Plan Overview

A Data Management Plan created using DMPonline

Title: HUGODECA

Creator: Laurence Noel

Principal Investigator: Laurence NOEL

Data Manager: Laurence NOEL

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 Template

ORCID iD: 0000-0001-7577-3794

Project abstract:

The single ground-breaking goal of the HUGODECA project is to describe the cellular composition and organization of the developing human gonads and to understand how it changes during sex determination into testes in males and ovaries in females. What are the underlying mechanisms and first molecular and cellular events that accompany the differentiation and divergence of embryonic gonads? How and when do male and female cell lineages diverge and specific traits emerge? HUGODECA focuses on healthy gonad development, but our reference model will be tested using specific ex vixo assays mimicking Differences/Disorders of Sex Development (DSD). To reach this ambitious goal, the HUGODECA consortium brings together leading European academic and industrial experts (from 5 different EU countries). It includes active contributors and executives of the Human Cell Atlas (HCA) which will ensure complementarity with other ongoing HCA efforts. The overall HUGODECA concept is grounded on the integration of multiple synergistic expertise and technologies: Single cell profiling, Spatial transcriptomics, 2D Mass cytometry and cyclic immunofluorescence and 3D imaging of optically cleared gonads. HUGODECA will implement novel tools, analytical and computational methods to process and integrate multidimensional OMICS and image data across different platforms. It will evaluate the accuracy of ex vivo culture models of human gonadal development and assess consequences of altering key signaling pathways. HUGODECA will build the first multiscale developmental cell atlas and unprecedented reference maps of male and female human gonads. It will implement an interactive and multi-dimensional online portal for clinicians, scientists and public including DSD patient associations. HUGODECA shall improve our understanding of DSD, which is a major pediatric concern, requiring complex and highly specialized medical treatment and psychosocial care.

ID: 55010

Start date: 01-01-2020

End date: 30-06-2022

Last modified: 02-05-2022

Grant number / URL: 874741

Copyright information:

HUGODECA - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

The goal of the *HUGODECA* project is to describe the organization of the developing human gonads and to understand the molecular and cellular mechanisms underlying sex determination. In order to do so, the project aims to make major advances in the following complementary areas:

- scRNA- and ATAC-seq of human fetal gonads, with assessment of cell surface markers by flow cytometry (WP1);
- ST and ISS leading to the generation of 2D and 3D expression maps of targeted genes at single-cell resolution (WP2);
- Spatial proteomics based on diverse imaging methods (WP3);
- Software development to enable the integration of large OMICS data with imaging data (WP4);
- Organotypic culture models to evaluate the consequences of signaling pathways directing sex-specific differentiation and endocrine disruptors on gonadal development (WP5).

This project will lead to the generation and process of multiple types of research data, that is data produced as the research is conducted, and which should be as re-usable as possible by the scientific community. Most of their characteristics, such as format and size, vary according to the techniques used to generate them. For ex vivo experiments, the input and the output of the assay are both biomaterials; for imaging techniques, the input will still be some biomaterials, but the output will be an image. In the case of a single-cell experiment, the data first generated correspond to a textual sequence referring to the genetic code. The frontier between the different types of data generated or processed, however, is not always so clear-cut and one of the stated goal of the HUGODECA project is actually to be able to integrate heterogeneous types of data into one common scaffold. For this reason, it will sometimes be useful to be able to distinguish the primary medium of information from the complementary media, which have been generated to describe the primary one. In the case of ST, for instance, the image is the primary medium, and OMICS expression data will be used as a means of giving complementary information on that medium. Actually, the way these complementary data is stored may also vary: they can be mentioned in a separate file that will be part of a data package, or correspond to embedded metadata, integrated in the primary media.

Some of the complementary data which can be found in sets of research data are specially intended to document the way experiments are carried out:

- Study, sample and assay metadata: information about the study, the materials used (from biomaterial to lab equipment) and the conditions of the experiment;
- Methodological data: the protocol in itself, with the different steps of the procedure, and/or descriptions of data workflows;

As diverse as the technologies involved may be, these research data also share some common process steps and we have classified them according to the following sub-types:

- Raw data, resulting directly from the technology with which they have been acquired (initial image captures or sequencing data, for instance);
- Pre-analysis data: raw data having undergone a transformation process deemed necessary to be able to proceed to a meaningful analysis or generated in order to be able to run the main analysis. As we will mention later on, these types of data are generally not kept in the long-term;
- Analysis data: data associated to quantitative and/or qualitative features;
- Interpretation data: lists of biological elements, which have been identified as being of interest, or computational data resulting from the interpretation of the analysis.

When considering data from the point of view of the process, another specific type of data comes into the picture: coding data, from scripts written to establish a process pipeline to the complete source code of a software solution specially developed for the project.

Finally, the HUGODECA project will also lead to the production of structured documents, such as publications and communication supports, with the obvious goal of sharing and promoting the results achieved by the HUGODECA Consortium among the scientific community.

In the following subsection, we have listed the different types of expected data by task, to get a more detailed view of their specific characteristics (as the production of publications and communication supports is not task-dependent, they are not mentioned in this list but we will discuss the question of their findability and accessibility in the next sections). All the data mentioned below are not intended to be published or kept in the long-term: the intent, here, is more to have a broad view of the different data generated or processed in the short-term, so as to be able to take them into account when defining naming conventions or characteristics such as the data volume. It is to be noted that this last point is a sensitive issue for the project as the technologies involved imply to store several TB of data (even for a single task) which make the processes of data storage, transfer, back-up and preservation more difficult to address than for datasets of smaller sizes.

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Study and assay metadata will be created to document the different experiments defined in WP 1, WP 2, WP 3 and WP 5 (this is not applicable to WP 4, since it deals with software development).During the first stage of the project, we will gather information about the program and the different studies that

are planned by using SEEK (D7.1). This application follows the Investigation, Study and Assay (ISA) specifications, and we will create one investigation by work package, and one study by project task.To become findable (and citable), datasets should get a unique, persistent identifier. Persistent identifiers like Digital Object Identifiers (DOI) have been designed as a solution to avoid link rot (that is, when a hyperlink stops referring to the original source because it was moved). These persistent identifiers thus ensure that the data is and will be findable. The use of SEEK will make it possible to generate a DOI for each study.

When it comes to versioning :

- **For documents** (.doc, .pdf, ...), versioning will be maintained by following this pattern : [document name]_[version number] _[status: DRAFT{0,1}]. Example : HUGODECA_DMP_7-4_V1.doc
- **For data that can be generated multiple times** , the creation date may be indicated instead of a version number. The date will then be expressed with this format: YYYY-MM-DD
- **For software and scripts**, versioning will be controlled via the use of dedicated tools (Subversion, Git...)

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

Datasets generated and processed by the HUGODECA consortium intend to be listed on the HUGODECA web portal (D6.4), and the links to download the datasets will then point towards the different open repositories. As already mentioned, the consortium members intend to take part in the HCA initiative, so analysis data may also become available on the HCA data portal.

For Software and tools, Genomics and transcriptomics data are in file formats which are widely used within the scientific community, that is fastq.gz files for raw sequencing data, bam files for aligned sequences and count matrices in HDMF5-loom, mtx or csv formats for processed data, making these fully readable and re-usable worldwide from a technical viewpoint. Raw images can be viewed with open-source softwares like the Fiji/ImageJ tools (Fiji is an open source project hosted in aGit version control repository and comprehensive documentation, ImageJ source code is available online, with its user manual). Open-source softwares like Cell Profiler and histoCAT can also be used to analyse imaging data. However, in order to visualize the fully annotated 2D and 3D cell maps, Keen Eye 3D viewer will be necessary: Keen Eye Platform™ is a proprietary and patented software-as-a-service platform and the viewer will be accessible through a cloud service, promoting high accessibility across a worldwide research community.

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in**

your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Interoperability denotes the ability of diverse systems and organizations to work together (inter-operate), to “plug together” components in larger complex systems. It is a key aspect to consider if the participants involved want to avoid the “Tower of Babel” effect, that is to avoid the breakdown of the tower-building effort because the terms used to communicate are not shared and understood by all. For that reason, this section will define what values those metadata can take to ensure the use of a common vocabulary among researchers, but also what types of metadata are required to make sure that the datasets may be accepted in other repositories.

Types of metadata required

For samples and processes, a special attention will be given to the metadata used within the HCA initiative.

Regarding samples, there should be metadata about:

- The organism donor (embryo) : unique anonymous identifier (as transmitted by the provider), biological sex (female, male, mixed or unknown), developmental stage (in PCW), species (“Homo sapiens”, NCBItaxon:9606), known diseases (it should default to "no disease", which is equivalent to "MONDO_0000001");
- The sample in itself: unique identifier and organ from which the sample was collected.

Regarding sample processing and data acquisition, the metadata depend on the technology used. When single-cell technologies are involved, metadata should include information about:

- The dissociation procedure
- The library preparation protocol :
  - Library construction method (e.g. “10x v3 sequencing”), library construction kit (name, manufacturer and catalogue number)
  - The input nucleic acid molecule (e.g. “polyA RNA")
  - The nucleic acid source (e.g. “single cell”)
  - End bias (e.g “3 prime tag", "3 prime end bias", "full length”)
  - Strand ("first", "second", "unstranded" or "not provided")
  - Optional properties from those listed in the metadata for HCA
- The sequencing protocol :
  - Instrument used for sequencing and model (e.g. Illumina HiSeq 4000), the sequencing design (e.g. 2x100 bp) and the number of raw reads
  - 10x specific metadata such as fastq creation method (e.g. “Cellranger mkfastq” or “Illumina bcl2fastq”), fastq creation method version (e.g. “Cellranger 3.1”)

When imaging technologies are involved, metadata should include information about:

- Tissue preparation and the imaging preparation protocol
  - Fixation
  - Image slide thickness
  - Final slicing method (cryosectioning) …
- The imaging protocol :
  - Microscopy (« fluorescence microscopy »)
  - Magnification, Numerical aperture, pixel size …

These metadata can be found in the HCA metadata dictionary. Those which are the most relevant for the project are : Specimen from organism, Imaged Specimen, Imaging preparation protocol, Library Preparation Protocol, Sequencing protocol.

For biological imaging data, the OME Model defines a set of metadata to include, such as XYZ

dimensions and pixels type, as well as extensive metadata on, for example, image acquisition, annotation, and regions of interest (ROIs).

Metadata values

In order for all participants to use the same terms, ontologies and controlled vocabularies are going to be used. Whenever possible these values should correspond to the terms defined in the HCA ontology. If this ontology is incomplete for our needs, the concepts used will preferably come from other existing ontologies, including but not restricted to:

- EFO (Experimental Factor Ontology) which provides information about many experimental variables available in EBI databases, while Eagle-i resource (ERO) is an ontology of research resources such as instruments, protocols, reagents, animal models and biospecimens
- the Biological Imaging Methods Ontology (fBBi) is an ontology dedicated to the terms used in biomedical research for imaging and visualization methods, and the BioAssay Ontology (BAO) describes biological screening assays and their results including high-throughput screening (HTS).
- Embryonic structure is described in UBERON and EHDAA2 provides a structured controlled vocabulary of stage-specific anatomical structures of the developing human and is linked to HSAPDB, which includes carnegie stages.
- Different ontologies describe properties and classes at the cell level : the cell ontology (CL) is a general ontology which applies to cell types in animals; and CPO structures the vocabulary related to morphological and physiological phenotypic characteristics of cells, cell components and cellular processes.

For the different metadata fields which are required, we will thus define the ontology or controlled list from which the values should be part

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

By giving access not only to raw and analysis data, but also to study metadata, methodological data, and scripts used to facilitate the process and analysis of the data, the HUGODECA consortium pave the way for data re-usability. The fact that the different protocols established by the consortium members will be published will enable to check that quality assurance processes are ensured. In the case of coding data, documentation and literary programming is promoted : whenever possible, tutorials in the form of notebooks (jupyter notebooks, R notebook, KnitR, Sweave) will be published to document analysis scripts (potentially with the use of binder), so as to integrate the code with the corresponding narrative and documentation. The use of container technology and virtualization, such as Docker, is also encouraged.

A CC BY license (requiring only attribution) will be the preferred option for scientific publications and reports: data users will be free to “reuse the material in any medium or format, and to remix, transform, and build upon the material, even commercially” but they will be reminded that they should cite the dataset and acknowledge the data producers in any publications and presentations that make use of the data.

In conformance with the principle of being "as open as possible, as closed as necessary”, the

HUGODECA consortium will also support partners in protection of their results, and help them secure future exploitation opportunities. Project members will implement a continuous and integrated strategy for monitoring *HUGODECA* results and the Innovation Management Committee will define those which are likely to be exploited, thus kept confidential. During the preliminary exploitation plan of the HUGODECA project, the deliverables D1.1, D 2.3, D4.1, D4.2, D4.3 have already been identified as potentially patentable.

In the case of bioimaging data (3.3) all requests for access will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. Commercial destinations will be prohibited.

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

12 person-months have been requested by P01c-INSERM for data management (WP7). A data manager (L. Noël, P01c-Inserm) has been recruited to this end. She will help manage the data with the help of a data contact person, identified for each task (7 person-months requested globally to this end). They will notably implement tools necessary for data management, provide the relevant metadata during the course of the project, keep track of the dissemination level of the different data sets, and refine the naming convention.

20 000 euros have also been requested to cover the costs related to open access publications (author-publishing fees).

Cloud data storage (100 TB / month over 18 months) for 3D images has been estimated at a cost of 54000 euros: this sum has to be taken into account when considering the long-term preservation of imaging data, since this contract will have to be renewed. For that reason, partnerships with European institutes (CINES; CC-IN2P3, EU Image Data) that can provide long-term support for data hosting are under negotiations, all as well as the possibility to deposit the images in IDR.

For OMICS data, the deposition of raw and analysis data in public repositories remain the best alternative to ensure their long-term preservation: those are trustworthy data repositories, with a certificate or explicitly adhering to archival standards.

The long-term preservation of HUGODECA data is particularly valuable since they deal with unique specimens, analyzed at crucial developmental stages. As the ethical aspects are particularly sensitive, these datasets will be a major source of information, if future regulations should come to restrict this work as in other major countries. The HUGODECA programme is already of a high value for researchers who can not carry out experiments on this type of specimens due to restrictive regulations in their home country.

The data generated for the HUGODECA project have also a high commercial and educational potential: the study of the cell atlas may eventually be included into university curriculum (medical, biology, etc.) and converted as a unique learning 3D tool for anatomy/surgery. It is also to be noted that the consortium members have already received many requests from documentary makers (TV, etc.) for anatomy content and human fetus discovery.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Initial storage of the data will be performed locally (with the exception of HUGODECA_DS_3-3, HUGODECA_DS_5-1, HUGODECA_DS_5-2, HUGODECA_DS_5-3, HUGODECA_DS_5-4 imaging data, for which a cloud data storage solution has been chosen). For academic groups involved in the project, institutional ICT facilities will ensure a secure environment, with firewall system in place, virus intruder protection, and access to digital files controlled with encryption and/or password protection and SSL encryption for data transfer. Industrial partners will observe the same types of security measures, and dedicate specific storage space for the data generated for this project.

For file transfer and backups, checksums will provide a simple way to compute the integrity of data files before and after file operations.

The sheer size of expected data make it difficult to follow back up best practices (at list 3 copies on at least two different media). As a general rule, we remind here that raw data are usually considered to be the master copy of any given record (or golden copy): therefore, there should be a back-up copy of all raw data. The use of the Dell EMC Isilon storage system (approx. 540 TB) (funded by INSERM) to store their raw data on at least one second media is under study. For images, the important point is also to keep track and preserve all biomaterials used to produce the image captures, since it is another way to ensure that these images can be reproduced.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethical aspects have already been addressed in the ethical section of the DoA as well as in the “D9.1 POPD Requirement number 3” document, since the intended use of human organs and tissues warrants serious consideration of ethical issues. As mentioned in that document, the partners will implement their research activities in full respect of the legal and ethical European / national / institutional requirements, and codes of practices. PIs from concerned laboratories (biobank and processing) have already full ethical clearance. Furthermore, the project management structure will include an Ethics Advisory Board (EAB), to provide guidance and ensure strict ethical governance relating to the documentation, application, material or any other aspect of the project that could have ethical implications.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Question not answered.

HUGODECA - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The goal of the *HUGODECA* project is to describe the organization of the developing human gonads and to understand the molecular and cellular mechanisms underlying sex determination. In order to do so, the project aims to make major advances in the following complementary areas:

- scRNA- and ATAC-seq of human fetal gonads, with assessment of cell surface markers by flow cytometry (WP1);
- ST and ISS leading to the generation of 2D and 3D expression maps of targeted genes at single-cell resolution (WP2);
- Spatial proteomics based on diverse imaging methods (WP3);
- Software development to enable the integration of large OMICS data with imaging data (WP4);
- Organotypic culture models to evaluate the consequences of signaling pathways directing sex-specific differentiation and endocrine disruptors on gonadal development (WP5).

Explain the relation to the objectives of the project

This project will lead to the generation and process of multiple types of research data, that is data produced as the research is conducted, and which should be as re-usable as possible by the scientific community. Most of their characteristics, such as format and size, vary according to the techniques used to generate them.

For ex vivo experiments, the input and the output of the assay are both biomaterials; for imaging techniques, the input will still be some biomaterials, but the output will be an image. In the case of a single-cell experiment, the data first generated correspond to a textual sequence referring to the genetic code. The frontier between the different types of data generated or processed, however, is not always so clear-cut and one of the stated goal of the HUGODECA project is actually to be able to integrate heterogeneous types of data into one common scaffold. For this reason, it will sometimes be useful to be able to distinguish the primary medium of information from the complementary media, which have been generated to describe the primary one. In the case of ST, for instance, the image is the primary medium, and OMICS expression data will be used as a means of giving complementary information on that medium. Actually, the way these complementary data is stored may also vary: they can be mentioned in a separate file that will be part of a data package, or correspond to embedded metadata, integrated in the primary media.

The goal of the DMP is thus to define the different types of data generated, with their specific characteristics, and to outline how they will be made as FAIR as possible.

Specify the types and formats of data generated/collected

As diverse as the technologies involved may be, the research data share some common process steps and we have classified them according to the following sub-types:

- Raw data, resulting directly from the technology with which they have been acquired (initial image captures or sequencing data, for instance);

- Pre-analysis data: raw data having undergone a transformation process deemed necessary to be able to proceed to a meaningful analysis or generated in order to be able to run the main analysis. As we will mention later on, these types of data are generally not kept in the long-term;
- Analysis data: data associated to quantitative and/or qualitative features;
- Interpretation data: lists of biological elements, which have been identified as being of interest, or computational data resulting from the interpretation of the analysis.

Some of the complementary data which can be found in sets of research data are specially intended to document the way experiments are carried out:

- Study, sample and assay metadata: information about the study, the materials used (from biomaterial to lab equipment) and the conditions of the experiment;
- Methodological data: the protocol in itself, with the different steps of the procedure, and/or descriptions of data workflows;

When considering data from the point of view of the process, another specific type of data comes into the picture: coding data, from scripts written to establish a process pipeline to the complete source code of a software solution specially developed for the project.

Finally, the HUGODECA project will also lead to the production of structured documents, such as publications and communication supports, with the obvious goal of sharing and promoting the results achieved by the HUGODECA Consortium among the scientific community.

