Msc Bioinformatics thesis

Study of Division of Labor in Pseudomonas throught single-cell RNA-seq

Valentin Goupille

Master 2 in Bioinformatics

Academic Year: 2024-2025

Internship conducted at Ecobio UMR 6553 CNRS-University of Rennes

Ecobio UMR 6553 CNRS-University of Rennes

Campus de Beaulieu, 35042 Rennes Cedex, France

Under the supervision of:

Solène Mauger-Franklin, Postdoctoral Researcher

Philippe Vandenkoornhuyse, Professor

Presented on 2025-07-01

Table of contents

Copyright notice

Produced on 17 June 2025.

Declaration

Statement of originality

Université de Rennes

I, the undersigned, **Valentin Goupille**, a student in the **Master's program in Bioinformatics**, hereby declare that I am fully aware that plagiarism of documents or parts of documents published on any type of medium, including the internet, constitutes a violation of copyright laws as well as an act of fraud.

As a result, I commit to citing all the sources I have used in the writing of this document.

Date : **01/04/2025**

Signature :

Reproducibility statement

This thesis is written using Quarto. All materials (including the data sets and source files) required to reproduce this document can be found at the Github repository `github.com/vgoupille/Internship_2025`.

Abstract

Study of Pseudomonas brassicacearum gene expression variation in environ-mental constraints, towards the validation of Division Of Labor.

Proin sodales neque erat, varius cursus diam tincidunt sit amet. Etiam scelerisque fringilla nisl eu venenatis. Donec sem ipsum, scelerisque ac venenatis quis, hendrerit vel mauris. Praesent semper erat sit amet purus condimentum, sit amet auctor mi feugiat. In hac habitasse platea dictumst. Nunc ac mauris in massa feugiat bibendum id in dui. Praesent accumsan urna at lacinia aliquet. Proin ultricies eu est quis pellentesque. In vel lorem at nisl rhoncus cursus eu quis mi. In eu rutrum ante, quis placerat justo. Etiam euismod nibh nibh, sed elementum nunc imperdiet in. Praesent gravida nunc vel odio lacinia, at tempus nisl placerat. Aenean id ipsum sed est sagittis hendrerit non in tortor.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis sagittis posuere ligula sit amet lacinia. Duis dignissim pellentesque magna, rhoncus congue sapien finibus mollis. Ut eu sem laoreet, vehicula ipsum in, convallis erat. Vestibulum magna sem, blandit pulvinar augue sit amet, auctor malesuada sapien. Nullam faucibus leo eget eros hendrerit, non laoreet ipsum lacinia. Curabitur cursus diam elit, non tempus ante volutpat a. Quisque hendrerit blandit purus non fringilla. Integer sit amet elit viverra ante dapibus semper. Vestibulum viverra rutrum enim, at luctus enim posuere eu. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

Keywords :

Single-cell RNA-seq, Pseudomonas brassicacearum, Division Of Labor, (4-5 keywords) bacterial population, metabolism, specialization, root colonization

Acknowledgements

I would like to thank ... Ecobio ANR Divide

In accordance with Chapter 7.1.4 of the research degrees handbook, if you have engaged the services of a professional editor, you must provide their name and a brief description of the service rendered. If the professional editor's current or former area of academic specialisation is similar your own, this too should be stated as it may suggest to examiners that the editor's advice to the student has extended beyond guidance on English expression to affect the substance and structure of the thesis.

If you have used generative artificial intelligence (AI) technologies, you must include a written acknowledgment of the use and its extent. Your acknowledgement should at a minimum specify which technology was used, include explicit description on how the information was generated, and explain how the output was used in your work. Below is a suggested format:

"I acknowledge the use of [insert AI system(s) and link] to [specific use of generative artificial intelligence]. The output from these was used to [explain use]."

Free text section for you to record your acknowledgment and gratitude for the more general academic input and support such as financial support from grants and scholarships and the non-academic support you have received during the course of your enrolment. If you are a recipient of the "Australian Government Research Training Program Scholarship", you are required to include the following statement:

> "This research was supported by an Australian Government Research Training Program (RTP) Scholarship."

You may also wish to acknowledge significant and substantial contribution made by others to the research, work and writing represented and/or reported in the thesis. These could include significant contributions to: the conception and design of the project; non-routine

technical work; analysis and interpretation of research data; drafting significant parts of the work or critically revising it to contribute to the interpretation.

« We are most grateful to the Genomics Core Facility GenoA, member of Biogenouest and France Genomique and to the Bioinformatics Core Facility BiRD, member of Biogenouest and Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013) for the use of their resources and their technical support »

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
ANR	Agence Nationale de la Recherche
DNA	Deoxyribonucleic Acid
DOL	Division Of Labor
NGS	Next Generation Sequencing
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
scRNA-seq	single-cell RNA sequencing

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Litterature review

deddfefde[1,2] Internship description

The survival of organisms in evolving environments is driven by their fitness. The cost-benefit ratio of traits is constantly balanced and gives rise to different populational evolutionary strategies. To succeed, organisms will have to compete, cooperate and/or specialize as a result of how fit their traits are considering their biotic and abiotic environment. Bacteria are unicellular organisms with therefore little option to specialize and give up certain traits production to limit their metabolic costs, unlike multicellular organisms that present many different forms of specialized cells in one single organism. However, [[auxotrophic bacteria]] (i.e bacteria lacking genes coding for a molecule essential for their survival) have been studied (Morris et al., 2012).[3]

Auxotroph bacteria can take advantage of leaky functions of helper's organisms to fulfill their needs in specific compounds (Morris et al., 2014, Estrela et al., 2016).[4,5] With a reduced genetic material, the beneficiary organism fitness is improved, at the risk of being dependent on the helpers presence in their environment. The conditions in which patterns of such [[division of labor (DOL)]] arise are still obscure, but its advantages for bacterial population are clear: DOL allows to diminish the cost associated to certain functions and the possibility of cohabitation of various mutants/specialized cells within the population to respond as a whole to environmental constraints, and thrive. New technologies allow us to access within-species diversity and study the possible metabolic specialization between cells. Single-cell -omics have been developed for this purpose in human health and are now applied to microbial systems. However, analyzing such datasets still requires custom pipelines to respond to the specificity of bacterial biology and technical challenges.

The goal of this internship is to explore [[scRNA-seq]] (single-cell RNA-seq) datasets of [[Pseudomonas brassicacearum]], a root colonizer. The student will analyse samples datasets from various nutritional conditions to determine if DOL can be detected within this species as a strategy for efficient root colonization. The intern will have to implement transcriptomic data analyses from ultra-high throughput sequence run(s). Thus the main aim of the intern will be to set up bioinformatic workflow(s) from existing tools to produce interpretable results.

