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

**MetaboMax-10: An Approach to Enhance Reproducibility of 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*](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 analysis remains a challenge. Reproducibility issue can be addressed by mitigating numerical instability. Previous study indicates different software can generate very different results for metabolomics data processing even if the used algorithms were the same. Research also shows using different operating systems to run software can also produce numerical instability in a bioinformatics study. However, no research has been conducted to explore whether the untargeted metabolomics data processing algorithm itself can cause numerical instability.

**Objective**: The goal of our study is to test if metabolomics data processing algorithm itself can produce numerical instability, and propose an approach to address it.

**Findings**: In our experiments, we used Red Hat Enterprise Linux Server release 7.6 to run MZmine-2.53 to process metabolomics data from the same research (Study MTBL146 from Metabolights) for multiple times, and see if we could get the exact same result each time. There were 443 samples in total. We increased the sample size from 2 to 443, and for each size we run peak detection algorithms for 10 times. We found as we increase the number of sample from 2 to 128 to 443), the standard deviation of number of detected peaks also becomes larger and larger (from 0 to 6.7 to 10.0), which means larger number of sample can cause stronger numerical instability. Since the data size in the future can become larger, the instability issue can become stronger and may become strong enough someday to make a research conclusion instable. To address this issue, we proposed MetaboMax-10 algorithm: 1) running peak detection algorithm for the same data for 10 times; 2) select the one with the highest peak number as final processing result. We ran MetaboMax-10 using all the 443 samples for 10 times, the standard deviation was reduced from … to ….

**Conclusion**: We tested the numerical instability produced by untargeted metabolomics data processing algorithm, and proposed an approach to mitigate this instability issue for the propose of improving reproducibility.