Msc Bioinformatics thesis

Study of Division of Labor in Pseudomonas through 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 19 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.

Nulla eget cursus ipsum. Vivamus porttitor leo diam, sed volutpat lectus facilisis sit amet. Maecenas et pulvinar metus. Ut at dignissim tellus. In in tincidunt elit. Etiam vulputate lobortis arcu, vel faucibus leo lobortis ac. Aliquam erat volutpat. In interdum orci ac est euismod euismod. Nunc eleifend tristique risus, at lacinia odio commodo in. Sed aliquet ligula odio, sed tempor neque ultricies sit amet.

Etiam quis tortor luctus, pellentesque ante a, finibus dolor. Phasellus in nibh et magna pulvinar malesuada. Ut nisl ex, sagittis at sollicitudin et, sollicitudin id nunc. In id porta urna. Proin porta dolor dolor, vel dapibus nisi lacinia in. Pellentesque ante mauris, ornare non euismod a, fermentum ut sapien. Proin sed vehicula enim. Aliquam tortor odio, vestibulum vitae odio in, tempor molestie justo. Praesent maximus lacus nec leo maximus blandit.

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

This method allows for the analysis of transcriptomic heterogeneity within bacterial populations at single-cell resolution, providing insights into gene expression patterns and cellular diversity that would be masked in bulk RNA-seq approaches.

operon , facteur sigma , new genome annotatipn ...

Chapter 2

Materials and Methods

2.1 Bacterial culture

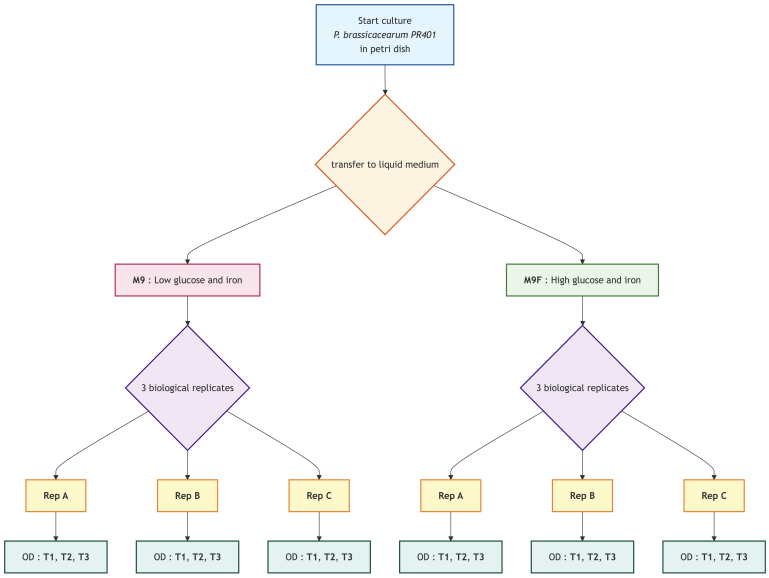


Figure 2.1: *Experimental design for bacterial culture*

The workflow illustrates the progression from initial culture of P. brassicacearum R401 on petri dishes to liquid culture in two different media: M9 (low glucose and iron) and M9F (high glucose and iron). Each medium condition was replicated three times (Rep A, B, C) and growth was monitored at three timepoints (T1, T2, T3) by optical density measurements. This design resulted in 18 total experimental conditions (2 media × 3 biological replicates × 3 timepoints) for subsequent single-cell RNA-seq analysis.

An isogenic[1] population of *P. brassicacearum* R401 was initially cultured on petri dishes and then transferred to different liquid media to investigate the effects of nutrient availability on bacterial

[1] An isogenic population refers to a group of organisms that are genetically identical, derived from a single ancestral cell or clone.

growth and gene expression (Figure 2.1).

Two distinct culture conditions were applied to the bacteria: M9 medium containing low glucose and low iron concentrations, and M9F medium containing high glucose and high iron concentrations (see Table A.1 for detailed concentrations). Each condition was replicated three times to ensure statistical robustness of the experimental results. The bacterial growth was monitored by measuring optical density (OD) at regular intervals. The growth curves obtained from these measurements are presented (Figure 2.2) and (Table A.2). This experimental design resulted in a total of 18 conditions: 2 media types × 3 biological replicates × 3 time points, providing comprehensive coverage of the growth dynamics under different nutrient conditions.

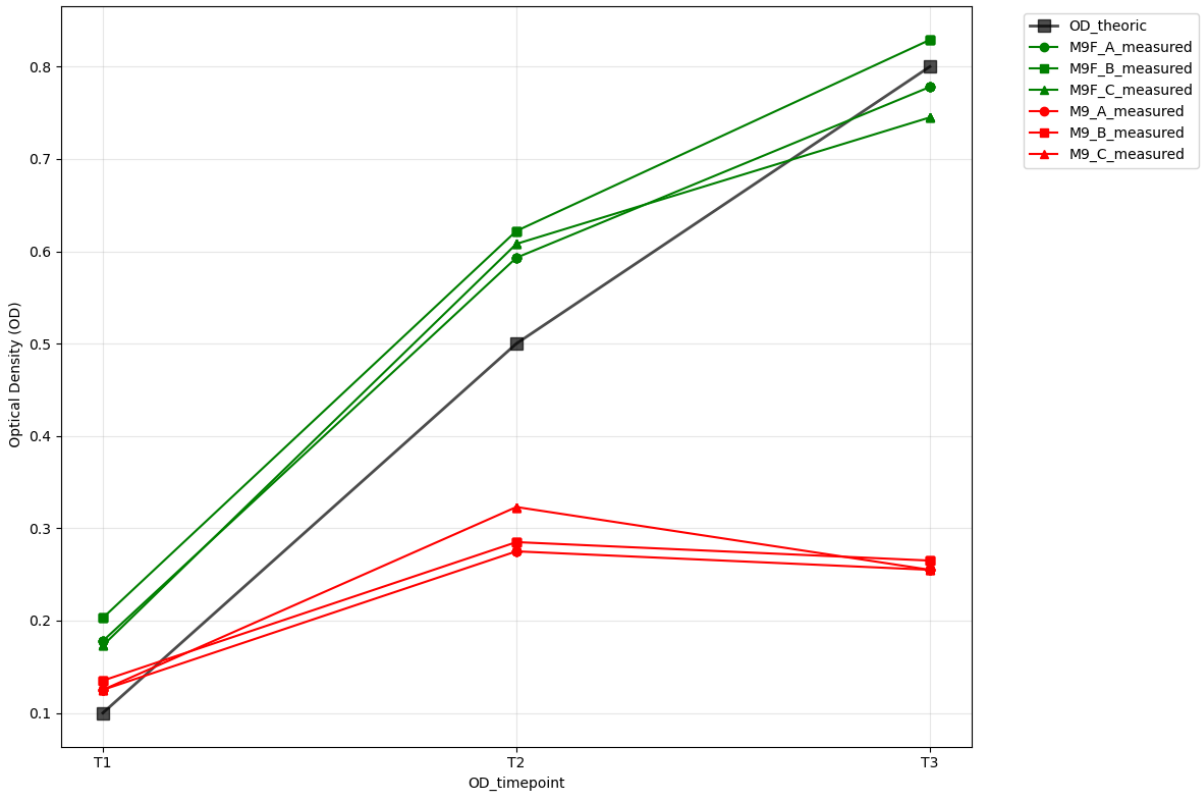


Figure 2.2: *Bacterial growth dynamics of* P. brassicacearum *R401 populations measured by optical density (OD600) cultured under two different nutrient conditions: M9 (low glucose/iron) and M9F (high glucose/iron).*

Measurements were taken at three timepoints (T1, T2, T3) for three biological replicates (Rep A, B, C).