- differentes methodes de single cell RNA seq (voir diapo et citations)
- voir annexes pour les differentes methodes
- nous focus sur microSPLiT stratégie qui est derivé de la SPLiTseq pour les eucaryotes (=> voir matériels et methodes pour l'explication de la méthode)[6] It's a high-throughput single-cell RNA sequencing method for bacteria. The microSPLiT technology was developed from SPLiT-seq16, a combinatorial split-pool scRNA-seq technology for eukaryotic cells.

-nombreux defi pour les bactéries : faire la liste ici

et pour le moment pas d'outils pour le moment vraiment adaptés ...[7]

- voir mon diapo pour toute l'introduction , les figures ET LA LOGIQUE

=> QUESTIONS BIOLOGIQUES IMPORTANT :

DOL 2 hypotheses : noises / specialized cells

- seurat object
- d'abord filtrer les cellules par Sample_ID (meilleurs signal probablement)
- ensuite recuperer seulement CDS

ouverture : est ce que methode permet de capturer efficacement en condition de stress

Chapter 2

Materials and Methods

2.1 Bacterial culture

-boite petri population isogenique , puis culture en liquid medium

Pseudomonas brassicacearum R401 was grown in liquid medium.... at ...°C.

Two different conditions were applied to the bacteria :

- Low glucose and low iron (M9 medium)
- High glucose and high iron (M9F medium)

For each condition, 3 replicates were grown (biological replicates) and des cellules ont été prelevées de ces cultures. The DO was measured during the growth à 3 DO_timepoints (OD 0.1, 0.2, 0.3) Which do a total of 18 biological Samples/ Conditions (2 medium * 3 biological replicates * 3 timepoints)

Culture medium | Biological replicates | OD_timepoint

fig : plot curve of growth for each condition (ciblé et reel)

mes questions : est ce que replica bio, sont les meme entre stress et non stress ?voir avec Solène comment sont appliqué les stress (des debut ou apres un certain temps)=> parler de ça en discussion => see annex for the media composition and more details

Col1	Col2	Col3

Tableau des conditions biologiques : - figure / tableau et explications des conditions biologiques

2.2 microSPLiT protocol

2.2.1 microSPLiT barcoding

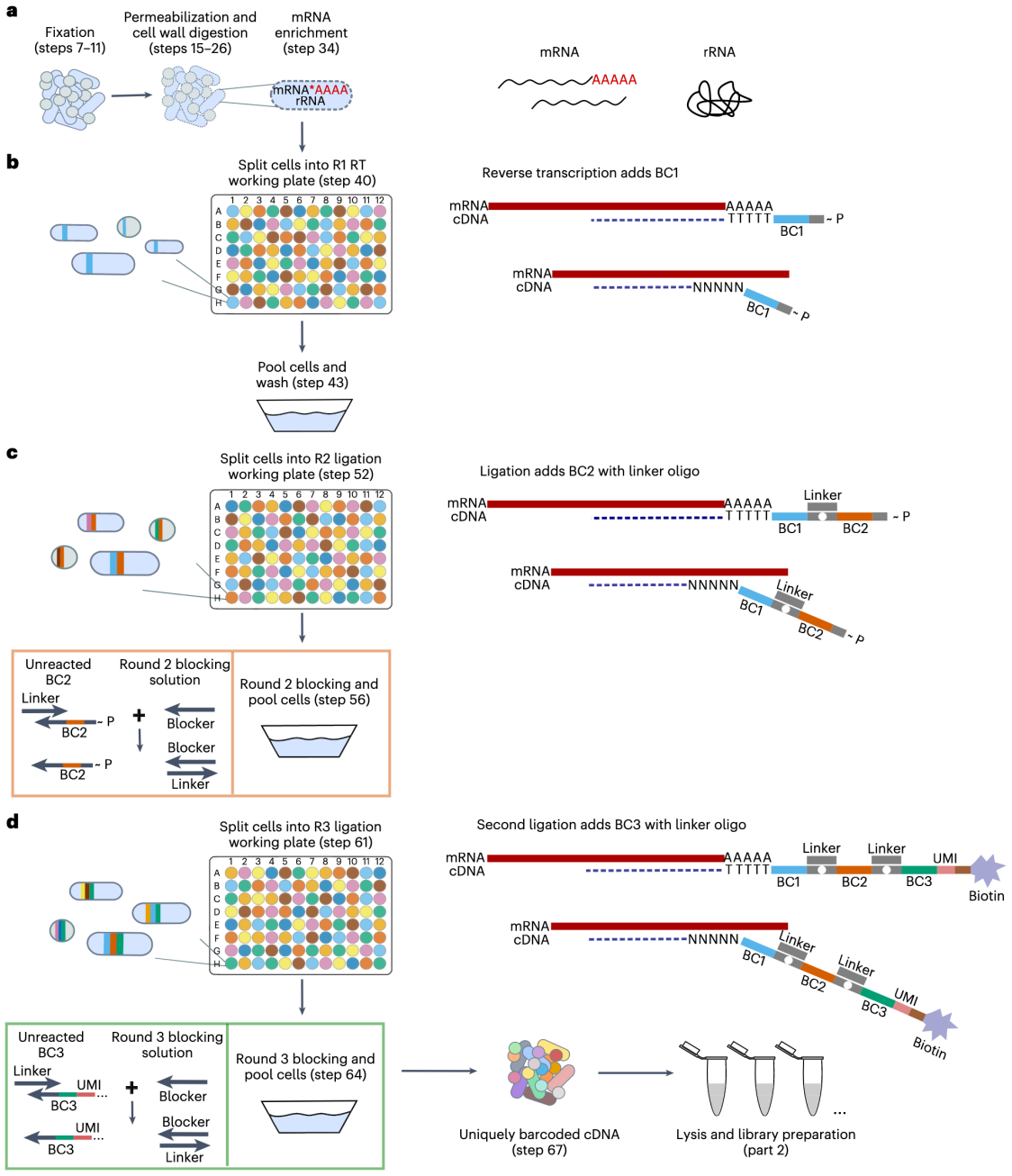

Figure 2.1: *MicroSPLiT Protocol*

MicroSPLiT in-cell cDNA barcoding scheme. a, Bacterial cells are fixed overnight and permeabilized (Part 1, Steps 7–26) before the mRNA is preferentially polyadenylated (Part 1, Step 34). After mRNA enrichment, cells may contain both polyadenylated and non-polyadenylated mRNA. b, Cells are

distributed into the first barcoding plate, and the mRNA is reverse transcribed by using a mixture of poly-dT and random hexamer primers carrying a barcode (barcode 1, BC1) and a 5' phosphate for future ligation at their 5' end (Part 1, Step 41). After the barcoding reaction, cells are pooled together and split again into the second barcoded plate (Part 1, Steps 43–52). c, Ligation adds a 5' phosphorylated barcode 2 (BC2) to BC1 with a linker strand. A blocking solution is then added to each of the wells of the second plate, preventing any unreacted BC2 from future ligation (Part 1, Step 56). Cells are pooled and split into the third and final barcoded plate (Part 1, Steps 58–61). d, A second ligation step adds barcode 3 (BC3) with another linker strand. BC3 also contains a 5' biotin, a primer binding site and a unique molecular identifier (UMI). A blocking solution for the R3 linker is added to each of the wells in the plate before the final pooling of cells (Part 1, Step 64). This results in uniquely barcoded cells that can be distributed in aliquots into sub-libraries and stored until future use or used immediately for library preparation. R1, round 1; R2, round 2; R3, round 3.

