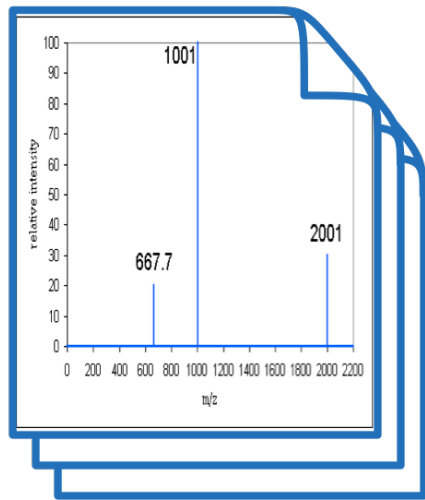


**GALAXY-BASED  
PEPTIDESHAKER TOOLS AND  
APPLICATIONS, *WITH A FOCUS  
ON DOWNSTREAM  
APPLICATIONS.***

**IRA COOKE**  
**PRATIK JAGTAP**



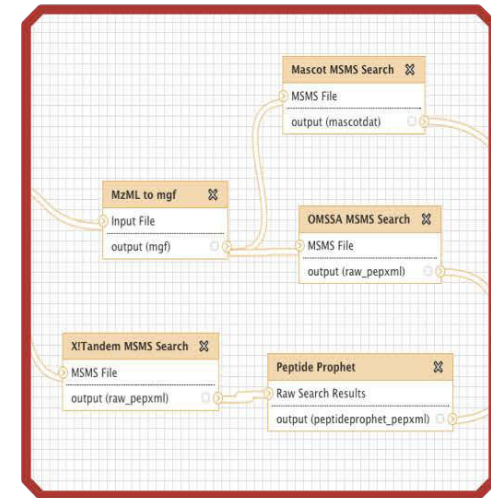
# Why PeptideShaker in Galaxy?



