Updated: 2019-02-26

Data Management Plan

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the partners with regard to all the data sets that will be generated by the project.

The DMP is not a fixed document, but evolves during the lifespan of the project. The DMP addresses the points below on a dataset by dataset basis and reflects the current status of reflection within the consortium about the data that will be produced. The DMP follows both H2020 and University of Helsinki research data policy guidelines.

Data set reference and name

1) Clinical and sample data for project HERCULES
2) Sequencing data (2a) bulk and 2b) single-cell) for project HERCULES
3) Cytometry (FACS and mass cytometry) data for project HERCULES
4) Drug screening data for project HERCULES
5) Histopathological image data for project HERCULES

Data set description

Description of the data that will be generated or collected, its origin, nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

For all datasets:
The purpose of the collection of the data is to publish the research results, however the datasets themselves will only be accessible to the researchers within the project due to patient confidentiality and IPR issues.

Data set description for 1) clinical and sample data

Patient clinical data are collected from electronic hospital records, surgical operation notes, and sample collection forms. The information is gathered partly manually and partly with data exports received from Centre for Clinical Informatics in the Turku University Hospital. Clinical data include several fields, such as patient characteristics, surgical procedures, imaging results, chemotherapy and other treatment cycles given to patient, blood sample results and information on patient’s treatment outcome and survival.

Data are collected systematically to an ovarian cancer database "Ovcabase" (File Maker Pro 17.0 Advanced software) stored at the Turku University Hospital server. Data fields have been constructed specifically for the collection of categorical and numeric data. Data on the number of collected tissue and blood samples, sample details and storage are collected to the same database. The data are utilized for analysis in correlation of the experimental data to produce scientific

publications of the consortium. The database consists of a single File Maker Pro 17 file approximately 5MB in size, and there are currently 1100 data fields and 140 000 records.

Data set description for 2) sequencing data, 3) cytometry data, 4) drug screening data and 5) histopathological imaging data

The data collected for data sets 2-5 is instrument measured data in digital formats.

Data calculated for 100 patients, which is a rough estimate for enrollment of high-grade serous patients.

2) Sequencing:

> 2a) 1-15 samples per patient including paraffin embedded and frozen tumor and normal tissue, blood, plasma and patient derived and commercial cell lines. On average 600 Gb per sample, 6 samples per patient, 600x6x100 = 360 TB.
>
> 2b) Expression quantification of living cells, 4 samples / patient, 80Gb / sample; 80GB*4*100 = 32TB

3) Cytometry: instrument-generated data in digital form; 4 samples / patient, 30Mb / sample; 30MB*4*100 = 12Gb (average size of 5 GB each in raw format)

4) Drug screening: Imaging and plate reader data, 30 GB per patient = 3T

5) Histopathological imaging: 4 slides / patient, 2GB / slide; 2GB*4*100 = 800GB

Standards and metadata

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.

Standards and metadata for 1) clinical and sample data

Clinical data is collected with current clinical standards. FIGO staging system 2014 for ovarian cancer is used to determine disease spread. Treatment response is evaluated with RECIST criteria 1.1. and GCIG criteria for imaging studies and serum CA 125 values. The data fields are named unequivocally, and data dictionaries on key fields are constructed for exported data to explain variable names and abbreviations to data end users. The constructed data dictionaries are available in the project intranet (wiki.helsinki), as well as in the E-duuni service environment through which the data exports are shared.

The process of sample and data collection in surgery has been documented. Metadata on e.g. numbers of collected samples is visible in the database. Standard file formats include .fmp (File Maker) and .xlsx, .csv.

Standards and metadata for 2) sequencing, 3) cytometry data, 4) drug screening data and 5) histopathological imaging data

Data fields in measurements use industry standard naming schemes. Data collection methods are documented, or follow instrument manufacturer specifications. Sample names used in sequence data files are anonymous and include only a running patient number, a unique identifier for the original tissue sample and the type of a sample (tissue, cell line, DNA, RNA etc).

