



DELIVERABLE D1.3

QUALITY CRITERIA FOR GLOBAL EXPANSION OF THE REPOSITORY OF MULTI-OMIC REFERENCE VALUES

WP1 – Multi-omics technologies for Personalised Medicine

Lead Beneficiary: UP/IMTM

WP Leader and Institution: Marian Hajduch (UP/IMTM)

Contributing Partner(s): UH/FIMM, RUMC, SERMAS, IMTM, EATRIS

Contractual Delivery Date: 31/12/2023

Actual Delivery Date: 08/12/2023

Authors of Deliverable: Andreas Scherer (UH/FIMM), Anna Niehues (EATRIS), Peter-Bram t'Hoen (RUMC), Laura Garcia Bermejo (SERMAS), Elisa Conde (SERMAS), Lukas Najdekr (IMTM)

Grant agreement no. 871096

Horizon 2020

H2020-INFRADEV-3

Type of Action: RIA



TABLE OF CONTENTS

Executive summary	3
Project objectives.....	3
Detailed report on the deliverable	3
Background	3
Description of work.....	4
Next steps	9
Abbreviations.....	9
Delivery and schedule.....	9
Adjustments made.....	9
Appendices.....	9



EXECUTIVE SUMMARY

This deliverable describes what precautions need to be taken by both the multi-omics data provider (EATRIS) as well as the data user, to ensure successful integration of the multi-omics data in new analysis, and the reusability of the multi-omics data as reference values. Keys are data FAIRification, analysis transparency, and the consideration of data quality parameters, including unwanted sources of noise.

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 research infrastructures for PM. The project spans 4 years (2020-2023), the consortium consists of 19 partners and the total budget is 4,9 mil euros. Activities are divided into 9 WPs overseen by the Steering Committee consisting of each WP leader. The main goals of the EATRIS-Plus project will be to:

- Consolidate EATRIS capacities in the field of personalised medicine (PM) (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.

DETAILED REPORT ON THE DELIVERABLE

BACKGROUND

The expansion and integration of existing datasets to create larger datasets with higher statistical power will increase the scientific and clinical use of the datasets. For the multi-omics data of more than 100 healthy individuals¹ were generated in EATRIS Plus WP1 to be viable and usable by any researcher with knowledge in the field, harmonised quality and IT criteria must be met by all parties. To combine datasets, data must be findable and accessible, and the samples must be characterised by sufficient metadata. Also, certain quality parameters must be met, which include both pre-analytical and analytical aspects, as well as sufficiently suited bioinformatics approaches. Hence, the entire workflow from sampling to analysis, (raw) data deposition, as well as analysis methods and codes must be well documented and accessible.

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

Here we outline which quality aspects we deem necessary to be considered, to facilitate the use and expansion and integration of the EATRIS-Plus multi-omics dataset. We make use of documentation in scientific literature, and our own expertise and experience in the field.

DESCRIPTION OF WORK

The work within this deliverable consisted of collection and representation of consensus knowledge, which was gained in the work on assessment and assurance of omics data quality, data management, FAIRification, and data integration. These aspects are integral to approaching successful data sharing and integration with external datasets. The results presented in this deliverable build on the experience and work from other work packages, which are presented and summarised in several deliverable documents, such as:

- D1.1 – Reference protocol/manual for individual omics methods²
- D1.2 – Sample analysis using individual omic techniques and data outputs³
- D2.1 – Report describing the EATRIS-Plus FAIRification template and workflows⁴
- D2.2 – Report on the evaluation of methods for improving comparability of cross-omics data from different cohorts⁵
- D3.1 – Report on proficiency testing for biological specimen processing (confidential)
- D3.2 – Eligibility for “EATRIS certificate of commitment to quality”⁶
- D3.3 – Guidelines for the establishment of reference values⁷
- D9.3 – Data Management Plan⁸

We refer to the outcome and findings of the deliverables at the appropriate locations within this deliverable.

REQUIREMENTS FOR DATA INTEGRATION

Transparency about samples, experimental methods, data, and data analyses are crucial for reproducibility of translational research and to enable and promote reuse of omics data. We aim to increase the value of data generated within EATRIS-Plus and conducted research by applying FAIR

² EATRIS-Plus deliverable 1.1: Technological Reference Protocols for Transcriptomic, Proteomic and Metabolomic Analysis in EATRIS-Plus Project. Scherer 2021. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.4659769>.

