DELIVERABLE D2.2

REPORT ON THE EVALUATION OF METHODS FOR IMPROVING COMPARABILITY OF CROSS-OMICS DATA FROM DIFFERENT COHORTS

WP2 – Data Stewardship and integration of omic research in Personalised Medicine

Lead Beneficiary: Radboudumc (RUMC)

WP Leader and Institution: Alain van Gool (RUMC), Peter-Bram ’t Hoen (RUMC)

Contributing Partner(s): UH, UP, SERMAS, UU, IBBL

Contractual Delivery Date: 31 December 2022 [M36], extended to 30 June 2023 [M42]

Actual Delivery Date: 28 June 2023

Authors of Deliverable: Anna Niehues (RUMC), Peter-Bram 't Hoen (RUMC), Casper de Visser (RUMC), Jana Vrbkova (UP), Lukas Najdekr (UP), Reka Toth (LIH), Val Fernandez (SERMAS)

Grant agreement no. 871096

Horizon 2020

H2020-INFRADEV-3

Type of Action: RIA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096

EXECUTIVE SUMMARY

The key scientific output of the EATRIS-Plus project is to develop a Multi-omic Toolbox available for researchers in order to have a better understanding of the molecular profiles in personalised medicine.

Within work packages 1, 2, and 3, we implement practices based on the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles to ensure reproducibility of our work and to make the multi-omics cohort data reusable for future translational research. The aim of this deliverable (D2.2) is to describe methods developed and applied to the Czech multi-omics cohort [1] that can improve comparability of data when integrating omics data from different sources.

PROJECT OBJECTIVES

Our flagship project EATRIS-Plus aims to build further capabilities and deliver innovative scientific tools to support the long-term sustainability strategy of EATRIS as one of Europe's key European research infrastructures for Personalised Medicine.

The main goals of the EATRIS-Plus will be to:

- Consolidate EATRIS capacities in the field of Personalised Medicine (particularly omics technologies) to better serve academia and industry and augment the number of EATRIS Innovation Hubs with large pharma;
- Drive patient empowerment through active involvement in the infrastructure's operations;
- Expand strategic partnerships with research infrastructures and other relevant stakeholders, and
- Further strengthen the long-term sustainability of the EATRIS financial model.
- Develop a Multi-omic Toolbox for researchers.

DETAILED REPORT ON THE DELIVERABLE

BACKGROUND

Integration of omics data from different sources can be challenging when data and data processing steps are not sufficiently annotated, or sources of biological and technical variation are unknown. The Czech multi-omics demonstrator cohort serves as a demonstrator for the development and application of computational workflows for omics preprocessing and integrative analyses.

DESCRIPTION OF WORK

[1] *Study subjects were informed and consented to provision of the data related to their person for research and development purposes to existing and future research partners of IMTM in an anonymous form (pseudonymized, from the definition of GDPR, while only IMTM holds the additional information) as described in Ethics Deliverable 10.1 V1.1- Appendix 4*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096

Comparability of data from different sources (samples, cohorts, omics types, platforms) is closely dependent on how data provenance, and in particular sample preparation, data generation, quality control procedures (including internal and external standards), data analysis and results, are reported. Standardised protocols for omics measurements (WP1, D1.1), FAIR data and metadata (WP2, D2.1) and standardized quality assessment using reference samples (WP3) contribute to transparency and reusability of omics data and can facilitate data integration across different sources. Here, we concentrate on best practices for reporting metadata from workflows used for data analysis and computational methods to increase comparability across datasets.

A major challenge when integrating processed (multi-)omics data sets is a lack of unambiguous descriptions of data processing workflows used to generate the data set. FAIR data analysis workflows are needed to apply the same data processing methods to different data sets, or to assess whether different processed data set can be integrated. We share a set of recommendations (de Visser et al. 2023, Ten quick tips for building FAIR workflows, under review in PLOS Computational Biology; poster available at: https://doi.org/10.5281/zenodo.7994135; preprint: https://doi.org/10.5281/zenodo.7994136) to implement FAIR Principles for Research Software (FAIR4RS) and apply them to computational workflows developed in WP2. While the EATRIS-Plus workflows are going to be published before December 31, 2023, we implemented these practices in collaboration with the Netherlands X-omics Initiative (Niehues et al. 2023, https://doi.org/10.1101/2023.06.07.543986) to demonstrate the implementation of FAIR4RS principles in a multi-omics data analysis workflow.

