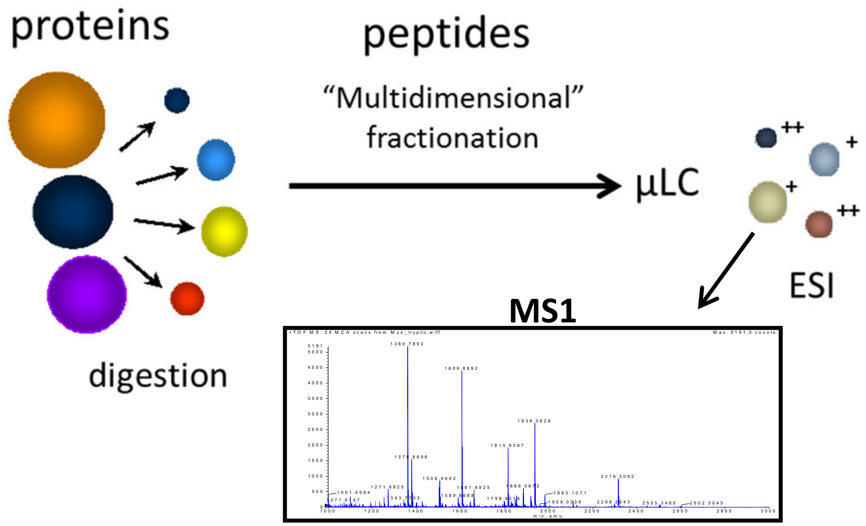
GalaxyP
galaxyp.org

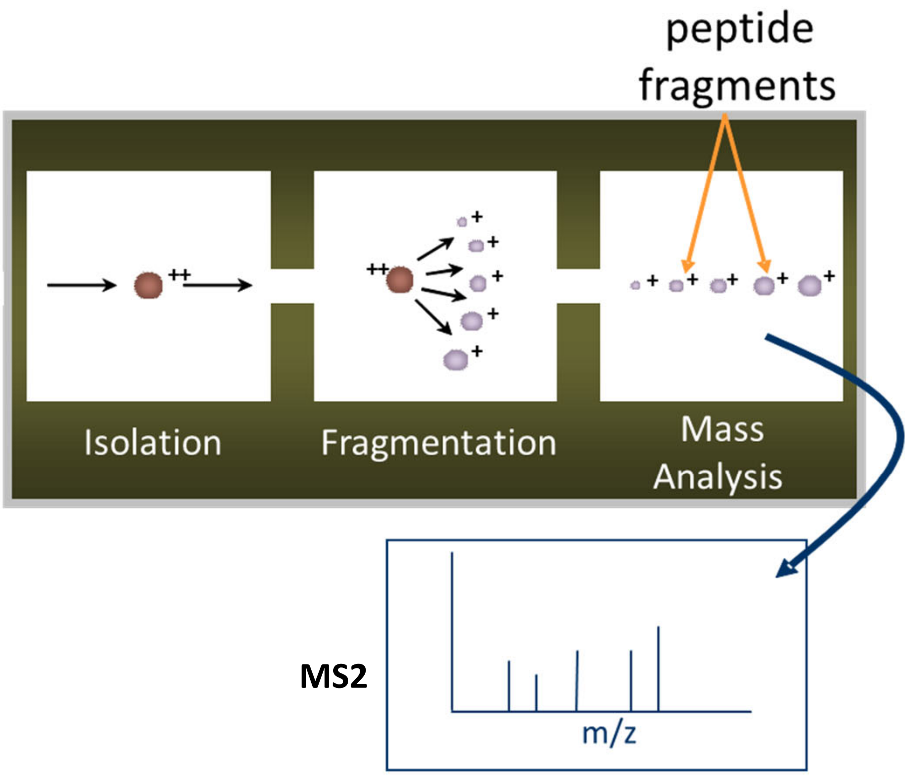
Outline: Proteogenomic workflows

- Background and informatics challenges
- Overview of existing software and workflows
- Access to the community