³ EATRIS-Plus deliverable 1.2: Sample Analysis Using Individual Omic Techniques and Data Outputs. Najdekr *et al.* 2023. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.8112434>.

⁴ EATRIS-Plus deliverable 2.1: Report Describing EATRIS-Plus FAIRification Template and Workflows. Niehues *et al.* 2022. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.6984721>.

⁵ EATRIS-Plus deliverable 2.2: Report on the Evaluation of Methods for Improving Comparability of Cross-Omics Data from Different Cohorts. Niehues *et al.* 2023. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.8112703>.

⁶ EATRIS-Plus deliverable 3.2: EATRIS Certificate of Commitment to Quality - Eligibility Criteria. Scherer 2023. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.8112381>.

⁷ EATRIS-Plus deliverable 3.3: Guidelines for the Establishment of Reference Values for Omics. Scherer *et al.* 2021. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.6984796>.

⁸ EATRIS-Plus deliverable 9.3: Data Management Plan. <https://zenodo.org/doi/10.5281/zenodo.10246048>

principles to data, metadata, and analysis workflows⁹. Another aim is to provide guidelines for other research projects in the field of personalised health and personalised medicine to make the generated multi-omics cohort data reusable and eligible for further global expansion for future translational research. The FAIRification strategy of EATRIS-Plus has been described along with templates that were used to collect metadata about multi-omics experiments⁴. Additionally, we are going to publicly share the project's Data Management Plan (D9.3)⁸. Aspects relevant to the expansion and integration of EATRIS-Plus multi-omics cohort data are outlined in the following.

In the case of multi-omics, two levels of data integration can be identified: 1) Data integration within a multi-omics dataset; 2) Data combination across multi-omics datasets. All recommendations below affect the analysis performance on both levels. In (1) a FAIRification step that maps measured features to existing ontologies helps to find biological connections between the different –omics levels. In (2) different FAIRification steps that make it possible to automatically identify which –omics features, and sample characteristics are shared between datasets largely facilitated data integration. For any data integration approach, rich metadata describing how the data were generated and processed, which quality controls were applied and which samples and –omics features were measured are of utmost importance.

METADATA

Multi-omics datasets require additional levels of quality control, to avoid samples from different sources (e.g., different patients) are labelled as originating from the same patient. Occasional mislabelling of samples can be a big issue in large studies, and, if undetected, lead to misguided conclusions. Detection of mislabelling events is an issue that deserves some attention, especially in personalised medicine scenarios, but not so much in large scale population health approaches. Multi-omics data can help identify mislabelled samples¹⁰. To this end, relationships between samples and data need to be described (metadata). EATRIS-Plus used existing metadata standards to make data interoperable. Biological sources and measured samples are described using the Investigation/Study/Assay (ISA) metadata framework. EATRIS-Plus additionally suggests using omics specific data standards including the one promoted by the metabolomics standard initiative (MSI), Minimum Information About a Next-generation Sequencing Experiment (MINSEQE) guidelines, minimum information about a proteomics experiment (MIAPE), or guidelines for domain-specific data archives (EBI MetaboLights; EBI European Nucleotide Archive ENA). Please see D9.3⁸ for more detail.

QUALITY CONTROL OF MULTI-OMICS DATA

Quality control parameters for the individual omics platforms are available (e.g., D3.3⁶). During the processing of large datasets, aliquots of the same reference samples should be processed in each batch, so that the resulting datasets can be adjusted for unwanted sources of variation (D2.2⁵).

⁹ The FAIR Guiding Principles for scientific data management and stewardship. Wilkinson *et al.* 2016. Scientific Data. <https://doi.org/10.1038/sdata.2016.18>

¹⁰ A community effort to identify and correct mislabeled samples in proteogenomic studies. Yoo *et al.* 2021. Patterns (N Y). <https://doi.org/10.1016/j.patter.2021.100245>.

