GigaScience

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-23-00162R2
Full Title:	A Multi-omics Data Analysis Workflow Packaged as a FAIR Digital Object
Article Type:	Technical Note
Funding Information:	
Abstract:	Background. Applying good data management and FAIR data principles (Findable, Accessible, Interoperable, and Reusable) in research projects can help disentangle knowledge discovery, study result reproducibility, and data reuse in future studies. Based on the concepts of the original FAIR principles for research data, FAIR principles for research software were recently proposed. FAIR Digital Objects enable discovery and reuse of Research Objects, including computational workflows for both humans and machines. Practical examples can help promote the adoption of FAIR practices for computational workflows in the research community. We developed a multi-omics data analysis workflow implementing FAIR practices to share it as a FAIR Digital Object.Findings. We conducted a case study investigating shared patterns between multi-omics data and childhood externalizing behavior. The analysis workflow was implemented as a modular pipeline in the workflow manager Nextflow, including containers with software dependencies. We adhered to software development practices like version control, documentation, and licensing. Finally, the workflow was described with rich semantic metadata, packaged as a Research Object Crate, and shared via WorkflowHub.Conclusions. Along with the packaged multi-omics data analysis workflow, we share our experiences adopting various FAIR practices and creating a FAIR Digital Object. We hope our experiences can help other researchers who develop omics data analysis workflows to turn FAIR principles into practice.
Corresponding Author:	Peter-Bram t' Hoen NETHERLANDS
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Anna Niehues
First Author Secondary Information:	
Order of Authors:	Anna Niehues
	Casper de Visser
	Fiona A. Hagenbeek
	Purva Kulkarni
	Rene Pool
	Naama Karu
	Alida S. D. Kindt
	Gurnoor Singh
	Robert R. J. M. Vermeiren
	Dorret I. Boomsma
	Jenny van Dongen
	Peter A. C. 't Hoen

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	Alain J. van Gool
Order of Authors Secondary Information:	
Response to Reviewers:	With all requested changes in the email from Friday December 8
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using	Yes

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

LaTeX - Main Document (TeX file)
Click here to access/download;LaTeX - Main Document (TeX file);main.tex

```
This is pdfTeX, Version 3.141592653-2.6-1.40.24 (TeX Live 2022)
(preloaded format=pdflatex 2023.3.8)  9 DEC 2023 10:16
entering extended mode
 restricted \write18 enabled.
 %&-line parsing enabled.
**main.tex
(./main.tex
LaTeX2e <2022-11-01> patch level 1
L3 programming layer <2023-02-22>

! LaTeX Error: File `oup-contemporary.cls' not found.

Type X to quit or <RETURN> to proceed,
or enter new name. (Default extension: cls)

Enter file name:
! Emergency stop.
<read *>

l.11 ^^M

*** (cannot \read from terminal in nonstop modes)


Here is how much of TeX's memory you used:
 23 strings out of 476024
 503 string characters out of 5794017
 1849382 words of memory out of 5000000
 20571 multiletter control sequences out of 15000+600000
 512287 words of font info for 32 fonts, out of 8000000 for 9000
 1141 hyphenation exceptions out of 8191
 19i,0n,29p,95b,17s stack positions out of
10000i,1000n,20000p,200000b,200000s
!  ==> Fatal error occurred, no output PDF file produced!
```

Manuscript

Click here to access/download;Manuscript;Manuscript_2023-12-08.pdf

OXFORD

(GIGA)n SCIENCE

GigaScience, 2023, 1–12

doi: xx.xxxx/xxxx
Manuscript in Preparation
Technical Note

TECHNICAL NOTE

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

Anna Niehues (0000-0002-9839-5439)[1,2,†], Casper de Visser (0000-0002-2812-5898)[1], Fiona A. Hagenbeek (0000-0002-8773-0430)[3,4,‡], Purva Kulkarni (0000-0002-4681-4582)[1,2,5], René Pool (0000-0001-5579-0933)[3,4], Naama Karu (0000-0001-8005-0726)[6,§], Alida S. D. Kindt (0000-0001-6551-6030)[6], Gurnoor Singh (0000-0003-1615-4197)[1], Robert R. J. M. Vermeiren (0000-0002-8673-2207)[7], Dorret I. Boomsma (0000-0002-7099-7972)[3,4,8], Jenny van Dongen (0000-0003-2063-8741)[3,4,8], Peter A. C. 't Hoen (0000-0003-4450-3112)[1,*] and Alain J. van Gool (0000-0003-0010-5286)[2]

[1]Department of Medical BioSciences, Radboud university medical center, Nijmegen, The Netherlands and [2]Translational Metabolic Laboratory, Department of Laboratory Medicine, Radboud university medical center, Nijmegen, The Netherlands and [3]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands and [4]Amsterdam Public Health research institute, Amsterdam, The Netherlands and [5]Department of Human Genetics, Radboud university medical center, Nijmegen, Netherlands and [6]Metabolomics and Analytics Centre, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands and [7]Department of Child and Adolescent Psychiatry, LUMC-Curium, Leiden University Medical Center, Leiden, The Netherlands and [8]Amsterdam Reproduction & Development (AR&D) research institute, Amsterdam, The Netherlands

[*] Peter-Bram.tHoen@radboudumc.nl
[†]Current affiliation: EATRIS ERIC, European Infrastructure for Translational Medicine, Amsterdam, The Netherlands
[‡]Current affiliation: Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland
[§]Current affiliation: TIMACS, Hobart, Tasmania, Australia

Abstract

Background Applying good data management and FAIR data principles (Findable, Accessible, Interoperable, and Reusable) in research projects can help disentangle knowledge discovery, study result reproducibility, and data reuse in future studies. Based on the concepts of the original FAIR principles for research data, FAIR principles for research software were recently proposed. FAIR Digital Objects enable discovery and reuse of Research Objects, including computational workflows for both humans and machines. Practical examples can help promote the adoption of FAIR practices for computational workflows in the research community. We developed a multi-omics data analysis workflow implementing FAIR practices to share it as a FAIR Digital Object. **Findings** We conducted a case study investigating shared patterns between multi-omics data and childhood externalizing behavior. The analysis workflow was implemented as a modular pipeline in the workflow manager Nextflow, including containers with software dependencies. We adhered to software development practices like version control, documentation, and licensing. Finally, the workflow was described with rich semantic metadata, packaged as a Research Object Crate, and shared via WorkflowHub. **Conclusions** Along with the packaged multi-omics data analysis workflow, we share our experiences adopting various FAIR practices and creating a FAIR Digital Object. We hope our experiences can help other researchers who develop omics data analysis workflows to turn FAIR principles into practice.

Key words: multi-omics; workflow; metadata; FAIR; RO-Crate; FDO

Key Points

- The FAIR4RS principles provide guidelines to enhance the discovery and reuse of research software.
- FAIR Digital Objects support Findability, Accessibility, Interoperability, and Reusability by both humans and machines.
- We here demonstrate an implementation of a multi-omics data analysis workflow and share it as a FAIR Digital Object.

Background

The FAIR principles for research data [1] were proposed to guide researchers to create research data that is findable, accessible, interoperable, and reusable (FAIR). Since these guidelines aim to enable researchers handling and navigating through the rapidly increasing amounts of data, special emphasis was put on concepts to make data not only usable by humans but also machine-actionable. In the past years, various standards [2, 3] and implementations [4, 5, 6, 7] of the FAIR principles have been introduced, and it has been demonstrated that FAIR data is beneficial to research and patients [8, 9, 10]. Reuse of research data and reproducibility of research results [11] are facilitated by good data provenance and this requires not only the data but also the data processing and analysis workflows to be FAIR. Consequently, guidelines and practices for FAIR research software have been proposed [12, 13, 14] (see Table 1) and the special case of computational workflows has been discussed [15, 16]. These guidelines aim to increase reproducibility not only at the experimental level but also at the data analysis level. It has been shown that the availability of data and code alone is not sufficient. They both need to be provided in an open and interoperable format and described by metadata [17].

Several practices recommended for research software development originate from general software engineering practices [12, 15, 19] which include version control, documentation, and licensing. Version control of source code facilitates collaborative development and monitoring changes [13]. Additionally, making the code publicly available on dedicated software repositories that support version control such as Git-based [20] GitHub [21], GitLab [22], or BitBucket [23] contributes to findability [24], accessibility [12], and reusability [13]. The documentation of research software includes multiple levels. First, a comprehensive end user documentation and usage examples enable reusability by other researchers [17, 25, 26, 24]. It should also include the documentation of workflow parameters [17, 16]. Second, source code documentation enables other developers to understand and build upon the software [17]. Documentation of code changes via a version control system helps document the development process [25, 19] and documentation of dependencies are prerequisite for software interoperability [24] and reusability [18]. Adding a clear and machine-readable [16] license is essential to allow for software reuse. It is recommended to choose a widely used and preferably open source license that is compatible with licenses of the dependencies [12, 25, 19, 24, 13, 14, 18]. Examples of open-source licenses with few restrictions are the Apache License 2.0 [27] and the MIT License [28].

