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.

Nunc ac dignissim magna. Vestibulum vitae egestas elit. Proin feugiat leo quis ante condimentum, eu ornare mauris feugiat. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris cursus laoreet ex, dignissim bibendum est posuere iaculis. Suspendisse et maximus elit. In fringilla gravida ornare. Aenean id lectus pulvinar, sagittis felis nec, rutrum risus. Nam vel neque eu arcu blandit fringilla et in quam. Aliquam luctus est sit amet vestibulum eleifend. Phasellus elementum sagittis molestie. Proin tempor lorem arcu, at condimentum purus volutpat eu. Fusce et pellentesque ligula. Pellentesque id tellus at erat luctus fringilla. Suspendisse potenti.

Etiam maximus accumsan gravida. Maecenas at nunc dignissim, euismod enim ac, bibendum ipsum. Maecenas vehicula velit in nisl aliquet ultricies. Nam eget massa interdum, maximus arcu vel, pretium erat. Maecenas sit amet tempor purus, vitae aliquet nunc. Vivamus cursus urna velit, eleifend dictum magna laoreet ut. Duis eu erat mollis, blandit magna id, tincidunt ipsum. Integer massa nibh, commodo eu ex vel, venenatis efficitur ligula. Integer convallis lacus elit, maximus eleifend lacus ornare ac. Vestibulum scelerisque viverra urna id lacinia. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia curae; Aenean eget enim at diam bibendum tincidunt eu non purus. Nullam id magna ultrices, sodales metus viverra, tempus turpis.

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

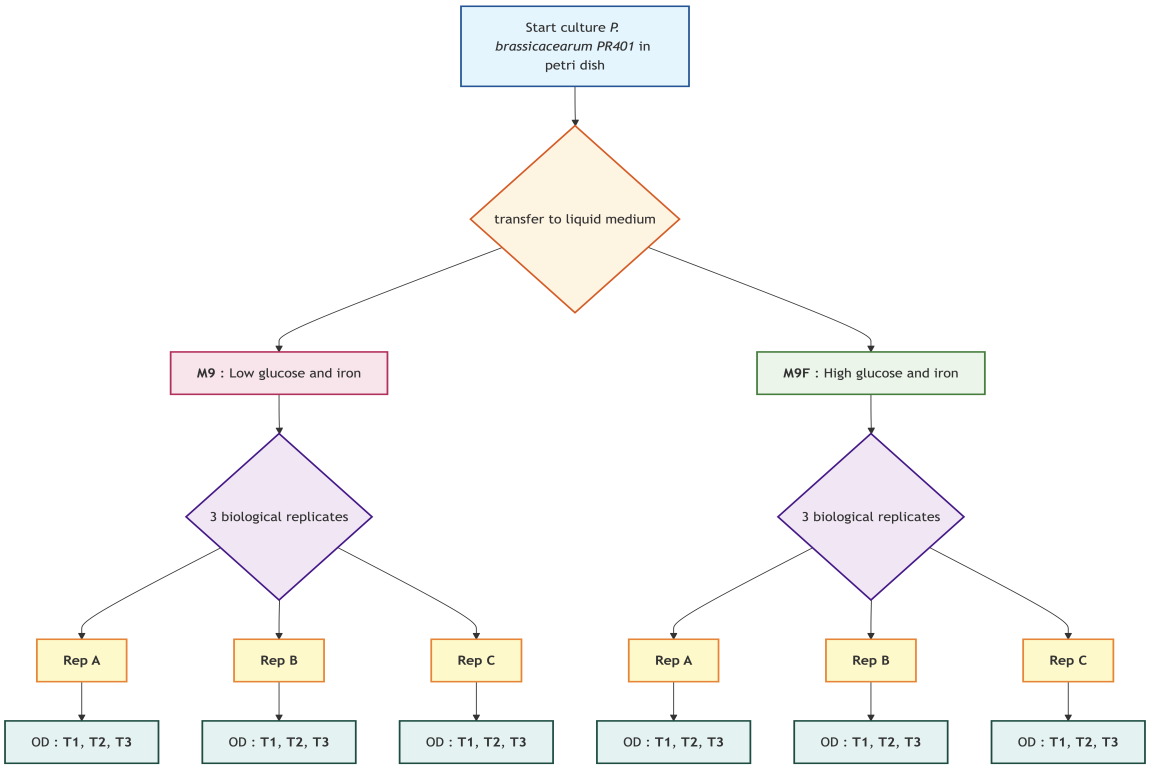

Figure 2.1: *Experimental design for bacterial culture*

An isogenic population of *P. brassicacearum* PR401 was initially cultured on petri dishes and then transferred to different liquid media to investigate the effects of nutrient availability on bacterial 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.

