SearchGUI and PeptideShaker deployed in the Galaxy framework: A powerful informatics platform for protein identification and beyond

Authors:

Cooke I. R.[1], Grüning B. A.[2], Barsnes H.[3], Vaudel, M.[3], Martens, L.[4], Johnson, J.[5], Guerrero, C.[5], Onsongo, G.[5], Chilton, J.[6], Jagtap, P.[5,7] and Griffin, T. J.[5,7]

Affiliations:

1. La Trobe University, Melbourne Australia
2. University of Freiburg, Germany
3. Proteomics Unit, University of Bergen, Norway
4. Computational Omics and Systems Biology Group, VIB, Gent University, Belgium
5. University of Minnesota, Minneapolis, USA
6. Penn State University, USA
7. Center for Mass Spectrometry and Proteomics, UMN, St.Paul, MN

Introduction

The SearchGUI and PeptideShaker (Vaudel et al, Nature Biotech. doi:10.1038/nbt.3109) pipeline is a suite of software that comprehensively covers all tasks in protein identification, from MS/MS database search to protein inference, and aids biological interpretation via Gene Ontology and UniProt. It stands to benefit greatly from integration into a workflow framework such as Galaxy given that protein identification is computationally intensive, and often forms a central component in complex multi-step tasks such as quantitative analysis, multi-omics applications (proteogenomics, metaproteomics) and DIA proteomics. Deploying SearchGUI and PeptideShaker in Galaxy has enabled its use in such complex workflows. Additionally, these workflows can now be executed within a scalable computing environment suitable for large datasets and multiple users.

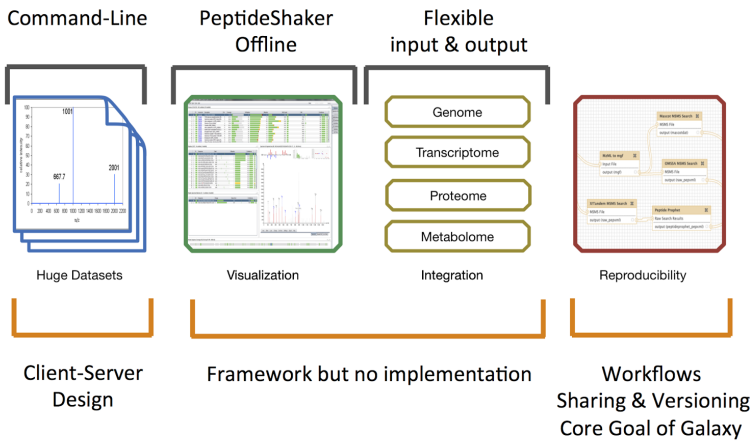

Figure 1: Addressing the needs of multi-omics analysis by combining PeptideShaker & SearchGUI with Galaxy. Central boxes represent needs. Annotations at bottom are contributions from Galaxy and annotations at top come from PeptideShaker/SearchGUI.

Methods

Integration of PeptideShaker and SearchGUI into Galaxy required a coordinated effort between PeptideShaker and SearchGUI developers, Galaxy tool developers, and end users (see http://bit.ly/galaxypforum). The PeptideShaker and SearchGUI developers adapted to run in a cluster environment, and added the ability to export a self contained package of results, spectra and protein sequences, while the Galaxy tool developers created a Galaxy user interface and corresponding Tool Shed packages to automate installation within Galaxy. Importantly, the entire software stack was subjected to automated testing, as well as real-world testing by end users at the University of Minnesota and La Trobe University. Feedback from these tests was used to identify bugs and missing features required to support specific uses.

