Plan Overview

A Data Management Plan created using DMPonline

Title: Genetic causes and underlying mechanisms in metabolic bone diseases

Creator: Sakshi Vats

Principal Investigator: Outi Mäkitie

Contributor: Sakshi Vats

Affiliation: Karolinska Institutet

Funder: Swedish Research Council

Template: Swedish Research Council Template

ORCID iD: 0000-0002-4547-001X

Project abstract:

Metabolic bone diseases are a diverse group of conditions that affect bone strength due to disrupted bone homeostasis or imbalances in calcium, phosphate, or vitamin D metabolism. These conditions are frequently caused by genetic defects, which may be ultra-rare, and impair skeletal patterning, growth, or long-term maintenance

This long-term research program aims to identify novel genes and pathogenic variants associated with metabolic bone diseases and to investigate the disease-specific cellular and molecular mechanisms. We have access to a valuable collection of previously obtained genetic and clinical data from patients and families affected by these disorders. Affected patients and family members are also actively recruited through Karolinska University Hospital and collaborating national, international centers with data sharing agreements in place.

To uncover the genetic causes, we analyze human cohorts or samples using a broad range of high-throughput sequencing methodologies and integrative analyses. As a standard, we employ whole-genome sequencing (WGS) supported by clinical and family-based archival data. We also use methods like whole-exome sequencing (WES), transcriptomics sequencing, methylation profiling, long read sequencing, and optical mapping where needed. Functional follow-up studies using patient-derived cells are guided by genetic findings.

Ultimately, this research will advance our understanding of the complex pathogenesis of metabolic bone diseases, support individualized treatment strategies, and contribute to the development of improved diagnostic tools and targeted therapies.

ID: 182533

Start date: 14-08-2025

End date: 14-08-2029

Last modified: 14-08-2025

Grant number / URL: 2022-00800

Copyright information:

Genetic causes and underlying mechanisms in metabolic bone diseases

General Information

Project Title

Genetic causes and underlying mechanisms in metabolic bone diseases

Project Leader

Outi Mäkitie

Registration number/corresponding

Project ID: 182533

Version

1.0

Date

2025-07-24

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

Clinical data from the individuals will be collected by the referring doctors and safely stored at their **Clinic/Hospital**. No personal data (e.g. patient's name and social security number) will be sent to **us. Anonymized information and codes will be used instead and saved in** Microsoft Excel master files in our institutional storage. A **summary PowerPoint presentation** with the pedigree of the family, the phenotypic information, and available genetic information will be sent to us and safely stored at our Department/Institutional safe data storage option. All the output from sequencing facilities will be transferred safely with the help of the IT service from the department. We use **departmental windows-based storage server** for archiving raw sequencing data and important

processed output. The data will be safely transferred (**sftp protocol**) to the **UPPMAX Bianca cluster** for analysis and created output will be again archived in our departmental server.

The existing clinical and genetic data are covered by specific ethical approvals and informed consents which restrict their use to **specified research questions**.

Data provenance will be documented using a combination of **electronic lab notebooks** (ELNs) and **structured metadata** (excel) files stored alongside the data in safe departmental storage with backup. In the metadata, we will describe the origin (sample ID, collection date, source institution, quality raw file names and storage paths, person responsible for analysis and progress related to and path to downstream analysis files. The raw data or combination data received from sequencing facility will be stored in **“read-only” directories** with checksums and backup. During analysis, **standardized README files** will be used to record the processing parameters (software versions etc), generated output. **All provenance records will be long-term stored** in the departmental secure storage server together with the data and updated whenever new processing steps are performed.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

The data collected and generated will include:

- **Numeric data**: Clinical measurements, laboratory test results (qPCR, ddPCR outputs), and variant annotation tables stored in spreadsheets or tabular formats.
- **Textual data**: Pedigrees, case summaries, phenotypic descriptions, analysis logs, and PowerPoint presentations integrating family, phenotype, and genetic information.
- **Image data**: Western blot images, microscopy images (including associated metadata), and histomorphometry images.
- **Genomic data**: Whole-genome sequencing (WGS) and Sanger sequencing files.
- **Mixed media**: Documents and presentations combining text, figures, and pedigree diagrams for internal review and reporting.

File formats used and justification:

- **Genomic data**: FASTQ, CRAM, and VCF are internationally recognised standard formats for sequencing data, ensuring interoperability with bioinformatics pipelines and repositories.
- **Electropherograms**: AB1 is the standard output format from Sanger sequencing instruments, preserving raw trace data for re-analysis.
- **Tabular and numeric data**: CSV and XLSX are widely supported and accessible, with CSV offering a non-proprietary format for long-term preservation, and XLSX providing enhanced formatting and ease of use for staff.
- **Images**: TIFF (.tif) and OME-TIFF (.ome.tif) are open, high-quality formats suitable for long-term storage and re-analysis, widely used in imaging and microscopy. Raw data from the imaging instruments like .scn file etc, will be stored in "read-only" folders separately.
- **Presentations and reports**: PPTX, PDF, and DOCX are widely used within the institution and among collaborators, balancing editability with fixed-format archiving.

These formats were selected for compatibility with sequencing facilities and laboratory equipment, widespread acceptance in the genomics and biomedical research community, and long-term accessibility. Staff expertise and existing institutional infrastructure support their continued use.

Details on the data volume:

- **Highthroughput sequencing raw data**: ~15–120 GB per sample (paired-end FASTQ); CRAM files ~20–80 GB per sample.
- **Variant data**: VCF and annotation files 50MB–5 GB per sample.