REFERENCE METHODS FOR INDIVIDUAL OMICS METHODS: PROFICIENCY TESTING AND SAMPLE PROCESSING METHODS (D1.1 AND D3.1)

An essential criterion for re-usability of biological data is data quality. Validated sample processing protocols should be applied. The validation process should include the quality assessment of pre-analytical methods which can be done in proficiency testing schemes, such as the one provided under EATRIS-Plus and offered by the Integrated BioBank of Luxembourg (IBBL). Briefly, IBBL have produced the standardised samples of the abovementioned PT schemes and have shipped them to the EATRIS-Plus participants. EATRIS-Plus participants have used their routine processing methods to extract the samples and have filled in an online questionnaire to provide pre-analytical data. IBBL have organised the returned dry ice shipment of the extracted samples from EATRIS-Plus participants' premises to IBBL. The extracted samples have been analysed in an isochronous testing. Test results from all participants (including non-EATRIS-Plus participants) have been gathered to perform a statistical analysis and benchmarking (details are described in D3.1).

Participation in comparative ring testing studies is recommended. The quality of reference sample processing outcome can be addressed by open comparison of the results within the same facility over time or to other facilities. EATRIS-Plus has described several efforts in D3.3⁷.

In an effort towards further harmonization of procedures, EATRIS-Plus shares standard operating procedures (SOP), and provides protocols, which have been developed, validated, and established over years and proven of high quality in both academic and industry partnership projects with returning users and customers. Several sites have also participated in one or more international quality assessment consortia projects, such as the FDA-guided SEQC. The protocols were also applied in WP1 of EATRIS-Plus (Deliverable 1.1²).

Analysis of the data should include a rigid fitness-for-use quality assessment, which depend on the omics technology used, and may include quality assessment of the raw data (e.g. FASTQC for DNA-seq and RNA-seq), and the processed data, using techniques such as principal component analysis (PCA), or other batch effect detection methods. Typical quality assuring steps include filtering out "bad" samples, noise correction, normalization and variance stabilization, removal of confounding sources of variation, dimensionality reduction, The methods are extensively described in Deliverable 1.2³.

FAIRIFICATION (D2.1)

Within EATRIS-Plus, we identified relevant data and metadata standards to report and describe data generated by different omics technologies. The existing ISA (Investigation, Study, Assay) metadata framework (<https://isa-tools.org/>) is used to capture experimental metadata. ISA-Tab metadata is used and supported by a wide range of single-omics data repositories (<https://www.isacommmons.org/>). We collected relevant fields needed to describe different omics experiments among the project partners. Based on this collection, a Jupyter notebook was developed to generate ISA-Tab and ISA-JSON files for the multi-omics data set that complies with reporting guidelines for individual omics types. Importantly, the ISA framework enables the use of controlled vocabularies and ontologies supporting both human-readability and machine-readability. The ISA-Tab template files being developed in EATRIS-Plus will serve as templates for collecting the required metadata for integration of multi-omics in the context of research projects around personalised health and personalised medicine. The ISA templates as well as workflow (Jupyter notebook) to create them

are available at <https://github.com/EATRIS/eatris-plus-isa>. They are also shared as part of the EATRIS-Plus Multi-omics Toolbox (MOTBX) launched in November 2023¹¹ and available at <https://motbx.eatris.eu/>.

In addition to the experimental metadata, the multi-omics data set is accompanied by phenotypic information. We chose the GA4GH Phenopackets model to capture Phenotype information in a FAIR format. Fields were mapped to ontology terms and a Phenopacket JSON files was created programmatically. The workflow is available at https://github.com/EATRIS/phenopackets_template.

The original FAIR principles apply to research data. However, to fully ensure reproducibility of conducted research, data analysis workflows need to be FAIR as well. We follow best practices recommended^{12,13} by different initiatives to improve FAIRness of computational research workflows as well. This includes version-control (Git), documentation, citation, adding license, opensource code sharing (<https://github.com/>), containerization (Docker, Singularity), workflow management and registration (WorkflowHub). Metadata about computational workflows will be captured using RO-Crate and workflows will be published on WorkflowHub and shared via the EATRIS-Plus Multi-omics Toolbox (<https://motbx.eatris.eu>). An example workflow annotated with machine-readable metadata was developed in collaboration with The Netherlands X-omics Initiative¹⁴. EATRIS-Plus has described its data access policy, the data repository,

