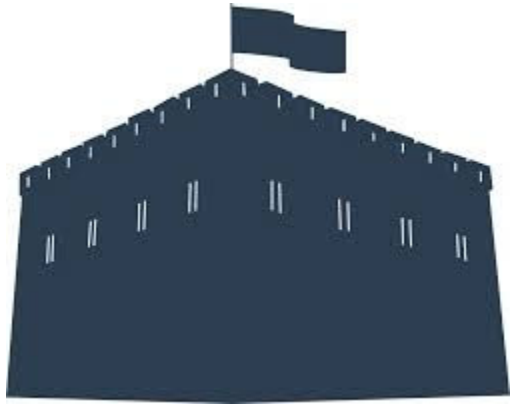
Mass Spectrometry-based Proteomics Data Analysis using Galaxy-P



GalaxyP

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

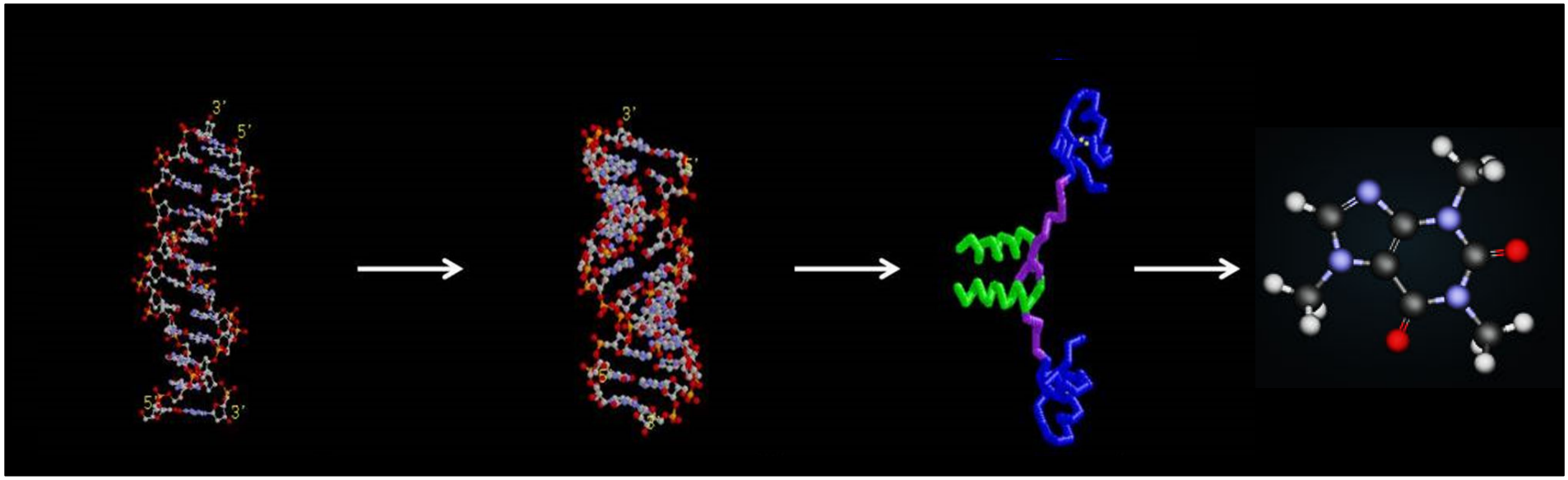
Documentation and Infrastructure Support

Kevin Murray Ray Sajulga Thomas McGowan

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



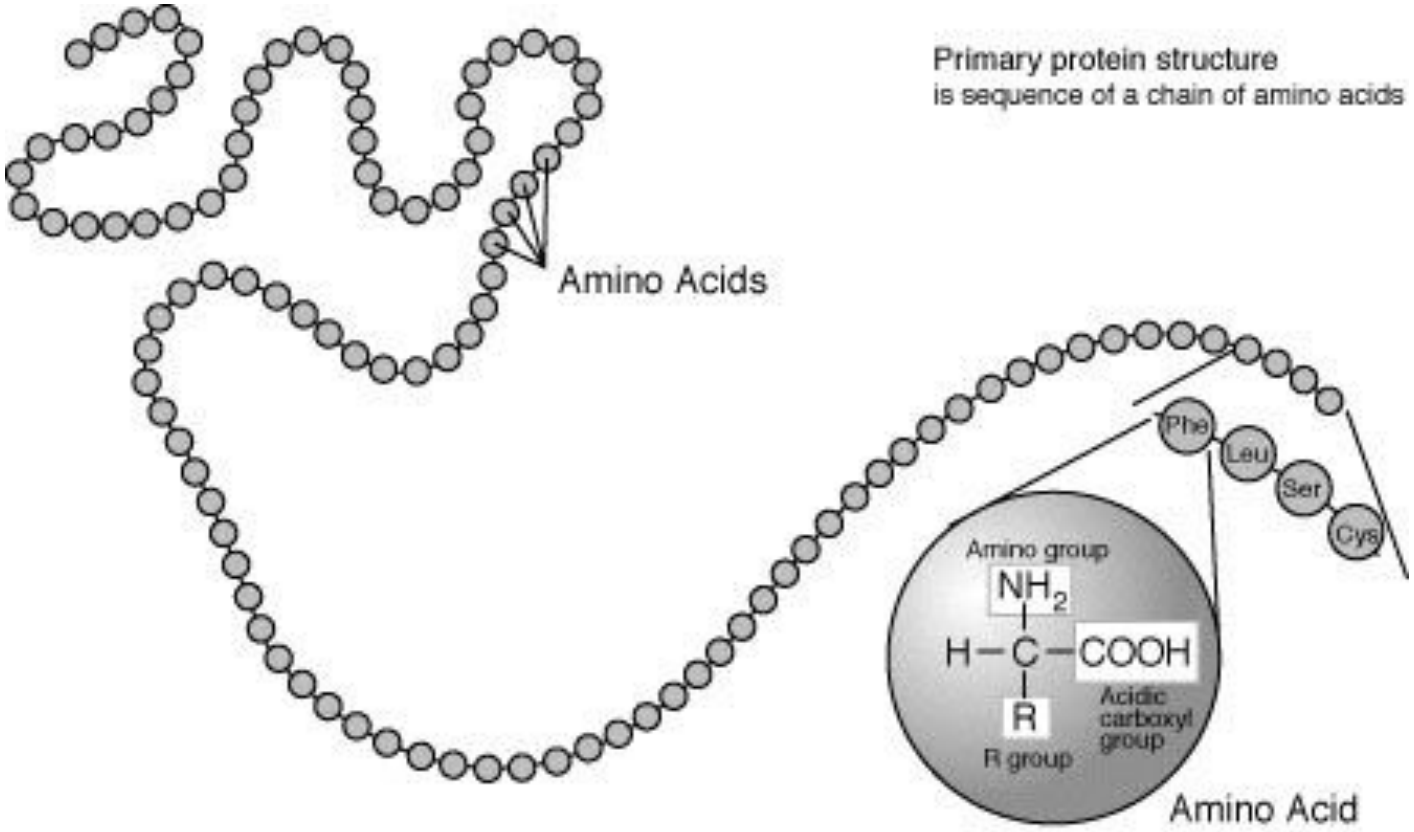
DNA *Genome* | RNA *Transcriptome* | Protein *Proteome* | Metabolite *Metabolome*

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.)

Quick reminder: primary protein structure



(20 naturally occurring)

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

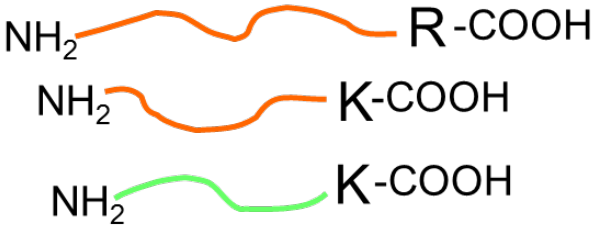
Measuring masses of proteins and peptides

1. Protein, MW = 10,000 +

"Top down proteomics"

digest into peptides (enzymatic, chemical)

2. Peptides, MW < 4,000



Trypsin

N-TH**K**.NCPHIVVGTPG**R**.IPD-C

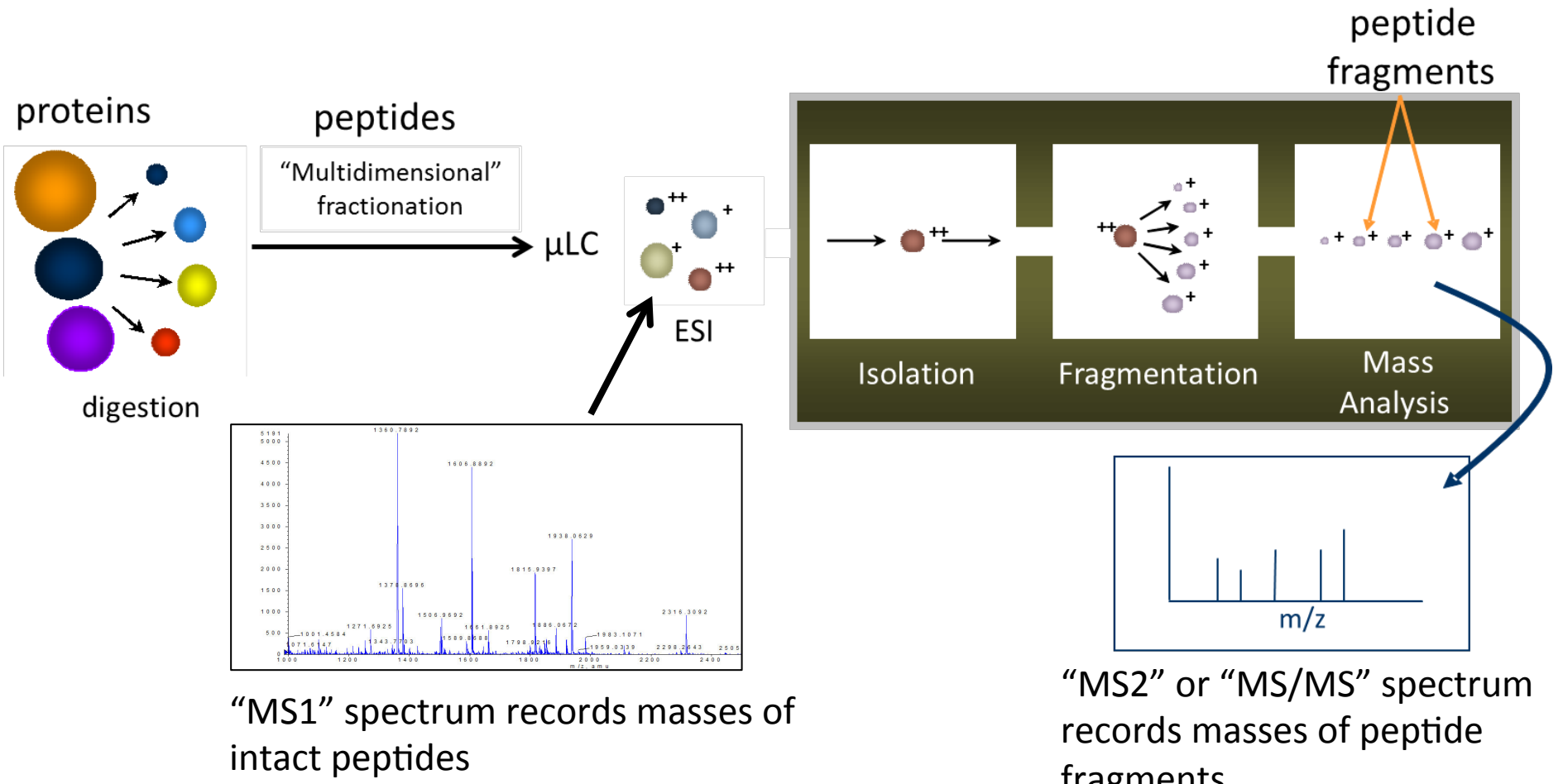
cleaves c-teminal side of arginine (R) and lysine (K)

3. Peptide fragments

"Bottom up proteomics"

High throughput protein identification by MS

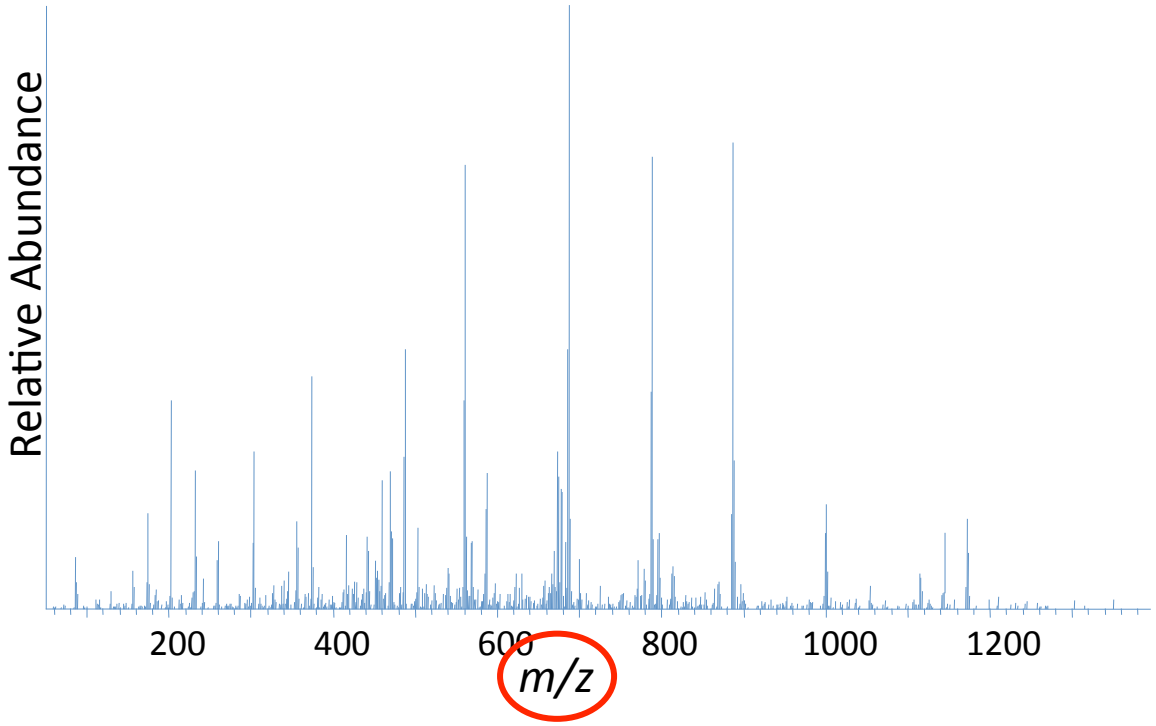
Peptide fractionation coupled to tandem mass spectrometry (MS/MS)



"MS1" spectrum records masses of intact peptides

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

Information “currency” of MS: mass spectra



Common mass spectrometry terms

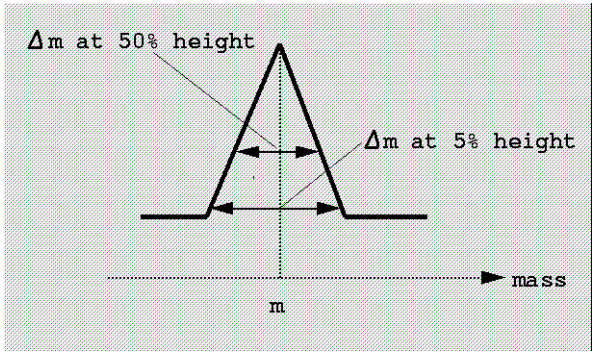
m/z = mass-to-charge = mass of analyte/number of retained charges

Mass accuracy (ppm) = [(measured m/z – actual m/z)/(actual m/z)] * 10^6

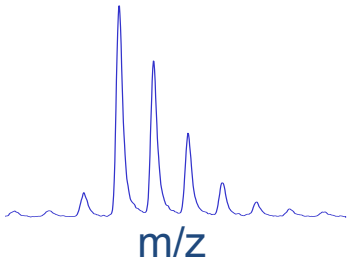
- Mass accuracy is usually expressed in units of parts per million (ppm)
- Generally a lower mass accuracy in ppm is preferred

Mass resolution = m/Δm
where m = analyte mass

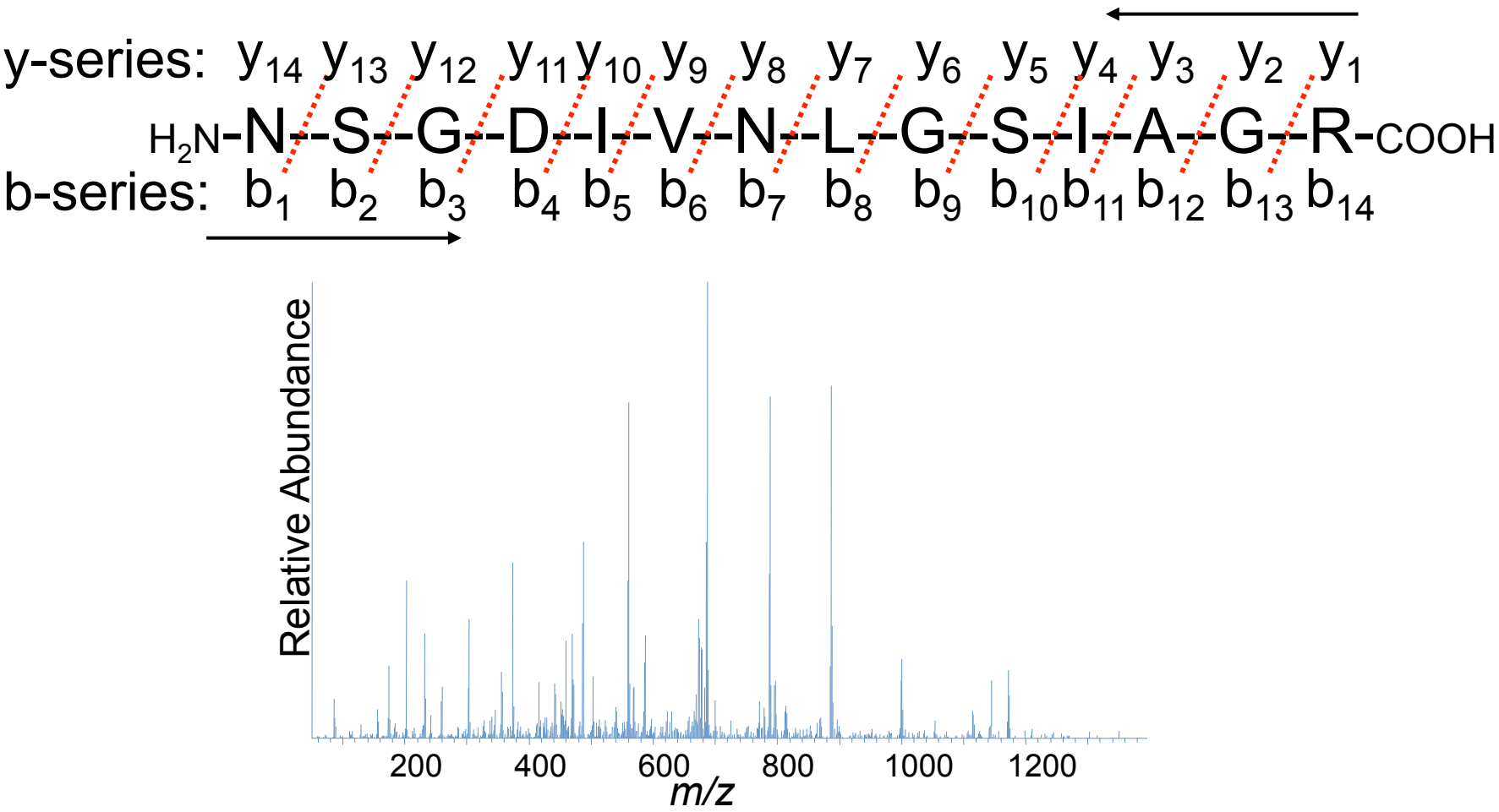
Δm = width of peak, generally at 50% of peak height (also called full-width half-max, or FWHM)



- Generally higher mass resolution is preferred



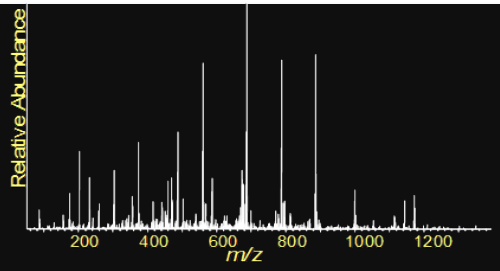
Tandem mass spectrometry and peptide sequence



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

High throughput identification of proteins

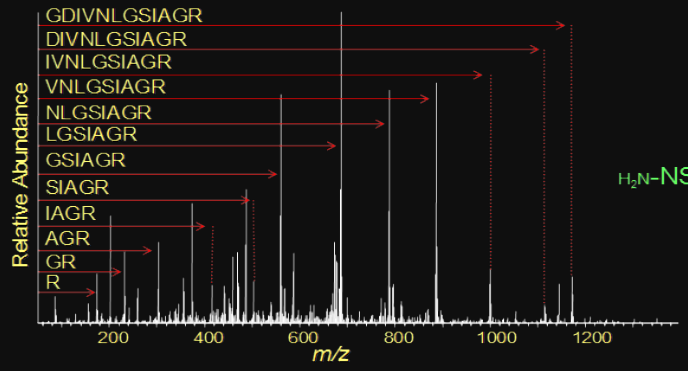
Raw MS/MS spectrum



Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

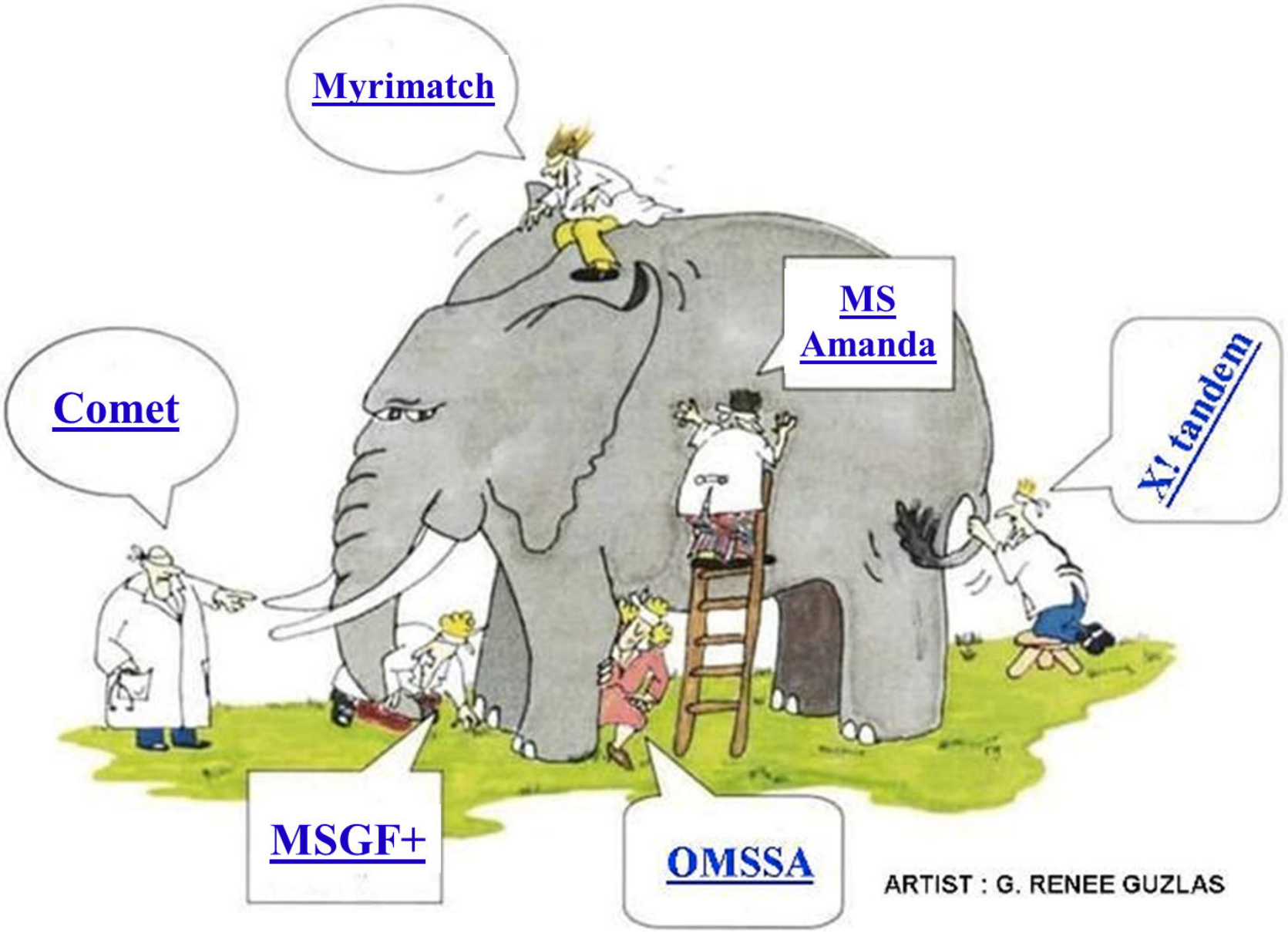


H_2N-NSGDIVNLGSIAGR-COOH

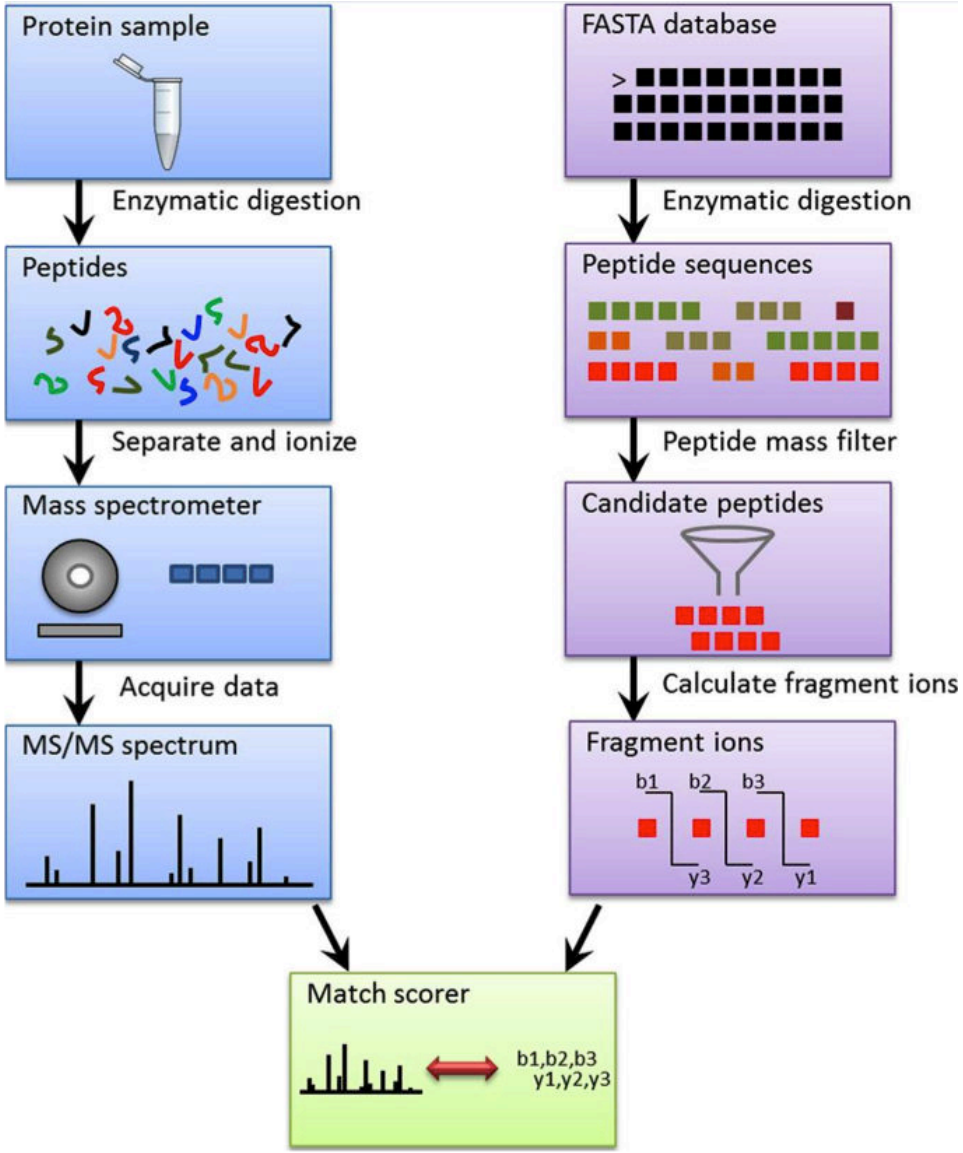
Protein identification

GalaxyP

Different choices for sequence database searching

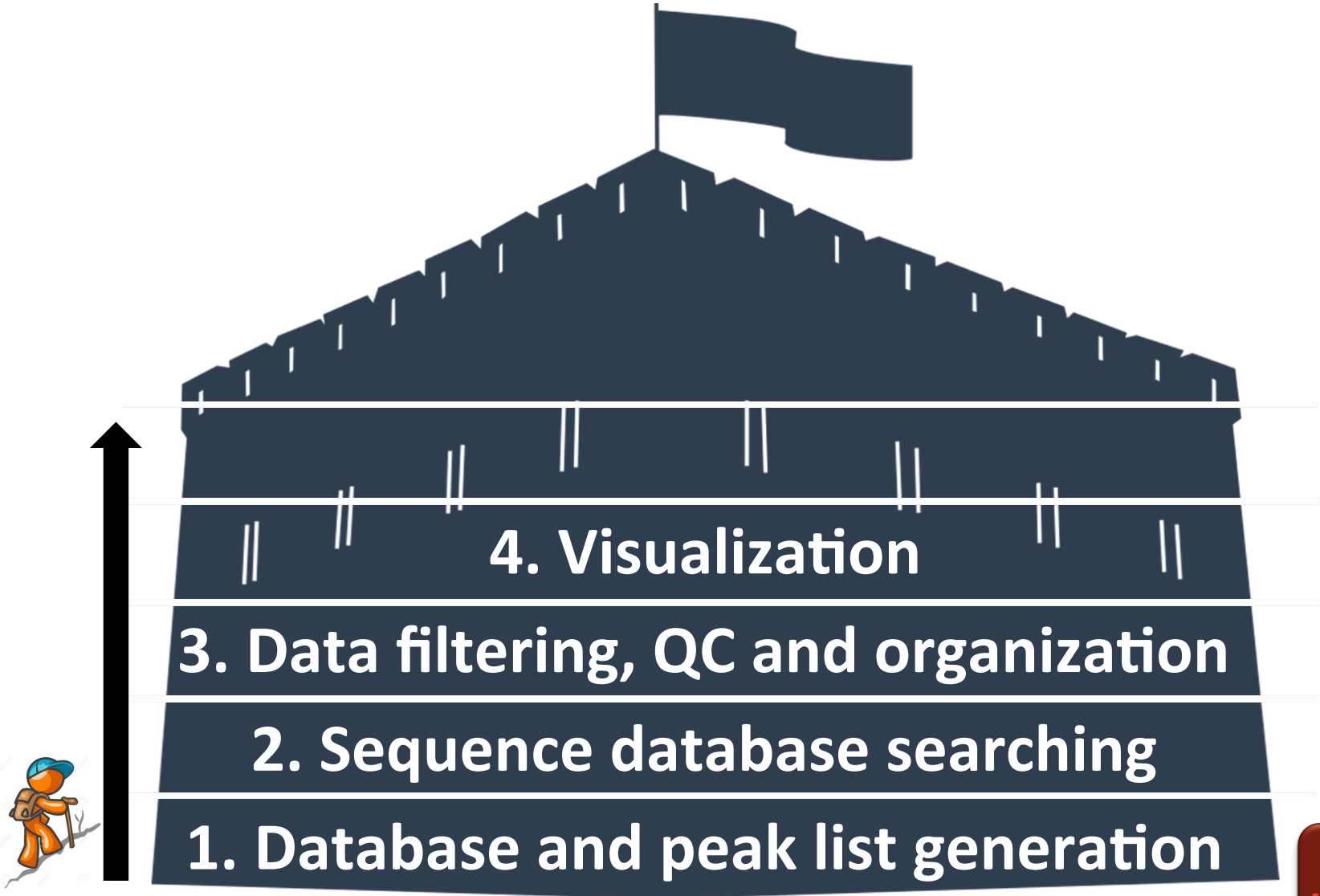


Experimental workflow in MS-based proteomics

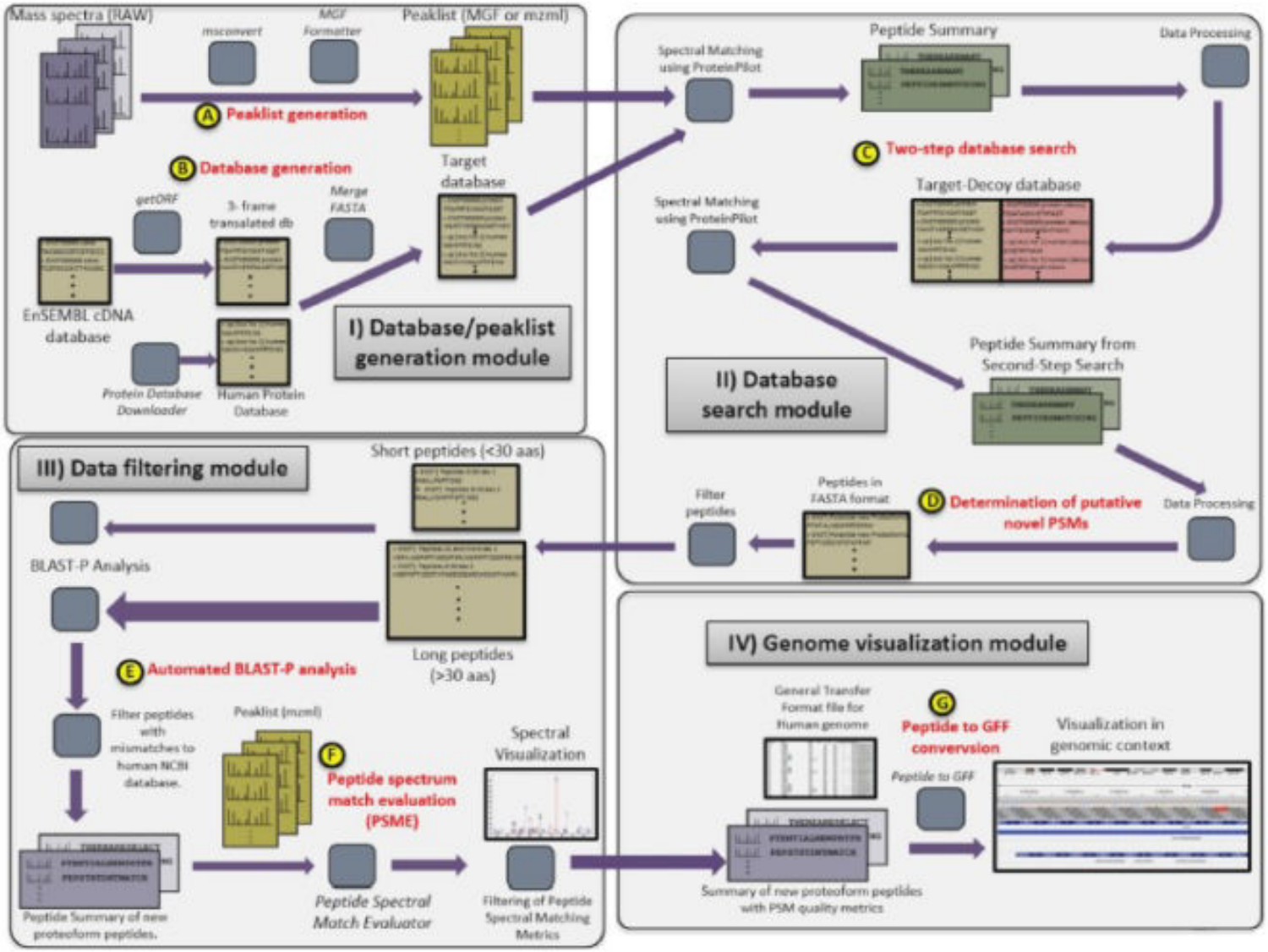


Eng et al 2011
Mol Cell Proteomics. 10(11): R111.009522.