Big Data

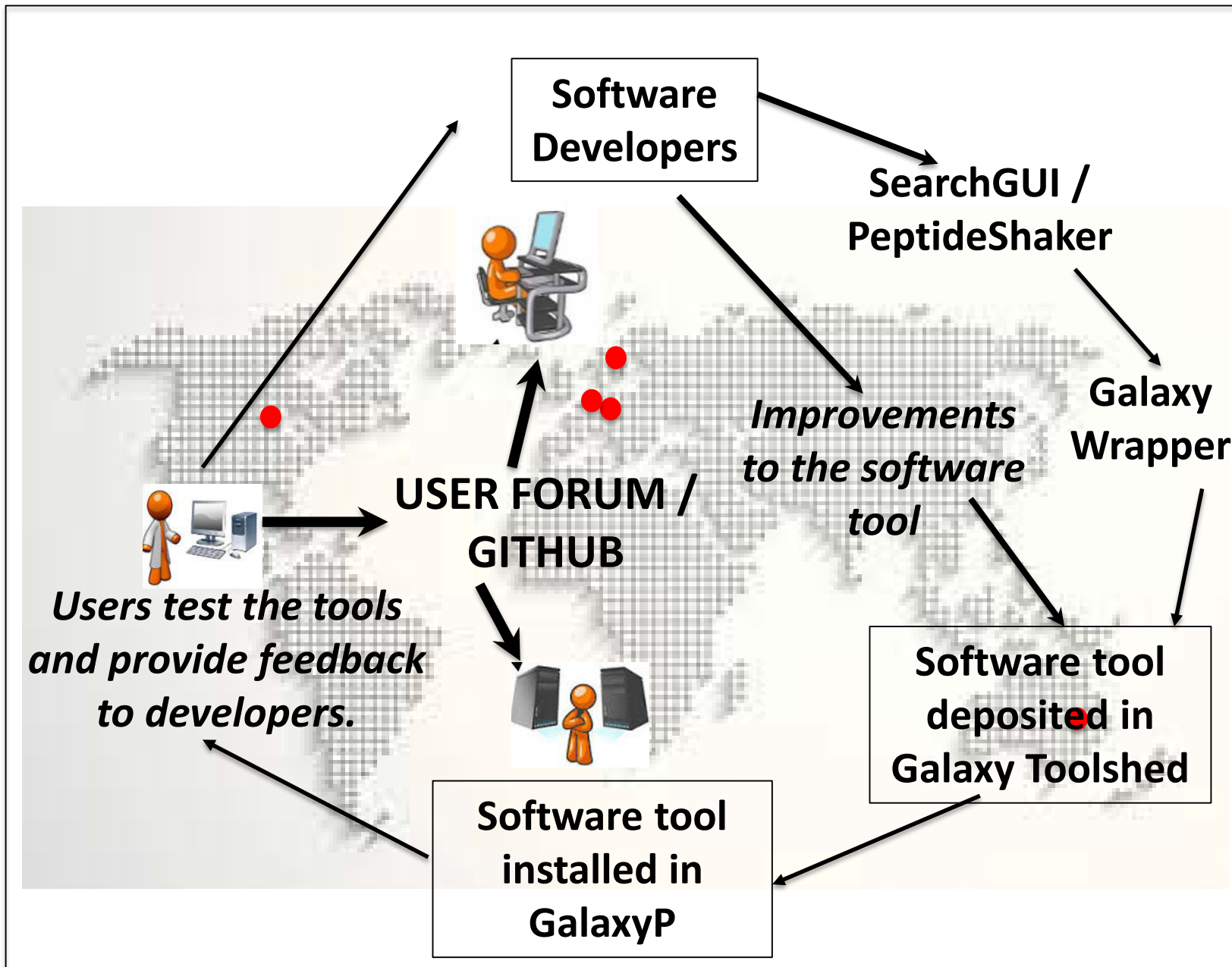


Visualisation

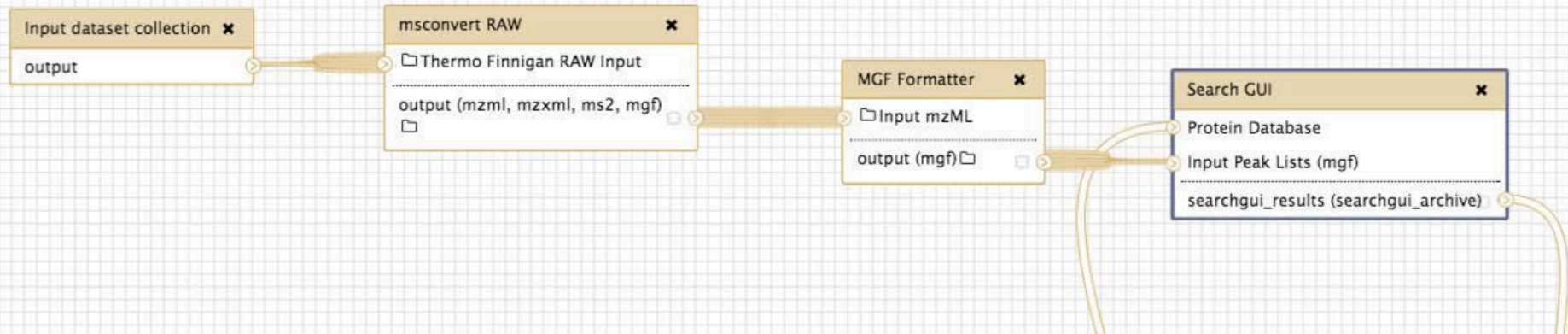


Workflows

# COMMUNITY-BASED SOFTWARE DEVELOPMENT

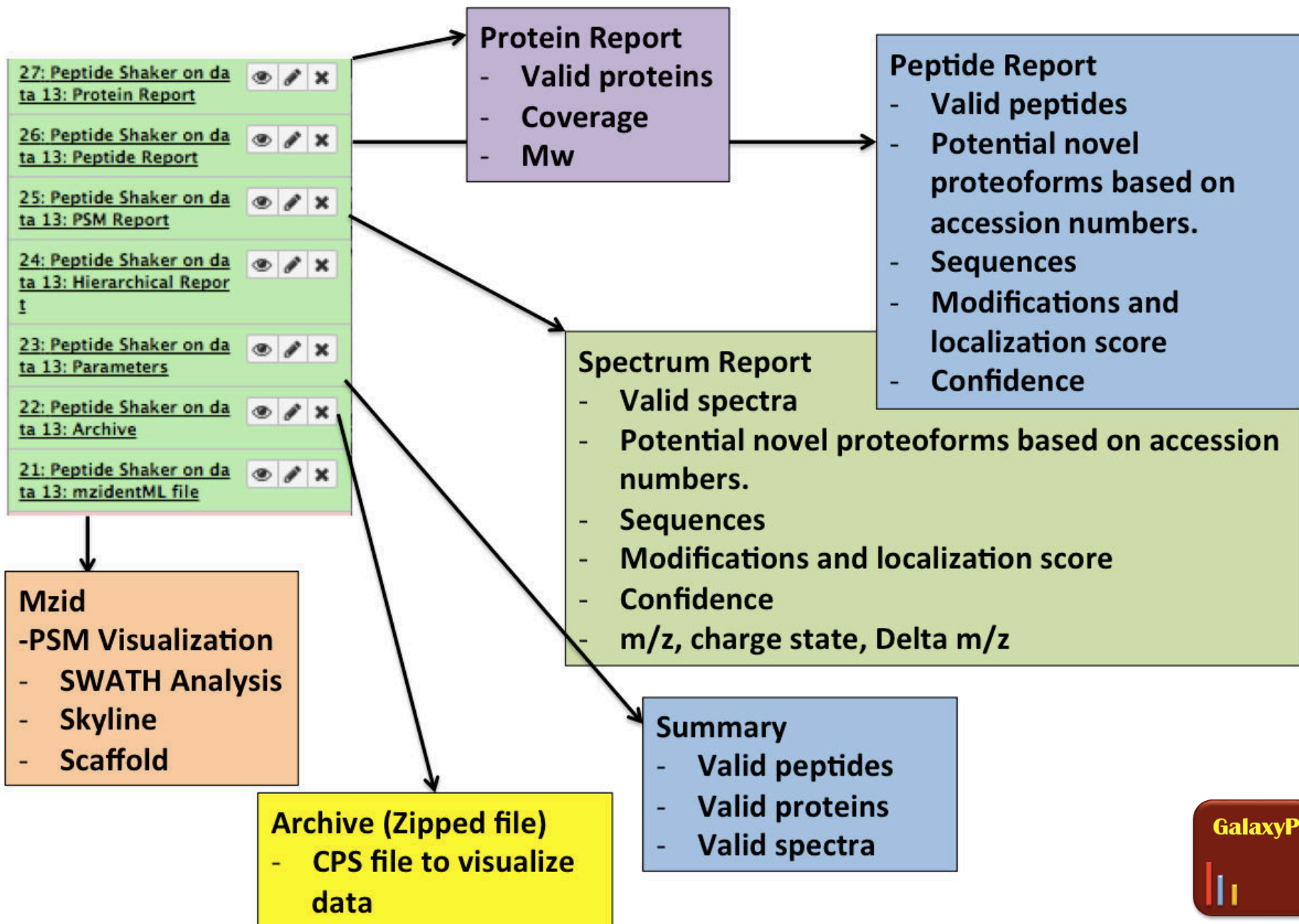


# GALAXY WORKFLOWS

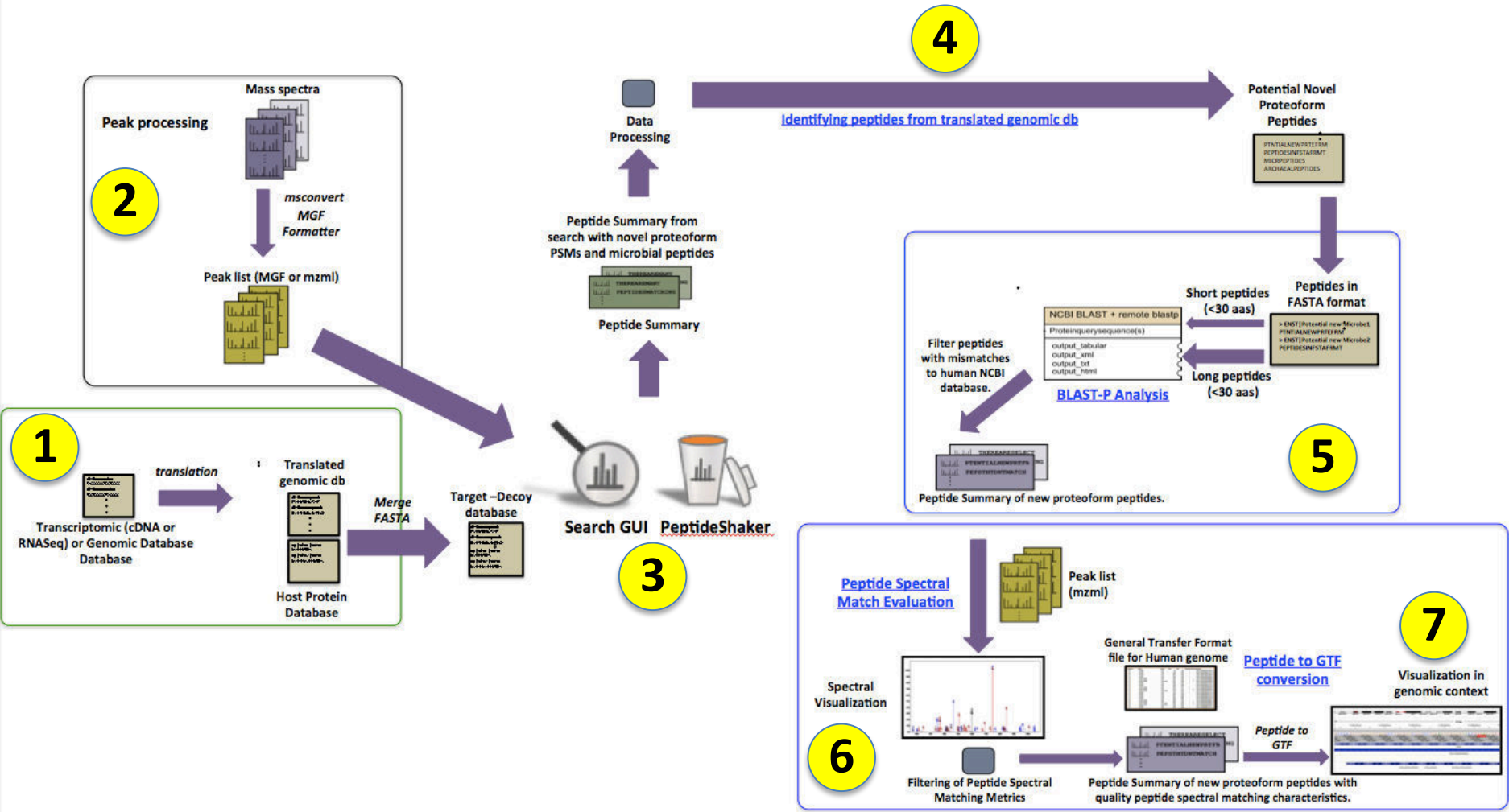


*Software tools can be used in a sequential manner to generate analytical workflows that can be reused, shared and creatively modified for multiple studies.*