There are differences between research software that implements a specific method as a standalone tool or a software library and complex analysis workflows [16]. Computational analysis workflows can comprise numerous steps that are integrated into pipelines [16] and are often developed in a specific project [29, 19]. With a multitude of analysis steps being combined into complex workflows, the documentation of the individual analyses and their dependencies can become challenging. To facilitate the automation of analysis tasks and their documentation, workflows can be described using workflow management systems such as Nextflow [30] or Snakemake [31]. Workflow managers that support the creation of reusable modules can help reduce complexity and promote the reuse of workflows or workflow modules [15, 16, 32]. Additionally, notebooks can apply the concept of literate programming and are a useful tool to add human-readable documentation next to code blocks [19]. Interoperability and reusability of workflows can be achieved using portable software containers such as Apptainer/Singularity [33] or Docker [34] that capture the runtime environment of a workflow or a workflow module [15, 35, 16, 26].

Computational workflows can be regarded as digital objects. The concept of FAIR Digital Objects (FDOs) was introduced to make digital objects fully FAIR [36]. FDOs comprise, among others, the digital object, a persistent identifier (PID), and metadata (title, authors, licenses, etc.) describing the object. The RO-Crate approach was specified to package digital research artefacts or Research Objects (RO) such as computational workflows [37]. The RO-Crate contains a PID that links to an RO, which is described by a structured JSON-LD RO-Crate metadata file. It contains all contextual and non-contextual related data to re-run the workflow. In case the actual data cannot be publicly shared due to privacy reasons, synthetic data can complement analysis workflows to demonstrate the computational procedure [16, 38]. To make an RO-Crate findable, it needs to be registered at a registry such as WorkflowHub [39, 40]. The WorkflowHub RO-Crate represents an approach to implementing the FDO concept [41, 42].

We here demonstrate the development of a FAIR Digital Object comprising a computational workflow that analyzes and integrates multi-omics and phenotype data and is associated with rich human and machine-readable metadata.

Findings

Workflow Implementation

To develop a reusable workflow, our input data and intermediate files were largely based on open and widely-used formats or community standards. For the metabolomics data and metadata, we adopted practices of the MetaboLights database [43] of the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL). Metabolite levels and annotations are reported in metabolite annotation/assignment files (MAF). The experimental metadata for omics measurements is reported using the Investigation/Study/Assay (ISA) metadata framework [44]. We employed Jupyter [45] and the Python ISA API [46] to create ISA-Tab and ISA-JSON files [47]. For machine-readable descriptions of the experiments, ontology terms were used. Ontologies are standardized taxonomies of entities of a specific subject (domain) including definitions of relationships between these entities. Ontology terms refer to these entities [48]. Based on recommended standards from FAIRgenomes [3] and Metabolights [43], we preferably employed the following ontologies: National Cancer Institute

Compiled on: December 8, 2023.
Draft manuscript prepared by the author.

Table 1. Overview of recommended FAIR practices for research data and software.

FAIR guiding principles [1]	Open source software recommendations [12]	Recommendations for FAIR software [13]	FAIR principles for research software [14, 18]
Findable			F: Software, and its associated metadata, is easy for both humans and machines to find.
F1: (Meta) data are assigned globally unique and persistent identifiers			F1. Software is assigned a globally unique and persistent identifier. F1.1. Components of the software representing levels of granularity are assigned distinct identifiers. F1.2. Different versions of the software are assigned distinct identifiers.
F2: Data are described with rich metadata			F2. Software is described with rich metadata.
F3: Metadata clearly and explicitly include the identifier of the data they describe			F3. Metadata clearly and explicitly include the identifier of the software they describe.
F4: (Meta)data are registered or indexed in a searchable resource	R2. Make software easy to discover by providing software metadata via a popular community registry	#3 Register your code in a community registry	F4. Metadata are FAIR, searchable and indexable.
Accessible	R1. Make source code publicly accessible from day one	#1 Use a publicly accessible repository with version control	A: Software, and its metadata, is retrievable via standardized protocols.
A1: (Meta)data are retrievable by their identifier using a standardised communication protocol			A1. Software is retrievable by its identifier using a standardized communications protocol.
A1.1: The protocol is open, free and universally implementable			A1.1. The protocol is open, free, and universally implementable.
A1.2: The protocol allows for an authentication and authorisation procedure where necessary			A1.2. The protocol allows for an authentication and authorization procedure, where necessary.
A2: Metadata should be accessible even when the data is no longer available			A2. Metadata are accessible, even when the software is no longer available.
Interoperable			I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation			I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards.
I2: (Meta)data use vocabularies that follow the FAIR principles			
I3: (Meta)data include qualified references to other (meta)data			I2. Software includes qualified references to other objects.
Reusable	R4. Define clear and transparent contribution, governance and communication processes	#4 Enable citation of the software; #5 Use a software quality checklist	R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).
R1: (Meta)data are richly described with a plurality of accurate and relevant attributes			R1. Software is described with a plurality of accurate and relevant attributes.
R1.1: (Meta)data are released with a clear and accessible data usage license	R3. Adopt a license and comply with the licence of third-party dependencies	#2 Add a license	R1.1. Software is given a clear and accessible license.
R1.2: (Meta)data are associated with detailed provenance			R1.2. Software is associated with detailed provenance. R2. Software includes qualified references to other software.
R1.3: (Meta)data meet domain-relevant community standards			R3. Software meets domain-relevant community standards.

Thesaurus (NCIT) [49], Experimental Factor Ontology (EFO) [50], Ontology for Biomedical Investigations (OBI) [51], Metabolomics Standards Initiative Ontology (MSIO) [52], Chemical Methods Ontology (CHMO) [53] and Chemical Entities of Biological Interest (ChEBI) [54]. The DNA methylation levels and associated metadata, behavioral data, and additional information about phenotypes or technical and biological covariates are stored as comma-separated values (CSV) files. This allows our computational workflow to be easily reusable and adaptable for other data sets. The workflow documentation [55] describes all input files used in the workflow and provides human-readable descriptions of every step of the workflow processing and integrating individual input data types. Each of these analysis steps is implemented in Python or R, and added as a module to the workflow. We employ Jupyter and R notebooks for implementing downstream analyses and visualization of results. We chose Nextflow as our workflow management system, since it allows flexible development, can be run on different platforms, supports containers, is well documented, and is already widely adopted by the bioinformatics community [32]. Each module of the workflow is provided with their own Docker container to ensure portability and eliminate the need for local software installations.

Finally, the Nextflow workflow is packaged as an RO-Crate. In addition to the workflow and a synthetic data set, it contains a structured metadata file with machine-readable descriptions of input files and analysis steps (`ro-crate-metadata.json`). We preferably used EDAM—Ontology of bioscientific data analysis and data management [56] as it is recommended for workflow RO-Crates [37]. For terms that were not available in EDAM, alternative ontologies such as NCIT [49], OBI [51], or the Semanticscience Integrated Ontology (SIO) [57] were used. We employed the Python package ro-crate-py [58] to create the RO-Crate metadata file. The RO-Crate further contains an image with an overview of the analysis steps. For findability, the packaged workflow (see Figure 2) is registered on WorkflowHub [39] and provided with a Digital Object Identifier (DOI) [https://doi.org/10.48546/workflowhub.workflow.402.8].

Case Study

Our workflow was developed to analyze and integrate DNA methylation and urine metabolomics profiles with behavioral data originating from the ACTION Biomarker Study (ACTION, Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies) [59, 60, 61] (see 'Case Study' in the 'Methods' section on page 6). Within ACTION, urine and buccal-cell samples were collected in a twin cohort from the Netherlands Twin Register (NTR), and in a cohort of children referred to an academic center for child and youth psychiatry in the Netherlands (LUMC-Curium). These children were also characterized for behavioral problems and here we look at externalizing problems. We purposely selected a case of complex human behavioral phenotype that is typically not caused by a single well-defined molecular defect, but originates from changes in multiple factors, and as such would benefit from a multi-omics analysis. Since we consider this data to be potentially personally identifiable information, we share a synthetic data set to demonstrate the workflow. The goal of the analysis is the identification of substructures in the multi-omics data and to determine if they correlate with behavioral data (see 'Unsupervised Data Analysis' on page 6). A team comprising members of the Netherlands X-omics Initiative [62] in collaboration with the Netherlands Twin Register (NTR) [63] developed the computational workflow. To uncover possible relationships between the multi-omics data and the behavioral data, we applied different unsupervised data-driven methods followed by downstream analyses including determining the effect of possible confounding factors sex and age. An overview of the main analysis steps is shown in Figure 1. An overview of data dimensions and types during different steps of the workflow is provided in [Additional File 10].

To identify underlying patterns in childhood externalizing behavior, we applied Multiple Correspondence Analysis (MCA) [64, 65] to the parent-rated responses on the externalizing behavior items of the Child Behavior Checklist (CBCL) of the Achenbach System of Empirically Based Assessment (ASEBA) [66] in both cohorts. In NTR participants, the first three MCA dimensions jointly explain 30% of the variation in 26 externalizing behavior items of the ASEBA CBCL [see Additional File 1]. Additional dimensions each explain <5% of the variation. The presence rather than the absence of externalizing behaviors characterized all of the first three dimensions, which reflects the answer options to items (a problem behavior is not present, a little, or a lot). Variables that contributed most to the first dimension, which explained 16% of the variation, represent temperamental behavior (frequent temper tantrums, stubbornness, screaming, and arguing). Variables contributing to the second dimension, which explained 9% of the variation, represent hostile aggressive behaviors (frequent vandalism, bullying, and cruelty). In LUMC-Curium participants, the first two MCA dimensions suffice to explain 30% of the variation in 18 items of the ASEBA CBCL [see Additional File 2]. Similar to NTR, these first dimensions in LUMC-Curium are characterized by the presence of aggressive behaviors.

