CUSTOMIZED, REPRODUCIBLE AND ACCESSIBLE TOOLS FOR TANDEM PROTEOGENOMIC-METAPROTEOMIC ANALYSIS

Pratik Jagtap[1], James Johnson[1], John Chilton[1,3], Trevor Wennblom[1], Bart Gottschalk[1], Sricharan Bandhakavi[1,2], Joel Rudney[1], Tim Griffin[1]

[1]University of Minnesota, Minneapolis, MN; [2]Bio-Rad Laboratories, Hercules, CA; [3]Pennsylvania State University, University Park, PA

INTRODUCTION

Proteogenomics and metaproteomics are rapidly emerging fields, based on integrated analysis of multi-omic data. For many studies of human health and disease, combining proteogenomic and metaproteomic analysis offers great promise for new discoveries. Proteogenomics integrates high throughput genomic/transcriptomic sequencing data with MS-based proteomic data, identifying novel, disease-associated proteoforms in the human host. Metaproteomics integrates metagenomic data with MS-based proteomic data, identifying disease-associated proteins expressed by microbial communities residing in the host. Informatic workflows for proteogenomics and metaproteomics are inherently complex, and not run in combination. Consequently, neither tandem proteogenomic-metaproteomic analysis, nor the new discoveries it offers, is accessible to most researchers. In response, we describe a solution - the Galaxy framework for bioinformatic workflow development. Galaxy offers many unique and powerful features for multi-omic applications, including the ability to integrate numerous disparate software and processing steps, into automated workflows. These complete workflows can be shared with others, promoting reproducibility and accessibility. Capitalizing on Galaxy's flexibility, we have developed a novel tandem proteogenomic-metaproteomic workflow, presented here.

Galaxy-P

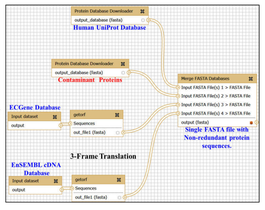
We are developing Galaxy for proteomics, or Galaxy-P, which is an extension of the Galaxy framework implementing software and tools for MS-based proteomic and multi-omic analysis applications. Unique benefits of Galaxy-P include:

- Flexibility for integrating established and new software tools into customized workflows
- Accessibility for biologists (usegalaxyp.org)
- Share-ability of tools, complete workflows and histories – promoting reproducibility, transparency and community evaluation

Galaxy-P Tools

Galaxy-P has multiple software tools – some specific to MS-based proteomics, others for genomic or transcriptomic data analysis

