TARGET-seq -Parallel evolution in CBF AML

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateur du PGD : Matthieu Duchmann

Affiliation du créateur principal : INSERM

Modèle du PGD : INRAE - General project template

Dernière modification du PGD : 07/11/2020

Financeur : ARC

Numéro de subvention : PGA1-RC20180206836

Résumé du projet :

Using TARGET-seq to explore transcriptome heterogeneity between subclone in CBF AML.

Chercheur Principal : Matthieu Duchmann

Contact pour les Données : Matthieu Duchmann

Produits de recherche :

1. scRNAseq : scRNAseq-training FASTQ files (Jeu de données)
2. scGenotyping : scGenotyping FASTQ files (Jeu de données)
3. scIndexsorting : scIndexsort matrix (Jeu de données)

TARGET-seq -Parallel evolution in CBF AML

Information concerning the management plan

Author of the DMP (if different from the principal investigator/researcher): name, email

Matthieu Duchmann, matthieu.duchmann@inserm.fr

Affiliation of the author of the DMP

Puissant Lab, INSERM U944/UMR7212, Institut Jean Bernard, Hôpital St Louis

Date of creation of DMP

20/06/2020

Current version: (n°, date)

02-01112020

Information on the research project

Identifier of the call for proposal

PGA1-RC20180206836

Project funder(s)

ARC

Name of research programme

PGA-RC

Reference of funding agreement

PGA1-RC20180206836

Project acronym

ICLAC

Name of research project

Interférence Clonale des mutations de signalisation dans les LAM CBF

Project leader institution, coordinator & beneficiary (name, country)

INSERM, France

Other partners (name, country, role of each partner other than the project leader institution)

Question sans réponse.

Unit to which project leader belongs

U944 / UMR7212

Project dates and duration

2019-2023

Brief presentation of project data

scRNAseq : scRNAseq-training FASTQ files

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

Experimental data from TARGET-seq, corresponds to demultiplexed single-cell RNAseq data (.fastq) from sorted hematopoietic stem and progenitor cells from diagnostic or relapse AML samples. Corresponds to a list of fastq file per patients, one fastq file per cell.

scGenotyping : scGenotyping FASTQ files

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

Experimental data from TARGET-seq, corresponds to demultiplexed single-cell genotyping data (.fastq) from sorted hematopoietic stem and progenitor cells from diagnostic or relapse AML samples. Corresponds to a list of fastq file per patients, one fastq file per cell.

scIndexsorting : scIndexsort matrix

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

Experimental data from TARGET-seq, corresponds to single-cell mean fluorescence intensity (MFI) and cell population annotation during index sorting of sorted hematopoietic stem and progenitor cells from diagnostic or relapse AML samples. Correponds to a matrix.

Intellectual property rights

Who owns the rights on data and other information created during the project?

Data belongs to Raphael Itzykson, Puissant Lab, U944/7212 INSERM.

Will material protected by specific rights be used during the project? In this case, who will deal with the formalities required, obtain the authorisations for use and possible dissemination?

All samples are anonymised, and patients have provided informed consent.
The project have been approved by INSERM ethical commitee.

Confidentiality

Identification of the confidential data sets

Fastq files and processed data

What are the measures taken and the norms that must be met to guarantee this confidentiality?

NA

If applicable, how will data confidentiality be guaranteed when the data will be shared or made available for second level analysis?

NA

Access and sharing of data at the end of the project

Is there an obligation to share data (or on the contrary a prohibition or restriction?

Raw data will be deposited on a public repository after publication.

What data will be shared at the end of the project? If all the data are not available in the same way, or at the same time, please specify

Yes, raw data will be deposited on a public repository after publication.

What are the potential reuses for these data?

Exploratory analyses, grouping with private data.

Does reading the data require specific software or tool? If so, which one?

No.

How will the data be shared?

Deposit on a public repository (SRA)

With whom? With what licence?

Open access

As from when?

Starting from publication date.

For how long?

10 years.

Will the data be identified by a permanent identifier (DOI or other)?

Yes, probably according to SRA nomenclature.

Which organisation will be responsible for requesting the identifier in the case of multi-partner projects?

INSERM

Description and organisation of data

scRNAseq : scRNAseq-training FASTQ files

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

Demultiplexed fastq files from Illumina Nextseq sequencing using bcl2fastq.

Fastq files will be pre-processed (filtering and trimming, using Fastp), and then map to reference genome (hg38) using STAR to produce bam files.

Bam files will then be processed with FeatureCounts to quantify gene expression, producing a gene count matrix (.tsv).

Gene count will then be filtered, and normalized to produce a gene count processed matrix (.tsv).

Documentation associated with the data

A large metadata file describing for each single cell :

- cell id
- plate id
- patient id
- date of processing
- library used
- sequencer
- run id

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

A large metadata file describing for each single cell :

- cell id
- plate id
- patient id
- date of processing
- library used
- sequencer
- run id

How will the metadata be produced?

RDA

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files…

Raw and processed data will be processed differentially.

.fastq -> fastq.preprocessed -> aligned.bam -> sorted.indexed.bam -> .counts -> filtered.normalized.counts

What is the quality control procedure of the data?

