**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 , (HiPerGator) were used to bundle Metabolomic data processing was completed usig We tested the performance of our metabolic pipeline using human milk as part of

**Findings**: Our results

Although we found we got the exact same number of peaks when using two different operating systems (Mac and Linux) for all four samples (paired t-test p-value is 1.000), previous study shows different operating systems can produce slightly different bioinformatics study results. Therefore it is good to have a docker container. Moreover, we found different MZmine version (we tested 2.11 and 2.53) can produce significantly different number of peaks for the four samples (paired t-test p-value is 0.002). However, by randomly reading recent 10 literature using MZmine (database: PubMed; search keyword: mzmine; sort by most recent), we found only 2 of them mentioned the detailed version of MZmine they used, others just said they used MZmine 2. When employing our Nextflow-based workflow, we got the exact same number of peaks (paired t-test p-value is 1.000) in two different host machines, even though their operating systems and MZmine versions are different.

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