Promoters

Prof. Nico Boon and Dr. Ruben Props

Center for Microbial Ecology and Technology (CMET), Department of Biotechnology, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

Members of the examination committee

Prof. Monica Höfte (Chair)

Laboratory of Phytopathology, Department of Plants and Crops, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

Prof. Andre Skirtach (Secretary)

Laboratory of Nano & Biotechnology, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

Dr. Luz Pérez Gómez de Cadiñanos

Research & Development Project Manager, Syngulon

Prof. Peter Vandamme

Laboratory of Microbiology, Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium

Prof. Huabing Yin

School of Engineering, University of Glasgow, Glasgow, United Kingdom

Dean of the Faculty of Bioscience Engineering

Prof. Marc Van Meirvenne

Rector of Ghent University

Prof. Rik Van de Walle

GHENT UNIVERSITY

FACULTY OF BIOSCIENCE ENGINEERING

Raman microscopy for phenotyping microorganisms

Raman microscopie voor fenotypering van micro-organismen

Cristina García Timermans

Thesis submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Bioscience Engineering

Please refer to this work as

García-Timermans, C. (2020). *Raman microscopy for phenotyping microorganisms*. PhD thesis, Ghent University, Belgium.

ISBN: 9789463573597

Funding

This work was supported by Qindao Beibao Marine Science & Technology Co. Ltd., Qingdao West-coast economic new area, China, the Flemish Fund for Scientific Research (FWO G020119N) and by the Geconcerteerde Onderzoeksacties (GOA) research grant from Ghent University (BOF15/GOA/006).

Table of Contents

Summary i

Samenvatting i

1 Introduction 1

1.1 How bacteria shape the world . . . 1

1.2 The tree of life . . . 2

1.3 Phenotypic heterogeneity in isogenic bacterial populations . . . 3

1.4 Single-cell tools to study intra-species phenotypic heterogeneity . . . 5

1.5 Raman microscopy . . . 8

1.5.1 Principle . . . 8

1.5.2 Data analysis . . . 13

1.5.3 Current applications in microbial ecology . . . 16

1.5.4 Limitations . . . 17

1.6 Flow cytometry . . . 18

1.6.1 Principle . . . 18

1.6.2 Current applications in microbial ecology . . . 18

1.7 Microbial diversity quantification . . . 20

2 Research objectives **23**

2.0.1 Standardization of label-free Raman microscopy 23

2.0.2 Comparing the resolution of Raman microscopy and FCM to identify single-cell phenotypes 24

2.0.3 Automatic identification of single cell phenotypes based on their Raman spectra 24

2.0.4 Hill numbers to quantify single-cell diversity with Raman spectra 25

2.0.5 Applications of Raman microscopy to estimate nutritionally valuable compounds and detect stress in bioproduction 25

3 Basis for label-free phenotyping with Raman microscopy **27**

3.1 Abstract 27

3.2 Introduction 28

3.3 Materials and methods 29

3.3.1 Inducing phenotypes with different media 29

3.3.2 General fixation procedure 29

3.3.3 The effect of storage time 30

3.3.4 Time on the slide and centrifugation 30

3.3.5 Raman microscopy 31

3.3.6 Data preprocessing 32

3.3.7 Multivariate analysis 32

3.4 Results and discussion 33

3.4.1 Multivariate analyses 34

3.4.2 Data standardization recording: a Raman checklist for microbial phenotyping 39

3.5 Conclusions 41

3.6 Appendix 42

3.6.1 Acknowledgements . . . 42
3.6.2 Conflicts of interest . . . 42
3.6.3 Supplementary information . . . 43

4 Comparing flow cytometry and Raman microscopy 47

4.1 Abstract . . . 47
4.2 Introduction . . . 48
4.3 Materials and methods . . . 51
4.3.1 Cell culture . . . 51
4.3.2 Sample preparation . . . 52
4.3.3 Flow cytometry . . . 52
4.3.3.1 Single-cell analysis . . . 53
4.3.3.2 Community analysis . . . 53
4.3.3.3 Principal component analysis and principal coordinate analysis . . . 54
4.3.4 Raman microscopy . . . 54
4.3.4.1 Raman spectra preprocessing . . . 54
4.3.4.2 Single-cell analysis . . . 55
4.4 Results . . . 57
4.4.1 Flow cytometry . . . 57
4.4.2 Raman microscopy: clustering results . . . 59
4.4.3 Raman microscopy: tentative region assignment . . . 61
4.4.4 Validation of single-cell analysis of Raman spectra . . . 65
4.5 Discussion . . . 66
4.5.1 Flow cytometry quantifies population shifts . . . 66

4.5.2 Raman spectroscopy allows to detect differences in biomolecules across samples . . . 70
4.5.3 The best of both worlds . . . 71
4.5.4 How to define a phenotypic population? . . . 71
4.6 Conclusions . . . 73
4.7 Acknowledgements . . . 73
4.8 Appendix . . . 74
4.8.1 Supplementary information . . . 74
4.8.2 Availability of data and material . . . 79
4.8.3 External dataset . . . 80
4.8.4 Conflicts of interest . . . 80
4.8.5 Author contributions . . . 80

5 Raman spectroscopy-based measurements of single-cell phenotypic diversity in microbial populations 81

5.1 Abstract . . . 81
5.2 Introduction . . . 82
5.3 Materials and methods . . . 84
5.3.1 Data sets . . . 84
5.3.2 Case studies: single-cell phenotypic diversity quantification in stress-induced phenotypes . . . 85
5.3.2.1 Population resolution: *E. coli* exposed to ethanol . . . 86
5.3.2.2 Subpopulation resolution: *S. cerevisiae* after nutrient limitation . . . 86
5.3.3 Raman microscopy . . . 87
5.3.4 Data analysis . . . 87
5.3.4.1 Preprocessing . . . 88

5.3.4.2 Single-cell phenotypic diversity calculation (sc-D_2) for single cells with Raman microscopy . . . 88

5.3.4.3 Statistical analysis . . . 89

5.3.4.4 Principal coordinate analysis (PCoA) . . . 89

5.3.4.5 Sampling size . . . 89

5.3.4.6 Subpopulation types . . . 90

5.3.4.7 Data availability . . . 90

5.4 Results . . . 91

5.4.1 Phenotypic diversity quantification of Raman spectra using Hill numbers . . . 91

5.4.2 Sample size dependence of phenotypic diversity (sc-D_2) measurements . . . 93

5.4.3 Case studies: phenotypic diversity quantification in stress-induced phenotypes 94

5.4.3.1 Tracking *E. coli* population diversification dynamics following exposure to ethanol stress . . . 95

5.4.3.2 Discriminating *S. cerevisiae* subpopulations following exposure to nutrient limitation . . . 96

5.5 Discussion . . . 98

5.6 Conclusions . . . 103

5.7 Appendix . . . 104

5.7.1 Acknowledgements . . . 104

5.7.2 Supplementary information . . . 105

5.7.3 Availability of data and material . . . 108

5.7.4 External data set . . . 108

5.7.5 Conflicts of interest . . . 109

5.7.6 Author contributions . . . 109

6 Raman spectroscopy as a tool for estimating nutritionally valuable compounds in in microbial protein production 111

6.1 Abstract . . . 111

6.2 Introduction . . . 112

6.3 Materials and methods . . . 116

6.3.1 Cell culture . . . 116

6.3.1.1 Calibration . . . 116

6.3.1.2 Enrichment cultures . . . 116

6.3.1.3 Cocultures . . . 117

6.3.2 Calibration . . . 118

6.3.2.1 Amino acid and protein quantification . . . 118

6.3.2.2 Raman dataset . . . 119

6.3.3 16S rRNA amplicon sequencing . . . 119

6.3.4 Flow cytometry . . . 120

6.3.5 Raman microscopy . . . 120

6.3.5.1 Preprocessing . . . 120

6.3.5.2 Biomolecules content . . . 121

6.3.5.3 Statistical analysis . . . 121

6.3.5.4 Minimal sampling size . . . 122

6.3.6 Data availability and reproducibility . . . 122

6.4 Results . . . 122

6.4.1 Calibration . . . 122

6.4.1.1 Amino acid and total protein content estimation . . . 122

6.4.1.2 Sample size . . . 125

6.4.2 Influence of the carbon source in the nutritional profile . . . 126

6.4.3 Influence of cocultivation on the nutritional profile 128
6.5 Discussion 137
6.5.1 Methodological limitations 139
6.5.2 General overview 141
6.5.3 Conclusions 142
6.6 Appendix 143
6.6.1 Acknowledgements 143
6.6.2 Conflicts of interest 143
6.6.3 Data availability 143
6.6.4 Author contributions 144
6.6.5 Supplementary information 145

7 General discussion **153**

7.1 Research outcomes 153
7.1.1 Standardization of label-free Raman measurements for better reproducibility 154
7.1.2 Comparing the resolution of Raman microscopy and FCM to identify single-cell phenotypes 155
7.1.3 Automatic identification of single cell phenotypes based on their Raman spectra 157
7.1.4 Hill numbers to quantify single-cell diversity with Raman spectra 160
7.1.5 Applications of Raman microscopy to estimate nutritionally valuable compounds in bioproduction 161
7.2 Raman based microbial diversity assessment 163
7.2.1 The relevance of microbial diversity 163
7.2.2 Defining phenotypes: how far does the rabbit hole go? 169
7.3 Raman spectroscopy applications in natural and engineered microbial ecosystems 171

7.3.1 Conclusion 175

General conclusions **i**

Afterword **iii**

Acknowledgements **v**

Curriculum vitae **xxxi**

List of Acronyms

A

AA	Acetic acid
A.U.	Arbitrary units
ARI	Adjusted Rand index
AUC	Area under the curve

B

BONCAT	Bioorthogonal non-canonical amino acid tagging

C

CARS	Coherent anti-Stokes Raman spectroscopy
CCD	Charge coupled device
COD	Chemical oxygen demand
CTC-FCM	5-cyano-2,3-ditolyl tetrazolium chloridecombination flow cytometry

D

DAPI	4',6-diamidino-2-phenylindole
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid

E

EDTA	Ethylenediaminetetraacetic acid
EPS	Extracellular polymeric substances

F

FA	Formic acid
FACS	Fluorescence-activated sorting
FISH	Fluorescence in situ hybridization
FCM	Flow cytometry
FCS	Forward scatter
FT-IR	Fourier-transform infrared spectroscopy
FT-MIR	Fourier-transform mid-infrared spectroscopy
FT-Raman	Fourier-transform Raman

G

GABA	Gamma-aminobutyrate
GFP	Green fluorescent protein

H

HOB	Hydrogen oxidizing bacteria

L

LB	Luria Bertani

M

MALDI-TOF MS	Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MERFISH	Multiplexed error-robust fluorescence in situ hybridization
MP	Microbial protein
mRNA	Messenger ribonucleic acid

N

NanoSIMS	Nanoscale secondary ion mass spectrometry
NB	Nutrient broth
NIR	Near-infrared spectroscopy
NMDS	Non-metric multidimensional scaling

O

OD	Optical density
OIU	Operational isolation units
OTU	Operational taxonomic unit
OPU	Operational phenotypic unit

P

PBS	Phosphate-buffered saline
PCA	Principal component analysis

PCoA	Principal coordinate analysis
PFA	Paraformaldehyde
PI	Propidium iodide
PCR	Polymerase chain reaction

R

RACS	Raman-activated cell sorting
RACE	Raman-activated cell ejection
RNA	Ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribosomal ribonucleic acid

S

SG	SybrGreen
SERS	Surface-enhanced Raman spectroscopy
SNIP	Sensitive nonlinear iterative peak
SSC	Side scatter

T

TIC	Total ion current
TERS	Tip enhanced Raman scattering
t-SNE	T-distributed Stochastic Neighbor Embedding

U

UV	Ultraviolet

V

VBNC	Viable but non-culturable

W

WGS	Whole genome sequencing

Summary

Single cell phenotypic differences arise even in monoclonal populations. This allows them to survive, increase their fitness or organize their spatial structure. However, the methods most commonly used to study microbial populations (*i.e.* sequencing techniques and OMICs) analyse all the cells of the same sample together in bulk. Although this is valuable information, bulk techniques only inform on the average behaviour of populations, masking single-cell heterogeneity.

In this manuscript, we discuss the use of the single-cell tool Raman microscopy to study microbial phenotypic heterogeneity. First, we explored how acquiring label-free Raman spectra can be affected due to the sample preparation and collection. We found how delays between fixaton and measuring, the time the sample spends on the slide or the centrifugations made to prepare the sample impact single-cell classification. Therefore, we proposed a standard way to collect Raman metadata for a better experimental interpretation and to increase experimental reproducibility. Secondly, we compared the resolution of Raman microscopy with another single-cell tool, flow cytometry, to identify single-cell phenotypes in isogenic populations. While Raman microscopy describes many variables per cell, it is much less high-throughput than flow cytometry. After testing the resolution of both instruments in retrieving phenotypes in isogenic populations, we found that flow cytometry can detect changes at the population level, whereas Raman microscopy has sufficient resolving power to identify separate phenotypes at the single-cell level. Thirdly, we proposed methods to automatically define phenotypes based on single-cell Raman spectra using dimensionality reduction and clustering algorithms. Then, we discussed how single-cell phenotypic heterogeneity can be quantified applying the information contained in the Raman spectra in the Hill diversity framework, and how this can be used to monitor stress-driven changes in microbial populations. Finally, we show how label-free Raman microscopy can be used to estimate nutritionally valuable compounds in bioproduction.

Raman microscopy presents an opportunity to study phenotypic heterogeneity at the single-cell level and to describe, explain, predict and manage microbial communities.

Samenvatting

Enkelcellige fenotypische verschillen ontstaan zelfs in monoklonale populaties. Dit stelt hen in staat om te overleven, hun conditie te verhogen of hun ruimtelijke structuur te organiseren. Echter, de methoden die het meest gebruikt worden om microbiële populaties te bestuderen (bijvoorbeeld sequentietechnieken en OMICs) analyseren alle cellen van hetzelfde monster samen in bulk. Hoewel dit waardevolle informatie is, informeren bulktechnieken alleen over het gemiddelde gedrag van populaties, waardoor de eencellige heterogeniteit wordt gemaskeerd.

In dit manuscript bespreken we het gebruik van de eencellige Raman microscopie om de microbiële fenotypische heterogeniteit te bestuderen. Eerst onderzoeken we hoe het verwerven van labelvrije Raman spectra beïnvloed kan worden door de monster-voorbereiding en -verzameling. We vonden hoe vertragingen tussen fixatie en meting, de tijd die het monster besteedt aan het objectglaasje of de centrifugaties die gemaakt zijn om het monster te prepareren, de eencellige classificatie beïnvloeden. Daarom stellen we een standaard manier voor om Raman metadata te verzamelen voor een betere experimentele interpretatie en om de experimentele reproduceerbaarheid te verhogen. Ten tweede vergelijken we de resolutie van de Raman microscopie met een ander eencellige tool, flowcytometrie, om eencellige fenotypes in isogene populaties te identificeren. Terwijl Raman microscopie veel variabelen per cel beschrijft, is het veel minder high-throughput dan flowcytometrie. Na het testen van de resolutie van beide instrumenten in het ophalen van fenotypes in isogene populaties, vonden we dat flowcytometrie veranderingen in de fenotypische heterogeniteit op populatieniveau kan detecteren, terwijl Raman microscopie voldoende oplossend vermogen heeft om afzonderlijke fenotypes op het niveau van de eencellige populatie te identificeren. Ten derde stellen we verschillende methoden voor om automatisch fenotypes te definiëren op basis van ééncellige Raman spectra met behulp van dimensionale reductie en clustering. Vervolgens bespreken we hoe enkelcellige fenotypische heterogeniteit kan worden gekwantificeerd met behulp van de informatie in de Raman spectra in het Hill diversiteitskader, en hoe dit kan worden gebruikt om stress-gedreven veranderingen in microbiële populaties te monitoren. Tot slot laten we zien hoe labelvrije Raman microscopie kan worden gebruikt om de voedingswaarde van verbindingen in de bioproductie in te schatten.

Raman microscopie biedt de mogelijkheid om fenotypische heterogeniteit op het niveau van één cel te bestuderen en om microbiële gemeenschappen te beschrijven, uit te leggen, te voorspellen en te beheren.

1

Introduction

1.1 How bacteria shape the world

Microorganisms appeared on Earth around 4 billion years ago and were its only inhabitants for the next 3 billion years. Since their appearance, they have had a dramatic influence on the environment. They changed the atmosphere by releasing great amounts of oxygen, allowing for the existence of multicellular life (Falkowski & Godfrey, 2008). Microorganisms participate in biogeochemical cycles, transforming organic and inorganic matter. Their role in the carbon, nitrogen or sulfur cycle -to name a few- is well known. Their relationship –symbiotic or otherwise- with plants and animals greatly influences the health and disease of their host (Schirawski & Perlin, 2018; McFall-Ngai *et al.*, 2013). These 'small living factories' are capable of transforming one product into another, which open the door to countless applications: food fermentation, crop protection, the production of microbial proteins, improving our gut health or helping to depollute the environment. To better understand how bacteria shape our environment, and to manage and use microbial communities, we need to understand how they organize and function.

1.2 The tree of life

A first step to describe microbial communities is to classify them into categories. This was first done based on physiological and morphological characteristics, until the appearance of molecular techniques changed the classification system in the late 1960s, when DNA-DNA hybridization (DDH) became the standard to measure similarity between genomes (Brenner *et al.*, 1969). It was then defined that a bacterial species consisted of microorganisms that shared at least 70% or more DNA-DNA relatedness (Wayne *et al.*, 1987).

The sequencing of the 16S rRNA gene has allowed to define the taxonomy of an array of organisms and it is also used to define species. This gene is ubiquitous in prokaryotes, has a high functional conservation that can be used to trace evolution, and it contains both conserved regions and hypervariable regions that are used to design amplification primers or to identify bacteria, respectively (Mizrahi-Man *et al.*, 2013). It is considered that a $\sim$97% similarity in the 16S rRNA gene corresponds to 70% DDH, although a higher threshold is needed to differentiate certain species that have a high level of 16S rRNA gene sequence similarity (Kim *et al.*, 2014). In certain cases, only a portion of the full 16S rRNA gene is sequenced, a technology known as 16S rRNA amplicon sequencing. One (or several) of the hypervariable regions are amplified and the sequencing reads are clustered into operational taxonomic units (OTUs) based on their similarity using a threshold of 97-98%. Then, the OTUs are identified by comparing their sequence to a database, such as SILVA (Quast *et al.*, 2012).

Further development of molecular techniques has facilitated sequencing the whole genome of microorganisms, leading to new proposals to classify bacteria. Average nucleotide identity (ANI), which represents the mean of identity/similarity values between homologous genomic regions shared by two genomes, has been proposed as the next-generation standard to delineate species (Kim *et al.*, 2014). An ANI value of $\sim$95% corresponds to a 70% DDH similarity (Vandamme, 2015). Sequencing whole genomes has allowed to find a previously unknown group called the candidate phyla radiation (CPR), that seems to be widespread -it has been found in the human microbiome, drinking wtaer, soil and other niches- and could constitute as much as 50% of all bacterial diversity (Méheust *et al.*, 2019). This phyla is still being defined, but it has already changed the

shape of the tree of life as we know it (Hug *et al.*, 2016).

Defining the genotype of bacteria is crucial to build a phylogenetic tree to understand their ancestry and evolution. On the other hand, phenotypic information is still crucial to generate a useful classification system that can answer (at least partially) what is the function of a certain group of bacteria.

1.3 Phenotypic heterogeneity in isogenic bacterial populations

Bacteria live in complex heterogeneous communities, usually formed by various species with a distinct genetic makeup, or genotype. Individuals from the same species might not be identical -carry the same function, or present the same morphology- as there is variation in the genetic expression of individual cells, known as the phenotype (Table 1.1). Phenotypic heterogeneity increases population survival or fitness, as it allows bacteria to divide tasks and cope with changing environments, important to organize the spatial structure of a community. These phenotypic variations can arise due to stochastic gene expression, periodic oscillations in cellular functions (such as the cell cycle), cellular age or cell-to-cell interactions. Perturbations or fluctuations in the environment also influence gene expression (Avery, 2006; Altschuler & Wu, 2010; Ackermann, 2015).

There are many examples of phenotypic heterogeneity within the same species. For instance, in populations of *Bacillus cereus* only a small subpopulation (1-2%) is responsible for the production of cytotoxin K (Ceuppens *et al.*, 2013). Another example is the “altruistic behavior” found in *Escherichia coli*, where a group of bacteria produce a protective molecule (indole, that turns on drug efflux pumps and oxidative-stress mechanisms) at their own individual cost for their non-resistant neighbours, resulting in an improvement of the overall population survival (Lee *et al.*, 2010). *Bacillus subtilis* can present a distinct lifestyle –a ‘swimming’ or ‘chilling’ phenotype- depending on the expression of a certain epigenetic switch: while in sessile cells the flagella and autolytic enzymes are off, in the motile cells they are on (Chai *et al.*, 2010).

Despite the importance of intra-species heterogeneity, most ecological studies focus on genetic differences. Bacterial populations are often described using 16S rRNA sequencing,

Table 1.1: Definitions of terms used throughout this work. DDH: DNA-DNA hybridization; ANI: average nucleotide identity.

Term	Definition
Phenotype	Observable chraracteristics of an organism
Genotype	The set of alleles of an organism (Van Rossum *et al.*, 2020)
Species	Coherent genomic cluster composed by organisms that have ≥70% DDH similarity (or ~95% ANI) (Kim *et al.*, 2014)
Strain	Set of genetically similar descendants of a single colony or cell. Depending on the field, it can be genetic or phenotypic based (Van Rossum *et al.*, 2020)
Community	Group of organisms that live in a particular location or ecological niche at a given time (modified from Van Rossum *et al.*, 2020)
Population	Group of organisms from the same species that live in a particular location or ecological niche at a given time (modified from Van Rossum *et al.*, 2020)
Isogenic population	Population of genetically identical individuals
Monoclonal population	Population grown from the same single cell through asexual reproduction
Subpopulation	Portion of a population or a community
Phenotyping	Describing observable characteristics or traits amongst bacteria

Next Generation Sequencing (NGS, a technique to study DNA sequences), transcriptomics, metabolomics and/or proteomics. While these techniques allow to understand who is present in the population they have certain limitations when studying bacterial functionalities, as they analyze all the cells from the sample together in bulk, giving an averaged result for the population. Because of the intra-species phenotypic differences, this averaged result can mask relevant information on the dynamics and composition on the population. For instance, it can hide the importance of the functionality of a small subpopulation, or it can define a non-existing averaged population (Fig. 1.1). In the case of transcriptomics, it can be risky to make inferences of genotype-phenotype based solely on mRNA data. The mRNA first has to be translated into a protein, and it is known that the correlation of mRNA and protein abundance is usually weak (Nie *et al.*, 2006): it was found in *E. coli* that a single cell's protein and mRNA copy numbers are uncorrelated for any given gene (Taniguchi *et al.*, 2010). These single-cell OMICs can be used at the single-cell

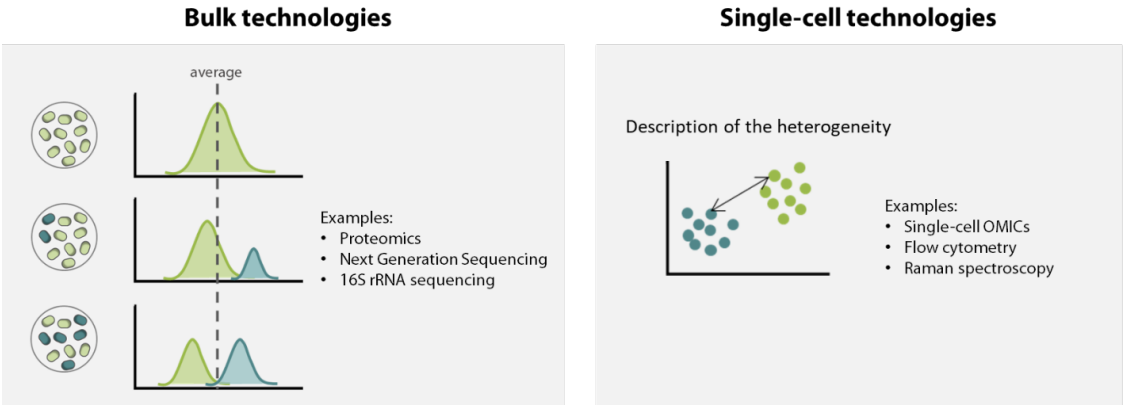


Figure 1.1: Averaging populations can be misleading. Bulk techniques (such as proteomics, Next Generation Sequencing or 16S rRNA sequencing) describe the average of the population, which can be misleading. Single-cell techniques (such as imaging techniques, flow cytometry, Raman spectroscopy or single-cell omics) represent more accurately bacterial phenotypic heterogeneity.

level, although this is technically challenging, as we will explain further in this chapter. Other technologies, such as flow cytometry or spectroscopy, also represent alternatives to bulk technologies.

1.4 Single-cell tools to study intra-species phenotypic heterogeneity

There are several tools available for describing single-cell heterogeneity (Table 1.2). Throughout this work, when referring to the description of observable characteristics or traits amongst microorganisms, we use the term 'phenotyping' (Table 1.1).

Imaging techniques detect a fluorescent label and thus require the use of fluorescent dyes or a tagged bacterial strain. This makes the technique less interesting to study environmental microbial communities, not only because one cannot use a tagged strain, but also because the members of a community are often unknown in advance, thus limiting the choice of labels to 'universal labels', such as nucleic acid stains. The use of (multiple) dyes for bacteria presents more challenges than mammalian dyes, as compared

to them, bacteria present robust cell walls, and have a high internal complexity, but low protein abundances in total (Endesfelder, 2019). It is possible to detect specific sequences in microorganisms using fluorescent in-situ hybridization (FISH). This technique uses a fluorescent probe that will bind to a complementary sequence, and can be used in combination with other single-cell techniques. However, only known dyes for mRNA transcripts can be used making it a less interesting choice for environmental or unknown samples. It is worth mentioning that the most throughout FISH method, called multiplexed error-robust FISH (MERFISH) only allows for the simultaneous identification of 1001 targets, when the expected value of transcript is around 12000 (Huber *et al.*, 2018). Also, a critical review of FISH has shown that the effectiveness of the detection of target cells varies widely from one experiment to another (Bouvier & del Giorgio, 2003).

Other imaging techniques include matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) and nanoscale secondary ion mass spectrometry (NanoSIMS). In MALDI-TOF-MS, the molecules in the sample are ionized, and then accelerated. The time it takes them to reach the detector is used to calculate its mass-to-charge ratio, and to identify the molecule (Nuñez *et al.*, 2018). NanoSIMS detects the secondary ions generated by the impact of a primary ion beam on the sample surface. This results in the ejection of ionized secondary ions that are separated in a mass spectrometer according to their mass-to-charge ratio (Kopp *et al.*, 2015). Although it can be used directly in certain samples (*e.g.*, discussing metal accumulation in microorganisms or looking for isotope fractionation as an indicative of the presence of microorganisms), a stable isotope-labelled species needs to be added first to the sample (Nuñez *et al.*, 2018). This measurement fragments the surface molecules significantly, not allowing to detect information of significant chemical bonds (Kopp *et al.*, 2015). MALDI-TOF-MS requires disruption of the cell wall and nanoSIMS significantly damages the surface of cells, precluding subsequent cultivation (van Belkum A, 2017; Gao *et al.*, 2016).

In the **molecular techniques** we find single-cell omics. 'Omics' is "the large-scale studies of genes (genomics and epigenomics), transcripts (transcriptomics), proteins (proteomics), metabolites (metabolomics), lipids (lipidomics) and interactions (interactomics)" (Wang & Bodovitz, 2010). While understanding single cells to this level of detail represents a breakthrough in microbial ecology, there are still many challenges to overcome. Not only in terms of making the techniques more affordable and faster -improvements that might

arrive over time- but also in terms of the technology itself. The omics techniques that target nucleic acids are largely based on PCR amplification; a technique that can produce errors. How to differentiate these errors from single-cell variants is still a challenge. Also, it is known that single-cell RNA-sequencing has a low detection efficiency (Zhang *et al.*, 2018). Finally, omics require to lyse the samples, making impossible their reuse for further analysis with other techniques.

Optical methods such as FT-IR (Fourier transformed infrared) or Raman spectroscopy are non-destructive. They can both provide fingerprints and (semi)quantitative information on the biomolecular content of single cells. In FT- IR, after a beam with several frequencies passes through the sample, it detects how much has been absorbed by the sample (Naumann *et al.*, 1991), while Raman spectroscopy uses a laser to excite the molecules present in the cell and records their inelastic scattering, that varies depending on their chemical structure (Huang *et al.*, 2010). However, FT-IR is sensitive to water and samples need to be dehydrated (usually at 55℃) prior to analysis. This makes it a poor candidate for online monitoring of cultures or in vivo analysis of aquatic environments (Butler *et al.*, 2016; Chisanga *et al.*, 2018). Another optical tool used to study microbial communities is flow cytometry, a high-throughput technique, able to record thousands of cells per second. Once cells are stained with a dye to extract relevant information (for example, nucleic acids or permeability), they are hydrodynamically focused and passed through a laser. The information about the fluorescence of the dyes, as well as the forward and side scatter of the cells is recorded.

The techniques mentioned in this section can be combined. For instance, fluorescent dyes and fluorescently labelled microbes can be detected by flow cytometry. FISH can be used with Raman spectroscopy, nanoSIMS or flow cytometry (Huang *et al.*, 2007; Musat *et al.*, 2016; Arrigucci *et al.*, 2017). In this work we explore two optical techniques to describe single-cell bacterial heterogeneity: Raman spectroscopy and flow cytometry.

Table 1.2: Summary of single-cell methods for bacterial phenotyping.

Imaging techniques	Samples can be reused	Notes
Labels	Yes	Known target or universal labels
FISH	Yes	Known target
MALDI-TOF-MS	Cell wall disrupted	Label-free
Nano-SIMS	Cell wall disrupted	Isotope labelling
Molecular techniques		
Single-cell OMICs	No	Technical limitations
Optical techniques		
FT-IR	Yes	No aqueous samples, label-free or labelled samples, difficult to interpret in complex samples
Raman spectroscopy	Yes	Label-free or labelled samples, difficult to interpret in complex samples
Flow cytometry	Yes	High throughput, needs fluorescent labelling

1.5 Raman microscopy

1.5.1 Principle

The Raman effect is named after C. V. Raman, who discovered it in 1928 with K. S. Krishnan (Raman & Krishnan, 1928). This effect was observed simultaneously by L. I. Mandelstam and G. S. Landsberg, two scientists from the URSS, and was predicted theoretically in 1923 by A. Smekal, an Austrian scientist that conducted his work in Germany. Only Raman was awarded the Nobel of Physics, causing controversy. In Russia, this effect is known as combination scattering (Singh & Riess, 2001; Masters, 2009).

Raman spectroscopy found its first uses in chemistry, and later found an application in microbial ecology. It does not require labeling cells, and it is non-destructive. This technique is based on the recording of Raman spectra, that results from the inelastic scattering of photons from a molecule. When the laser hits the sample, the molecules of the

sample will be excited to a virtual state. Depending on the vibrational mode of the molecule –its atoms and its molecular bonds- the photons from the laser will gain (anti-Stokes) or lose (Stokes) energy (Clarke & Goodacre, 2003; Butler *et al.*, 2016) (Fig. 1.2). The result will be a spectrum with several peaks that correspond to a particular chemical bond and their vibrations (Fig. 1.3). Conventional Raman spectroscopes are based on Stokes Raman scattering, which is relatively weak as only 1 in 10^6–10^8 photons undergo inelastic Raman scattering (Chisanga *et al.*, 2018). These are the most commonly used instruments at the moment, and it is the one used throughout this research. It is possible to obtain higher vibrational signals using coherent anti-Stokes Raman spectroscopy or (CARS). In this process, two pump beams interact generating a strong anti-Stokes signal (more details on this effect can be found in Evans & Xie (2008)).

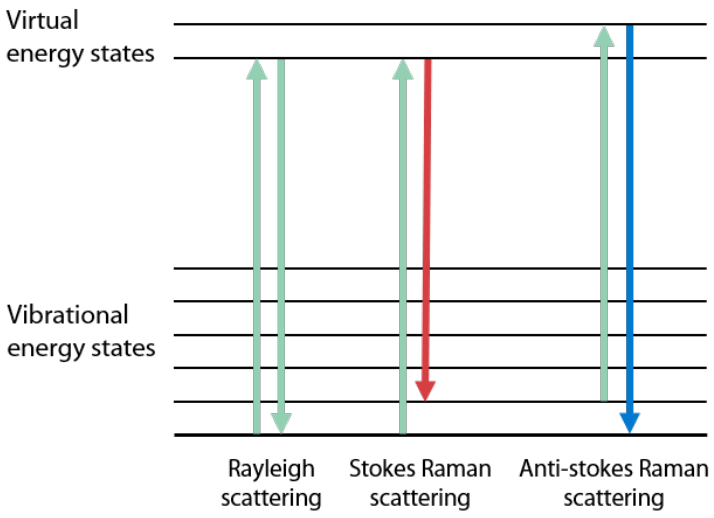


Figure 1.2: Energy-level diagram that showing elastic (Rayleigh) scattering and Raman scattering .

The resulting Raman spectra can be used as a fingerprint to identify bacteria (Goodacre *et al.*, 1998; Willemse-Erix *et al.*, 2009; Kusić *et al.*, 2014). The spectral region used for bacterial fingerprinting is around 500-2000 cm^{-1} (Huang *et al.*, 2010). The spectral information can be also used to obtain semi-quantitative information about the components of the cell (Butler *et al.*, 2016), that can be quantitative if a standard for the molecule(s) of interest is made (Lowery *et al.*, 2017).