In the following subsection, we have listed the different types of expected data by task, to get a more detailed view of their specific characteristics (as the production of publications and communication supports is not task-dependent, they are not mentioned in this list but we will discuss the question of their findability and accessibility in the next sections). All the data mentioned below are not intended to be published or kept in the long-term: the intent, here, is more to have a broad view of the different data generated or processed in the short-term, so as to be able to take them into account when defining naming conventions or characteristics such as the data volume. It is to be noted that this last point is a sensitive issue for the project as the technologies involved imply to store several TB of data (even for a single task) which make the processes of data storage, transfer, back-up and preservation more difficult to address than for datasets of smaller sizes.

LIST OF DATASETS

Reference	HUGODECA_DS_11
Task	1.1 Single-cell RNA-sequencing of human fetal gonads
Task leader	Antoine Rolland (antoine.rolland@univ-rennes1.fr)
Data contact person	Paul RIVAUD (paul.rivaud@inserm.fr) Antoine ROLLAND (antoine.rolland@univ-rennes1.fr)
Description	This dataset corresponds to the data generated via single-cell RNA-sequencing (Chromium technology, 10X Genomics) and subsequently analyzed to provide a transcriptional characterization of human fetal gonad development, at a single cell resolution.

Types and formats	■ Raw sequencing data in *fastq.gz* format (de-multiplexed data, if needed) ■ Pre-analysis data alignment data compressed in binary format (BAM), with a BAM index (.bai) ■ Analysis data : o expression files in *mtx* format with associated tsv files (genes.tsv, barcodes.tsv) as generated by CellRanger pipeline, o sparse matrix data file in *HDF5* format (.h5 or *.loom*) or R format (.rds) containing the combined expression matrix, associated metadata (a metadata file can also be stored separately in .txt, .csv or .tsv , one row per feature/annotation) and cell coordinates (a coordinate file can also be stored separately : it will contain t-SNE or UMAP coordinates with at minimum three columns, which are "cell names" - as specified in the metadata file - "x", "y") ■ Interpretation data: list of marker genes that can be used to characterize the cells at distinct developmental stages (.txt) ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf and/or .html) ■ Coding data in bash, python or R
Origin	Gonads and mesonephroi from male and female embryos will be collected during the first and second semesters of development. Samples will be collected and processed at the Irset institute (Rennes, France) for partner P01c-INSERM and at the Wellcome Sanger Institute (Cambridge, United Kingdom) for partner P05-GRL. Sequencing data will be generated by a subcontractor for the Irset Institute and on-site for the Wellcome Sanger Institute. Data processing will be performed at both the Irset institute and the Wellcome Sanger Institute. Single-cell RNA-sequencing data of 30 human fetal gonads (from female and male embryos aged 6 to 12 post-conception weeks) previously generated by partner P01c-INSERM (SINERGIA grant from the Swiss National Science Foundation) will be re-used for this task. Briefly, these will be combined with expression data for single gonadal cells from second trimester embryos as well as with expression data for single mesonephric cells from first and second trimester embryos.
Expected size	For a given sample (4-5k cells) raw sequencing data represents 20 to 30GB, while the corresponding expression data is about 100-200MB (mtx format) and 600-800MB (tsv format). The total number of targeted cells being 400-450k cells, the overall size of the data will range between 20 and 30TB.

Utility	The analysis of resulting data will allow to i) identify all cell types present in human fetal gonads, ii) characterize how distinct cell lineages differentiate during gonad development, and iii) provide new candidate markers genes for spatial characterization of gonadogenesis in Humans (WP2 and WP3).
Shareable data	Upon P01c-INSERM DPO's validation, the following data will be made accessible: · scAnalysis of Human fetal Gonads : o sequencing data (fastqs) will be submitted to EGA o expression matrix (loom) will be published on the HUGODECA data portal (May 2022) · scAnalysis of human fetal gonads and mesonephros o sequencing data (fastqs) will be submitted to EGA o expression matrix (loom) will be published on the HUGODECA data portal (May 2022) Upon P05-GRL DPO's validation, the following data will be made accessible: · Sequencing data: has been submitted to ArrayExpress/ENA (**private status** until end of january). Project identifier: E-MTAB-10551 · Analysis data: to be published on the HUGODECA data portal (May 2022) o expression matrix (HGDC_VENTO_scRNAseq_aggregated.rds, converted to loom) will be published on the HUGODECA data portal The protocols followed by P05-GRL are published on protocols.io (and the links will be listed on the HUGODECA data portal) o doi:10.17504/protocols.io.66fhhbn o doi:10.17504/protocols.io.bwcipaue The following article has been submitted to Research Gate by P05-GRL: https://doi.org/10.21203/rs.3.rs-496470/v1 The scRNAseq pipeline established by P01c-INSERM is publically available on github : https://github.com/umr1085-irset/scrnaseq_standard_pipes

Reference	HUGODECA_DS_12
Task	1.2 Single-cell ATAC-sequencing of human fetal gonads

Task leader	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Data contact person	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Description	This dataset corresponds to the data generated via single-cell ATAC-sequencing and subsequently analyzed to assess genome-wide chromatin accessibility of XX and XY human gonadal cells across developmental stages (4 to 20 PCW, ~250,000 cells, P01c-Inserm, P05-GRL). Differential accessibility analyses will be conducted on pseudo-bulk ATAC-seq profiles to highlight open chromatin peaks in each group of cells. These potential active regulatory regions will be used to identify enriched DNA motifs and corresponding transcription factors (TFs) for each cell type. The dynamics of chromatin accessibility will then be mapped across the RNA-defined trajectories of the different gonadal cell lineages.
Types and formats	■ Raw sequencing data in *fastq.gz* format ■ Pre-analysis data : alignment data compressed in binary format (BAM), with a BAM index (.bai) ■ Analysis data : o bed file of all called peak locations (peaks.bed) o raw and filtered peak barcode matrix in HDF5 , MEX format or dense CSV format o loupe cell browser files (.cloupe) for visualization o peak annotations in tsv format o motif finding results, in text format (as a position weight matrix PWM, for instance) ■ Interpretation data : o TF-gene interaction network in .tsv format o List of marker regulators (.txt) ■ Study, assay and sample metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf and/or .html) ■ Coding data in bash, python or R
Origin	The samples used for scATAC-seq will be collected and processed at the Irset institute (Rennes, France) for partner P01c-INSERM and at the Wellcome Sanger Institute (Cambridge, United Kingdom) for partner P05-GRL. Sequencing data will be generated by a subcontractor for the Irset Institute and on-site for the Wellcome Sanger Institute. Data processing will be performed at both the Irset institute and the Wellcome Sanger Institute. scRNA-seq dataset (WP 1.1) will be re-used to perform computational pairing with scATAC-seq datasets and transfer RNA-based cell labels (i.e. cell type identification) onto chromatin accessibility data.
Expected size	10-50 TB

Utility	By integrating known TF-gene interactions for the identified cis-regulatory elements, we will reconstruct gene-regulatory networks and predict "master" regulators driving cell fate transition as the human gonad differentiates into a testis or an ovary.
Shareable data	Upon P05-GRL DPO's validation, the following data will be made accessible: · Sequencing data: has been submitted to ArrayExpress/ENA (**private status** until end of january). Project identifier: E-MTAB-10570 · Expression matrix (HGDC_VENTO_scATACseq_aggregated.rds, converted to loom) will be published on the HUGODECA data portal (May 2022) (NB : same protocols as in WP 1.1) Upon P01c-INSERM DPO's validation, the following data will be made accessible: · sequencing data (fastqs) will be submitted to EGA · expression matrix (loom) will be published on the HUGODECA data portal (May 2022)

Reference	HUGODECA_DS_13
Task	1.3 Flow cytometry based surface protein marker screen of human fetal gonads
Task leader	Andreas Bosio (andreasbo@miltenyibiotec.de)
Data contact person	Silvia Rüberg (Sylvia@miltenyi.com)
Description	The data collected for this task will be acquired by flow cytometry : they will result from the processing of freshly dissociated cells with MACSQuant® Instruments and MACS® Marker Screen, which will allow to identify cell surface protein markers and assess their level of expression.
Types and formats	■ raw data acquired via MACSQuant® Instruments will be stored in the .mqd file format. These mqd files can be exported as FCS (data file standard for the reading and writing of data from flow cytometry experiments - 2.0, 3.0, 3.1 compatible) or CSV files. ■ Interpretation data : o List of main cell surface protein markers (.txt) ■ Study, sample and assay metadata (database/spreadsheet)
Origin	Samples are collected by P01c-Inserm lab members (4 pairs of testes from XY individuals aged 11-12 PCW) Screen of 133 antigens (MB)
Expected size	~500 MB
Utility	Identified markers or combinations of markers will serve to better characterize human gonad cells, to develop target cell isolation strategies for cells of interest (antibody-based cell surface protein profiling will determine marker combinations for sorting of live cells) and as starting library for the analysis of gonad development using 2D/3D microscopy in WP3.
Shareable data	[To be confirmed] : list of cell surface markers to be published on the HUGODECA data portal

Reference	HUGODECA_DS_21
Task	2.1 - Generation of 2D ST maps of human gonads development
Task leader	Joakim Lundeberg (joakim.lundeberg@scilifelab.se)
Data contact person	Ludvig Larsson (ludvig.larsson@scilifelab.se)

Description	The data generated for task 2.1 correspond to transcriptome-wide 2D maps of the anatomical landscape of human developing gonads by using spatially resolved mRNA expression analysis technologies developed and managed at SciLifeLab Sweden (http://www.scilifelab.se/). The generation of high quality cDNA libraries with precise positional information for RNA-seq is obtained by placing histological sections on glass slides with arrayed oligonucleotides containing positional barcodes.
Types and formats	■ Raw data: full resolution images (24-bit color TIFF, 16-bit grayscale TIFF, or JPEG) of Hematoxylin and Eosin (H&E) stained developmental gonad tissue sections and raw sequencing data stored in compressed FASTQ format (fastq.gz). ■ Pre-analysis data: o QC images in jpg format o Alignment data compressed in binary format (BAM), with a BAM index (.bai) o Lower resolution images (downscaled version of full resolution H&E images) stored in PNG format (.png) o Tissue position lists stored as comma separated values (.csv) with capture areas coordinates stored in each row for each spot. o Scalefactors for conversion of capture area coordinates between full resolution to lower resolution images stored in JSON format (.json). ■ Analysis data: o Count matrices in mtx and HDF5 format (.h5) with genes in rows and spots in columns. o R format objects (.rds) including normalized expression data, dimensionality reduction results and meta data. o Loupe files (.cloupe) for interactive visualization of data. ■ Interpretation data o Table of differentially expressed marker genes formatted as tab separated values (.tsv). o Table of enriched pathways formatted as tab separated values (.tsv). ■ Study, sample and assay metadata (database/spreadsheet)

Origin	Fresh frozen embedded embryonic abdominal tissue (containing the gonad) will be provided by and P01b-Inserm and P01c-Inserm. Tissue samples of both genders will be collected from embryos aged from 7 PCW and 14 PCW. The tissues will be cryo-sectioned for Spatial Transcriptomics analysis. Optimal Spatial Transcriptomics protocol conditions for gonad tissue will be established, using an in-house developed quantitative assay (the incorporation of fluorescently labeled nucleotides will be measured to quantify the cDNA synthesis on the array). Optimal conditions will next be used on spatially barcoded arrays to generate transcriptome-wide 2D maps of the anatomical landscape of human developing gonads.
Expected size	10-20 TB
Utility	The 2D ST maps will enable to characterize the organization of different cell types in gonadal tissues: this analysis will serve to make a first spatial assembly of the cell types defined by scRNAseq (WP1). It can also find spatially interesting gene expression patterns that has not come out of the scRNAseq data. These genes will be more specifically targeted for WP 2.2 and ST maps are also part of the materials needed as a pre-requisite for WP 4.1
Shareable data	· ST images : to publish on the KeenEye platform (authentication required) · Count matrix (.h5) and high resolution image (tif/jpg): to be published on the HUGODECA data portal (if validated by DPO) · Article (not ready yet)

Reference	HUGODECA_DS_22
Task	2.2 Generation of 2D ISS maps of human gonads development
Task leader	Initially : Malte Kuhnemund (malte@cartana.se) Potentially re-assigned to : Mats Nilsson (mats.nilsson@scilifelab.se) [to be confirmed after amendment of the contract)
Data contact person	Initially : Ivan Hernandez (ivan@cartana.se) Potentially re-assigned to : Marco Grillo (marco.grillo@scilifelab.se)

Description	Cartana's In Situ Sequencing (ISS) kits and service will be used to generate 2D expression maps of 100-200 targeted genes at single-cell resolution (https://www.cartana.se/). The tissue samples will be prepared according to CARTANA Library Preparation Kit protocol. Probes that have specifically interacted with the targeted transcripts will then be amplified by Rolling Circle Amplification (RCA) reaction. Specific barcode sequence in the probe for each targeted gene is then decoded using CARTANA in situ sequencing chemistry and for at least 6 imaging cycles. The signals are detected by fluorescence microscopy. Images are then processed to establish the coordinate of each signal within the tissue and generate the gene expression map.
Types and formats	■ Raw data: initial raw images acquired with an epifluorescence microscope in .czi, .nd2 or .lif formats ■ Pre-analysis data : stitched aligned images in TIFF format ■ Analysis data : o CSV files for x-y coordinates map of each signal, and each cell nucleus. o Cell maps (MATLAB format or ome-tiff) ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	Fresh frozen embedded embryonic abdominal tissue (containing the gonad) will be provided by and P01b-Inserm and P01c-Inserm. Tissue samples of both genders will be collected from embryos aged from 7 PCW and 14 PCW.. Tissue/cell line samples will be sectioned and collected on standard microscopy slides and pretreated according to CARTANA sample handling guidelines. The genes that will be targeted will be those defined as marker genes by scRNA-seq (WP 1.1) and those with spatially interesting expression patterns found by Spatial Transcriptomics (SP) (WP 2.1).
Expected size	For a given sample of 1 cm2, the initial microscope file will be around 200 GB. After stitching and alignment there will be *5* images of 2 GB each/10 GB in total. The size of the expression map will be 2 GB. The overall data size for a 1 cm2 tissue section will be around 250 GB.
Utility	The 2D ST maps will enable to characterize the organization of different cell types in gonadal tissues: this analysis will serve to make a first spatial assembly of the cell types defined by scRNAseq (WP1). It can also find spatially interesting gene expression patterns that have not come out of the scRNAseq data. ST maps are also part of the materials needed as a pre-requisite for WP 4.1.

Shareable data	· 2D expression maps (analysis data) will be made accessible on the KeenEye platform after an authentication process · 1 or 2 images in lower resolution will be displayed on the HUGODECA portal to present the study (available in march) · The list of markers will be accessible on the HUGODECA portal (probably only after article publication) · Pipelines: the pipelines used are accessible on gitlab (https://github.com/Moldia) (private access for now, to become public in march). A link to this page will be added on the HUGODECA data portal. · Articles (planned but not submitted anywhere yet)

Reference	HUGODECA_DS_23
Task	2.3 - Optimization of ST and ISS protocols on cleared tissue samples for the generation of 3D OMICS maps of the same specimen
Task leader	Mats Nilsson (mats.nilsson@scilifelab.se)
Data contact person	Marco Grillo (marco.grillo@scilifelab.se)
Description	The goal of WP 2.3 is to assess the feasibility of ST assay and ISS on sections cut from optically cleared samples after light-sheet imaging. This will tell whether the RNA molecules are intact in cleared samples to provide additional markers potentially useful for spatial analysis. ISS method will be adapted to allow processing of thicker sections to create 3D data and enable spatial maps of cells in 3D. Finally, we will test the robustness, standardize and implement the developed ISS assays for optically cleared sections.
Types and formats	■ Raw data: raw images of cleared tissue samples acquired via light-sheet imaging in TIFF format ■ Pre-analysis data : stitched aligned images in TIFF format ■ Analysis data : o CSV files for x-y-z coordinates map. o 3D omics maps (ome-tiff) ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	Fresh and fixed mouse brain samples have been used to perform an initial benchmark. The samples used for the completion of this dataset will be the same as the ones used for WP 2.1 and WP 2.2. Images of cleared tissue samples will be provided by P01a-Inserm.
Expected size	Image dataset (to be reused in WP4) + protocol : ~20 TB
Utility	This task will allow to define new protocols for generating ST expression data and spatial ISS expression data in optically cleared tissues. The images produced will be re-used in WP4 to generate 3D models of transcriptionally defined cells of entire developing gonads.
Shareable data	The 3DISCO clearing protocol used by P01a-Inserm has the following DOID: (https://doi.org/10.1038/nprot.2012.119) [ISS protocol: confidential status to be confirmed by CARTANA]

Reference	HUGODECA_DS_24

Task	Profile the spatial coordinates of germ and somatic cells using NanoString Whole Transcriptome Atlas (WTA) technology.
Task leader	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Data contact person	Roser Vento (rv4@sanger.ac.uk), Luz Garcia-Alonso (lg18@sanger.ac.uk)
Description	This dataset will be obtained by using GeoMx NanoString technology to provide a Whole Transcriptome Atlas (WTA) of the developing gonads.We will use marker genes obtained from the gonadal single-cell transcriptomic atlas (WP1) to study the cell-extrinsic factors mediating germ cell differentiation and maturation in intact tissues. We have developed an image analysis pipeline that segments germ cells and adjacent somatic cells. Our image analysis pipeline prepares masks that guides the transcriptome collection in a spatially controlled fashion. The transcriptome of the collected material will undergo sequencing and data analysis to quantify potential transcriptomic programs supporting the interaction of somatic and germ cells. Image features such as marker intensity, number and area of segmented cells, relative location in the intact tissue will be quantified post imaging and used to classify the spatially interacting cells.
Types and formats	■ Raw data: raw images of fixed tissue sections, sequencing data ■ Analysis data : o Annotated images of tissue sections o Data matrices with marker intensity, number and area of segmented cells, location in the tissue ■ Study, sample and assay metadata (spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	· Fixed tissue sections will be collected from HDBR covering second trimester female gonads (14-21pcw). · The marker genes used to study the cell-extrinsic factors mediating germ cell differentiation and maturation in intact tissues will be those resulting from WP1 study.
Expected size	~3Gb per image
Utility	The WTA experiment will allow us to study the dialogue between somatic and germ cells during sex specification and gonadal differentiation.

