UMPIRE: A Reproducible Pipeline for Scalable Metabolomics Data Analysis

Xinsong Du1, Luran Manfio1, William Hogan1, Timothy J. Garrett2, Dominick Lemas1

1 Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida

2 Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida

**Background:** Metabolomics includes the study of small molecules (molecular weight < 1500Da) acting as middle or end products of cellular metabolism. Large-scale epidemiological studies are increasingly collecting metabolomics data in biologically relevant tissues to understand metabolic variation in the development of health and disease resulting in a rich array of complex, high-dimensional data that must be handled through computational means. However, reproducibility is vital in science, and metabolomics data analysis remains an area of high concern, particularly cross platform. Although an increasing number of open source software packages have been developed to complete metabolomics analysis, most of these programs were designed for single desktop users and lack high-throughput scalability and reproducibility. Currently, software containers are able to package all codes and dependencies of an application to ensure portability, infrastructure flexibility and reproducibility. Applications running with the container will depend on the environment pre-built in the container regardless of the environment in the host machine. Nextflow is a pipeline development tool supporting containerization and high performance computing, which improves the scalability and reproducibility of bioinformatics research

**Objective**: The goal of our project is to develop an open-source tool using Nextflow to facilitate reproducible and scalable metabolomic data analysis.

**Findings**: UMPIRE can be executed on any UNIX-like systems with a specific focus on implementing UMPIRE within large-scale computing environments. We have developed UMPIRE to run in paralleled common metabolites analysis software packages including MzMine and XCMS for peak detection to enhance precision. Moreover, UMPIRE provides users multiple statistical analysis algorithms including SECIMtools, principle component analysis, volcano plots and hierarchical clustering. To facilitate dynamic and transparent data processing, we have included MultiQC reports to visualize the result of data processing and summarize metabolomics output. Furthermore, we will test the relationship between input size and computing resource use, and provide suggestions for computing resource allocation so that users can manage their resources efficiently.

**Conclusion**: UMPIRE is a container-based platform that has potential to facilitate high-throughput and scalable metabolomic data analysis with high levels of reproducibility and transparency.