The growth curves reveal distinct patterns between the two culture conditions. Bacteria grown in M9F medium (high glucose and iron) exhibited significantly higher growth rates and reached higher optical densities (OD 0.17-0.21 at T1, 0.59-0.63 at T2, 0.74-0.83 at T3) compared to M9 medium (low glucose and iron) which showed limited growth (OD 0.13 at T1, 0.28-0.33 at T2, 0.26 at T3). While M9F cultures showed continued growth from T2 to T3, the growth rate slowed down

during this period, indicating the beginning of transition towards stationary phase. The M9 cultures appeared to reach a growth plateau by T3, while M9F cultures maintained higher densities despite the growth deceleration, suggesting nutrient limitation in the M9 condition. Biological replicates showed excellent reproducibility validating the experimental design.

Cells were collected at each timepoint (T1, T2, T3) from all biological replicates for subsequent single-cell RNA-seq analysis using the Microbial split-pool ligation transcriptomics (microSPLiT) protocol[1,2].

2.2 microSPLiT protocol

MicroSPLiT[1,2] is a high-throughput single-cell RNA sequencing method for bacteria, capable of profiling transcriptional states in hundreds of thousands of cells per experiment without the need for specialized equipment[2,6]. Unlike other single-cell RNA-seq approaches that require physical isolation of individual cells (e.g., plate-based or droplet-based methods), microSPLiT uses a split-pool barcoding strategy to uniquely label transcripts within each cell. (see Figure A.1 for an overview of single-cell RNA-seq methods in bacteria)

> Information
>
> The microSPLiT strategy will not be described in detail here; for more information, see Gaisser protocol.[2]. Only the key steps necessary for a general understanding of the method are presented below.

a, Bacterial cells are fixed overnight and permeabilized before the mRNA is preferentially polyadenylated. 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. After the barcoding reaction, cells are pooled together and split again into the second barcoded plate. 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. Cells are pooled and split into the third and final barcoded plate. 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. 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).[2]

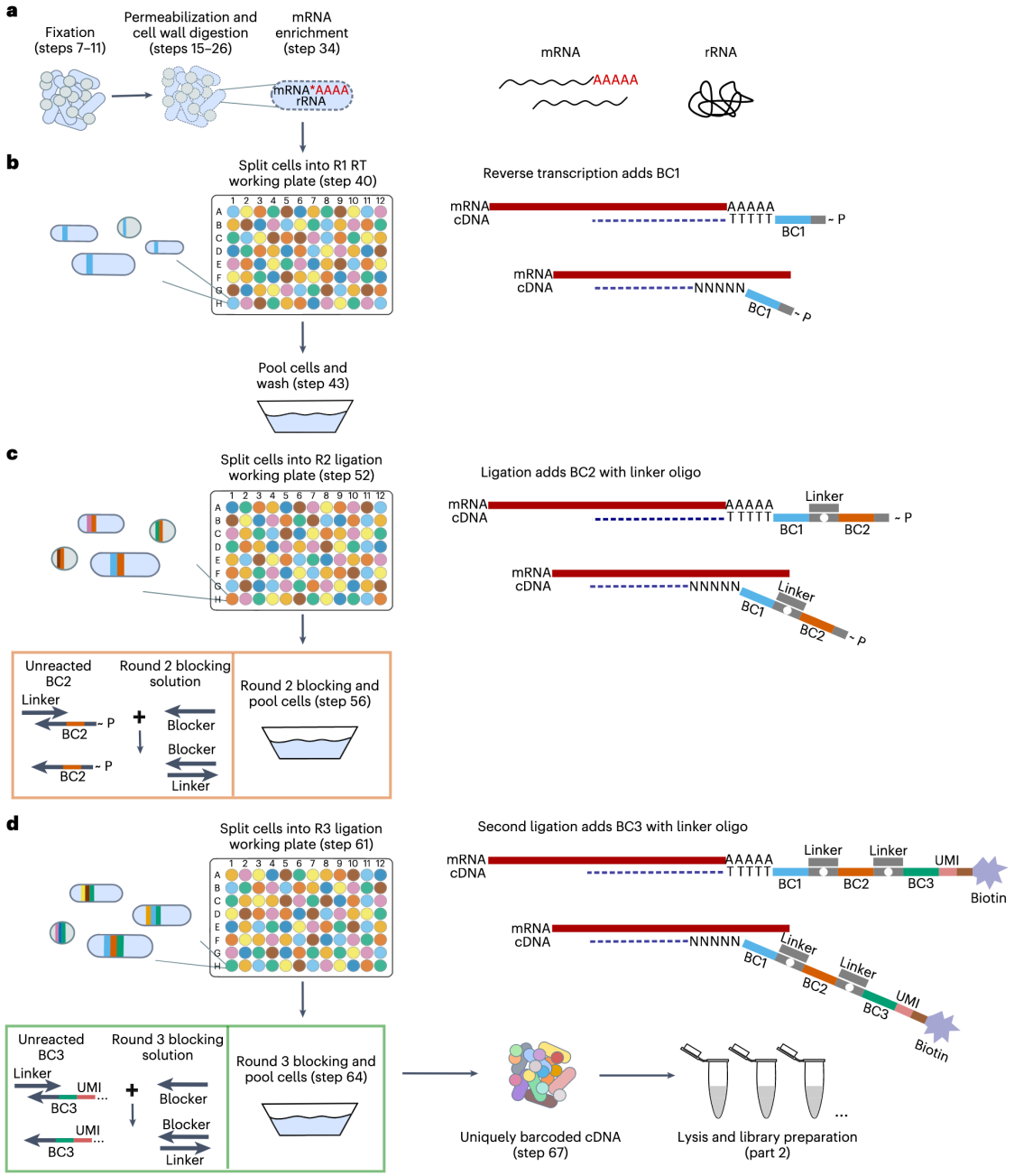


Figure 2.3: *MicroSPLiT in-cell cDNA barcoding scheme.*

2.2.1 Fixation and permeabilization

The first step is fixation of the bacterial suspension with formaldehyde Figure 2.3 immediately after sampling the 18 conditions Figure 2.1. This preserves the transcriptomic state and cross-links RNA to proteins, preventing leakage of each cell's transcriptome. Next, cells are permeabilized using mild detergent and lysozyme, allowing enzymes and oligonucleotides to access intracellular RNA for barcoding.