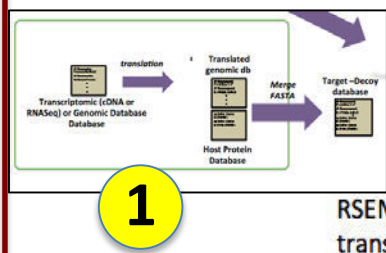
# OUTPUTS FROM SEARCHGUI / PEPTIDESHAKER



# PROTEOGENOMICS WORKFLOWS

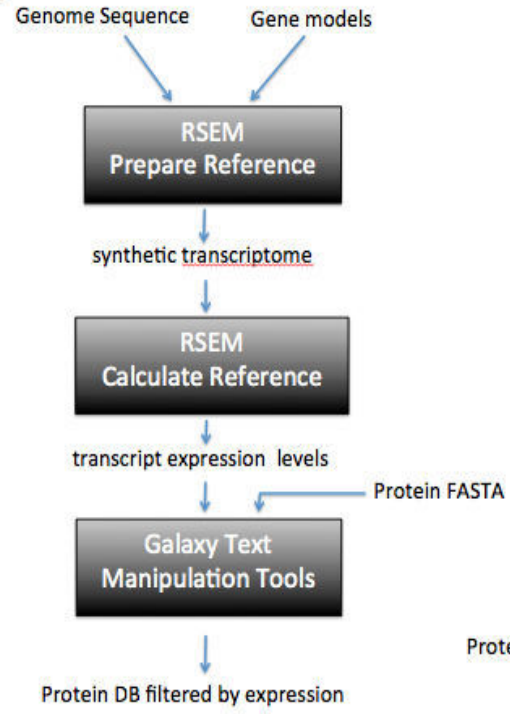


# RNA-SEQ DERIVED PROTEOMICS DATABASES



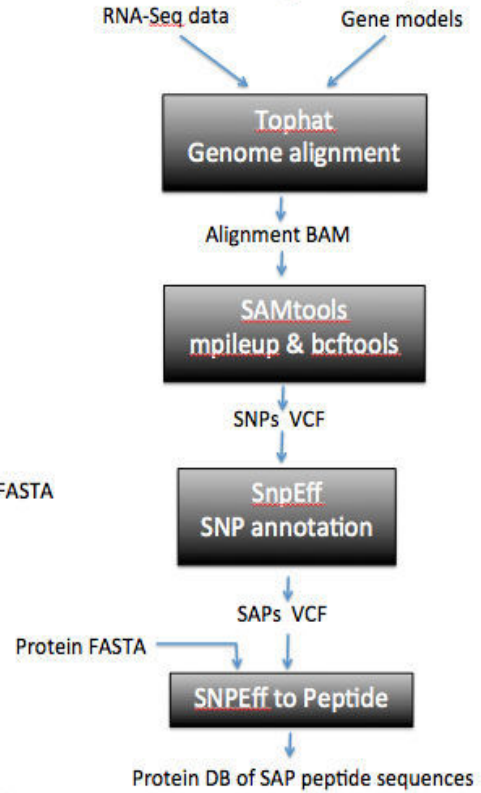
## Reduced Database

RSEM determines the RNA-Seq transcripts expressed at detectable levels. Proteins from transcripts that are not expressed are filtered out.