We applied Multi-Omics Factor Analysis (MOFA) [67] in both cohorts to obtain ten factors to describe the buccal DNA methylation (Illumina EPIC array) and urine metabolomics data. For this analysis, we selected the top 10% most variable probes from DNA methylation data. Cumulatively, the ten factors explained 22.5% and 74.9% of variation in the DNA methylation data and 0.001% and 1.89% in the urine metabolomics data in NTR [see Additional File 3] and LUMC-Curium [see Additional File 4], respectively. We observed no evidence that any of the factors captured sources of variation in both the DNA methylation and urine metabolomics data in NTR and LUMC-Curium. Particularly factors 1 and 2 in NTR and factor 1 in LUMC-Curium were specific to the DNA methylation data. To help elucidate the etiology of the ten MOFA factors, we selected for each factor the top 100 CpGs with the largest weights and performed enrichment analyses within the Epigenome-Wide Association Study (EWAS) atlas [68]. Multiple factors in both cohorts [see Additional File 5 for ACTION-NTR and Additional Files 6 for LUMC-Curium cohort] showed enrichment of CpGs associated with glucocorticoid exposure (i.e., administration of corticosteroid medication [69]), CpGs associated with ageing, and CpGs associated with immune-related traits, such as psoriasis. Apart from these robustly enriched traits, additional significant enrichments were found but were often based on ≤ 5 overlapping CpGs between the factor results and the original studies. A limitation of the enrichment analysis is that the majority of previous EWAS studies included in this analysis were conducted on blood samples from adult populations with the Illumina 450K BeadChip. In the factor weights for metabolites, we observe that for both NTR [Additional File 3] and LUMC-Curium [Additional File 4] many of the factors are characterized by only one or few metabolites. We note that in both cohorts the factors explained only a small amount of variation in the metabolomics data. To investigate whether the omics factors were associated with behavioral dimensions (MCA), we ran gee models adjusting for relatedness in NTR, and correlation analyses in Curium [see Additional File 3 for ACTION-NTR and Additional Files 4 for LUMC-Curium cohort]. None of the omics factors were significantly associated with the behavioral dimensions in NTR or LUMC-Curium, nor did we observe significant associations of sex-and-age-specific T-scores for aggressive behavior with the omics factors. In previous multi-omics analyses of high versus low levels of childhood aggression [70] and Attention-Deficit/Hyperactivity Disorder [71] we applied supervised analyses in these cohorts while applying unsupervised analyses here. In these previous supervised analyses, where we also included an additional omics layer—polygenic scores—we found that although multi-omics models had low predictive value, they revealed some

connections of omics traits with externalizing problems, that suggested biological plausibility.

We also constructed integrated similarity networks with Similarity Network Fusion (SNF) [72] to identify subgroups of individuals based on omics data. In both NTR and LUMC-Curium, we defined integrated similarity networks based on two and four clusters. The two clusters in NTR were characterized by differences in age whereas the two clusters in LUMC-Curium were characterized by differences in the proportion of boys and girls. To investigate whether the omics clusters were associated with externalizing behavior, we compared the behavioral dimension scores from MCA between children in the different clusters. In both NTR and LUMC-Curium, we observed no significant differences in the behavioral dimensions across the two omics clusters after correction for multiple testing [see Additional File 7 for NTR and Additional File 8 for LUMC-Curium cohort]. Similarly, no differences in behavioral dimensions were observed between the four omics clusters in NTR, however, in LUMC-Curium, behavioral dimension 6 differed significantly between the four omics clusters. In LUMC-Curium, dimension 6 explained 3.9% of the variance in childhood externalizing behavior and the strongest contributors to this dimension comprised higher frequencies of parent-rated tendencies to be suspicious and loud. Such forms of direct aggressive behavior, particularly physical aggressive behavior, are common in early childhood in both boys and girls, and while overall levels of aggression decline with age and are roughly similar for boys and girls [73], boys are more likely to engage in direct and physical forms of aggression by age 11 [74]. Thus, this finding aligns with the observation that the two omics clusters differ in the proportion of males and females and in the age composition.

Our data-driven approach to identifying possible relationships between multi-omics and behavioral data did not reveal significant findings that could not also be explained by potential confounding factors sex or age. Since we here focused on latent dimensions representing the largest variations between individuals (after correcting for known confounders), it is possible that relationships between omics and aggressive behavior can be found in lower dimensions that reflect only a small amount of variation in the cohorts. However, including more (latent) variables in the correlation analysis will also increase the chance of false positive findings.

Discussion

In this collaborative research project, partners from the Netherlands X-omics Initiative co-developed a workflow to analyze a complex multimodal data set. Developing workflows with partners across multiple institutions can pose a challenge and we experienced that a secure shared computing environment was key to the success of this project. Additionally, practices aiming to increase FAIRness of the shared workflow such as version control with Git and a modular workflow structure allowed for transparent and target-oriented workflow development. Therefore, while the use of technologies like Git or workflow management systems might require initial training of researchers, we believe this to be worthwhile not only for future reuse, but also during workflow development.

To make the workflow findable, we registered it in WorkflowHub [39] which is part of the European Open Science Cloud (EOSC) [75]. Since this was the first workflow we registered in WorkflowHub, we profited from its documentation and active community. The registry allowed us to assign a globally unique and persistent identifier to the workflow [76] and its versions. Metadata could be added using the open RO-Crate standard and is searchable in the registry. The workflow page [76] links to the publicly accessible and version-controlled source code on GitHub [77].

Several FAIR practices for workflows include existing best practices of software development, for example, version control and good documentation. Adoption of these practices, along with the use of workflow managers and software containers, aims to contribute to better interoperability, reusability and reproducibility of analysis workflows and research results. While we experienced the adoption of these technologies to be straightforward, fully FAIR, and especially interoperable, data or software requires also machine-understandable semantic metadata. Specifications like the ISA metadata framework and RO-Crate allow ontology-based annotations of omics experiments and analysis workflows, respectively. Our choice of ontologies was mainly guided by the documented submission requirements or recommendations provided by services such as the MetaboLights archive or WorkflowHub. However, when recommended ontologies do not comprise suitable terms, choosing appropriate ones from ontologies can be challenging. For example, no exact match to the generic term *sample collection* that is part of the ISA schema can be found in any ontology available in EBI's Ontology Lookup Serivce (OLS) [78]. To describe workflow steps in RO-Crate with *unsupervised learning* we had to employ the eNanoMapper Ontology [79] as no matching term was available in the recommended EDAM ontology. Consequently, we recognize the importance of teams dedicated to ontology curation, active user communities, and training of researchers in using semantic technologies. This is especially important for multi-omics research that spans multiple research domains.

While machine-actionability supported by standardized metadata is relevant for interoperability, the workflow also needs to be usable and reusable by humans. We added software containers that are referenced by the workflow metadata. They enable portability and thereby reusability. A user documentation was added to help understand the workflow steps and facilitate reuse. Enabling richer workflow annotation with RO-Crate in combination with additional tooling that enable automated generation of user documentation could potentially reduce the efforts of manual workflow documentation in the future.

For reproducibility of research results, it is essential that data are shared along with the workflow. However, privacy regulations prohibit sharing of potentially personally identifiable data such as omics measurements or clinical information. To demonstrate the functionality of the workflow, we shared a synthetic data set that emulates the structure of the case study data set. Current developments in the areas of federated data storage and analysis such as Federated European Genome-phenome Archive (EGA) [80] and the Personal Health Train [81] have the potential to allow fully FAIR and reproducible data analysis workflows while maintaining privacy regulation compliance.

Implementing these FAIR practices required us to use various tools, some of which we used for the first time. While this required some time and openness to getting familiarized with these tools, we experienced that the tools were generally well documented and could quickly be adopted. Open online resources such the ELIXIR's [82] community-driven FAIR Cookbook [83, 84] provide guides and examples that can help researchers implement FAIR practices. Existing Python libraries such as the ISA API[46] and ro-crate-py [58] were very useful when implementing metadata standards as they can help ensure compliance with the standards as well as automating creation of metadata files. However, it would have been useful if more use cases implementing FAIR practices for scientific computational workflows were available as examples or tutorials. We experienced that implementing FAIR practices from the start helped us create a transparent multi-omics analysis workflow. Additionally, we are convinced that FAIR workflows are key not only to reproducible but also efficient research as workflows or sub-workflows can be reused in new contexts, thereby saving time. Therefore, we hope our experiences help other researchers who develop multi-omics data analysis workflows choosing and implementing practices that makes their research more FAIR.