> **Note**
>
> Adequate permeabilization is essential for efficient barcoding, but over-permeabilization can compromise cell integrity. For successful single-cell resolution, cells must remain intact after permeabilization to allow multiple split-pool steps and retain cross-linked RNA. Figure 2.3

2.2.2 mRNA enrichment

After permeabilization, the transcripts in the fixed and permeabilized cells undergo in situ polyadenylation with the addition of a 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 ribosomique RNA (rRNA) Figure 2.3.

2.2.3 Barcoding

> **Tip**
>
> **The protocol utilizes several rounds of split-pool barcoding where cells are distributed into 96-well plates, barcoded, pooled, and redistributed for subsequent rounds, creating unique barcode combinations that identify individual cells.**

Barcoding round 1 (R1) to identify the condition and the technical replicate

Each of the 18 samples is split into 5 technical replicates for barcoding, resulting in 90 subsamples. These technical replicates are then distributed into individual wells of a 96-well plate (6 wells not used), with each well containing uniquely barcoded primers Figure 2.3. In each well, mRNA is reverse transcribed into cDNA using a mix of barcoded poly(T) and random hexamer primers. The primers used in each well contain either a dT15 sequence to capture polyadenylated mRNA or six random nucleotides to bind any RNA, followed by a universal sequence for subsequent ligation steps. By assigning each technical replicate to a specific well, all cells in the same well receive the same unique barcode during reverse transcription. This allows each technical replicate and condition to be identified later based on the first barcode.

Barcoding rounds 2 (R2) and 3 (R3) for unique cell and transcript identification

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 Figure 2.3. Due to the random redistribution of cells, each well of the second-round plate is likely to contain a mix of cells with different first-round barcodes, resulting in highly diverse barcode combinations. The ligation reaction is carried out by the T4 DNA ligase, which requires double-stranded DNA. Therefore, in the second

barcoding plate, each barcode is first hybridized to a short linker oligonucleotide whose overhang is complementary to the universal sequence at the 5' end of the RT barcodes. Figure 2.3.

> **Note**
>
> After the ligation step, some barcodes may remain unreacted in the solution. To prevent these free barcodes from attaching non-specifically to DNA from other cells during pooling, a blocker strand is added. This blocker has a longer complementary region to the linker, allowing it to displace any unreacted barcodes from the linker and thus ensures that only correctly ligated barcodes remain attached to the cDNA. Figure 2.3

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 universal PCR handle and a 5' biotin[2] molecule. The ligation reaction is stopped by adding a second blocker strand and EDTA.

> **Warning**
>
> In our experiment, only 95 out of the 96 wells of the R3 plate are used to minimize potential bias in cell distribution. This setup allows for $90 \times 96 \times 95 = \mathbf{820{,}800}$ **possible barcode combinations**, enabling the identification of up to 820,800 individual cells.

2.2.4 Sub-library and sequencing preparation

The pooled cells are washed, counted, and divided into multiple sub-libraries. Only sub-libraries containing approximately 3,000 cells were selected for sequencing, in order to maximize sequencing depth per cell and minimize barcode collision rates which is the probability that two cells receive the same barcode combination.

After lysis and cDNA purification on streptavidin beads (Figure A.2), a second reverse transcription is performed to improve cDNA yield, during which a template switch oligo (TSO) is added to introduce a 3' adapter. The resulting cDNA is then amplified by PCR. Following amplification, a size selection step removes short byproducts such as adapter or barcode dimers, ensuring that only high-quality cDNA fragments are retained for sequencing.

To optimize sequencing depth, the final library was split into four sub-libraries, each receiving a distinct index during adapter ligation: BC_0076 (CAGATC), BC_0077 (ACTTGA), BC_0078 (TAGCTT),

[2] Biotin is a small vitamin molecule that binds with extremely high affinity to streptavidin. This biotin-streptavidin interaction is used for the selective capture and purification of biotinylated cDNA molecules on streptavidin-coated beads during the library preparation process.

and BC_0079 (GGCTAC). These indexes were used solely to improve sequencing quality and balance on the NovaSeq platform, without introducing any experimental or technical variation between sub-libraries.

2.2.5 Sequencing and demultiplexing sub-libraries

Sequencing was performed on a NovaSeq™ X plus instrument at GenoBIRD platform in paired-end mode. The library pool was loaded onto all lanes of the flowcell at a final concentration of 200 pM with 20% PhiX[3]. The sequencing program consisted of 241 cycles for Read 1, 6 cycles for Index i7 and 91 cycles for Read 2. The sequencing facility performed demultiplexing of sub-libraries, resulting in eight FASTQ files (R1 and R2 for each index). R1 files contain the cDNA sequences, while R2 files contain the cell barcodes (from the three split-pool rounds) and unique molecular identifiers (UMIs).

- For each index, two paired-end FASTQ files were generated :
  - **R1** contains the cDNA sequence of interest (transcriptome).
  - **R2** contains the cell barcodes and unique molecular identifiers (UMIs).

