Flexible, accessible & reproducible workflows for tandem proteogenomic and metaproteomic analysis using the GalaxyP platform.

Pratik Jagtap[1]; Brian Sandri[2]; Julie Yang[2]; Kevin Murray[2]; Joel Kooren[2]; James Johnson[3]; Getiria Onsongo[3]; Joel Rudney[2]; Christine Wendt[2] and Tim Griffin1[1,2]

1. Center for Mass Spectrometry and Proteomics, UMN, St. Paul, MN; 2. University of Minnesota, Minneapolis, MN; 3. Minnesota Supercomputing Institute, Minneapolis, MN.

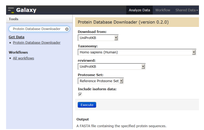
University of Minnesota Driven to Discover℠

INTRODUCTION

- **Proteogenomics (for identifying unannotated proteoforms) and metaproteomics (for characterizing non-host/multi-organism proteomes) are research areas that extend discoveries beyond the reference proteome.**
- **For biomedically-relevant proteomics studies, tandem proteogenomic and metaproteomic analysis offers great promise for new discoveries.**
- **We describe effective and accessible bioinformatic analytical workflows, amenable to creative customization and sharing to foster collaborative research efforts.**

GALAXYP

GalaxyP Tools



GalaxyP Workflow

GalaxyP 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 / GalaxyP:

- *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 GalaxyP (usegalaxyp.org). 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). After optimization & testing, multiple workflows were used in a sequential manner to generate inputs for the subsequent workflow. Microbial peptides were identified after using metaproteomic workflows & novel proteoforms were identified after using proteogenomic workflows.

UNLABELED SAMPLE:

- **Oral pre-malignant lesion (OPML) dataset was collected as oral exudate using PerioPaper strip method (Kooren *et al* 2011) from an individual with pre-malignant lesion & a matched control sample from adjacent oral cavity.**

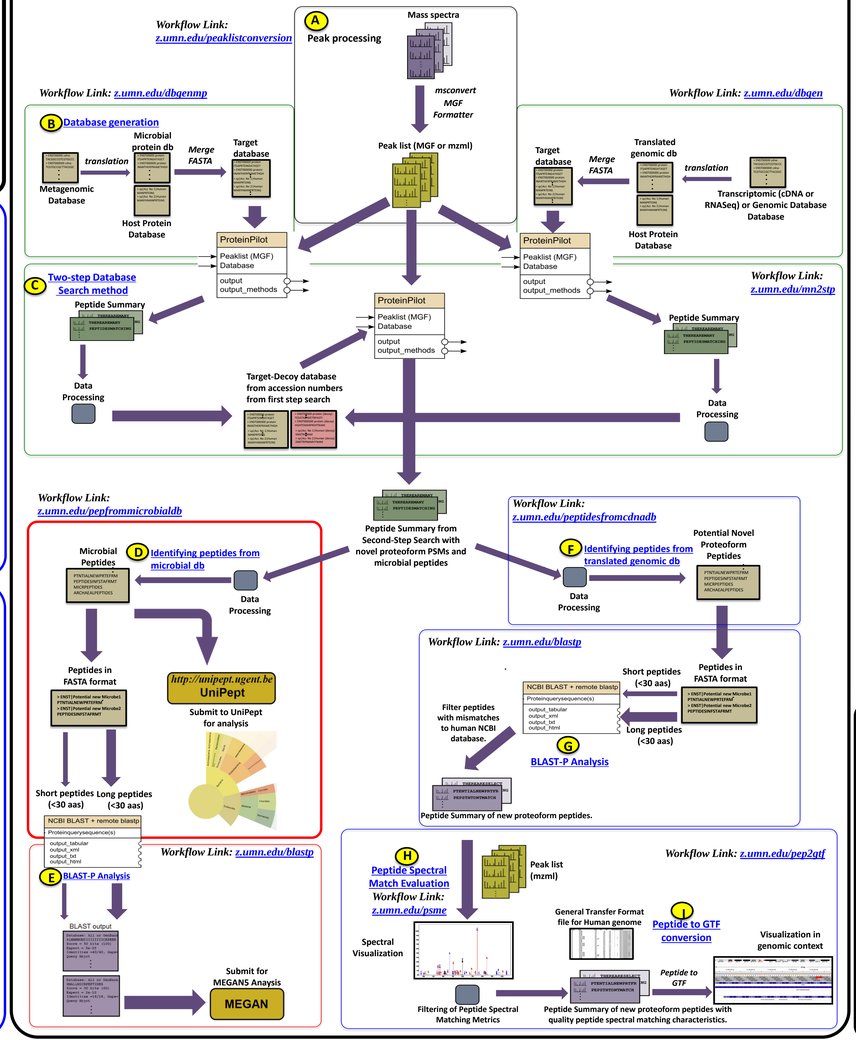
4-plex iTRAQ LABELED SAMPLE:

- **Brush biopsies were collected from patients diagnosed with OPML and from patients with Oral Squamous Cell Carcinoma (OSCC). For each patient, brush biopsies from the lesion & the healthy mucosa of corresponding contralateral area were collected (Yang et al 2014).**

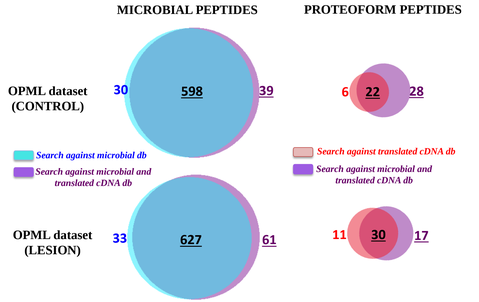
8-plex iTRAQ LABELED SAMPLE:

- **Chronic Obstructive Pulmonary Disease (COPD) – linked lung cancer tissue samples were collected & subjected to iTRAQ labeling and 2D LC-MS. Ten replicates of this dataset were searched against the 3-frame translated cDNA database & human oral microbiome database (HOMD) using the two-step method.**

WORKFLOWS FOR TANDEM PROTEOGENOMIC AND METAPROTEOMIC ANALYSIS.