- **Sanger sequencing**: ~1–2 MB per .ab1 file.
- **qPCR/ddPCR data**: Typically <5 MB per run.
- **Image files**: Western blot membranes: ~10–20 MB per .tif file. Microscopy and histomorphometry images: ~50–200 MB per .ome.tif file.
- **Clinical spreadsheets and documentation**: Typically <5 MB per file.

Documentation and data quality

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

All materials will be documented using a combination of **ELNs, structured metadata files, and analysis logs** stored alongside the data. Each dataset will be accompanied by a **README** file describing the content, file format, origin, and collection method. For sequencing data, metadata will include sample IDs, collection date, source institution, sequencing facility, platform, library preparation method, and file **checksums**. Analysis metadata will record software names, versions, parameters, and dates of processing, ensuring reproducibility.

Laboratory validation data (qPCR, ddPCR, Western blot, microscopy) will include experimental design details, triplicate replicate structure, reagents used, equipment settings, and image acquisition parameters. File naming conventions will follow a consistent standard reflecting project name, sample ID, date, and data type. All metadata will be stored in non-proprietary formats where possible (e.g. CSV, TXT) and archived alongside the corresponding raw and processed data in secure institutional storage.

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

Data quality will be safeguarded through **standardised experimental protocol**s, inclusion of **appropriate controls, and replication**. Functional and molecular experiments (e.g. qPCR, ddPCR, Western blot, microscopy) will be performed in **triplicates with positive and negative controls**. Sequencing **data quality will be assessed using standard metrics** (e.g. read quality scores, coverage depth, duplication rates), with failed or low-quality runs repeated. **Imaging** experiments will follow **calibrated** instrument settings, with acquisition parameters and magnifications recorded. All metadata will capture experimental design, collection methods, reagents, equipment, QC outcomes. Data integrity will be verified using **checksums**, and both raw and processed data will be stored in secure institutional storage server with **structured metadata files** (e.g. CSV, TXT) describing content, format, and provenance.

Storage and backup

How is storage and backup of data and metadata safeguarded during the research process?

Multiple storage and backup systems will be put on place to guarantee a safe storage of the data during the entire research process. These systems/workflows are in place after discussions with the **infrastructure manager**(s) and are actively maintained by the **department personnel responsible**.

We will use **KI ELN** to keep track of all the samples that are analyzed and the experiments that have been performed. Additionally, a **master record** will be maintained in our secure departmental storage server. During analysis on Bianca, scripts, raw data and some important output will also be saved in the backup area to make sure that we safeguard data during the research process.

Each researcher involved in the project will also maintain some anonymized analytic data, scripts, other general output and documents on their **backed-up KI OneDrive cloud** .

For long term storage, archives would be maintained on our departmental storage server.

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

All the samples we receive will be **coded and anonymized**. **Clinical/phenotype information** will be saved on our shared institutional/departmental storage and **access will be restricted to personnel/researchers involved in the project.** All the c**onsent forms** signed by the patients or their legal guardians will be safely stored in a **security box at out Department.**

Multiple storage and backup systems will be put on place to guarantee a safe storage of the data during the entire research process.

Legal and ethical aspects

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

- Sensitive personal data will be handled according to GDPR.
- Data will be anonymized and the codes will be kept separately from the data.
- An ethical permit to collect samples from international centers has already been obtained. The referring doctors will use anonymized codes for the patients and their family members.
- Data Transfer/Processing agreements will be performed between our research group and collaborators for data transfer, previously approved by KI's legal department.

How is correct data handling according to ethical aspects safeguarded?

- Our study will be performed in accordance with the ethical principles of the World Medical Association (WMA) Declaration of Helsinki and aims to follow Good Clinical Practice (GCP) guidelines.
- Ethical approvals/potential amendments and informed consent forms for the project are safely stored by the group members.
- A consent form is signed by the patients and/or their legal guardians before sample collection.
- Clinical data from the patients and their families will be anonymized.
- Our research study respects the confidentiality and the privacy of the information in order to

protect the participants from uncomfortable situations and guarantee their anonymity during the research work and public talks/publications.

Accessibility and long-term storage

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

The results of this research projects will be stored in form of scientific publications.

All the group members will manage, process and document the material during the project as well as prepare the material for long term preservation and possibly dissemination.

Metadata and aggregated data is published openly, underlying raw data is made available upon request after ensuring compliance with relevant legislation and KI guidelines.

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

Long-term storage will take place at the server at the Institution and in ELN and in the KI server. Data will be stored at least 10 years after publication. The data will include raw data and the final data analysis file.

Prof. Outi Mäkitie is responsible for guarantying that the public records from the project are archived and kept for at least 10 years after the project is completed.

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

The data can be read by using different software that can read .xls, .doc, .ppt, .tif, .jpeg, .cvs documents.

Eventual codes that are necessary to process and interpret the data will be deposited on GitHub.

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

A DOI will be obtained through PUBMED or other public sources.

Responsibility and resources

Who is responsible for data management and (possibly) supports the work with this while

the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

Data management is performed by the members of our research group who are involved in this project. The project leader, Professor Outi Mäkitie, will also be responsible for ensuring that everyone in this group has received the necessary training and follows the same standardized practices.

The PI and the Department of Molecular Medicine and Surgery (MMK) will be responsible for the data management and long-term storage after the research project has ended.

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

The main resources that will be used for data management and fulfilling FAIR principles, include the KI ELN, Institutional servers and UPPMAX compute project. All raw data, scripts and necessary output will also be backed-up on the institutional storage with the help of departmental personnel responsible. These costs are already accounted in the grant proposal.

All the project members will make sure to make the data easily accessible and findable, following the standardized data management plan. A written summary explaining how the data are stored will be created during the project and when the project ends in order to guarantee a safe longterm storage.