MassOmics: An R package of a cross-platform data processing pipeline for large-scale GC-MS untargeted metabolomics datasets.

Guangyu GUO1#, Elizabeth J. McKenzie2#, M. Beatrix Jones3, Erica Zarate4, Jamie V. de Seymour5, Philip N. Baker2, Silas G. Villas-Bôas4, Ting-Li Han6\*

1Mass spectrometry hub, The University of Auckland, New Zealand

2Liggins Institute, the University of Auckland, New Zealand

3Department of Statistics, The University of Auckland, New Zealand

4School of Biological Sciences, University of Auckland, New Zealand

5Massey University, Auckland, New Zealand

6Department of Obstetrics and Gynaecology, the First Affiliated Hospital of Chongqing Medical University, China

#Co-first author

\*Corresponding author: Ting-Li Han E-mail address: morgan\_han\_0816@hotmail.com

Large-scale gas chromatography-mass spectrometry (GC-MS) based untargeted metabolomics, where hundreds or thousands of samples are analysed over a period of weeks or months, has specific challenges. These include variation in instrument performance, signal intensity loss due to column ageing, the build-up of contaminants in the ion source, and sample handling variability. In addition, the computational challenges of compound identification are intensified when dealing with a large number of samples. A data processing software package to address these problems is required. Our software, MassOmics, is designed to bring together R packages and scripts for GC-MS data processing to rapidly integrate and annotate peaks in large-scale datasets, all within a graphical user interface. This package also provides identification of background contaminants, data scaling and transformation, various batch effect removal methods, machine learning-powered grouping of metabolites, and metabolite importance analysis. With these functions, MassOmics can parse and summarize library batch search results from ChemStation and MassHunter, and produce an output, which is compatible with various downstream statistical and metabolic pathway analysis tools. The module-based design and intermediate data transferring approach enable MassOmics to work with data integration platforms such as KNIME to generate an adaptive and customizable processing workflow. The MassOmics package is designed for researchers with little experience using R, and substantially improves GC-MS data extraction efficiency and accuracy, as well as reducing the time required for manual checking and re-integration.