Mass Spectrometry-based Proteomics Data Analysis using Galaxy-P







The Galaxy-P Project



Extending Galaxy for MS-based proteomics data analysis applications



UNIVERSITY OF MINNESOTA SUPERCOMPUTING INSTITUTE







James Johnson (JJ)

Pratik Jagtap

Tim Griffin

Documentation: z.umn.edu/gcc2015gp

Documentation and Infrastructure Support Kevin Murray Ray Sajulga Tl

Thomas McGowan

2

Objectives for workshop

- Basics of data analysis for proteomics/proteogenomics
- Learn about tools available in Galaxy for proteomics
- Learn how to build some workflows and use tools



'Omic technologies and the molecular biology paradigm



High throughput sequencing technologies

Biological Mass Spectrometry

- Proteomics complements genomics/transcriptomics
- Direct information on molecular "effectors" of cell functions (enzymes, transporters etc.)
- Properties not predicted by genes/transcripts (PTMs, protein complexes etc.)

GalaxvP

Quick reminder: primary protein structure



⁽²⁰ naturally occurring)

http://www.scienceprofonline.com/chemistry/what-is-organic-chemistry-carbohydrates-proteins-lipids-nucleic-acids.html



Measuring masses of proteins and peptides



High throughput protein identification by MS

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



intact peptides

"MS2" or "MS/MS" spectrum records masses of peptide fragments



Information "currency" of MS: mass spectra



Common mass spectrometry terms

m/z = mass-to-charge = mass of analyte/number of retained charges

Mass accuracy (ppm) = [(measured m/z – actual m/z)/(actual m/z)] * 10⁶

- Mass accuracy is usually expressed in units of parts per million (ppm)
- Generally a lower mass accuracy in ppm is preferred

Mass resolution = $m/\Delta m$



Generally <u>higher</u> mass resolution is preferred



Tandem mass spectrometry and peptide sequence



• An MS/MS spectrum contains a mixture of b and y ions

High throughput identification of proteins



11

Different choices for sequence database searching



Experimental workflow in MS-based proteomics



Eng et al 2011 Mol Cell Proteomics. 10(11): R111.009522.



Data analysis modules for protein identification



Why use Galaxy for proteomics data analysis?



J Proteome Res. 2014 13:5898-908

15

GalaxvP

Proteomics tools accessible via the Tool Shed

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

🚆 Galaxy Tool Shed		Repositories Help+ User+			
3374 valid tools on Jul 04, 2015 Search	Repositories in Category	Proteomics			
Search for Valid Cools Search for workflows Valid Colour Utilities		Market	letadata	Tools or	0
<u>Tools</u>	<u>Name</u> ;	<u>synopsis</u> Type Ri	evisions	Verified	<u>owner</u>
<u>Custom datatypes</u>	appendfdr	Add false discovery rate to tabular data. Unrestricted 0	(2013-05-10)	no	galaxyp
<u>Repository dependency definitions</u> Tool dependency definitions	<u>blast plus remote blastp</u>	NCBI BLAST+ remote blastp Unrestricted 4	4 (2015-05-04) 🔻	no	<u>qalaxvp</u>
All Repositories	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
Browse by category Available Actions	dbbuilder	This tool allows users to download protein databases from common sources. Unrestricted	4 (2014-09-26) 🔻	no	galaxyp
Login to create a repository	<u>decovfasta</u>	Galaxy tool wrapper for the transproteomic pipeline decoyFASTA tool. Unrestricted	5 (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
	<u>acms lcms analysis</u>	GCMS and LCMS workflows Unrestricted 0	(2015-05-15)	n/a	proteomisc
	idpgonvert	Bumbershoot idpQonvert, a part of Bumbershoot IDPicker. Unrestricted 2	(2014-09-30)	no	galaxyp
	<u>ltg iquant cli</u>	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	<u>iracooke</u>
	mascot	Mascot MS/MS Search Unrestricted S	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
	<u>ms data converter</u>	AB SCIEX MS Data Converter Unrestricted 1	(2015-03-11)	no	<u>galaxyp</u>
	<u>msafplus</u>	MSGF+ Galaxy Wrapper Unrestricted 1	15 (2015-03-26) 🔻	no	<u>iracooke</u>
	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

https://github.com/galaxyproteomics/tools-galaxyp

Basics of proteogenomics



Peptide sequence variants and "proteoforms"



| VOL.10 NO.3 | MARCH 2013 | NATURE METHODS

What will this workshop cover?

- Basics of MS-based proteomic data analysis, with a showcasing of proteogenomic workflows in Galaxy.
- The first workflow module for merging search databases will be actually 'run' while others we will 'pretend run' to avoid slow down on the cloud.
- The different modules can be run as a complete workflow for the sake of a tutorial we will run modules separately.
- We have included documentation for each portion of this workshop for reference.



INPUTS : PEAKLISTS and SEARCH db₂₀



July 8th 2015: Session 7 : 2:05 PM: Extending Galaxy's reach: recent progress towards complete multi-omic data analysis workflows. Timothy J Griffin

GalaxvP

21

INPUTS : PEAKLISTS and SEARCH db₂







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

Galaxy

23



Getting Started GCC2015 GalaxyP Tutorial

Open a web browser and navigate to the GalaxyP cloud instance – http://54.205.17.20/

At the top of the screen select User and Register and Log in for your Cloud username (your email) and password.

2.2 Getting Started GCC2015 GalaxyP Tutorial

A. Open a web browser and navigate to the GalaxyP cloud instance - http://54.205.17.20/B. At the top of the screen select User and Register and Log in for your Cloud username (your email) and password.

C. At the top of the screen select **Shared Data** then migrate to **Published Histories**. D. Select **History 1** from the list of published histories.

E. Select Import history to add the selected history to your user histories.

F. On the confirmation screen select **start using this history** to navigate to this history in the Galaxy view.

G. At the top of the screen select Shared Data then migrate to **Published Workflows** .

H. Select Workflow 1A: Workflow for History 1 from the list of published workflows and choose import to copy the workflow into your user workflows.I. On the confirmation screen, select start using this workflow to navigate to your user

workflows.

J. In the workflows menu select **Run** Workflow 1A: Workflow for History 1 from the drop down menu.

K. Appropriately assign each input database from History 1 to the corresponding input or the workflow.



PROTEOMICS WORKFLOW





INPUTS : PEAKLISTS and SEARCH db

INPUTS : MASS SPECTRAL DATA AND SEARCH DATABASE.



The dataset will be searched against FASTA database with human proteins, contaminant proteins, spiked in proteins and a subset of 3frame translated cDNA database from EnSEMBL.

INPUTS: a) MGF formatter MGF files. (dataset collection) b) ABRF-Spike4: FASTA sequences of 4 spiked in proteins. c) FASTA File from EnSEMBL Searches: Subset of 3-frame translated cDNA database from EnSEMBL (our template for identifying novel proteoforms). d) Human UniProt FASTA file + contaminant proteins.



SEARCH DATABASES



INPUTS : PEAKLISTS and SEARCH db

PROTEOMIC DATABASES

UniProt

Swiss-Prot is the <u>manually annotated</u> and reviewed section of the UniProt Knowledgebase (UniProtKB). It is a high quality annotated and nonredundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.

http://en.wikipedia.org/wiki/Swiss-Prot

TrEMBL contains high-quality computationally analyzed records, which are enriched with automatic annotation.

The translations of annotated coding sequences in the EMBL-Bank/GenBank/ DDBJ nucleotide sequence database are automatically processed and entered in TrEMBL.

http://en.wikipedia.org/wiki/TrEMBL



INPUTS : PEAKLISTS and SEARCH db

CUSTOMIZED PROTEOMIC DATABASES



LOOKING BEYOND THE KNOWN PROTEOME



WORFLOW 1A



INPUTS : PEAKLISTS and SEARCH db

PROTEOMICS WORKFLOW





INPUTS : PEAKLISTS and SEARCH db

INPUTS : MASS SPECTRAL DATA AND SEARCH DATABASE.



The dataset will be searched against FASTA database with human proteins, contaminant proteins, spiked in proteins and a subset of 3frame translated cDNA database from EnSEMBL.

INPUTS: a) MGF formatter MGF files. (dataset collection) b) ABRF-Spike4: FASTA sequences of 4 spiked in proteins. c) FASTA File from EnSEMBL Searches: Subset of 3-frame translated cDNA database from EnSEMBL (our template for identifying novel proteoforms). d) Human UniProt FASTA file + contaminant proteins.







Eng et al 2011 Mol Cell Proteomics. 10(11): R111.009522.



