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

A Reproducible Pipeline for Scalable Untargeted Metabolomics Data Analysis

Xinsong Du(1), **Luran Manfio**(1), **Alexander Kirpich**(32, **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*

(3)*Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida. 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 analysis remains a challenge. Although an increasing number of open source software packages have been developed to complete untargeted metabolomics analysis, study shows different software or even software version can produce very different results. 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 embedded a common metabolomics analysis software – Mzmine, to Nextflow for the purpose of enhancing reproducibility. Moreover, we used Python for statistical tests, as well as providing users multiple visualization methods including principle component analysis, and hierarchical clustering. To facilitate dynamic and transparent data processing, we have included MultiQC interactive reports to visualize the result of data processing and summarize metabolomics output. To test our pipeline, 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 machine. We got very different peak numbers from MZmine of the two machines, but the exact same peak numbers when using 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.