- Microbial split-pool ligation transcriptomics (microSPLiT) protocol was performed as described in[2] and[1]. It's a high-throughput single-cell RNA sequencing method for bacteria can profile transcriptional states in hundreds of thousands of bacteria in a single experiment without specialized equipment.

As bacterial samples are fixed and permeabilized before barcoding, they can be collected and stored ahead of time.

Contrary to other single-cell RNA sequencing methods, microSPLiT does not require the isolation of individual cells.

Instead of lysing bacteria and releasing the transcripts from each cell into a barcoding reaction vessel, in microSPLiT, each cell is the vessel enclosing its own transcripts. microSPLiT preserves the intact bacterial cell as a reaction compartment, enabling in situ barcoding of intracellular RNA.

Fixation and permeabilization

Fixation For each sample, Bacterial cells are collected in bulk and fixed using formaldehyde. Fixation has two essential roles: - It preserves the transcriptomic state of each cell at the time of sampling. - It creates covalent cross-links between RNA and intracellular proteins, thereby preventing RNA leakage during later processing. - Importantly, fixation must retain the physical integrity of the cells to ensure downstream single-cell analysis is possible.

Permeabilization Permeabilization is the process of making the cell membrane permeable to allow the entry of molecules. In microSPLiT, the cell wall/cell membrane is permeabilized using a combination of detergents and enzymes. The aim is to permeabilize the cell envelope without disrupting cell structure. This allows external enzymes (e.g., poly(A) polymerase, reverse transcriptase, ligase) and

oligonucleotides to enter the cell. A key balance must be achieved: sufficient permeabilization for enzyme access, but minimal structural damage to maintain single-cell resolution.

In-cell polyadenylation

After permeabilization, the transcripts in the fixed and permeabilized cells undergo in situ polyadenylation with the addition of Escherichia coli poly(A) polymerase (PAP) and ATP. This step enriches for mRNA in the total barcoded RNA pool because, under these conditions, PAP preferentially polyadenylates mRNA as opposed to rRNA.

During the first barcoding round, the fixed and permeabilized bacteria are distributed into a 96-well plate, where their transcripts are reverse transcribed into cDNA and labeled with the first well-specific barcode inside the cells. The cells are mixed and redistributed two more times into new 96-well plates, where the second and third barcodes are appended to the cDNA via in-cell ligation reactions. Finally, the cells are mixed and divided into aliquot sub-libraries, which can be stored until future use or prepared for sequencing with the addition of a fourth barcode. It takes 4 days to generate sequencing-ready libraries, including 1 day for collection and overnight fixation of samples.

microSPLiT barcoding

Instead of lysing bacteria and releasing the transcripts from each cell into a barcoding reaction vessel, in microSPLiT, each cell is the vessel enclosing its own transcripts. The procedure starts with the collection of cells in bulk and the fixation of the bacterial suspension with formaldehyde. It then proceeds with permeabilization by using sequential mild detergent and lysozyme treatments (Fig. 1a). Fixation is critical because it both preserves the cellular transcriptomic state and covalently cross-links the transcripts with the proteins inside the cells to prevent leakage after permeabilization.

The permeabilization step ensures that the externally supplied enzymes and oligonucleotides can access the RNA transcripts in the fixed intracellular milieu. While sufficient permeabilization is crucial to the efficiency of barcoding, it is also critical to preserve the physical integrity of the fixed cells to maintain the single-cell resolution of the method.

We emphasize that for a successful microSPLiT experiment, the cells, after permeabilization, must still exist as intact, individual units to permit several split and pool steps and hold together the cross-linked RNA. After permeabilization, the transcripts in the fixed and permeabilized cells undergo in situ polyadenylation with the addition of Escherichia coli poly(A) polymerase (PAP) and ATP. This step enriches for mRNA in the total barcoded RNA pool because, under these conditions, PAP preferentially polyadenylates mRNA as opposed to rRNA.

round barcoding

on split les 18 samples , en 5 pour avoir des replicats techni donc on obtient 90 samples , permettra d'evaluer la variance technique

depot meme quantité de cellules dans chaque puit

round 1 barcoding les 90 samples sont reparties dans 90 puits distincts, chacun contenant un unique primer barcodé

In the next step, the cell suspension and each sample is distributed into a 96-well plate with uniquely barcoded primers in each well (Fig. 1b, round 1 (R1) reverse transcription (RT) working plate).

The mRNA is then converted to cDNA through in-cell RT with a mixture of barcoded poly(T) and random hexamer primers. cette etape de barcoding permet demarquer les cellules par condition

round 2 barcoding Cells are then pooled, washed and randomly redistributed into a new 96-well plate (round 2 (R2) ligation working plate) containing a second set of well-specific barcodes, which are appended to the first barcode on the cDNA through an in-cell ligation reaction (Fig. 1c). Because of the random cell distribution, there is a high chance that each well of the secondround plate will contain cells with a mixture of different first-round barcodes, creating diverse barcode combinations.

round 3 barcoding Cells are then pooled again, and a split-ligation-pool cycle is repeated for the second time. Cells are randomly distributed into a third 96-well plate (round 3 (R3) ligation working plate), which is loaded with barcoded oligonucleotides containing the third cell barcode annealed with a linker, a 10-base unique molecular identifier (UMI), a common PCR handle and a 5' biotin molecule . en réalité ici 95 a la place de 96 pref separer

909695 = 820800 combinations de barcodes possibles => autant de cellules individuelles possibles

The pooled cells are washed, counted and divided into sub-libraries of variable sizes, which can be stored at -80 °C for >=6 months before proceeding with sequencing library preparation. Dividing sub-libraries into aliquots has two main advantages. First, it allows fine control over the number of cells in the final sequencing libraries. The size of a sub-library can be chosen so that the number of cells that receive the same barcode combination by chance does not exceed the desired collision rate (Table 1). It also permits multiplexing several libraries, potentially even from different experiments, in a single sequencing run.

