1.1 What data will you collect, observe, generate or reuse?

Questions you might want to consider:

- What type, format and volume of data will you collect, observe, generate or reuse?
- Which existing data (yours or third-party) will you reuse?
- Briefly describe the data you will collect, observe or generate. Also mention any existing data that will be (re)used. The descriptions should include the type, format and content of each dataset. Furthermore, provide an estimation of the volume of the generated data sets. (This relates to the FAIR Data Principles F2, I3, R1 & R1.2.)

This project will generate different types of digital data of important size, which will be produced using several analysis devices:

1. RNA sequencing analysis data will be generated and saved as .bam, .gz and/or .txt files, for a total of about 150 GB.
2. Flow cytometry data will be generated and saved as .fcs data that include analysis settings, for a total of about 100 GB.
3. Movies and pictures will be generated using a miniature endoscope purchased in the frame of the SNSF project #XXXX (*XXXXX*) and saved as .mpg, .jpeg or .tiff data, for a total of about 50 GB.
4. Histology sections will be scanned for analysis using different scanners and saved as .mrxs or .dat data for a total of about 50 GB
5. Metabolomics data will be generated at the Functional Genomics Center Zurich (FGCZ) in a NetCDF format and processed in a tabular text format. We estimate that a total data volume ranging between 25-30 GB will be generated during the course of the project.

Other derived data (measurements, quantifications, graphical representations, statistical analysis) are not expected to exceed 100MB. These data include among others data in spreadsheets (stored as .csv or .xlsx), data in freetext documents (stored as .txt, .docx, or .pdf), data represented as graphs, with linked statistical analysis between groups (stored as pzfx.).

In this project, we may reuse RNA sequencing analysis data generated within the SNSF project #XXXX (*XXXX*) and that have been saved as .bam, .gz and/or .txt files, for a total of 50 GB.
We may also reuse public data from The Cancer Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/docs/publications/coadread_2012/; RNA sequencing data from XXXX), which we also used in the frame of the SNSF project #XXXX.

Metabolomic data: we will generate experimental mass spectra and in-silico computer generated mass spectra as well as chromatograms. We will include standardization and quantification data. The raw data files will be processed by Progenesis QI software (Nonlinear Dynamics / Waters) and / or will be converted to NetCDF format and processed using cosmiq (FGCZ, available at Bioconductor, https://bioconductor.org/packages/release/bioc/html/cosmiq.html).
Raw metabolomic data will be stored in original and converted formats. Processed data will be stored in a tabular text format.

1.2 How will the data be collected, observed or generated?

Questions you might want to consider:

- What standards, methodologies or quality assurance processes will you use?
- How will you organize your files and handle versioning?
- Explain how the data will be collected, observed or generated. Describe how you plan to control and document the consistency and quality of the collected data: calibration processes, repeated measurements, data recording standards, usage of controlled vocabularies, data entry validation, data peer review, etc. Discuss how the data management will be handled during the project, mentioning for example naming conventions, version control and folder structures. (This relates to the FAIR Data Principle R1.)

All samples from which data is to be collected will be prepared according to published standard protocols in the field of immunology/microbiology/cellular biology or after optimizing them for the specific needs of the project. All the instruments that will be used including sequencers, flow cytometers, scanners, endoscopes, etc. are regularly serviced, calibrated or controlled.
Cell lines or bacteria strains have been or will be purchased from repositories or authentified by STRs analysis or specific sequencing.
Mice used in these experiments are maintained according to Swiss regulations, in an animal facility approved by the Cantonal Veterinary Authorities of Bern. In particular, state-of-the-art standards of care will be provided to ensure optimal and species-specific housing and environmental conditions, as well as daily clinical wellbeing and health monitoring by dedicated staff and quarterly microbiological screening according to the FELASA recommendations. All genetically modified mouse lines used in experiments are regularly genotyped and assessed for phenotyping traits if this applies.
Experiments will be conducted using $n \geq 3$ per group and repeated at least 3 times. Proper negative and positive controls will be run in each experiment. Experiments and protocol will be recorded in laboratory journals or described on a text document that will be stored on the server of our Institute, which is backed up every day.
If required for the study, patient-derived samples will be obtained via the Tissue Bank Bern (TBB). The TBB fulfils the Vita standard from the Swiss Biobanking Platform (SBP) and has interfaces accredited with the ISO/IEC 17025 und ISO 15189 standards.
Files and dataset will be named with information including date and name or number of the experiment. Any deviation from the protocol will be annotated and mentioned in the above-described text document. Microscope images capture a range of metadata (field size, magnification, lens phase) with each image.
Sequencing data will be performed according to the standard of the Next Generation Sequencing (NGS) Platform of the University of Bern.
Mice experiments will be conducted according to the Animal Protection Act and the Animal Protection Ordinance, after protocol approval by the Cantonal Veterinary Office. To minimize intra- and inter-experiment variations, we will always strive to compare gender and aged-matched animals within an experiment and between several experiments that are similar.