Enclose the quality insurance plan if possible

Use of integrity tools

scGenotyping : scGenotyping FASTQ files

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

Demultiplexed fastq files from Illumina Nextseq sequencing using bcl2fastq.

Fastq files will be pre-processed (filtering and trimming, using Fastp), and then map to reference genome (hg38) using STAR to produce bam files.

Variant calling will be performed using mpileup from Samtools, and the Varscan to generate vcf with mutational status for each single cell. cDNA and gDNA genotyping will be then summarized in a consensus genotype in a tsv file using vcfR.

Documentation associated with the data

A unique metadata file per TARGET-seq experiment.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

A large metadata file describing for each single cell :

- cell id
- plate id
- patient id
- date of processing
- library used
- sequencer
- run id

Hg38 will be used as the reference genome.

How will the metadata be produced?

Manualy

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files…

Version control by a shared github repository.
Raw data will be separated from processed data.
File name will correspond to plateid_wellid_patientid.fastq and then change in extension.
fastq -> processed.fastq -> .aligned.bam -> .sorted.filtered.cDNA.bam/.sorted.filtered.gDNA.bam -> .mpileup -> .vcf -> .tsv

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Use of integrity tool (ETL)

scIndexsorting : scIndexsort matrix

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

Index sorting will be available from the single cell sorting using the Sony MA900 cell sorter. Each single cell will have the MFI for the 6 labels (lineage, CD34 CD38 CD90 CD45RA CD123) in a large matrix (.tsv)

Documentation associated with the data

A unique metadata file per TARGET-seq experiment.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

A large metadata file describing for each single cell :
- cell id
- plate id
- patient id
- date of processing
- library used
- sequencer
- run id

Hg38 will be used as the reference genome.

How will the metadata be produced?

Use if RDA

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files…

Version control by a shared github repository.
Raw data will be separated from processed data.
File name will correspond to plateid_wellid_patientid.fastq and then change in extension.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Use of integrity tool (type ETL).

Data storage and backup during the project

Storage: what media will be used for data during the project?

Hard drive and local server during analysis

Storage: What types of flows will be used by the data during the project?

Sequencing facility -> Matthieu Duchmann Hard drive / Puissant Lab Server (after sequencing)
Puissant Lab Server -> Puissant Lab Server (finished work)
Puissant Lab Server -> Matthieu Duchmann Hard drive (finished work)
Puissant Lab Server -> Public access repository (publication process)

Storage: What is the estimated amount of data?

Raw data :
- 20GB per samples
- 20GB x 10 patients x 2 timepoints = 400 GB.

Processed data :
- .bam + processed matrix = 400 GB.

A total of 1TB has been allocated to this project.

Storage: Where will the data be stored, on what type of host?

During work : hard drive and local cluster
After work : hard drive + local server
After publication : hard drive and public repository

Storage: Where will the data be located geographically?

INSERM, U944

Security: Does the entity physically hosting the data have a security policy for its information system?

Yes

Security - Confidentiality: will the data de exchanged or shared with third parties?

Not before publication

Security - Confidentiality: how are rights of access to data determined during the research project?

No access during research project

Security - Confidentiality: how will all the project partner researchers have access to data during the project?

No access.

Security – Integrity – Traceability: what measures of protection will be taken to monitor data production and analysis during the project?

Shared project repository locally. Standard procedure to name files.
Outputs data will have released number.
Shared github repository for logs and results produce during secondary analysis

Data archiving and conservation after the end of the project

scRNAseq : scRNAseq-training FASTQ files

What data will be conserved in the medium and long term and what data will be destroyed?

fastq files, raw and processed count matrix will be conserved.
pre-processed fastq, bam files and other intermediate files will be removed.

On what permanent archive platform will the data that are to be conserved long-term be archived? What procedures will be set up for long-term conservation?

SRA
Hard-drive

What is the duration of data conservation?

To be determined

Who will be responsible for long-term conservation?
Name an individual contact

Raphael itzykson

What will be the volume of these data?

10GB

What funding guarantees will cover the costs of long-term conservation?

To be determined.

scGenotyping : scGenotyping FASTQ files

What data will be conserved in the medium and long term and what data will be destroyed?

fastq files, vcf and annotation file will be conserved.
pre-processed fastq, bam files and other intermediate files will be removed.

On what permanent archive platform will the data that are to be conserved long-term be archived? What procedures will be set up for long-term conservation?

SRA
Hard-drive

What is the duration of data conservation?

To be determined

Who will be responsible for long-term conservation?
Name an individual contact

Raphael Itzykson

What will be the volume of these data?

10GB

What funding guarantees will cover the costs of long-term conservation?

To be determined.

scIndexsorting : scIndexsort matrix

What data will be conserved in the medium and long term and what data will be destroyed?

Raw data files will be conserved (Fastq)
Count matrix, Index Sorting matrix and vcf will be conserved.
Intermediate files (.bam, .processed.fastq) will be destroyed.

On what permanent archive platform will the data that are to be conserved long-term be archived?
What procedures will be set up for long-term conservation?

SRA
Hard drive

What is the duration of data conservation?

10 years.

Who will be responsible for long-term conservation?
Name an individual contact

Raphael Itzykson

What will be the volume of these data?

400GB.

What funding guarantees will cover the costs of long-term conservation?

NA