2.3 Pipeline for microSPLiT data processing

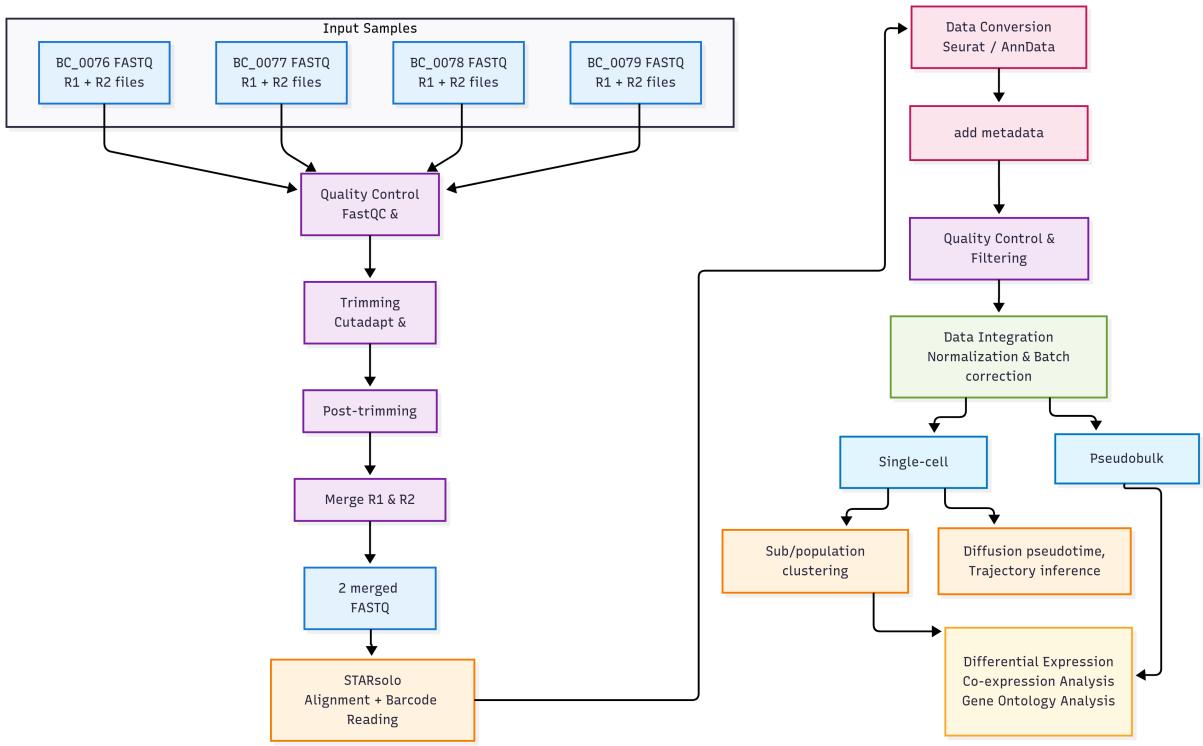


Figure 2.4: *Comprehensive pipeline for microSPLiT single-cell RNA-seq data processing.*

[3]PhiX is a control library containing a known viral genome sequence that is spiked into sequencing runs to monitor sequencing quality, calibrate base calling, and provide a reference for quality control metrics. It helps ensure accurate sequencing performance and data quality assessment.

The workflow encompasses the complete analytical process from raw sequencing data to biological interpretation, including quality control and preprocessing of FASTQ files, alignment and quantification using STARsolo, data structuring with metadata assignment, quality filtering and integration, single-cell and pseudobulk analysis approaches, population characterization through clustering and trajectory inference, and downstream expression analysis including differential expression, co-expression networks, and gene ontology enrichment. Figure 2.4

2.3.1 Preprocessing of the sequencing data

All quality control, trimming, alignment, barcode reading and generation of cell-gene count matrix steps were performed on the GenOuest high-performance computing cluster using SLURM job scripts and parallelization to ensure efficient and reproducible analysis of large-scale sequencing data.

Quality control and trimming

Read quality was initially assessed for all four libraries (R1 and R2) using FastQC[8] and MultiQC[9]. Trimming was then performed with Cutadapt[10] and Fastp[11] to clean the sequencing data. For R2 files, trimming focused on filtering for valid barcodes. For R1 files, trimming removed various artifacts: template-switching oligo (TSO) sequences at the 5' end, adapter sequences, and 3' artifacts including polyG stretches (NovaSeq-specific artifacts) and potential R1 complement sequences when cDNA was short. Only reads with a minimum length of 25 bp were kept. The detailed trimming pipeline is described in Appendix Section Section A.6. After trimming, read quality was reassessed with FastQC[8] and MultiQC[11], and results were merged into unique files (R1 and R2).

Quality control after trimming

After trimming, a quality control step was performed to ensure that the remaining reads were of high quality and suitable for downstream analysis. This step involved checking the distribution of read lengths, GC content, and other relevant metrics.

Merge files

After quality control, the files from all four libraries (R1 and R2 for each index) were merged into a single file for each index. This step ensured that all cells from all conditions and technical replicates were included in the analysis.

Alignment, barcode reading and generation of cell-gene count matrix

The alignment and quantification pipeline was implemented using STARsolo[12,13], an extension of the STAR aligner specifically designed for single-cell RNA-seq data. STARsolo was chosen based on benchmarking studies showing it offers the best combination of speed and reproducibility for SPLiT-seq / microSPLiT data analysis[14]. The implementation followed the recommendations outlined

in Gaisser et al[2] for optimal microSPLiT data processing. Complete pipeline scripts and parameters are detailed in Appendix Section Section A.7.

Reference genome and annotation. The reference genome of *Pseudomonas brassicacearum* R401: ASM3006410v1 (GCA_030064105.1) and its annotation were downloaded from GenBank. The GFF3 annotation file was converted to GTF format using gffread (Cufflinks[15] package) for compatibility with STARsolo.

Correcting GTF file for compatibility with STAR. The conversion was verified to ensure all required fields were present, particularly confirming that genes were labeled as 'exon' features rather than 'CDS' descriptors, and that chromosome names matched between reference sequence and annotation files. This correction was performed to ensure proper compatibility with STARsolo.

Alignment parameters. The pipeline used optimized parameters for microSPLiT data: minimum 50 matching bases for valid alignment and 1 mismatch tolerance for both barcode and UMI matching. The complex barcode structure (R2) was configured with positions 0_10_0_17, 0_48_0_55, and 0_78_0_85 for the three barcoding rounds, and UMI position 0_0_0_9.

Output matrices. STARsolo generated count matrices using `GeneFull` feature counting and `UniqueAndMult-Uniform` mapping strategy (distributes multi-mapped reads uniformly). Although bacteria lack introns, `GeneFull` was chosen to include reads that may map to intergenic regions or incompletely annotated gene boundaries, which is common in bacterial genomes. The `UniqueAndMult-Uniform` strategy is particularly important for bacterial genomes due to the presence of paralogous genes, repetitive sequences, and operon structures that can result in reads mapping to multiple genomic locations. Raw data matrices (unfiltered barcodes) were used for downstream analysis[2], with cell filtering applied later in the processing pipeline.

Quality control and output files. After STARsolo analysis, quality control was performed using the Log.final.out and summary.csv files. The main output files for downstream analysis included barcode.tsv (cell identifiers), features.tsv (gene identifiers), and UniqueAndMult-Uniform.mtx (count matrix).

2.3.2 Single-cell data processing

All downstream analyses were performed locally using a reproducible development container environment (Docker[16] and Rocker Project) with Visual Studio Code dev containers to ensure consistent software versions and analysis reproducibility, including version control for R and Python packages.

Data conversion and metadata assignment.

Raw count matrix was converted to Seurat v5 objects[17] in R. Two types of metadata were assigned:

- **Cell metadata** based on barcode combinations, linking each cell to its experimental condition (medium type, biological and technical replicate, timepoint, and well plate position at each barcoding round)
- **Gene metadata** including sequence type and gene symbols for downstream analysis.

For Python-based analyses, Seurat objects were converted to AnnData objects[18] for use with Scanpy[19].

Quality control and filtering.

The data was processed through quality control and filtering steps. Quality control metrics were calculated for each cell, including total UMI counts and number of detected genes. Cells with low quality metrics or potential contamination were filtered out to ensure robust downstream analysis. Because the reads depth or cell expression is different between : Stress condition (M9F) and non-stress condition (M9) et also between the 3 timepoints (T1, T2, T3) (see suppl figure) , we decided to filter the cells based on the number of UMI counts per cell (nCounts) and the number of detected genes (nFeatures) per sample.

