**Research Reproducibility 2020  
Educating for Reproducibility: Pathways to Research Integrity  
University of Florida, Gainesville, FL, USA**

**A Highly-Reproducible Workflow for Untargeted Metabolomics Data Processing**

**Xinsong Du**(1)**, Luran Manfio**(1)**, Alexander Kirpich**(2**, William R. Hogan**(1)**, Timothy J. Garrett**(3)**, Dominick J. Lemas**(1)

(1)*Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida. {xinsongdu, manfiol, hoganwr, djlemas}@ufl.edu*

(2)*Department of Population Health Sciences, School of Public Health, Georgia State University.* [*akirpich@gsu.edu*](mailto:akirpich@gsu.edu)

(3)*Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida.* [*tgarrett@ufl.edu*](mailto:tgarrett@ufl.edu)

##### ABSTRACT

**Background:** Reproducibility of untargeted metabolomics data processing remains a challenge.

**Objective**: The goal of our study is to build a containerized workflow that can improve reproducibility of untargeted metabolomics data processing.

**Methods**: We used Nextflow, a pipeline development tool supporting containerization, and high performance computing (HiPerGator) to develop the workflow. Docker container were used to bundle all codes and dependencies. Metabolomic data processing was completed using MZmine-2.53. We tested the performance of our metabolic pipeline using human milk samples collected as part of the Breastfeeding and Early Child Health (BEACH) Study.

**Findings**: To illustrate how our workflow can be used, the numerical instability caused by different versions of data processing software (i.e., MZmine) was evaluated. Significant variations of detected number of peaks of the four samples across different software versions (MZmine-2.11 versus MZmine-2.53) were observed (paired t-test p-value is 0.002), which contrasts with the deterministic behavior produced by each of the individual versions of MZmine. As a fixed version of MZmine (version 2.53) and all its dependencies were dockerized and incorporated in our pipeline, we are able to obtain stable readouts.

**Conclusion**: Our results demonstrate a Nextflow-based framework for untargeted metabolomics data processing has potential to improve reproducibility at the level of data processing.