Plan Overview

A Data Management Plan created using DMPonline

Title: Reconstructing the cell evolution of metastasis in high-risk neuroblastomas

Creator: Oscar Bedoya Reina

Principal Investigator: Oscar Bedoya Reina

Data Manager: Oscar Bedoya Reina

Project Administrator: Oscar Bedoya Reina

Affiliation: Karolinska Institutet

Template: KIs Template for Swedish Research Council DMP

ORCID iD: 0009-0001-1703-2258

Project abstract:

Neuroblastoma (NB) is one of the most lethal pediatric cancers and accounts for approximately 15% of deaths associated with childhood malignancies. Patients diagnosed after 18 months of age frequently develop metastatic disease, exhibit poor responses to therapy, and have survival rates below 50%. Previous work from our group identified previously unrecognized progenitor-like cell populations enriched in high-risk neuroblastoma and associated with migratory and mesenchymal transcriptional programs, suggesting a potential cellular origin for metastatic dissemination. The objective of this project is to reconstruct the cellular and molecular evolution of metastasis in high-risk neuroblastoma by integrating single-nucleus transcriptomics, genomic variation, evolutionary reconstruction methods, and artificial intelligence. We will identify cell populations associated with metastatic progression, characterize molecular changes driving disease evolution, and develop predictive models of metastasis and treatment resistance.

ID: 214205

Start date: 15-11-2023

End date: 15-11-2029

Last modified: 02-11-2026

Grant number / URL: TJ2021-0137

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Reconstructing the cell evolution of metastasis in high-risk neuroblastomas

Description of data

How will data be collected, created or reused?

The project will combine newly generated sequencing data from neuroblastoma patient samples, including paired primary tumors and metastatic lesions, with existing datasets from publicly available and controlled-access repositories. In particular, data from the TARGET initiative will be reused alongside publicly available single-cell and single-nucleus RNA sequencing datasets from human and mouse development, including normal tissue references.

New data will be generated through sequencing of patient-derived samples, following relevant ethical approvals and established experimental protocols. Data acquisition and processing will involve quality control, read alignment, gene expression quantification, and single-cell transcriptomic analysis. Computational methods will be applied to reconstruct tumor evolutionary trajectories, identify genetic alterations, and characterize transcriptional changes associated with metastasis and drug resistance.

What types of data will be created and/or collected, in terms of data format? Include version numbers if applicable.

The project will generate and process multiple data types, including raw sequencing files (FASTQ), aligned reads (BAM), variant calls (VCF), gene annotations (GTF), expression matrices, and integrated single-cell datasets (HDF5/H5AD). Further computational analyses will produce evolutionary reconstructions, AI-derived feature matrices and embeddings, trained model checkpoints, and reproducible workflow outputs.

Existing datasets will be integrated with newly generated data using standardized computational pipelines to ensure comparability and reproducibility.

What volumes of data will be created and/or collected?

- > 100 TB

The project will utilize approximately 250 TB of active storage, comprising raw sequencing files, aligned reads, gene expression matrices, integrated single-cell datasets, evolutionary reconstructions, AI-derived feature matrices, model checkpoints, and reproducible computational workflow outputs.

Data volumes are expected to increase as additional patient samples, publicly available datasets, and developmental reference datasets are incorporated into the analyses. Furthermore, computational processing, large-scale data integration, and AI-based modeling will generate substantial intermediate and processed datasets.

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, file naming-format-versioning, etc

All research data will be systematically documented using standardized formats and metadata to ensure traceability, reproducibility, and interoperability. The project will follow established practices for genomic and transcriptomic data management, including single-cell and single-nucleus RNA sequencing, genetic variant analysis, evolutionary reconstruction, and AI-based modeling.

Data will be stored in widely adopted formats, including FASTQ for raw sequencing reads, BAM for aligned sequences, VCF for genetic variants, GTF for genomic annotations, and HDF5/H5AD for processed single-cell datasets. Expression matrices, computational outputs, AI-derived embeddings, and model checkpoints will be maintained in appropriate standardized formats.

Metadata will document sample origin (primary tumor, metastasis, or normal tissue), biological characteristics, sequencing methods, experimental protocols, reference genomes, quality control metrics, and processing parameters. For publicly available datasets, including TARGET and human and mouse developmental references, original accession numbers, data sources, and relevant metadata will be preserved. Clinical metadata will be pseudonymized and handled according to applicable ethical and data protection requirements.

