Plan Overview

A Data Management Plan created using DMPonline

Title: Proteogenomics of local recurrences in breast cancer

Creator: Tommaso De Marchi

Principal Investigator: Emma Niméus

Data Manager: Tommaso De Marchi

Project Administrator: Tommaso De Marchi, Emma Niméus

Affiliation: Lund University

Template: LU standard template for data management plans

ORCID iD: 0000-0003-0262-9083

Project abstract:

Breast cancer is the most common malignancy among women, and recurrent breast malignancies are its major cause of death due to therapy resistance and metastatic spread. Immunotherapies are an effective way of leveraging the patient own immune system to halt tumor progression and cure the disease, but are restricted by the low amount of targets discovered so far. Our group has developed several genomic and proteomic approaches to characterize recurrent breast cancers and detect their changes when compared to their primary tumors of origin. More recently, we have developed a combined proteogenomic approach to characterize recurrent breast cancers molecular features predictive of tumor evolution and prognosis. Additionally, we have implemented a method for the analysis of surface-exposed proteins on cancer cells. In this proposal, we aim to employ these approaches to define hypermutated cancers, detect cancer-specific epitopes expressed on the cancer cell surfaces, and use the obtained information to design new candidates for immunotherapy.

ID: 152919

Start date: 29-05-2024

End date: 31-12-2027

Last modified: 26-09-2024

Proteogenomics of local recurrences in breast cancer

1. Data description

1.1 What method of data collection will you be using?

- using existing data (archival data, data from previous project – your own or other researchers')

We plan to extract DNA and RNA from archival breast cancer tissues (and from normal tissue adjacent to each cancer) which are part of an archival collection, collected for a randomized clinical trial in Sweden (SweBCG91RT). All samples are stored in a biobank and no new collection is needed.

1.2 Describe how data will be collected, created or reused.

Every tissue sample - stored as a paraffin-embedded block - will be inspected by a resident pathology specialist. Here, areas of tumor and adjacent normal tissues will be marked and excised for DNA and RNA extraction using QIAGEN FFPE kits. Resulting DNA and RNA eluates will be processed at Lund University Center for Translational Genomics (part of SciLife laboratories) for DNA/RNA QC and sequencing (DNA: whole exome sequencing; RNA: RNA sequencing). Data will be then uploaded on a secure encrypted server. Backup with equal security level will also be put in place.

1.3 What type of material (physical or digital) will you use (e.g. text, images, measurement data)? In which file formats will you save your data?

Data will be primarily constituted of sequencing files (.fastq), as they are the resulting raw files from Illumina sequencers. Additional processed files (e.g. .mgf tables) will also be stored. Clinical metadata will be stored as text files.

1.4. According to your estimation, how large is the maximum storage capacity will you need throughout the project (primary data and revisions of processed data)?

- >1 TB

We plan to analyze a cohort of breast cancers by whole exome sequencing (WES) and RNA sequencing (RNAseq):
- primary breast tumors with no recurrence + normal adjacent tissues (NATs; n = 192 + 192)
- primary breast tumors with recurrence + NATs (n = 192 + 192)
- locally recurrent breast tumors + NATs (n = 192 +192)
TOTAL number of specimens: 1,152 (WES) + 1,152 (RNAseq)
WES data size: ~6 GB per sample (only raw data) --> ~20 TB total storage (including intermediate files, results, and metadata)
RNAseq data size: ~5 GB per sample --> ~15 TB total storage (including intermediate files, results, and metadata)
TOTAL allocation memory: ~35 TB

2. Documentation and data quality

2.1 How will your data/your material be documented and described with metadata, take collection

method, content, structure, standards and formats in consideration; in order for you and other researchers or computer software to read and be able to interpret the data correctly?

Each sequencing file will be connected to already collected metadata, that is clinical characteristics recorded at diagnosis and histo-pathological features used for diagnosis. Metadata is stored electronically as tabular text files.

Metadata will be employed to group sample sets and guide the analysis of specific subgroups of patients.

We have already obtained access to the metadata files of the sample cohort and already performed studies on the larger dataset (e.g. PMID: 36975842).

2.2 How will the quality of the research data be ensured and documented (for example by repeated measurements, data entry validation etc.)?

The metadata files containing the clinical and histo-pathological features of the tumors included in this study have already been verified and transcribed from official clinical records.

With regards to sequencing data, quality control prior sequencing analyses are routinely run for each sample by the Center for Translational Genomics. Additionally, fastQC is routinely used to assess sequencing data quality post hoc. These tests assess, among others, sequence quality, GC content, and sequence duplication levels. All of these results will be stored and employed for selection of only high quality results.

3. Storage and backup

3.1 How will you ensure integrity of storage and backup of data and metadata during the research process?

The NAISS SENS platform (Bianca and Kastor resources) will be employed for storage and processing during the research process. These are resources hosted by UPPMAX, part of the University of Uppsala. Encrypted backups will be set in place at Lund University through Lund Data Center and at UPPMAX (backup of Kastor resource).

Data managers (Lund University: Tommaso De Marchi, Aaron Scott, and Elisabeth Hjortswang; UPPMAX: to be defined) will manage the data and be responsible for backup at each location (Lund and Uppsala). For each file, a total of 3 copies will be kept up to results publication. After publication, a backed up copy will be kept within Lund University encrypted servers.

Data will be backed up once a year.

Version control will not be implemented.

3.2 How is information security and access to data controlled, for example in reference to sensitive data and personal data?

Data security measures will encompass the use of encryption for data storage and the use of two-factor authentication for authorized data managers.

