PRENATAL MICROBIAL IMPACTS ON THE DEVELOPMENT OF THE INTESTINAL IMMUNE SYSTEM

Data Management Plan created using DMPTuuli
Creator: Mikael Niku, University of Helsinki (ORCID ID: 0000-0003-1087-9859), 23.9.2019

1. GENERAL DESCRIPTION OF DATA

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Give a rough estimate of the size of the data.

Data type	Source	Format	Size
high-throughput sequencing data: RNA-Seq, virome, sequencing, 16S rRNA gene amplicon sequencing, immunoglobulin sequencing	produced in the project	standard fastq and fasta sequence data files, and csv files	1-100 Gb / dataset, total ~0,5 Tb
previously published sequencing data: 16S rRNA gene reference sequences, immunoglobulin reference sequences	reused from public databases	standard fasta sequence data files	
metabolomics data (nontargeted and targeted LC-MS)	produced in the project	Excel, converted to csv files for long term storage; original raw data is processed by the core facility	100 Mb
imaging data (microscopic imaging of immunohisto-chemical and immunofluorescence analyses)	produced in the project	TIFF and JPEG image files and MIRAX histology slide scanner files; metadata stored in an Access relational database	1 Tb
flow cytometry data	produced in the project	Excel, converted to csv files for long term storage	100 Mb
laboratory notebooks and electronic laboratory notes.	produced in the project	paper notebooks, stored in the laboratory archives; digital notebook formats, converted to pdf for long term storage	1 Gb

1.2 How will the consistency and quality of data be controlled?

The following challenges to the data integrity and quality can be identified:

- Data is often produced as a collaborative effort of experts of different topics, with limited expertise on the other persons' work.
- Bioinformatic processing involves complex steps and transformations of large data volumes between formats.
- Image data is often poorly structured, as it consists of large numbers of individual image files produced by various imaging instruments.

We ensure the integrity and quality by the following procedures:

- Structured data (such as relational databases) and jointly agreed conventions (file system structures, file naming practices) are utilized.
- The PI ensures that all personnel is sufficiently trained in methodology and the laboratory protocols.
- All results are discussed and reviewed in research team meetings.
- Data is collected and processed using established standard operating procedures, where available, or other published methods, as described in the Research plan.
- High-throughput data is generated in established high-quality core facilities.
- Checksums are used when data is transferred.

2. ETHICAL AND LEGAL COMPLIANCE

2.1 What ethical issues are related to your data management, for example, in handling sensitive data, protecting the identity of participants, or gaining consent for data sharing?

The project does not use data originating from humans. Privately owned animals are included in the research material. The information regarding the ownership of animals is not relevant to the research and is not included in the research data. The research data does include confidential or sensitive information.
The ethical committee of the Viikki campus has approved the animal sampling performed by ourselves. The collaborators and animal core facilities have obtained animal experimentation permits for their operations.

2.2 How will data ownership, copyright and IPR issues be managed? Are there any copyrights, licences or other restrictions that prevent you from using or sharing the data?

The PI concludes contracts of ownership and rights of use of data within the research team and with the collaborators, with the support of the university legal advisors. The publication policy is thoroughly discussed with team members and collaboration partners involved in data generation, with signed agreements. A separate agreement has been signed with the collaborating commercial company, specifying the use of research data and commercial IPRs.
The existing data required in the project is available through public repositories.

3. DOCUMENTATION AND METADATA

3.1 How will you document your data to make them findable, accessible, interoperable and reusable for you and others? What kinds of metadata standards, README files or other documentation will you use to help others understand and use your data?

- High-throughput sequencing generates data in standardized formats and requires consistent metadata from the beginning. The original raw sequencing data files contain individual identifiers for each biological sample of origin. The data files are stored together with csv files containing all collected metadata (in a structured format) from the individual animals and the identifiers connecting with the sequence data files, as well as brief description of the project and data ownership and the metadata variables. The raw data from each subproject is stored as a compressed tar archive using a filename identifying the project.

