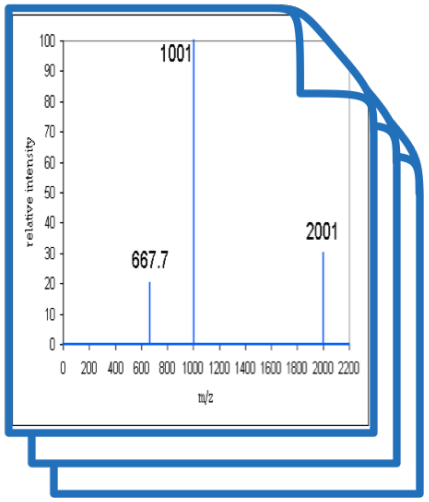
Galaxy-based PeptideShaker tools and applications, *with a focus on DownStream Applications.*

Ira Cooke
Pratik Jagtap

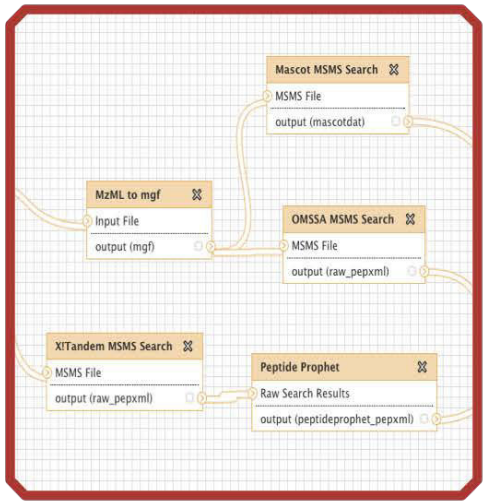
GalaxyP

Why PeptideShaker in Galaxy?