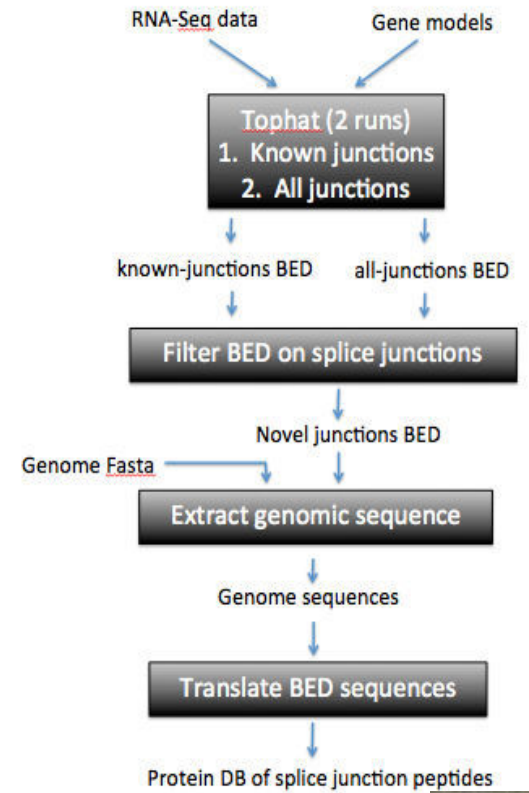
## SAP Database

RNA-Seq reads are aligned to the reference genome with tophat. SAMtools identifies variant DNA bases. SNPEff annotates the variants with effects to genes and proteins.



## Splice Database

Tophat alignments are used to find evidence of novel splice variant transcripts. The novel splice junctions are translated into a protein database.



“Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations.”  
 Sheynkman G et al BMC Genomics. doi: 10.1186/1471-2164-15-703.



# DATABASE SEARCH

Galaxy / GalaxyP

Analyze Data Workflow Shared Data Visualization Help User

Using 1.5.1

Tools

Workflow Canvas | GCC Workshop (RAW->mzml->Mascotmgf->SG\_0\_MSGF\_XTA->PS)

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

- Fetch Sequences
- Fetch Alignments
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Variant

EMBOSS

- Blast
- Picard

MISC

- Misc
- Beta Test Tools
- OpenMS

Workflow control

Inputs

Input dataset collection

msconvert RAW

MGF Formatter

Protein Database Downloader

FASTA Merge Files and Filter Unique Sequences

Search GUI

Peptide Shaker

Protein Database

Input Peak Lists (mgf)

searchgui\_results (searchgui\_archive)

Compressed SearchGUI results

mzidentML (mzid)

output\_cps (peptideshaker\_archive)

output\_zip (zip)

output\_certificate (txt)

output\_hierarchical (tabular)

output\_psm\_phosphorylation (tabular)

output\_psm (tabular)

output\_peptides\_phosphorylation (tabular)

output\_peptides (tabular)

output\_proteins\_phosphorylation (tabular)

output\_proteins (tabular)

output\_database (fasta)

Input dataset

Input FASTA File(s) 1 > FASTA File

Input FASTA File(s) 2 > FASTA File

Input FASTA File(s) 3 > FASTA File

Input FASTA File(s) 4 > FASTA File

output (fasta)

output\_database (fasta)

Peptide Summary from search with novel proteoform PSMs and microbial peptides

Peptide Summary

Target-Decoy database

Search GUI

PeptideShaker

3

History

search datasets

May 7 - GCC Workshop (RAW->mzml->Mascotmgf->SG\_0\_MSGF\_XTA->PS)

14 shown, 13 deleted, 6 hidden

818.7 MB

- 27: Peptide Shaker on data 13: Protein Report
- 26: Peptide Shaker on data 13: Peptide Report
- 25: Peptide Shaker on data 13: PSM Report
- 24: Peptide Shaker on data 13: Hierarchical Report
- 23: Peptide Shaker on data 13: Parameters
- 22: Peptide Shaker on data 13: Archive
- 21: Peptide Shaker on data 13: mzidentML file
- 13: Search GUI on data 9, data 11, and data 10
- 12: MGF Formatter across collection 8
- 9: Merged and Filtered FASTA from data 5, data 2, and data 4
- 8: msconvert RAW across collection 3
- 4: Regex Replace on data 11
- 3: ABRF-Spike4.fasta
- 1: New Dataset List



# SEARCH GUI / PEPTIDESHAKER SEARCH

https://galaxy.msi.umn.edu

Galaxy / GalaxyP

Analyze Data Workflow Shared Data Visualization Help User

Using 1.5 TB

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
  - Fetch Sequences
  - Fetch Alignments
  - NGS: Mapping
  - NGS: RNA Analysis
  - NGS: SAM Tools
  - NGS: Variant
  - EMBOSS
  - Blast
  - Picard
- MISC
  - Misc
  - Beta Test Tools
  - OpenMS
- Workflows
  - 795 NGS For Rodney datasets
  - Workflow for paired metaproteomics comparison studies
  - MPL: Workflow for paired metaproteomics comparison studies - HOMD db search (imported from uploaded file)
  - All workflows

## Basic or Advanced Search options

Advanced

### Run X!Tandem search

Search with X!Tandem

**X!Tandem: Total Peaks**  
50  
Maximum number of peaks to be used from a spectrum

**X!Tandem: Min Peaks**  
15  
Minimum number of peaks required for a spectrum to be considered

**X!Tandem: Min Frag m/z**  
200  
Fragment mass peaks with m/z less than this value will be discarded

**X!Tandem: Min Precursor Mass**  
200  
Minimum mass of 1+ mass of parent ion to be considered

**X!Tandem: Noise Suppression**  
Yes No  
Use noise suppression

**X!Tandem peptide model refinement**  
Don't refine

### Run OMSSA search

Search with OMSSA

**OMSSA: Hit List Length**  
25

**OMSSA: Remove Precursor**  
Yes No

**OMSSA: Scale Precursor Mass**  
Yes No

**OMSSA: Estimate Charge**  
Yes No

### Run MSGF search

Search with MSGF

**Search Decoys**  
Yes No  
If yes then a decoy database will be generated and searched. Assumed input database contains no decoys

**Minimum Peptide Length**  
6  
Minimum length for a peptide to be considered

**Maximum Peptide Length**  
50  
Maximum length for a peptide to be considered

**Number of tolerable termini**  
2 (ie fully-tryptic cleavage)  
Searches will take much longer if selecting a value other than 2

**Max PTMs per peptide**  
2

### Peptide Shaker Perform protein identification using various search engines based on results from SearchGUI (Galaxy Tool Version 0.37.0)

Compressed SearchGUI results  
13: Search GUI on data 9, data 11, and data 10  
SearchGUI Results from History