Data and Methods

Case Study

Our case study comprises data from two cohorts that took part in the ACTION Biomarker Study (ACTION, Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies) [59, 60, 61]. The ACTION Biomarker Study collected buccal DNA samples for large-scale genome-wide and epigenome-wide association studies [85, 86] and first-morning urine samples to investigate the association of urine biomarkers and metabolites with childhood aggression [61]. These urine and buccal-cell samples were collected in a twin cohort from the Netherlands Twin Register (NTR) [87], where twin pairs were selected on their longitudinal concordance or discordance for childhood aggression, and in a cohort of children referred to an academic center for child and youth psychiatry in the Netherlands (LUMC-Curium). The DNA methylation, genotype, metabolomics, and behavioural data from these cohorts were previously used for multi-omics analyses of aggressive behaviour [70] and Attention Deficit Hyperactivity Disorder (ADHD) [71]. Detailed information on the study populations and study protocol is available at protocols.io [88].

Data

Genome-wide DNA methylation data in buccal DNA samples were measured on the Infinium MethylationEPIC BeadChip kit (Illumina, San Diego, CA, USA [89]) by the Human Genotyping Facility (HuGeF) of ErasmusMC (the Netherlands; [90]). The ZymoResearch EZ DNA Methylation kit (Zymo Research Corp, Irvine, CA, USA) was used for bisulfite treatment of 500 ng of genomic DNA obtained from buccal swabs. The Infinium HD Methylation Assay was performed according to the manufacturer's specification. Good Biomarker Sciences Leiden (Leiden, the Netherlands) measured the specific gravity (by refractomertry), levels of creatinine (by colorimertry), blood traces, markers of leukocytes, proteins, glucose and nitrites (the latter five by dipstick) of each urine sample. The Metabolomics Facility of the University of Leiden (Leiden, the Netherlands) quantified urine metabolites using three platforms: a liquid chromatography mass spectrometry (LC-MS) platform targeting amines (66 biomarkers), an LC-MS platform targeting steroid hormones (13 biomarkers), and a gas chromatography (GC)-MS platform targeting organic acids (21 biomarkers). Behavioral data comprises the 115 items of the Dutch version of the Achenbach System of Empirically Based Assessment (ASEBA) Child Behavior Checklist (CBCL) for school-aged children (6–18 years) [66]. For participants of the NTR cohort we used the mother-rated CBCL as completed at the time of biological sample collection, and for participants of the LUMC-Curium cohort we used the parent-rated (90% mother ratings) CBCL as completed in a 6-month window surrounding the biological sample collection. Again, details on the data generation are available in Ref. [88].

Synthetic Data and Metadata

The purpose of the synthetic data set that is part of the RO-Crate is to demonstrate how the workflow can be run. It resembles the structure of the files of the cohort data. The values were randomly sampled from the observed values in the NTR cohort without preserving any correlations. While creation of ISA metadata is not part of this workflow, we share the Jupyter notebook employing the Python ISA API [46] that was used to create the metadata for the synthetic data set [47].

Data Processing

To ensure the urine sample metabolic integrity and to minimize bias contributed by health conditions, we excluded samples from the metabolomics data from *(1)* subjects who have started menstruating, *(2)* subjects where the time between urine sample collection and storing in the freezer was > 2 hours, *(3)* subjects where severe violations to the sampling protocol occurred (e.g. not putting a lid on the container), *(4)* subjects where the leukocyte count was above trace, *(5)* subjects where the nitrites level was "positive high", *(6)* subjects where the protein level was > 0.3, *(7)* subjects with glucose levels above trace, *(8)* subjects with blood levels above trace, *(9)* subjects having the flu, *(10)* subjects reporting inflammation, *(11)* subjects reporting vomiting, *(12)* subjects reporting abdominal pain and *(12)* subjects reporting general health problems. Note that the above criteria 4–8 are based on the dipstick marker estimation performed separately from the metabolomics measurements on the same samples [88], while the other criteria are based on questionnaire data at the time of sampling.

The metabolomics features were filtered based on missing values. Missing values were reported for cases where the metabolite concentration is below the limit of quantification. Samples and metabolites with 15% or more of missing values were discarded. Sample-wise normalization to correct for urine concentration was conducted by adjusting metabolite intensities to the sample creatinine levels [88]. This was followed by metabolite-wise Pareto scaling [91] to statistically account for large differences in reported values.

Quality Control (QC) and normalization of the DNA methylation array data have been previously described [85] and were carried out with a pipeline developed by the Biobank-based Integrative Omics Study (BIOS) consortium [92]. From the 787,711 autosomal methylation probes that survived QC, the top 10% most variable probes were included in the analyses. Cellular proportions of buccal samples were predicted with Hierarchical Epigenetic Dissection of Intra-Sample-Heterogeneity (HepiDISH) with the RPC method (reduced partial correlation), as described Zheng et al. [93] and implemented in the R/Bioconductor package EpiDISH. Median imputation was carried out on the epigenetics data. Residual methylation levels were obtained by regressing the effects of percentages of epithelial and natural killer cells, EPIC array row, and bisulfite sample plate, from the methylation beta-values.

Missing values in the externalizing behavior items were imputed with the non-parametric random forests method from the R library missForest (1.4) [94].

Unsupervised Data Analysis

Each cohort was analyzed separately. We applied Multi-Omics Factor Analysis (MOFA) using the R/Bioconductor library MOFA2 (1.3.4) [67, 95] to obtain factors for the buccal DNA methylation and urine metabolomics data and applied Multiple Correspondence Analysis (MCA) [65] using the R library FactoMineR (2.4) [64] to obtain factors for the behavioral data.

To identify subgroups of individuals based on their buccal DNA methylation and urine metabolomics data we constructed integrated similarity networks with Similarity Network Fusion (SNF) [72]. The optimal numbers of clusters were determined using a built-in function of the Python library SNFpy [96] that uses the eigengap method [97] to find the optimal number of clusters. SNF first constructs sample similarity networks for each available data type and then fuses these into a single network comprising both the shared and unique information from each data type. The final fused network thus captures how each data type contributes to the similarity amongst the samples. We tested whether the behavioral dimension scores from MCA differ between children in the different SNF clusters, using Mann-Whitney U tests (two clusters) or Kruskal-Wallis tests (four clusters) in the Curium cohort, and with generalized estimation equation (GEE) models (with cluster as predictor and behavioral dimension score as outcome) in NTR.

We determined correlations amongst the obtained factors capturing the omics and behavioral data, respectively, using Spear-

man's rank correlation, and additionally in the NTR cohort using GEE models. All GEE models were fitted with the R package GEE, with the following specifications: Gaussian link function (for continuous data), 100 iterations, and the 'exchangeable' option to account for the correlations in twin pairs. Statistical test were adjusted for multiple testing using false discovery rate (FDR) [98].

Workflow Implementation

Availability of source code and requirements

- Project name: X-omics ACTION demonstrator multi-omics analysis workflow
- Project home page: [77]
- Operating system(s): Platform independent
- Programming language: Python, R
- Other requirements: Nextflow (22.04.0), Docker (19.03.1), Singularity (3.8.0)
- License: MIT
- SciCrunch: RRID:SCR_024719

Additional Files

- File name: Additional File 1
- File format: .html
- Title: Multiple Correspondence Analysis of CBCL behavioral data
- Description: Overview of externalizing behavior items of the Child Behavior Checklist (CBCL) of the Achenbach System of Empirically Based Assessment (ASEBA) in the ACTION-NTR cohort before and after imputation of missing values using random forests and results and visualizations of Multiple Correspondence Analysis (MCA).

- File name: Additional File 2
- File format: .html
- Title: Multiple Correspondence Analysis of CBCL behavioral data
- Description: Overview of externalizing behavior items of the Child Behavior Checklist (CBCL) of the Achenbach System of Empirically Based Assessment (ASEBA) in the LUMC-Curium cohort before and after imputation of missing values using random forests and results and visualizations of Multiple Correspondence Analysis (MCA).

- File name: Additional File 3
- File format: .html
- Title: MOFA downstream analysis report
- Description: Visualizations of Multi-omics Factor Analysis (MOFA) of buccal DNA methylation (Illumina EPIC array) and urine metabolomics data of the ACTION-NTR cohort. Also, associations between the MOFA factors and phenotypic data are tested with GEE models.

- File name: Additional File 4
- File format: .html
- Title: MOFA downstream analysis report
- Description: Visualizations of Multi-omics Factor Analysis (MOFA) of buccal DNA methylation (Illumina EPIC array) and urine metabolomics data of the LUMC-Curium cohort.

- File name: Additional File 5
- File format: .xlsx
- Title: EWAS atlas enrichment analysis
- Description: Enriched traits for CpGs with top 100 largest weights of ACTION-NTR MOFA factors 1-10.

- File name: Additional File 6
- File format: .xlsx
- Title: EWAS atlas enrichment analysis
- Description: Enriched traits for CpGs with top 100 largest weights of LUMC-Curium MOFA factors 1-10.

- File name: Additional File 7
- File format: .html
- Title: Similarity Network Fusion downstream analysis
- Description: Visualizations of Similarity Network Fusions (SNF) and subsequent spectral clustering of buccal DNA methylation (Illumina EPIC array) and urine metabolomics data of the ACTION-NTR cohort.