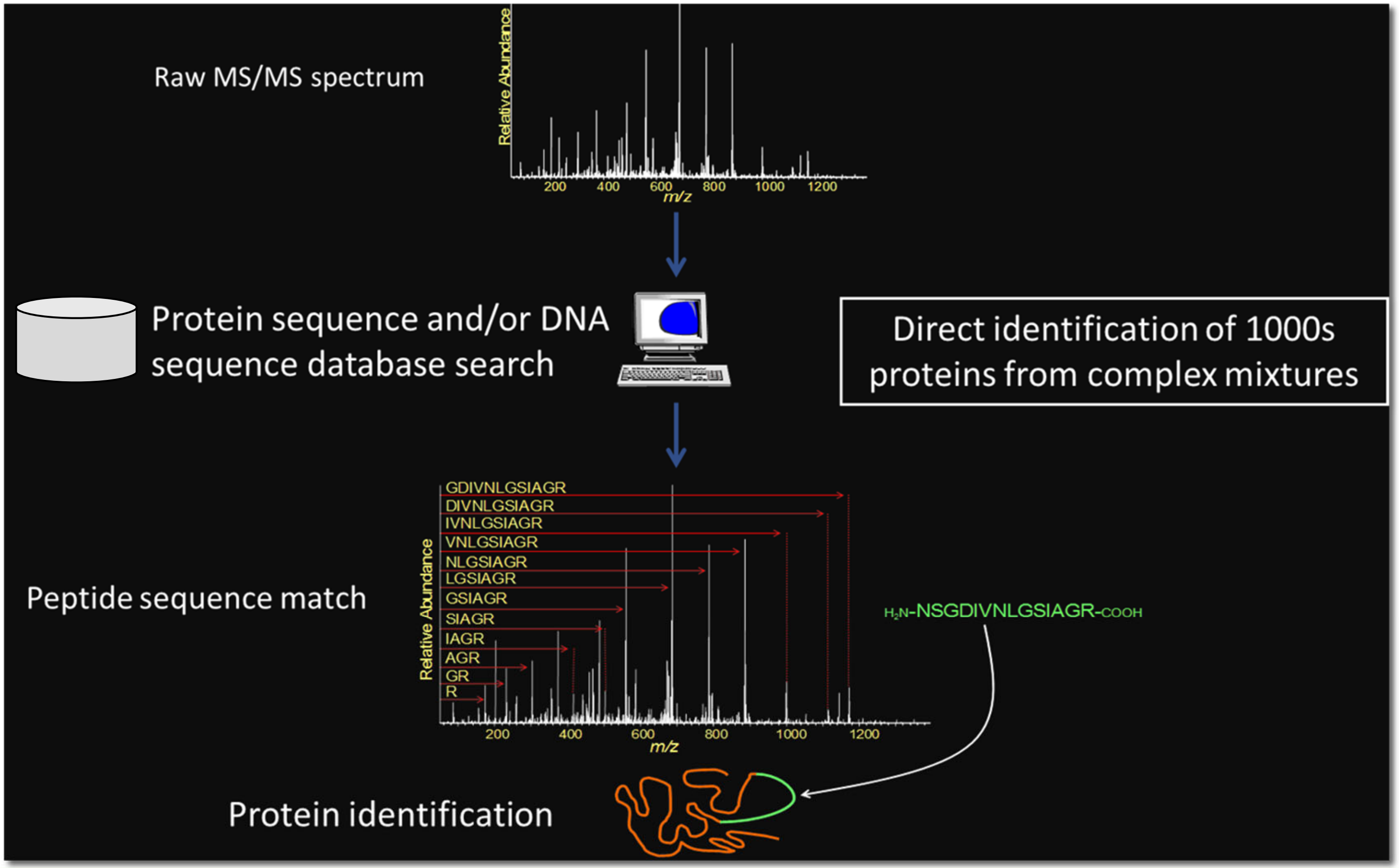
Proteogenomics: A primer

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





Matching amino acid sequences to MS/MS data



Raw MS/MS spectrum

Protein sequence and/or DNA sequence database search

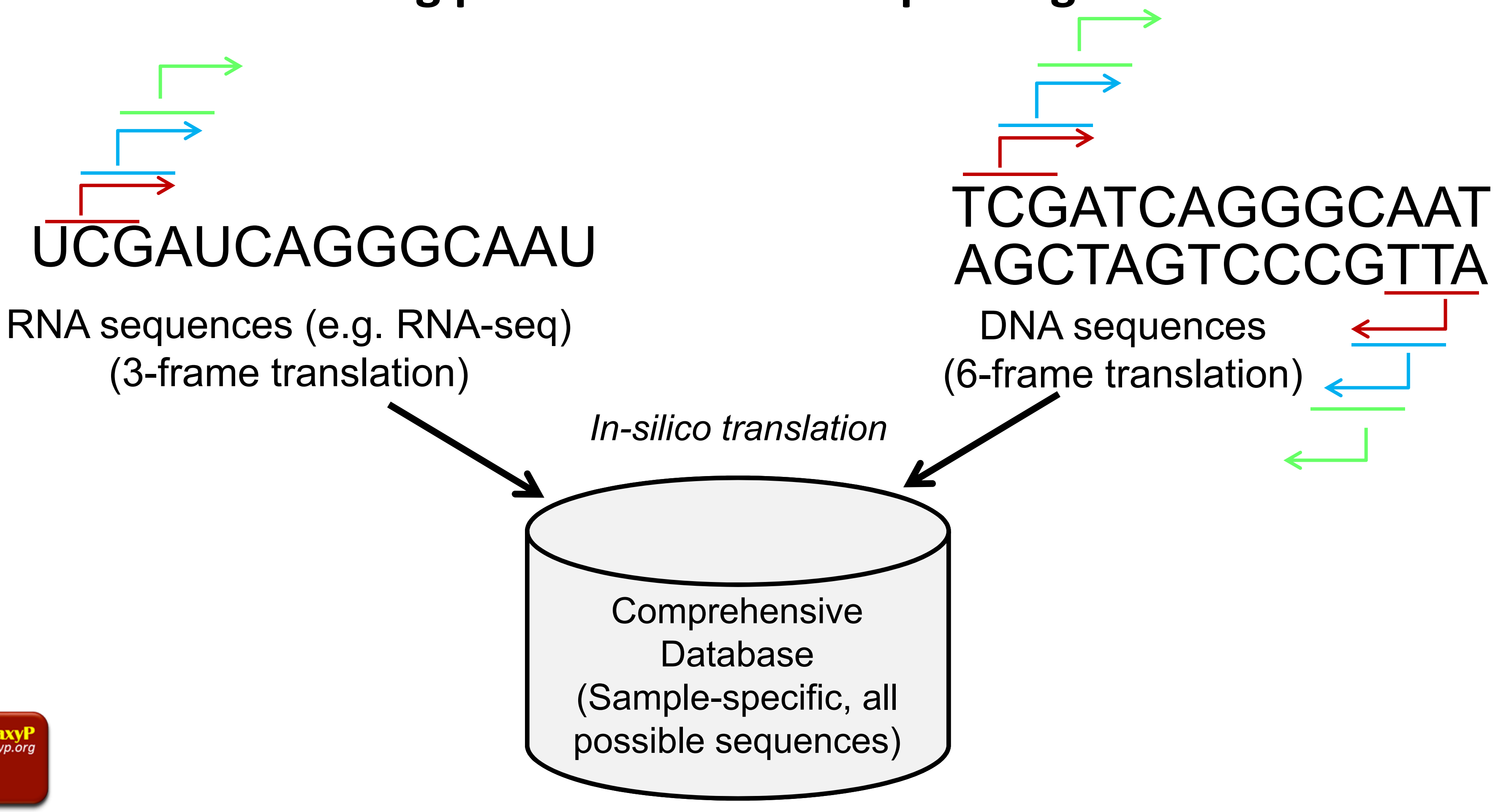
Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification

GalaxyP
galaxyp.org

Detecting protein variants via proteogenomics



UCGAUCAGGGCAAU

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

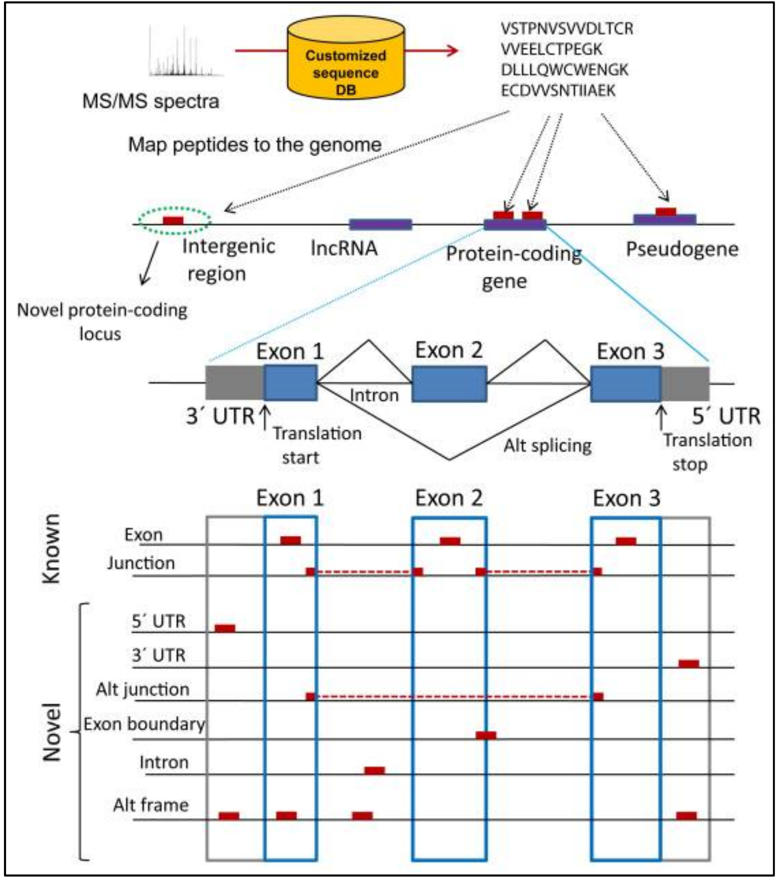
TCGATCAGGGCAAT
AGCTAGTCCCGTTA

DNA sequences
(6-frame translation)

In-silico translation

Comprehensive
Database
(Sample-specific, all
possible sequences)

Proteogenomic outcomes



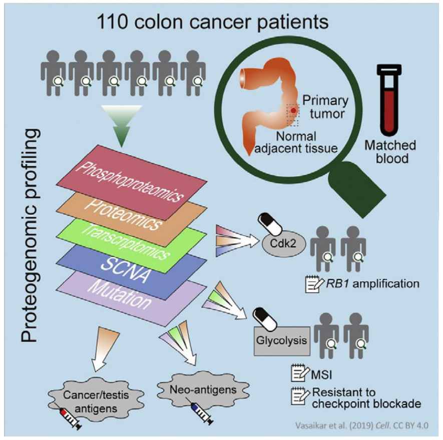
VOL.11 |NO.11 | NOVEMBER 2014 | **NATURE METHODS**

- ✓ *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

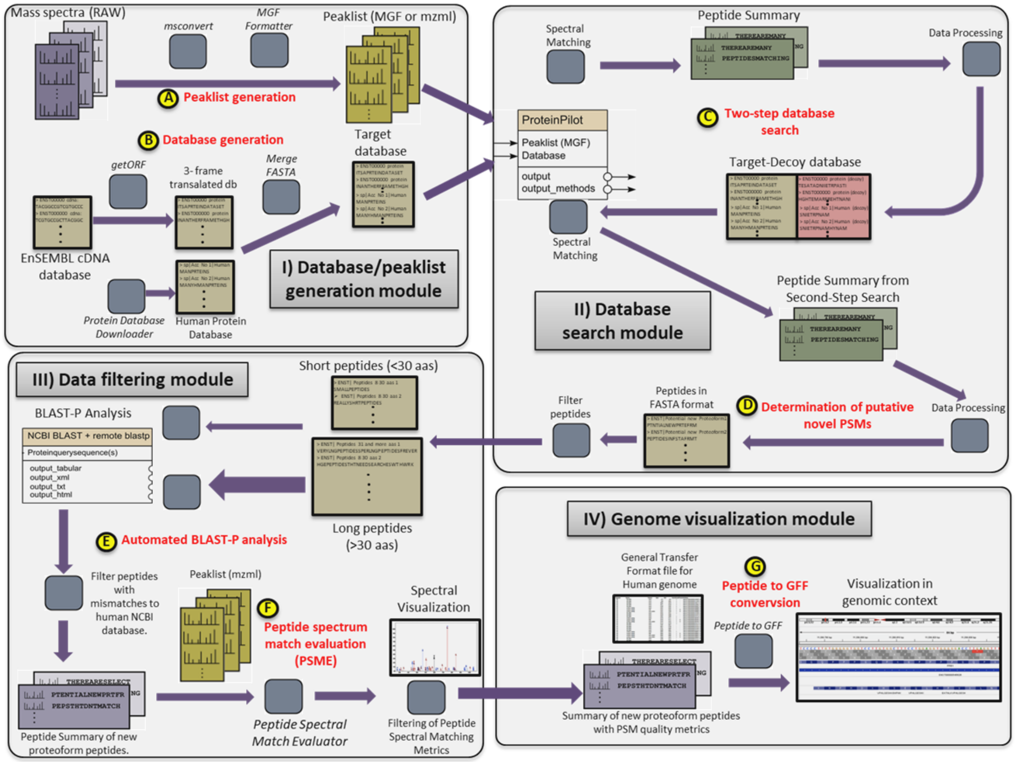


Vasaikar et al. (2019) *Cell.* CC BY 4.0

GalaxyP
galaxyp.org

Bringing proteogenomics to the masses: informatics challenges

- Many software tools, integration, automation….