Big Data

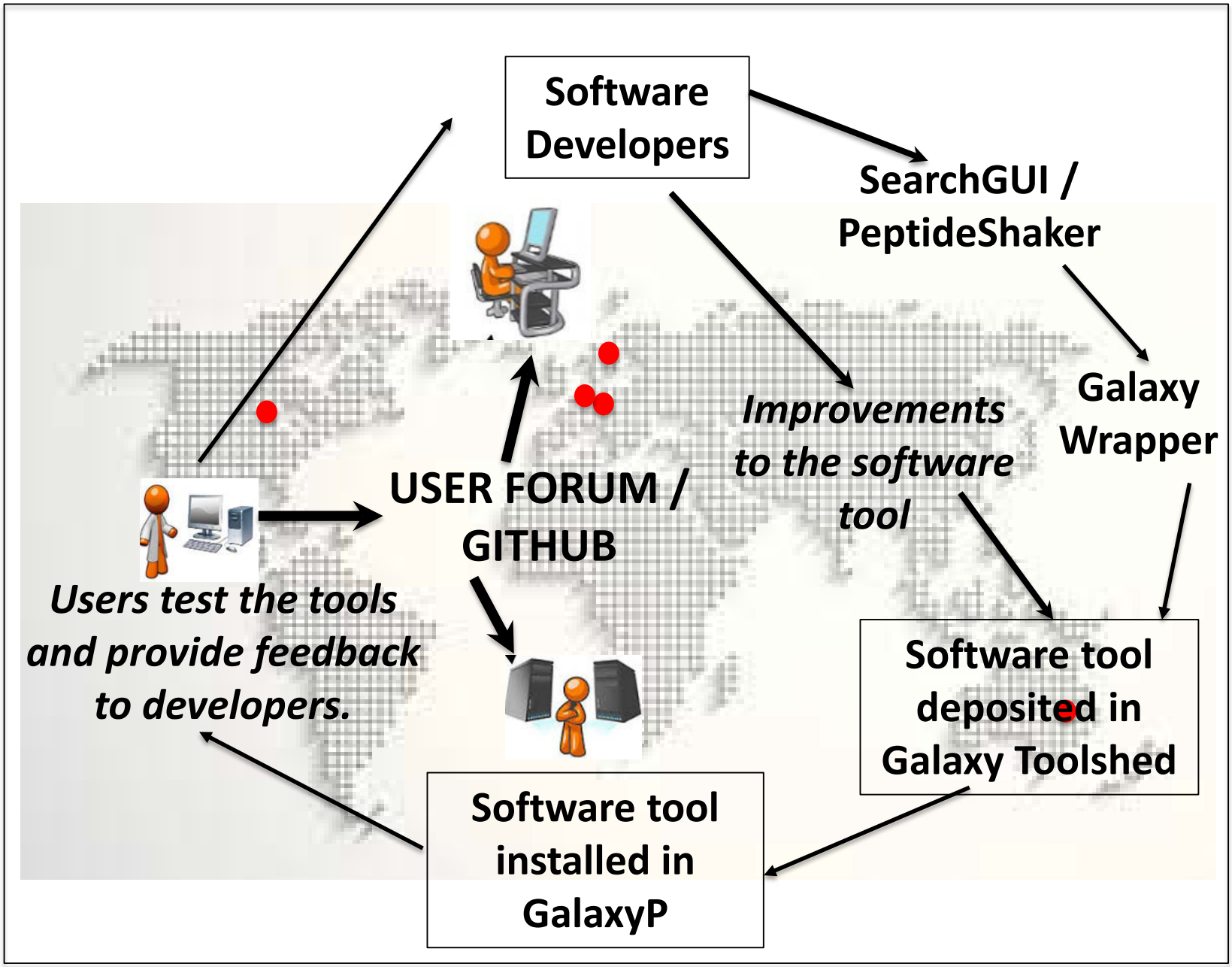
Visualisation



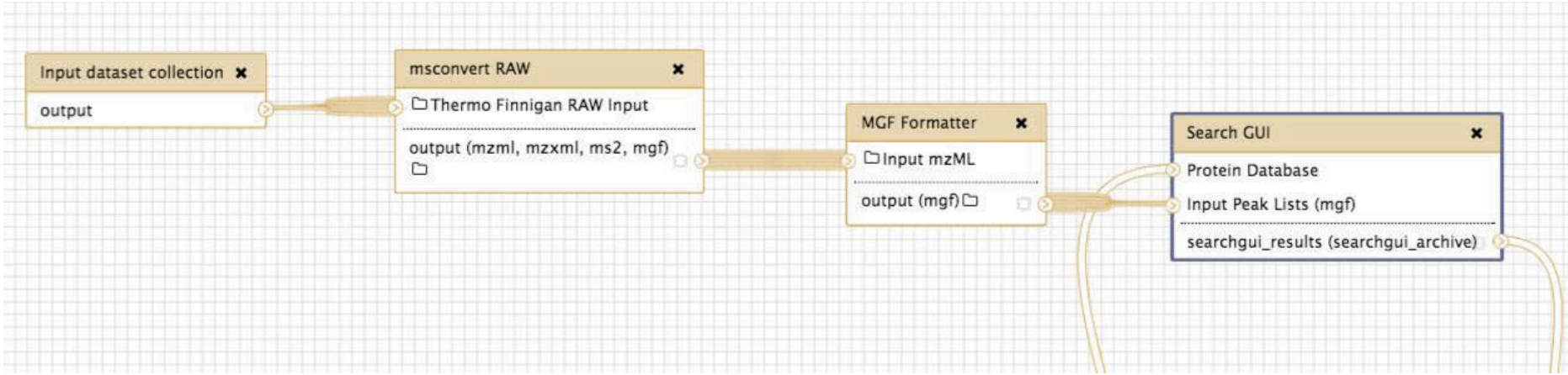
Workflows

GalaxyP

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.*

GalaxyP

