

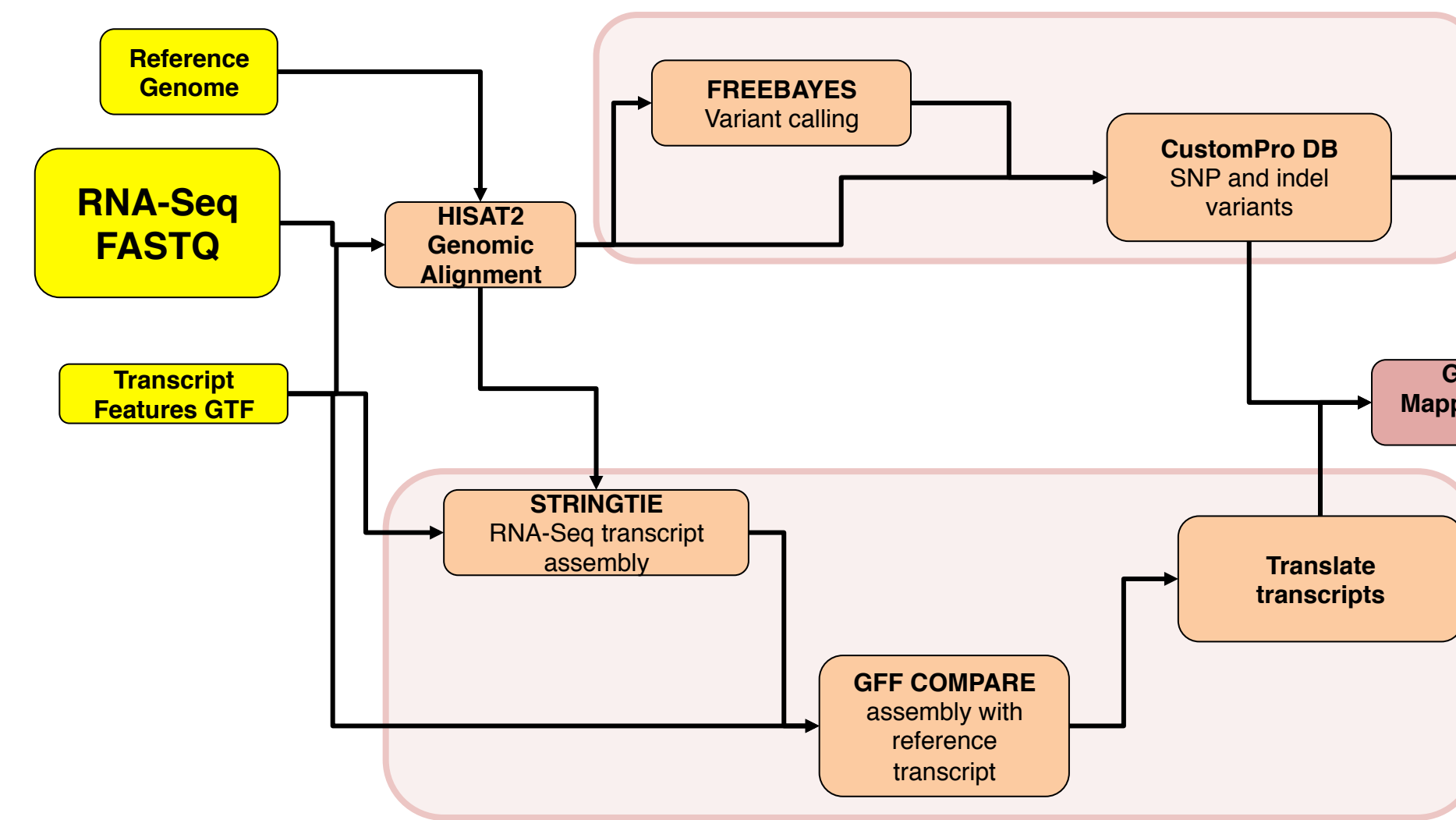
# From Raw Data to Results on Your Screen

A suite of accessible software for comprehensive proteogenomic informatics

James Johnson<sup>1</sup>; Tom McGowan<sup>1</sup>; Matthew Chambers<sup>2</sup>; Praveen Kumar<sup>3</sup>; Subina Mehta<sup>2</sup>; Pratik Jagtap<sup>2</sup>; Tim Griffin<sup>2</sup>

1. Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN; 2. Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN; 3. Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN

## RNAseq Analysis



## Proteomics Search

## Novel Protein Identification

## Visualization

### MVP Multi-omic Visualization Platform

The MVP Multi-omic Visualization Platform interface shows a Peptide Overview table with columns for Sequence, Spectra Count, and Protein Count. The table lists several peptides, including AVDPSSAEASGLR, AVDPSSAEASGLRAGQR, QDQLENPVLVSGAV, and QDQAGSGLSEASAAR. A Filtering Sequences dialog box is open, showing a list of peptides. The interface also includes a 'Load from Galaxy' button and a 'Render' dropdown menu.

The MVP Galaxy visualization plugin allows the user to browse and select proteomics search results.

Using a SQLite database allows results to be filtered and aggregated in a variety of ways, as well as providing quick retrieval of details for selected items.

Detailed Peptide Spectral Match (PSM) data can be displayed for selected peptides, and a lorikeet spectral view can be displayed for each PSM.