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

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*

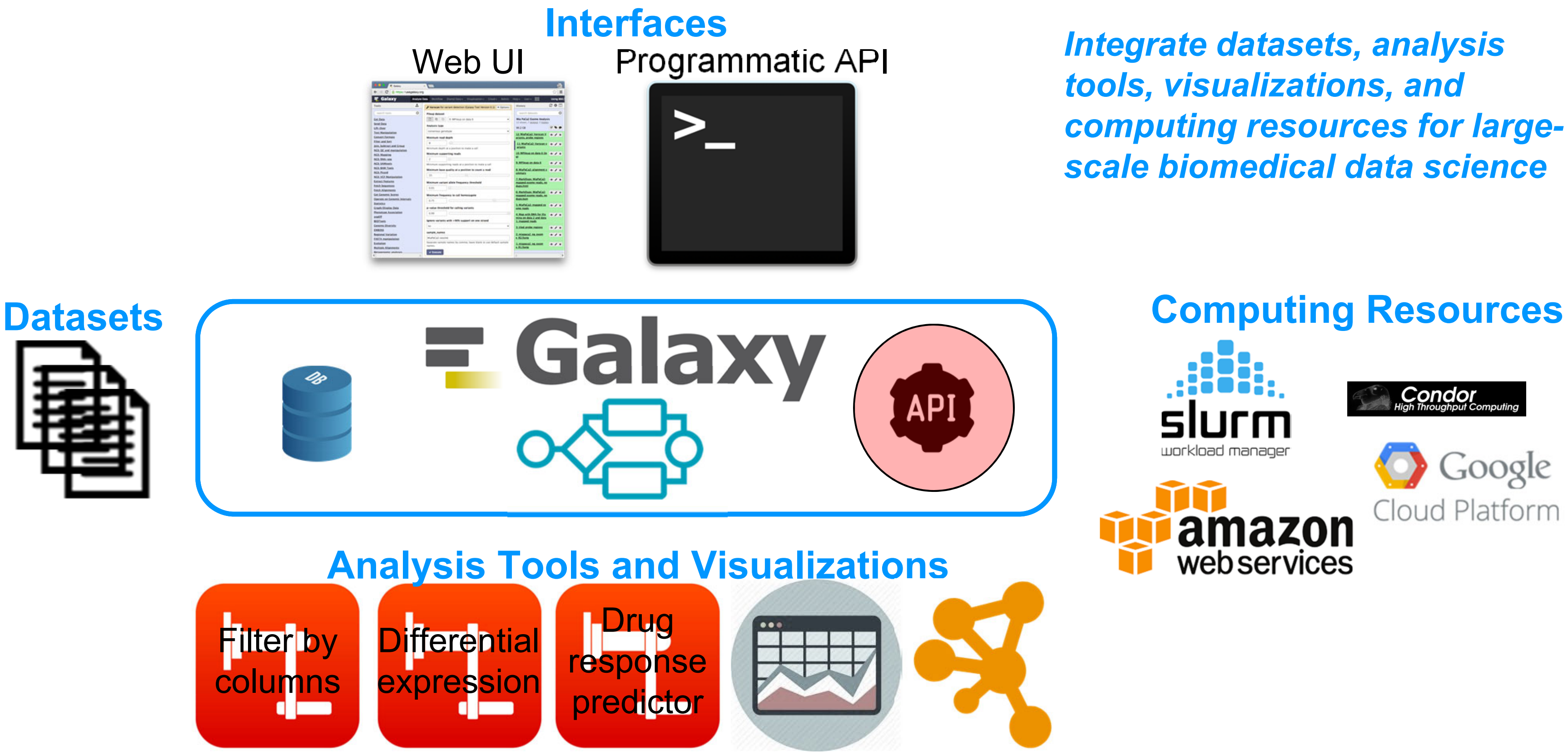
One workflow solution: Galaxy

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:

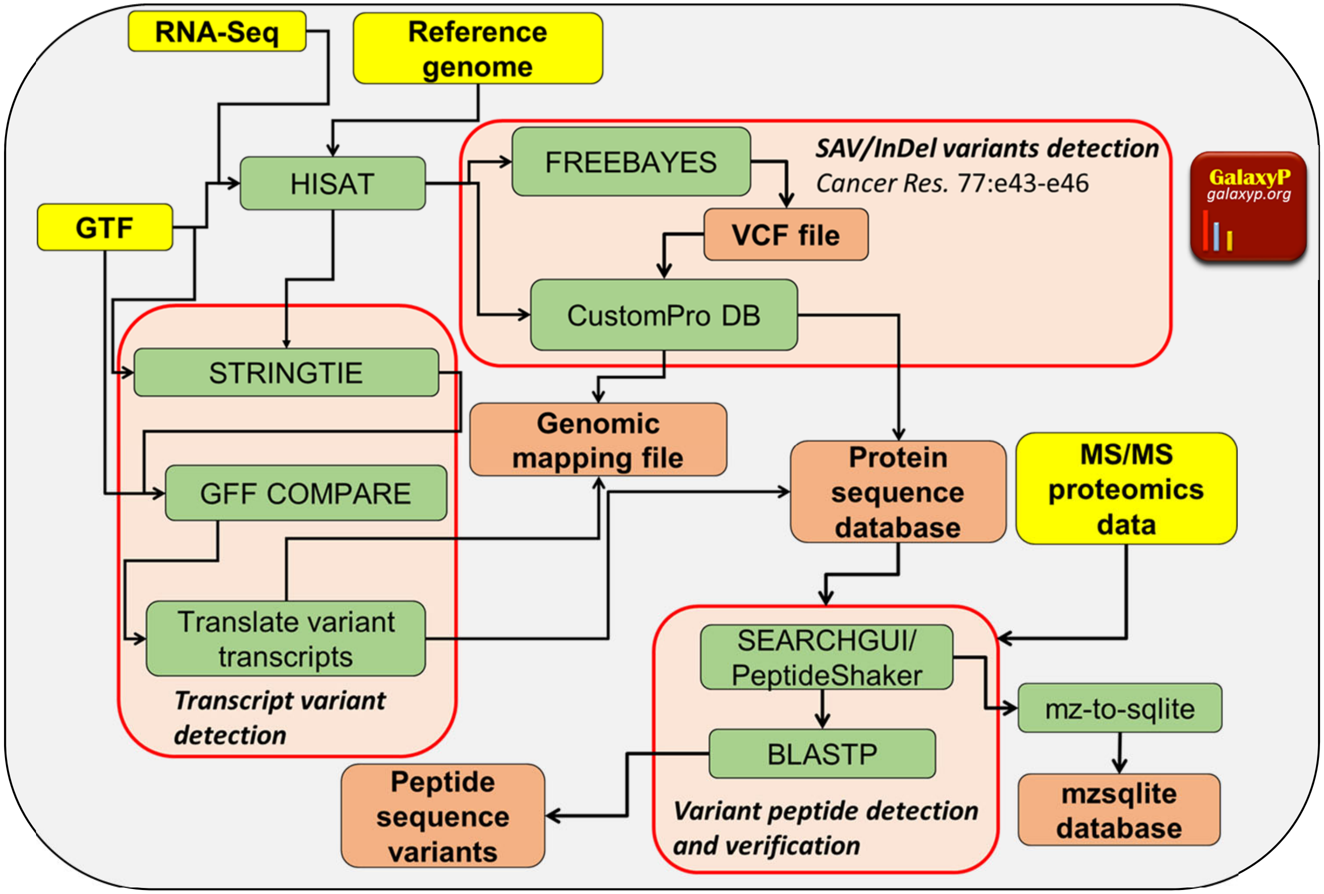
Galaxy: an integrative workbench well-suited for multi-omics