DATA INTEGRATION

Confounding factors also affect comparability between datasets or studies. Knowledge of concepts of design of experiments is essential to minimize the risk of unwanted variation and needs for post-hoc data manipulation. 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 considered in the analysis (as we have done for sex in the analysis of the Czech cohort multi-omics data, D1.1²). To account for technical confounders, we recommend the implementation of statistical methods to identify batch effects in single omics data sets visually (e.g. Principal Component Analysis score plots, correlation plots) and statistically (e.g. Redundancy Analysis, Analysis of Variance). In D2.2⁵, we evaluated several batch effect corrections.

Holistic representations of biological samples can be obtained only with analyses across modalities (i.e., technology platforms, such as omics and imaging), whereby several modalities of the same sample are jointly examined. Although advancements in experimental assays allow for the paired measurements of many modality combinations, different modalities are still commonly measured

¹¹ EATRIS-Plus deliverable 3.5: The Multi-omics Toolbox (MOTBX): Empowering Multi-omics Research and Analysis for the Translational Medicine Community. Oldoni *et al.* 2023. Zenodo. <https://zenodo.org/doi/10.5281/zenodo.10141669>

¹² Software engineering for scientific big data analysis. Grüning *et al.* 2019. GigaScience. <https://doi.org/10.1093/gigascience/giz054>.

¹³ Four simple recommendations to encourage best practices in research software. Jimenez *et al.* 2017. F1000Research. <https://doi.org/10.12688/f1000research.11407.1>

¹⁴ A Multi-omics Data Analysis Workflow Packaged as a FAIR Digital Object. Niehues *et al.* 2023. bioRxiv. <https://doi.org/10.1101/2023.06.07.543986>. Submitted to GigaScience.

independently, resulting in unpaired data. These data sets need to be properly integrated to obtain an informative low-dimensional embedding that can be used to visualize properties of interest.

Three types of multi-omics data integration approaches have been identified in the literature (e.g. Heumos *et al.*¹⁵ : early, late, and intermediate integration. In early integration, the different omics datasets are combined into one table or graph-based representation which is then used as input to a Machine Learning (ML) model. In late integration, models are applied to each dataset independently. Then, a second model combines their predictions. Finally, in intermediate integration, a model learns a joint representation of the datasets. According to Heumos *et al.*,¹⁵ the methods can be described as follows:

COMBINING JOINTLY MEASURED MODALITIES: PAIRED INTEGRATION

Paired integration can be conducted with linear approaches such as factor analysis implemented in MOFA+ to obtain a joint, interpretable latent space. This approach requires size factor normalization to ensure that the first factors are not dominated by differences in total expression per sample. Alternatively, weighted nearest-neighbour (WNN)³ analysis learns cell-specific modality weights that reflect the modality information content to determine the importance of modalities in downstream analyses in the form of a neighbour graph. This graph can be reused for the calculation of embeddings or distance metrics.

INTEGRATING DISJOINT MEASUREMENTS: UNPAIRED INTEGRATION

The main difficulty in integrating unpaired multi-omics data lies in the distinct feature spaces. Initial approaches that map multimodal data into a common feature space based on prior knowledge may result in information loss. Heumos *et al.* note that nonlinear manifold alignment approaches such as optimal transport-based methods such as SCOT²⁴⁹ or UnionCom²⁵⁰ do not require prior knowledge and could therefore reduce the inter-modality information loss.

INTEGRATING JOINT AND DISJOINT MEASUREMENTS: MOSAIC INTEGRATION

Capture of several modalities from the same sample, e.g., a cell, simultaneously is still challenging despite advancements in experimental assays. Profiling individual modalities on different populations of cells from the same biological sample is more common, leading to completely missing data matrices. The integration of data in such set-ups is known as ‘mosaic integration’, for which tools recently started to emerge.