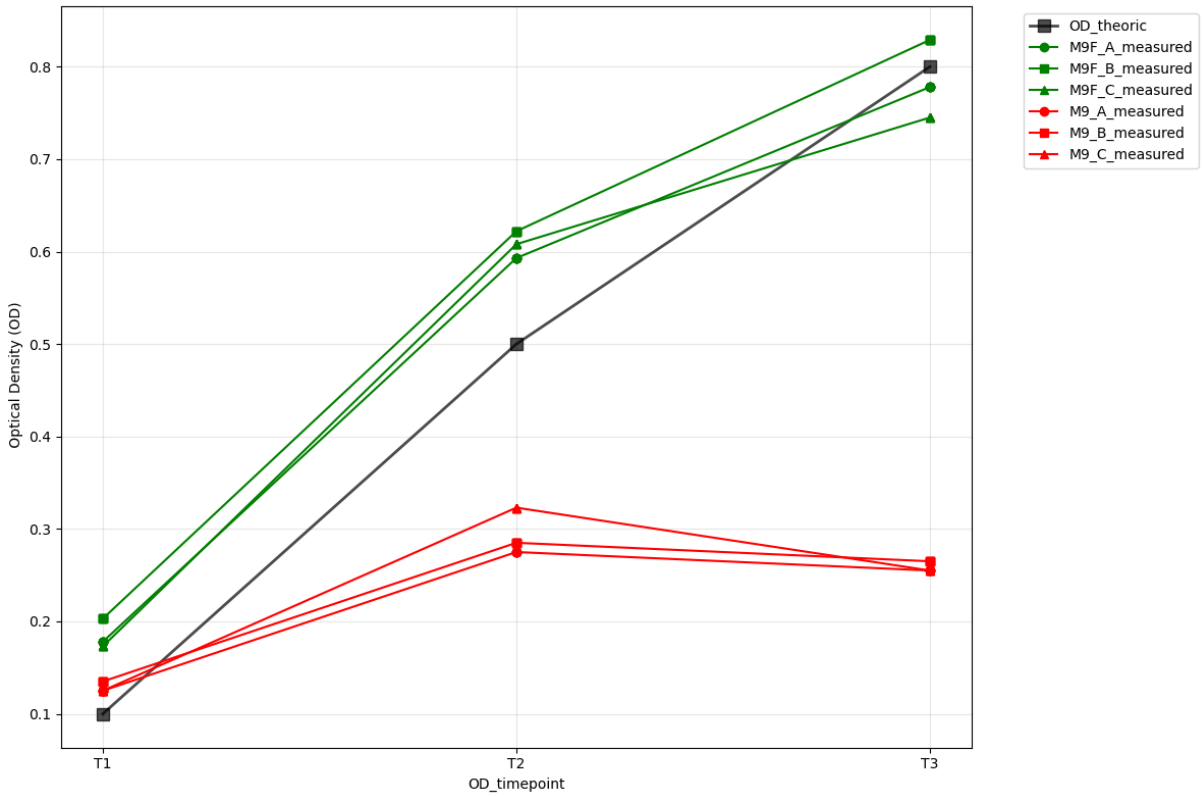

Figure 2.2: *Growth of the different bacterial populations over time*

Growth curves for each condition showing the optical density (OD) measurements over time for M9 and M9F media across three biological replicates (Rep A, B, C) at three time points (T1, T2, T3).

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 microSPLiT protocol.

