From Start to Finish: a complete proteogenomic informatics environment implemented in the Galaxy platform

Getiria Onsongo[1*], Pratik D.Jagtap[2], James E. Johnson[1], Thomas McGowan[1], Mohammad Heydarian[3], Karen Reddy[3],Timothy J. Griffin[4]

[1]Minnesota Supercomputing Institute, UMN, Minneapolis, MN; [2]Center for Mass Spectrometry and Proteomics, UMN, St. Paul, MN; [3]John Hopkins University, Baltimore, MD;

[3]Department of Biochemistry, Molecular Biology and Biophysics, UMN, Minneapolis, MN

University of Minnesota
Driven to Discover℠

* presenter : **onson001@umn.edu**

Overview

- We have developed a complete proteogenomic informatics environment that seamlessly combines genomics, transcriptomics and proteomics data.
- We developed new Visualization tools for evaluating and interpreting results
- Implementation was done on the Galaxy-P framework, an extension of Galaxy. Galaxy is an open, web-based platform for data intensive biomedical research.

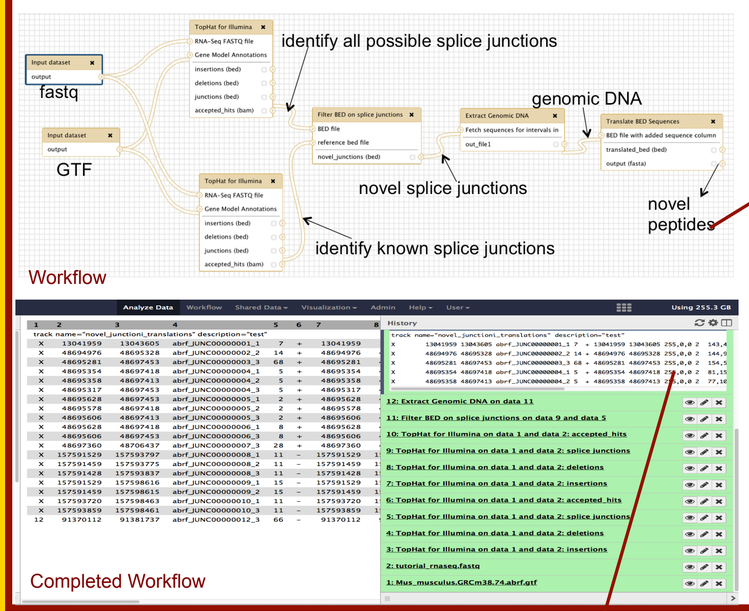
Methods

- RNA-Seq data used to generate sample specific protein database of potential novel proteoforms
- SearchGUI/PeptideShaker match MS/MS data to database to verify presence of variant peptides
- Multi-Omics Visualization Platform (MVP) used to validate novelty and quality of identified peptides
- Peptides of interest mapped to the genome and viewed using Integrated Genomics Viewer (IGV) to assess quality of associated RNA-Seq data