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

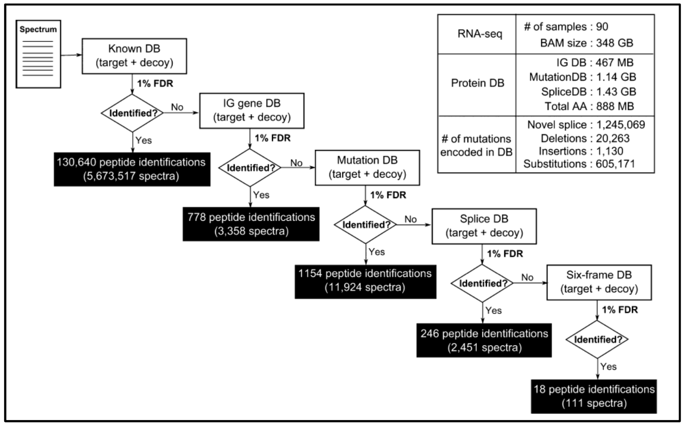
Courtesy Jeremy Goecks, OHSU

Integrative data processing: RNA-Seq + proteomics



Best practices for peptide spectrum matches (PSMs) in proteogenomics

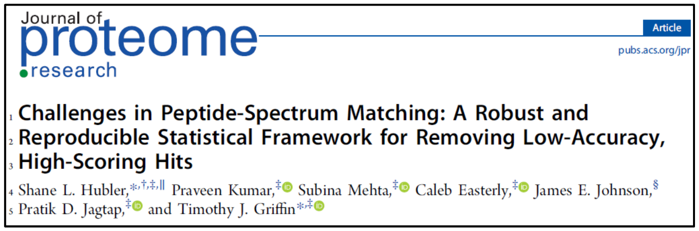
- *Utilize multiple database searching programs (e.g. SearchGUI)*
- *Multi-stage database search to obtain variant specific FDR estimates*


Spectrum
Known DB (target + decoy)
1% FDR
Identified?
No
IG gene DB (target + decoy)
Yes
130,640 peptide identifications (5,673,517 spectra)
1% FDR
Identified?
No
Mutation DB (target + decoy)
Yes
778 peptide identifications (3,358 spectra)
1% FDR
Identified?
No
Splice DB (target + decoy)
Yes
1154 peptide identifications (11,924 spectra)
1% FDR
Identified?
No
Six-frame DB (target + decoy)
Yes
246 peptide identifications (2,451 spectra)
1% FDR
Identified?
Yes
18 peptide identifications (111 spectra)

RNA-seq	# of samples : 90 BAM size : 348 GB
Protein DB	IG DB : 467 MB MutationDB : 1.14 GB SpliceDB : 1.43 GB Total AA : 888 MB
# of mutations encoded in DB	Novel splice : 1,245,069 Deletions : 20,263 Insertions : 1,130 Substitutions : 605,171