Shareable data	Upon P05-GRL DPO’s validation, the following data will be made accessible: · Sequencing data: will be submitted to ENA · Analysis data : will be published on the HUGODECA data portal - AnnData (.h5ad) - Metadata: .csv The protocols followed will be published on protocols.io An article will be submitted for this study (biorxiv)

Reference	HUGODECA_DS_31 [**SUSPENDED**: Hyperyon access is restricted due to ongoing COVID-19]
Task	3.1 - Imaging mass cytometry of developing gonads
Task leader	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Data contact person	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Description	The data generated for this task will be acquired by Imaging Mass Cytometry, which empowers simultaneous imaging of up to 37 protein markers at a time : this technique will be used to identify gonadal cell types and correlate the findings from WP1 with the spatial location of those cells in situ, thus providing information on the proximity of cells within the gonad.
Types and formats	■ Raw data: images acquired with Fluidigm Hyperion Imaging platform can be saved in MCD TIFF or OME-TIFF format (File headers then contain an OME-XML metadata block compatible with Open Microscopy Environment) ■ Analysis data o Measurement data in CSV files or as records of a database o 2D maps in MATLAB format ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	The samples used will be collected at the Wellcome Sanger Institute (Cambridge, United Kingdom). There will be n=3 samples from 4-6 PCW, n=6 from 6-10 PCW and n=6 from 10-20 PCW male and female gonads.
Utility	Data resulting from this task will be evaluated with those found with WP 3.2 and 3.3 to assess the distribution of distinct cell types in intact gonads as they interact during sex determination and differentiation
Shareable data	/

Reference	HUGODECA_DS_32
Task	3.2 - Single cell cyclic immunofluorescence protein expression analysis of human fetal gonads
Task leader	Andreas Bosio (andreasbo@miltenyibiotec.de)
Data contact person	Werner Muller (wernerm@miltenyi.com) Paurush Praveen (paurushp@miltenyi.com)
Description	The data will be generated by using Miltenyi Biotec's MACsima technology, which enables the simultaneous analysis of multiple markers on a single sample based on fluorescence microscopy. It uses the principle of iterative staining with different fluorochrome-conjugated antibodies to acquire microscopy data for a multitude of parameters. The iterative process comprises three automated main steps: fluorescent staining, image acquisition, and erasure of the fluorescence signal.
Types and formats	■ Analysis data o 2D image stacks in TIFF formats ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	Protein expression and cellular composition will be assessed with a set of more than 100 antibodies on frozen sections of gonads at three developmental stages (in triplicate). The samples used will be collected by INSERM partners (P01a-INSERM, P01b-INSERM).
Expected size	50 TB during first year
Utility	Data resulting from this task will be evaluated with those found with WP 3.1 and 3.3 to assess the distribution of distinct cell types in intact gonads as they interact during sex determination and differentiation

Reference	HUGODECA_DS_33
Task	3.3 - A 3D cellular atlas of the developing human
Task leader	Alain Chédotal (alain.chedotal@inserm.fr)
Data contact person	Yorick Gitton (yorick.gitton@inserm.fr)
Description	The data generated for this task will be acquired with Light-Sheet Fluorescence Microscopy (LSFM) from solvent-cleared human gonads and genital ducts.
Types and formats	■ Analysis data o 3D image stacks in TIFF formats ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	Entire organs and human embryos between 5 and 12 PCW (at least 20 cases of both sexes) will be collected by P01-bInserm. Strong candidates as makers of gonadal determination have already been identified in preliminary experiments by P01a-Inserm and P01b-Inserm and genes identified in WP1 and WP2 will also be taken into account.
Expected size	Initial assessments on the new ultramicroscope have produced 2.5TB from 22 scans over four weeks. Expecting to reach **15TB/month** during 1st year.
Utility	This dataset will enable to have a better understanding of gonadal cell lineage relationships in entire organs and human embryos between 5 and 12 PCW and to determine the organization of vasculature and innervation during sex determination. The technology and the protocol that will be defined during this task will also be applied to study gonad samples in WP5.
Shareable data	· 3D images will be viewable via KeenEye's platform (authentication process required) · Low resolution image (2D) will be displayed on the HUGODECA data portal to present the study [To be confirmed]

Reference	HUGODECA_DS_40
Tasks	4.1 - Simultaneous visualization of H&E and immunofluorescent stained tissue and OMICS data 4.2 - Fully 3D rendering/viewer 4.3 - Use existing data for the development of novel 3D integration tools 4.4 - Integration of 2D Cyclic immunofluorescent data in 3D images
WP leader	Sylvain Berlemont (sylvain.berlemont@keeneye.tech)
Data contact person	Sylvain Berlemont (sylvain.berlemont@keeneye.tech)
Description	The data generated for the different tasks of WP 4 correspond to the source code of a 2D/3D viewer developed by Keen Eye Technologies. This viewer will be developed to visualize data from WP 1, 2 and 3 in a combined way: LSFM data, 2D cyclic immunofluorescence, ST, *ISS* as well as single-cell approaches will all be incorporated into one scaffold. Computational models will also be defined and tested to check the feasibility of integrating datasets from different individuals at a similar development stage.
Types and formats	■ Coding data in Python and C++
Origin	The data used to test the viewer will come from WP 1, WP 2 and WP 3. Existing datasets from spatial platforms will also be used in an initial step for the development of new bioinformatics and 3D integration tools. This model will be built upon a training data set using another organ system (embryonic heart) on which transcriptomics data (ST and ISS) and protein stains on cleared embryos (https://transparent-human-embryo.com/) already exist.
Expected size	several MB
Utility	The software solution will allow to deliver OMICS spatial 2D and 3D maps of gonad development. This technology will also be used for WP 5, to better study the effects of alterations on human gonad development.

Reference	HUGODECA_DS_51
Task	5.1 - Manipulation of the intrinsic *WNT/β-catenin* signaling pathway
Task leader	Anne Jorgensen (Anne.Joergensen.02@regionh.dk)
Data contact person	Anne Jorgensen (Anne.Joergensen.02@regionh.dk)

Description	The data generated for this task are diverse since both ex-vivo techniques and imaging technologies will be used to understand the effects of the manipulation of the WNT/β-catenin signaling pathway in fetal testes and ovaries (with a focus on the specification of Sertoli and granulosa cells): inhibition of WNT signaling will be done using IWR-1 in fetal ovaries, while its stimulation in fetal testes will be achieved with CHIR 99021 in combination with recombinant WNT4. Subsequent analyses of organotypic cultures will include 3D-cellular maps, combining whole-mount immunohistochemistry (IHC) and 3D imaging of solvent-cleared organs (3DISCO) with light-sheet fluorescence microscopy (LSFM) (P01a-Inserm, P01b-Inserm), in addition to classic IHC investigation of cell lineage markers and mass spectrometry measurement of sex hormones (P04-RegionH).
Types and formats	For ex vivo techniques: Analysis data o Mass spectrometry measurements of sex hormones (.csv) o Expression data for granulosa and Sertoli cell markers (TIFF) Interpretation data: o List of altered cell lineage markers (.doc, .pdf or .html) o Hormone levels (.xls) Study, sample and assay metadata (database/spreadsheet) Methodological data (.doc, .pdf or .html) For 3D cellular maps, the types and formats of the generated data will be the same as those mentioned for the task 3.3.
Origin	Manipulations will be done for 2 weeks in organotypic culture of fetal gonads aged 6- 9 PCW (P04-RegionH) and compared with vehicle-treated control samples.
Expected size	5-10 TB for 3D imaging
Utility	Abnormally developing gonads, either due to genetic or environmental alterations, cannot be obtained from in vivo specimens. Using organotypic ex vivo models will allow to recapitulate both normal and dysregulated human fetal gonad development. This dataset will allow to better understand the role in the case of the intrinsic WNT/þ-catenin signaling pathway during sex determination in Humans. This study will also be a demonstrative use case of the implementation of 3D cellular maps.

Shareable data	· 3D images will be viewable via KeenEye's platform (authentication process required) · Low resolution image (2D) will be displayed on the HUGODECA data portal to present the study [To be confirmed] · The publication of an article is planned

Reference	HUGODECA_DS_52
Task	5.2 - Manipulation of the intrinsic DHH signaling pathway
Task leader	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Data contact person	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Description	The data generated for this task will result from the application of multiple techniques: ex-vivo techniques, 3D-imaging technologies and scRNAseq. The DHH signaling pathway will be manipulated by using organotypic cultures of 7 PCW testes. The manipulation will be done with known receptor agonists (e.g. SAG) and inhibitors (e.g. cyclopamine) of the different steps of this pathway (P01c-Inserm). The different Leydig cell populations will be stained in cultivated testes and further processed for volume-imaging and 3D-automated cell quantification (P01a-Inserm, P01b-Inserm), hormone levels will be measured in the media (P01c-Inserm, P04-RegionH), and scRNA-seq will be performed to understand how cell lineage differentiation is affected (P01c-Inserm).
Types and formats	For ex vivo techniques : ■ Analysis data o Mass spectrometry measurements of sex hormones (.csv) ■ Interpretation data o Hormone levels (.csv) ■ Study, sample and assay metadata (database/spreadsheet) ■ Methodological data (.doc, .pdf or .html) For scRNA-seq, the types and formats of the generated data will be the same as those mentioned for the task 1.1. For imaging data, the types and formats will be the same as those mentioned for task 2.3.
Origin	Manipulations will be done in organotypic cultures of 7 PCW testes (P01c-Inserm). The bioinformatics analysis (scRNA-seq) will include a comparison with uncultivated testes of the same age.

Expected size	5-10 TB for 3D imaging 2-5 TB for scRNA-seq
Utility	The data in this study will enable to understand the involvement of the DHH signaling in the complex patterning of the Leydig cell lineage in human fetal testes. This study will also be a demonstrative use case of the implementation of 3D cellular maps in combination with scRNA-seq data.
Shareable data	Upon P01c-INSERM DPO's validation, the following data will be made accessible: · sequencing data (fastqs) will be submitted to EGA · expression matrix (loom) will be published on the HUGODECA data portal

Reference	HUGODECA_DS_53
Task	5.3 - Manipulation of Nrp1 signaling pathway
Task leader	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Data contact person	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Description	The data generated for this task will mainly result from the application of 3D imaging technologies on 7 PCW testes in organotypic cultures, which will have been treated with a validated NRP1 blocking function antibody.
Types and formats	Same as those mentioned for tasks 1.1 and 3.3
Origin	Organotypic cultures of 7 PCW testes (P01c-Inserm)
Expected size	5-10 TB for 3D imaging
Utility	The data generated for this task will enable to understand the involvement of the NRP1 signaling in the establishment of a sexually dimorphic vasculature during human gonadogenesis. This dataset will also demonstrate the utility of 3D-imaging technologies as a means of observing changes in vascularization after intrinsic alterations on human gonad development.
Shareable data	· 3D images will be viewable via KeenEye's platform (authentication process required) · Low resolution image (2D) will be displayed on the HUGODECA data portal to present the study [To be confirmed]

Reference	HUGODECA_DS_54
Task	5.4 - Ibuprofen as a prototype of extrinsic factor known to alter the differentiation of both the human ovary and testis
Task leader	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Data contact person	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Description	This data generated for this task will result from the application of scRNA-seq following organotypic cultures of fetal testes (8-10 PCW) and ovaries (10-12 PCW) with or without Ibuprofen.
Types and formats	Same as those mentioned for task 1.1
Origin	Manipulations will be done in organotypic cultures of 8-10 PCW testes and 10-12 PCW ovaries (P01c-Inserm). The bioinformatics analysis (scRNA-seq) will include a comparison with uncultivated testes and ovaries of the same age.
Expected size	2-5 TB
Utility	The data in this study will allow to identify the cellular targets of Ibuprofen and the corresponding gene pathways altered. This study will also be a demonstrative use case of the utility of reference scRNA-seq dataset (uncultivated testes and ovaries) for the understanding of toxicological effects.
Shareable data	Upon P01c-INSERM DPO’s validation, the following data will be made accessible: · sequencing data (fastqs) will be submitted to EGA · expression matrix (loom) will be published on the HUGODECA data portal

Specify if existing data is being re-used (if any)

See the "origin" field in the list of datasets mentionned above

Specify the origin of the data

See the "origin" field in the list of datasets mentionned above

State the expected size of the data (if known)

See the "size" field in the list of datasets mentionned above

Outline the data utility: to whom will it be useful

See the "utility" field in the list of datasets mentionned above

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

Study and assay metadata will be created to document the different experiments defined in WP 1, WP 2, WP 3 and WP 5 (this is not applicable to WP 4, since it deals with software development).
During the first stage of the project, we will gather information about the program and the different studies that are planned by using SEEK (D7.1) (https://studies.hgdc.genouest.org/) This application follows the Investigation, Study and Assay (ISA) specifications, and we will create one investigation by work package, and one study by project task. Metadata templates (conformant with the HCA metadata model) have been created to define the different sample types used in the assays (https://studies.hgdc.genouest.org/sample_types) and the protocols applied. Complementary information about the metadata collected is given in the part about data interoperability.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

To become findable (and citable), datasets should get a unique, persistent identifier. Persistent identifiers like Digital Object Identifiers (DOI) have been designed as a solution to avoid link rot (that is, when a hyperlink stops referring to the original source because it was moved). These persistent identifiers thus ensure that the data is and will be findable. For single cell datasets, all raw and processed data deposited on public repositories can also be assigned a DOI. An issue that remains to be addressed is that of the assignment of a DOI to the different image stacks.

If they do not already have one, the HUGODECA consortium members will also be encouraged to get a persistent author identifier via ORCID, so as to increase the connections towards the research work.

Outline naming conventions used

Naming files, folders and records consistently will facilitate their retrieval, and will enable users to browse the data more effectively and efficiently. For this reason, some basic naming convention should be agreed upon by the consortium members.

Preliminary remarks
Special characters (& * % $ £] { ! @) and spaces should be avoided and that the full stop should

preferably be reserved for indicating the separation between the file name and the file extension.

Codes and index used in sample and file names

Project code
HUGODECA will be used when the project code is used for folders and communication documents
HGDC will be used when the project code is used in sample names or generated file names.

WP/Task index
If a WP or task index needs to be mentioned in a file or dataset name, only the numbers are used (the dot is not included). If the implementation level is a whole WP, the index will be the WP number followed by 0
Examples:
index for Task 1.1 is 11
index for WP 1 is 10

Site Code

Partner	site Code	Country
P01b-Inserm	LI	France
P01c-Inserm	RE	France
P04-RegionH	RI	Denmark
P05-GRL	WS	UK

Codes to identify protocol types

Code	Protocol types
COLLEC	Collection protocol
DIFFER	Differentiation protocol
DISSOC	Dissociation protocol
IMGPREP	Imaging preparation protocol
IMAGING	Imaging protocol
SEQUEN	Sequencing protocol
LIBPREP	Library preparation protocol
ANALYS	Analysis protocol

Naming convention in the metadata

- For donor organism (embryos), the pattern is: HGDC_%siteCode%%alphanumericIndex%

Examples : HGDC_RE2345

For derived samples (tissues, organs..), the pattern is: %donorOrganismCode%_%derivedInfoCode%
Derived samples always start with the donor organism (embryo or fetus) code, followed by descriptor codes that either refer to the organ (ie GON1), a cell suspension code (ie SC1) or molecules applied for an organotypic culture.

Examples: HGDC_RIh2143_CHIR: code corresponding to a tissue cultured with CHIR99021

Naming convention for sequencing files and gene expression matrix

Fastq files will be named according to Illumina nomenclature, with the the sample number, which is a numeric assignment based on the order that the sample is listed in the sample sheet. *Example:* HGDC_RE2345_RNA1_L001_R1_001.fastq.gz

Naming convention for images and image datasets

With the actual naming convention, the name of images and image datasets starts with the identifier of the sample used as an input to generate the data. Filenames carry exhaustive descriptors corresponding to the different channels, laser sources, and tiles (ie: HDC_RE2341_ac217-12-waist_561-TH_647-pax2_790-periph-0o63X-2um_LR).

For datasets that will be made accessible via the keenEye platform, the goal is to obtain shorter filenames, generated according to the following pattern: FileNumber_Marker_PCWindex (ie: 000905_Pax2_pcw12) since information about other descriptors will be stored as metadata. The images are going to be organized in directories / subdirectories as follow:

- HGDCidentifier_marker_PCWindex
    - CanalA
        - Images

NB : there will be one ‘image stack’ by canal

Outline the approach towards search keyword

Keywords associated to a dataset should correspond to ontology terms, whenever possible. They should at least mention the organ, the developmental stage, the sex of the embryo, the different technologies used, and the main biological entities studied (tissue type, genes…)

Outline the approach for clear versioning

For documents (.doc, .pdf, …), versioning will be maintained by following this pattern :
[document name]_[version number] _[status: DRAFT{0,1}]
Example : HUGODECA_DMP_7-4_V1.doc

For data that can be generated multiple times , the creation date may be indicated instead of a version number. The date will then be expressed with this format: YYYY-MM-DD

For software and scripts, versioning will be controlled via the use of dedicated tools (Subversion, Git…)

(As it may be confusing and storage-consuming to have too many similar or related files: a good file versioning practise is also to discard and erase intermediate working files - and keep only raw data, last version of processed data, and the definitive copy of the analysis data)

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Naming cell types is an issue that remains to be adressed by a dedicated group, gathering ontologists and scientific experts.

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

As regards OMICS data generated during the research phase, only raw and analysis data will be published (ie sequencing data and expression matrices with their associated metadata). There is no point in giving public access to pre-analysis data as long as raw data and analysis data are available. As for interpretation data, they will be an integrated part of the scientific publications.

It is to be noted that sequencing data corresponds to genetic data (according to Recital 34 of the GDPR, genetic data includes chromosomal, DNA or RNA analysis, or any other type of analysis that enables you to obtain equivalent information) and as such, it is considered special category data (ie personal data requiring extra protection). In order to lawfully process special category data, a lawful basis under Article 6 and Article 9 of the GDPR is required (see section about ethics below): open (or controlled) availability of this data will here be dependent on the fact that the individual has given clear consent for the HUGODECA research teams to process this type of data for the research purpose indicated in the consent forms (mention of secondary use for unspecified research will also be required for open accessibility, and mention of secondary use for a specified type of research will be required for controlled accessibility).

2D and 3D images will be made available in a controlled way via the KeenEye platform (authentication required). The size of 3D datasets ranges from tens to hundreds of GB, and thus required specific softwares and services to be processed (to be displayed online or to be transferred for instance). For 3D datasets, all requests for access will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. For ST data (2D), high resolution images used for the analysis will be made openly available on the HUGODECA website.

Protocols and pipelines used or defined during the project will be made openly available, whenever it is possible: during the preliminary exploitation plan of the HUGODECA project, the protocols for generating spatial ISS expression data in optically cleared tissues (part of HUGODECA_DS_2-3) have been identified as a potential candidate for patent-filling. This protocol may thus remain confidential (to be confirmed by CARTANA).

Code / software: The data generated for the different tasks of WP 4 correspond to the source code of a 2D/3D viewer developed by Keen Eye Technologies, and will remain confidential: a patent will be filled in order to secure the commercial exploitation of this innovative solution on the market (Information relating to the patent that has been registered must be submitted under the 'IPR' section of the EU Participant Portal)

Scientific publications will be either gold or green open access (free of charge, online access for any user):

- Gold open access means that the publication is available by the scientific publisher as open access. Some journals require an author-processing fee for publishing open access.
- Green open access or self-archiving means that the published article or the final peer-reviewed manuscript is archived by the researcher itself in an online repository, in most cases after its publication in the journal. The journal must grant the researcher the permission to self-archive the final peer-reviewed article, at the latest,12 months after publication.

For the HUGODECA project, gold open access will be the preferred way of publishing the results.

Specify how the data will be made available

OMICS and image datasets

Expression matrices intend to be listed and made viewable on the HUGODECA web portal (D6.4), with a link to download the datasets from external repositories when applicable. As already mentioned, the consortium members intend to take part in the HCA initiative, so analysis data may also become available on the HCA data portal, if all legal and ethical requirements are met.

