INSTITUT COCHIN ENTITY DMP - GENOM'IC Core Facility

Franck Letourneur, Thomas Guilbert

INSTITUT COCHIN ENTITY DMP - GENOM'IC

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Science Europe: structured template for research entities" fourni par DMP OPIDoR.

Plan Details

Plan title	INSTITUT COCHIN ENTITY DMP - GENOM'IC
Deliverable	
Version	First version
Plan purpose/scope	The aim of this DMP is to describe as precisely as possible the various research products produced by the services offered by the core facility. This DMP is part of a wider framework, that of making available all the DMPs of the 9 Institut Cochin core facilities (MOUSET'IC, PROTEOM'IC, PIV, PIME, METABOL'IC, CYBIO, IMAG'IC, GENOM'IC and HISTIM) as well as the DMP of the CID database dedicated to the transfer, storage, consultation and sometimes processing of most of the data produced by the Institut Cochin's core facilities. The purpose of this DMP entity is to provide drafting assistance for researchers wishing to draw up a DMP for their research projects and wishing to call on the expertise of our core facilities.
Fields of science and technology (from OECD classification)	1.6 Biological sciences
Language	eng
Creation date	2024-07-18
Last modification date	2024-10-25
Identifier	
Identifier type	DOI
License	**Name** Creative Commons Attribution 4.0 International **URL** http://spdx.org/licenses/CC-BY-4.0.json

Structure Details

Entity's name	Institut Cochin Core Facility - GENOM'IC
Acronym	GENOM'IC
Identifier	https://institutcochin.fr/plateformes/genomic
Identifier type	URL
Description	GENOM'IC is one of the 9 core facilities of the Cochin Institute. It offers its expertise in the field of transcriptome analysis. It brings together cutting-edge skills and technologies in molecular biology and bioinformatics that it makes available to academic and industrial scientific communities. GENOM'IC offers ISO9001 certified services combining the production of data from high-throughput sequencing and their statistical analysis. It specializes its activities in the study of gene expression and regulation (RNA-seq, ChIP-seq, ATAC-seq) on tissue, cellular (single cell RNA-seq) and spatial (spatial transcriptomics) approaches. GENOM'IC supports teams from assistance with experimental design and choice of the adapted technique to each project until scientific publication. GENOM'IC ensures constant technological monitoring allowing it to develop its technical approaches in order to offer them to research teams.
Creation date	1995-01-01

Research outputs :

1. raw and analysed Next Generation Sequencing data (Dataset)

Contributors

Name	Affiliation	Roles
Adoux Lucie - 0009-0004-8737-3877	Institut Cochin - 051sk4035	• Data Collector • Data Collector • Data Collector • Project Member
Guilbert Thomas - 0000-0001-5069-0730	Institut Cochin - 051sk4035	• DMP manager
Hamroune Juliette - 0000-0002-1812-2853	Institut Cochin - 051sk4035	• Data Collector • Data Manager • Data Manager • Data Manager • Project Member
Izac Brigitte - 0009-0004-5083-3681	Institut Cochin - 051sk4035	• Data Collector • Data Collector • Data Collector • Project Member
Jacques Sébastien - 0000-0003-0669-0586	Institut Cochin - 051sk4035	• Data Collector • Data Collector • Data Collector • Project Member
Letourneur Franck - 0000-0001-9465-4774	Institut Cochin - 051sk4035	• Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager • Entity's Manager
Martin Yoann	Institut Cochin - 051sk4035	• Data Collector • Data Collector • Data Collector • Project Member
Saintpierre Benjamin - 0009-0004-9440-5902	Institut Cochin - 051sk4035	• Data Collector • Data Collector • Data Collector • Project Member
Younsi Lilia - linkedin.com/in/lilia-y-484613173	Institut Cochin - 051sk4035	• Data Collector • Data Collector • Data Collector • Project Member