INPUTS : PEAKLISTS and SEARCH db
RAW DATA CONVERSION TOOL



INPUTS : PEAKLISTS and SEARCH db

Target -Decoy database Search GUI P 1B

INPUTS : PEAKLISTS and SEARCH db SEARCHGUI

SEARCHGUI : SEARCH ALGORITHMS

INPUTS : PEAKLISTS and SEARCH db_®



INPUTS : PEAKLISTS and SEARCH db SEARCHGUI

GalaxyP

3.2 SearchGUI in GalaxyP

- From Published Workflows, import Workflow 1B: Workflow for History 1 to History 2.
- On the confirmation screen, select start using this workflow to navigate to your user workflows.
- Run Workflow 1B on your current history (History 1).
- Appropriately select each dataset from History 1 that corresponds to input within the workflow.
- In Step 3: SearchGUI some parameters must be set at run time:
- Choose to run X!Tandem, MyriMatch, MS-GF+, OMSSA, and Comet
- Change the Fragment Tolerance to 0.01 Daltons
- Select the box Send results to a new history named: and enter History 2 into the field provided.
- Select Run and navigate to History 2 in the normal Galaxy view. For the purposes of this tutorial, a completed History 2 may be imported from Published Histories.

3.2 SearchGUI in GalaxyP





PROTEOMICS WORKFLOW



GalaxyP

DATABASE SEARCH



43

DATABASE SEARCH



Nesvizhskii et al Nature Methods - 4, 787 - 797 (2007)

DATABASE SEARCH



Nesvizhskii *et al Nature Methods* - **4**, 787 - 797 (2007)

SEARCHGUI

SearchGUI 1.26.2				
File Edit Help				
Input & Output				
Spectrum File(s)		1	file(s) selected	Add Clear
Search Settings	tutorial.parameters Edit			Edit Load
Output Folder			C:\Users\Public	Browse
Search Engines				
V	x! tandem	<i>l</i> # € ∆	X!Tandem Search Algorithm - XITandem web page	o
V	MyriMatch	AU 🛆	MyriMatch Search Algorithm - MyriMatch web page	0
V	MS Amanda	Az 🛎 🛆	MS Amanda Search Algorithm - <u>MS Amanda web page</u>	0
V	MS-GF+	<i>1</i> 1 € ∆	MS-GF+ Search Algorithm - <u>MS-GF+ web page</u>	0
	OMSSA	<i>4</i> 7 € ∆	OMSSA Search Algorithm - OMSSA web page	0
V	Comet	AU &	Comet Search Algorithm - Comet web page	0
V	Tide	<i>l</i> # € ∆	Tide Search Algorithm - <u>Tide web page</u>	0
Post Processing				
Ø	THE D	47 € ∆	PeptideShaker - <u>Visualize the results in PeptideShaker</u>	o
	Please cite SearchG	UI as <u>Vaudel et al.:</u>	Proteomics 2011;11(5):996-9.	Start the Search!

Vaudel M. et al Proteomics (2011) 11(5)

https://code.google.com/p/searchgui/



MULTIPLE SEARCH ALGORITHMS



MULTIPLE SEARCH ALGORITHMS





Visualizing parameters for SearchGUI analysis

	1113101	
Merged and Filtered FASTA file 14: Merged and Filtered FASTA from data 12, data 13, and others type to filter	search datasets imported: History 2	0
Step 2: Input dataset collection MGF files LTQ/ Orbitrap acquired dataset.	384.3 MB	•
Mascot formatted MGF Files: Input Dataset Collection (MGF Files)	15: Search GUI on data 3, data 4, and data 14 179.4 MB format: searchgui_archive ?	database:
Step 3: Search GUI (version 1.28.0)	Creating decoy database.	
Protein Database Output dataset 'output' from step 1 Create a concatenated target/decoy database before running PeptideShaker True 🕫	Reindexing: input_databas 10% 20% 30% 40% 50% 60% 90%Input: /mnt/galaxy/tmp/job_wor	e.fasta. 6 70% 80% king_directc
Input Peak Lists (mgf) Output dataset 'output' from step 2	Name: input_database	
DB-Search Engines ✓ X!Tandem	Decoy Tag: null Type: UniProt Last modified:	
MS Amanda	DAC	
✓ MS_GF+		• •
	Compressed binary file	
Comet	14: Merged and Filtered	@ # ¥
🗋 Tide	FASTA from data 12, dat	
Precursor Ion Tolerance Units	a 13, and others	
Parts per million (ppm) 🕜	13: Regex Find And Repl	
Percursor Ion Tolerance	ace on data 2	
10.0	12: Protein Database	• / ×
Fragment Tolerance (Daltons)	11: Protein Database	• / ×
	8: Regex Find And Repla	• / ×

Visualizing parameters for SearchGUI analysis



INPUTS SEARCHGUI PEPTIDESHAKER BLAST-P PSM Visualization





INPUTS : PEAKLISTS and SEARCH db SEARCHGUI PEPTIDESHAKER

PEPTIDESHAKER

SEARCHGUI : SEARCH ALGORITHMS

INPUTS : PEAKLISTS and SEARCH db₁

PEPTIDESHAKER



Vaudel et al Nature Biotechnology, 33, (2015)

http://galaxyproteomics.github.io/peptideshaker/

GalaxyP

52

High throughput identification of proteins



53

PEPTIDESHAKER : PROTEIN INFERENCE Shotgun Protein Identification



Slide from Alexey Nesvizshkii talk at http://www.scivee.tv/node/12671

4.3 Peptide Shaker in GalaxyP

A. At the top of the screen, click Shared Data then select the option: Published Workflows .

B. Click 'Workflow2: Workflow for History 2 to History 3' and choose import.

C. Select start using this workflow.

D. Click imported Workflow 2: Workflow for History 2 to History 3' and choose edit to

open the workflow editor.

E. To view the whole workflow (two tools*), navigate to the right and down by clicking and dragging the grid...

- i. ...or use the minimap in the lower right corner.
- ii. Note: Galaxy workflows are usually more complex with many more tools and Connections; for the sake of this tutorial we will start simple.
- F. Click on the PeptideShaker tool to view its details in the right panel.
- G. To run, click the cogwheel at the top right corner of the middle frame and select run from the resulting menu.
- a. The input for this workflow under SearchGUI Results should be 8: SearchGUI On data 7 data2 and data 1 from History 3. If this is not the case then refer to Section 2.2 on how to import a complete history.

4.3 Peptide Shaker in GalaxyP



4.3 Peptide Shaker in GalaxyP

	Q	Details
		Tool: Peptide Shaker
		Version: 0.40.0
		Compressed SearchGUI results Data input 'searchgui_input' (searchgui_archive) The species type to use for the
		No species restriction \$
		Specify Advanced PeptideShaker Processing Options:
Input dataset 🗙		Advanced Processing Option \$
output		FDR at the protein level: V
λ		1.0
		FDR at the peptide level: V
		1.0
	Peptide Shaker 🗙	FDR at the PSM level: V
	Compressed SearchGUI results	1.0
	mzidentML (mzid) output_cps (peptideshaker_archive)	Minimum confidence required for a protein in the fraction MW plot:
	output_zip (zip) 🔹 🔉	95.0
	output_certificate (txt)	The PTM probabilistic score to us for PTM localization:
	output nem phosphorylation (tabular)	A-score
	output_psm (tabular)	Specify Advanced Filtering Options:
	output_peptides_phosphorylation	Advanced Filtering Options 🗘
	output poptider (tabular)	Minimum Peptide Length: 🔻
	output proteins phosphon/ation	6
	(tabular)	Maximum Peptide Length: V
	output_proteins (tabular) 🔹 📀	40
		Maximum Precursor Error: V
		10.0
		Maximum Precursor Error Type:
		ppm \$

PEPTIDESHAKER : TARGET-DECOY SEARCH REVERSE DATABASE SEARCH



>IPI:IPI00205563.1|Gene_Symbol=Tmsbl1 thymosin beta-like protein MSDKPDLSEVETFDKSKLKKTNTEEKNTLPSKETIQQEKEYNQRS

>IPI:REV_IPI00205563.1|Gene_Symbol=Tmsbl1 thymosin beta-like protein SRQNYKEEQQITKESPLTKNEETNKKTKLKSDFTEVESLDKPDSM

GalaxvP

58

PEPTIDESHAKER : TARGET-DECOY SEARCH

FALSE DISCOVERY RATE ANALYSIS



59

PEPTIDESHAKER: OUTPUTS



PEPTIDESHAKER: OUTPUTS

GCC5: PeptideShaker Outputs 7 shown	
359.1 MB	2 9 9
7: Peptide Shaker on data 8: Protein Report	• / ×
6: Peptide Shaker on data 8: Peptide Report	• / ×
4,372 lines format: tabular , database: <u>?</u>	
Path configuration completed. Thu Jun 11 16:50:55 CDT 2015 Unzipping searchgui_in 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%	put.zip.
Thu Jun 11 16:51:26 CDT 2015 Import process for Galaxy_Experiment_2015061116501434059408 (Samp Sample_20150611165014340	le:
8 0 2 Lul	۲
1 2	
Protein(s) 1 H0YMR4; H0YN07; Q8TEX9; Q8TEX9-2 2 A0A0A0MSD5; A0A0A0MSS8; A0A0A0MT30; B4 (16) 2 2 100.0 0 3 ENST00000291565_14; F2Z2Y4; 000764; 000764-2 4 A0A0A0MSM0; Q92598; Q92598-2; Q92598-3; Q925	4DK69; HØY Confide 2; 000764-: 598-4
5: Peptide Shaker on data 8: PSM Report	• / ×
4: Peptide Shaker on data 8: Parameters	• / ×
3: Peptide Shaker on data 8: Archive	• / ×
2: Peptide Shaker on data 8: mzidentML file	• / ×
1: Search GUI on data 7, data 2, and data 1	• / ×