The species type to use for the gene annotation

#### Specify Advanced PeptideShaker Processing Options

Advanced Processing Options

**FDR at the protein level**  
1  
In percent (default 1% FDR: '1')

**FDR at the peptide level**  
1  
In percent (default 1% FDR: '1')

**FDR at the PSM level**

Maximum Precursor Error  
10  
Next option specifies units (Da or ppm)

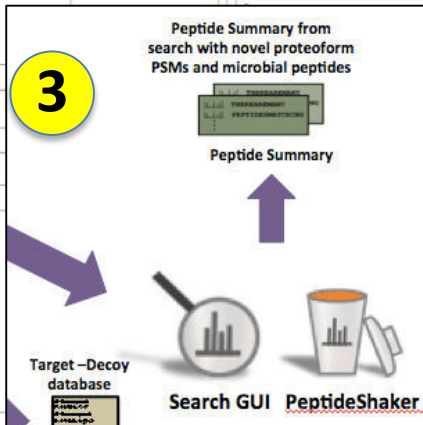
**Maximum Precursor Error Type**  
ppm

**Exclude Unknown PTMs**  
Yes No

#### Output options

Select/Unselect all

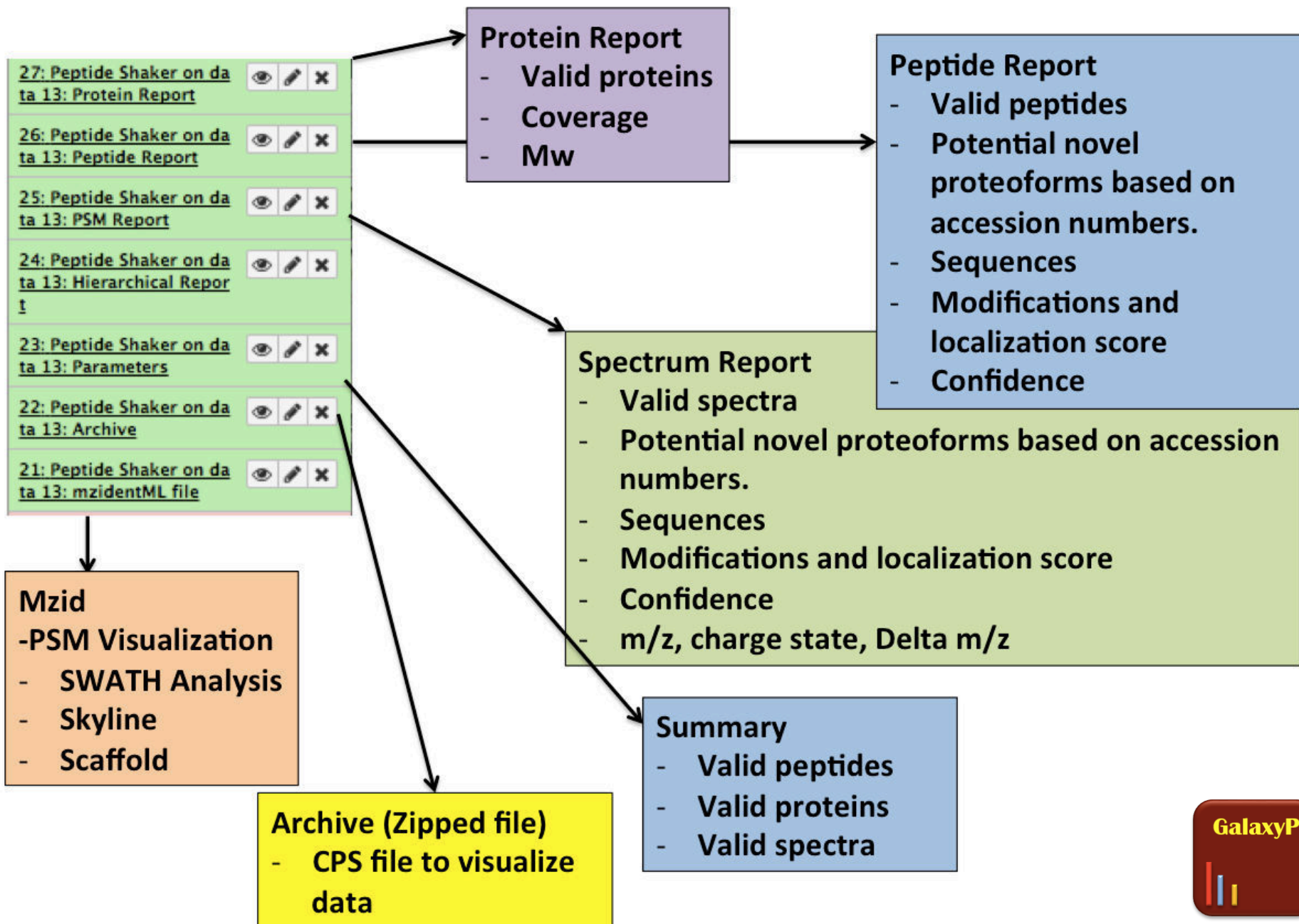
- Zip File for import to Desktop App
- mzidentML File
- PSM Report
- Peptide Phosphorylation Report
- Peptide Report



History

- search datasets
- May 7 - GCC Workshop (RAW->xml->Mascotmgf->SG\_O\_MSGF\_XTA->PS) 19 shown, 13 deleted, 6 history 818.7 MB
- 27: Peptide Shaker on data 13: Protein Report
- 26: Peptide Shaker on data 13: Peptide Report
- 25: Peptide Shaker on data 13: PSM Report
- 24: Peptide Shaker on data 13: Hierarchical Report
- 23: Peptide Shaker on data 13: Parameters
- 22: Peptide Shaker on data 13: Archive
- 21: Peptide Shaker on data 13: mzidentML file
- 13: Search GUI on data 9, data 11, and data 10 154.7 MB  
format: searchgui\_archive, database: 2  
Creating decoy database. Rendering: input\_database.fasta 10% 20% 30% 40% 50% 60% 70% 80% 90%Input: /panfs/roo/groups/1/msc/pep/apps/ga  
Name: input\_database.fasta Version: 7.5.2015 Decoy Tag:
- 12: MGF Formatter across collection 8
- 9: Merged and Filtered FASTA from data 5, data 2, and data 4
- 8: msconvert RAW across collection 3
- 4: Regex Replace on data 11
- 3: ABRF-Spike4.fasta
- 1: New Dataset List