on choisi de sequencer la plus petite librairie, celle avec 3000 cellules afin d'avoir de maximiser la profondeur de sequençage et limiter le nombre de collision rate (0.34 quand 969696) et avoir un nombre suffisant de cellules pour avoir un signal

90 samples see annex for the plate with barcoded primers

2.2.2 microSPLiT sequencing library preparation

2.3 other remarks

Bacterial culture (Biological conditions) Before to explain the microSPLiT the experimental methods are presented.

see the protocol of - deux conditions env : stress (low_glucose_low fer) / pas stress

- suivi temporel à 3 timepoints (peut etre discuter de ce point après car les mesures de do ont été fait a des temps précis et variation de DO entre les reelles et attendus)
- 3 replicats biologiques par condition
- 5 replicats techniques par replicat biologique

=> un nombre cellules visés de 3000 cellules au totale (voir annexes pour le choix cela)

=> voir annexe pour le plan de plaques

=> donc 90 echantillons differents

=> rep bio : voir si variation entre les populations => replica techn : permettre d'estimer si variance dans le nombre - utilisation du nombre d'UMIs dans le round1 pour estimer cela => voir resultats : pooler ensemble

dans l'hypothèse equilibre parfaite a noté que 33 cellules par conditions ce qui est tres faible => peut etre soumis à des variations individuels au qui pourrait rendre difficule evalutaion variation (tirage aléatoire, possible cellules avec faible activi ou inverse (desequilibre))

=> discussion : renvoie vers l'outils Shiny pour voir les conditions biologiques permet une visualisation plus rapide des donnees

2.4 MicroSPLiT

figure du protocole microsplit

- explication de la méthode microSPLiT[1,8]
- pour faire simple : differentes etape : fixation ; . . . ;
- contrairement aux autres methodes, pas d'isolation individuelle des cellules
- 3 rounds de split-pool : le premier round on affecte les differentes conditions biologiques

- un 4eme round pour rajouté un UMI pour les differentes libraires de sequençage

-envoie d'un pool de librairies à la plateforme de sequençage - sequençage de type NovaSeq™ X Plus par la plateforme GenoBIRD , et demulteplexé permet d'amelioré la qualité de sequençage

- renvoie vers le protocole de kuchina 2021 et bretner 2024[1,8] pour l'explication de la méthode en detail
- => discussion des limites de cette methode dans la partie discussion
- ce qui est importantes de comprendre c'est que cela repose sur un methodes mathematiques combinatoires mais pas methodes d'isolation en tant que tel .
- un grand nombre de cellules ont été barcodés : plusieurs dizaines ou 100aines de milliers mais seulemtnn pres de 3000 cellules ont été choisi pour le sequencçage :
  voir annex pour le tableau de choix du nombre de cellules pour la librairies (est un compromis pour avoir suffissament de cellules potentielles mais pas trop pour ne pas avoir des librairies trop grandes qui pourrait entrainer un profondeur de sequençage pas suffisante) => discussion sur le nombre de cellules choisi pour la librairie

2.5 Librairies structures

-stucture de la librairie

- figure : Final librairies structure
- R1 contient la sequence d'interet
- R2 contient les barcodes
- polyA ou random_hexamer

-> key point : dans chauque puit polyA et random

=> voir annex pour la structure complete avec TSO… => discussion sur TSO

-STARsolo permet l'alignements des reads et lectures des barcodes - details de methode d'alignements partielles ou non … - tailles des reads que j'ai alignés

Chapter 3

Pipeline of the analysis

- figure of the pipeline

Preprocesing sur le cluster genouest, la suite en local et sur le cluster aussi (peut etre) tous les scripts sont dispo sur differents depots githubs

- differentes etapes :
  - demultiplexage des index de librairies (réalisé par la plateforme de sequençage)
  - QC control des données avec Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ et avec[9]
  - trimming des données de seqeunçage avec Fastp[10] et Cutadapt[11]
  - aligment des data sur le genome de reference de *Pseudomonas brassicacearum* grace à STARsolo un amelioration de l'outils STAR pour les données single-cell [[12]][13]
  - different other tools existe comme pour alignement et lecture des barcodes SPLiTseq/ microSPliT comme Kallisto[14],[15] mais d'apres le benchmarking de le plus rapide, reproductible est starsolo[16]

    Although single-cell sequencing approaches have been developed for several molecular modalities, single-cell transcriptome sequencing is the most prevalent and widely applied technique. SPLiT-seq (split-pool ligation-based transcriptome sequencing) is one of these single-cell transcriptome techniques that applies a unique combinatorial-barcoding approach by splitting and pooling cells into multi-well plates containing barcodes. This unique approach required the development of dedicated computational tools to preprocess

the data and extract the count matrices. Here we compare eight bioinformatic pipelines (alevin-fry splitp, LR-splitpipe, SCSit, splitpipe, splitpipeline, SPLiTseq-demultiplex, STARsolo and zUMI) that have been developed to process SPLiT-seq data.

- Metadata assignation to seura

=> d'autres étapes ou autres outils pourrait etre ajouter dans le pipeline (voir la partie disccussion)

- differnetes steps of the pipeline
-
- assignation des metadonnées (utilisation d'un genome de reference => discussion)

3.0.1 Demultiplexing

3.0.2

3.0.3

3.0.4 STARsolo

500 go

attention je vais devoir dire les version des outils utilisés

github,

differentes etapes mais pas detailler tous ici renvoie vers le codes commenter et les readme pour comprendre en details telechangement des fichiers, decompression .zip ; 4 fichier , analyse indenpendant de la qualité des 4 librair, trimming

parallelisation pour ganger du temps ...

Final effective command line of STARsolo: STAR

–runThreadN 64

–genomeDir /path/to/genome_index

–readFilesIn

/path/to/input/merged_trimmed-R1.fastq.gz

/path/to/input/merged_trimmed-R2.fastq.gz

–readFilesCommand gunzip -c

–outFileNamePrefix /path/to/output/starsolo_output/

–outSAMtype BAM Unsorted

–outFilterScoreMinOverLread 0

–outFilterMatchNmin 50

–outFilterMatchNminOverLread 0

–alignSJoverhangMin 1000

–alignSJDBoverhangMin 1000

–soloType CB_UMI_Complex

–soloCBwhitelist

/path/to/barcodes/barcode_round3.txt

/path/to/barcodes/barcode_round2.txt

/path/to/barcodes/barcode_round1.txt

–soloFeatures Gene GeneFull

–soloUMIdedup 1MM_All

–soloCBmatchWLtype 1MM

–soloCBposition 0_10_0_17 0_48_0_55 0_78_0_85

–soloUMIposition 0_0_0_9

–soloMultiMappers Uniform

3.0.5 STARsolo Parameters Explanation

This section details the key parameters used in our STARsolo analysis and their significance:

General STAR Parameters

- `--runThreadN 64` : Use of 64 threads for parallel alignment
- `--genomeDir` : Path to the reference genome index
- `--readFilesIn` : Input FASTQ files (R1 and R2)
- `--readFilesCommand gunzip -c` : Command to decompress FASTQ.gz files
- `--outFileNamePrefix` : Prefix for output files
- `--outSAMtype BAM Unsorted` : Unsorted BAM output format

Filtering Parameters

- `--outFilterScoreMinOverLread 0` : Minimum filtering score relative to read length
- `--outFilterMatchNmin 50` : Minimum number of matching bases for a valid alignment
- `--outFilterMatchNminOverLread 0` : Minimum match ratio relative to read length
- `--alignSJoverhangMin 1000` and `--alignSJDBoverhangMin 1000` : Strict parameters for splice junction detection

STARsolo-specific Parameters

- `--soloType CB_UMI_Complex` : Analysis type for cell barcodes (CB) and complex UMIs
- `--soloCBwhitelist` : List of valid cell barcodes for the three barcoding rounds
- `--soloFeatures Gene GeneFull` : Analysis of features at both gene and full transcript levels
- `--soloUMIdedup 1MM_All` : UMI deduplication with one mutation tolerance

- `--soloCBmatchWLtype 1MM` : Cell barcode matching with one mutation tolerance
- `--soloCBposition` : Cell barcode positions in reads (3 rounds)
  - Round 1: 0_10_0_17
  - Round 2: 0_48_0_55
  - Round 3: 0_78_0_85
- `--soloUMIposition 0_0_0_9` : UMI position in reads
- `--soloMultiMappers Uniform` : Uniform distribution of multi-mapped reads

These parameters were chosen to optimize single-cell detection while maintaining high alignment quality and accounting for the complexity of our three-round barcoding protocol.

Chapter 4

Results

- 4 libraries of unequal sizes (see Solène's explanation for why they were not exactly equivalent)
- This impacts library efficiency
- => Recommendation: balance the libraries for optimal results

4.1 Stats sur les reads R1 et R2 :

- On a subset of 1,000,000 reads:
  - Percentage of reads containing TSO
  - Percentage of reads containing polyA
  - Percentage of reads containing adapter
  - Percentage of reads containing linker -> possibly in appendix

Reminder about saturation calculation method:

4.2 Trimming

ici j'ai suivi les recommendations de kuchina , seulement presenté les resultats Genfull , mettre plus tard les resultats de Gene

- renvoie vers l'annexe pour les multiqc (fastp, cutadapt) avant et apres trimming
-
- mais au final on obtient des resultats tres interessants et propres

=> peut etre il aurait été interressant de faire un trimming comme kuchina 2021 pour comparer les resultats => renvoie vers la discussion pour le trimming fait dans l'article de[1] - avait presque 90% de saturation => a verifier si c'est vraiment le cas - et 10 % des sequences avec TSO (voir Annexe)

- Summary table of Starsolo results
- numbers of reads before and after trimming
- taux de saturation ...
- => renvoie vers la discussion pour les taux de saturation

4.3 STARsolo

-pour starsolo renvoie un gene qui serait mal annoté donc est automatiquement ignoré

commande starsolo suivante comme dans bretn permet UMI a une erreur... possition des barcodes

4.4 Genome

4.5 Transcriptome

Chapter 5

Fitrage des cellules

- reflexion sur le filtrages des cellules est complexe : comme mentionnée dans l'article de il existe des methodes de filtrages plus ou moins complexes
  -filtrage sur le types de reads, globales ou seuils differents entre les differentes conditions biologiques -filtrage sur le nombre de reads par cellule -filtrage sur le nombre de genes exprimés par cellule -filtrage sur le nombre de reads par gene
- filtrage avec un threashold
- dans l'article de kuchina 2021 , ils ont utilisé un seuil de 200 UMI par cellule
-
- preprint de … pourrait etre interessant de se focus aussi sur rRNA (lien avec growth rate)

Chapter 6

housekeeping genes

choix ou non de pooler les replicas techniques ensembles

6.1 Summary of results

6.2 MultiQC Quality Reports

Detailed sequence quality reports are available below. Click on each report to view it.

MultiQC Report R1 Before Trimming

Click to view MultiQC report R1 before trimming

MultiQC Report R1 After Trimming

Click to view MultiQC report R1 after trimming

MultiQC Report R2 Before Trimming

Click to view MultiQC report R2 before trimming

MultiQC Report R2 After Trimming

Click to view MultiQC report R2 after trimming

Key quality metrics are summarized in the tables below.

6.2.1 Barcode Statistics Interpretation

The analysis of cell barcodes reveals several important points about our data quality:

- **Barcode Quality**:

Table 6.1: *Before trimming*

Sample Name	Dups	GC	Median len	Seqs
BC_0076_R1	94.5%	55.0%	241bp	631.4M
BC_0077_R1	94.1%	53.0%	241bp	325.5M
BC_0079_R1	93.8%	53.0%	241bp	379.1M
BC_0080_R1	94.6%	54.0%	241bp	397.7M
Total	-	-	-	**1,733.7M**

Table 6.3: *After trimming*

Sample Name	Dups	GC	Median len	Seqs
BC_0076_R1	98.7%	51.0%	127bp	450.8M
BC_0077_R1	98.6%	51.0%	157bp	248.4M
BC_0079_R1	98.6%	51.0%	152bp	285.2M
BC_0080_R1	98.7%	51.0%	132bp	300.1M
Total	-	-	-	**1,284.5M**

Table 6.5: *Summary of sequence metrics before and after trimming, including percentage changes*

Sample Name	Before Trimming		After Trimming		Change in	
	Median len	Seqs	Median len	Seqs	Length	Sequences
BC_0076_R1	241bp	631.4M	127bp	450.8M	-47.3%	-28.6%
BC_0077_R1	241bp	325.5M	157bp	248.4M	-34.9%	-23.7%
BC_0079_R1	241bp	379.1M	152bp	285.2M	-36.9%	-24.8%
BC_0080_R1	241bp	397.7M	132bp	300.1M	-45.2%	-24.5%
Mean	**241bp**	**433.4M**	**142bp**	**321.1M**	-	-
Total	-	**1733.7M**	-	**1284.5M**	**-41.1%**	**-25.4%**

Table 6.7: *STARsolo barcode statistics*

Metric	Count
nNoAdapter	0
nNoUMI	0
nNoCB	67,177
nNinCB	0
nNinUMI	16,893,550
nUMIhomopolymer	1,697,129
nTooMany	0
nNoMatch	163,166,830
nMismatchesInMultCB	3,323,192
nExactMatch	1,046,121,284
nMismatchOneWL	53,206,471
nMismatchToMultWL	0