OUTPUTS FROM SEARCHGUI / PEPTIDESHAKER

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

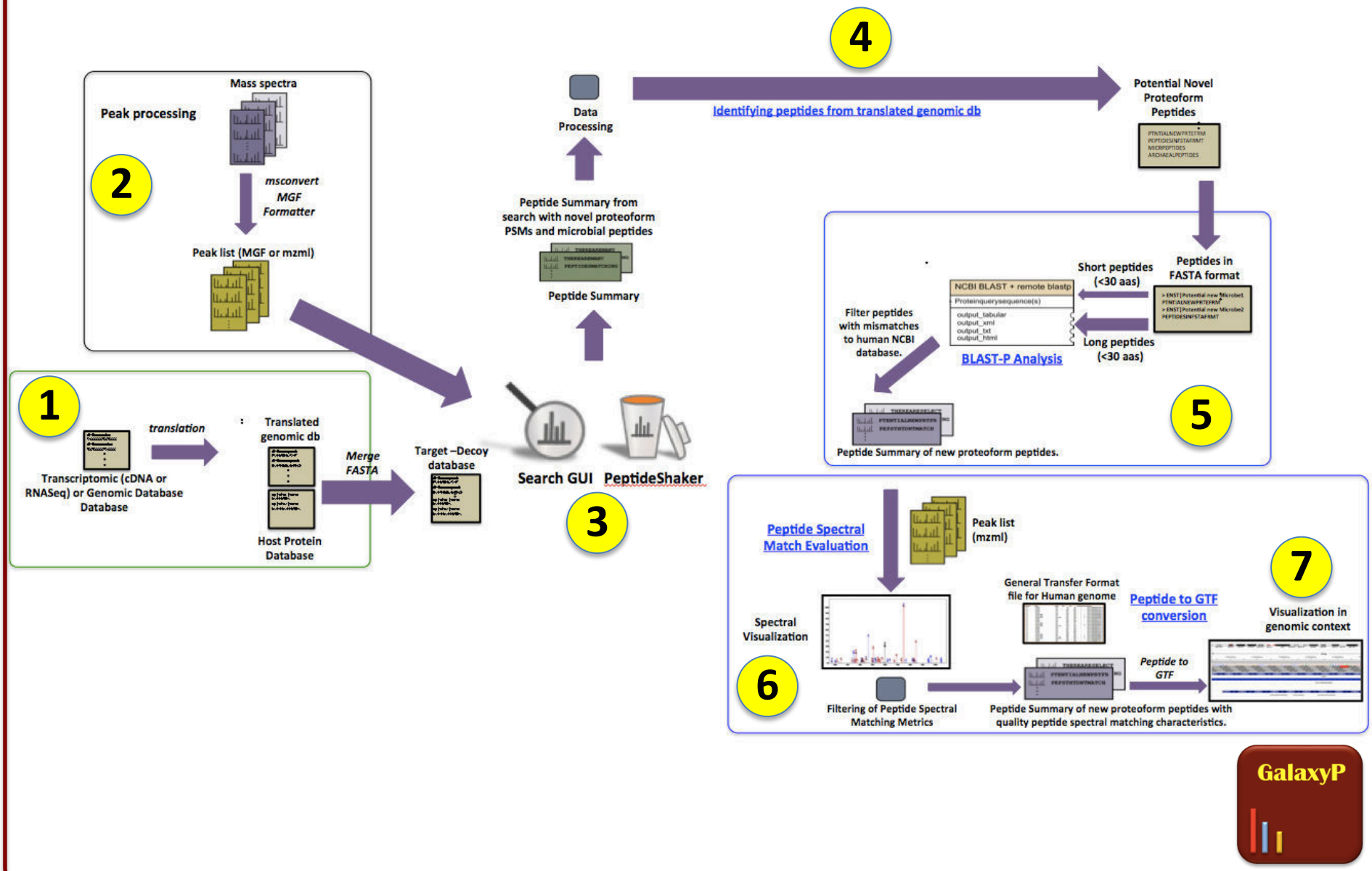
Mzid
-PSM Visualization
- SWATH Analysis
- Skyline
- Scaffold