Data analysis modules for protein identification



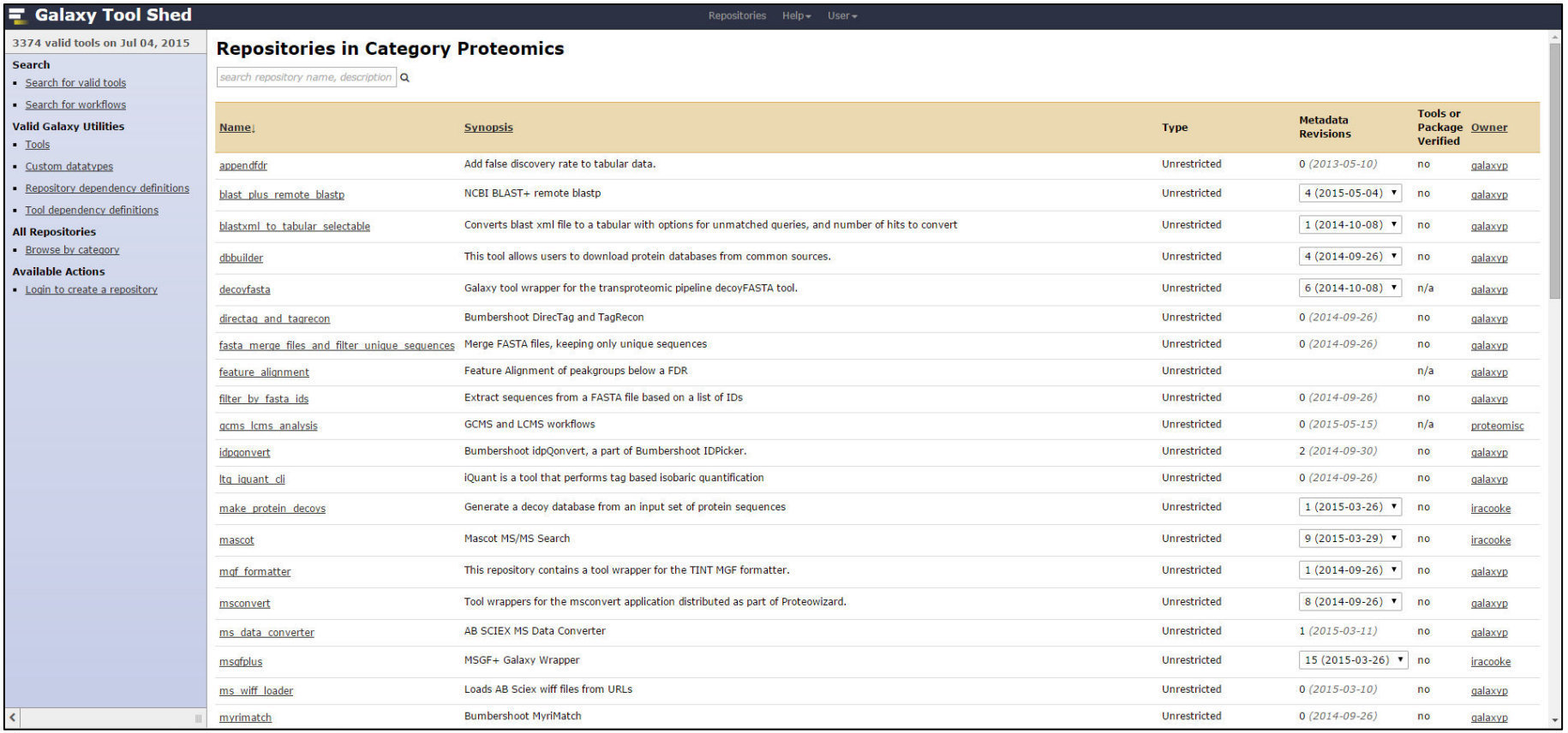
Why use Galaxy for proteomics data analysis?



J Proteome Res. 2014 13:5898-908

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
- Search for valid tools
- Search for workflows

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

All Repositories
- Browse by category

Available Actions
- Login to create a repository

Repositories in Category Proteomics

search repository name, description

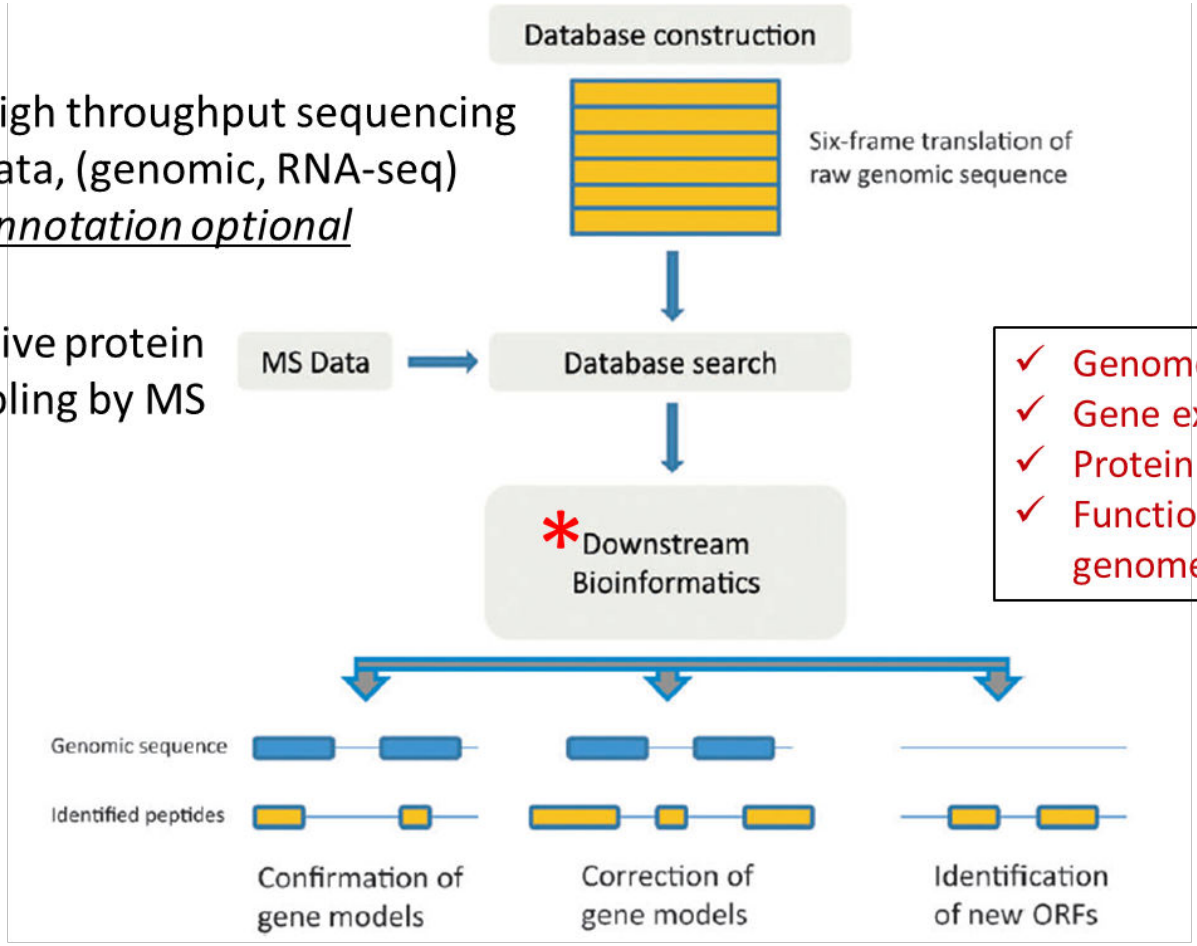
Name	Synopsis	Type	Metadata Revisions	Tools or Package Verified	Owner
appendfdr	Add false discovery rate to tabular data.	Unrestricted	0 *(2013-05-10)*	no	galaxyp
blast_plus_remote_blastp	NCBI BLAST+ remote blastp	Unrestricted	4 (2015-05-04)	no	galaxyp
blastxml_to_tabular_selectable	Converts blast xml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

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

Basics of proteogenomics

High throughput sequencing data, (genomic, RNA-seq) *annotation optional*

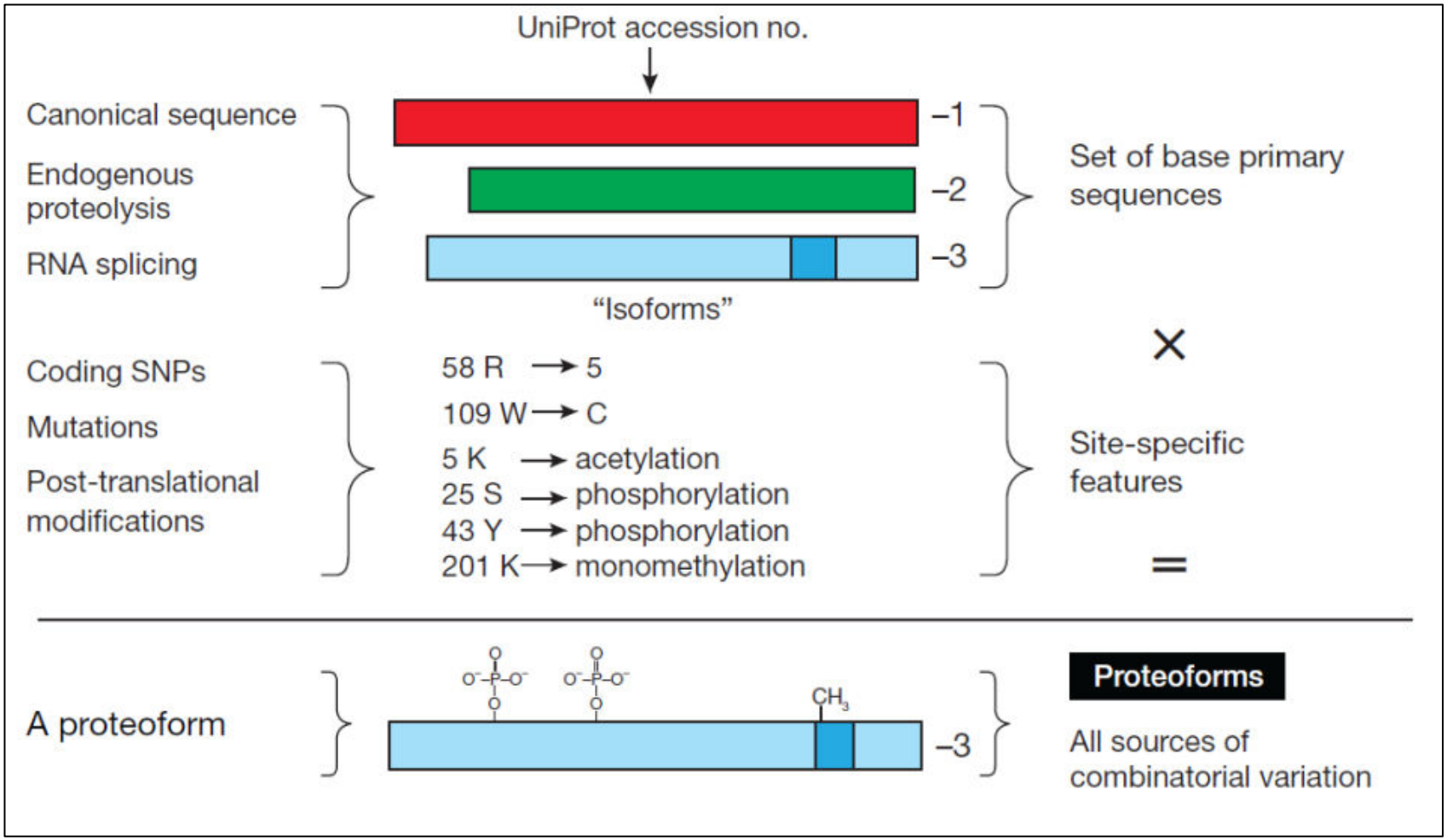
Comprehensive protein sampling by MS



- ✓ Genome annotation
- ✓ Gene expression regulation
- ✓ Protein variants in disease
- ✓ Functional outcomes of genome mutation

Mol. BioSyst., 2011, 7, 284–291

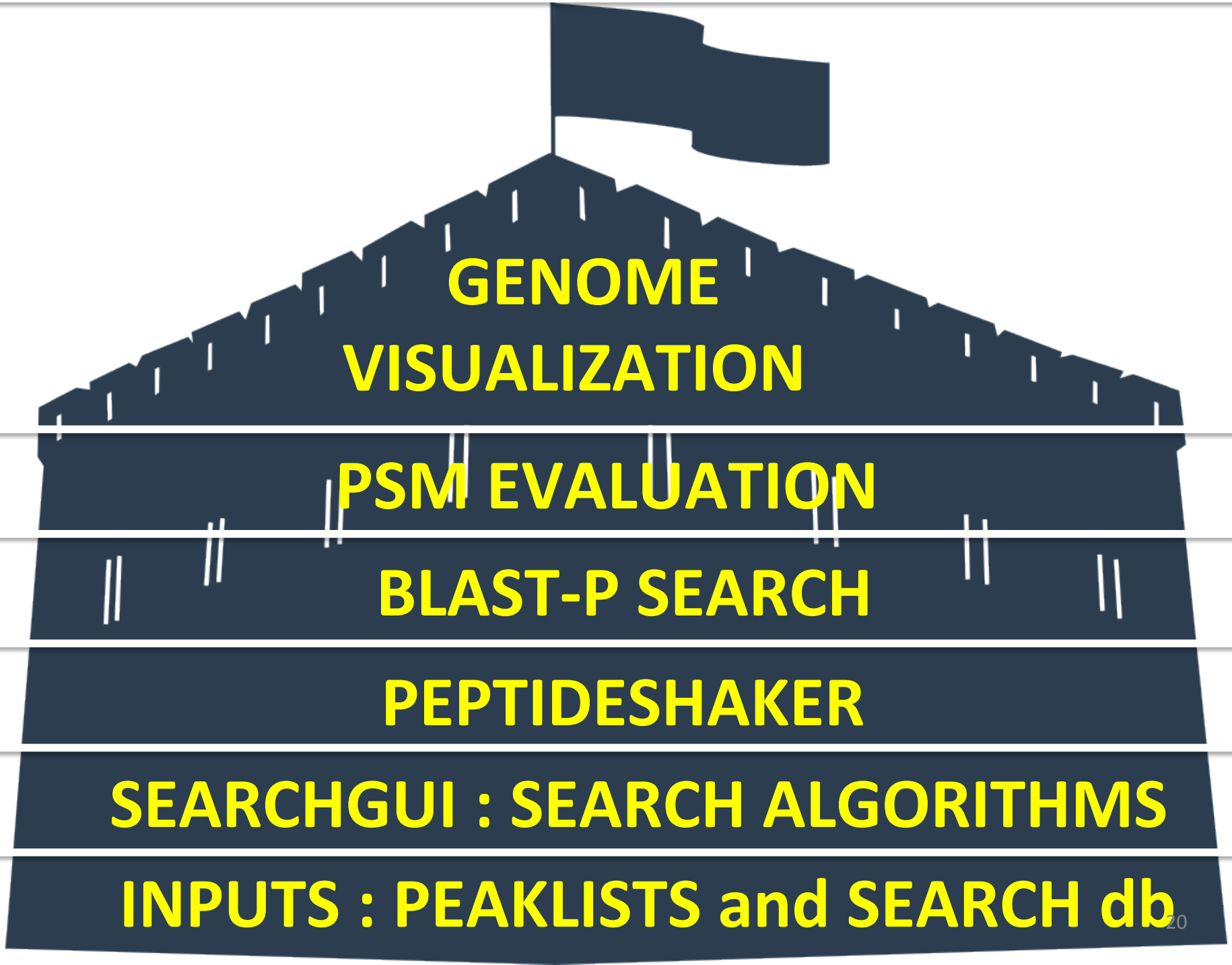
Peptide sequence variants and “proteoforms”



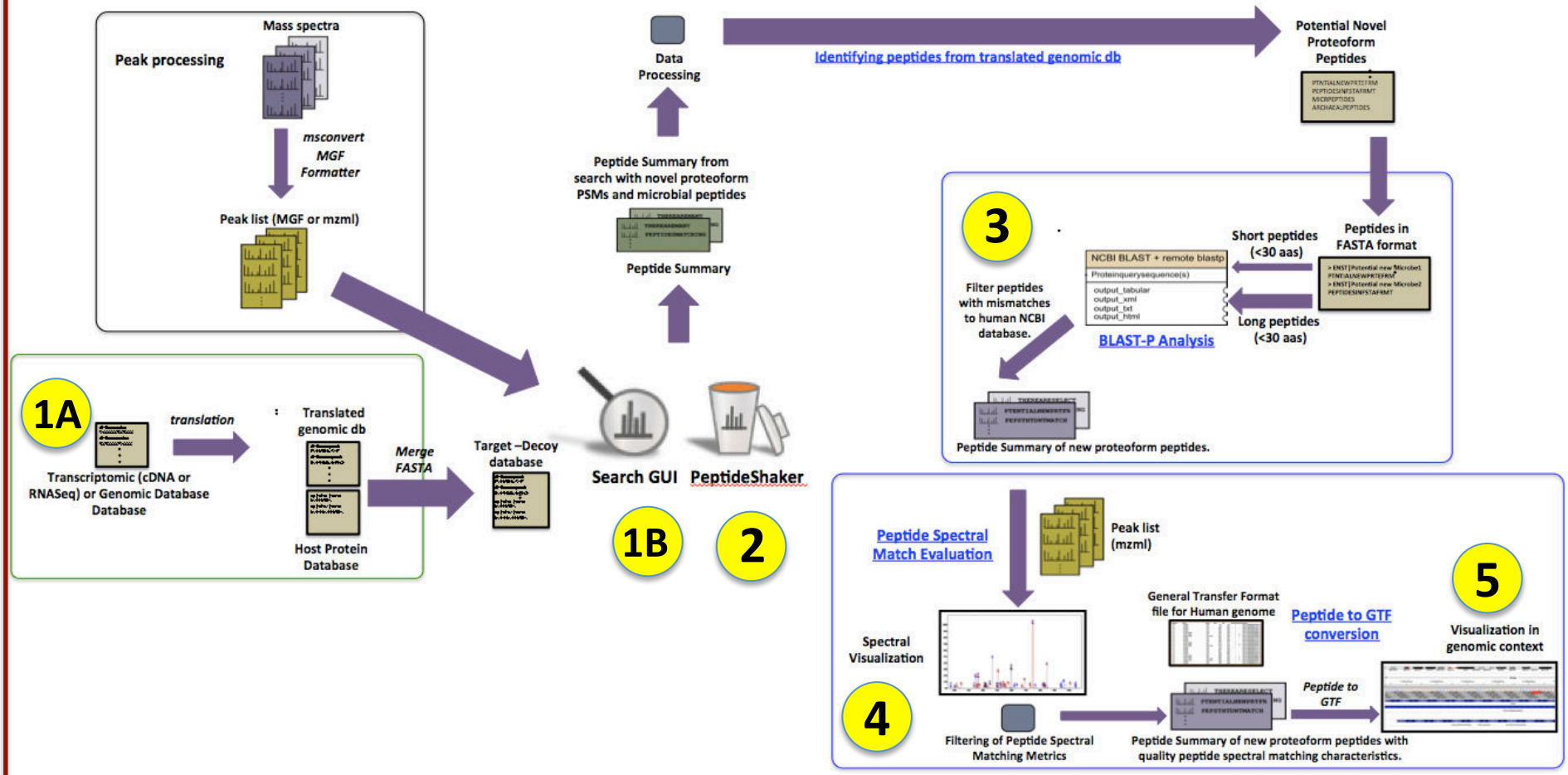
| VOL.10 NO.3 | MARCH 2013 | NATURE METHODS

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.


GENOME VISUALIZATION
PSM EVALUATION
BLAST-P SEARCH
PEPTIDESHAKER
SEARCHGUI : SEARCH ALGORITHMS
INPUTS : PEAKLISTS and SEARCH db

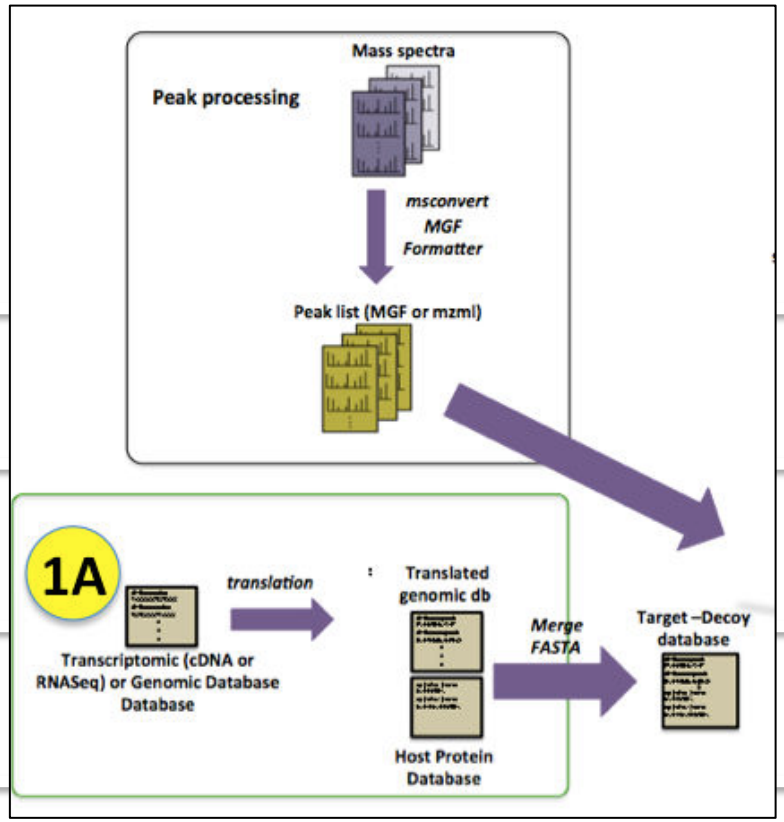
PROTEOGENOMICS WORKFLOW



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

Timothy J Griffin

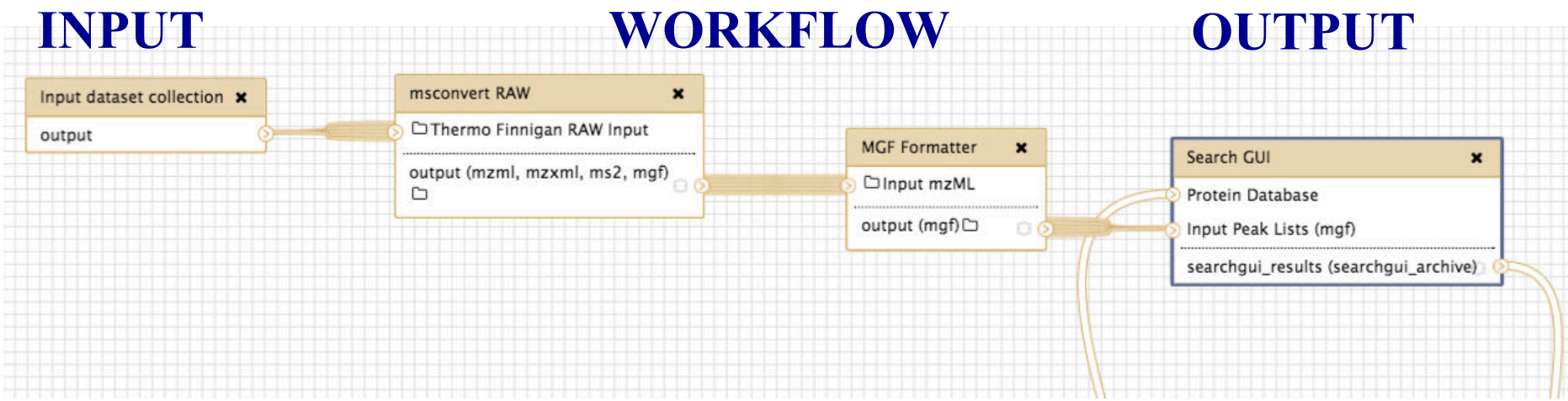
GalaxyP



INPUTS : PEAKLISTS and SEARCH db

INPUTS : PEAKLISTS and SEARCH db

GALAXY TOOLS AND WORKFLOWS



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

INPUTS : PEAKLISTS and SEARCH db

GalaxyP

Getting Started GCC2015 GalaxyP Tutorial

Open a web browser and navigate to the GalaxyP cloud instance –

http://54.205.17.20/

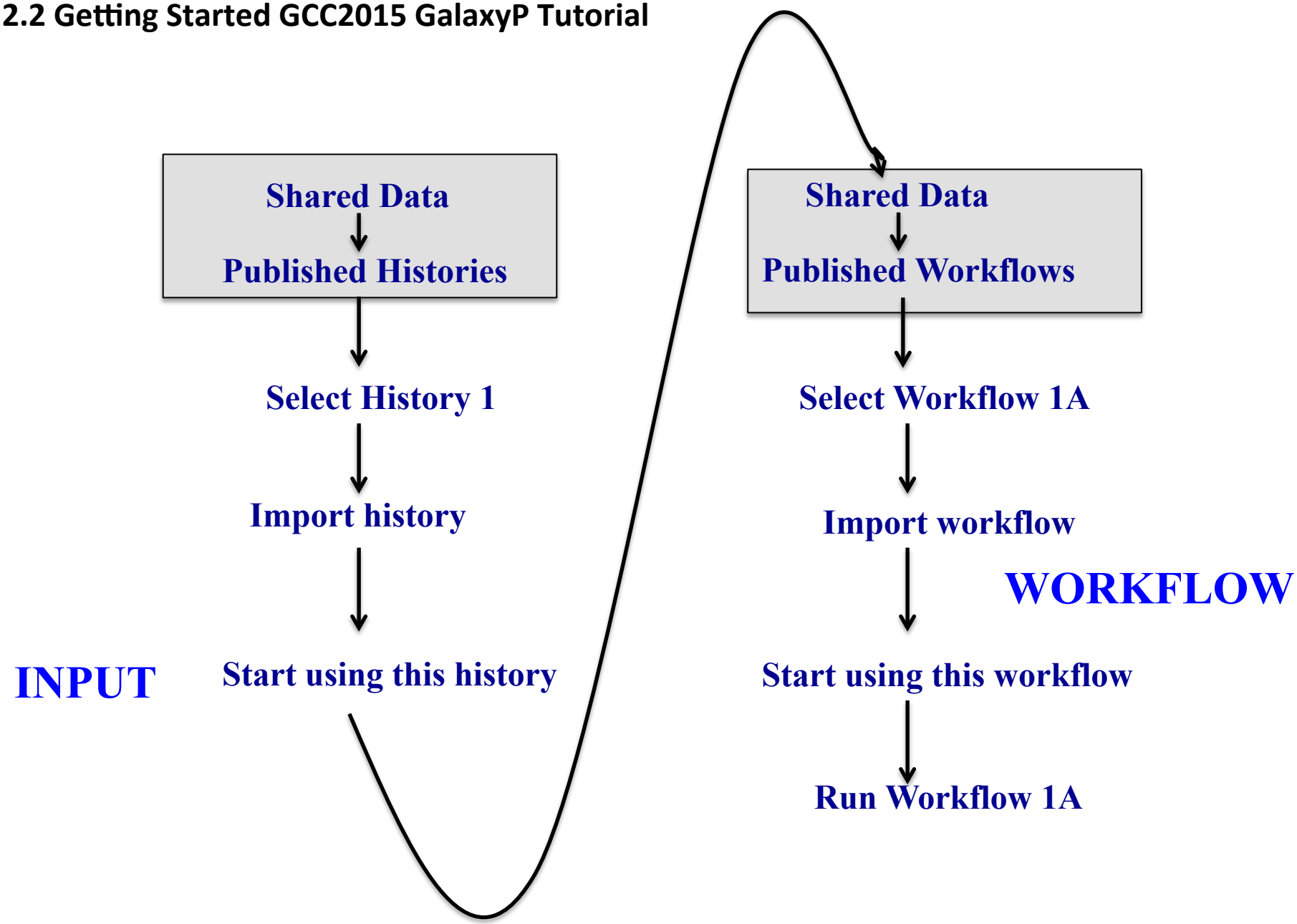
At the top of the screen select User and Register and Log in for your Cloud username (your email) and password.

2.2 Getting Started GCC2015 GalaxyP Tutorial

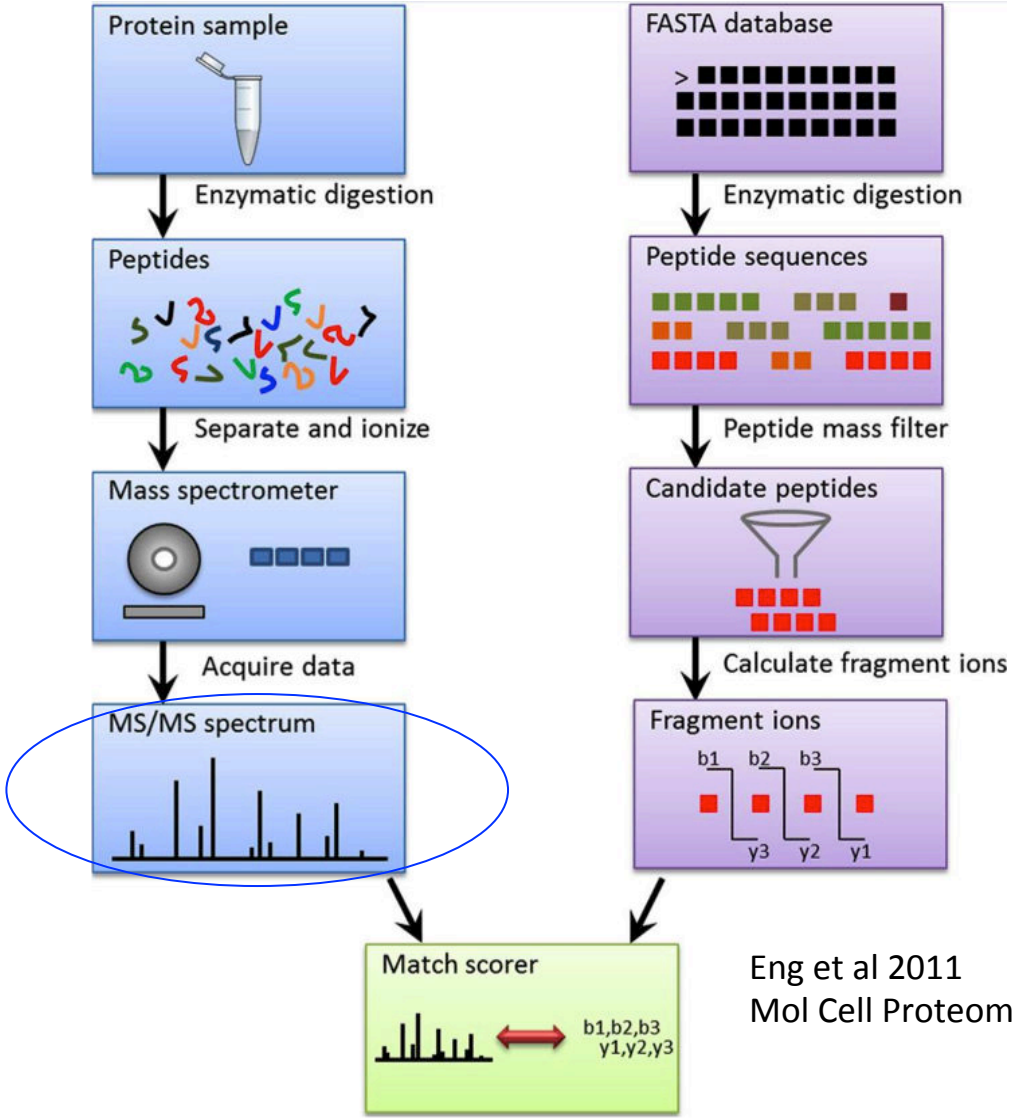
A. Open a web browser and navigate to the GalaxyP cloud instance - http://54.205.17.20/
B. At the top of the screen select User and Register and **Log in** for your Cloud username (your email) and password.

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.

INPUTS : PEAKLISTS and SEARCH db


Shared Data
Published Histories
Select History 1
Import history
Start using this history
INPUT
Shared Data
Published Workflows
Select Workflow 1A
Import workflow
Start using this workflow
Run Workflow 1A
WORKFLOW

PROTEOMICS WORKFLOW

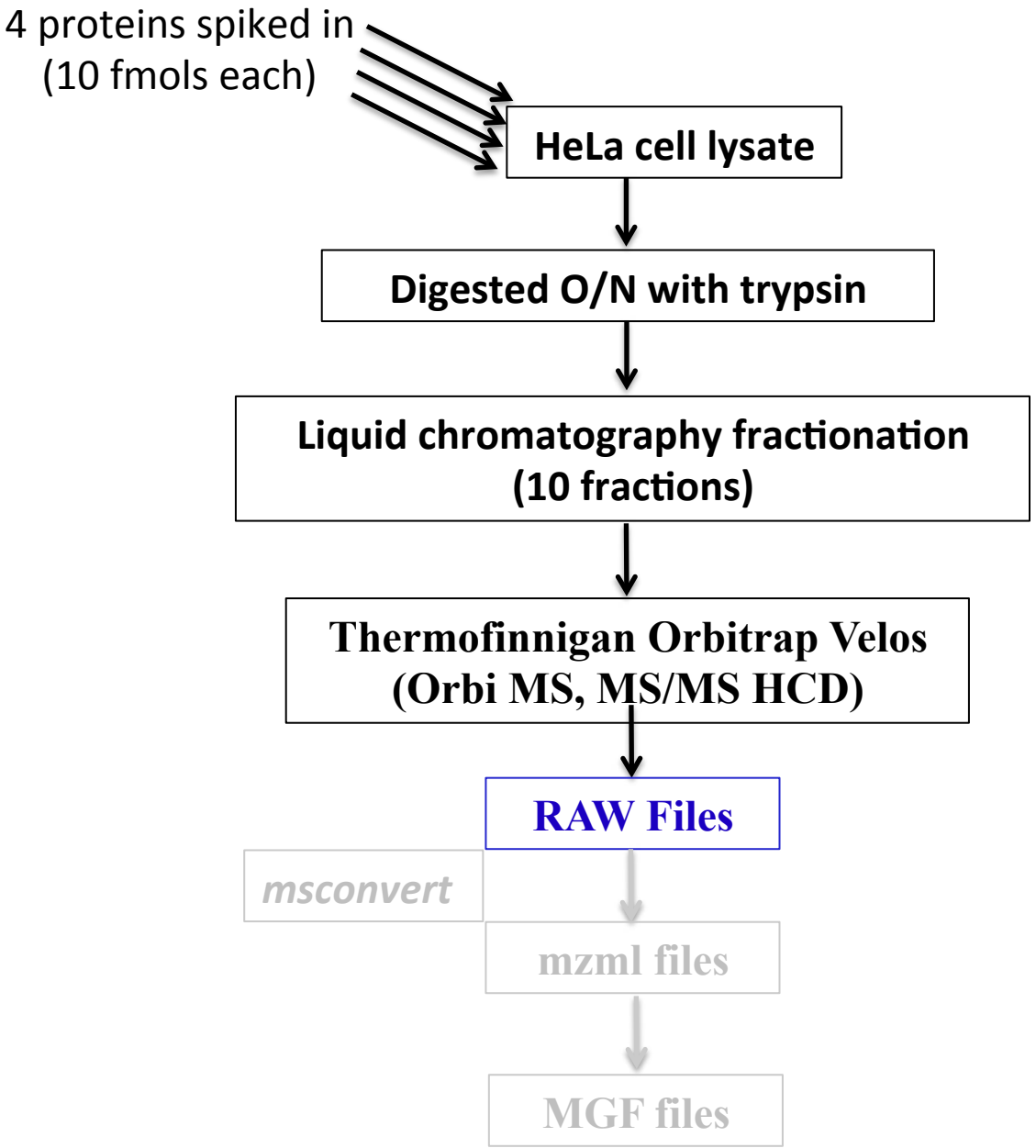


Eng et al 2011
Mol Cell Proteomics. 10(11): R111.009522.

