Functional analysis of non-coding variation in human craniofacial-related disorders

A Data Management Plan created using DMPonline.be

Creator: Catia Attanasio

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Catia Attanasio

Grant number / URL: C14/23/139

ID: 204099

Start date: 01-10-2023

End date: 30-09-2027

Project abstract:

While craniofacial malformations are among the most common congenital anomalies in humans, their genetic causes remain largely unknown. Previous studies have shown that non-coding (NC) variation within craniofacial enhancers can lead to aberrant gene expression and craniofacial anomalies. Translating the occurrence of NC variation into biological consequences is, however, still challenging as our knowledge of the function of NC DNA is limited. Here we propose to use a combination of chromatin structure, epigenomics and transcriptomics analyses in human cranial neural crest cells and their derivatives to establish temporal maps of active craniofacial enhancers and identify the gene(s) they control. This will inform us on their molecular functions and allow us to better understand how variation in their sequence might affect their activity. We will then use these datasets to revisit unsolved craniofacial-related clinical cases, focusing on NC variation. This study will generate critical data for the integration of NC variation analysis in clinical practice.

Last modified: 27-03-2024

Functional analysis of non-coding variation in human craniofacial-related disorders

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	
Promoter Capture Hi-C	Chromatin interactions maps in CNCCs, Chondrocytes and Osteoblasts	N	D	N,T	FASTQ, BAM, BEDM TXT, XLS, HTML, PDF	<5TB	
Cell line RNA-seq	CNCCs, chondrocytes and osteoblasts	N	D	N, T	FASTQ,TSV, XLS, TXT, BIGWIG	<100GB	
H3K27ac ChIP-seq	CNCCs, chondrocytes and osteoblasts	N	D	N, T	FASTQ, XLS, TXT, BIGWIG	<100GB	
Variant analysis in patients		N	D	N, T, I	VCF, TXT, CVS, SVG	<1GB	
Luciferase assay	ESC, CNCCs, chondrocytes and osteoblasts	N	D	N,T	TXT, XLS	<1GB	
CRISPR-Cas9 variant analysis	ESC, CNCCs, chondrocytes and osteoblasts	N	D	N,T	FASTQ,TXT, XLS,TSV, BED	<100GB	
omics publicly available data	publicly available transcription or chromatin state data for CNCCs, chondrocytes or osteoblasts	E	D	N,T	BAM, TSV, BIGBED, BIGWIG, BED	<1TB	

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:

We will use previously published datasets from other groups. We will access that data from the publication directly (Supplementary tables) or from the GEO repository (https://www.ncbi.nlm.nih.gov/geo/) where published genomic data are often deposited.
We will also download data from:

- FACEBASE (https://www.facebase.org/): data resource for craniofacial research
- VISTA enhancer (https://enhancer.lbl.gov/)

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)

H9 embryonic stem cells related data : S67823
Secondary use of data - reanalysis of whole genomes from unsolved (BeSolveRD) patients: S64603

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)

Genetic data of patients recruited in the context of the Belgian Genome Resource to Resolve Rare Diseases Consortium (S64603). We receive preprocessed anonymized variant calling data from our collaborators along with health data such as diagnoses and symptoms.
PRET: G-2023-7031
S67823

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.

- Yes

To purchase the commercial cell line (H9 ESC from WiCell), an MTA had to be signed, which was followed up by LRD. The MTA restrict some of the usage of ESC cells (which are not in the scope of this research) and sharing of the cells with other research groups without WiCell consent. Genetic data sharing also require third parties to agree with the restriction of use of WiCell.

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).

We will use the ManGO active research data repository to depose our metadata and linked them to the individual data files.
Data processing is dependent on the data type but for each data type we have a set of standard (published) pipelines that are used. We will anyway consult with the ManGO team on best practices to ensure recording of specific parameters and reproducible data processing.

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

We will use the GEO metadata standard form for high-throughput sequencing data (https://www.ncbi.nlm.nih.gov/geo/info/seq.html).
For non high-throughput data (patients' variant calling and luciferase) we will develop a standard README.txt file inspired from the GEO and/or FACEBASE metadata files. Alternatively we will look into standards available through the ManGO platform.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Shared network drive (J-drive)
- OneDrive (KU Leuven)
- ManGO
- Large Volume Storage

To preserve our sensible data, we will store it on KU Leuven drives (J- and K- drives, short and long term storage fitted for 'strictly confidential data' storage) and/or OneDrive linked to a KU Leuven account, which is suitable for storing sensitive data when set up with Multi Factor Authentication, via the KU Leuven Authenticator.
We will also use the ManGo active research data management platform of KU Leuven. Once the analyses are completed and access to the files are not needed anymore, we will transfer them to long-term storage e.g. KU Leuven K- drive.
Concerning WGS data of patients, note that we do not ourselves have access to that data, we only receive preprocessed anonymized variant calling data. For data storage, back up and sharing of WGS from the BeSolveRD project see ethical approval S67695.

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.

- No (explain solution below)

We have not yet set it up, however, we will prioritize it as first datasets will be available soon. We will consult with ManGO and ICTS teams.

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

The data will only be available to authorized personnel through the ManGO user identification system.
On KU Leuven drives the data will only be available to authorized persons by setting specific user permissions.

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

The cost will be covered by our C1 grant.

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

Omics data will be preserved for 10 years according to KU Leuven RDM policy.
BeSolveRD patients data will be preserved for 25 years according to EC Research UZ/KU Leuven. This is managed by our collaborators under their ethical approval S64603.

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

- Other (specify below)
- Large Volume Storage (longterm for large volumes)

To preserve data during and up to 10 years after the project, we will store the data on KU Leuven servers or on secured cloud-based platforms.
Upon publication, we will also deposit our data in the open-access non-profit GEO (https://www.ncbi.nlm.nih.gov/geo), which will ensure its longevity and its accessibility to the scientific community.

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

We have budgeted long-term cold archiving storage (95 Euro/TB/year) in our C1 grant.

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)

All data generated from H9 ESC will be made openly available.
Variant analysis in patients will be restricted.

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

Genetic data related to patients can only be shared upon approval from the PI of the BeSolveRD project (S64603) and myself (S67823). After approval, the data can be shared without additional contracts.
Note that we are also considering depositing patients' data in the EGA (repository for storing genomic and/or genetic data) after publication. In that case, to access the data, the applicant will need to make an application to the Data Access Committee using the dataset page.

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, ethical aspects

Genetic data from patients are considered as sensitive data and their access is consequently restricted.

Where will the data be made available?

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

- Other data repository (specify below)
- KU Leuven RDR (Research Data Repository)

- Omics data such as RNA-seq, Promoter Capture-HiC, ChIP-seq, etc. will be deposited on the GEO public functional genomics data repository (https://www.ncbi.nlm.nih.gov/geo/).
- Genome variant analysis: we will probably deposit variants analyses in EGA (the European Genome-phenome Archive). The genetic data will be pseudonymously deposited in EGA with controlled access, meaning that a third party can obtain access to the sequencing data only following approval by the KU Leuven/UZ Leuven Data Access Committee.

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.

- CC-BY 4.0 (data)
- 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?

We don’t expect additional costs for data sharing.

Responsibilities

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

The research staff under the supervision of the PI.

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

The research staff will ensure data storage and back up under the supervision of the PI.

Who will manage data preservation and sharing?

The PI: Catia Attanasio

Who will update and implement this DMP?

The PI: Catia Attanasio