Files will be named according to a pre-agreed naming conventions in the group.
For each experiment, a description of the experiment design, the raw data with the metadata of each sample, an analysis and a graphical representation of the data will all be saved in an experiment subfolder of our research group folder on the server of our Institute, which is backed up every day. This should allow the data to be understood by other members of our research group and add contextual value to the dataset if it should be reused in the future.

For metabolomics data:
All samples will be managed at the FGCZ using the Bb-Fabric platform (Türker, C., Akal, F., & Schlapbach, R. (2011). Life sciences data and application integration with B-Fabric. Journal of Integrative Bioinformatics, 8(2)). B-Fabric is the web-based management system for projects, samples and data, that is developed and operated by the FGCZ. The B-Fabric data system manages also the unambiguous naming and placing of data files with unique identifiers.
All raw data will be generated at the FGCZ respecting the applicable standard operating procedures (SOPs). The chromatography systems and mass spectrometers will be operated by expert personnel. Raw data will be imported directly from the instrument workstations to B-Fabric and are accessible by all project members by secure internet or intranet connections.
Quality of analytical data will be guaranteed through calibration of devices and comparison with internal standards.
Appropriate experimental design, data recording and data validation (controls, randomization/blinding, sampling/replicates, experimental versus hypothesis driven-protocol) ensuring internal validity.
The produced raw data will be processed using the following tools: Progenesis QI (Nonlinear Dynamics / Waters), cosmiq (FGCZ available at Bioconductor, https://bioconductor.org/packages/release/bioc/html/cosmiq.html), Quanbrowser (Xcalibur, Thermo Fisher) or QuanLynx (MassLynx ,Waters). Processed raw data will be analysed by using custom R scripts and open source software including Lipid Data Analyser, mummichog and MetaboAnalyst as well as defined spreadsheet files.
The version of each software and library will be documented for each set of results in B-Fabric. Analysis scripts, if applied, will be version controlled with using FGCZ's git server (https://gitlab.bfabric.org)
R Markdown / Knitr will be used to generate reproducible analysis reports.
All the aforementioned steps should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future.

1.3 What documentation and metadata will you provide with the data?

Questions you might want to consider:

- What information is required for computers or humans to read and interpret the data in the future?
- How will you generate this documentation?
- What community standards (if any) will be used to annotate the (meta)data?
- Describe all types of documentation (README files, metadata, etc.) you will provide to help secondary users to understand and reuse your data.
  Metadata should at least include basic details allowing other users (computer or human) to find the data. This includes at least a name and a persistent identifier for each file, the name of the person who collected or contributed to the data, the date of collection and the conditions to access the data. Furthermore, the documentation may include details on the methodology used, information about the performed processing and analytical steps, variable definitions, references to vocabularies used, as well as units of measurement. Wherever possible, the documentation should follow existing community standards and guidelines. Explain how you will prepare and share this information. (This relates to the FAIR Data Principles I1, I2, I3, R1, R1.2 & R1.3.)

Files and dataset will be named with information including date, name and number of the experiments.
Data will be classified and stored in separate folders that follow a defined organization or structure for each method or subproject, for each staff member. Within each method or subproject, folders named with the method and date of experiment will be prepared. All files (raw and analysis) for the respective experiment will be named with the method or subproject and the date of the experiment and saved in the respective folder.
Methods and dates of experiments are reported in the lab journal book and are therefore linked with the data. Any deviation from the protocol will be annotated.
Sequencing files will be saved according to the barcode related to each run. A metadata including a description of all the available data and the location of the data will be made. The metadata will include: name and identifier of each file, name of the person who collected the data and executed the experiment, date and methods.