In this section, we list the different repositories that have been identified as potential places where the data could be submitted to guarantee their accessibility in the long-term (actual submission of OMICS data to these repositories is dependent on the validation of the DPO of the research team)

For raw sequencing data, the repositories of reference are the European Nucleotide Archive[1] (ENA) (open access) and the European Genome-phenome Archive (EGA) (controlled access). Both are ELIXIR Core Data Resource: they are part of a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data. This means that they have been evaluated according to : their scientific focus and quality of science, the community served by the resource, the quality of service, the legal and funding infrastructure and governance, the impact and translational stories.

EGA

This repository enables to publish personally identifiable genetic and phenotypic data resulting from biomedical research projects in a controlled way. The EGA is co-managed by EMBL-EBI and CRG (Centre for Genomic Regulation). Since the introduction of GDPR in May 2018, EMBL has established an internal policy on General Data Protection (IP68) and the CRG operates within the EU and so fully complies with the GDPR. With regard to GDPR, EGA is a data processor as it processes data as instructed by the data controller, who is the person who submits the data to EGA (implementation of the GDPR is explained in more detailed in the document listed below).

Information or service provided	Link to page or service (last accessed 2022/01/24)
EGA Helpdesk (contact)	helpdesk@ega-archive.org
EGA submission guidelines and FAQs	https://ega-archive.org/submission
EGA submitter portal	https://ega-archive.org/submission/tools/submitter-portal
EGA security document	https://ega-archive.org/files/European_Genome_phenome_Archive_Security_Overview.pdf
EGA and implementation of the GDPR	https://ega-archive.org/files/EGA_GDPR.pdf
EGA data processing agreement for the submission and distribution of personal data	https://ega-archive.org/files/EGA_Data_Processing_Agreement_v1.1.pdf
Data use conditions in EGA	https://ega-archive.org/data-use-conditions

ENA

The ENA accepts all kinds of nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. It is an open access repository, and as such, data submitters must ensure that the datasets they deposit can be reused for unspecified research. The ENA is developed and maintained at the EMBL-EBI under the guidance of the INSDC (International Advisory Committee and a Scientific Advisory Board).

Information or service provided	Link to page or service (last accessed 2022/01/24)
ENA Helpdesk	https://www.ebi.ac.uk/ena/browser/support
ENA submission guidelines and FAQs	https://ena-docs.readthedocs.io/en/latest/
ENA submission tools	https://www.ebi.ac.uk/ena/browser/submit
ENA policy	https://www.ebi.ac.uk/ena/browser/about/policies

Single cell experiments can also be submitted on ArrayExpress ArrayExpress accepts all functional genomics data generated from microarray or next-generation sequencing (NGS) platforms: expression matrices and sequencing data can be submitted at the same time (if sequencing data are submitted, they are then brokered to the ENA). ArrayExpress is an ELIXIR Core Data Resource. Just like the ENA, it is an open access repository maintained at the EMBL-EBI.

Information or service provided	Link to page or service (last accessed 2022/01/24)
ArrayExpress single cell submission guide	https://www.ebi.ac.uk/arrayexpress/help/single-cell_submission_guide.html
ArrayExpress submission tools	https://www.ebi.ac.uk/fg/annotare/login/
ArrayExpress FAQ	https://www.ebi.ac.uk/arrayexpress/help/FAQ.html
ArrayExpress Single cell atlas search page	https://www.ebi.ac.uk/gxa/sc/home

HCA (Human Cell Atlas) is an international collaborative consortium that charts the cell types in the healthy body, across time from development to adulthood, and eventually to old age. It is building a reference map of all cells in the human body through the creation of an online database made up of gene expression data (also called transcriptomic data). HCA is an open access repository : the HCA DCP currently does not accept any data that requires controlled access, but datasets with legal/ethical constraints could eventually be included in next phases of development of the HCA DCP. As HCA is hosted in the US, it is to be noted that data transfer has to conform to the principles mentioned in Chapter V of the GDPR (https://gdpr-info.eu/chapter-5/).

Information or service provided	Link to page or service (last accessed 2022/01/24)
HCA submission guidelines	https://data.humancellatlas.org/contribute
HCA contact for submission	wrangler-team@data.humancellatlas.org.
HCA entry page to access the data	https://data.humancellatlas.org/explore/
HCA Data coordination platform	https://www.humancellatlas.org/data-coordination-2/
HCA template consent form for developmental atlas	https://drive.google.com/file/d/1O26gT8p3_hBOdAPDLhNz0pOKxo5epGD7/view
HCA metadata types	https://data.humancellatlas.org/metadata

For bioimaging data in general, there is an open repository for images, the Image Data Resource (IDR), but the deposit of HUGODECA image datasets is still being discussed as the volume needed to store the stacks of images is way over the normal expected size.

Information or service provided	Link to service or page (last accessed 2022/01/24)
IDR submission guidelines	https://idr.openmicroscopy.org/about/submission.html
IDR FAQ	https://idr.openmicroscopy.org/about/faq/
IDR search page, to access the data	https://idr.openmicroscopy.org/cell/ https://idr.openmicroscopy.org/tissue/

Methodological data

Protocols.io is on web platform for developing and sharing reproducible methods. Every new protocol starts out private: it may then either remain private (5 private protocols maximum per workspace with a free account) or become public to be shared with all the scientific community. Protocols then receive a DOI (they are archived with CLOCKSS to ensure long-term preservation of the knowledge)..As part of the goal to be part of the HCA initiative, HUGODECA consortium members are encouraged to join the Human Cell Atlas Method Development Community (https://www.protocols.io/groups/hca).
Protocols used or defined by the HUGODECA consortium will also be listed on the HUGODECA data portal.

Pipelines and softwares

Partners that are part of an academic institution will make their scripts publically available on a dedicated platform. GitHub is already used by several of the research groups involved in the project for that purpose (e.g.https://github.com/Moldia,https://github.com/SpatialTranscriptomicsResearch/st_pipeline, https://github.com/umr1085-irset).

Information or service provided	Link to service or page (last accessed 2022/01/24)
Github presentation page	https://github.com/about
Github privacy statement	https://docs.github.com/en/github/site-policy/github-privacy-statement
Github and data protection	https://docs.github.com/en/github/understanding-how-github-uses-and-protects-your-data
Github status page	https://www.githubstatus.com/
Github information page on licensing a repository	https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/licensing-a-repository

Scientific publications

The pre-publication of results is encouraged so as to make research results available as soon as possible to the public community, notably by posting pre-prints on bioRxiv.

Information or service provided	Link to page or service (last accessed 2022/01/24)
bioRxiv submission guidelines	https://www.biorxiv.org/submit-a-manuscript
bioRxiv contact for submission	https://submit.biorxiv.org/
bioRxiv search page to access data	https://www.biorxiv.org/search

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Genomics and transcriptomics data are in file formats which are widely used within the scientific community, that is fastq.gz files for raw sequencing data, bam files for aligned sequences and count matrices in HDMF5-loom, mtx or csv formats for processed data, making these fully readable and re-usable worldwide from a technical viewpoint.

Raw images can be viewed with open-source softwares like the Fiji/ImageJ tools (Fiji is an open source project hosted in a Git version control repository and comprehensive documentation, ImageJ source code is available online, with its user manual). Open-source softwares like Cell Profiler and histoCAT can also be used to analyse imaging data. However, in order to visualize the fully annotated 2D and 3D cell maps, Keen Eye 3D viewer will be necessary: Keen Eye Platform™ is a proprietary and patented software-as-a-service platform and the viewer will be accessible through a cloud service, promoting high accessibility across a worldwide research community.

Specify where the data and associated metadata, documentation and code are deposited

Information about the way each dataset is intended to be deposited is given in the “shareable data” entry of the list of datasets in the data summary. We list here the deposition status according to the origin of the samples, as permissions to publish mainly depend on the consent forms that have been used and on the validation of the DPO of each institute.

Origin and use	Information about deposit
Samples collected by P01-Inserm and used by P01-Inserm, P02-KTH, P03-SU	No datasets have been submitted to external repositories as for now. Intended deposit : § Sequencing data generated by P01c-Inserm (WP1.1, WP1.2, WP5.2, WP5.4) will be submitted to the EGA (controlled access) with specific data use conditions § Expression matrices generated by P01c-Inserm and P02-KTH (WP1.1, WP1.2, WP2.1, WP5.2, WP5.4) will be viewable on the HUGODECA data portal § Sequencing and analysis data will be made available in HCA when a controlled access deposit will be implemented (and upon validation that data transfer to this non-EU repository is permitted) § 2D/3D high resolution cell maps (P01a-Inserm, P01b-Inserm, P02-KTH, P03-SU) will be deposited on the cloud-based storage server used by KEENEYE
Samples collected by HDBR and used by P05-GRL	§ Sequencing data generated by P05-GRL (WP1.1 and WP1.2) have been submitted to ArrayExpress/ENA and are in private status for now (to become public upon the validation by DPO / Ethics group). § Sequencing data generated by P05-GRL (WP2.4) intend to be submitted to ArrayExpress/ENA. § After the data are published on ArrayExpress/ENA, sequencing data will be made available in HCA (upon validation that data transfer to this non-EU repository is permitted)
Samples collected and used by P04-RegionH	2D/3D high resolution cell maps (P01a-Inserm, P01b-Inserm, P02-KTH, P03-SU) will be deposited on the cloud-based storage server used by KEENEYE

Specify how access will be provided in case there are any restrictions

- The sequencing data generated by P01c-Inserm require the following data use conditions: “population origins or ancestry research prohibited”, “ethics approval required”, “health/medical/biomedical research and clinical care” (These conditions are to be confirmed by the DPO)
- Access to the KeenEye platform will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving

institution hosting the requesting PI user. Commercial destinations will be prohibited

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

Types of metadata required

For samples and processes, a special attention will be given to the metadata used within the HCA initiative.

Regarding samples, there should be metadata about:

- The donor organism (**embryo**) : unique anonymous identifier (as transmitted by the provider), biological sex (female, male, mixed or unknown), developmental stage (in PCW), species (“Homo sapiens”, NCBItaxon:9606), known diseases (it should default to "no disease", which is equivalent to "MONDO_0000001");
- The sample in itself: unique identifier and organ from which the sample was collected.

NB : **no personal data about the mother** (indirect or direct) will be collected in the metadata for submission to external repositories.

Regarding sample processing and data acquisition, the metadata depend on the technology used. When single-cell technologies are involved, metadata should include information about:

- The dissociation procedure
- The library preparation protocol :
  - Library construction method (e.g. “10x v3 sequencing”), library construction kit (name, manufacturer and catalogue number)
  - The input nucleic acid molecule (e.g. “polyA RNA")
  - The nucleic acid source (e.g. “single cell”)
  - End bias (e.g “3 prime tag", "3 prime end bias", "full length”)
  - Strand ("first", "second", "unstranded" or "not provided")
  - Optional properties from those listed in the metadata for HCA
- The sequencing protocol :
  - Instrument used for sequencing and model (e.g. Illumina HiSeq 4000), the sequencing design (e.g. 2x100 bp) and the number of raw reads
  - 10x specific metadata such as fastq creation method (e.g. “Cellranger mkfastq” or “Illumina bcl2fastq”), fastq creation method version (e.g. “Cellranger 3.1”)

When imaging technologies are involved, metadata should include information about:

- Tissue preparation and the imaging preparation protocol (in the case of cryosectioning only) :
  - Fixation
  - Image slide thickness
  - Final slicing method (cryosectioning) …
- The imaging protocol :
  - Microscopy (« fluorescence microscopy »)
  - Magnification, Numerical aperture, pixel size …

These metadata can be found in the HCA metadata dictionary. Those which are the most relevant for the project are : Specimen from organism, Imaged Specimen, Imaging preparation protocol, Library Preparation Protocol, Sequencing protocol.

For biological imaging data, the OME Model defines a set of metadata to include, such as XYZ dimensions and pixels type, as well as extensive metadata on, for example, image acquisition, annotation, and regions of interest (ROIs).

Metadata values

In order for all participants to use the same terms, ontologies and controlled vocabularies are going to be used. Whenever possible these values should correspond to the terms defined in the HCA ontology. If this ontology is incomplete for our needs, the concepts used will preferably come from other existing ontologies, including but not restricted to:

- EFO (Experimental Factor Ontology) which provides information about many experimental variables available in EBI databases, while Eagle-i resource (ERO) is an ontology of research resources such as instruments, protocols, reagents, animal models and biospecimens
- the Biological Imaging Methods Ontology (fBBi) is an ontology dedicated to the terms used in biomedical research for imaging and visualization methods, and the BioAssay Ontology (BAO) describes biological screening assays and their results including high-throughput screening (HTS).
- Embryonic structure is described in UBERON and EHDAA2 provides a structured controlled vocabulary of stage-specific anatomical structures of the developing human and is linked to HSAPDB, which includes carnegie stages.
- Different ontologies describe properties and classes at the cell level : the cell ontology (CL) is a general ontology which applies to cell types in animals; and CPO structures the vocabulary related to morphological and physiological phenotypic characteristics of cells, cell components and cellular processes.

For the different metadata fields which are required, we will thus define the ontology or controlled list from which the values should be part, as described in the table below (this table is intended for illustrative purpose only, it is not an exhaustive account of the different metadata fields).

Metadata name	Type	Description
species	ontology	ontology identifier from NCBItaxon (9606)
disease	ontology	ontology identifier from MONDO (if no disease, current accepted proxy for "no disease" is "MONDO_0000001")
organ	ontology	ontology identifier from Uberon
development stage	ontology	ontology identifier from HsapDv (e.g HsapDv_000007)
sex	controlled list	one of ["male", "female", "mixed", "unknown"]
sample_type	controlled list	one of ["cell line", "organoid", "direct from donor - fresh", "direct from donor - frozen", "cultured primary cells"]
tissues	Ontology	ontology identifier from Uberon
cell type	Ontology	ontology identifier from CL (e.g. CL:1001610)
gene ID	Controlled list	Gene identifier from Ensembl
gene name	Ontology	as referenced in the ontology version of the human gene nomenclature (HGNC)
imaging method	Ontology	ontology identifier from fbbi

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Almost all metadata can be mapped to commonly used ontologies, but as mentioned beforehands, naming cell types in astandardized vocabulary remains an issue.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The ENA, EGA and ArrayExpress are managed by EMBL-EBI which tries to minimize barriers to reuse of data by adopting the Creative Commons (CC) licence framework across all its data resources. The licence intends to be explicitly stated on the resource and at the record level, in both humand and machine-readable formats (it is not however implemented yet).

In the case of coding data, documentation and literary programming is promoted : whenever possible, tutorials in the form of notebooks (jupyter notebooks, R notebook) will be published to document analysis scripts so as to integrate the code with the corresponding narrative and documentation. Different pipelines intend to be published on github, which provides guidance on licensing a repository (https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/licensing-a-repository, last accessed 2022/01/24). The MIT licence is the licensing option advised for the HUGODECA project.

In conformance with the principle of being "as open as possible, as closed as necessary", the HUGODECA consortium will also support partners in protection of their results, and help them secure future exploitation opportunities.During the preliminary exploitation plan of the HUGODECA project, the deliverables D4.1, D4.2, D4.3 had been identified as potentially patentable [*patent status to be confirmed by the end of the project in June*]

In the case of bioimaging data, all requests for access will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. Commercial destinations will be prohibited.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

Data will be made available upon project completion (or article publication) to protect scientific information and guarantee the rights of consortium members to be the first to present or publish large-scale analyses of their results.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

OMICS data intend to be published on external repositories, with restrictions based on permissions

given in the consent forms.

Describe data quality assurance processes

By giving access not only to raw and analysis data, but also to study metadata, methodological data, and scripts used to facilitate the process and analysis of the data, the HUGODECA consortium pave the way for data re-usability. The fact that the different protocols established by the consortium members will be published will enable to check that quality assurance processes are ensured. In the case of coding data, documentation and literary programming is promoted : whenever possible, tutorials in the form of notebooks (jupyter notebooks, R notebook, KnitR, Sweave) will be published to document analysis scripts.

Specify the length of time for which the data will remain re-usable

Datasets published on external repositories such as ENA, EGA, ArrayExpress and HCA will be stored for permanent archiving.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

12 person-months have been requested by P01c-INSERM for data management (WP7). A data manager (L. Noël, P01c-Inserm) has been recruited to this end. She will help manage the data with the help of a data contact person, identified for each task (7 person-months requested globally to this end). They will notably implement tools necessary for data management, provide the relevant metadata during the course of the project, keep track of the dissemination level of the different data sets, and refine the naming convention.

20 000 euros have also been requested to cover the costs related to open access publications (author-publishing fees).

Cloud data storage (100 TB / month over 18 months) for 3D images has been estimated at a cost of 54000 euros: this sum has to be taken into account when considering the long-term preservation of imaging data, since this contract will have to be renewed. For that reason, partnerships with European institutes (CINES; CC-IN2P3, EU Image Data) that can provide long-term support for data hosting are under negotiations, all as well as the possibility to deposit the images in IDR.

Clearly identify responsibilities for data management in your project

A data manager (L. Noël, P01c-Inserm) has been recruited to manage the data with the help of a data contact person, identified for each task (7 person-months requested globally to this end). They will notably implement tools necessary for data management, provide the relevant metadata during the course of the project, keep track of the dissemination level of the different data sets, and refine the naming convention.

Describe costs and potential value of long term preservation

Cloud storage has to be taken into account when considering the long-term preservation of imaging data, since this contract will have to be renewed. For that reason, partnerships with European institutes (CINES; CC-IN2P3, EU Image Data) that can provide long-term support for data hosting are under negotiations, all as well as the possibility to deposit the images in IDR.

For OMICS data, the deposition of raw and analysis data in public repositories remain the best alternative to ensure their long-term preservation: those are trustworthy data repositories, with a certificate or explicitly adhering to archival standards.

The long-term preservation of HUGODECA data is particularly valuable since they deal with unique specimens, analyzed at crucial developmental stages. As the ethical aspects are particularly sensitive, these datasets will be a major source of information, if future regulations should come to restrict this work as in other major countries. The HUGODECA programme is already of a high value for researchers who can not carry out experiments on this type of specimens due to restrictive regulations in their home country.

The data generated for the HUGODECA project have also a high commercial and educational potential: the study of the cell atlas may eventually be included into university curriculum (medical, biology, etc.) and converted as a unique learning 3D tool for anatomy/surgery. It is also to be noted that the consortium members have already received many requests from documentary makers (TV, etc.) for anatomy content and human fetus discovery.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Data storage

Initial storage of the data will be performed locally. For academic groups involved in the project, institutional ICT facilities will ensure a secure environment, with firewall system in place, virus intruder protection, and access to digital files controlled with encryption and/or password protection and SSL encryption for data transfer. Industrial partners will observe the same types of security measures, and dedicate specific storage space for the data generated for this project.