EFFECT OF SEARCH DATABASES ON MICROBIAL & NOVEL PROTEOFORM IDENTIFICATIONS.

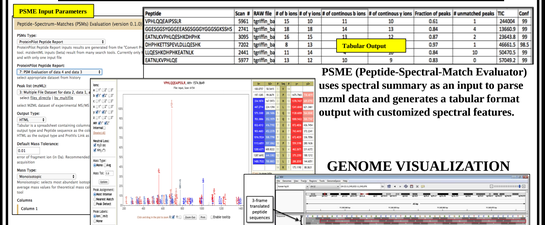


Using correct search databases that contribute to the proteome under study (both metaproteomic & proteogenomic databases in tandem) help in confident identification of microbial peptides & novel proteoforms.

RESULTS SUMMARY

Dataset	RAW Files	Distinct peptides of microbial origin	Number of unique peptides (Species Identified)	Novel proteoform peptides
OPML Control (unlabeled)	7	637	136 (6)	50
OPML Lesion (non-labeled)	7	688	136 (3)	47
Brush Biopsy OSCC (4-plex iTRAQ)	15	1118	166 (6)	6
COPD (10 replicates) (8-plex iTRAQ)	150	87	9 (9)	7

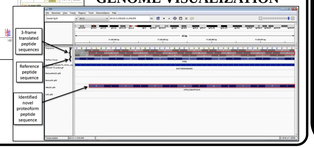
PSM EVALUATION & GENOME VISUALIZATION



PSME (Peptide-Spectral-Match Evaluator) uses spectral summary as an input to parse mzml data and generates a tabular format output with customized spectral features.

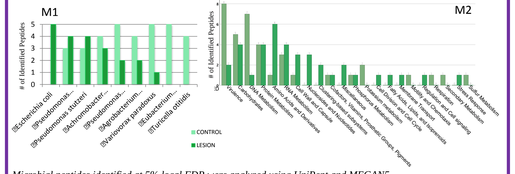
PSME also generates HTML links that are used to visualize spectral assignments. Interactive boxes can be used to change ion assignments and other parameters.

GENOME VISUALIZATION



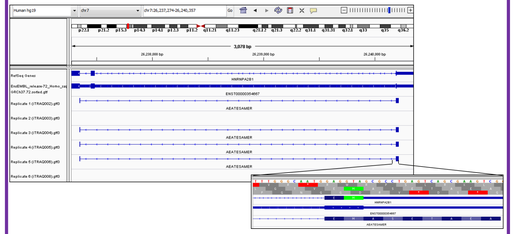
RESULTS

METAPROTEOMICS



Microbial peptides identified at 5% local FDR were analyzed using UniPept and MEGAN5.
***Fig M1:** Organisms were identified at species level only when assigned 3 unique peptides or more.*
***Fig M2:** Functional groups were assigned using SEED program within MEGAN5.*
In COPD dataset, five lung-infecting organisms were identified. Actinomyces viscosus – a bacterium that causes Actinomycosis (granulomatous infection with the formation of abscesses) in the lungs was identified in five replicates.

PROTEOGENOMICS



***Screenshot of a novel proteoform peptide within Integrated Genomic Viewer.** Novel proteoform peptides identified at 5% local FDR were used to generate GTF file so that the genomic localization of these peptides can be viewed in genomic context. Proteoform corresponding to HNRNPA2B1 was identified to have an alternative start site and was identified in 4 replicates.*

CONCLUSIONS

- **We demonstrate the use of a complete platform for tandem metaproteomic / proteogenomic analysis. Workflow for each module/step have been shared for use within Galaxy environment.**
- **Using both metaproteomic & proteogenomic databases in tandem help in confident identification of microbial peptides & novel proteoforms.**
- **Using this platform, we have identified microbial peptides & novel proteoforms from both labeled & unlabeled datasets. For example for COPD datasets, the number of identifications from both proteogenomic & metaproteomic databases is limited yet consistent across replicates..**
- **For metaproteomics studies, identified microbial peptides were used for taxonomic classification (Unipept and MEGAN5) for functional classification (*MEGAN5*).**
- **For proteogenomic studies, identified novel proteoforms were validated using PSM evaluation tool and visualizing peptides against the genome.**

***ACKNOWLEDGEMENTS:** GalaxyP is supported through the National Science Foundation Grant 1147079. Many thanks to John Chilton (PennState) for GalaxyP development. Also thanks to LeeAnn Higgins, Todd Markowski (CMSP, UMN), Bart Gottschalk and Anne Lamblin (MSI), Katie Vermillion (UMD-Duluth) and Gloria Shenykman (UW-Madison) for helpful discussions.*