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 2025-07-10

Table of contents

Copyright notice

Produced on 5 juillet 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.

Division of labor (DOL) represents a fundamental biological strategy enhancing collective performance through task specialization. While interspecific DOL is well-documented in microbial communities, intraspecific DOL within clonal bacterial populations remains underexplored. This study investigated whether genetically identical bacterial cells exhibit functional specialization under iron limitation, using *Pseudomonas brassicacearum* R401 as a model system.

We employed microSPLiT (microbial Split-Pool Ligation Transcriptomics) technology to perform single-cell RNA sequencing on *P. brassicacearum* R401 populations grown under contrasting iron conditions: iron-limited (M9) and iron-replete (M9F) media. Bacterial cultures were sampled at three timepoints with three biological replicates per condition, resulting in 18 experimental samples.

The microSPLiT methodology successfully generated high-quality single-cell transcriptomic data, with 85.58% of sequencing reads containing valid barcodes and detection of 6,035 genes (96.6% of the annotated genome). Following quality control and filtering, we analyzed approximately 3,000 cells. Principal component analysis revealed clear transcriptional distinctions between culture conditions, with iron-limited cells at later timepoints showing reduced transcriptional activity and distinct gene expression patterns.

Differential expression analysis identified genes associated with translational regulation, iron metabolism, and stress response as key contributors to condition-specific programs. Cells under iron limitation exhibited downregulation of ribosomal protein genes (e.g., RplA) and upregulation of storage metabolism genes (e.g., phasin) and siderophore biosynthesis regulators, suggesting adaptive metabolic reorganization under iron stress.

This study successfully established the technical foundation for bacterial single-cell transcriptomics using microSPLiT. The methodological advances and initial transcriptomic insights provide a solid basis for future investigations of bacterial population heterogeneity and potential cellular specialization under environmental constraints.

Keywords :

Single-cell RNA-seq, *Pseudomonas brassicacearum*, Division Of Labor, microSPLiT, iron limitation

Acknowledgements

I would like to thank my supervisors Solène and Philippe for their continuous support and interesting discussions throughout this internship. Their guidance and expertise were invaluable in developing my understanding of bacterial single-cell transcriptomics and division of labor concepts.

Special thanks to my fellow interns Killian and Manon for their camaraderie and mutual support during this research period.

I also extend my gratitude to the Ecobio laboratory and the ANR Divide project for providing the research framework and resources necessary for this work.

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
BacSC	Bacterial Single-Cell pipeline
BC	Barcode
BiRD	Bioinformatics Core Facility
CB	Cell Barcode
CNRS	Centre National de la Recherche Scientifique
DNA	Deoxyribonucleic Acid
DOL	Division Of Labor
FASTQ	FASTQ file format
FDR	False Discovery Rate
GenoA	Genomics Core Facility
IFB	Institut Français de Bioinformatique
M9	Minimal medium
M9F	Minimal medium with iron
MCP	Methyl-accepting Chemotaxis Protein
microSPLiT	microbial Split-Pool Ligation Transcriptomics
mRNA	messenger RNA
ncRNA	non-coding RNA
NGS	Next Generation Sequencing
OD	Optical Density
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PHA	Polyhydroxyalkanoate
PhiX	PhiX control library

Abbreviation	Definition
PsR401	*Pseudomonas brassicacearum* R401
qRT-PCR	quantitative Reverse Transcription PCR
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
scRNA-seq	single-cell RNA sequencing
STARsolo	STAR aligner for single-cell data
TSB	Tryptone Soy Broth
TSO	Template Switching Oligo
tRNA	transfer RNA
UMI	Unique Molecular Identifier
UMAP	Uniform Manifold Approximation and Projection

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Division of Labor: A Fundamental Biological Principle

The survival of organisms in evolving environments is driven by their fitness[1], where the cost-benefit ratio of traits is constantly balanced and gives rise to different populational evolutionary strategies. To succeed, organisms must compete, cooperate, and/or specialize based on how well their traits enable resource acquisition and utilization in their biotic and abiotic environment. Division of Labor (DoL) represents one such strategy for resource use optimisation. Division of labour occurs when different individuals, cells or tissues become specialised to perform complementary tasks that benefit the whole organism or social group[1] and improves collective performance[2]. This fundamental principle operates throughout biological systems, from molecular evolution where gene duplication enables enzyme specialization, to multicellular organisms where cellular differentiation creates specialized tissues, to eusocial insect societies with their reproductive and functional caste systems[3].

Among the diverse biological systems where DoL operates, microbial communities provide particularly compelling examples. Microbial communitie's dynamics demonstrate how DoL can emerge even among unicellular organisms, where individual cells can specialize in different metabolic functions while cooperating for collective benefit. Giri and colleagues have attempted to define the concept of DoL specifically within microbial communities, identifying key criteria that distinguish DoL from other types of ecological interactions[3]. Microbial interactions can be classified based on their directionality and the species involved. Interactions can be unidirectional or bidirectional, and can occur within the same species (intraspecific) or between different species (interspecific). For an interaction to qualify as DoL, it must involve reciprocal fitness benefits between partners, where both participants

[1] Fitness refers to the ability of an organism to survive and reproduce in its environment, measured by its reproductive success and contribution to future generations.

gain from the interaction. DoL between microbial species has been well-characterized in various ecosystems, with numerous examples of cross-feeding and mutualistic interactions documented in the human gut microbiome and soil communities[4].

1.2 Division of Labor in Root Microbiota

Building on this broader understanding of microbial DoL, the root microbiota represents a particularly well-studied example of interspecific DoL[5]. Roots of plants host diverse bacteria that are collectively referred to as the bacterial root microbiota. Unlike multicellular organisms that evolved diverse cell types to achieve distinct biological functions and promote a division of labour, unicellular organisms such as bacteria rely on limited metabolic specialisation possibilities as a unit. Indeed, recent reports[6,7], indicate that metabolic interdependencies and cross-feeding exchanges are widespread among taxonomically diverse bacteria and likely drive microbial co-existence within complex bacterial communities[8,9].

While interspecific DoL has been extensively studied[3], intraspecific DoL within clonal/isogenic[2] bacterial populations remains less explored, despite its potential importance for understanding population-level adaptation and functional diversity. The traditional view of biological populations assumed that all individuals within a clonal population behave identically. However, evidence has accumulated over decades showing that even genetically identical organisms can exhibit functional heterogeneity, leading to population-level benefits through task specialization[10].

A major unsolved question is whether populations of genetically identical bacteria can minimise energetically costly processes by each executing different metabolic tasks at the intra-population level. Here, we hypothesise that metabolic cooperation within bacterial populations plays a key role in modulating population dynamics, competitiveness, and persistence at the root–soil interface. This hypothesis builds on the idea that bacteria are subject to stochastic, noisy gene expression[11–13]. In such systems, not all individuals respond identically to environmental cues, leading to phenotypic heterogeneity. This noise-driven diversity can be beneficial at the population level. The theory of Noise-Averaging Cooperation (NAC)[14] further proposes that metabolic noise—exacerbated by the small size of bacteria—can constrain individual growth, but can be mitigated through metabolite sharing among related cells. This “leaky function”, forming a “metabolic marketplace”, allows populations to buffer stochastic fluctuations and improve collective fitness. In this context, cross-feeding interactions may emerge as a result of gene expression variability and/or the selection of advantageous mutations, fostering metabolic interdependencies within the population (see Figure 1.1).

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

Finally, environmental regulation also shapes this process: microenvironmental cues and spatial heterogeneity can induce context-dependent gene expression patterns[10,15], which, if consistently beneficial, may become genetically encoded over evolutionary time. Altogether, these mechanisms point toward an intra-population division of labour emerging from the interplay between gene expression noise, environmental signals, and evolutionary selection.

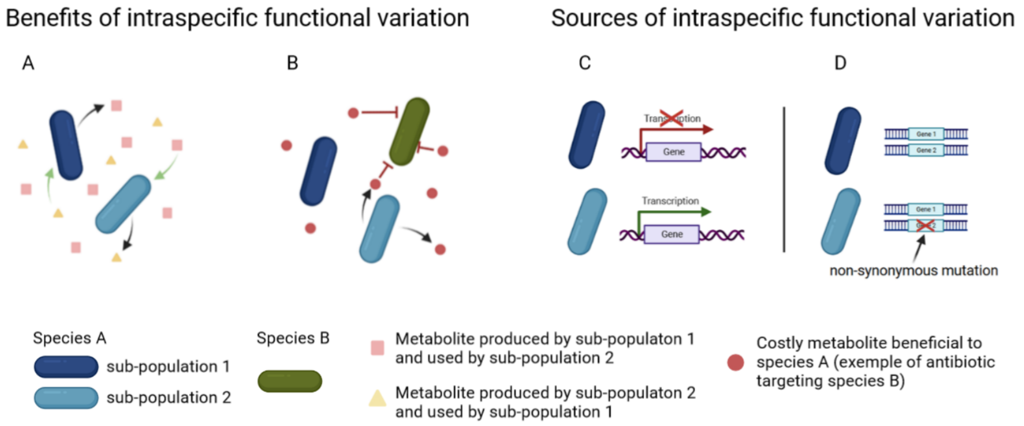

Benefits of intraspecific functional variation
Sources of intraspecific functional variation
A
B
C
D
Transcription
Gene
Transcription
Gene
Gene 1
Gene 2
Gene 1
Gene 2
non-synonymous mutation
Species A
sub-population 1
sub-population 2
Species B
Metabolite produced by sub-populaton 1 and used by sub-population 2
Metabolite produced by sub-populaton 2 and used by sub-population 1
Costly metabolite beneficial to species A (exemple of antibiotic targeting species B)

Figure 1.1: *Examples of benefits and sources of intraspecific division of labour between intra-populations based on metabolic complementation and costly "common good" metabolite production.*

From the left to the right, co-metabolism within an isogenic bacterial population (A) and production of antimicrobials targeting other bacteria (B) can be considered as division of labour in a population. These mechanisms of division of labour can be based on differential transcription between bacteria (C) within a population and/or genetic variations between intra-populations (D).

1.3 *PsR401*: A Model System for Studying Intraspecific DoL

To investigate these theoretical frameworks of intraspecific DoL, we require a well-characterized bacterial model system that exhibits the complex metabolic interactions described above. The root environment provides an ideal context for studying such interactions, as successful bacterial establishment at roots requires the coordination of multiple independent biological processes. These include both host-microbe interactions (signal recognition, chemotaxis, surface attachment, biofilm formation, virulence factors)[10] and microbe-microbe interactions (production of antimicrobials or public goods)[16]. Given the energetic costs associated with simultaneously activating these diverse processes, we postulate that cooperation between genetically identical strains is key for promoting bacterial pervasiveness at roots.

Consistent with this hypothesis, the robust root coloniser *Pseudomonas brassicacearum R401* (hereafter referred to as *PsR401*) provides an excellent model system for studying intraspecific DoL. This bacterium deploys multiple independent strategies that co-function to promote colonisation and persistence at roots[17]. Unlike many pathogenic bacteria, *PsR401* lacks genes for a type III secretion system (T3SS) and does not overgrow or suppress plant immune responses[18,19]. Instead, this opportunistic pathogen[20] of the plant model *Arabidopsis thaliana* acts as a potent antagonist that relies on the combined action of three distinct exometabolites to suppress competitors and promote root colonization.

The three key exometabolites produced by *PsR401* each serve distinct but complementary functions. First, this Gram-negative bacterium produces Brassicapeptin[20,21], a phytotoxin that promotes both pathogenicity and root colonization in mono-association experiments with *Arabidopsis thaliana*. Second, it synthesizes 2,4-diacetylphloroglucinol (DAPG)[17], an antimicrobial compound that directly inhibits competing microbes. Third, it produces pyoverdine[17], a siderophore that chelates iron from the environment—an essential but scarce micronutrient in the rhizosphere[22].

The production of pyoverdine highlights the central role of iron availability in microbial interactions at the root–soil interface. Iron functions as a major micronutrient that modulates strain competitiveness and proliferation at roots[23,24]. In iron-limited environments, siderophore-mediated iron scavenging confers a strong competitive advantage by depriving rival microbes of access to this vital resource. This resource competition, especially for iron, represents a key mechanism of indirect microbial antagonism. Beyond simply acquiring nutrients, microbes may also sequester them, preventing uptake by others and modulating community composition and function[25].

Biological Hypothesis

Given that iron functions as a public good whose availability becomes rate-limiting in the root compartment and that production of the above-mentioned processes are all modulated by iron availability,[26] we propose that division of labour (DoL) among genetically identical *PsR401* cells may be reinforced under iron-limiting conditions, such as those found in the root habitat. In such scenarios, phenotypic heterogeneity—whether driven by gene expression noise or environmentally induced regulation—could lead to subpopulations specialising in complementary tasks, such as toxin production, antimicrobial defense, or siderophore-mediated iron acquisition, thereby enhancing population-level fitness.