2.2 microSPLiT protocol

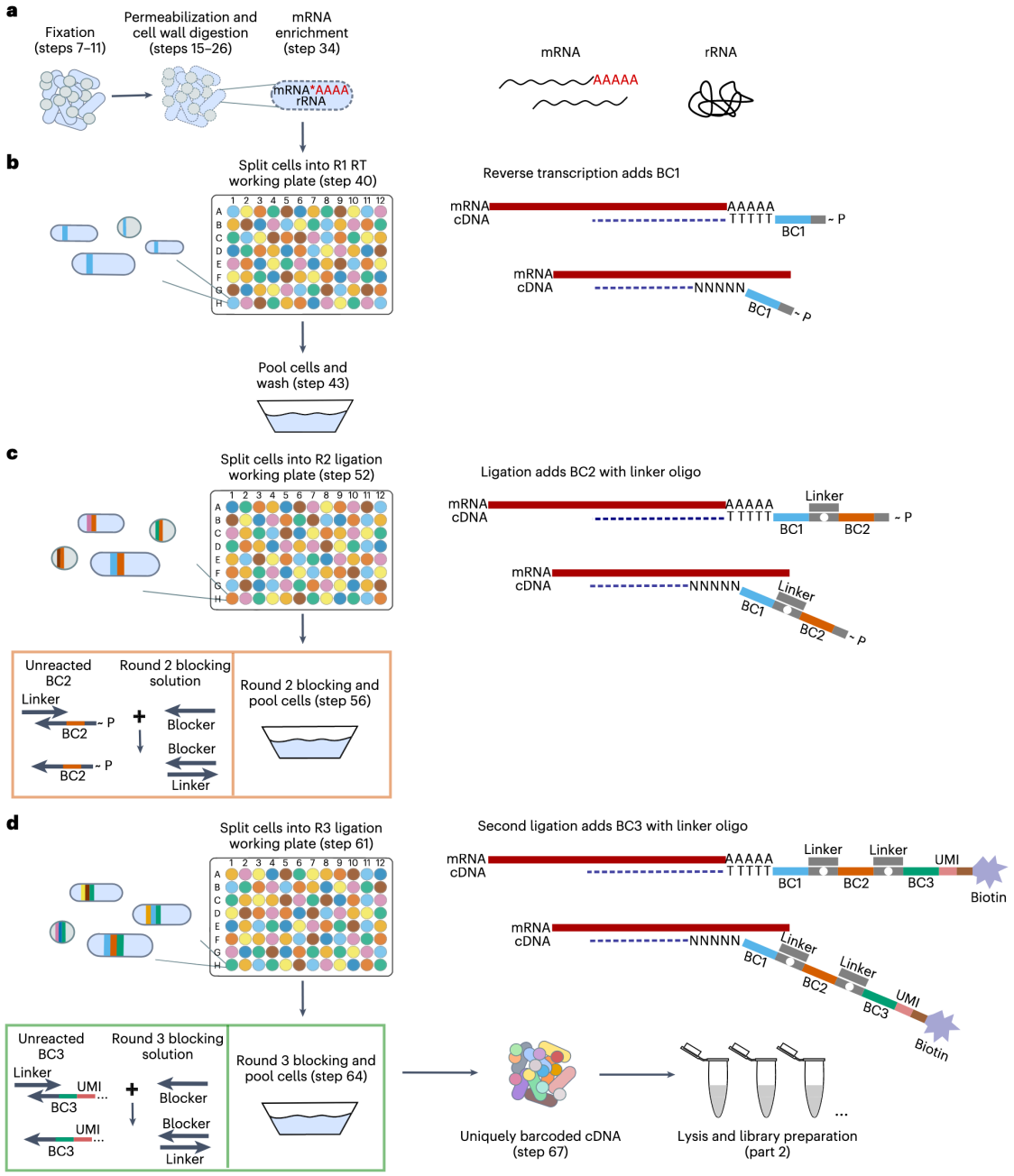

Figure 2.3: *Schematic representation of the microSPLiT protocol showing the barcoding workflow for single-cell RNA-seq.*[2]

pas fou rexpliquer en mieux et simple , juste focus sur elements importants The microSPLiT (microbial Split-Pool Ligation Transcriptomics) protocol enables single-cell RNA sequencing of bacteria by performing barcoding reactions inside individual cells rather than in separate reaction vessels .

This approach maintains single-cell resolution throughout the entire process, from cell fixation to library preparation. The protocol involves multiple 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. 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.

2.2.1 microSPLiT library and read structure

```
5'- AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX...XXX(pA)NN
3'

NNNNNNNNCCACAGTCTCAAGCACGTG
```

Legend: NNNNNNNN = Round 1 barcode (8 bp), NNNNNNNN = Round 2 barcode (8 bp), NNNNNNNNNNNNNNNNNN = Round 3 barcode + UMI (18 bp), NNNNNNN = i7 index (6 bp)

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

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).

- 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[8],

pour la qualité et contamination : centriguge et recentrifuge[9,10]

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

7.0.1 Division of Labor Mechanisms

7.0.2 Biological Significance

7.0.3 Technical Considerations

7.1 Comparison with Existing Literature

7.1.1 Similarities with Previous Studies

7.1.2 Novel Insights

7.1.3 Discrepancies and Their Implications

7.2 Methodological Strengths and Limitations

7.2.1 Technical Advantages

7.2.2 Potential Limitations

7.2.3 Future Methodological Improvements

7.3 Biological Implications

7.3.1 Ecological Significance

7.3.2 Evolutionary Perspectives

7.3.3 Potential Applications

7.4 Future Research Directions

7.4.1 Open Questions

7.4.2 Suggested Follow-up Studies

7.4.3 Technical Improvements

7.5 Conclusion

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. Rossello, M., Tandonnet, S. & Almudi, I. BarQC: Quality Control and Preprocessing for SPLiT-Seq Data. doi:10.1101/2025.02.04.635005.
9. Martí, J. M. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Computational Biology* **15**, e1006967 (2019).
10. 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

- 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.