INSTITUT COCHIN ENTITY DMP - GENOM'IC

Data description and collection or re-use of existing data

Research output description

Name	raw and analysed Next Generation Sequencing data
Description	GENOM'IC provide its expertise to research teams in the transcriptomic analysis of their project. This is realized through production and analyses of Next generation sequencing data. For each new project, GENOM'IC define an experimental design to answear to the scientific question asked by the research team : number / quality of samples, type of library preparation and sequencing mode, bio-informatic methodology. GENOM'IC delivers then raw and analyzed data. We distinguish two steps in the new data generation : • Raw data produced by the NGS Illumina sequencers (Nextseq 500, 2000 and Miseq) on biological samples : data relative to each sequencing run (composed of different samples) are first produced under a proprietary format (Bcl). Once obtained, software solutions convert it to generate data for each individual sample in a standardized open format (fastq). • Analyzed data obtained from various processes of the raw data linked to the type of data produced and to the scientific question asked in the project. These processes generate intermediate and final analyzed data and need descriptive data of each sample obtained by the research team. Whatever its nature, informations concerning each data (ex : raw data production date, type of technical treatment, versions of algorithms used for analysis, ...) are directly formalized during the production process (with a project file resuming all intermediary results and tracing all reagents used during the process) and the analysis process (with an analysis report resuming methodology, main results and tracing all versions used for the analysis). .GENOM'IC can also re-use existing data coming from • previous partner project : minimal informations are provided by the research team. • public database from litterature
Type	Dataset
Workpackage	
Keywords	• ADN (Thésaurus INRAE) • ARN (Thésaurus INRAE) • ARN messager (Thésaurus INRAE)
Keywords (free-text)	
Language	eng
Issued Date	
Persistent identifier	The persistent identifier is the unique name of the project, indicating the year of production and a number.
Identifier type	Identifiant local
May contain personal data?	No
May contain sensible data?	No
May take ethical issues into account?	Yes

Will existing data be reused?

Justification	GENOM'IC can also re-use existing data coming from • previous treated project by the platform with the agreement of the data owner research team • public database from litterature in accordance with the license associated with the publication of the data concerned. Data are the property of the supervisory authority on which the research team depends (INSERM, CNRS, University, ...).

How new data will be collected or produced?

Name of the method	Production of sequence files and analysis results
Description	Raw sequencing data is produced by Illumina sequencers owned by GENOM'IC. Raw files are in .bcl binary format. Processed sequence files are in .fastq format, and produced by the Aozan tool, using bcl-convert software. Mapping files are in .bam format, and produced by STAR aligner. Count files are in .tsv format, and produced by RSEM tool. Files from the differential analysis are in various formats (.tsv, .png, .rds), all generated with the R package DESeq2.
Data Nature	Experimental Data
Equipments, technical platforms used	• GENOM'IC : https://cat.opidor.fr/index.php/GENOM'IC
Related references	• Aozan : https://www.outils.genomique.biologie.ens.fr/aozan/index.html • STAR: ultrafast universal RNA-seq aligner : https://github.com/alexdobin/STAR • RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome : https://github.com/deweylab/RSEM • Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 : https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Documentation and data quality

What metadata and documentation (for example way of organising data) will accompagny the data?

Description	During all processes, original sample names are conserved and all associated elements are stored in a local SQL database, to retrieve information whenever needed. We identified two main types of metadata linked to the two main types of data (raw and analyzed). Each project is identified by a code "NGS-year of production - number". All metadata of a project are conserved in the directory file of the project. For raw data, we identified • metadata coming from the scientific project (administrative informations, experimental design, objectives of the study, type and names of samples, organism studied, ...) : they are conserved in the project file filled by the team leader. Samples processed by GENOM'IC arrive with sample names given by the research lab they come from. GENOM'IC transcribes these names as filenames of the FASTQ files previously described. • metadata coming from the production step (quality control of libraries and sequencing, fluorimetry measures, sequencing mode, ...) : they are conserved in the project directory in a folluw-up file and in the systems ("cahier de laboratoire electronique") used to produce these data (instruments for QC control of libraries or RNA, sequencers). For analyzed data, we identified information about reference genome used for the analysis (reference files from the Ensembl database (ensembl.org)), and about the different tools used for the analysis, for which GENOM'IC provides bibliographical references and used versions of the tools, to match the FAIR recommendations. For each project, a report file resuming the different analysis steps and their results is given.
Related references	• Metadata, FAIR principles, and their importance in genomics : https://genestack.com/assets/pdfs/The%20importance%20of%20metadata%20in%20genomics%20and%20the%20FAIR%20principles%20ebook.pdf • FAIR Genomes : https://fairgenomes.org/ • FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research : 10.1038/s41597-022-01265-x
Metadata/data standards	• Genome Metadata : https://rdamsc.bath.ac.uk/api2/m19
Metadata language code	eng

What methods will be used to ensure their scientific quality?

