**A Reproducible Pipeline for Scalable Untargeted Metabolomics Data Analysis**

Xinsong Du1, Luran Manfio1, Alexander Kirpich2, William Hogan1, Timothy J. Garrett3, Dominick J. Lemas1

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

2 Department of Population Health Sciences, School of Public Health, Georgia State University

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

**Background:** Untargeted metabolomics data is increasingly collected by epidemiological studies to investigate population-level variation in the development of health and disease. Reproducibility of untargeted metabolomics data analysis remains a challenge. Although an increasing number of open source software packages have been developed to complete untargeted 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 in 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 untargeted metabolomic data analysis.

**Findings**: The pipeline can be executed on any UNIX-like systems with a specific focus on implementation within large-scale computing environments. We have developed the pipeline to run parallelized common metabolites analysis software packages including MzMine and XCMS for peak detection to enhance precision. Moreover, we used Python for statistical tests such as ANOVA and student t-test, as well as providing users multiple visualization methods including principle component analysis, volcano plots, hierarchical clustering and others. To facilitate dynamic and transparent data processing, we have included MultiQC reports to visualize the result of data processing and summarize metabolomics output. Furthermore, the relationship between the input file size and the use of corresponding computing resources is monitored within the pipeline. This provides useful suggestions for further computing resource allocation and efficient use. In a pilot experiment, we used two different operating systems in which the MZmine versions were also different to do peak detection for four sample metabolomics data. This simulates the situation that one researcher wants to reproduce the other researcher’s published work with a different operating system and software version. We got very different peak numbers from MZmine of the two host machines, but the exact same peak numbers when employing our Nextflow pipeline.

**Conclusion**: We have developed a container-based platform that has potential to facilitate high-throughput and scalable untargeted metabolomics data analysis with high levels of reproducibility and transparency.