The typical configuration of a Raman microscope is shown in Figure 1.4. The laser light is focused on the sample by the objective lens, that collects the scattered light. Filters only allow the Raman scattered light to pass, blocking out other scattered light. Then, the spectrometer separates the light into its components and the signal is collected by a charge-coupled device (CCD) (Schmid & Dariz, 2019). Different excitation wavalengths can be used in Raman spectroscopy: from UV (200-260 nm), to visible light (380-630 nm) to near-infrared (630-1060 nm). Because the Raman scattering intensity is inversely proportional to the fourth power of the excitation wavelength, the higher the excitation frequency, the higher the Raman signal (Tuschel, 2016). Although UV (that has a high frequency) gives a high Raman signal, the radiation can damage the sample. Also, fluorescence occurs mostly when exciting with visible light, therefore choosing a laser in the near infrared can suppress this effect providing a good signal-to-noise ratio (De Gelder *et al.*, 2008). Throughout this manuscript, we used a 785 nm excitation wavelength and a microscope to be able to find the location of the microorganisms.

Raman microscopy presents certain advantages to study both natural and synthetic bacterial communities when compared to other spectroscopic techniques. While FT-IR cannot measure aqueous samples, and requires sample preparation, Raman spectroscopy allows to directly measure bacteria in suspension or in a biofilm, or measure them after fixation (Chisanga *et al.*, 2018). Another popular tool for bacterial fingerprinting is MALDI-TOF MS (Matrix-assisted laser desorption/ionization-Time of Flight Mass Spectrometry). This tool has been proposed for bacterial identification (Singhal *et al.*, 2015), and even phenotypic discrimination in *Staphylococcus aureus* (Majcherczyk *et al.*, 2006), but it has been reported as not sensitive enough for strain-level discrimination in closely related bacterial species such as *Acetinobacter* strains (Rim *et al.*, 2015) *Streptococci* strains (Seng *et al.*, 2009; van Veen *et al.*, 2010) or to discriminate *E. coli* and *Shigella spp.* (Bizzini *et al.*, 2010). Another advantage of Raman spectroscopy is that is it non-destructive, so it can be combined with other methods for further analysis.

However, due to the weak nature of Raman scattering, obtaining the fingerprint of an unlabeled cell is time consuming compared to other techniques (about 30 sec per cell). This signal can be enhanced using metallic nanoparticles, mainly gold and silver but also copper and aluminum. Laser excitation of the nanoparticles creates an enhanced light field, and a large enhancement of the Raman signal of molecules close to this field will follow

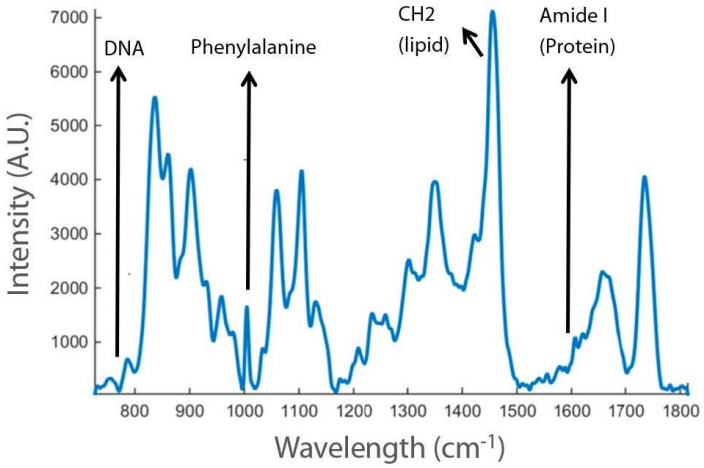


Figure 1.3: Raman spectra of a bacteria and what the peaks correspond to. Modified from Samek *et al.* 2016

(Pilot *et al.*, 2019). These metals can be used in suspension, on a surface (both known as SERS or surface-enhanced Raman spectroscopy), or on the tip of the scanning probe (tip-enhanced Raman spectroscopy, or TERS) (Table 1.3). These techniques increase the Raman signal by 10^6–10^{14} (Lombardi & Birke, 2009), allowing to scan cells in 1-3 sec (Liu *et al.*, 2016). It has been shown how the type of SERS and the protocol followed can greatly influence SERS spectra of bacteria (Mosier-Boss, 2017). Instead of conventional Raman spectroscopy, coherent anti-Stokes Raman spectroscopy (CARS) can be used. This technique uses two laser beams to enhance the Raman signal and increase the signal-to-noise ratio, and allows to use Raman spectroscopy at the sub-micron scale (Chan *et al.*, 2005).

Table 1.3: Summary of methods to enhance the Raman signal.

Acronym	Technique	Description
SERS	Surface-enhanced Raman spectroscopy	Use of metallic nanoparticles in suspension or on a surface to enhance the Raman signal
TERS	Tip-enhanced Raman scattering	Use of metallic nanoparticles on the tip of the scanning probe to enhance the Raman signal
CARS	Coherent anti-Stokes Raman spectroscopy	Two laser beams are used to produce anti-Stokes Raman scattering, which has a more intense signal than Stokes Raman scattering

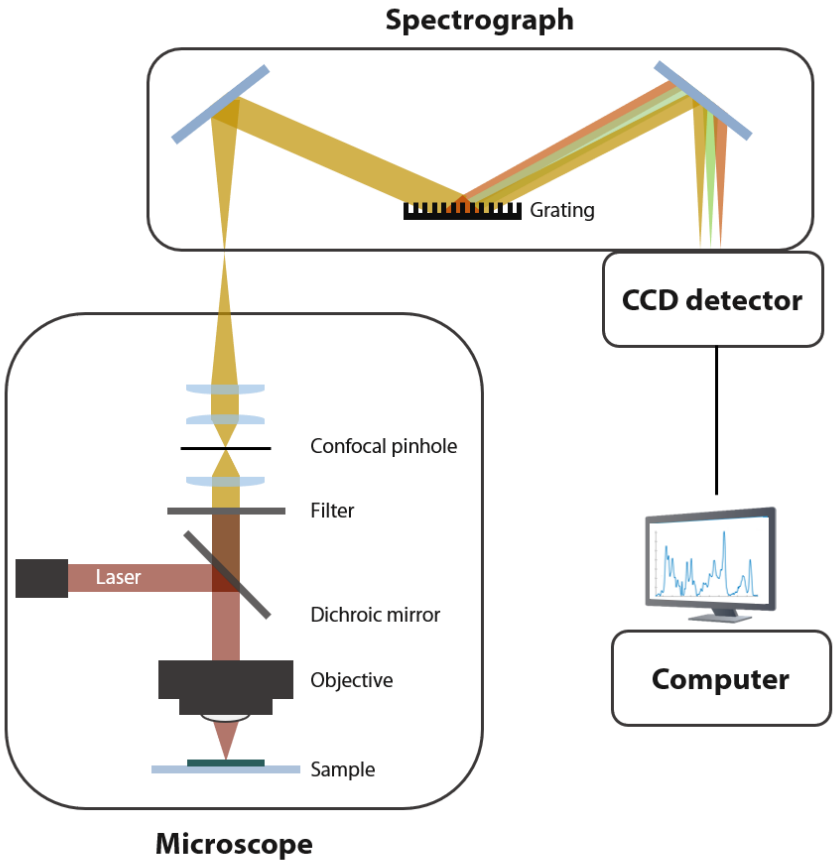


Figure 1.4: Simplified scheme of the configuration of a standard Raman spectroscope. Modified from Schmid & Dariz (2019).

One of the advantages of Raman spectroscopy is that it is non-destructive, so bacteria with a certain spectrum can be sorted out for cultivation or molecular analysis. When a Raman spectroscope is coupled to a sorting system, is referred to as Raman-activated cell sorting (RACS). The cell isolation can happen in a solution using optical tweezers to trap the individual bacteria (Raman tweezers), with a microfluidic chip (microfluidic based RACS) or on a surface (Raman-activated cell ejection or RACE) (Song *et al.*, 2016) (Table 1.4). Raman tweezers can be used in combination with a microfluidic system to move the bacteria of interest into a special reservoir for further evaluation. Lee *et al.* (2019) used this

technique in cells labelled with isotopes, and could sort 3.3–8.3 cells per min. Microfluidic RACS is a faster alternative that can sort between 5-100 cells per sec. The sample needs to be in an aqueous solution, and the cells will pass through a laser one at a time (this technique is analogue to FACS, or fluorescence-activated activated sorting, although FACS can measure thousands of cells per second) (Song *et al.*, 2016). RACE allows to sort in a non-aqueous sample, such as a biofilm, a tissue sample or a solid surface. This method gives laser pulses through a transparent substrate onto a light-absorbing layer (such as water) to disintegrate the layer (evaporate the water) and generate energy to eject the cell. The process takes about 1 sec per cell (Wang *et al.*, 2013).

Table 1.4: Methods for Raman-activated cell sorting (RACS).

Technique	Operated in	Notes
Raman tweezers	Solution	Raman spectroscope coupled to optical tweezers to trap individual bacteria
Microfluidics RACS	Flow	Raman spectroscope coupled to a microfluidic chip
Raman-activated cell ejection (RACE)	Surface	This method gives laser pulses through a transparent substrate onto a light-absorbing layer (such as water) to disintegrate the layer (evaporate the water) and generate energy to eject the cell.

1.5.2 Data analysis

The raw Raman spectra need to be preprocessed before doing any metrics with them (Fig. 1.5). The aim of this step is to take as much noise as possible out of the spectra, to be able to extract relevant biological information from it. First, cosmic rays need to be removed. They come from outer space, and when hitting the atmosphere they produce a cascade of particles that can be detected by CCD cameras, and generate spikes in the Raman spectra (Uckert & Michel, 2019). There are methods to automatically remove these spikes in Raman datasets (Tian & Burch, 2016; Barton & Hennelly, 2019; Uckert & Michel, 2019), but these spectra can be also removed manually. Then, the baseline needs to be corrected. The spectral baseline can be degraded due to instrument fluctuations or background-signal influence (Liu *et al.*, 2015). The spectra also need to be normalized to avoid that the absolute intensity masks the variation of signals of interest (Beattie *et al.*, 2009). There are methods to do the baseline correction and normalization in a single step (Liu *et al.*, 2015). It is possible to smooth the spectra, but this is not without its risks,

as small points in the spectra will be erased. Because there are small deviations in the instrument over time (García-Timermans *et al.*, 2018), spectra can be aligned. However, this step might introduce noise (*e.g.*, by misplacing Raman signals) and should be carefully considered.

Once the spectra have been preprocessed, different information can be extracted. For example, peaks of interest can be selected for semi-quantitative analysis or quantitative analysis using a calibration curve (Butler *et al.*, 2016). Also, the whole spectra can be used to classify cells using several dimensionality reduction and/or clustering methods, such as principal component analysis, principal coordinate analysis or non-metric multidimensional scaling. The distances between cells can be used to construct dendrograms (García-Timermans *et al.*, 2018).

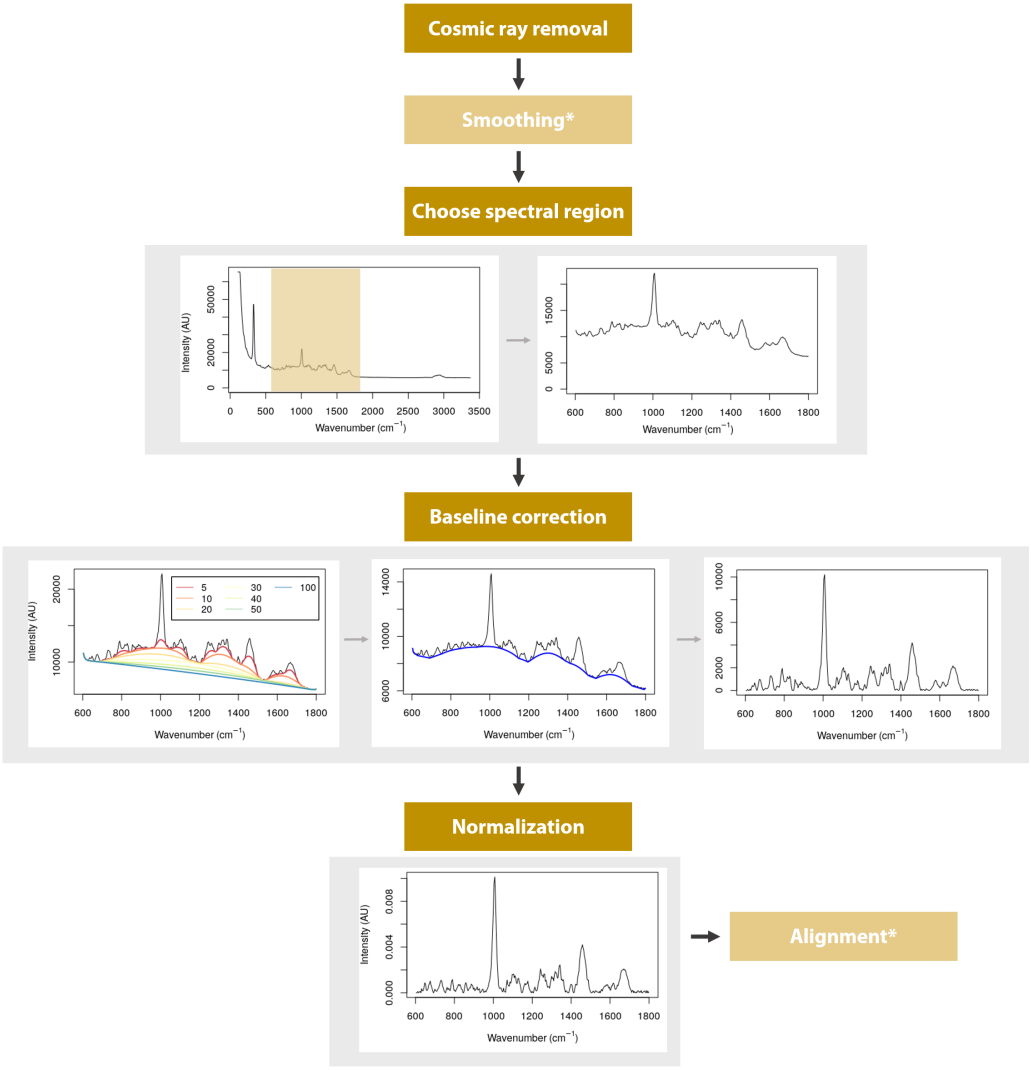


Figure 1.5: Summary of the preprocessing of the Raman spectra. First, the spectra are baseline corrected and normalized. Smoothing and alignment steps can be included. However, smoothing can erase potentially relevant information, and should be carefully considered. Similarly, alignment can produce faulty spectra by displacing the signal, and thus needs to be used reasonably.

1.5.3 Current applications in microbial ecology

The information of the Raman spectra can be used to observe the physiological state of a cell, and determine the production of a certain biomolecule in a (semi)quantitative way. This can be done in unlabeled bacteria (Teng *et al.*, 2016), or using isotope probing (Wang *et al.*, 2016). For instance, it is common to study the production of unlabeled compounds that have a strong Raman signal, such as chlorophylls, carotenoids and other pigments (Jehlička *et al.*, 2014). Also, labelled molecules such as ^{13}C, ^{15}N or deuterium can be used to study respectively the carbon or nitrogen metabolism, or the metabolic rate, in natural or synthetic communities (Muhamadali *et al.*, 2015; Berry *et al.*, 2015). Isotope probing can be coupled to cell sorting to further characterize cells that have a certain metabolism or produce a specific molecule. For example Jing *et al.* (2018) sorted a natural community from the ocean based on the CO_2 fixation capacity of single cells, and then sequenced these subpopulations. This experiment resulted the finding of new CO_2 fixation pathways (Jing *et al.*, 2018).

The Raman fingerprint of cells is often used to identify what strain they belong to. In the public-health field this is useful to detect pathogenic bacteria. For example, Kearns *et al.* (2017) have developed an assay to trap and identify multiple bacteria using SERS to detect food poisoning, and van de Vossenberg *et al.* (2013) have used it in drinking water to discriminate between *Legionella* strains and between *E. coli* and coliform strains. Strain identification is also useful in armed forced operations, to identify potential bioweapons (Pearman & Fountain, 2006), or in space missions. For instance, a Raman spectroscope is included as part of an operation of the European Space Agency to Mars to look for life outside the Earth (Rull & Martínez-Frías, 2006). This tool is a good candidate as samples do not need to be treated or labeled, and the laser does not need to contact the studied rock, diminishing the risk of contamination.

Raman spectroscopy can also be used to identify microbial phenotypes. It is able to discriminate cells from the same population that have been treated with different stressors such as alcohol, metals and antibiotics, or that have been grown in different conditions (Zu *et al.*, 2014; Teng *et al.*, 2016; Tanniche *et al.*, 2020). The spectra from cells treated with antibiotics had enough resolution to distinguish between profiles induced by antibiotics belonging to the same class, making this a powerful tool to predict the functional

class of an unknown antibiotic, identify individual antibiotics that elicit similar phenotypic responses (Athamneh *et al.*, 2014) and determine the antibiotic susceptibility of bacteria (Novelli-Rousseau *et al.*, 2018).

1.5.4 Limitations

Raman spectroscopy presents certain technical disadvantages. Although they are further discussed in the different chapters of this work, we briefly summarize them here.

First, the nature of the Raman effect makes the signal inherently weak. There are certain chemical bonds that do not present a strong Raman signal, while other chemical bonds are present in several molecules, making the discerning of certain metabolites a daunting task. This disadvantage is especially important when handling complex samples, such as microorganisms. Furthermore, certain chemical bonds can have a strong Raman signal and can be overrepresented in the spectra. These limitations are explained in **chapter 6 - Methodological limitations**. Secondly, there can be a shift when measuring the same spectra in different instruments. In **chapter 3** this is further discussed, and a method to record the metadata is proposed. Also, recording the Raman spectra of unlabelled individual cells takes around 30 sec, which is relatively slow compared to other single-cell technologies (such as flow cytometry). In **chapter 4** this difficulty and ways to make faster Raman measurements are discussed.

We propose that there are limitations in the state of the art when clustering microorganisms using their Raman spectra. **Chapters 4 and 5** are dedicated to exploring dimensionality reduction and clustering algorithms that had not yet been used with Raman data. Finally, Raman microscopy can present difficulties for monitoring the dynamics of a population, because when microorganisms grow together, they influence each other's phenotype (Heyse *et al.*, 2019), making difficult the use of databases of axenic cultures in cocultures. This issue is discussed in **chapter 7 - Raman spectroscopy applications in natural and engineered microbial ecosystems**.

1.6 Flow cytometry

1.6.1 Principle

Flow cytometry is a single-cell tool that can be used for bacterial phenotyping. It is a high-throughput technology, able to analyze thousands of bacteria per second. Each cell passes through a laser, and then several detectors collect information on the light scattering -the forward scatter (FSC), and the side scatter (SSC)- or the fluorescence of a specific probe (Davey & Kell, 1996) (Fig. 1.6).

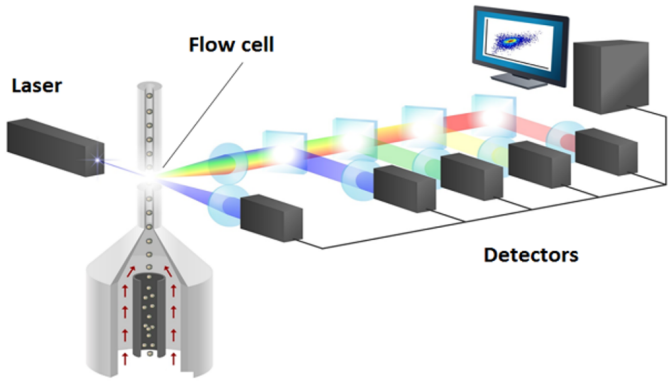


Figure 1.6: Configuration of a standard flow cytometer. Modified from De Roy (2014).

1.6.2 Current applications in microbial ecology

General nucleic acid stains –such as 4',6-diamidino-2-phenylindole (DAPI) , SYTO 9 or SYBR Green -are often used in flow cytometry to quantify bacteria (Button & Robertson, 2001; Virta *et al.*, 1998; Van Nevel *et al.*, 2013). The information derived from these dyes can also be used for bacterial fingerprinting. By applying different gates to the flow cytometric cloud, a 'phenotype' can be defined. Some approaches rely on the manual drawing of these gates (for example, the Dalmatian Plot or CyBar) (Koch *et al.*, 2014).

While manual gating is successful in retrieving functional groups, it is more sensitive to individual experience and error (Koch *et al.*, 2014). Automated gating present a solution that is not only more reproducible, but also makes the analysis less time consuming. There are several tools available, to name a few, flowEMMi , FlowFP or PhenoFlow (Ludwig *et al.*, 2019; Holyst & Rogers, version 4.0; Props *et al.*, 2016). Throughout this work, we used PhenoFlow, whose diversity estimations have been shown to correlate well with 16S rRNA amplicon sequencing results. This method divides the space into an equally spaced grid, and assigns to each bin a density value that corresponds to the probability of one cell having this specific phenotype in the defined bivariate parameter space (Props *et al.*, 2016). The obtained phenotypic fingerprint is then used to calculate diversity metrics, or to identify a certain population (Fig. 1.7).

Cell viability can be determined using FCM. This is often done using a combination of a nucleic acid stain, that will detect all the cells such as SYBR Green, that has a green emission spectrum, with Propidium Iodide, a molecule with a red emission spectrum that will only penetrate the cells whose membrane is permeable (and thus are considered non-viable) (Berney *et al.*, 2007). Other dyes can be used to measure activity, redox potential or membrane potential. For example, bioorthogonal noncanonical amino acid tagging (BONCAT) allows to label newly synthesised proteins and can be used to visualize this synthesis with imaging techniques (Dieterich *et al.*, 2006) and with flow cytometry or FACS (Hatzenpichler *et al.*, 2016).

Flow cytometry can be coupled to cell sorting (known as Fluorescent-activated cell sorting, FACS) to sample a subpopulation out of a community for further study. For example, after finding subpopulations with a distinct expression pattern in the FCM fingerprint of a community, they can be sorted out to do a proteomic analysis, and link them to a certain function (Jahn *et al.*, 2013) did this in a prokaryotic population), or they could be further cultured or analyzed.

In this research, we used flow cytometry to estimate cell density, fingerprint microbial cells and to sort out subpopulations based on the single-cell expression of a fluorescent-labelled reporter.

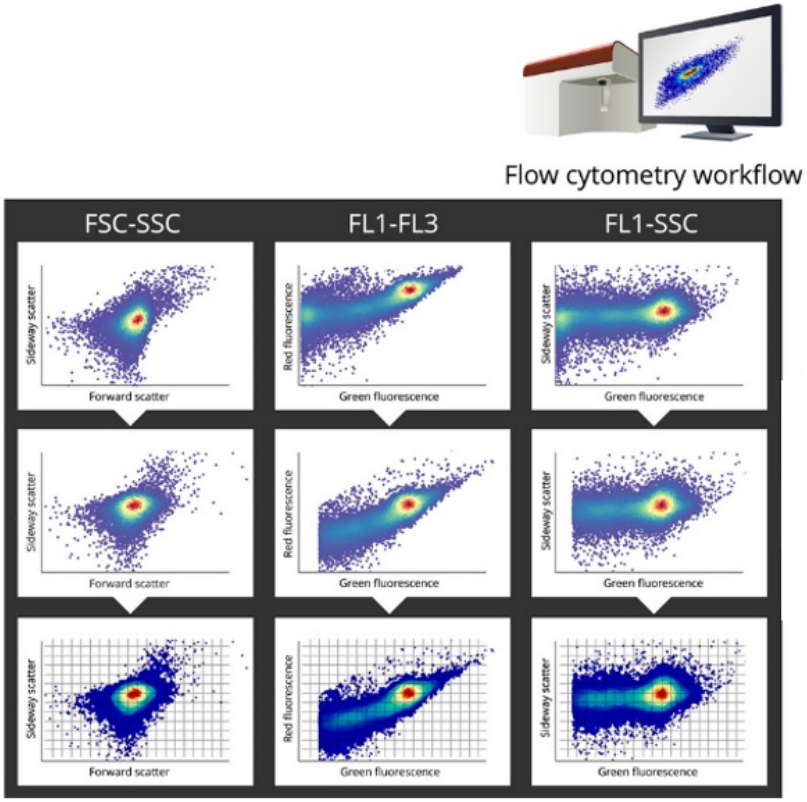


Figure 1.7: PhenoFlow workflow. The flow cytometric data is plotted base on the forward (FSC) and scatter (SSC) signal, the green (FL1) or red (FL3) fluorescence, or FL1 and SSC. Then, the space is divided into a grid and the probability of a cell having that specific phenotype is calculated. Extracted from Props *et al.* 2016.

1.7 Microbial diversity quantification

Microbial ecology studies the relationships of microbes and their environment to describe, explain, predict and control microbial species (Konopka, 2009). Studying microbial diversity -and how it evolves over time- is key to understand community composition, structure, functionality and group dynamics in bacteria.

A common way to measure microbial diversity are sequencing technologies, such as 16S amplicon rRNA sequencing or whole-genome-sequencing (WGS). Although these

tools are increasingly cheaper, they are prone to user bias, as their results are highly dependent on DNA extraction methods, the choice of primers or the analysis pipeline (Fouhy *et al.*, 2016). They are also relatively time-consuming, slow and expensive. As we have seen, other tools such as flow cytometry or mass spectrometry, have been proposed to study diversity in microbial populations (Props *et al.*, 2016; Dumolin *et al.*, 2019a).

Once acquired, the single-cell data allows to study the microbial diversity of the sample. Usually, the multiple single-cell parameters are used to find the cells that are more similar/dissimilar to each other and to cluster them accordingly. This is done through dimensionality reduction and clustering algorithms, and the most common include principal component analysis (PCA), principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS). PCA reduces the multi-dimensional space into a two-dimensional space, to visualize the variance. The other methods first calculate a dissimilarity matrix, and then either calculate the PCA (in the case of PCoA) or look to non-parametric relationships between the points (in the case of NMDS) (Ramette, 2007). The most commonly used metrics in microbial ecology are the Bray-Curtis and the Jaccard dissimilarity, that can handle zeroes (absence of species) in a dataset, and will not consider shared absences as being similar. Bray-Curtis describes community overlap as the fractional minimum abundance of shared taxa between samples, while Jaccard describes the ratio of shared taxa among all observed taxa without considering the abundance information (Schmidt *et al.*, 2016).

While these algorithms give insightful information, they do not provide a quantitative description of the phenotypic diversity in microbial populations: they are respectively dimensionality reduction or clustering tools. A widely used set of metrics to quantify the diversity of microbial communities are Hill numbers, also known as the effective number of species, as they express in intuitive units the number of equally abundant species that are needed to give the same value of the diversity measure. Hill numbers respect important ecological principles, such as the replication principle, that states that in a group with N equally diverse groups that have no species in common, the diversity of the pooled groups must be the N times the diversity of a single group (Chao *et al.*, 2014; Daly *et al.*, 2018). They are commonly used to quantify microbial diversity based on 16S rRNA amplicon sequencing techniques but have also been applied to flow cytometry yielding similar results (Props *et al.*, 2016). These metrics reflect two components that are considered essential

in the description of microbial diversity: richness (*i.e.*, the number of different types of taxa) and evenness (*i.e.*,the distribution of the abundances of the taxa) (Roberts, 2019). The general Hill equation is:

$$D_q = (\sum p_i^q)^{1/(1-q)} \tag{1.1}$$

Where *p* represents the relative abundance of an i number of taxa, and *q* is the sensitivity parameter, also known as the diversity order, which can be 0, 1 or 2. The diversity index of order 0 (D_0, when *q*=0) corresponds to the taxon richness (is insensitive to the species evenness), D_1 weighs each taxon proportionally to their abundance, and D_2 considers both richness and evenness. When *q*=0, D_0 corresponds to the total number of species in the sample. For *q*=1, the result is undefined, but its limit as *q* tends to 1 is

$$D_1 = exp(-\sum p_i \ln p_i) \tag{1.2}$$

The proof of this can be found in (Hill, 1973). And for *q*=2 the Hill equation is

$$D_2 = \frac{1}{\sum p_i^2} \tag{1.3}$$

More information on the diversity measures used in microbial ecology and the advantages of Hill numbers can be found in Chao *et al.* (2014) and Daly *et al.* (2018).

Throughout this work, we use the Hill numbers -specially D_2, as it considers both richness and evenness- to quantify diversity using flow cytometric data. This is the classic use of Hill numbers, that describe the diversity of a population or community. We also propose in **chapter 5** the definition of single-cell phenotypic diversity using the multiparametric information on Raman spectra on the Hill numbers framework.

2

Research objectives

Single-cell microbial ecology presents an opportunity to better understand, describe and steer the interactions and functionalities of microorganisms. This research explores the power of Raman spectroscopy as a tool for single-cell bacterial phenotyping in isogenic populations (Fig. 2.1). The abundant information that Raman spectroscopy can gather of individual cells without the use of any dyes makes it an interesting candidate for describing heterogeneity in microbial populations.

2.0.1 Standardization of label-free Raman microscopy

Problem statement: label-free Raman spectroscopy can record noise as relevant biological information. When detecting single-cell phenotypes with Raman spectroscopy, small spectral differences are being compared. It is known that factors such as fixation or the instrumental variation can affect the Raman spectra.

In **chapter 3**, we test how technical manipulations of an *E. coli* LMG 2092 sample (*i.e.*, storage time, time on slide and centrifugation steps) affect Raman spectra using different multivariate statistical techniques.

2.0.2 Comparing the resolution of Raman microscopy and FCM to identify single-cell phenotypes

Problem statement: flow cytometry and Raman microscopy are two optical tools used to study single-cell phenotypic heterogeneity in bacterial populations. While flow cytometry can record more cells in less time, the Raman spectra contains more single cell information. Also, it is possible that the noise-to-signal ratio of Raman spectroscopy is higher than that of flow cytometry, due to the weaker nature of Raman scattering.

In **chapter 4** we compare the resolution of these optical tools and propose when they should be used. We also propose a computational workflow to automatically distinguish phenotypic populations using Raman microscopy and validate it using an external dataset.

2.0.3 Automatic identification of single cell phenotypes based on their Raman spectra

Problem statement: currently many tools are proposed for dimensionality reduction and clustering Raman spectra, but there is a lack of algorithms to automatically retrieve phenotypes.

In **chapter 4** we identified the phenotypes of *E. coli* LMG 2092 cells in the lag, log and stationary phase using t-SNE. In **chapter 5**, we are able to differentiate the phenotypes of metabolically stressed cells and non-stressed cells using PCA.

2.0.4 Hill numbers to quantify single-cell diversity with Raman spectra

Problem statement: there is a lack of a quantitative methods to compare differences in the metabolic diversity amongst single cells using their Raman spectra.

In **chapter 5**, we introduce a method for describing single-cell phenotypic diversity using the Hill diversity framework with Raman spectroscopy data. Using the biomolecular profile of individual cells, we obtain a metric to compare cellular states and use it to study stress-induced changes in *E. coli* DH5α and *S. cerevisiae* CENPK 113-7D.

2.0.5 Applications of Raman microscopy to estimate nutritionally valuable compounds and detect stress in bioproduction

Problem statement: the bulk quantification of amino acids in microbial protein remains slow and time-consuming.

In **chapter 6** we explore the use of Raman spectroscopy as a single-cell alternative to quantify total protein content and content of the indispensable amino acids. We study how different conditions in microbial protein production (*i.e.*, carbon source and cocultivation) affect the nutritional profile of the final product. In **chapter 5** we use the tools developed in previous chapters tools to identify stress and non-stressed phenotypes and to quantify stress-driven phenotypic heterogeneity, as well as to study the different molecular composition of stressed and non-stressed (sub)populations in a dataset them in two strains commonly used for bioproduction -*E. coli* DH5α and *S. cerevisiae* CENPK 113-7D.

Research objectives

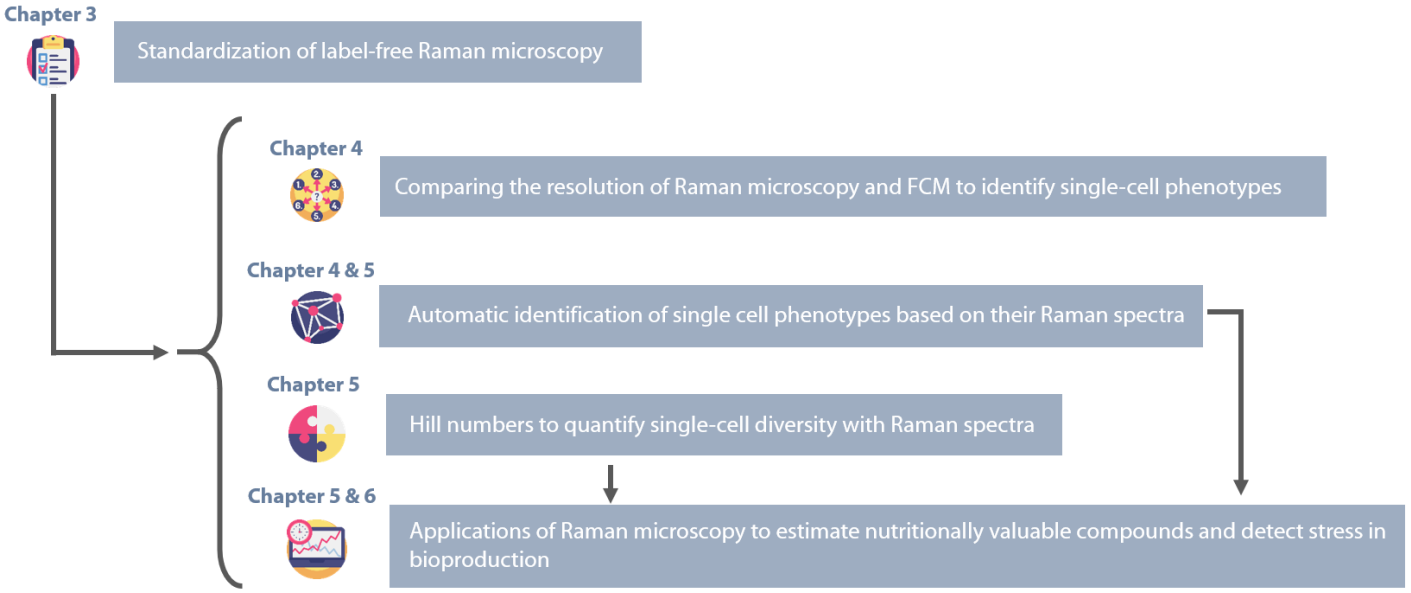


Figure 2.1: Summary of the research objectives and chapters developed in this thesis.