INPUTS : PEAKLISTS and SEARCH db

GalaxyP

INPUTS : Mass spectral data and search database.

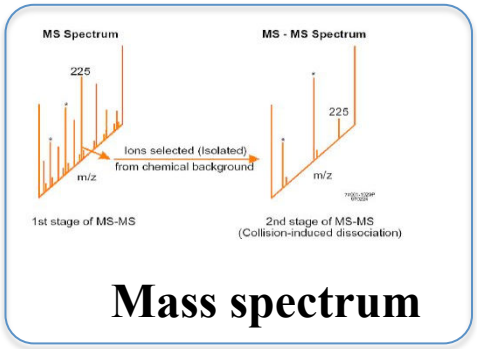


The dataset will be searched against FASTA database with human proteins, contaminant proteins, spiked in proteins and a subset of 3-frame translated cDNA database from EnSEMBL.

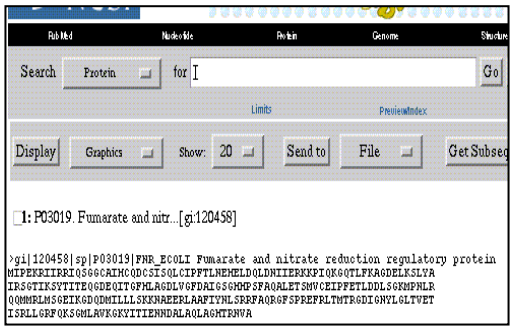
INPUTS: a) MGF formatter MGF files. (dataset collection)
b) ABRF-Spike4: FASTA sequences of 4 spiked in proteins.
c) FASTA File from EnSEMBL Searches: Subset of 3-frame translated cDNA database from EnSEMBL (our template for identifying novel proteoforms).
d) Human UniProt FASTA file + contaminant proteins.

INPUTS : PEAKLISTS and SEARCH db

SEARCH DATABASES

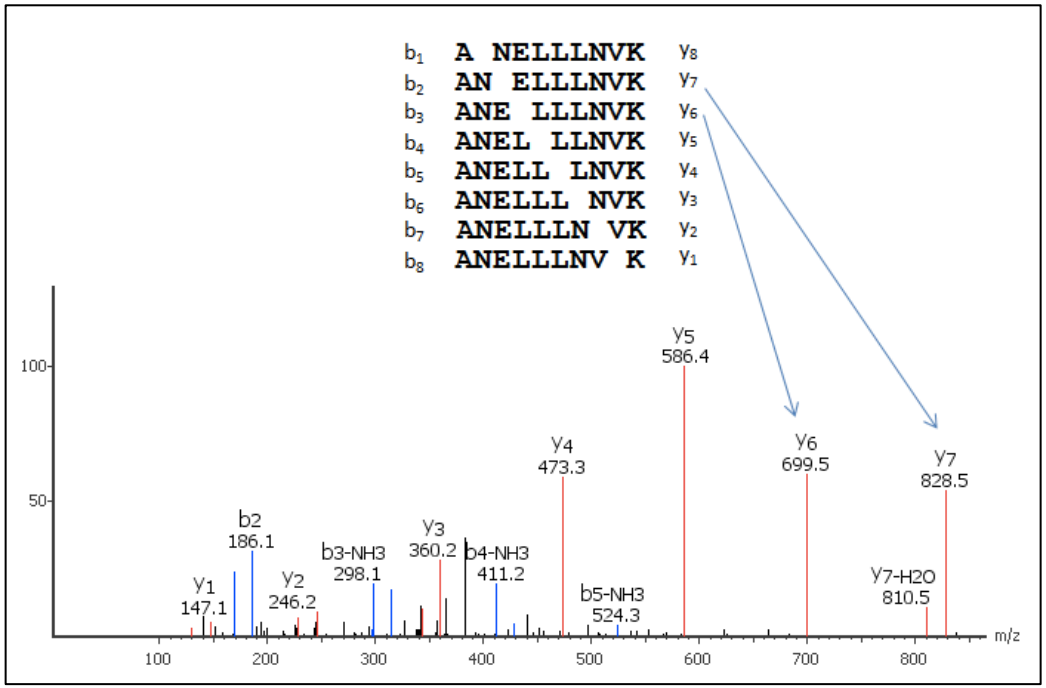


Mass spectrum



Reference Protein Database from genomic annotation

Peptide Spectral Match



INPUTS : PEAKLISTS and SEARCH db

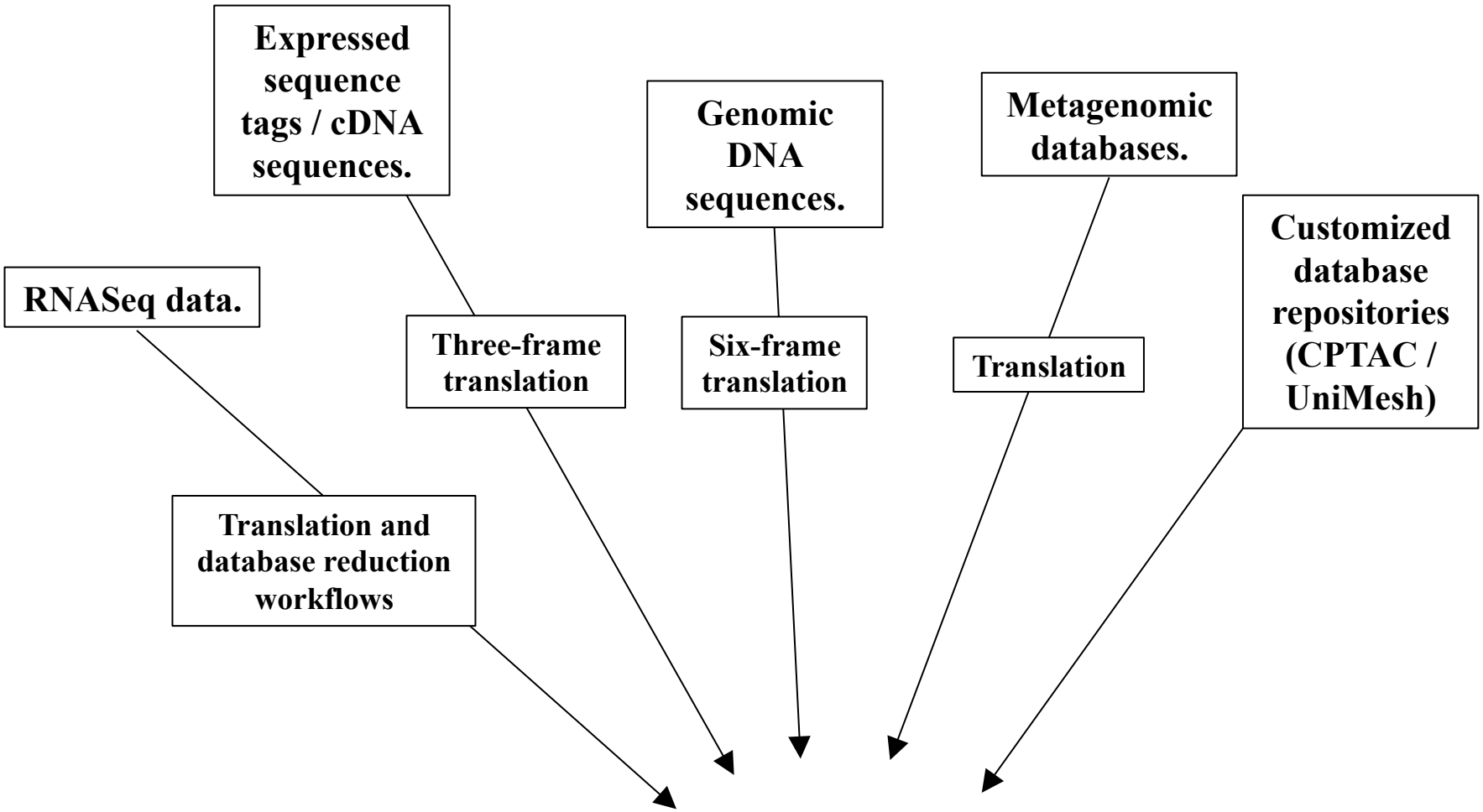
PROTEOMIC DATABASES

Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB).
It is a high quality annotated and non-redundant 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

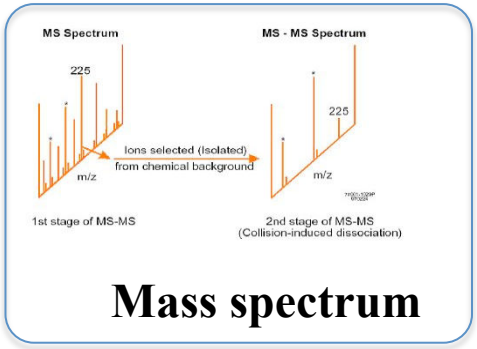
INPUTS : PEAKLISTS and SEARCH db

CUSTOMIZED PROTEOMIC DATABASES

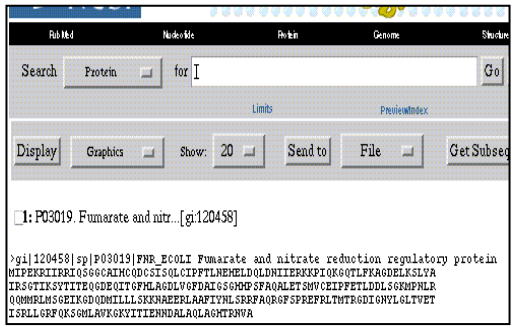


INPUTS : PEAKLISTS and SEARCH db

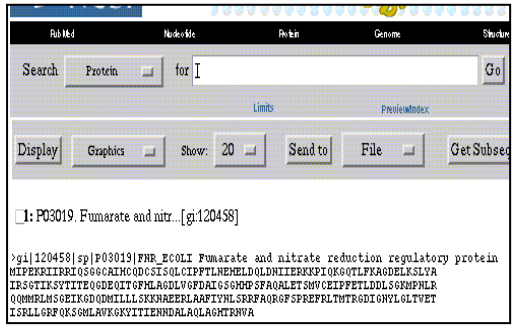
LOOKING BEYOND THE KNOWN PROTEOME



Mass spectrum



Reference Protein Database from genomic annotation



6-frame DNA sequences.
3-frame cDNA sequences.

Identification of peptides corresponding to novel proteoforms.

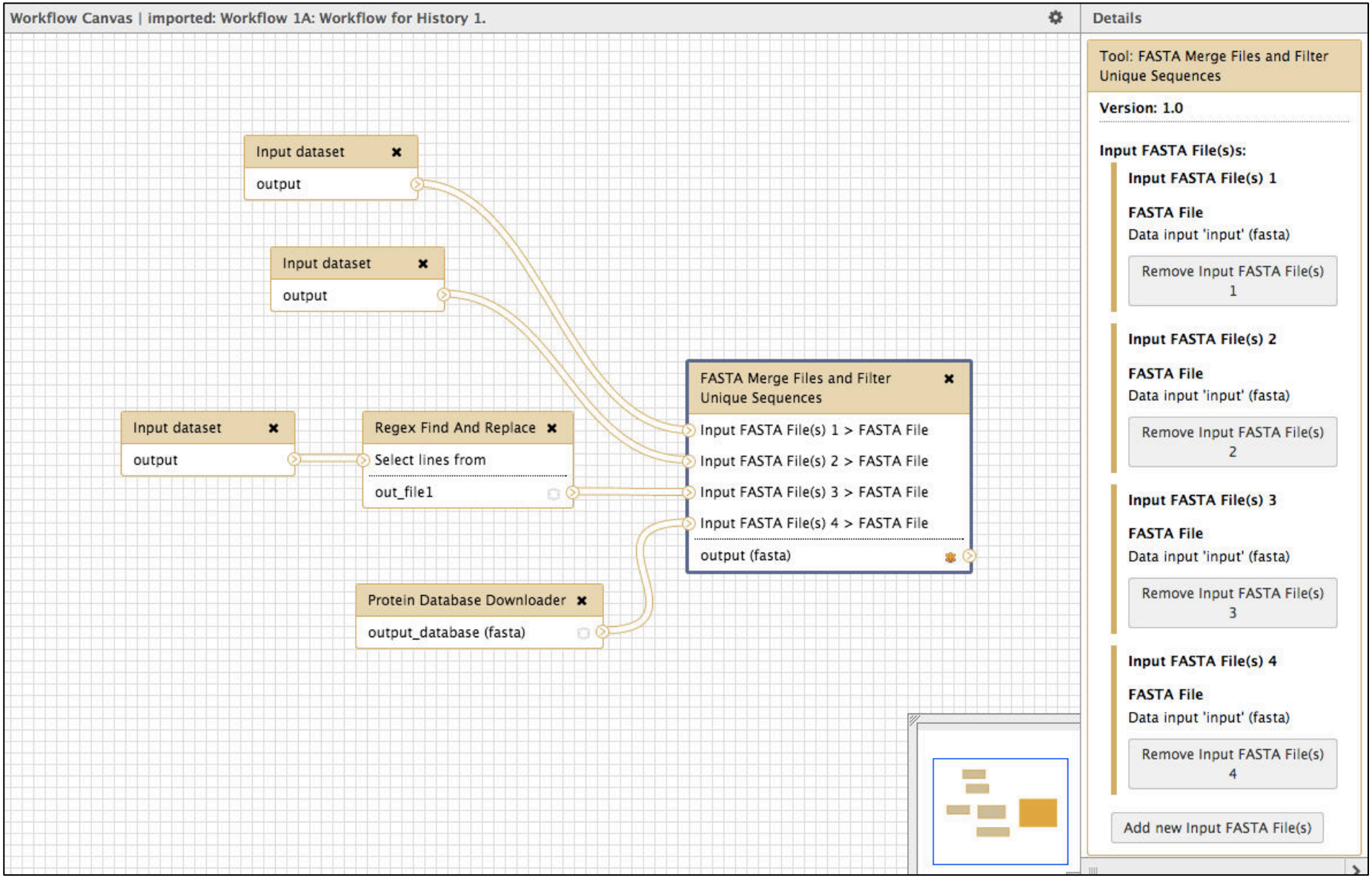
Cancer / Disease related Databases such as COSMIC, IARC p53, OMIM…

Deep genome sequencing data from ICGC, TCGA and CPTAC

RNASeq data (Customized OR Combined)

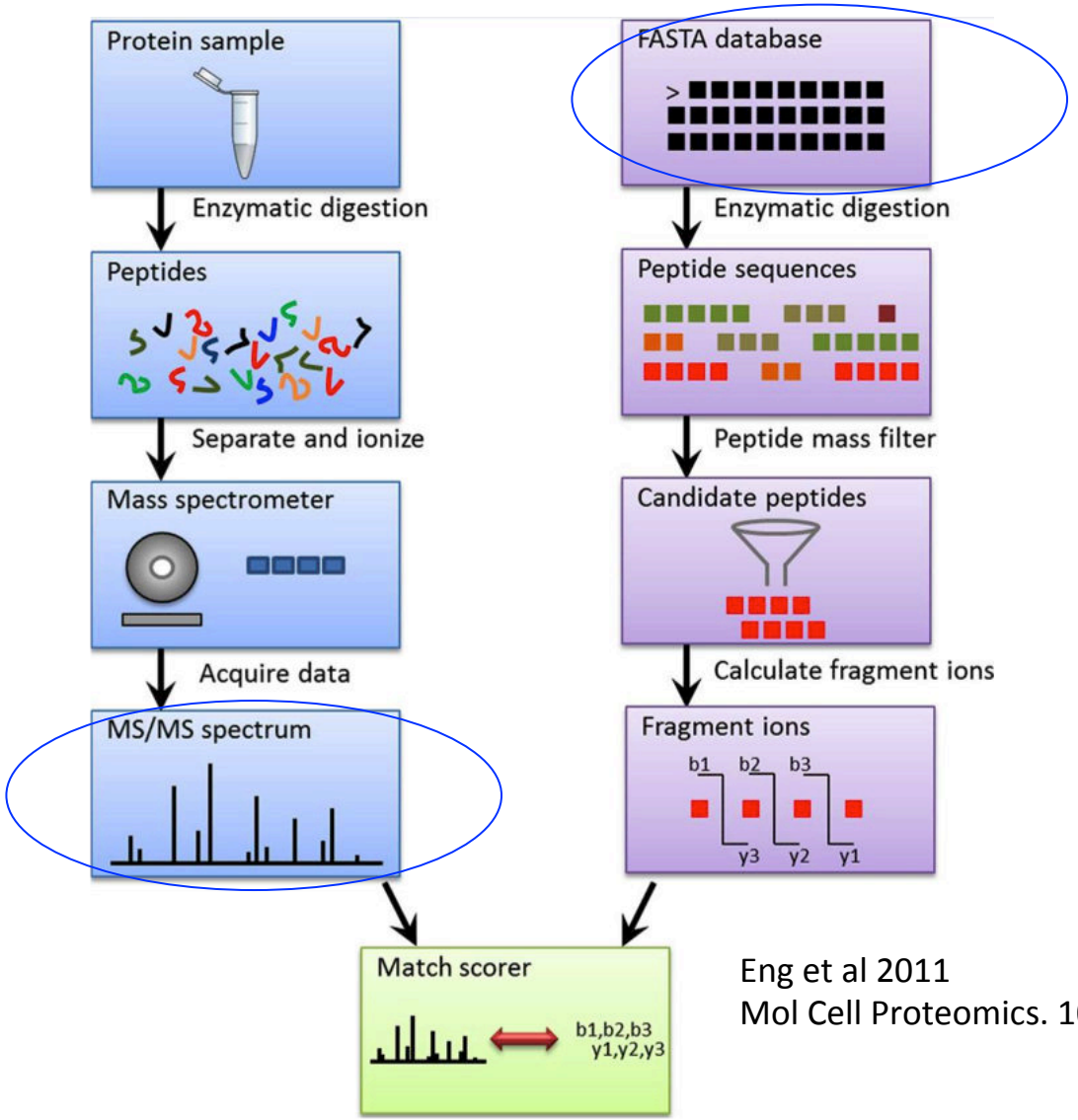
INPUTS : PEAKLISTS and SEARCH db

WORFLOW 1A



INPUTS : PEAKLISTS and SEARCH db

PROTEOMICS WORKFLOW

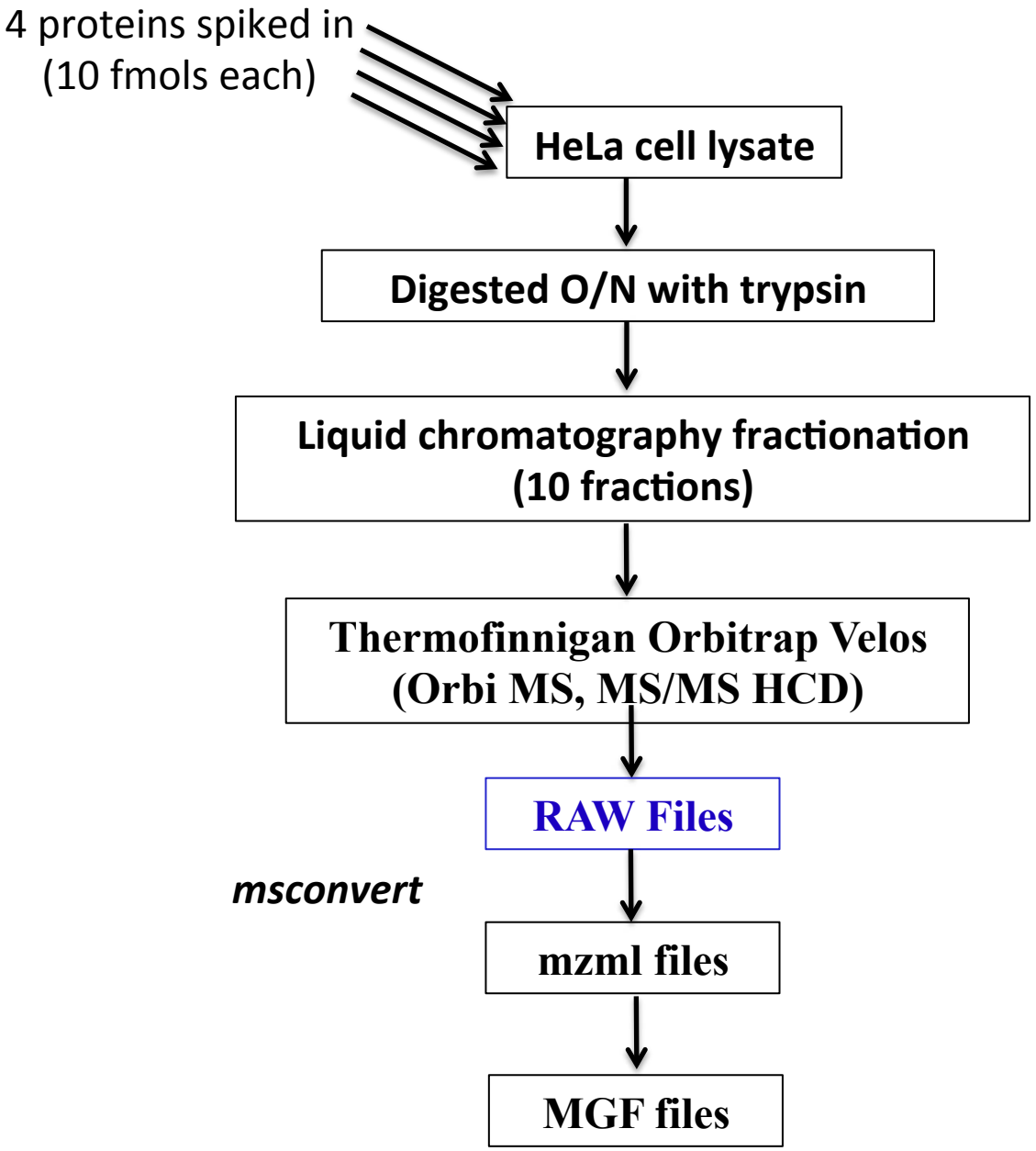


Eng et al 2011
Mol Cell Proteomics. 10(11): R111.009522.

INPUTS : PEAKLISTS and SEARCH db

GalaxyP

INPUTS : Mass spectral data and search database.

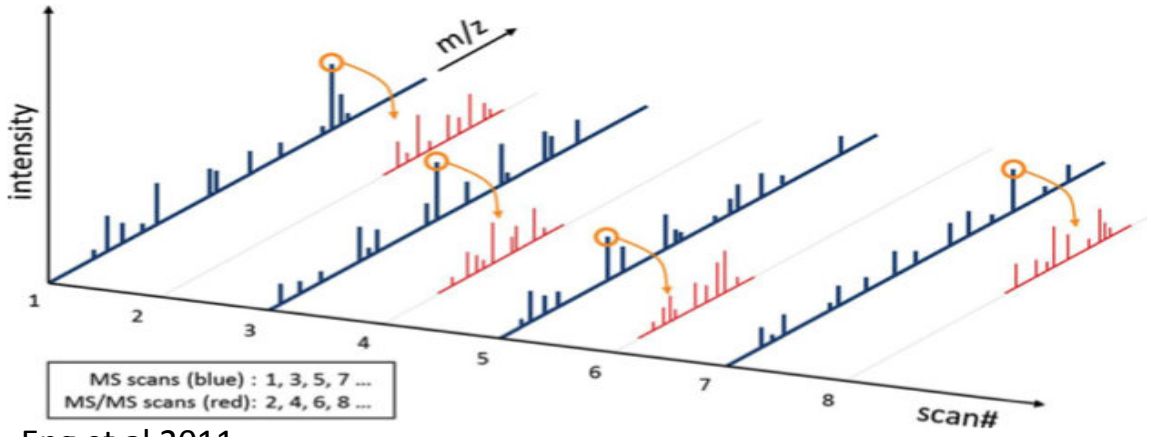


The dataset will be searched against FASTA database with human proteins, contaminant proteins, spiked in proteins and a subset of 3-frame translated cDNA database from EnSEMBL.

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.

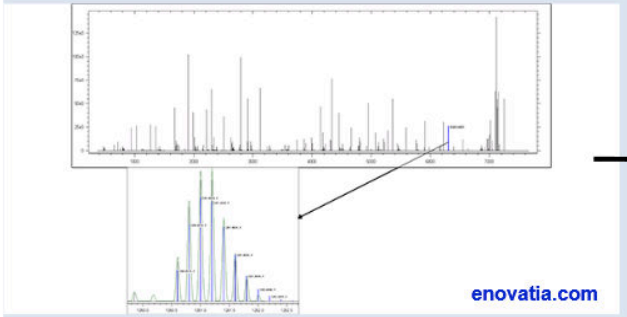
INPUTS : PEAKLISTS and SEARCH db

MASS SPECTRAL DATA



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

RAW DATA AND PEAKLIST



Rat9_SILAC (MaxQuant generated mgf)

- Peak detection
- Noise removal
- Baseline correction
- Monoisotoping
- Charge state

Raw Data
Peaks with wider distribution (Area under curve)

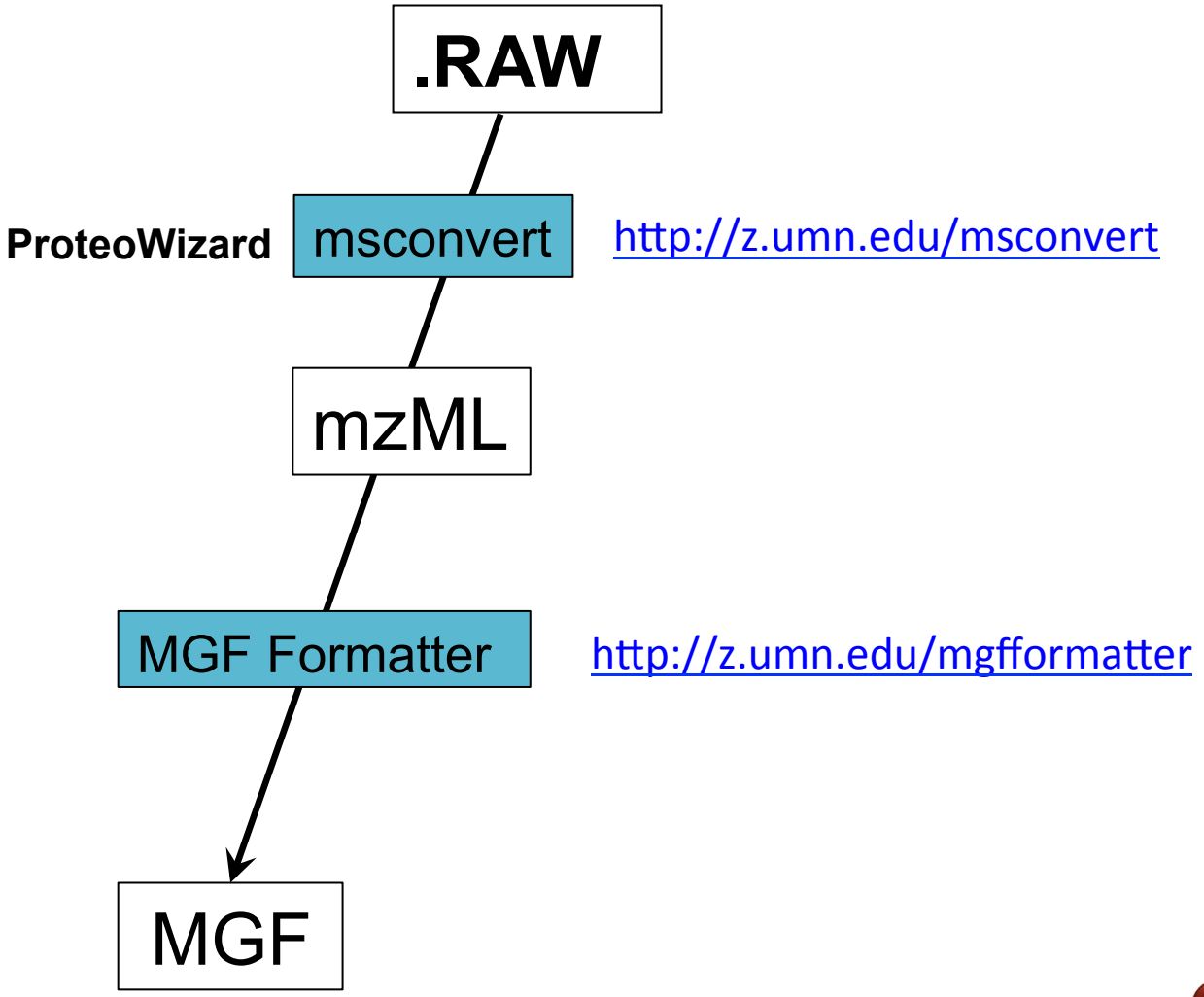
Additional information (Noise Peaks)

PeakList (Processed)
Peaks (centroid) with intensity values.

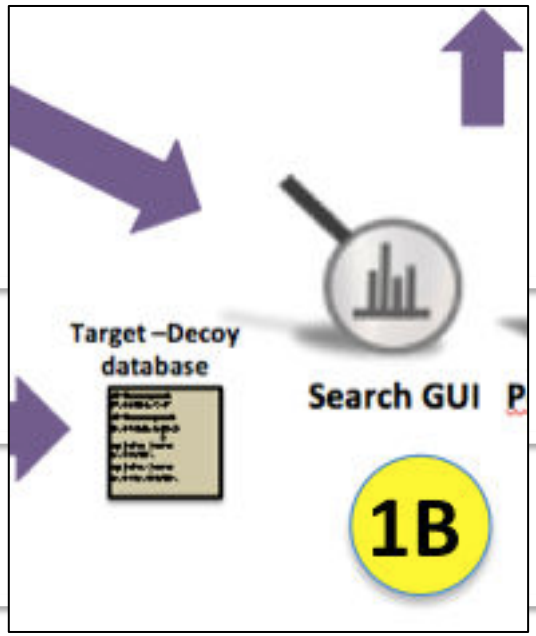
Removal of noise peaks.

INPUTS : PEAKLISTS and SEARCH db

RAW DATA CONVERSION TOOL



INPUTS : PEAKLISTS and SEARCH db

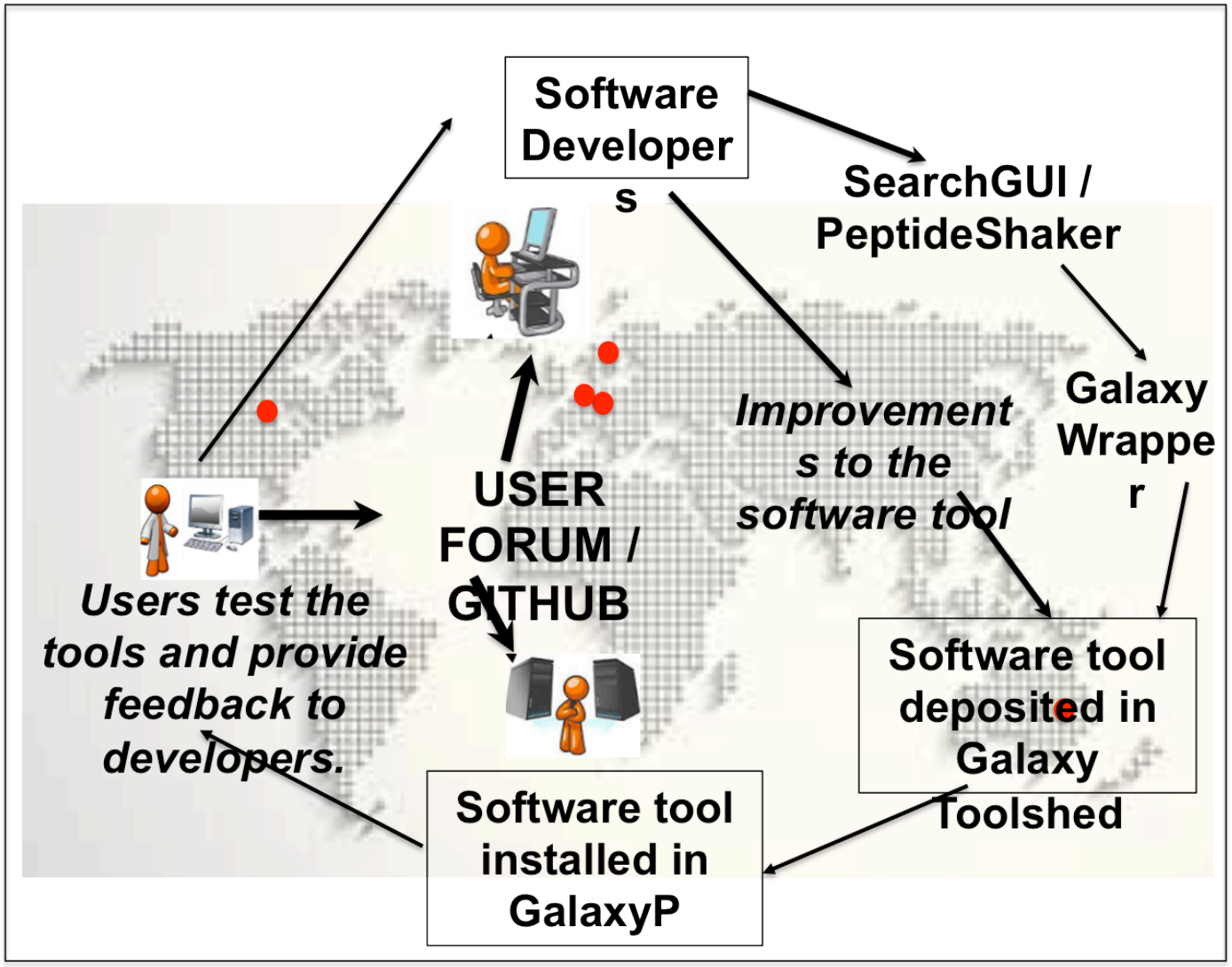


INPUTS : PEAKLISTS and SEARCH db → SEARCHGUI

SEARCHGUI : SEARCH ALGORITHMS

INPUTS : PEAKLISTS and SEARCH db

COMMUNITY-BASED SOFTWARE DEVELOPMENT

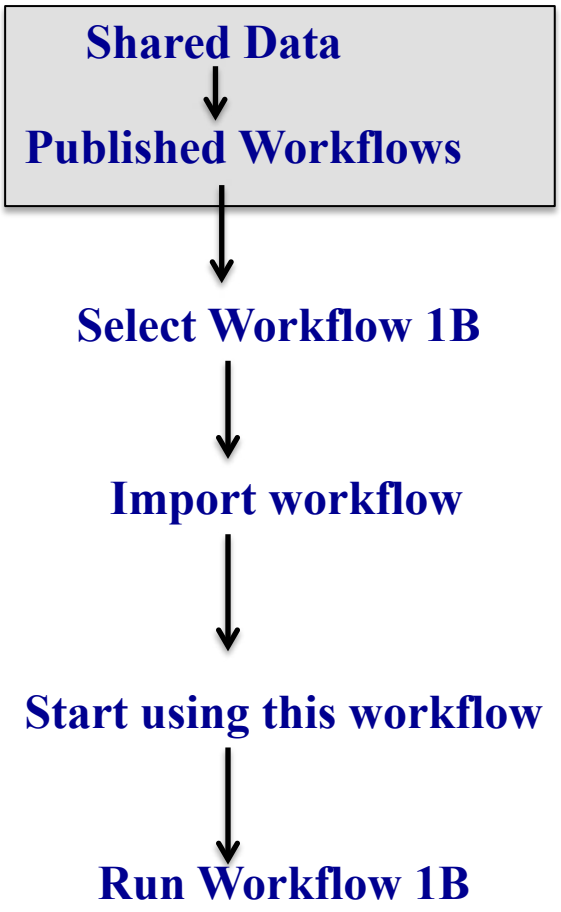


INPUTS : PEAKLISTS and SEARCH db → SEARCHGUI

3.2 SearchGUI in GalaxyP

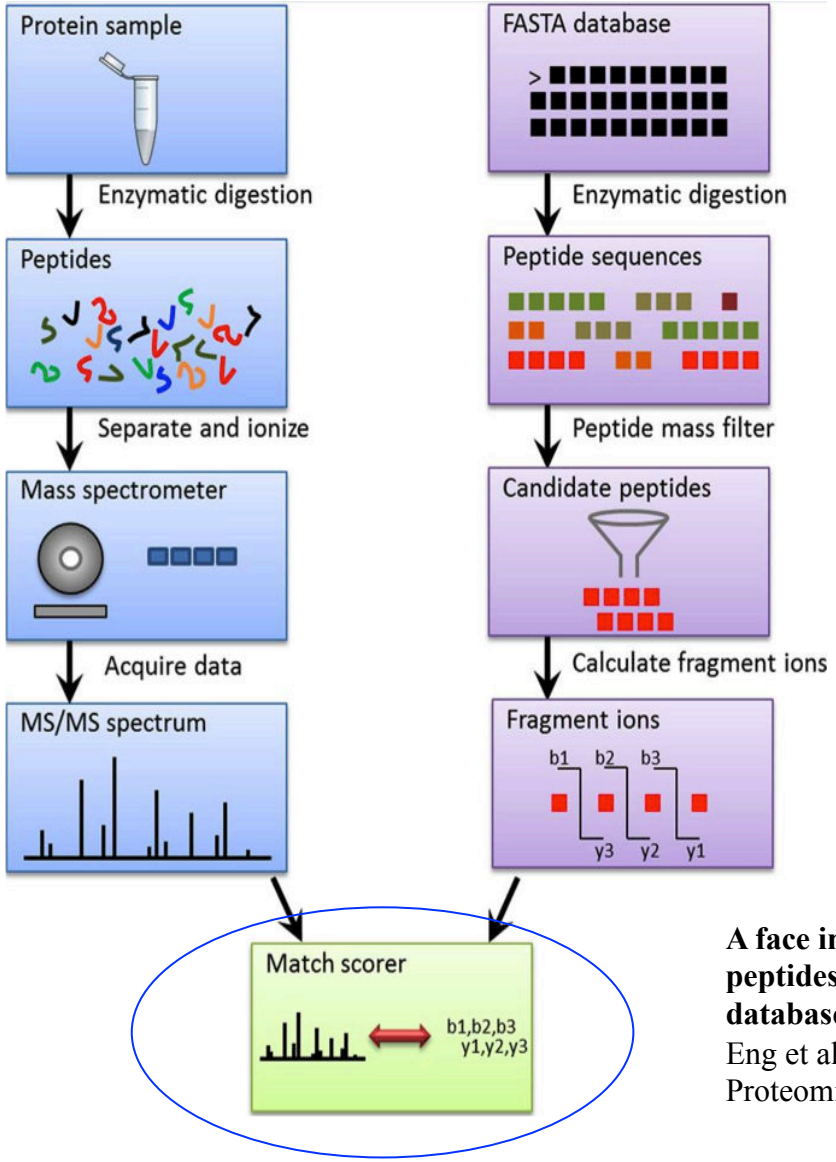
- From Published Workflows, import Workflow 1B: Workflow for History 1 to History 2.
- On the confirmation screen, select start using this workflow to navigate to your user workflows.
- Run Workflow 1B on your current history (History 1).
- Appropriately select each dataset from History 1 that corresponds 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