# OUTPUTS FROM SEARCHGUI / PEPTIDESHAKER





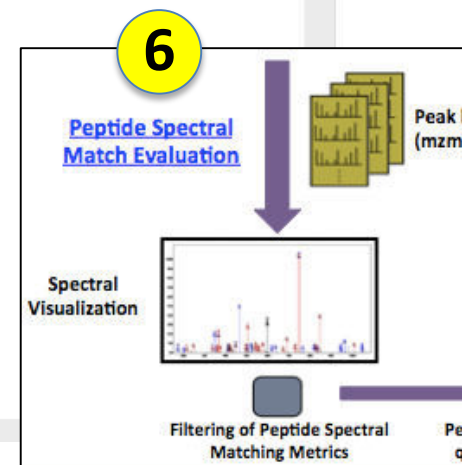
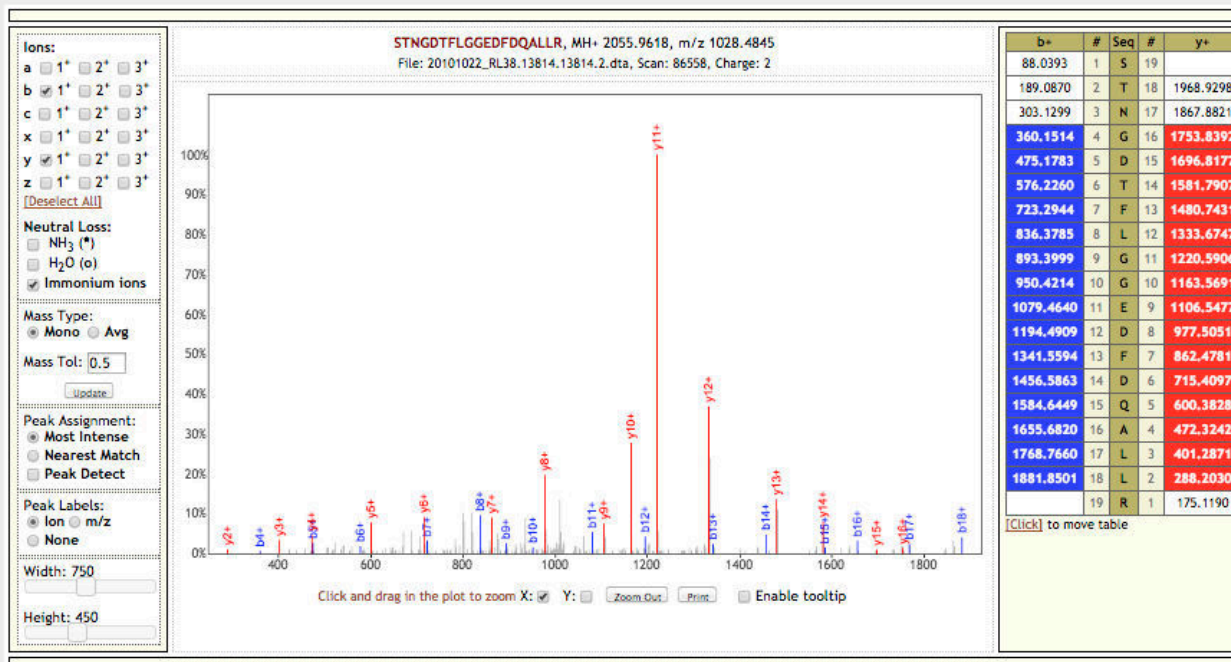
# PSM EVALUATION

Search:

Mascot:identity threshold	Mascot:score	Scaffold: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	MKETAENYLGHAK	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.

