Plugging Proteomics Peptide-Spectral-Match Visualization into Galaxy

James Johnson[1]; Tom McGowan[1]; Ira Cooke[2]; John Chilton[3]; Pratik Jagtap[4]; Tim Griffin[1]

1. University of Minnesota, Minneapolis, MN; 2. La Trobe University, Melbourne, Australia; 3. Penn State University, State College, PA; 4. Center for Mass Spectrometry and Proteomics, UMN, St. Paul, MN

Introduction

Galaxy is a scalable computing environment suitable for managing large datasets among multiple users. Many proteomics applications are available as Galaxy tools enabling complex multi-step analysis from a web browser (https://usegalaxyp.org/). Visualization of proteomics data is often required by users to verify analytic results. The Galaxy framework allows plugins for visualizing a dataset based on its datatype. For interactive web browser viewing of the data, datatype associated dataproviders respond quickly to incremental data requests. We have defined a proteomics schema as a Galaxy SQLite datatype, provided a Galaxy tool to convert peptide identification and spectrum files into a SQLite database, and developed a Galaxy visualization plugin to query peptide spectral matches and visualize them in a spectral viewer.

Overview

Why Galaxy

Scalable

Galaxy is a scalable computing environment suitable for managing large datasets among multiple users. Users access Galaxy via a web browser or web applications. Galaxy can be configured to distribute jobs to heterogeneous compute nodes, including windows.

Collaborative

Galaxy items: histories, datasets, workflows, visualizations can be shared among users so that results can verified by others.

Reproducible

Galaxy records all inputs, outputs, options, and versions for all applications run, enabling others to reproduce results exactly or experiment with alternative settings.

Extensible

While Galaxy was originally deployed for Genomics analysis, it can be extended to new application areas by adding datatypes and tools. This allows for combining analysis from different research areas on a single platform, e.g. proteogenomics analysis combining sequencing and mass spec sample data.

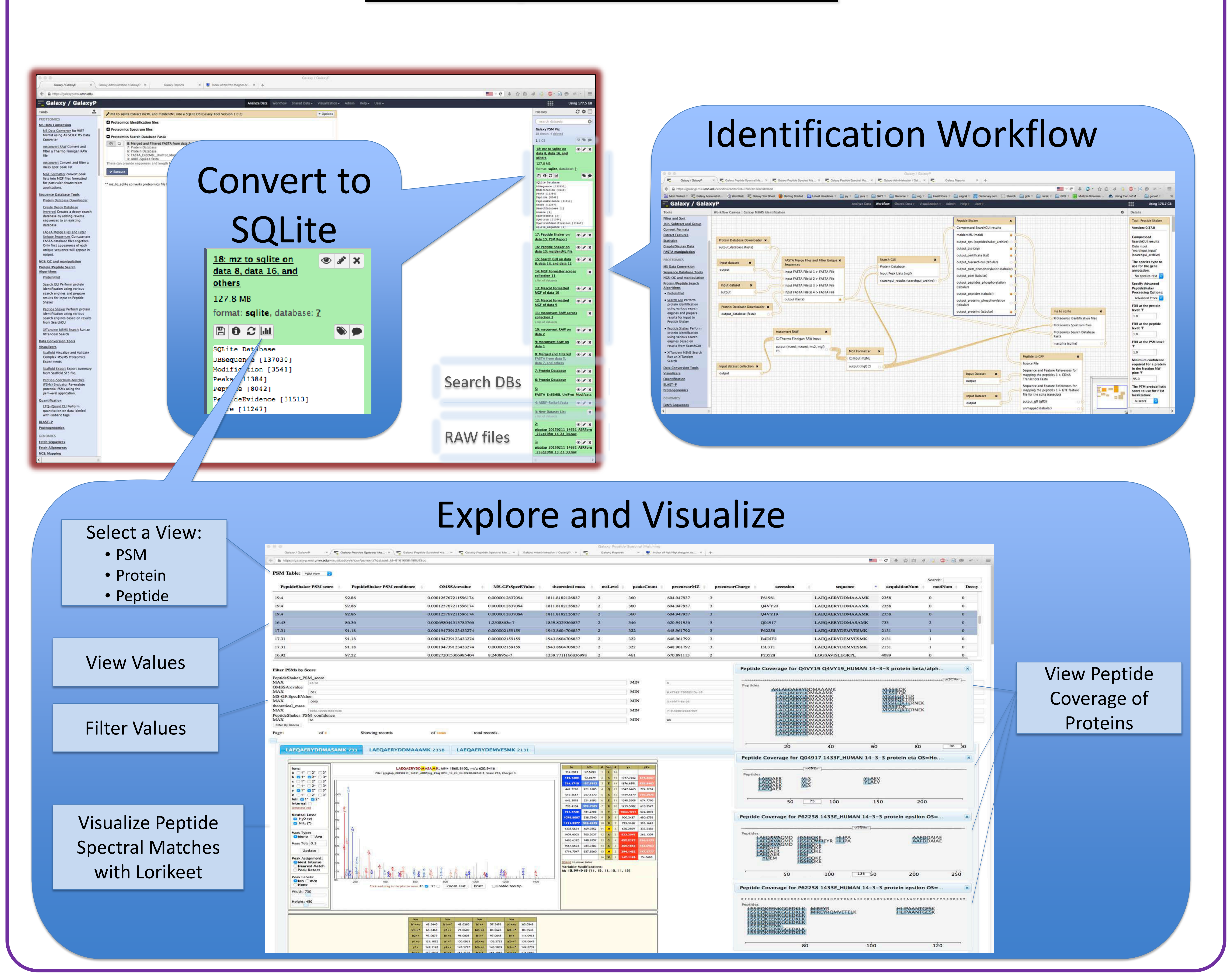
Adding a Galaxy Proteomics Visualization