3

Basis for label-free phenotyping with Raman microscopy

3.1 Abstract

Raman spectroscopy has gained relevance in single-cell microbiology for its ability to detect bacterial (sub)populations in a non-destructive and label-free way. However, the Raman spectrum of a bacterium can be heavily affected by abiotic factors, which may influence the interpretation of experimental results. Additionally, there is no publicly available standard for the annotation of metadata describing sample preparation and acquisition of Raman spectra. This chapter explores the importance of sample manipulations when measuring bacterial subpopulations using Raman microscopy. Based on the results of this study and previous findings in literature we propose a Raman metadata standard that incorporates the minimum information that is required to be reported in order to correctly interpret data from Raman spectroscopy experiments. Its aim is twofold: 1) mitigate technical noise due to sample preparation and manipulation and 2) improve reproducibility in Raman spectroscopy experiments studying microbial communities.

Chapter written after: **Cristina García-Timermans**, Peter Rubbens, Frederiek-Maarten Kerckhof, Benjamin Buysschaert, Dmitry Khalenkow, Willem Waegeman, Andre G. Skirtach, Nico Boon. Label-free Raman characterization of bacteria calls for standardized procedures.Journal of Microbiological Methods (2018) doi: 10.1016/j.mimet.2018.05.027

3.2 Introduction

Single-cell technologies have been proposed to observe and characterize phenotypic heterogeneity (Davis & Isberg, 2016). For example, flow cytometry offers high throughput measurements and the possibility to employ numerous dyes that can be used to characterize bacteria (Ambriz-Avina *et al.*, 2014). Imaging techniques can be used to detect gene expression (Ceuppens *et al.*, 2013; Li *et al.*, 2008). Spectroscopic methods, such as Fourier-transform infrared spectroscopy (FT-IR) or Raman spectroscopy are also used to identify bacteria subpopulations (Athamneh *et al.*, 2014; Wehrli *et al.*, 2014).

Raman spectroscopy is an advantageous technology as it can be used without labelling the sample, is rapid and non-destructive, allowing to keep the bacteria alive after the analysis. It detects the inelastic scattering of the molecules present in the sample, resulting in a molecular fingerprint that gives information about lipids, carbohydrates, proteins and nucleic acid content of the bacteria (Huang *et al.*, 2010). With this information, both the structure and metabolic state of individual cells, bacterial species, subspecies and phenotypes can be identified (Davis & Isberg, 2016; Lorenz *et al.*, 2017). The potential of Raman spectroscopy to identify bacteria has aroused interest of the medical, pharmaceutical and defense field (Hakonen *et al.*, 2015; Neugebauer *et al.*, 2015).

The Raman signal is weak – it is estimated that only 1 in 10^8 incident photons are Raman scattered (Jarvis & Goodacre, 2004). To enhance the signal, bacteria can be labelled (*e.g.*, deuterium or isotope probing) and techniques such as Surface Enhanced Raman Spectroscopy (SERS) can be used (Berry *et al.*, 2015; Taylor *et al.*, 2017). However, the signal-to-noise ratio might be too low in unlabelled samples to detect biologically relevant information. Especially when measuring phenotypes in unlabelled samples, this noise could mislead in the result interpretation. It is known from literature that parameters such as laser power, acquisition time and fixation can affect the Raman spectra. While progress has been made towards standardization (Butler *et al.*, 2016; Chen *et al.*, 2014; Guo *et al.*, 2017; Hutsebaut *et al.*, 2005; Rodriguez *et al.*, 2011), there is currently no general protocol available on how to optimally handle bacterial cells for the purpose of identification of subpopulations using a label-free Raman approach. Neither is there a publicly available standard for the annotation of metadata describing the acquisition of Raman spectra.

This study outlines the standardization of label-free bacterial phenotypic identification and investigates the impact of sample manipulations on the analysis of Raman spectra. We evaluated how the different steps in a standard protocol for measuring a sample with Raman microscopy (*i.e.*, the effect of storage time, the time on the slide or the influence of different centrifugation and resuspension steps) influence the spectra. Multivariate statistical techniques, with and without prior knowledge of sample manipulations (*i.e.*, using supervised or unsupervised methods) were used for this. We show that these manipulations induced 'phenotypes' that had no biological relevance, but were identified as separate groups in both the supervised and unsupervised setting. To assist researchers with the annotation of metadata, we combined our results with existing literature on Raman standardization and created a Raman metadata recording tool.

3.3 Materials and methods

3.3.1 Inducing phenotypes with different media

Escherichia coli DSM 2092 was grown in Nutrient Broth (NB, Oxoid) or in Luria Bertani broth (LB, Oxoid) in a shaking incubator at 120 rpm at 28 ℃. Cells were harvested in the stationary phase. To determine the stationary phase, 10^6 cells/mL were inoculated in the media and samples were incubated in the dark for 30 h at 28 ℃, during which optical density measurements were automatically collected each hour using a microtiter plate reader (OD, λ = 620 nm, Tecan Infinite M200 Pro; Tecan UK, Reading, United Kingdom). The growth phases were visually determined after plotting OD over time. The stationary phase was reached in both cultures after 24 h, with a final concentration of approximately 10^8 cells/mL. Three replicates of the cell culture were analyzed for each condition (LB or NB media).

3.3.2 General fixation procedure

After the cultures reached the stationary phase (24 h), bacteria were fixed in 4% formaldehyde (Sigma- Aldrich) dissolved in phosphate-buffered saline (PBS) (protocol

from Bio-Techno Ltd. Belgium). Formaldehyde was chosen as fixation method to preserve the physical characteristics of the cell (Read and Whiteley, 2015). First, 1 mL of the cell suspension was centrifuged for 5 min at room temperature and 1957 g. The supernatant was discarded and cells were suspended in filtered and cold (4 °C) PBS (Thermo-Fisher). The samples were again centrifuged at 1957 g for 5 min at room temperature. The supernatant was discarded and the pellet was resuspended in 0.2 µL filtered formaldehyde 4% (RC Minisart filter, Sigma-Aldrich). The cells were fixed for 1 h at room temperature. Subsequently, the samples were centrifuged at 1957 g for 5 min at room temperature and washed twice with equal volumes of cold PBS. Then, samples were resuspended in Milli-Q water (Merck-Millipore) and four 5 µL drops were put on the CaF_2 slide (grade 13 mm diameter by 0.5 mm polished disc, Crystran Ltd.) and allowed to dry until complete evaporation at room temperature. Samples were resuspended in 1 mL of PBS and stored at 4 °C.

3.3.3 The effect of storage time

To assess how many days bacteria can be stored without inducing changes in their Raman spectra, a sample grown in Luria Bertani (LB) and another sample grown in Nutrient Broth (NB) were harvested and fixed immediately (time 0 h) and measured on that day, after 5 days and after 12 days. They were resuspended in 100 µL of Milli-Q water and four 5 µL drops were put on the CaF_2 slide and allowed to dry until complete evaporation. After sampling, bacteria were resuspended in 1 mL of PBS and stored at 4 °C.

3.3.4 Time on the slide and centrifugation

To investigate the effect of the drying time on the slide of the sample, four 5 µL drops were dried on a CaF_2 slide for 15 min. The slide was kept at room temperature and measured again after 3 h and 6 h. One sample from this batch was centrifuged at 1957 g for 5 min and resuspended in 1 mL of PBS six additional times.

An overview of the different technical manipulations is given in Table 3.1.

Table 3.1: Sample description. Description of the samples produced for every condition. Replicates of the cell culture were made for bacteria grown in Luria Bertani (LB) and nutrient broth (NB). Different storage days, time on the slide and centrifugations were tested.

Growth medium	Replicate number	Days stored	Time on slide	Cells analyzed	Centrifugations
LB*	1	0 days	0 h	45	Standard
		5 days	0 h	38	Standard
		12 days	0 h	39	Standard
NB*	1	0 days	0 h	45	Standard
		5 days	0 h	38	Standard
		12 days	0 h	39	Standard
LB	2	0 days	0 h	45	Standard
LB	3	0 days	0 h	44	Standard
NB	2	0 days	0 h	44	Standard
NB	3	0 days	0 h	45	Standard
LB*	4	0 days	0 h	40	Standard
			3 h	39	
			6 h	40	
		0 days	0 h	40	Extra centrifugations

3.3.5 Raman microscopy

The spectra were measured with a WITec Alpha300R+ spectroscope using a 785 nm laser (Toptica). As a control for the instrument performance, a silicon piece (IMEC, Belgium) was measured with a grating of 600 g/mm, with a 1 sec of acquisition time and 10 accumulations. The intensity of the peak around 520 cm^{-1} was monitored over time. Laser power was also monitored to detect possible variations. Bacteria were measured with a grating of 300 g/mm, with a 40 sec of acquisition time and 1 accumulation. We have found 30-40 sec of acquisition to be most optimal for acquiring the spectra of label-free cells (data not shown). More information on the Raman microscope and data collection is included in the Supplementary Information (see Supplementary Table 3.4).

Three replicates of the cell culture were made for cells grown in in Luria Bertani (LB) or nutrient broth (NB). They are labelled as replicate 1, 2 and 3 respectively. The samples ‘LB replicate 1’ and ‘NB replicate 3’ were stored at 4℃ and analyzed after 5 and 12 days. The sample ‘LB replicate 4’ was spotted on a slide an measured after 3 h and 6 h. From the sample ‘LB replicate 4’ two aliquots were made: one was treated following our standard protocol (see ‘General fixation procedure’), the second followed extra centrifugation steps. We measured the Raman spectra of as many cells as possible in a space of 3 h and manually removed those who had cosmic rays.

3.3.6 Data preprocessing

The obtained spectra were imported as SPC files in R (R Foundation for Statistical Computing, version 3.4.4 (Team, 2015) for preprocessing and analysis. We used the workflow as shown in Fig. 1.5. After manually removing the cosmic rays, the region between 600 and 1800 cm^{-1} -that has most biological significance- was selected using the Hyperspec package v0.98.20161118 (Beleites, 2017). Next, the baseline was estimated using the Sensitive Nonlinear Iterative Peak (SNIP) algorithm with ten iterations and corrected by subtraction. This algorithm gradually corrects a Raman region by replacing its values with the minima in that region. Its advantage is that it allows for (semi)automated background substraction and that it can be used for a variety of background shapes (Tomoyori *et al.*, 2015). Then, the data was normalized using the Total Ion Current (TIC), where the intensity of each point of the spectra is by the mean of all peak intensities of the dataset. Both functions are implemented in the MALDIquant package v1.16.2 (Gibb & Strimmer, 2012).

Raw data can be found in the GitHub repository 'MicroRaman' (Kerchkof *et al.*, 2017).

3.3.7 Multivariate analysis

To investigate the impact of technical manipulations in the Raman spectra, two analyses were performed. The first one in a supervised setting, where the algorithm knows to what group each cell corresponds to, and the second in an unsupervised setting, where the algorithm is naïve to this knowledge. For the supervised setting, a random forest model was used (Breiman, 2001; Boulesteix *et al.*, 2012). First, the classifier was trained on 75% of the data (training set) and predictions were evaluated on the other 25% of the data that was left out (test set). This was done four times, for different non-overlapping sets of held-out data, in order to cover all observations by the test set. In other words, a 4-fold cross-validation scheme was used. The performance was expressed in terms of the accuracy, which is calculated as the fraction of cells that were classified correctly in function of the treatment. Random forests were implemented using default settings, using the R machine learning package 'mlr' (Bischl *et al.*, 2016). The function *randomForest()* was used to evaluate the feature importances, which tells what regions of the spectra were

most important for the random forest classification.

In the unsupervised setting, the analysis was based on the similarity between all spectra. This was calculated based on the spectral contrast angle that measures the angle between two vectors corresponding to closely related spectra to measure whether they are the same or not (Wan *et al.*, 2002). The resulting similarities were clustered by agglomerative clustering with Ward's minimum variance method (ward.D2 from the 'hclust' package) with default settings as linkage from the stats package (Beleites & Salzer, 2008) and visualized as a dendrogram with the iToL interface (Letunic & Bork, 2016). The aforementioned methods were implemented as described in the 'MicroRaman' package (Kerchkof *et al.*, 2017).

To assess whether the entire spectrum was affected because of technical manipulations, the groups were analyzed with the contrast function *ram_contrast()* from 'MicroRaman' (Kerchkof *et al.*, 2017). This function subtracts the intensities for every wavelength across two groups, *a* and *b* (average intensity of group *a* minus *b*). Resulting values were visualized in function of the wavelength.

3.4 Results and discussion

This chapter reports how sources of variation (growth medium, storage time, time on the slide and centrifugation) introduce noise in Raman spectra, influencing bacterial identification. The impact of this bias on multivariate statistical techniques was evaluated in two settings: I) single-cell classification using a supervised machine learning method (random forests) and II) an unsupervised analysis of single-cell Raman spectra using clustering (hierarchical clustering) and dimensionality reduction (PCA).

Three replicates of the cell culture were made for the bacteria grown in Luria Bertani (LB) and nutrient broth (NB), to account for biological variation. To make sure the changes observed were the consequence of technical manipulations, and not due to biological changes (*i.e.*, change in growth phase, nutrient depravation, temperature change or stress), cells were fixed in formaldehyde 4%.

3.4.1 Multivariate analyses

The contrast function shows the difference between the average intensities for each treatment, thus highlighting the regions of the spectra that shift after every treatment (Fig. 3.1). The most intense differences in the spectra (intensity shift $>$ 0.1 A.U.) are shown in Table 3.2. These regions vary across samples, making it difficult to associate technical variations to a specific spectral area. The samples tested for the effect of centrifugation do not show a strong shift.

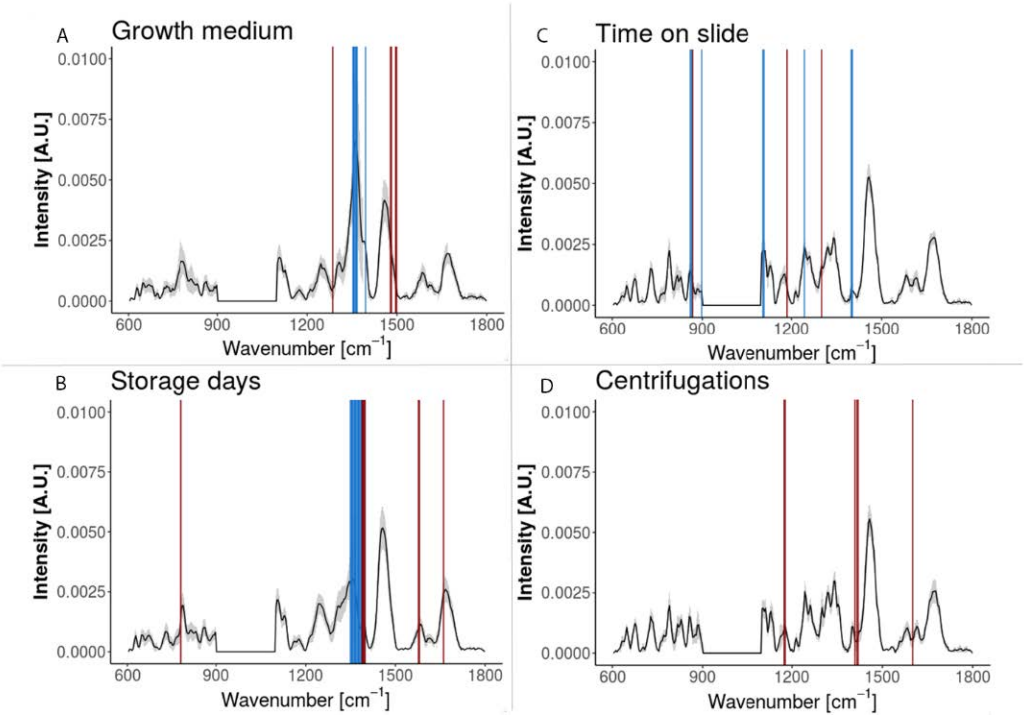


Figure 3.1: Visualization of the random forest and contrast function. This figure represents the spectral areas described in Table 3.2. The random forest results (red lines) are for the five most important regions for classification. The results for the contrast function (blue lines) show the most intense shifts in the spectra ($>$1% intensity). A.U.: arbitrary units.

Random forest could accurately ($>$ 99%) identify variations due to growth media, sample storage, time on the slide and centrifugation. In Table 3.2, the five most important regions for the classification are shown. Most of the regions used by the random forest algorithm to classify groups are also close to those present in the contrast function (Fig. 3.1). While the samples tested for alterations due to centrifugation do not show an intense peak shift in the contrast function, they were recognized by the random forest algorithm. The regions found in the centrifuged samples are similar to those found in those tested for the time on the slide.

Despite the small sample size ($\sim$ 40 cells per condition), the random forest classifier was still able to extract patterns from this data and identify unseen spectra with high accuracy ($>$ 99%). Thus, growth medium, storage, time on the slide and centrifugation produced a marked effect on the spectra. The effect of sample size in single-cell classification is further explored in **chapter 6**, where $\sim$450 Raman spectra are measured in axenic cultures. Although $\sim$50 spectra already give a result close to the population average for most axenic cultures, certain populations -presumably with a higher diversity- need at least 300 measurements (Fig. 6.11).

Hierarchical clustering could identify most of the groups. For the bacteria grown in different media and their replicates, this classification is less accurate and shows dispersion. For the other variables –storage time, time on the slide and centrifugation- the groups are more clearly separated. Especially the time on the slide and centrifugation seem to have great influence over hierarchical clustering analysis.

The results of the principal component analysis (PCA) are in line with the hierarchical cluster analysis (Fig. 3.2). For the variability resulting from different growth media, the groups (LB and NB) can be differentiated, but a great dispersion can be seen, especially for the NB samples (Fig. 3.3A). Cells measured 5 and 12 days after the fixation cluster together, and independently of the non-stored samples (Fig. 3.3B). A shift of the cluster in PCA can be observed for cells that spent 6 h on the slide. The time on the slide and centrifugation seem to have great influence on hierarchical clustering analysis. Yet the explained variance in first two components of PCA is lower. This can be explained with the results from the random forest and the contrast function, that point to various regions of the Raman spectra as a source of variation (Fig. 3.3C and D).

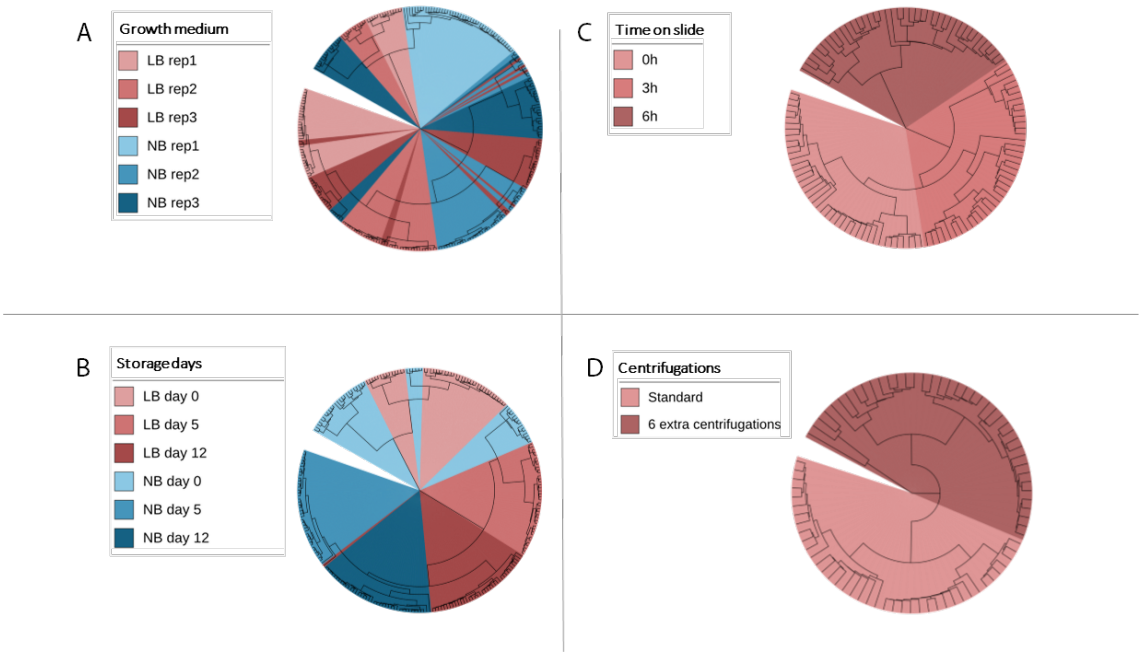


Figure 3.2: Hierarchical clustering analysis for *E. coli* after various treatments. A) 'Growth medium' shows the results for *E. coli* grown in LB and NB, and respective replicates of the cell culture. B) 'Storage days' is the analysis fixed cells that were stored at 4 ℃ and measured again. C) 'Time on slide' shows the results for cells immediately analyzed after being dried on the slide, and 3 h and 6 h later. D) 'Centrifugations' shows the results of cells that received the standard treatment and cells that underwent 6 extra centrifugation steps.

Table 3.2: Summary of random forest results and contrast function. A.U.: arbitrary units.

	Random forest			Contrast function	
Variables	**Prediction accuracy**	**Spectral regions (cm^{-1})**	**Features importance (A.U.)**	**Spectral regions (cm^{-1}) (intensity difference >0.1%)**	**Intensity difference (A.U.)**
Growth medium	99.3%	1286, 1289, 1478, 1496, 1499	3.66, 3.21, 3.09, 3.21, 3.24	1353,1357,1360, 1364,1367,1395	0.12,0.15,0.13, 0.12,0.1,0.11
Sample storage	100 %	778, 782,1395, 1577,1581	10.8,9.31,10.34, 11.33,10.75	1349,1353,1357, 1360,1364,1367, 1371,1374,1378, 1381,1385,1388, 1392,1395,1399	0.12,0.21,0.32, 0.37,0.36,0.33, 0.28,0.21,0.16, 0.12,0.14,0.18, 0.21,0.1,0.16
Drying time on slide	100%	867, 1174, 1185, 1402, 1409	10.02, 9.70, 11.72, 9.61, 9.40	860,863,867, 898,1104,1108, 1243,1399,1402	0.12,0.17,0.15, 0.1,0.12,0.11, 0.13,0.11,0.11
Centrifugation	100%	886,1174, 1409, 1416, 1620	1.30, 1.06, 1.83, 1.56, 1.08	-	-

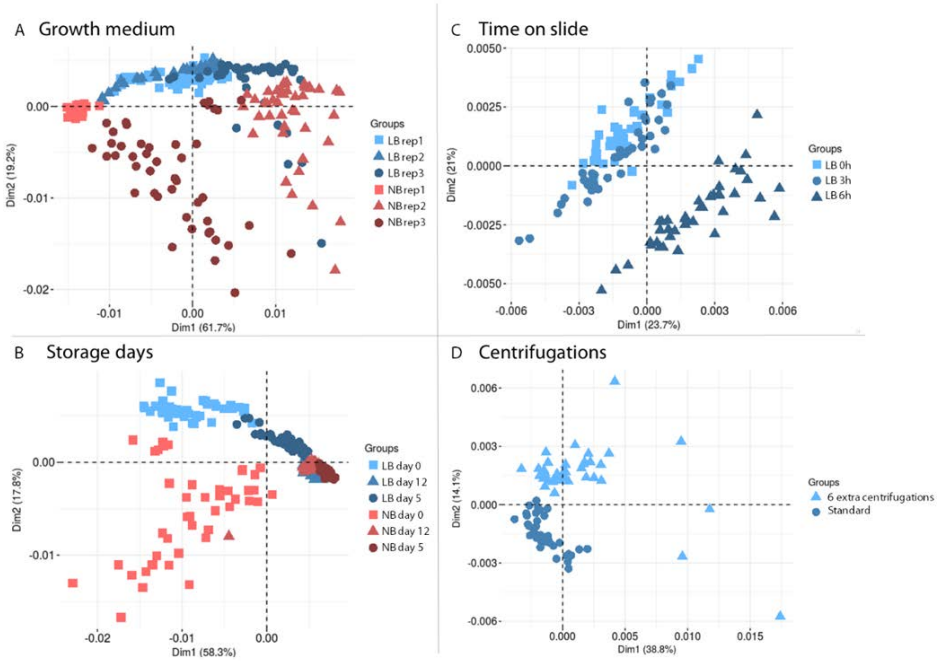


Figure 3.3: PCA of *E. coli* samples under technical variations. A) 'Growth medium' shows the results for *E. coli* grown in LB and NB, and their respective replicates of the cell culture. B) 'Storage days' refers to cells fixed and stored at 4℃. C) 'Time on slide' shows similarity in cells measured after several days. D) 'Centrifugations' show the results of cells treated standardly, or submitted to 6 extra centrifugations.

It could be argued that the effect seen when measuring the same sample over time ('time on the slide') could be due to a slight shift of the Raman spectroscope over the day. It is known that changes in the laser power can influence the spectra (Butler *et al.*, 2016). However, this effect is not observed in the LB and NB replicates, measured on the same day. Although the instrument variations cannot be discarded as a source of variation –and should be taken into account when performing an experiment- it is also reasonable to think that changes in the structure of the cell, such as drying, can be detected by the Raman spectroscope.

During the Raman spectra acquisition, a peak doubling the average intensity showed up at 900–1100 cm^{-1} in a number of the measured cells. This was observed in one vial

which was treated standardly. When preparing other cultures according to the same protocol, it did not show up again. This effect has never been explained in the literature to the best of our knowledge. The corresponding region was therefore assessed as unreliable and removed for further analysis. This should not prohibit the analysis of Rama spectra for bacterial identification, as Kampe and colleagues (Kampe *et al.*, 2017) have shown that a few regions of the spectra should be enough for classification using a supervised algorithm.

3.4.2 Data standardization recording: a Raman checklist for microbial phenotyping

This chapter shows how sample preparation and manipulation prior to analysis can influence the Raman spectra when analyzing label-free bacteria. There are many other publications on the standardization of Raman spectroscopy. Butler and colleagues (Butler *et al.*, 2016) present an extensive protocol proposing sample preparation for different biological samples, as well as instrumental setup, spectra acquisition and data preprocessing. As shown by Read and Whiteley (Read & Whiteley, 2015), it is preferable to use formaldehyde or sodium azide when fixing bacteria for Raman analysis. The work by Hutsebaut *et al.* (2005) describes a calibration protocol for the spectroscope. There are different proposals to reduce the differences amongst instruments (Butler *et al.*, 2016; Chen *et al.*, 2014; Hutsebaut *et al.*, 2005) or across databases (Guo *et al.*, 2017).

Combining the findings of this work and a literature search on sample preparation, instrument setup and data analysis, we present a Raman metadata aid (Table 3.3). It aims to facilitate reporting and improve the reuse of Raman data in further studies and across groups.

Table 3.3: Raman metadata aid. A filled table with the information for this experiment can be found in the Supplementary Information (Supplementary Table 3.4)

Experiment overview	Instrument
1.Hypothesis 2.Variable(s) tested 3.Conclusions 4.Quality control (internal/external)	1.Laser power 2.Silicon piece (quality control) 3.Objective used (magnification) / Numeric aperture (NA) 4.Camera 5.Dry/water/oil objective 6.Model of spectroscope 7.Other specifications (chromatic/flat field correction/other)
Samples and sample acquisition	**Data analysis**
1.Material and source 2.Growing conditions/sampling 3.Filename format 4.Label in the samples 5.Fixation method 6.Integration time 7.Accumulations 8.Grid	1.Background substraction method (if used) 2.Normaliation method (peak/min-max/ area under the curve/other) 3.Smoothing and interpolation (if done) 4.Statistics/machine learning algorithm 5.Accessibility 6.Other relevant information

3.5 Conclusions

When looking for small changes in the Raman spectra of isogenic populations, there is the risk of classifying noise as phenotypic heterogeneity. This chapter proves how changes in the steps used for sample preparation and collection can lead to an incorrect classification of phenotypes. Using a supervised (*i.e.*, random forest) and unsupervised (*i.e.*, hierarchical clustering) algorithm, the impact of the growth medium, sample storage, time on the slide and centrifugation were tested. The delays between cell fixation and testing, the time on the slide and the centrifugations greatly impacted the hierarchical clustering analysis. The effect of the growth medium (LB or NB) had a minor effect. The random forest could identify all groups with high accuracy ($>$ 99%). Taking into account these results, along with the existing literature on Raman spectroscopy standardization, we propose a metadata aid to facilitate reporting and improve data sharing amongst users. Although this is not an extensive list of all the factors that potentially influence Raman spectra, it is a first step for metadata recording in Raman spectroscopy.

3.6 Appendix

3.6.1 Acknowledgements

The authors would like to thank the funding that made possible this research. CGT is funded by Qindao Beibao Marine Science & Technology Co. Ltd., Qingdao West-coast economic new area, China. PR is funded by Special Research Fund (BOF-STA2015000501) from Ghent University. FMK is funded by the Belgian Federal Science Policy (BelSpo) under the inter-university attraction programme " μ -manager" (IUAP P7/25, www.mrm.ugent.be). BB is funded by project grant SB-131370 of the Agency for Innovation by Science and Technology (IWT Flanders). DK is funded by Bijzonder Onderzoeksfonds (BOF) of UGent (24J201400010). AGS acknowledges support of BOF UGent. We would like to thank Ruben Props for supporting the data analysis, and Jasmine Heyse and Charlotte de Rudder, whose suggestions helped improve and clarify this chapter.

3.6.2 Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.6.3 Supplementary information

Table 3.4: Raman metadata aid

Experiment overview	
Hypothesis	Applying different protocols to process samples will lead to different spectra
Variable(s) tested	Replicates (reproducibility) in different media (phenotypes): spinning time, drying time on the slide; storage time (time spent at 4℃ before analysis)
Conclusions	Hypothesis supported by the data. Researchers studying phenotypes need to carefully follow the same protocol, and keep the least space between analysis possible.
Quality control (internal/external)	Silicon piece check
Samples and sample acquisition	
Material and source	*Escherichia coli* DSM 2092: the number of cells can be found in Table 3.1
Growing conditions/sampling	Cells were grown at 28℃, 120 rpm
Filename format:	<cellnumber>_<bacterium type or strain number>_ <treatment or condition>_ <dayrecorded>_ <integration time>.spc Avoid spaces and non-alphanumerical characters
Label in the samples	No label used
Fixation method	Filtered PFA 4%
Integration time	40 second
Accumulations	1
Grid	300 g/mm
Instrument	
Laser power	785 nm excitation diode laser (Toptica)
Silicon piece (quality control)	Before objective/after objective or just the total laser power: stability regarding other days was checked
Objective used (magnification) / Numeric aperture (NA)	100x/0.9 NA (Nikon)
Camera	-70℃ cooled CCD camera (iDus 401 BR-DD, ANDOR)
Dry/water/oil objective	Dried samples
Model of spectroscope	WITec Alpha300R+
Other specifications (chromatic/flat field correction/other)	
Data analysis	
Background subtraction method (if used)	No. Repeated measurements with cosmic rays
Normalization method (peak /min-max /area under-curve /other)	Area under the curve ('Total Ion Current')
Smoothing and interpolation (if done)	Baseline correction
Statistics/Machine learning algorithm	'MicroRaman' package (GitHub). Spectral contrast angle, ward.D2 dissimilarity and hierarchical clustering. Random forest
Accessibility	Repository: 'MicroRaman', GitHub
Other relevant information	Peak selection 900-1100 cm^{-1} removed for unexplained variation

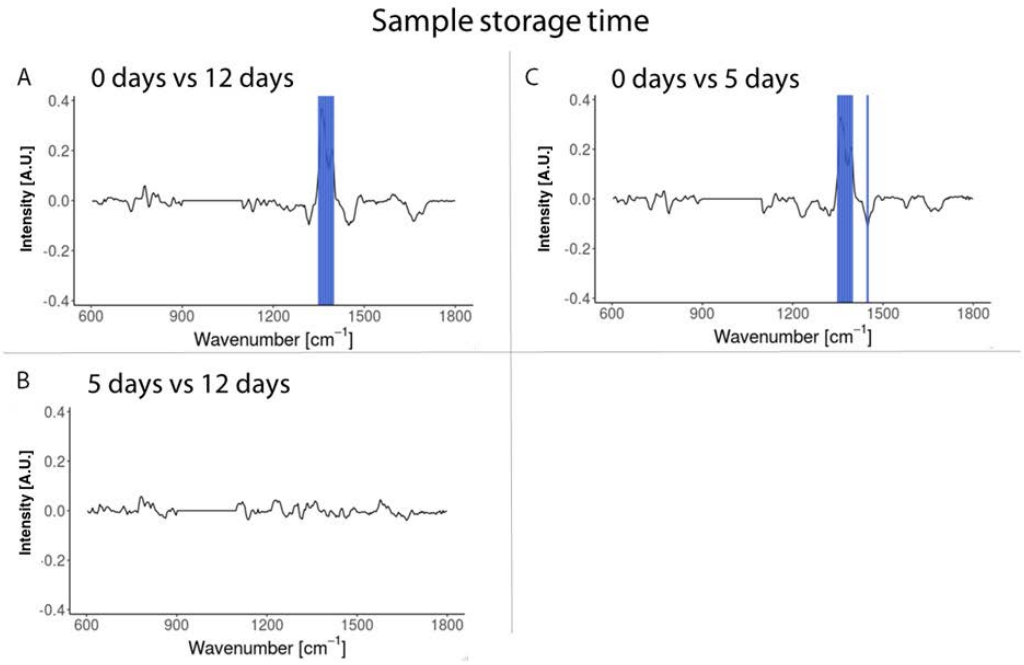


Figure 3.4: Contrast function to highlight the spectral regions that differ when storing and re-analyzing the same sample. Results for: A) 0 days vs 12 days ; B) 5 days vs 12 days ; C) 0 days vs 5 days.