Archive (Zipped file)
- CPS file to visualize data

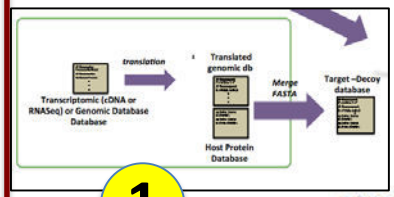
Summary
- Valid peptides
- Valid proteins
- Valid spectra

GalaxyP

PROTEOGENOMICS WORKFLOWS



RNA-SEQ DERIVED PROTEOMICS DATABASES



1

Reduced Database

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

Genome Sequence + Gene models → RSEM Prepare Reference → synthetic transcriptome → RSEM Calculate Reference → transcript expression levels + Protein FASTA → Galaxy Text Manipulation Tools → Protein DB filtered by expression

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.

RNA-Seq data + Gene models → Tophat Genome alignment → Alignment BAM → SAMtools mpileup & bcftools → SNPs VCF → SnpEff SNP annotation → SAPs VCF + Protein FASTA → SNPEff to Peptide → Protein DB of SAP peptide sequences

Splice Database

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

RNA-Seq data + Gene models → Tophat (2 runs) 1. Known junctions 2. All junctions → known-junctions BED + all-junctions BED → Filter BED on splice junctions → Novel junctions BED + Genome Fasta → Extract genomic sequence → Genome sequences → Translate BED sequences → Protein DB of splice junction peptides

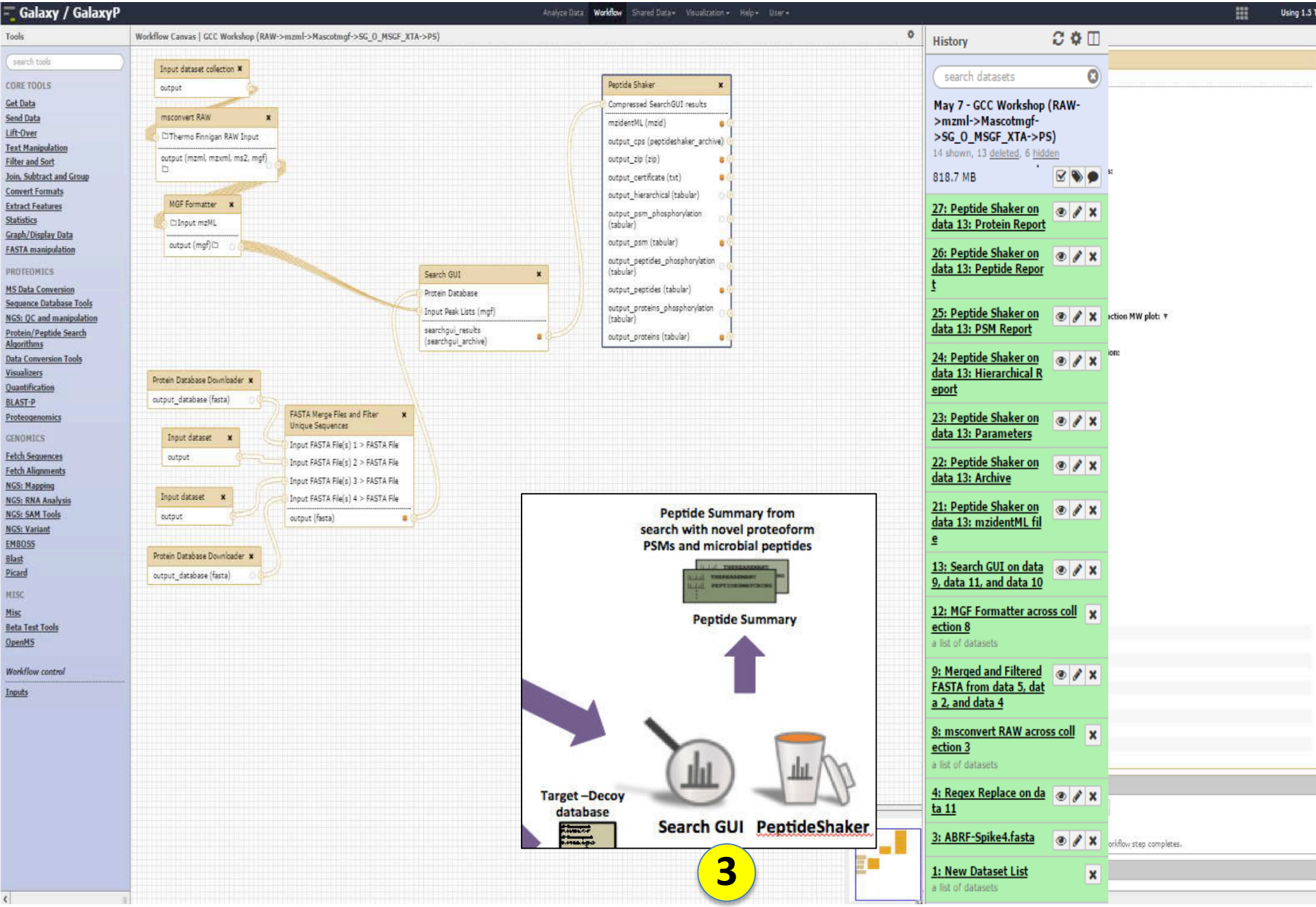
"Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations."

