Assessing the role of developmental enhancers in craniofacial development and disease

A Data Management Plan created using DMPonline.be

Creator: Katerina Rapti

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Catia Attanasio

Project Administrator: Katerina Rapti

Grant number / URL: 11Q3V24N

ID: 204392

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

Craniofacial disorders constitute 1/3 of all congenital defects. Despite their evident genetic component, a big proportion of affected patients lack molecular diagnosis, after following the standard of care. This advises towards the implication of variants in the non-coding genome. GWAS studies are also pointing to this direction, as most detected variants associated with face morphology lie in non-coding regions. Enhancers are non-coding cis-regulatory elements that orchestrate gene expression during craniofacial development, in a spaciotemporal manner, by interacting, through chromatin looping, with their target gene. Variants within enhancers can alter these interactions and cause aberrant gene expression and disease. However, modern genetics stand far from functionally interpreting non-coding variants and integrating them in the diagnostic setting. Here, using a cell differentiation model of human face progenitor cells, we propose to map chromatin interactions through time, combine this data with gene expression analysis and epigenomic data, and use our datasets to functionally revisit unsolved craniofacial clinical cases. Our project will shed light on the regulatory networks that drive the development of human face and disease and link non-coding variation to specific phenotypes, paving the way towards understanding the clinical significance of the non-coding genome.

Last modified: 23-04-2024

Assessing the role of developmental enhancers in craniofacial development and disease
FWO DMP (Flemish Standard DMP)

1. 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.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset name / ID	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
			• Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Promoter capture Hi-C	3D chromatin interaction maps in cranial neural crest cells (CNCCs), chondrocytes and osteoblasts	Generate new data	Digital	Experimental	fastq, bed, txt, xls, bam, pdf, html,	<5TB	
cell line RNA-seq	RNA-seq in H9 Embryonic Stem Cells (ESCs) CNCCs, chondrocytes, osteoblasts	Generate new data	Digital	Experimental	fastq, txt, xlx, tsv, bigwig	<100GB	
Patients' Variants re-analysis	Prioritization and selection of single nucleotide variants found in patients	Generate new data	Digital	Experimental	vcf, xls, txt, svg	<1GB	
cell-line H3K27ac ChIP-seq	H3K27ac ChIP-seq in CNCCs, chondrocytes, osteoblasts	Generate new data	Digital	Experimental	fastq, xls, txt, bigwig	<1TB	
multi-omics publicly available data	publicly available -omics datasets from hESCs, CNCCs, chondrocytes (ATAC-seq, CHIP-seq...)	Reused data	Digital	Experimental	txt, xls, bam, bigbed, bigwig, bigbed, tsv	<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 access publicly available datasets, published by different research groups, either directly from the publications (supplementary tables) or from the GEO repository (https://www.ncbi.nlm.nih.gov/geo/). We will also download data from FACEBASE (https://www.facebase.org/), a data repository for craniofacial researchers.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

H9 human embryonic stem cells (ESCs) will be differentiated into cranial neural crest cells (CNCCs) and downstream derivatives, and these cell types will be used to generate the aforementioned datasets.
Ethical approval number related to ESCs cell culture and related data: S67823
We will receive and re-analyse anonymized genetic data (single nucleotide variants/SNVs detected in patients) from unsolved patients included in the Belgian Genome Resource to Resolve Rare Diseases (BeSolveRD) Consortium. (https://beshg.be/workgroups/besolverd). Ethical approval number: S64603

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We will receive preprocessed anonymized variant calling data generated from patients recruited in the study BeSolveRD. We will also receive health data (diagnosis/symptoms/phenotypic characteristics) of the same patients.

PRET G-number: G-2023-7031
EC S-number: 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/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 H9 ESCs from WiCell, an MTA had to be signed, followed up by LRD. The MTA restricts some of the usage of the ESCs, however none of the restricted uses are in the scope of this research. Sharing of the cells with other research groups without WiCell consent is restricted. Genetic data sharing also require third parties to agree with the restrictions 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

2. 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).