Confounding factors also affect comparability between datasets or studies. Confounders can be of technical nature, such as batch effects, or of biological nature, such as gender. A balanced study design can minimize such effects. Rich metadata can help identifying the biological confounders, which can subsequently be taken into account in the analysis (as we have done for sex in the analysis of the Czech cohort multi-omics data). To account for technical confounders, we implemented several methods to identify batch effects in single omics data sets visually (Principal Component Analysis score plots, correlation plots) and statistically (Redundancy Analysis, Analysis of Variance). We evaluated several batch effect corrections. ComBat (Johnson et al. 2007) is a popular method to adjust for batch effects in RNA-seq and metabolomics data. It can be applied in such a way that known biological effects are not corrected for. We showed this by including sex, the main source of biological variation in our dataset, as covariate in ComBat, and were also able to account for unbalanced batch assignment of male and female samples. To choose the best method for batch effect adjustment, we evaluated the ability of supervised methods (Support Vector Machines, LASSO) to discriminate between 1. sex and 2. batches in the adjusted data sets. The chosen model (ComBat with sex as covariate) performed best at discriminating between males and females while not being able to discriminate between batches. Another advantage of using ComBat is that this method does not require the inclusion of reference samples in each batch, which are not always available and take up measurement capacity.

If QC samples are prepared by a strategy of pooling biological material, other methods can be employed. For the untargeted lipidomics analyses, pooled quality control (QC) samples were injected as every sixth sample through the whole batch. Batch correction was done by SERRF QC Correction (Systematical Error Removal using Random Forest) using these QC samples (Fan et al., 2019, https://pubs.acs.org/doi/10.1021/acs.analchem.8b05592).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096

Many multi-omics integration methods are implemented in R or Python. Employing data structures that comprise multi-omics data, phenotype, and associated metadata for molecular entities, measurements, and individuals can facilitate comparisons across multi-omics studies. We use data structures from the R/Bioconductor library MultiAssayExperiment (Ramos et al., 2017, https://doi.org/10.1158/0008-5472.can-17-0344) and the Python library MuData (Bredikhin et al., 2022, https://doi.org/10.1186/s13059-021-02577-8). Integrative multi-omics analysis workflows use these objects as input. This reduced the time needed for data preparation.

We implemented various single- and multi-omics data analysis methods (Table 1) to identify technical and biological sources of variability that are shared or unique to the different -omics layers. This information is represented in latent dimensions or factors. We determined whether these latent omics dimensions were correlated with phenotypic information or enriched for biological pathways to interpret dimensions biologically.

Table 1: Overview of omics analysis methods implemented in WP2.

	Single omics analyses	*Pairwise omics*	*Multi-omics analyses*
Unsupervised analyses	Principal Component Analysis (PCA), Independent Component Analysis (ICA)	(sparse) Partial Least Squares (PLS)	Multi-omics Factor Analysis (MOFA), Hierarchical Variational Autoencoder, Similarity Network Fusion (SNF)
Supervised analyses	Robust linear models, (s)PLS-DA, Redundancy Analysis (RDA), SVM, LASSO		Multiblock (s)PLS-DA (DIABLO)
Biological interpretation	Correlation analysis, Pathway enrichment analysis (PEA)		

We evaluated the ability of different methods to capture biologically interpretable dimensions and distinguish between variation associated with unique biological factors. We observed that in the Czech demonstrator cohort, body mass index (BMI) is confounded by sex. We found that integrative analyses like MOFA or hierarchical VAE can distinguish between variation associated with sex and variation associated with BMI in this multi-omics data set.

Omics data from different sources can be integrated directly using, e.g., early (concatenation of data sets) or late (separate analyses per data set followed by integrative analysis) integration. Another possibility to transfer information from cohort to another is the application of omics-based predictors. These can be used to enrich existing and new data sets. We described the application of metabolomic predictors of phenotypic traits to complement measured clinical variables in population-scale expression studies (Niehues et al. 2022, https://doi.org/10.1186/s12864-022-08771-7). In the EATRIS-Plus cohort, we applied existing predictors for biological age based on methylation profiles (Levine, 2018 doi: 10.18632/aging.101414). We also used these profiles to measure the samples' cell type composition, using a large set of reference methylomes (Houseman

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096

et al. 2012, https://doi.org/10.1186/1471-2105-13-86, Loyfer et al. https://doi.org/10.1038/s41586-022-05580-6).

NEXT STEPS

We are drafting a scientific publication on the comparison of different batch correction methods. This will include a comparison and evaluation of different methods (PCA, ICA, MOFA, VAE) in their ability to extract biologically interpretable factors from the omics data will be performed, while accounting for technical variation.

Computational workflows developed in the context of this deliverable and WP2 will be published on WorkflowHub and shared via the EATRIS-Plus Multi-omics Toolbox. We plan to investigate how differences between chronological age and methylation age are reflected in other omics types and confirm that in external cohorts of collaborators.

ABBREVIATIONS

BMI	body mass index
FAIR	Findable, Accessible, Interoperable, Reusable
ICA	Independent Component Analysis
MOFA	Multi-omics Factor Analysis
PCA	Principal Component Analysis
VAE	Variational Autoencoder

DELIVERY AND SCHEDULE

The begin of cohort data preparation and analysis was delayed due to initial delays of sample shipments and data acquisition.

A 6-month extension to deliverable requested on 22/11/2022, approved by the PO on 24/11/22 for the deliverable to be submitted on 30/06/2023. Deliverable submitted on 28/06/2023.

ADJUSTMENTS MADE

n/a

APPENDICES

n/a

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871096