1.4 Leveraging Single-Cell Transcriptomics to Study DoL

To test our biological hypothesis regarding intraspecific DoL in *PsR401*, we need to examine transcriptional heterogeneity at the single-cell level. Traditionally, studies of bacterial gene expression have relied on bulk RNA sequencing methods, which provide an average view of the transcriptome across a population.[27] However, these approaches mask the underlying cell-to-cell variability that is critical for understanding complex bacterial behaviors and adaptations. The advent of single-cell RNA sequencing (scRNA-seq) and now multi-omics technologies has provided unprecedented insights into cellular heterogeneity across various biological systems[28–30]. Although eukaryotic cells have benefited from scRNA-seq technology since 2009, prokaryotic systems have faced significant implementation delays owing to distinct technical obstacles. These challenges include low RNA content in individual cells, the absence of poly-A tails on bacterial mRNAs, and diverse cell wall structures.[27] These factors necessitate the development of specialized techniques for efficient cell lysis, RNA extraction, and mRNA enrichment in bacterial systems. Despite these challenges, recent years have seen significant progress in developing and refining bacterial scRNA-seq methods[27,31]. These methodological improvements have created novel opportunities to explore bacterial physiology, stress responses, and population dynamics at single-cell resolution.

1.5 Research Objectives and Approach

1.5.1 Our Focus: microSPLiT Technology

To address our biological hypothesis regarding intraspecific DoL in *PsR401*, we will leverage the microSPLiT (microbial Split-Pool Ligation Transcriptomics) technology[32,33]. This cutting-edge bacterial scRNA-seq method enables high-throughput profiling of individual bacterial cells, providing the resolution necessary to detect transcriptional heterogeneity within clonal populations. By analyzing *PsR401* cells under contrasting nutrient conditions—particularly iron-limited versus iron-replete environments—we hope to identify potential subpopulations that specialize in different metabolic tasks (see Figure 1.2).

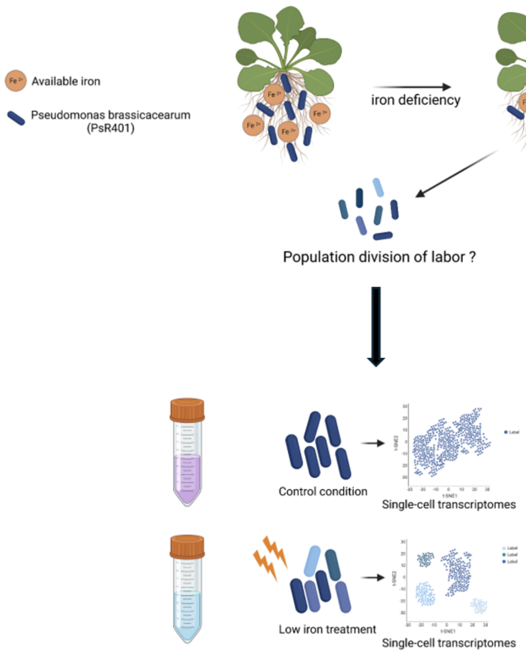

Available iron
Pseudomonas brassicacearum (PsR401)
iron deficiency
Population division of labor ?
Control condition
Single-cell transcriptomes
Low iron treatment
Single-cell transcriptomes

Figure 1.2: *Hypothesis of DoL in* PsR401 *under iron-limiting conditions*

Testing the division of labor hypothesis in PsR401 under two contrasting conditions: iron-limited and iron-replete environments. Single-cell RNA sequencing (scRNA-seq) with microSPLiT will enable detection of transcriptional heterogeneity within the population, revealing whether this specialization is driven by noisy gene expression or environmentally induced metabolic specialization.

1.5.2 Internship Goals and Expected Outcomes

Building upon the theoretical framework of intraspecific division of labor and the biological characteristics of *PsR401*, this internship aims to leverage cutting-edge single-cell transcriptomics to investigate metabolic cooperation within clonal bacterial populations. The primary goal is to validate the microSPLiT technology for bacterial systems while testing our hypothesis that iron limitation promotes functional specialization among genetically identical cells.

Through systematic analysis of transcriptional heterogeneity under contrasting iron conditions in culture medium, we expect to uncover whether *PsR401* populations exhibit distinct subpopulations with specialized metabolic functions or whether cooperation emerges through noisy gene expression regulation across the population. By examining temporal dynamics of gene expression patterns, we will gain insights into how these cooperative behaviors evolve and stabilize over time. This investigation will provide critical insights into the mechanisms underlying population-level adaptation and resilience, particularly in the context of fluctuating iron availability.

The expected outcomes of this research extend beyond understanding *PsR401* biology. Ultimately, this project will advance our understanding of how phenotypic diversity among clonal bacterial populations facilitates ecological success and resilience during root colonization, while establishing methodological foundations for future studies of bacterial population dynamics at single-cell resolution. By establishing robust methodologies for bacterial single-cell transcriptomics, this work will contribute to the broader field of microbial population dynamics.

Chapter 2

Materials and Methods

2.1 Bacterial culture

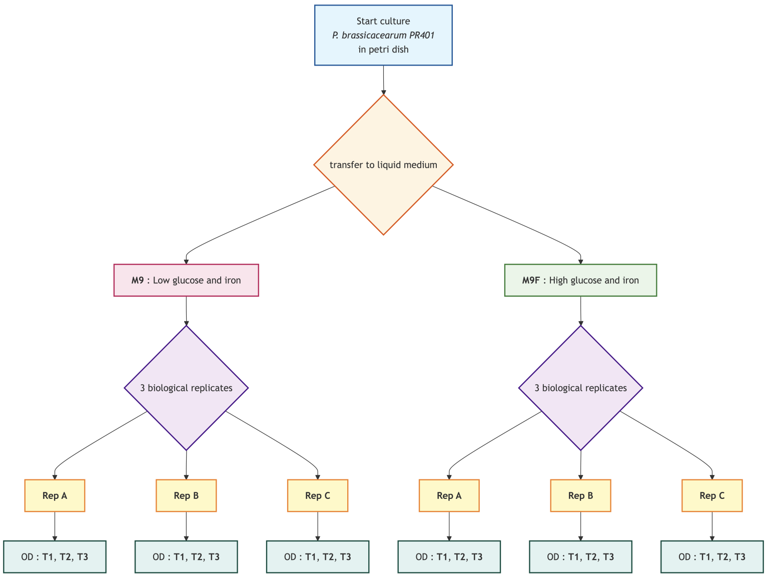


Figure 2.1: *Experimental design for bacterial culture*

The workflow illustrates the P. brassicacearum R401 cells preparation in two different media: i) a first stringent medium (M9) with low glucose availability and no iron supplementation) and ii) a less stressful medium (M9F) with regular glucose and iron supplementation. Each medium condition was replicated three times (Rep A, B, C) and bacterial growth was monitored regularly via optical density measured at 600nm. Bacterial cells were sampled three specific timepoints (T1, T2, T3) for each medium. This design resulted in 18 samples (2 media × 3 biological replicates × 3 timepoints) for subsequent single-cell RNA-seq library preparation and analysis.

An isogenic population of *P. brassicacearum R401* was initially cultured in a rich medium (Tryptone Soy Broth; "TSB") before being 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:a stringent medium containing low glucose and no iron supplementation (M9), and a medium containing regular glucose concentration (20 mM) and high FeCl3 concentrations (100 µM) (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 (OD600nm) at regular intervals. The growth curves obtained from these measurements are presented (Figure 2.2). This experimental design resulted in a total of 18 cell samples: 2 media types × 3 biological replicates × 3 time points, providing comprehensive coverage of the growth dynamics under different nutrient conditions.

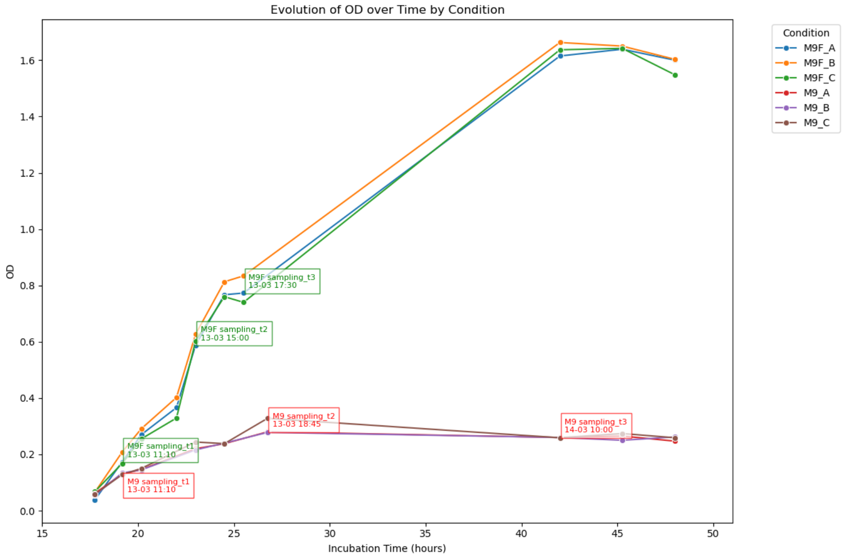


Figure 2.2: *Bacterial growth dynamics of* P. brassicacearum *R401 populations measured by optical density (OD600) cultured under two different nutrient conditions: M9 (low glucose/iron) and M9F (high glucose/iron).*

Measurements were taken at three timepoints (T1, T2, T3) for three biological replicates (Rep A, B, C) for both culture media (M9 and M9F). T1 timepoints are identical for both conditions, while T2 and T3 timepoints differ between M9 and M9F culture conditions.

The growth curves reveal distinct patterns between the two culture conditions. Both M9 and M9F cultures were sampled at the same initial timepoint (T1), showing similar optical densities (OD 0.13-0.21). However, subsequent sampling timepoints (T2 and T3) were selected based on the specific growth dynamics of each condition. M9F cultures exhibited significantly higher growth rates and reached higher optical densities (OD 0.59-0.63 at T2, 0.74-0.83 at T3), while M9 cultures showed limited growth with lower densities (OD 0.28-0.33 at T2, 0.26 at T3). The different timepoints reflect the distinct growth kinetics: M9F cultures showed continued active growth from T2 to T3, maintaining exponential growth phase, while M9 cultures reached a growth plateau by T2, suggesting nutrient limitation in the M9 condition. Biological replicates showed consistent results, validating the reproducibility of this growth pattern.Cells were collected at each timepoint (T1, T2, T3) from all biological replicates for subsequent single-cell RNA-seq library preparation using the Microbial split-pool ligation transcriptomics (microSPLiT) protocol[32,33].

2.2 microSPLiT protocol

MicroSPLiT[32,33] is a high-throughput single-cell RNA sequencing plate-based method for bacteria, allowing the profiling of hundreds of thousands of cells' transcriptional states per experiment without the need for specialized equipment[27,33]. Unlike other single-cell RNA-seq approaches that require physical isolation of individual cells (e.g., droplet-based methods), microSPLiT uses a split-pool barcoding strategy to uniquely label transcripts within each cell.

Information

The microSPLiT strategy will not be described in detail here; for more information, see Gaisser protocol.[33]. Only the key steps necessary for a general understanding of the method are presented below.

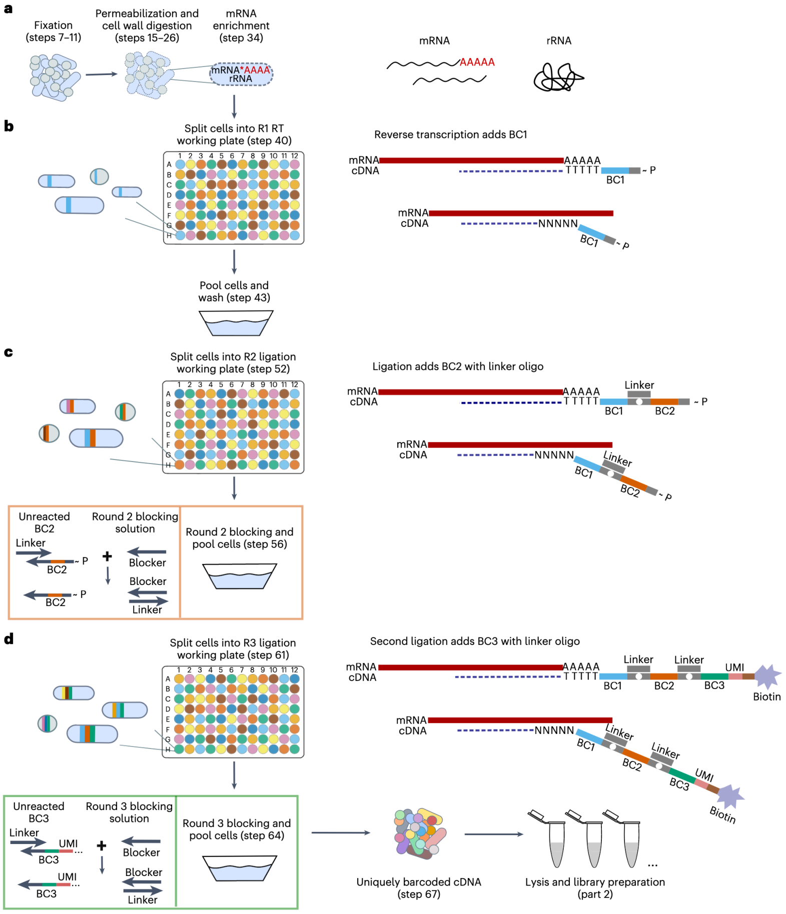

a
Fixation (steps 7–11)
Permeabilization and cell wall digestion (steps 15–26)
mRNA enrichment (step 34)
mRNA AAAA
rRNA
mRNA
AAAAA
rRNA
b
Split cells into R1 RT working plate (step 40)
Reverse transcription adds BC1
mRNA
cDNA
AAAAA
TTTTT
BC1
P
NNNNN
Pool cells and wash (step 43)
c
Split cells into R2 ligation working plate (step 52)
Ligation adds BC2 with linker oligo
Linker
BC1
BC2
Unreacted BC2
Round 2 blocking solution
Blocker
Round 2 blocking and pool cells (step 56)
d
Split cells into R3 ligation working plate (step 61)
Second ligation adds BC3 with linker oligo
BC3
UMI
Biotin
Unreacted BC3
Round 3 blocking solution
Round 3 blocking and pool cells (step 64)
Uniquely barcoded cDNA (step 67)
Lysis and library preparation (part 2)

Figure 2.3: *MicroSPLiT in-cell cDNA barcoding scheme (from Gaisser et al. 2024)*[33]

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)[33].