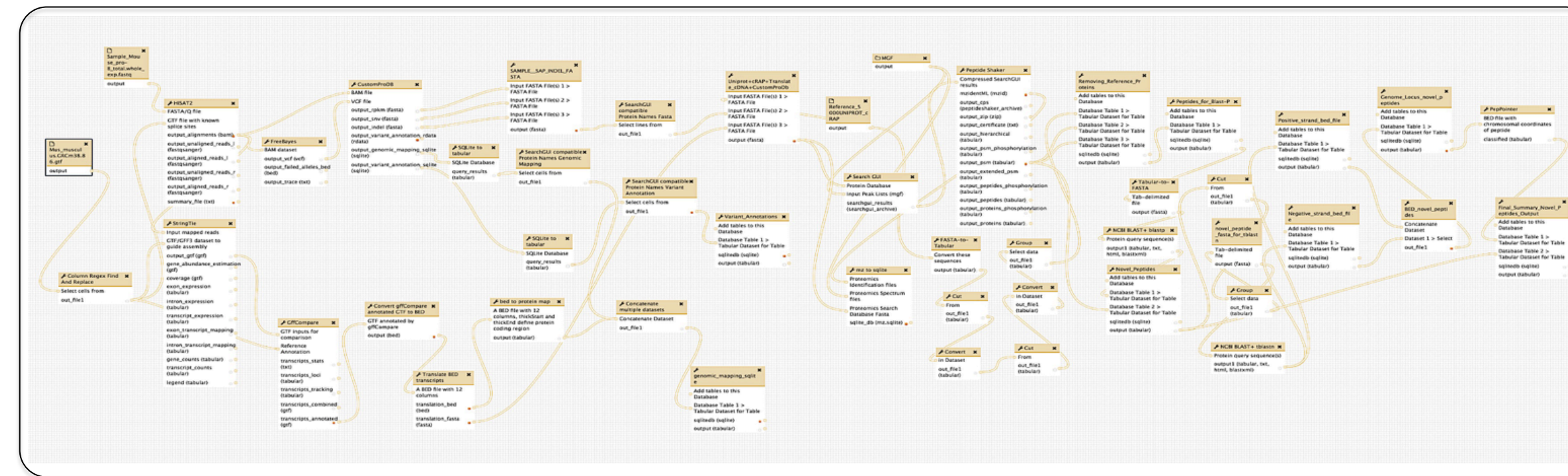
A selected peptide can also be displayed in a protein view. The protein view shows the peptide in context with all peptide identifications to the search protein.

A variant search protein can have an entry in the variant\_annotation table, then MVP will add the "search protein matches reference protein" line above the protein, and highlight the variant residues.

When genomic mapping is available for a search protein in the feature\_cds\_map table, an exon track is displayed above the protein. Clicking on an exon bar opens an IGV.js genome browser plugin to that location in the genome. This can provide an integrated view of the sequencing and proteomic data in their genomic context.

- Explore Peptide Identifications
- Verify Spectral Match in Lorikeet
- View Peptide in Protein Context
- View Genomic context in IGV.js

## As an integrated workflow on the Galaxy platform



RNA sequencing data allows for the generation of a sample-specific search database for improved identification of proteins.

The applications required for analysis are packaged as Galaxy tools installable on a Galaxy server from the Galaxy toolshed.

We employed several workflows to produce a sample specific search database.

SNPs and indels were identified using HISAT2 and FreeBayes; customProDB generates the variant proteins.

Variant transcripts were identified using HISAT2, StringTie, and gffCompare; the resulting GTF file is converted to BED format and translated to variant proteins by translate\_bed.

A common feature of the RNAseq search database generation workflows is that each produces a protein fasta to be added to the search database and also generates a genomic mapping output that is converted to a SQLite database that allows the PSM results to later be viewed on a genome browser.

The genome mapping schema can also be used for structural variants.

- Acknowledgements:
- The Galaxy Project: <http://galaxyproject.org/>
  - Galaxy-P: <https://github.com/galaxyproteomics>
  - Compomics Utilities: [Barnes et al. BMC Bioinformatics. 2011 Mar 8;12\(1\):70.](https://github.com/compomics)
  - Lorikeet: <http://uwpr.github.io/Lorikeet/>
  - IGV.js: <https://github.com/genome/genome>



NSF grant 1458524, and NIH grant 1U24CA199347 for funding support to the Galaxy-P team at the University of Minnesota.

Minnesota Supercomputing Institute, University of MN, for providing computing infrastructure

The Lorikeet Spectrum Viewer interface shows a mass spectrum plot for the peptide ESRRALVEPTSESPRPALAR. The x-axis represents the mass-to-charge ratio (m/z) and the y-axis represents relative intensity. The plot shows a base peak at m/z 411. The interface also includes a table of peaks with columns for m/z, relative intensity, and other parameters. The peptide sequence is displayed at the top: ESRRALVEPTSESPRPALAR.

The IGV.js Genome Browser interface shows a genomic track for chromosome 11. The track displays various genomic features, including a reference transcript, peptide genomic coordinates, and RNA-seq reads. The interface includes a search bar, a track control panel, and a detailed view of the selected region. The peptide sequence ESRRALVEPTSESPRPALAR is highlighted in the reference transcript track.

The Galaxy history panel shows a list of workflow steps and their outputs. The steps include: ASRF\_History\_Input, Final\_Summary\_Novel\_Peptides\_Output, 59\_PepPointer\_on\_data\_58, 58\_Peptide\_genomic\_coordinate, 57\_Novel\_Peptides, 56\_Extracting\_Novel\_Peptides, 55\_BlastP\_output, 53\_Peptides\_for\_BlastP\_analysis, 47\_mz\_to\_sqlite\_on\_data\_29\_data\_36\_and\_others, and 68.5\_MG. The outputs are listed with their respective file names and sizes.