PEPTIDESHAKER: OUTPUTS



62

INPUTS : PEAKLISTS and SEARCH db₃

SEARCHGUI : SEARCH ALGORITHMS

PEPTIDESHAKER





5.2 BLAST-P Search

a. At the top of the screen, click Shared Data then select the option: Published Workflows .

b. Click Workflow 3: Workflow for History 3 to History 4 and choose import.

c. Select start using this workflow .

d. To run, click 'Workflow 3: Workflow from History 3 to History 4 and choose run .

i. Change your input file to 6: Peptide Shaker on data 8: Peptide Report.

ii. Notice the number of steps in the workflow (refer to Section 5.3 for more detail)

iii. Check Send results to a new history at the bottom and name it History 4.



5.2 BLAST-P Search



GalaxyP

INPUTS SEARCHGUI PEPTIDESHAKER BLAST-P

BLAST-P SEARCH

Ero m		From		
				Tool: NCBI BLAST+ remote blast
out_file1 (tabul	liar)	out_file1 (tabular)		Version: 1.0
	1			Protein guery sequence(s)
		Tabular-to-FASTA X		Data input 'query' (fasta)
		Tab-delimited file		Subject database/sequences:
	/	output (fasta)		NCBI Remote Database 💲
	// /			Protein BLAST database: 🔻
	//	Compute sequence length X		Non-redundant protein seq 🛊
		Compute length for these	Count ×	Search Organism Restrictions:
		outout (tabular)	from dataset	Search Organism Restrictio
	//	and an (manual)	out_file1 (tabular) 🔹 😥	
				NCBI Taxon ID: V
	//			9000
		Tab-delimited file		Exclude this NCBI Taxon II
	//	output (fasta)		
				Ramova Search Organism
Filters	sequences by length		Filter sequences by length x	Remove Search Organism Restriction 1
Filter s	sequences by length ×		Filter sequences by length 🗙	Remove Search Organism Restriction 1
Filter s	sequences by length X file		Filter sequences by length x Fasta file output (fasta)	Remove Search Organism Restriction 1 Add new Search Organism
Filter s Fasta f output	sequences by length × file	,	Filter sequences by length × Fasta file output (fasta)	Remove Search Organism Restriction 1 Add new Search Organism Restriction
Filter s Fasta f output Reg	equences by length × file t (fasta) 0 0	,	Filter sequences by length × Fasta file output (fasta) Regex Find And Replace ×	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V
Filter s Fasta f output Reg	sequences by length × file t (fasta) 0 0 ex Find And Replace × ect lines from	,	Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp
Filter s Fasta f output Reg Sele	sequences by length × file t (fasta)		Filter sequences by length X Fasta file output (fasta) Regex Find And Replace X Select lines from out_file1	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: ♥
Filter s Fasta f output Reg Sele out	equences by length × file t (fasta)		Filter sequences by length X Fasta file output (fasta) Regex Find And Replace X Select lines from out_file1 NCBI BLAST+ remote blastp X	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V © blastp © blastp © blastp-short Set expectation value cutoff: V 200000.0
Filter s Fasta f output Reg Sele out_ NCI	eequences by length × file t (fasta)		Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from out_file1 NCBI BLAST+ remote blastp × Protein guery sequence(s)	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp blastp-short Set expectation value cutoff: V 20000.0
Filter s Fasta f output Reg Sele out, NCI	sequences by length × file t (fasta) tex Find And Replace × ect lines from file1 BI BLAST+ remote blastp × otein query sequence(s)		Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from out_file1 NCBI BLAST+ remote blastp × Protein query sequence(s) output_tabular (fabular)	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp blastp-short Set expectation value cutoff: V 20000.0 Output format: BLAST XML
Filter s Fasta f output Reg Sele out NCI	sequences by length × file t (fasta) ext Find And Replace × ext lines from file 1 BI BLAST+ remote blastp × otein query sequence(s) tput_tabular (tabular)		Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from out_file1 NCBI BLAST+ remote blastp × Protein query sequence(s) output_tabular (tabular)	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp blastp-short Set expectation value cutoff: V 20000.0 Output format: BLAST XML \$
Filter s Fasta f output Reg. Sele out NCI Pro out out	sequences by length × file t (fasta) • ex Find And Replace × ect lines from file1 • BI BLAST+ remote blastp × otein query sequence(s) tput_tabular (tabular) • tput_xml (blastxml)		Filter sequences by length X Fasta file output (fasta) Regex Find And Replace X Select lines from out_file1 NCBI BLAST+ remote blastp X Protein query sequence(s) output_tabular (tabular) output_xml (blas)	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V © blastp © blastp © blastp-short Set expectation value cutoff: V 200000.0 Output format: BLAST XML \$ None: V
Filter s Fasta f output Reg Sele out_ NCI Pro out out out	sequences by length × file t (fasta) ex Find And Replace × ect lines from file1 BI BLAST+ remote blastp × stein query sequence(s) tput_tabular (tabular) tput_xml (blastxml) tput_txt (txt)		Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from out_file1 NCBI BLAST+ remote blastp × Protein query sequence(s) output_tabular (tabular) output_tabular (tabular) output_tabular (tabular)	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp-short Set expectation value cutoff: V 200000.0 Output format: BLAST XML \$ None: V Maximum hits to show: V
Filter s Fasta f output Sele out, NCC Pro out out out out	eequences by length × file t (fasta) ex Find And Replace × ect lines from file1 BI BLAST+ remote blastp × otein query sequence(s) tput_tabular (tabular) tput_tabular (tabular) tput_txt (txt) tput_ttml (html)		Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from out_file1 NCBI BLAST+ remote blastp × Protein query sequence(s) output_tabular (tabular) output_tabular (tabular) output_txt (txt) output_html (htm	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp blastp-short Set expectation value cutoff: V 20000.0 Output format: BLAST XML \$ None: V Maximum hits to show: V 1
Filter s Fasta f output Reg Sele out, NCI Pro out out out out	sequences by length × file t (fasta) ex Find And Replace × ect lines from file1 BI BLAST+ remote blastp × otein query sequence(s) tput_tabular (tabular) tput_tabular (tabular) tput_txt (txt) tput_txt (txt) tput_html (html)		Filter sequences by length × Fasta file output (fasta) Regex Find And Replace × Select lines from out_file1 NCBI BLAST+ remote blastp × Protein query sequence(s) output_tabular (tabular) output_tabular (tabular) output_txtl (tabular) output_txtl (blas output_txtl (blas output_txtl (blas	Remove Search Organism Restriction 1 Add new Search Organism Restriction Type of BLAST: V blastp blastp blastp-short Set expectation value cutoff: V 20000.0 Output format: BLAST XML \$ None: V Maximum hits to show: V 1 Advanced Options:



INPUTS SEARCHGUI PEPTIDESHAKER BLAST-P

INPUTS : PEAKLISTS and SEARCH db₈

SEARCHGUI : SEARCH ALGORITHMS

PEPTIDESHAKER





6.2 Generating a sqlite Database

a. From Shared Data at the top of the screen, import History 5 and Workflow 4: History 4 to History 5.

b. From the workflows menu , run Workflow 4: History 4 to History 5 on History 4.

c. In the workflow run options, check that all inputs are correct and select the box Send results to a new history named: and enter History 5 into the field provided.

d. Select Run and navigate to History 5 in the normal Galaxy view. For the purposes of this tutorial, a completed History 5 may be imported from Published Histories.





6.3 PSM Evaluator in GalaxyP



INPUTS SEARCHGUI PEPTIDESHAKER BLAST-P PSM Visualization

GalaxyP

71

6.3 PSM Evaluator in GalaxyP

a. From Shared Data at the top of the screen, import History 5b. Click on mz_to_sqlite dataset to expand, select Visualize in PSM Viewer .

c. Select Peptide View from PSM Table at the top of the screen.

d. To view the spectra of the novel peptides input the peptide sequences (LLSVGGLR, SPVLKPSR) into the filter peptides by sequence(s) field and select the sequence to generate the spectral data.

e. Within the viewer users can utilize numerous organization tools to sort through data of each peptide.