***J Proteome Res*.** 2015 14:3555-67

- *Post-processing filtering for extra scrutiny of putatively confident PSMs*


Journal of proteome research

Article

pubs.acs.org/jpr

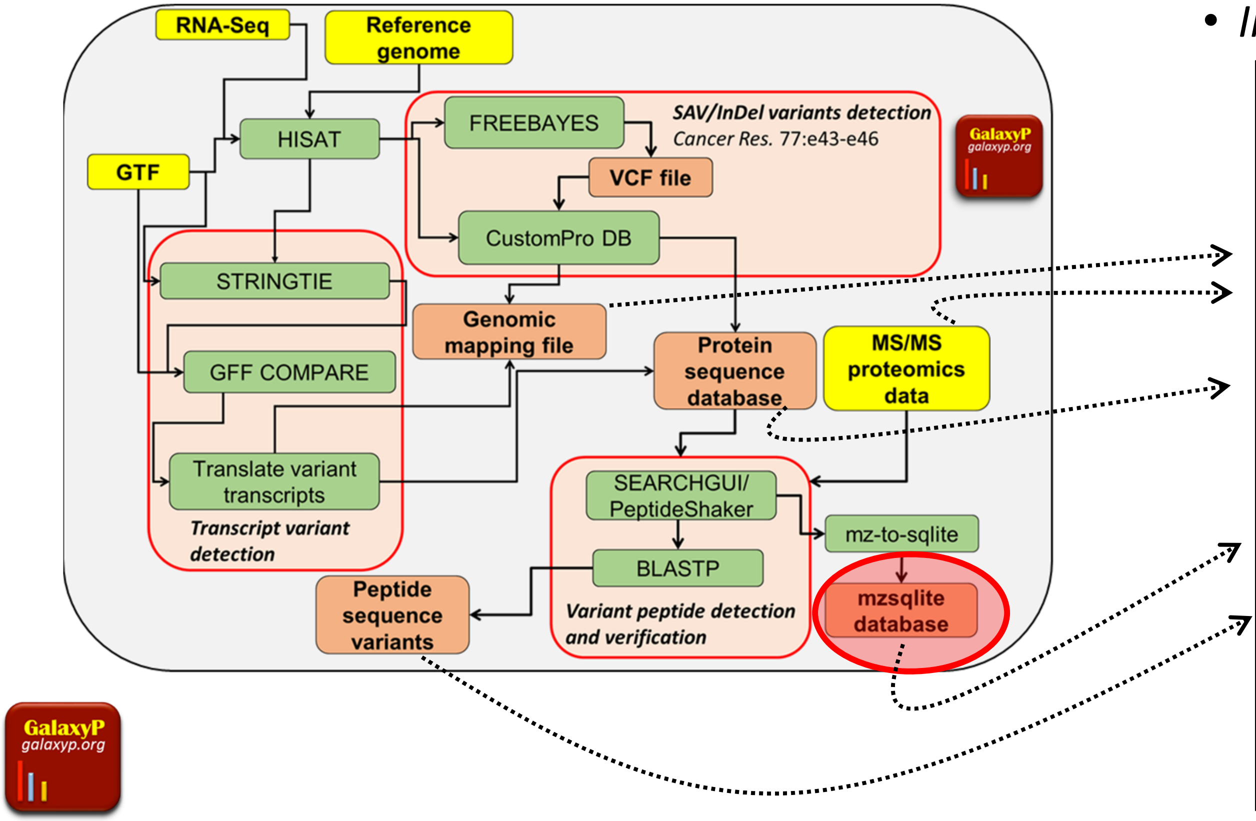
Challenges in Peptide-Spectrum Matching: A Robust and Reproducible Statistical Framework for Removing Low-Accuracy, High-Scoring Hits

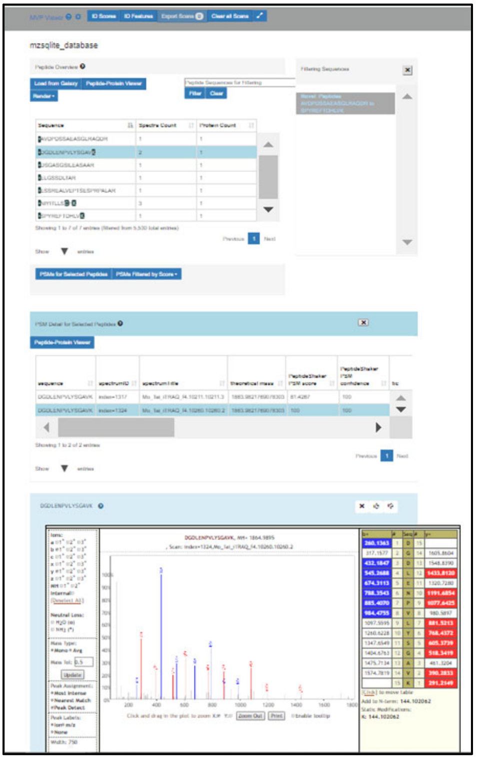
Shane L. Hubler,[*,†,‡,∥] Praveen Kumar,[‡] Subina Mehta,[‡] Caleb Easterly,[‡] James E. Johnson,[§] Pratik D. Jagtap,[‡] and Timothy J. Griffin[*,‡]

What's next? Beyond a big list….

Multi-Omics Visualization Platform:
Characterizing the nature of detected variants

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





mzsqlite_database

Novel Peptides AVDPDSSAEASGLRAQDR to SPYREFTDHLVR

Sequence	Spectra Count	Protein Count
AVDPDSSAEASGLRAQDR	1	1
DGDLENPVLYSGAVK	2	1
DSGASGSILEASAAR	1	1
ELGSSDLIAR	1	1
ESSREALVEPTSESPRPALAR	1	1
NYITLLSGK	3	1
SPYREFTDHLVR	1	1

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

Show entries

sequence	spectrumID	spectrumTitle	theoretical mass	PeptideShaker PSM score	PeptideShaker PSM confidence	bc
DGDLENPVLYSGAVK	index=1317	Mo_Tai_iTRAQ_f4.10211.10211.3	1863.98217690783	81.4267	100	
DGDLENPVLYSGAVK	index=1324	Mo_Tai_iTRAQ_f4.10260.10260.2	1863.98217690783	100	100	

