DELIVERABLE D2.1

REPORT DESCRIBING THE EATRIS-PLUS FAIRIFICATION TEMPLATE AND WORKFLOWS

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 2021 [M24]

Actual Delivery Date: 21. December 2021

Authors of Deliverable: Anna Niehues (RUMC), Peter-Bram 't Hoen (RUMC), Alain van Gool (RUMC), Elisa Conde Moreno (SERMAS), Maria Laura Garcia Bermejo (SERMAS), Bishwa R. Ghimire (UH/FIMM), Bhagwan Yadav (UH/FIMM), Jessica Nordlund (UU), Sara Ekberg (UU), Jarmila Staňková (UP/IMTM), Patricia Žižkovicičová (UP/IMTM), Petr Džubák (UP/IMTM)

Grant agreement no. 871096

Horizon 2020

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

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 (D2.1) is to describe the EATRIS-Plus FAIRification process, and list standards, formats, and tools used in this process.

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

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 principles to data, metadata, and analysis workflows. This should set an important example and provide guidelines for other research projects in the field of personalized health and personalized medicine.

DESCRIPTION OF WORK

As part of this deliverable (D2.1), we describe the EATRIS-Plus FAIRification strategy (Appendix 1: EATRIS-Plus FAIRification strategy) to make the generated multi-omics cohort data reusable for future translational research.

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

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/)[1] is used to capture experimental metadata. ISA-Tab metadata is used and supported by a wide range of single-omics data repositories (https://www.isacommons.org/). We develop a Jupyter notebook to generate ISA-Tab 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 personalized health and personalized medicine. They will be published and shared and as part of the EATRIS-Plus multi-omics toolbox.

NEXT STEPS

We will use the developed template to collect metadata for the EATRIS-Plus multi-omics cohort. We keep monitoring developments in different omics communities and collaborating with other initiatives. The EATRIS-Plus ISA template and corresponding Jupyter notebook will be updated accordingly and shared with the public as part of the EATRIS-Plus multi-omics toolbox.

ABBREVIATIONS

FAIR	acronym for Findability, Accessibility, Interoperability, and Reusability
ISA	Investigation, Study, Assay (metadata framework)

DELIVERY AND SCHEDULE

The deliverable is submitted within the contractual delivery date.

ADJUSTMENTS MADE

APPENDICES

- Appendix 1: EATRIS-Plus FAIRification strategy

[1] Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., … Hide, W. (2012). Toward interoperable bioscience data. Nature Genetics, 44(2), 121–126. https://doi.org/10.1038/ng.1054

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

APPENDIX 1: EATRIS-PLUS FAIRIFICATION STRATEGY

The EATRIS-Plus multi-omics cohort data set comprises different types of molecular profiles (genomic variation and DNA methylation, mRNA, miRNA, protein, and metabolite levels) and phenotypic information of >100 healthy individuals. Guided by the FAIR principles that were first introduced in 2016 (Wilkinson et al., 2016), we implement data management practices to ensure transparency and reproducibility of our research, and reuse of the data generated within EATRIS-Plus. Our FAIRification strategy is aligned with international standards and guidelines related to data stewardship in general and to specific omics platforms.

The acronym FAIR stands for Findability, Accessibility, Interoperability, and Reusability. Different recommended practices (https://www.go-fair.org/fair-principles/) contribute to improve each of these aspects. The addition of rich metadata improves both findability and reusability. The use of standardized vocabularies, ontologies and unique and persistent identifiers for metadata and features enables machine-readability and thereby enhances interoperability. This is important for both integration of different types of omics data (mapping of identifiers) as well as integration with other data sets.

The ISA (Investigation, Study, Assay) metadata framework (Sansone et al., 2012) allows to capture experimental metadata. This includes metadata on samples and sample processing, omics measurements including sample preparation, measurement protocols, and analysis protocols. Different fields used to describe the metadata are presented in the EATRIS-Plus example ISA-Tab files. We also develop a Jupyter notebook employing the ISA-API (Johnson et al., 2021) which can be used to create the ISA-Tab and ISA-JSON files, and will be made publicly available as part of the EATRIS-Plus multi-omics toolbox. Minimal information standards and guidelines for submission to omics data repositories are being followed so that data submission is facilitated and the reuse of data in those repositories is enhanced. The standards and guidelines are described below.

For **metabolomics**, we follow the guidelines of the EBI MetaboLights repository (Haug et al., 2019) which adheres to Metabolomics Standard Initiative (MSI) standards (Salek et al., 2015; Spicer et al., 2017; van Rijswijk et al., 2017). These include reporting sample processing such as extraction and labeling, measurement specifications such as chromatography and mass spectrometry, and data analysis protocols. We link to the applied protocols (SOPs) using unique and persistent identifiers.