- i. Users may select which columns they prefer to see in their PSM Tables.
- ii. Users may search the table for a specific piece of data.

iii. Users may sort data in each column and organize the order in which columns appear.

f. To view the spectra of LLSVGGLR select anywhere with the peptide spectral data table.

Galaxv


PSM EVALUATION

Mascot:identity _ threshold	Mascot:score	Scoffold:Peptide Probability	acquisitionNum	msLevel	polarity	peaksCount	sequence	precursorMZ	precursorCharge	totIonCurr
39.525017	33.52	0.95	26957	2		300	AQFEGIVTDLIRR	506.72560001939905	3	
40.916668	41.64	0.95	87885	2		300	SQVFSTAADGQTQVEIK	904.7714000290986	2	
40.918777	29.78	0.8760495	32450	2		300	MKETAENYLGHTAK	531.4867000193991	3	
41.337624	89.47	0.95	86558	2		300	STNGDTFLGGEDFDQALLR	1028.3580000290983	2	
41.48911	70.94	0.95	63125	2		300	AQFEGIVTDLIR	682.0270000290985	2	
41.723988	105.71	0.95	93899	2		300	STNGDTFLGGEDFDQALLR	1028.8940000290984	2	
41.81472	29.33	0.69156307	11016	2		300	LVGMPAKR	436.27940002909855	2	
41.81472	27.29	0.5512762	11006	2		300	LVGMPAKR	436.27380002909854	2	
	·									

Page 6

of 12313 total records.

STNGDTFLGGEDFDQALLR 86558

of 247

Showing records



July 7th 2015: Session II : 11:20 AM : Proteomics Visualization in Galaxy

James E. Johnson

INPUTS SEARCHGUI PEPTIDESHAKER BLAST-P PSM Visualization

PSM EVALUATION

- Go to 'Peptide' View.
- Type in 'LLSVGGLR, SPVLKPSR' in Filter box. OR

Copy the peptide sequences from item #52 in history 5.

• Follow the instructor for demo.

- Go to 'Protein' View.
- Type in 'ABRF' in Filter box.
- Follow the instructor for demo.

July 7th 2015: Session II : 11:20 AM : Proteomics Visualization in Galaxy James E. Johnson



INPUTS : PEAKLISTS and SEARCH db₅

7. Mapping Peptides to a Genome

- Map the peptide to the SearchDB protein
- Find the mapping of the protein to the reference genome. An Ensembl GTF file shows how the exons of the protein sequence are mapped to the genome sequence
- Each SearchDB construction may need its own mapping to genome method.



7. Mapping Peptides to a Genome

7.2 Generating a GFF file

a. From Shared Data at the top of the screen, import Input for History 6 and Workflow 5 : Workflow to History 6 from Published Histories and Workflows, respectively.

b. From the workflow menu, Run Workflow 5 on Inputs for History 6.

- c. Select the appropriate inputs for Workflow 5 from Inputs to History 6.
- d. Check to send the results to a new history named History 6 and Run the workflow.

For the purposes of this tutorial, a completed History 6 may be imported from Published Histories.

BLAST-P

PSM Viz

7.3 From GFF to IGV Browser

SEARCHGUI PEPTIDESHAKER

INPUTS





7.2 Generating a GFF File.



INPUTS SEARCHGUI PEPTIDESHAKER BLAST-P PSM Visualization

GalaxyP



INPUTS

SEARCHGUI PEPTIDESHAKER BLAST-P

PSM Viz



INPUTS



Genome Visualization

iiii vim v java v py v ru	by v GWT v Galaxy v MSI v docker v News v DB v bio v prot v html v Mothur v RickRack v reg v Norsk v Apple v Hea	IlthCare V UniSci V 2008 V FMS CFS V
Ga	laxy / GalaxyP https://galaxyp.msi.umn.edu/datasets/20e4f86aacb54c7f/show_params	Untitled -
🧧 Galaxy / Gala	KYP Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Using 183.9 GB
Fools	Peptides to GTF for genome visualization"	History 2 🌣 🗆
search tools	Step 1: Input dataset	search datasets
CORE TOOLS	Peptide Report for Novel Proteoform Peptides (Potential)	novel proteoforms
Get Data	T: bLAS1-P_rintered_Peptide_Report.tabular	200 have
iend Data		200 bytes
.ift-Over	Step 2: Input dataset	3: Homo_sapiens.GRC @ * *
Fext Manipulation		h <u>37_canon.73.gtt</u>
ilter and Sort	2: Homo sapiens GBCb37.73.cdna.all.fa	~2,000,000 lines
oin, Subtract and Group	type to filter	ionnali gri, database ng 15
Convert Formats		uploaded gtf file
Extract Features	Step 3: Input dataset	B 0 2 III 🛛 🔊 🗩
<u>statistics</u>		display in IGB View
Graph/Display Data	3: Homo sapiens.GRCh37 canon.73.otf	display with IGV web current local
ASTA manipulation	type to filter	display at Ensembl <u>Current</u>
PROTEOMICS		uispiay at ocse main
MS Data Conversion	Step 4: Column Regex Find And Replace (version 0.1.0)	1. Seqname 2. Source 3
equence Database Tools		1 processed_transcript e
NGS: QC and manipulation	Step 5: Peptide to GFF (version 1.0)	1 processed_transcript e
Protein/Peptide Search	Peptide Source Format	1 unprocessed_pseudogene ε
Data Conversion Tools	Generic Tabular (with peptide and accession columns)	1 unprocessed_pseudogene e
/isualizers	Source File	1 unprocessed_pseudogene e
Duantification	Output dataset 'out_file1' from step 4	
LAST-P	Peptide Column	2: Homo_sapiens.GRC
Proteogenomics	4	n37.73.cona.all.ta
	Accession Identifier Column	1: BLAST-P_Filtered_P
SENOMICS		eptide_Report.tabular

Genome Visualization



Genome Visualization



INPUTS

iiii vim v java v py v ruby v	T 🗸 Galaxy 🖌 MSI 🗸 docker 🖌 News 🗸	DB v bio v prot v htm	ml 🛩 Mothur 🛩 RickRack 🛩 reg 🛩 Norsk 🛩 Apple 🛩 H	lealthCare - UniSci - 2008 - FMS CFS - >
Galaxy	yP https://g	galaxyp.msi.umn.edu/datasets/	/20e4f86aacb54c7f/show_params	Untitled -
🗧 Galaxy / Galaxy	Analyze Data Wor	rkflow Shared Data + V	Visualization - Admin Help - User -	Using 183.9 GB
Fools	Seqid Source Type Start	End Score Stra	andPhaseAttributes	History C 🔅 🗆
search tools	##gff-version 3			Er nontidor off?
search cools	##sequence-region 4 1 155412208			<u>5. peptides.girs</u>
ORE TOOLS	4 MassSpec peptide 155412185	155412208 10.0 -	0 ID=LLSVGGLR	4 lines, 3 comments
Jet Data	4 MassSpec CDS 155412185	155412208	0 Parent=LLSVGGLR;transcript_id=ENS1000003394	452 format: gff3, database: hg19
end Data	##sequence-region 11 1 128838509	128828560 10.0		Mapped 2 entries
ft-Over	11 MassSpec (DS 128838546	128838569 -	0 Parent=SPVI KPSR transcript id=ENST000003103	
ext Manipulation	11 Massiper ebs 120050540	120030303		
ilter and Sort				display with IGV web current local
oin, Subtract and Group				display at UCSC <u>main</u>
Convert Formats				1.Seqid 2.Source 3.Type 4.Star
xtract Features				##gff-version 3
itatistics				##sequence-region 4 1 155412208
Graph/Display Data				4 MassSpec peptide 1554121
ASTA manipulation				4 MassSpec CDS 1554121
POTEOMICS				##sequence-region 11 1 128838569
AS Data Conversion				11 MassSpec peptide 1288385
Sequence Database Tools				
ICS: OC and manipulation				3: Homo_sapiens.GRC
Protein /Pentide Search				
Algorithms				format: atf . database: ha19
ata Conversion Tools				
isualizers				uploaded gtf file
Quantification				🖺 0 2 📖 📎 🗩
LAST-P				display in IGB View
Proteogenomics				display with IGV web current local
ENOMICS				display at Ensembl Current
IENOMICS				display at UCSC main

