THE GALAXY FRAMEWORK AS A UNIFYING BIOINFORMATICS SOLUTION FOR 'OMICS' CORE FACILITIES

Pratik Jagtap[1]; James Johnson[2]; Bart Gottchalk[2]; Getiria Onsongo[2]; Sricharan Bandhakavi[3]; Joel Kooren[4]; Brian Sandri[4]; Ebbing de Jong[1]; Todd Markowski[1]; LeeAnn Higgins[1]; Chris Wendt[4]; Joel Rudney[4] and Timothy Griffin[4]

1.Center for Mass Spectrometry and Proteomics 2. Minnesota Supercomputing Institute 3. Bio-Rad Laboratories, Richmond, CA 4. University of Minnesota, Minneapolis, MN 55455

INTRODUCTION

- **Integration of different 'omics' data reveals novel discoveries into biological systems.**
- **However - the need for multiple, disparate software makes the integration of multi-omic data a serious challenge.**
- **Extension of Galaxy (a web-based bioinformatics data analysis platform) for mass spectrometric-based proteomics software enables advanced multi-omic applications such as proteogenomics, metaproteomics and quantitative proteomics.**

GALAXY-P

Galaxy-P Tools

Galaxy-P Workflow

Galaxy-P has multiple software tools - some proteomics-specific - and others from the genomics Galaxy framework. → **Tools can be used in a sequential manner to generate workflows that can be reused, shared and creatively modified for multiple studies.**

Benefits of Galaxy / Galaxy-P:

- ***Software accessibility and usability.***
- ***Share-ability of tools, workflows and histories.***
- ***Reproducibility and ability to test and compare results after using multiple parameters.***
- ***Analytical transparency***
- ***Scalability of data***

METHODS & DATASETS

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

METAPROTEOMICS

- **Severe Early Childhood Caries (SECC) dataset for clinical comparison of oral microcosm biofilms grown from plaque either in presence or absence of sucrose.**
- **Salivary supernatant dataset - 3D-fractionated with or without ProteoMiner treatment (Bandhakavi *et al* 2009). 200 RAW files were acquired on LTQ/Orbitrap XL. Both the datasets were searched against the human oral microbiome database (HOMD) using the two-step method (Jagtap *et al* 2013).**

PROTEOGENOMICS

- **Salivary supernatant - same as in metaproteomics study above.**
- **Oral pre-malignant lesion (OPML) dataset was collected as brush biopsy sample from six individuals with pre-malignant lesions and a matched control sample from adjacent oral cavity (Kooren *et al* unpublished). Both the datasets were searched against 3-frame translated cDNA database and the human oral microbial database by using two-step method (Jagtap *et al* 2013).**

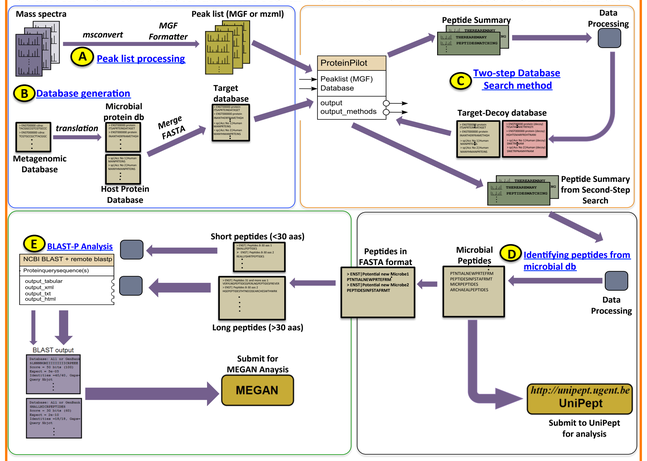
QUANTITATIVE PROTEOMICS

- **Chronic Obstructive Pulmonary Disease (COPD) – linked lung cancer tissue samples were collected and subjected to iTRAQ labeling and 2D LC-MS. The dataset was searched against Human UniProt database.**

METAPROTEOMICS

SECC and Salivary datasets

OVERVIEW OF MODULES AND ANALYTICAL WORKFLOWS FOR METAPROTEOMIC ANALYSIS.



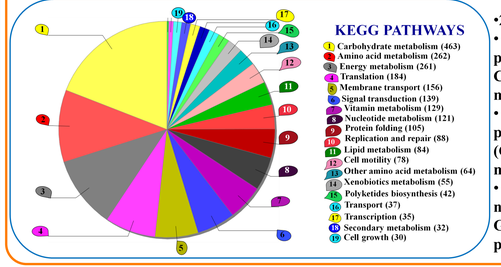
Shareable workflows: (A) z.umn.edu/peaklistconversion (B) z.umn.edu/dbgenmp (C) z.umn.edu/mn2stp (D) z.umn.edu/pepfrommicrobialdb (E) z.umn.edu/blastp

All together : z.umn.edu/mp65

Results Summary

Dataset	Total spectra	Distinct peptides of microbial origin	Phyla*	Genera*	Species*
Whole human salivary supernatant	988,974	1926	12	65	123
SECC without sucrose	153,019	28,126	5	33	56
SECC with sucrose	139,759	23,029	5	13	33