Galaxy-P Workflow



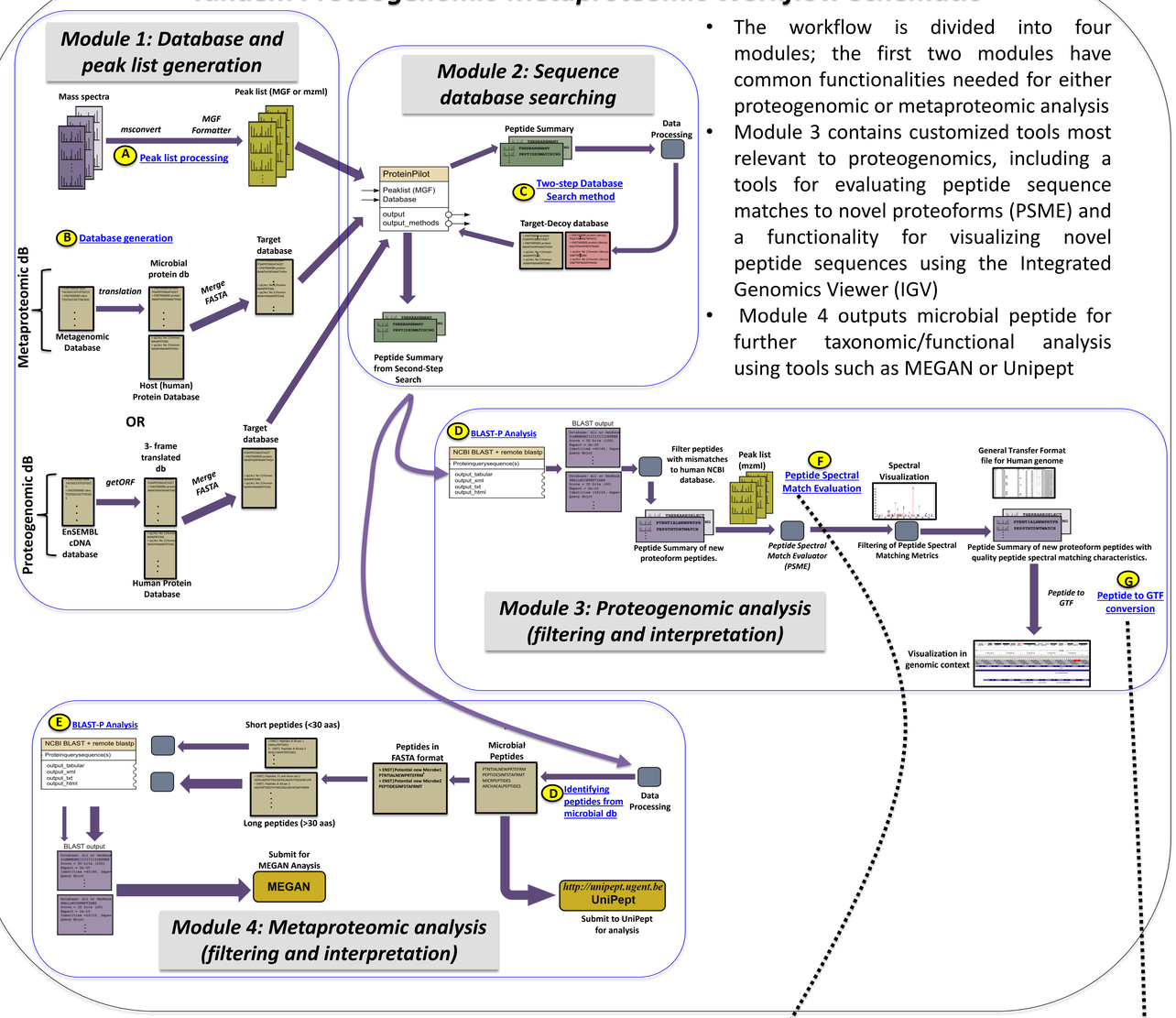
Tools can be integrated to generate modular workflows that can be reused, shared and creatively modified for multiple studies.

METHODS

RAW files from multiple datasets (see below) were generated from Orbitrap XL instrument. The processed peak lists were searched using ProteinPilot ™ version 4.5 (AB Sciex) within Galaxy-P. After optimization and testing, multiple workflows were used in a sequential manner to generate inputs for the subsequent workflow.

- Salivary supernatant was 3D-fractionated with or without ProteoMiner treatment (Bandhakavi *et al* 2009) was used. 200 RAW files were acquired on LTQ/Orbitrap XL. Both the datasets were searched against the human oral microbiome database (HOMD) or the 3-frame translated human cDNA database using the "Minnesota two-step" method (Jagtap et al 2013)

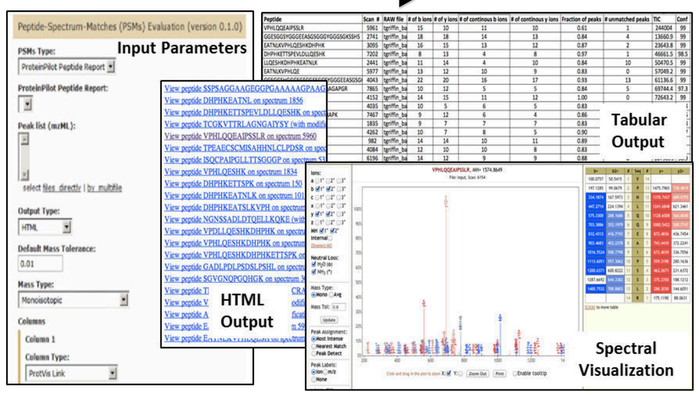
Tandem Proteogenomic-Metaproteomic Workflow Schematic