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.

> ℹ 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 ribosomal RNA (rRNA) Figure 2.3.

2.2.3 Barcoding

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, and distributed into individual wells of a 96-well plate containing uniquely barcoded oligos Figure 2.3. In each well, mRNA is reverse transcribed into cDNA using a mix of poly(T) and random hexamer oligos with the same barcode. The oligos used in each well contain either a dT15 sequence to capture polyadenylated mRNA previously enriched or six random nucleotides to bind any RNA, followed by a universal sequence for subsequent ligation steps. All cells in the same well receive the same unique barcode during reverse transcription, which allows sample identification 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.

> **i** 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[1] 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 $90 \times 96 \times 95 =$ **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. A sub-library containing approximately 3,000 cells was 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[33].

After cell lysis and cDNA purification on streptavidin beads, 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.

[1] 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.

To increase index diversity necessary for the well-being of the sequencing, 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), and BC_0079 (GGCTAC). These indices 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$^{\text{TM}}$ X plus instrument at GenoA 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[2]. 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 of interest (transcriptome), while R2 files contain the cell barcodes (from the three split-pool rounds) and unique molecular identifiers (UMIs).

2.3 Pipeline for microSPLiT data processing

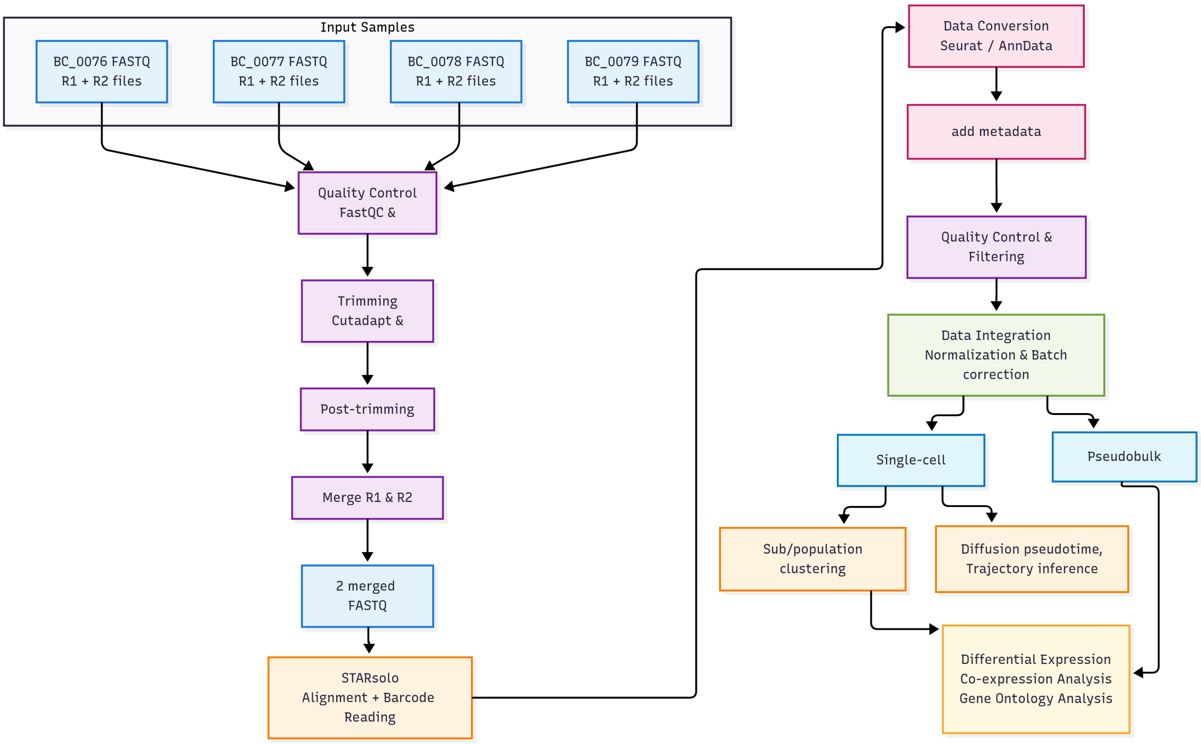

Input Samples
BC_0076 FASTQ R1 + R2 files
BC_0077 FASTQ R1 + R2 files
BC_0078 FASTQ R1 + R2 files
BC_0079 FASTQ R1 + R2 files
Quality Control FastQC &
Trimming Cutadapt &
Post-trimming
Merge R1 & R2
2 merged FASTQ
STARsolo Alignment + Barcode Reading
Data Conversion Seurat / AnnData
add metadata
Quality Control & Filtering
Data Integration Normalization & Batch correction
Single-cell
Pseudobulk
Sub/population clustering
Diffusion pseudotime, Trajectory inference
Differential Expression Co-expression Analysis Gene Ontology Analysis

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

[2] 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.

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 (see left part of Figure 2.4).

Quality control and trimming

Read quality was initially assessed for all four libraries (R1 and R2) using FastQC[34] and MultiQC[35]. Trimming was then performed with Cutadapt[36] and Fastp[37] 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 conserved. The detailed trimming pipeline is described in Appendix Section Section A.2.

Quality control and file merging

After trimming, read quality was reassessed with FastQC[34] and MultiQC[35] 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, duplication rates, quality scores (Q30), adapter content, and other relevant metrics. Following quality control, the files from all four libraries (R1 and R2 for each index) were merged into unique files (R1 and R2), ensuring 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[38,39], 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[40]. The implementation followed the recommendations outlined in Gaisser et al., 2024[33] for optimal microSPLiT data processing. Complete pipeline scripts and parameters are detailed in Appendix Section Section A.3.

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[41] 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 of gene counts for each cell (N-by-K matrix, with N cells and K genes) using `GeneFull` feature counting and `UniqueAndMult-Uniform` mapping strategy (which 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[33], 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 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[42] in R and AnnData objects[43] for use with Scanpy[44]. 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.

Quality control and filtering.

A multi-step filtering strategy was implemented to ensure data quality and remove technical artifacts while preserving biological variation. The filtering pipeline was designed to address specific challenges of microSPLiT data and experimental design considerations.

An initial UMI-based filtering was performed to remove cells with fewer than 100 unique molecular identifiers (UMIs) from the analysis Figure A.1. This threshold was chosen based on preliminary analysis showing clear differences in UMI distributions between the two culture conditions (M9 vs M9F), while also serving as a quality control metric to identify potentially failed technical replicates. To identify potential doublets, the top-performing cells from each technical replicate were first filtered based on UMI counts. Doublets are technical artifacts that occur when two or more cells are incorrectly assigned the same barcode combination, resulting in a mixed transcriptomic profile that can confound single-cell analysis. The distribution of UMI counts per cell was examined within these high-quality cells to detect outliers that deviated significantly from the expected distribution, which typically represent potential doublets or technical artifacts (visual inspection of the distribution of UMI counts per cell). A comprehensive list of these putative doublets was compiled and subsequently removed from the initial dataset to ensure data quality.

To ensure robust representation of each experimental condition, filtering was applied at the biological replicate level to obtain approximately 165 cells per condition, representing a total of 3000 cells. For each biological replicate at each optical density timepoint (T1, T2, T3) for both culture conditions (M9, M9F), only the most deeply sequenced cells were retained. This approach allowed for the selection of the best-performing technical replicates. Then, in each of the 18 conditions, technical replicates containing fewer than 5 cells were removed.Transcript type filtering was applied to retain only mRNA transcripts for analysis, filtering out ribosomal RNA (rRNA), transfer RNA (tRNA), and other non-coding RNA species. Gene expression filtering was performed to remove genes expressed in fewer than 5 cells across the entire dataset, eliminating low-quality or spurious gene detection events and focusing on robustly detected transcripts.

2.3.3 Single-cell analysis

For this first approach to single-cell RNA-seq analysis, the complete dataset with all conditions pooled together was analyzed, rather than focusing on specific sub-conditions (Culture Medium × Biological Replicate × Sampling Time). The primary objective was to validate that the dataset contains sufficient signal to distinguish between culture media conditions at a global level before conducting more detailed condition-specific analyses. While the main biological question is to investigate Division of Labor (DoL) within the bacterial population, this approach did not directly address this question but served as a foundation for understanding data quality and taking a broad exploratory view of the data.

General analytical approach

The analysis workflow encompassed several key steps: Normalization and scaling, feature selection, dimensionality reduction, clustering, and differential expression analysis. We systematically tested different parameters and methodologies to optimize each step of the pipeline. We followed the Scanpy tutorial for the general approach and implemented the BacSC protocol (preprint[45]), a computational pipeline designed to limit methodological biases in bacterial single-cell analysis. We tested BacSC specifically for normalization, scaling and feature selection, and its output was then used for subsequent dimensionality reduction and clustering steps.

Normalization, scaling and feature selection. We implemented the BacSC protocol (currently a preprint[45]), a fully data-driven computational pipeline based on the Python version of Seurat's SCTransform tool. SCTransform is an advanced normalization method that replaces the classical steps of normalization + log-transformation + variable gene selection. It performs variance stabilizing transformation (VST) to correct for differences in sequencing depth between cells, stabilizes variance across genes, and automatically identifies the most informative genes for biological variation. This automated pipeline selects optimal features for normalization and feature selection without requiring manual intervention.

Principal component analysis (PCA). PCA was computed using `sc.pp.pca`, with the number of principal components systematically tested at multiple levels (i.e. 3, 10, 20, 50) depending on the data characteristics. The optimal number of components was determined using Scanpy's `sc.pl.pca_variance_ratio` plot, which displays the explained variance ratio for each principal component, allowing us to identify the elbow point where additional components provide diminishing returns in variance explanation.

Clustering and visualization. For our analysis that will be presented, we constructed the neighborhood graph using the first 5 principal components (PCs) with 10 neighbors per cell, capturing local similarity between cells in the reduced dimensional space. This approach is essential for downstream clustering and visualization analyses. We then applied Leiden clustering with two different resolution parameters (0.1 and 0.25) to identify distinct cell populations. Uniform manifold approximation and projection (UMAP) was computed using `sc.tl.umap` for visualization, with parameters optimized for bacterial single-cell data. For comparison and validation purposes, we also performed the same analytical workflow using Seurat v5[42], which yielded similar results and performance, confirming the robustness of our analytical approach across different computational frameworks.

Differential expression analysis

Differential gene expression analysis was performed using Scanpy's gene ranking functions (`sc.tl.rank_genes_groups` and `sc.get.rank_genes_groups_df`) to compare expression patterns between experimental conditions. Statistical outputs included gene names, z-scores, log fold changes, p-values, and adjusted p-values for robust identification of differentially expressed genes.

2.3.4 Analytical scope and limitations

As mentioned in the pipeline description (Figure 2.4), this study focuses on the fundamental analytical steps: data quality assessment, basic clustering and visualization, and initial differential expression analysis between culture conditions. The more advanced analytical approaches such as pseudobulk analysis, trajectory inference, gene co-expression networks, and gene ontology enrichment were not pursued at this exploratory stage.