For **RNAseq** (mRNA and microRNA sequencing), we follow MINSEQE (Minimum Information About a Next-generation Sequencing Experiment) guidelines (Brazma et al., 2012) to provide information about the sequencing experiment.

For **WGS**, we follow the General Guide On ENA Data Submission (https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html) to provide information about the sequencing experiment.

For **proteomics**, we follow guidelines of the HUPO-PSI, specifically MIAPE – The Minimum Information About a Proteomics Experiment (Taylor et al., 2007) in following parts, MIAPE: Mass Spectrometry version 2.98, MIAPE: Column Chromatography version 1.1 and MIAPE: Mass Spectrometry Informatics version 1.1. and mzTab v1.0.0. We have described our sample preparation protocol in our applied SOPs (Scherer, 2021) in detail.

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

For **microRNA qRT-PCR** determination, we follow the general recommendations for qPCR included in the MIQE guideline (Bustin et al., 2009). Since we used the miRCURY LNA miRNA Focus PCR Panel from Qiagen in serum samples, we follow Qiagen's guidelines for profiling miRNAs in body fluids (Guidelines for Profiling Biofluid miRNAs – QIAGEN (Qiagen, 2019)). For data normalization, we follow the recommendations described by Mestdagh et al., 2009 and Marabita et al.,2016 (Marabita et al., 2016; Mestdagh et al., 2009). These references describe procedures for each of the main steps required for the miRNA determination by qPCR including sample obtaining and processing, RNA isolation from fluids, cDNA generation by reverse transcription, determination of miRNAs panel by quantitative PCR, quality controls and data normalization.

For **DNA methylation sequencing (EM-seq)**, we follow the ENA Metadata Model to provide meta data about the library and sequencing experiment: https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html

For reporting of measured omics data, we use community standard formats recommended by the aforementioned reporting guidelines. These are listed in Table 1. Database identifiers to report which molecular features were measured are summarized in Table 2.

Table 1: Used formats for raw and processed data per omics technology.

Omics technology	Raw data format	Processed data format(s)	Format used for multi-omics data integration
Targeted metabolomics	.raw (proprietary format), .mzML (open format)	.tsv - metabolite annotation file (MAF)	.tsv - metabolite annotation file (MAF)
Proteomics	.RAW file (Thermo XCalibur)	mzTab (proteomic standard data format)	mzTab and .csv with peptide and protein lists exported from ProteomeDiscoverer
mRNA sequencing	.fastq.gz	.bam, .bam.bai additional file formats in the pre-processed data are: .html, .txt, .zip, .csv, .sf, .tsv, .pdf, .r and .gtf	.tsv
microRNA sequencing	.fastq.gz	.bam additional file formats in the pre-processed data are: .bai, .html, .txt, .zip, .gz, .csv, .tsv, .pdf, .gff and .gtf	.tsv
microRNA qRT-PCR	.txt	Normalized Ct (Crossing Threshold) .xls	.txt and .xls following NCBI GEO database standard
DNA methylation sequencing	.fastq	.bam .bedGraph	.bedgraph
WGS	.fastq.gz	.bam, .vcf	.tsv

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

Table 2: Database identifiers for molecular features per omics technology.

Omics technology	Entity	Database identifier(s)
Targeted metabolomics	Metabolite	CHEBI, HMDB
Proteomics	Peptide Protein	UniProtKB Sequence UniProtKB Accession Number
mRNA sequencing	Transcript	Ensembl ID
microRNA sequencing	microRNA	miRBase ID and Accession number
microRNA qRT-PCR	microRNA	miRBAse ID (release 20)
DNA methylation sequencing	Genomic coordinate	Unique identifier based on chromosome (referred to by GenBank ID and version) and genomic coordinate (GRCh38) corresponding to CpG site
WGS	Gene	HUGO Gene Nomenclature (HGNC), Ensembl Gene ID in combination with a genomic coordinates (GRCh38)

The original FAIR principles apply to research data. However, in order fully ensure reproducibility of conducted research, data analysis workflows need to be FAIR as well. We follow best practices recommended (Gruening et al., 2018; Jiménez et al., 2017) by different initiatives to improve FAIRness of computational research workflows as well. This includes version-control (git), documentation, citation, adding license, open source code sharing (https://github.com/), containerization (Docker, Singularity), workflow management and registration (https://workflowhub.eu/).

