TRANSTEM

D 8.8 - Data Management Plan Submission

Project acronym	TRANSTEM
Project name	TRANSTEM: ERA Chair in Translational Stem Cell Biology
Project type	Coordination and Support Action
Start date of the project	01 / 10 / 2019
End date of the project	01 / 10 / 2024
Contributing WP	WP8 - Project Management
Deliverable identifier	D 8.8.
Contractual delivery date	31 / 03 / 2020
Actual delivery date	31 / 03 / 2020
Deliverable type	ORDP: Open Research Data Pilot
Dissemination level	PU
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 856871.	

Table of content

1. Introduction

This document represents the Data Management Plan for the TRANSTEM project. The plan details what data the project will generate, how it will be exploited and made accessible for verification and re-use, and how it will be curated and preserved. The underlying principles informing this plan are that the data should be managed so that it is findable, accessible, interoperable and reusable (FAIR) – Guidelines on FAIR Data Management in Horizon 2020.

2. Types of Data

The TRANSTEM project will generate research products on Environmental Metagenomics and its application to ecological problems. There are five main data types that this project will produce:

2.1. Nucleic acid sequence data;

2.2. Patient information and data;

2.3. Molecular biological and field collection methods;

2.4. Imaging database and associated metadata.

Each data type will be archived and disseminated in the ways that are appropriate to its specific qualities.

3. Data Archiving

3.1. *Nucleic acid sequence data*

Nucleic acid sequence data can be archived in many ways depending upon the degree of analysis and annotation of features that has been undertaken on it. Raw reads will be archived either in the NCBI Short Read Archive; or similar appropriate archives of un-analysed sequence data; or a generic data archive such as DRYAD (http://datadryad.org/). Sequences that have been carefully inspected and associated with a clearly defined taxon verified by a recognized taxonomic expert will be deposited in the GenBank (https://www.ncbi.nlm.nih.gov/genbank/) or DRYAD. These approaches are the international standard for the field ensuring data is openly accessible. This is also the main way that nucleic acid sequence data is re-used by other researchers.

3.2. Patient information and data

The consent forms (hard copies) will be approved by the Ethics Committee of the Medical University – Varna (EC-MUV). The signed forms will be kept in secure locked storage in a research office for 10 years. The data will only be available to the principal investigator and appointed by the principal investigator members of the research team.

Pseudo-anonymous electronic data sent to MUV from the clinicians will be stored as a non-patient identifiable database on encrypted MUV computers for 10 years.

Pseudo-anonymous samples of tissue donated to MUV will be destroyed on site once analysis has completed unless the patient consents to donate the sample to the biobank of MUV.

The principal investigator and selected by her/him team members will have access to the data if required. It will be carefully explained to patients that their data will be retained for up to 10 years (as opposed to the usual 5) to complete the analyses on their donated tissue and to answer research questions that arise from the analyses.

Storage and backup

The hard copies of consent forms will be stored in a locked filing cabinet in a locked room at each hospital site.

All electronic data will be stored on backed-up MUV servers.

All human tissue samples will be stored in an appropriate laboratory and labeled. The tissue will be stored in its own area and will not be mixed with those of other projects.

3.3. *Molecular biological or field methods*

Molecular biological or field collection procedures for biological materials will be achieved either through peer-reviewed publication in Open Access journals or by making a transcript or video of the methods and making this available on the TRANSTEM website. This follows the H2020 principle of Open Research Data (ORD) publication.

Sequencing data

We store the raw sequencing data in our own servers. These files will be organized by the date of the sequencing run and that will enable us to easily match corresponding sequencing data with sample data. Since the primary analysis of the sequencing data is performed by using MySeq™ Genomic Analysis Software, the structure of the resulting output files is also used for long term storage of analysis results on the same servers. These results are logically structured into projects and analyses by the software and this file system is very easy to navigate.

Some of the sequencing will be outsourced to external sequencing facilities and after receiving the raw data, we will store the raw sequencing data on our own servers. After publishing a peer-reviewed publication in Open Access journals the data will be made available to other researchers through dedicated scientific sequencing databases. This again is in accordance to the H2020 principle of Open Research Data (ORD) publication.

3.4. *Imaging database and associated metadata*

The project will involve digital images which will be available in a public database (www.monkey-niche.org). The images will be converted into a zoomable image pyramid format (Zoomify http://www.zoomify.com/) with Libvips (https://libvips.github.io/libvips/), and the OpenSlide library (https://openslide.org/). The metadata will be extracted and stored in a MySQL Database together with the standard image data. The database will accessed using a web interface and the images will be rendered via a custom implementation of the Openseadragon viewer (https://openseadragon.github.io/) which additionally features an image comparison mode.

4. Data Dissemination

Pseudo-anonymized group data will be made publicly available via the university recommended repositories at the end of the project. The GDPR will be adhered to with respect to data sharing.

Research on multiple pseudo-anonymous samples will be reported through peer reviewed academic journals. The findings will be made public through public engagement events. There is no public database for registering this form of research.

To comply with the Open Access Policy, all published research will be made publicly accessible via green or gold access and a Data Access Statement will be made available with the publication. Journal restrictions will be adhered to.

Only anonymized data will be shared to the public.

5. Conclusions

Ethical considerations for our data are all covered in the TRANSTEM Deliverables 9.1 - 9.3 that are due at M10 of the project. The DMP is panned to be a living document - it will be updated as soon as more details are available and a more detailed and elaborated version of the DMP will be delivered at later stages of the project.