FAIR data cube and NTR-ACTION demonstrator

X-omics/ACTION FAIR and data analysis working group

Fernanda de Andrade[1], Jasmin Böhmer[2], Dorret Boomsma[3], Jenny van Dongen[3], Fiona Hagenbeek[3], Peter-Bram 't Hoen[5], Naama Karu[4], Alida Kindt[4], Purva Kulkarni[5], XiaoFeng Liao[5], Leon Mei[6], Anna Niehues[5], René Pool[3], Dieuwke Roelofs-Prins[1], Gurnoor Singh[5], Morris Swertz[1], Joeri van der Velde[1], Casper de Visser[5], Michael van Vliet[4], Gerben van der Vries[1]

[1]UMCG, [2]UMCU, [3]VU, [4]LU, [5]Radboudumc, [6]LUMC

Anna.Niehues@radboudumc.nl, XiaoFeng.Liao@radboudumc.nl, j.van.dongen@vu.nl, Peter-Bram.tHoen@radboudumc.nl

Introduction

The data analysis, integration and stewardship pillar of the Netherlands X-omics initiative aims to contribute to the realization of an integrated X-omics infrastructure and to facilitate multi-omics research by providing means for the creation, analysis and integration of FAIR -omics data. In addition to standardization of data and metadata, we envision a FAIR data cube that combines individual -omics data sets or pointers to these data sets with associated linked metadata. The FAIR data cube should provide an interface to query/search rich human- and machine-understandable metadata and extract relevant molecular data for subsequent analysis. This will aid the integration of different types of omics data, and also promote the integration of -omics data from different sources, as well as facilitate submission to relevant data archives.

Within the X-omics/NTR-ACTION demonstrator project, we are establishing a first use case of the X-omics FAIR data cube, while investigating the predictive value of multi-omics data for classification of aggressive behavior in children.

The ACTION study

The NTR-ACTION demonstrator dataset comprises genomic data (SNP arrays), epigenomic data (Illumina EPIC DNA methylation array data from buccal cells), and biomarker and metabolomic data (from three platforms; amines, organic acids, and steroids) assessed in first morning urine. Samples were collected from approximately 1300 twins around age 9 (Table 1). Data were collected by the Netherlands Twin Register (Ligthart et al. 2019) as part of the ACTION (Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies) project (Hagenbeek et al 2020, van Dongen et al. 2021), with the goal to identify biomarkers for childhood aggression. Twins were invited for participation in the biomarker study based on their longitudinal data on aggressive behavior at ages 3, 7, and/or 9/10 years. At, or around these ages, parents of twins completed the Achenbach System of Empirically Based Assessment (ASEBA) Child Behavior Checklist (CBCL). and teachers of twins completed the ASEBA Teacher Rating Form (TRF). A design was chosen that selected twin pairs who were concordant low-low, concordant high-high, or discordant low-high on indices of childhood aggression, with an oversampling of monozygotic (MZ) pairs.

Table 1. Participant characteristics of the twins (n = 1,347)

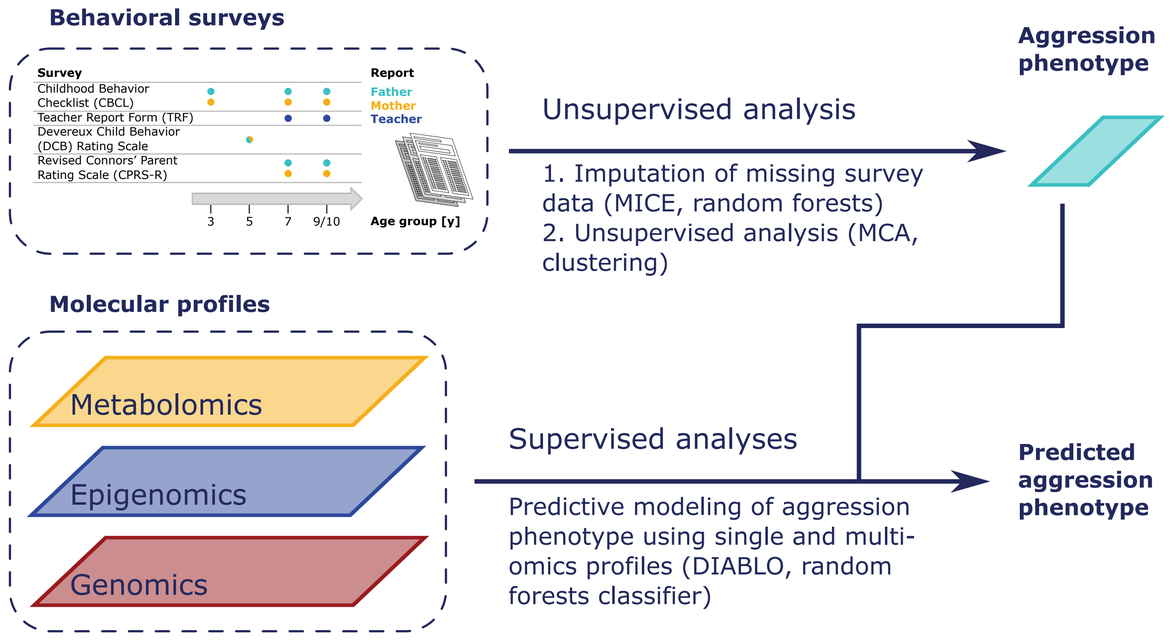
	Twins			
	Concordant low	Discordant		Concordant high
	n = 605	Low (*n* = 189)	High (*n* = 189)	*n* = 364
N complete twin pairs	302	189		182
Mean (SD) age sample collection	9.4 (1.9)	10.1 (1.7)		9.5 (1.8)
Range age sample collection	5.6–12.6	6.1–12.7		5.8–12.9
N (%) MZ twins	469 (77.5%)	306 (81.0%)		330 (90.7%)
N (%) females	323 (53.4%)	85 (45.0%)	79 (41.8%)	159 (43.7%)
CBCL mother (SD) aggression score	2.7 (3.8)	4.4 (4.4)	6.2 (5.8)	7.6 (6.0)
Current psychotropic medication use				
Stimulants	10 (1.7%)	7 (3.7%)	13 (6.9%)	25 (7.0%)
Analgesics	1 (0.2%)	1 (0.5%)	3 (1.6%)	1 (0.3%)
Antipsychotics	1 (0.2%)	0 (0.0%)	1 (0.5%)	3 (0.8%)
Hypnotics/sedatives	7 (1.2%)	1 (0.5%)	2 (1.1%)	6 (1.7%)