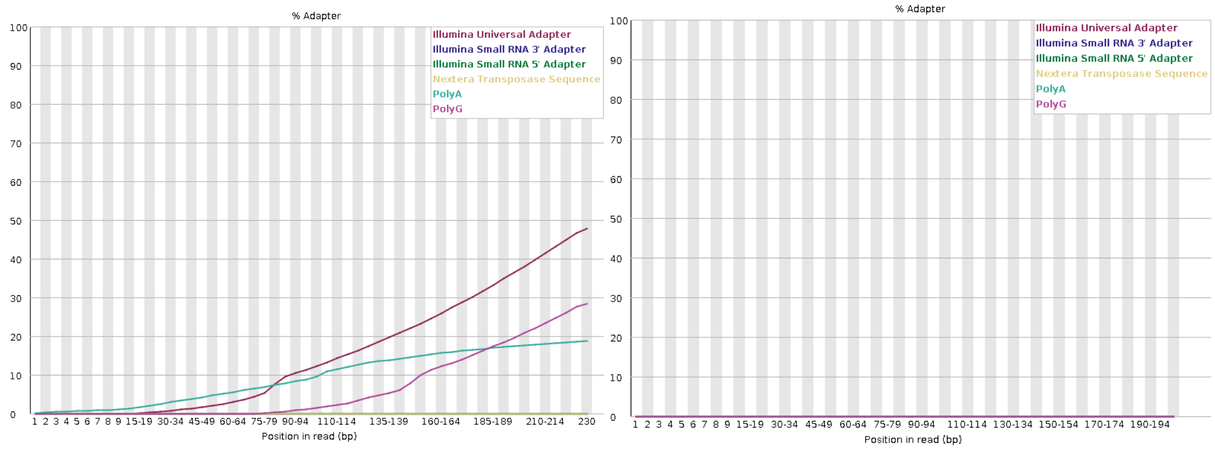
Chapter 3

Results

This chapter presents the findings of our single-cell RNA-seq analysis of *P. brassicacearum* R401. We first established data quality through preprocessing and quality control steps, then applied filtering strategies to ensure reliable cell identification and gene expression quantification.

3.1 Preprocessing and Quality Control

3.1.1 Trimming and Quality Control of FASTQ Files



(a) *before trimming*

(b) *after trimming*

Figure 3.1: *Artifacts content of the sublibrary BC_0076 before and after trimming with Cutadapt and Fastp*

The adapter content analysis shows the presence of various sequencing artifacts in the raw data (example of BC_0076). Before trimming (Figure 3.1a), distinct patterns are visible: polyG stretches (pink), polyA tails (cyan), and Illumina adapter sequences (purple). After trimming (Figure 3.1b), these artifacts are effectively removed, resulting in clean sequence data suitable for downstream analysis.

Table 3.1: *Summary of sequence metrics before and after trimming of each sublibrary with Cutadapt and Fastp (R1 read length metrics)*

Sample Name	BC_0076	BC_0077	BC_0079	BC_0080	$Mean$/Total
R1 length before trimming	241bp	241bp	241bp	241bp	$241bp$
R1 length after trimming	127bp	157bp	152bp	132bp	$142bp$
R1 number of sequences before trimming	631.4M	325.5M	379.1M	397.7M	**1733.7M**
R1 number of sequences after trimming	450.8M	248.4M	285.2M	300.1M	**1284.5M**
Change in number of sequences	-28.6%	-23.7%	-24.8%	-24.5%	**-25.4%**

The trimming process successfully removed various sequencing artifacts and improved data quality for downstream analysis. The preprocessing pipeline eliminated template-switching oligo (TSO) sequences, polyG stretches (NovaSeq-specific artifacts), polyA tails, and adapter sequences from the raw sequencing data (Figure 3.1, Table 3.1). This cleaning step was essential for accurate gene expression quantification and reliable single-cell analysis. The trimming process resulted in an average reduction of 25.4% in the total number of sequences across all sublibraries, from 1,733.7 million to 1,284.5 million reads (Table 3.1).

Table 3.3: *Summary of metrics obtained with STARsolo after barcode-UMI reading and alignment with the reference genome of P. brassicacearum R401*

Metric	Count/Percentage
Total Reads Processed	1,284,475,633
N_reads with valid barcodes	1,099,327,755 (**85.58%**)
- Exact Barcode Match	1,046,121,284 (**81.44%**)
- Single Mismatch Barcode	53,206,471 (**4.14%**)
Q30 Bases in RNA reads	95.79%
Q30 Bases in CB+UMI	95.51%
Unique Gene Mapping	34,593,349 (2.69%)
Unique + Multiple Gene Mapping	971,519,404 (75.64%)
Total Genes Detected	6,035 on the total of 6,249 genes (96.57%)
Cell Barcodes Detected	699,355
N_umi	36,565,214

3.1.2 Alignment and Quantification with STARsolo

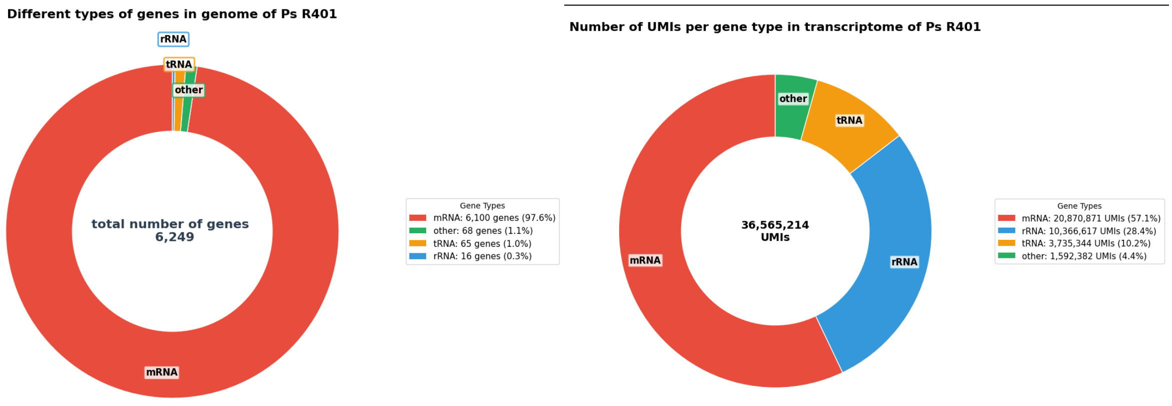
Following the successful preprocessing and trimming of our sequencing data, we performed alignment against the reference genome of *P. brassicacearum R401* (Figure 3.2a). The barcode reading and quantification by STARsolo[38,39] provided comprehensive quality metrics and mapping statistics.

The STARsolo analysis demonstrated excellent barcode quality with 85.58% of reads having valid barcodes, falling within the expected range of 70-90% for successful experiments[33]. The sequencing quality was outstanding with Q30 scores above 95% for both RNA reads and barcode/UMI sequences.

Gene mapping analysis revealed that 75.64% of reads mapped to genes, with 2.69% showing unique mapping, which is typical for bacterial genomes with overlapping genes and repeated sequences[33]. The fraction of uniquely aligned reads falls within the expected range of 3-12% for bacterial samples[33]. A total of 6,035 genes were detected across 699,355 unique cell barcodes, representing 96.6% of the 6,249 genes annotated in the *P. brassicacearum R401* genome.

Genome and Transcriptome Composition

The genome and transcriptome composition analysis reveals the distribution of different RNA types in our *P. brassicacearum* sample (Figure 3.2). In the annotated genome of *P. brassicacearum* R401, we identified 6,100 mRNA genes (97.6% of total genes), 65 tRNA genes (1.1% of total genes), 16 rRNA genes (0.3%) and 68 other genes. However, in the transcriptome analysis of our 36,565,214 UMIs Table 3.3, the distribution shows a different pattern: 57.1% correspond to mRNA transcripts (21 millions), while 28.4% are rRNA transcripts (representing approximately 10 millions UMIs), 10.2% are tRNA transcripts (representing approximately 4 millions UMIs) and 4.4% are other genes (representing approximately 1.5 millions UMIs). This distribution concords with the well-established



(a) *genome*

(b) *transcriptome*

Figure 3.2: *Compositional analysis of* P. brassicacearum *R401 genome annotation and transcriptome distribution of different RNA types*

observation that rRNA and tRNA transcripts represent a large proportion of the bacterial transcriptome despite constituting only a small fraction of the genome, due to their high transcriptional activity and stability[27].

3.2 Cell Quality Filtering and Dataset Characterization

STARsolo[38,39] detected 699,355 out of 820,800 possible barcode combinations, significantly more than the approximately 3,000 cells targeted in our sublibrary design Table 3.3. This large discrepancy indicates the presence of contaminating barcodes that can occur throughout the protocol, including potential contamination in the source oligonucleotide plates. To proceed with single-cell analysis, we needed to identify and retain only genuine cells among this large barcode population. STARsolo employs a knee plot strategy to identify "real" cells[46]. However, as recommended in the literature (Gaisser et al. 2024[33]), we chose to apply our own filtering criteria rather than using STARsolo's default KneePant method, which would have resulted in approximately 27,000 cells. This decision was based on the consideration that our experimental design included cells grown under different conditions (culture medium composition) and at different growth stages, which could result in varying transcriptional activity levels. Therefore, applying a single threshold across all conditions might eliminate cells with naturally lower but biologically relevant expression profiles. Instead, we chose to apply different thresholds to retain the best cells from each biological replicate, ensuring balanced representation across experimental conditions. This approach was validated by our initial UMI-based filtering results Section A.4, which revealed clear differences between culture conditions with M9 medium (nutrient-limited) showing fewer retained barcodes compared to M9F (nutrient-rich) when we used a global threshold of 100 UMIs per cell, confirming that different thresholds per samples were indeed necessary to capture cells from all experimental conditions.

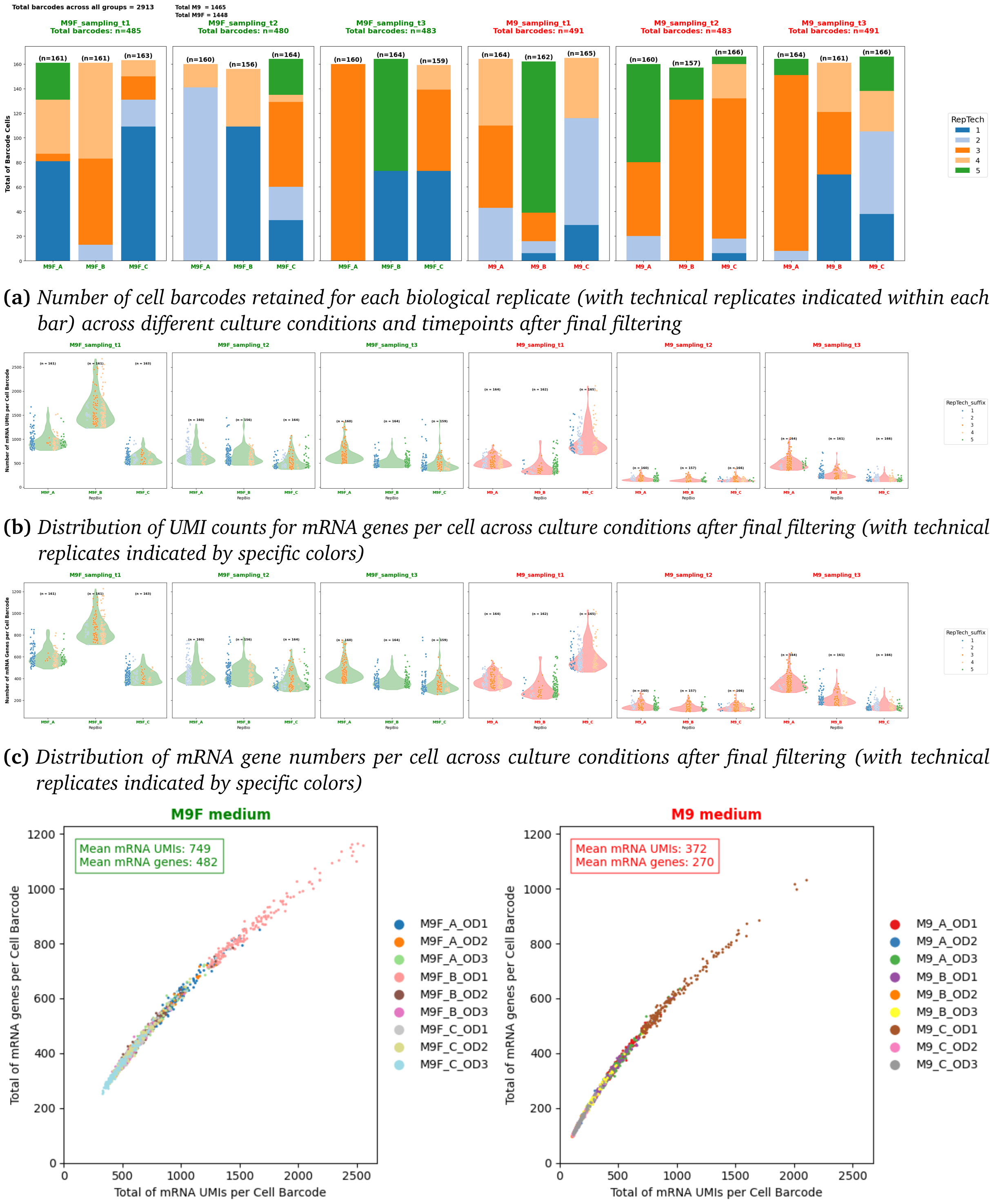
Following the initial UMI-based filtering, additional quality control steps were applied including doublet removal, biological replicate-based filtering, and mRNA type filtering. The final dataset contained approximately 3,000 cells, representing the highest-quality cells selected from each biological replicate to ensure balanced representation across experimental conditions. The filtering by biological replicate shows that some technical replicates were partially or completely eliminated to retain only the best-performing ones. The distribution of retained cells shows approximately 160 cells per biological replicate (Figure 3.3a). When few technical replicate categories remain for a biological replicate, it indicates high variability between technical replicates, likely due to differential barcode efficiency since each technical replicate possesses different barcode 1 combinations (Figure 3.3a). Some technical replicates were more efficient than others (e.g., M9F A sampling T2 had one technical replicate that was much more efficient than its group and therefore almost the only one for which cells passed the filter). Following this filtering, technical replicates are no longer considered for subsequent analyses, and we focus on the biological replicate level.