WORKFLOW

PROTEOMICS WORKFLOW

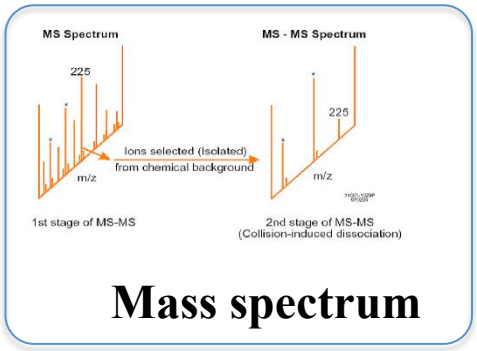


A face in the crowd: recognizing peptides through database search.
Eng et al 2011 Mol Cell Proteomics. 10(11)

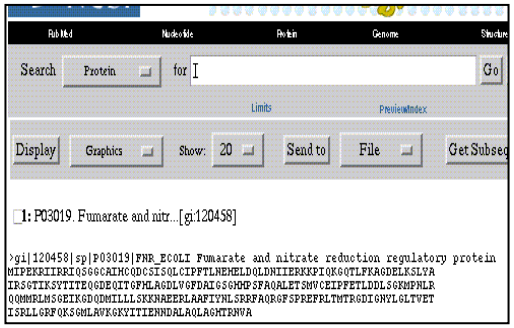
INPUTS : PEAKLISTS and SEARCH db **SEARCHGUI**

GalaxyP

DATABASE SEARCH

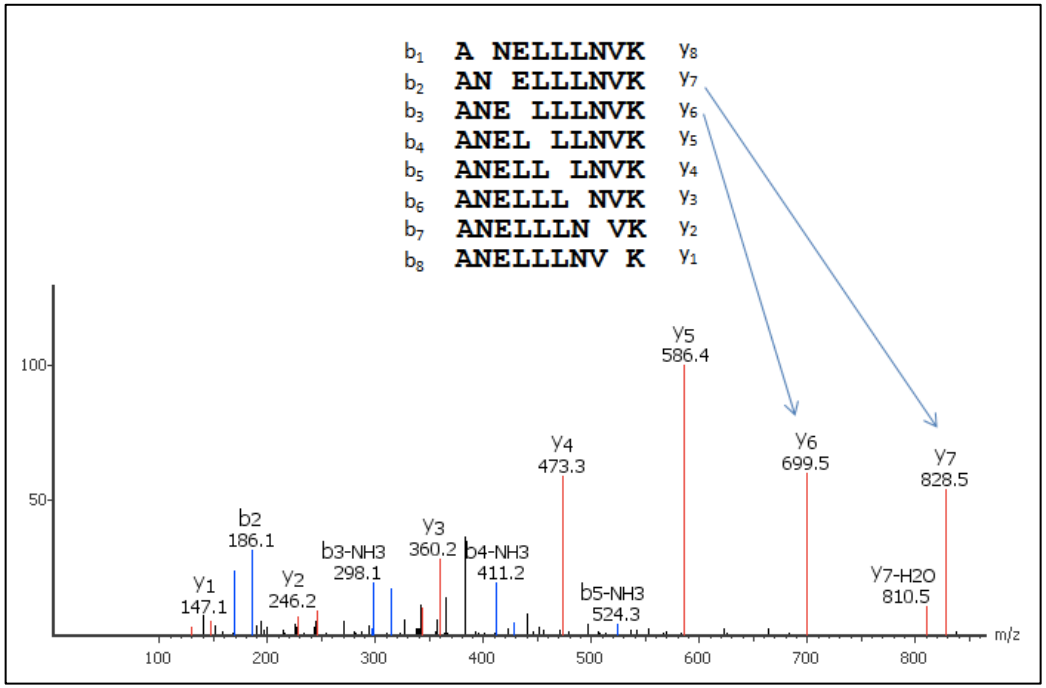


Mass spectrum



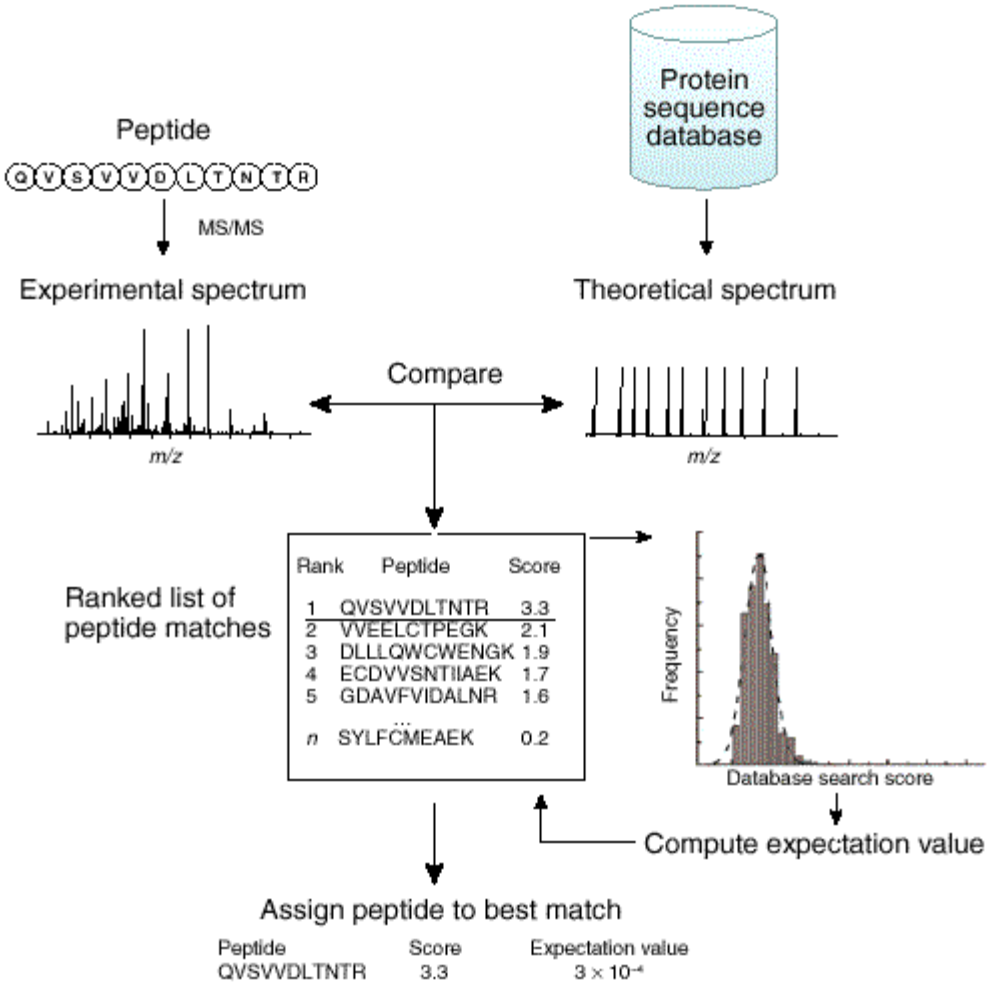
Reference Protein Database from genomic annotation

Peptide Spectral Match



INPUTS : PEAKLISTS and SEARCH db **SEARCHGUI**

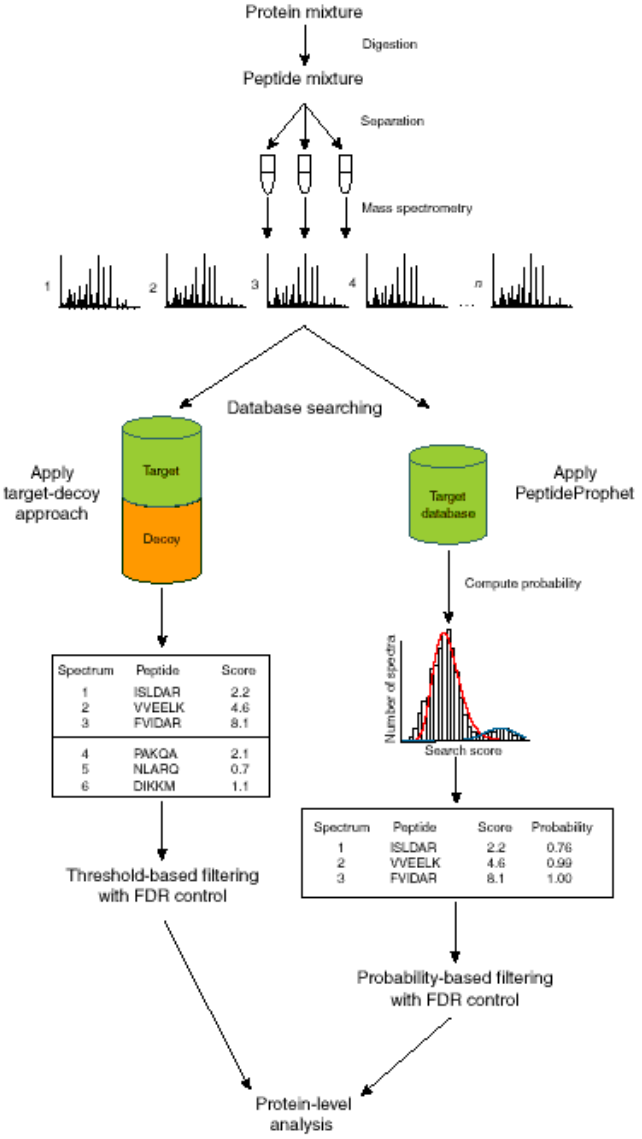
DATABASE SEARCH



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

INPUTS : PEAKLISTS and SEARCH db **SEARCHGUI**

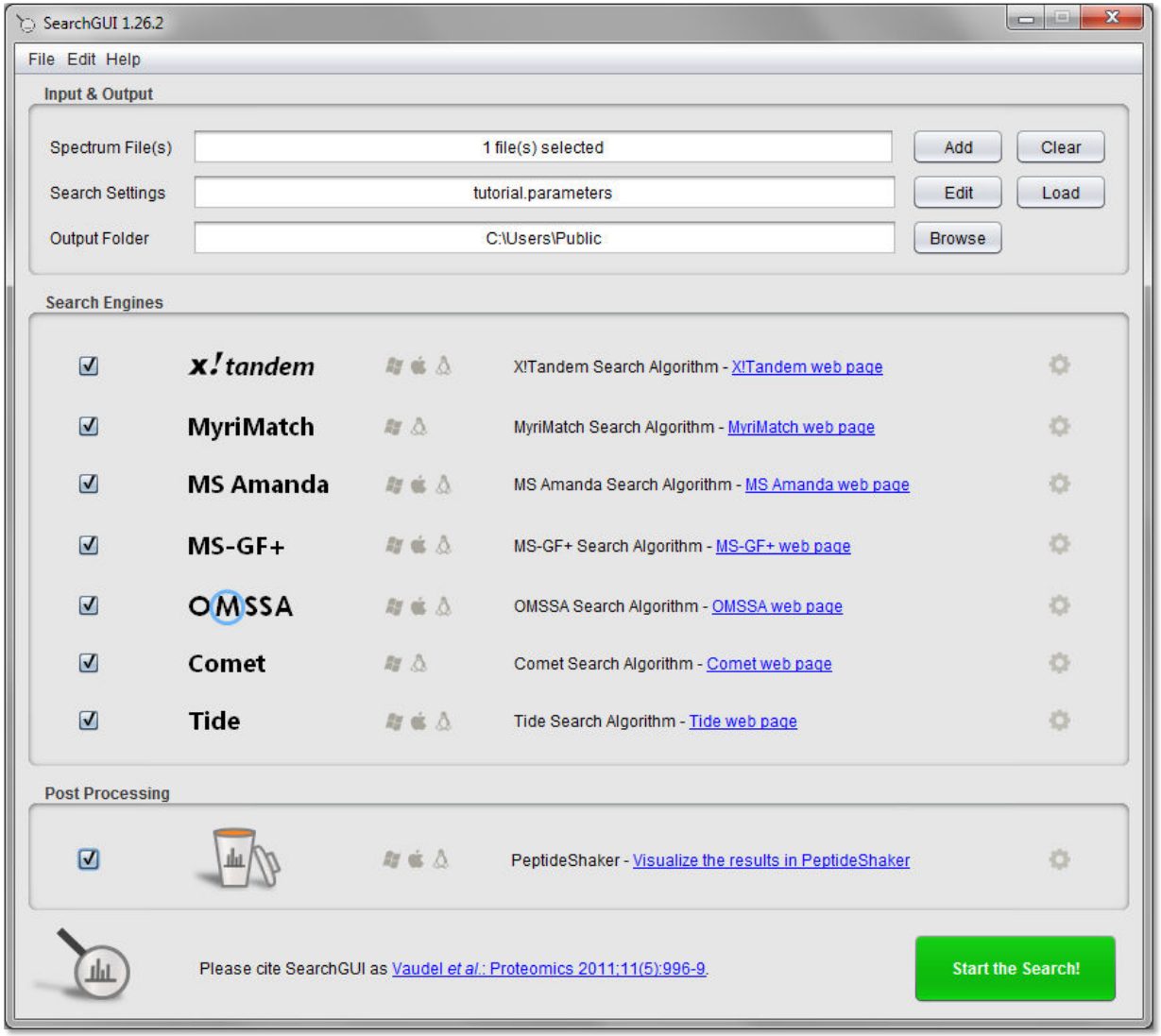
Database Search



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

INPUTS : PEAKLISTS and SEARCH db **SEARCHGUI**

SEARCHGUI

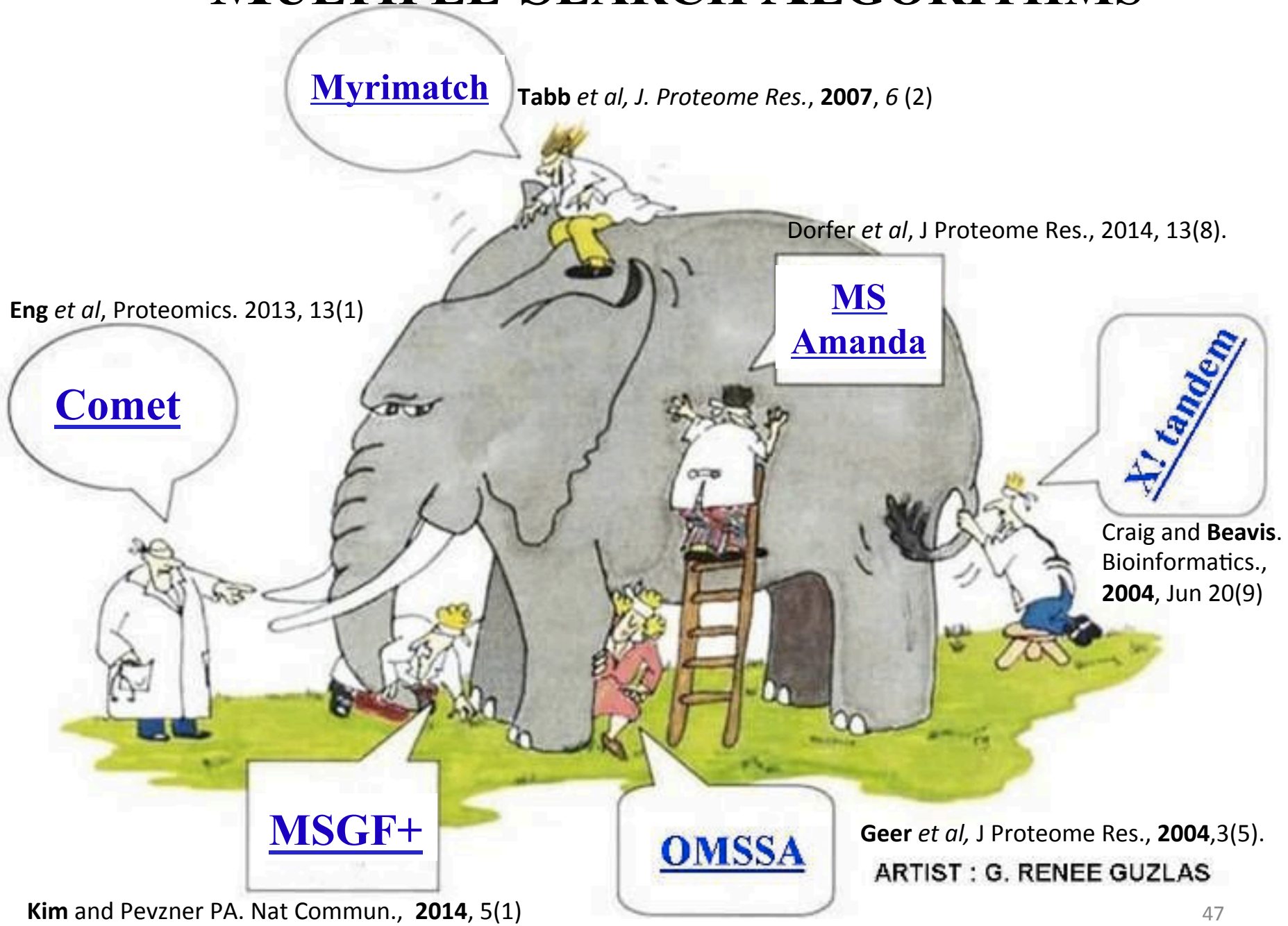


Vaudel M. *et al* Proteomics (2011) 11(5)

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

INPUTS : PEAKLISTS and SEARCH db **SEARCHGUI**

Multiple Search Algorithms



Tabb *et al, J. Proteome Res.*, **2007**, *6* (2)

Dorfer *et al*, J Proteome Res., 2014, 13(8).

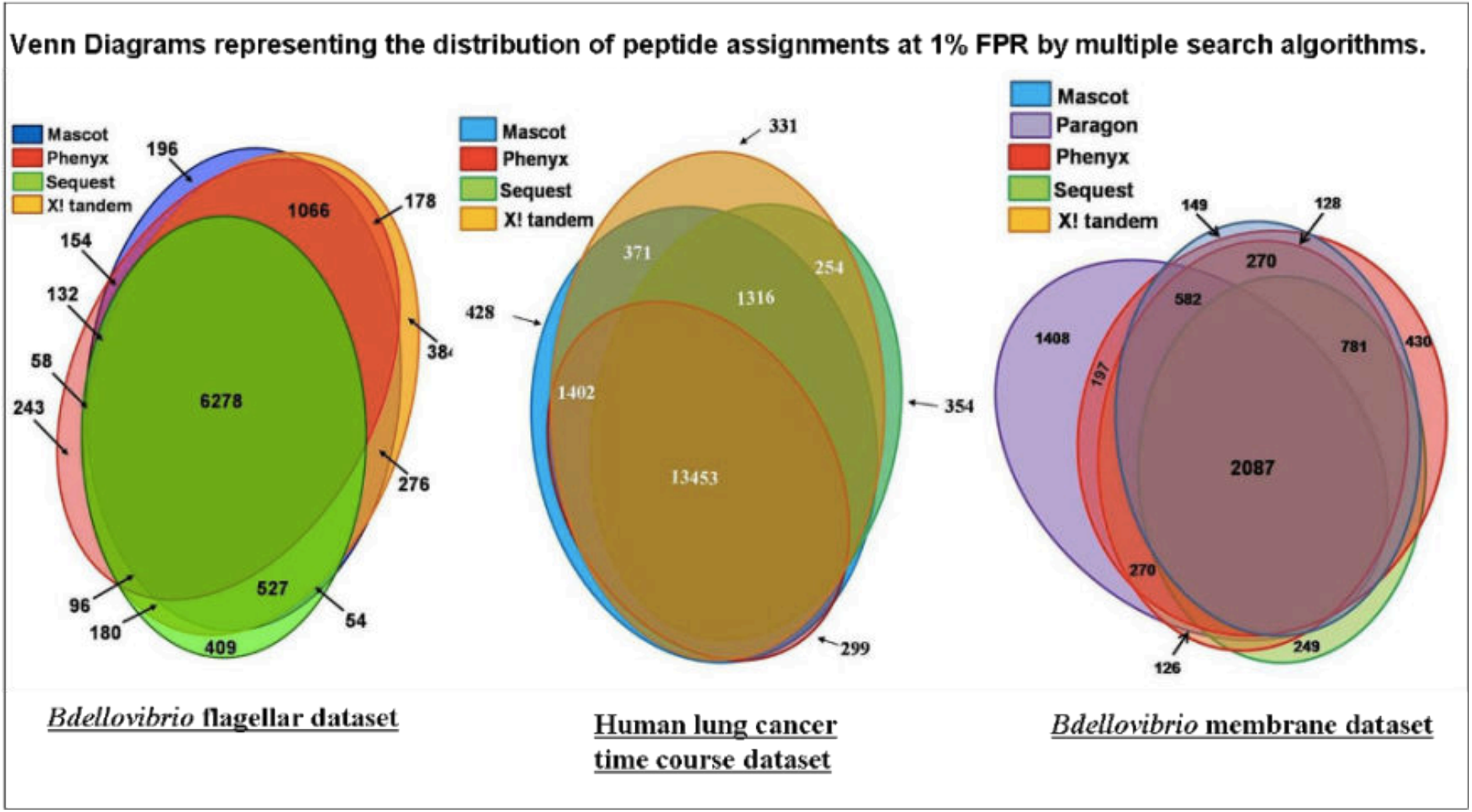
Eng *et al*, Proteomics. 2013, 13(1)

Craig and **Beavis**. Bioinformatics., **2004**, Jun 20(9)

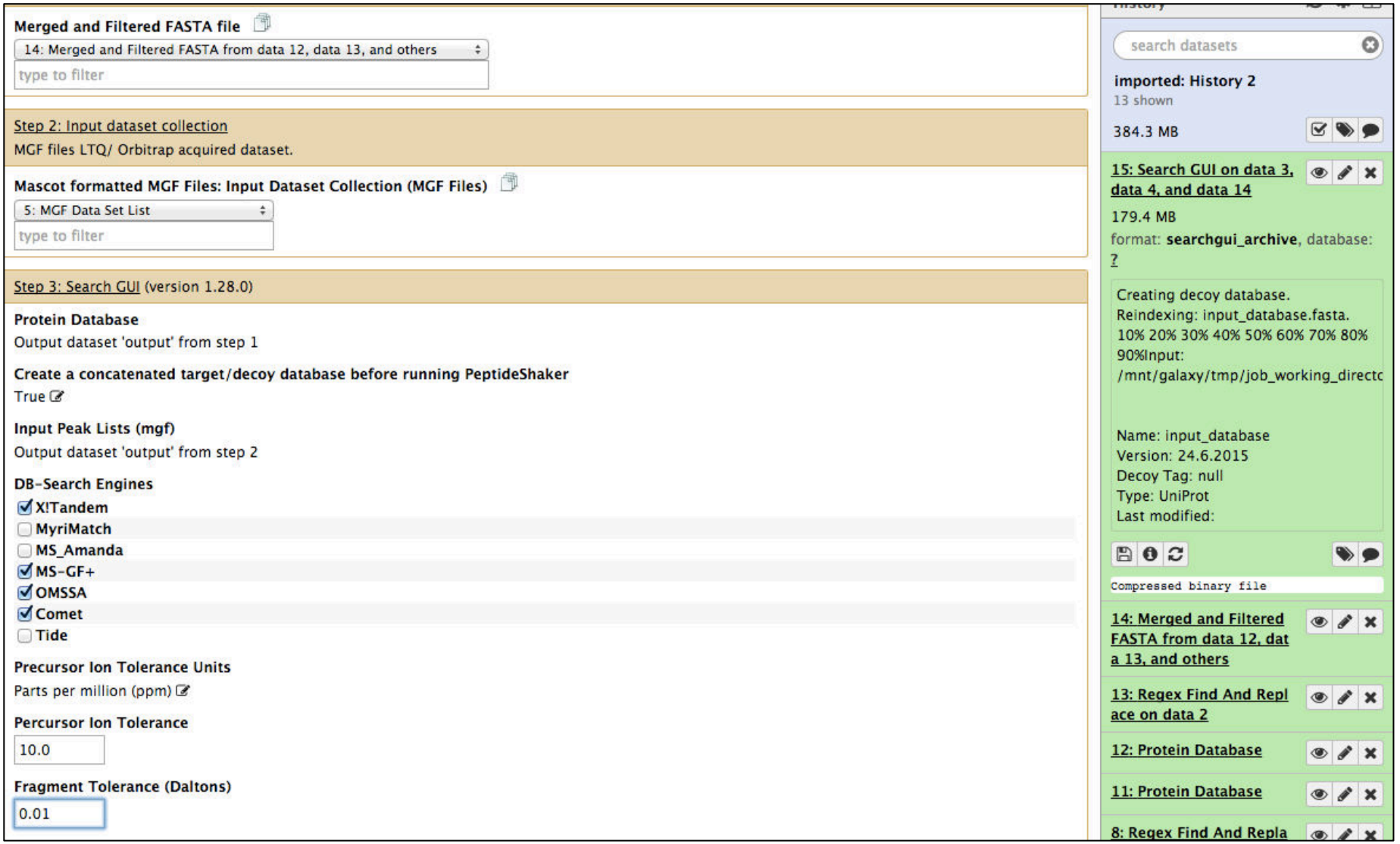
Geer *et al,* J Proteome Res., **2004**,3(5).

Kim and Pevzner PA. Nat Commun., **2014**, 5(1)

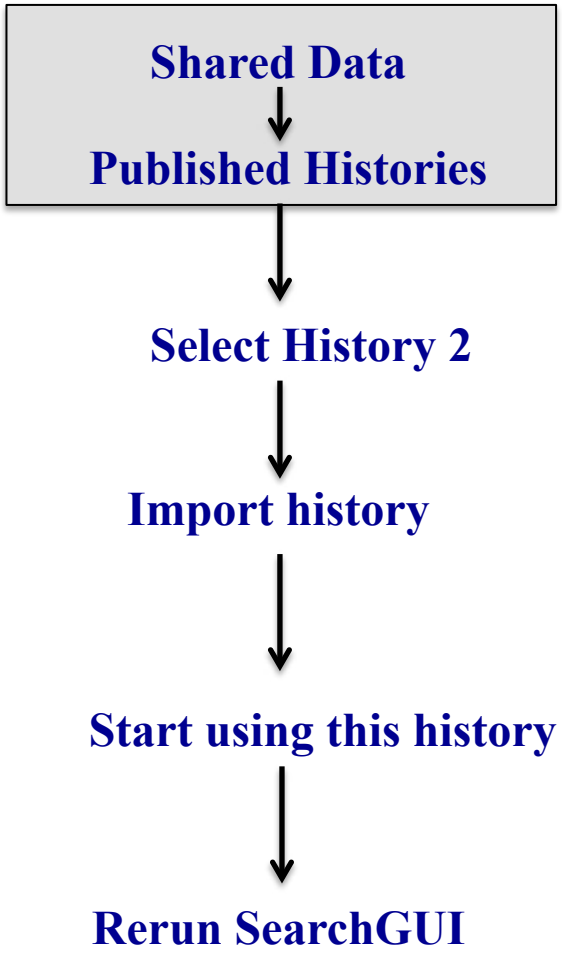
Multiple Search Algorithms

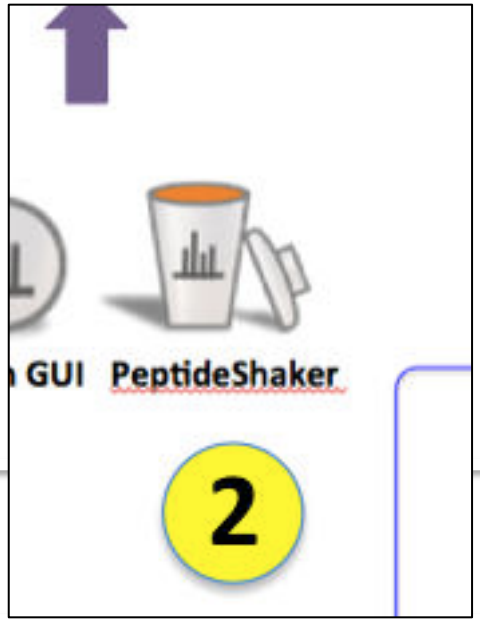


Visualizing parameters for SearchGUI analysis



Visualizing parameters for SearchGUI analysis





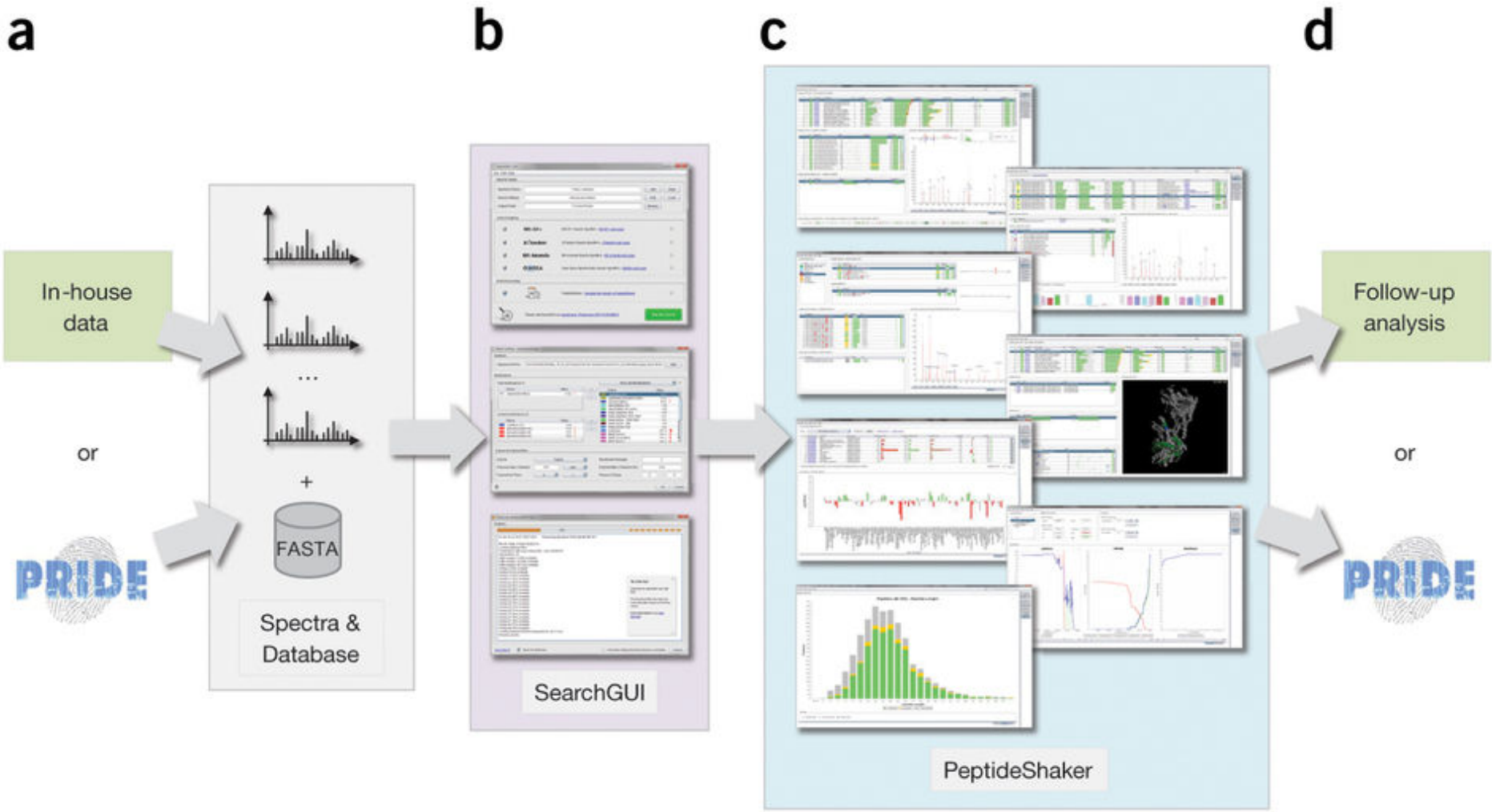
INPUTS : PEAKLISTS and SEARCH db → SEARCHGUI → PEPTIDESHAKER

