**[Name]: A Container-based Reproducible Untargeted Metabolomics Data Processing Pipeline**

Xinsong Du1, Luran Manfio1, Alexander Kirpich2, William Hogan1, Timothy J. Garrett3, Dominick J. Lemas1

1 Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida

2 Department of Population Health Sciences, School of Public Health, Georgia State University

3 Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida

Keywords: febrile neutropenia; machine learning; mortality; cancer; HCUP

Author to whom correspondence should be addressed:

Dominick J. Lemas, Ph.D

Assistant Professor

Department of Health Outcomes and Biomedical Informatics

University of Florida College of Medicine

2004 Mowry Road- Clinical and Translational Research Building

Gainesville, FL 32608

Ph: 352-294-5971

Email: [djlemas@ufl.edu](mailto:djlemas@ufl.edu)

Abstract word count: 113

Abstract word limit: 250

Target Journal: GigaScience

Reference limit: not mention in the guideline

Reference count:

Total word limit: not mention in the guideline

Current word count:

**ABSTRACT**

***Background:***Untargeted metabolomics data processing remains a challenge due to the large size of metabolomics data files as well as the extensive steps and computational power required to process the data. The goal of our study is to build a containerized workflow that can enhance computational reproducibility of untargeted metabolomics data processing.

***Findings****:* Nextflow, a pipeline development tool supporting containerization, parallelization and high performance computing, was used to build the pipeline. Several tools and codes were connected and incorporated into the pipeline. [Name] can be executed on any UNIX-like system and job schedulers for high-performance computing are supported.

***Conclusions****:* A highly-reproducible containerized untargeted metabolomics data processing tool was developed to facilitate collaborative research.

**1.1 Background**

The development of high-throughput technology enables us to analyze large amount of data cost-effectively. A large number of research collecting and processing big volume data appears, which opens challenges for research reproducibility. A survey of 1,576 researchers indicates more than 50% were unable to reproduce their own experiments while 70% failed to reproduce other researcher’s results. Irreproducibility can be caused by data unavailability [cite 2-4 in YAMP paper], code or software variation [cite Dr. Du’s paper] and operating system difference [cite Nextflow paper].

**1.2 Findings**

1.2.1 The [Name] workflow

(A figure describing the workflow)

1.2.2 Implementation

1.2.3 [Name] input/output

(A screenshot of log file)

(A screenshot of execution profile)

1.2.4 Results

1.2.4.1 Simulation study

Reproducibility test among different operating systems

|  |  |  |  |
| --- | --- | --- | --- |
| Operating system | Experiment ID | M/Z value of creatinine | T-test for M/Z values |
| Linux (UF HiPerGator) | L1 | 114.06658299763997 |  |
| L2 | 114.06658299763997 |
| L3 | 114.06658299763997 |
| L4 | 114.06658299763997 |
| L5 | 114.06658299763997 |
| MacOS | M1 | 114.066580772399 |
| M2 | 114.066580772399 |
| M3 |  |
| M4 |  |
| M5 |  |
| Linux  +  [Name] | L6 |  |  |
| L7 |  |
| L8 |  |
| L9 |  |
| L10 |  |
| MacOS  +  [Name] | M6 |  |
| M7 |  |
| M8 |  |
| M9 |  |
| M10 |  |

The difference of M/Z value with the increase of sample size (figure).

1.2.4.2 Real-world case study

(Is it a good idea to use human-bovine study sample after the human-bovine paper is published?)

1.2.5 Discussion

**1.3 Methods**

1.3.1 FAIR criteria

|  |  |  |
| --- | --- | --- |
|  | Data | Tool |
| (F)indability | Stored in an indexable database (e.g. MetaboLights) | [Name] was open sourced and indexed in domain relevant software repositories -Github |
| (A)ccessibility | Data management and access follow e.g., data use ontology. | [Name] is publicly available and entitled with GNU license. |
| (I)nteroperability | Standardized data format (i.e., mzXML files) | Containerization enables [Name] to be run on different host machines. |
| (R)eusability | Data was deposited in indexable database and sites (e.g., OmicsDI). | [Name] has a use guideline. |

1.3.1 Data availability

1.3.1.1 Simulation study

1.3.1.2 Real-world case study

1.3.2 Data analysis

1.3.2.1 Simulation study

1.3.2.2 Real-world study

**1.7 Availability of source code**

Project name: [Name]

Project homepage: <https://github.com/GalaxyDream/metabolomics_data_processing>

Operating systems: UNIX-like system.

Programming language: Python, Bash, Nextflow

Other requirements: Java, Singularity/Docker

License: GNU GPL v3

Any restrictions to use by non-academics: None

Documentation and tutorials:

Scicrunch.org

**1.8 Availability of supporting data**

**1.9 Abbreviations**

**1.10 Competing interests**

The authors have no financial or personal relationships with other people or organizations that could inappropriately influence (bias) their work.

**1.11 Funding**

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases [K01DK115632] and the University of Florida Clinical and Translational Science Institute [UL1TR001427]. The content is solely the responsibility of the authors and does not necessarily represent the official views the University of Florida’s Clinical and Translational Science Institute, or the National Institutes of Health.

**1.7 Author contributions**

**1.8 Acknowledgements**

**1.8 References**

**FIGURES**

Screenshot of log file

Relationship between input size and memory cost

Relationship between input number of samples and memory cost

**TABLES**

**SUPPLEMENTARY MATERIAL**