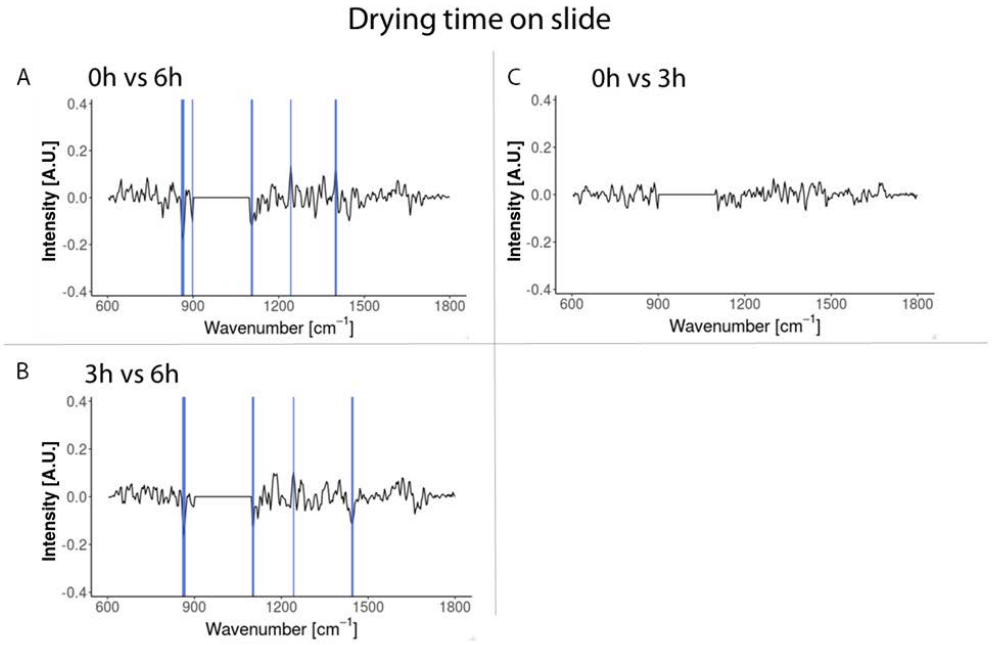


Figure 3.5: Contrast function to highlight the spectral regions that differ according to the time the samples were on the slide. Results for: A) 0 h vs 6 h ; B) 3 h vs 6 h ; C) 0 h vs 3 h.

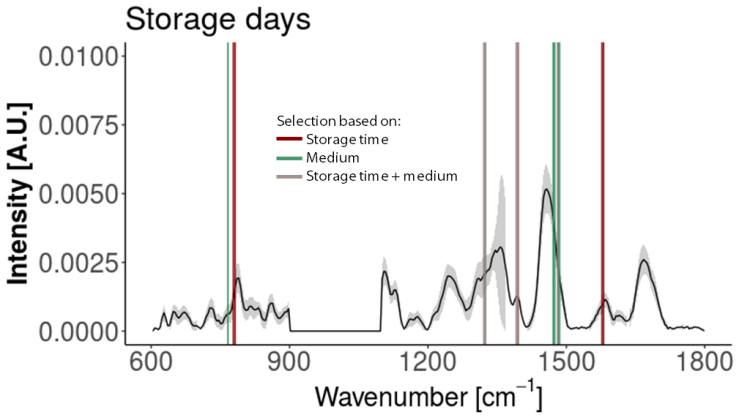


Figure 3.6: Five most important regions for classification according to the random forest classifier. Classification made according to storage time (red), medium (LB or NB, green) or storage time and medium (grey).

4

Comparing flow cytometry and Raman microscopy

4.1 Abstract

Investigating phenotypic heterogeneity can help to better understand and manage microbial communities. However, characterizing phenotypic heterogeneity remains a challenge, as there is no standardized analysis framework. Several optical tools are available, such as flow cytometry and Raman microscopy, which describe optical properties of the individual cell. In this work, we compare Raman microscopy and flow cytometry to study phenotypic heterogeneity in bacterial populations. The growth stages of three replicate *Escherichia coli* populations were characterized using both technologies. Our findings show that flow cytometry detects and quantifies shifts in phenotypic heterogeneity at the population level due to its high-throughput nature. Raman spectroscopy, on the other hand, offers a much higher resolution at the single cell level (*i.e.*, more biochemical information is recorded). Therefore, it can identify distinct phenotypic populations when coupled with analyses tailored toward single-cell data. In addition, it provides information about biomolecules that are present, which can be linked to cell functionality. We propose a computational workflow to distinguish between bacterial phenotypic populations using Raman microscopy and validated this approach with an external dataset. We recommend using flow cytometry to quantify phenotypic heterogeneity at the population level, and Raman spectroscopy to perform a more in-depth analysis of heterogeneity at the single cell level.

Chapter written after: **Cristina García-Timermans***, Peter Rubbens*, Frederiek-Maarten Kerckhof, Benjamin Buysschaert, Dmitry Khalenkow, Willem Waegeman, Andre G. Skirtach, Nico Boon. Discriminating Bacterial Phenotypes at the Population and Single-Cell Level: A Comparison of Flow Cytometry and Raman Spectroscopy Fingerprinting. Cyto A (2019) doi:10.1002/cyto.a.23952

** The authors contributed equally to the manuscript*

4.2 Introduction

Single-cell phenotypic differences arise even in genetically identical cultures (Govers *et al.*, 2017). A definition of a phenotypic population is an observed cellular state within a given genetic and environmental background. It arises due to epigenetic variations, stochastic gene expression, cellular age or oscillations such as the cell cycle. This is one of the strategies that bacteria use to adapt to a changing environment, as well as to divide the labor within the community or population (Avery, 2006; Ackermann, 2015).

Phenotypic heterogeneity in laboratory cultures is well documented. For example, it has been studied in bacterial subpopulations that could tolerate antibiotics (known as persisters) (Dhar & McKinney, 2007), in the production of cytotoxin K in *Bacillus cereus* (Ceuppens *et al.*, 2013) or in the differential expression of flagellin in *Salmonella typhimurium* (Stewart *et al.*, 2011). The challenge remains to find tools to measure and quantify this heterogeneity (*i.e.*, phenotypic populations), in order to be able to link it with bacterial functionality. This would allow to manage—and potentially steer—microbial communities in order to optimize bioprocesses.

Several tools are available for single-cell phenotyping (Davis & Isberg, 2016). Imaging techniques are popular, but they require tagged bacterial cells or a probe to visualize the bacteria or the molecule of interest (Taniguchi *et al.*, 2010; Anetzberger *et al.*, 2012), making them less suitable to study environmental communities. There are other techniques that do not require a probe, such as intrinsic fluorescence (Georgakoudi *et al.*, 2007) or the detection of autofluorescent NAD(P)H (Bhattacharjee *et al.*, 2017). However, the amount of information that can be gathered is limited compared to other techniques, such as transcriptomics, flow cytometry, or spectroscopy techniques. Single-cell transcriptomics are also an option for bacterial phenotyping, but a few hundred cells are needed and only about 15–25% of the expressed mRNAs can be detected (Tang *et al.*, 2011). This analysis requires for bacteria to be lysed, and it was found in *E. coli* that a single cells' protein and mRNA copy numbers are uncorrelated for any given gene (Taniguchi *et al.*, 2010).

A more high-throughput option for single-cell analysis is flow cytometry, that can measure thousands of bacterial cells per second. Individual cells pass through a laser, after which detectors collect information on the scattered laser light -forward scatter (FSC)

and side scatter (SSC)- and on autofluorescence and/or emission of specific fluorescent probes (Davey & Kell, 1996). To detect bacteria, general nucleic acid stains (such as SYBR Green I or 40,6-diamidino-2-phenylindole (DAPI)) can be used (Koch & Müller, 2018). Flow cytometry allows to quantify cells and to identify different phenotypes in bacterial populations. For example, this technique allowed Sanchez-Romero & Casadesus (2014) to find a differential expression of a GFP-tagged gene related to antibiotic resistance in a *Salmonella enterica* population and Cronin & Wilkinson (2008) to detect a heterogeneous response of *Bacillus cereus* endospores to different heat treatments. Furthermore, the information derived from the FCM measurements can be transformed into a fingerprint and can be used to calculate inter- and intra-species variations in bacterial communities and populations (De Roy *et al.*, 2012; Koch *et al.*, 2014; Props *et al.*, 2016). FCM can also be used for bioprocess monitoring, as it allows to quantify the number of cells present in a reactor, their viability and activity, as well as their membrane potential over time (Díaz *et al.*, 2010). When this technique is coupled to cell sorting (also known as FACS or fluorescence-activated cell sorting), a follow-up analysis on the subpopulations can be made. For example, by doing a proteomic analysis to link these phenotypes to a certain functionality (Jahn *et al.*, 2013), to further culture the cells, or by doing single-cell microscopy analysis (Nebe-von Caron *et al.*, 2000).

Raman spectroscopy is another single-cell technology that has been proposed to study phenotypic heterogeneity. It does not require labeling and it is nondestructive. The laser excites individual cells, which leads to inelastic scattering, which in turn is collected in the form of Raman spectra. Because Raman scattering is weak (only 1 in 10^8 incident photons are Raman scattered) (Jarvis & Goodacre, 2004), collection times can be high (around 30 sec per cell or more). The Raman signal can be enhanced with metallic particles (known as surfaced-enhanced Raman spectroscopy or SERS), which reduces the acquisition time to 1–3 sec per cell (Liu *et al.*, 2016). The resulting spectrum contains biochemical information of the molecules that are present in the cell— for example, lipids, carbohydrates, nucleic acids, and proteins—and can be used to classify bacteria according to phylogeny (Goodacre *et al.*, 1998). This information can be quantitative if an internal standard for the molecule(s) of interest is made. For example, Cowcher et al. (Cowcher *et al.*, 2013) quantified the dipicolinate (DPA) biomarker for *Bacillus* spores, and Samek *et al.* (2016) quantified polyhydroxyalkanoates produced by *Cupriavidus necator* H16. Raman spectroscopy can also be linked to cell sorting, known as Raman activated cell

sorting, to further study phenotypic subpopulations (Zhang *et al.*, 2015).

Raman spectroscopy can be used for the monitoring of bioprocesses, as it can measure compounds that are present in the supernatant such as glucose, protein production, or others over time (Lee *et al.*, 2004), as well as some Raman reactive compounds present in the bacteria, such as chlorophylls, carotenoids, and other pigments (Jehlička *et al.*, 2014). Although this technique is used for the identification of bacterial strains and species (Huang *et al.*, 2010; Almarashi *et al.*, 2012; Strola *et al.*, 2014; Pahlow *et al.*, 2015), the potential of Raman spectroscopy to automatically differentiate unknown phenotypic subpopulations remains relatively little explored.

Both flow cytometry and Raman spectroscopy give rise to data that need distinct preprocessing and analysis (O'Neill *et al.*, 2013; Saeys *et al.*, 2016; García-Timermans *et al.*, 2018; Ryabchykov *et al.*, 2018). While microbial flow cytometry is rather limited in its phenotypic resolution (*i.e.*, only a few properties are measured per cell), Raman spectroscopy characterizes many more biochemical properties of bacterial cells. It therefore requires analysis of high-dimensional data, which can be challenging, but it allows to characterize phenotypic heterogeneity at a much higher resolution. It is worth mentioning that although Raman has more parameters, its signal-to-noise ratio could be lower than that of flow cytometry (due to the weaker nature of Raman scattering) and this should be also considered when comparing resolutions.

In this chapter, we have analyzed bacterial cells from nine phenotypic populations—with a different growth stage and/or from a different replicate—using flow cytometry and Raman spectroscopy. We first compare the resolution of two visualization tools that can reduce the dimensionality of datasets to find clusters (phenotypes): principal component analysis (PCA), a linear tool that is widely used in microbial ecology and t-distributed stochastic neighborhood embedding (t-SNE), that reduces the dimension of each point so that similar objects are modelled by nearby points. We also propose the clustering algorithm t-SNE, that can automatically retrieve the number of phenotypes and classify the cells. Once populations have been determined, we illustrate how the Raman spectra of cells can be used to perform metabolic inference using a machine learning based variable selection algorithm. Finally, the advantages and disadvantages of these tools for microbial phenotyping are discussed. We motivate that, in its current form, microbial flow

cytometry can be used to quantify phenotypic heterogeneity and describe population-level dynamics, while Raman spectroscopy can be applied to describe single-cell heterogeneity and possibly identify distinct phenotypic subpopulations. We include a recommendation for microbiologists on how to employ Raman microscopy and flow cytometry in future phenotyping studies.

4.3 Materials and methods

4.3.1 Cell culture

To determine the growth stages of the cell culture (lag, log, and early stationary phase), *E. coli* LMG 2092 was grown in nutrient broth (NB; Oxoid, United Kingdom) at 28℃, 120 rpm shaking and then inoculated in the same medium and conditions in three replicates. Cultures had an initial concentration of 106 cells/ml, measured with a BD Accuri C6 flow cytometer (BD Biosciences), following the protocol from Van Nevel (Van Nevel *et al.*, 2013). The samples were incubated in the dark for 30 h at 28℃ in a 96-well plate, during which optical density (OD, λ = 620 nm) measurements were automatically collected each hour using a microtiter plate reader (Tecan Infinite M200 Pro; Tecan UK, Reading, United Kingdom). To avoid evaporation, the wells around the cultures were filled with phosphate-buffered saline (PBS). The growth phases were assigned after fitting the results with the function *SummarizeGrowth()* from the "Growthcurver v0.30" R package (Sprouffske & Wagner, 2016). This package fits the growth curve data to the equation displayed, where N_t is the number of cells (or the absorbance reading) at time t, N_0 is the initial cell count (or absorbance reading), K is the carrying capacity, and r is the growth rate.

$$N_t = \frac{N_0 K}{N_0 + (K - N_0)e^{-rt}}$$

Cells were harvested 1 h, 7 h 30 min, and 24 h after inoculation, visually labeled as the lag, log, and early stationary phases of *E. coli* (see Supporting Information Fig. 4.7). Nutrient broth was included as a negative control.

4.3.2 Sample preparation

Samples were measured immediately in the flow cytometer after sampling. For Raman microscopy, samples were harvested and fixed in formaldehyde 4% (Sigma-Aldrich, Merck KGaA, Darmstadt, Germany) dissolved in PBS (protocol from Bio-Techno Ltd., Belgium) following the protocol from the study by García-Timermans et al. (García-Timermans *et al.*, 2018). Paraformaldehyde was chosen as it preserves spectral features better than ethanol and glutaraldehyde (Read & Whiteley, 2015). First, 1 ml of the cell suspension was centrifuged for 5 min at room temperature and 1,957 g. For the samples in the lag phase, up to 10 ml were suspended until a pellet could be seen. The supernatant was discarded and cells were suspended in filtered and cold PBS (4℃). The samples were again centrifuged at 1,957 g for 5 min at room temperature. The supernatant was discarded and the pellet was re-suspended in filtered formaldehyde 4%. The cells were allowed to fix for 1 h at room temperature (21℃). Then, the samples were centrifuged at 1,957 g for 5 min at room temperature and washed twice with cold PBS (4℃). Cells were stored at 4℃ and analyzed with the Raman spectroscope within 1 week.

4.3.3 Flow cytometry

Fresh samples taken at the lag, log, and early stationary phase were diluted in filtered PBS and stained with SYBR Green I 1% (Thermo Fisher) during 13 min at 37℃. They were measured with the flow cytometer BD Accuri C6 (BD Biosciences). This resulted in a multivariate description of each cell by four fluorescence detectors (FL1: 533/30 nm, FL2: 585/40 nm, FL3: $>$ 670 nm long pass, FL4: 675/25 nm), of which the FL1 detector was targeted by SYBR Green I, and two scatter detectors (forward scatter, FSC and side scatter, SSC). The channels FSC-H, SSC-H, FL1-H, and FL3-H were used for data analysis.

4.3.3.1 Single-cell analysis

t-distributed stochastic neighborhood embedding (t-SNE) t-SNE is a dimensionality reduction technique developed for the visualization of high-dimensional data (Van Der Maaten & Hinton, 2008). The *TSNE()* function from the 'scikit-learn' machine learning library was used (Pedregosa *et al.*, 2011), v0.19.1. PCA was set as initialization method. TSNE was run with default settings unless reported otherwise. Data were first transformed by the function *f(x) = asinh(x)*, and normalized so that each channel has a mean of zero and standard deviation of one.

Principal component analysis Flow cytometric single-cell data were analyzed with the *PCA ()* function from the scikit-learn machine learning library after normalization. Data were first transformed by the function *f(x) = asinh(x)* and normalized so that each channel has a mean of zero and standard deviation of 1.

4.3.3.2 Community analysis

The PhenoFlow R package (Props *et al.*, 2016) was used for the analysis. Four channels (FL1-H, FL3-H, FSC-H, and SSC-H) were selected to derive a phenotypic fingerprint for each sample. Bacteria were gated to differentiate them from background noise as shown in the Supporting Information Figure 4.8. As quality control, the stability of the FL1 signal over time was checked. A 128 × 128 binning grid was constructed for each pairwise combination of these channels (resulting in six in total). Next, a bin a kernel density estimation was performed to determine the density per bin (with a Gaussian kernel density bandwidth of 0.01). Then, all bins are concatenated to a one-dimensional vector, representing the cytometric fingerprint. Data were transformed by the function *f(x) = asinh(x)* transformation. At least 10.000 cells were measured per sample.

4.3.3.3 Principal component analysis and principal coordinate analysis

The pulled information for every group was analyzed with the function *fviz_pca_ind()* from the R package "factoextra" (Kassambara & Mundt, 2017).

PCoA, also known as multidimensional scaling, was calculated based on the Bray–Curtis dissimilarities between all fingerprints. The function *beta_div_fcm()* from the R package "PhenoFlow" was used (Props *et al.*, 2016).

4.3.4 Raman microscopy

Fixed samples were centrifuged at 1,957 g for 5 min at room temperature and re-suspended in cold Milli-Q water (MerckMillipore) (4 ℃). Then, a 5 µL drop was allowed to dry until evaporation on a CaF_2 slide (grade 13 mm diameter by 0.5 mm polished disk; Crystran Ltd). At least 60 single cells were measured per biological replicate. As control for the instrument performance, a silicon piece (IMEC, Belgium) was measured with a grating of 600 g/mm, with a 1 sec acquisition time and 10 accumulations. The intensity of the peak around 520 cm^{-1} was monitored over time. Laser power was also monitored to detect possible variations. Bacteria were measured with a grating of 300 g/mm, with a 40 sec acquisition time and 1 accumulation. More information on the Raman microscope and data collection are included in the Raman aid. The metadata were reported following the guidelines from García-Timermans *et al.* (2018) and can be found in the Supporting Information Table 4.3.

4.3.4.1 Raman spectra preprocessing

The Raman spectra were analyzed in the 600–1800 cm_{-1} region, and baseline correction using the SNIP algorithm (10 iterations) and normalization were performed. The area under the curve (AUC) normalization was calculated with the MALDIquant package (v1.16.2) (Gibb & Strimmer, 2012).

4.3.4.2 Single-cell analysis

t-distributed stochastic neighborhood embedding (t-SNE) Raman single-cell data were analyzed using t-SNE. The *TSNE()* function from the scikit-learn machine learning library was used. PCA was set as initialization method. TSNE was run with default settings unless reported otherwise. Each region in the spectra was normalized to have zero mean and standard deviation of 1.

Principal component analysis Single-cell Raman spectra were analyzed with the *PCA()* function from the 'scikit-learn' machine learning library after normalization of the spectra, so that each region has a mean of zero and standard deviation of 1.

Hierarchical clustering To measure how dissimilar the samples were, we calculated the spectral contrast angle (Wan *et al.*, 2002) between individual cells based on Raman spectra. Then, clusters were determined in an agglomerative way, through Ward's method (ward.D2) from the R package 'fastcluster' (Müllner, 2013). Hierarchical clustering was implemented using the *hclust()* function from the stats package (R Core Team, 2018).

PhenoGraph PhenoGraph is a clustering algorithm specifically designed for the analysis of high-dimensional flow- or mass-cytometry data (Levine *et al.*, 2015). It employs a two-step approach, in which for every cell its k-nearest cells of similar phenotypic populations are grouped together. This means that, if N denotes the number of cells, N neighborhoods are created. Next, a weighted graph is created on these sets of cells. The weight between nodes scales with the number of neighbors that are shared. The Louvain community detection method is implemented to cluster the graph by maximizing the modularity of different groupings of the nodes (Blondel *et al.*, 2008). The PhenoGraph algorithm was run with default settings, in which k was evaluated for different values between five and 100 (github.com/jacoblevine/PhenoGraph). PhenoGraph was run after normalization of the spectra to have zero mean and standard deviation of 1.

Adjusted Rand Index Clustering results from both hierarchical clustering and PhenoGraph were quantified by the ARI (Hubert & Arabie, 1985). The ARI was calculated with the *adjusted_rand_score()* function from the 'scikit-learn' machine-learning library (v0.19.1) (Pedregosa *et al.*, 2011). The Rand index is defined as the number of pairs of instances that are in the same group or in different groups based on two partitions, which is divided by the total number of pairs of instances. This index is then corrected for the expected index, which is based on random clustering in which the elements per cluster are shuffled between clusters. A value of 1 resembles the perfect match between cluster assignments and ground truth labels, a value of 0 resembles random clustering and a negative value (up to -1) resembles arbitrarily worse clustering.

Boruta variable selection The Boruta variable selection extends on traditional variable selection using Random Forest-based variable importance measures (Kursa & Rudnicki, 2010). The method includes shadow variables, which are copies of original variables that have been permutated. In order to achieve a more stable variable importance score compared to a traditional score derived from a Random Forest model, multiple models (in our manuscript 100) are fitted to the data. Doing this, one can decide by means of a statistical test, in this case a t-test with correction for multiple hypothesis testing, which variables have a statistically significant higher importance score compared to the most relevant shadow variable. The Boruta algorithm from the Boruta R package was run, using the default settings (v6.0.0) (Kursa & Rudnicki, 2010).

Statistical test on Boruta outcome The ten most relevant regions for classification according to the Boruta algorithm were selected. The intensity of these peaks among the growth phases was compared with the Wilcoxon rank sum test with a Benjamini-Hochberg correction (upon rejection of the null hypothesis). The functions *pairwise.wilcox.test()* and *p.adjust()* from the R package stats v3.5.1 (R Core Team, 2018) were used.

4.4 Results

In this work, we define a “phenotypic population” as a group of bacteria grown under the same environmental conditions (*i.e.*, cells from the same biological replicate at a certain growth stage). This population will share morphological and/or metabolic traits that can be detected by FCM and Raman microscopy. Samples of *E. coli* LMG 2092 were measured in the lag, log, and early stationary phase. For every condition, triplicates of the cell culture were made. Thus, we expected to retrieve nine phenotypic populations. As it will be argued in the discussion, this does not exclude the presence of additional fine-scale phenotypic heterogeneity in the so called phenotypic subpopulations.

4.4.1 Flow cytometry

Three biological replicates of *E. coli* LMG 2092 were measured in the lag, log, and early stationary phase through flow cytometry (Supporting Information Fig. 4.7). Data were analyzed at two levels: (1) the single-cell level (*i.e.*, cells were analyzed as individual instances) and (2) the cell population level (*i.e.*, cytometric fingerprints were constructed to describe population dynamics) (Fig. 1). t-distributed stochastic neighborhood embedding (t-SNE) and principal component analysis (PCA) were used to visualize the data at the single-cell level (Fig. 4.1a–d and Supporting Information Fig. 4.9). Principal coordinate analysis (PCoA) was applied to visualize the differences of the phenotypic populations based on Bray–Curtis dissimilarities (Fig. 4.1f). As a validation, t-SNE was performed on the population level as well (Fig. 4.1e).

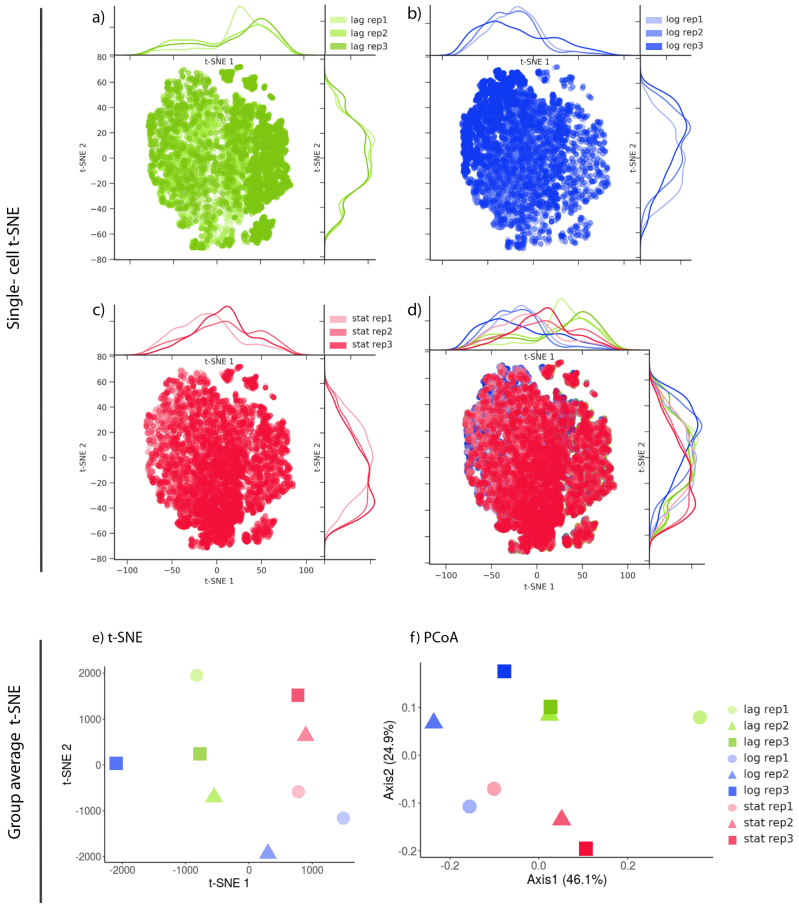


Figure 4.1: *E. coli* measured with flow cytometry and analyzed at the single cell level (a-d) and population level (e and f). t-SNE was performed on the aggregation of all samples (d), and visualized separately for each growth phase, to allow for easier interpretation (a-c). Distributions on the side represent the t-SNE distributions separately visualised for each growth phase/replicate to allow for easier interpretation. (e-f) Visualization of cytometric fingerprints at the sample level, using t-SNE (e) or PCoA (f).

No separated subpopulations could be distinguished based on cytometric single-cell data (Fig. 4.1a–d). Yet shifts in the distribution of cells were clear, both between different growth phases and replicates, as can be seen from the marginal distributions. Therefore, by creating cytometric fingerprints, which are vectorizations of the cell counts per bin, these differences could be quantified and visualized at the community level

(Fig. 4.1e,f). Differences between fingerprints were calculated using the Bray–Curtis dissimilarity. Average dissimilarities per growth phase and replicate were summarized in Table 4.1. The average Bray–Curtis was smaller compared to samples that originated from the same replicate (Table 4.1). The lag phase for replicate 1 was quite different from the other samples (Fig. 4.1f).

Table 4.1: Average Bray-Curtis distance between the samples based on their growth phase or their replicate. A.U.= arbitrary units.

	Average distance between samples (A.U., n=3)	Standard deviation (n=3)
Lag phase	0.33	0.08
Log phase	0.33	0.06
Stationary phase	0.21	0.06
Replicate 1	0.41	0.18
Replicate 2	0.35	0.04
Replicate 3	0.35	0.07

4.4.2 Raman microscopy: clustering results

The samples used for flow cytometric analysis were fixed and analyzed using label-free Raman microscopy following the protocol from the study by García-Timermans *et al.* (2018). To identify phenotypic populations, two clustering methods were used. First, an agglomerative hierarchical clustering approach and second, the PhenoGraph algorithm—a tool originally developed for the analysis of high-dimensional cytometry data. To determine the hierarchical clustering, the spectral contrast angle between samples was calculated (a measure of the spectral similarity). Next, phenotypic populations could be delineated by setting a threshold upon inspection of the resulting dendrogram after hierarchical clustering (Fig. 4.2). On the other hand, PhenoGraph makes use of a k-nearest-neighbor weighted graph and clustering, in order to determine groups of similar cells, and as such, phenotypic populations. In other words, *k* expresses the amount of local information that is included when cells are grouped according to similar spectra. *k* will therefore, in a similar way as the threshold used in hierarchical clustering, impact the number of phenotypic populations

that are defined (Fig. 4.3).

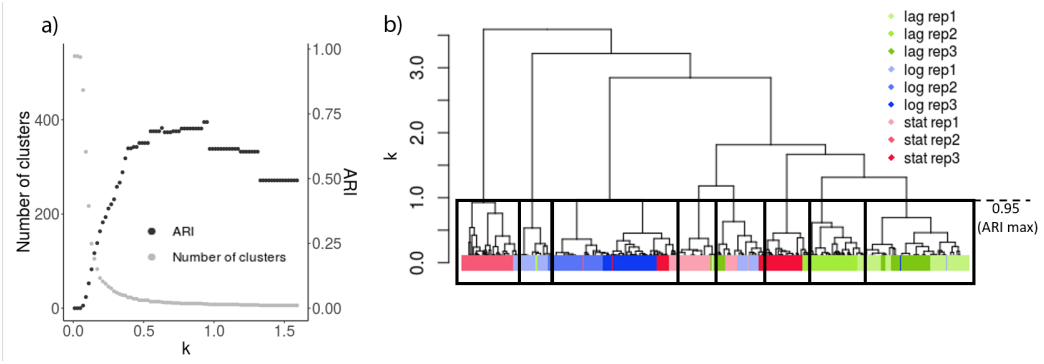


Figure 4.2: Hierarchical clustering of the Raman spectra from the *E. coli* cultures. Cells were measured in the log, lag and early stationary phase using Raman microscopy. (a) Left axis, grey: Visualization of the number of clusters we can obtain by cutting the cluster at different heights (k). Right axis, black: adjusted Rand index (ARI), which quantifies how many cells are identified as the expected phenotypic population when clusters are made at different heights (k). (b) Dendrogram representing the distance (calculated as the spectral contrast angle) amongst the Raman spectra. The black boxes dived the phenotypic populations when the dendrogram is cut at k=0.9. This is the ARI, as calculated in Fig. 4.2a.

The adjusted Rand index (ARI) was used to quantify similarity between the clusters that were determined by hierarchical clustering and Phenograph and the known phenotypic populations (*i.e.*, growth phase and replicate). An ARI of 1 indicates perfect grouping of the data. The PhenoGraph algorithm resulted in a higher ARI compared to hierarchical clustering based on the spectral contrast angle (Fig. 2a *v.s.* Fig. 3a). Inspecting the PhenoGraph results, there is a stable region for k that retrieves clustering according to both growth phase and replicate (*i.e.*, nine clusters were found for k = 20, . . . , 60). A value of k=24 or 26 resulted in an optimal clustering (Fig. 4.3a). Smaller k allowed to inspect phenotypic populations at smaller scales and investigate the heterogeneity accordingly. See, for example, the clustering results for k = 15, which resulted in 11 different groups of cells (Supporting Information Fig. 4.11). Additional clusters that emerged were the result of splitting two clusters into two smaller ones. Likewise, larger k will result in larger clusters. For example, for k = 100, data are grouped in five clusters (Supporting Information Fig. 4.11). Structure in the data is retained, as clusters are merged either according to growth

phase (Clusters 0, 2, and 3) or replicate (Cluster 1).

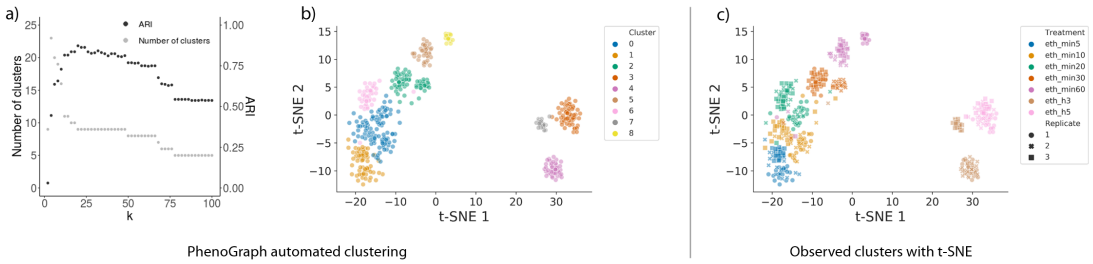


Figure 4.3: PhenoGraph clustering of the Raman spectra derived from the *E. coli* culture. a) Influence of hyperparameter *k* on the automated identification of phenotypic populations. Left axis, grey: Visualization of the number of clusters in function of *k*. Right axis, black: Adjusted Rand index (ARI), which quantifies how many cells are identified as the expected phenotypic population when clusters are made at different levels (*k*). b) t-SNE visualization, colored according to PhenoGraph clustering with optimal ARI (*k*=22). c) t-SNE results colored according to growth phases and replicates.