Data formats are instrument specific, and if not accessible by generally accessible software, then converted to standard file formats as seen convenient.

Standard file formats used:

- 2) FASTQ, Bam alignment files (http://genome.sph.umich.edu/wiki/BAM), standard Variant Call Format (VCF and gVCF, http://vcftools.sourceforge.net), allele specific copy number calls (https://www.crick.ac.uk/peter-van-loo/software/ASCAT), standard ANNOVAR format for functional annotation of genetic variants (http://annovar.openbioinformatics.org/), CSV Comma Separated Values, XLS Excel Spreadsheets spesific tissue samples.
- 3) FCS 2.0 Flow Cytometry Standard
- 4) CSV Comma Separated Values, XLS Excel Spreadsheets
- 5) TIFF Tagged Image File Format, MRXS Mirax Virtual Slide File

Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).

In case the dataset cannot be shared, the reasons for this are mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

Data sharing for 1) clinical and sample data

All data in the clinical database "Ovcabase" is available to key personnel in the HERCULES responsible for the data collection. Ovcabase is double-secured in the Turku University Hospital server, with limited access with Turku University Hospital account and a personalized access code to the actual database. All the persons who have access to Ovcabase have agreed on confidentiality by the Turku University Hospital. The access codes are available to only limited numbers of personnel, and access is granted by Turku University Hospital Head of department and executed by persons responsible for the database (Johanna Hynninen, Veli-Matti isoviita). In order to protect patient privacy, study patients names or personal identification codes are visible only to persons involved in patient recruitment or tissue/blood sampling. Other users handle data with pseudonyms.

The clinical and sample data are available to HERCULES group members. Access to both samples and the data related to samples are limited to partners included in present project, not to other groups, studies or purposes.

The data are exported from the database by responsible investigators. The exported/shared data are pseudonymous; codes are used in order to secure patient identity. All kinds of clinical and experimental data are available, except for information that identifies the patient. Related clinical data is shared after identification of specific samples/details via internal networks. The exported clinical data are stored at the E-duuni (https://info.eduuni.fi). Eduuni is a collaboration service environment for flexible and secure collaboration across organization and ecosystem boundaries provided by the CSC IT center for Science. The Eduuni service environment is maintained by the Finnish state security regulation increased level (Vahti 2/2010), which is ensured with regular audits by CSC and external auditor. On these basis E-duuni service environment can be used for material that is in protection level IV (Restricted).

Access to the data in E-duuni is granted and controlled by the project manager (Tiia Pelkonen). Access is granted to members of the groups contributing to the project after they have signed a confidentiality agreement and sent it to the project manager.

The data are exported to E-duuni from the database by regular (3 monthly) basis or upon request.

Clinical data include sensitive personal data. The Hospital District of Southwest Finland (HDSWF) is the responsible party for lawful data management. Sample and clinical data are owned by the HDSWF. The clinical investigators of Seija Grénman's group (Turku University Hospital, Dept. of Obstetrics and Gynecology) define the Clinical data that is shared. Data sharing principles have been distributed and agreed upon by all members in the consortium.

Data sharing for 2) Sequencing data (bulk and single-cell) for project HERCULES

Data are stored into files which are labeled with sample identification number. Logical organization of the files is ensured by progressive sample identification number and date of acquisition. Access to data is restricted to persons involved in the project. The rights to use the data are regulated by the HERCULES Consortium Agreement. Raw sequencing data will be used and analyzed mainly by Sampsa Hautaniemi's group at University of Helsinki. Files are generated in a standard non-proprietary format. Data can be shared to other HERCULES partners via secure transfer of files after all personnel involved in the data processing have signed the project NDA and data handling instructions.