- File name: Additional File 8
- File format: .html
- Title: Similarity Network Fusion downstream analysis
- Description: Visualizations of Similarity Network Fusions (SNF) and subsequent spectral clustering of buccal DNA methylation (Illumina EPIC array) and urine metabolomics data of the LUMC-Curium cohort.

- File name: Additional File 9
- File format: .html
- Title: Similarity Network Fusion downstream analysis with GEE models
- Description: Associations between the SNF clusters and phenotypic data are tested with GEE models in the ACTION-NTR cohort.

- File name: Additional File 10
- File format: .pdf
- Title: Overview of data dimensions for ACTION-NTR and LUMC-CURIUM cohort
- Description: Table with dimensions (variables by observations) of input and intermediate data types.

Data Availability

The data of the Netherlands Twin Register (NTR) ACTION Biomarker Study may be accessed, upon approval of the data access committee, through the NTR [99].

A synthetic data set representing the structure of the ACTION Biomarker Study data set is available as part of the workflow RO-Crate available at WorkflowHub [76]. An archival copy of the workflow is also available via the GigaScience database GigaDB [100].

Declarations

List of abbreviations

ACTION: Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies; ASEBA: Achenbach System of Empirically Based Assessment; EBI: European Bioinformatics Institute; EGA: European Genome-phenome Archive; EMBL: European Molecular Biology Laboratory; EWAS: epigenome-wide association study; FAIR: Findable, Accessible, Interoperable, Reusable; GC-MS: gas chromatography–mass spectrometry; ISA: Investigation, Study, Assay; LC-MS: liquid chromatography–mass spectrometry; MOFA: Multi-Omics Factor Analysis; NTR: Netherlands Twin Register; RO-Crate: Research Object Crate; SNF: Similarity Network Fusion

Ethical Approval

The ACTION Biomarker study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Central Ethics Committee on Research Involving Human Subjects of the VU

University Medical Center, Amsterdam (NTR 25th of May 2007 and ACTION 2013/41 and 2014.252), an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance- FWA00017598; IRB/institute codes), and the Medical Ethical Committee of Leiden University Medical Center (B17.031, B17.032 and B17.040).

Parents provided written informed consent for the participation of their children.

Consent for publication

Not applicable

Competing Interests

The author(s) declare that they have no competing interests.

Funding

A.J.G. and P.A.C.H. received funding from The Netherlands X-omics Initiative, which is (partially) funded by the Dutch Research Council (NWO), project 184.034.019, and from EATRIS-Plus which has received funding the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096. D.I.B and R.R.J.M.V received funding from "Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies" (ACTION), which is (partially) funded by the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no 602768. D.I.B. received funding from "Consortium on Individual Development" (CID) funded by the Gravitation Program of the Dutch Ministry of Education, Culture, and Science and the Dutch Research Council (NWO) under grant agreement no 024-001-003 and is supported by the Royal Netherlands Academy of Arts and Sciences (KNAW) Professor Award (PAH/6635).

Authors' Contributions

A.J.G, P.A.C.H., D.I.B. and J.D. formulated overarching research aims (Conceptualization). R.P. and N.K. maintained and annotated the metabolomics data of the ACTION-NTR cohort (Data curation). C.V., A.N., P.K., and JD applied statistical methods to analyze the study data (Formal Analysis). A.J.G., P.A.C.H., R.R.J.M.V and D.I.B. acquired financial support for the projects leading to this publication (Funding acquisition). A.N., C.V., F.A.H., N.K., A.S.D.K., P.K., R.P. and J.D. conducted the research process (Investigation). A.N., C.V., F.A.H., N.K., A.S.D.K., P.K., R.P. and J.D. designed the methodology (Methodology). A.N. coordinated the research activity planning and execution (Project administration). F.A.H., R.P., J.D., R.R.J.M.V. and D.B. provided study samples and acquired data (Resources). C.V., P.K., R.P., J.D. and A.N. implemented the source code (Software). D.I.B., A.J.G. and P.A.C.H. provided mentorship to the research team (Supervision). A.N. wrote the initial draft of the manuscript (Writing – original draft). All authors critically reviewed and contributed to writing the manuscript (Writing – review & editing).

Acknowledgements

We acknowledge the ACTION Consortium and thank the participants of the ACTION Biomarker Study. We thank Michael van Vliet for discussions on the FAIRification strategy and Philippe Rocca-Serra and his team for their support on ISA.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2016 dec;3(1):160018. https://doi.org/10.1038/sdata.2016.18.
2. Lawson J, Cabili MN, Kerry G, Boughtwood T, Thorogood A, Alper P, et al. The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genomics 2021 nov;1(2):100028. https://doi.org/10.1016/j.xgen.2021.100028.
3. van der Velde KJ, Singh G, Kaliyaperumal R, Liao X, de Ridder S, Rebers S, et al. FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. Scientific Data 2022 dec;9(1):169. https://doi.org/10.1038/s41597-022-01265-x.
4. Peters K, Bradbury J, Bergmann S, Capuccini M, Cascante M, de Atauri P, et al. PhenoMeNal: processing and analysis of metabolomics data in the cloud. GigaScience 2019 feb;8(2):1–12. https://doi.org/10.1093/gigascience/giy149.
5. Berrios DC, Galazka J, Grigorev K, Gebre S, Costes SV. NASA GeneLab: interfaces for the exploration of space omics data. Nucleic Acids Research 2021 jan;49(D1):D1515–D1522. https://doi.org/10.1093/nar/gkaa887.
6. Courtot M, Gupta D, Liyanage I, Xu F, Burdett T. BioSamples database: FAIRer samples metadata to accelerate research data management. Nucleic Acids Research 2022 jan;50(D1):D1500–D1507. https://doi.org/10.1093/nar/gkab1046.
7. da Silva Santos LOB, Burger K, Kaliyaperumal R, Wilkinson MD. FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. Data Intelligence 2022 sep;p. 1–21. https://doi.org/10.1162/dint_a_00160.
8. van Lin N, Paliouras G, Vroom E, 't Hoen PAC, Roos M. How Patient Organizations Can Drive FAIR Data Efforts to Facilitate Research and Health Care: A Report of the Virtual Second International Meeting on Duchenne Data Sharing, March 3, 2021. Journal of Neuromuscular Diseases 2021 nov;8(6):1097–1108. http://doi.org/10.3233/JND-210721.
9. Lalout N, ERN EURO-NMD and Duchenne Parent Project begin leveraging their FAIR data – EURO-NMD Registry Website (ern-euro-nmd.eu);. https://registry.ern-euro-nmd.eu/ern-euro-nmd-and-duchenne-parent-project-begin-leveraging-their-fai accessed 18 Nov 2022.
10. Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, Long Q, Joosten SA, van der Wijk HJ, et al. Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. Journal of Biomedical Semantics 2022 dec;13(1):12. https://doi.org/10.1186/s13326-022-00263-7.
11. Baker M. 1,500 scientists lift the lid on reproducibility. Nature 2016 may;533(7604):452–454. https://doi.org/10.1038/533452a.
12. Jiménez RC, Kuzak M, Alhamdoosh M, Barker M, Batut B, Borg M, et al. Four simple recommendations to encourage best practices in research software. F1000Research 2017 jun;6:876. https://doi.org/10.12688/f1000research.11407.1.
13. The Netherlands eScience Center, Dutch Data Archiving and Networked Services (DANS), FAIR | FAIR (fair-software.nl);. https://fair-software.nl/, accessed 21 Nov 2022.
14. Barker M, Chue Hong NP, Katz DS, Lamprecht AL, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. Scientific Data 2022;9(1):622. https://doi.org/10.1038/s41597-022-01710-x.
15. Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, et al. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Generation Computer Systems 2017 oct;75:284–298. https://doi.org/10.1016/j.future.2017.01.012.

16. Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe MR, et al. FAIR Computational Workflows. Data Intelligence 2020 jan;2(1-2):108–121. https://doi.org/10.1162/dint_a_00033.
17. Kim YM, Poline JB, Dumas G. Experimenting with reproducibility: A case study of robustness in bioinformatics. GigaScience 2018;7(7):1–8. https://doi.org/10.1093/gigascience/giy077.
18. Chue Hong NP, Katz DS, Barker M, Lamprecht AL, Martinez C, Psomopoulos FE, et al. FAIR Principles for Research Software version 1.0. (FAIR4RS Principles v1.0). Research Data Alliance 2022;https://doi.org/10.15497/RDA00068.
19. Stoudt S, Vásquez VN, Martinez CC. Principles for data analysis workflows. PLoS Computational Biology 2021;17(3):1–26. https://doi.org/10.1371/journal.pcbi.1008770.
20. Software Freedom Conservancy, Git;. https://git-scm.com/, accessed 31 Oct 2023.
21. GitHub, Inc , GitHub;. https://github.com/, accessed 21 Nov 2022.
22. GitLab B V , The One DevOps Platform | GitLab;. https://gitlab.com/, accessed 21 Nov 2022.
23. Atlassian Pty Ltd, Bitbucket | Git solution for teams using Jira;. https://bitbucket.org/, accessed 21 Nov 2022.
24. Martín del Pico E, Gelpi JL, Capella-Gutiérrez S. FAIRsoft – A practical implementation of FAIR principles for research software. bioRxiv 2022;p. 2022.05.04.490563. https://doi.org/10.1101/2022.05.04.490563.
25. Grüning BA, Lampa S, Vaudel M, Blankenberg D. Software engineering for scientific big data analysis. GigaScience 2019;8(5):1–6. https://doi.org/10.1093/gigascience/giz054.
26. Marx V. When computational pipelines go 'clank'. Nature Methods 2020 jul;17(7):659–662. https://doi.org/10.1038/s41592-020-0886-9.
27. The Apache Software Foundation, APACHE LICENSE, VERSION 2.0;. https://www.apache.org/licenses/LICENSE-2.0, accessed 2 Feb 2023.
28. Opensource org, The MIT License;. https://opensource.org/license/mit/, accessed 2 Feb 2023.
29. Reiter T, Brooks† PT, Irber† L, Joslin† SEK, Reid† CM, Scott† C, et al. Streamlining data-intensive biology with workflow systems. GigaScience 2021;10(1):1–19. https://doi.org/10.1093/gigascience/giaa140.
30. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology 2017;35(4):316–319. https://doi.org/10.1038/nbt.3820.
31. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research 2021 apr;10(May):33. https://doi.org/10.12688/f1000research.29032.2.
32. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nature Methods 2021 oct;18(10):1161–1168. https://www.nature.com/articles/s41592-021-01254-9.
33. Contributors to the Apptainer project, established as Apptainer a Series of LF Projects LLC, Home | Apptainer;. https://apptainer.org/, accessed 22 Nov 2022.
34. Docker Inc , Docker: Accelerated, Containerized Application Development;. https://www.docker.com/, accessed 22 Nov 2022.
35. Gruening B, Sallou O, Moreno P, da Veiga Leprevost F, Ménager H, Søndergaard D, et al. Recommendations for the packaging and containerizing of bioinformatics software. F1000Research 2019 mar;7:742. https://doi.org/10.12688/f1000research.15140.2.
36. De Smedt K, Koureas D, Wittenburg P. FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 2020 apr;8(2):21. https://doi.org/10.3390/publications8020021.
37. Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández JM, et al. Packaging research artefacts with RO-Crate. Data Science 2022 jul;5(2):97–138. https://doi.org/10.3233/DS-210053.
38. Krassowski M, Das V, Sahu SK, Misra BB. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Frontiers in Genetics 2020 dec;11(December):1–17. https://doi.org/10.3389/fgene.2020.610798.
39. The University of Manchester, HITS gGmbH, WorkflowHub;. https://workflowhub.eu/, accessed 10 Nov 2022.
40. Goble C, Soiland-reyes S, Bacall F, Owen S, Williams A, Eguinoa I, et al. Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo 2021;https://doi.org/10.5281/zenodo.4605653.
41. Soiland-Reyes S, Sefton P, Castro LJ, Coppens F, Garijo D, Leo S, et al. Creating lightweight FAIR Digital Objects with RO-Crate. Research Ideas and Outcomes 2022;8:e93937. https://doi.org/10.3897/rio.8.e93937.
42. Soiland-Reyes S, Goble C, Groth P, Evaluating FAIR Digital Object and Linked Data as distributed object systems; 2023. https://doi.org/10.48550/arXiv.2306.07436.
43. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Research 2020 nov;48(D1):D440–D444. https://doi.org/10.1093/nar/gkz1019.
44. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. Nature Genetics 2012 feb;44(2):121–126. https://doi.org/10.1038/ng.1054.
45. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016 2016;p. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87.
46. Johnson D, Batista D, Cochrane K, Davey RP, Etuk A, Gonzalez-Beltran A, et al. ISA API: An open platform for interoperable life science experimental metadata. GigaScience 2021 sep;10(9):1–13. https://doi.org/10.1093/gigascience/giab060.
47. The Netherlands X-omics Initiative, X-omics ISA-ACTION-Template;. https://github.com/Xomics/ISA-ACTION-Template/tree/v1.0.0, accessed 9 Nov 2022.
48. Harris MA. In: Keith JM, editor. Developing an Ontology Totowa, NJ: Humana Press; 2008. p. 111–124. https://doi.org/10.1007/978-1-60327-159-2_5.
49. National Cancer Institute, NCI Thesaurus OBO Edition;. https://github.com/NCI-Thesaurus/thesaurus-obo-edition/, accessed 21 Nov 2022.
50. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 2010 03;26(8):1112–1118. https://doi.org/10.1093/bioinformatics/btq099.
51. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. PLOS ONE 2016 04;11(4):1–19. https://doi.org/10.1371/journal.pone.0154556.
52. Metabolomics Standards Initiative, Metabolomics Standards Initiative Ontology;. https://github.com/MSI-Metabolomics-Standards-Initiative/MSIO/, accessed 21 Nov 2022.
53. Royal Society of Chemistry, Chemical Methods Ontology;. http://purl.obolibrary.org/obo/chmo.owl, accessed 21 Nov 2022.

54. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 2015 10;44(D1):D1214–D1219. https://doi.org/10.1093/nar/gkv1031.
55. The Netherlands X-omics Initiative, X-omics ACTION demonstrator multi-omics analysis workflow Documentation;. https://github.com/Xomics/ACTIONdemonstrator_workflow/blob/main/Documentation.md, accessed 9 Nov 2022.
56. Ison J, Kalaš M, Jonassen I, Bolser D, Uludag M, McWilliam H, et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 2013 03;29(10):1325–1332. https://doi.org/10.1093/bioinformatics/btt113.
57. Dumontier M, Baker C, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of biomedical semantics 2014 03;5:14. https://doi.org/10.1186/2041-1480-5-14.
58. De Geest P, Droesbeke B, Eguinoa I, Gaignard A, Huber S, Leo S, et al., ro-crate-py. Zenodo; 2022. https://doi.org/10.5281/zenodo.6522015.
59. Boomsma DI. Aggression in children: unravelling the interplay of genes and environment through (epi) genetics and metabolomics. Journal of Pediatric and Neonatal Individualized Medicine 2015;4(2):e040251. https://doi.org/10.7363/040251.
60. Bartels M, Hendriks A, Mauri M, Krapohl E, Whipp A, Bolhuis K, et al. 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 2018 sep;27(9):1105–1121. https://doi.org/10.1007/s00787-018-1169-1.
61. Hagenbeek FA, Roetman PJ, Pool R, Kluft C, Harms AC, van Dongen J, et al. Urinary Amine and Organic Acid Metabolites Evaluated as Markers for Childhood Aggression: The ACTION Biomarker Study. Frontiers in Psychiatry 2020 mar;11(165). https://doi.org/10.3389/fpsyt.2020.00165.
62. The Netherlands X-omics Initiative, X-omics - Home;. https://www.x-omics.nl/, accessed 24 Nov 2022.
63. Vrije Universiteit Amsterdam, Information for Researchers | Nederlands Tweelingen Register (vu.nl);. https://tweelingenregister.vu.nl/information_for_researchers/information-for-researchers, accessed 24 Nov 2022.
64. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software 2008;25(1):1–18. https://doi.org/10.18637/jss.v025.i01.
65. Husson F, Lê S, Pagès J. Multiple Correspondence Analysis (MCA). In: Exploratory Multivariate Analysis by Example Using R, 2nd ed. Chapman and Hall/CRC; 2017. https://doi.org/10.1201/b21874.
66. Achenbach TM, Ivanova MY, Rescorla LA. Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. Comprehensive Psychiatry 2017;79:4–18. https://doi.org/10.1016/j.comppsych.2017.03.006.
67. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology 2018;14(6):e8124. https://doi.org/10.15252/msb.20178124.
68. Xiong Z, Yang F, Li M, Ma Y, Zhao W, Wang G, et al. EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. Nucleic Acids Research 2022 jan;50(D1):D1004–D1009. https://doi.org/10.1093/nar/gkab972.
69. Braun PR, Tanaka-Sahker M, Chan AC, Jellison SS, Klisares MJ, Hing BW, et al. Genome-wide DNA methylation investigation of glucocorticoid exposure within buccal samples. Psychiatry and Clinical Neurosciences 2019 jun;73(6):323–330. https://doi.org/10.1111/pcn.12835.
70. Hagenbeek FA, van Dongen J, Pool R, Roetman PJ, Harms AC, Hottenga JJ, et al. Integrative Multi-omics Analysis of Childhood Aggressive Behavior. Behavior Genetics 2023;53:101–117. https://doi.org/10.1007/s10519-022-10126-7.
71. Hubers N, Hagenbeek FA, Pool R, Déjean S, Harms AC, Roetman PJ, et al. Integrative multi-omics analysis of genomic, epigenomic, and metabolomics data leads to new insights for Attention-Deficit/Hyperactivity Disorder. medRxiv 2022;https://doi.org/10.1101/2022.07.21.22277887.
72. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 2014;11(3):333–337. https://doi.org/10.1038/nmeth.2810.
73. Côté S, Vaillancourt T, LeBlanc JC, Nagin DS, Tremblay RE. The Development of Physical Aggression from Toddlerhood to Pre-Adolescence: A Nation Wide Longitudinal Study of Canadian Children. Journal of Abnormal Child Psychology 2006 feb;34(1):68–82. https://doi.org/10.1007/s10802-005-9001-z.
74. Archer J. Does sexual selection explain human sex differences in aggression? Behavioral and Brain Sciences 2009 aug;32(3-4):249–266. https://doi.org/10.1017/s0140525x09990951.
75. EOSC Future, EOSC Association, EOSC Portal (eosc-portal.eu);. https://eosc-portal.eu/, accessed 31 Oct 2023.
76. The Netherlands X-omics Initiative, X-omics ACTION demonstrator multi-omics analysis workflow;. https://doi.org/10.48546/workflowhub.workflow.402.8, accessed 08 Dec 2023.
77. The Netherlands X-omics Initiative, X-omics ACTION demonstrator multi-omics analysis workflow;. https://github.com/Xomics/ACTIONdemonstrator_workflow, accessed 9 Nov 2022.
78. EMBL-EBI, Ontology Lookup Service < EMBL-EBI;. https://www.ebi.ac.uk/ols/index, accessed 9 Mar 2023.
79. Hastings J, Jeliazkova N, Owen G, Tsiliki G, Munteanu CR, Steinbeck C, et al. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. Journal of Biomedical Semantics 2015 dec;6(1):10. https://doi.org/10.1186/s13326-015-0005-5.
80. EGA Consortium, Federated EGA - EGA European Genome-Phenome Archive (ega-archive.org);. https://ega-archive.org/federated, accessed 27 Feb 2023.
81. Health-RI, The Personal Health Train Network | The Personal Health Train (health-ri.nl);. https://pht.health-ri.nl/, accessed 27 Feb 2023.
82. ELIXIR, ELIXIR | A distributed infrastructure for life-science information (elixir-europe.org);. https://elixir-europe.org/, accessed 31 Oct 2023.
83. ELIXIR, FAIRplus, FAIR Cookbook (elixir-europe.org);. https://faircookbook.elixir-europe.org/, accessed 31 Oct 2023.
84. Rocca-Serra P, Gu W, Ioannidis V, Abbassi-Daloii T, Capella-Gutierrez S, Chandramouliswaran I, et al. The FAIR Cookbook - the essential resource for and by FAIR doers. Scientific Data 2023;10:292. https://doi.org/10.1038/s41597-023-02166-3.
85. van Dongen J, Hagenbeek FA, Suderman M, Roetman PJ, Sugden K, Chiocchetti AG, et al. DNA methylation signatures of aggression and closely related constructs: A meta-analysis of epigenome-wide studies across the lifespan. Molecular Psychiatry 2021 jan;p. 2020.07.22.215939. https://doi.org/10.1038/s41380-020-00987-x.
86. Ip HF, van der Laan CM, Krapohl EML, Brikell I, Sánchez-Mora C, Nolte IM, et al. Genetic association study of childhood aggression across raters, instruments, and age. Translational Psychiatry 2021 dec;11(1):413. https://doi.org/10.