A consistent file-naming convention will be applied, incorporating dataset identifiers, sample identifiers, data types, processing stages, and version numbers. Directory structures will distinguish raw data, intermediate files, processed datasets, and final analytical outputs.

Computational workflows will be documented using version-controlled scripts, software versions, reference databases, and analysis parameters. Data processing steps, including quality control, alignment, quantification, single-cell integration, evolutionary

reconstruction, and AI model training, will be recorded to ensure reproducibility. Model checkpoints and associated training configurations will be retained where appropriate.

Documentation and metadata will be maintained alongside the datasets, following FAIR principles (Findable, Accessible, Interoperable, and Reusable).

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

Data quality will be ensured through standardized quality control procedures, validated analysis pipelines, and systematic documentation. Sequencing data will be assessed for key quality metrics, while single-cell datasets will undergo filtering for low-quality cells, doublets, and batch effects. Public datasets will be evaluated for completeness, consistency, and correct annotation before integration. Computational analyses will include validation, statistical quality assessments, and, where applicable, independent validation or cross-validation to ensure robustness. All processing steps, software versions, parameters, and quality metrics will be documented in reproducible, version-controlled workflows, ensuring data reliability, traceability, and reproducibility.

Storage and backup

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

- KI OwnCloud

Research data and associated metadata will be securely stored using an institutionally managed ownCloud platform, providing centralized storage, controlled access, and synchronization across authorized research environments.

Raw sequencing data, processed datasets, computational outputs, and associated metadata will be systematically organized and documented to ensure traceability and reproducibility. Storage capacity and backup requirements will be regularly reviewed to accommodate increasing data volumes throughout the project.

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

Data security will be maintained through secure encrypted connections, institutional authentication mechanisms, and compliance with applicable data protection regulations, including GDPR. Patient-derived data will be pseudonymized before storage and analysis.

Access to KI OwnCloud will be restricted to authorized project members through individual user accounts and appropriate access permissions, ensuring the confidentiality and integrity of sensitive genomic, transcriptomic, and clinical data.

Human sequencing data, including data generated will be processed and temporarily stored on secure computing infrastructure provided through NAISS-SENS, including the Bianca cluster at UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science) and Arrhenius SENS Storage (NAISS, Linköping University). This infrastructure is specifically designed for sensitive human genomic data and incorporates multiple layers of security, including isolated project environments, two-factor authentication, restricted network access, and controlled data transfers.

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 pseudonymized and a key will be kept separately from the data.

If necessary, data transfer or data processing agreement 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?

Patient data is pseudonymized by the clinical collaborator and the code is not accessible to researchers in our research group. The

material will arrive to KI coded, and the original code will be saved by the collaborators.
Consent has been acquired from human participants to process/share data.
Results will only be presented on aggregated level without any possibility of backward identification.

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, licenses and limitations on the access to and reuse of data?

Only metadata is published openly, underlying data is made available upon request after ensuring compliance with relevant legislation and KI guidelines.
Analysis scripts and other developed code will be uploaded to Github.

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 KI. Data will be stored at least 10 years after publication. The data will include raw data and the final data analysis file.

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

Code necessary to process and interpret the data will be deposited on GitHub.

How will unique and persistent identifiers for the research data, such as a Digital Object Identifier (DOI), be obtained?

A DOI will be assigned to the dataset by the data repository (e.g. SND).

Responsibility and resources

Who is responsible for data management while the research project is in progress?

Data management is performed by a dedicated data manager in the research group, who is an experienced researcher with a PhD, and by the PI.

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

The PI is responsible for data management and the archive function will be responsible for long-term storage

What resources (costs, labour or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)?

Access to the departmental server is required; if additional external storage is needed, the associated costs will be covered by the project.

What resources will be needed to ensure that data fulfil the FAIR principles?

We plan to make our datasets FINDABLE by uploading rich metadata to a searchable resource (a data repository) and having a

persistent identifier assigned to the data by the repository,

We plan to make our datasets ACCESSIBLE by ensuring that following the persistent identifier will lead to the data or associated metadata

We plan to make our datasets INTEROPERABLE by using controlled vocabularies, keywords or ontologies where possible and by using open file formats

We plan to make our datasets RESUSABLE by assuring high data quality, by providing all documentation needed to support data interpretation and reuse and by clearly licensing the data via the repository so that others know what kinds of reuse are permitted.