Sheynkman G et al BMC Genomics. doi: 10.1186/1471-2164-15-703.

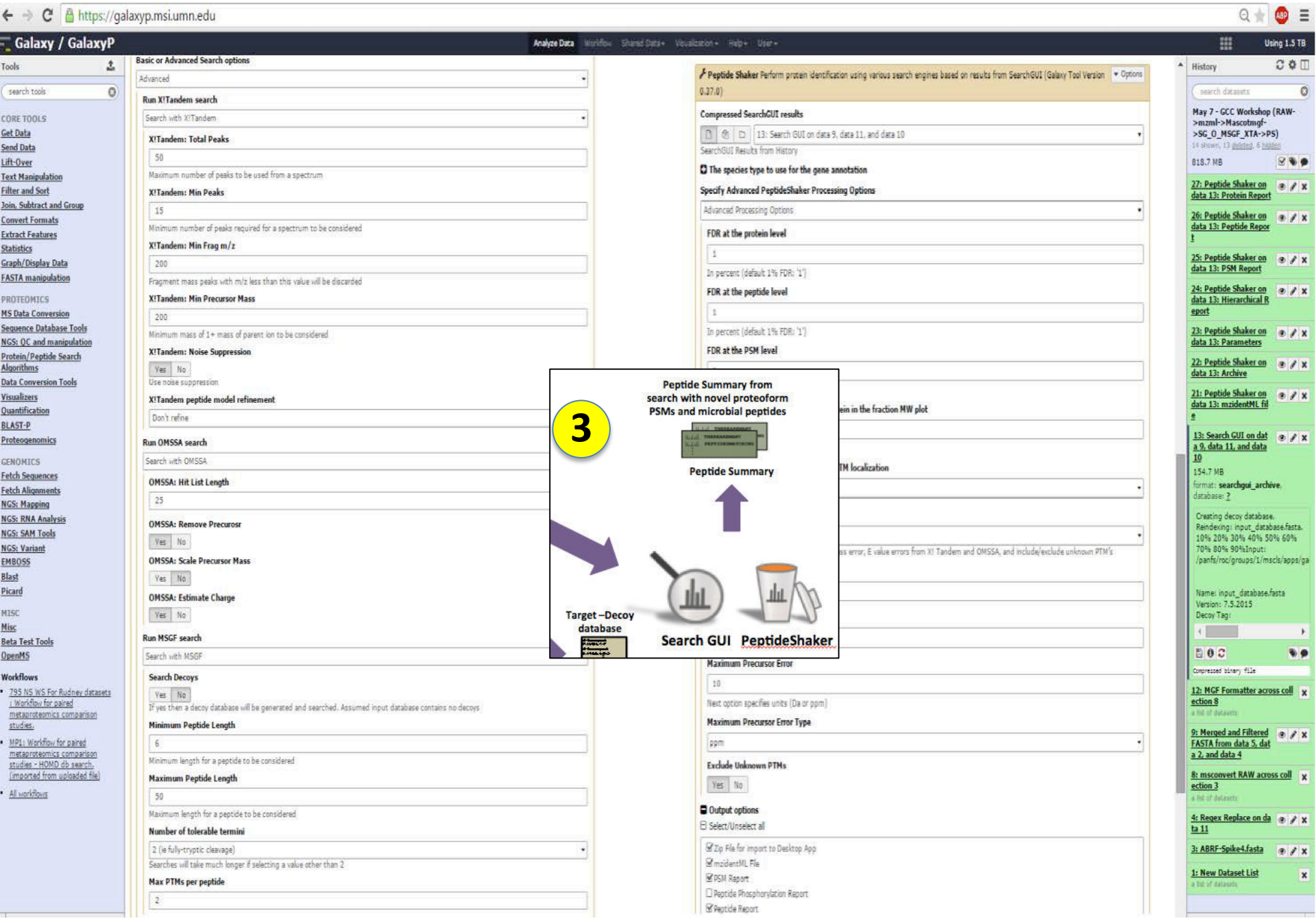

Gloria Sheynkman


James Johnson

DATABASE SEARCH



SEARCH GUI / PEPTIDESHAKER SEARCH



OUTPUTS FROM SEARCHGUI / PEPTIDESHAKER

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

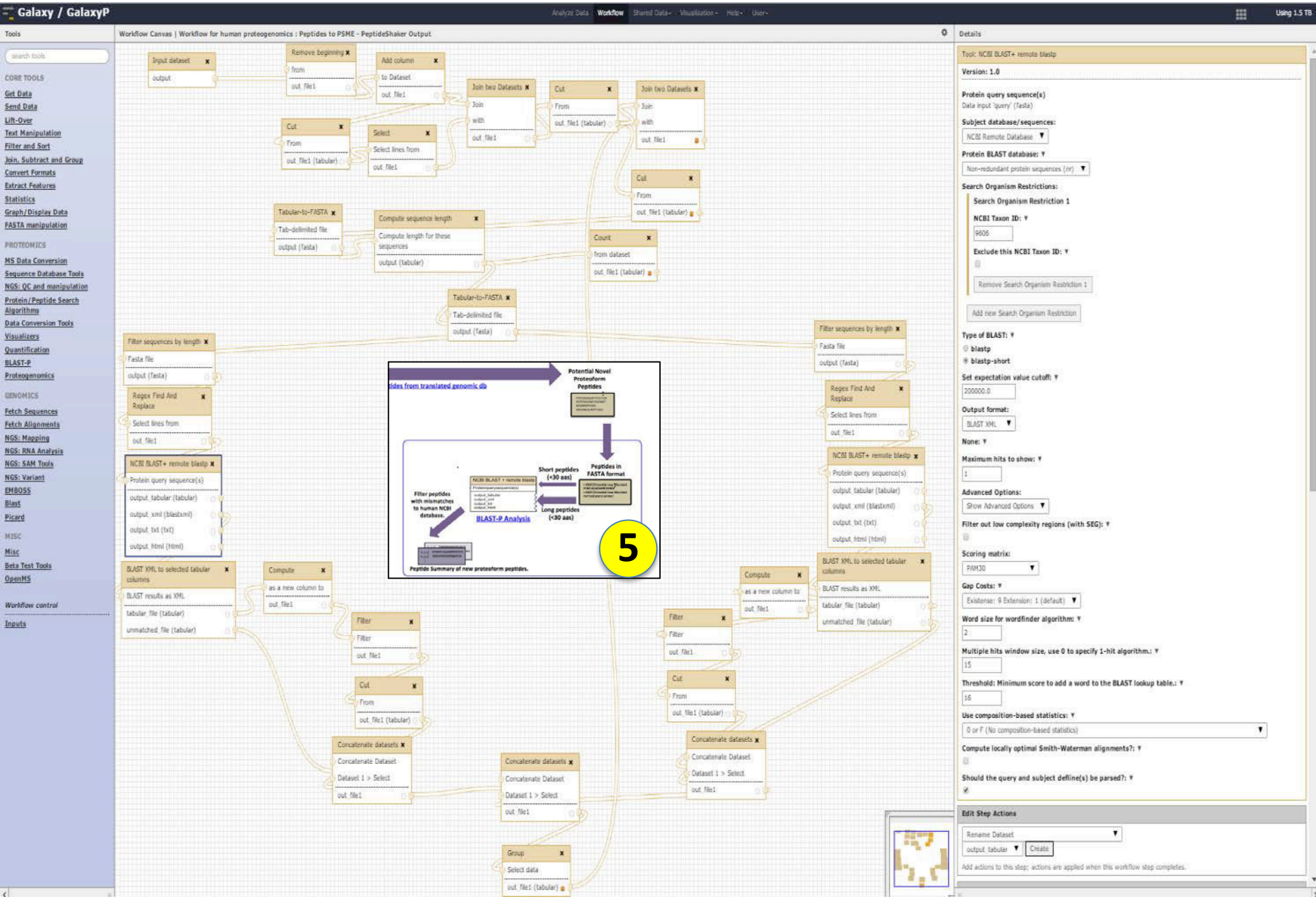
Mzid
-PSM Visualization
- SWATH Analysis
- Skyline
- Scaffold