1038/s41398-021-01480-x.
87. Ligthart L, van Beijsterveldt CEM, Kevenaar ST, de Zeeuw E, van Bergen E, Bruins S, et al. The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. Twin Research and Human Genetics 2019 dec;22(6):623–636. https://doi.org/10.1017/thg.2019.93.
88. Hagenbeek FA, van Dongen J, Roetman PJ, Ehli EA, Bartels M, Vermeiren RRJM, et al. ACTION Biomarker Study. protocolsio 2023;https://dx.doi.org/10.17504/protocols.io.eq2ly7qkwlx9/v1.
89. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics 2016;8(3):389–399. https://doi.org/10.2217/epi.15.114.
90. Human Genomics Facility, Laboratory for Population Genomics of the Department of Internal Medicine at Erasmus MC, HuGe-F (glimdna.org);. http://www.glimdna.org/, accessed 8 Nov 2022.
91. Eriksson L. Introduction to multi-and megavariate data analysis using projection methods (PCA & PLS). Umetrics AB; 1999.
92. Sinke L, van Iterson M, Cats D, Slieker R, Heijmans B, DNAmArray: Streamlined workflow for the quality control, normalization, and analysis of Illumina methylation array data;. https://molepi.github.io/DNAmArray_workflow/.
93. Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, et al. A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. Epigenomics 2018 jul;10(7):925–940. https://doi.org/10.2217/epi-2018-0037.
94. Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 2012 jan;28(1):112–118. https://doi.org/10.1093/bioinformatics/btr597.
95. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology 2020 dec;21(1):111. https://doi.org/10.1186/s13059-020-02015-1.
96. Ross Markello, SNFpy;. https://github.com/rmarkello/snfpy, accessed 17 Feb 2023.
97. John CR, Watson D, Barnes MR, Pitzalis C, Lewis MJ. Spectrum: fast density-aware spectral clustering for single and multi-omic data. Bioinformatics 2019 09;36(4):1159–1166. https://doi.org/10.1093/bioinformatics/btz704.
98. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
99. Vrije Universiteit Amsterdam, Working with NTR data - Working with NTR data (vu.nl);. https://ntr-data-request.psy.vu.nl/, accessed 24 Nov 2022.
100. Niehues A, de Visser C, Hagenbeek FA, Kulkarni P, Pool R, Karu N, et al. Supporting data for "A Multi-omics Data Analysis Workflow Packaged as a FAIR Digital Object". GigaScience Database 2023;http://dx.doi.org/10.5524/102488.

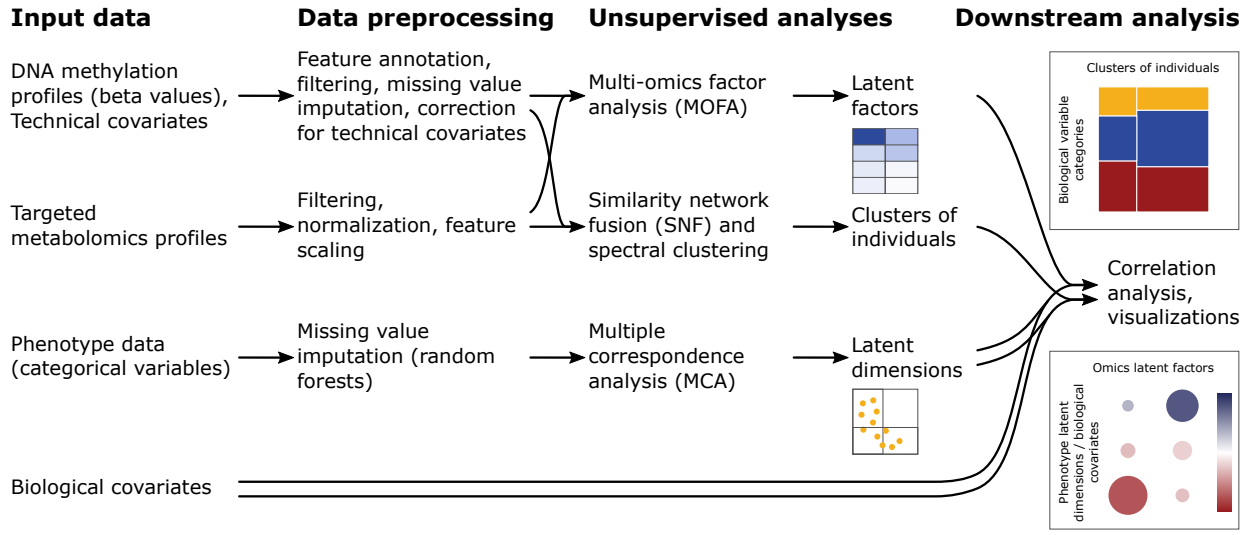


Figure 1. Overview of analysis steps.