Cell filtering strategy. For each of the 90 samples (18 conditions $\times$ 5 technical replicates), cells were filtered to retain only the top 10% of cells with the highest number of expressed genes (nFeatures) per sample, resulting in 82,080 cells from the initial 820,800 possible cells. This selection strategy ensures that only the highest quality cells from each sample are retained for downstream analysis.

UMI-based filtering. Following the initial gene-based filtering, cells were further filtered based on UMI counts per cell (nCounts). From each sample, only the best cells were retained to achieve a target of 3,000 cells total across all conditions. This approach maximizes sequencing depth per cell while maintaining representation across all experimental conditions.

Doublet estimation. Based on the doublet rate of 0.34% reported by Kuchina et al.[2], we estimated approximately 10.2 potential doublets among the 3,000 selected cells ($0.34 \times 3000/100$). To remove potential doublets, cells were filtered using arbitrary thresholds: cells with nCount $>$ 5000 or nFeature $>$ 1500 were excluded, as these thresholds typically indicate doublet contamination in single-cell RNA-seq data (3 cells removed).

Gene filtering. Only mRNA sequences were retained for analysis, excluding ribosomal RNA and transfer RNA sequences etc. Additionally, genes were required to be expressed in a minimum of 5 cells to ensure statistical reliability in downstream analyses.

Final dataset. After applying these filtering criteria, the final processed dataset contained high-quality single-cell transcriptomes ready for differential expression analysis and cell state characterization.

> ⚠ Warning
>
> The filtering thresholds used in this study were chosen arbitrarily, as each single-cell RNA-seq study typically defines its own filtering criteria. These choices may have implications for the interpretation of results, particularly regarding the representation of different cell states and the detection of rare cell populations. The potential biases introduced by these filtering strategies will be discussed in the Results and Discussion sections, as this is a common limitation across single-cell RNA-seq studies in the field.

-retiré un replica tech aussi (voir annexe) au total on conserve : x cellules , avec x genes

Chapter 3

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

3.1 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:

3.2 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

3.3 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

3.4 Genome

3.5 Transcriptome

Chapter 4

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 5

housekeeping genes

choix ou non de pooler les replicas techniques ensembles

5.1 Summary of results

5.2 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.

5.2.1 Barcode Statistics Interpretation

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

- **Barcode Quality**:

Table 5.1: *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 5.3: *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 5.5: *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 5.7: *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.

5.3 Genfull

5.3.1 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 5.9: *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 5.11: *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.

5.4 Genefull summary stats

5.5 Interpretation of STARsolo Results

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

5.5.1 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 ...

5.5.2 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

5.5.3 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

5.5.4 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

5.5.5 filter

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

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

5.6.1 genome

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

5.6.2

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 6

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

le choix de filtration sera discuté apres

because the reads depth is different between : Stress condition (M9F) and non-stress condition (M9) et also between the 3 timepoints (T1, T2, T3) (voir figure, reuslttas comme deja observé chez)

Single-cell analysis Two separate libraries were concatenated after filtering, log-normalized by 600 counts per cell, and scaled to unit variance and zero mean. Subsequently, ComBat(70) was run for batch effect correction. Normalized expression data was dimensionally reduced using principal component analysis (PCA). Shared neighbor graphs and uniform manifold approximation representations (UMAP(71)) were calculated with the first 12 principal components. All subsequent calculations were run in Python using Scanpy(72) documentation for single-cell analysis. Differential gene expression analysis Scanpy gene ranking functions (sc.tl.rank_genes_groups and sc.get.rank_genes_groups_df) were used to analyze and retrieve statistical data between two groups of interest within the annotated data object. The output parameters included names of all genes, z-score, log fold change, p-values, and adjusted p-values. To create a volcano plot from this data, either the |score| or -log(adjusted p-value) was plotted on the y-axis against the log-fold change on the x axis. Lists of defense genes from DefenseFinder(39), host response genes upregulated after phage treatment, and CPS genes were used to assign colors to each point. Host response genes were taken as the top 15 distinct

genes identified in the phage-treated sample only clustering analysis relative to the untreated sample. Initial CPS gene annotations were mapped by and received from Dr. Laurie Comstock (University of Chicago). Calculation of co-expression scores To assess the co-expression between two genes A and B, first, probabilities of expression of individual genes were calculated as fractions of cells having above-zero normalized non-scaled expression values p(A) and p(B), respectively. Then, the probability of simultaneous expression of the two genes, p(A&B), was calculated as a fraction of cells having above-zero normalized non-scaled expression values for both inspected genes. Co-expression score was calculated as a ratio of p(A&B) to the multiplication of p(A) and p(B). Values close to 1 indicate independent expression of genes; values above 1 indicate co-expression, and values below 1 indicate mutually exclusive expression of genes. To define a border of a CPS operon, mean co-expression values were calculated between genes adjacent to the CPS operon and core genes of a CPS operon. Co-expression between gene sets (e.g., CPS operons) was calculated similarly to genes, with probability of expression calculated as a fraction of cells having above-zero normalized non-scaled expression values for any of genes within a set. Diffusion pseudotime analysis The data was first subsetted into the phage-treated sample alone. A root cell was defined using adata.uns['iroot'] = np.flatnonzero(adata.obs['louvan']==0)[1], which selected a random indexed cell from the untreated cluster within all phage-treated cells. Scanpy diffusion maps were created prior to running the existing diffusion pseudotime tool with 0 branchings and 10 diffusion components. For downstream analysis using pseudotime values, 5 bins of equal size were created to group cells into pseudotime ranges (0.0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, and 0.8-1.0). Differential expression analysis was run to identify the top 10 distinct B. fragilis genes within each pseudotime bin, and the Clusters of Orthologous Groups (COGs) database was used to define functional categories for these genes. Duplicate genes were removed. Raw mean expression for each of these genes, grouped by functional category, was calculated using the adata.raw.X matrix.

[20][21 22 23]

24

6.1 Overview

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

6.2 Single-cell RNA-seq Analysis

6.2.1 Data Quality and Preprocessing

6.2.2 Cell Type Identification

6.2.3 Differential Expression Analysis

6.2.4 Division of Labor Patterns

6.3 Functional Analysis

6.3.1 Pathway Enrichment

6.3.2 Gene Set Analysis

6.3.3 Regulatory Network Analysis

6.4 Integration with Previous Studies

6.5 Summary of Key Findings

Chapter 7

Discussion

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

-d'autres methodes de filtrage , log log comme dans...

While other tools exist for alignment and barcode reading in SPLiT-seq/microSPLiT protocols, such as Kallisto[25,26], The pipeline could be further improved with additional quality verification steps and implementation of workflow management systems such as Nextflow and nf-core for enhanced reproducibility and scalability.