File Genomes View	ew Tracks Regions	Tools GenomeSpace	Help	IGV			
Human hg19	¢ Chr4	¢ chr4		Go 👚 🔺	▶ 🏟 🖪 X 📮 I		
	p16.2 p15.33	p15.2 p15.1 p14 p12	q12 q13.1 q13. 3	3 q21.21 q22.1 190	q23 q25 q26 q28.1 q28.3 q3	i1.21 q31.3 q32.2	q33 q34.3 q35.
	nb 20 mb	, 40 mb	60 mb	80 mb	100 mb 120 mb 140 mb 	160 mb 	180 mb
RefSeq Genes peptides.gff3	595 CPZ BST1 M	IIR573 ARAP2 YIPF7 S	SGCB LPHN3 HTN3	FRAS1 AFF1 UN	C5C CXXC4 TIFA PRDM5 PCDH10 IL1	5 DCLK2 FNIP2 ANX LLSVGGLR	A10 NEIL3 IRF2
4 tracks chr	r4:102.623,155						
	java v py v ruby v	∎ galaxyp.msi.umn.edu/root?worki GWT ~ Galaxy ~ MSi ~ docke	now_ia=ua57a5e9c811b11b 3r × News × DB × bio ×	v prot v html v Mot	Lu' V hur v RickRack v reg v Norsk v Apple v HealthC	iare v UniSci v 2008 v FM	4,462M of 5,797M
	i v java v py v ruby v i Galaxy / Ga	∎ galaxyp.msr.umn.edu/root?work GWT ← Galaxy ← MSI ← docke ilaxyP	now_id=0d57d5e9c6f1b11b er × News × DB × bio × https://galaxyp.msi.umi	v prot v html v Mot	Lu View Versen verse ver	iare - UniSci - 2008 - FM Untitled	4,462M of 5,797M
III vim	java py ruby Galaxy/Gi alaxy / GalaxyP	galaxyp:mst.umn.eou/root/work GWT ~ Galaxy ~ MSI ~ docke alaxyP Analy	now_ld=udb/dbescentb11b er × News × DB × bio × https://galaxyp.msi.umi yze Data Workflow Share	v prot v html v Mot nn.edu/datasets/20e4f86aa red Data v Visualizatio	hur v RickRack v reg v Norsk v Apple v HealthC cb54c7f/show_params on v Admin Help v User v	iare Y UniSci Y 2008 Y FM Untitled	4,462M of 5,797M IS CFS ~ >> + sing 183.9 CB
	i java py ruby Galaxy/Galaxy/GalaxyP	Galaxyp.msi.umn.eou/root /work GWT ~ Galaxy ~ MSI ~ docki alaxyP Analy Seqid Source Type	arow_lo=udo/doescaribiiib er × News × DB × bio × https://galaxyp.msi.um yze Data Workflow Sharo Start End	r prot v html v Mot in.edu/datasets/20e4f86aa red Data v Visualizatio Score StrandPhase	hur v RickRack v reg v Norsk v Apple v HealthC cb54c7f/show_params on v Admin Help v User v Attributes	are Y UniSci Y 2008 Y FM Untitled U: History	4,462M of 5,797M IS CFS
Tools Search t CORE TOO Get Data Send Data	iv java v pyv rubyv r Galaxy / Galaxy / G ilaxy / Galaxy P tools	GWT ~ Galaxy ~ MSI ~ dock alaxyP Seqid Source Type ##gff-version 3 ##sequence-region 4 1 1554 4 MassSpec peptide 4 MassSpec CDS #tenuence-region 11 1 128	How_ld=udb/dbeston FDT FD er × News × DB × bio × https://galaxyp.msi.um yze Data Workflow Start End 412208 155412185 155412208 155412185 155412208 185838569 138838569	v prot v html v Mot in.edu/datasets/20e4f86aa red Data v Visualizatie Score StrandPhase 8 10.0 - 0 8 0 9 . 10.0 - 0	hur v RickRack v reg v Norsk v Apple v HealthC cb54c7f/show_params on v Admin Help Vser v Attributes ID=LLSVGGLR Parent=LLSVGGLR;transcript_id=ENST00000339452 ID=SB4 krssp	are V UniSci V 2008 V FM Untitled U: History 5: peptides.gff3 4 lines, 3 comments format: gff3, database: H Mapped 2 entries	4,462M of 5,797M IS CFS ~ >> + sing 183.9 CB C & P & sing 183.9 CB C & C & Sing 183.9 CB C & Sing 183.9 CB C & Sing 183.9 CB C & Sing 183.9 CB C & Sing 183.9 CB Sing 183.

INPUTS

File Genomes View	IGV Tracks Regions Tools GenomeSpace Help	
Human hg19		
	p16.2 p15.33 p15.2 p15.1 p14 p12 q12 q13.1 q13.3 q21.21 q22.1 q23 q25 q26 q28.1 q28.3 q31.	21 q31.3 q32.2 q33 q34.3 q35.
	nb 20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb 	160 mb 180 mb
RefSeq Genes peptides.gff3	595 CPZ BST1 MIR573 ARAP2 YIPF7 SGCB LPHN3 HTN3 FRAS1 AFF1 UNC5C CXXC4 TIFA PRDM5 PCDH10 IL15	DCLK2 FNIP2 ANXA10 NEIL3 IRF2
4 tracks chr4	102.623,155	4,462M of 5,797M
Statistics Graph/Disj FASTA mar PROTEOMI MS Data Cd Sequence I NGS: QC ar Protein/Pe Algorithms Data Conver Visualizers Quantificat BLAST-P Proteogene GENOMICS	May Data ipulation CS inversion batabase Tools d manipulation stide Search irsion Tools ion mics	<pre>##sequence-region 4 1 155412288 4 MassSpec peptide 1554121 4 MassSpec CDS 1554121 ##sequence-region 11 1 128838569 11 MassSpec peptide 1288385 3: Homo_sapiens.GRC</pre>

SEARCHGUI PEPTIDESHAKER BLAST-P

INPUTS

Genome Visualization

PSM Viz

● ● ● File Genomes View Tr	IGV acks Regions Tools GenomeSpace Help	
Human hg19 🛟	chr4 Co [⊕] [⊕] [⊕] [□] [×] [□] [×] [□] [×] [□] [×] [□] [×] [×] [□] [×]	-
p		1.21 q31.3 q32.2 q33 q34.3 q35.
nb	- 190 mb	160 mb 180 mb
RefSeq Genes	CPZ BST1 MIR573 ARAP2 YIPF7 SGCB LPHN3 HTN3 FRAS1 AFF1 UNC5C CXXC4 TIFA PRDM5 PCDH10 IL1	Image: Second state Image: Second state
Homo_sapiens.GRCh37_ca 73.gtf D059	6885 ENST0000503823 ENST00000319592 ENST00000401208 ENST0000264399 ENST00000504592 ENST00000264499 ENST0000	LLSVGGLR 00513000 ENST00000512831 ENST000005008
4 tracks chr4:94,665	aon iii galaxyp.msi.umn.edu/root?workflow_id=0d57d5e9c8f1b11b C III D V	4,523M of 5,797M
iiii vim v java	yy ruby GWT Galaxy MSI docker News DB bio prot html Mothur RickRack reg Norsk Apple HealthCo Galaxy Galaxy Galaxy https://galaxyo.msi.umn.edu/datasets/20e4/86aacb54c7f/show_params	are ~ UniSci ~ 2008 ~ FMS CFS ~ >> Untitled +
- Galaxy	/ GalaxyP Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Using 183.9 GB
Tools	Seqid Source Type Start End Score Strand Phase Attributes	History 📿 🌣 🖽
Search tools	##gff-version 3 ##sequence-region 4 1 155412208 4 MassSpec peptide 155412208 10.0 - 0 ID=LLSVGGLR 4 MassSpec CDS 155412185 155412208 - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452	5: peptides.gff3 ③ 2 × 4 lines, 3 comments format: gff3, database: hg19
Get Data Send Data	##sequence-region 11 1 128838569	Mapped 2 entries
Lift-Over Text Manipulation Filter and Sort Join, Subtract and Convert Formats	Group Interspect ppplate 120030310 120030300 1000 0 100-artExtSR Interspect Dpplate 128838546 128838569 - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343	display with IGV web current local display at Ensembl <u>Current</u> display at UCSC <u>main</u>