The following visualizations also show the distribution of mRNA gene expression patterns across the filtered cell population, specifically the number of mRNA genes detected per cell and the number of UMIs associated with mRNA genes per cell Figure 3.3b, Figure 3.3c, Figure 3.3d.

M9F medium consistently showed higher values compared to M9 (Figure 3.3d), with mean mRNA counts of 749 for M9F versus 372 for M9, and mean gene counts of 462 for M9F versus 270 for M9.

Biological replicate variability was particularly pronounced at T1 for both conditions and at T3 for M9 conditions Figure 3.3b, Figure 3.3c. Under nutrient-limited conditions (M9), cells showed markedly reduced transcriptional activity at T2 and T3 timepoints, with some replicates displaying very low UMI and gene counts, suggesting either reduced cellular activity or technical challenges in recovering stressed cells. The scatter plot (Figure 3.3d) reveals a strong positive correlation between UMI counts and mRNA gene numbers, confirming the quality of the filtered dataset and the relationship between sequencing depth and gene detection.

These results demonstrate that the filtering strategy effectively removed poor-quality barcodes to retain only potential “real cells” from each biological replicate that can be analyzed for downstream investigation.



(a) *Number of cell barcodes retained for each biological replicate (with technical replicates indicated within each bar) across different culture conditions and timepoints after final filtering*

(b) *Distribution of UMI counts for mRNA genes per cell across culture conditions after final filtering (with technical replicates indicated by specific colors)*

(c) *Distribution of mRNA gene numbers per cell across culture conditions after final filtering (with technical replicates indicated by specific colors)*

(d) *Relationship between UMI counts and mRNA gene counts per cell in the final filtered dataset*

Figure 3.3: *Metrics of filtered cells across experimental conditions (Culture Medium × Biological Replicate × Sampling Time) in the single-cell RNA-seq experiment.*

3.3 Single-Cell Analysis and Transcriptional Heterogeneity

Following the quality filtering steps, we performed dimensionality reduction and clustering analysis to explore transcriptional heterogeneity across cell populations and assess global differences between culture conditions. The analysis began with data transformation using the BacSC method (see Figure A.2b for detailed about it), followed by Principal Component Analysis (PCA) to identify the optimal number of dimensions for downstream analysis. UMAP visualization and clustering were then used to detect global transcriptional differences between experimental conditions.

3.3.1 Principal Component Analysis

The PCA analysis revealed a characteristic pattern of low variance explained by individual principal components (Figure A.2b), which is typical for bacterial single-cell RNA-seq data, particularly under stress conditions. PC1 captured approximately 0.6% of the total variance and dominated the other components, which showed even lower contributions. This indicates that the transcriptional variability is diffuse or weakly structured in the dataset. In this context, for the next analysis we chose to retain the first 5 PCs for UMAP embedding to capture the few potentially meaningful signals carried by the leading components while avoiding the integration of excessive noise. This selection provides a balance between maintaining minimal expressiveness for dimensionality reduction and exercising caution given the low inherent structure of the data.

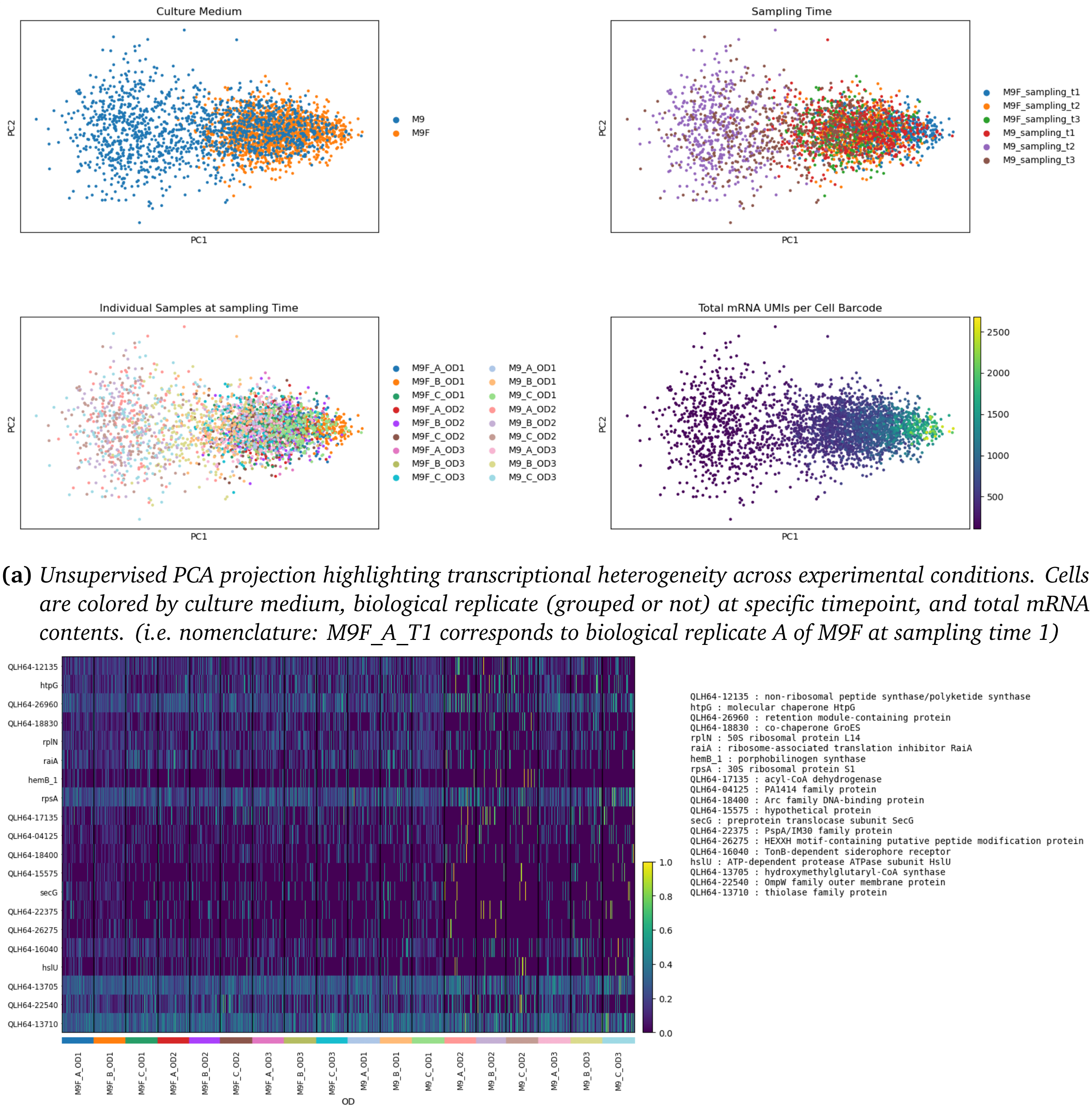
PC1 successfully distinguished between the two culture conditions (M9 and M9F), revealing distinct transcriptional patterns (Figure 3.4a). In fact, we can see that M9 T2 and M9 T3 samples (brown and purple points in the Sampling Time facet, top right) were clearly separated from other conditions, while M9 T1 showed no significant difference from M9F conditions. We can also observe this on the heatmaps (Figure 3.4b).

Notably, M9 samples positioned on the left side of PC1 showed greater dispersion compared to those on the right. Furthermore, PC1 axis alignment followed a gradient of total mRNA concentration per cell, suggesting that this component primarily captures differences in overall transcriptional activity between conditions.

Gene Contribution Analysis to PC1

The heatmap analysis with z-score scaling (ranging from 0 to 1) revealed the genes contributing most significantly to PC1 variance (Figure 3.4b). Each row represents a gene and each column represents a cell, with the legend at the bottom corresponding to biological replicates.

The genes contributing most to PC1 variance were primarily associated with translational regulation and iron metabolism, reflecting the stress response to iron limitation in M9 medium. Among the



(a) *Unsupervised PCA projection highlighting transcriptional heterogeneity across experimental conditions. Cells are colored by culture medium, biological replicate (grouped or not) at specific timepoint, and total mRNA contents. (i.e. nomenclature: M9F_A_T1 corresponds to biological replicate A of M9F at sampling time 1)*

(b) *Heatmap of the top contributing genes (top 20) to the first principal components (PC1), based on scaled expression values (z-score per gene). Rows represent genes, columns represent single cells collected at each timepoints for all biological replicates. The legend at the right side of the heatmap shows the full name of the genes.*

Figure 3.4: *Principal Component Analysis (PCA) results showing transcriptional heterogeneity and gene contribution patterns in* P. brassicacearum *populations*

translational regulators, we identified rpsA (30S ribosomal protein S1), raiA (ribosome-associated translation inhibitor), and htpG (HSP90 protein chaperone). HtpG, as a molecular chaperone, plays an indirect but essential role in the biosynthesis and functionality of siderophores. For example, HtpG facilitates the correct folding and stability of enzymes involved in siderophore synthesis, such as those participating in yersiniabactine production in *Yersinia* species[47].Notably, porphobilinogen synthase was also identified among the top 20 genes contributing to PC1 variance. This finding is particularly relevant given our experimental design using iron-limited M9 medium. In bacteria, the link between porphobilinogen (PBG) and iron is essential for heme synthesis, a cofactor indispensable for many proteins involved in metabolism, cellular respiration, and electron transport. The biosynthesis of heme in bacteria follows a pathway similar to eukaryotes, with PBG being a key intermediate formed by the condensation of two delta-aminolevulinic acid (ALA) molecules, catalyzed by porphobilinogen synthase. This process is regulated by iron availability in bacterial cells. The differential expression of porphobilinogen synthase between M9 and M9F conditions thus reflects the cellular adaptation to iron stress[48]. The analysis also identified TonB-dependent siderophore receptors, which are essential for high-affinity iron acquisition in gram-negative bacteria[49,50]. These outer membrane-localized proteins bind iron chelates at the cell surface and promote their uptake. Together with porphobilinogen synthase, these genes form a coherent transcriptional response to iron limitation, explaining the clear separation of M9 conditions along PC1.

While PC1 primarily captured differences in transcriptional activity and stress response, PC2 revealed distinct patterns related to cellular motility and chemotaxis (Figure A.2d). For example, methyl-accepting chemotaxis proteins (MCPs) and flagellin genes were identified as major contributors to PC2 variance, suggesting a role in chemotaxis and motility responses to environmental conditions.[1]

3.3.2 UMAP Visualization and Clustering Analysis

Following the PCA analysis, we performed UMAP visualization and clustering analysis to explore transcriptional heterogeneity across cell populations and assess global differences between culture conditions.

The clustering analysis with resolution 0.1 identified two main clusters, while resolution 0.25 revealed three clusters (Figure 3.5b). Both clustering approaches successfully separated the experimental conditions, with M9 T2 and M9 T3 samples clearly distinguished from other conditions. This separation aligns with the PCA results, confirming the distinct transcriptional profiles of cells under iron-limited conditions at later timepoints.

Differential Expression Analysis Between Clusters

To characterize the transcriptional differences between the identified clusters, we performed differential expression analysis between the two clusters obtained with resolution 0.1. We used the Wilcoxon rank-sum test with FDR correction ($p < 0.05$) to identify significantly differentially expressed genes. The complete results of this analysis are presented in the appendix (Figure A.3a).

Group 1 (primarily M9 T2 and T3 cells) exhibited globally lower expression levels compared to

[1] **Chemotaxis** is a biological process by which cells move in response to chemical gradients in their environment, enabling navigation toward favorable conditions and away from unfavorable ones.