Archive (Zipped file)
- CPS file to visualize data

Summary
- Valid peptides
- Valid proteins
- Valid spectra

GalaxyP

BLAST-P SEARCH



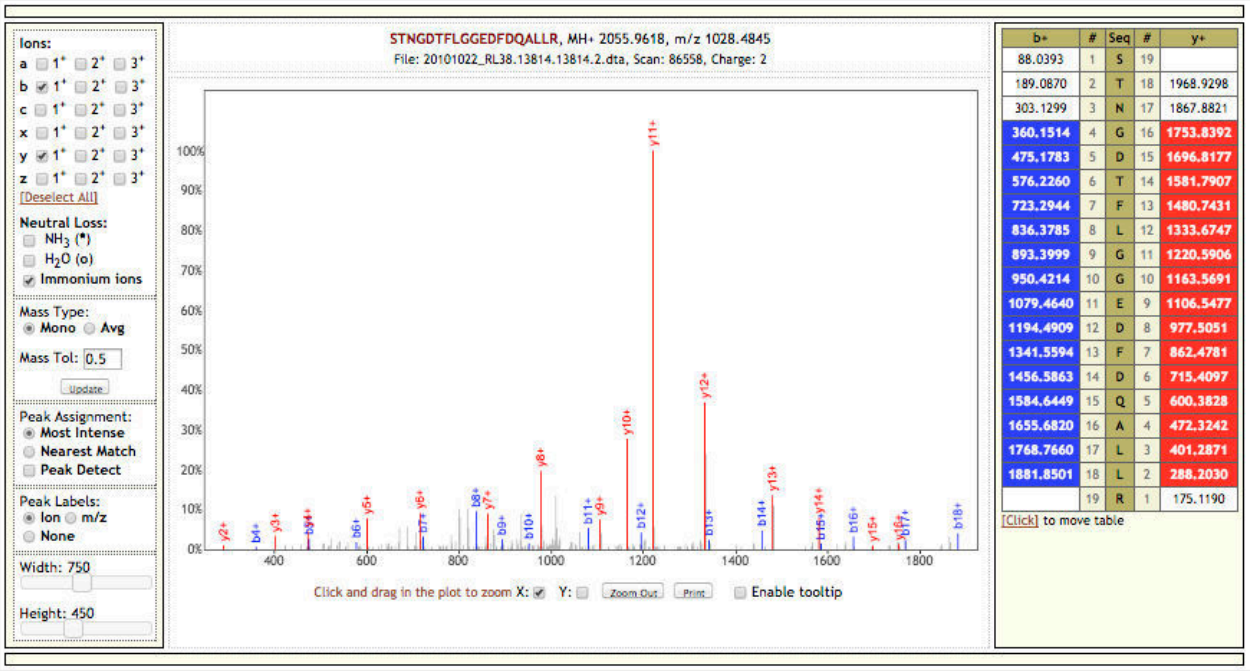
PSM EVALUATION

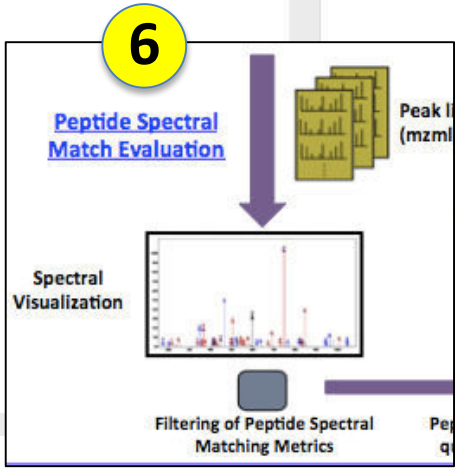
Search:

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

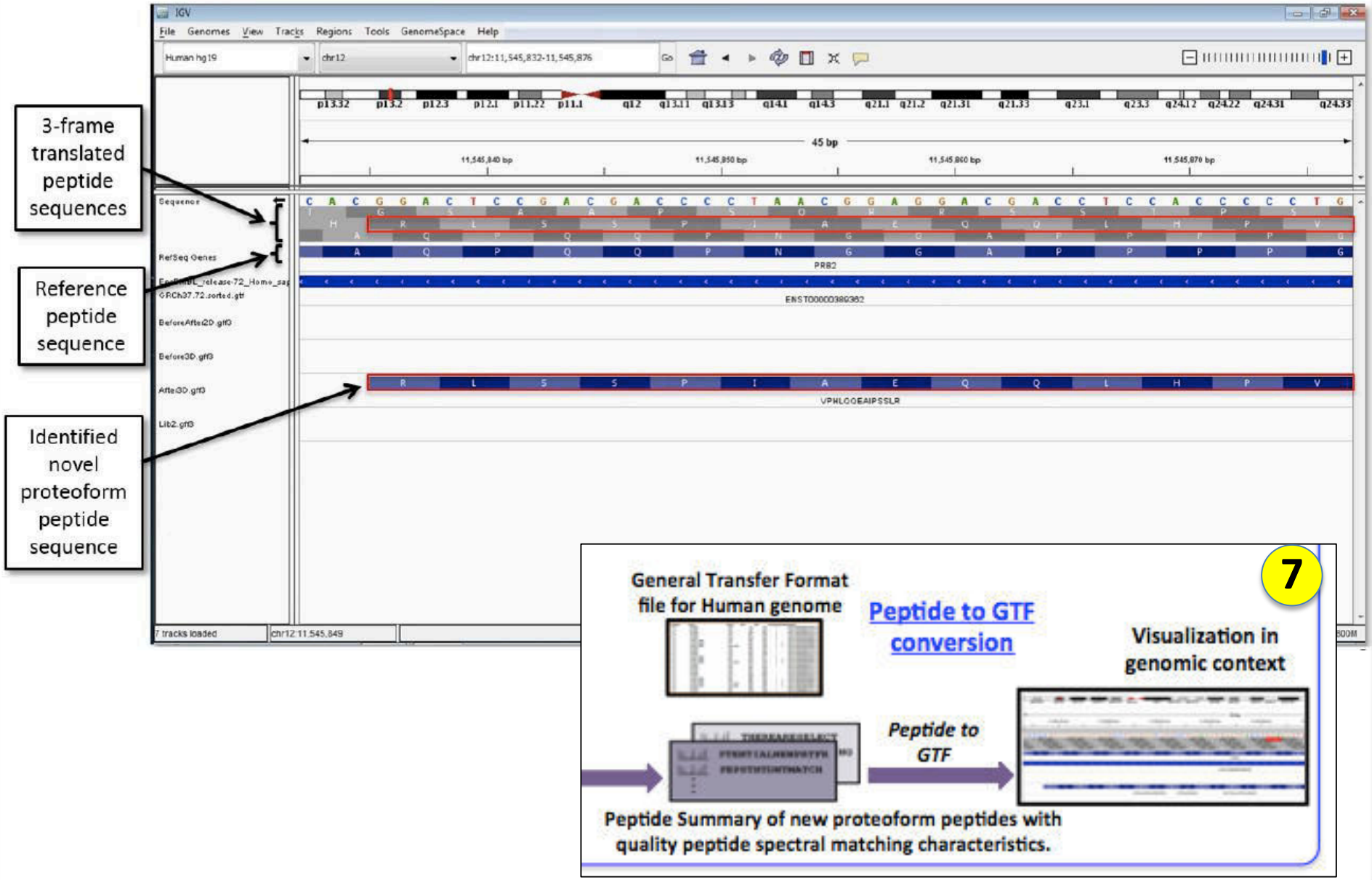




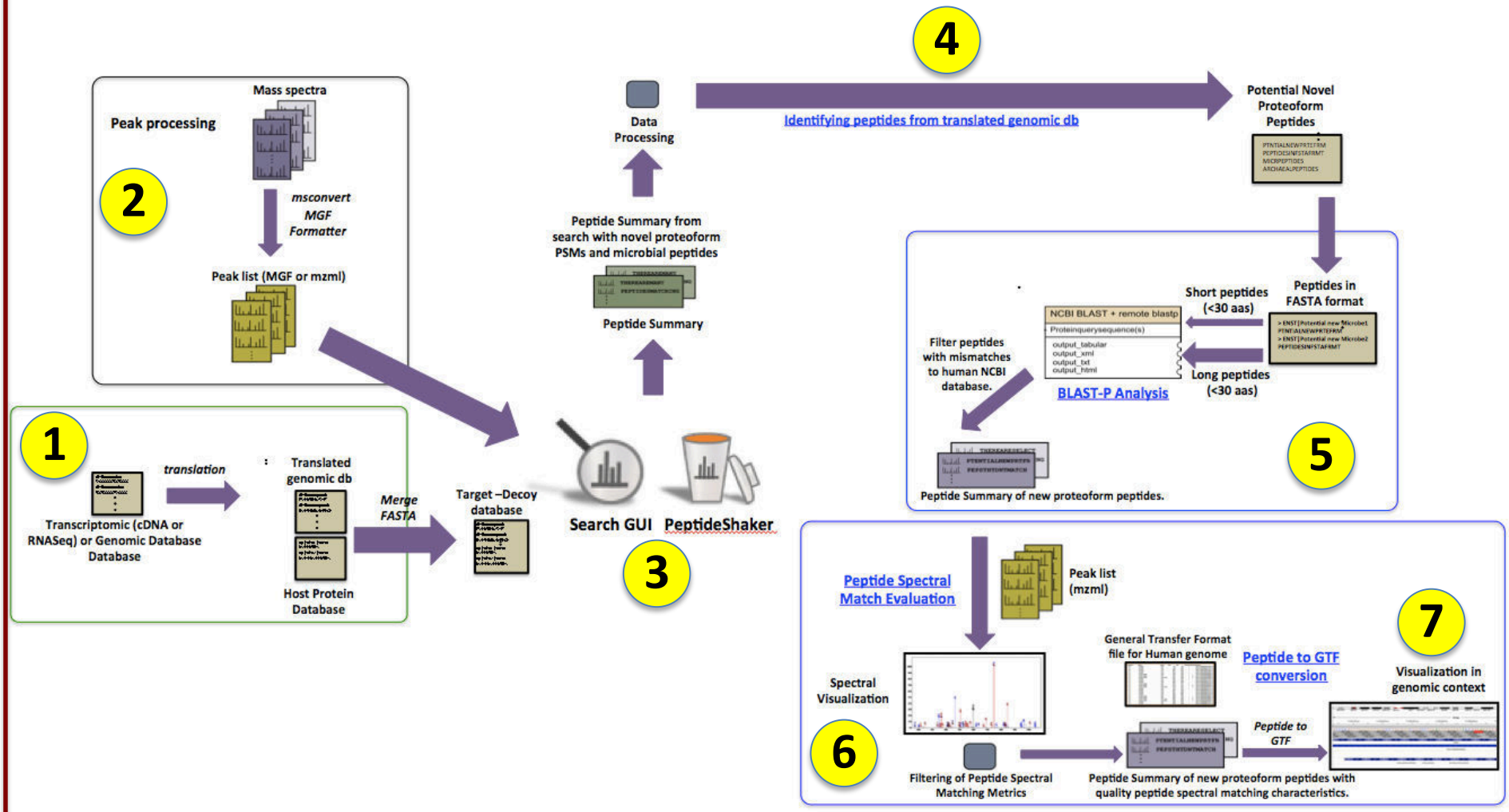
June 2nd 2015: Tuesday

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

GENOME VISUALIZATION USING IGV BROWSER



PROTEOGENOMICS WORKFLOWS

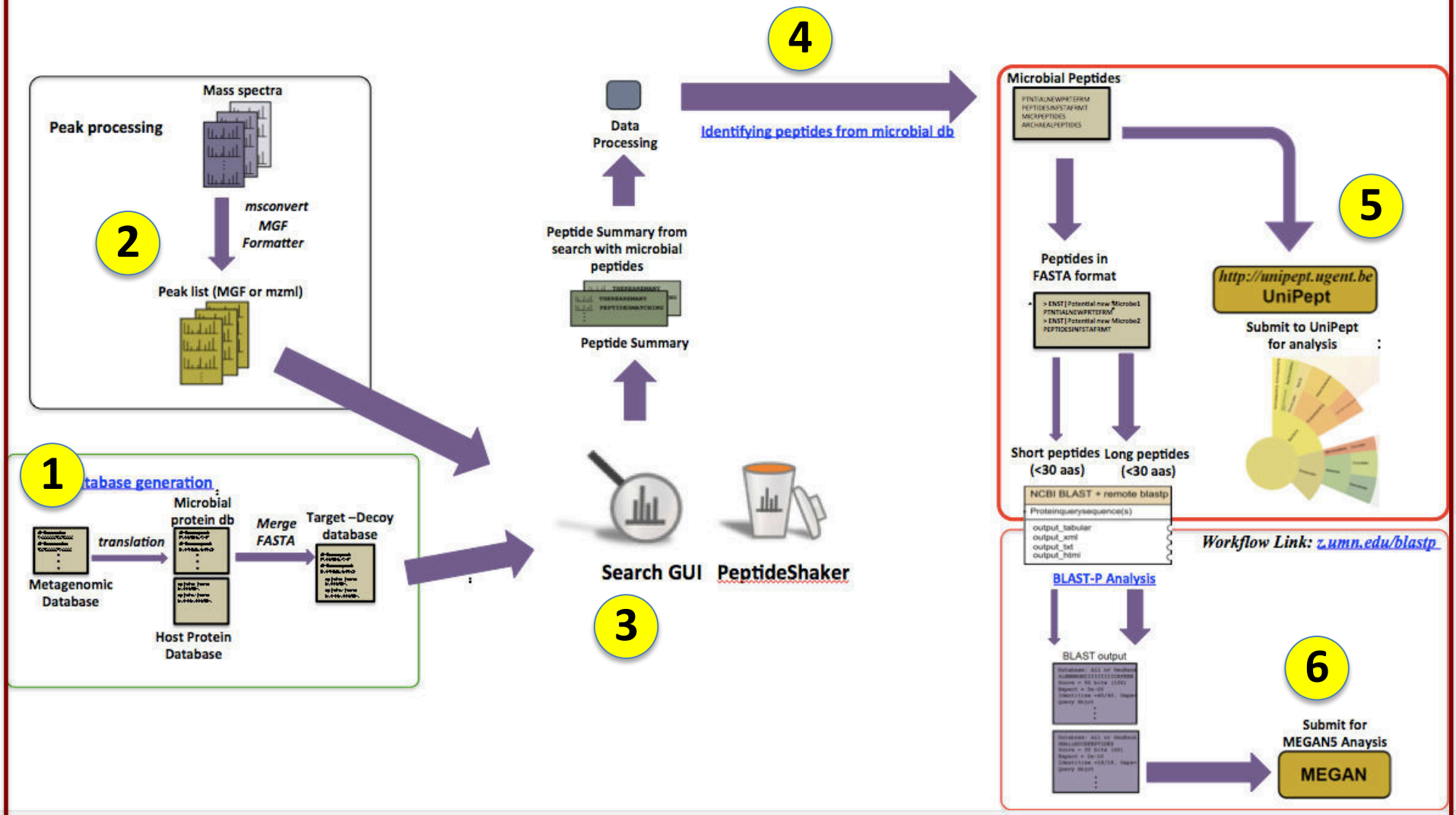


Jagtap et al *J Proteome Res.* 2014 13:5898-908

http://z.umn.edu/pgfirstlook

GalaxyP

METAPROTEOMICS WORKFLOWS



Jagtap et al *Proteomics* (2015)
http://dx.doi.org/10.1002/pmic.201500074

GalaxyP

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

  June 2nd 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 4th 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 2nd 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