Protocols and details related to data collection and processing will be recorded in physical lab books and transcribed to Word or Excel files.

All our final datasets will be accompanied by README.txt files containing all necessary information for future (re)use of the data.

We will use the ManGO active data management platform to depose our metadata and link them to the individual data files. Data processing dependents on the data type but for each data type we have a set of standard (published) pipelines that are used.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) 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) 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.

3. Data storage & back-up during the research project

Where will the data be stored?

We will ensure data storage and back up under the supervision of the PI and with the support from ICTS and HPC staff. To preserve our data, we will store them on KU Leuven drives (J-drive), on OneDrive linked to a KU Leuven account, on the Flemish Supercomputer Center (VSC) and on 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.

How will the data be backed up?

Data will be backed up by standard back-up provided by KU Leuven ICTS.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- No

We are currently setting this up as we are still generating our datasets.
We will consult with ManGO and the 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 data storage and backup costs will be covered by a C1 grant (KU Leuven internal funds) that my group's PI (Catia Attanasio) has obtained.

4. Data preservation after the end of the research project

Which data will be retained for at least five 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...).

PCHiC, RNA-seq and ChIP-seq data will be preserved for at least 5 years according to KU Leuven RDM policy.

Data related tp the patients from BeSolveRD 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)?

For long-term data preservation, data will be stored on KU Leuven servers or on secured cloud-based platforms. Upon publication, we will also deposit our data in the open-access GEO (https://www.ncbi.nlm.nih.gov/geo). This way we ensure that the generated data remain accessible to the scientific community.

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

The group's PI (Catia Attanasio) has obtained a C1 grant, which will be used to cover the expected costs for data preservation (95 Euro/TB/year).

5. Data sharing and reuse

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

- Yes, in a restricted access repository (after approval, institutional access only, …)
- Yes, in an Open Access repository

All datasets (PCHiC, ChIP-seq, RNA-seq) generated from ESCs and derivative cell types will be openly available. Files generated during patients' variants analysis 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.

The data sharing responsible and the PI of the project BeSolveRD are responsible for sharing and accessing genetic data of patients. Once approval is granted, the data can be shared without further agreements.

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 in the comment section per dataset or data type where appropriate.

- Yes, Ethical aspects

The genetic data that we will receive and analyze come from patients and are anonymized. Thus, access to this information will be restricted, and will be granted only after approval of the PI of BeSolveRD.
The existing data that we will use are already publicly available and before re-using and re-sharing these and derived data, we will verify that we are allowed to.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Our data will be made available on the KU Leuven RDR (Research Data Repository) as well as other data repositories.

Promoter Capture-HiC, RNA-seq, ChIP-seq datasets will be deposited on the GEO public functional genomics data repository (https://www.ncbi.nlm.nih.gov/geo/).

Data generated from patients' variant analysis will be deposited in EGA (the European Genome-phenome Archive). The genetic data will be pseudonymously deposited in EGA with controlled access. This way a third party can gain access to this data only following approval by the KU Leuven/UZ Leuven Data Access Committee.

When will the data be made available?

All research outputs will be made available at the time of publication.

Which data usage licenses are you going to provide? If none, please explain why.

PCHiC, RNA-seq, ChIP-seq data: CC-BY 4.0
Restricted data: Data transfer agreement

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

Upon deposit in a data repository, a PID will be added.

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

No additional costs are expected for data sharing

6. Responsibilities

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

The PhD researcher (Aikaterini Rapti) under the supervision of the PI (Catia Attanasio)

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

The PhD researcher (Aikaterini Rapti) under the supervision of the PI (Catia Attanasio)

Who will manage data preservation and sharing?

The PI (Catia Attanasio)

Who will update and implement this DMP?

The PI (Catia Attanasio) bears the end responsibility of updating and implementing this DMP