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 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.

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 ProToMiner 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 microbe database (HOMID) or the 3-frame translated human cDNA database using the “Minnesota two-step” method (Jagtap et al. 2013)

ANALYSIS RESULTS

- 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 tool for evaluating peptide sequence matches to novel protein sequences (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 UniPept or MEGAN.

CONCLUDING REMARKS
- Galaxy-P 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