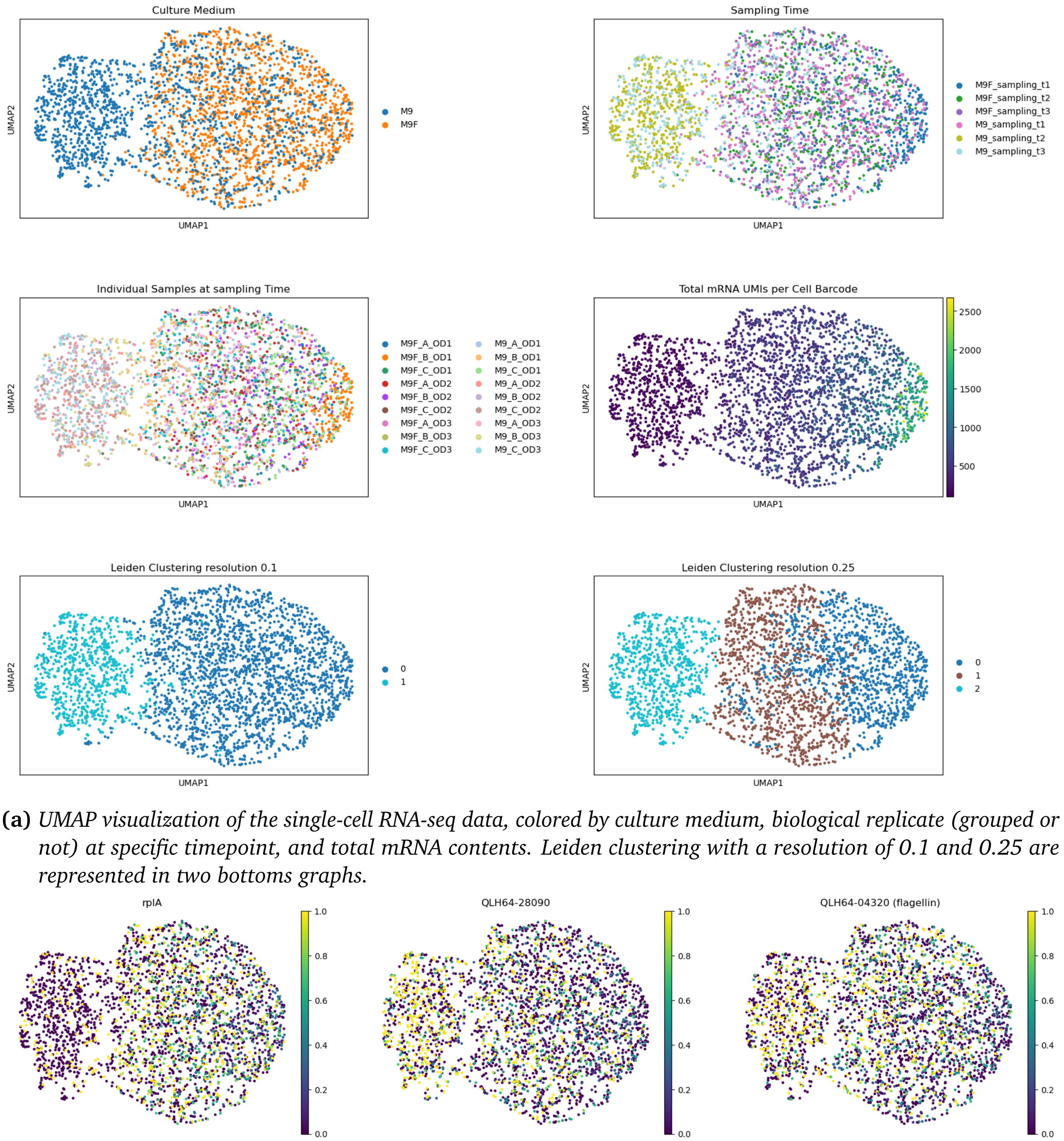
Group 0 (primarily M9F and M9 T1 cells), which showed a more heterogeneous expression gradient (Figure 3.5b). This pattern reflects the reduced transcriptional activity observed in cells under iron stress conditions.

Characterization of Differentially Expressed Genes

Among the significantly differentially expressed genes, we identified three representative examples that illustrate the distinct transcriptional programs between clusters. The *RplA* gene that codes for ribosomal protein RplA (30S ribosomal protein L1) showed lower expression in M9 T2 and T3 cells compared to M9F and M9 T1 cells. RplA is constitutively expressed as it is essential for protein synthesis, with expression levels reflecting the cell's translational activity. Higher expression typically indicates active protein synthesis during rapid growth, while lower expression suggests reduced translational activity under stress conditions. The *QLH64-28090* (phasin) gene showed an opposite pattern, with higher expression in M9 T2 and T3 cells. Phasins are multifunctional proteins associated with polyhydroxyalkanoate (PHA) granules that play crucial roles in stress response and energy metabolism[51]. They form an interface between the hydrophobic PHA granules and the hydrophilic cytoplasm, regulating both PHA accumulation and utilization. Under stress conditions, phasins can activate PHA depolymerization to release energy metabolites, increase PHA synthase activity . Additionally, some phasins exhibit chaperone-like properties, protecting cellular proteins against stress-induced denaturation and oxidative damage. The *Flagellin* gene, which is a housekeeping gene, was the most highly expressed gene in the raw data matrix, also showed differential expression between clusters, with higher expression in M9 T2 and T3 cells compared to M9F and M9 T1.

Global Expression Patterns and Cluster Heterogeneity

However, the analysis revealed significant heterogeneity within each cluster, with groups of cells showing distinct expression patterns for specific genes compared to the global cluster average. This heterogeneity is particularly evident in the violin plots presented in the appendix Figure A.3c, which show variable distributions of gene expression within clusters. In the heatmap Figure A.3b representing top significant differentially expressed genes, we can see that some genes are expressed at very high levels, particularly among the significant genes for cluster 1. This observation could suggest the presence of specialized subpopulations within the broader transcriptional groups, potentially indicating fine-grained division of labor mechanisms at the cellular level.



(a) *UMAP visualization of the single-cell RNA-seq data, colored by culture medium, biological replicate (grouped or not) at specific timepoint, and total mRNA contents. Leiden clustering with a resolution of 0.1 and 0.25 are represented in two bottoms graphs.*

(b) *UMAP visualization colored by expression of examples top differentially expressed genes (FDR < 0.05) between clusters: RplA (30S ribosomal protein L1), QLH64-28090 (phasin), and flagellin. Color intensity indicates relative expression levels (yellow: high expression, blue: low expression).*

Figure 3.5: *Single-cell RNA-seq analysis results showing gene expression patterns and cellular heterogeneity in* P. brassicacearum *populations*

Chapter 4

Discussion

This study aimed to validate the microSPLiT technology for exploring division of labor (DOL) within bacterial populations, specifically *Pseudomonas brassicacearum* R401. While a comprehensive DOL analysis remains to be conducted, our work has successfully established a robust analysis pipeline and generated promising initial data that provide valuable insights into bacterial single-cell transcriptomics.

4.1 Technical Validation of microSPLiT Methodology

The microSPLiT methodology performed as anticipated, with sequencing and trimming steps effectively removing adapters and yielding high-quality reads suitable for alignment to the *PsR401* genome using STARsolo. The successful implementation of this technology represents a significant step forward in bacterial single-cell analysis, particularly given the technical challenges associated with prokaryotic systems compared to eukaryotic counterparts[27].

While the current pipeline achieved good results (Figure 3.1 Table 3.1 Table 3.3), there are opportunities to further optimize read recovery. STARsolo's reliance on fixed barcode positions could be complemented by implementing BarQC[52], which would enable recovery of shifted barcodes utilizing CIGAR motif analysis and thus increase the overall yield of usable reads. BarQC would also provide visual representations of read proportions in each well at every barcoding round, facilitating the assessment of potential biases in cell distribution or barcoding efficiency.

Furthermore, implementing comprehensive quality control measures, including contamination screening tools such as FastQ Screen[53], Centrifuge[54] or Recentrifuge[55], would improve data reliability and ensure the absence of cross-contamination between samples.

The presence of a significant proportion of ribosomal RNA reads (∼28%) Figure 3.2b was expected but highlights an opportunity for optimization[27]. Implementing upstream rRNA depletion could substantially increase mRNA yield, thereby enhancing the biological signal available for transcriptomic analysis. Additionally, the recent availability of an updated genome annotation presents an opportunity to improve the accuracy of gene quantification in future analyses.

4.2 Transcriptomic Responses to Iron Limitation

Our analysis revealed clear transcriptomic distinctions between experimental conditions, with cells grown in iron-limited medium at timepoints 2 and 3 showing fewer detected genes and lower total UMI counts compared to iron-rich conditions Figure 3.3. This observation necessitated the adaptation of filtering thresholds per condition to maintain comparable cell distributions across samples, ultimately retaining 160 cells per condition for a total of 3,000 cells.

The PCA and UMAP analyses demonstrated that, despite the first principal component explaining only a small fraction of variance (∼0.6%), typical for bacterial scRNA-seq datasets, it effectively distinguished between M9 (iron-poor) and M9F (iron-rich) conditions Figure 3.4 Figure 3.4b. The distinct transcriptomic profiles observed in cells from M9 T2 and T3 compared to M9 T1 or M9F support the hypothesis that T1 cells in depleted medium had not yet exhausted residual iron resources. This interpretation is corroborated by optical density measurements, which plateaued at T2 and T3, indicating growth limitation.

4.2.1 Gene Expression Patterns Under Iron Stress

Analysis of the most contributive genes Figure 3.4b to condition separation revealed an enrichment of genes related to translation machinery, protein chaperones, siderophore biosynthesis regulators, and iron acquisition receptors. This pattern indicates an adaptive metabolic reorganization under iron stress, consistent with previous studies demonstrating that iron limitation triggers comprehensive cellular reprogramming in bacteria[26].

Differential expression analysis confirmed the downregulation of ribosomal protein genes, including RplA, under iron stress conditions, alongside the relative upregulation of genes involved in storage metabolism and stress response mechanisms. This transcriptional shift reflects the cellular transition from active growth to survival-oriented metabolism, characterized by reduced protein synthesis and enhanced stress protection mechanisms. Such responses align with established paradigms of bacterial stress adaptation, where resource limitation drives cells toward quiescence and dormancy states[56].

The coarse-grained clustering approach employed in this study, while revealing global transcriptional differences between major experimental groups, may mask important biological heterogeneity that could indicate DOL mechanisms. To address our original biological question about cellular specialization within bacterial populations, future analyses must examine finer-scale heterogeneity by analyzing individual culture conditions and biological replicates at each timepoint. This approach would be necessary to identify potential subpopulations that might reveal specialized cellular functions and DOL patterns within *P. brassicacearum* populations.

4.3 Methodological Considerations and Limitations

4.3.1 Cell Quality Assessment Challenges

The estimation of unique cell numbers through barcode analysis presents ongoing challenges, as the distinction between genuine single cells and cell aggregates or artifacts remains difficult to establish definitively. The subjective nature of filtering methods compounds this issue, as cells deemed “low quality” in non-stressed conditions might represent biologically relevant states, such as dormant or stress-adapted phenotypes that could contribute to population-level DOL.

4.3.2 Dataset Scale and Computational Considerations

Compared to eukaryotic single-cell studies that routinely analyze hundreds of thousands to millions of cells, our prokaryotic dataset appears modest with 3,000 cells analyzed. However, the smaller genome size of bacteria offers advantages for local computational analysis and may provide pedagogical value for method development and training purposes. The manageable dataset size facilitates thorough exploration of analytical approaches and parameter optimization, which is particularly valuable for emerging technologies like bacterial single-cell RNA sequencing.

4.4 Future Directions and Recommendations

Several technological and methodological improvements could enhance future investigations of bacterial DOL using single-cell approaches. If the experiment were to be repeated, it would be beneficial to include bulk RNA-seq and blank samples to improve statistical power and provide additional validation of single-cell results[57].

The complete testing of the bacSC pipeline[45] would also be valuable, as this specialized tool for bacterial single-cell analysis may provide improved clustering specifically tailored for prokaryotic systems. Future analyses should focus on performing DOL analysis by studying each sample independently, allowing for the identification of condition-specific cellular states and specialized subpopulations.

Subsequently, comprehensive analyses including Gene Ontology enrichment, cell trajectory inference, metabolomic pathway analysis, and pseudobulk analysis would provide deeper insights into bacterial population dynamics and the molecular mechanisms underlying potential DOL phenomena. These approaches would enable a more complete understanding of how bacterial populations coordinate responses to environmental stress and whether specialized cellular roles emerge under challenging conditions.

In conclusion, while this study has successfully established the technical foundation for bacterial single-cell transcriptomics using microSPLiT, the full exploration of division of labor in *P. brassicacearum* populations remains an exciting avenue for future research. The methodological advances and initial findings presented here provide a solid basis for more detailed investigations of bacterial population heterogeneity and specialization.