Metabolomic data:
B-Fabric keeps track of all samples and meta-information regarding sample treatment and processing in the laboratory. Any data analysis results will be imported in b-Fabric and will be documented, reproducible and reusable, since the results are always accompanied with the analysis scripts, version numbers of the software used and the parameters that were used.
The final datasets will be deposited additionally in EMBL's public repository MetaboLights.

2.1 How will ethical issues be addressed and handled?

Questions you might want to consider:

- What is the relevant protection standard for your data? Are you bound by a confidentiality agreement?
- Do you have the necessary permission to obtain, process, preserve and share the data? Have the people whose data you are using been informed or did they give their consent?
- What methods will you use to ensure the protection of personal or other sensitive data?
- Ethical issues in research projects demand for an adaptation of research data management practices, e.g. how data is stored, who can access/reuse the data and how long the data is stored. Methods to manage ethical concerns may include: anonymization of data; gain approval by ethics committees; formal consent agreements. You should outline that all ethical issues in your project have been identified, including the corresponding measures in data management.

Provided patient-derived material is necessary, these samples will be collected according to the Human Research Act (HRA) and Human Research Ordinance (HRO), and only for those patients who have signed an institutional general consent. Ethical approval to collect patient-derived material has been given by the Cantonal Ethics Committee of Bern (permissions no xxxx and xxxx). Patient data will be coded and stored in a database conform to the Human Research Act (HRA) and Human Research Ordinance (HRO), and as described in the document approved by the Cantonal Ethics Committee of Bern.
The IT system of our Institute ensures the safety of the data according to ISO-standards ISO 15189 und ISO/IEC 17025.
Mice experiments will be conducted according to the Animal Protection Act and the Animal Protection Ordinance under the approval of the Cantonal Veterinary Office of Bern. The experiments will be conducted under the current licence numbers XXX and XXX, which include the type of the experiments described in the project.

2.2 How will data access and security be managed?

Questions you might want to consider:

- What are the main concerns regarding data security, what are the levels of risk and what measures are in place to handle security risks?
- How will you regulate data access rights/permissions to ensure the security of the data?
- How will personal or other sensitive data be handled to ensure safe data storage and -transfer?
- If you work with personal or other sensitive data you should outline the security measures in order to protect the data. Please list formal standards which will be adopted in your study. An example is ISO 27001-Information security management. Furthermore, describe the main processes or facilities for storage and processing of personal or other sensitive data.

Data are saved on internal server of our Institute. Access to the server is limited only to the people involved in the project; it is fully managed and audited by our IT Department every six months. The IT infrastructure of our Institute ensures the safety of the data according to ISO-standard ISO 15189 und ISO/IEC 17025.
Since all the human-derived samples and patient information will be obtained via the Tissue Bank Bern (TBB), the TBB staff will assign a code to the received specimen, for the purpose of linking the specimen's patient-related data to this study. Only this code, which per se allows no deduction of any patient-related data, is made available to the study investigators. All code keys will be maintained by the TBB and researcher will only receive coded data. This way of proceeding is in agreement with the Human Research Act (HRA) and Human Research Ordinance (HRO), and has been approved by the Cantonal Ethics Committee of Bern.

Metabolomic data: Data access and security will be managed by B-Fabric. B-Fabric provides project-specific access to databases and data files. Additionally within each project permission are granted based on the user and his role in the project.
All data is physically stored by the University of Zurich and access is managed by FGCZ personnel.

2.3 How will you handle copyright and Intellectual Property Rights issues?

Questions you might want to consider:

- Who will be the owner of the data?
- Which licenses will be applied to the data?
- What restrictions apply to the reuse of third-party data?
- Outline the owners of the copyright and Intellectual Property Right (IPR) of all data that will be collected and generated, including the licence(s). For consortia, an IPR ownership agreement might be necessary. You should comply with relevant funder, institutional, departmental or group policies on copyright or IPR. Furthermore, clarify what permissions are required should third-party data be reused

Copyrights belong to the University of Bern. Any findings that generate intellectual property or that are relevant for commercial applications will be handled by Unitectra, which manages the IP for the University of Bern. Unitectra is a non-profit organization focused on issues related to technology transfer, which provides this IP-related support to the University of Bern.
Data will be available under a Creative Commons Attribution-Share-Alike (CC BY-SA) license.
Data from the TCGA data portal will handled according to the requested use policies and publications guidelines for investigators.
Data from the Greengenes and Ribosomal Database Project (RDP) databases will be retrieved under a CC BY-SA license and be cited in publications according to the requested publications guidelines for investigators.