INPUTS

Eile Canomas Via	w Tracks Pagions	Tools Canomasa	co Holo	4	GV			
Human hg19	chr4	toois Genomespa	54 813 641-15	971 141 Go	4	• 🕅 🖬 ¥ 🥅 I		
			54,015,041 15.	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			·	
	p16.2 p15.33	p15.2 p15.1 p14 p1	2 q12 q13	.1 q13.3 q21.	21 q22.1	q23 q25 q26 q28.1	q28.3 q31.21 q31.3 q	32.2 q33 q34.3 q35.1
	kb I	155,000 kb 	155,2 	00 kb	1,1 1	55 kb	00 kb 155, 	800 kb
RefSeq Genes			₩₩	<	DCHS2	DCHS2 FGB FGG	<mark>∦ →<mark> </mark>→ LRAT RBM46</mark>	
peptides.gff3						L SVGGL R		
Homo_sapiens.GRCh37_ca	a	F					→ + → + + → + →	
tracks chr-	4:154,817,656	r galaxyp.msr.umm.eou/root≁ GWT ∽ Galaxy ∽ MSI ∽	worknow_ra=uasras dacker ~ News ~	escentration DB v bio v prot	→ html → M	Lin V othur v RickRack v reg v Norsk v A	pple ∽ HealthCare ∽ UniSci ∽ 200	4,566M of 5,797M
tracks chr	4:154,817,656 → java → py → ruby → G Galaxy / Ga	r galaxyp.msi.umn.eou/root / GWT ∽ Galaxy ∽ MSI ∽ ilaxyP	vorkiiow_ia=uaszas docker ∽ News ∽ https://g	escerrorro DB ∽ bio ∽ prot jalaxyp.msi.umn.edu/di	→ html → M atasets/20e4f86	LP ♥ othur ▼ RickRack ▼ reg ▼ Norsk ▼ A aacb54c7f/show_params	pple ∽ HealthCare ∽ UniSci ∽ 200 Untitled	4,566M of 5,797M 8 × FMS CFS × >> +
tracks chr.	4:154,817,656 java	r galaxyp:msi.umn.eou/root≁ GWT × Galaxy × MSI × JlaxyP	docker V News V https://g Analyze Data Wor	escorrorro DB × bio × prot Ialaxyp.msi.umn.edu/d: kflow Shared Data	→ html → M atasets/20e4186/ → Visualiza	LP ● othur → RickRack → reg → Norsk → Aj acob54c7f/show_params tion → Admin Help → User →	pple × HealthCare × UniSci × 200 Untitled	4,566M of 5,797M 8 Y FMS CFS Y >> + Using 183.9 GB
tracks chr iiii vim Gal Tools	4:154,817,656 • java • py • ruby • 0 Galaxy / Ga laxy / GalaxyP	I galaxyp.msi.umn.eou/root/ GWT ← Galaxy ← MSI ← IlaxyP Seqid Source Type	docker V News V https://g Analyze Data Wor Start	e9CBTTDTTD DB × bio × prot jalaxyp.msi.umn.edu/di kflow Shared Data End Score	 → html ~ M atasets/20e4/86/ ✓ Visualiza ◆ StrandPhase 	othur ∽ RickRack ∽ reg ∽ Norsk ∽ A aacb54c7f/show_params tion ← Admin Help ← User ← eAttributes	pple - HealthCare - UniSci - 200 Untitled History	4,566M of 5,797M 8 × FMS CFS × >> + Using 183.9 GB
tracks chr iii vim Cal Tools search t	4:154,817,656 ↓ java ↓ py ↓ ruby ↓ 0 Galaxy / Galaxy / Galaxy P tools ③	Galaxyp.msi.umn.edu/roor/ GWT × Galaxy × MSI × alaxyP Seqid Source Type ##gff-version 3 ##sequence-region 4 1	vorkilow_id=005705 docker v News v https://g Analyze Data Wor Start 155412208	e9C6TTDTTD DB × bio × prot Ialaxyp.msi.umn.edu/da kflow Shared Data End Score	 → html ~ M → Msualiza → Visualiza ⇒ StrandPhas 	othur v RickRack v reg v Norsk v Aj aacb54c7f/show_params tion v Admin Help v User v eAttributes	pple - HealthCare - UniSci - 200 Untitled History 5: peptides.gff3	4,566M of 5,797M 8 × FMS CFS × >> + Using 183.9 GB 2 * 1
tracks chr IIII vim Gal Tools Search t CORE TOO	4:154,817,656 • java • py • ruby • 0 Galaxy / Galaxy / Galaxy P tools ③	Galaxyp.msi.umn.edu/root/ GWT × Galaxy × MSI × ilaxyP Seqid Source Type ##gff-version 3 ##sequence-region 4 1 4 MassSpec pep 4 MassSpec CDS	Vorkilow_id=005705 docker v News v https://g Analyze Data Wor Start 155412208 ide 155412185 155412185	e9C6110110 DB × bio × prot Ialaxyp.msi.umn.edu/da kflow Shared Data End Score 155412208 10. 155412208 .	 html × M html × M tasets/20e4f86 Visualiza StrandPhase 0 - 0 - 0 	Dothur Y RickRack Y reg Y Norsk Y Aj hacb54c7f/show_params tlon Admin Help V User Y eAttributes ID=LLSVGGLR Parent=LLSVGGLR;transcript_id=ENSTO	pple × HealthCare × UniSci × 200 Untitled History 5: peptides.gff3 4 lines, 3 comm format: gff3, da	4.566M of 5,797M 8 × FMS CFS × >> + Using 183.9 GB 2 * total action of the second se
tracks chr iii vim Gal Tools Search t CORE TOO Get Data Send Data	4:154,817,656 ↓ java ↓ py ↓ ruby ↓ or Galaxy / Galaxy P ↓ tools ③ OLS	galaxyp.msi.umn.edu/root / GWT × Galaxy × MSI × llaxyP Segid Source Type ##eqff-version 3 ##sequence-region 4 1 4 MassSpec pep 4 MassSpec CDS ##sequence-region 11	Vorknow_id=000700 docker × News × https://g Analyze Data Wor Start 155412208 ide 155412185 155412185 1128838569	escondination DB × bio × prot alaxyp.msi.umn.edu/di kflow Shared Data End Score 155412208 10. 155412208 .	 html × M html × M visualiza StrandPhas 0 - 0 - 0 	Dothur ← RickRack ← reg ← Norsk ← Aj hacb54c7f/show_params tlon ← Admin Help ← User ← eAttributes ID=LLSVGGLR Parent=LLSVGGLR;transcript_id=ENSTO	pple × HealthCare × UniSci × 200 Untitled History 5: peptides.gff3 4 lines, 3 comm format: gff3, da Mapped 2 entri	4,566M of 5,797N 8 × FMS CFS × >> + Using 183.9 GB 2 & 11 4 ents tabase: hg19 es
tracks chr iiii vim Ga Tools search t CORE TOO Get Data Send Data Lift-Over	4:154,817,656 v java v py v ruby v f Galaxy / Galaxy / Galaxy P tools C	galaxyp.msi.umn.edu/root / GWT × Galaxy × MSI × ilaxyP Seqid Source Type ##gff-version 3 ##sequence-region 4 1 4 MassSpec 4 MassSpec ##sequence-region 11 11 MassSpec 11 MassSpec 11 MassSpec 11 MassSpec 11 MassSpec 12 MassSpec 13 MassSpec 14 MassSpec 15 ##sequence-region 11	Vorkilow_id=005705 docker V News V https://c Analyze Data Wor Start 155412208 ide 155412185 155412185 1128838569 ide 128838546	e9corrorror DB × bio × prot alaxyp.msi.umn.edu/dr kflow Shared Data End Score 155412208 10./ 155412208 . 128838569 10./ 128838569 .	 ✓ LJ × html × M × Visualiza × Strand Phase 0 - 0 - 0 - 0 - 0 - 0 	othur × RickRack × reg × Norsk × Ar aacb54c7f/show_params tion - Admin Help × User - eAttributes ID=LLSVGGLR Parent=LLSVGGLR;transcript_id=ENST0 ID=SPVLKPSR Parent=SPVLKPSR:transcript_id=ENST00	pple × HealthCare × UniSci × 200 Untitled History 5: peptides.gff3 4 lines, 3 comm format: gff3, da Mapped 2 entri 00000310343	4,566M of 5,797M 8 × FMS CFS × >> + Using 183.9 GB 2 & 11 2 & 11 2 & 12 10 10 10 10 10 10 10 10 10 10
tracks chr iiii vim Cols Search t CORE TOO Get Data Send Data Lift-Over Text Mani Filter and Join, Subtr	4:154,817,656 v java v py v ruby v (Galaxy / GalaxyP tools (2) tools (2) oLS a : ipulation I Sort tract and Group	galaxyp.msi.umin.edu/root/ GWT × Galaxy × MSI × ilaxyP Seqid Source Type ##gff-version 3 ##sequence-region 4 1 4 MassSpec CDS ##sequence-region 11 11 MassSpec CDS ##sequence-region 11 11 MassSpec CDS #1 MassSpec CDS	Vorkilow_id=005705 docker < News < https://c Analyze Data Wor Start 155412208 ide 155412185 155412185 1128838569 ide 128838546 128838546	e9C8T10T10 DB × bio × prot alaxyp.msi.umn.edu/da kflow Shared Data End Score 155412208 10. 155412208 . 128838569 10. 128838569 .	 html < M html < M visualiza Strand Phase 0 - 0 - 0 - 0 - 0 - 0 	othur v RickRack v reg v Norsk v Ar aacb54c7t/show_params tion v Admin Help v User v eAttributes ID=LLSVGGLR Parent=LLSVGGLR;transcript_id=ENSTO ID=SPVLKPSR Parent=SPVLKPSR;transcript_id=ENSTO	pple × HealthCare × UniSci × 200 Untitled History 5: peptides.gff3 4 lines, 3 comm format: gff3, da Mapped 2 entri 00000310343 display with ICV display at Ensen display at UCSC	4,566M of 5,797M 8 Y FMS CFS Y >> + Using 183.9 GB C * 1 tabase: hg19 es web current local hbl <u>Current local</u> hbl <u>Current local</u>
tracks chr.	<pre>4:154,817,656</pre>	Galaxyp.msr.umn.edu/root/ GWT × Galaxy × MSI × MaxyP Seqid Source Type ##gff-version 3 ##sequence-region 4 1 4 MassSpec Pep 4 MassSpec CDS ##sequence-region 11 11 MassSpec Pep 11 MassSpec CDS	Vorkilow_id=UU3703 docker < News < https://g Analyze Data Wor Start 155412208 ide 155412185 155412185 1128838569 ide 128838546 128838546	escention in in DB × bio × prot ialaxyp.msi.umn.edu/di ialaxyp.msi.umn.edu/di kflow Shared Data End 155412208 10. 155412208 . 128838569 10. 128838569 .	 html ~ M html ~ Visualiza StrandPhas 0 - 0 - 0 0 - 0 - 0 	othur ~ RickRack ~ reg ~ Norsk ~ A aacb54c7f/show_params tion ~ Admin Help ~ User ~ eAttributes ID=LLSVGGLR Parent=LLSVGGLR;transcript_id=ENSTO ID=SPVLKPSR Parent=SPVLKPSR;transcript_id=ENSTO	pple V HealthCare V UniSci V 200 Untitled History 5: peptides.gff3 4 lines, 3 comm format: gff3, da Mapped 2 entri 00000310343 display with ICV display at Ensen display at Ensen display at CSC	4,566M of 5,797M 8 × FMS CFS × >> + Using 183.9 GB C * II es tabase: hg19 es web current local hbl Current main rcc 3.Type 4.Stor