Showing 1 to 2 of 2 entries

Show entries

DGDLENPVLYSGAVK

DGDLENPVLYSGAVK, MH+ 1864.9895

, Scan: index=1324,Mo_Tai_iTRAQ_f4.10260.10260.2

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

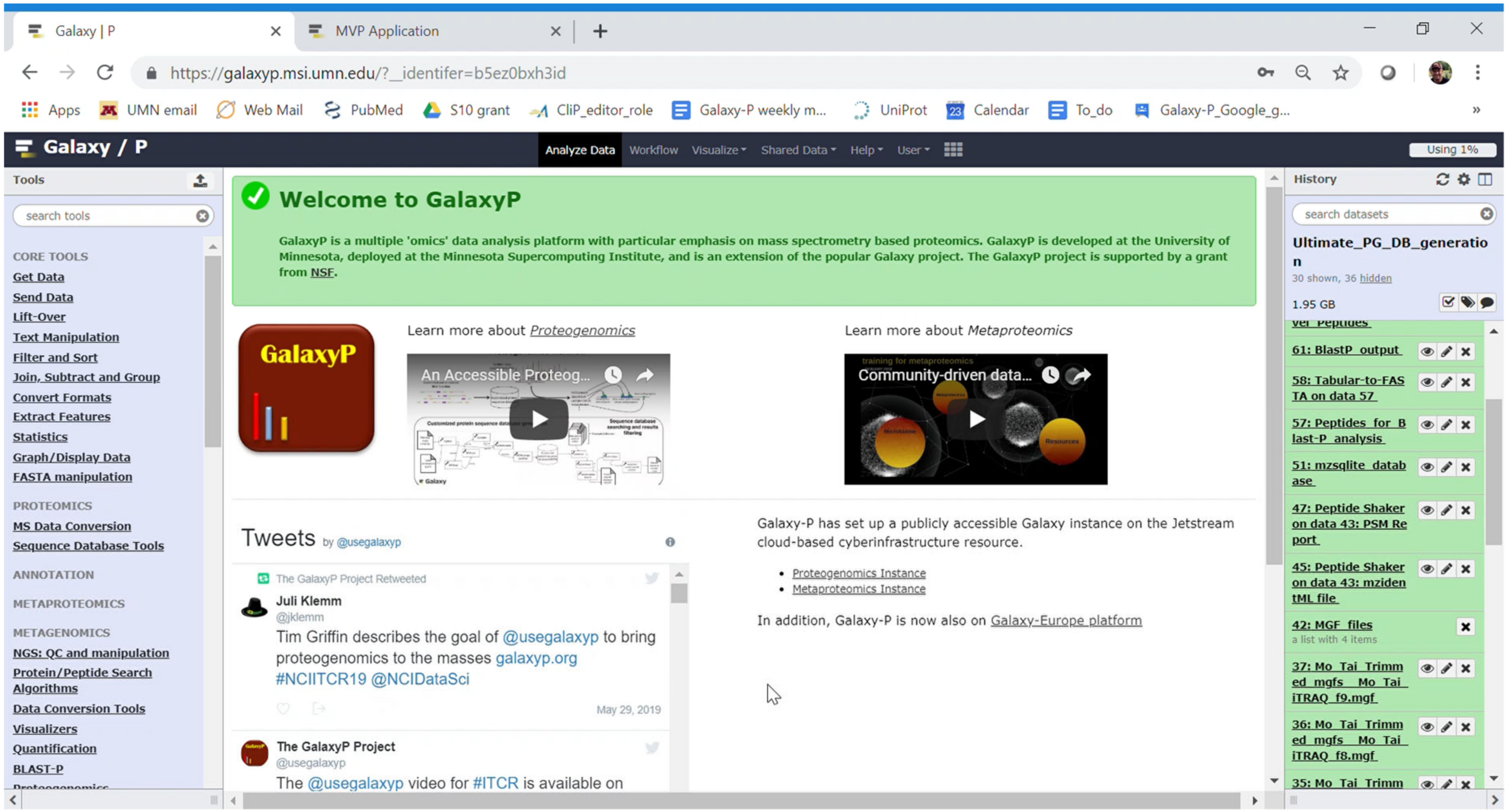
Width: 750

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

Multi-Omics Visualization Platform: Characterizing the nature of detected variants



Welcome to GalaxyP

GalaxyP is a multiple 'omics' data analysis platform with particular emphasis on mass spectrometry based proteomics. GalaxyP is developed at the University of Minnesota, deployed at the Minnesota Supercomputing Institute, and is an extension of the popular Galaxy project. The GalaxyP project is supported by a grant from NSF.

Learn more about *Proteogenomics*

Learn more about *Metaproteomics*

Tweets by @usegalaxyp

The GalaxyP Project Retweeted

Juli Klemm
@jklemm

Tim Griffin describes the goal of @usegalaxyp to bring proteogenomics to the masses galaxyp.org #NCIITCR19 @NCIDataSci

May 29, 2019

The GalaxyP Project
@usegalaxyp

The @usegalaxyp video for #ITCR is available on

Galaxy-P has set up a publicly accessible Galaxy instance on the Jetstream cloud-based cyberinfrastructure resource.

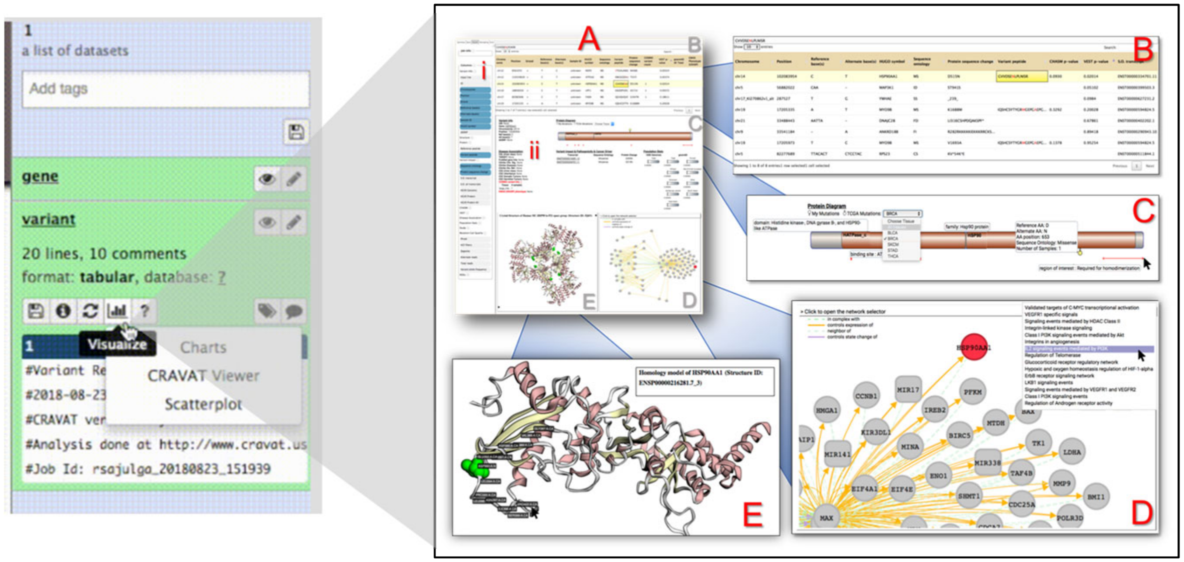
- Proteogenomics Instance
- Metaproteomics Instance