- The workflow is divided into four modules; the first two modules have common functionalities needed for either proteogenomic or metaproteomic analysis
- Module 3 contains customized tools most relevant to proteogenomics, including a tools for evaluating peptide sequence matches to novel proteoforms (PSME) and a functionality for visualizing novel peptide sequences using the Integrated Genomics Viewer (IGV)
- Module 4 outputs microbial peptide for further taxonomic/functional analysis using tools such as MEGAN or Unipept

Customized Tools in Galaxy-P

- The **peptide sequence match evaluator (PSME) tool** provides stringent evaluation and visualization of PSMs to novel peptide sequences; users can select levels of stringency and submit batches of PSMs

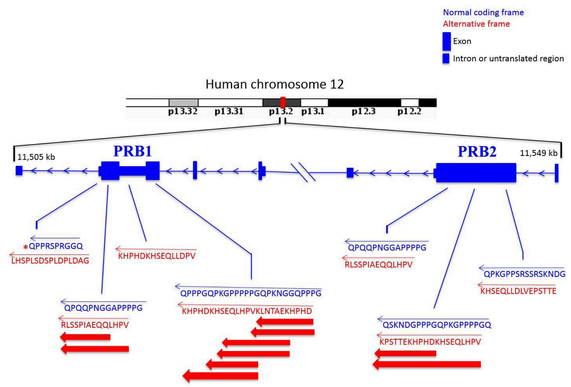


- After filtering and evaluation, novel peptides of interest are **converted to a GTF file for viewing with the popular IGV software**; novel identified peptide sequences can be compared to reference genome or protein sequences

ANALYSIS RESULTS

Proteogenomics

Total spectra	Novel proteoform peptides	Novel proteoform peptides (filtered)	Distinct, peptide sequences mapped to genome
988,974	254	105	52

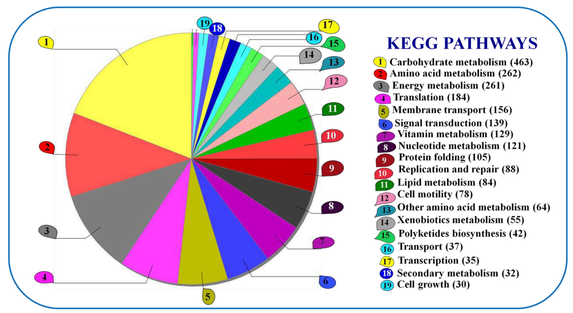


Novel peptide sequences identified and mapping to the proline-rich basic proteins 1 (PRB1) and 2 (PRB2). Peptides in red were identified as frameshifted sequences; peptides in blue represent expected sequences from the normal reading frame. Arrows show sequence direction (N- to C-terminus). Block red arrows depict distinct peptides identified within each continuous frameshifted peptide.

* = stop codon position

Metaproteomics

Total spectra	Distinct peptides of microbial origin	Phyla	Genera	Species
988,974	1926	12	65	123



Functional analysis of microbial proteins in whole saliva analyzed via the MEGAN software for taxonomic and bioinformatic analysis of large-scale metagenomic or metaproteomic data.

CONCLUDING REMARKS

- GalaxyP provides unsurpassed flexibility for integrating software and building powerful workflows for parallel proteogenomic and metaproteomic analysis
- Customized tools such as automated BLASTP searching and PSME provide a means to ensure high quality results from these analyses
- Output is compatible with IGV for visualizing novel proteoforms (proteogenomics) and MEGAN or UniPept taxonomic/functional analysis (metaproteomics)
- Complex workflows (~150 steps) can be run with little user intervention and shared in their completeness with others
- This work was funded by NSF grant 1147079