Data sharing for 3) Cytometry (FACS and mass cytometry) data for project HERCULES

Data are stored into files which are labeled with sample identification number. Logical organization of the files is ensured by progressive sample identification number and date of acquisition. Access to data containing PCs is restricted by using password available only to persons involved in the project. The rights to use the data are regulated by the HERCULES Consortium Agreement. Data will be used by the Istituto Superiore di Sanita group and by Sampsa Hautaniemi's group. Files are generated in a standard non-proprietary format.

Data sharing for 4) Drug screening data for project HERCULES

Primary measurement data from drug screening (TIFF image files and XLS Excel spreadsheet files) as well as processed image data (CSV comma separated value or XLS Excel spreadsheet files) will be stored at servers at Institute of Molecular Medicine Finland (FIMM) at University

of Helsinki. Primary data can be shared to other HERCULES partners. Secure transfer of files may be organized by request.

Fully processed drug response data (dose response data for identified clonal responses) can be made available to all HERCULES partners through E-duuni in the form of XLS Excel spreadsheet files.

Data sharing for 5) Histopathological image data for project HERCULES

Data from measurement instruments include the numerical data from the measured sample, and the sample identifier to connect the measurements to clinical information.

Data from measurement instruments is stored on Sampsa Hautaniemi's group's storage servers. The storage servers are only accessible by Sampsa Hautaniemi's group employees, who are also the main users of the data. Anyone accessing the data must agree on the group's data security principles.

Measurement data can be shared to other HERCULES partners via secure transfer of files.

File formats for these data are industry standards, and they all have free readers available.

All data sets:
The samples and all derivatives of them are owned at all times by the Hospital district of Southwest Finland as according to the Material Transfer Agreements made between the project participants; ownership of the research results is determined in the HERCULES consortium agreement (results are owned by the party/parties that produce them). None of the data sets will be shared publicly at this stage due to ethical (patient confidentiality) reasons. All patients participating in the study sign a consent form for use of their samples for research purposes and are informed that the data collected about them will be processed anonymously so that they cannot be identified from the data; the consent and information forms have been approved by the Ethical Committee of Hospital Districts of Southwest Finland. In addition, commercially exploitable results are expected from the project, therefore any potential IPR issues must be resolved before considering publication of the data.

Archiving and preservation

Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

Archiving and preservation for 1) clinical and sample data

The database is stored physically in the Turku University Hospital server, in Turku for at least as long as the project is ongoing. The data are intended to be stored for at least 10 years after the project is finished and are expected to be available for research use also during that time, depending on ethical permissions. Funding for the long term storage has been secured. After completion of the project, the stored blood and tissue samples will be transferred to Auria Biobank with the patient consent (Informed consent by Auria Biobank). After transfer, access to

the samples will be controlled by the Biobank and can be applied by other researchers for other projects according to Auria Biobanks application procedure.

Backups of the database server are done once every day, and the backed up files are saved for 2 months during which they can be restored.

The collected paper documents are stored in a locked cabinet in study nurses office in Turku University Hospital. They will be destroyed ten years after completion of the project.

Archiving and preservation for 2) sequencing data

Data is received on external hard drives, which are stored as a backup in locked cabinets. Data from measurement instruments are transferred to servers with high hardware failure tolerance.The large data servers are not backed up due to high costs. The planned life cycle time of the storage is the duration of the project, with extension capabilities. At the end of the project, the sequencing data will be stored into Finnish genome repositories (biobank or genome center) for those patients who have given a biobank consent, where it will be available to the research community according to biobank regulations.

Archiving and preservation for 3) cytometry data

Data will be periodically exported to suitable memory storage units to generate backup copies. Storage of data is ensured for 10 years after the completion of the project. No specific funding is necessary for data storage.

Archiving and preservation for 4) drug screening data

Primary measurement data are stored on back-up servers. After the completion of the project, the data is planned to be archived as compressed image files to save space.

Archiving and preservation for 5) histopathological imaging data

Data from measurement instruments are stored on servers with high hardware failure tolerance. The large data stores are not backed up due to high costs. The life cycle time of the storage is set to 5 years from now, with extension capabilities adding at least 10 years more.