  - The absence of reads without adapter (`nNoAdapter = 0`) and without UMI (`nNoUMI = 0`) indicates excellent library preparation quality
  - The relatively low number of reads without cell barcode (`nNoCB = 67,177`) represents less than 0.01% of total reads, which is excellent

- **Barcode Accuracy**:
  - The majority of reads (1,046,121,284) have a perfectly aligned barcode (`nExactMatch`)
  - Approximately 53 million reads show a single mismatch (`nMismatchOneWL`)
  - The absence of reads with multiple matches (`nMismatchToMultWL = 0`) suggests good barcode specificity

- **UMI Quality**:
  - The number of invalid UMIs (`nNinUMI = 16,893,550`) represents a relatively small proportion of total reads and might be primarily due to sequencing errors at the beginning of reads, which is a common observation in Illumina sequencing
  - The presence of homopolymers in UMIs (`nUMIhomopolymer = 1,697,129`) is a known phenomenon that can affect molecular counting accuracy, but the relatively low number suggests this is not a major concern

- **Overall Matching**:
  - The significant number of unmatched reads (`nNoMatch = 163,166,830`) suggests that a substantial portion of reads do not match expected barcodes
  - This could be due to sequencing errors or potential contamination

These results indicate overall good library preparation quality, with excellent cell barcode specificity, although some improvements could be made regarding UMI quality.

6.3 Genfull

6.3.1 Detailed Interpretation of STARsolo Statistics

Based on the official STAR documentation and explanations from Alex Dobin (STAR developer) [?], here is a detailed interpretation of our STARsolo statistics:

Barcode Statistics (Barcodes.stats)

Statistics with the “no” prefix indicate reads that are not used for quantification:

- **Barcode Quality** :
  - `nNoAdapter` : Reads without adapter
  - `nNoUMI` : Reads without valid UMI

Table 6.9: *STARsolo feature mapping statistics*

Metric	Count
nUnmapped	114,628,761
nNoFeature	13,153,735
nAmbigFeature	936,951,032
nAmbigFeatureMultimap	935,077,136
nTooMany	0
nNoExactMatch	125,216
nExactMatch	4,471,174,098
nMatch	971,519,404
nMatchUnique	34,593,349
nCellBarcodes	699,355
nUMIs	34,258,961

  - `nNoCB` : Reads without valid cell barcode
  - `nNinCB` : Reads with 'N' bases in cell barcode
  - `nNinUMI` : Reads with 'N' bases in UMI
  - `nUMIhomopolymer` : Reads with homopolymeric UMI

Mapping Statistics (Features.stats)

These statistics refer to the number of reads, except for `nCellBarcodes` and `nUMIs` which represent the number of valid cell barcodes and UMIs respectively.

- **General Mapping** :
  - `nUnmapped` : Reads not mapped to the genome
  - `nNoFeature` : Reads not mapped to an annotated feature
  - `nAmbigFeature` : Reads mapped to multiple features
  - `nAmbigFeatureMultimap` : Subset of `nAmbigFeature` where reads are mapped to multiple genomic loci
- **Mapping Quality** :
  - `nExactMatch` : Reads with exact mapping
  - `nMatch` : Total mapped reads (unique + multiple)
  - `nMatchUnique` : Reads with unique mapping

Sequencing Saturation

Sequencing saturation is calculated as follows:

```
Saturation = 1 - (N_umi / N_reads)
```

where: - N_umi = number of unique CB/UMI/gene combinations - N_reads = number of reads with valid CB/UMI/gene

Table 6.11: *STARsolo summary statistics*

Metric	Value
Number of Reads	1,284,475,633
Reads With Valid Barcodes	85.58%
Sequencing Saturation	0.97%
Q30 Bases in CB+UMI	95.51%
Q30 Bases in RNA read	95.79%
Reads Mapped to Genome: Unique+Multiple	89.57%
Reads Mapped to Genome: Unique	3.64%
Reads Mapped to GeneFull: Unique+Multiple	75.64%
Reads Mapped to GeneFull: Unique	2.69%
Estimated Number of Cells	27,203
Unique Reads in Cells Mapped to GeneFull	12,794,311
Fraction of Unique Reads in Cells	36.98%
Mean Reads per Cell	470
Median Reads per Cell	381
UMIs in Cells	12,663,144
Mean UMI per Cell	465
Median UMI per Cell	378
Mean GeneFull per Cell	296
Median GeneFull per Cell	258
Total GeneFull Detected	6,035

In our case, the very low saturation (0.97%) indicates that we could sequence deeper to capture more unique molecules.

Key Points of Our Analysis

- The high number of ambiguous mappings (`nAmbigFeature = 936,951,032`) is typical for bacterial data due to the compact nature of the genome
- The majority of reads have exact mapping (`nExactMatch = 4,471,174,098`), indicating good mapping quality. This possibly includes both unique and multi-mapped reads that match exactly to their reference locations
- The number of detected cell barcodes (`nCellBarcodes = 699,355`) is high, suggesting good cellular diversity
- The number of UMIs (`nUMIs = 34,258,961`) indicates good molecular coverage

These metrics suggest that our data is of good technical quality, although the low saturation indicates potential for deeper sequencing.

6.4 Genefull summary stats

6.5 Interpretation of STARsolo Results

The STARsolo analysis revealed several key insights about our single-cell RNA-seq data:

6.5.1 Sequencing Quality and Mapping

- The sequencing quality is excellent, with over 95% of bases having Q30 quality scores in both barcode/UMI and RNA reads
- A high proportion (85.58%) of reads contained valid cell barcodes, indicating good library preparation
- The mapping rates are robust:
  - 89.57% of reads mapped to the genome (unique + multiple)
  - 75.64% of reads mapped to genes (unique + multiple)
- The low unique mapping rate (3.64% to genome, 2.69% to genes) is typical for bacterial RNA-seq due to ...