- `different other tools existe comme pour alignement et lecture des barcodes SPLiTseq/`

pleins parametre pour etre change, unique alignement peut de rRNA je trouve par rapport a autre articles

To remove empty and low gene detection barcodes, we applied the "knee" detection filter previously described in Brettner et al. 202435. These quality-thresholded gene-by-barcode matrices were then converted to the R datatype, Seurat Objects, using the Seurat R package for further analyses83.

- biais de la methode
- coloration fluorescences pour voir marqueur , cell states , i

-agotation oiu pas

-discuter du pipeline bac -voir

- 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[27],

pour la qualité et contamination : centriguge et recentrifuge[28,29]

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

Chapter 8

Conclusion and Future Work

8.1 Summary of Main Findings

8.1.1 Key Discoveries

8.1.2 Methodological Contributions

8.1.3 Biological Insights

8.2 Impact on the Field

8.2.1 Contribution to Single-cell RNA-seq Methodology

8.2.2 Contribution to Pseudomonas Research

8.2.3 Broader Implications for Microbial Ecology

8.3 Future Research Directions

8.3.1 Technical Improvements

8.3.2 Biological Questions to Address

8.3.3 Potential Applications

8.4 Final Remarks

8.5 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. Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data.
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. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
11. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
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. Kuijpers, L. *et al.* Split pool ligation-based single-cell transcriptome sequencing (SPLiT-seq) data processing pipeline comparison. *BMC Genomics* **25**, 361 (2024).

15. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511–515 (2010).

16. Merkel, D. Docker: Lightweight linux containers for consistent development and deployment. *Linux J.* **2014**, 2:2 (2014).

17. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* **42**, 293–304 (2024).

18. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Access and store annotated data matrices. *Journal of Open Source Software* **9**, 4371 (2024).

19. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).

20. Gupta, A. *et al.* Combinatorial phenotypic landscape enables bacterial resistance to phage infection. doi:10.1101/2025.01.13.632860.

21. Cyriaque, V. *et al.* Single-cell RNA sequencing reveals plasmid constrains bacterial population heterogeneity and identifies a non-conjugating subpopulation. *Nature Communications* **15**, 5853 (2024).

22. Brettner, L., Eder, R., Schmidlin, K. & Geiler-Samerotte, K. An ultra high-throughput, massively multiplexable, single-cell RNA-seq platform in yeasts. *Yeast* **41**, 242–255 (2024).

23. 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.

24. Korshoj, L. E. & Kielian, T. Bacterial single-cell RNA sequencing captures biofilm transcriptional heterogeneity and differential responses to immune pressure. *Nature Communications* **15**, 10184 (2024).

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

26. 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).

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

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

29. 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

Appendix A

A.1 Media composition

The following table details the composition of the culture media used in this study.

Table A.1: *Media composition for bacterial culture experiments*

Component	M9F (mL)	M9 (mL)
Base M9	125	125
Glucose 1M	2.5	0.25
MgSO4 1M	0.25	0.25
CaCl2 1M	0.0125	0.0125
FeCl3 100mM	0.1277	0
Vf (mL)	**127.7625**	**125.5125**

The M9 medium represents low nutrient conditions with minimal glucose and iron concentrations, while M9F medium provides high nutrient availability with elevated glucose and iron levels.

A.2 Growth curves data

The following table presents the optical density (OD) measurements for each culture condition and biological replicate at the three timepoints.

Table A.2: *Growth curves data for bacterial culture experiments*

Culture Medium	Rep Bio	OD1 (T1)	OD2 (T2)	OD3 (T3)
M9	A	0.130	0.280	0.260
M9	B	0.130	0.280	0.260
M9	C	0.130	0.328	0.260
M9F	A	0.173	0.588	0.773
M9F	B	0.208	0.627	0.834
M9F	C	0.168	0.603	0.740

A.3 Overview of single-cell RNA-seq methods in bacteria

A.4 MicroSPLiT sequencing library preparation

A.5 TSO removal statistics

The following table summarizes the number of R1 reads before and after TSO removal for each sample, as well as the corresponding percentage.

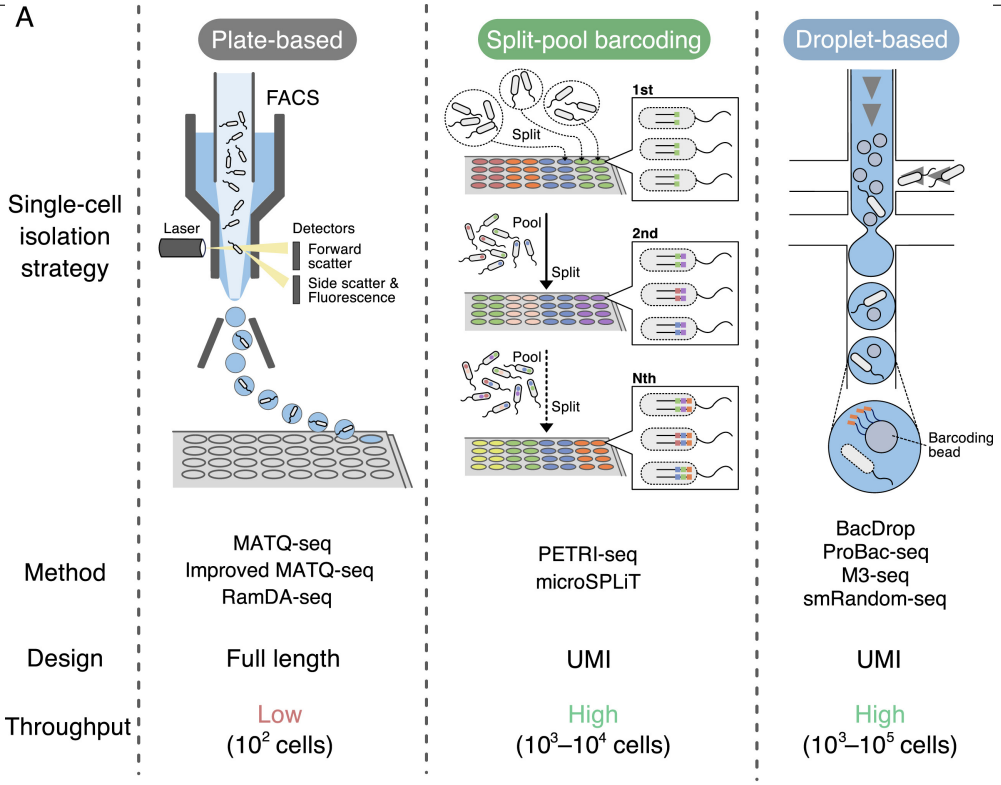
Table A.3: *TSO removal statistics for each sample. The table shows the total number of R1 reads, the number of R1 reads after TSO sequence removal, and the corresponding percentage.*

Sample	Total R1	R1 with TSO removed	Percentage (%)
BC_0076	631,393,326	153,472,373	24.3
BC_0077	325,495,590	90,142,002	27.7
BC_0079	379,108,253	102,386,519	27.0
BC_0080	397,654,767	99,430,515	25.0

// Table added to summarize TSO removal efficiency for each sample. // ... existing code ...

