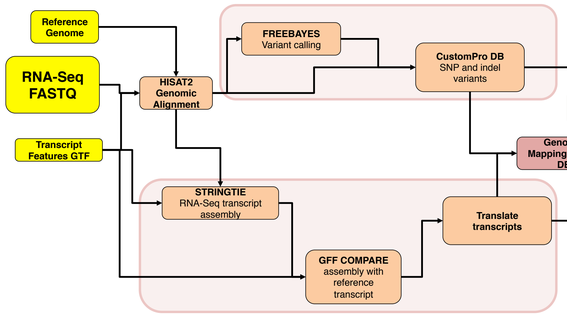
From Raw Data to Results on Your Screen

A suite of accessible software for comprehensive proteogenomic informatics

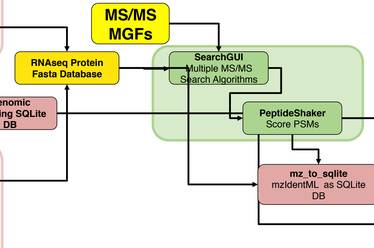
James Johnson[1]; Tom McGowan[1]; Matthew Chambers[2]; Praveen Kumar[3]; Subina Mehta[2]; Pratik Jagtap[2]; Tim Griffin[2]

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

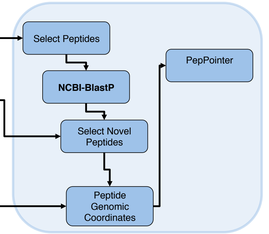
RNAseq Analysis



Proteomics Search

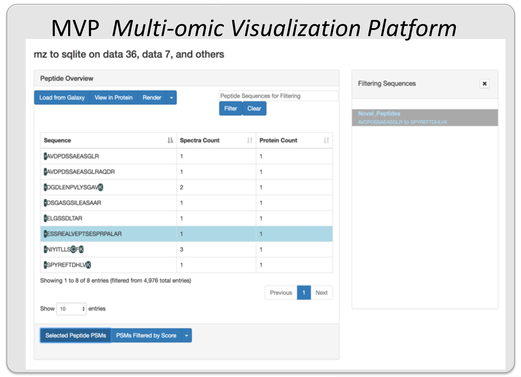


Novel Protein Identification



Visualization

MVP *Multi-omic Visualization Platform*



The MVP 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**

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.

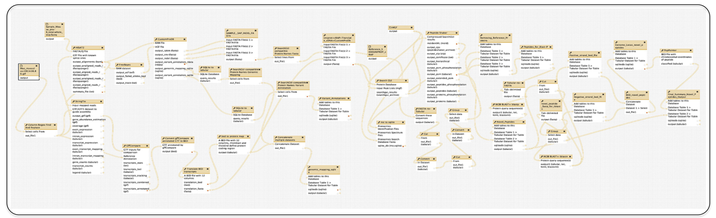
We employed several workflows to produce a sample specific search database.

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

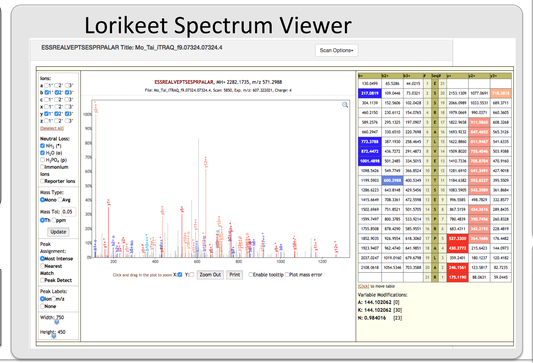
Variant transcripts were identified using HISAT2, StringTie, and gffCompare; the resulting GTF file is converted to BED format and translated to variant proteins by translate_bed.

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 the PSM results to later be viewed on a genome browser.
The genome mapping schema can also be used for structural variants.

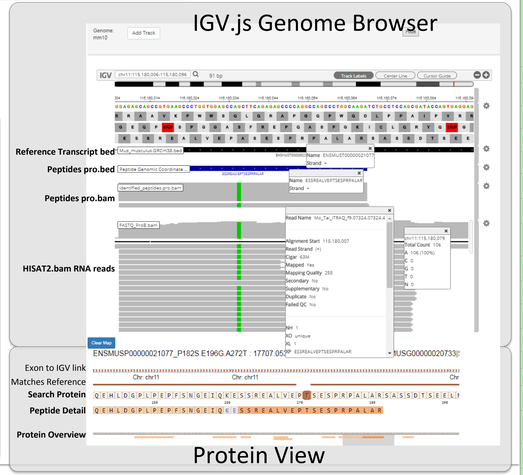
As an integrated workflow on the Galaxy platform



Lorikeet Spectrum Viewer



IGV.js Genome Browser



Protein View

Acknowledgements:

- The Galaxy Project: http://galaxyproject.org/
- Galaxy-P: https://github.com/galaxyproteomics
- Compomics Utilities: Barsnes et al: BMC Bioinformatics. 2011 Mar 8;12(1):70.
- Lorikeet: http://uwpr.github.io/Lorikeet/
- IGV.js: https://github.com/igvteam/igv.js/wiki

NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

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