** Analysis using MEGAN.*

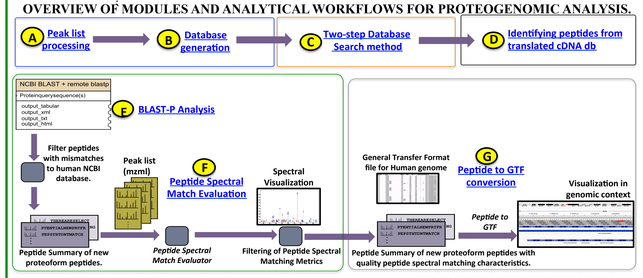


- **20 KEGG pathways.**
- **Most prevalent pathway : Carbohydrate metabolism.**
- **'Best-populated' pathway : Glycolysis (Carbohydrate metabolism).**
- **Protein with highest number of reads: Glyceraldehyde-3-phosphate.**

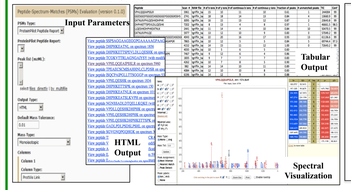
PROTEOGENOMICS

Salivary and OPML datasets

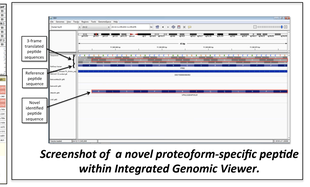
OVERVIEW OF MODULES AND ANALYTICAL WORKFLOWS FOR PROTEOGENOMIC ANALYSIS.



PEPTIDE SPECTRAL MATCH EVALUATION



GENOMIC CONTEXT VISUALIZATION



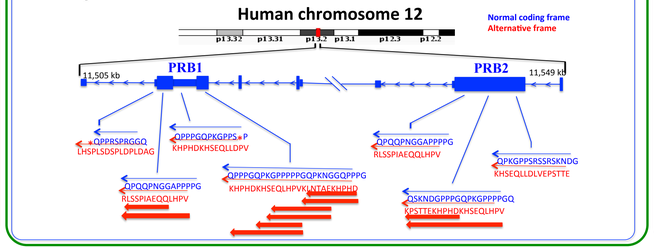
Screenshot of a novel proteoform-specific peptide within Integrated Genomic Viewer.

Shareable workflows: (A) z.umn.edu/peaklistconversion (B) z.umn.edu/dbgen (C) z.umn.edu/mn2stp (D) z.umn.edu/peptidesfromcdnadb (E) z.umn.edu/blastp (F) z.umn.edu/psme (G) z.umn.edu/pep2gtf All together : z.umn.edu/pg140

Results Summary

Dataset	Number of spectra	Novel proteoform peptides	Novel proteoform peptides filtered after PSM evaluation	Number of distinct peptides after visualization and for genome localization.
Salivary supernatant	988,974	254	105	52
OPML Control	156,405	904	34	17
OPML Lesion	157,299	887	29	21

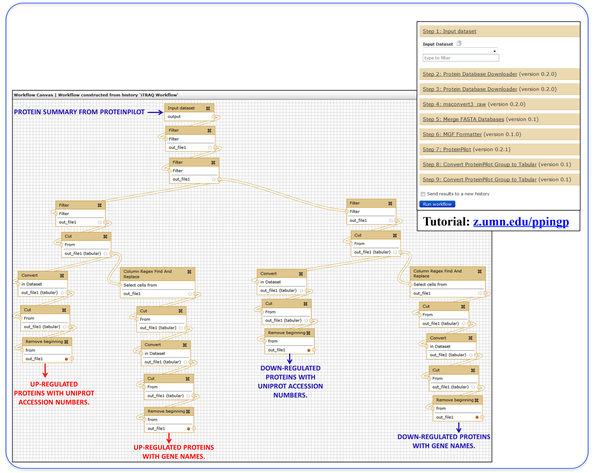
Representation of genomic organization of identified novel proteoform-specific peptides from PRB1 and PRB2 genes on chromosome 12.



QUANTITATIVE PROTEOMICS

COPD dataset

WORKFLOW AND TUTORIAL



CONCLUDING REMARKS

- **Salivary ProteoGenomics : 52 novel proteoforms identified (Notably - alternate frame translation for PRB1 and PRB2 proteins**
- **SECC Metaproteomics: Analysis of outputs from Galaxy-P workflows and MEGAN analysis is currently ongoing for three replicates**
- **Quantitative proteomics : Workflows have been used on multiple replicates. Reproducibility analysis and RNASeq data integration in works.**

- **IMMEDIATE PLANS:**
  - **Working along with genomics research community and Genomics Center to develop on integration of RNASeq derived workflows for database generation.**
  - **Working closely with metagenomics / microbiome research community to develop functional pathway analytical workflows using the metaproteomics data.**
  - **Working on correlating RNASeq quantitative information with quantitative iTRAQ proteomic information for both model and non-model organisms.**

- **FUTURE PLANS:**
  - **Installation and testing of open-source tools (such as MS-GF+ and PeptideShaker). The installation and testing is being carried out through and international collaboration between developers and users.**
  - **Improving on current metaproteomic and proteogenomic workflows.**

Acknowledgements : This work was funded by NSF grant 1147079. Also many thanks to John Chilton (Penn State) for his work in the first year of the project.