When a single cell was wrongly clustered, it was due to a misclassification of the growth phase, rather than the expected replicate (Fig. 4.3c). The samples in the lag phase seem to have a single cell that is already in the log phase, and in the cultures in the log phase, we find a cell in the early stationary phase (in Replicate 3) and one cell in the lag and in the early stationary phase (in Replicates 1 and 2). It is also worth noting that some cells from replicate 3 seem to be between the log and the early stationary phase.

4.4.3 Raman microscopy: tentative region assignment

The Boruta algorithm, a variable selection algorithm based on Random Forests, was used to associate the most distinctive regions in the Raman spectrum with cluster assignments according to the hierarchical clustering and PhenoGraph algorithm. The cluster labels that resulted in an optimal ARI were used. Regions were linked with different molecules based on the manuscript from Wang and colleagues (2016). In this way, metabolic associations could be inferred that contained predictive power as a function of different phenotypic populations (Table 4.2) (Yu *et al.*, 2016).

Table 4.2: Tentative assignment of Raman spectra using the Boruta algorithm based on phenotypic identification using hierarchical clustering and PhenoGraph. The 10 highest ranked areas are shown. When there is no known compound in the spectral region, either the closest compound or a blank is shown. A.U.=arbitrary units.

Wavelength (cm^{-1})	Features importance		Tentative assignment (cm^{-1})
	PhenoGraph (k=22, 9 clusters) (A.U.)	hclust (h=1, 8 clusters) (A.U.)	
1042	9.8	9.8	Carbohydrates, Proline (1043)
971	9.5	9.7	v(C-C) wagging (971) Phosphate monoester groups of phosphorylated proteins and cellular nucleic acids (970)
945	9	8.9	vs (CH3) of proteins (α-helix) (951)
1057	8.9	8.8	Lipids (1057) Carbohydrates (1030-1130)
1294	8.5	8.5	CH2 deformation (1295)
1050	8.5	8.6	Nucleic acids, CO stretching; protein, C-N stretching, PHB (1054) Carbohydrates (1030-1130)
1053	8.4	8.4	Nucleic acid (1054) Carbohydrates (1030-1130)
1127	8.3	8.4	v(C-N), protein (1127) Carbohydrates (1030-1130)
1046	8.1	8.2	v3PO4 3– (symmetric stretching vibration of v3PO43– of HA) (1044)
641	8	7.9	C-C twisting of tyrosine (642)

To understand how the molecules in Table 4.1 vary from one group to another, the distribution of intensities of these Raman regions was plotted for every growth phase (Fig. 4.4). The ten highest ranked variables according to the Boruta algorithm are visualized after performing a Wilcokon rank sum test with a Benjamini-Hochberg correction (Fig. 4.4).

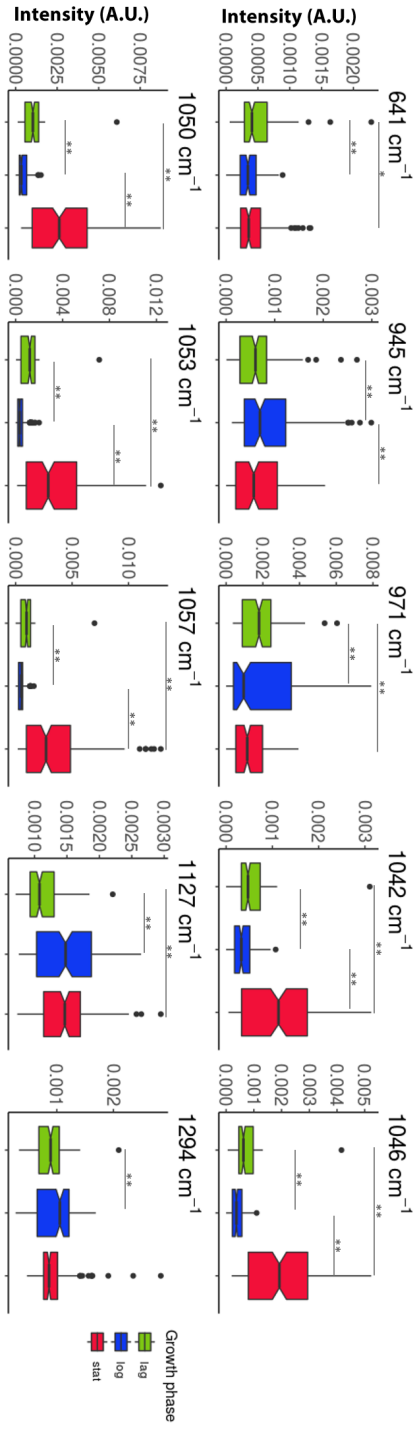


Figure 4.4: Distribution of intensities of the most relevant regions associated with phenotypic populations according to the Boruta algorithm. Boxplots represent the growth phases (replicates were pooled together). Groups were made according to the growth phase: lag (green), logarithmic (blue) or early stationary (red). For every spectral region, a Wilcoxon rank sum test was made, with a Benjamini-Hochberg correction (upon rejection of the null hypothesis). Groups with significantly different peaks are signalled with (*) ($p<0.05$) or (**) ($p<0.01$).

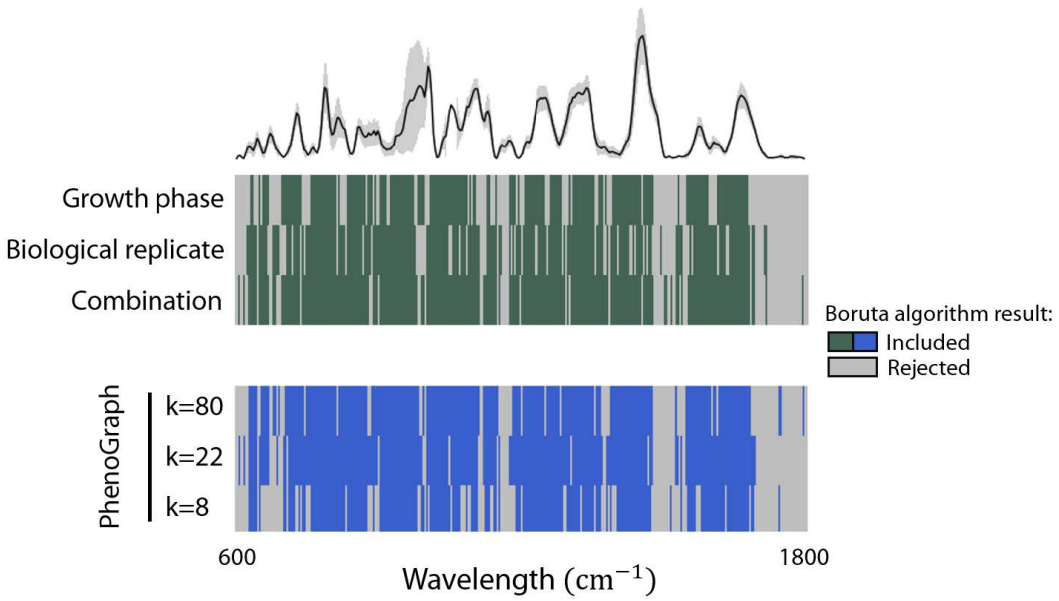


Figure 4.5: Spectral regions relevant for phenotypic classification at different levels, according to the Boruta algorithm. The top heatmap shows the results for the growth phase and replicates. At the bottom, the results for PhenoGraph with the clustering hyperparameter *k* as 80, 22 or 8 are shown. In green and blue, spectral regions confirmed by the Boruta algorithm as relevant, in grey the rejected. The average of all spectra is also plotted; the grey areas in the average spectrum correspond to the standard deviation.

To better understand what regions of the Raman spectra (and therefore, what biomolecules) were making these phenotypic populations different, we defined phenotypic populations at different levels (changing the *k* parameter in the PhenoGraph algorithm) and then used the Boruta algorithm to identify the most relevant regions.

When more phenotypic populations were distinguished (*i.e.*, setting the value of *k* lower) more regions in the Raman spectrum were associated with differences in phenotypic populations. As shown in Figure 4.5, to find the phenotypic populations with different growth phases, 59% of the regions are included (green); to find the biological replicates, it is 67%; and to find both categories, 77%. Although this result was expected, a large number of Raman regions (48%) were relevant for all levels of classification.

4.4.4 Validation of single-cell analysis of Raman spectra

To validate our workflow for the analysis of microbial single-cell Raman data, the dataset from Teng *et al.* (2016) was used. In this work, *E. coli* was exposed to different chemicals (ethanol, antibiotics, n-butanol, or heavy metals) and the spectra of the bacteria were measured at several time points after the treatment (5, 10, 20, 30 and 60 min, 3 h, and 5 h). Three replicates of the cell culture were made for each treatment. Here, we show the results for cells treated with ethanol (Fig. 4.6), representative for what is observed in the other groups (Supporting Information Fig. 4.12).

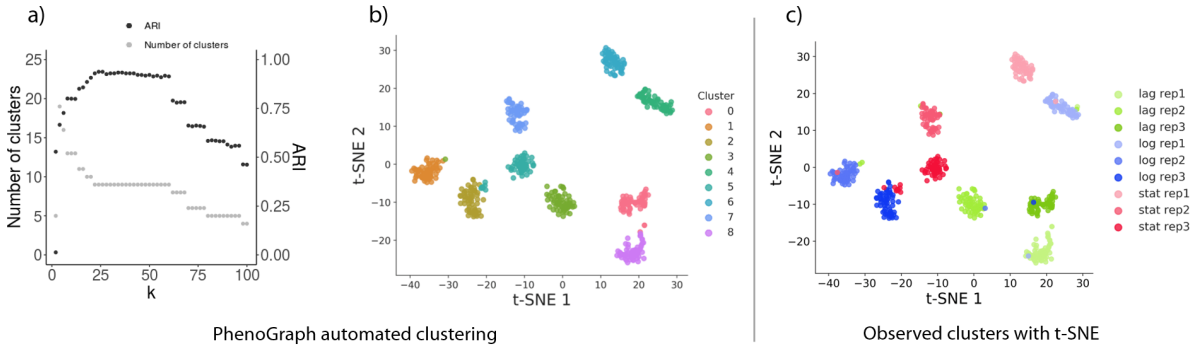


Figure 4.6: External dataset clustered using t-SNE. Raman fingerprint of *E. coli* treated with ethanol and measured at time points 5, 10, 20, 30 and 60 min, 3 h and 5 h. A) Influence of hyperparameter *k* on the automated identification of clusters. Left axis, grey: Visualization of the number of clusters as a function of *k*. Right axis, black: Adjusted Rand index (ARI), which quantifies how many cells are identified as the expected phenotype when clusters are made at different levels (*k*). B) Using the maximum ARI, samples were automatically clustered using PhenoGraph. C) t-SNE was performed on the dataset and samples were labelled according to their treatment. The shapes represent the sample replicate.

t-SNE was able to visualize groups of bacteria that received different treatments at different points in time. Furthermore, two subpopulations are seen in every group. They correspond to the replicates, where two replicate samples are separated, and the third replicate is either assigned to one of the two or divided amongst the two subpopulations. The optimal ARI is lower than the one reported for our own work but still considerably higher

than zero. This means that although the clusters assigned according to PhenoGraph have a better match with the treatments induced in our own dataset compared to this one, the clustering is still meaningful.

4.5 Discussion

4.5.1 Flow cytometry quantifies population shifts

Flow cytometry is a high-throughput technique, able to rapidly measure hundreds to thousands of individual cells per second. By applying fingerprinting approaches to cytometry data, differences between microbial populations at the population level can be assessed and quantified. In this work, gradual shifts could be detected in the flow cytometric data at the level of individual cells, while at the sample-level (*i.e.*, the population distribution level), differences between communities could be quantified (*e.g.*, using the Bray–Curtis dissimilarity) and separated accordingly.

Single-cell flow cytometry measurements of the different phenotypic populations overlapped and did not form separate clusters, as shown by both the t-SNE and the PCA (Fig. 4.1a and Supporting Information Fig. 4.9). However, in the t-SNE plot, a consistent shift in the cells distribution could be observed in response to the different growth phases (Fig. 4.1a–d). In other words, gradual shifts in the structure of the phenotypic population, that is, the phenotypic heterogeneity, could be detected, although individual cells could not be separated according to growth phase or replicate. The differences in phenotypic heterogeneity at the population level could be described by constructing cytometric fingerprints (Koch *et al.*, 2014; Props *et al.*, 2016) for which the Bray– Curtis dissimilarity was used. The average Bray–Curtis dissimilarity showed that the effect of the replicates exceeded the effect of the growth phase (except for the lag phase, Table 4.1). This implies that the differences between *E. coli* cells in different growth stages are comparable or smaller to the differences among replicates in the same growth stage.

It is worth noting that in this work, the effect of using additional or more specific labels for cytometric analysis has not been explored, which might improve the resolution. It

is possible to add stains to target specific substrates (see the review of Léonard et al. (Léonard *et al.*, 2016) on the use of individual and double stains and an example of a three-color analysis by Barbesti et al. (Barbesti *et al.*, 2000)), but the number of markers describing microbial cells using flow cytometry will never be of the same order as that of Raman spectroscopy. Steen & Boye (1980) could differentiate growth stages using flow cytometry in *E. coli* K-12 using a combination of ethidium bromide and mithramycin. Müller and Nebe-von-Caron (Nebe-von Caron *et al.*, 2000) recommend in their review the use of DAPI dyes for DNA staining; however, a UV laser is needed for this, which is less common in routine flow cytometers due to its price. This work only explores the use of SYBR Green I with a 488 nm laser, a more affordable tool for microbial phenotyping, but the effect of alternative dyes on the resolution of the data has not been explored. In eukaryotic flow cytometry, where the tagging of specific antibodies is much more feasible, 19-parameter flow cytometry is routinely used (17 fluorescence and two scatters) (Perfetto *et al.*, 2004) and 30-parameter flow cytometry has just recently been published (Mair & Prlic, 2018). However, the dimensionality of cytometry data in these settings is still much lower than the number of variables derived from Raman spectroscopy. Even in the best-case scenario, the dimensionality of flow cytometry data cannot get close to the number of parameters that Raman spectroscopy exhibits. On the other hand, depending on the research question, a high-dimensional tool might not be needed. For example, biomolecules that are associated with a phenotypic population might be known, and there could be a dye available to target these molecules. In this case, flow cytometry could be more suitable for phenotyping than Raman microscopy, provided the proper parameters are chosen to differentiate among treatments. In this work, we compared phenotypic populations that are not differentiated by (a) specific, known molecule(s) but used a general marker to characterize the DNA content. The nucleic acids were labeled using SYBR Green, a widely used dye for flow cytometric bacterial quantification and fingerprinting. These factors might explain why flow cytometry did not have enough resolution to differentiate the phenotypic populations. Raman spectroscopy detects phenotypic populations at the single-cell level.

Raman microscopy is lower in throughput for single-cell analysis when compared to flow cytometry, but it can retrieve much more information per cell. Its resolution is enough to conduct research at the single-cell level. The study of bacterial phenotypes using Raman spectroscopy has been conducted by other research groups as well, for example to identify different growth stages in *L. casei* (Ren *et al.*, 2017), stress-induced phenotypic

populations (Teng *et al.*, 2016), bacterial phenotypes with different antibiotic responses (Athamneh *et al.*, 2014) or with different antibiotic susceptibility (Novelli-Rousseau *et al.*, 2018). It has also been used to discriminate between different *Acinetobacter* (Maquelin *et al.*, 2006) or different *E. coli* strains (Jarvis & Goodacre, 2004) among other examples. In these studies, the expected phenotypes where known in advance. However, how to define what a phenotype is in a less-known system or a natural environment?

In this chapter, we proposed and validated the use of PhenoGraph. The PhenoGraph algorithm was originally developed for mass cytometry data (Levine *et al.*, 2015), a variation of flow cytometry which makes use of heavy metal ion tags instead of fluorochromes, resulting in more observed variables but at a lower acquisition speed (Spitzer & Nolan, 2016). PhenoGraph demonstrated to be highly effective for clustering purposes of single-cell Raman data and returned a higher clustering performance compared to a more traditional hierarchical clustering approach. However, hierarchical clustering allows to inspect which cells are most similar to each other, a characteristic which is lost when using PhenoGraph. Therefore, we want to reiterate that, as proposed by Andrews & Hemberg (2018) for the analysis of single-cell RNA sequencing data, "[l]ikewise, no computational methods for dimensionality reduction, feature selection and unsupervised clustering will be optimal in all situations." **The algorithm of choice** depends on the needs of the user. If a researcher wants to visualize subpopulations, we recommend the use of t-SNE. If identification of phenotypic populations is needed in an automated way, PhenoGraph is more appropriate. It is possible to use them together, as we show in this manuscript, as the first is a visualization tool and the second a clustering algorithm. To assess which individual cells are phenotypically closest, hierarchical clustering can be used, as it will quantify the similarity between the cells and allow to construct dendrograms. Further investigation of the analysis of Raman data is needed, but investigating additional algorithms specifically developed for high-dimensional single-cell data might further support the impact of the use of Raman spectroscopy.

The sample size for the Raman measurements in this experiment is 60 cells per replicate and 3 replicates. This choice was motivated by two factors (1) numerous papers that conduct microbial ecology studies with Raman spectroscopy measure 3–30 cells per replicate (*e.g.*, Huang *et al.* (Huang *et al.*, 2010) investigate the discrimination of species and growth stages based on Raman spectroscopy, for which they measure three cells

per replicate and three replicates; Ren *et al.* (Ren *et al.*, 2017) describe growth stages in *Lactobacillus* measuring 10 cells per replicate and 3 replicates; Huang *et al.* (Huang *et al.*, 2010) study the metabolism of *Pseudomonas* measuring four to eight spectra per replicate and two to three replicates; Teng *et al.* (Teng *et al.*, 2016) study stress responses in *E. coli* measuring 20 cells from each replicate and 3 replicates); (2) if individuals cells are measured over a long period of time, there might be more noise introduced by the instrument variability. Thus, we estimate that 60 cells per condition and 3 replicates is enough to make a relevant case. However, when discussing phenotypic heterogeneity, sample size should be more deeply investigated, as it is likely that increasing it would result in better classification. Findings from Ali *et al.* (2018) demonstrate how such a study can be done and recommend (in their example) to measure 7 replicates and 142 cells per replicate. This is far from the typical 3–30 cells measured in Raman experiments or from our 60 measurements per replicate. As mentioned in **chapter 3**, the importance of sample size in Raman microscopy measurements is explored in **chapter 6**, where we show how measuring $\sim$50 spectra already gives a result close to the population average for most axenic cultures, but how certain populations -presumably with a higher diversity- need at least 300 measurements (Fig. 6.11).

The main downside of the use of label-free Raman microscopy is that the time of measurement is long: in this experiment, for single-cell label-free measurements, we used an acquisition time of 40 sec per cell. Even when the acquisition time is lower —for instance Liu *et al.* (2016) reported a 1–3 sec acquisition time to detect antibiotic susceptibility using surface enhanced Raman spectroscopic (SERS) biomarkers—the speed of Raman spectroscopy cannot match the high throughput nature of flow cytometry for single-cell analysis. Other strategies to enhance the Raman scattering is the use of metallic substrates (SERS) combined with microfluidic chips (McIlvenna *et al.*, 2016) or alone. Another disadvantage is that the Raman signals of certain compounds can be quite weak, making them difficult to detect or undetectable. The Raman signal of certain compounds can be composed of several peaks, or be unknown, making the identification of these compounds difficult. Furthermore, the background of samples can interfere with the Raman signal of bacteria. Finally, the equipment can be quite costly, depending on the type of Raman spectroscope.

An advantage of Raman spectroscopy is that it can be applied without the use of labels.

This allows to analyze the biochemistry of samples even without knowing their nature. This is especially useful when studying natural communities. Also, Raman spectroscopy offers more parameters per cell compared to flow cytometry (hundreds vs typically three or four for microbial experiments). Thus, individual bacteria are described in a much larger multivariate space and can therefore be clustered into separate phenotypic populations. This explains why bacterial subpopulations can be visualized at the single-cell level using t-SNE (Fig. 4.3). The t-SNE results were confirmed with a PCA (Supporting Information Fig. 4.9).

4.5.2 Raman spectroscopy allows to detect differences in biomolecules across samples

Raman spectroscopy allows to detect biomolecules present in single cells. Therefore, after identification of phenotypic populations, one can use the phenotypic groups to perform a variable selection strategy to select important regions in a data-driven way. We illustrated this approach using the Boruta algorithm, that was recently evaluated as one of the state-of-the-art variable selection methods using Random Forests for omics datasets (Degenhardt *et al.*, 2017). We found that a majority of selected spectral regions were the same according to treatment and automated phenotypic population identification using PhenoGraph (Fig. 4.3). This information can be used to infer how phenotypic populations are different at the level of their metabolism. To do so, we have based ourselves on a recent literature survey summarizing associations between Raman regions and certain biological compounds (Wang *et al.*, 2016). The 10 most important regions in function of phenotypic identification are listed in Table 4.2, along with the distribution of their intensities (Fig. 4.4). These regions correspond to carbohydrates and nucleic acids. An increase in the carbohydrate band (peaks 1042, 1046, 1050, and 1057 cm^{-1}) was observed for the early stationary phase. The band at 1053 cm^{-1} could also be a nucleic acid peak, expected at 1054 cm^{-1}. Nevertheless, these assignments for the Raman bands are tentative and based solely on a literature research, and thus proper validation of these results would have to be made in future experiments.

4.5.3 The best of both worlds

It is possible to combine the rapidity of flow cytometry with the resolution of Raman spectroscopy. For instance, there are Raman spectral flow cytometers that combine high throughput with the detection of Raman scattering. This increase in throughput is mostly due to the use of cells with strong Raman signals or SERS—as discussed in this paper, this can cut the measuring time from 30–40 sec to 1–3 sec and the addition of the flow allows to localize and focus the cells in a fast and automated way. Some of them do not only measure Raman scattering, but also the typical FCM parameters (light scattering and fluorescence). As discussed previously, Raman scattering is weak, so this tool can be used for Raman active molecules that have a strong signal (*e.g.*, carotene) or needs to be combined with metals (SERS) or other dyes (such as deuterium or ^{12}C) (Goddard *et al.*, 2006; Watson *et al.*, 2008). It is possible to combine these tools in a microfluidic chip, and even to separate cells rapidly according to their Raman spectra—a tool named RACS or Raman-activated cell sorting after its analogous FACS (Song *et al.*, 2016). Another possible combination of these technologies could be to use FACS to sort a high number of cells based on a certain characteristic (nucleic acid content, activity label such as BONCAT or other stains) and subsequently analyze these subpopulations using Raman spectroscopy.

4.5.4 How to define a phenotypic population?

In this chapter, we have steered microbial communities toward a certain growth stage, expecting that they would express a certain phenotype that could be retrieved using flow cytometry and Raman microscopy. However, in each one of these isogenic populations, there might be subpopulations, as shown in Figure 4.2b.

We acknowledge the difficulty in defining what a phenotypic population is and setting a threshold to determine when one phenotypic population ends and another begins and propose a definition based on single cell similarities (after setting a similarity threshold) and their ecology (their relationship with one another and with their environment). Similarity can be quantified in a data-driven way, by means of, for example, clustering at the resolution that is required for the specific research. This operational definition allows to define

phenotypic populations depending on the research question, as long as researchers motivate and validate their choice. However, using this operational definition means that results cannot be compared across experiments or labs. This is why we reiterate the need to find a more standard way to define "basic phenotypic units," that would allow to measure phenotypic traits and determine whether bacteria belong to the same phenotypic population. This idea is further explored in **chapter 7 - Defining phenotypes: how far does the rabbit hole go?**.

In this chapter, we propose to use algorithms—such as hierarchical clustering, t-SNE or PhenoGraph, applied throughout this article—to define, visualize, and characterize phenotypic populations. t-SNE is a well-known technique to visualize high-dimensional single-cell data, being commonly applied to visualize, for example, cytometry and single-cell RNA sequencing data (Andrews & Hemberg, 2018; Amir *et al.*, 2013). Our results confirm that it can be used as an "off-the-shelf" visualization method to detect phenotypic populations in Raman data when applied to microorganisms. It must be noted that PCA retains the global structure of different clusters (*e.g.*, cells measured in the lag phase lie closer together than cells originating from the early stationary phase), which is not the case for t-SNE.

Bacteria were grown in nine different conditions (three replicate cultures of three growth stage conditions) to steer the same *E. coli* population to a different morphological and/or metabolic state—to steer them into nine phenotypic populations. While hierarchical clustering was able to find eight of these phenotypic populations, PhenoGraph was able to retrieve all nine of them, resulting in a higher ARI as well.

t-SNE and PhenoGraph were also applied to an external dataset from Teng *et al.* (2016), consisting of *E. coli* that had been treated with different agents, and measured at several time points. We showed that PhenoGraph was capable of differentiating the time points per treatment. Interestingly, two subpopulations were identified per treatment, although samples were measured in triplicate. These corresponded to two replicates, where the third was either assigned to one subpopulation or divided between both (Fig. 4.6 and Supporting Information Fig. 4.12). Our group has previously shown how small technical variations can create subpopulations that have no biological meaning (García-Timermans *et al.*, 2018), which might explain these findings.

4.6 Conclusions

The results of this research suggest that:

- Flow cytometry is a more high-throughput technology than label-free Raman microscopy, but Raman describes bacterial cells in many more variables, without the need for staining.
- Flow cytometry can be applied in isogenic populations to quantify differences in phenotypic heterogeneity at the population level, whereas Raman spectroscopy has sufficient resolving power to identify separate phenotypic populations at the single-cell level.
- Raman spectroscopy provides the possibility to infer which metabolic properties define different phenotypic populations and potentially exploit this information for bioprocess monitoring.
- We propose a computational workflow to automatically identify bacterial phenotypes, based on Raman spectral data. We also recommend t-SNE to visualize Raman data.
- From a broader perspective, one can motivate that the definition of phenotypic populations using algorithms is highly dependent on the similarity threshold that is used to distinguish these subpopulations. We therefore suggest that researchers include validation controls in their experimental setup, so that the detected populations are ecologically meaningful and not merely arbitrarily defined groups.

4.7 Acknowledgements

The authors would like to thank Dmitry Khalenkow for his help setting up the Raman microscope. The authors would like to thank the funding that made possible this research. CGT is funded by Qindao Beibao Marine Science & Technology Co. Ltd., Qingdao West-coast economic new area, China. PR is funded by Special Research

Fund (BOFSTA2015000501) from Ghent University. RP is funded by by Ghent University (BOFDOC2015000601). JH is funded by the Flemish Fund for Scientific research (FWO-Vlaanderen, 1S80618N). AGS acknowledges support of BOF UGent (BOF14/IOP/003, BAS094-18, 01IO3618) and FWO-Vlaanderen (G043219). This work was supported through the Geconcerteerde Onderzoeksactie (GOA) from Ghent University (BOF15/GOA/006) and the MICCAS project (project grant no. 3G020119 of the FWO Flanders).

4.8 Appendix

4.8.1 Supplementary information

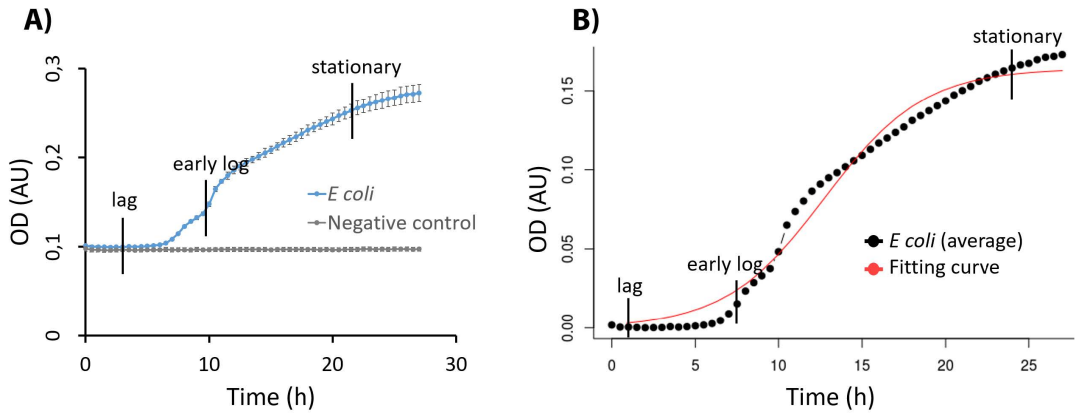


Figure 4.7: Growth curve for *E. coli* in nutrient broth at 28°C, 120 rpm shaking. Three replicates of the cell culture were made. A) OD results. In blue, the results for the *E. coli*; in grey, the negative control (medium). B) Fitting model for the *E. coli* OD results (after background subtraction) to assign the lag, log and early stationary phases.

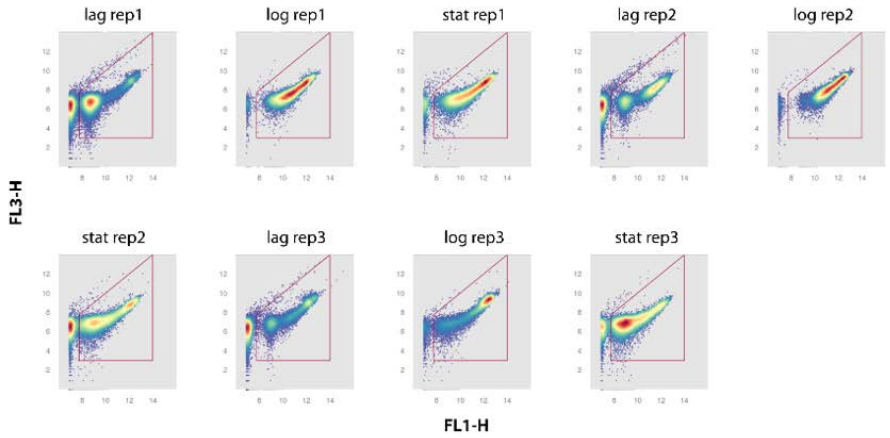


Figure 4.8: Gating strategy for the flow cytometric data. Arcsinh transformed data of FL1 (green channel)/FL3 (red channel) in a density plot (see Materials and Methods).

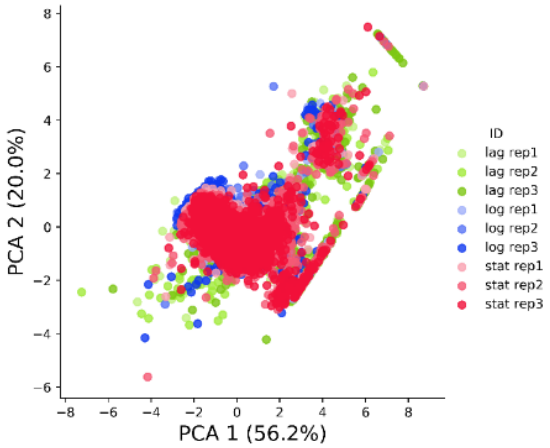


Figure 4.9: First two components of a PCA for single-cell flow cytometric data. Colors correspond to growth phases and color shades to replicates.

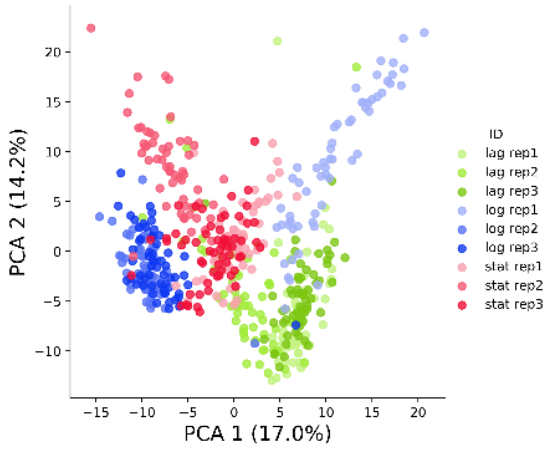


Figure 4.10: First two components of the PCA for Raman spectra of individual cells. Colors correspond to growth phases and color shades to replicates.

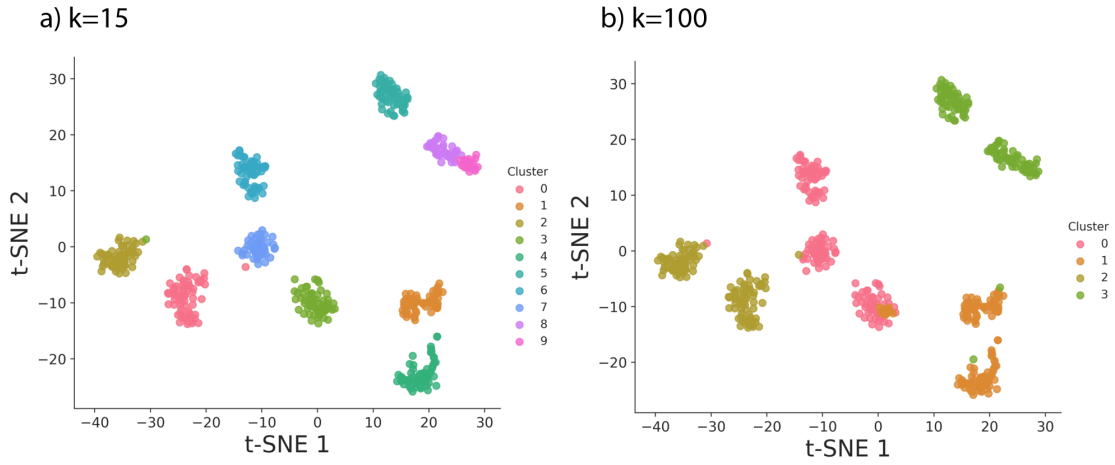


Figure 4.11: Visualization of PhenoGraph clustering results of the Raman data derived from the *E. coli* culture presented in Figure 4.3 for different k: (a) k=15, (b): k=100.