PEPTIDESHAKER

SEARCHGUI : SEARCH ALGORITHMS

INPUTS : PEAKLISTS and SEARCH db

PEPTIDESHAKER

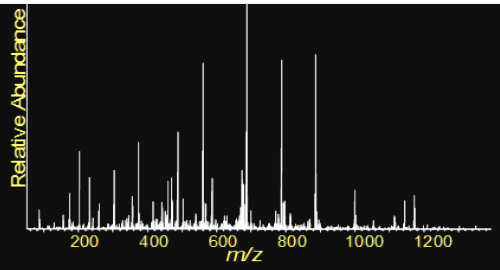


Vaudel *et al Nature Biotechnology,* 33, (2015)

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

High throughput identification of proteins

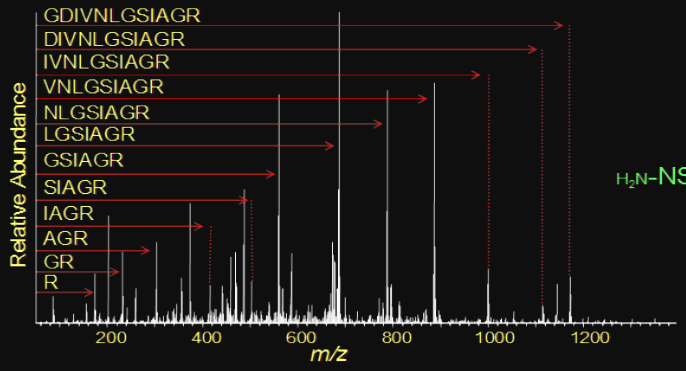
Raw MS/MS spectrum



Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match



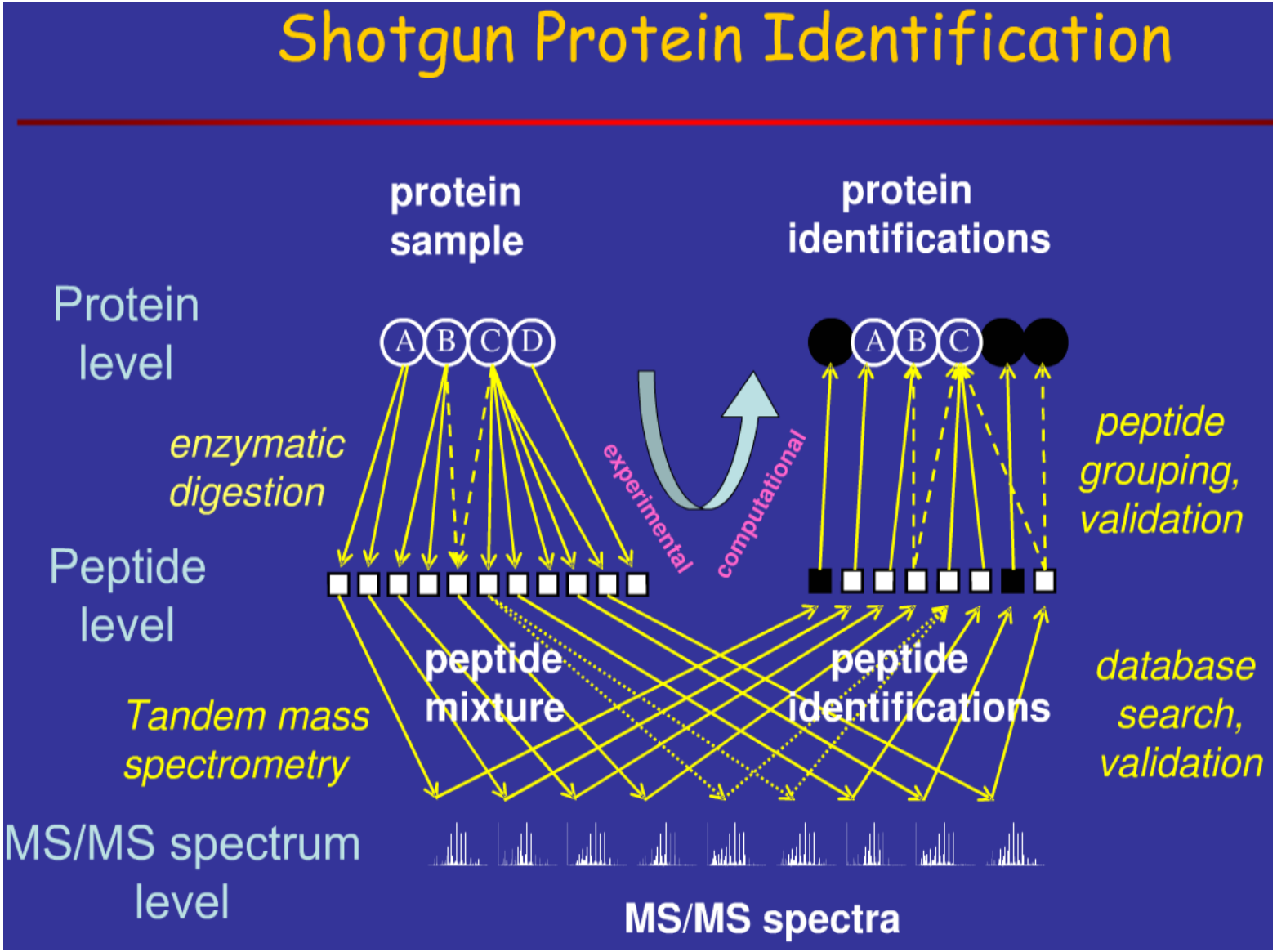
H_2N-NSGDIVNLGSIAGR-COOH

Protein identification

INPUTS : PEAKLISTS and SEARCH db → SEARCHGUI → PEPTIDESHAKER

GalaxyP

PEPTIDESHAKER : PROTEIN INFERENCE

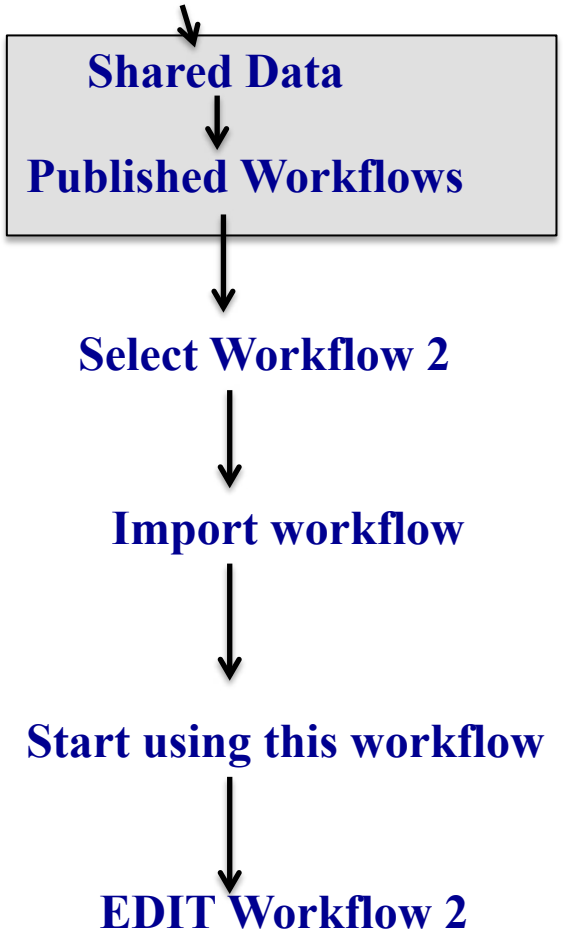


Slide from Alexey Nesvizshkii talk at http://www.scivee.tv/node/12671

4.3 Peptide Shaker in GalaxyP

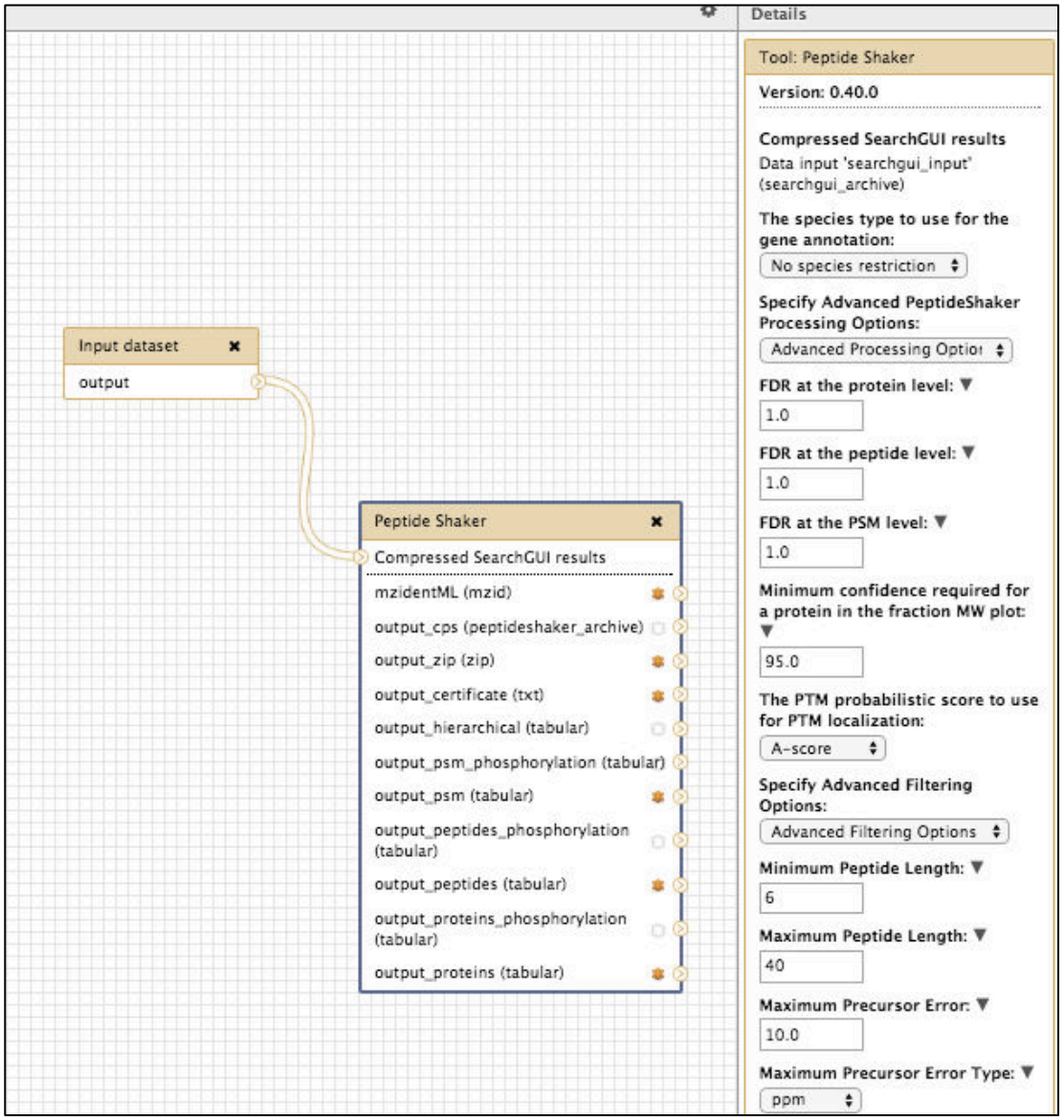
A. At the top of the screen, click Shared Data then select the option: Published Workflows .
B. Click 'Workflow2: Workflow for History 2 to History 3' and choose import.
C. Select start using this workflow.
D. Click imported Workflow 2: Workflow for History 2 to History 3' and choose edit to
open the 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



WORKFLOW

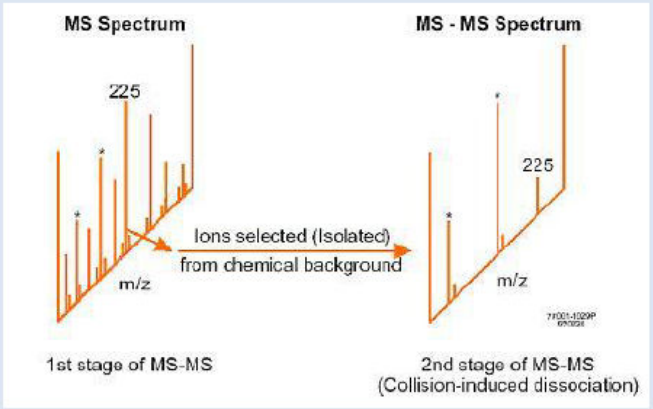
4.3 Peptide Shaker in GalaxyP



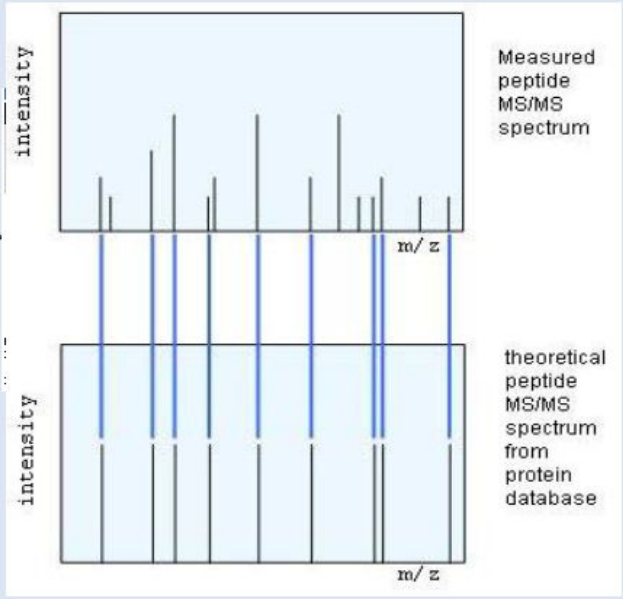
INPUTS : PEAKLISTS and SEARCH db → SEARCHGUI → PEPTIDESHAKER

PEPTIDESHAKER : TARGET-DECOY SEARCH

REVERSE DATABASE SEARCH



Mass spectrum

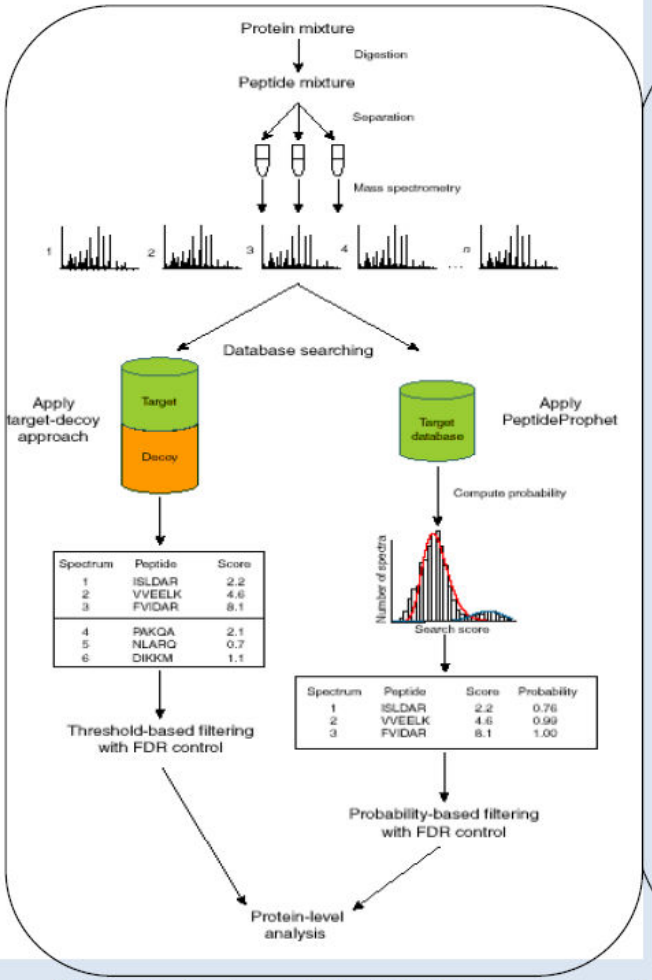


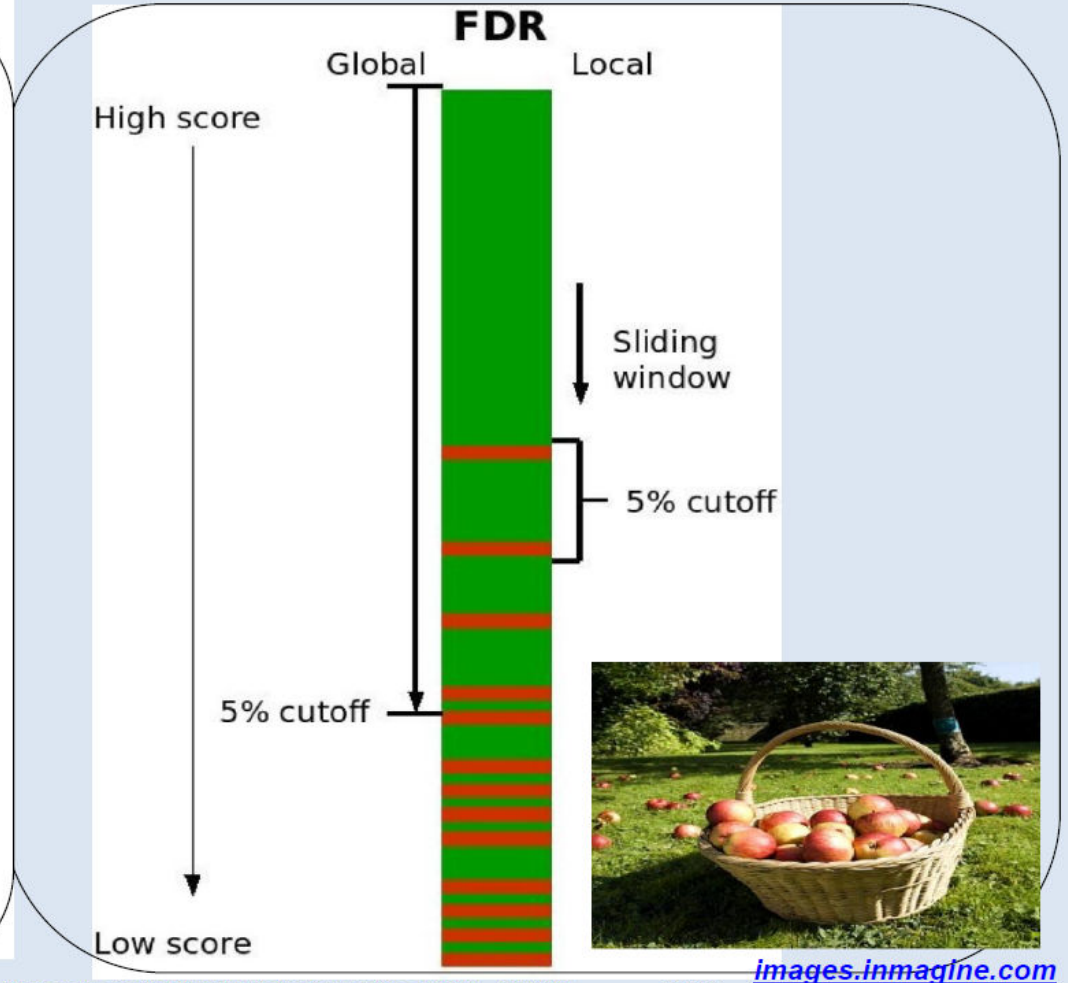
✓ >IPI:IPI00205563.1|Gene_Symbol=Tmsbl1 thymosin beta-like protein
MSDKPDLSEVETFDKSKLKKTNTEEKNTLPSKETIQQEKEYNQRS

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

PEPTIDESHAKER : TARGET-DECOY SEARCH

FALSE DISCOVERY RATE ANALYSIS





Nature Methods - 4, 787 - 797 (2007)

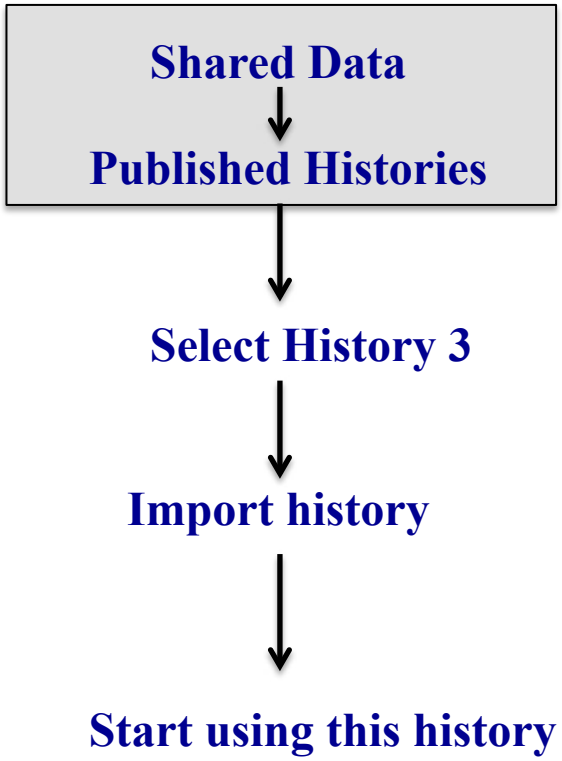
Sennels et al BMC Bioinformatics 2009, 10:179

images.inmagine.com

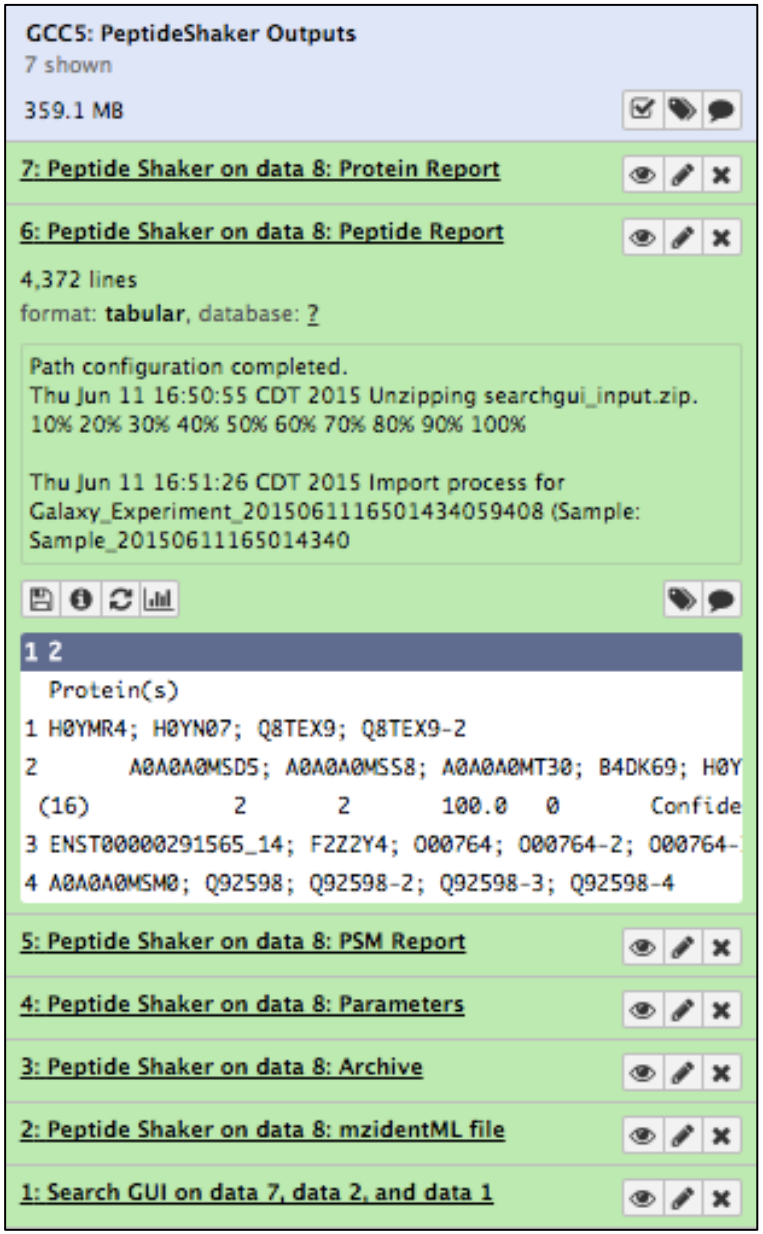
INPUTS : PEAKLISTS and SEARCH db → SEARCHGUI → PEPTIDESHAKER

GalaxyP

PEPTIDESHAKER: OUTPUTS



PEPTIDESHAKER: OUTPUTS



PEPTIDESHAKER: OUTPUTS

Protein Report
- Valid proteins
- Coverage
- Mw

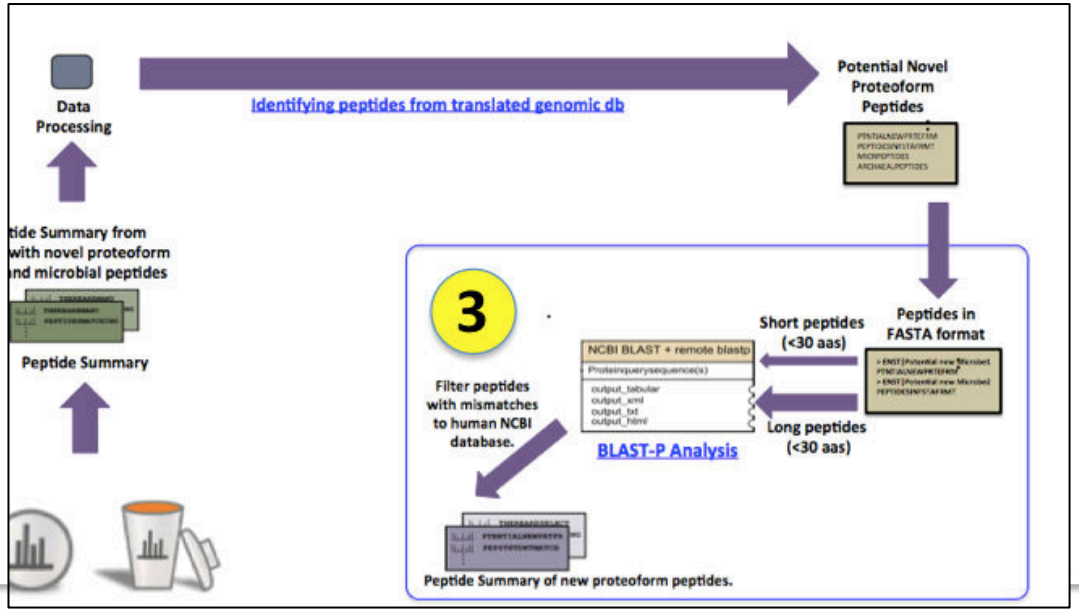
Peptide Report
- Valid peptides
- Potential novel proteoforms based on accession numbers.
- Sequences
- Modifications and localization score
- Confidence

Spectrum Report
- Valid spectra
- Potential novel proteoforms based on accession numbers.
- Sequences
- Modifications and localization score
- Confidence
- m/z, charge state, Delta m/z

