Pan-omics Informatics and Advanced Analysis Pipelines

Gordon A. Anderson* (gordo@pnnl.gov), Ronald J. Moore, Joshua N. Adkins, Kenneth J. Auberry, Mikhail E. Belov, Kristin E. Burnum, Kevin L. Crowell, Stephen J. Callister, Therese R.W. Clauss, Kim K. Hixson, Gary R. Kiebel, Brian L. LaMarche, Mary S. Lipton, Da Meng, Thomas O. Metz, Matthew E. Monroe, Heather M. Brewer, Carrie D. Nicora, Angela D. Norbeck, Daniel J. Orton, Ljiljana Paša-Tolić, Samuel O. Purvine, John D. Sandoval, Anuj Shah, Yufeng Shen, Anil K. Shukla, Mudita Singhal, Gordon W. Slysz, Aleksey V. Tolmachev, Nikola Tolić, Karl Weitz, Aaron T. Wright, Rui Zhang, Rui Zhao, and **Richard D. Smith** (PI; rds@pnnl.gov)

Biological Sciences Division and Environmental Molecular Sciences Laboratory Pacific Northwest National Laboratory, Richland, WA 99352

As part of our Proteomics and Pan-Omics project in support of GSP Systems Biology Research, we are developing an advanced informatics framework to enable application of pan-omics measurements that afford the comprehensive global molecular characterization necessary for understanding, modeling, and potentially manipulating biological systems. Pan-omics measurement capabilities are based on essentially identical separations-MS measurement platforms and similar data processing/informatics pipelines: metabolomics, lipidomics, and glycomics measurements in addition to expanded proteomics measurements. A key element of the pan-omics informatics framework are tools and methodologies for integrating data from various omics measurements and determining data quality and ambiguities (e.g., the confidence in peptide and protein identifications, modification sites, and abundance levels), as well as approaches for managing and communicating data.

We are leveraging the extensive experience and capabilities developed to date at PNNL to extend high throughput proteomics measurements to multiple omics measurements and provide a framework for evaluating and controlling data quality; processing and integrating data from the various analysis streams; and disseminating data and information to collaborators, users, and the broader scientific community. The Pan-omics Research Informatics Storage and Management (PRISM) system has been enhanced to support advanced pan-omics analysis pipelines that allow automated high level data analysis and integration of hundreds to thousands of individual datasets into one comprehensive research result report. The framework further supports integration of genomics data from public repositories and aims to provide the needed infrastructure to interoperate with the GTL Knowledgebase.

Current efforts are focused on developing a suite of data analysis tools, data consolidation applications, and statistical packages, as well as visualization software for data interpretation and integration. New and advanced tools support metabolomics, glycomics, and top-down proteomics pipelines, as well as pan-omics data integration. Each pipeline produces a self-documented, tested, and repeatable analysis, complete with visual inspection plots produced for each step in the pipeline. The advanced automation and critical infrastructure elements demonstrate high throughput and automated file processing, as well as data filtering and result aggregation.

Acknowledgements: This research is supported by the Office of Biological and Environmental Research (OBER) of the U.S. Department of Energy (DOE). Portions of this research were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's OBER. Pacific Northwest National Laboratory is operated for the DOE by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.