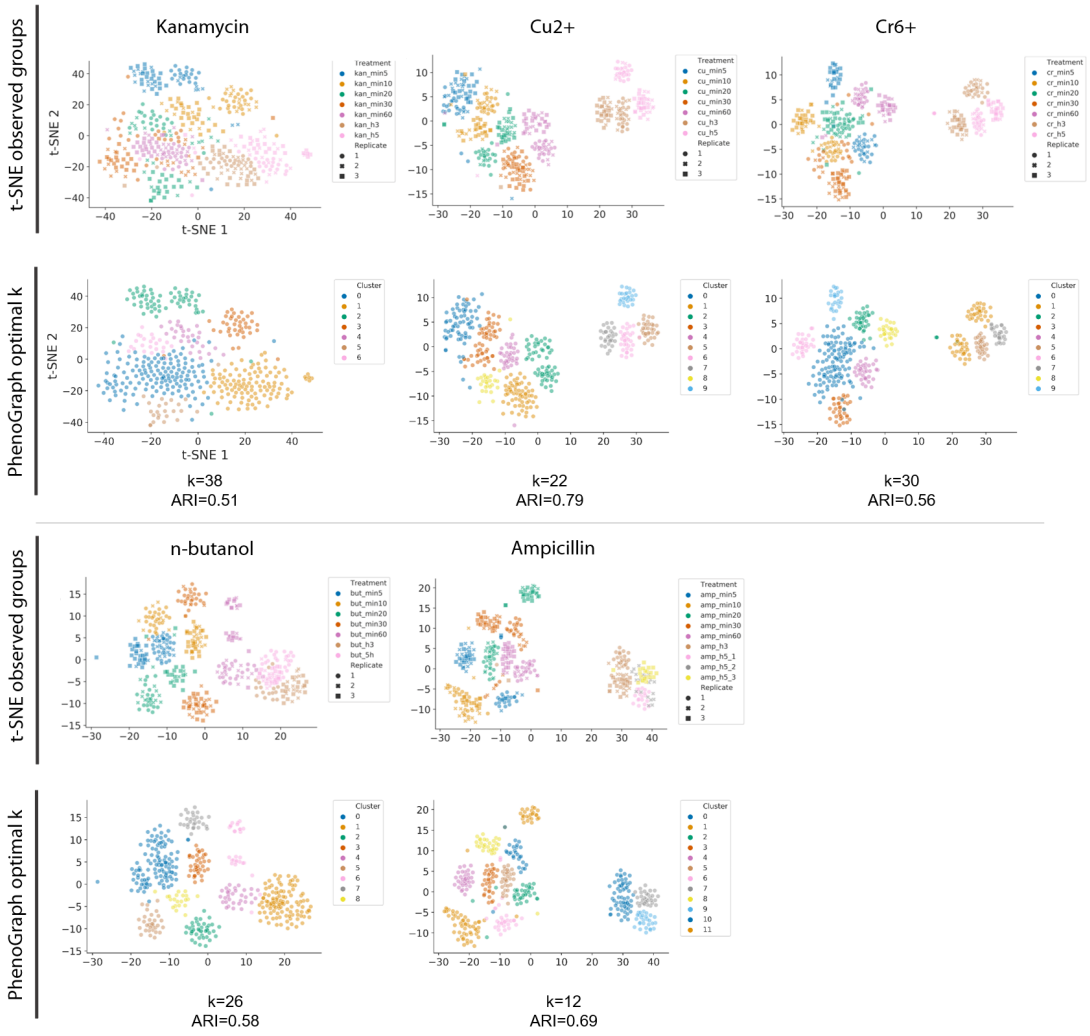


Figure 4.12: Ramanone results for all the groups. On top, the t-SNE for the observed phenotypic groups. At the bottom, the PhenoGraph results after calculating the optimal number of clusters (highest ARI).

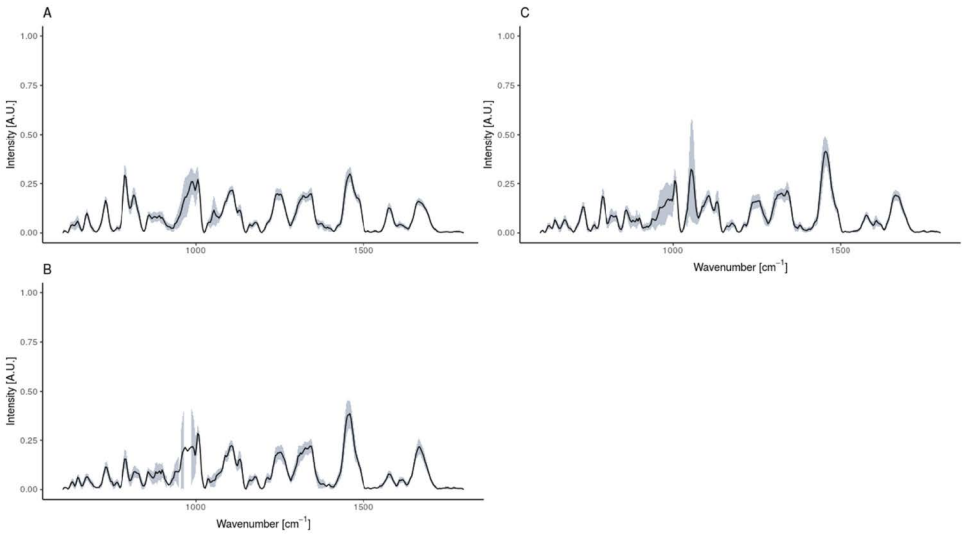


Figure 4.13: Raman spectra of the cells according to growth stage. *E. coli* harvested in the A) lag B) log and C) early stationary phase. In lighter color, the standard deviation is represented.

Table 4.3: Metadata aid for Raman spectra

Experiment overview	
Hypothesis	*E. coli* in different growth stages and/or cells from different replicates will differ in their Raman spectra
Variable(s) tested	*E.coli* were fixed in log, lag and early stationary phase. Three replicates of the cell culture.
Conclusions	Differences in the Raman spectra were found when cells differed in the growth stage and/or belonged to another culture
Quality control (internal/external)	Silicon check
Material and source	*Escherichia coli DSM 2092*
Growing conditions/sampling	Cells were grown in LB, 25℃, 120 rpm shaking
Filename format:	<Replicate number>_<Treatment>_<Cell number>
Label in the samples	No label used
Fixation method	Filtered 4% formaldehyde solution from PFA
Integration time	40 sec
Accumulations	1
Grid	300 g/mm
Instrument	
Laser	785 nm excitation diode laser (Toptica). 175 mW of power before the objective.
Quality control	A silicon piece (IMEC, Belgium) sample was measured with a grating of 600 g/mm, with a 1 second time acquisition and 10 accumulations. Laser power was also monitored to detect possible variations.
Objective used (magnification) / Numeric aperture (NA)	100x/0.9 NA (Nikon)
Camera	-70℃ cooled CCD camera (iDus 401 BR-DD, ANDOR)
Dry/water/oil objective	Dried samples
Model of spectroscope	WITec Alpha300R+
Other specifications (chromatic/flat field correction/other)	
Data analysis	
Background subtraction method (if used)	No. Measurements with cosmic rays were deleted
Normalization method (peak /min-max /area under-curve /other)	Area under the curve ('Total Ion Current')
Smoothing and interpolation (if done)	Baseline correction
Statistics/Machine learning algorithm	'MicroRaman' package (GitHub). Spectral contrast angle, ward.D2 dissimilarity and hierarchical clustering. Boruta.
Accessibility	https://github.com/CMET-Ugent/FCMvsRaman.
Other relevant information	

4.8.2 Availability of data and material

Data and code to reproduce analysis is available on the following repository: https://github.com/CMET-Ugent/ FCMvsRaman. Data analysis was conducted using the program R (R Core Team, 2018), RStudio (RStudio, 2016) and Python. The flow cytometry data are available under the Flow Repository ID FR-FCM-ZYV6.

4.8.3 External dataset

We included the dataset from Teng *et al.* (2016) in order to validate the generalizability of the PhenoGraph and t-SNE algorithms for the analysis of label-free bacterial Raman data. As described in their article, they tested the stress response of *E. coli* to six chemical stressors at different time intervals with label-free Raman spectroscopy: ethanol, antibiotics ampicillin and kanamycin, n-butanol or heavy metals Cu^{2+} ($CuSO_4$) and Cr^{6+} (K_2CrO_4). Teng *et al.* showed that each of these treatments resulted in a different phenotype. In other words, each treatment resulted in a unique Raman characterization of cells, which should group together upon analysis. These treatments were therefore used as label according to which PhenoGraph or t-SNE should group the cells. Three biological replicates of the cell culture were made, and 20 cells were tested per replicate. Bacteria were sampled at different stages of the cell growth. The Raman spectra of the stressed cells were collected after the treatment (5, 10, 20, 30 and 60 min, 3 h and 5 h).

4.8.4 Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.8.5 Author contributions

CGT and PR co-wrote the paper with contributions from JH, FMK, RP, A.S., W.W., and NB. CGT collected the data. PR, CGT, and RP performed the data analysis. CGT, PR, JH, FMK, RP, WW, and NB designed the study.

5

Raman spectroscopy-based measurements of single-cell phenotypic diversity in microbial populations

5.1 Abstract

Microbial cells experience physiological changes due to environmental change, such as pH and temperature, the release of bactericidal agents, or nutrient limitation. This, has been shown to affect community assembly and other processes such as stress tolerance, virulence or cell physiology. Metabolic stress is one such physiological changes and is typically quantified by measuring community phenotypic properties such as biomass growth, reactive oxygen species or cell permeability. However, community measurements do not take into account single-cell phenotypic diversity, important for a better understanding and management of microbial populations. Raman spectroscopy is a non-destructive alternative that provides detailed information on the biochemical make-up of each individual cell. Here, we introduce a method for describing single-cell phenotypic diversity using the Hill diversity framework of Raman spectra. Using the biomolecular profile of individual cells, we obtained a metric to compare cellular states and used it to study stress-induced changes. First, in two *Escherichia coli* populations either treated with ethanol or non-treated. Then, in two *Saccharomyces cerevisiae* subpopulations with either high or low expression of a stress reporter. In both cases, we were able to quantify single-cell phenotypic diversity and to discriminate metabolically stressed cells using a clustering algorithm. We also described how the lipid, protein and nucleic acid composition changed after the exposure to the stressor using information from the Raman spectra. Our results show that Raman spectroscopy delivers the necessary resolution to quantify phenotypic diversity within individual cells and that this information can be used to study stress-driven metabolic diversity in microbial populations.

Chapter written after:

Cristina García-Timermans, Ruben Props, Boris Zacchetti, Myrsini Sakarika, Frank Delvigne and Nico Boon. Raman spectroscopy-based measurements of single-cell phenotypic diversity in microbial populations (in revision) https://doi.org/10.1101/2020.05.21.109934

5.2 Introduction

Monoclonal microbial populations can exhibit heterogeneous genetic expression, which underlies phenotypic differences between cells. Phenotypic diversity has been shown to increase population survival or fitness in a changing environment and allows microorganisms to divide tasks and organize as a group. This differential gene expression can arise due to environmental pressure, stochastic events, periodic oscillations or cell-to-cell interactions (Ackermann, 2015; Altschuler & Wu, 2010; Avery, 2006). When a deviation from optimal growth conditions occurs such as changes in temperature, pH, nutrients salts and/or oxygen levels, a stress response is triggered in microorganisms (both prokaryotes and eukaryotes), resulting in a biochemical cascade to promote stress tolerance, virulence or other physiological changes. These strategies can result in enhanced survival, virulence, cross-protection or cell death Ron (2013); Święciło (2016); Wesche *et al.* (2009). Usually, microorganisms show mixed behavioural strategies, maximizing the chances of survival Lowery *et al.* (2017), making phenotypic diversity a crucial characteristic of stress-driven phenotypes. However, cellular stress is often measured at the community level using bulk technologies, such as cell concentration, quantity of reactive oxygen species (ROS), cell permeability or protein content. While these methods reveal important information, they provide the average information for the whole population, failing to describe cell-to-cell variability and bet-hedging strategies Veening *et al.* (2008). To better understand stress-driven changes, single-cell technologies provide new opportunities.

There are several single cell technologies available to study the response of individual cells to stress. For example, fluorescent labels that tag certain cellular functions (membrane potential, intracellular enzyme activity, a stress reporter) can be used in combination with flow cytometry Delvigne *et al.* (2015); Porter *et al.* (1995) or imaging techniques Benomar *et al.* (2015). Single-cell (multi)-OMICs open the door to a very detailed understanding of the metabolism of individual cells, although it is a low-throughput technique that still presents many challenges in its accuracy Bock *et al.* (2016). Raman spectroscopy is an alternative single-cell tool that can detect individual phenotypes without the use of fluorescent probes. It is an optical method that uses a laser to excite the molecules present in the cell and records their inelastic scattering, thereby generating a single-cell fingerprint that contains (semi)quantitative information on its constituent molecules, such

as nucleic acids, proteins, lipids and carbohydrates. This technique has been used to study stress-induced phenotypic differences of the cyanobacterium *Synechocystis sp.* (Tanniche *et al.*, 2020): the fingerprints of cells treated with different concentrations of acetate or NaCl and non-treated cells were differentiable using discriminant analysis or principal component analysis (PCA). Also, Teng *et al.* (2016) found that *Escherichia coli* cells exposed to several antibiotics, alcohols and chemicals had distinct Raman fingerprints. However, there are currently no quantitative methods to describe phenotypic diversity in single cells using their unlabelled Raman spectra.

A widely used set of metrics to quantify the diversity of microbial communities are Hill numbers, also known as the effective number of species, as they express in intuitive units the number of equally abundant species that are needed to give the same value of the diversity measure. Hill numbers respect other important ecological principles, such as the replication principle, that states that in a group with N equally diverse groups that have no species in common, the diversity of the pooled groups must be the N times the diversity of a single group (Chao *et al.*, 2014; Daly *et al.*, 2018). They are commonly used to quantify microbial diversity based on sequencing techniques but have also been applied to flow cytometry yielding similar results (Props *et al.*, 2016). However, phenotypic diversity at the single-cell level - defined as the diversity of observable characteristics or traits in single cells - has not yet been described. This would require multiparametric information of individual cells, something Raman spectroscopy can provide.

Quantifying phenotypic diversity at the single-cell level could be useful to follow and manage stress in bioproduction: to maintain high bioproduction rates, it is important to find or create stress-tolerant organisms. For instance, in microbial production of alcohol (considered a sustainable alternative source for chemicals and fuels), one of the major limitations is the toxicity and/or growth inhibition caused by the alcohol that is produced. The alcohol increases the fluidity of the cell membrane and causes a disruption on the phospholipid components that inhibits growth and can lead to death. It also affects nutrient uptake and ion transport. Therefore, there have been efforts in evolutionary and synthetic engineering to increase alcohol tolerance in several organisms, for example, *E. coli* and *S. cerevisiae*, widely used in bioproduction (Jia *et al.*, 2010).

In this chapter, we aim to quantify single-cell phenotypic diversity using Raman microscopy based on the Hill diversity framework. We describe the necessary steps to preprocess Raman spectra and demonstrate its integration into the Hill diversity framework. The necessary functionalities are also embedded in the open source MicroRaman package (Kerchkof *et al.*, 2017). To illustrate the use of this method, we applied it in two popular strains in bioproduction. First, we compared an *E. coli* population in stress conditions (cultivated with ethanol) with a control population. Secondly, we separated two subpopulations of a *S. cerevisiae* culture that was under nutrient-limiting conditions using a GFP tag and analyzed them using Raman microscopy. In both cases, we show how the stress-induced single-cell phenotypic diversity can be quantified using the Raman spectra of the single cells, and how this information can be used to detect a shift in the phenotype of the population. Finally, we used this information to explain how the molecular profile of the cells changes after being exposed to the stressors.

5.3 Materials and methods

5.3.1 Data sets

The cells listed in Table 5.1 were cultured in 2L Erlenmeyer flasks with a working volume of 1L, at 28℃ with 120 rpm orbital shaking. They were chosen because they are microorganisms with different shapes and characteristics that can be used for microbial protein production (Kunasundari *et al.*, 2013; Hardy *et al.*, 2018; Yan *et al.*, 2018; Cereghino & Cregg, 2000).

All cultures were aseptically inoculated in the corresponding rich liquid medium (Table 5.1), and re-cultivated every 24 to 48 h during 2 months to get sufficient biomass for the amino acid analysis (*i.e.*, 100 g of wet biomass). Briefly, 10% v/v of the cultures (100 mL) was used as inoculum for the subsequent cultivation, while the remaining culture (900 mL) was harvested via centrifugation at 6603 g for 5 min, washed with 0.1M phosphate buffer saline (PBS) and stored at -20℃ until sufficient amount of biomass was collected.

Table 5.1: List of organisms and medium used to grow them

Organism	Liquid medium	Characteristics
Cupriavidus necator LMG 1199	Nutrient broth (Oxoid Ltd, England)	Hydrogen-oxidizing bacterium
Methylobacterium extorquens DSM 1338	Nutrient Broth (Oxoid Ltd, England) with 1% methanol	Gram negative bacterium, methylotrophic
Yarrowia lipolytica ATCC 20362	YM Broth (catalogue number 271120, BD Biosciences, USA)	Fungi, can grow in hydrophobic environments
Komagataella phaffii ATCC 76273	Sabouraud Broth (catalogue number 238230, BD Biosciences, USA)	Methylotrophic yeast

5.3.2 Case studies: single-cell phenotypic diversity quantification in stress-induced phenotypes

To test the capacity of the single-cell phenotypic diversity (sc-D_2) calculation to identify metabolic changes, we used two case studies. First, we studied two *E. coli* populations that had been grown together in different conditions: one was treated with ethanol while the other was not. Secondly, a *S. cerevisiae* culture was grown in nutrient limiting conditions, which resulted in differential expression of the chimeric stress reporter (tagged with eGFP). The two subpopulations (high expressing and low expressing eGFP) were isolated (Fig. 5.1).

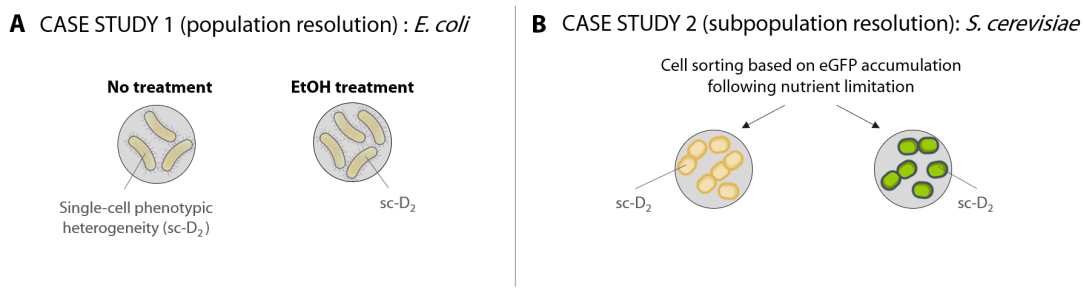


Figure 5.1: Overview of the case studies. A) Study of two *E. coli* populations grown separately with ethanol in the medium or non-treated. B) Two subpopulations were isolated from a *S. cerevisiae* culture based on the expression of the GFP marked chimeric stress reporter after nutrient limitation. The Raman spectra of single cells were used to calculate their phenotypic diversity (sc-D_2).

5.3.2.1 Population resolution: *E. coli* exposed to ethanol

The dataset from Teng *et al.* 2016 was used to validate the diversity calculations. According to their manuscript, this dataset consists of Raman spectra of *Escherichia coli* in different time intervals (5, 10, 20, 30 and 60 min, 3 h and 5 h) after being cultured with different chemical stressors. We used the ethanol-treated samples and the controls to illustrate our point. The dataset consists of three biological replicates of the cell culture and measured 20 cells per replicate.

5.3.2.2 Subpopulation resolution: *S. cerevisiae* after nutrient limitation

The prototrophic haploid yeast strain *Saccharomyces cerevisiae* CENPK 113-7D was used in this study (Nijkamp *et al.*, 2012). eGFP was produced under the control of a chimeric promoter composed of fragments of the HSP26 and GLC3 promoters. The promoter sequence was previously published (chimaera 2 in Zid & O'Shea (2014)). A synthetic construct containing the promoter, the eGFP gene and the G418 resistance marker was integrated in the genome via homologous recombination at the uga1 site. The correct insertion was confirmed via PCR analysis and lack of growth on gamma-aminobutyrate (GABA) as the sole nitrogen source.

Samples were collected after 10 residence times in a continuous culture operated at D=0.1 h-1 in a 2-liter stirred-tank bioreactor with 1 L operating volume. Defined yeast mineral medium containing 7.5 g/l was used (Verduyn *et al.*, 1992). The culture temperature was maintained at 30 °C, the stirrer speed at 1000 rpm and the air provision at 1 vvm. The culture pH was controlled at 5.0 through the automated addition of either 25% KOH or 25% M H_3PO_4.

Before cell sorting, samples were fixed in formaldehyde 4%, following the protocol from García-Timermans *et al.*, 2018. Paraformaldehyde is known to preserve the Raman spectral features better than other fixatives, such as ethanol or glutaraldehyde (Read & Whiteley, 2015). Upon reaching steady-state in nutrient limited continuous culture, yeast population was sorted in two distinct sub-populations, *i.e.*, the first one exhibited a high GFP content (high GFP) and the second one exhibiting a low GFP content (low GFP). Then,

the high GFP and low GFP subpopulations were separated using Fluorescence-activated cell sorting (FACS). For this purpose, cell suspension collected from the bioreactor was diluted 10 times in PBS (ThermoFischer scientific, Belgium) and was further analyzed and sorted with a FACSaria (Becton Dickinson, Belgium). Cells have been collected following an enrichment sorting mode. Fractions containing 10^6 cells of each subpopulation were collected. Gating details used for cell sorting can be found in the Supplementary Information.

5.3.3 Raman microscopy

For the *S. cerevisiae* samples, three drops of 2 µL were placed on a CaF_2 slide (grade 11 mm diameter by 0.5 mm polished disc, Crystran Ltd.). In each drop, 65 points were measured using a WITec Alpha300R+ with a 785nm excitation diode laser (Topotica) and a 100x/0.9 NA objective (Nikon) with 40 sec of exposure and 1 accumulation using a 300 g/mm grating.

For the samples from*C. necator, M. extorquens, Y. lipolytica* and *K. phaffi*, ∼450 points were measured using 5 sec of exposure and 1 accumulation with a 300 g/mm grating.

As a control for the instrument performance, a silicon piece (IMEC, Belgium) slide was measured with a grating of 300 g/mm, with a 1 sec time exposure and 10 accumulations. The intensity of the peak around 520 cm^{-1} was monitored over time. Laser power was monitored to detect possible variations. More information can be found in the Raman metadata aid (see Supplementary Table 5.2) collected following the guidelines from García-Timermans (2018).

5.3.4 Data analysis

The data analysis was conducted using R (R Core Team 3.6.2, 2019) in RStudio version 1.2.1335 (RStudio, 2019). Plots were produced using the package ggplot2 and ggpubr (Kassambara & Mundt, 2017; Villanueva *et al.*, 2016).

5.3.4.1 Preprocessing

We manually eliminated the spectra that contained cosmic rays. The remaning spectra were preprocessed using the R packages 'MALDIquant' (v1.16.2) (Gibb & Strimmer, 2012) or 'HyperSpec' (Beleites & Sergo, 2012). To reduce the noise in the spectra, we smoothed it using the *spc.loess()* function. The 400-1800 cm^{-1} region of the spectrum (which contains the biological information in bacteria) was selected for fingerprint. The baseline was corrected for instrumental fluctuations or background noise using the Sensitive Nonlinear Iterative Peak (SNIP) algorithm (using ten iterations) and spectra were normalized using the Total Ion Current (TIC). Then, the spectra were normalized using the *calibrateIntensity()* function and aligned per group with the *alignedSpectra()* function. These preprocessed data were used to calculate the single-cell phenotypic diversity and principal coordinate analysis. We decided to align per group in this chapter because we select certain peaks to quantify (in a semi-quantitaive way) the levels of different biomolecules present in the individual cells.

5.3.4.2 Single-cell phenotypic diversity calculation (sc-D_2) for single cells with Raman microscopy

The Hill equations were adapted in this chapter to quantify the phenotypic diversity of single cells using preprocessed Raman spectra. Every Raman signal corresponds to a single or multiple chemical bond that correspond to a metabolite(s). We have called these regions components (x). The relative abundance of each component was normalized, by calculating their relative abundance. Then, they were used in the Hill equation as described in the Results section.

Hill numbers are commonly used to calculate microbial diversity. They are also known as the effective number of species, as they express in intuitive units the number of equally abundant species that are needed to give the same value of the diversity measure. Hill numbers respect other important ecological principles, such as the replication principle, that states that in a group with N equally diverse groups that have no species in common, the diversity of the pooled groups must be the N times the diversity of a single group. In the general introduction, the equation can be found. In this chapter, we focused on sc-D_2,

as it considers both richness and evenness.

5.3.4.3 Statistical analysis

Normality was studied using *ggdensity()* and *ggqqplot()* from the package 'ggpubr' (Villanueva *et al.*, 2016).

The *E. coli* samples followed a non-parametric distribution, and thus to compare multiple groups (the mean phenotypic diversity (sc-D_2) of ethanol and the control group over time) we did an ANOVA test using the function *aov()*. Post-hoc testing of pairwise differences was done using *Tukey_HSD()*. Both functions are from the package 'stats'.

The expression of the biomolecules in the two *S. cerevisiae* subpopulations followed an non-parametric distribution, and thus were analysed using Wilcoxon test with the function *wilcox.test()* from the package 'stats'.

5.3.4.4 Principal coordinate analysis (PCoA)

The principal coordinate analysis (PCoA) was calculated as the eigenvalues divided by the sum of the eigenvalues.

5.3.4.5 Sampling size

We used a dataset of 4 axenic cultures (described in Table 5.1) and measured $\sim$450 Raman spectra per sample, for which we calculated their single-cell phenotypic diversity (sc-D_2). Then, we did 1000 simulations were the data were permuted, and calculated the average D_2 when using a increasing number of spectra. The average and standard deviation of these 1000 simulations were plotted.

5.3.4.6 Subpopulation types

Subpopulation types were calculated by adapting the code for flow cytometry data. The method was originally intended to separate sample clusters, while in its application for Raman spectroscopy we aim to identify and differentiate cell clusters (Props *et al.*, 2016).

First a PCA is performed to reduce the dimensionality of the data. A reduced dataset with the principal components that explain the majority of the variance ($>$40%) are used to calculate the optimal number of clusters using the silhouette index, and then used partitioning around medoids (PAM) as a clustering algorithm to determine to which cluster cells belong to. This was done using the *pam()* function from the package 'cluster' (version 2.1.0). Once every cell was assigned to a phenotype (cluster), the median phenotype to which the (sub)population corresponds to was calculated.

5.3.4.7 Data availability

The analysis pipeline and the raw data can be found in https://github.com/CMET-UGent/Raman_PhenoDiv.

5.4 Results

5.4.1 Phenotypic diversity quantification of Raman spectra using Hill numbers

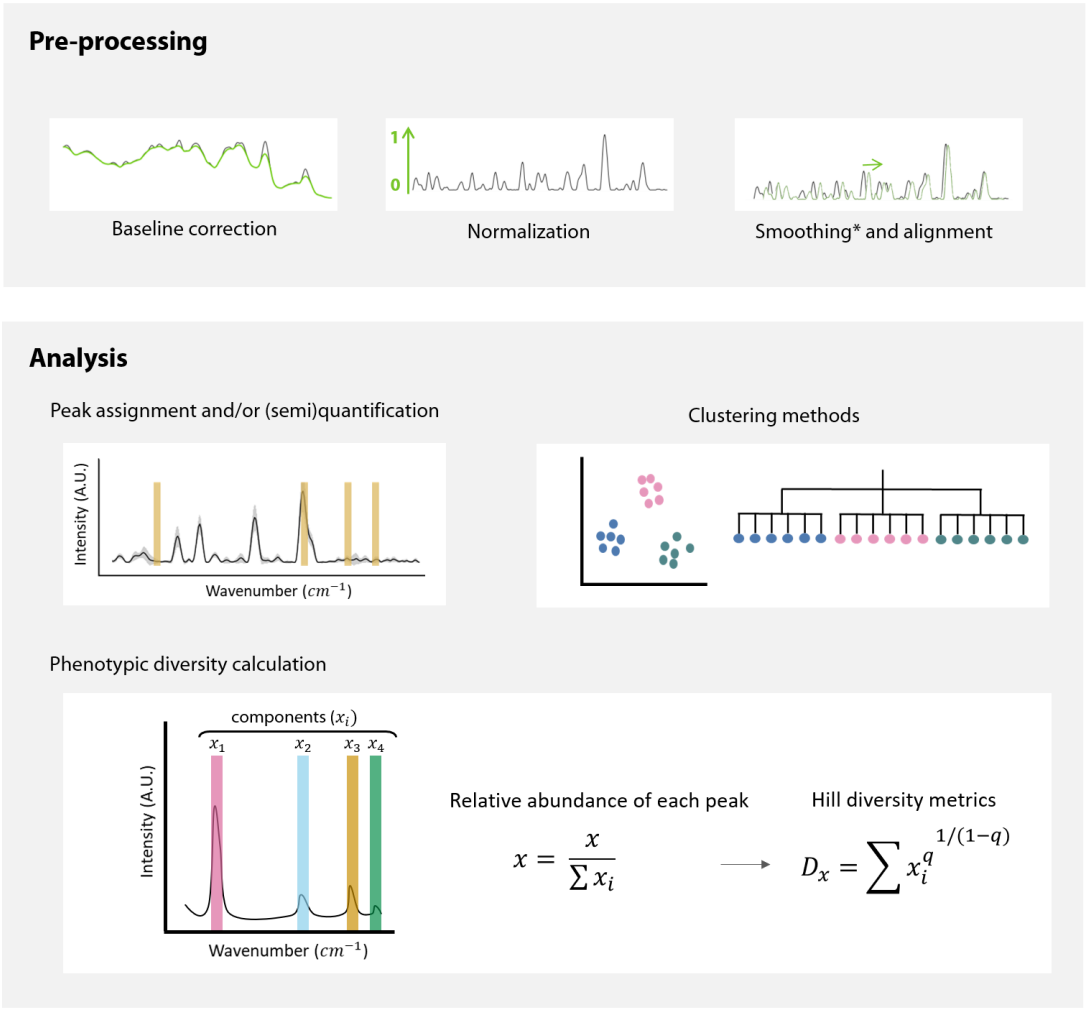


Figure 5.2: Summary of the preprocessing and analysis of the Raman spectra. After removing the cosmic rays, the baseline is corrected and the spectra are normalized. Spectra can be smoothed and aligned; however, smoothing can erase potentially relevant information, and should be carefully considered. Similarly, alignment can produce faulty spectra by displacing the signal, and thus need to be used reasonably. Once the spectra are preprocessed, it is possible to (1) extract (semi)quantitative information (2) cluster cells or create phenotypic trees or (3) calculate the single-cell phenotypic diversity. For the latter, Raman peaks that correspond to one or several metabolites are considered as components. The intensity of these components (x) is used to quantify phenotypic diversity. The order of diversity (q) can be 0, 1 or 2, meaning respectively that richness, evenness or both parameters are considered in the metric. This equation considers richness and evenness of metabolites in a single cell.

Single-cell phenotypic changes can be captured by Raman spectroscopy, by which information is collected on the (bio)molecules present in individual cells. Once the Raman spectra are acquired, the raw data need to be preprocessed (Fig. 5.2- Preprocessing). This step aims to remove noise from spectra and to be able to extract meaningful biological information. First, the spectra that contain cosmic rays need to be removed manually or automatically (Wahl *et al.*, 2020). Then, we select the spectral region that is most relevant for microbial fingerprinting, around 500-2000 cm^{-1} (Huang *et al.*, 2010). Once this region of the spectra is selected, the first step in the preprocessing is to correct the baseline, that can be degraded due to instrument fluctuations or background-signal influence (Liu *et al.*, 2015; Wahl *et al.*, 2020). Then, the spectra are normalized to avoid that the absolute intensity masks the variation of signals of interest (Beattie *et al.*, 2009; Gautam *et al.*, 2015). It is also possible to align and/or smooth the Raman signal, but these steps can introduce noise to the measurements and should be carefully considered.

After the spectra have been preprocessed, different information can be extracted (Fig. 5.2-Analysis). For example, peaks of interest can be selected for semi-quantitative analysis or quantitative analysis using a calibration curve Butler *et al.* (2016). Also, the whole spectra can be used to classify cells using several clustering methods, such as principal component analysis, principal coordinate analysis, non-metric multidimensional scaling or T-distributed stochastic neighbour embedding. This information can also be used to construct dendrograms (García-Timermans *et al.*, 2018). Here we used the preprocessed spectra to quantify the single-cell phenotypic diversity using Hill numbers. Every Raman peak corresponds to a different metabolite or a combination of metabolites, called components (x) (Fig. 5.2). To calculate the relative abundance of each peak, the intensity of the signal of each component was normalized by the sum of all intensities, and this information was then used in the Hill equations.

The order of diversity (q) can be 0, 1 or 2, meaning that richness, evenness or both richness and evenness are taken into account in the metric. sc-D_0 contains information about the number of components (x_i) in the Raman spectra, and is calculated as shown at the end of Figure 5.2. sc- D_1 informs about the evenness of each component. In this chapter, we mostly focus on single-cell D_2 (sc-D_2) (q=2) as it takes both richness and evenness of the Raman components into account.

