FAIR for automatic federated omics analysis

Daphne Wijnbergen[1,*], Georgios Malamas[1], Marco Roos[1] and Eleni Mina[1]

[1]*Human Genetics, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands*

Abstract

In this work, we create a workflow to apply federated gene expression meta-analysis in the Virtual Platform of the EJP RD. Based on this workflow, we identify which metadata is needed to make the data machine actionable. We then present a metadata schema that is based on the EJP RD metadata schema and consists of scientific, biological and file metadata.

Keywords

FAIR, Federated analysis, metadata, machine actionability

1. Introduction

In the analysis of biomedical data, a large amount of time and effort is spent on finding datasets, mapping identifiers, and data munging. An initiative that can help mitigate these issues is FAIR [1]. With FAIR, machines can increasingly perform actions on data without human intervention, if machine-actionable metadata is provided.

Another factor that hinders the application of data analysis in biomedical research is privacy. Human data, such as genomic data, is privacy sensitive and can not be fully anonymized. Consequently, data often cannot be accessed and analyzed from outside the institute where it was generated. Multiple efforts are ongoing to create infrastructures that enable federated analysis of data. In this paradigm, an analysis method can be sent from one institute to the data of another and executed, if approved. The results are then sent back to the first institute. This ensures that the analysis can be performed, while privacy is preserved. One such effort is the development of the "Virtual Platform" (VP) network of FAIR resources by the European Joint Programme on Rare Diseases (EJP RD). Currently, various resources relevant for rare diseases are FAIRified and connected within the VP. One goal of the EJP RD is to enable automated, federated analysis over the resources in the VP.

In our project, we created a workflow to apply federated analysis on omics data for rare diseases. To achieve this, we have identified what metadata is needed to perform this analysis by machines in an automated way.

SWAT4HCLS: 15th International SWAT4HCLS Conference, February 26–29, 2024, Leiden, The Netherlands

d.wijnbergen@lumc.nl (D. Wijnbergen); yiorgos.malamas@student.uva.nl (G. Malamas); m.roos@lumc.nl (M. Roos); e.mina@lumc.nl (E. Mina)

0000-0002-7449-6657 (D. Wijnbergen); 0009-0001-9556-2158 (G. Malamas); 0000-0002-8691-772X (M. Roos); 0000-0002-8972-9206 (E. Mina)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Methods

We implemented a workflow for gene expression analysis in Inclusion Body Myositis to serve as the basis of our use case. This workflow consists of four main steps: (1) Identifying transcriptomics datasets of interest (2) Applying differential gene expression analysis on these datasets (3) Mapping identifiers between datasets for data integration, (4) applying meta-analysis (analysis of multiple analysis results) to determine which genes are differentially expressed in multiple studies in Inclusion Body Myositis.

We identified what metadata is necessary for the data to be machine actionable for the purpose of this use case. The VP metadata schema and an extension of the Data Catalog Vocabulary (DCAT) [2] were analyzed and extended with metadata elements needed for our use case.

3. Results

We defined a metadata schema that extends the EJP RD and DCAT metadata schemas. This schema contains metadata in three categories: 1.Scientific metadata such as the measurement type, measurement device, and study design, that help find datasets that are measuring variables of interest E.g. measurements of gene expression in case vs control. 2. Biological metadata such as the disease, species and tissue, that are needed to select datasets that are biologically relevant for the research question; e.g. Selecting datasets for Inclusion Body Myositis. 3. Metadata about the data file itself, such as the download URL, the format, the media type, and a domain specific file specification, that are needed for the machine to understand how to use the data.

4. Discussion

In this work, we created a workflow for detecting differential gene expression in various transcriptomics datasets together with a metadata schema to make these datasets machine actionable. Our work enables a machine to automatically run this workflow in a federated manner on (privacy-sensitive) omics datasets for various rare diseases in the VP.

Acknowledgments

We would like to thank Luiz Bonino, Mark Wilkinson, Andra Waagmeester, Henriette Harmse, Sunil Rodger, Alberto Camara Ballesteros, Wolmar Akerstrom, Eric Prud'Hommeaux and Alexandra Tataru for helpful discussions. This initiative received funding from the EU Horizon 2020, grant agreements 825575 (EJP RD) and 824087 (EOSC-Life), and ELIXIR.

References

[1] M. D. Wilkinson et al., Comment: The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 1–9. doi:10.1038/sdata.2016.18.

[2] R. Albertoni et al., Data catalog vocabulary (DCAT) - version 2, 2020. URL: https://www.w3.org/TR/vocab-dcat-2/.