INPUTS



INPUTS

Eile Conomer View Tracks Perions	IGV	
Human halo		
ruman ng19 🐳 Cnr4	Chr4:155,412,131-155,412,229 G0 ■ ■ ₩ ₩ ₩ ₩ ₩ ₩ ₩	
p16.2 p15.33 p	15.2 p15.1 p14 p12 q12 q13.1 q13.3 q21.21 q22.1 q23 q25 q26 q28.1 q28.3 q31.7	21 q31.3 q32.2 q33 q34.3 q35.1
	99 bp	
30 bp 155,412,140 b 	p 155,412,150 bp 155,412,160 bp 155,412,170 bp 155,412,180 bp 155,412,190 bp 155,412,200 bp 	155,412,210 bp 155,412,220 bp 155,4
Sequence → CG GAT GAT GC CG A D D R M M R C R RefSeq Genes R I	GTGTCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGCCTCCGGACAGAAGAGG G C T G V L R Q K V Q Q R G V L G V L R Q K V S G C T W K S S S S G E S S E S S D R K C P G A R G S R P A A G S H R S P P T E R T D P H V H F D D L L P S D D S D E S L F	AGCCGCTCCCCTCCTGCTGCTGC E A A P L L L K P L P S C C C S R S P P A A A F G S G E Q Q Q
nentides off3	R L G G V S L	L
Harris espises OBOK27 and	LLSVGGLR	LLSVGGLR
73.gtf	< < < < < < < < < < < < < < < < < < <	<pre></pre>
		ID: LLSVGGLR
4 tracks chr4:155,412,208		Exon number: 1 M of 5,797M chr4:155412185-155412208
		Parent: LLSVGGLR
Galaxy / Gal	VT v Galaxy v MSI v docker v News v DB v bio v prot v html v Mothur v RickRack v reg v Norsk v Apple v HealthCan https://galaxyp.msi.ump.edu/datasets/20e4186aacb54c7f/show_params	untitle
- Galaxy / GalaxyP	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Using 183.9 GB
Taols	Conid Courses Tumo Start End Course Changel Inco Attributes	History C Ö 🗍
	##gff-version 3	
search tools	##sequence-region 4 1 155412208	5: peptides.gff3 ③ 🖋 🗙
CORE TOOLS	4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR 4 MassSpec CDS 155412185 155412208 - 0 Parent=LLSVGGLR:transcript_id=ENST00000339452	4 lines, 3 comments
Get Data	##sequence-region 11 1 128838569	ionnat. gris, database. iigis
Send Data	11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR	Mapped 2 entries
Lift-Over Text Manipulation	11 MassSpec CDS 128838546 128838569 0 Parent=SPVLKPSR;transcript_id=ENST00000310343	
Filter and Sort		display with IGV web current local
Join, Subtract and Group		display at UCSC <u>main</u>
Convert Formats		1.Seqid 2.Source 3.Type 4.Star
Extract Features	1	##off-version 3

INPUTS



INPUTS

e e File Genomes Vie	w Tracks Reg	ons Tools	GenomeSpace	Help		IGV							
Human hg19	the second seco		¢ chr11			Go 👚	< →	 \$\vee\$ \$\vee\$ \$\vee\$ \$\vee\$ \$\vee\$ \$\vee\$ \$\vee\$ \$\vee\$ \$\vee\$ 	X 🏳				+
	p15.4 p1	5.3 p15.1 p	14.3 p14.1 p13	p12	011.2 p11.11	q12.2	q1 3.	2 q13.4	q14.1 q	14.2 q2	1 q22.1	q22.3 q23.1	q23.3 q24.1 q24.3
	nb	20 mb 	I	40 mb 	I	60 mb 	— 134 r	mb	80 mb 		100 mb 	Click a	Ind drag to zoom in. 120 mb
RefSeq Genes	28326 ILK US	247 TPH1 I	UZP2 DCDC5		DDB2 TR	IM48 ZP1	SIPA1	RNF121 GA	AB2 DLG2	FAT3	CNTN5	GRIA4 BTG4	SIK3 BLID ETS1 JAN
peptides.gff3				Linto io			0				0.1110		
Homo_sapiens.GRCh37_c 73.gtf	ca 00519787 ENST	00000534211 E	ENST00000513853	ENST00000	533565 ENST	0000052798	5 ENS	T00000351960	D ENSTODO	00534163 EN	IST0000047	76452 ENST0000	00517061 ENST0000028143
4 tracks	r11:117,559,881	e calaxyo.ms	s.umn.edu/root/workt	10w 1d=0d57d5	e9c8f1b11b	6							4,698M of 5,797
4 tracks chr	r11:117,559,881	iii galaxyp.ms × GWT × Gala	ii.umn.edu/root?workt axy ∽ MSI ∽ docke	iow_id=0d57d5 ar ~ News ~	e9c8f1D11b DB ~ bio ~	C prot ~ html •	₩ • Moth	ur v RickRack	≺ reg ∽ No	orsk 🗸 Apple v	HealthCare	e v UniSci v 2004	4,698M of 5,797
4 tracks	r11:117,559,881	i galaxyp.ms → GWT → Gala y / GalaxyP	a.umn.edu/root?workt axy ∽ MSI ∽ docke	low_id=0d57d5 er × News × https://g	e9c8f1b11b DB × bio × jalaxyp.msi.umn.e	C prot ∽ html 1 adú/datasets/20	≝∟ ∽ Moth e4f86aacl	ur v RickRack	❤ reg ❤ Ni ams	orsk ∽ Apple ∾	· HealthCare	e × UniSci × 2000 Untitled	4,698M of 5,797 8 × FMS CFS × >> +
4 tracks chr	r11:117,559,881 v java v py v ruby Gala laxy / Galax	■ galaxyp.ms × GWT × Gala y / GalaxyP /P	il.umn.edu/root?workt axy ~ MSI ~ docke Analy	low_id=Ud57d5 er V News V https://j /ze Data Wo	e9c8f1D11D DB × bio × jalaxyp.msi.umn.e rkflow Shared	o prot Y html v edu/datasets/20 I Data → Vist	Moth e4f86aacl	ur v RickRack b54c7l/show_part n Admin	× reg × No ams Help → Use	orsk Y Apple Y	· HealthCare	e v UniSci v 2004 Untitled	4,698M of 5,797 8 × FMS CFS × >> + Using 183.9 GB
4 tracks chr	r11:117,559,881	GWT ~ Gala GWT ~ Gala y / GalaxyP /P Seqid Sou ##off_v	a.umn.edu/root?worki axy × MSI × docke Analy urce Type	iow_id=0d57d5 ar Views V https://f yze Data Wo Start	e96811D11D DB × bio × Jalaxyp.msi.umn.e kflow Shared End !	prot Y html \ adu/datasets/20 I Data → Visi Score Strand	₩ ✓ Moth e4f86aacl ualization dPhaseA	ur ∨ RickRack b54c7f/show_parc n - Admin attributes	× reg × No ams Help ← Use	orsk 👻 Apple 🗸	r HealthCare	e v UniSci v 2000 Untitled History	4,698M of 5,797 8 × FMS CFS × >> + Using 183.9 CB C * 10
4 tracks chr IIII vim Gal Tools Search t CORE TOO	r11:117,559,881	GWT ~ Gala GWT ~ Gala y / GalaxyP P Seqid Sou ##gff-v ##sequ 4 M 4 M	ii.umn.edu/root?worki axy × MSI × docke Analy urce Type rersion 3 ience-region 4 1 1554 hassSpec peptide hassSpec CDS	IOW_Id=0d5/d5 ar V News V https://q yze Data Wo Start 412208 155412185 155412185	e9c8(1D11b DB > bio > jalaxyp.msi.umn.e kflow Shared End : 155412208 155412208	prot × html v adu/datasets/20 I Data → Visi Score Strand 10.0 - 	⊥⊥ ✓ Moth e4f86aacl ualization dPhase A 0 0	ur × RickRack b54c7t/show_parc n~ Admin ttributes ID=LLSVCGLR Parent=LLSVCG	× reg × No ams Help ← Use 5LR;transcript_	orsk × Apple ×	HealthCare	e v UniSci v 2004 Untitled History 5: peptides.gff3 4 lines, 3 comm format: gff3, dat	4,698M of 5,797 8 × FMS CFS × >> + Using 183.9 CB C * 1 to 2
4 tracks chr iiii vim Gal Tools Search t CORE TOO Get Data Send Data	r11:117,559,881	iii galaxyp.ms V GWT V Gala V / GalaxyP P Seqid Soo ##gff-v ##seque 4 M 4 M ##seque	axy × MSI × docke Analy urce Type rersion 3 rence-region 4 1 1554 AassSpec peptide AassSpec CDS ence-region 11 124	low_id=0d57d5 ar Vews V https:// yze Data Wo Start 412208 155412185 155412185 155412185	e9c8(1b) 1b DB × bio × jalaxyp.msi.umn.e. tkflow Shared End : 155412208 155412208	prot v html v adu/datasets/20 I Data v Vist Score Stranc 10.0 - 	Moth e4f86aaci yalizatior dPhase A 0 0	Ur V RickRack b54c7f/show_pare n Admin tttributes ID=LLSVCGLR Parent=LLSVCG	Y reg Y No ams Help → User 5LR;transcript_	orsk × Apple × r • jd=ENST00000	HealthCare	e V UniSci V 2000 Untitled History 5: peptides.gff3 4 lines, 3 comme format: gff3, dat Mapped 2 entrie	4,698M of 5,797 8 × FMS CFS × >> + Using 183.9 CB C & t t t t t t t t t t t t t