6.5.2 Mapping Terminology

- **Gene mapping**: Refers to reads mapped to annotated coding sequences (CDS) only
- **GeneFull mapping**: Includes reads mapped to all annotated features including:
  - Coding sequences (CDS)
  - Untranslated regions (UTRs)
  - Non-coding RNAs
  - Intergenic regions
  - This broader mapping approach is particularly relevant for bacterial transcriptomics as it captures the full complexity of the transcriptome

6.5.3 Cell Recovery and Expression

- We estimated 27,203 cells in our dataset
- The sequencing saturation is very low (0.97%), suggesting we could sequence deeper if needed
- Cell-level metrics show good coverage:
  - Mean/median reads per cell: 470/381
  - Mean/median UMIs per cell: 465/378
  - Mean/median genes per cell: 296/258
- We detected 6,035 genes in total across all cells

6.5.4 Data Quality Assessment

- The high Q30 scores and mapping rates indicate good technical quality
- The cell-level metrics suggest sufficient coverage for downstream analysis

- The low sequencing saturation suggests potential for deeper sequencing if needed
- The high proportion of reads with valid barcodes (85.58%) indicates good library preparation

6.5.5 filter

Nous on prend tout les barcodes pas ceux qui sont filtré donc les 820, 800 barcodes

6.6 apres starsolo mettre le nombre de reads avec valid barcodes dans la table

6.6.1 genome

circular representation of c-bacterial genome and read alignment voir comment faire ce type de figure et l'article qui l'avait fait

6.6.2

Attention j'ai filtré pour ne garder que les CDS mais certains pas annotés plutot exclure les tRNA et rRNA

In addition, we kept the highest-scored multimapping reads, assigning a fractional count based on the number of equally good alignments, because bacterial genomes are known to contain overlapping coding sequences. We then generated a matrix of gene counts for each cell (N-by-K matrix, with N cells and K genes).

dans l'article de kuchina ils ont filtré les cellules en fonction du nombre de reads et de genes :

"Processing of data from the heat shock experiment Clustering and data analysis for the speciesmixing experiment with heat shock treatment was performed using Scanpy (59). We only kept transcriptomes that had a **number of total reads higher than 200**. Then, we removed the **ribosomal and tRNA reads from the data**, retaining only reads that represented the mRNA counts for both species. We further filtered cells based on the mRNA counts, **retaining cells expressing >100 reads and >100 genes**, and additionally filtered the genes, retaining the **genes expressed in >5 cells**. We then applied standard Scanpy normalization and scaling, dimensionality reduction, and clustering, as described in the Scanpy tutorial (59, 60). The clusters were produced by the Louvain graph-clustering method and manually inspected for the top differentially expressed genes. After inspection, three pairs of transcriptionally similar clusters with fewer differentially expressed genes were merged, resulting in clusters 1, 2, and 3 in Fig. 1D." (Kuchina et al., 2021, p. 8) (pdf)

Chapter 7

Initialize the Seurat object with the raw (non-normalized data).

pbmc <- CreateSeuratObject(counts = pbmc.data, project = “pbmc3k”, min.cells = 3, min.features = 200)

https://rnabioco.github.io/cellar/previous/2019/docs/2_filtering_QC.html

7.1 Overview

This chapter presents the findings of our single-cell RNA-seq analysis of Pseudomonas, focusing on the division of labor within bacterial populations.

7.2 Single-cell RNA-seq Analysis

7.2.1 Data Quality and Preprocessing

7.2.2 Cell Type Identification

7.2.3 Differential Expression Analysis

7.2.4 Division of Labor Patterns

7.3 Functional Analysis

7.3.1 Pathway Enrichment

7.3.2 Gene Set Analysis

7.3.3 Regulatory Network Analysis

7.4 Integration with Previous Studies

7.5 Summary of Key Findings

Chapter 8

Discussion

We applied microSPLiT to P. brassicacearum growing in two different conditions in rich medium (M9F) and in minimal medium (M9).

- recepetion tardives des resultats
- mis beaucoup de temps pour le trimming (1mois) le temps de comprendre la structure de la librairie et
-
-
- analyse temporelle , metabolique , bulkRNAseq
- utilisation pour capturer specifique mRNA (voir article : 2 methodes existes ; et apres aussi peut etre fait)
- mais je pense deja bioinformatiquement on peut faire des choses pour ameliorer reads utilisables
- comparer avec differentes methodes de single cell RNA seq, voir si on observe toujours la meme chose ou pas
- versionnement des outils utilisés (renv , singularity, conda)
- rapport fait un template pour rendu propre
-

autres outils pourrait etre ajouter dans le pipeline comme BarQC alternative à Starsolo pour meilleur la lecture des barcodes (en considerant utilisant des positions non fixe (CIGAR motif) et evaluer la qualité UMIs et repartitions[17],

pour la qualité et contamination : centriguge et recentrifuge[18,19]

meme si moins de risque de contamination car cellules fixé ... (deve

- Nextflow pour le trimming, QC , et STARsolo serait une bonne idée , et barQC ; pourrait etre utile pour la communauté

-autono -gestion des datas tailles des ## Interpretation of Key Findings

8.0.1 Division of Labor Mechanisms

8.0.2 Biological Significance

8.0.3 Technical Considerations

8.1 Comparison with Existing Literature

8.1.1 Similarities with Previous Studies

8.1.2 Novel Insights

8.1.3 Discrepancies and Their Implications

8.2 Methodological Strengths and Limitations

8.2.1 Technical Advantages

8.2.2 Potential Limitations

8.2.3 Future Methodological Improvements

8.3 Biological Implications

8.3.1 Ecological Significance

8.3.2 Evolutionary Perspectives

8.3.3 Potential Applications

8.4 Future Research Directions

8.4.1 Open Questions

8.4.2 Suggested Follow-up Studies

8.4.3 Technical Improvements

8.5 Conclusion

Chapter 9

Conclusion and Future Work

9.1 Summary of Main Findings

9.1.1 Key Discoveries

9.1.2 Methodological Contributions

9.1.3 Biological Insights

9.2 Impact on the Field

9.2.1 Contribution to Single-cell RNA-seq Methodology

9.2.2 Contribution to Pseudomonas Research

9.2.3 Broader Implications for Microbial Ecology

9.3 Future Research Directions

9.3.1 Technical Improvements

9.3.2 Biological Questions to Address

9.3.3 Potential Applications

9.4 Final Remarks

9.5 References

Bibliography

1. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**, eaba5257 (2021).
2. Gaisser, K. D. *et al.* High-throughput single-cell transcriptomics of bacteria using combinatorial barcoding. *Nature Protocols* **19**, 3048–3084 (2024).
3. Morris, J. J., Lenski, R. E. & Zinser, E. R. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036–12 (2012).
4. Morris, E. K. *et al.* Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution* **4**, 3514–3524 (2014).
5. Estrela, S., Kerr, B. & Morris, J. J. Transitions in individuality through symbiosis. *Current Opinion in Microbiology* **31**, 191–198 (2016).
6. Nishimura, M., Takahashi, K. & Hosokawa, M. Recent advances in single-cell RNA sequencing of bacteria: Techniques, challenges, and applications. *Journal of Bioscience and Bioengineering* (2025) doi:10.1016/j.jbiosc.2025.01.008.
7. Ostner, J. *et al.* BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis. doi:10.1101/2024.06.22.600071.
8. Brettner, L. & Geiler-Samerotte, K. Single-cell heterogeneity in ribosome content and the consequences for the growth laws. *bioRxiv: The Preprint Server for Biology* 2024.04.19.590370 (2024) doi:10.1101/2024.04.19.590370.
9. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)* **32**, 3047–3048 (2016).
10. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
11. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

12. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).

13. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. doi:10.1101/2021.05.05.442755.

14. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016).

15. Sullivan, D. K. *et al.* kallisto, bustools and kb-python for quantifying bulk, single-cell and single-nucleus RNA-seq. *Nature Protocols* **20**, 587–607 (2025).

16. Kuijpers, L. *et al.* Split pool ligation-based single-cell transcriptome sequencing (SPLiT-seq) data processing pipeline comparison. *BMC Genomics* **25**, 361 (2024).

17. Rossello, M., Tandonnet, S. & Almudi, I. BarQC: Quality Control and Preprocessing for SPLiT-Seq Data. doi:10.1101/2025.02.04.635005.

18. Martí, J. M. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Computational Biology* **15**, e1006967 (2019).

19. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research* **26**, 1721–1729 (2016).

Appendix A

- plan de plaque
- librairies avec TSO
- tableau choix de profondeur / nombre de cellules
- mettre difference entre experience de kuhina et la notre pour les resultats de Starsolo

Appendix B

Annexe B: erferfrefref

B.1 summary stats and features.stats for Gene

nUnmapped	114628761
nNoFeature	20579292
nAmbigFeature	934096737
nAmbigFeatureMultimap	934096737
nTooMany	0
nNoExactMatch	124864
nExactMatch	4458742628
nMatch	964094047
nMatchUnique	30022209
nCellBarcodes	689818
nUMIs	29709734

Number of Reads,1284475633 Reads With Valid Barcodes,0.85576 Sequencing Saturation,0.0104081 Q30 Bases in CB+UMI,0.955089 Q30 Bases in RNA read,0.957923 Reads Mapped to Genome: Unique+Multiple,0.895694 Reads Mapped to Genome: Unique,0.0363836 Reads Mapped to Gene: Unique+Multipe Gene,0.750574 Reads Mapped to Gene: Unique Gene,0.0233731 Estimated Number of Cells,27268 Unique Reads in Cells Mapped to Gene,11121465 Fraction of Unique Reads in Cells,0.370441 Mean Reads per Cell,407 Median Reads per Cell,331 UMIs in Cells,10998607 Mean UMI per Cell,403 Median UMI per Cell,327 Mean Gene per Cell,253 Median Gene per Cell,221 Total Gene Detected,5894

B.2 warning

!!!!! WARNING: while processing sjdbGTFfile=/projects/microsplit/data/processed_data/STARsolo_result/merged_tr line: CP125962.1 Genbank exon 298557 300953 . - 0 transcript_id “gene-QLH64_29550”; gene_id “gene-QLH64_29550”; gene_name “QLH64_29550”; exon end = 300953 is larger than the chromosome CP125962.1 length = 299955 , will skip this exon

Log of the STARsolo run : Alignment statistics: ———————— Number of input reads | 1284475633 Average input read length | 135 Uniquely mapped reads number | 46733841 Uniquely mapped reads % | 3.64% Number of reads mapped to multiple loci | 1103762730 % of reads mapped to multiple loci | 85.93% Number of reads unmapped: other | 128049614 % of reads unmapped: other | 9.97% Mismatch rate per base, % | 0.38% Fri May 30 15:39:42 CEST 2025 - Pipeline completed!

B.2.1 Pretest STARSolo on BC_0077 without trimming :

nNoAdapter	0
nNoUMI	0
nNoCB	0
nNinCB	0
nNinUMI	4467771
nUMIhomopolymer	4325324
nTooMany	0
nNoMatch	66837523
nMismatchesInMultCB	1680783
nExactMatch	234544811
nMismatchOneWL	13639378
nMismatchToMultWL	0

barcodes stats

Genfull summary stats

Metric	Count
nUnmapped	89,425,075
nNoFeature	1,000,851
nAmbigFeature	152,909,678
nAmbigFeatureMultimap	152,443,034
nTooMany	0
nNoExactMatch	185,805

Metric	Count
nExactMatch	729,180,878
nMatch	157,719,964
nMatchUnique	4,847,425
nCellBarcodes	168,346
nUMIs	305,287

Metric	Value
Number of Reads	325,495,590
Reads With Valid Barcodes	76.19%
Sequencing Saturation	93.70%
Q30 Bases in CB+UMI	92.28%
Q30 Bases in RNA read	86.19%
Reads Mapped to Genome: Unique+Multiple	58.16%
Reads Mapped to Genome: Unique	2.15%
Reads Mapped to GeneFull: Unique+Multiple	48.46%
Reads Mapped to GeneFull: Unique	1.49%
Estimated Number of Cells	66,026
Unique Reads in Cells Mapped to GeneFull	3,538,648
Fraction of Unique Reads in Cells	73.00%
Mean Reads per Cell	53
Median Reads per Cell	36
UMIs in Cells	202,967
Mean UMI per Cell	3
Median UMI per Cell	2
Mean GeneFull per Cell	2
Median GeneFull per Cell	2
Total GeneFull Detected	5,295

Gene summary stats

```
nUnmapped       89425075
nNoFeature       7350074
nAmbigFeature  147588031
```

nAmbigFeatureMultimap	147588029
nTooMany	0
nNoExactMatch	182600
nExactMatch	704353731
nMatch	151371640
nMatchUnique	3820029
nCellBarcodes	135433
nUMIs	224743

Number of Reads,325495590 Reads With Valid Barcodes,0.76192 Sequencing Saturation,0.941167 Q30 Bases in CB+UMI,0.922758 Q30 Bases in RNA read,0.861863 Reads Mapped to Genome: Unique+Multiple,0.581583 Reads Mapped to Genome: Unique,0.0215405 Reads Mapped to Gene: Unique+Multipe Gene,0.46505 Reads Mapped to Gene: Unique Gene,0.011736 Estimated Number of Cells,47264 Unique Reads in Cells Mapped to Gene,2582244 Fraction of Unique Reads in Cells,0.675975 Mean Reads per Cell,54 Median Reads per Cell,40 UMIs in Cells,136574 Mean UMI per Cell,2 Median UMI per Cell,2 Mean Gene per Cell,2 Median Gene per Cell,2 Total Gene Detected,4838

Appendix C

Annexe C: codcefe

Master's Thesis in Bioinformatics

University of Rennes

ECOBIO
Rennes

This thesis was conducted in the framework of the Master's program in Bioinformatics at the University of Rennes. The research presented here contributes to the field of computational biology and bioinformatics.