The sheer size of expected data makes it difficult to follow back up best practices (at list 3 copies on at least two different media). As a general rule, we remind here that raw data are usually considered to be the master copy of any given record (or golden copy), therefore, there should be a back-up copy of all raw data. Submitting raw sequencing data to external repositories is a way to ensure their long-term storage: upon the different DPOs' validation, sequencing data produced by P01-cInserm intend to be submitted to EGA (controlled access is required since unspecified research use is not permitted) and sequencing data produced by P05-GRL intend to be submitted to ENA (open access). The EGA contributes and helps to defines guidelines, best practices, and standards for building and operating an infrastructure that promotes responsible data sharing in accordance with the Global Alliance for Genomics and Health (GA4GH) Privacy and Security Policy (cf. https://ega-archive.org/files/European_Genome_phenome_Archive_Security_Overview.pdf, last accessed 2022/01/24)

For images, the important point is also to keep track and preserve all biomaterials used to produce the

image captures, since it is another way to make sure that these images can be reproduced. The use of a Dell EMC Isilon storage system (approx. 540 TB) enables to archive image datasets. This server is hosted by Genouest which has been ISO9001:2015 certified for Software development, Bioinformatics Expertise and Bioinformatics environment provisioning. Once data are archived, they are in read-only permission to prevent any unintended deletion. Image datasets also need to be viewable with KeenEye's platform: data are then stored in a secured cloud-based environment (hosted in the EU).

Data transfer

Data transfer between HUGODECA members

Files can be transferred to the Isilon server and to KeenEye by FTP (with an authentication process). In the case of the Isilon server, checksums (sah256) provide a simple way to compute the integrity of data files before and after file operations.

Transfer of personal data to the US (HCA)

According to chapter V of the GDPR, data transfers from the EU to a third country have to be framed according to specific principles. If the EU has made an 'adequacy decision' (adequacy meaning here that an equivalent level of protection for personal data is guaranteed) in relation to the country or territory where the receiver is located or a sector which covers the receiver, the transfer can be made. In case of the USA, the adequacy finding for this country is only for personal data transfers covered by the EU-US Privacy Shield framework.

The Privacy Shield places requirements on US companies certified by the scheme to protect personal data and provides for redress mechanisms for individuals. US Government departments such as the Department of Commerce oversee certification under the scheme. The list of US organisations under the Privacy Shield is available online (https://www.privacyshield.gov/list, last accessed 2022/01/24).

As HCA is not listed on the Privacy Shield list, transfer of personal data (ie sequencing data) requires an explicit consent from the data subject.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethical aspects are addressed in the ethical section of the DoA as well as in the "D9.1 POPD Requirement number 3" document, since the intended use of human organs and tissues warrants serious consideration of ethical issues. As mentioned in that document, the partners will implement their research activities in full respect of the legal and ethical European / national / institutional requirements, and codes of practices. PIs from concerned laboratories (biobank and processing) must have full ethical clearance for the research project and for the submission of generated data to other repositories.

GDPR and data processing operations

In the context of the General Data Protection Regulation (GDPR), personal data is information that relates to an identified or identifiable individual.

- Directly identifying data: surname, forename, address, photo, voice, etc.
- Indirectly identifying data: a telephone number, a spatiotemporally identified activity, a set of unique physical features, ...

Among personal data, several are "sensitive" data, notably health data, and they imply specific data protection. In order to lawfully process special category data, you must identify both a lawful basis under Article 6 and a separate condition for processing special category data under Article 9. These do not have to be linked. Explicit consent from the data subject can legitimize use of special category data

Genetic data as personal data

A genetic sample itself is not personal data until you analyze it to produce some data.**Genetic analysis which includes enough genetic markers to be unique to an individual is personal data and special category genetic data, even if you have removed other names or identifiers**. And any genetic test results which are linked to a specific biological sample are usually personal data, even if the results themselves are not unique to the individual, because the sample is by its nature specific to an individual and provides the link back to their specific genetic identity.

Collection of personal data about the mothers : anonymization and pseudonymization

A data processing operation is any operation involving personal data, whatever the process, the medium used, regardless of whether it is computerized (it is technologically neutral) The data are used to meet objectives/purposes. The processing of data in the sense of "protection of personal data" goes beyond the analysis or exploitation of the data, it also covers the collection, analysis, reuse of data, archiving, etc. (Article 4 of the GDPR)

Situation 1 : if data are anonymized, they do not fall within the scope of the GDPR

Irreversibly anonymised data, whereby a person can no longer be re-identified, are not subject to the laws and regulations on the protection of personal data.

Whatever the technique used, anonymisation must lead to compliance with three criteria:

- Total inability to single out an individual
- Total inability to link records relating to two individuals together
- Impossibility to deduce information about an individual

Situation 2 : if data are pseudonymized, they are subject the GDPR

Pseudonymised data are personal data that can no longer be directly attributed to the data subject. However, the use of additional information, such as a correspondence table, can be used to re-identify the person. In this case, the General Data Protection Regulation (GDPR) applies: if there is a pseudonymization process, it does not mean that the data cannot be used for research purposes, but it entails to provide safeguards to protect the privacy of the people involved in the data collection (Article 5 of the GDPR and Article 89 of the GDPR).

Pseudonymization has two major impacts for the data collectors :

- The consent forms have to clearly mention that personal data are collected
- Each collecting center is required to document the processing of personal data in a registry and maintain up-to-date records of processing operations, keeping track in particular of:
    - The purposes of the processing operation
    - The categories of data subjects and related data
    - The recipients of the data
    - Information on the use of data, their storage and the rights of the data subjects
    - The names and contact details of the controller (ie the person, public authority or body that determines the purpose and means of the processing operation)

Recommendations about consent forms

The table below lists different recommendations about the different types of mentions that should be present in the consent forms. These recommendations are not exhaustive, and in any case, only the validation from the DPO of the research institution will ensure that personal data are processed in fully GDPR-compliant way, and that research data can be published on external repositories to enable secondary use.

Situations	Mentions to check in consent forms

You collect personal data (ie in an identified OR pseudonomized way)	It should be clear that this consent is freely given (the data subject had the choice not to accept to sign the consent form) Consent must be specific : The request for consent to personal data collection shall be presented in a manner which is clearly distinguishable from the other matters (article 6 of the GDPR should be mentioned) Consent must be informed : Informed consent means the data subject knows your identity, what data processing activities you intend to conduct, the purpose of the data processing. There should be precise information on the processing, purpose, use of the data. Storage period must be provided to the data subjects. (Article 12 of the GDPR) Consent can be revoked : Normally, a person should have the right to rectify or erase the data collected about them (right to erasure see Article 17)/ It should be clear who they can address to apply their rights (DPO). The right to erasure, however, is not absolute and only applies in certain circumstances (the personal data is no longer necessary for the purpose which you originally collected or processed it for; you are relying on consent as your lawful basis for holding the data, and the individual withdraws their consent; The request may not be granted if exercising this right is likely to make impossible or seriously impair the achievement of the processing objectives…) In the case of biospecimen collection, guidelines recommend that participants must be allowed to withdraw the remainder of their specimen, but that samples and data that have been distributed can not be recalled
+ You collect data for genetic / genomic analysis (or genetic / genomic analysis is a possible secondary use)	A description of any reasonably foreseeable risks or discomforts to the subject, as well as a description of any reasonably expected benefits to the subject or to others should be mentioned

<table>
<tr><td>+ You want to submit to an open access repository (for instance ENA or HCA)</td><td>Consent forms should contain a mention about unspecified secondary use that could be made of the data by different types of institutions or organizations (including potentially use of data for commercial purposes for instance)</td></tr>
<tr><td>+ You want to be able to submit to a controlled access repository (for instance EGA)</td><td>Consent forms should contain information about potential specific secondary use (research on a particular domain, public institutions only, etc.)</td></tr>
<tr><td>+ You want to submit to a non-European repository (international transfer => HCA)</td><td>The GDPR restricts transfers of personal data outside the EEA, or the protection of the GDPR, unless the rights of the individuals in respect of their personal data is protected in another way, or one of a limited number of exceptions applies<table><tr><td>“In the absence of an adequacy decision pursuant to Article 45(3), or of appropriate safeguards pursuant to Article 46, including binding corporate rules, a transfer or a set of transfers of personal data to a third country or an international organization”, one of the conditions for international data transfer is if the data subject has explicitly consented to the proposed transfer, after having been informed of the possible risks of such transfers for the data subject due to the absence of an adequacy decision and appropriate safeguards;</td></tr><tr><td></td></tr></table></td></tr>
</table>

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Question not answered.

HUGODECA - Final review DMP

1. Data summary

State the purpose of the data collection/generation

The goal of the *HUGODECA* project is to describe the organization of the developing human gonads and to understand the molecular and cellular mechanisms underlying sex determination. In order to do so, the project aims to make major advances in the following complementary areas:

- scRNA- and ATAC-seq of human fetal gonads, with assessment of cell surface markers by flow cytometry (WP1);
- ST and ISS leading to the generation of 2D and 3D expression maps of targeted genes at single-cell resolution (WP2);
- Spatial proteomics based on diverse imaging methods (WP3);
- Software development to enable the integration of large OMICS data with imaging data (WP4);
- Organotypic culture models to evaluate the consequences of signaling pathways directing sex-specific differentiation and endocrine disruptors on gonadal development (WP5).

Explain the relation to the objectives of the project

This project will lead to the generation and process of multiple types of research data, that is data produced as the research is conducted, and which should be as re-usable as possible by the scientific community. Most of their characteristics, such as format and size, vary according to the techniques used to generate them.

For ex vivo experiments, the input and the output of the assay are both biomaterials; for imaging techniques, the input will still be some biomaterials, but the output will be an image. In the case of a single-cell experiment, the data first generated correspond to a textual sequence referring to the genetic code. The frontier between the different types of data generated or processed, however, is not always so clear-cut and one of the stated goal of the HUGODECA project is actually to be able to integrate heterogeneous types of data into one common scaffold. For this reason, it will sometimes be useful to be able to distinguish the primary medium of information from the complementary media, which have been generated to describe the primary one. In the case of ST, for instance, the image is the primary medium, and OMICS expression data will be used as a means of giving complementary information on that medium. Actually, the way these complementary data is stored may also vary: they can be mentioned in a separate file that will be part of a data package, or correspond to embedded metadata, integrated in the primary media.

The goal of the DMP is thus to define the different types of data generated, with their specific characteristics, and to outline how they will be made as FAIR as possible.

Specify the types and formats of data generated/collected

As diverse as the technologies involved may be, the research data share some common process steps and we have classified them according to the following sub-types:

- Raw data, resulting directly from the technology with which they have been acquired (initial image captures or sequencing data, for instance);

- Pre-analysis data: raw data having undergone a transformation process deemed necessary to be able to proceed to a meaningful analysis or generated in order to be able to run the main analysis. As we will mention later on, these types of data are generally not kept in the long-term;
- Analysis data: data associated to quantitative and/or qualitative features;
- Interpretation data: lists of biological elements, which have been identified as being of interest, or computational data resulting from the interpretation of the analysis.

Some of the complementary data which can be found in sets of research data are specially intended to document the way experiments are carried out:

- Study, sample and assay metadata: information about the study, the materials used (from biomaterial to lab equipment) and the conditions of the experiment;
- Methodological data: the protocol in itself, with the different steps of the procedure, and/or descriptions of data workflows;

When considering data from the point of view of the process, another specific type of data comes into the picture: coding data, from scripts written to establish a process pipeline to the complete source code of a software solution specially developed for the project.

Finally, the HUGODECA project will also lead to the production of structured documents, such as publications and communication supports, with the obvious goal of sharing and promoting the results achieved by the HUGODECA Consortium among the scientific community.

In the following subsection, we have listed the different types of expected data by task, to get a more detailed view of their specific characteristics (as the production of publications and communication supports is not task-dependent, they are not mentioned in this list but we will discuss the question of their findability and accessibility in the next sections). All the data mentioned below are not intended to be published or kept in the long-term: the intent, here, is more to have a broad view of the different data generated or processed in the short-term, so as to be able to take them into account when defining naming conventions or characteristics such as the data volume. It is to be noted that this last point is a sensitive issue for the project as the technologies involved imply to store several TB of data (even for a single task) which make the processes of data storage, transfer, back-up and preservation more difficult to address than for datasets of smaller sizes.

LIST OF DATASETS

Reference	HUGODECA_DS_11
Task	1.1 Single-cell RNA-sequencing of human fetal gonads
Task leader	Antoine Rolland (antoine.rolland@univ-rennes1.fr)
Data contact person	Paul RIVAUD (paul.rivaud@inserm.fr) Antoine ROLLAND (antoine.rolland@univ-rennes1.fr)
Description	This dataset corresponds to the data generated via single-cell RNA-sequencing (Chromium technology, 10X Genomics) and subsequently analyzed to provide a transcriptional characterization of human fetal gonad development, at a single cell resolution.

Types and formats	■ Raw sequencing data in *fastq.gz* format (de-multiplexed data, if needed) ■ Pre-analysis data alignment data compressed in binary format (BAM), with a BAM index (.bai) ■ Analysis data : o expression files in *mtx* format with associated tsv files (genes.tsv, barcodes.tsv) as generated by CellRanger pipeline, o sparse matrix data file in *HDF5* format (.h5 or *.loom*) or R format (.rds) containing the combined expression matrix, associated metadata (a metadata file can also be stored separately in .txt, .csv or .tsv , one row per feature/annotation) and cell coordinates (a coordinate file can also be stored separately : it will contain t-SNE or UMAP coordinates with at minimum three columns, which are “cell names” - as specified in the metadata file - “x”, “y”) ■ Interpretation data: list of marker genes that can be used to characterize the cells at distinct developmental stages (.txt) ■ Study, sample and assay metadata (database or csv format) ■ Methodological data (.doc, .pdf and/or .html) ■ Coding data in bash, python or R
Origin	Gonads and mesonephroi from male and female embryos will be collected during the first and second semesters of development. Samples will be collected and processed at the Irset institute (Rennes, France) for partner P01c-INSERM and at the Wellcome Sanger Institute (Cambridge, United Kingdom) for partner P05-GRL. Sequencing data will be generated by a subcontractor (to be identified) for the Irset Institute and on-site for the Wellcome Sanger Institute. Data processing will be performed at both the Irset institute and the Wellcome Sanger Institute. Single-cell RNA-sequencing data of 30 human fetal gonads (from female and male embryos aged 6 to 12 post-conception weeks) previously generated by partner P01c-INSERM (SINERGIA grant from the Swiss National Science Foundation) will be re-used for this task. Briefly, these will be combined with expression data for single gonadal cells from second trimester embryos as well as with expression data for single mesonephric cells from first and second trimester embryos.
Expected size	For a given sample (4-5k cells) raw sequencing data represents 20 to 30GB, while the corresponding expression data is about 100-200MB (mtx format) and 600-800MB (tsv format). The total number of targeted cells being 400-450k cells, the overall size of the data will range between 20 and 30TB.

Utility	The analysis of resulting data will allow to i) identify all cell types present in human fetal gonads, ii) characterize how distinct cell lineages differentiate during gonad development, and iii) provide new candidate markers genes for spatial characterization of gonadogenesis in Humans (WP2 and WP3).

Reference	HUGODECA_DS_12
Task	1.2 Single-cell ATAC-sequencing of human fetal gonads
Task leader	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Data contact person	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Description	This dataset corresponds to the data generated via single-cell ATAC-sequencing and subsequently analyzed to assess genome-wide chromatin accessibility of XX and XY human gonadal cells across developmental stages (4 to 20 PCW, ~250,000 cells, P01c-Inserm, P05-GRL). Differential accessibility analyses will be conducted on pseudo-bulk ATAC-seq profiles to highlight open chromatin peaks in each group of cells. These potential active regulatory regions will be used to identify enriched DNA motifs and corresponding transcription factors (TFs) for each cell type. The dynamics of chromatin accessibility will then be mapped across the RNA-defined trajectories of the different gonadal cell lineages.
Types and formats	■ Raw sequencing data in *fastq.gz* format ■ Pre-analysis data : alignment data compressed in binary format (BAM), with a BAM index (.bai) ■ Analysis data : o bed file of all called peak locations (peaks.bed) o raw and filtered peak barcode matrix in HDF5 , MEX format or dense CSV format o loupe cell browser files (.cloupe) for visualization o peak annotations in tsv format o motif finding results, in text format (as a position weight matrix PWM, for instance) ■ Interpretation data : o TF-gene interaction network in .tsv format o List of marker regulators (.txt) ■ Study, assay and sample metadata (database or csv format) ■ Methodological data (.doc, .pdf and/or .html) ■ Coding data in bash, python or R

Origin	The samples used for scATAC-seq will be collected and processed at the Irset institute (Rennes, France) for partner P01c-INSERM and at the Wellcome Sanger Institute (Cambridge, United Kingdom) for partner P05-GRL. Sequencing data will be generated by a subcontractor (to be identified) for the Irset Institute and on-site for the Wellcome Sanger Institute. Data processing will be performed at both the Irset institute and the Wellcome Sanger Institute. scRNA-seq dataset (WP 1.1) will be re-used to perform computational pairing with scATAC-seq datasets and transfer RNA-based cell labels (i.e. cell type identification) onto chromatin accessibility data.
Expected size	10-50 TB
Utility	By integrating known TF-gene interactions for the identified cis-regulatory elements, we will reconstruct gene-regulatory networks and predict "master" regulators driving cell fate transition as the human gonad differentiates into a testis or an ovary.

Reference	HUGODECA_DS_13
Task	1.3 Flow cytometry based surface protein marker screen of human fetal gonads
Task leader	Andreas Bosio (andreasbo@miltenyibiotec.de)
Data contact person	Werner Muller (wernerm@miltenyi.com) Paurush Praveen (paurushp@miltenyi.com)
Description	The data collected for this task will be acquired by flow cytometry : they will result from the processing of freshly dissociated cells with MACSQuant® Instruments and MACS® Marker Screen, which will allow to identify cell surface protein markers and assess their level of expression.
Types and formats	■ raw data acquired via MACSQuant® Instruments will be stored in the .mqd file format. These mqd files can be exported as FCS (data file standard for the reading and writing of data from flow cytometry experiments - 2.0, 3.0, 3.1 compatible) or CSV files. ■ Interpretation data : o List of main cell surface protein markers (.txt) ■ Study, sample and assay metadata (database or csv format)
Origin	The data processed will come from freshly dissociated cells of 2 to 4 gonads of each sex, at 3 relevant developmental stages (as determined from preliminary data and from Task1.1) Pre-titrated antibodies are used to screen for markers in a standardized and high throughput manner (371 human markers).
Expected size	~500 MB
Utility	Identified markers or combinations of markers will serve to better characterize human gonad cells, to develop target cell isolation strategies for cells of interest (antibody-based cell surface protein profiling will determine marker combinations for sorting of live cells) and as starting library for the analysis of gonad development using 2D/3D microscopy in WP3.