INPUTS

e enomes Viev	w Tracks Regions	IGV	
Human hg19	chr11	+ chr11:128,834,505-128,844,004 Go ²	- +
	p15.4 p15.3	p15.1 p14.3 p14.1 p13 p12 p11.2 p11.11 q12.2 q13.2 q13.4 q14.1 q14.2 q21 q22.1	q22.3 q23.1 q23.3 q24.1 q24.3
	128,835,000 bp	9,484 bp	28,842,000 bp 128,843,000 bp 128,844 I I I I
RefSeq Genes		ARHGAP32	· · · · · · · · · · · · · · · · · · ·
peptides.gff3		SPVLKPSR	
73.gtf		ENST00000310343 ENST00000526162	
tracks	1:128,840,313		4,713M of 5,797M
IIII vim v	java 🛩 py 🛩 ruby 🛩	Galaxy Y MSI Y docker Y News Y DB Y bio Y prot Y html Y Mothur Y RickRack Y reg Y Norsk Y Apple Y HealthCare	• VniSci v 2008 v FMS CFS v >>
	Galaxy / G	laxyP https://galaxyp.msi.umn.edu/datasets/20e4f86aacb54c7f/show_params	Jntitled +
- Gala	axy / GalaxyP	Analyze Data Workflow Shared Data + Visualization + Admin Help + User +	Using 183.9 GB
Tools	1	Seqid Source Type Start End Score Strand Phase Attributes	History C 🌣 о
Search to	ols 🕑	##gff-version 3 ##sequence-region 4 1 155412208 4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR	5: peptides.gff3
<u>Get Data</u> Send Data		4 MassSpec CDS 155412185 155412208 - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452 ##sequence-region 11 1 128838569	format: gff3, database: hg19 Mapped 2 entries
Lift-Over Text Manip Filter and S Join, Subtr	ulation jort act and Group	11 MassSpec CDS 128838546 128838569 0 Parent=SPVLKPSR;transcript_id=ENST00000310343	display with IGV web current local display at Ensembl <u>Current</u> display at UCSC <u>main</u>

INPUTS



INPUTS



SEARCHGUI PEPTIDESHAKER BLAST-P **PSM Visualization**

ENTIRE PROTEOGENOMICS WORKFLOW

NPUT	History 8A: Input for Entire Workflow 151.1 MB search datasets Dataset L: ABRF-Spike5fasta fasta 2: FASTA. File. from. EnSEMBL. Searches.fasta 3: Mascot formatted MGF of data. 5.mgf 4: Mascot. formatted. MGF. of. data. 5.mgf 5: MGF. Data. Set. List	WORKFLOW Published Workflows <i>search name, annotation, owner, and tag</i> Q Advanced Search	Publishe	d Histories	٩
	a list of datasets 6: Homo sapiens.GRCh37.73.cdna.chr 4 17.fa	Name	Nama		
	Z: Homo_sapiens.GRCh37_canon.73.chr.4_17.gtf	Workflow 8: Entire Proteogenomics Workflow.	Entire History		
	Running workflow "Workflow 8: Entire	Proteogenomics Workflow."	nd All Collapse	History	2 🌣 🗆
	Step 1: Input dataset			search datasets	0
	Subset of 3-frame translated database			imported: Entire History 57 shown	(mark)
	type to filter			907.4 MB	
	Step 2: Protein Database Downloader (version 0.2.0)			60: Peptide to GFF pepti des.gff3	● / ×
	Step 3: Input dataset			59: Peptide to GFF pepti des.unmapped	● / ×
	Spiked in proteins 1: ABRF-Spike4.fasta.fasta == == == == == == == == == == == == ==			57: Column Regex Find A nd Replace on data 56	● / ×
	type to filter			56: Homo sapiens.GRCh 37 canon.73.chr 4 17.gt	● / ×
	Step 4: Protein Database Downloader (version 0.2.0)			I	
	Step 5: Input dataset collection			37.73.cdna.chr 4 17.fa	• / ×
	MGF Files (dataset collection)			54: mz to sqlite on data 14, data 3, and others	• / ×
	type to filter			53: BLAST-P Filtered Pep tide Report	• / ×
	cDNA database			52: BLAST-P Filtered Pep tides	• / ×
	55: Homo_sapiens.GRCh37.73.cdna.chr_4_17.fa *) type to filter			51: Concatenate dataset s on data 49 and data 50	• / ×
	Step 7: Input dataset			50: Concatenate dataset s on data 48 and data 42	• / ×
	GTF File 56: Homo_sapiens.GRCh37_canon.73.chr_4_17.gtf			49: Concatenate dataset s on data 40 and data 47	● / ×
	type to filter			48: Cut on data 46	• / ×
	Step 8: Regex Find And Replace (version 0.1.0)			47: Cut on data 45	• / ×
	Step 9: FASTA Merge Files and Filter Unique Sequences (version 1.0)				

OUTPUT

SEARCHGUI PEPTIDESHAKER BLAST-P

INPUTS

Genome Visualization

PSM Viz



<u>Biochemistry, Molecular Biology &</u> <u>Biophysics</u> Kevin Murray Ray Sajulga Candace Guerrero

<u>Center for Mass Spectrometry</u> and Proteomics

Ebbing de Jong LeeAnn Higgins Todd Markowski



July 7th 2015: Session II : 11:20 AM <u>Proteomics Visualization in Galaxy</u> James E. Johnson

July 8th 2015: Session 7 : 2:05 PM: Extending Galaxy's reach: recent progress towards complete multi-omic data analysis workflows. Timothy J Griffin

UNIVERSITY OF MINNESOTA SUPERCOMPUTING INSTITUTE

Tom McGowan

Trevor Wennblom Getiria Onsongo Bill Gallip Ben Lynch

COMMUNITY BASED SOFTWARE DEVELOPMENT

Harald Barsnes and Marc Vaudel

University of Bergen, Bergen, Norway Bjoern Gruening University of Freiburg, Freiburg, Germany Lennart Martens VIB Department of Medical Protein Research, Ugent, Belgium Ira Cooke La Trobe University, Melbourne , Australia John Chilton Galaxy Team Penn State University



