Omics Discovery Index – Discovering and Linking Public 'Omics' Datasets

Henning Hermjakob, EMBL-EBI; Yasset Perez-Riverol, EMBL-EBI; Mingze Bai, EMBL-EBI and National Center for Protein Sciences; Gaurhari Dass, EMBL-EBI; Rodrigo Lopez, EMBL-EBI; Peipei Ping, Departments of Physiology, Medicine, and Bioinformatics, University of California, Los Angeles

Biomedical data, in particular omics datasets are being generated at an unprecedented rate. As a result, the number of deposited datasets in public repositories originating from various omics approaches has increased dramatically in recent years. However, this also means that discovery of all relevant datasets for a given scientific question is non-trivial. Here, we introduce the Omics Discovery Index (OmicsDI), an integrated and open source platform facilitating the access and dissemination of omics datasets. OmicsDI provides a unique infrastructure to integrate datasets coming from multiple omics studies, including at present proteomics, genomics transcriptomics, and metabolomics, as a globally distributed resource.

As of October 2016, OmicsDI provides a lightweight discovery tool including more than 80,000 omics datasets from ten different repositories, four different omics types, and three continents. While advanced metadata-based browsing and indexing supports dataset findability, the lightweight approach avoids the development of redundant concepts and infrastructure. The original datasets are not replicated, but referenced. In the interest of sustainability, the responsibility for provision of a well-formatted metadata records lies with the original data providers, similarly to the concept of publisher data provision to PubMed or EuroPMC.

OmicsDI supports full text search as well as ontology-based search extensions. In addition, we use the concept of biological dataset similarity, based on the number of shared biological entities, for example protein identifications, among datasets. This allows us to suggest potential relationships among datasets even if they don't share sufficient metadata annotation.

OmicsDI is accessible at http://omicsdi.org.