REFERENCES

Brazma, A., Ball, C., Bumgarner, R., Furlanello, C., Miller, M., Quackenbush, J., Reich, M., Rustici, G., Stoeckert, C., Trutane, S. C., & Taylor, R. C. (2012). *MINSEQE: Minimum Information about a high-throughput Nucleotide SeQuencing Experiment - a proposal for standards in functional genomic data reporting*. https://doi.org/10.5281/ZENODO.5706412

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, *55*(4), 611–622. https://doi.org/10.1373/clinchem.2008.112797

Gruening, B., Sallou, O., Moreno, P., da Veiga Leprevost, F., Ménager, H., Søndergaard, D., Röst, H., Sachsenberg, T., O'Connor, B., Madeira, F., Dominguez Del Angel, V., Crusoe, M. R., Varma, S., Blankenberg, D., Jimenez, R. C., & Perez-Riverol, Y. (2018). Recommendations for the packaging and containerizing of bioinformatics software. *F1000Research*, *7*, 742. https://doi.org/10.12688/f1000research.15140.1

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., & O'Donovan, C.

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

(2019). MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Research*, *48*(D1), D440–D444. https://doi.org/10.1093/nar/gkz1019

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., Capella-Gutierrez, S., Chue Hong, N., Cook, M., Corpas, M., Flannery, M., Garcia, L., Gelpí, J. L., Gladman, S., Goble, C., González Ferreiro, M., Gonzalez-Beltran, A., Griffin, P. C., Grüning, B., … Crouch, S. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research*, *6*, 876. https://doi.org/10.12688/f1000research.11407.1

Johnson, D., Batista, D., Cochrane, K., Davey, R. P., Etuk, A., Gonzalez-Beltran, A., Haug, K., Izzo, M., Larralde, M., Lawson, T. N., Minotto, A., Moreno, P., Nainala, V. C., O'Donovan, C., Pireddu, L., Roger, P., Shaw, F., Steinbeck, C., Weber, R. J. M., … Rocca-Serra, P. (2021). ISA API: An open platform for interoperable life science experimental metadata. *GigaScience*, *10*(9), 1–37. https://doi.org/10.1093/gigascience/giab060

Marabita, F., de Candia, P., Torri, A., Tegnér, J., Abrignani, S., & Rossi, R. L. (2016). Normalization of circulating microRNA expression data obtained by quantitative real-time RT-PCR. *Briefings in Bioinformatics*, *17*(2), 204–212. https://doi.org/10.1093/bib/bbv056

Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F., Speleman, F., & Vandesompele, J. (2009). A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biology*, *10*(6), R64. https://doi.org/10.1186/gb-2009-10-6-r64

Qiagen. (2019). *Guidelines for Profiling Biofluid miRNAs* (pp. 1–38). Qiagen. https://www.qiagen.com/-/media/project/qiagen/qiagen-home/content-worlds/pcr/l2-real-time-qpcr/guidelines_for_profiling_biofluid_mirnas.pdf

Salek, R. M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., Correa, E., Reijmers, T., Rosato, A., Tenori, L., Turano, P., Marin, S., Deborde, C., Jacob, D., Rolin, D., Dartigues, B., Conesa, P., Haug, K., Rocca-Serra, P., … Steinbeck, C. (2015). COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics*, *11*(6), 1587–1597. https://doi.org/10.1007/s11306-015-0810-y

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., … Hide, W. (2012). Toward interoperable bioscience data. *Nature Genetics*, *44*(2), 121–126. https://doi.org/10.1038/ng.1054

Scherer, A. (2021). *Technological Reference Protocols for Transcriptomic, Proteomic and Metabolomic Analysis in EATRIS-Plus Project*. Zenodo. https://doi.org/10.5281/zenodo.5513674

Spicer, R. A., Salek, R., & Steinbeck, C. (2017). Compliance with minimum information guidelines in public metabolomics repositories. *Scientific Data*, *4*(1), 170137. https://doi.org/10.1038/sdata.2017.137

Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., … Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, *25*(8), 887–893. https://doi.org/10.1038/nbt1329

van Rijswijk, M., Beirnaert, C., Caron, C., Cascante, M., Dominguez, V., Dunn, W. B., Ebbels, T. M. D., Giacomoni, F., Gonzalez-Beltran, A., Hankemeier, T., Haug, K., Izquierdo-Garcia, J. L., Jimenez,

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

R. C., Jourdan, F., Kale, N., Klapa, M. I., Kohlbacher, O., Koort, K., Kultima, K., … Steinbeck, C. (2017). The future of metabolomics in ELIXIR. *F1000Research*, *6*, 1649. https://doi.org/10.12688/f1000research.12342.2

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

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