Reference	HUGODECA_DS_21
Task	2.1 - Generation of 2D ST maps of human gonads development
Task leader	Joakim Lundeberg (joakim.lundeberg@scilifelab.se)
Data contact person	Ludvig Larsson (ludvig.larsson@scilifelab.se)

Description	The data generated for task 2.1 correspond to transcriptome-wide 2D maps of the anatomical landscape of human developing gonads by using spatially resolved mRNA expression analysis technologies developed and managed at SciLifeLab Sweden (http://www.scilifelab.se/). The generation of high quality cDNA libraries with precise positional information for RNA-seq is obtained by placing histological sections on glass slides with arrayed oligonucleotides containing positional barcodes.
Types and formats	■ Raw data: full resolution images (24-bit color TIFF, 16-bit grayscale TIFF, or JPEG ?) of Hematoxylin and Eosin (H&E) stained developmental gonad tissue sections and raw sequencing data stored in compressed FASTQ format (fastq.gz). ■ Pre-analysis data: o QC images in jpg format o Alignment data compressed in binary format (BAM), with a BAM index (.bai) o Lower resolution images (downscaled version of full resolution H&E images) stored in PNG format (.png) o Tissue position lists stored as comma separated values (.csv) with capture areas coordinates stored in each row for each spot. o Scalefactors for conversion of capture area coordinates between full resolution to lower resolution images stored in JSON format (.json). ■ Analysis data: o Count matrices in mtx and HDF5 format (.h5) with genes in rows and spots in columns. o R format objects (.rds) including normalized expression data, dimensionality reduction results and meta data. o Loupe files (.cloupe) for interactive visualization of data. ■ Interpretation data o Table of differentially expressed marker genes formatted as tab separated values (.tsv). o Table of enriched pathways formatted as tab separated values (.tsv). ■ Study, sample and assay metadata (database or csv format)

Origin	Fresh frozen embedded embryonic abdominal tissue (containing the gonad) will be collected at the Karolinska University Hospital, Stockholm, Danderyd Hospital, Danderyd, Sweden and HDBR/Sanger Institute. At least two tissue samples of both genders will be collected at 6 PCW and 8/9 PCW. The tissues will be cryo-sectioned for Spatial Transcriptomics analysis. Optimal Spatial Transcriptomics protocol conditions for gonad tissue will be established, using an in-house developed quantitative assay (the incorporation of fluorescently labeled nucleotides will be measured to quantify the cDNA synthesis on the array). Optimal conditions will next be used on spatially barcoded arrays to generate transcriptome-wide 2D maps of the anatomical landscape of human developing gonads.
Expected size	10-20 TB
Utility	The 2D ST maps will enable to characterize the organization of different cell types in gonadal tissues: this analysis will serve to make a first spatial assembly of the cell types defined by scRNAseq (WP1). It can also find spatially interesting gene expression patterns that has not come out of the scRNAseq data. These genes will be more specifically targeted for WP 2.2 and ST maps are also part of the materials needed as a pre-requisite for WP 4.1

Reference	HUGODECA_DS_23
Task	2.3 - Optimization of ST and ISS protocols on cleared tissue samples for the generation of 3D OMICS maps of the same specimen
Task leader	Mats Nilsson (mats.nilsson@scilifelab.se)
Data contact person	Marco Grillo (marco.grillo@scilifelab.se)
Description	The goal of WP 2.3 is to assess the feasibility of ST assay and ISS on sections cut from optically cleared samples after light-sheet imaging. This will tell whether the RNA molecules are intact in cleared samples to provide additional markers potentially useful for spatial analysis. ISS method will be adapted to allow processing of thicker sections to create 3D data and enable spatial maps of cells in 3D. Finally, we will test the robustness, standardize and implement the developed ISS assays for optically cleared sections.
Types and formats	■ Raw data: raw images of cleared tissue samples acquired via light-sheet imaging in TIFF format ■ Pre-analysis data : stitched aligned images in TIFF format ■ Analysis data : o CSV files for x-y-z coordinates map. o 3D omics maps in (mat file ?) ■ Study, sample and assay metadata (database or spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	The specimen used will be the same as the ones used for WP 2.1 and WP 2.2. Images of cleared tissue samples will be provided by P01a-Inserm.
Expected size	Expected to be in TB, but to be re-evaluated
Utility	This task will allow to define new protocols for generating ST expression data and spatial ISS expression data in optically cleared tissues. The images produced will be re-used in WP4 to generate 3D models of transcriptionally defined cells of entire developing gonads.

Reference	HUGODECA_DS_31
Task	3.1 - Imaging mass cytometry of developing gonads
Task leader	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Data contact person	Muzlifah Haniffa (m.a.haniffa@ncl.ac.uk)
Description	The data generated for this task will be acquired by Imaging Mass Cytometry, which empowers simultaneous imaging of up to 37 protein markers at a time : this technique will be used to identify gonadal cell types and correlate the findings from WP1 with the spatial location of those cells in situ, thus providing information on the proximity of cells within the gonad.
Types and formats	■ Raw data: images acquired with Fluidigm Hyperion Imaging platform can be saved in MCD TIFF or OME-TIFF format (File headers then contain an OME-XML metadata block compatible with Open Microscopy Environment) ■ Analysis data o Measurement data in CSV files or as records of a database o 2D maps in MATLAB format ■ Study, sample and assay metadata (database or spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	The samples used will be collected at the Wellcome Sanger Institute (Cambridge, United Kingdom). There will be n=3 samples from 4-6 PCW, n=6 from 6-10 PCW and n=6 from 10-20 PCW male and female gonads. P05-GRL recently established protocols (including for custom antibody metal conjugations) for 15-plex mapping of protein expression in formalin-fixed paraffin embedded human fetal liver tissue sections using the Fluidigm Hyperion Imaging Mass Cytometre (Nature, in revision).
Expected size	Around 50 TB during first year
Utility	Data resulting from this task will be evaluated with those found with WP 3.2 and 3.3 to assess the distribution of distinct cell types in intact gonads as they interact during sex determination and differentiation

Reference	HUGODECA_DS_32
Task	3.2 - Single cell cyclic immunofluorescence protein expression analysis of human fetal gonads
Task leader	Andreas Bosio (andreasbo@miltenyibiotec.de)
Data contact person	Werner Muller (wernerm@miltenyi.com) Paurush Praveen (paurushp@miltenyi.com)
Description	The data will be generated by using Miltenyi Biotec's MACsima technology, which enables the simultaneous analysis of multiple markers on a single sample based on fluorescence microscopy. It uses the principle of iterative staining with different fluorochrome-conjugated antibodies to acquire microscopy data for a multitude of parameters. The iterative process comprises three automated main steps: fluorescent staining, image acquisition, and erasure of the fluorescence signal.
Types and formats	■ Analysis data o 2D image stacks in TIFF formats ■ Study, sample and assay metadata (database or spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	Protein expression and cellular composition will be assessed with a set of more than 100 antibodies on frozen sections of gonads at three developmental stages (in triplicate). The samples used will be collected by INSERM partners (P01a-INSERM, P01b-INSERM).
Expected size	50 TB during first year
Utility	Data resulting from this task will be evaluated with those found with WP 3.1 and 3.3 to assess the distribution of distinct cell types in intact gonads as they interact during sex determination and differentiation

Reference	HUGODECA_DS_33
Task	3.3 - A 3D cellular atlas of the developing human
Task leader	Alain Chédotal (alain.chedotal@inserm.fr)
Data contact person	Yorick Gitton (yorick.gitton@inserm.fr)
Description	The data generated for this task will be acquired with Light-Sheet Fluorescence Microscopy (LSFM) from solvent-cleared human gonads and genital ducts.
Types and formats	■ Analysis data o 3D image stacks in TIFF formats ■ Study, sample and assay metadata (database or spreadsheet) ■ Methodological data (.doc, .pdf or .html)
Origin	Entire organs and human embryos between 5 and 12 PCW (at least 20 cases of both sexes) will be collected by P01-bInserm. Strong candidates as makers of gonadal determination have already been identified in preliminary experiments by P01a-Inserm and P01b-Inserm and genes identified in WP1 and WP2 will also be taken into account.
Expected size	Initial assessments on the new ultramicroscope have produced 2.5TB from 22 scans over four weeks. Expecting to reach **15TB/month** during 1st year.
Utility	This dataset will enable to have a better understanding of gonadal cell lineage relationships in entire organs and human embryos between 5 and 12 PCW and to determine the organization of vasculature and innervation during sex determination. The technology and the protocol that will be defined during this task will also be applied to study gonad samples in WP5.

Reference	HUGODECA_DS_40
Tasks	4.1 - Simultaneous visualization of H&E and immunofluorescent stained tissue and OMICS data 4.2 - Fully 3D rendering/viewer 4.3 - Use existing data for the development of novel 3D integration tools 4.4 - Integration of 2D Cyclic immunofluorescent data in 3D images
WP leader	Sylvain Berlemont (sylvain.berlemont@keeneye.tech)
Data contact person	Sylvain Berlemont (sylvain.berlemont@keeneye.tech)
Description	The data generated for the different tasks of WP 4 correspond to the source code of a 2D/3D viewer developed by Keen Eye Technologies. This viewer will be developed to visualize data from WP 1, 2 and 3 in a combined way: LSFM data, 2D cyclic immunofluorescence, ST, *ISS* as well as single-cell approaches will all be incorporated into one scaffold. Computational models will also be defined and tested to check the feasibility of integrating datasets from different individuals at a similar development stage.
Types and formats	■ Coding data in Python and C++
Origin	The data used to test the viewer will come from WP 1, WP 2 and WP 3. Existing datasets from spatial platforms will also be used in an initial step for the development of new bioinformatics and 3D integration tools. This model will be built upon a training data set using another organ system (embryonic heart) on which transcriptomics data (ST and ISS) and protein stains on cleared embryos (https://transparent-human-embryo.com/) already exist.
Expected size	several MB
Utility	The software solution will allow to deliver OMICS spatial 2D and 3D maps of gonad development. This technology will also be used for WP 5, to better study the effects of alterations on human gonad development.

Reference	HUGODECA_DS_51
Task	5.1 - Manipulation of the intrinsic *WNT/β-catenin* signaling pathway
Task leader	Anne Jorgensen (Anne.Joergensen.02@regionh.dk)
Data contact person	Anne Jorgensen (Anne.Joergensen.02@regionh.dk)

Description	The data generated for this task are diverse since both ex-vivo techniques and imaging technologies will be used to understand the effects of the manipulation of the WNT/β-catenin signaling pathway in fetal testes and ovaries (with a focus on the specification of Sertoli and granulosa cells): inhibition of WNT signaling will be done using IWR-1 in fetal ovaries, while its stimulation in fetal testes will be achieved with CHIR 99021 in combination with recombinant WNT4. Subsequent analyses of organotypic cultures will include 3D-cellular maps, combining whole-mount immunohistochemistry (IHC) and 3D imaging of solvent-cleared organs (3DISCO) with light-sheet fluorescence microscopy (LSFM) (P01a-Inserm, P01b-Inserm), in addition to classic IHC investigation of cell lineage markers and mass spectrometry measurement of sex hormones (P04-RegionH).
Types and formats	For ex vivo techniques: Analysis data o Mass spectrometry measurements of sex hormones (.csv) o Expression data for granulosa and Sertoli cell markers (TIFF) Interpretation data: o List of altered cell lineage markers (.doc, .pdf or .html) o Hormone levels (.xls) Study, sample and assay metadata (database or spreadsheet) Methodological data (.doc, .pdf or .html) For 3D cellular maps, the types and formats of the generated data will be the same as those mentioned for the task 3.3.
Origin	Manipulations will be done for 2 weeks in organotypic culture of fetal gonads aged 6- 9 PCW (P04-RegionH) and compared with vehicle-treated control samples.
Expected size	5-10 TB for 3D imaging
Utility	Abnormally developing gonads, either due to genetic or environmental alterations, cannot be obtained from in vivo specimens. Using organotypic ex vivo models will allow to recapitulate both normal and dysregulated human fetal gonad development. This dataset will allow to better understand the role in the case of the intrinsic WNT/þ-catenin signaling pathway during sex determination in Humans. This study will also be a demonstrative use case of the implementation of 3D cellular maps.

Reference	HUGODECA_DS_52
Task	5.2 - Manipulation of the intrinsic DHH signaling pathway
Task leader	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Data contact person	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Description	The data generated for this task will result from the application of multiple techniques: ex-vivo techniques, 3D-imaging technologies and scRNAseq. The DHH signaling pathway will be manipulated by using organotypic cultures of 7 PCW testes. The manipulation will be done with known receptor agonists (e.g. SAG) and inhibitors (e.g. cyclopamine) of the different steps of this pathway (P01c-Inserm). The different Leydig cell populations will be stained in cultivated testes and further processed for volume-imaging and 3D-automated cell quantification (P01a-Inserm, P01b-Inserm), hormone levels will be measured in the media (P01c-Inserm, P04-RegionH), and scRNA-seq will be performed to understand how cell lineage differentiation is affected (P01c-Inserm).
Types and formats	For ex vivo techniques : ■ Analysis data o Mass spectrometry measurements of sex hormones (.csv) ■ Interpretation data o Hormone levels (.csv) ■ Study, sample and assay metadata (database or spreadsheet) ■ Methodological data (.doc, .pdf or .html) For scRNA-seq, the types and formats of the generated data will be the same as those mentioned for the task 1.1. For imaging data, the types and formats will be the same as those mentioned for task 2.3.
Origin	Manipulations will be done in organotypic cultures of 7 PCW testes (P01c-Inserm). The bioinformatics analysis (scRNA-seq) will include a comparison with uncultivated testes of the same age.
Expected size	5-10 TB for 3D imaging 2-5 TB for scRNA-seq
Utility	The data in this study will enable to understand the involvement of the DHH signaling in the complex patterning of the Leydig cell lineage in human fetal testes. This study will also be a demonstrative use case of the implementation of 3D cellular maps in combination with scRNA-seq data.

Reference	HUGODECA_DS_53
Task	5.3 - Manipulation of Nrp1 signaling pathway
Task leader	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Data contact person	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Description	The data generated for this task will mainly result from the application of 3D imaging technologies on 7 PCW testes in organotypic cultures, which will have been treated with a validated NRP1 blocking function antibody.
Types and formats	Same as those mentioned for tasks 1.1 and 3.3
Origin	Organotypic cultures of 7 PCW testes (P01c-Inserm)
Expected size	5-10 TB for 3D imaging
Utility	The data generated for this task will enable to understand the involvement of the NRP1 signaling in the establishment of a sexually dimorphic vasculature during human gonadogenesis. This dataset will also demonstrate the utility of 3D-imaging technologies as a means of observing changes in vascularization after intrinsic alterations on human gonad development.

Reference	HUGODECA_DS_54
Task	5.4 - Ibuprofen as a prototype of extrinsic factor known to alter the differentiation of both the human ovary and testis
Task leader	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Data contact person	Séverine Mazaud-Guittot (severine.mazaud@univ-rennes1.fr)
Description	This data generated for this task will result from the application of scRNA-seq following organotypic cultures of fetal testes (8-10 PCW) and ovaries (10-12 PCW) with or without Ibuprofen.
Types and formats	Same as those mentioned for task 1.1
Origin	Manipulations will be done in organotypic cultures of 8-10 PCW testes and 10-12 PCW ovaries (P01c-Inserm). The bioinformatics analysis (scRNA-seq) will include a comparison with uncultivated testes and ovaries of the same age.
Expected size	2-5 TB
Utility	The data in this study will allow to identify the cellular targets of Ibuprofen and the corresponding gene pathways altered. This study will also be a demonstrative use case of the utility of reference scRNA-seq dataset (uncultivated testes and ovaries) for the understanding of toxicological effects.

Specify if existing data is being re-used (if any)

For Task 1.1, single-cell RNA-sequencing data of 30 human fetal gonads (from female and male embryos aged 6 to 12 post-conception weeks) previously generated by partner P01c-INSERM (SINERGIA grant from the Swiss National Science Foundation) will be re-used

Specify the origin of the data

See the "origin" field in the list of datasets mentionned above

State the expected size of the data (if known)

See the "size" field in the list of datasets mentionned above

Outline the data utility: to whom will it be useful

See the "utility" field in the list of datasets mentionned above

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

Study and assay metadata will be created to document the different experiments defined in WP 1, WP 2, WP 3 and WP 5 (this is not applicable to WP 4, since it deals with software development).
During the first stage of the project, we will gather information about the program and the different studies that are planned by using SEEK (D7.1) (https://studies.hgdc.genouest.org/) This application follows the Investigation, Study and Assay (ISA) specifications, and we will create one investigation by work package, and one study by project task. Metadata templates (conformant with the HCA metadata model) have been created to define the different sample types used in the assays (https://studies.hgdc.genouest.org/sample_types) and the protocols applied. Complementary information about the metadata collected is given in the part about data interoperability.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

To become findable (and citable), datasets should get a unique, persistent identifier. Persistent identifiers like Digital Object Identifiers (DOI) have been designed as a solution to avoid link rot (that is, when a hyperlink stops referring to the original source because it was moved). These persistent identifiers thus ensure that the data is and will be findable. The use of SEEK will make it possible to generate a DOI for each study. For single cell datasets, all raw and processed data deposited on public repositories can also be assigned a DOI. An issue that will be addressed in a subsequent version of the DMP is that of the assignment of a DOI to the image stacks.

If they do not already have one, the HUGODECA consortium members will also be encouraged to get a persistent author identifier via ORCID, so as to increase the connections towards the research work.

Outline naming conventions used

Naming files, folders and records consistently will facilitate their retrieval, and will enable users to browse the data more effectively and efficiently. For this reason, some basic naming convention should be agreed upon by the consortium members.

Preliminary remarks

Special characters (& * % $ £] { ! @) and spaces should be avoided and that the full stop should preferably be reserved for indicating the separation between the file name and the file extension.

Codes and index used in sample and file names

Project code

HUGODECA will be used when the project code is used for folders and communication documents

HGDC will be used when the project code is used in sample names or generated file names.

WP/Task index

If a WP or task index needs to be mentioned in a file or dataset name, only the numbers are used (the dot is not included). If the implementation level is a whole WP, the index will be the WP number followed by 0

Examples:

index for Task 1.1 is 11

index for WP 1 is 10

Site Code

Partner	site Code	Country
P01b-Inserm	LI	France
P01c-Inserm	RE	France
P04-RegionH	RI	Denmark
P05-GRL	WS	UK

Codes to identify protocol types

Code	Protocol types
COLLEC	Collection protocol
DIFFER	Differentiation protocol
DISSOC	Dissociation protocol
IMGPREP	Imaging preparation protocol
IMAGING	Imaging protocol
SEQUEN	Sequencing protocol
LIBPREP	Library preparation protocol
ANALYS	Analysis protocol

Naming convention in the metadata

- For donor organism (embryos), the pattern is: HGDC_%siteCode%%alphanumericIndex%

Examples : HGDC_RE2345

For derived samples (tissues, organs..), the pattern is: %donorOrganismCode%_%derivedInfoCode%
Derived samples always start with the donor organism (embryo or fetus) code, followed by descriptor codes that either refer to the organ (ie GON1), a cell suspension code (ie SC1) or molecules applied for an organotypic culture.

Examples: HGDC_RIh2143_CHIR: code corresponding to a tissue cultured with CHIR99021

Naming convention for sequencing files and gene expression matrix

Fastq files will be named according to Illumina nomenclature, with the the sample number, which is a numeric assignment based on the order that the sample is listed in the sample sheet. *Example:* HGDC_RE2345_RNA1_L001_R1_001.fastq.gz

Naming convention for images and image datasets

With the actual naming convention, the name of images and image datasets starts with the identifier of the sample used as an input to generate the data. Filenames carry exhaustive descriptors corresponding to the different channels, laser sources, and tiles (ie: HDC_RE2341_ac217-12-waist_561-TH_647-pax2_790-periph-0o63X-2um_LR).