Mzid
-PSM Visualization
- SWATH Analysis
- Skyline
- Scaffold

Summary
- Valid peptides
- Valid proteins
- Valid spectra

Archive (Zipped file)
- CPS file to visualize data



INPUTS → SEARCHGUI → PEPTIDESHAKER → BLAST-P

BLAST-P SEARCH

PEPTIDESHAKER

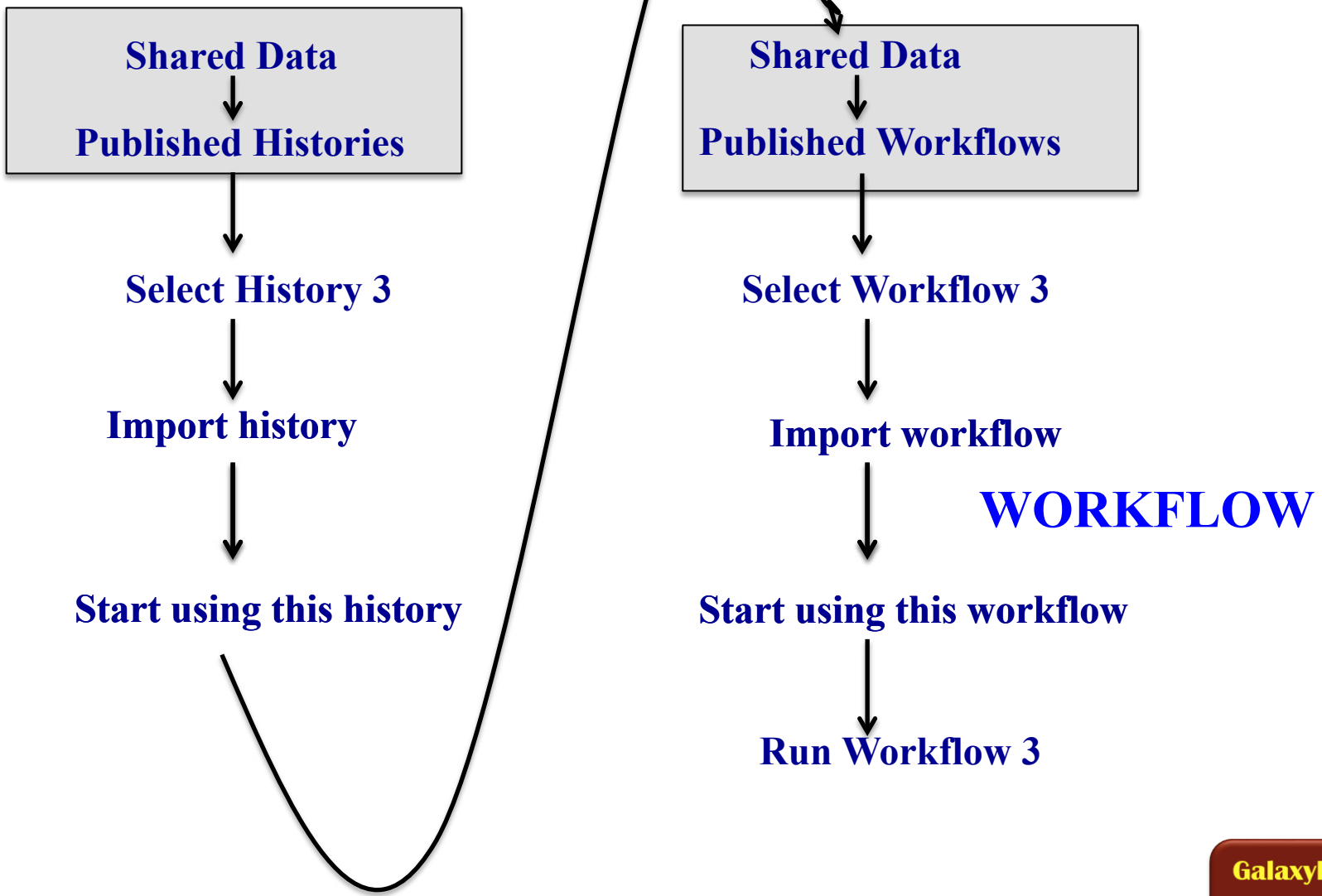
SEARCHGUI : SEARCH ALGORITHMS

INPUTS : PEAKLISTS and SEARCH db

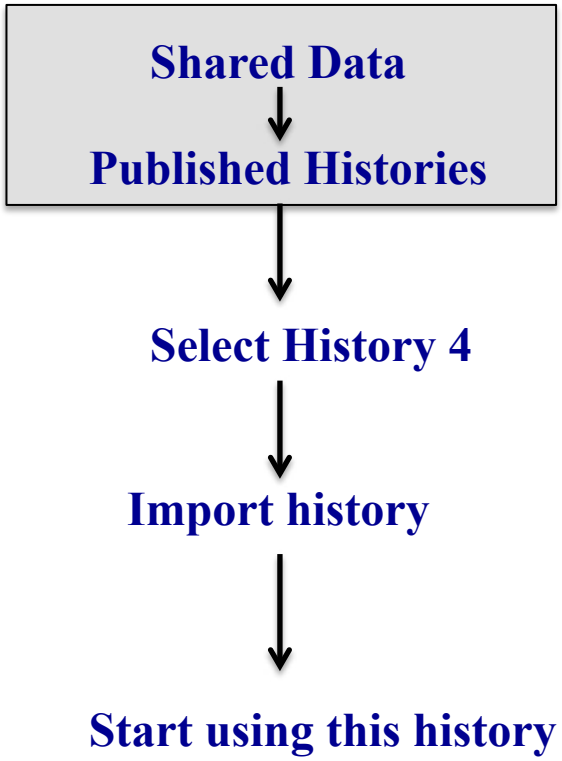
5.2 BLAST-P Search

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

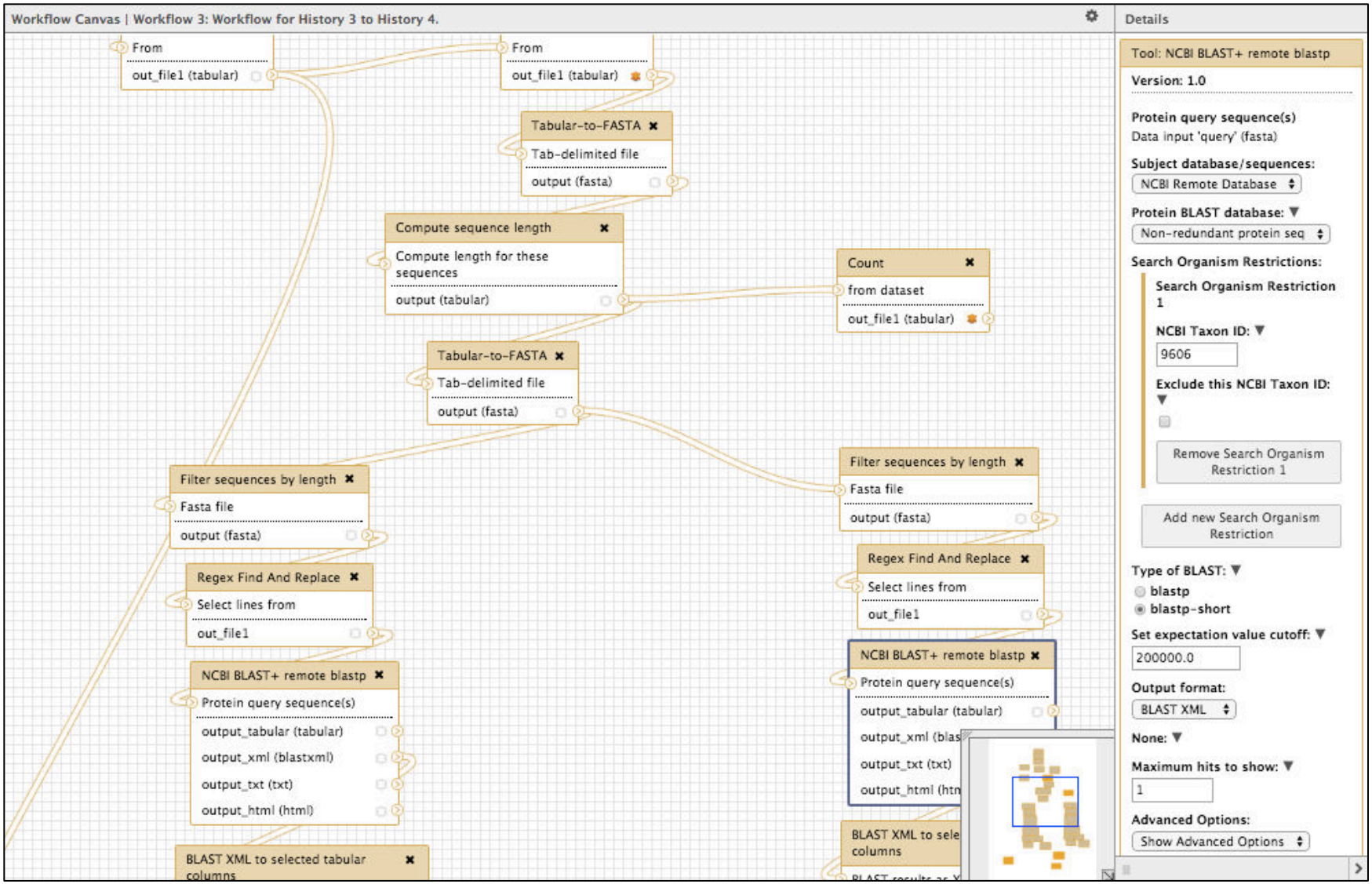


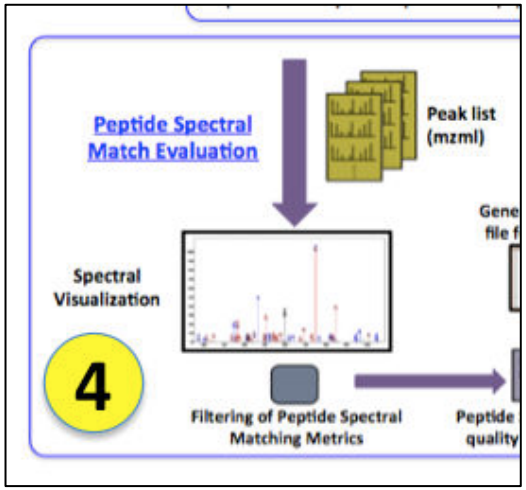
5.2 BLAST-P Search



OUTPUT

BLAST-P SEARCH





INPUTS → SEARCHGUI → PEPTIDESHAKER → BLAST-P → PSM Visualization

PSM EVALUATION

BLAST-P SEARCH

PEPTIDESHAKER

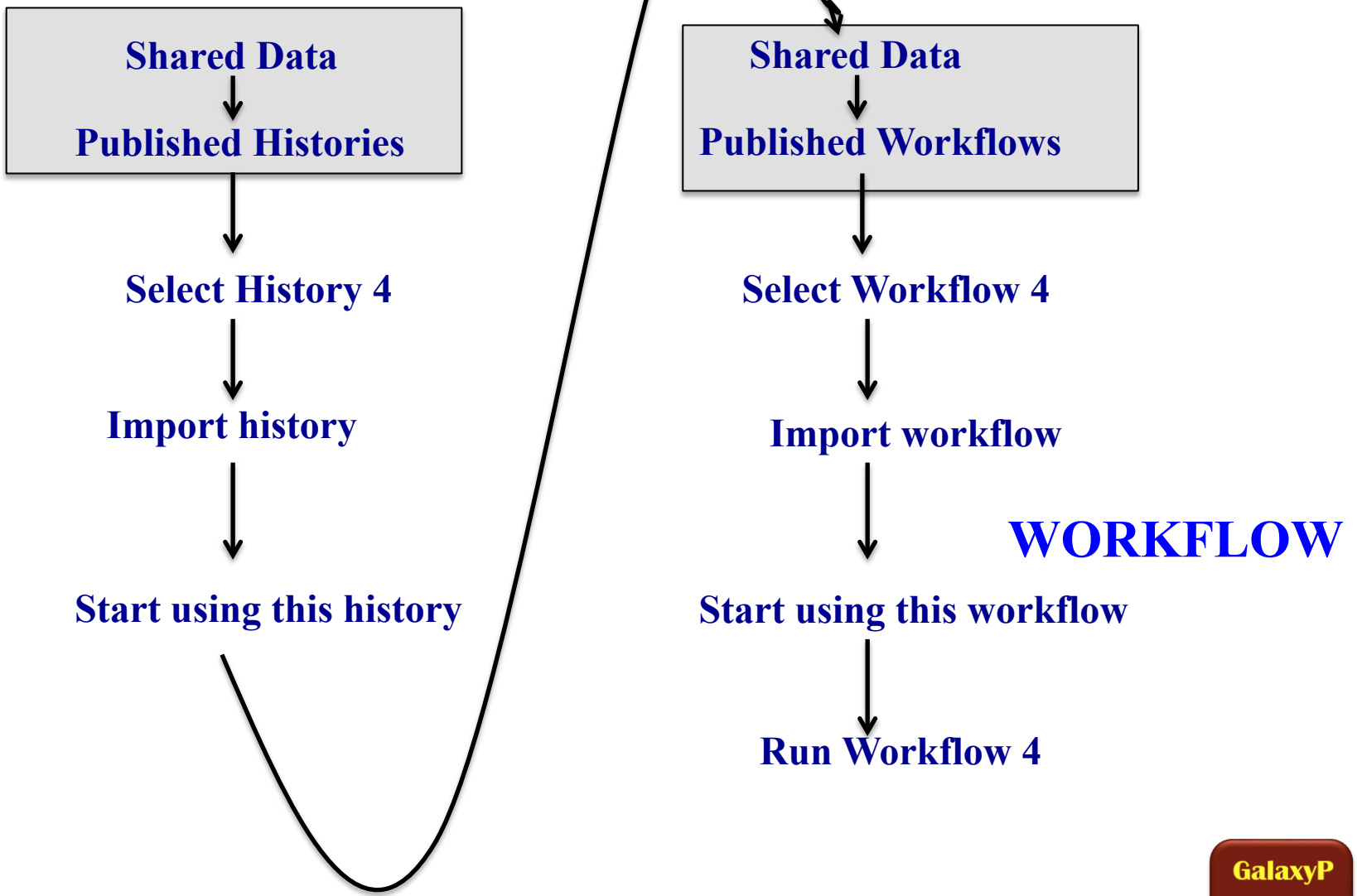
SEARCHGUI : SEARCH ALGORITHMS

INPUTS : PEAKLISTS and SEARCH db

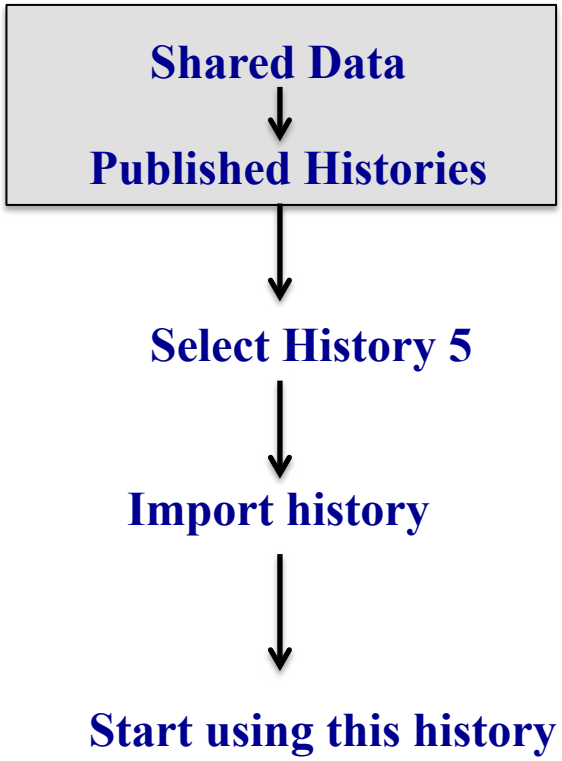
6.2 Generating a sqlite Database

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.

6.2 Generating a sqlite Database



6.3 PSM Evaluator in GalaxyP



OUTPUT

6.3 PSM Evaluator in GalaxyP

a. From Shared Data at the top of the screen, import History 5
b. 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.

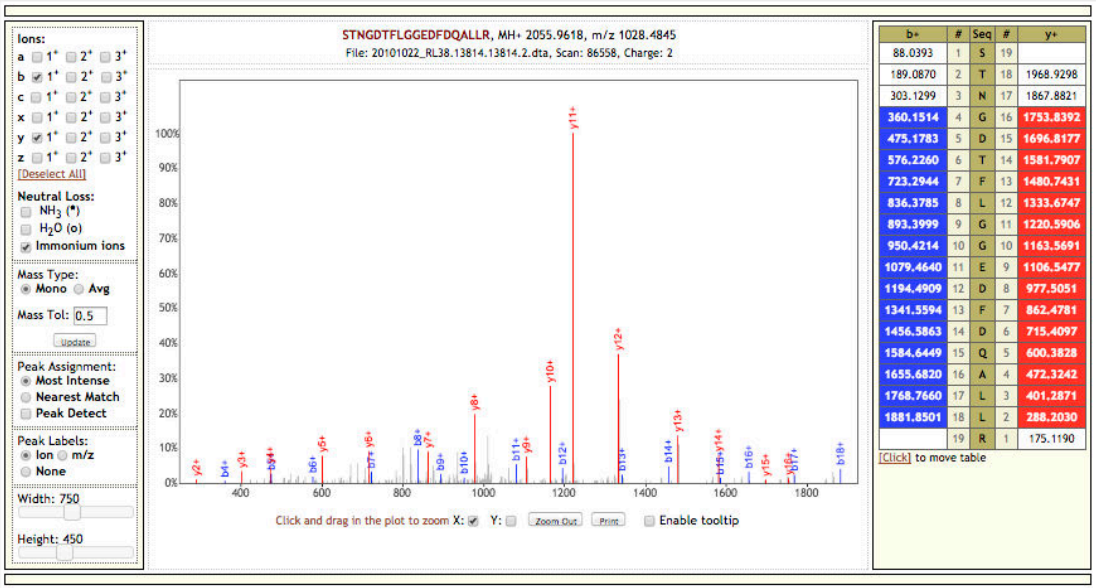
PSM EVALUATION

Search:

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 247 Showing records of 12313 total records.

STNGDTFLGGEDFDQALLR 86558



b+	#	Seq	#	y+
88.0393	1	S	19	
189.0870	2	T	18	1968.9298
303.1299	3	N	17	1867.8821
360.1514	4	G	16	1753.8392
475.1783	5	D	15	1696.8177
576.2260	6	T	14	1581.7907
723.2944	7	F	13	1480.7431
836.3785	8	L	12	1333.6747
893.3999	9	G	11	1220.5906
950.4214	10	G	10	1163.5691
1079.4640	11	E	9	1106.5477
1194.4909	12	D	8	977.5051
1341.5594	13	F	7	862.4781
1456.5863	14	D	6	715.4097
1584.6449	15	Q	5	600.3828
1655.6820	16	A	4	472.3242
1768.7660	17	L	3	401.2871
1881.8501	18	L	2	288.2030
	19	R	1	175.1190

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

James E. Johnson

PSM EVALUATION

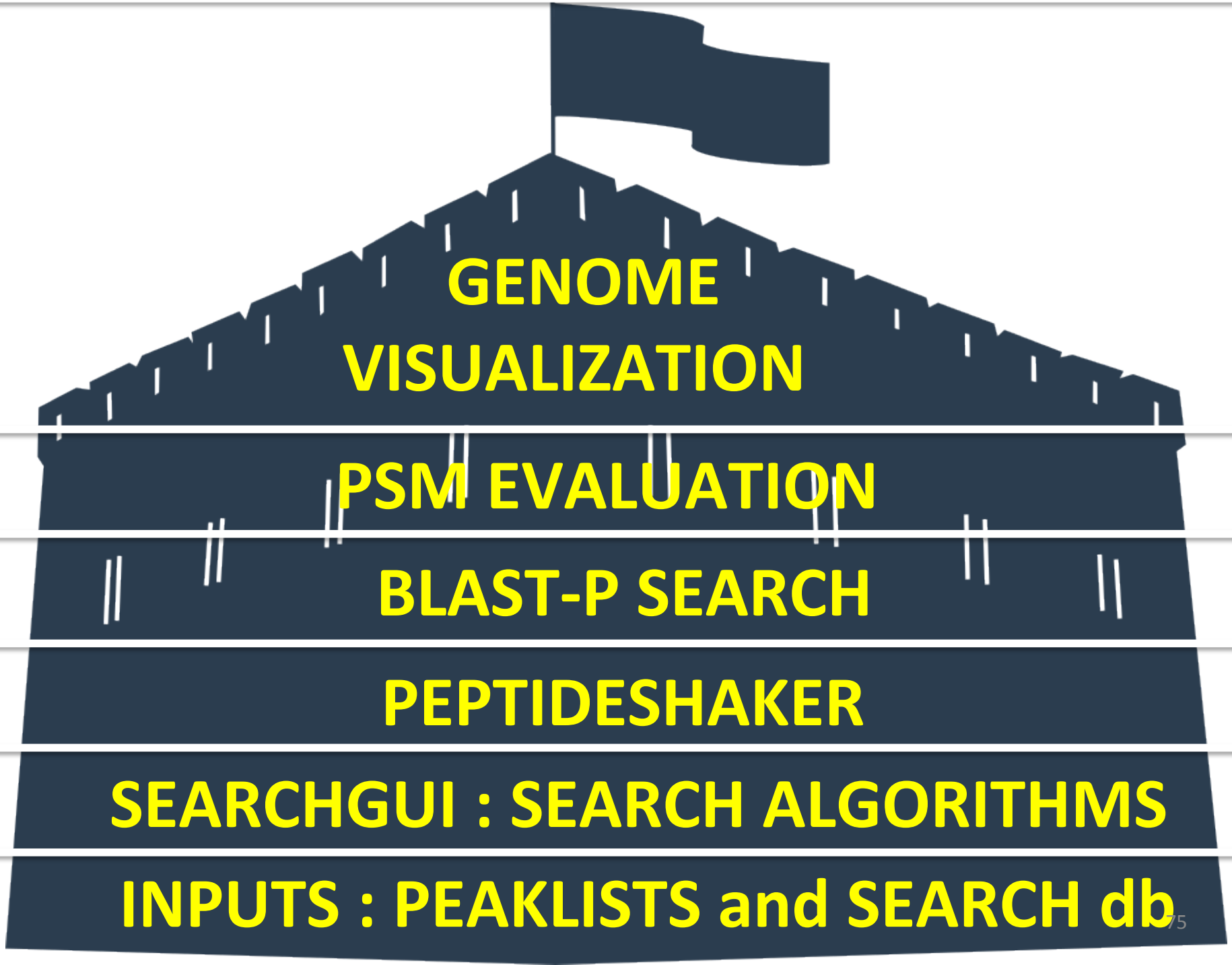
- **Go to 'Peptide' View.**
- **Type in 'LLSVGGLR, SPVLKPSR' in Filter box.**

  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


GENOME VISUALIZATION
PSM EVALUATION
BLAST-P SEARCH
PEPTIDESHAKER
SEARCHGUI : SEARCH ALGORITHMS
INPUTS : PEAKLISTS and SEARCH db

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.

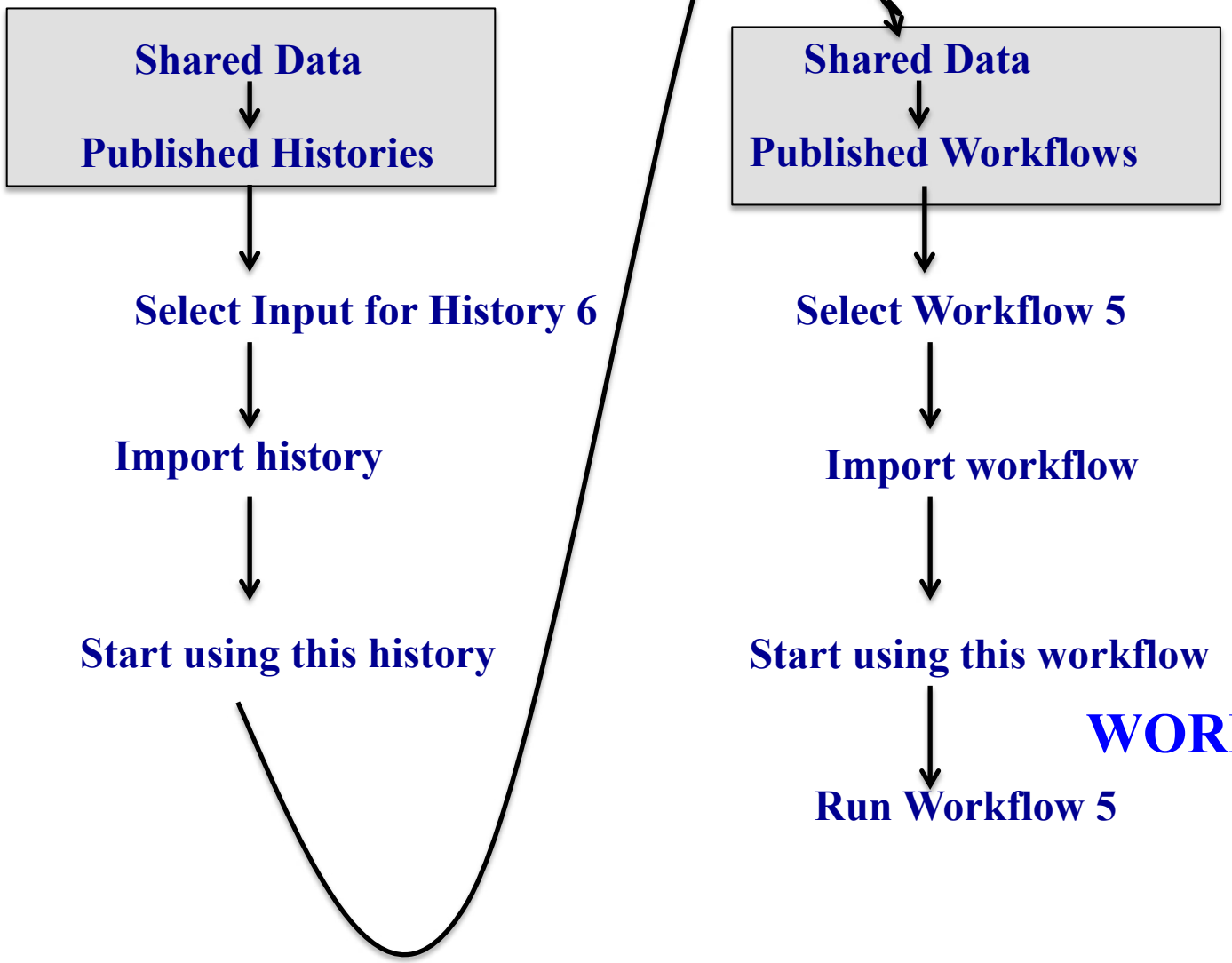
c. Select the appropriate inputs for Workflow 5 from Inputs to History 6.

d. Check to send the results to a new history named History 6 and Run the workflow.

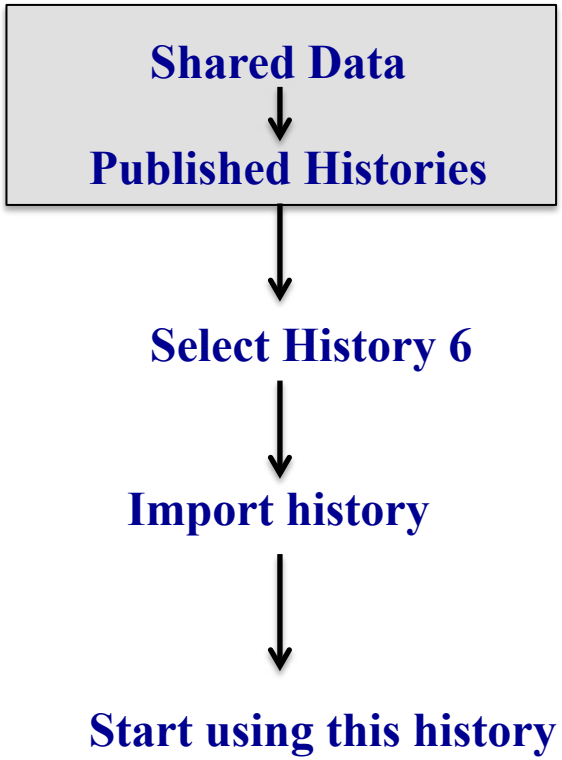
For the purposes of this tutorial, a completed History 6 may be imported from Published Histories.

7.3 From GFF to IGV Browser

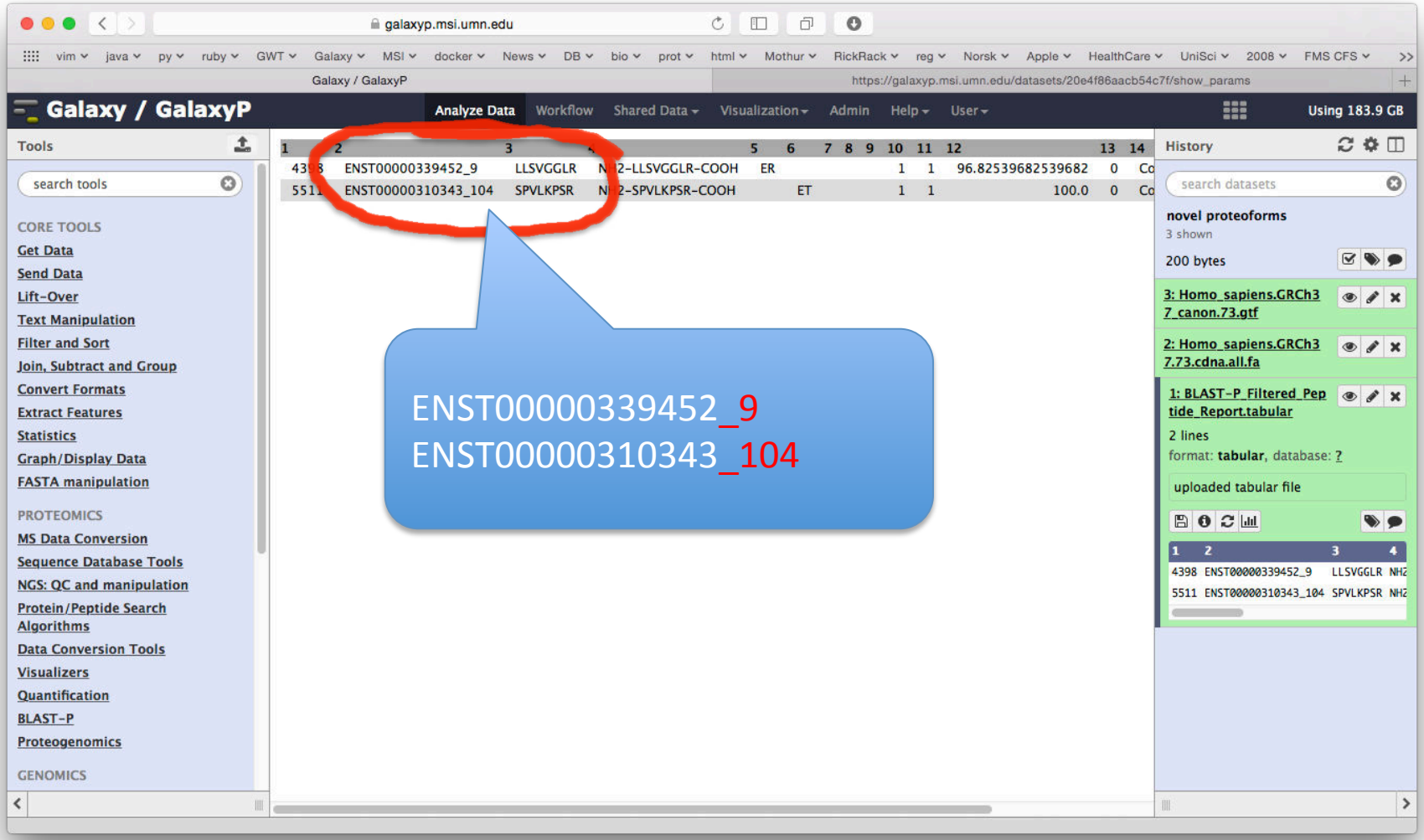
7.2 Generating a GFF File.

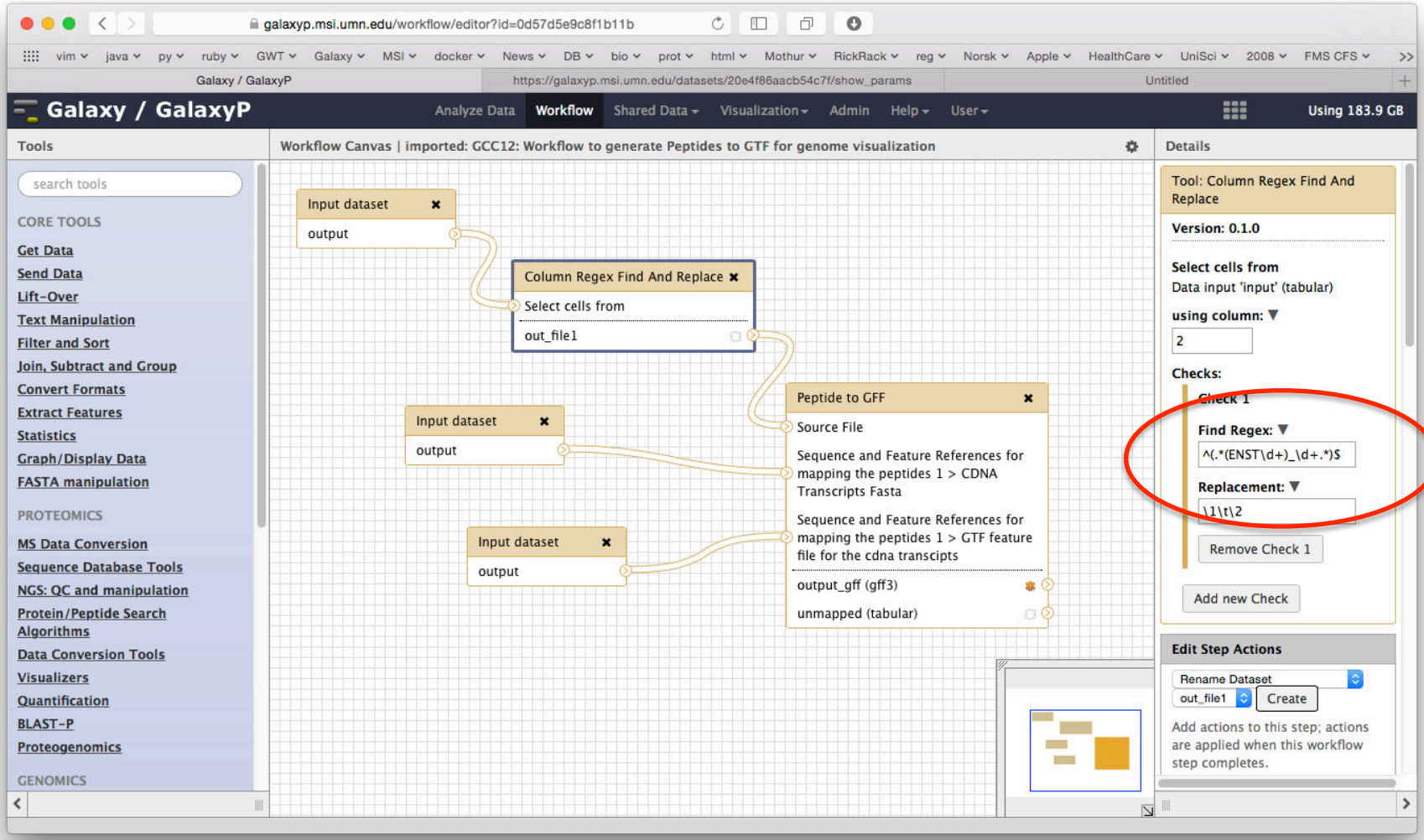


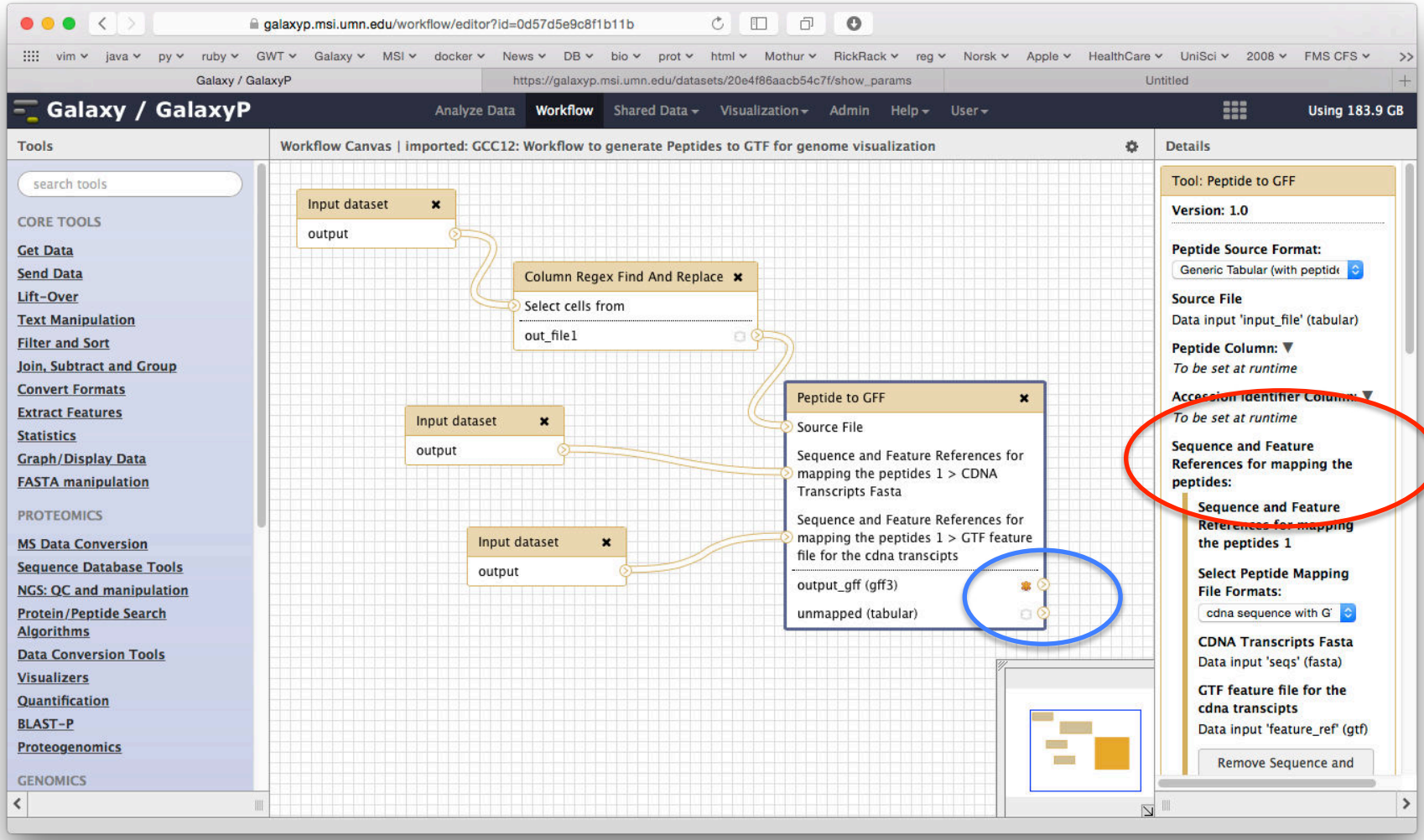
7.2 Generating a GFF File.