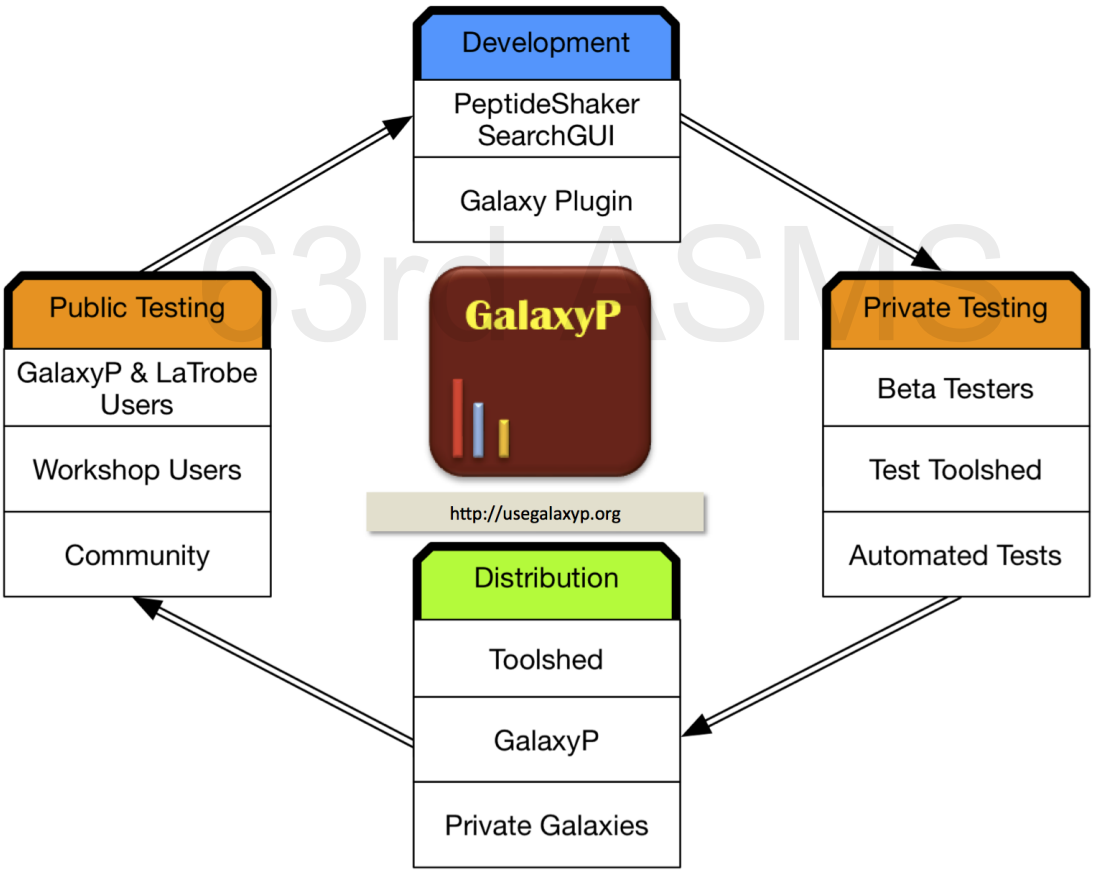

Figure 2: Development workflow for integrating SearchGUI and PeptideShaker with Galaxy

Results

We have successfully integrated SearchGUI and PeptideShaker into Galaxy, and demonstrate the benefits of this platform with diverse use cases from our own research and training activities: from conventional quantitative proteomics studies, to applications integrating multi-omic data types (proteogenomics, metaproteomics), to emerging applications requiring spectral library generation (e.g. DIA SWATH). All of these applications benefit from the increased confidence of peptide and protein identification enabled by use of multiple database search engines via SearchGUI and PeptideShaker. Our Galaxy wrapper takes advantage of the diverse export capabilities of PeptideShaker to enable direct integration with a wide variety of downstream tasks. For example, standard mzIdentML export permits integration with downstream tools for spectral library creation as a precursor for DIA analysis using programs such as OpenSWATH. Output can also be seamlessly passed as input to existing Galaxy tools that further validate peptide sequence matches to novel sequence variants (proteogenomics) or microbiome-expressed sequences (metaproteomics), as well as tools for visualizing results (e.g. mapping peptides to genomic coordinates for proteogenomics). Alternatively, results can be exported for offline examination, or further processing as tabular outputs with R or IPython notebook, both of which can be run directly from within Galaxy. PeptideShaker also offers the ability to reprocess data deposited in public repositories using its “Reshake’ feature, which will help researchers getting additional information from archived datasets. The scalability of the SearchGUI and PeptideShaker Galaxy platform offers an enterprise-level solution for proteomic data analysis -- useful for institutions needing a platform accessible by multiple researchers. Such a platform also enables group training and use in a workshop setting. Finally, the Galaxy-based tools and workflows are easily disseminated via the Galaxy Tool Shed, enabling their use in local instances of the framework at any institution.

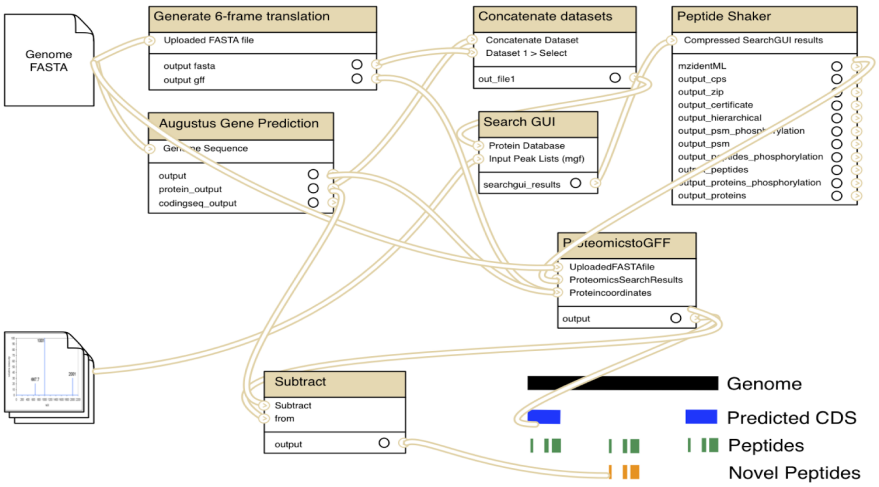

Figure 3: Example workflow demonstrating the use of SearchGUI and PeptideShaker along with other Galaxy tools to find novel peptides.