Bibliography

1. Taborsky, M., Fewell, J. H., Gilles, R. & Taborsky, B. Division of labour as key driver of social evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **380**, 20230261 (2025).
2. Cooper, G. A. & West, S. A. Division of labour and the evolution of extreme specialization. *Nature Ecology & Evolution* **2**, 1161–1167 (2018).
3. Giri, S., Waschina, S., Kaleta, C. & Kost, C. Defining division of labor in microbial communities. *Journal of Molecular Biology* **431**, 4712–4731 (2019).
4. Rafieenia, R., Atkinson, E. & Ledesma-Amaro, R. Division of labor for substrate utilization in natural and synthetic microbial communities. *Current Opinion in Biotechnology* **75**, 102706 (2022).
5. Durán, P. *et al.* Microbial interkingdom interactions in roots promote arabidopsis survival. *Cell* **175**, 973–983.e14 (2018).
6. Mataigne, V., Vannier, N., Vandenkoornhuyse, P. & Hacquard, S. Microbial systems ecology to understand cross-feeding in microbiomes. *Frontiers in Microbiology* **12**, (2021).
7. Mataigne, V., Vannier, N., Vandenkoornhuyse, P. & Hacquard, S. Multi-genome metabolic modeling predicts functional inter-dependencies in the arabidopsis root microbiome. *Microbiome* **10**, 217 (2022).
8. Estrela, S., Kerr, B. & Morris, J. J. Transitions in individuality through symbiosis. *Current Opinion in Microbiology* **31**, 191–198 (2016).
9. Adkins-Jablonsky, S. J., Clark, C. M., Papoulis, S. E., Kuhl, M. D. & Morris, J. J. Market forces determine the distribution of a leaky function in a simple microbial community. *Proceedings of the National Academy of Sciences of the United States of America* **118**, e2109813118 (2021).
10. López-Pagán, N. *et al.* Pseudomonas syringae subpopulations cooperate by coordinating flagellar and type III secretion spatiotemporal dynamics to facilitate plant infection. *Nature Microbiology* **10**, 958–972 (2025).

11. Raj, A. & Oudenaarden, A. van. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).

12. Keren, L. *et al.* Noise in gene expression is coupled to growth rate. *Genome Research* **25**, 1893–1902 (2015).

13. Chowdhury, D., Wang, C., Lu, A. & Zhu, H. Cis-regulatory logic produces gene-expression noise describing phenotypic heterogeneity in bacteria. *Frontiers in Genetics* **12**, (2021).

14. Lopez, J. G. & Wingreen, N. S. Noisy metabolism can promote microbial cross-feeding. *eLife* **11**, e70694 (2022).

15. 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).

16. Knights, H. E., Jorrin, B., Haskett, T. L. & Poole, P. S. Deciphering bacterial mechanisms of root colonization. *Environmental Microbiology Reports* **13**, 428–444 (2021).

17. Getzke, F. *et al.* Cofunctioning of bacterial exometabolites drives root microbiota establishment. *Proceedings of the National Academy of Sciences* **120**, e2221508120 (2023).

18. Jian, Y. *et al.* How plants manage pathogen infection. *EMBO reports* **25**, 31–44 (2024).

19. Dodds, P. N., Chen, J. & Outram, M. A. Pathogen perception and signaling in plant immunity. *The Plant Cell* **36**, 1465–1481 (2024).

20. Getzke, F. *et al.* Physiochemical interaction between osmotic stress and a bacterial exometabolite promotes plant disease. *Nature Communications* **15**, 4438 (2024).

21. Chesneau, G., Herpell, J., Wolf, S. M., Perin, S. & Hacquard, S. MetaFlowTrain: a highly parallelized and modular fluidic system for studying exometabolite-mediated inter-organismal interactions. *Nature Communications* **16**, 3310 (2025).

22. Cao, M. *et al.* Spatial IMA1 regulation restricts root iron acquisition on MAMP perception. *Nature* **625**, 750–759 (2024).

23. Gu, S. *et al.* Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. *Nature Microbiology* **5**, 1002–1010 (2020).

24. Harbort, C. J. *et al.* Root-secreted coumarins and the microbiota interact to improve iron nutrition in arabidopsis. *Cell Host & Microbe* **28**, 825–837.e6 (2020).

25. Mesny, F., Hacquard, S. & Thomma, B. P. Co-evolution within the plant holobiont drives host performance. *EMBO reports* **24**, e57455 (2023).

26. Lim, C. K., Hassan, K. A., Tetu, S. G., Loper, J. E. & Paulsen, I. T. The Effect of Iron Limitation on the Transcriptome and Proteome of Pseudomonas fluorescens Pf-5. *PLOS ONE* **7**, e39139 (2012).

27. 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.

28. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics* **24**, 494–515 (2023).

29. Nobori, T. *et al.* A rare PRIMER cell state in plant immunity. *Nature* 1–9 (2025) doi:10.1038/s41586-024-08383-z.

30. Nobori, T. Exploring the untapped potential of single-cell and spatial omics in plant biology. *New Phytologist* **n/a**,.

31. Sarfatis, A., Wang, Y., Twumasi-Ankrah, N. & Moffitt, J. R. Highly multiplexed spatial transcriptomics in bacteria. *Science* **387**, eadr0932 (2025).

32. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**, eaba5257 (2021).

33. Gaisser, K. D. *et al.* High-throughput single-cell transcriptomics of bacteria using combinatorial barcoding. *Nature Protocols* **19**, 3048–3084 (2024).

34. Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data.

35. 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).

36. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

37. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

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

39. 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.

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

41. 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).

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

43. 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).

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

45. Ostner, J. *et al.* BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis. doi:10.1101/2024.06.22.600071.

46. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology* **15**, e8746 (2019).

47. Mangla, N., Singh, R. & Agarwal, N. HtpG is a metal-dependent chaperone which assists the DnaK/DnaJ/GrpE chaperone system of mycobacterium tuberculosis via direct association with DnaJ2. *Microbiology Spectrum* **11**, e00312–23.

48. Choby, J. E. & Skaar, E. P. Heme synthesis and acquisition in bacterial pathogens. *Journal of molecular biology* **428**, 3408–3428 (2016).

49. Moeck, G. S. & Coulton, J. W. TonB-dependent iron acquisition: mechanisms of siderophore-mediated active transport. *Molecular Microbiology* **28**, 675–681 (1998).

50. Fujita, M. *et al.* A TonB-dependent receptor constitutes the outer membrane transport system for a lignin-derived aromatic compound. *Communications Biology* **2**, 432 (2019).

51. Mezzina, M. P. & Pettinari, M. J. Phasins, Multifaceted Polyhydroxyalkanoate Granule-Associated Proteins. *Applied and Environmental Microbiology* **82**, 5060–5067 (2016).

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

53. Wingett, S. W. & Andrews, S. FastQ screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).

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

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

56. Rittershaus, E. S. C., Baek, S.-H. & Sassetti, C. M. The normalcy of dormancy: Common themes in microbial quiescence. *Cell Host & Microbe* **13**, 643–651 (2013).

57. 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).

Appendix A

Appendix

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 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.

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 ${SLURM_CPUS_PER_TASK} \
    -g "AAGCAGTGGTATCAACGCAGAGTGAATGGG; min_overlap=6; max_errors=0.2" \
    -g "CAGAGTGAATGGG; min_overlap=6; max_errors=0.2" \
    --pair-filter=both \
    -m 20: \
    --too-short-output
     ↪  "${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.3 STARsolo supplementary information

A.3.1 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.3.2 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.3.3 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

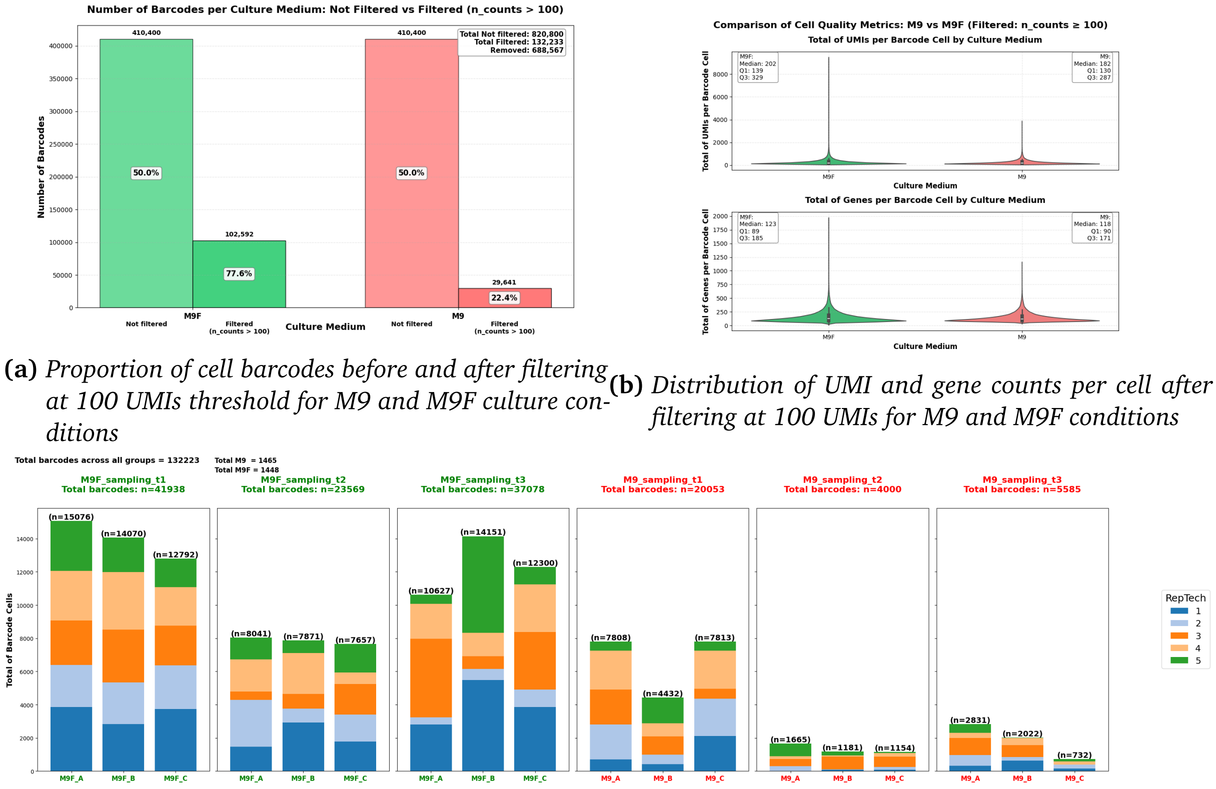
- `--soloType CB_UMI_Complex` : Analysis type for cell barcodes (CB) and complex UMIs
- `--soloCBwhitelist` : List of valid cell barcodes for the three barcoding rounds
- `--soloFeatures Gene GeneFull` : Analysis of features at both gene and full transcript levels
- `--soloUMIdedup 1MM_All` : UMI deduplication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

A.4 Initial filtering with minimum 100 UMI per cell

The first step involved filtering cells based on unique molecular identifier (UMI) counts with a minimum of 100 UMIs per cell.



(a) *Proportion of cell barcodes before and after filtering at 100 UMIs threshold for M9 and M9F culture conditions*

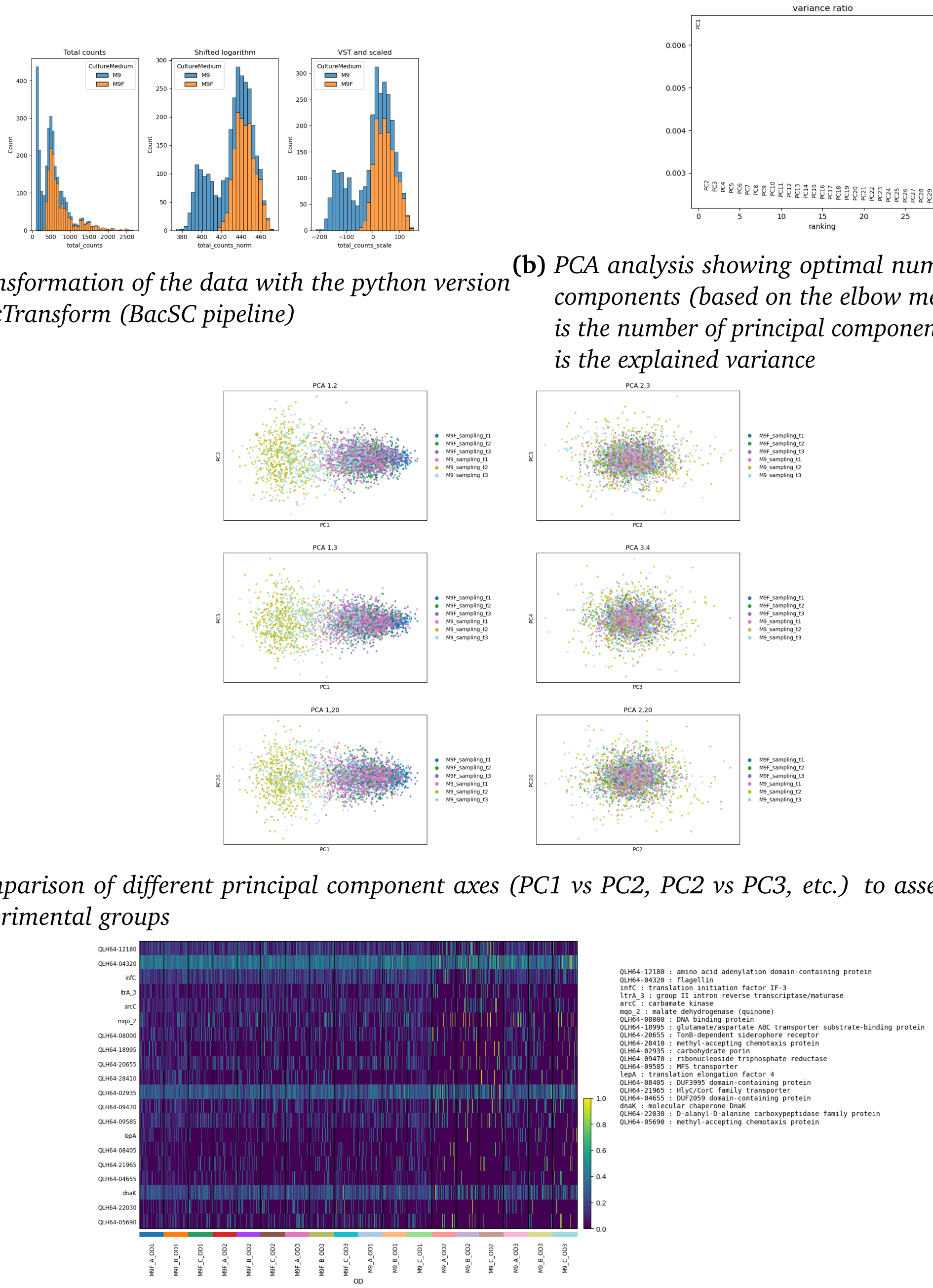
(b) *Distribution of UMI and gene counts per cell after filtering at 100 UMIs for M9 and M9F conditions*

