**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:** 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 processing remains a challenge, which can result from the use of instrument, different data processing software and different operating system. Therefore, it is necessary and feasible to create a workflow that can increase reproducibility and facilitate collaborative research by controlling the use of data processing software and operating system.

**Objective**: The goal of our study is to build up a workflow that can improve reproducibility of untargeted metabolomics data processing, and test its performance.

**Findings**: Nextflow, a pipeline development tool supporting containerization and high performance computing, was used to develop the workflow. Docker container, which packages operating system as well as all codes and dependencies, was created specifically for the workflow. MZmine-2.53 was included in the workflow as the data processing software. Four milk samples collected from the Breastfeeding and Early Child Health (BEACH) Study were used for the experiments. During tests, 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**: We proposed a Nextflow-based framework for untargeted metabolomics data processing and tested its reproducibility. We found our framework can effectively improve the reproducibility thus facilitate collaborative research.