- The bioinformatics pipelines used in the project automatically generate detailed log files and modify the file names, thus unequivocally documenting each processing step and version.

- The pre-processed sequencing data (such as OTU/ASV and genus tables generated from 16S data) and essential output files from bioinformatic and statistical analyses are stored as csv files, together with the metadata files.

- The metabolomics data is formatted and documented by the service provider.

- The image files are identified by their filenames. The filename contains an individual identifier connecting the image file to a relational database which we have built to store the metadata for each histological slide (such as antibodies and protocols used in immunostaining) and for each tissue section on the slides (animal and tissue of origin and preservation protocols), as well as brief description of the imaging protocol. The database includes identifiers for each animal, and the metadata for the animals is stored in another relational database. The databases are built using Microsoft Access.

- Flow cytometry data and other types of data are stored as Excel and csv files containing sufficient metadata.

4. STORAGE AND BACKUP DURING THE RESEARCH PROJECT

4.1 Where will your data be stored, and how will they be backed up?

Most of our data is stored on the network drives provided by the university, with automated backup, and processed either on local workstations or on the servers of the Finnish IT Center for Science (CSC). After each subproject, the data is preserved as described later.

The virome sequencing data is stored during the subproject in the Illumina cloud storage provided by the sequencing instrument company.

Large-scale MIRAX images (1-10Gb each) produced by digital slide scanners are stored on local workstations during the analyses, and backed up on portable hard disk drives, as they do not contain any sensitive data. The original histological sections are stored as the hard copy original data, reducing

the requirements for digital storage security. Selected subsets of the scanner images are also stored in the digital format on the university network drives and submitted to public data repositories.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

The PI will control access to the data. The access is granted to all research group members, and to selected collaborators where necessary.
All workstations, network drives and servers are password protected. Access to the university network drives and CSC storage services is only granted to research team members. Local workstations, portable hard disks and physical sample collections are located in offices which are locked when not in use. These practices are sufficient as we do not generate sensitive data.

5. OPENING, PUBLISHING AND ARCHIVING THE DATA AFTER THE RESEARCH PROJECT

5.1 What part of the data can be made openly available or published? Where and when will the data, or their metadata, be made available?

All relevant data (primarily metabolomics and high-throughput sequencing data) will be made openly available upon publication, using established field-specific public repositories, according to their recommendations. The published data is linked to the publications.
Comprehensive metadata is provided to allow efficient re-use of the data. The services use persistent identifiers or local identifiers of the data archive.
The next-generation sequencing data is deposited in the European Nucleotide Archive (ENA). Metabolomics data is deposited in the EMBL-EBI MetaboLights. The other types of potentially re-usable data are submitted to the European OpenAire funded research data archive Zenodo.

5.2 Where will data with long-term value be archived, and for how long?

The metabolomics data, most of the high-throughput sequencing data and selected imaging data have long-term reuse value. The policies and repositories described above ensure the preservation of data as long as it is practically relevant.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6.1. Who will be responsible for specific tasks of data management during the research project life cycle? Estimate also the resources (e.g. financial, time and effort) required for data management.

Each researcher is responsible for documentation and secure storage of data generated by her/him, as instructed by the PI. The PI is responsible for data publication, data protection and information security. Experts of the university research data management service are consulted to ensure proper practices. Our primary data types are by nature consistently formatted and documented throughout the workflow, minimizing the extra effort required for publication and archiving. The additional documentation and re-formatting required by the public repositories takes an maximum average of one workday per a dataset generated in a subproject (total ~2 work weeks for the entire project).

The estimated costs of data management are included in the project budget. We mostly utilize the computing resources and storage services provided free to researchers by CSC and University of Helsinki. Additional services can be obtained from these organizations at moderate costs. The time and personnel resources needed for metadata preparation are intimately included in our routine processes, as this is required by the public data repositories we are asked to use by most journals in our field.