Plan Overview

A Data Management Plan created using DMPonline.be

Title: DeepHeartNet: a computational and functional characterization of the genetic architecture of non-syndromic congenital heart disease

Creator: Alejandro Sifrim

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Project abstract:

Isolated congenital heart disease (CHD) affects about 1 in 150 newborns and arises from both genetic and environmental factors. Extensive DNA sequencing by us and others revealed a significant burden of inherited rare variants in known and novel genes, but many genes remain undiscovered due to lack of statistical power. Here, we will remedy this in 2 ways. Firstly, we will contribute to the largest whole-exome sequencing CHD cohort to date. While more powered, many variants will likely still fail to achieve genome- wide significance. In a second stage, we will therefore build novel computational deep learning approaches for gene prioritization based on single-cell multi-omics datasets. Importantly, prioritized genes will also be validated, by implementing high-throughput gene perturbation in an in vitro cardiac progenitor differentiation model. Together, this project will improve the diagnostic yield in CHD patients, and thus ameliorate genetic counselling for one of the leading causes of non-infectious child mortality.

ID: 213910

Start date: 01-10-2024

End date: 30-09-2028

Last modified: 11-04-2025

DeepHeartNet: a computational and functional characterization of the genetic architecture of non-syndromic congenital heart disease

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N****(ew data) or* ***E****(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
mus_MESP1+	single-cell multi-omics data in mus musculus	E	D	T	FASTA/BAM/RDS	<1TB	VSC (stg_00077)/MANGO/External USB HDD
human_MESP1+_CROPSEQ	CROPSeq data in human cell line	N	D	T	FASTA/BAM/RDS	<5TB	VSC (stg_00077)/MANGO
analysis_code	Analysis code (Python notebooks)	N	D	T	Python	<1GB	Github
sample_metadata	Sample metadata	N	D	T	TSV	<1GB	VSC (stg_00077)/MANGO
ai_models	Trained AI models	N	D	N	Pickled python objects	<1TB	VSC (stg_00077)/MANGO
human_sequencing_data	Human clinical WGS data	N	D	T	FASTA/BAM/VCF	<10TB	VSC (stg_00077)/MANGO

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Question not answered.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- Yes, human subject data (Provide SMEC or EC approval number below)

Sequencing whole-genome sequencing data (**Ethical approval pending)**

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- Yes (Provide PRET G-number or EC S-number below)

S70151

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

- Sample metadata will be kept both in a metadata spreadsheet, as well as catalogued inside the MANGO metadata system.
- Analysis code will be kept in Jupyter notebooks and version controlled using Git. Trained AI models will also be versioned and kept in MANGO with associated code repositories being referenced in the MANGO metadata.

Will a metadata standard be used to make it easier to find and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- Yes

- Human phenotypes will be encoded using the Human Phenotype Ontology (HPO)
- Sequencing instrument metadata outputted by the instruments will also be kept and catalogued

Data Storage & Back-up during the Research Project

Where will the data be stored?

- ManGO
- Other (specify below)

We will also keep active research data on our VSC volume (stg_00077) and cold archived in the FRIGO system.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Only authorized persons will have access to the folder where the data will be stored (using the Linux/MANGO user permission systems).

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

We envision <20TB will be generated/archived. At current costs of 30 Euro/TB/year this will amount to 2400 Euro (4 years of the project). This will be funded by a VLIR infrastructure grant for data storage infrastructure obtained by the Leuven Institute of Single-cell Omics and has also been taken into account in the funding budget.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy
- All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans

Where will these data be archived (stored and curated for the long-term)?

- Other (specify below)

We will archive the data in cold storage solutions being worked out by KUL ICTS (FRIGO system). Clinical sequencing data will be

archived as part of the Center of Human Genetics data repository for clinical sequencing data.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

20 Euro/TB/Year of cold storage x 20TB x 10 years = 4000 Euro. These costs will be covered by a VLIR infrastructure grant (PRISMO) for data storage which was obtained by the Leuven Institute of Single Cell Omics.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- Yes, as open data
- Yes, as restricted data (upon approval, or institutional access only)

- Single-cell data will be made publicly available through public repositories such as the Gene Expression Omnibus (GEO).
- WGS sequencing data of patients will be made available through restricted access via a Data Access Committee through the European Genome-Phenome Archive (EGA) repository.

If access is restricted, please specify who will be able to access the data and under what conditions.

- Only credentialed researchers in relevant research fields will be able to access the clinical sequencing data after approval by the DAC.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

- Yes, privacy aspects
- Yes, ethical aspects

- Clinical sequencing data is identifiable and as such data sharing needs to be appropriately consented and restricted according to the guidelines set by the ethical commission.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other data repository (specify below)

- Single-cell data will be made publicly available through public repositories such as the Gene Expression Omnibus (GEO).
- WGS sequencing data of patients will be made available through restricted access via a Data Access Committee through the European Genome-Phenome Archive (EGA) repository.

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- Data Transfer Agreement (restricted data)

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

What are the expected costs for data sharing? How will these costs be covered?

The proposed public repositories provide this service free of charge.

Responsibilities

Who will manage data documentation and metadata during the research project?

- Fatemeh Hoseinkhani (PhD Student)
- Saptarshi Chakrabarti (Data Engineer)

Who will manage data storage and backup during the research project?

- Saptarshi Chakrabarti (Data Engineer)

Who will manage data preservation and sharing?

- Saptarshi Chakrabarti (Data Engineer)

Who will update and implement this DMP?

- Fatemeh Hoseinkhani (PhD Student)
- Saptarshi Chakrabarti (Data Engineer)
- Alejandro Sifrim (PI)