Title: Enabling Metaproteomics Research

Ernesto S Nakayasu[1], Meagan C Burnett[1], Grant M Fujimoto[1], Joon-Yong Lee[1], **Samuel H Payne[1]* (samuel.payne@pnnl.gov)**

1. Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA

Project Goals: This project is focused on improving algorithms and methods for mass spectrometry data analysis of metaproteomics data. Recent advances in mass spectrometry and biological separations have dramatically increased the depth of proteomic discovery. Unfortunately, traditional computational workflows are in many cases preventing researchers from realizing these benefits for microbial communities. We propose to create a new generation of computational workflows to overcome the sensitivity limitations inherent in status quo data processing schemes.

To advance our ability to annotate tandem mass spectrometry data from microbial communities, our project has been developing algorithms to match spectra from metaproteomics experiments to a library of annotated spectra. This year we made significant improvements in both the breadth of coverage for proteomics data of environmental bacteria and also the efficiency of algorithms for peptide annotation. We have expanded our proteomic coverage to 50 new organisms, focusing on organisms found in the soil and the terrestrial/aquatic interface. The availability of proteomics data for many previously unstudied organisms in these niches helps to elucidate protein functional regulation, e.g. by studying conserved post-translational modifications and conserved gene expression networks across orthologs. With the expansion of proteomics libraries growing to hundreds of species and millions of spectra, annotation algorithms face a major hurdle in computational efficiency. To overcome this, we created a new algorithm called "FLASH" to dramatically improve the speed of annotation. FLASH is a hybrid library assisted annotation algorithm, meaning that it leverages annotated libraries but is not limited by them. This flexibility is necessary for the diverse biomes that are sampled in metaproteomics data.

This project was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Early Career Research Program.