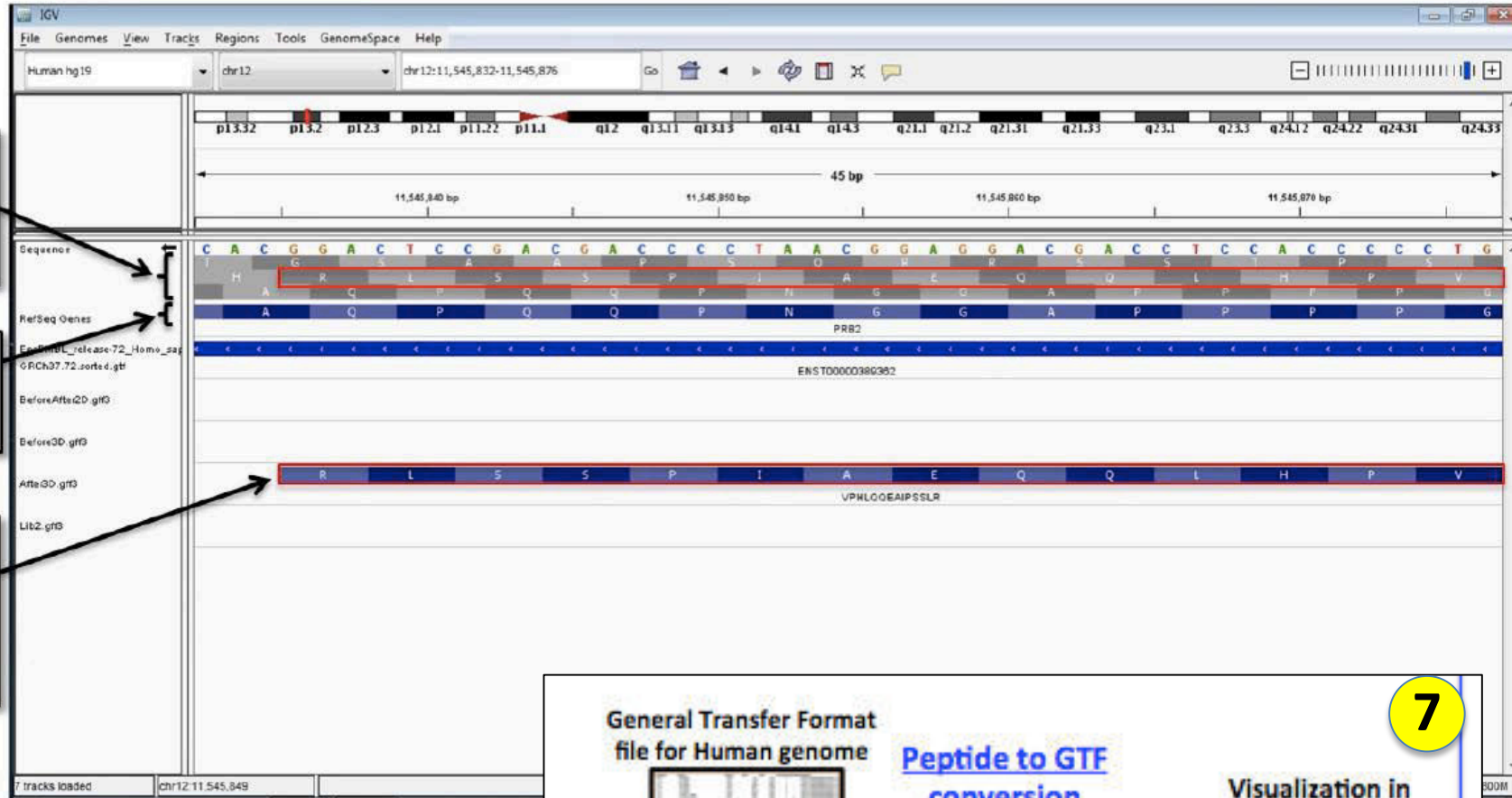
STNGDTFLGGEDFDQALLR 86558



June 2<sup>nd</sup> 2015: Tuesday

POSTER 131: Plugging Proteomics Peptide-Spectral Match Visualization into Galaxy. (Johnson et al)