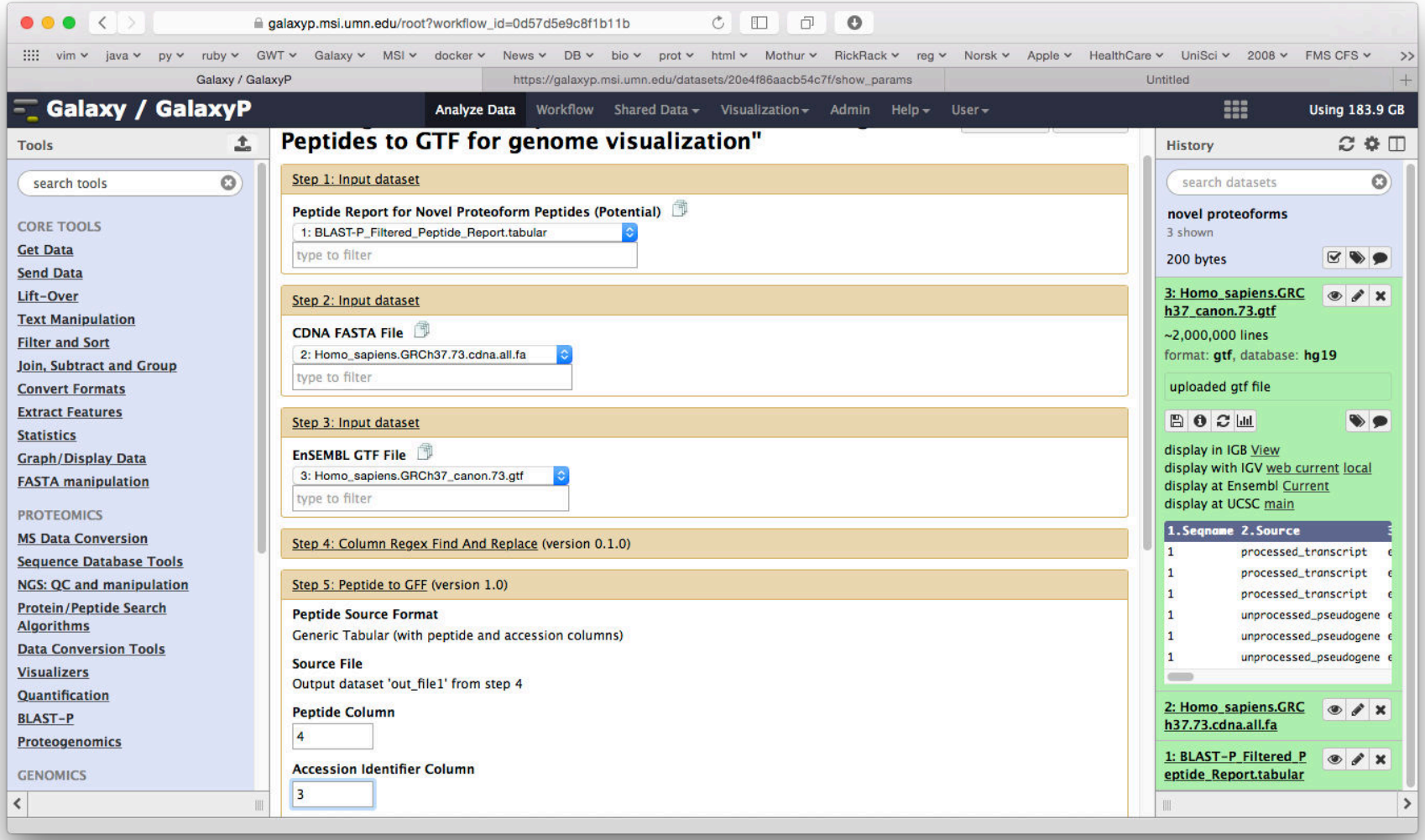


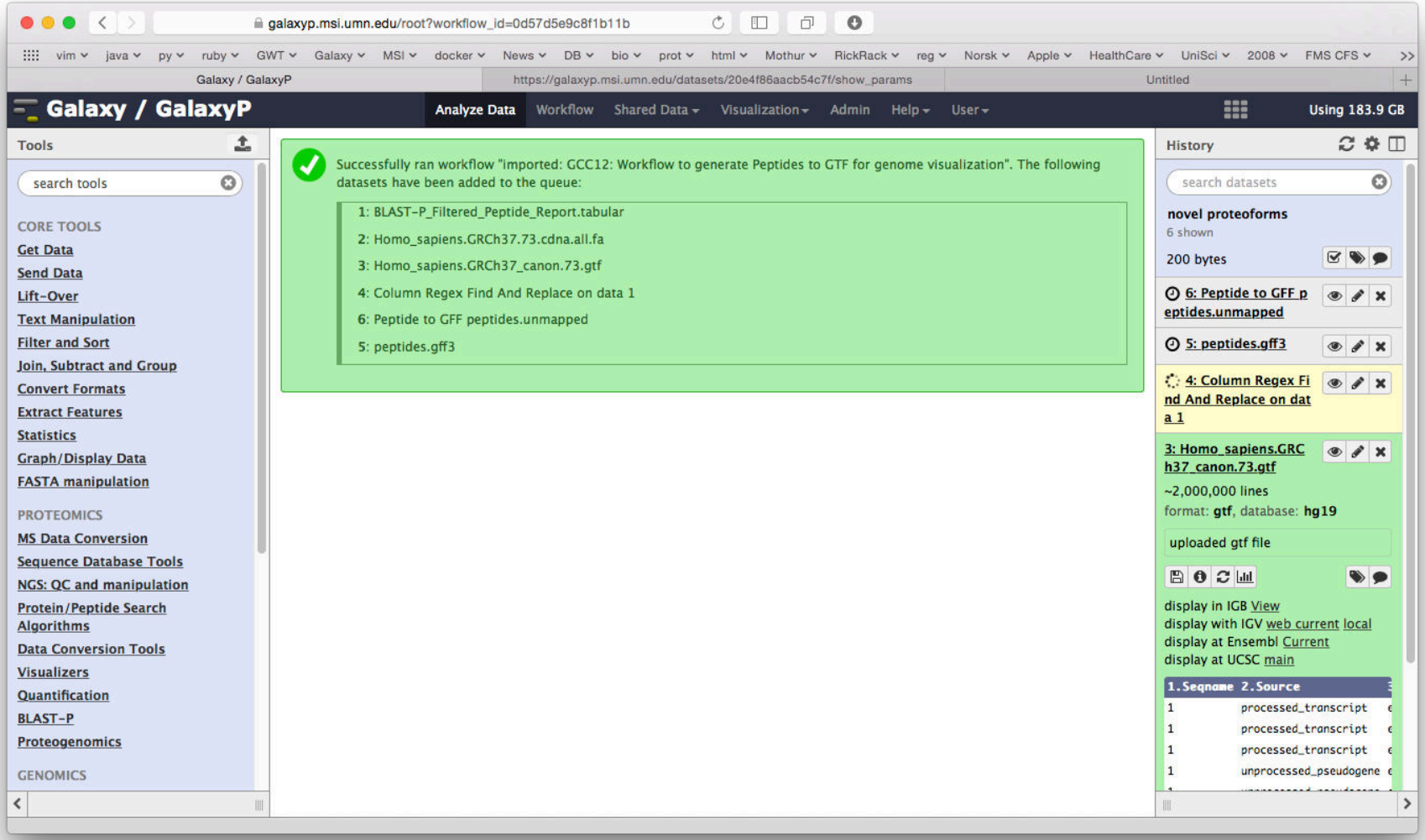
OUTPUT

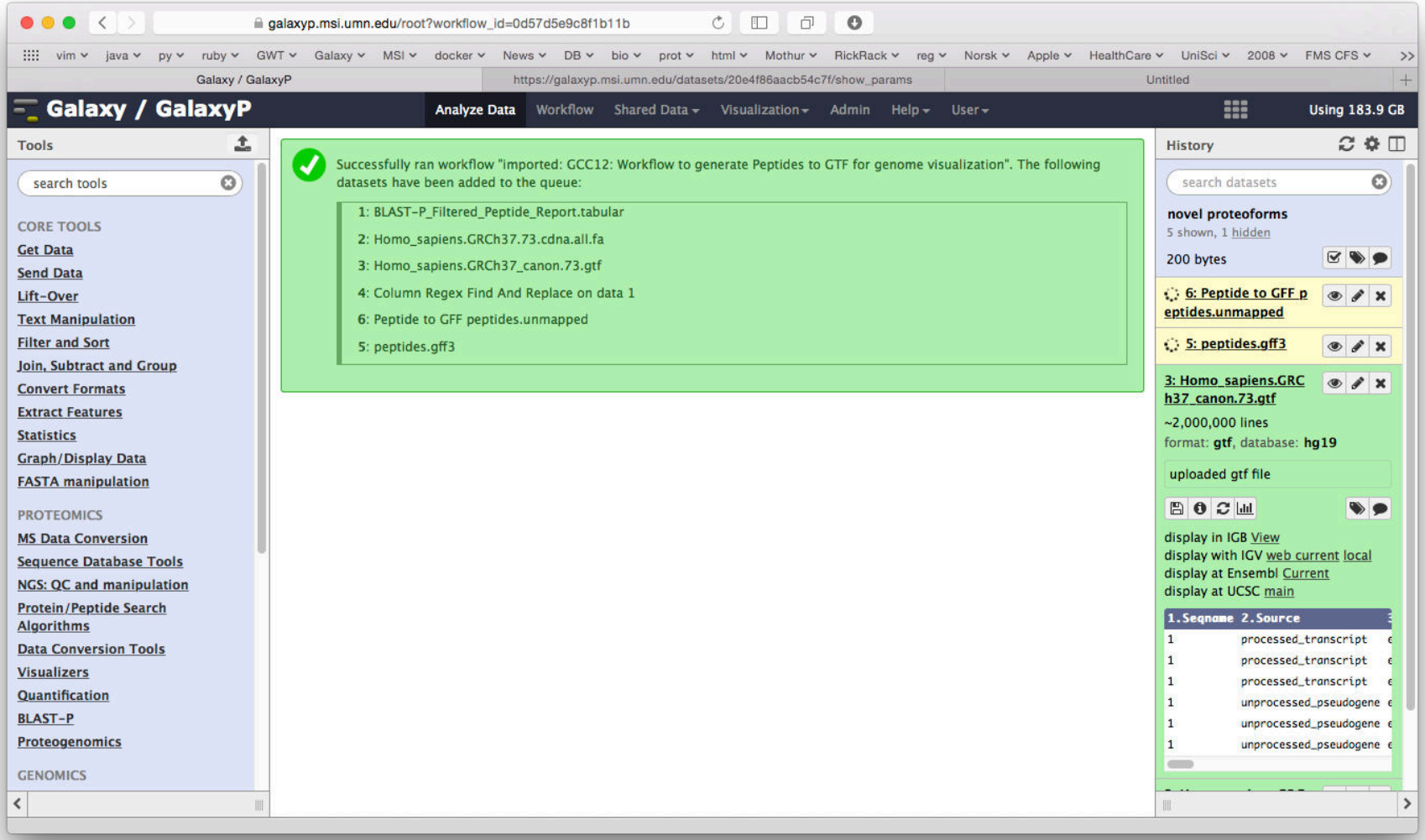

galaxyp.msi.umn.edu
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
ENST00000339452_9 LLSVGGLR NH2-LLSVGGLR-COOH ER 1 1 96.82539682539682 0
ENST00000310343_104 SPVLKPSR NH2-SPVLKPSR-COOH ET 1 1 100.0 0
ENST00000339452_9
ENST00000310343_104
History
search datasets
novel proteoforms
3 shown
200 bytes
3: Homo_sapiens.GRCh37_canon.73.gtf
2: Homo_sapiens.GRCh37.73.cdna.all.fa
1: BLAST-P_Filtered_Peptide_Report.tabular
2 lines
format: tabular, database: ?
uploaded tabular file
4398 ENST00000339452_9 LLSVGGLR NH2
5511 ENST00000310343_104 SPVLKPSR NH2

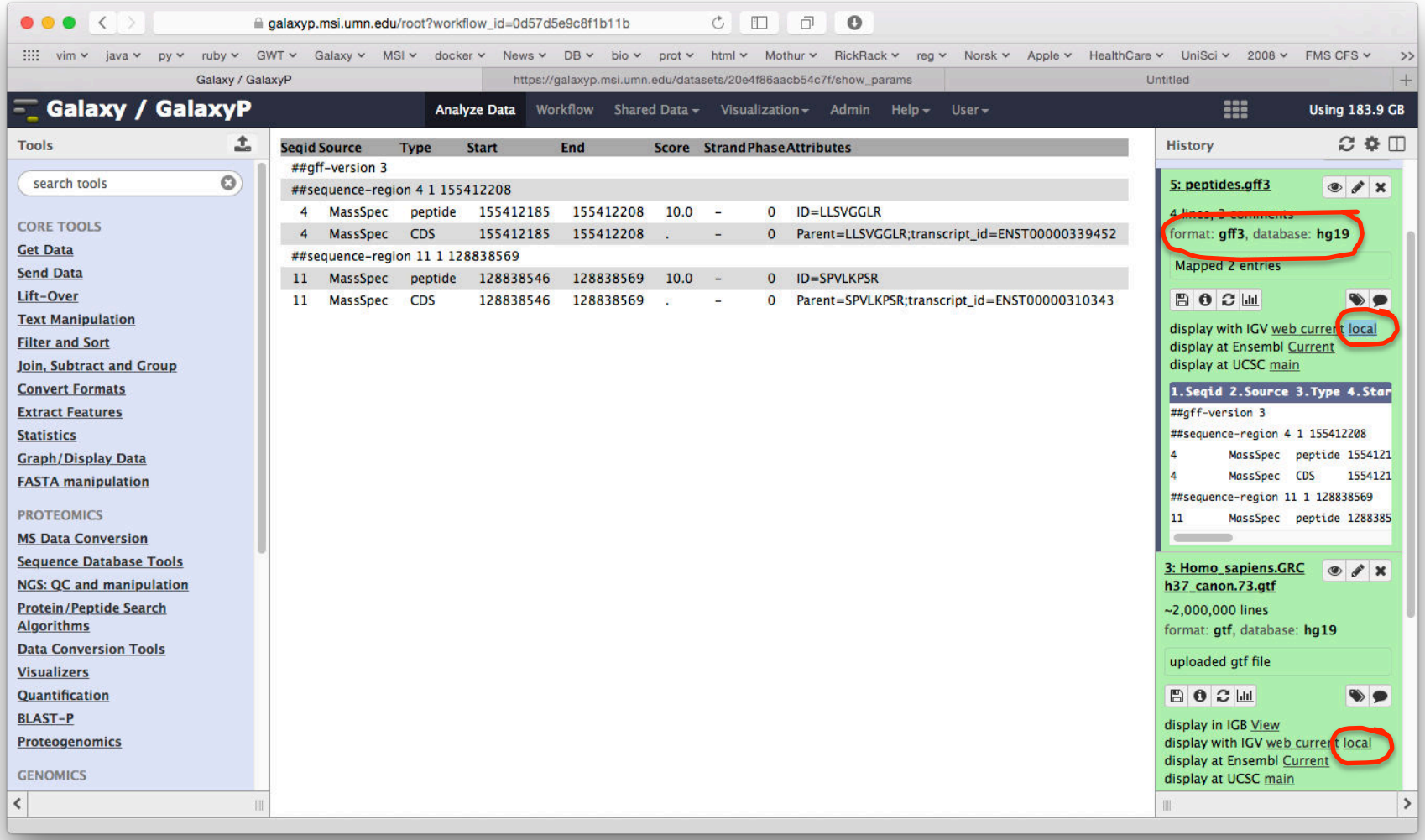

galaxyp.msi.umn.edu/workflow/editor?id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
https://galaxyp.msi.umn.edu/datasets/20e4f86aacb54c7f/show_params
Untitled
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
Workflow Canvas | imported: GCC12: Workflow to generate Peptides to GTF for genome visualization
Input dataset
output
Column Regex Find And Replace
Select cells from
out_file1
Input dataset
output
Input dataset
output
Peptide to GFF
Source File
Sequence and Feature References for mapping the peptides 1 > CDNA Transcripts Fasta
Sequence and Feature References for mapping the peptides 1 > GTF feature file for the cdna transcipts
output_gff (gff3)
unmapped (tabular)
Details
Tool: Column Regex Find And Replace
Version: 0.1.0
Select cells from
Data input 'input' (tabular)
using column:
2
Checks:
Check 1
Find Regex:
^(.*(ENST\d+)_\d+.*)$
Replacement:
\1\t\2
Remove Check 1
Add new Check
Edit Step Actions
Rename Dataset
out_file1
Create
Add actions to this step; actions are applied when this workflow step completes.

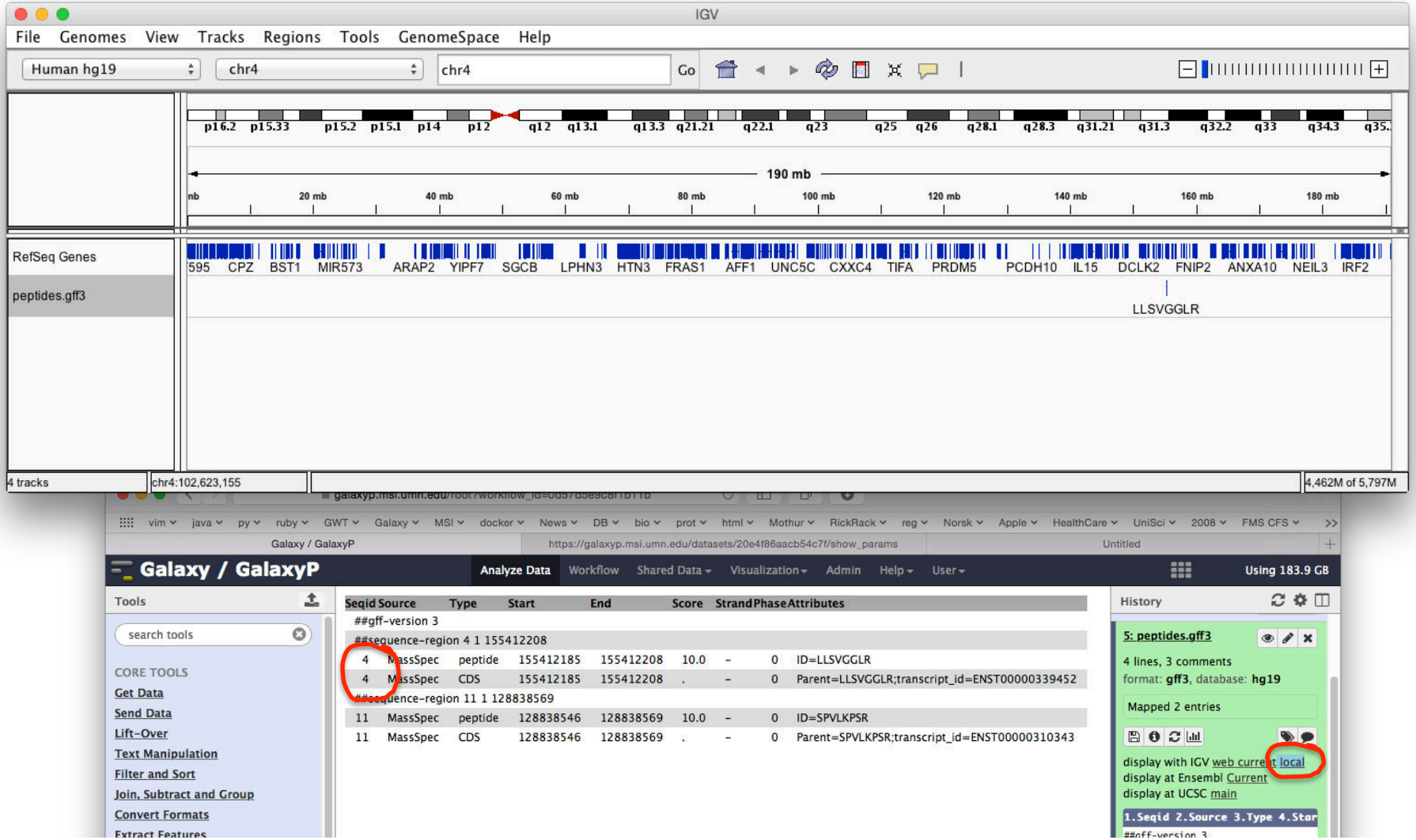

galaxyp.msi.umn.edu/workflow/editor?id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
Workflow Canvas | imported: GCC12: Workflow to generate Peptides to GTF for genome visualization
Input dataset
output
Column Regex Find And Replace
Select cells from
out_file1
Input dataset
output
Input dataset
output
Peptide to GFF
Source File
Sequence and Feature References for mapping the peptides 1 > CDNA Transcripts Fasta
Sequence and Feature References for mapping the peptides 1 > GTF feature file for the cdna transcipts
output_gff (gff3)
unmapped (tabular)
Details
Tool: Peptide to GFF
Version: 1.0
Peptide Source Format:
Source File
Data input 'input_file' (tabular)
Peptide Column:
To be set at runtime
To be set at runtime
Sequence and Feature References for mapping the peptides:
Sequence and Feature References for mapping the peptides 1
Select Peptide Mapping File Formats:
CDNA Transcripts Fasta
Data input 'seqs' (fasta)
GTF feature file for the cdna transcipts
Data input 'feature_ref' (gtf)
Remove Sequence and

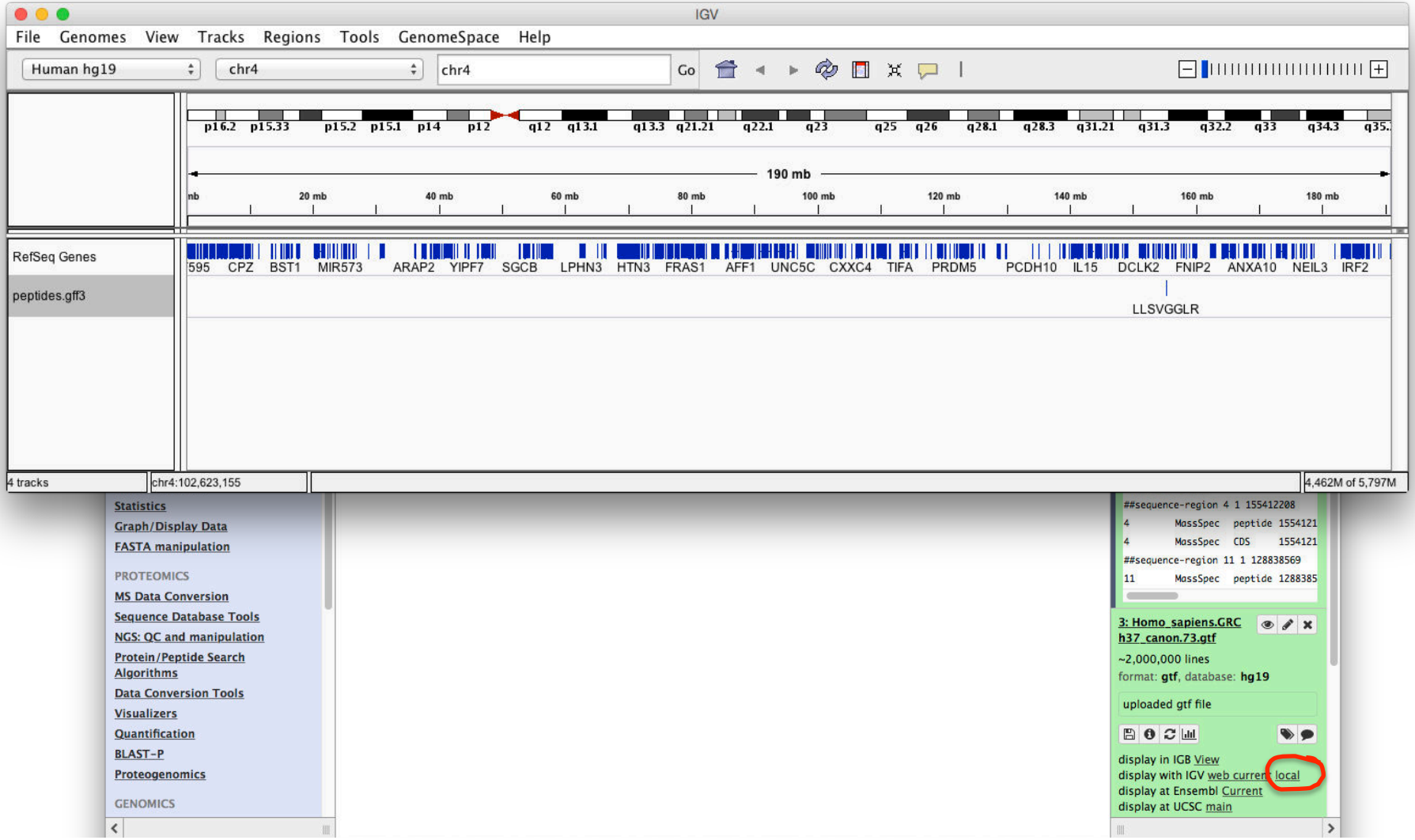

galaxyp.msi.umn.edu/root?workflow_id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
Peptides to GTF for genome visualization"
Step 1: Input dataset
Peptide Report for Novel Proteoform Peptides (Potential)
1: BLAST-P_Filtered_Peptide_Report.tabular
type to filter
Step 2: Input dataset
CDNA FASTA File
2: Homo_sapiens.GRCh37.73.cdna.all.fa
type to filter
Step 3: Input dataset
EnSEMBL GTF File
3: Homo_sapiens.GRCh37_canon.73.gtf
type to filter
Step 4: Column Regex Find And Replace (version 0.1.0)
Step 5: Peptide to GFF (version 1.0)
Peptide Source Format
Generic Tabular (with peptide and accession columns)
Source File
Output dataset 'out_file1' from step 4
Peptide Column
4
Accession Identifier Column
3
History
search datasets
novel proteoforms
3 shown
200 bytes
3: Homo_sapiens.GRCh37_canon.73.gtf
~2,000,000 lines
format: gtf, database: hg19
uploaded gtf file
display in IGB View
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqname 2.Source
1 processed_transcript
1 processed_transcript
1 processed_transcript
1 unprocessed_pseudogene
1 unprocessed_pseudogene
1 unprocessed_pseudogene
2: Homo_sapiens.GRCh37.73.cdna.all.fa
1: BLAST-P_Filtered_Peptide_Report.tabular

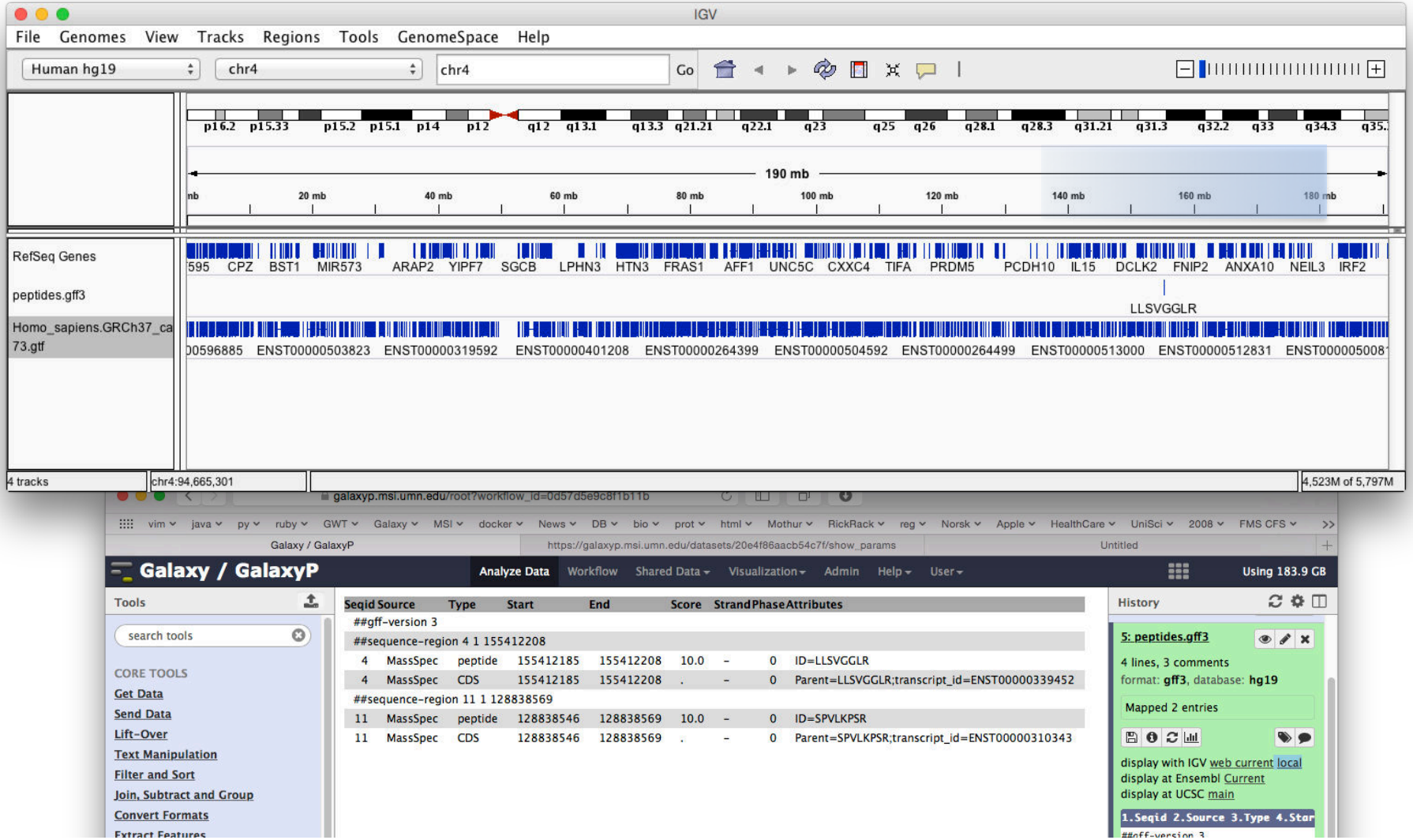

galaxyp.msi.umn.edu/root?workflow_id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
https://galaxyp.msi.umn.edu/datasets/20e4f86aacb54c7f/show_params
Untitled
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
Successfully ran workflow "imported: GCC12: Workflow to generate Peptides to GTF for genome visualization". The following datasets have been added to the queue:
1: BLAST-P_Filtered_Peptide_Report.tabular
2: Homo_sapiens.GRCh37.73.cdna.all.fa
3: Homo_sapiens.GRCh37_canon.73.gtf
4: Column Regex Find And Replace on data 1
6: Peptide to GFF peptides.unmapped
5: peptides.gff3
History
search datasets
novel proteoforms
6 shown
200 bytes
6: Peptide to GFF peptides.unmapped
5: peptides.gff3
4: Column Regex Find And Replace on data 1
3: Homo_sapiens.GRCh37_canon.73.gtf
~2,000,000 lines
format: gtf, database: hg19
uploaded gtf file
display in IGB View
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqname 2.Source
1 processed_transcript
1 processed_transcript
1 processed_transcript
1 unprocessed_pseudogene


galaxyp.msi.umn.edu/root?workflow_id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
https://galaxyp.msi.umn.edu/datasets/20e4f86aacb54c7f/show_params
Untitled
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
Successfully ran workflow "imported: GCC12: Workflow to generate Peptides to GTF for genome visualization". The following datasets have been added to the queue:
1: BLAST-P_Filtered_Peptide_Report.tabular
2: Homo_sapiens.GRCh37.73.cdna.all.fa
3: Homo_sapiens.GRCh37_canon.73.gtf
4: Column Regex Find And Replace on data 1
6: Peptide to GFF peptides.unmapped
5: peptides.gff3
History
search datasets
novel proteoforms
5 shown, 1 hidden
200 bytes
6: Peptide to GFF peptides.unmapped
5: peptides.gff3
3: Homo_sapiens.GRCh37_canon.73.gtf
~2,000,000 lines
format: gtf, database: hg19
uploaded gtf file
display in IGB View
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqname 2.Source
1 processed_transcript
1 processed_transcript
1 processed_transcript
1 unprocessed_pseudogene
1 unprocessed_pseudogene
1 unprocessed_pseudogene

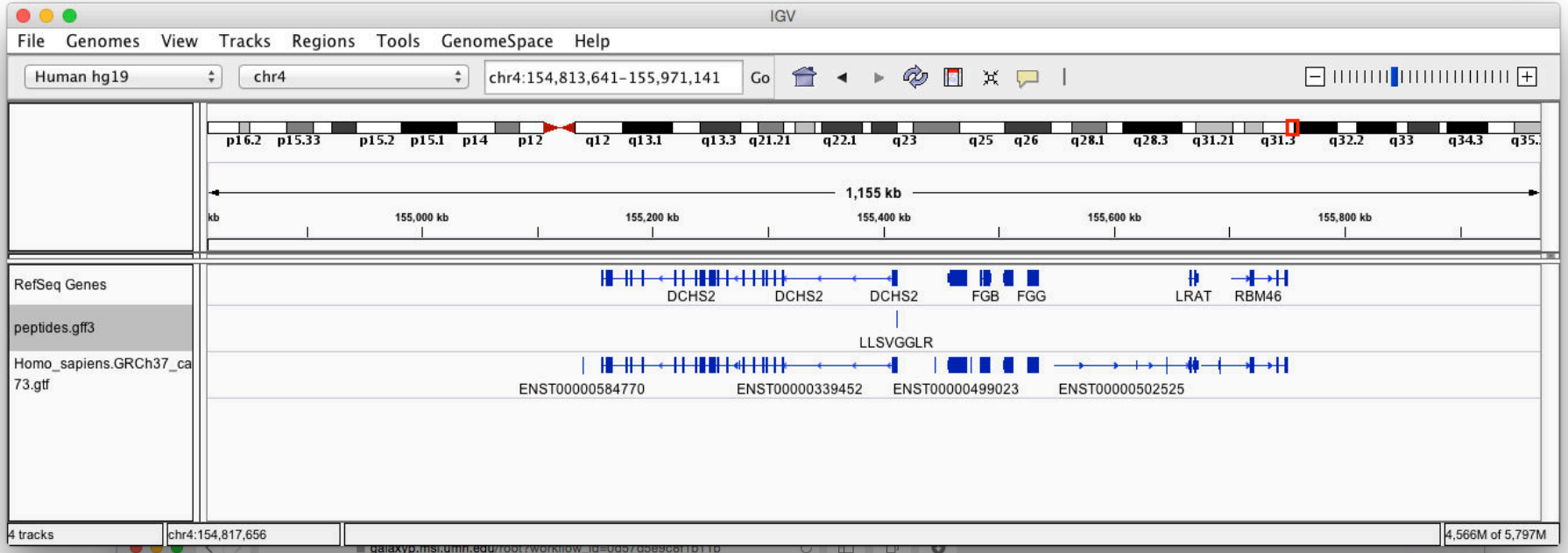

galaxyp.msi.umn.edu/root?workflow_id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
https://galaxyp.msi.umn.edu/datasets/20e4f86aacb54c7f/show_params
Untitled
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqid 2.Source 3.Type 4.Star
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 1554121
4 MassSpec CDS 1554121
##sequence-region 11 1 128838569
11 MassSpec peptide 1288385
3: Homo_sapiens.GRCh37_canon.73.gtf
~2,000,000 lines
format: gtf, database: hg19
uploaded gtf file
display in IGB View
display with IGV web current local
display at Ensembl Current
display at UCSC main

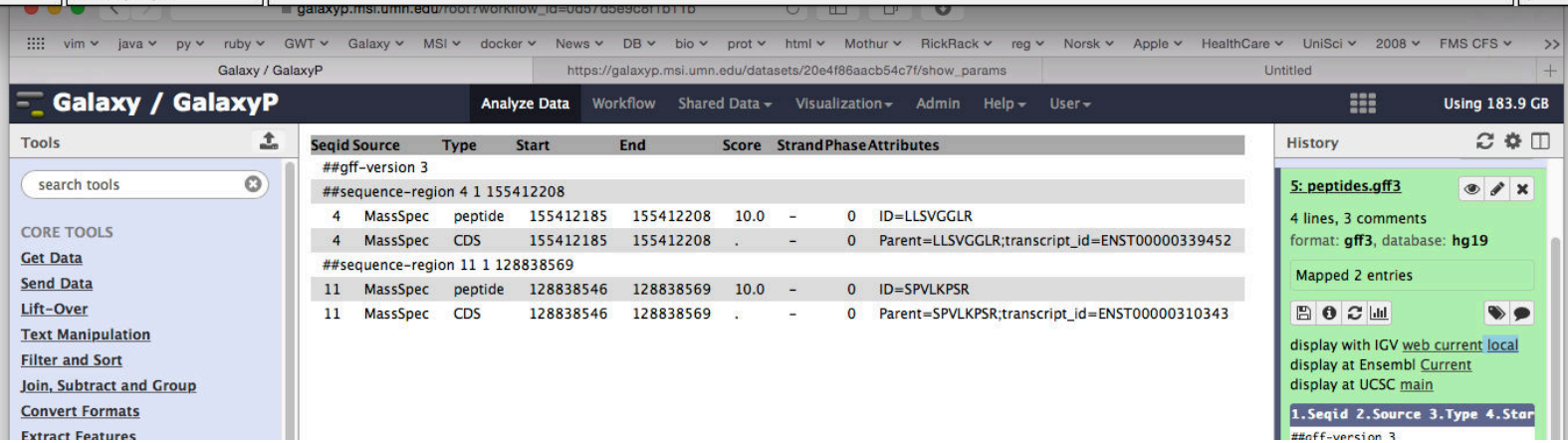

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr4
chr4
Go
p16.2 p15.33 p15.2 p15.1 p14 p12 q12 q13.1 q13.3 q21.21 q22.1 q23 q25 q26 q28.1 q28.3 q31.21 q31.3 q32.2 q33 q34.3 q35.
190 mb
nb 20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb 160 mb 180 mb
RefSeq Genes
595 CPZ BST1 MIR573 ARAP2 YIPF7 SGCB LPHN3 HTN3 FRAS1 AFF1 UNC5C CXXC4 TIFA PRDM5 PCDH10 IL15 DCLK2 FNIP2 ANXA10 NEIL3 IRF2
peptides.gff3
LLSVGGLR
4 tracks
chr4:102,623,155
4,462M of 5,797M
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
4 lines, 3 comments
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqid 2.Source 3.Type 4.Star

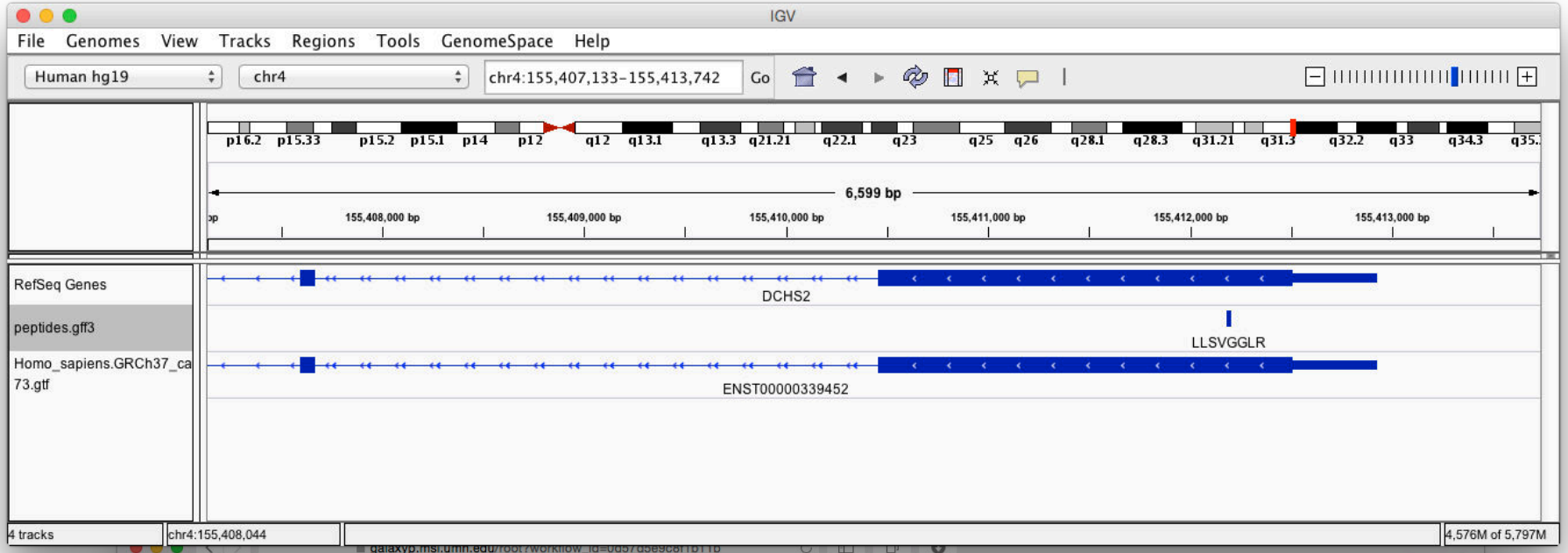

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr4
chr4
Go
p16.2 p15.33 p15.2 p15.1 p14 p12 q12 q13.1 q13.3 q21.21 q22.1 q23 q25 q26 q28.1 q28.3 q31.21 q31.3 q32.2 q33 q34.3 q35.
190 mb
nb 20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb 160 mb 180 mb
RefSeq Genes
peptides.gff3
595 CPZ BST1 MIR573 ARAP2 YIPF7 SGCB LPHN3 HTN3 FRAS1 AFF1 UNC5C CXXC4 TIFA PRDM5 PCDH10 IL15 DCLK2 FNIP2 ANXA10 NEIL3 IRF2
LLSVGGLR
4 tracks
chr4:102,623,155
4,462M of 5,797M
Statistics
Graph/Display Data
FASTA manipulation
PROTEOMICS
MS Data Conversion
Sequence Database Tools
NGS: QC and manipulation
Protein/Peptide Search Algorithms
Data Conversion Tools
Visualizers
Quantification
BLAST-P
Proteogenomics
GENOMICS
##sequence-region 4 1 155412208
4 MassSpec peptide 1554121
4 MassSpec CDS 1554121
##sequence-region 11 1 128838569
11 MassSpec peptide 1288385
3: Homo_sapiens.GRCh37_canon.73.gtf
~2,000,000 lines
format: gtf, database: hg19
uploaded gtf file
display in IGB View
display with IGV web current local
display at Ensembl Current
display at UCSC main