5.4.2 Sample size dependence of phenotypic diversity (sc-D_2) measurements

To understand the distribution of single-cell phenotypic diversity in a population, we did ∼450 measurements in 4 axenic cultures of *C. necator, M. extorquens, Y. lipolytica* and *K. phaffi*. We calculated the average diversity estimation for an increasing number of spectra and bootstrapped 1000 times. The average of the total number of measurements is plotted in grey, and the 5% of this average is represented with a dotted grey line.

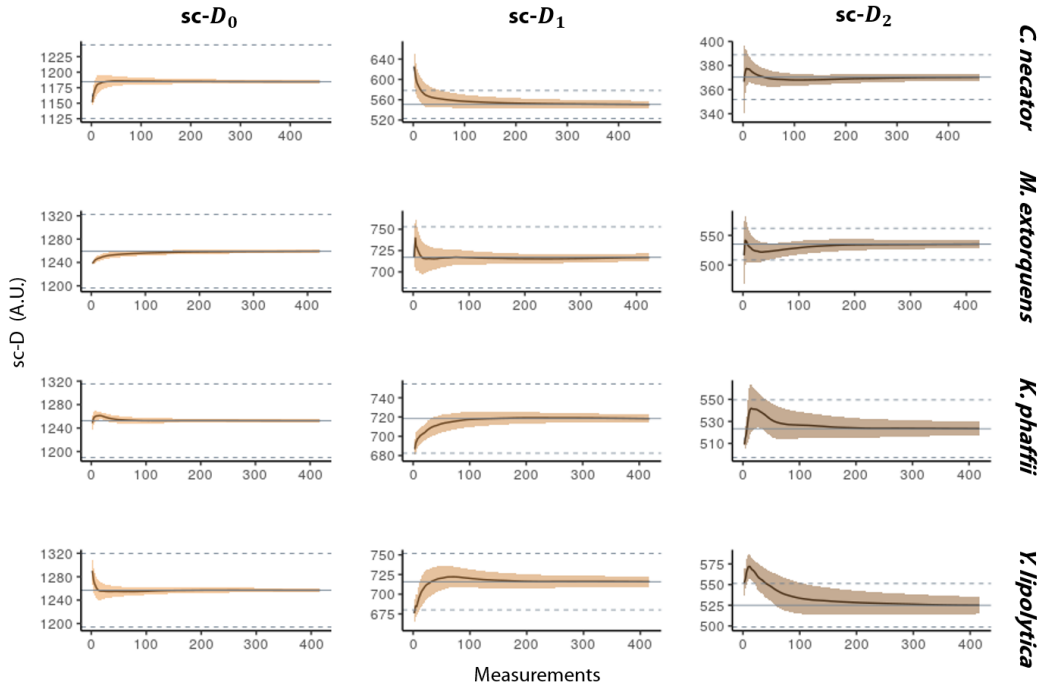


Figure 5.3: Effect of sampling size on the single-cell phenotypic diversity average. We calculated the average single-cell phenotypic diversity using the Hill equations (single-cell D_0, D_1 and D_2) for an increasing number of measurements and repeated the calculation picking spectra randomly 1000 times. We used the Raman spectra of four pure cultures and ∼450 measurements on each. The smear represents the standard deviation. The grey line represents the average sc-D value of the total population, and the dashed lines a 5% deviation from the mean.

We looked at how many measurements were needed to calculate the population average (grey line) and how many are needed to have an accurate estimation (95%, dashed lines). For the estimation of sc-D_0, few measurements ($\sim$10-50) were needed to obtain the population average. The sc-D_1 calculation grants a greater weight to high-intensity wavenumber and/or peaks of these components, and required $\sim$100 measurements. Although *M. extorquens* reaches it after $\sim$20 measurements. The sc-D_2 estimation takes both the number of components and their abundance into account and needed between $\sim$50 (*C. necator*) to $\sim$180 (*Y. lipolytica*) measurements to estimate the population average.

5.4.3 Case studies: phenotypic diversity quantification in stress-induced phenotypes

When stress is applied in a microorganism, a set of genes and proteins are expressed, changing the metabolic phenotype of the cell. This metabolic change can be captured by Raman spectroscopy, that collects information on the (bio)molecules present in individual cells. To compare stressed and non-stressed cells, we quantified their phenotypic diversity using our proposed methodology, as shown in Figure 5.1. First, we compared two *E. coli* cultures growing in different conditions: with ethanol (stressed) or non-treated (control). Then, we compared two subpopulations of the same *S. cerevisiae* culture, separated based on their expression of the GFP stress reporter in nutrient-limiting conditions.

5.4.3.1 Tracking *E. coli* population diversification dynamics following exposure to ethanol stress

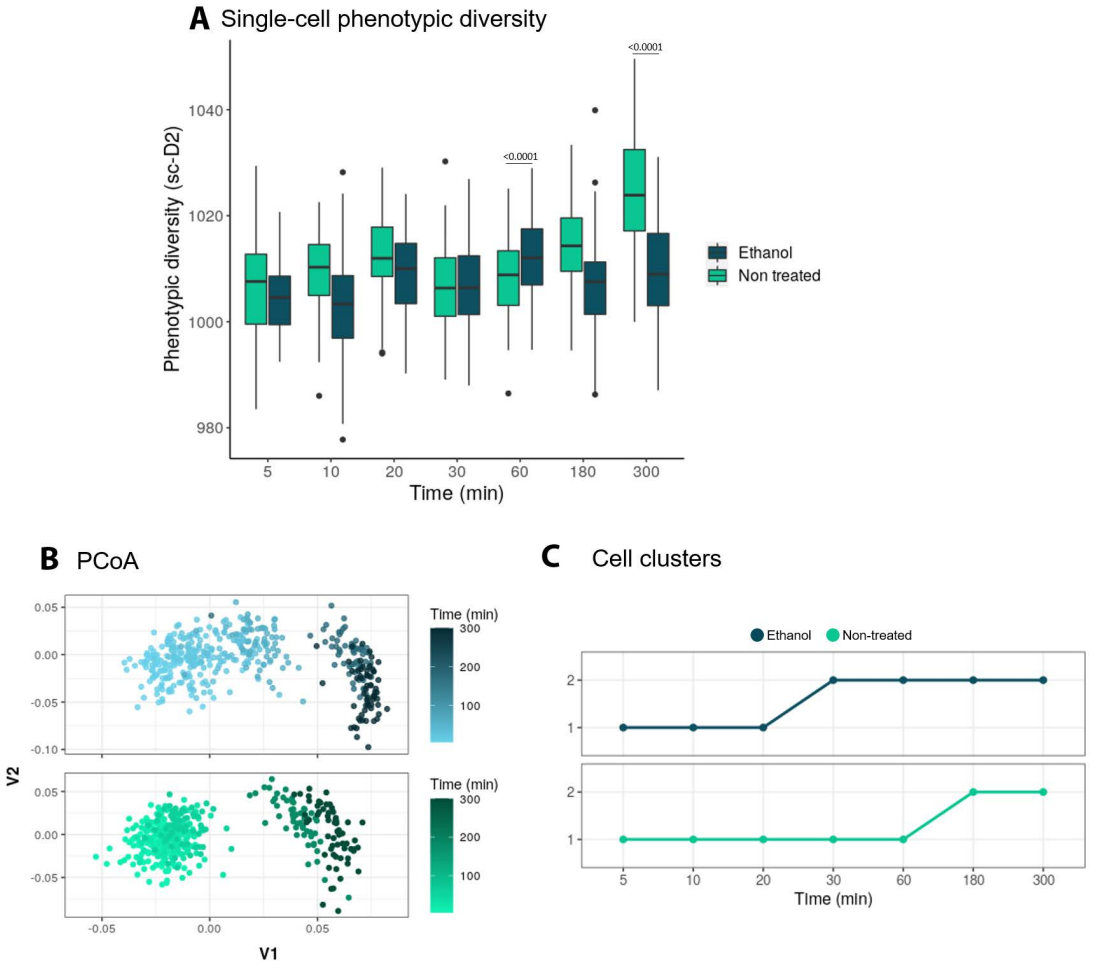


Figure 5.4: A) Single-cell phenotypic diversity (sc-D_2) of the stressed (ethanol treated) and non-stressed (non-treated) *E. coli* populations. Treatments and treatments over time are significantly different (two-way ANOVA, $p \leq 0.0001$). A post-hoc Tuckey test showed that the ethanol and control groups are significantly different on timepoint 60 min and 180 min ($p \leq 0.0001$). B) Raman fingerprint of the stressed (ethanol treated) and non-stressed (non-treated) *E. coli* populations, plotted using principal component analysis (PCoA). The time progression is represented with a darker colour. Every point represents a single cell. C) The clustering algorithm shows the phenotypic shift happens after 20 min for the ethanol-treated population and after 180 min for the control. Two phenotypes were found. Every point represents the average "phenotypic type" of the population. N=60

We used a dataset from Teng *et al.* (2016), consisting of spectra of *Escherichia coli* sampled at different time points (5, 10, 20, 30 and 60 min, 3 h and 5 h) after being cultured in standard conditions or with ethanol. There were three biological replicates of the cell culture and 20 cells were measured per replicate.

The stress-induced metabolic diversity of single cells was quantified using the sc-D_2 Hill equation and the average diversity for each population (stress and non-stressed) was plotted (Fig. 5.4A). After testing for normality, a two-way ANOVA test showed a significant difference between treatments and treatments over time ($p \leq 0.0001$). A post-hoc Tukey test showed that the ethanol and control groups were significantly different at time point 60 min and 180 min ($p \leq 0.0001$). Then we used PCoA, a common clustering method to visualize the dissimilarities in the fingerprints. The Raman fingerprint of the stressed and control cells is similar at the beginning and then shift over time (Fig. 5.4B). We used a clustering algorithm to define exactly when this shift takes place: after 20 min for the ethanol-treated population and 180 min for the control population Fig. 5.4C).

5.4.3.2 Discriminating *S. cerevisiae* subpopulations following exposure to nutrient limitation

A *S. cerevisiae* population was cultured in nutrient-limiting conditions. Based on GFP expression as an indicator of stress activation, we separated two subpopulations (one that activated the stress reporter, and one that did not) using FACS. Then, we analyzed 65 cells in each subpopulation using Raman microscopy.

First, we calculated the single-cell phenotypic diversity (sc-D_2) of the subpopulations with high (+) or low(-) stress reporter expression. To prove that sc-D_2 calculations are quantitative, we also created an *in silico* group by mixing the two subpopulations (Fig. 5.5A). The *in silico* mix group was expected to have an average sc-D_2. Then, we checked the dissimilarity of the fingerprints using PCoA (Fig. 5.5B). Two clusters are differentiated depending on the reporter expression.

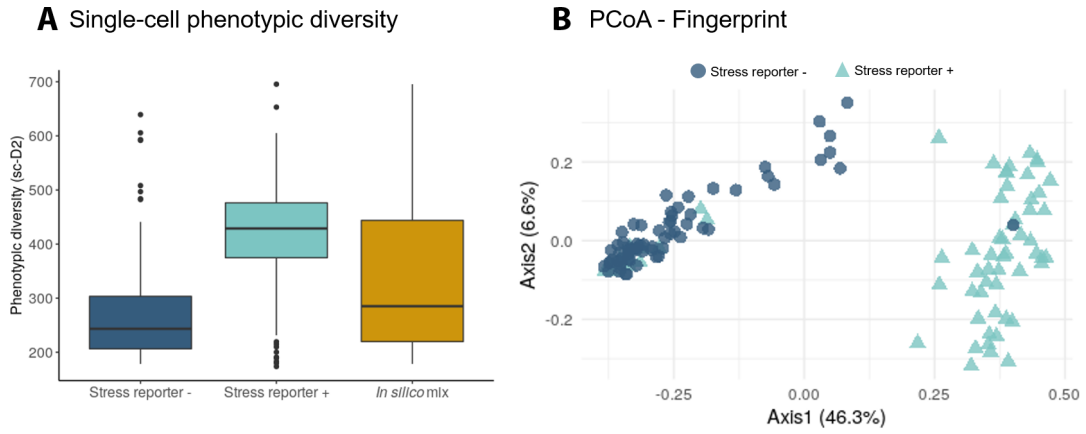


Figure 5.5: A) Single-cell phenotypic diversity of a *S. cerevisiae* subpopulations with high or low stress reporter expression and an *in silico* mix of both groups. The *in silico* is a random selection of cells coming from the stressed and non-stressed population B) Visualization of the stress-induced phenotypic change of *Saccharomyces cerevisiae* subpopulations with high or low stress reporter expression using principal coordinates analysis (PCoA). Every dot is a single cell. The size of the dot corresponds to the single-cell phenotypic diversity (sc-D_2). N= 65.

The information of the Raman spectra from each group was used to understand the effect of the stress reporter activation on the metabolic response of *S. cerevisiae*. Using a tentative assignment based on Teng *et al.* (2016), we estimated the protein (1006 cm^{-1}), total lipid (1450 cm^{-1}), nucleic acid (786 cm^{-1}) and saturated lipid (1132 cm^{-1}) content in in the subpopulations with a high or low stress-reporter expression (Fig. 5.6). There can be spectral shifts between databases, due to the use of a different laser and instrument, and/or because of the handling of the sample. To examine these phenomena, we took as a reference the 1002 cm^{-1} peak, that corresponds to the aromatic amino acids phenylalanine and/or tyrosine. It is a very prominent band that is usually recognizable in biological samples (De Gelder *et al.*, 2007). We found that both groups have a significantly different metabolism: the subpopulation with a high (+) expression of the stress reporter had a higher protein content, but contained less total and saturated lipids, and nucleic acids (Wilcoxon rank-sum test, $p \leq 0.0001$). However, this peak assignment is only tentative as it is based on the literature, and should be validated using another technique. We cannot claim with certainty the exact molecular identity corresponding to each Raman

wavenumber, as further explained in the discussion section.

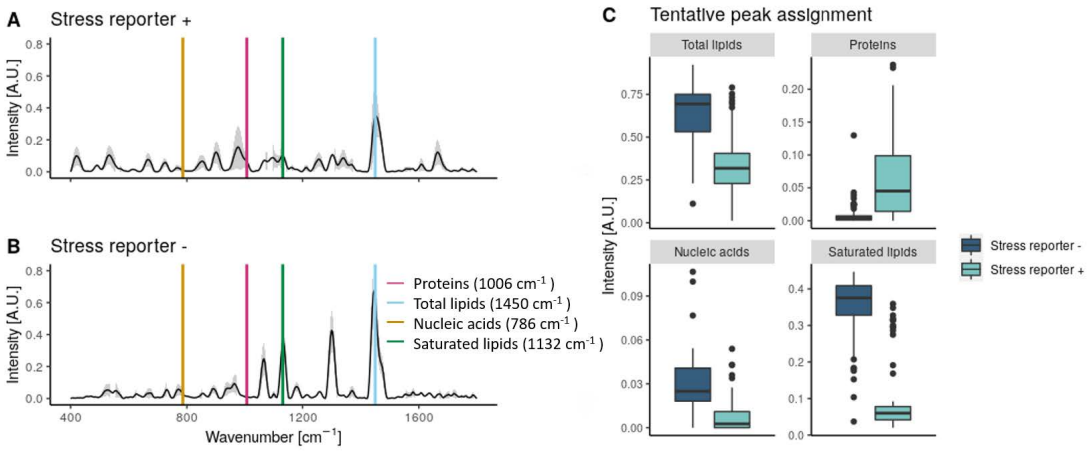


Figure 5.6: Raman spectra of *S. cerevisiae* subpopulations with high (+) or low (-) expression of the stress reporter (A,B). The average of the spectra is plotted with a black line and the standard deviation in grey. The putative peaks corresponding to proteins, lipids, nucleic acids and unsaturated lipids according to Teng and colleagues (2016) are plotted over the spectra. C) The intensity of the metabolic peaks highlighted in plot A and B for the subpopulations with high or low expression of the stress reporter. The *p* values for the Wilcoxon test for every metabolite is shown. N=65.

5.5 Discussion

This work shows how Raman spectra data can be used to study stress-driven metabolic heterogeneity at the single-cell level. The laboratory and computational workflow is relatively fast and non-destructive, and can provide (semi)quantitative information about the biomolecular composition of cells. Although Raman spectroscopy has been previously used to detect stress-driven phenotypes (Tanniche *et al.*, 2020), we argue that there is a need for quantitative single-cell measurements for phenotypic diversity and propose the use of Hill numbers. We chose Hill numbers for our calculations because they are widely used in microbial ecology.

To estimate phenotypic diversity using Hill numbers, we considered that each Raman

signal corresponds to a component (a single or multiple chemical bonds), and that the intensity of these components is correlated with their quantity (Tang *et al.*, 2013; Wu *et al.*, 2011). After normalizing the components, they were used in the Hill equations (Fig. 5.2). Although we chose to use the whole spectrum for this calculation, it is possible to select only the peaks. This could influence the resolution: algorithms for peak detection typically divide the spectrum according to a certain window size and look for the local maximum (Gibb & Strimmer, 2012). Using this algorithm would not take into account the width of the peaks, which is a characteristic of the chemical bonds. Also, some components with a close signal would be ignored, and the choice of window size would affect the final result. The way in which the Raman spectra are preprocessed will have an impact on the results. The region used for fingerprinting needs to be considered so that all the relevant biomolecules to address the hypothesis are reported. Both the baseline correction and normalization will have an impact on the intensity reported for the different components. Smoothing functions assume spectra are noise, and erase certain signal. Finally, aligning spectra when unnecessary can misplace the signals. Using the same preprocessing steps when comparing samples is crucial, as well as detailing the preprocessing steps and providing the raw data.

To explore the importance of the sample size in these estimations, we used a large dataset consisting of $\sim$450 Raman spectra from 2 axenic bacterial cultures (*C. necator* and *M. extorquens*) and 2 axenic yeast cultures (*Y. lipolytica* and *K. phaffi*). Then, the effect of the sampling size on the average single-cell phenotypic diversity and its standard deviation was calculated. Our results show that this is highly population-dependent: for example, while *C. necator* only needed 15 spectra to approach the expected sc-D_2 average, *Y. lipolytica* needed more than 150 measurements (5.3). This could be due to a different degree of phenotypic diversity in the populations. Sample size should be explored for every experiment, to make sure that the estimations are representative.

After developing the methodology to quantify single-cell phenotypic diversity, we applied it to two case studies to demonstrate its use. We focused on sc-D_2, as it considers how many components are being expressed per cell, and their evenness. In the first case study, we compared an ethanol-treated and a control *E. coli* population. We found that when *E. coli* is grown in standard conditions, there is a phenotypic shift after 60 min. This shift happens earlier in stressed cells (20 min) (Fig. 5.4C). The shift in the fingerprint in the

control group could be due to the entering in the log phase. Our group previously showed how *E. coli* start their log phase after $\sim$1h of cultivation in rich medium, and how at different growth stages bacteria change their phenotype (García-Timermans *et al.*, 2019). Although both the ethanol-treated and the control populations end up having a similar phenotype after 60 min, the stressed population has a lower metabolic diversity (Fig. 5.4A), a lower nucleic acid content and a higher protein and lipid content. Clustering algorithms are useful to automatically identify phenotypes and quickly asses when the phenotype of a population has changed in a reproducible way. While here we use PCA, other metrics can be used, such as non-metric multidimensional scaling (NMDS), t-distributed Stochastic Neighbor Embedding (t-SNE) and other clustering methods. The choice of the clustering method should be based on the hypothesis, and how important it is to conserve the distances between the cells and the relative size of the cluster.

In the second case study, we analyzed the response of two *S. cerevisiae* subpopulations. When in nutrient-limiting conditions, *S. cerevisiae* resorts to a bet-hedging strategy where some yeasts will enter a quiescent state, while others will activate a stress-induced response (Gray *et al.*, 2004). The strain used in this experiment produces GFP upon activation of nutritional stress, so when the *S. cerevisiae* culture diversified into two populations -with either high or low expression of the stress reporter- these were separated using FACS and analyzed with Raman microscopy. Because the Raman spectroscope used has a 785 nm laser, we do not expect the fluorescent signal (excited at 510 nm) to be picked up with this instrument. Single-cell phenotypic diversity (sc- D_2) in the stressed subpopulation is higher than the non-stressed (Fig. 5.5A). As expected, the *in silico* mix shows a diversity that is close to the average of both subpopulations. We then checked that the subpopulations with high and low stress reporter expression have a very distinct fingerprint using PCoA, a tool widely used for Raman spectra in microbial ecology (Fig. 5.5B). Using the metabolic information contained in the Raman spectra, we found a higher nucleic acid content in the non-stressed subpopulation (in line with the findings of Teng (2016) in stressed *E. coli* cells). This could be explained by the higher ribosome content in non-stressed cells. We also found that the stress response triggered by the activation of the chimeric promoter results in a raise of protein production (Fig. 5.6), similar to the results found in stressed *E. coli* cells. We cannot, on the other hand, exclude that GFP production may have somewhat influenced the molecular fingerprint of cells (*e.g.*, via depletion of amino acids pools, reducing ribosomal availability). Also, differences in protein

abundance between the stressed and non-stressed subpopulations could be (at least partially) due to the GFP protein itself. To explore these possibilities, a proteomics and/or transcriptomics analysis at the single-cell level would be required. The choice of this promoter based on a fusion of *glc3* and *hsp26* as a single proxy to define a metabolically stressed population is cross validated by these findings, that show two clearly metabolically distinct subpopulations. It is important to mention that these metabolic estimations were made using an external database, and should be considered as tentative assignments. To confirm these results, a second technique should be used. Other authors that have previously explored stress-induced responses in yeast using Raman spectroscopy, found that the 1602 cm^{-1} band - that corresponds mainly to ergosterol production (Chiu *et al.*, 2012) - can indirectly measure oxidative stress and cellular metabolism after atmospheric or nutrient changes (Huang *et al.*, 2007). This band can be used as a label-free in vivo activity indicator in both *S. cerevisiae* and *S. pombe*.

Finally, we explored whether the number of cells measured in both case studies was enough to capture the diversity of the cultures. In *S. cerevisiae*, 65 cells were enough to estimate single-cell diversity, and most biomolecules (Supplementary Figures 5.8 and 5.9). However, to properly estimate the protein content in the non-stressed subpopulation more cells would have been needed. The laser spot used for these measurements had a diameter of 1.7 µm, and thus the spectra of the yeasts could have varied depending on the position of the laser inside the yeast. To avoid this variation, the operator aimed at the center of the yeast to the best of her ability. Also, Supplementary Figure 5.9 shows how after measuring 40-50 cells the measurements become quite representative of the population. However, a space-resolved experiment would help in gaining an insight into the effect of stress in the cell wall or other structures of *S. cerevisiae*. In the *E. coli* population, we tested the sample size in the ethanol-treated population at timepoint 5 min and 300 min. Very few cells are needed to have a representative single-cell diversity estimation: the sc-D_0 is the same for all cells (Supplementary Figure 5.10). This metric looks at the number of components present in each cell, which in this case seem to be the same for all individuals. It could be that these cells express the same molecules, but different amounts, and/or an artefact of the pre-processing carried out by Teng and colleagues, that could have erased some of the smaller peaks. This highlights the importance of making the raw data available, following the trends of other disciplines such as new generation sequencing (NGS) or flow cytometry.

Inferring metabolic expression from Raman spectra in microbial cells is not without challenges. For instance, many databases propose different peaks to identify the same biomolecules. In this chapter, we have chosen those presented in Teng *et al.* (2016) to be able to compare the results they found in *E. coli* and we found in *S. cerevisiae*. To account for the wavelength shift between these two databases (due to the use of a different laser and instrument, and/or because of the handling of the sample), we took as a reference the 1002 cm^{-1} peak, that corresponds to the aromatic amino acids phenylalanine and/or tyrosine. It is a very prominent band that is usually recognizable in biological samples (De Gelder *et al.*, 2007). This is at 1006 cm^{-1} in the *S. cerevisiae* dataset, so we accounted for a 4 cm^{-1} shift between datasets. It is worth mentioning that in **chapter 6** we found that using two peaks that correspond to aromatic peaks correlates best with total protein content. However, here we decided to use only one peak, following the findings of Teng *et al.* (2016). When studying the content of biomolecules with Raman spectroscopy, we also need to consider that some molecules are not Raman active (*i.e.*, their chemical bonds have a weak signal), and thus will not be reflected in the spectra. Conversely, some Raman active molecules can be overrepresented in the analysis because certain chemical bonds exhibit a strong Raman signal. Also, Raman peaks can correspond to several molecules due to the presence of shared chemical bonds. These limitations should be considered when using Raman spectroscopy for microbial ecology. A better assignment of the Raman signals will also contribute to an improved understanding of the metabolic changes driving single-cell phenotypic heterogeneity.

Most ecological studies use low-dimensional physiological data, or use single marker-gene expression to understand microbial populations. Raman spectroscopy is a promising single-cell technology able to quantify phenotypic diversity in individual cells, identify changes in phenotypes, and infer metabolic information (semi)quantitatively. This tool will allow microbial ecologists to go beyond community measurements, and shed light on how heterogeneity shapes populations.

5.6 Conclusions

- Raman microscopy can be used to quantify single-cell stress-driven phenotypic diversity in microbial populations.
- Raman spectral points correspond to different chemical bonds (or to multiple ones), that are expressed with a certain intensity and evenness. Using this information in the Hill diversity framework, we can estimate the phenotypic diversity in single cells. We show that these methods work to study changes at the population and subpopulation level in both prokaryotes and eukaryotes.
- The Raman spectra contain information about the biomolecules present in a cell, and can be used to study the metabolic shift in stressed cells.
- We propose an automatic classification of phenotypes using clustering methods. This is a useful tool to track changes in singe-cell physiology.
- As Raman spectroscopy can detect stressed phenotypes, we propose it as a tool for monitoring microbial populations in bioproduction.

5.7 Appendix

5.7.1 Acknowledgements

The authors thank the funding that made this research possible. CGT is funded by the Flemish Fund for Scientific Research (FWO G020119N) and by the Geconcerteerde Onderzoeksacties (GOA) research grant from Ghent University (BOF15/GOA/006). RP is supported by the Flemish Fund for Scientific Research (FWO). BZ is supported by a post-doctoral grant through an Era-Cobiotech project (”ComRaDes” Computation for Rational Design: From Lab to Production with Success). MS is supported by the Catalisti cluster SBO project CO2PERATE (“All renewable CCU based on formic acid integrated in an industrial microgrid”), with the financial support of VLAIO, Belgium (Flemish Agency for Innovation and Entrepreneurship). This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant 722361.

5.7.2 Supplementary information

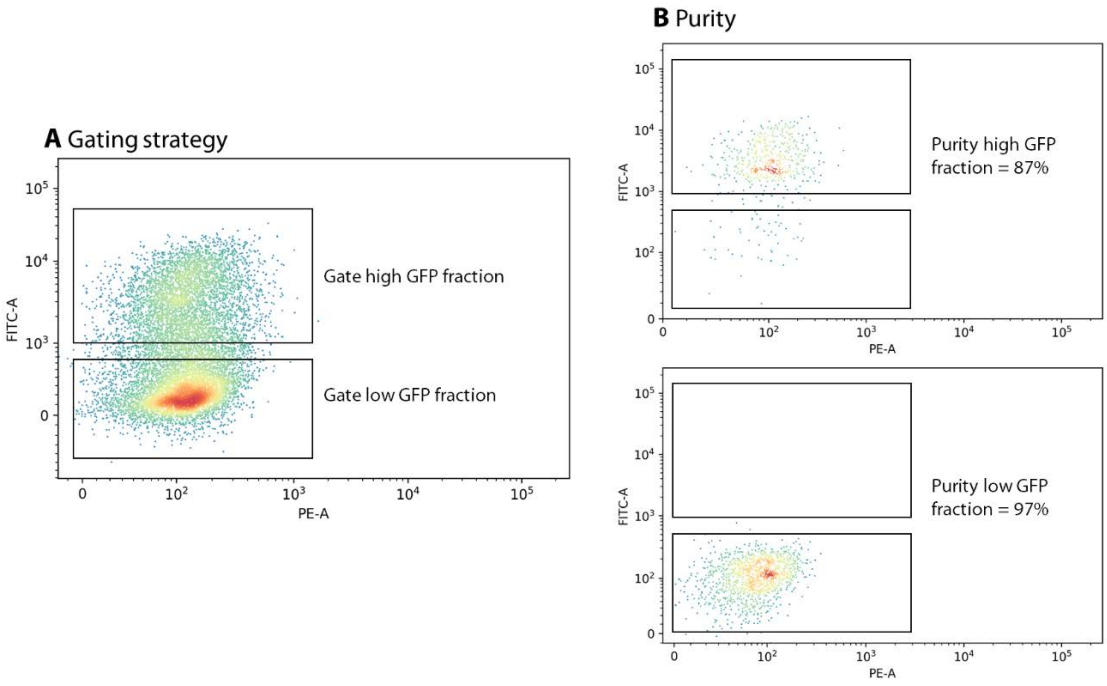


Figure 5.7: A) FACS gating strategy. B) Purity of the high GFP and low GFP fractions.

Table 5.2: Metadata aid for Raman spectra

Experiment overview	
Hypothesis	The *Saccharomyces cerevisiae* strain CENPK 113-7D with an eGFP tag in its chimeric promoter was used in this experiment. After growing the cells, they were fixed in formaldehyde 4% and sorted into two groups depending on their GFP expression (high or low) using fluorescence-activated cell sorting (FACS). Then 65 cells of each group were measured. To understand the number of measurements that need to be made by sample,we did $\sim$ measurements in 4 axenic cultures of *C. necator, M. extorquens, Y. lipolytica* and *K. phaffi*.
Variable(s) tested	Subpopulation differences
Conclusions	Our pipeline can discriminate the two subpopulations
Quality control (internal/external)	Silicon check
Samples and sample acquisition	
Material and source	*Saccharomyces cerevisiae*. *C. necator, M. extorquens, Y. lipolytica* and *K. phaffi*
Growing conditions/sampling	See description in materials and methods
Filename format:	<Replicate number>_<Treatment>_<Cell number>
Label in the samples	No label used
Fixation method	Filtered 4% formaldehyde solution from PFA
Integration time and accumulations	For the *S. cerevisiae* samples, 40 sec time exposure and 1 accumulation. For the samples from *C. necator, M. extorquens Y. lipolytica* and *K. phaffi*, $\sim$450 points were measured using 5 sec of exposure and 1 accumulation.
Grid	300 g/mm
Instrument	
Laser	785 nm excitation diode laser (Toptica). 175 mW of power before the objective.
Quality control	A silicon piece (IMEC, Belgium) sample was measured with a grating of 600 g/mm, with a 1 sec time exposure and 10 accumulations. Laser power was also monitored to detect possible variations.
Objective used (magnification) / Numeric aperture (NA)	100x/0.9 NA (Nikon)
Camera	-70℃ cooled CCD camera (iDus 401 BR-DD, ANDOR)
Dry/water/oil objective	Dried samples
Model of spectroscope	WITec Alpha300R+
Other specifications (chromatic/flat field correction/other)	
Data analysis	
Background subtraction method (if used)	No. Measurements with cosmic rays were deleted
Normalization method (peak /min-max /area under-curve /other)	Area under the curve ('Total Ion Current')
Smoothing and interpolation (if done)	Smoothing, baseline correction, normalization and alignment (per group)
Statistics/Machine learning algorithm	Wilcoxon test for pairwise comparisons between two groups. Boruta
Accessibility	https://github.com/CMET-UGent/Raman_PhenoDiv
Other relevant information	

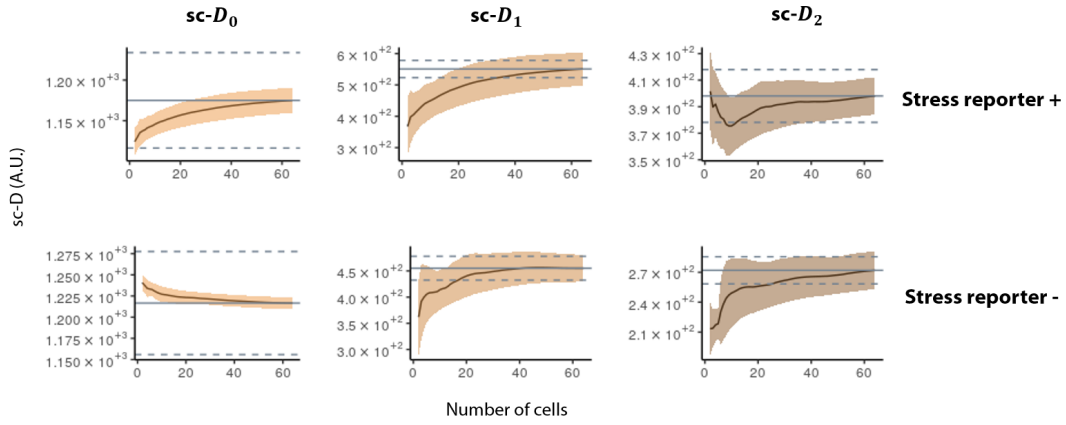


Figure 5.8: Effect of sampling size on the single-cell phenotypic diversity average of *S. cerevisiae*. We calculated the average single-cell phenotypic diversity using the Hill equations (single-cell D_0, D_1 and D_2) for an increasing number of cells in two *S. cerevisiae* subpopulations, with either high or low stress reporter expression. We repeated the calculation picking cells randomly 1000 times. The smear represents the standard deviation. The grey line represents the average sc-D value of the total population, and the dashed lines a 5% deviation from the mean. N=65.