Galaxy visualizations are web browser applications that display views of Galaxy datasets. For interactive data exploration, they need to be able to selectively and quickly retrieve data incrementally from Galaxy via web requests to a Galaxy dataset dataprovider. Common proteomic file formats do not afford that level of interactive access. Thus, we store the proteomics data in a SQLite database so that any information can be retrieved by a SQL query. The **psmeviz** visualization web application manages the paging and filtering of data in its SQL query construction, which enables it to scale to millions of data items.

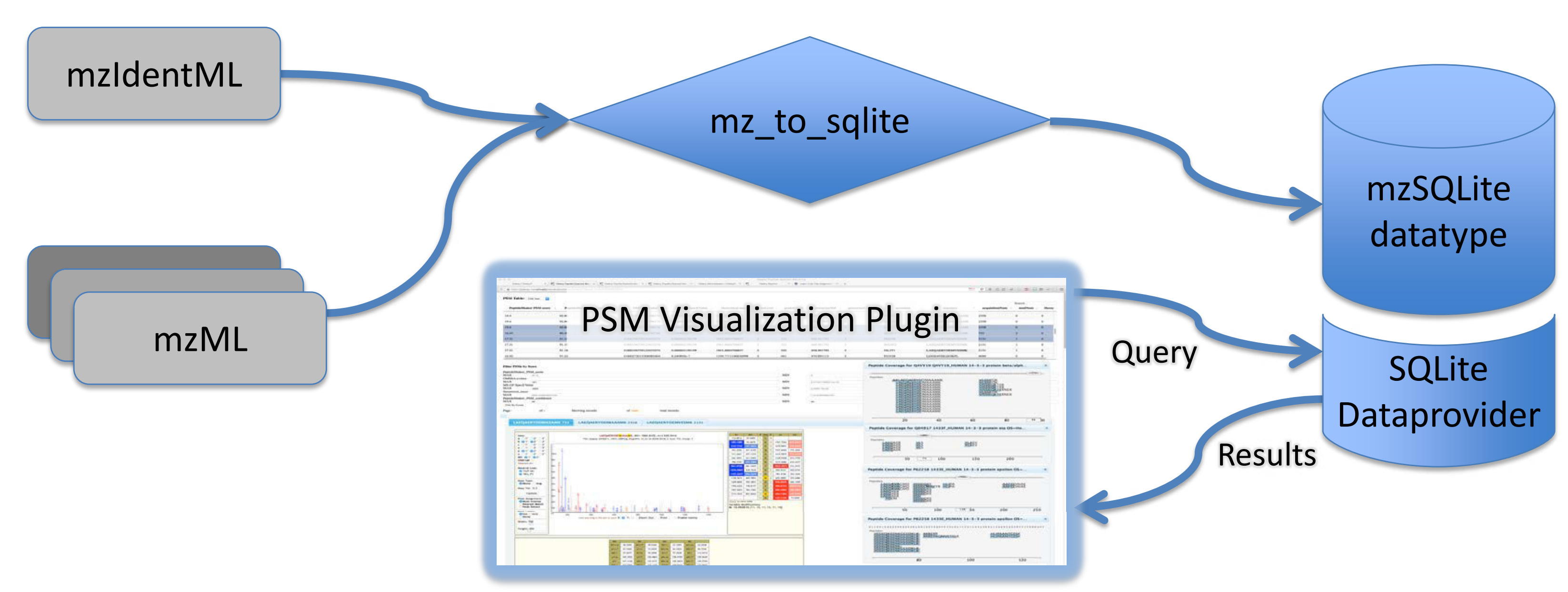
Galaxy terms:

- **Dataset** – The input and output files for Tools
- **Datatype** – A file format for a Dataset, it may contain metadata about the dataset.
- **Dataprovider** – Galaxy server python code that selects and returns data from a dataset in response to a web request.
- **Tool** – A wrapper for an application which specifies the inputs, outputs, options to run the application.
- **History** – A collection of datasets for an investigation, a tool run in a history adds its output datasets to the history
- **Workflow** – a network of tools to run.
- **Visualization** – a client-side interactive view of a dataset via the dataprovider

Galaxy Proteomics



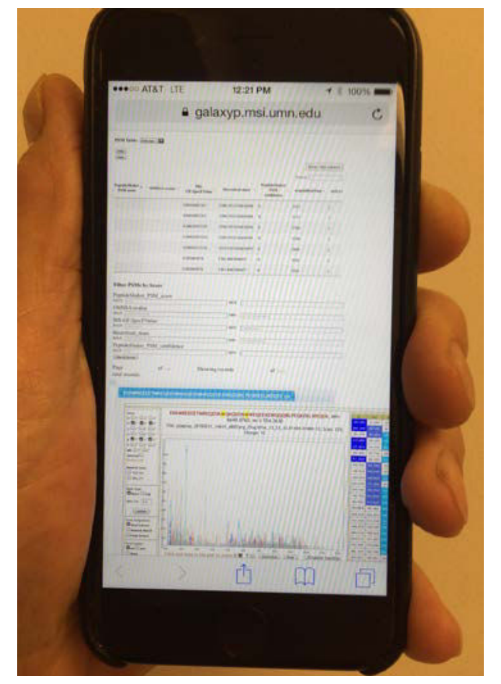
The Proteomics Visualization Plugin



1. **Add Galaxy Datatypes:**
   - Define the **SQLite** Galaxy datatype to Galaxy, along with associated dataproviders that respond with a result table when presented with a SQL query.
     - https://github.com/galaxyproject/galaxy/blob/dev/lib/galaxy/datatypes/binary.py
     - https://github.com/galaxyproject/galaxy/blob/dev/lib/galaxy/datatypes/dataproviders/dataset.py
   - Design a SQLite database schema that represents the data in mzIdentML and mzML file formats
   - Define the **mzSQLite** Galaxy
2. **Provide tools forDataset conversion to mzSQLite:**
   - Galaxy tool, **mz_to_sqlite**, parses mzIdentML, mzML, MGF, and Search DB fasta Galaxy datasets, and populates the mzSQLite schema tables.
     - https://github.com/galaxyproteomics/tools-galaxyp/tree/master/tools/mz_to_sqlite
3. **Design Proteomics Visualization Plugin:**
   - Multiple Views: Protein, Peptide, Peptide Spectral Match
   - Generate SQL queries to retrieve data
   - Provide filters for narrowing the query results
   - Manage paging and filtering in the construction of SQL queries
   - Display peptide coverage for selected proteins
   - Display a Lorikeet spectral viewer for selected PSMs
     - https://github.com/galaxyproteomics/tools-galaxyp/tree/master/visualizations/psmeviz

Conclusions

Visualizing proteomics via Galaxy allows the user to access large quantities of data from any web browser, large or small.



Future Directions:

- **Enable saving and sharing the visualization on Galaxy**
- **Save the visualization to an external site**
- **Enable filtering by values retrieved from another history dataset**
- **Add a generic SQLite query view that allows free form SQL queries**
- **Add additional conversion tools, e.g. add a mzSQLite output to psm_eval tool**

Acknowledgements:

- The Galaxy Project: http://galaxyproject.org/
- John Chilton for Galaxy-P development that adapted Galaxy for proteomics
- Genetech pviz library: http://research-pub.gene.com/pviz/app/proteomics-3d/
- Compomics Utilities: Barsnes et al: BMC Bioinformatics. 2011 Mar 8;12(1):70.
- Lorikeet: http://uwpr.github.io/Lorikeet/

- Galaxy-P is supported through the National Science Foundation; DBI-ABI award 1147079 "Galaxy-P: A new community-based informatics paradigm for MS-based proteomics"
- Minnesota Supercomputing Institute, University of MN, for providing computing infrastructure