A.6 Trimming pipeline steps

The following steps were performed sequentially for read trimming, as implemented in the custom pipeline (see process_sample.sh). Each step is performed in paired-end mode to maintain synchronization between R1 and R2 files.



B

Method	Year	Features										
		Single-cell sorting strategy			cDNA generation & amplification		rRNA depeletion			Single-cell barcoding		
		FACS	Split-pool	Droplet	in situ generation	Random displacement amplification	RNaseH	Cas9	Other (Customed probe)	Split-pool	Droplet	Hybrid
PETRI-seq	2020		•		•					•		
MATQ-seq	2020	•										
microSPLiT	2021		•		•					•		
BacDrop	2023			•	•		•				•	•
ProBac-seq	2023			•	•				•		•	
M3-seq	2023			•	•		•				•	•
Improved MATQ-seq	2023	•						•				
smRandom-seq	2023			•	•			•			•	
RamDA-seq	2023	•				•		•				

Figure A.1: *Overview of bacterial single-cell RNA sequencing approaches. (A) Schematic summary of single-cell isolation strategies employed in bacterial single-cell RNA-seq, highlighting the key features and distinctions of each approach. FACS, fluorescent activated cell sorting; UMI, unique molecular identifier. (B) Summary of features of each bacterial single-cell RNA-seq method, including MATQ-seq, RamDA-seq, PETRI-seq, microSPLiT, BacDrop, ProBac-seq, M3-seq, and smRandom-seq*[6]

1. **TSO trimming (Cutadapt):**

   Removal of template-switching oligo (TSO) sequences from R1 using Cutadapt. This step targets TSO sequences at the 5' end of cDNA reads to eliminate technical artifacts.

```
cutadapt -j ${SLURM_CPUS_PER_TASK} \
    -g "AAGCAGTGGTATCAACGCAGAGTGAATGGG; min_overlap=6; max_errors=0.2" \
    -g "CAGAGTGAATGGG; min_overlap=6; max_errors=0.2" \
    --pair-filter=both \
```

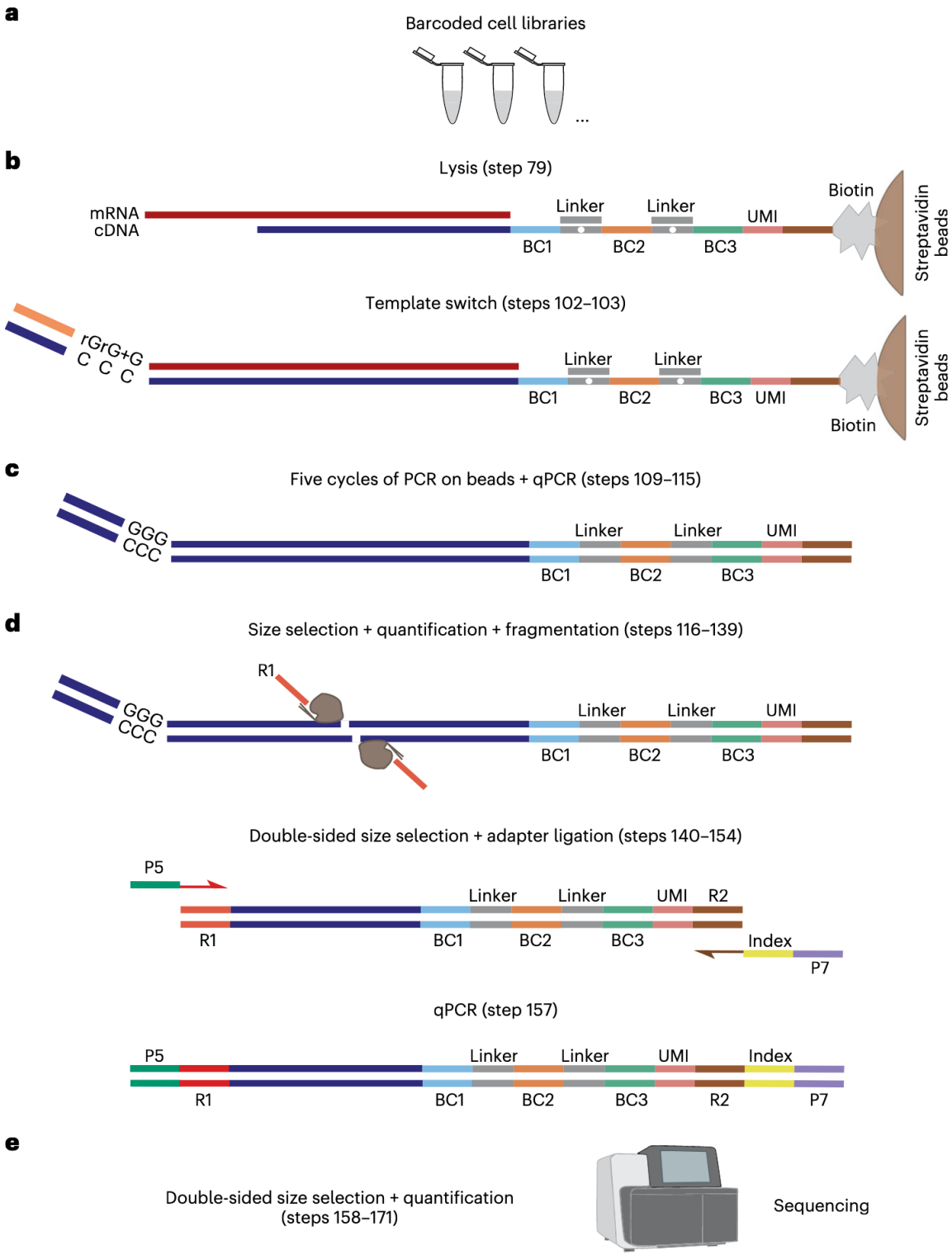


Figure A.2: *MicroSPLiT sequencing library preparation. a, Selected sub-libraries with barcoded cells are lysed. Because cDNA molecules primed with both random hexamer and poly-dT primers undergo the same downstream reactions, only one of them is shown for clarity. b, After lysis, cDNA is purified via streptavidin beads. The cells then undergo an additional RT and template switching step. The template switch primer has two RNA G bases and a locked nucleic acid G base ('rGrG+G') sequence to facilitate the binding. c, cDNA is amplified, and size is selected to eliminate the unwanted short product ('dimer') from the cDNA amplification product. At this point, the size and concentration of the cDNA product are quantified (Part 2, Step 127). d, The library then undergoes fragmentation and adapter ligation. The desired sequencing product containing the barcodes is amplified with the primers for both the third barcode adapter and the ligation adapter, which contain Read 1 (R1) and Read 2 (R2) sequences. Illumina P5 and P7 sequence adapters and a final sub-library index are also appended at this final PCR step . e, A 0.5–0.7× double size selection then selects out unwanted fragments. The final product's concentration and size are measured before sequencing .*

```
    -m 20: \
    --too-short-output
     ↪  "${output_dir}/${sample_name}_R1_too_short.fastq.gz" \
    --too-short-paired-output
     ↪  "${output_dir}/${sample_name}_R2_too_short.fastq.gz" \
    -o "${r1_output}" \
    -p "${r2_output}" \
    "${r1_input}" "${r2_input}" \
    --report=full \
    --json "${output_dir}/${sample_name}_stats.json"
```