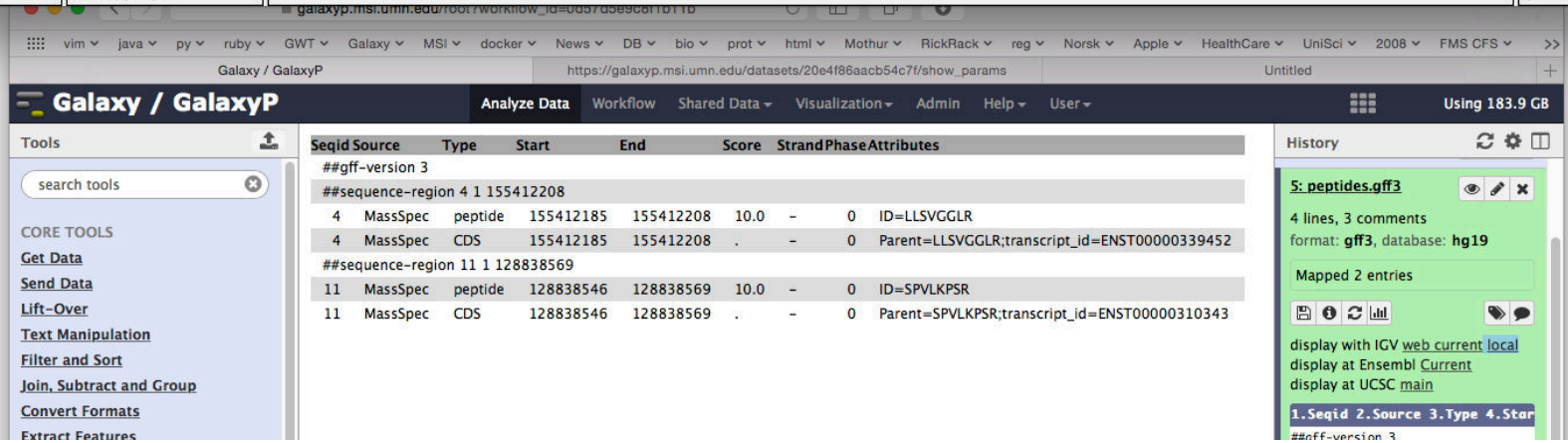


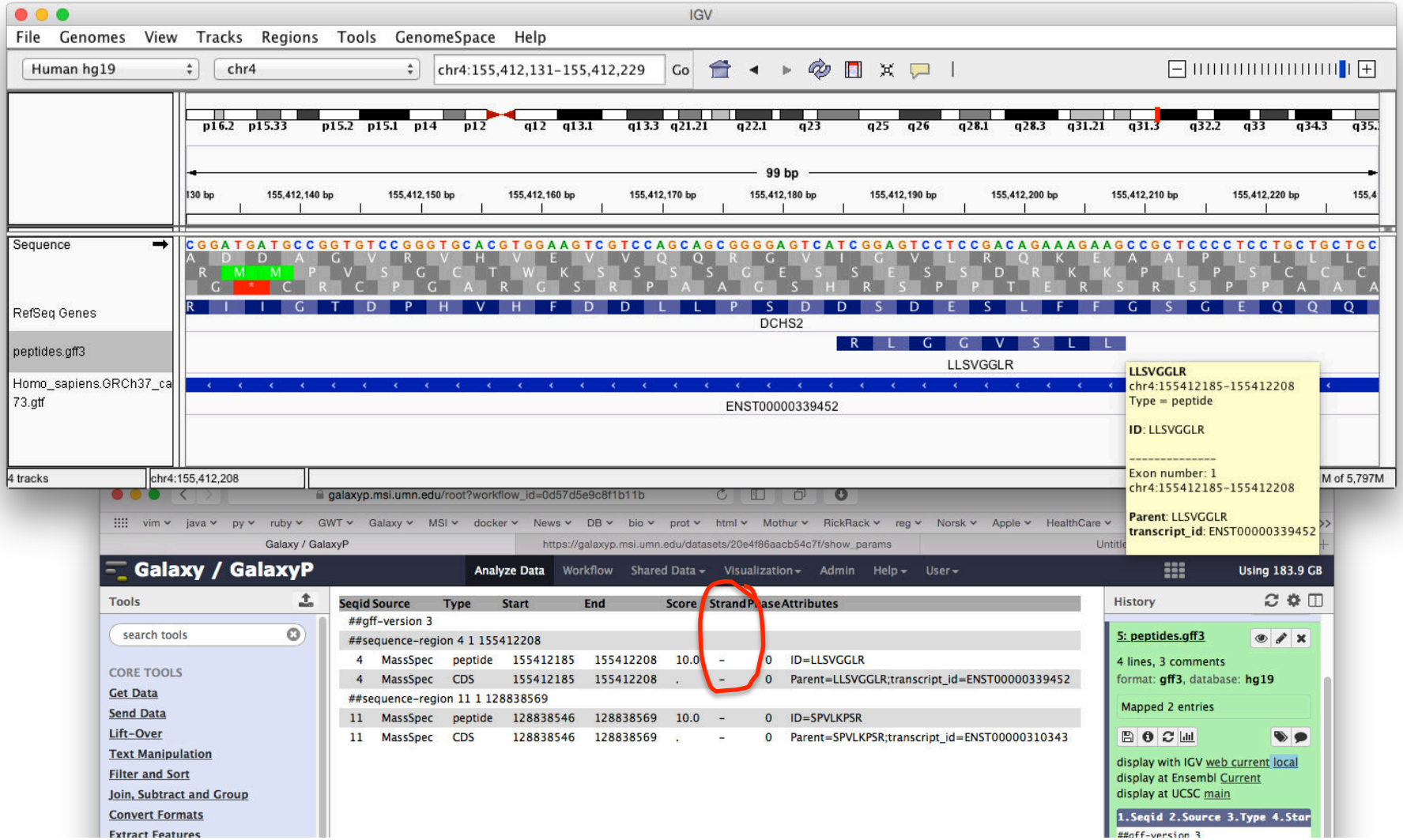
Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
##sequence-region 4 1 155412208								
4	MassSpec	peptide	155412185	155412208	10.0	-	0	ID=LLSVGGLR
4	MassSpec	CDS	155412185	155412208	.	-	0	Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569								
11	MassSpec	peptide	128838546	128838569	10.0	-	0	ID=SPVLKPSR
11	MassSpec	CDS	128838546	128838569	.	-	0	Parent=SPVLKPSR;transcript_id=ENST00000310343

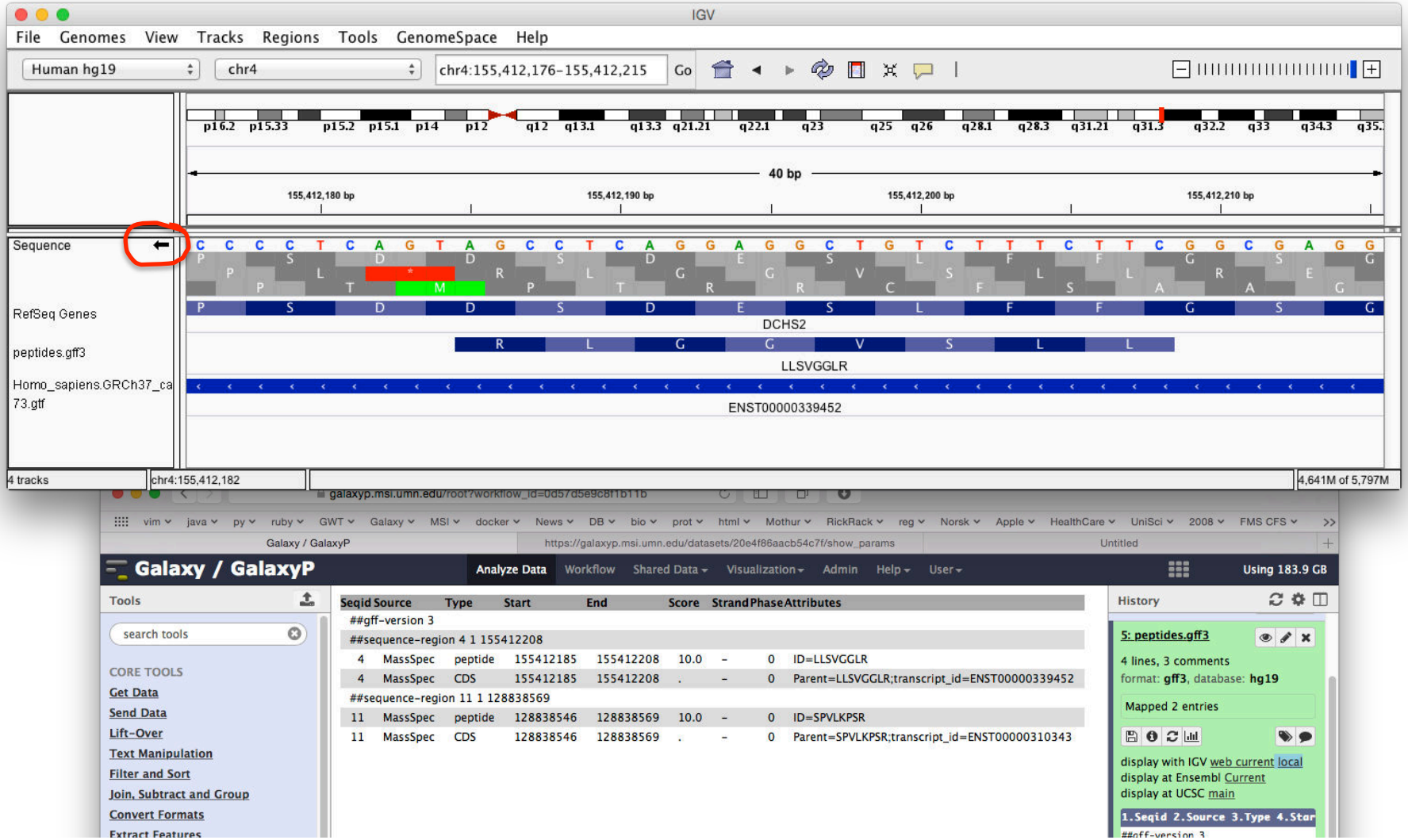

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr4
chr4:154,813,641-155,971,141
Go
1,155 kb
155,000 kb
155,200 kb
155,400 kb
155,600 kb
155,800 kb
RefSeq Genes
peptides.gff3
Homo_sapiens.GRCh37_ca
73.gtf
DCHS2
DCHS2
DCHS2
FGB
FGG
LRAT
RBM46
LLSVGGLR
ENST00000584770
ENST00000339452
ENST00000499023
ENST00000502525
4 tracks
chr4:154,817,656
4,566M of 5,797M


Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
4 lines, 3 comments
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main

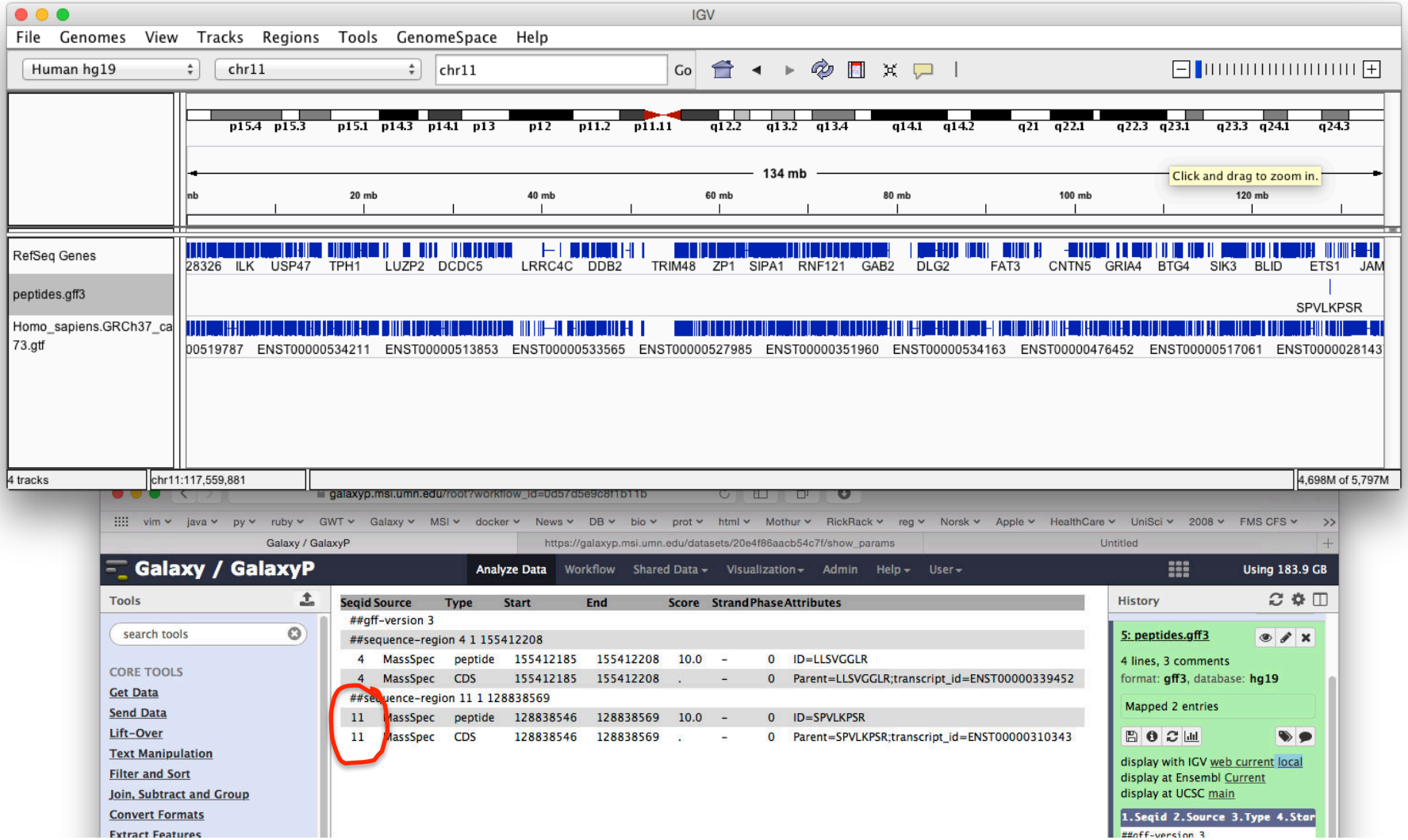

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr4
chr4:155,407,133-155,413,742
Go
6,599 bp
155,408,000 bp
155,409,000 bp
155,410,000 bp
155,411,000 bp
155,412,000 bp
155,413,000 bp
RefSeq Genes
peptides.gff3
Homo_sapiens.GRCh37_ca
73.gtf
DCHS2
LLSVGGLR
ENST00000339452
4 tracks
chr4:155,408,044
4,576M of 5,797M


Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
4 lines, 3 comments
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main

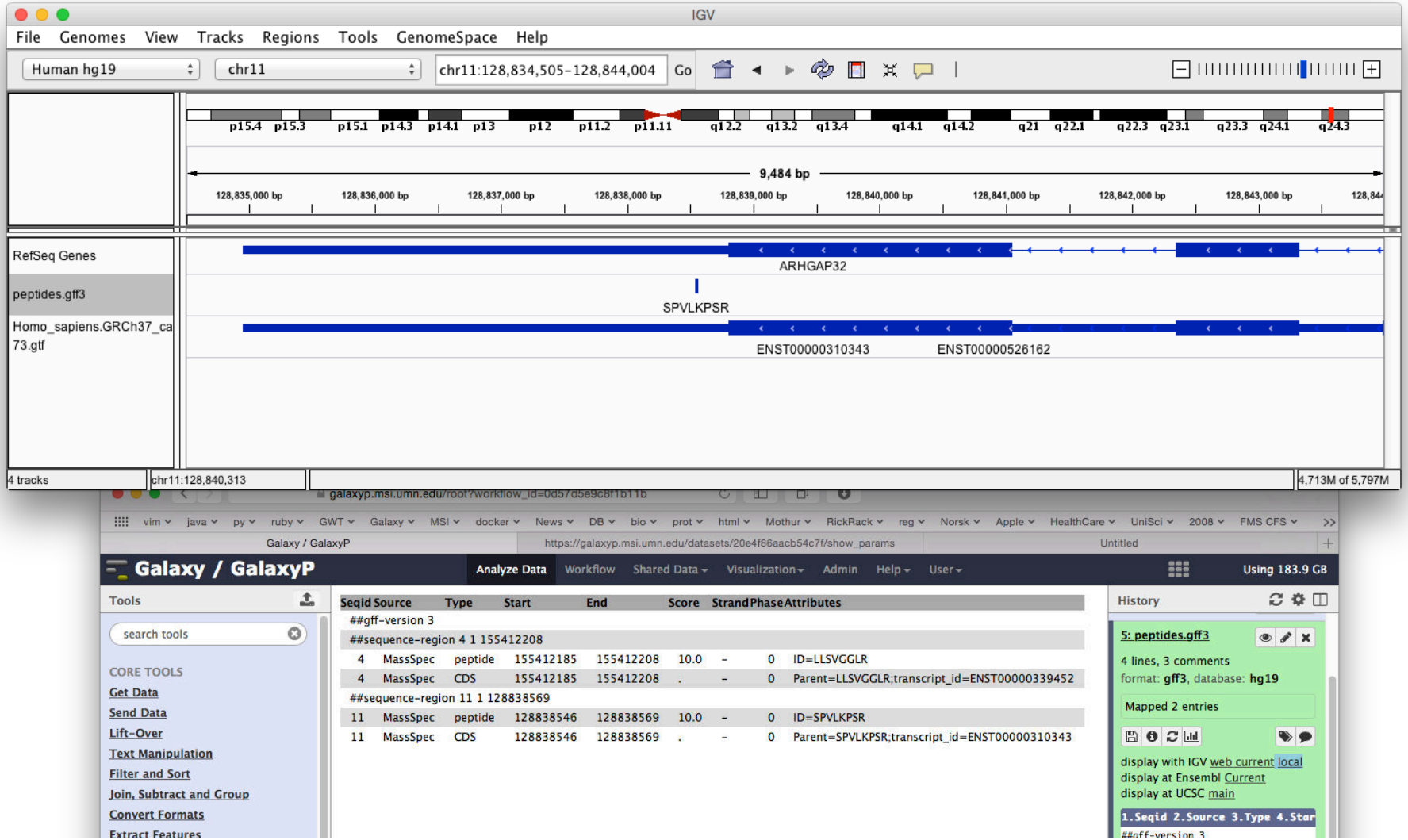

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr4
chr4:155,412,131-155,412,229
Go
99 bp
Sequence
RefSeq Genes
peptides.gff3
Homo_sapiens.GRCh37_ca
73.gtf
DCHS2
LLSVGGLR
ENST00000339452
LLSVGGLR
chr4:155412185-155412208
Type = peptide
ID: LLSVGGLR
Exon number: 1
chr4:155412185-155412208
Parent: LLSVGGLR
transcript_id: ENST00000339452
4 tracks
chr4:155,412,208
galaxyp.msi.umn.edu/root?workflow_id=0d57d5e9c8f1b11b
Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
4 lines, 3 comments
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main



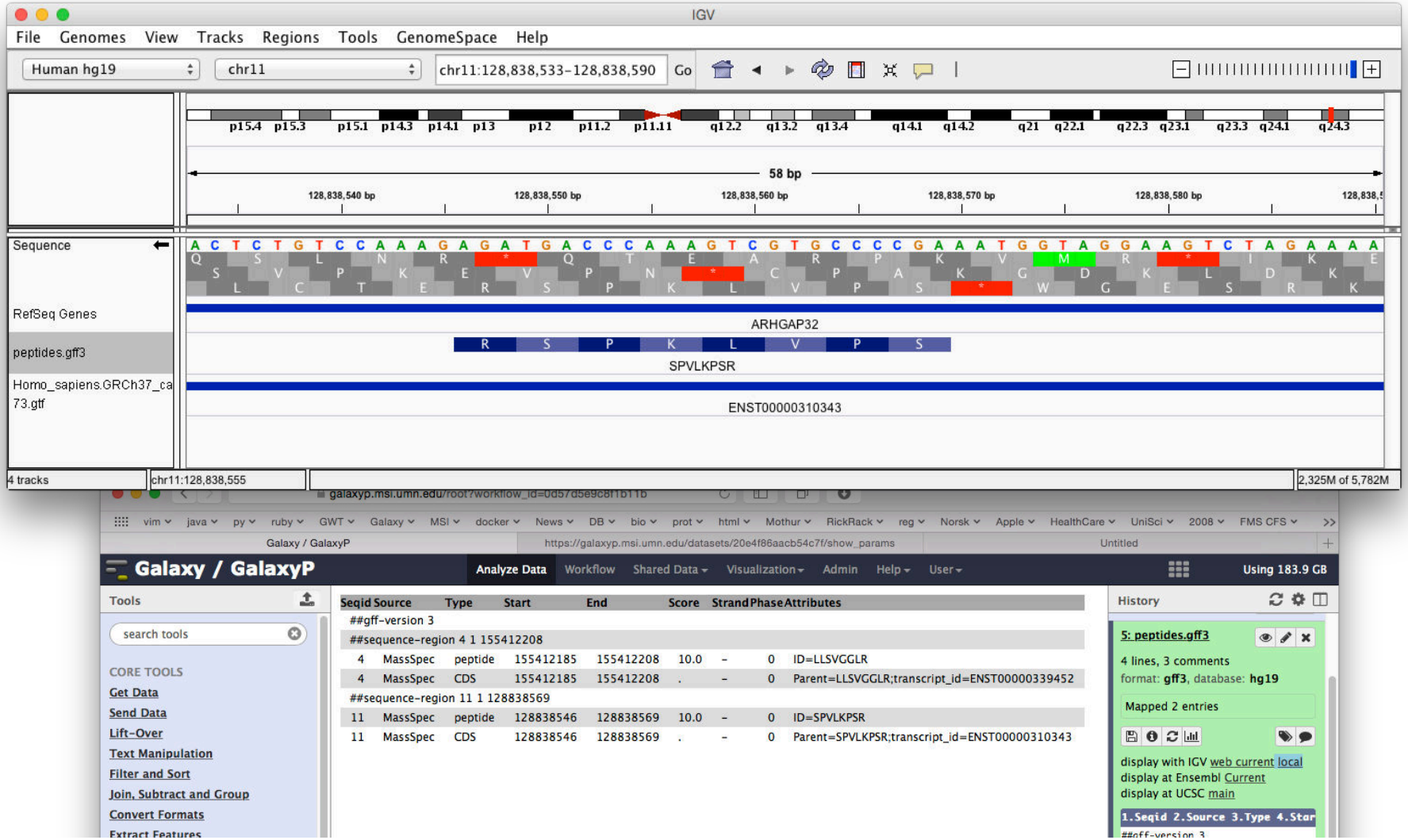
Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
##sequence-region 4 1 155412208								
4	MassSpec	peptide	155412185	155412208	10.0	-	0	ID=LLSVGGLR
4	MassSpec	CDS	155412185	155412208	.	-	0	Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569								
11	MassSpec	peptide	128838546	128838569	10.0	-	0	ID=SPVLKPSR
11	MassSpec	CDS	128838546	128838569	.	-	0	Parent=SPVLKPSR;transcript_id=ENST00000310343


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr11
chr11
Go
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
134 mb
Click and drag to zoom in.
nb 20 mb 40 mb 60 mb 80 mb 100 mb 120 mb
RefSeq Genes
peptides.gff3
Homo_sapiens.GRCh37_ca
73.gtf
28326 ILK USP47 TPH1 LUZP2 DCDC5 LRRC4C DDB2 TRIM48 ZP1 SIPA1 RNF121 GAB2 DLG2 FAT3 CNTN5 GRIA4 BTG4 SIK3 BLID ETS1 JAM
SPVLKPSR
00519787 ENST00000534211 ENST00000513853 ENST00000533565 ENST00000527985 ENST00000351960 ENST00000534163 ENST00000476452 ENST00000517061 ENST0000028143
4 tracks
chr11:117,559,881
4,698M of 5,797M

Galaxy / GalaxyP
Analyze Data Workflow Shared Data Visualization Admin Help User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
4 lines, 3 comments
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqid 2.Source 3.Type 4.Star

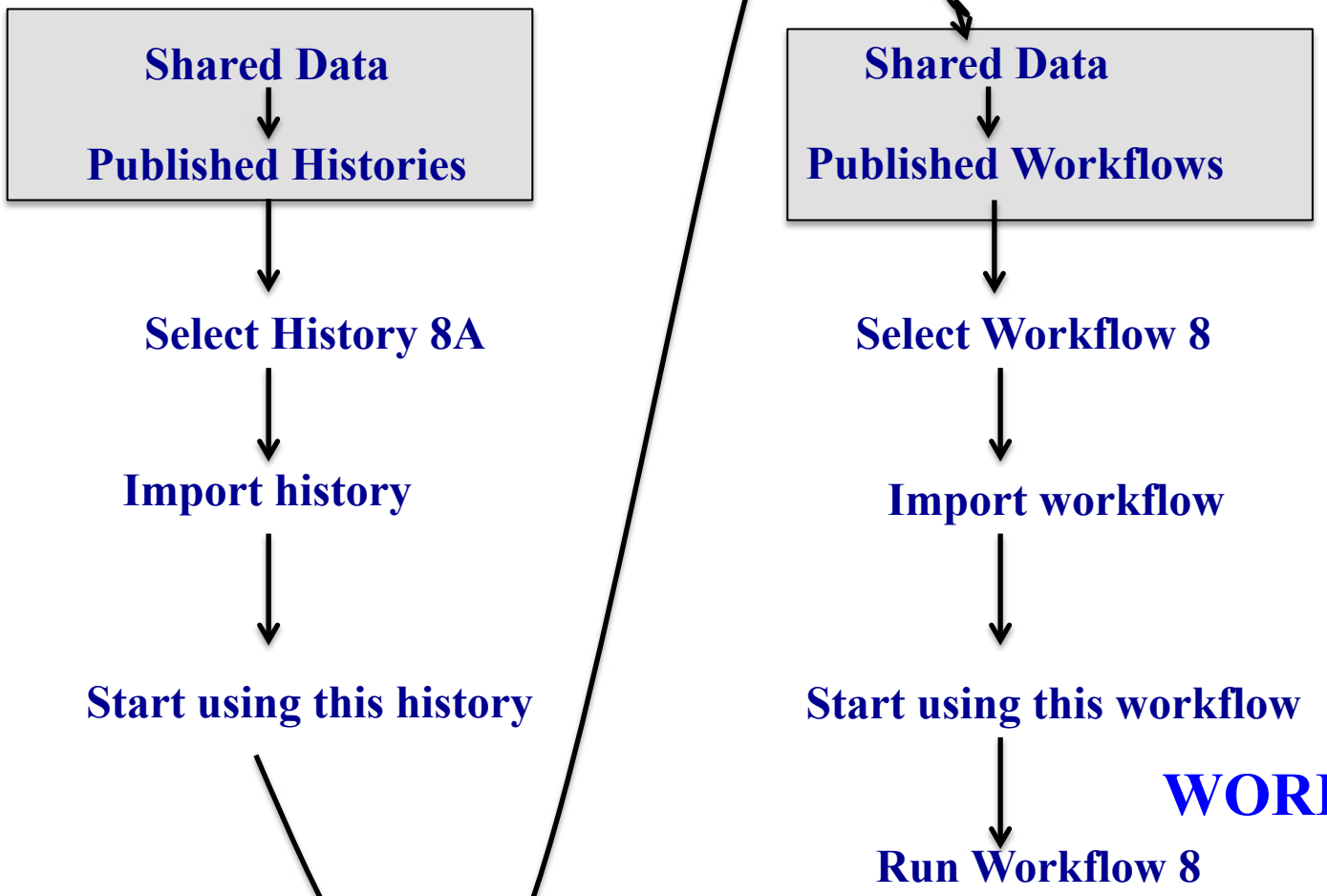

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr11
chr11:128,834,505-128,844,004
Go
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.11 q12.2 q13.2 q13.4 q14.1 q14.2 q21 q22.1 q22.3 q23.1 q23.3 q24.1 q24.3
9,484 bp
128,835,000 bp 128,836,000 bp 128,837,000 bp 128,838,000 bp 128,839,000 bp 128,840,000 bp 128,841,000 bp 128,842,000 bp 128,843,000 bp
RefSeq Genes
peptides.gff3
Homo_sapiens.GRCh37_ca
73.gtf
ARHGAP32
SPVLKPSR
ENST00000310343
ENST00000526162
4 tracks
chr11:128,840,313
4,713M of 5,797M

Galaxy / GalaxyP
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 183.9 GB
Tools
search tools
CORE TOOLS
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Seqid Source Type Start End Score Strand Phase Attributes
##gff-version 3
##sequence-region 4 1 155412208
4 MassSpec peptide 155412185 155412208 10.0 - 0 ID=LLSVGGLR
4 MassSpec CDS 155412185 155412208 . - 0 Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569
11 MassSpec peptide 128838546 128838569 10.0 - 0 ID=SPVLKPSR
11 MassSpec CDS 128838546 128838569 . - 0 Parent=SPVLKPSR;transcript_id=ENST00000310343
History
5: peptides.gff3
4 lines, 3 comments
format: gff3, database: hg19
Mapped 2 entries
display with IGV web current local
display at Ensembl Current
display at UCSC main
1.Seqid 2.Source 3.Type 4.Star



Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
##sequence-region 4 1 155412208								
4	MassSpec	peptide	155412185	155412208	10.0	-	0	ID=LLSVGGLR
4	MassSpec	CDS	155412185	155412208	.	-	0	Parent=LLSVGGLR;transcript_id=ENST00000339452
##sequence-region 11 1 128838569								
11	MassSpec	peptide	128838546	128838569	10.0	-	0	ID=SPVLKPSR
11	MassSpec	CDS	128838546	128838569	.	-	0	Parent=SPVLKPSR;transcript_id=ENST00000310343

8 Running the Entire Workflow

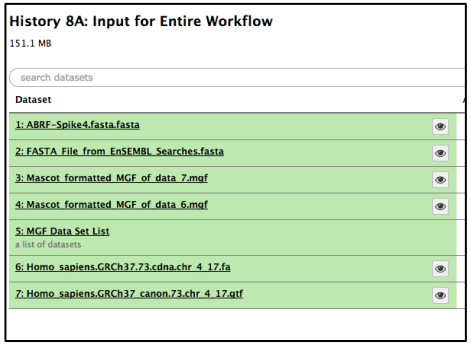


INPUT

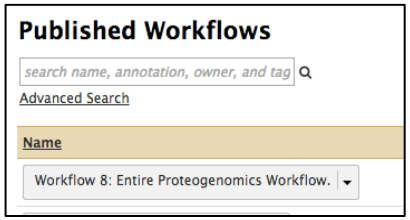
WORKFLOW

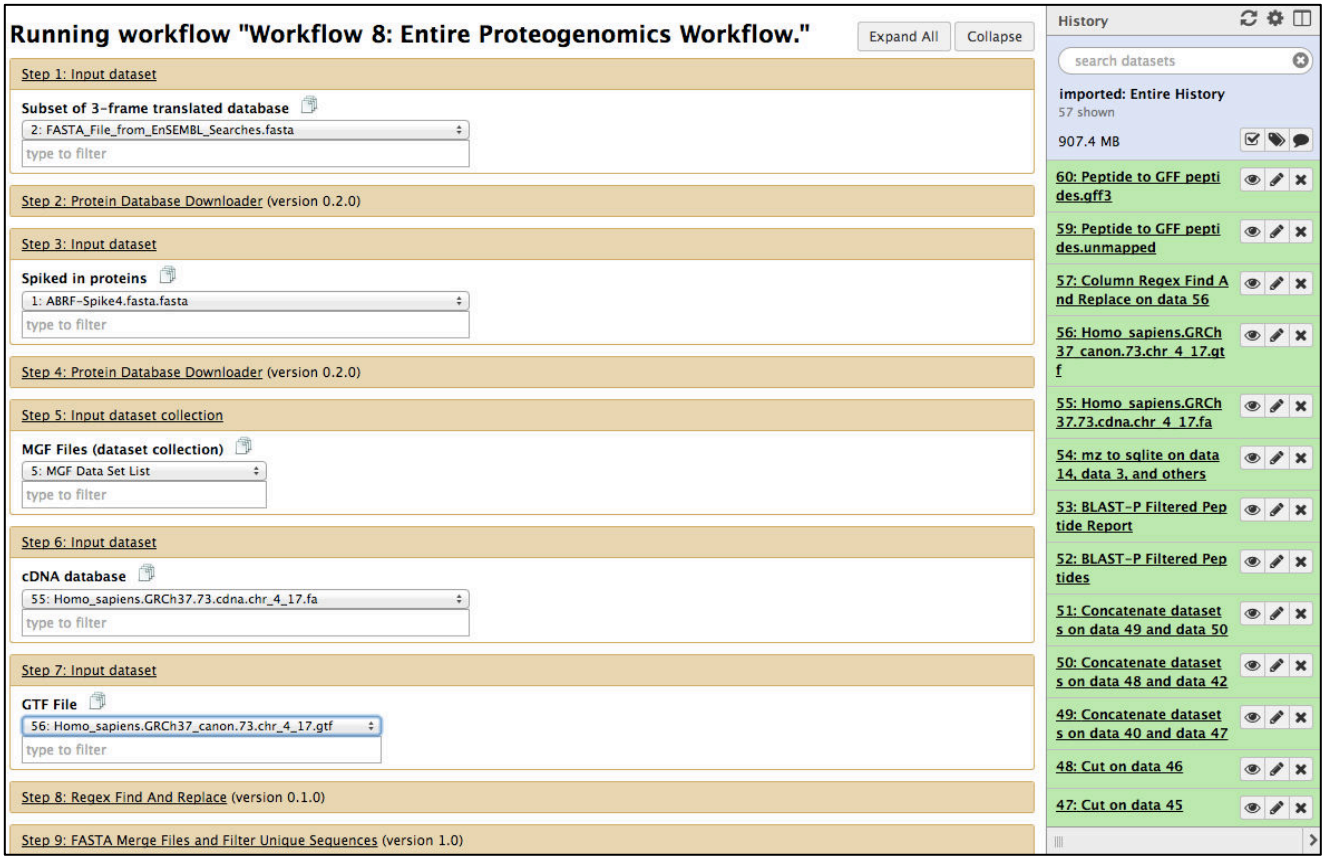
ENTIRE PROTEOGENOMICS WORKFLOW

INPUT



WORKFLOW





OUTPUT

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Biochemistry, Molecular Biology & Biophysics

Kevin Murray

Ray Sajulga

Candace Guerrero

UNIVERSITY OF MINNESOTA
SUPERCOMPUTING INSTITUTE

Tom McGowan

Trevor Wennblom

Getiria Onsongo

Bill Gallip

Ben Lynch

Center for Mass Spectrometry and Proteomics

Ebbing de Jong

LeeAnn Higgins

Todd Markowski

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

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

Funding

NSF, NIH