For datasets that will be made accessible via the keenEye platform, the goal is to obtain shorter filenames, generated according to the following pattern: FileNumber_Marker_PCWindex (ie: 000905_Pax2_pcw12) since information about other descriptors will be stored as metadata. The images are going to be organized in directories / subdirectories as follow:

- HGDCidentifier_marker_PCWindex
  - CanalA
    - Images

NB : there will be one ‘image stack’ by canal

Outline the approach towards search keyword

Keywords associated to a dataset should correspond to ontology terms, whenever possible. They should at least mention the organ, the developmental stage, the sex of the embryo, the different technologies used, and the main biological entities studied (tissue type, genes…)

Outline the approach for clear versioning

For documents (.doc, .pdf, …), versioning will be maintained by following this pattern : [document name]_[version number] _[status: DRAFT{0,1}]. Example : HUGODECA_DMP_7-4_V1.doc

For data that can be generated multiple times , the creation date may be indicated instead of a version number. The date will then be expressed with this format: YYYY-MM-DD

For software and scripts, versioning will be controlled via the use of dedicated tools (Subversion, Git…)

(As it may be confusing and storage-consuming to have too many similar or related files: a good file versioning practise is also to discard and erase intermediate working files - and keep only raw data, last version of processed data, and the definitive copy of the analysis data)

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Metadata creation will be made according to the HCA metadata requirements

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

As regards OMICS data generated during the research phase, only raw and analysis data will be published (ie sequencing data and expression matrices with their associated metadata). There is no point in giving public access to pre-analysis data as long as raw data and analysis data are available. As for interpretation data, they will be an integrated part of the scientific publications.

It is to be noted that sequencing data corresponds to genetic data (according to Recital 34 of the GDPR, genetic data includes chromosomal, DNA or RNA analysis, or any other type of analysis that enables you to obtain equivalent information) and as such, it is considered special category data (ie personal data requiring extra protection : in order to lawfully process special category data, a lawful basis under Article 6 and Article 9 of the GDPR is required). Open (or controlled) availability of this data will here be dependent on the fact that the individual has given clear consent for the HUGODECA research teams to process this type of data for the research purpose indicated in the consent forms. Mention of secondary use for unspecified research will also be required for open accessibility, and mention of secondary use for a specified type of research will be required for controlled accessibility.

2D and 3D images will be made available in a controlled way via the KeenEye platform (authentication required). The size of 3D datasets ranges from tens to hundreds of GB, and thus required specific softwares and services to be processed (to be displayed online or to be transferred for instance). For 3D datasets, all requests for access will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. For ST data (2D), high resolution images used for the analysis will be made openly available on the HUGODECA website.

Protocols and pipelines used or defined during the project will be made openly available, whenever it is possible: during the preliminary exploitation plan of the HUGODECA project, the protocols for generating spatial ISS expression data in optically cleared tissues (part of HUGODECA_DS_2-3) have been identified as a potential candidate for patent-filling. This protocol may thus remain confidential (to be confirmed by CARTANA).

Code / software: The data generated for the different tasks of WP 4 correspond to the source code of a 2D/3D viewer developed by Keen Eye Technologies, and will remain confidential: a patent will be filled in order to secure the commercial exploitation of this innovative solution on the market (Information relating to the patent that has been registered must be submitted under the 'IPR' section of the EU Participant Portal)

Scientific publications will be either gold or green open access (free of charge, online access for any user):

- Gold open access means that the publication is available by the scientific publisher as open access. Some journals require an author-processing fee for publishing open access.
- Green open access or self-archiving means that the published article or the final peer-reviewed manuscript is archived by the researcher itself in an online repository, in most cases after its

publication in the journal. The journal must grant the researcher the permission to self-archive the final peer-reviewed article, at the latest,12 months after publication.

For the HUGODECA project, gold open access will be the preferred way of publishing the results.

Specify how the data will be made available

Gene count matrices intend to be listed and made viewable on the HUGODECA web portal (D6.4), with a link to download the datasets from external repositories when applicable. As already mentioned, the consortium members intend to take part in the HCA initiative, so analysis data may also become available on the HCA data portal, if all legal and ethical requirements are met.

In this section, we list the different repositories that have been identified as potential places where the data could be submitted to guarantee their accessibility in the long-term (actual submission of OMICS data to these repositories is dependent on the validation of the DPO of the research team).

For raw sequencing data, the repositories of reference are the European Nucleotide Archive[1] (ENA) (open access) and the European Genome-phenome Archive (EGA) (controlled access). Both are ELIXIR Core Data Resource: they are part of a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data. This means that they have been evaluated according to : their scientific focus and quality of science, the community served by the resource, the quality of service, the legal and funding infrastructure and governance, the impact and translational stories. Both are also managed by EMBL: since the introduction of GDPR in May 2018, EMBL has established an internal policy on General Data Protection (IP68).

EGA

This repository enables to publish personally identifiable genetic and phenotypic data resulting from biomedical research projects in a controlled way. The EGA is co-managed by EMBL-EBI and CRG (Centre for Genomic Regulation). The CRG operates within the EU and fully complies with the GDPR. With regard to the data process agreement, EGA is a data processor and the data controller is the person who submits the data to EGA (implementation of the GDPR, and the data processing agreement are explained in more detailed in the document listed below). In its role as a data processor, EGA requires all submitters to sign a Data Processing Agreement (DPA) when the submission account is first created. Furthermore, the EGA implements several key points in terms of security strategy (regular risk assessment, identity and authorization management, audit logs, cryptography and communication security) to prevent any unauthorized access to the data and to guaranty the confidentiality, integrity and availability of the data.

Information or service provided	Link to page or service (last accessed 2022/01/24)
EGA Helpdesk (contact)	helpdesk@ega-archive.org
EGA submission guidelines and FAQs	https://ega-archive.org/submission
EGA submitter portal	https://ega-archive.org/submission/tools/submitter-portal
EGA security document	https://ega-archive.org/files/European_Genome_phenome_Archive_Security_Overview.pdf
EGA and implementation of the GDPR	https://ega-archive.org/files/EGA_GDPR.pdf
EGA data processing agreement for the submission and distribution of personal data	https://ega-archive.org/files/EGA_Data_Processing_Agreement_v1.1.pdf
EGA data access agreement (to be signed before submission)	https://ega-archive.org/submission/dac/documentation#DAA
Data use conditions in EGA	https://ega-archive.org/data-use-conditions

ENA

The ENA accepts all kinds of nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. It is an open access repository: data submitters must ensure that the datasets are fully anonymised. It is also the data submitters' duty to make sure that they have obtained broad consent from the donors for future research use of the data and that the donors have been informed of the benefits and risks of sharing data in an open database (notably, the risk that their genetic data could be matched to other genetic data held in other databases). The ENA is developed and maintained at the EMBL-EBI under the guidance of the INSDC (International Advisory Committee and a Scientific Advisory Board). The ENA is an ELIXIR Core Data Resource and is maintained by the EMBL-EBI, which means it is subject to the measures for data protection described by that intergovernmental institution, which notably states that "data entrusted to EMBL will be subject to adequate technical and organisational security measures" and which also mentions that EMBL "deems its updated framework on data protection to be 'adequate' in the sense of GDPR."

Information or service provided	Link to page or service (last accessed 2022/01/24)
ENA Helpdesk	https://www.ebi.ac.uk/ena/browser/support
ENA submission guidelines and FAQs	https://ena-docs.readthedocs.io/en/latest/
ENA submission tools	https://www.ebi.ac.uk/ena/browser/submit
ENA policy	https://www.ebi.ac.uk/ena/browser/about/policies

Single cell studies can also be submitted on ArrayExpress ArrayExpress accepts all functional genomics data generated from microarray or next-generation sequencing (NGS) platforms: expression matrices and sequencing data can be submitted at the same time (if sequencing data are submitted, they are then brokered to the ENA). ArrayExpress is an ELIXIR Core Data Resource. Just like the ENA, it is an open access repository maintained at the EMBL-EBI.

Information or service provided	Link to page or service (last accessed 2022/01/24)
ArrayExpress single cell submission guide	https://www.ebi.ac.uk/arrayexpress/help/single-cell_submission_guide.html
ArrayExpress submission tools	https://www.ebi.ac.uk/fg/annotare/login/
ArrayExpress FAQ	https://www.ebi.ac.uk/arrayexpress/help/FAQ.html
ArrayExpress Single cell atlas search page	https://www.ebi.ac.uk/gxa/sc/home

HCA (Human Cell Atlas) is an international collaborative consortium that charts the cell types in the healthy body, across time from development to adulthood, and eventually to old age. It is building a reference map of all cells in the human body through the creation of an online database made up of gene expression data (also called transcriptomic data). HCA is an open access repository : the HCA DCP currently does not accept any data that requires controlled access, but datasets with legal/ethical constraints could eventually be included in next phases of development of the HCA DCP. When it comes to coding procedures, the HCA requires that any directly identifying information (donor name, contact information, hospital identification number) be removed. The possibility for the donor to withdraw data means that, as a data contributor, the HUGODECA consortium is responsible to follow-up on requests from tissue donors to have their data withdrawn from the HCA DCP. If a donor contacts a data contributor to have their data withdrawn, it is the responsibility of the data submittor to contact HCA and specify the dataset that must be removed from the HCA database. As HCA is hosted in the US, it is to be noted that data transfer has to conform to the principles mentioned in Chapter V of the GDPR (https://gdpr-info.eu/chapter-5/) and may rely on a derogation to transfer data for important reasons of public interest Article 49 (d) of the GDPR (or explicit consent from the donor [Article 49(a)]
As the HCA DCP is not fully GDPR-compliant, the DCP teams recommend to submit the raw data (sequencing data) to the EGA (if controlled access is required) or to the ENA (if the data may be in open-access). Only summary level data, such as gene count matrices, may be submitted to the HCA DCP, and only as long as the metadata templates associated to the datasets do not contain any information about the mothers.

Information or service provided	Link to page or service (last accessed 2022/01/24)
HCA submission guidelines	https://data.humancellatlas.org/contribute
HCA contact for submission	wrangler-team@data.humancellatlas.org.
HCA entry page to access the data	https://data.humancellatlas.org/explore/
HCA Data coordination platform	https://www.humancellatlas.org/data-coordination-2/
HCA template consent form for developmental atlas	https://drive.google.com/file/d/1O26gT8p3_hBOdAPDLhNz0pOKxo5epGD7/view
HCA metadata types	https://data.humancellatlas.org/metadata

For bioimaging data in general, there is an open repository for images, the Image Data Resource (IDR), but the deposit of HUGODECA image datasets is still being discussed as the volume needed to store the stacks of images is way over the normal expected size.

Information or service provided	Link to service or page (last accessed 2022/01/24)
IDR submission guidelines	https://idr.openmicroscopy.org/about/submission.html
IDR FAQ	https://idr.openmicroscopy.org/about/faq/
IDR search page, to access the data	https://idr.openmicroscopy.org/cell/ https://idr.openmicroscopy.org/tissue/

Software and tools - Genomics and transcriptomics data are in file formats which are widely used within the scientific community, that is fastq.gz files for raw sequencing data, bam files for aligned sequences and count matrices in HDMF5-loom, mtx or csv formats for processed data, making these fully readable and re-usable worldwide from a technical viewpoint.

Raw images can be viewed with open-source softwares like the Fiji/ImageJ tools (Fiji is an open source project hosted in a Git version control repository and comprehensive documentation, ImageJ source code is available online, with its user manual). Open-source softwares like Cell Profiler and histoCAT can also be used to analyse imaging data. However, in order to visualize the fully annotated 2D and 3D

cell maps, Keen Eye 3D viewer will be necessary: Keen Eye Platform™ is a proprietary and patented software-as-a-service platform and the viewer will be accessible through a cloud service, promoting high accessibility across a worldwide research community.

Methodological data

Protocols.io is on web platform for developing and sharing reproducible methods. Every new protocol starts out private: it may then either remain private (5 private protocols maximum per workspace with a free account) or become public to be shared with all the scientific community. Protocols then receive a DOI (they are archived with CLOCKSS to ensure long-term preservation of the knowledge)..As part of the goal to be part of the HCA initiative, HUGODECA consortium members are encouraged to join the Human Cell Atlas Method Development Community (https://www.protocols.io/groups/hca).

Protocols used or defined by the HUGODECA consortium will also be listed on the HUGODECA data portal.

Pipelines and softwares

Partners that are part of an academic institution will make their scripts publically available on a dedicated platform. GitHub is already used by several of the research groups involved in the project for that purpose (e.g.https://github.com/Moldia,https://github.com/SpatialTranscriptomicsResearch/st_pipeline, https://github.com/umr1085-irset).

Information or service provided	Link to service or page (last accessed 2022/01/24)
Github presentation page	https://github.com/about
Github privacy statement	https://docs.github.com/en/github/site-policy/github-privacy-statement
Github and data protection	https://docs.github.com/en/github/understanding-how-github-uses-and-protects-your-data
Github status page	https://www.githubstatus.com/
Github information page on licensing a repository	https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/licensing-a-repository

Scientific publications

The pre-publication of results is encouraged so as to make research results available as soon as possible to the public community, notably by posting pre-prints on bioRxiv.

Information or service provided	Link to page or service (last accessed 2022/01/24)
bioRxiv submission guidelines	https://www.biorxiv.org/submit-a-manuscript
bioRxiv contact for submission	https://submit.biorxiv.org/
bioRxiv search page to access data	https://www.biorxiv.org/search

[1] https://www.ebi.ac.uk/ena

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Study metadata will be made available via the study management platform (SEEK - documentation online)

For genomic data, expression matrices will be made available via the visualization application.

In the case of the 2D/3D viewer that will be developed by KeenEye Technologies, the source code will remain confidential, and a patent will be filled in order to secure the commercial exploitation of this innovative solution on the market (Information relating to the patent that has been registered must be submitted under the 'IPR' section of the EU Participant Portal).

Specify where the data and associated metadata, documentation and code are deposited

Information about the way each dataset is intended to be deposited is given in the "shareable data" entry of the list of datasets in the data summary. We list here the deposition status according to the origin of the samples, as permissions to publish mainly depend on the consent forms that have been used and on the validation of the DPO of each institute.

Origin and use	Information about deposit
Samples collected by P01b/c-Inserm and used by P01a/b/c-Inserm, P02-KTH, P03-SU	No datasets have been submitted to external repositories as for now. Intended deposit: § Sequencing data generated by P01c-Inserm (WP1.1, WP1.2, WP5.2, WP5.4) will be submitted to the EGA (controlled access) with specific data use conditions § Gene count matrices generated by P01c-Inserm and P02-KTH (WP1.1, WP1.2, WP2.1, WP5.2, WP5.4) will be viewable on the HUGODECA data portal § Gene count matrices may be submitted to HCA § 2D/3D high resolution cell maps (P01a-Inserm, P01b-Inserm, P02-KTH, P03-SU) will be deposited on the cloud-based storage server used by KEENEYE
Samples collected by HDBR and used by P05-GRL	§ Sequencing data generated by P05-GRL (WP1.1 and WP1.2) have been submitted to ArrayExpress/ENA and are in private status for now (to become public upon the validation by DPO / Ethics group). § Raw gene count matrices have been submitted to ArrayExpress, and intend to be submitted to HCA. § Sequencing data generated by P05-GRL (WP2.4) intend to be submitted to ArrayExpress/ENA.
Samples collected and used by P04-RegionH	3D images (P01a-Inserm, P01b-Inserm) will be deposited on the cloud-based storage server used by KEENEYE

Specify how access will be provided in case there are any restrictions

- The sequencing data generated by P01c-Inserm require the following data use conditions: “population origins or ancestry research prohibited”, “ethics approval required”, “health/medical/biomedical research and clinical care” (These conditions are to be confirmed by the DPO)
- Access to the KeenEye platform will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. Commercial destinations will be prohibited

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

Types of metadata required

For samples and processes, a special attention will be given to the metadata used within the HCA initiative.

Regarding samples, there should be metadata about:

- The organism donor (embryo) : unique anonymous identifier (as transmitted by the provider), biological sex (female, male, mixed or unknown), developmental stage (in PCW), species (“Homo sapiens”, NCBItaxon:9606), known diseases (it should default to "no disease", which is equivalent to "MONDO_0000001");
- The sample in itself: unique identifier and organ from which the sample was collected.

Regarding sample processing and data acquisition, the metadata depend on the technology used. When single-cell technologies are involved, metadata should include information about:

- The dissociation procedure
- The library preparation protocol :
  - Library construction method (e.g. “10x v3 sequencing”), library construction kit (name, manufacturer and catalogue number)
  - The input nucleic acid molecule (e.g. “polyA RNA")
  - The nucleic acid source (e.g. “single cell”)
  - End bias (e.g “3 prime tag", "3 prime end bias", "full length”)
  - Strand ("first", "second", "unstranded" or "not provided")
  - Optional properties from those listed in the metadata for HCA
- The sequencing protocol :
  - Instrument used for sequencing and model (e.g. Illumina HiSeq 4000), the sequencing design (e.g. 2x100 bp) and the number of raw reads
  - 10x specific metadata such as fastq creation method (e.g. “Cellranger mkfastq” or “Illumina bcl2fastq”), fastq creation method version (e.g. “Cellranger 3.1”)

When imaging technologies are involved, metadata should include information about:

- Tissue preparation and the imaging preparation protocol
  - Fixation
  - Image slide thickness
  - Final slicing method (cryosectioning) …
- The imaging protocol :
  - Microscopy (« fluorescence microscopy »)
  - Magnification, Numerical aperture, pixel size …

These metadata can be found in the HCA metadata dictionary. Those which are the most relevant for the project are : Specimen from organism, Imaged Specimen, Imaging preparation protocol, Library Preparation Protocol, Sequencing protocol.

For biological imaging data, the OME Model defines a set of metadata to include, such as XYZ dimensions and pixels type, as well as extensive metadata on, for example, image acquisition, annotation, and regions of interest (ROIs).

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Metadata values

In order for all participants to use the same terms, ontologies and controlled vocabularies are going to be used. Whenever possible these values should correspond to the terms defined in the HCA ontology. If this ontology is incomplete for our needs, the concepts used will preferably come from other existing ontologies, including but not restricted to:

- EFO (Experimental Factor Ontology) which provides information about many experimental variables available in EBI databases, while Eagle-i resource (ERO) is an ontology of research resources such as instruments, protocols, reagents, animal models and biospecimens
- the Biological Imaging Methods Ontology (fBBi) is an ontology dedicated to the terms used in biomedical research for imaging and visualization methods, and the BioAssay Ontology (BAO) describes biological screening assays and their results including high-throughput screening (HTS).
- Embryonic structure is described in UBERON and EHDAA2 provides a structured controlled vocabulary of stage-specific anatomical structures of the developing human and is linked to HSAPDB, which includes carnegie stages.
- Different ontologies describe properties and classes at the cell level : the cell ontology (CL) is a general ontology which applies to cell types in animals; and CPO structures the vocabulary related to morphological and physiological phenotypic characteristics of cells, cell components and cellular processes.

For the different metadata fields which are required, we will thus define the ontology or controlled list from which the values should be part, as described in the table below (this table is intended for illustrative purpose only, it is not an exhaustive account of the different metadata fields).