Ethical Review Board authorizations (Diary numbers: LU-2010/127 and LU-2001/240) have been already obtained. Data will not be shared nor employed for any purpose other than the one stated in the research project aims.

Access to data will be highly selective, with only dataset administrators being granted access. Two-factor authentication and access monitoring will be put in place.

We do not plan to allow external collaborators to access the data. If the need arises, data access will be restricted (e.g. one-time password or restricted access time span) and will only be possible via UPPMAX. No copy of the data will be possible.

Data transfer to secure data storage after acquisition will be performed through secure protocols (e.g. FTPS).

4. Legal and ethical requirements

4.1 Will the project be processing personal data?

- Yes – in that case you must report this in Pulu (https://pulu.adm.lu.se)

PULU project name: "Randomized Clinical Trial - SweBCG-91-RT"

4.2. How will you ensure that data is processed according to the regulations concerning for example personal record handling, confidentiality and intellectual property rights?

Data will be only stored on secure encrypted location. Two-factor authentication will be available, but only to data managers.
Consent for use of specimens for research purposes was given at the time of specimen collection, thus no new consent request is needed.
Data will be pseudo anonymized for project leaders and anonymized for authorized data managers. Personal information such as personal numbers will not be accessibile.

4.3. In what way will you ensure that data is handled correctly from an ethical standpoint?

Data is under approval from the Ethical Review Board (Etikprövningsnämnden) with numbers DNR LU-2010/127 and LU-2001/240

4.4. Collaborative research projects involving external parties, may require an agreement between participants/principals of the study regarding processing, storage, ownership and aspects of intellectual property rights. Is this the case in your study?

- No, LU is the sole party

5. Data sharing and long-term preservation

5.1 Will research data and/or information on data (metadata) be made publicly available?

- No, restrictions do exist

As with previous projects, no personal information will be disclosed to the public or to the scientific community. Disclosed information will be non-personal and samples will be given anonymized IDs. Sequencing files will not be shared.
Following publication of results, a copy of the data will be kept on Lund University secure servers.

5.2 If so, how, when and where will data and/or metadata be made available? Are there any limitations (legal and/or ethical) that prevents sharing or reuse of it?

Sequencing data will not be shared at any moment.
Metadata may be shared upon request but always in anonymized form. This will be available upon publication of the dataset results in the form of a research paper. Contact information for data request will be published along with study results. Sensitive data will not be shared, while non-sensitive data will be handled directly.
We do not plan to have a PID for our data.

5.3 If you plan to make data/metadata publicly available, will you use a unique and persistent identifier (PID) such as a DOI?

Data will not be made available if not in anonymized form upon publication.

5.4 If data has been created or collected, is there a reason for keeping these forever or may they be destroyed after 10-20 years? What would the reasons be for preservation?

- My current assessment is that data probably may be destroyed.

Data will be stored securely since it constitutes a precious resource for the scientific community, but with constant technological advances it may become obsolete and thus be destroyed.

6. Responsibilities and resources

6.1 Who is responsible for the data management and assists with the data management during the project? Who is responsible for data management, keeping of records and long-term preservation, after the project finished?

The applicant (Tommaso De Marchi, Lund University), two bioinformaticians (Aaron Scott and Elisabeth Hjortswang, Lund University), and the designated person from NAISS SENS (University of Uppsala, to be defined) will be responsible for data management on each site, including storage and processing.
Data ownership rests with the PI, Emma Niméus.

6.2 What resources (cost, labor and miscellaneous costs) will be allocated to data management (including storage, backup, data sharing and long-time preservation preparation) within the project?

Resources available to the project:

- NAISS sens Bianca (processing) and Kastor (storage) will be employed. Bianca will be used continuously for data analysis, while Kastor just for storage. Usage will be terminated upon publication.
- Skilled personnel: Deputy data managers (Aaron Scott and Elisabeth Hjortswang) will be performing analyses and answer specific technical requests from NAISS platform or collaborators after previous consultation with the applicant and main data manager (Tommaso De Marchi). External technical support for using the platform will be provided by UPPMAX (e.g. storage encryption) while support for analyses will be provided (application pending) by the National Bionformatic Infrastructure Sweden (NBIS, pasrt of SciLife).
- No special equipment outside of encrypted servers is needed for the project.

Planned Research Outputs

Collection - "Whole exome sequencing data"

Analysis results from whole exome sequencing data will be organized in tabular format as follows:

- Single nucleotide variants: this table will contain information whether specific genes (ID: ENSEMBL, HUGO) are mutated in each tumor sample after filtering out the variants from normal tissue counterparts. Only cancer-specific mutations will be displayed.
- Structural and copy number variants: this table will contain gene copy number information when deviating from normal (i.e. cancer-specific). Gene inversions and translocations will also be included here.

Sample names will be anonymized so not to disclose personal information.

Collection - "RNA sequencing data"

Analysis results from RNA sequencing data will be organized in tabular format as follows:

- Single nucleotide variants: this table will contain information whether specific genes (ID: ENSEMBL, HUGO) are mutated in each tumor sample after filtering out the variants from normal tissue counterparts. Only cancer-specific mutations will be displayed.
- Summarized gene tables: summarized gene count tables (i.e. transcript abundance) will be included.

Sample names will be anonymized so not to disclose personal information.

Planned research output details

Title	DOI	Type	Release date	Access level	Repository(ies)	File size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Whole exome sequencing data		Collection	2025-07-30	Restricted	None specified	4 GB	Creative Commons Attribution 4.0 International	None specified	No	No
RNA sequencing data		Collection	2025-07-30	Restricted	None specified	1,000 MB	Creative Commons Attribution 4.0 International	None specified	No	No