Tarazano *et al.*¹⁶ provide a set of harmonised Figures of Merit (FoM) as quality descriptors applicable to different omics data types. Meeting the quality metrics of each omics is a prerequisite for expanding or merging datasets. Tarazano proposes the use of Figures of Merit to compare omics platforms, and as way to integrate multiple omics studies if commonly applied. FoM have been traditionally used in analytical chemistry to describe performance of instruments, however, Tarazano proposes to apply

¹⁵ Best practices for single-cell analysis across modalities. Heumos *et al.* 2023. Nat Rev Genet. <https://www.ncbi.nlm.nih.gov/pubmed/37002403>

¹⁶ Harmonization of quality metrics and power calculation in multi-omic studies. Tarazona *et al.* 2020. Nat Commun. <https://doi.org/10.1038/s41467-020-16937-8>.

them to omics technology platforms where they are found to differ from platform to platform. Some of the features are detection limit, selectivity, coverage, and identification.

REQUIREMENTS FOR IT INFRASTRUCTURE

With the increasing complexity and size of multi-omics data, requirements for a suitable IT environment on which the data are hosted for use are challenging. To the best of our knowledge, no public repository specifically designed for multi-omics datasets is available. Rather, specialised public repositories are built and maintained to host specific types of data and metadata, which must be accessed separately. Although most repositories allow for cross-referencing to the location of other data of the study in other repositories, the compilation of all available data of a study into a single dataset for integrative analysis is tedious, and prone to human error. EMBL-EBI's recommended option archive multi-omics data that requires controlled access is the European Genome-Phenome Archive. However, national legislation / implementation of GDPR can prohibit data from being archived in other countries. There are multiple ongoing initiatives working toward federated solutions (Federated EGA, European Genomic Data Infrastructure). Since these are not operational yet, we addressed this issue for our reference dataset by developing a data repository which hosts all omics data generated for this dataset. Data of the EATRIS-Plus healthy multi-omics cohort will be available in a data repository hosted in the Czech Republic (<https://clindata.imtm.cz>). To ensure findability, a study description, accompanied by information about the data access policy will be made available in a searchable registry.

NEXT STEPS

Recommendations on data integration requirements and strategies are being made available to the community via the Multi-omics Toolbox (<https://motbx.eatris.eu/>) that is the main scientific outcome of WPs 1, 2 and 3 and was already launched¹¹ to the community on 20 November 2023.

ABBREVIATIONS

EGA: European Genome-phenome Archive

FoM: Figures of Merit

GDPR: General Data Protection Regulation

MOTBX: Multi-omics Toolbox

DELIVERY AND SCHEDULE

Deliverable 1.3 is submitted in due time prior to the timeline of 31/12/2023.

ADJUSTMENTS MADE

n/a

APPENDICES



The Deliverables mentioned in the text are all available publicly on Zenodo. For the sake of brevity, they are not added as appendices to this Deliverable, but can be found below with their respective persistent identifiers.

D1.1. REFERENCE PROTOCOLS/MANUAL FOR INDIVIDUAL OMICS METHODS

(Technological Reference Protocols for Transcriptomic, Proteomic and Metabolomic Analysis in EATRIS-Plus Project. <https://zenodo.org/doi/10.5281/zenodo.4659769>)

D1.2 (Sample Analysis Using Individual Omic Techniques and Data Outputs.

<https://zenodo.org/doi/10.5281/zenodo.8112434>)

D2.1 (Report Describing EATRIS-Plus FAIRification Template and Workflows.

<https://zenodo.org/doi/10.5281/zenodo.6984721>)

D2.2 (Report on the Evaluation of Methods for Improving Comparability of Cross-Omics Data from Different Cohorts.

<https://zenodo.org/doi/10.5281/zenodo.8112703>)

D3.1 REPORT ON PROFICIENCY TESTING FOR BIOLOGICAL SPECIMEN PROCESSING (confidential)

D3.2 ELIGIBILITY FOR “EATRIS CERTIFICATE OF COMMITMENT TO QUALITY” (EATRIS Certificate of Commitment to Quality - Eligibility Criteria.

<https://zenodo.org/doi/10.5281/zenodo.8112381>)

D3.3 GUIDELINES FOR THE ESTABLISHMENT OF REFERENCE VALUES (Guidelines for the

Establishment of Reference Values for Omics. <https://zenodo.org/doi/10.5281/zenodo.6984796>)

D9.3 DATA MANAGEMENT PLAN

(<https://zenodo.org/doi/10.5281/zenodo.10246048>)