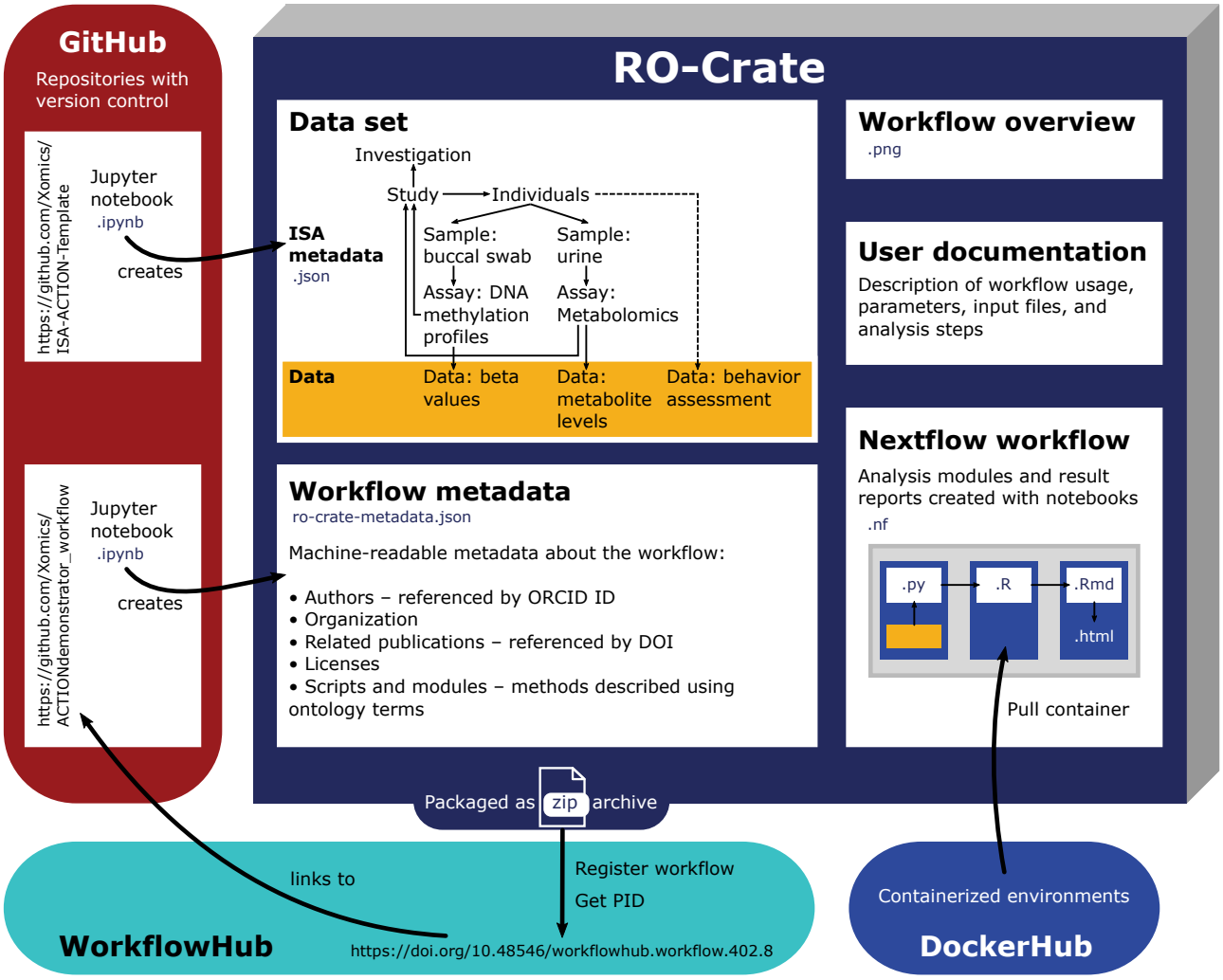


Figure 2. Schematic overview of packaged workflow.

Figure 1

Click here to access/download;Figure;fig1.pdf

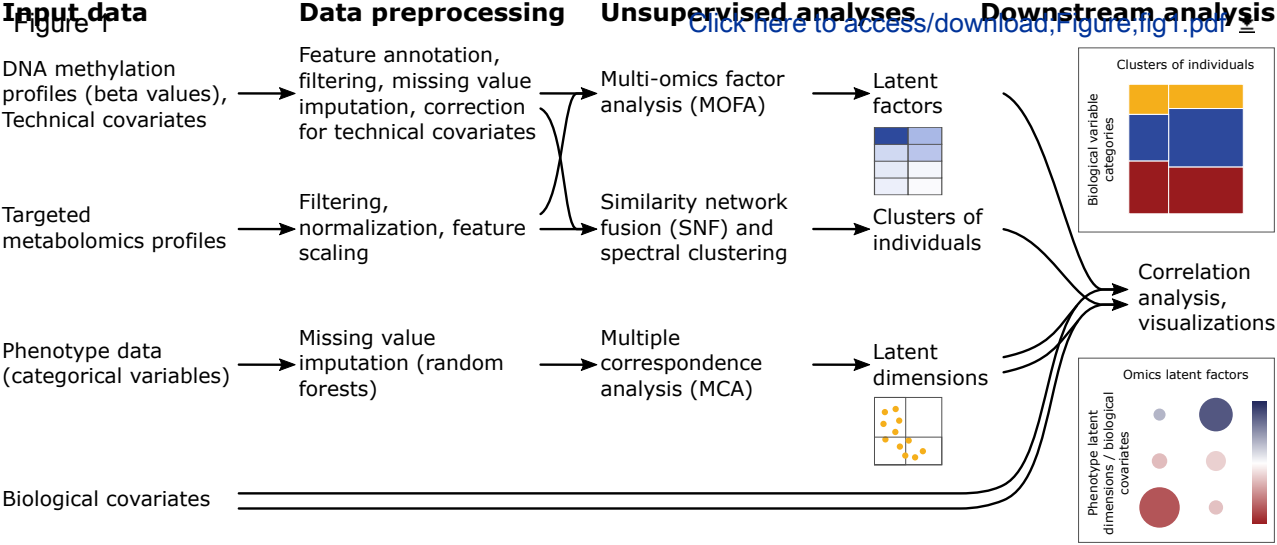
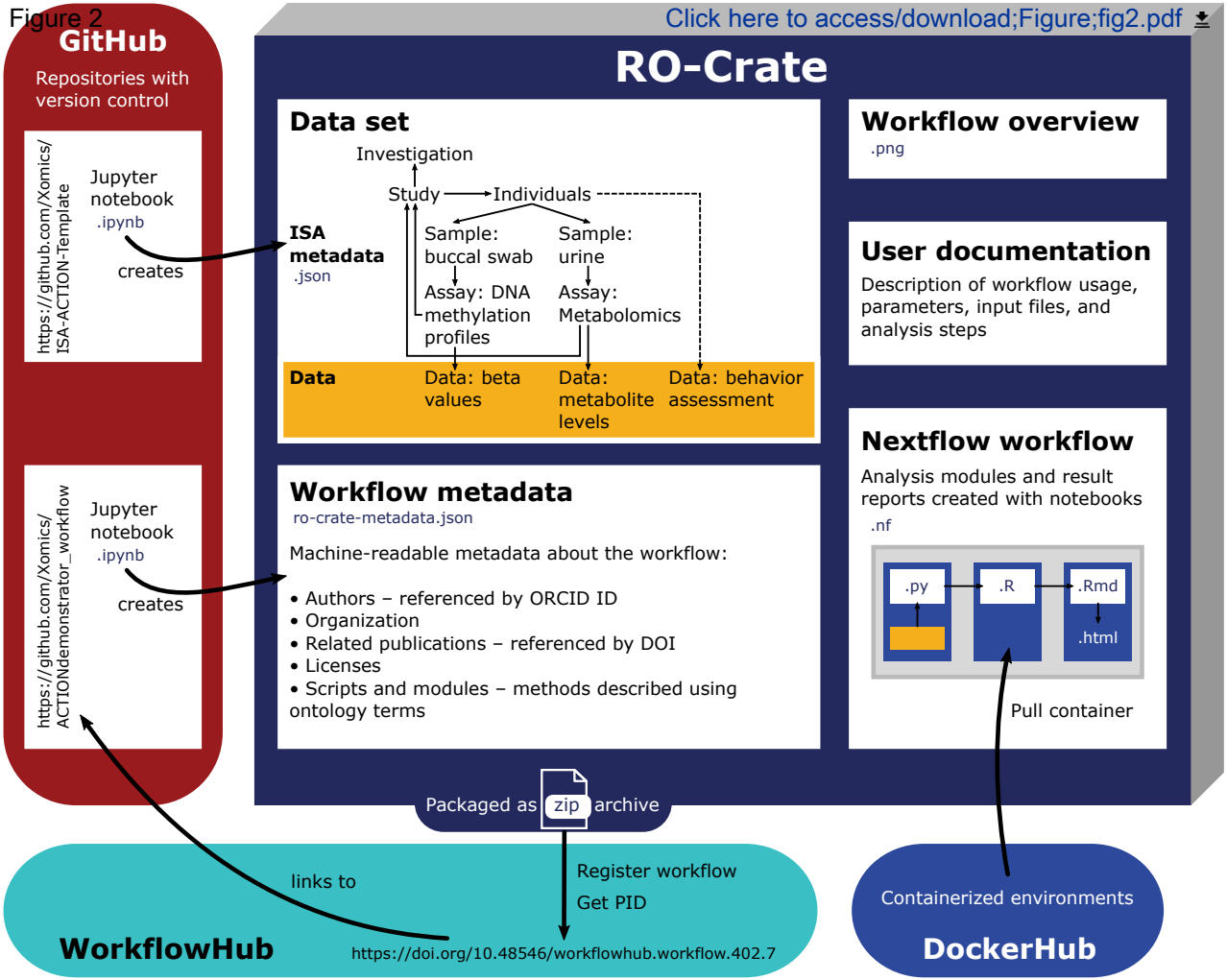


Figure 2

Click here to access/download;Figure;fig2.pdf



Click here to access/download
Supplementary Material
Additional File 1.html

Click here to access/download
Supplementary Material
Additional File 2.html

Click here to access/download
Supplementary Material
Additional File 3.html

Click here to access/download
Supplementary Material
Additional File 4.html

Click here to access/download
Supplementary Material
Additional File 5.xlsx

Click here to access/download
Supplementary Material
Additional File 6.xlsx

Click here to access/download
Supplementary Material
Additional File 7.html

Click here to access/download
Supplementary Material
Additional File 8.html

Click here to access/download
Supplementary Material
Additional File 9.html

Click here to access/download
Supplementary Material
Additional File 10.pdf