In addition, Galaxy-P is now also on Galaxy-Europe platform

GalaxyP
galaxyp.org

Other visualization and interpretation tools for protegenomics

- *Leveraging other tools and knowledgebases to assses impact of protein sequence variants (depends on good API development)*



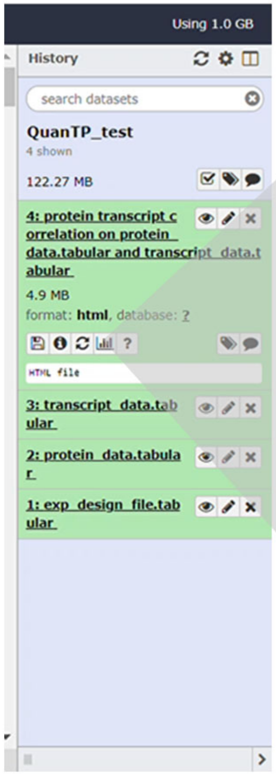
1
a list of datasets
Add tags
gene
variant
20 lines, 10 comments
format: tabular, database: ?

```
#Variant Re
#2018-08-23
#CRAVAT ver
#Analysis done at http://www.cravat.us
#Job Id: rsajulga_20180823_151939
```

“CRAVAT-P”
J Proteome Res. 2018 ,17:4329-4336

Other visualization and interpretation tools for protegenomics

- *Quantitative proteo-transcriptomics: comparing RNA and protein response*



Using 1.0 GB

History

search datasets

QuanTP_test

4 shown

122.27 MB

4: protein transcript correlation on protein data.tabular and transcript data.tabular

4.9 MB

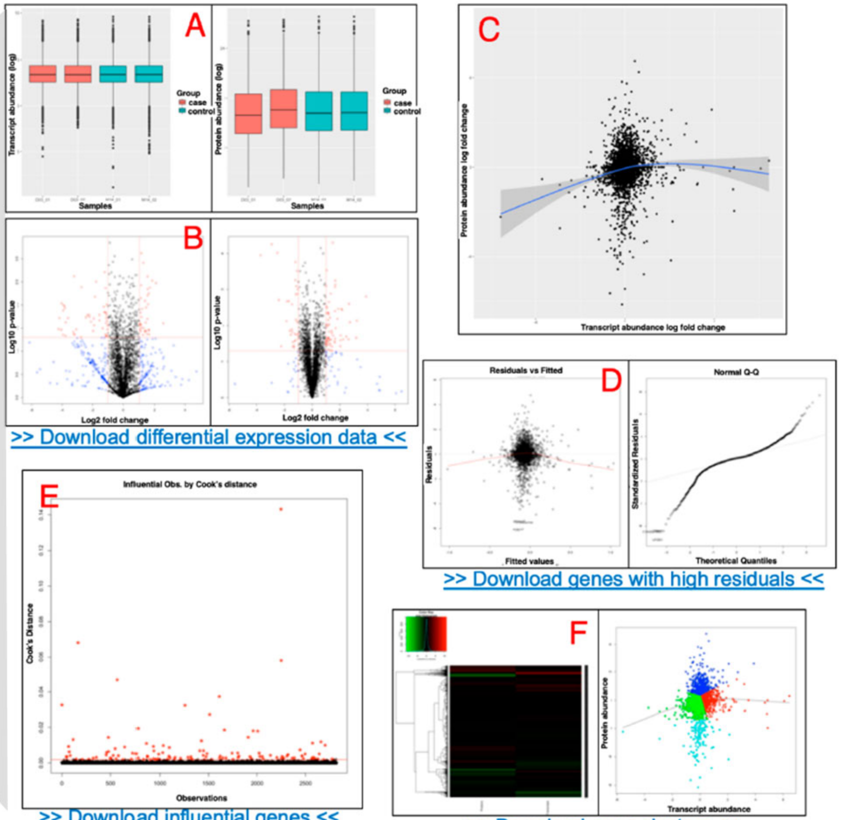
format: html, database: ?

HTML file

3: transcript_data.tabular

2: protein_data.tabular

1: exp_design_file.tabular



>> Download differential expression data <<

>> Download genes with high residuals <<

>> Download influential genes <<

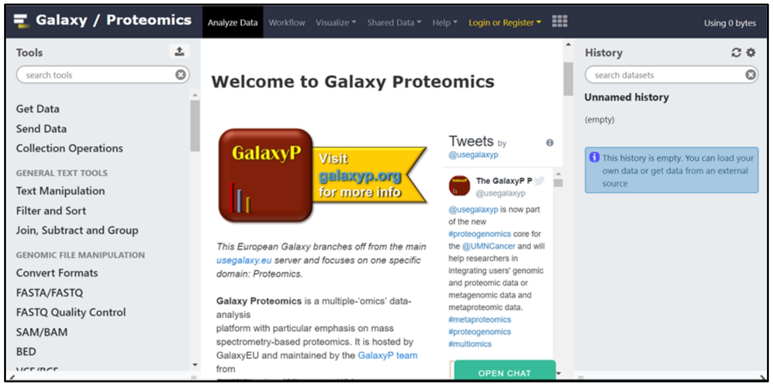
>> Download gene clusters <<

QuanTP

J Proteome Res. 2019, 18:782-790.

Providing access to the research community

Tools and Workflows available on : https://proteomics.usegalaxy.eu/



Welcome to Galaxy Proteomics

This European Galaxy branches off from the main usegalaxy.eu server and focuses on one specific domain: Proteomics.

Galaxy Proteomics is a multiple-'omics' data-analysis platform with particular emphasis on mass spectrometry-based proteomics. It is hosted by GalaxyEU and maintained by the GalaxyP team from

@usegalaxyp is now part of the new #proteogenomics core for the @UMNCancer and will help researchers in integrating users' genomic and proteomic data or metagenomic data and metaproteomic data. #metaproteomics #proteogenomics #multiomics

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

Proteogenomics Gateway: z.umn.edu/proteogenomicsgateway

Training!

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