(c) *Total number of cell barcodes retained for each biological replicate (with technical replicates indicated within each bar) across different culture conditions and timepoints after filtering at 100 UMIs*

Figure A.1: *Filtering of low-quality cell barcodes with fewer than 100 UMIs per cell across all biological replicates in the single-cell RNA-seq experiment.*

The initial filtering step removed cell barcodes with fewer than 100 UMIs to eliminate artifacts. This threshold was chosen to remove very low-quality barcodes and to demonstrate clear differences in UMI distributions between the two culture conditions (M9 vs M9F), with M9 losing more BCs at this threshold, while also serving as a quality control metric to identify potentially failed technical replicates. The filtering revealed significant heterogeneity between conditions, with M9 medium showing fewer retained barcodes compared to M9F. This difference could reflect lower transcriptional activity under nutrient-limited conditions, but may also be related to other factors such as cell wall modifications affecting permeabilization efficiency, or increased cell death leading to differential recovery during washing steps. Additionally, technical replicates showed varying sensitivity to filtering, indicating heterogeneity in the ability to recover reads across different experimental batches. Notably, one technical replicate (replicate 3 of M9_C at T3) appeared to have failed, likely due to a pipetting error, and was consequently eliminated by this threshold.

A.5 PCA analysis Supplementary Figures



(a) *Transformation of the data with the python version of scTransform (BacSC pipeline)*

(b) *PCA analysis showing optimal number of principal components (based on the elbow method) the x axis is the number of principal components and the y axis is the explained variance*

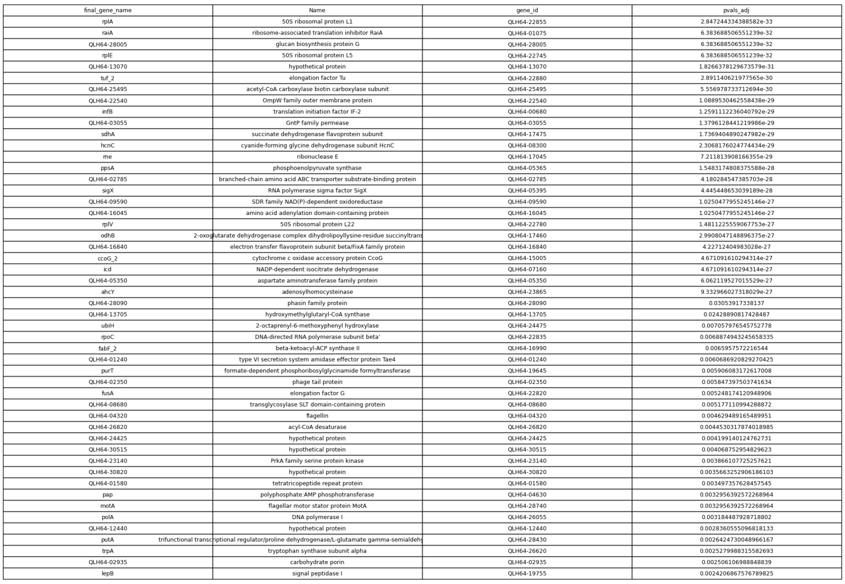
(c) *Comparison of different principal component axes (PC1 vs PC2, PC2 vs PC3, etc.) to assess separation of experimental groups*

(d) *Heatmap of the top contributing genes (top 20) to the first principal components (PC2), based on scaled expression values (z-score per gene). Rows represent genes, columns represent single cells collected at each timepoints for all biological replicates. The legend at the right side of the heatmap shows the full name of the genes.*

Figure A.2: *Supplementary figures for the PCA analysis*

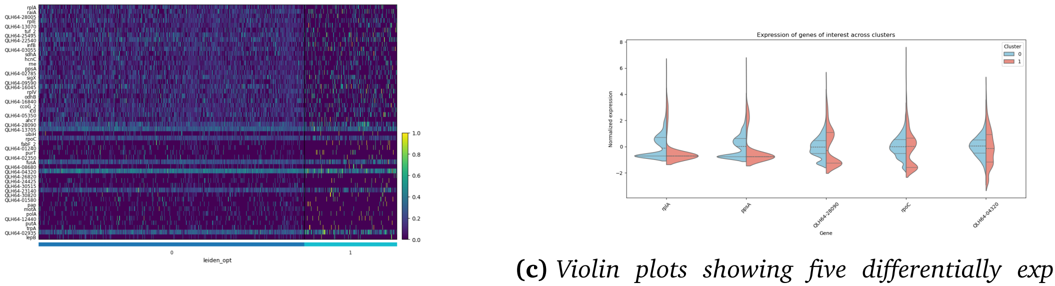
The PCA analysis supplementary figures provide detailed insights into the dimensionality reduction process. The scaling transformation (Figure A.2a) shows data normalization using scTransform from the BacSC pipeline. The elbow plot (Figure A.2b) demonstrates the optimal number of principal components selection based on explained variance. The PC comparison (Figure A.2c) evaluates different principal component *combinations for group separation. The heatmap (Figure A.2d) identifies the top contributing genes to PC2, revealing key transcriptional drivers of the observed separation between experimental conditions.*

A.6 UMAP Analysis Supplementary Figures



final_gene_name	Name	gene_id	pvals_adj
rplA	50S ribosomal protein L1	QLH64-22855	2.847244334388582e-33
raiA	ribosome-associated translation inhibitor RaiA	QLH64-01075	6.383688506551239e-32
QLH64-28005	glucan biosynthesis protein G	QLH64-28005	6.383688506551239e-32
rplE	50S ribosomal protein L5	QLH64-22745	6.383688506551239e-32
QLH64-13070	hypothetical protein	QLH64-13070	1.8266378129673579e-31
tuf_2	elongation factor Tu	QLH64-22880	2.891140621977565e-30
QLH64-25495	acetyl-CoA carboxylase biotin carboxylase subunit	QLH64-25495	5.556978733712694e-30
QLH64-22540	OmpW family outer membrane protein	QLH64-22540	1.0889530462558438e-29
infB	translation initiation factor IF-2	QLH64-00680	1.2591112236040792e-29
QLH64-03055	GntP family permease	QLH64-03055	1.3796128441219986e-29
sdhA	succinate dehydrogenase flavoprotein subunit	QLH64-17475	1.7369404890247982e-29
hcnC	cyanide-forming glycine dehydrogenase subunit HcnC	QLH64-08300	2.3068176024774434e-29
rne	ribonuclease E	QLH64-17045	7.211813908166355e-29
ppsA	phosphoenolpyruvate synthase	QLH64-05365	1.5483174808375588e-28
QLH64-02785	branched-chain amino acid ABC transporter substrate-binding protein	QLH64-02785	4.180284547385703e-28
sigX	RNA polymerase sigma factor SigX	QLH64-05395	4.445448653039189e-28
QLH64-09590	SDR family NAD(P)-dependent oxidoreductase	QLH64-09590	1.0250477955245146e-27
QLH64-16045	amino acid adenylation domain-containing protein	QLH64-16045	1.0250477955245146e-27
rplV	50S ribosomal protein L22	QLH64-22780	1.4811225559067753e-27
odhB	2-oxoglutarate dehydrogenase complex dihydrolipoyllysine-residue succinyltrans	QLH64-17460	2.9908047148896375e-27
QLH64-16840	electron transfer flavoprotein subunit beta/FixA family protein	QLH64-16840	4.22712404983028e-27
ccoG_2	cytochrome c oxidase accessory protein CcoG	QLH64-15005	4.671091610294314e-27
icd	NADP-dependent isocitrate dehydrogenase	QLH64-07160	4.671091610294314e-27
QLH64-05350	aspartate aminotransferase family protein	QLH64-05350	6.062119527015529e-27
ahcY	adenosylhomocysteinase	QLH64-23865	9.332966027318029e-27
QLH64-28090	phasin family protein	QLH64-28090	0.03053917338137
QLH64-13705	hydroxymethylglutaryl-CoA synthase	QLH64-13705	0.02428890817428487
ubiH	2-octaprenyl-6-methoxyphenyl hydroxylase	QLH64-24475	0.007057976545752778
rpoC	DNA-directed RNA polymerase subunit beta'	QLH64-22835	0.0068874943245658335
fabF_2	beta-ketoacyl-ACP synthase II	QLH64-16990	0.0065957572216544
QLH64-01240	type VI secretion system amidase effector protein Tae4	QLH64-01240	0.0060686920829270425
purT	formate-dependent phosphoribosylglycinamide formyltransferase	QLH64-19645	0.005906083172617008
QLH64-02350	phage tail protein	QLH64-02350	0.005847397503741634
fusA	elongation factor G	QLH64-22820	0.005248174120948906
QLH64-08680	transglycosylase SLT domain-containing protein	QLH64-08680	0.005177110994288872
QLH64-04320	flagellin	QLH64-04320	0.004629489165489951
QLH64-26820	acyl-CoA desaturase	QLH64-26820	0.0044530317874018985
QLH64-24425	hypothetical protein	QLH64-24425	0.004199140124762731
QLH64-30515	hypothetical protein	QLH64-30515	0.004068752954829623
QLH64-23140	PrkA family serine protein kinase	QLH64-23140	0.003866107725257621
QLH64-30820	hypothetical protein	QLH64-30820	0.0035663252906186103
QLH64-01580	tetratricopeptide repeat protein	QLH64-01580	0.003497357628457545
pap	polyphosphate:AMP phosphotransferase	QLH64-04630	0.0032956392572268964
motA	flagellar motor stator protein MotA	QLH64-28740	0.0032956392572268964
polA	DNA polymerase I	QLH64-26055	0.003184487928718802
QLH64-12440	hypothetical protein	QLH64-12440	0.0028360555096818133
putA	trifunctional transcriptional regulator/proline dehydrogenase/L-glutamate gamma-semialdehy	QLH64-28430	0.0026424730048966167
trpA	tryptophan synthase subunit alpha	QLH64-26620	0.0025279988315582693
QLH64-02935	carbohydrate porin	QLH64-02935	0.002506106988848839
lepB	signal peptidase I	QLH64-19755	0.0024206867576789825

(a) *Table showing the top 50 most significant genes among differentially expressed genes between the two clusters identified by Leiden clustering, with gene names and significance thresholds (p-value adj) after FDR correction*



(b) *Heatmap with z-score scaling (ranging from 0 to 1) of the top 50 most significant genes among differentially expressed genes between the two clusters identified by Leiden clustering. Rows represent genes, columns represent single cells of both leiden clusters*

(c) *Violin plots showing five differentially expressed genes (rplA (30S ribosomal protein L1), ppsA, QLH64_28090 (phasin), rpoC, QLH64_04320 (flagellin)) between the two clusters identified by Leiden clustering*

Figure A.3: *Supplementary figures from UMAP clustering analysis revealing transcriptional heterogeneity in P.* brassicacearum *populations under iron stress.*

Differential expression analysis identified significantly differentially expressed genes (Wilcoxon test with FDR correction, $p < 0.05$) between the two main clusters identified using Leiden clustering with resolution 0.1. The heatmap (Figure A.3b) shows expression patterns of the top 50 most significant genes, while violin plots (Figure A.3c) illustrate the distribution of five representative genes, revealing that despite statistical significance, considerable variability exists within clusters, which may suggest specialized cellular functions or heterogeneous physiological states.

Master’s Thesis in Bioinformatics

University of 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.