3.1 How will your data be stored and backed-up during the research?

Questions you might want to consider:

- What are your storage capacity and where will the data be stored?
- What are the back-up procedures?
- Please mention what the needs are in terms of data storage and where the data will be stored. Please consider that data storage on laptops or hard drives, for example, is risky. Storage through IT teams is safer. If external services are asked for, it is important that this does not conflict with the policy of each entity involved in the project, especially concerning the issue of sensitive data. Please specify your back-up procedure (frequency of updates, responsibilities, automatic/manual process, security measures, etc.

All digital data is daily backed up on the server of our Institute, and replicated on an off-site location. The capacity is expandable to the needs.

Metabolomic data: All data will be stored in B-Fabric and all data files will be stored using the storage system provided by the University of Zurich. All disks and databases are backed up daily. The FGCZ has access to 400 TB of extendable storage.

3.2 What is your data preservation plan?

Questions you might want to consider:

- What procedures would be used to select data to be preserved?
- What file formats will be used for preservation?
- Please specify which data will be retained, shared and archived after the completion of the project and the corresponding data selection procedure (e.g. long-term value, potential value for reuse, obligations to destroy some data, etc.). Please outline a long-term preservation plan for the datasets beyond the life-time of the project. In particular, comment on the choice of file formats and the use of community standards.

Sequencing data will be shared upon publication of the corresponding manuscript on publicly accessible platforms, including the NCBI Gene Expression Omnibus (GEO) repository, the European Nucleotide Archive (ENA), or the ArrayExpress Archive of Functional Genomics Data from the European Bioinformatics Institute (EMBL-EBI). These platforms allow data to be publicly searchable. The URLs associated with the datasets in the repository will be included as part of a data citation in publications, allowing the datasets underpinning a manuscript or publication to be identified and accessed.
Sequencing data will be stored for long-term to allow re-use. Raw data will be stored and preserved for at least 10 years after publication in raw data files as described in 1.1. After expiration of the working contract of researcher, all electronic data will be mirrored on a location of the Institute server, which is accessible by the P.I. and the group members working on follow-up projects.
Raw data from sequencing analyses will be stored as in the standard format of these data as bam, .gz and/or as .txt files and made public in data portals including GEO, ENA, or ArrayExpress.

Metabolomic data will be deposited in the public Metabolights repository at EMBL. The final datasets as deposited in the chosen data repositories will be accompanied by the relevant metadata documentation.
Major revisions of data analysis scripts will be stored in B-Fabric and/or in Zenodo with a stable DOI. B-Fabric will provide long term storage of meta-information associated with the project. Bad quality data will be permanently discarded at the end of the project.
Metabolomic data will be preserved as raw and additionaly in NetCDF and/or mzXML file types.

4.1 How and where will the data be shared?

Questions you might want to consider

- On which repository do you plan to share your data?
- How will potential users find out about your data?
- Consider how and on which repository the data will be made available. The methods applied to data sharing will depend on several factors such as the type, size, complexity and sensitivity of data. Please also consider how the reuse of your data will be valued and acknowledged by other researchers.

Data will be shared upon publication of the corresponding manuscript. Data from sequencing analyses will be made public in data portals including GEO, ENA, or ArrayExpress.
Metabolomic data will be deposited in MetaboLights (https://www.ebi.ac.uk/metabolights).
For other types of published results, and whenever it applies, corresponding data necessary for the better understanding of these results will be uploaded on the open-access repository Zenodo (https://zenodo.org/). In dependence of their nature and according to the standards in

the field, these data may be provided as an annotated list of raw values, as normalized data or in a processed form. Alternatively, we may use for this aim more specific repositories that are listed in the Registry of Research Data Repositories (Re3data, https://www.re3data.org/), provided they are compatible with the FAIR Guiding Principles.
Information about and links to these different repositories will be provided in the published manuscripts.

4.2 Are there any necessary limitations to protect sensitive data?

Questions you might want to consider:

- Under which conditions will the data be made available (timing of data release, reason for delay if applicable)?
- Data have to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output. Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.

At the time of manuscript submission, data will be made available to peer-reviewers using a code.
Data will be made publicly available at the time of publication of a corresponding manuscript. We do not expect any specific restriction or limitation for the data generated to be made publicly available.