A Multi-omics Data Analysis Workflow Packaged as a FAIR Digital Object

Anna Niehues[‡], Casper de Visser[‡], Fiona A. Hagenbeek[§], Naama Karu[|], Alida S. D. Kindt[|], Purva Kulkarni[‡], René Pool[§], Dorret I. Boomsma[§], Jenny van Dongen[§], Alain J. van Gool[‡], Peter A. C. 't Hoen[‡]

‡ Radboud University Medical Center, Nijmegen, Netherlands
§ Vrije Universiteit Amsterdam, Amsterdam, Netherlands
| Leiden University, Leiden, Netherlands

Corresponding author: Anna Niehues (anna.niehues@radboudumc.nl)

Abstract

In current biomedical and complex trait research, increasing numbers of large molecular profiling (omics) data sets are being generated. At the same time, many studies fail to be reproduced (Baker 2016, Kim 2018). In order to improve study reproducibility and data reuse, including integration of data sets of different types and origins, it is imperative to work with omics data that is findable, accessible, interoperable, and reusable (FAIR, Wilkinson 2016) at the source. The data analysis, integration and stewardship pillar of the Netherlands X-omics Initiative aims to facilitate multi-omics research by providing tools to create, analyze and integrate FAIR omics data. We here report a joint activity of X-omics and the Netherlands Twin Register demonstrating the FAIRification of a multi-omics data set and the development of a FAIR multi-omics data analysis workflow.

The implementation of FAIR principles (Wilkinson 2016) can improve scientific transparency and facilitate data reuse. However, Kim (2018) showed in a case study that the availability of data and code are required but not sufficient to reproduce data analyses. They highlighted the importance of interoperable and open formats, and structured metadata. In order to increase research reproducibility on the data analysis level, additional practices such as version-control, code licensing, and documentation have been proposed. These include recommendations for FAIR software by the Netherlands eScience Center and the Dutch Data Archiving and Networked Services (DANS), and FAIR principles for research software proposed by the Research Data Alliance (Chue Hong 2022). Data analysis in biomedical research usually comprises multiple steps often resulting in complex data analysis workflows and requiring additional practices, such as containerization, to ensure transparency and reproducibility (Goble 2020, Stoudt 2021).

We apply these practices to a multi-omics data set that comprises genome-wide DNA methylation profiles, targeted metabolomics, and behavioral data of two cohorts that participated in the ACTION Biomarker Study (ACTION, Aggression in Children: Unraveling

gene-environment interplay to inform Treatment and InterventiON strategies, see consortium members in Suppl. material 1) (Boomsma 2015, Bartels 2018, Hagenbeek 2020, van Dongen 2021, Hagenbeek 2022). The ACTION-NTR cohort consists of twins that are either longitudinally concordant or discordant for childhood aggression. The ACTION-Curium-LUMC cohort consists of children referred to the Dutch LUMC Curium academic center for child and youth psychiatry. With the joint analysis of multi-omics data and behavioral data, we aim to identify substructures in the ACTION-NTR cohort and link them to aggressive behavior. First, the individuals are clustered using Similarity Network Fusion (SNF, Wang 2014), and latent feature dimensions are uncovered using different unsupervised methods including Multi-Omics Factor Analysis (MOFA) (Argelaguet 2018) and Multiple Correspondence Analysis (MCA, Lê 2008, Husson 2017). In a second step, we determine correlations between -omics and phenotype dimensions, and use them to explain the subgroups of individuals from the ACTION-NTR cohort. In order to validate the results, we project data of the ACTION-Curium-LUMC cohort onto the latent dimensions and determine if correlations between omics and phenotype data can be reproduced.

Integration of data across cohorts and across data types, requires interoperability. We applied different practices to make the data FAIR, including conversion of files to community-standard formats, and capturing experimental metadata using the ISA (Investigation, Study, Assay) metadata framework (Johnson 2021) and ontology-based annotations. All data analysis steps including pre-processing of different omics data types were implemented in either R or Python and combined in a modular Nextflow (Di Tommaso 2017) workflow, where the environment for each step is provided as a Singularity (Kurtzer 2017) container. The analysis workflow is packaged in a Research Object Crate (RO-Crate) (Soiland-Reyes 2022). The RO-Crate is a FAIR digital object that contains the Nextflow workflow including ontology-based annotations of each analysis step. Since omics data is considered to be potentially personally identifiable, the packaged workflow contains a minimal synthetic data set resembling the original data structure. Finally, the code is made available on GitHub and the workflow is registered at Workflowhub (Goble 2021). Since our Nextflow workflow is set up in a modular manner, the individual analysis steps can be reused in other workflows. We demonstrate this replicability by applying different sub-workflows to data from two different cohorts.

