
Subject Section

Federated OmicsDI: Cloud-based architecture for omics data discovery

Gaurhari Dass¹, Manh-Tu Vu¹, Pablo Moreno¹, David Ocana¹, Pan Xu², Weimin Zhu^{2,3}, Steven Newhouse¹, Henning Hermjakob^{1,2} & Yasset Perez-Riverol^{1,*}

¹ European Molecular Biology Laboratory, EMBL-European Bioinformatics Institute (EMBL-EBI), Cambridge, CB10 1S

² State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing 102206, China

³ CAS Key Laboratory of Computational Biology, BioMed Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences.

* To whom correspondence should be addressed. E-mail: Yasset Perez-Riverol (yperez@ebi.ac.uk)

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Omics Discovery Index (OmicsDI - www.omicsdi.org) is an integrated and open-source platform to facilitate the discovery and dissemination of omics datasets metadata. It provides a unique infrastructure to integrate datasets coming from multiple omics studies, including at present proteomics, genomics, transcriptomics, metabolomics, and systems biology. The OmicsDI architecture was originally implemented and deployed in a dedicated high-performance computing cluster, limiting scalability and dynamic allocation of resources by the data processing pipelines. In addition, the original OmicsDI resource could not be reused by independent laboratories and research groups to share and disseminate their data.

Results: Here, we present a new version of OmicsDI that can be easily deployed in cloud architectures and local infrastructures enabling the development of a Federated OmicsDI. The new architecture can be automatically synchronized with the main OmicsDI resource, increasing the integration with other omics data providers. Also, the proposed Cloud-based architecture is more scalable, providing better capabilities to manage the increase of data providers and datasets.

Availability and implementation: The software is freely available at <https://github.com/OmicsDI>

Contact: yperez@ebi.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

OmicsDI (www.omicsdi.org) (Perez-Riverol, et al., 2017; Perez-Riverol, et al., 2019) is a key open-source platform to integrate datasets metadata from many different omics databases into a single framework and interface. It makes life-science data more findable, accessible, and interoperable for both researchers and machines. OmicsDI used an efficient indexing system, which can integrate the metadata of each dataset and also the relevant biological entities associated with it like identified proteins, or genes changing their expression. A set of pipelines update the resource every

time a new dataset becomes publicly available in the contributing repository. By January 2021, OmicsDI provides metadata for over 2.3 million datasets, and the number will keep growing every month. In addition to indexing datasets from major omics archives such as the Gene Expression Omnibus (GEO), OmicsDI was designed to integrate datasets from smaller independent research laboratories. By supporting independent research groups and laboratories, OmicsDI aims to cover data providers that can't make their data public in major archives due to multiple reasons including volume or clinical and human protected datasets. However, the original implementation of the framework was linked to a specific high-performance computing architecture (HPC); limiting the possibility to reuse the OmicsDI framework by independent laboratories and consortiums