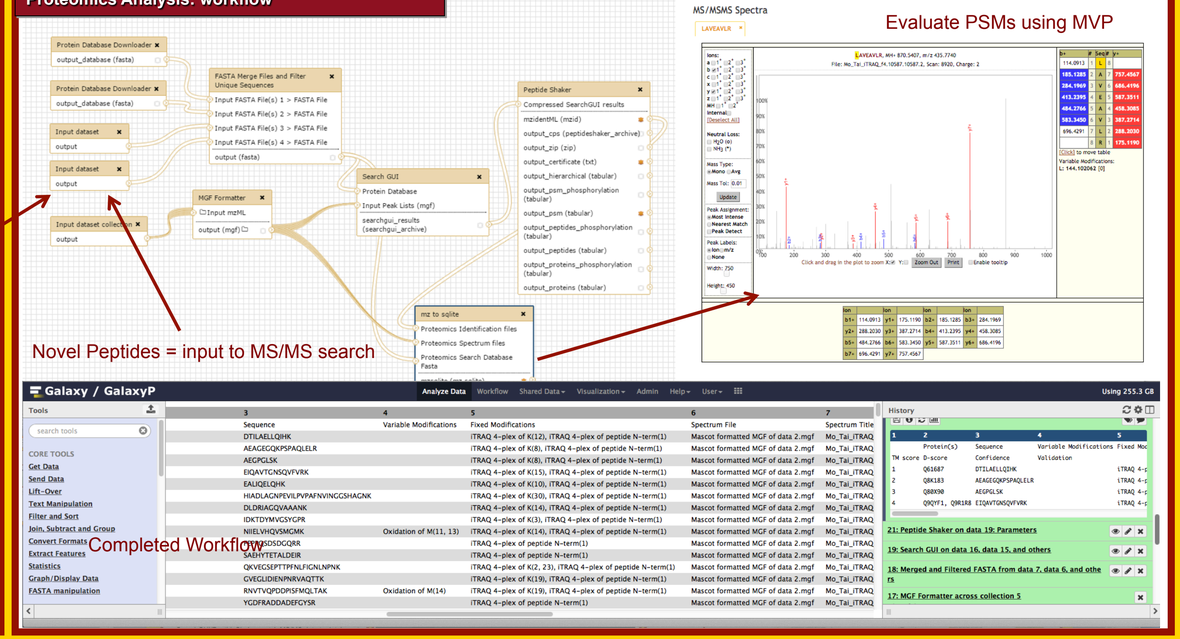
Results

- Complete, flexible and accessible proteogenomic informatics environment geared towards quality assessment and visualization of results.

RNA-Seq Analysis: workflow



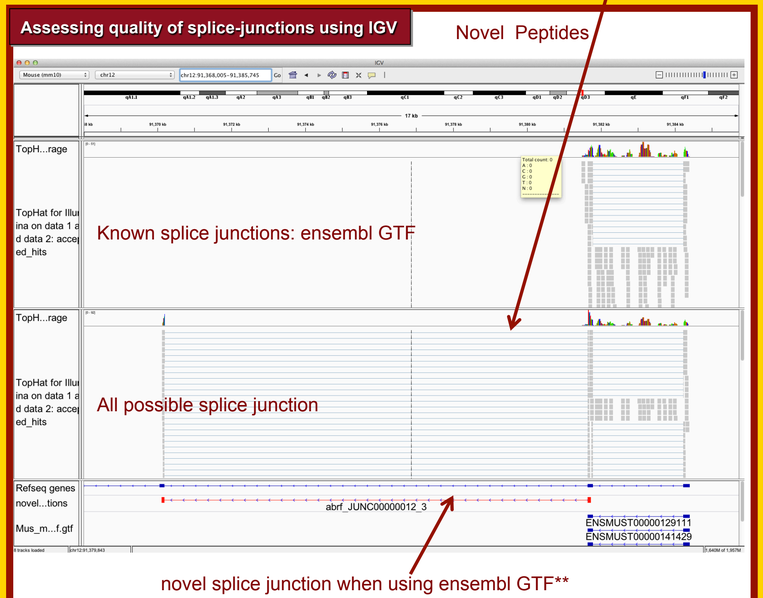
Proteomics Analysis: workflow



Introduction

- Proteogenomics enables new insights into complex biological systems, most commonly integrating RNA-Seq and proteomic data.
- Effective proteogenomics requires seamless integration of disparate analysis tools for genomics and proteomics data.
- We present a first-of-its-kind, complete proteogenomics informatics environment implemented in the Galaxy-P platform.
- This environment enables seamless navigation between genomics, transcriptomics & proteomics data and analysis tools, building customizable, easily shared workflows usable by bench researchers.
- It also provides filtering and visualization tools for evaluating and interpreting results.
- Results from a proteogenomic analysis of B-cell development, highlighting the many advantages of this platform

Assessing quality of splice-junctions using IGV



** NOTE: even though this junction is novel according to ensembl annotation, it is present in RefSeq

Results

- We leveraged existing tools in Galaxy for genomic/transcriptomic data analysis, to create an informatics environment for complete proteogenomic analysis.
- Assembled transcriptome data seamlessly integrates with tools to enable construction of protein databases containing possible variant sequences such as single amino acid polymorphisms (SAPs) and splice isoforms .
- MS/MS spectra are matched against the variant and canonical sequences using the Galaxy-deployed SearchGUI/PeptideShaker tools.
- Our recently developed MVP tool acts as a Galaxy plugin, and provides interactive viewing and filtering of annotated MS/MS data via our Peptide Sequence Match Evaluator (PSME).
- New extensions to MVP automatically link to IGV, enabling visualization of variant peptide matches against reference genomic and proteomic information.
- We have applied this informatics environment to analysis of RNA-Seq and proteomic data generated from a study of early B Cell development resulting in discovery of numerous novel splice variants, SAPs, as well as a number of novel proteins encoded within long non-coding RNAs.

Conclusion

- Our framework provides a complete, flexible and accessible proteogenomic informatics environment geared towards quality assessment and visualization of results.