Metadata name	Type	Description
species	ontology	ontology identifier from NCBItaxon (9606)
disease	ontology	ontology identifier from MONDO (if no disease, current accepted proxy for "no disease" is "MONDO_0000001")
organ	ontology	ontology identifier from Uberon
development stage	ontology	ontology identifier from HsapDv (e.g HsapDv_000007)
sex	controlled list	one of ["male", "female", "mixed", "unknown"]
sample_type	controlled list	one of ["cell line", "organoid", "direct from donor - fresh", "direct from donor - frozen", "cultured primary cells"]
tissues	Ontology	ontology identifier from Uberon
cell type	Ontology	ontology identifier from CL (e.g. CL:1001610)
gene ID	Controlled list	Gene identifier from Ensembl
gene name	Ontology	as referenced in the ontology version of the human gene nomenclature (HGNC)
imaging method	Ontology	ontology identifier from fbbi

New terms will be created only if no adequate concept can be found. They will then be mapped to the HCA ontology.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

At the moment, SC and ST pipelines are available with MIT licence.
A CC BY license (requiring only attribution) will be the preferred option for scientific publications and reports: data users will be free to “reuse the material in any medium or format, and to remix, transform, and build upon the material, even commercially” but they will be reminded that they should cite the dataset and acknowledge the data producers in any publications and presentations that make use of the data.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

Data will be made available upon project completion to protect scientific information and guarantee the rights of consortium members to be the first to present or publish large-scale analyses of their results.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

In the case of bioimaging data (3.3) all requests for access will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. Commercial destinations will be prohibited.

Describe data quality assurance processes

By giving access not only to raw and analysis data, but also to study metadata, methodological data, and scripts used to facilitate the process and analysis of the data, the HUGODECA consortium pave the way for data re-usability. The fact that the different protocols established by the consortium members will be published will enable to check that quality assurance processes are ensured. In the case of coding data, documentation and literary programming is promoted : whenever possible, tutorials in the form of notebooks (jupyter notebooks, R notebook, KnitR, Sweave) will be published to document analysis scripts (potentially with the use of binder), so as to integrate the code with the corresponding narrative and documentation

Specify the length of time for which the data will remain re-usable

External repositories such as ENA, EGA and HCA aimed at being resources for **permanent** secure archiving and sharing of datasets.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

12 person-months have been requested by P01c-INSERM for data management (WP7).

20 000 euros have also been requested to cover the costs related to open access publications (author-publishing fees).

Cloud data storage (100 TB / month over 18 months) for 3D images has been estimated at a cost of 54000 euros

Clearly identify responsibilities for data management in your project

A data manager (P01c-Inserm) has been recruited to manage the data with the help of a data contact person, identified for each task (7 person-months requested globally to this end). They will notably implement tools necessary for data management, provide the relevant metadata during the course of the project, keep track of the dissemination level of the different data sets, and refine the naming convention.

Describe costs and potential value of long term preservation

Cloud storage has to be taken into account when considering the long-term preservation of imaging data, since this contract will have to be renewed. For that reason, partnerships with European institutes (CINES; CC-IN2P3, EU Image Data) that can provide long-term support for data hosting are under negotiations, all as well as the possibility to deposit the images in IDR.

For OMICS data, the deposition of raw and analysis data in public repositories remain the best alternative to ensure their long-term preservation: those are trustworthy data repositories, with a certificate or explicitly adhering to archival standards.

The long-term preservation of HUGODECA data is particularly valuable since they deal with unique specimens, analyzed at crucial developmental stages. As the ethical aspects are particularly sensitive, these datasets will be a major source of information, if future regulations should come to restrict this work as in other major countries. The HUGODECA programme is already of a high value for researchers who can not carry out experiments on this type of specimens due to restrictive regulations in their home country.

The data generated for the HUGODECA project have also a high commercial and educational potential: the study of the cell atlas may eventually be included into university curriculum (medical, biology, etc.) and converted as a unique learning 3D tool for anatomy/surgery. It is also to be noted that the consortium members have already received many requests from documentary makers (TV, etc.) for anatomy content and human fetus discovery.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Storage & transfer of OMICS data

Initial storage of the data will be performed locally. Institutional ICT facilities will ensure a secure environment, with firewall system in place, virus intruder protection, and access to digital files controlled with encryption and/or password protection.

Two types of data require specific attention since they are sensitive data:

- The personal data collected about the mothers
- The genetic material generated from the embryonic/fetal tissues

Personal data collected about the mothers

For the HUGODECA project, embryonic/fetal tissues are initially handled by P01b-Inserm, P01c-Inserm, P04-RegionH and P05-GRL.

P04-RegionH and P05-GRL do not store any information about the mothers: they receive tissues respectively from the Biobank at Dept. Growth and Reproduction (Rigshospitalet), and the Human and the MRC-Wellcome Trust Human Developmental Biology Resource, in a fully anonymized way, with the guarantee that the rights of the mothers are also safeguarded by those institutes.

P01b-Inserm and P01c-Inserm are the only partners who collect personal data about the mothers, in a pseudonymized way, when receiving the embryonic/fetal tissues. P01b-Inserm collects the age, height and weight of the mother, and P01c-Inserm collects the age of the mother, the medication she received, and her tobacco exposure. In order to ensure the confidentiality of the data, identity and authorization management is implemented: only declared members of P01b-Inserm and P01c-Inserm have the rights to access the pseudonymized data. The data are stored on a server hosted by the Genouest platform with appropriate security measures (firewall protection, data encryption). When derived samples are transferred to other members of the HUGODECA consortium, the only information that may be communicated about the mother is her age, which makes it impossible to link the samples back to a specific individual.

Genetic material generated from embryonic/fetal material

According to Article 4 (13), “genetic data” means “*personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question*”; and according to Recital 34 of the GDPR, genetic data includes chromosomal, DNA or RNA analysis, or any other type of analysis enabling to obtain equivalent information.

Sequencing data generated in the course of the HUGODECA project do correspond to genetic data. However, 1) it corresponds to the fetus/embryo’s genetic sequence, not that of the mother, and 2) a very small portion of the initial genetic information is captured (only 3’ RNA sequencing is carried out for single cell and spatial transcriptomics analyses, not a whole transcript method). The genetic data thus contain only partial genetic sequences of the mothers’, and it would in fact be extremely difficult to link it back to the genetic identity of the mother. However, as special category data require extra protection and safeguards, those risks have been taken into account by the research teams who generate/store fastq data (P01c-Inserm, P02-KTH, P05-GRL).

Sequencing data generated by P01c-Inserm are stored on a server hosted in France by the Genouest Platform (https://www.genouest.org/) which has been certified (ISO9001:2015) for “Software development, Bioinformatics Expertise and Bioinformatics environment provisioning”. P02-KTH and P05-GRL store the data on servers administered by their respective institutes. Only declared members have the rights to access the data on the servers used by the research teams, with secure remote login and file transfer (SSH protocol, SSL encryption) and firewall protection.

NB: For scRNA-seq and Spatial Transcriptomics data, the status of gene count matrices has also been discussed by the ethics group of the HUGODECA project. In those matrices, each column corresponds to a cell (or a spot in case of ST), each row corresponds to a gene and each entry in the matrix represents the number of reads in a cell (or spot) originating from the corresponding gene. As those data are summary level data and as it would not be possible to link the information back to a specific individual, these gene count matrices are not considered personal data

Long term preservation

Submitting raw sequencing data to external repositories is a way to ensure their long-term preservation. At the end of the project, sequencing data generated by P01c-Inserm intend to be submitted to the European Genome-Phenome Archive (EGA), a controlled access repository, with data use conditions specifying that population origins or ancestry research is prohibited. Sequencing data generated from sample collected by HDBR (P05-GRL) intend to be submitted to ArrayExpress/ENA, an open-access registry, as the donors have given consent for their genetic data to be shared in an open-access database for other researchers to use in future studies.

Data transfer to external repositories

The ENA/ArrayExpress and the EGA are both hosted in the UK and managed by the EMBL. On 28 June 2021 the EU Commission adopted decisions on the UK’s adequacy under the EU’s General Data Protection Regulation (EU GDPR) and Law Enforcement Directive (LED). In both cases, the European Commission has found the UK to be adequate. This means that most data can continue to flow from the EU without the need for additional safeguards.

As a further reassurance, EMBL has a mandate to conduct world-class basic research and to enable international co-operation, as laid down in its founding act of 1973, ratified by 20 of the 28 member states of the European Union; and the mandate of the European Union under Article 179(2) of the Treaty on the Functioning of the European Union to encourage research centers and universities in their research activities of high quality and to support their free cross-border cooperation as important reasons of public interest. What is more, the EGA contributes and helps to define guidelines, best practices, and standards for building and operating an infrastructure that promotes responsible data sharing in accordance with the Global Alliance for Genomics and Health (GA4GH) Privacy and Security Policy (cf. https://ega-archive.org/files/European_Genome_phenome_Archive_Security_Overview.pdf, last accessed 2022/01/24)

The HCA is hosted in the US. In its judgment of 16 July 2020 (Case C-311/18), the Court of Justice of the European Union invalidated the adequacy decision based on the EU-US Privacy Shield. In the absence of an adequacy decision pursuant to Article 45(3), or of appropriate safeguards pursuant to Article 46, a transfer or a set of transfers of personal data to a third country or an international organization may only take place on some special derogations, such as an explicit consent from the data subject [Article 49 (a)] or the derogation mentioned in Article 49 (d) (“transfer is necessary for important reasons of public interest”). As it is, P01-Inserm and P05-GRL will not transfer any sequencing data to HCA. Only gene count matrices (summary level data, that cannot be linked back to a specific individual) may be submitted to this repository.

Storage & transfer of image datasets

3D image datasets are generated by P01a-Inserm and stored locally. They are transferred (via GridFTP with an authentication process) to a Dell EMC Isilon storage system (approx. 540 TB) to be archived. This server is hosted by the Genouest platform which has been ISO9001:2015 certified for Software development, Bioinformatics Expertise and Bioinformatics environment provisioning. Once data are archived, they are in read-only permission to prevent any unintended deletion. For images, the important point is also to keep track and preserve all biomaterials used to produce the image captures, since it is another way to make sure that these images can be reproduced.

2D cell maps (H&E images with gene expression data) and 3D images that intend to become viewable on the Keen Eye platform are stored in a secured cloud-based environment (Simple Storage Service from Amazon Web Services). Under the Shared Responsibility Model, AWS is responsible for providing secure infrastructure and services, while Keen Eye is responsible for architecting and securing their applications and solutions deployed in the AWS cloud.

- AWS implements technical and physical controls and processes designed to prevent unauthorized access or disclosure of customer data. In accordance with the AWS GDPR DPA, customers can select the AWS Region in which they store their customer data. In the case of Keen Eye, the AWS regions correspond to eu-west-3 (Europe Paris) and eu-west-1 (Europe Ireland). The data centers in these two regions are compliant with different certifications and attestations, such as ISO 27001 about security management controls, ISO 27017 about cloud specific controls and ISO/IEC 27018:2019 which is a code of practice that focuses on protection of personal data.
- Keen Eye makes use of AWS Identity and Access Management (IAM) to control access to resources and to manage permissions. The KeenEye plateform, in itself, is only accessible via an authentication process.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethical aspects are addressed in the ethical section of the DoA as well as in the “D9.1 POPD Requirement number 3” document, since the intended use of human organs and tissues warrants serious consideration of ethical issues. Embryonic and fetal tissues used for the HUGODECA project is donated voluntarily by women who have had a termination of pregnancy from hospitals collaborating either directly with the research teams or with the biobank that subsequently provide the biological material to the research teams. In all cases, women seeking elective, medical terminations of normally processed pregnancies are informed by nurses and medical doctors on the possibility to give their termination product to research during their first appointment at the hospital. In addition to explanations given by the medical staff, women who are initially interested receive written information material summarizing the research projects included in the protocol, which the women will be able to take home with them for reflection. Patients are in no way persuaded, induced or coerced into participating in this study: refusal to give consent has no impact on the treatment the woman receives and the giving of consent does not lead to any changes in treatment

As mentioned in D9.1 POPD Requirement number 3, the partners will implement their research activities in full respect of the legal and ethical European / national / institutional requirements, and codes of practices. PIs from concerned laboratories (biobank and processing) must have full ethical clearance for the research project and for the submission of generated data to other repositories. The consortium partners, as data controllers, take responsibility for the opinions provided by their DPOs, and also confirm that DPO opinions cover all data processing activities taking place under the consortium partners’ control.

GDPR and data processing operations

In the context of the General Data Protection Regulation (GDPR), personal data is information that relates to an identified or identifiable individual.

- Directly identifying data: surname, forename, address, photo, voice, etc.
- Indirectly identifying data: a telephone number, a spatiotemporally identified activity, a set of unique physical features, ...

Among personal data, several are "sensitive" data, notably health data, and they imply specific data protection. In order to lawfully process special category data, you must identify both a lawful basis under Article 6 and a separate condition for processing special category data under Article 9. These do not have to be linked. Explicit consent from the data subject can legitimize use of special category data

Genetic data as personal data

A genetic sample itself is not personal data until you analyze it to produce some data.**Genetic analysis which includes enough genetic markers to be unique to an individual is personal data and special category genetic data, even if you have removed other names or identifiers**. And any genetic test results which are linked to a specific biological sample are usually personal data, even if the results themselves are not unique to the individual, because the sample is by its nature specific to an individual and provides the link back to their specific genetic identity.

As mentioned in the previous sections:

- sequencing data are sensitive data, and they fall under the scope of the GPDR
- aggregated gene count matrices are outside the scope of the GDPR since they are summary level data and can not be used to re-identify a specific individual.

Collection of personal data about the mothers: anonymization and pseudonymization

A data processing operation is any operation involving personal data, whatever the process, the medium used, regardless of whether it is computerized (it is technologically neutral) The data are used to meet objectives/purposes. The processing of data in the sense of "protection of personal data" goes beyond the analysis or exploitation of the data, it also covers the collection, analysis, reuse of data, archiving, etc. (Article 4 of the GDPR)

Situation 1 : if data are anonymized, they do not fall within the scope of the GDPR
Irreversibly anonymized data, whereby a person can no longer be re-identified, are not subject to the laws and regulations on the protection of personal data.

Whatever the technique used, anonymization must lead to compliance with three criteria:

- Total inability to single out an individual
- Total inability to link records relating to two individuals together

- Impossibility to deduce information about an individual

P04-RegionH and P05-GRL receive fully anonymized samples, respectively from the Biobank at Dept. Growth and Reproduction (Rigshospitalet), and the Human and the MRC-Wellcome Trust Human Developmental Biology Resource, with the guarantee that the rights of the mothers are also safeguarded by those institutes.

P04-RegionH only have information about the city/hospital where the sample was collected and the age of the embryo/fetus.

1. The sample is collected from a consenting mother in the hospital, at this point the sample is given a laboratory sample number which is not listed in the patient record form, against the mother's personal data nor linked to the consent form. Thus, there is no linkage key back to the donor.
2. The sample is sent to Rigshospitalet. The only information that follows the sample is the laboratory sample number which is only used in the laboratory during experiments. The only information included in the sample database are the laboratory sample number, hospital of collection and the age of the fetus.
3. Importantly, the samples are completely anonymized in this collection with no linkage key back to the donor. Also, no genetic data are generated using human fetal gonad tissue collected by RegionH and thus no genetic information from these samples will be deposited in public repositories.

The samples used by P05-GRL come from the MRC-Wellcome Trust Human Developmental Biology Resource (HDBR), which received approval to function as a research Tissue Bank by the national Research Ethics Service (RES). Tissues from the bank can only be distributed to a research project if this project meets specific conditions, one of them being that all samples and any associated clinical information must be non-identifiable to the researcher at the point of release. For the HUGODECA project, P05-GRL does not receive any data associated with the mother, and they have no way of linking the sample or ID to the mother or any other living person. The samples are provided after the following stages:

- Stage one: The sample is collected from a consenting mother in hospital. At that stage, the sample is given a unique identifier, with no linkage key back to the donor.: iIn the information noticesheet, the mothers are are informed that the genetic data obtained from the collected tissues may be shared in a database for other researchers to use in future studies (open access). They are also informed that there is a remote risk that the genetic data could be matched to their genetic data held in other databases: if this database includes their personal details this could identify that they have donated fetal tissue. Participants have been made aware of this and have acknowledged and accepted this in the consent procedure

- Stage two: The sample is sent to the tissue bank (HDBR). The unique identifier attributed at the hospital is replaced with a new unique number, this number is not held against the original identifier and there is no linkage key from the new number back to the hospital unique identifier. The sample is then known by this final number, and the only data attached to it is age, height, weight and medication used for termination (if available).

- Stage three: On arrival at Sanger, the samples are given a third number: this number is listed against the number generated at the tissue bank, should we need to return the samples to the bank. This third number is two steps away from the original sample, and no researcher could trace the sample to the mother with the available information. No metadata about the mother is

received or stored at Sanger: the only data that P05-GRL receives alongside the sample is an anonymized ID created by HDBR and the age of the sample.

Situation 2 : if data are pseudonymized, they are subject the GDPR

Pseudonymized data are personal data that can no longer be directly attributed to the data subject. However, the use of additional information, such as a correspondence table, can be used to re-identify the person. In this case, the General Data Protection Regulation (GDPR) applies: if there is a pseudonymization process, it does not mean that the data cannot be used for research purposes, but it entails to provide safeguards to protect the privacy of the people involved in the data collection (Article 5 of the GDPR and Article 89 of the GDPR).

Pseudonymization has two major impacts for the data collectors:

- The consent forms have to clearly mention that personal data are collected
- The processing of personal data is documented in a registry to keep track in particular of:
    - The purposes of the processing operation
    - The categories of data subjects and related data
    - The recipients of the data
    - Information on the use of data, their storage and the rights of the data subjects
    - The names and contact details of the controller (ie the person, public authority or body that determines the purpose and means of the processing operation)

The samples collected by P01b-Inserm and P01c-Inserm are pseudonymized. P01b-Inserm collects the age, height and weight of the mother, and P01c-Inserm collects the age of the mother, the medication she received, and her tobacco exposure. P01b-Inserm and P01c-Inserm have taken the following measures to make sure that the rights and freedoms of the mothers are protected:

- The data are pseudonymized and subject to de-identification procedures: a first code is attributed to the sample at the hospital, and a second code is then attributed by the research teams when they collect the tissues (there is no table of correspondence between the two codes). As P01b-Inserm and P01c-Inserm are also part of the HuDeCA project, which aims at unifying embryonic research in France and homogenizing fetal tissue collection procedures, a third code is attributed to the samples when they are registered in the HuDeCA database.
- As required by the GDPR, the mothers are informed, in clear and plain language, that personal data will be collected about them when/if they accept to give embryonic/fetal tissues. The information sheet associated to the consent form lets the mothers know about their rights as regards the processing of their data (the right of access, the right to rectification, the right to erasure, the right to restrict processing and the right to object) and who they may contact to exercise those rights. They are also informed that the tissues and data collected may be transferred and used for planned secondary use. However, the data can only be reused upon approval from the HuDeCA consortium, and under specific conditions (ancestry research, for instance, is prohibited). Consequently, datasets generated from the samples collected via this informed consent procedures cannot be submitted to public open access repositories. The information sheet also mentions that the data may be transferred to other organisms or institutions, but only according to an established conventions or contracts that will guarantee the confidentiality of the data.
- In order to ensure the confidentiality of the data, identity and authorization management is implemented (cf. section about data security)
- When derived samples are transferred to other members of the HUGODECA consortium, the only information that may be communicated about the mother is her age, which makes it impossible

to link the samples back to a specific individual.
- Absolutely no information about the mothers is communicated in the metadata sheets/templates required for dataset submission to external repositories. The data provided concern solely the embryo/fetus.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Question not answered.