CBCL, Child Behavior Checklist; MZ, monozygotic.

Workflow implementation

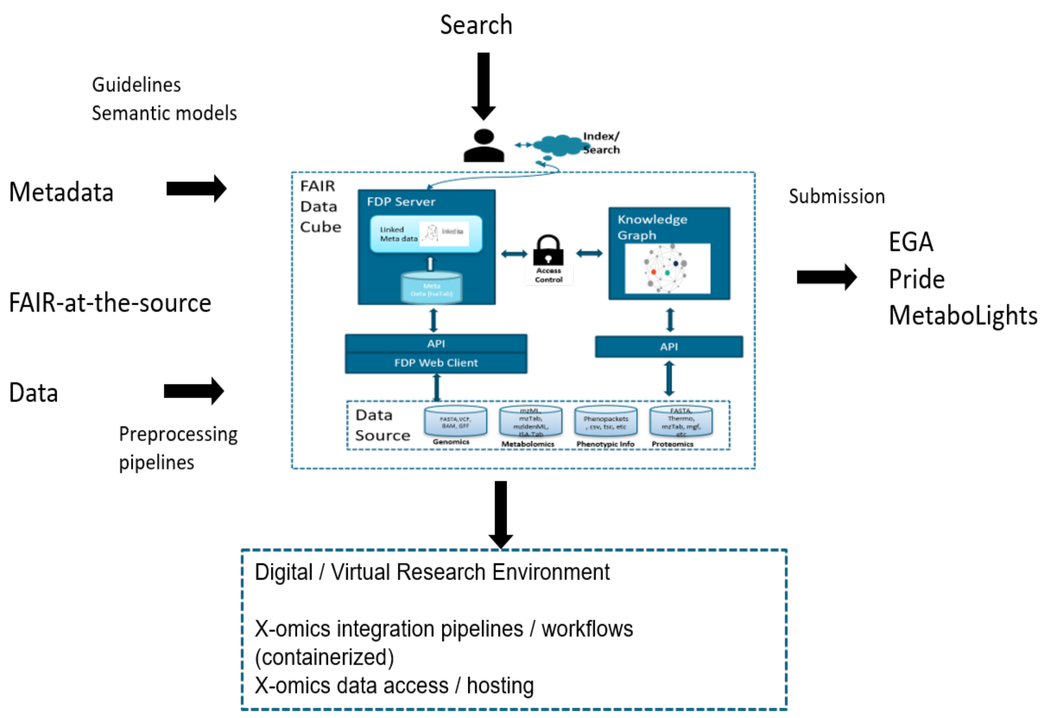
We follow FAIR software guidelines (https://fair-software.nl/) for implementation and sharing of our data analysis workflows. Input data formats are following community standards and are compatible with commonly used data archives.

Predicting childhood aggression with multi-omics data



Initial grouping of study participants is based on longitudinal behavioral surveys. Aggression outcome variables extracted by unsupervised analysis is then modeled using single and multi-omics data. The goal is to identify most suitable omics data types for extraction of aggression-related biomarkers and investigate the possible added value of multi-omics data in the classification of the studied phenotypes.

The FAIR data cube



The FAIR data cube, bears the principle that data should be "as open as possible and as closed as necessary".

By incorporating a FAIR Data Point component internally, the data can be as open as possible and be FAIR-at-the-source. The metadata contents (in our case, ISA metadata) are generated semi-automatically from the data source and exposed to user query in the form of linked metadata (in our case linked ISA). Resources are publicly accessible by anyone. Administration is needed to create/edit resources. The user control is also implemented in the FDP part.

By setting up a knowledge graph inside an internal "cube" and be accessed via an access control mechanism, the data can be as closed as necessary, which enables properly addressing aspects of legislation, privacy, and ethics. The knowledge graph content is generated via bespoken tools and managed internally.

References

Ligthart et al 2019 PMID 31666148, Hagenbeek et al 2020 PMID 32296350, van Dongen et al 2020 PMID 33420481

Acknowledgements

This research was (partially) funded by NWO, project 184.034.019. ACTION received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no 602768.