2. **Initial quality and adapter trimming (Fastp):**

   Removal of low-quality bases, polyG/polyX tails, and adapter sequences using Fastp. This step also removes the TruSeq Read 2 adapter and I7 adapter at the end of R1 if present.

```
fastp \
    -i "${r1_input}" \
    -I "${r2_input}" \
    -o "${r1_output}" \
    -O "${r2_output}" \
    --html "${output_dir}/${sample_name}_report.html" \
    --json "${output_dir}/${sample_name}_report.json" \
    --report_title "microSplit Initial Fastp Report - ${sample_name}" \
    --compression 4 \
    --verbose \
    --unpaired1 "${unpaired1}" \
    --unpaired2 "${unpaired2}" \
    --length_required 91 \
    --dont_overwrite \
    --trim_front1 0 \
    --trim_front2 0 \
    --trim_tail1 0 \
    --trim_tail2 0 \
    --trim_poly_g \
```

```
    --poly_g_min_len 10 \
    --trim_poly_x \
    --poly_x_min_len 12 \
    --detect_adapter_for_pe \
    --adapter_sequence=ATCTCGTATGCCGTCTTCTGCTTGA \
    --adapter_sequence=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
```

3. **PolyA trimming (Cutadapt):**

   Removal of polyA stretches (>=12 nt) and all downstream sequences from R1 using Cutadapt, targeting polyA sequences introduced during library preparation. This step cleans reads with short cDNA that extend into the R2 complementary region, using polyA as a repeat sequence (read_polyA from the library).

```
cutadapt -j ${SLURM_CPUS_PER_TASK} \
    -a "A{12}; min_overlap=12; max_errors=0.2" \
    --pair-filter=both \
    -m 20: \
    --too-short-output
     ↪  "${output_dir}/${sample_name}_R1_too_short.fastq.gz" \
    --too-short-paired-output
     ↪  "${output_dir}/${sample_name}_R2_too_short.fastq.gz" \
    -o "${r1_output}" \
    -p "${r2_output}" \
    "${r1_input}" "${r2_input}" \
    --report=full \
    --json "${output_dir}/${sample_name}_stats.json"
```

   This step trims polyA15 and longer stretches that may remain after the previous steps.

4. **Specific adapter trimming (Cutadapt):**

   Removal of the specific adapter sequence CCACAGTCTCAAGCAC from R1 using Cutadapt (corresponds to the round 2 linker sequence). This step uses the round 2 linker barcode as a reference point and eliminates everything behind it, particularly useful for cleaning random hexamer sequences with short cDNA that extend into R2 complementary sequences.

```
cutadapt -j ${SLURM_CPUS_PER_TASK} \
    -a "CCACAGTCTCAAGCAC; min_overlap=6; max_errors=0.1" \
    --pair-filter=both \
    -m 20: \
    --too-short-output
     ↪  "${output_dir}/${sample_name}_R1_too_short.fastq.gz" \
    --too-short-paired-output
     ↪  "${output_dir}/${sample_name}_R2_too_short.fastq.gz" \
    -o "${r1_output}" \
    -p "${r2_output}" \
    "${r1_input}" "${r2_input}" \
    --report=full \
    --json "${output_dir}/${sample_name}_stats.json"
```

5. **Linker and additional adapter trimming (Cutadapt):**

   Removal of linker and additional adapter sequences from R1 using Cutadapt, to further clean the reads. This includes TruSeq Read 2 adapter (AGATCGGAAGAGCACACGTCTGAACTCCAGTCA), Round 3 linker (AGTCGTACGCCGATGCGAAACATCGGCCAC), and Round 2 linker (CCACAGTCTCAAGCACGTGGAT).

   This step ensures that any remaining linker or adapter sequences are removed for certain libraries.

```
cutadapt -j ${SLURM_CPUS_PER_TASK} \
    -a "CCACAGTCTCAAGCACGTGGAT; min_overlap=6; max_errors=0.2" \
    -a "AGTCGTACGCCGATGCGAAACATCGGCCAC; min_overlap=6; max_errors=0.2" \
    -a "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA; min_overlap=6;
     ↪  max_errors=0.2" \
    --pair-filter=both \
    -m 20: \
    --too-short-output
     ↪  "${output_dir}/${sample_name}_R1_too_short.fastq.gz" \
    --too-short-paired-output
     ↪  "${output_dir}/${sample_name}_R2_too_short.fastq.gz" \
    -o "${r1_output}" \
```

```
    -p "${r2_output}" \
    "${r1_input}" "${r2_input}" \
    --report=full \
    --json "${output_dir}/${sample_name}_stats.json"
```

6. **Final quality and length filtering (Fastp):**

   Final trimming with Fastp, including additional adapter removal, trimming of fixed bases from the 5' and 3' ends, and filtering for minimum read length to ensure high-quality output for downstream analysis.

   This step trims R1 at both 5' and 3' ends to keep only cDNA and ensure clean sequences for downstream analysis.

```
fastp \
    -i "${r1_input}" \
    -I "${r2_input}" \
    -o "${r1_output}" \
    -O "${r2_output}" \
    --trim_front1 10 \
    --trim_front2 0 \
    --trim_tail1 16 \
    --trim_tail2 0 \
    --length_required 25 \
    --detect_adapter_for_pe \
    --adapter_sequence=AAGCAGTGGTATCAACGCAGAGTGAATGGG \
    --adapter_sequence=CCACAGTCTCAAGCACGTGGAT \
    --adapter_sequence=AGTCGTACGCCGATGCGAAACATCGGCCAC \
    --adapter_sequence=AGATCGGAAGAGCACACGTCTGAACTCCAGTCA \
    --html "${output_dir}/${sample_name}_report.html" \
    --json "${output_dir}/${sample_name}_report.json" \
    --report_title "microSplit Final Fastp Report - ${sample_name}" \
    --compression 4 \
    --verbose
```

A.7 Final effective command line of STARsolo

A.7.1 Computing Environment

The STARsolo analysis was performed on the GenOuest high-performance computing cluster using the following specifications: - **Node type**: bigmem (high-memory node) - **Memory allocation**: 500GB RAM - **CPU threads**: 64 parallel threads

A.7.2 STARsolo Command Line

```
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
```

A.7.3 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` : Maximum values for splice junction detection (set to maximum since bacterial genomes lack splicing)

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.

Each step is performed in paired-end mode to ensure synchronization between R1 and R2 files. See the pipeline script for implementation details.

- 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.