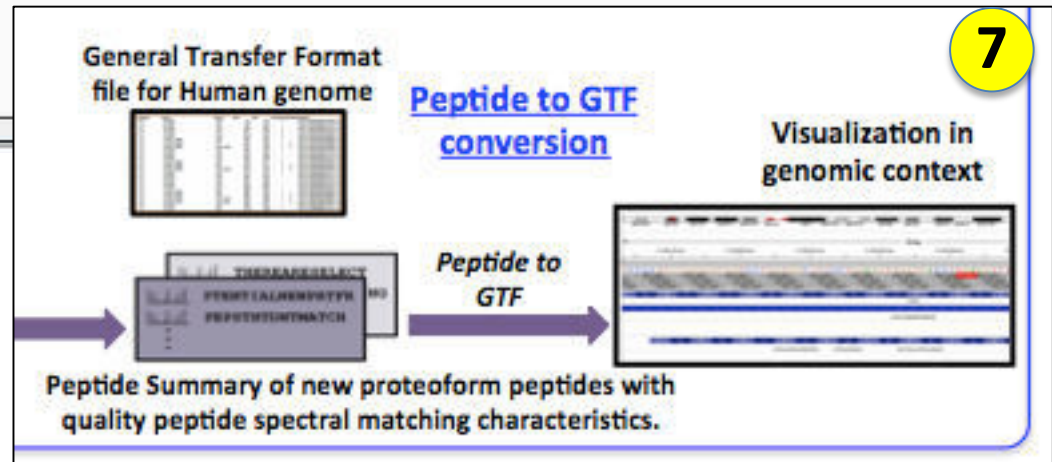
# GENOME VISUALIZATION USING IGV BROWSER



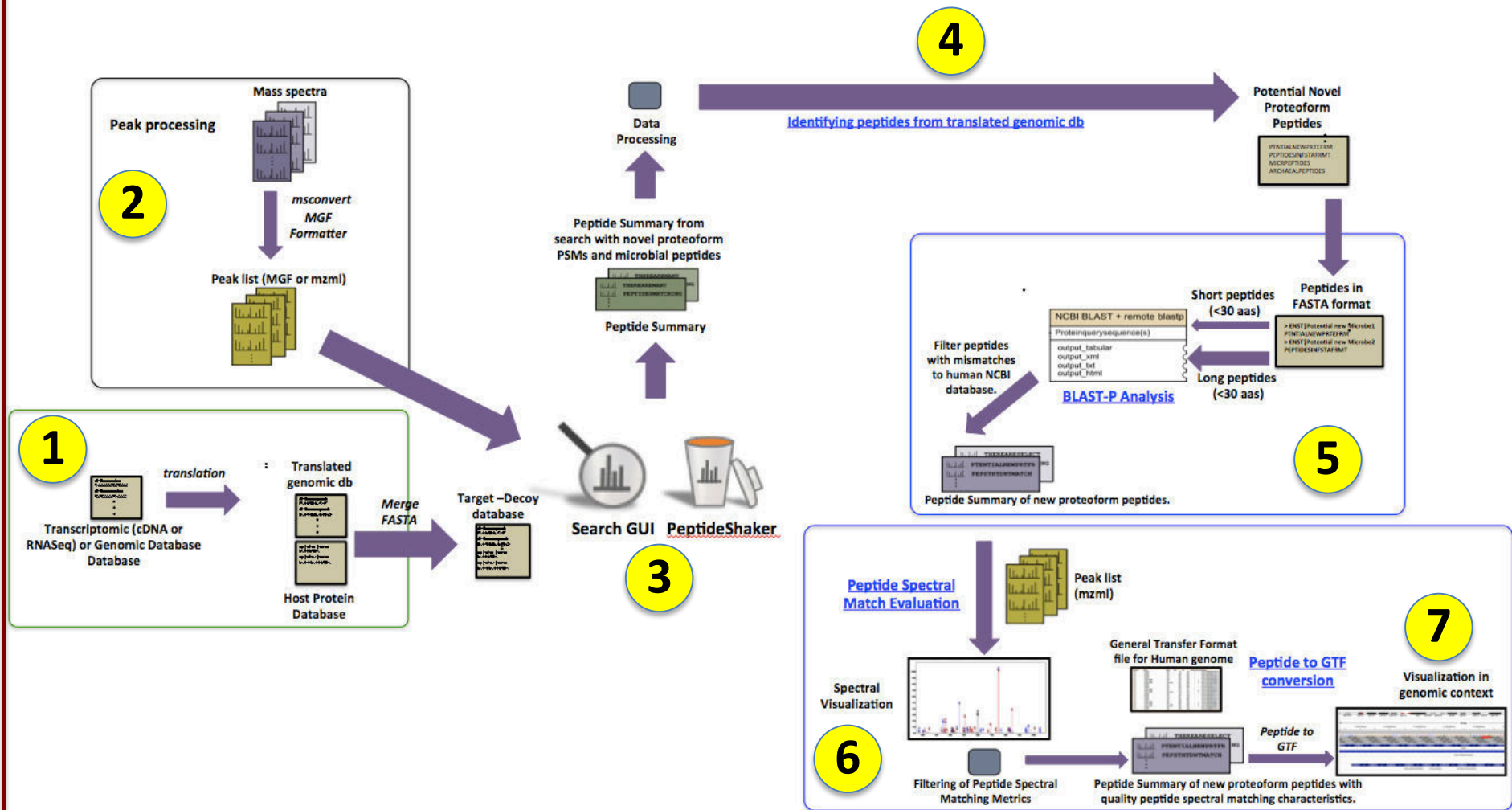
3-frame translated peptide sequences

Reference peptide sequence

Identified novel proteoform peptide sequence



# PROTEOGENOMICS WORKFLOWS

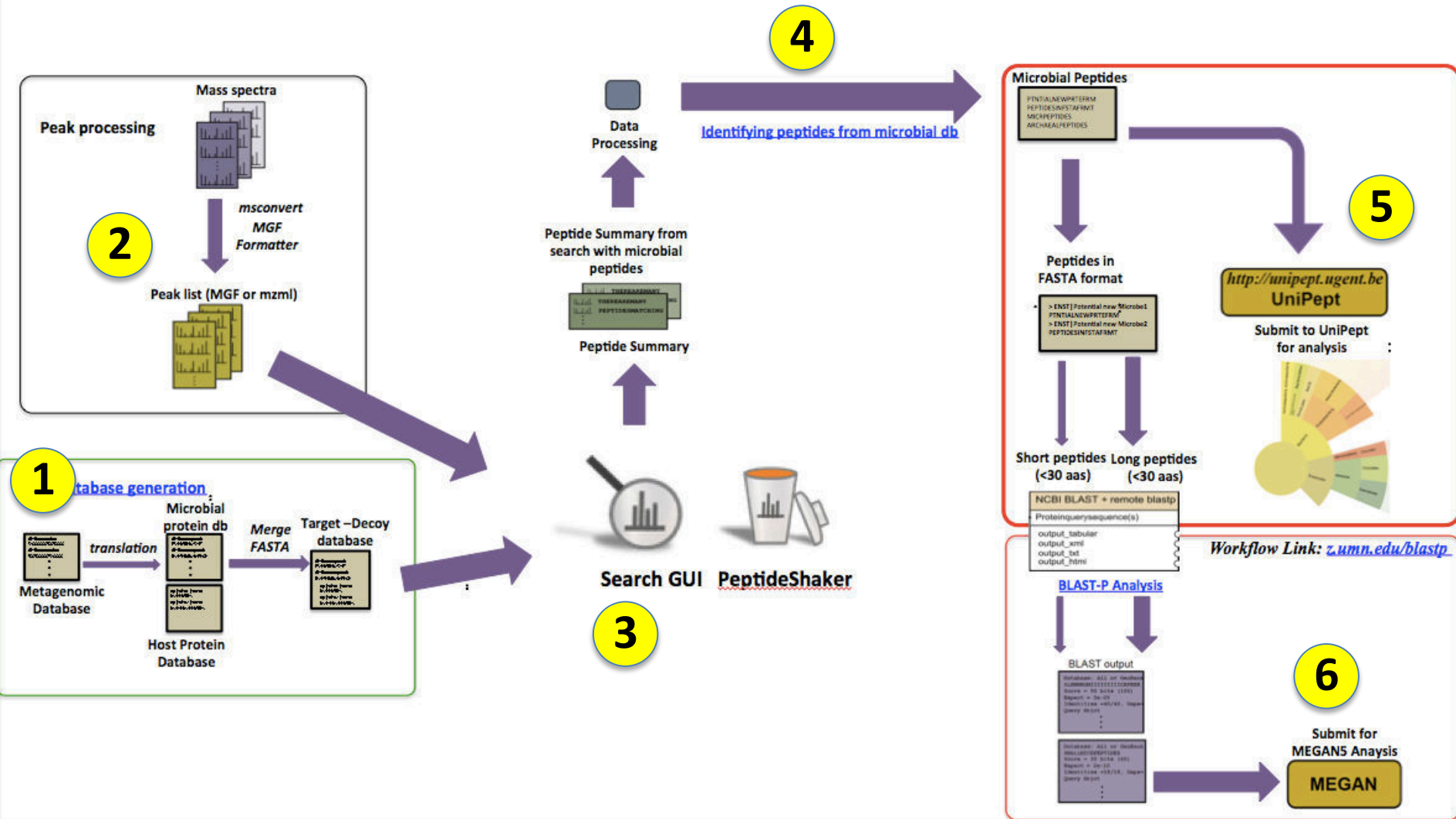


Jagtap et al *J Proteome Res.* 2014 13:5898-908

<http://z.umn.edu/pgfirstlook>



# METAPROTEOMICS WORKFLOWS



- **SearchGUI/PeptideShaker is an excellent open-source resource that was generated through community-based development and generates multiple inputs for analytical workflows.**

June 2<sup>nd</sup> 2015: Tuesday

**ORAL PRESENTATION: 9:50 AM ; Room 130/132** SearchGUI and PeptideShaker deployed in the Galaxy framework: A powerful informatics platform for protein identification and beyond. (**Cooke et al**)

- **GalaxyP workflows for proteogenomics and metaproteomics analysis are available and used in projects.**

**POSTER 131:** Plugging Proteomics Peptide-Spectral Match Visualization into Galaxy. (**Johnson et al**)

**POSTER 638:** Revealing Pathways In COPD-Associated Lung Cancer Large-Scale Quantitative Multiomic Analysis. (**Sandri et al**)

**POSTER 366:** Metaproteomic analysis of human cervical-vaginal fluid in residual Pap tests: Insights into the cervical microbiome. (**Griffin et al**)

June 4<sup>th</sup> 2015: Thursday

**POSTER 455:** A Novel Analytical-Informatics Platform Reveals the Hidden Tryptic Peptidome and Improves Multi-omic Applications. (**Guerrero et al**)

- **We are planning to integrate complex workflows such as OpenSWATH within GalaxyP.**

June 2<sup>nd</sup> 2015: Tuesday

**POSTER 127:** Democratizing and expanding the reach of DIA Mass Spectrometry: Developing OpenSWATH tools and workflows within user-friendly Galaxy-P platform. (**Jagtap et al**)

[tinyurl.com/gpasms15](http://tinyurl.com/gpasms15)