Description	Each instrument used in the pipeline of the data production (quality control of libraries, sequencer) is annualy controled to ensure a perfect data quality. Each protocol used for data production follow processes described in our quality management system (certified Iso9001 V2015) For raw sequencing data, processed FASTQ files generated by Aozan are simultaneously checked for quality using the FastQC and MultiQC modules. Different metrics quality data are avalaible (% of flowcel occupancy, % of Q30) and used as quality indicators in our ISO9001 quality management system. For analyzed data, the quality control is based on the expertise of the bioinformatician in charge of the analysis, using different tools. For example, the unsupervised approach (via PCA Principal Component Analysis) is used to confirm the belonging of each sample to its potential biological group. The alignment merics are also one of the way to control quality data.
Related references	• Aozan: an automated post-sequencing data-processing pipeline : https://www.outils.genomique.biologie.ens.fr/aozan/index.html • A quality control tool for high throughput sequence data. : https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Legal and ethical requirements, codes of conduct

How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Description

The research organism financing the project is the owner of the data and results generated by the facility and of their intellectual property. GenomIC produce and analyze the data and store them (with metadata) for 6 months in its data server.

These conditions are described in the general conditions of use of the facility.

What ethical issues and codes of conduct are there, and how will they be taken into account?

Description

Main ethical issues are related to human studies. These data are sensible data because they include genetic informations usable for individual recognition. To prevent this, samples are pseudonymized. Data are stored in servers under the responsabilty of the DSI INSERM. Higher storage protections are evaluated (double authentification access / isolated server from internet).

Data processing and analysis

How and with what resources will the data be processed / analyzed?

Description

.bcl file coming directly from the sequencers are converted in FASTQ files using BCL converter software integrated in Aozan software.

.fastq files are then processed and analyzed as follows :

- alignment on the reference genome
- counting genes
- differential analysis

The alignment step is done with the STAR aligner (see related references) using genome reference files from the Ensembl database. Output files are in BAM binary format.

The counting of the genes is done with the RSEM tool. Output files are in .tsv format with gene IDs and associated raw and normalized values.

Differential analysis is done with the DESeq2 R package. Various output files are produced : normalized data table in .tsv format, results data table with statistical output also in .tsv format, different graphs in .png format, and an R output file in .rds format, along with an HTML report.

Pipelines are development tools from GenomIC with various coding languages like R or Python. They are regularly updated on the Git of the lab.

Related references

- STAR: ultrafast universal RNA-seq aligner : https://github.com/alexdobin/STAR
- RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome : https://github.com/deweylab/RSEM
- Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 : https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Equipments, technical platforms

- GENOM'IC : https://cat.opidor.fr/index.php/GENOM'IC

Storage and backup during the research process

How will data be stored and backed up during the research?

Storage needs

Raw data coming from sequencers are directly stored on storage servers at the end of their production in a dedicated directory.

Actual data storage is based on two RAID servers (Nettapp serveur with 58To/67 - data from 2015 to 2020 - and a new Dell server with 51To/233 from 2021 to 2024). Raw data storage need is around 10-15 To per year.

Storage requirements are indicated for an average year of data generation.

Estimated volume of data

12

Unit

TB

Equipments, technical platforms

- CID : https://cat.opidor.fr/index.php/CID
- GENOM'IC : https://cat.opidor.fr/index.php/GENOM'IC :

Measures taken for data security

Experimental raw and analyzed data produced by GenomIC are stored on dedicated servers, located in the room servers of the institute Cochin. Data security is actually ensured by the security level of INSERM servers which host them.

Access to the storage servers is reserved to GenomIC via a single authentification (login + pswd).

Data sharing and long-term preservation

How will data be shared?

Modalities of sharing	At the end of the production phase, raw and analyzed data are shared with the team leader: raw data are loaded in the CID (https://imagerie.cochin.inserm.fr/sis4web/login.php) or sent by filesender (https://filesender.renater.fr/), or copied on a hard disk. At this point, the team leader is responsible of the long term storage of its data. At the end of the publication phase, raw data need to be shared with the scientific community via their loading on dedicated databases (GEO, SRA).
Reusability	YES
Data repository/catalogs	• GEO : https://www.ncbi.nlm.nih.gov/geo/ ()

How will data be long-term preserved? Which data?

Justification	Long term preservation of raw data is justified by their possible future use in a different scientific context of the one used to produce them. These data are fastq sequencing files extracted for each sample from the sequencing run. GenomIC hasn't the capacity to store all raw data during many years. So long term storage (unlimited store, depending of the storing tool) is under the responsability of the team leader. Beside storage on RAID servers, raw data are also backed up on external hard disks. Once published, data are stored on dedicated database (GEO) and accessible to the scientific community.
Estimated volume of data	10
Unit	TB
Start date	
End date	
Final dispositions	