Keywords

Nextflow, ISA-API, RO-Crate, Singularity, metadata

Presenting author

Anna Niehues

Presented at

First International Conference on FAIR Digital Objects, poster

Acknowledgements

We acknowledge the ACTION Consortium and thank the participants of the ACTION Biomarker Study.

Funding program

The Netherlands X-omics Initiative is (partially) funded by the Dutch Research Council (NWO), project 184.034.019. "Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies" (ACTION) received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no 602768.

Conflicts of interest

References

- Argelaguet R, et al. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology 14: e8124. https://doi.org/10.15252/msb.20178124
- Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature 533: 452-454. https://doi.org/10.1038/533452a
- Bartels M, et al. (2018) Childhood aggression and the co-occurrence of behavioural and emotional problems: results across ages 3–16 years from multiple raters in six cohorts in the EU-ACTION project. European Child & Adolescent Psychiatry 27: 1105-1121. https://doi.org/10.1007/s00787-018-1169-1
- Boomsma D (2015) Aggression in children: unravelling the interplay of genes and environment through (epi)genetics and metabolomics. Journal of Pediatric and Neonatal Individualized Medicine (JPNIM) 4 (2): e040251. https://doi.org/10.7363/040251
- Chue Hong NP, et al. (2022) FAIR Principles for Research Software version 1.0. (FAIR4RS Principles v1.0). Research Data Alliance. https://doi.org/10.15497/RDA00068
- Di Tommaso P, et al. (2017) Nextflow enables reproducible computational workflows. Nature Biotechnology 35: 316-319. https://doi.org/10.1038/nbt.3820
- Goble C, et al. (2020) FAIR Computational Workflows. Data Intelligence 2 (1-2): 108-121. https://doi.org/10.1162/dint_a_00033
- Goble C, et al. (2021) Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo https://doi.org/10.5281/zenodo.4605653

- Hagenbeek F, et al. (2020) Urinary Amine and Organic Acid Metabolites Evaluated as Markers for Childhood Aggression: The ACTION Biomarker Study. Frontiers in Psychiatry 11: 165. https://doi.org/10.3389/fpsyt.2020.00165
- Hagenbeek F, et al. (2022) Heritability of Urinary Amines, Organic Acids, and Steroid Hormones in Children. Metabolites 12 (6): 474. https://doi.org/10.3390/metabo12060474
- Husson F, et al. (2017) Exploratory Multivariate Analysis by Example Using R. 2nd Edition. Chapman and Hall/CRC, New York, 262 pp. [ISBN 9780429225437] https://doi.org/10.1201/b21874
- Johnson D, et al. (2021) ISA API: An open platform for interoperable life science experimental metadata. GigaScience 10 (9): giab060. https://doi.org/10.1093/gigascience/giab060
- Kim Y, et al. (2018) Experimenting with reproducibility: a case study of robustness in bioinformatics. GigaScience 7 (7): giy077. https://doi.org/10.1093/gigascience/giy077
- Kurtzer G, et al. (2017) Singularity: Scientific containers for mobility of compute. PLOS ONE 12 (5): e0177459. https://doi.org/10.1371/journal.pone.0177459
- Lê S, et al. (2008) FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software 25 (1): 1-18. https://doi.org/10.18637/jss.v025.i01
- Soiland-Reyes S, et al. (2022) Packaging research artefacts with RO-Crate. Data Science1-42. https://doi.org/10.3233/ds-210053
- Stoudt S, et al. (2021) Principles for data analysis workflows. PLOS Computational Biology 17 (3): e1008770. https://doi.org/10.1371/journal.pcbi.1008770
- van Dongen J, et al. (2021) DNA methylation signatures of aggression and closely related constructs: A meta-analysis of epigenome-wide studies across the lifespan. Molecular Psychiatry 26 (6): 2148-2162. https://doi.org/10.1038/s41380-020-00987-x
- Wang B, et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 11 (3): 333-337. https://doi.org/10.1038/nmeth.2810
- Wilkinson M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18

Supplementary material

Suppl. material 1: Members of the ACTION Consortium

Authors: ACTION Consortium
Data type: consortium members
Download file (87.73 kb)