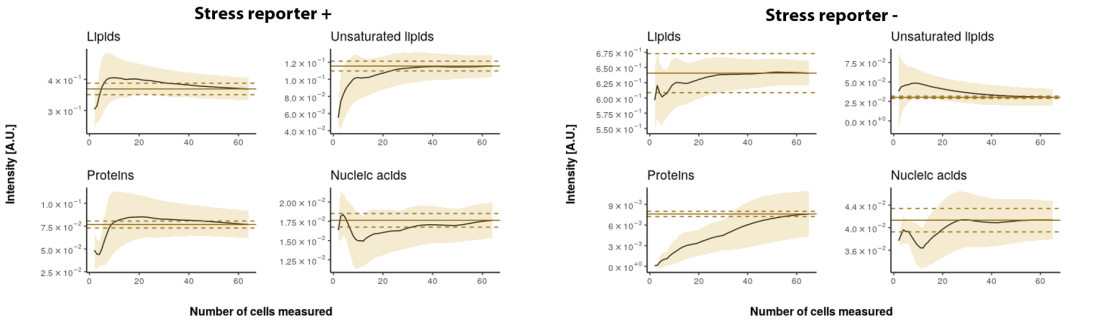


Figure 5.10: Effect of sampling size on the single-cell phenotypic diversity average of *E. coli*. We calculated the average single-cell phenotypic diversity in an *E. coli* population after being exposed to ethanol for 5 and 300 min. We used the Hill equations (single-cell D_0, D_1 and D_2) for an increasing number of cells, and repeated the calculation picking cells randomly 1000 times. The smear represents the standard deviation. The grey line represents the average sc-D value of the total population, and the dashed lines a 5% deviation from the mean. N=60.

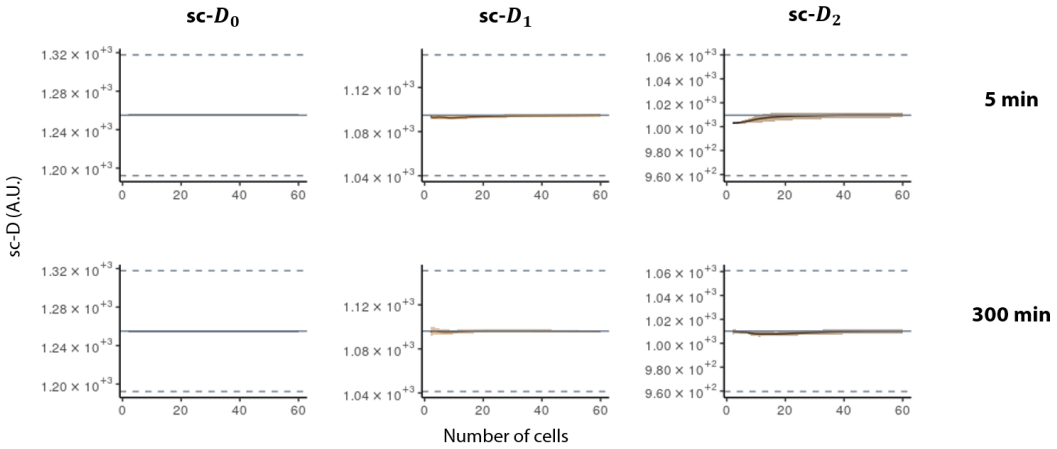


Figure 5.9: Effect of sampling size on the estimation of biomolecules in *S. cerevisiae*. We calculated the average of lipids, unsaturated lipids, proteins and nucleic acids in two *S. cerevisiae* subpopulations two subpopulations, with either high or low stress reporter expression for an increasing number of cells. We repeated the calculation picking cells randomly 1000 times. The smear represents the standard deviation. The grey line represents the average sc-D value of the total population, and the dashed lines a 5% deviation from the mean. N=65

5.7.3 Availability of data and material

The raw data and code to reproduce the analysis shown in this chapter can be found in the repository https://github.com/CMET-UGent/Raman_PhenoDiv

The dataset from Teng *et al.* 2016 was used to validate alpha and beta-diversity calculations, as well as the 'subpopulation type' definition.

5.7.4 External data set

We included the data set from Teng *et al.* in order to validate the generalizability of the PhenoGraph and t-SNE algorithms for the analysis of label-free bacterial Raman data. As described in their article, they tested the stress response of *E. coli* to six chemical stressors at different time intervals with label-free Raman spectroscopy: ethanol, antibiotics ampicillin and kanamycin, n-butanol or heavy metals Cu^{2+} ($CuSO_4$) and Cr^{6+} (K_2CrO_4).

et al. showed that each of these treatments resulted in a different phenotype. In other words, each treatment resulted in a unique Raman characterization of cells, which should group together upon analysis. These treatments were therefore used as label according to which PhenoGraph or t-SNE should group the cells. Three biological replicates of the cell culture were made, and 20 cells were tested per replicate. Bacteria were sampled at different stages of the cell growth. The Raman spectra of the stressed cells were collected after the treatment (5, 10, 20, 30 and 60 min, 3 h and 5 h).

5.7.5 Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

5.7.6 Author contributions

CGT wrote the paper with contributions from RP, BZ, FD and NB. BZ and FD cultivated, harvested and sorted the S. cerevisiae cells. MS cultivated and harvested *C. necator, M. extorquens, Y. lipolytica* and *K. phaffi*. CGT collected the Raman data. CGT performed the data analysis with the help of RP. CGT, RP, BZ, FD and NB designed the study. All authors read and approved the final version of the chapter.

6

Raman spectroscopy as a tool for estimating nutritionally valuable compounds in in microbial protein production

6.1 Abstract

Microbial protein (MP) is an alternative protein source with a potentially lower environmental footprint than conventional protein sources. It allows transforming side-streams into highly nutritional biomass, making it an interesting product in the context of the circular economy. When developing strategies for MP production, it is important to monitor its nutritional value by measuring the protein content and defining the amino acid profile, as well as the presence of other added-value compounds. Usually, these analyses are cumbersome, requiring harsh conditions such as extreme pH and or temperature, while their detection can be compromised by the presence of oxygen (in the case of amino acids) or light (in carotenoid detection). Furthermore, these processes are time-consuming, and the outcome can vary depending on the method used. In this chapter, we propose the use of Raman spectroscopy as a fast alternative to estimate indispensable amino acids and other added-value compounds such as unsaturated fatty acids and carotenoids, as well as the amino acid content, that should be the lowest possible for a safe consumption. First, we calibrated our method identifying the Raman signal(s) that best corresponded with the amino acid content of 4 axenic cultures. Then, this method was applied to two MP production set-ups, one for the batch cultivation of enriched bacterial cultures and one for the batch cultivation of cocultures, to determine under which condition -carbon source or microorganism- the nutritional profile improved. Our results show that Raman spectroscopy could estimate most indispensable amino acids (histidine, leucine, lysine, methionine, tryptophan, phenylalanine, valine and cysteine) and protein content. Additionally, we calculated the presence of other biomolecules of interest in the cultures, such as carotenoids. We propose that Raman spectroscopy has potential as a tool for rapidly estimating nutritionally valuable compounds in in microbial protein production to find the optimal culture condition (*e.g.*, substrate, organisms, pH, temperature), and as an online monitoring tool.

6.2 Introduction

Microorganisms have been reported as part of the human diet since 2500 B.C., first as fermented foods, and later in the 16t century directly as a major source of protein (*Spirulina*) (Bhatia & Nangul, 2013). Microbial protein (MP) was first developed during World War I, when surplus brewer's yeast was used for food (Ugalde & Castrillo, 2002). Nowadays, these alternative sources of protein are being looked at from the perspective of circular economy: producing protein-rich foods using side-streams that would otherwise be wasted. Also, the capacity of microorganisms to synthesize other nutritional compounds -certain vitamins, unsaturated fatty acids or carotenoids- could be exploited to develop a more nutritionally complete product (Anupama & Ravindra, 2000; Matassa *et al.*, 2016). In this chapter, we focus on two strategies for producing MP: HOB cultures and enrichments in acetic acid and formic acid.

Hydrogen-oxidizing bacteria (HOB) are a promising source of MP. First, they can alternatively grow heterotrophically and autotrophically, what gives them some flexibility: they grow heterotrophically using NH_4^+-N and organic carbon, and autotrophically using H_2 as an electron donor and O_2 as an electron aceptor (Matassa *et al.*, 2015). This in return means that their growth is not limited by light, and that there is an opportunity for growing them in a more sustainable way, by using ambient CO_2 as the carbon source and producing H_2 by electrolysis (splitting water into hydrogen and oxygen) using a renewable source of energy such as solar or wind (Hu, 2020). Secondly, HOBs produce polyhydroxybutyrate, a product that can serve as prebiotic (Matassa *et al.*, 2015). Finally, HOBs are also interesting for their high protein content (60-70% of the cell dry weight), and the quality of their protein, that has an amino acid composition and assimilation similar to that of the animal protein casein (Volova & Barashkov, 2010).

MP can also be produced using formate and acetate as a carbon source. These organic acids can be produced by converting H_2 and CO_2 into formate and acetate through a physicochemical reaction or through gas fermentation, respectively (Wang *et al.*, 2018; Takors *et al.*, 2018). However, high concentrations of formate and acetate can be toxic for bacteria (Sillman *et al.*, 2019; Pinhal *et al.*, 2019; Chen *et al.*, 2016).

When choosing a substrate or organism for microbial protein production, one of the

things that should be considered is the total amount of protein produced, as well as the single amino acid content, especially of essential amino acids. To quantify the total protein content, cells are first lysed to extract the proteins. The method of choice to lyse the cells can have an impact on the results and needs to be chosen carefully (De Mey *et al.*, 2008). Then, the total protein content can be measured by UV absorbance at 280 nm. The proteins absorb UV in proportion to their aromatic amino acid content (as they exhibit a strong UV-light absorption), making this technique less indicated for samples that contain a mixture of proteins (Noble, 2014). Other methods can be used for protein quantification. For example, indirectly by estimating the protein reduction of copper using the Lowry protein assay or Biuret reagents (Sapan *et al.*, 1999); using colorimetric assays such as the Bradford procedure, where specific dyes attach to the proteins and result into a color change after the binding (Noble, 2014); or with fluorescent dyes that attach to the proteins (Noble *et al.*, 2007). To quantify individual amino acids, the proteins first need to be hydrolysed. Then, they are separated using chromatography or electrophoresis and detected using absorbance, fluorescence or mass spectrometry (Kambhampati *et al.*, 2019). These methods are time-intensive, and need specific pipelines to quantify methionine, cysteine and tryptophan, as these amino acids are susceptible to acid hydrolysis (Rutherfurd *et al.*, 2007; Yust *et al.*, 2004).

Raman spectroscopy has been proposed as a fast alternative to detect metabolites in microorganisms. This optical tool detects the Raman scattering of the photons from the molecules present in the sample, generating a spectra with information on the lipid, carbohydrate, lipid and nucleic acid content present in the sample, amongst other molecules (Huang *et al.*, 2010; Tanniche *et al.*, 2020; Bunaciu *et al.*, 2015). Unlike the aforementioned methods for protein or amino acid quantification, this technique is non-destructive and can be used in fresh or fixed samples, meaning it needs little to no sample preparation and can be used for monitoring the MP production process online. It also allows to measure the lipid and carbohydrate content and detect other molecules of nutritional interest such as carotenoids.

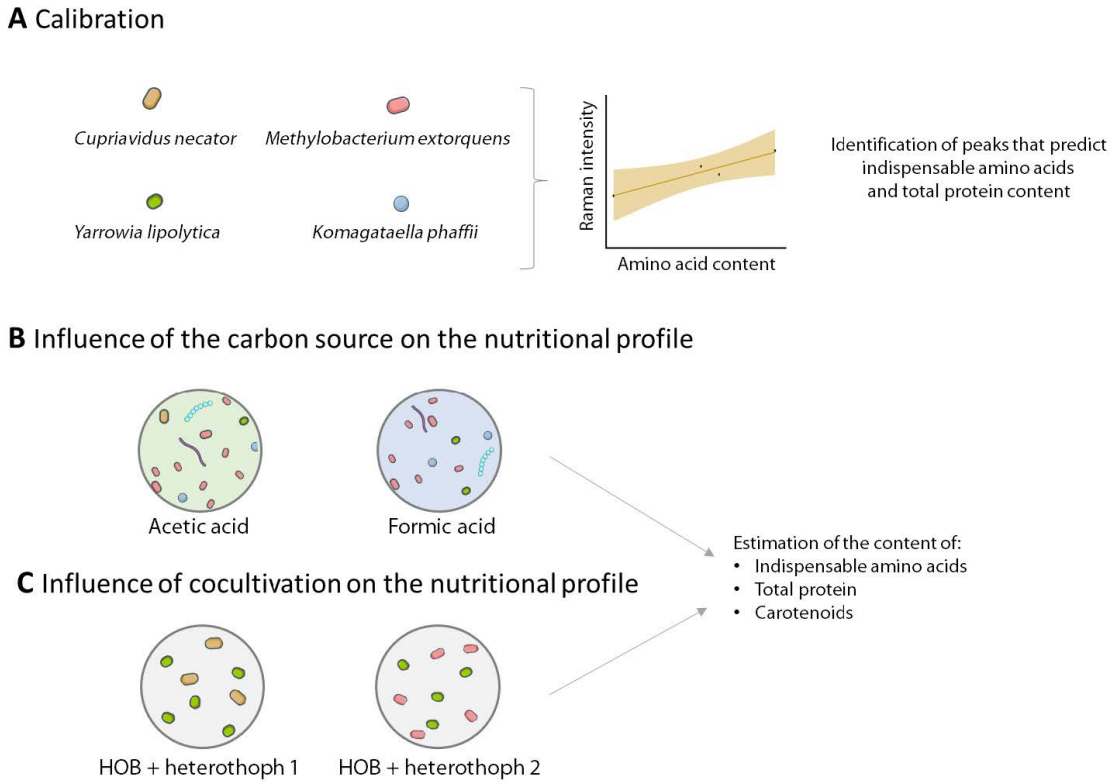


Figure 6.1: Experimental set-up. A) Calibration: four strains were analyzed with Raman spectroscopy and their amino acid content and total protein content was measured by a certified laboratory. Using a multi-point validation, we found the Raman regions that can be used for amino acid quantification. Using these spectral regions, we estimated the amino acid and total protein content in B) enrichments grown with either acetic acid or formic acid as a carbon source and C) cocultures of HOB and different heterotrophic bacteria.

In this chapter, we explore the use of Raman microscopy for quality control of microbial protein production (Fig 6.1). First, we calibrate the method using four axenic cultures and identifying the Raman signal(s) that best correlate with the expression of amino acids. We used the strains *Cupriavidus necator* LMG 1199, *Methylobacterium extorquens* DSM 1338, *Yarrowia lipolytica* ATCC 20362 and *Komagataella phaffii* ATCC 76273 because they are organisms used for microbial protein production with different morphology and characteristics. Then, we applied this method in two setups. The first setup produced microbial protein in enriched mixed cultures growing in formate or acetate. Using Raman microscopy, we compared the difference in the nutritional profile when using either carbon

source. The second setup consisted of cocultures of heterotrophs (*Chryseobacterium sp.*, *Microbacterium hominis* and *Sphingopyxis terrae*) and HOBs (*Xanthobacter agilis* and *Pinisolibacter sp.*) isolated from an enrichment. We tested their capacity to produce to produce microbial protein, and tested if cocultuing HOBs with heterotrophs would change their growth rate and/or nutritional profile, as it has been shown to be the case for methane-oxidizing bacteria (Veraart *et al.*, 2018).

6.3 Materials and methods

6.3.1 Cell culture

6.3.1.1 Calibration

We used the dataset from García-Timermans 2020, described in **chapter 5**. The cells listed in Table 6.1 were cultured in 2L Erlenmeyer flasks with a working volume of 1L, at 28℃ with 120 rpm orbital shaking. They were chosen because they are microorganisms with different shapes and characteristics that can be used for microbial protein production (Kunasundari *et al.*, 2013; Hardy *et al.*, 2018; Yan *et al.*, 2018; Cereghino & Cregg, 2000).

All cultures were aseptically inoculated in the corresponding rich liquid medium (Table 6.1), and re-cultivated every 24 to 48 h during 2 months to get sufficient biomass for the amino acid analysis (*i.e.*, 100 g of wet biomass). Briefly, 10% v/v of the cultures (100 mL) was used as inoculum for the subsequent cultivation, while the remaining culture (900 mL) was harvested via centrifugation at 6603 g for 5 min, washed with 0.1M phosphate buffer saline (PBS) and stored at -20℃ until sufficient amount of biomass was collected.

Table 6.1: List of organisms and medium used to grow them

Organism	**Liquid medium**	**Characteristics**
Cupriavidus necator LMG 1199	Nutrient broth (Oxoid Ltd, England)	Hydrogen-oxidizing bacterium
Methylobacterium extorquens DSM 1338	Nutrient Broth (Oxoid Ltd, England) with 1% methanol	Gram negative bacterium, methylotrophic
Yarrowia lipolytica ATCC 20362	YM Broth (catalogue number 271120, BD Biosciences, USA)	Fungi, can grow in hydrophobic environments
Komagataella phaffii ATCC 76273	Sabouraud Broth (catalogue number 238230, BD Biosciences, USA)	Methylotrophic yeast

6.3.1.2 Enrichment cultures

A culture used to produce microbial protein from potato processing side-streams (ValProMic, Avecom, Belgium) was used to obtain the enrichment cultures. For the pre-cultivation, this culture was grown in nutrient broth (Oxoid Ltd, England) for 24 h, in 1 L Erlenmeyer flasks with a working volume of 400 mL, at 28℃, under orbital shaking

(120 rpm). To obtain a dense culture, it was harvested through centrifugation at 6603 g for 5 min, and resuspended in fresh nutrient broth. After an additional 24 h of growth, this culture was divided in half and used for the enrichment on the substrate of interest. These cultures were centrifuged at 6603 g for 5 min, washed with 0.1M phosphate buffer saline (PBS) and resuspended in ammonium mineral medium with vitamins (details can be seen in Supplementary Table 6.4) with 40 mM of either sodium formate or sodium acetate and cultivated at 28℃ and 120 rpm shaking. During 2 months, 2-3 times per week 10% (v/v) of the cultures were inoculated into fresh medium. A total of 19 transfers were made.

These enrichments were used for the experiment described in the chapter. The cultures were under the same conditions except that the acetate and formate concentration was adapted to provide the same amount of chemical oxygen demand (COD), *i.e.*, 0.8 g COD /L (Supplementary Table 6.4). Samples were taken at the lag, log and stationary phase as estimated by OD_{600}. This was 0, 3 and 5 h in the case of acetic acid, and 0, 5 and 10 h in the case of formic acid. Samples were fixed using PFA 4% following the protocol described in García-Timermans *et al.* (2018). These fixed samples were then analysed using flow cytometry and Raman spectroscopy.

6.3.1.3 Cocultures

The strains used for the cocultures are described in Table 6.2. They can be found in the research collection of LM-UGent under the R number mentioned in Table 6.2.

These strains come originally from the HOB described in Hu (2020). After isolating the strains, Hu (2020) did 16S rRNA amplicon sequencing analysis, and we selected the most abundant OTUs for the experiment described in this chapter (*i.e.*, total abundance 97.29 .23%, relative abundance $\geq$ 0.19 ± 0.03%).

To determine whether these isolates were HOBs or not, Hu (2020) did the following experiment. First, she revitalized the colonies from the glycerol stock by incubation on solid R2A medium at 28 ℃ for three to four days. Single colonies were then transferred to R2A broth and incubated at 28 ℃ and 100 rpm for two days. The cultures were centrifuged at 5000 g for 3 min to get biomass pellets. To eliminate residual organics of the growth medium, the pellets were resuspended in phosphate-buffered saline and the supernatant

was removed after centrifugation at 5000 g for three minutes, which was repeated three times. with the nitrogen-free mineral medium. Two groups of 96-well plates filled with 180 L ammonium mineral medium were inoculated with 20 L washed cells in triplicates. One group was incubated in 2% O_2, 10% CO_2, 10% H_2 and 78% N_2 while the other one was incubated in air as a negative control without an external energy source. All the incubation was performed without shaking at 28 ℃ and the growth was checked by OD_{600} measurement after 7 days. The strains that grew in the presence of H_2 were classified as HOB.

Then, we chose the most abundant HOBs and heterotrophs. In the case of the HOBs, *Xanthobacter agilis* was the most abundant OTU (37.96 ± 2.90%) and *Pinisolibacter sp.* was from a less dominant but still abundant OTU (8.14 ± 0.41%). The three most abundant heterotrophs were *Chryseobacterium sp.* (2.41 ± 0.67%), *Microbacterium hominis* (1.08 ± 0.65%) and *Sphingopyxis terrae* (0.19 ± 0.03%).

First, the isolates were grown in a plate of R2A and gellan gum (10g/L, Thermo Fisher Scientific, Belgium) for 4 days at 28℃. Then, single colonies of each strain were transferred to liquid R2A medium and grown for 32 h at 28℃, 120 rpm shaking. Then, 10 mL of each culture were centrifuged at 6603 g for 5 min and washed 3 times with 4 mL of 0.22 µL filtered PBS. Finally, the pellets were resuspended in the medium described in Supplementary Table 6.5 to a final OD_{600} of 0.1 in a final volume of 30 mL. Cocultures were made mixing an HOB and heterotroph culture in equal parts (1.5 mL of each), in a 9 multi-well plate, 28℃, static, in a jar with continuous gas flow of composition 2% O_2, 10% H_2, 10% CO_2 and 78% N_2. Three replicates were made for the axenic cultures and cocultures. Samples were fixed as described in the ‘Enrichment’ section and used for flow cytometry and Raman microscopy analysis.

6.3.2 Calibration

6.3.2.1 Amino acid and protein quantification

Amino acid quantification was done by an external accredited laboratory (Eurofins Denmark A/S, Denmark). The protocol ISO 13903:2005 (EU 152/2009 (F)) was used to

Table 6.2: Strains used for the HOB-heterotroph cocultures

Hydrogen-oxidizing bacteria		
Strain	**Code**	**R number**
Xanthobacter sp.	**HOB1**	R-75750
Pinisolibacter sp.	**HOB2**	R-75754
Heterotrophs		
Strain	**Code**	
Sphingopyxis sp.	Het1	R-75763
Chryseobacterium sp.	Het2	R-75752
Microbacterium hominis	Het3	R-75761

measure cysteine and methionine, EU 152/2009 for tryptophan and ISO 13903:2005 (EU 152/2009 (F)) for the rest. The results are shown in Supplementary Table 6.4.

6.3.2.2 Raman dataset

The samples were processed as specified in García-Timermans *et al.* (2020). Then, we measured $\sim$450 points in every sample using a WITec Alpha300R+ spectroscope using a 785 nm laser (Toptica) with a 100x 0.9 NA Nikon objective, 5 sec of acquisition time and 1 accumulation. As a control for the instrument performance, a silicon piece (IMEC, Belgium) was measured with a grating of 600 g/mm, with a 1 sec of acquisition time and 10 accumulations. The intensity of the peak around 520 cm^{-1} was monitored over time. Laser power was also monitored to detect possible variations.

6.3.3 16S rRNA amplicon sequencing

We took samples from the replicate 1 of the enrichment cultures in the lag, log and stationary phase, at timepoints 0 h, 8 h and 11 h for AA enrichments and 0 h, 23 h and 36 h for FA enrichments. The DNA of these samples was extracted, and the results are analyzed following the protocol described in De Paepe *et al.* (2017).

6.3.4 Flow cytometry

Samples were diluted in filtered PBS and stained with SYBR Green I 1% (Thermo Fisher) for 13 min at 37℃ (Van Nevel *et al.*, 2013). They were measured with the flow cytometer BD Accuri C6 (BD Biosciences, USA) equipped with a blue 488 nm laser and 4 fluorescence detectors, namely FL1: 533/30 nm, FL2: 585/40 nm, FL3: $>$ 670 nm long pass and FL4: 675/25 nm), of which the FL1 detector was targeted by SYBR Green I, and two scatter detectors (forward scatter, FSC and side scatter, SSC). The channels FSC-H, SSC-H, FL1-H, and FL3-H were used for data analysis. Cell numbers were estimated using the PhenoFlow package (Props *et al.*, 2016).

6.3.5 Raman microscopy

In the samples from the enrichment cultures, we measured 20 single cells per sample. In the samples from the cocultre experiment, we measured 30 single cells per sample. The acquisition time was 40 sec and 1 accumulation was used. More details can be found in the Raman metadata aid (Supplementary Table 6.7).

6.3.5.1 Preprocessing

The 400-1800 cm^{-1} region of the spectrum was selected for fingerprint. The packages 'MALDIquant' (v1.16.2) (Gibb & Strimmer, 2012) or 'HyperSpec' (Beleites & Sergo, 2012) were used for preprocessing. The baseline correction was done using the Sensitive Nonlinear Iterative Peak (SNIP) algorithm (ten iterations) and spectra were normalized using the Total Ion Current (TIC). Then, the spectra were normalized using the *calibrateIntensity()* function and aligned using *alignSpectra()*.The hyperspec object was smoothed using the *spc.loess()* function. We decided to use this function because it had a small effect when estimating the biomolecules content, but made the plots less noisy and easier to interpret. The function *hs_contrast()* from the MicroRaman package (Kerchkof *et al.*, 2017) was used to compare the spectra of the two enrichment cultures.

6.3.5.2 Biomolecules content

We correlated the Raman intensity of the whole spectra with the total protein and amino acid content. The total protein content was calculated as the summary of all amino acids (see Supplementary Table 6.7) multiplied by the factor 100/116. This calculation corrects the amino acid sum to the corresponding weight of polypeptides, as it considers the water added during hydrolysis to individual amino acids (Feng *et al.*, 2016). This total protein calculation follows the recommendation from the FAO & agriculture organization of the United Nations (2003).

We first correlated all the regions with the amino acid or total protein content using the *ggscatter()* function of the 'ggpubr' package (Villanueva *et al.*, 2016). Then, we kept the regions that were best correlated and (1) were described as being part of that amino acid (Zhu *et al.*, 2011) and (for the amino acids) (2) were not regions that could describe the other amino acids. Then, for every amino acid we checked if the correlation improved ($>R^2$) using more regions. The resulting regions are described in Table 6.3. The same process was repeated for the total protein content.

The regions used for the correlations of unsaturated fatty acids and nucleic acids were found in Zhu (2011). For the unsaturated fatty acids, we did not use the 1448 cm^{-1} but the 1658 cm^{-1} instead because of the closeness of the first to ther compounds that were being estimated. We noticed a shift between our instrument and this database: when measuring pure phenylalanine, the characteristic 1005 cm^{-1} region was at 1012 cm^{-1}. This shift was considered when choosing the spectral regions for the correlations. The reference for carotenoids was found in the literature (Jehlička *et al.*, 2014).

6.3.5.3 Statistical analysis

Normality was tested with the Shapiro test using the function *shapiro.test()* from the package 'stats' (R Core Team 3.6.2, 2019). Pairwise comparisons of the groups were calculated and plotted using the function *stat_compare_means()* from the package 'ggpubr', that uses the Kruskal-Wallis test.

6.3.5.4 Minimal sampling size

To estimate the minimum number of spectra that need to be taken from a single sample to obtain a reliable result, we calculated the average intensity after adding one spectrum at a time. We did 1000 permutations on this dataset, and plotted the average result and the standard deviation. Also, we plotted a 5% deviation from the average using a dotted line.

6.3.6 Data availability and reproducibility

The analysis pipeline and the raw data can be found in https://github.com/CMET-UGent/RamanMP

6.4 Results

6.4.1 Calibration

6.4.1.1 Amino acid and total protein content estimation

The Raman signal of each amino acid consists of a series of signals that correspond to different molecular bonds (Zhu *et al.*, 2011). Some of these signals can be weak and/or be confounded with other molecules. We identified the Raman region(s) that corresponded best with each amino acid by extensively measuring four axenic cultures of *C. necator, M. extorquens, K. phaffii* and *Y. lipolytica*, and making a calibration between the Raman signal and the amino acid content as quantified by a certified laboratory. The regions that correspond to the amino acids of interest were selected, and the points that could correspond to several amino acids were discarded. Lastly, using a forward selection process, we identified the number of regions needed to obtain the best correlation as estimated by R^2 (Tables 6.3 and 6.2). We focused on the amino acid requirements in adults as determined by the World Health Organisation (WHO, 2017): essential amino acids (histidine, isoleucine, leucine, lysine, methionine, threonine, tryptophan, phenylalanine,

Table 6.3: Raman signals that correlate best with amino acid identification

Amino acid	**Raman regions used for quantification (cm^{-1})**
Histidine	555, 659
Isoleucine	671
Leucine	-
Lysine	1188
Methionine	1174, 1177
Threonine	1553
Tryptophan	770
Phenylalanine	1022, 1026
Valine	1502
Cystein	518
Tyrosine	-
Total protein content	1015, 1026

valine and cystein), cysteine and tyrosine.

We found unique Raman regions for all the tested amino acids, except for leucine and tyrosine (Table 6.3). Although the 687 cm^{-1} region had a good correlation with leucine, it could have also corresponded to phenylalanine. Similarly, the 1363 cm^{-1} region correlated with both tyrosine and lysine content. The calibrations had an $R^2 \geq 0.93$, except for isoleucine and threonine (Fig. 6.2A). To visualize what part of the spectra these regions represent, and their variation within the samples, we plotted the average spectra for all the cultures, and showed the amino acids that had a significant correlation in blue (Fig. 6.2).

Then, we identified the Raman regions that best correlated with the total protein content and did a forward selection to understand how many variables we needed in order to have the most accurate correlation (Table 6.3). We found that the peaks that had the best correlations (*i.e.*, 1015 and 1026 cm^{-1}) are aromatic regions that have been previously proposed by Teng (2016) to estimate the total protein content. While they found the aromatic peak in 1002 cm^{-1} region, there is a shift between their database and ours, as we found this peak in 1015 cm^{-1}.

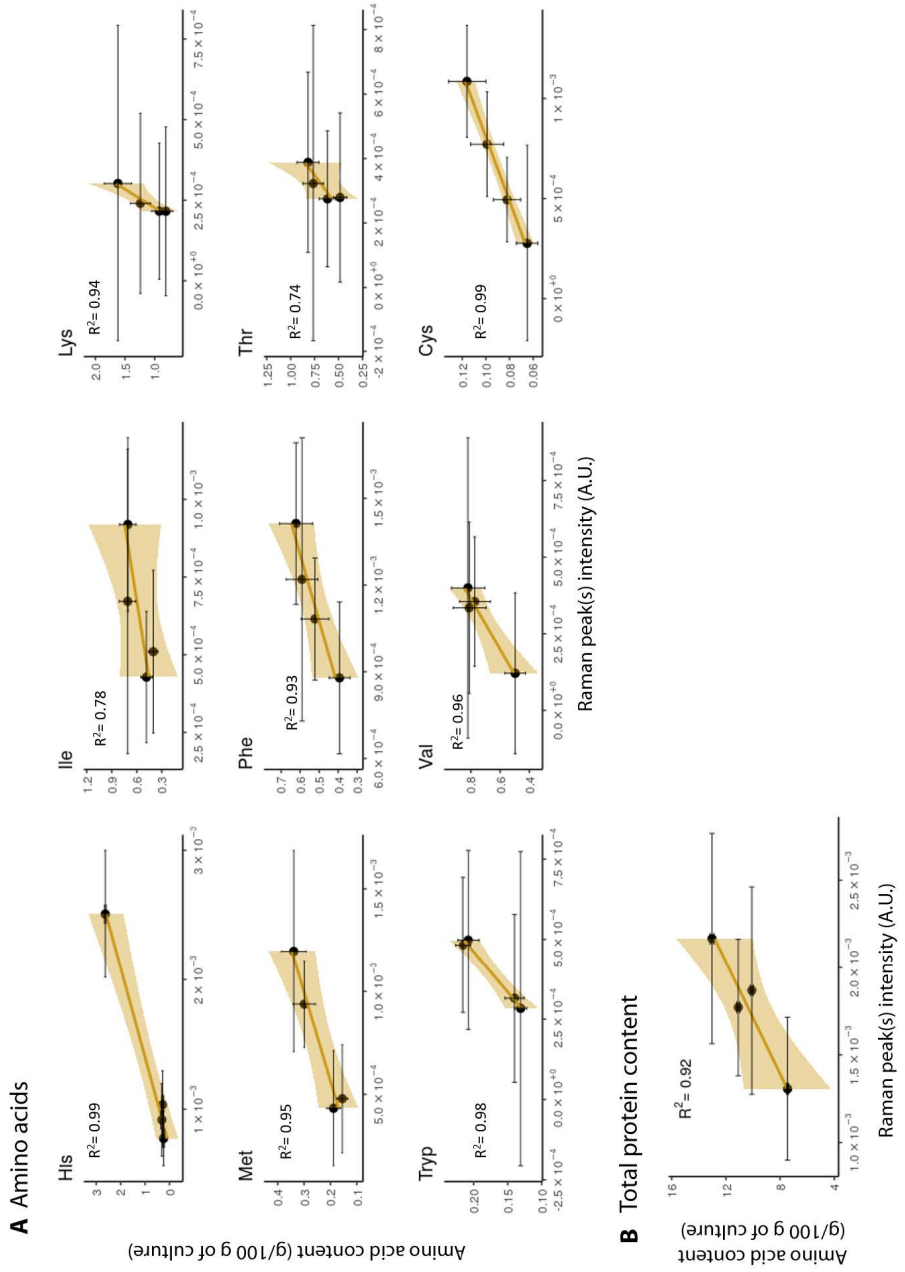


Figure 6.2: Spectral region selection: Correlation of amino acid content and the mean Raman intensity. The Raman regions used for these correlations are described in Table 6.3. In order, histidine, isoleucine, lysine, methionine, phenylalanine, threonine, tyrosine, valine and cysteine. Each point represents the mean value of the Raman intensity (∼450 measurements) and the mean amino acid content as calculated by a certified laboratory, with their standard deviation. The total protein content was calculated as the summary of all amino acids x 100/116, to correct for the water added during hydrolysis to individual amino acids. To calculate the standard deviation of the total protein content in the y axis, the squared root of the averaged variance of the standard deviation of individual amino acids was used. The scale in x and y axis is different for every compound.

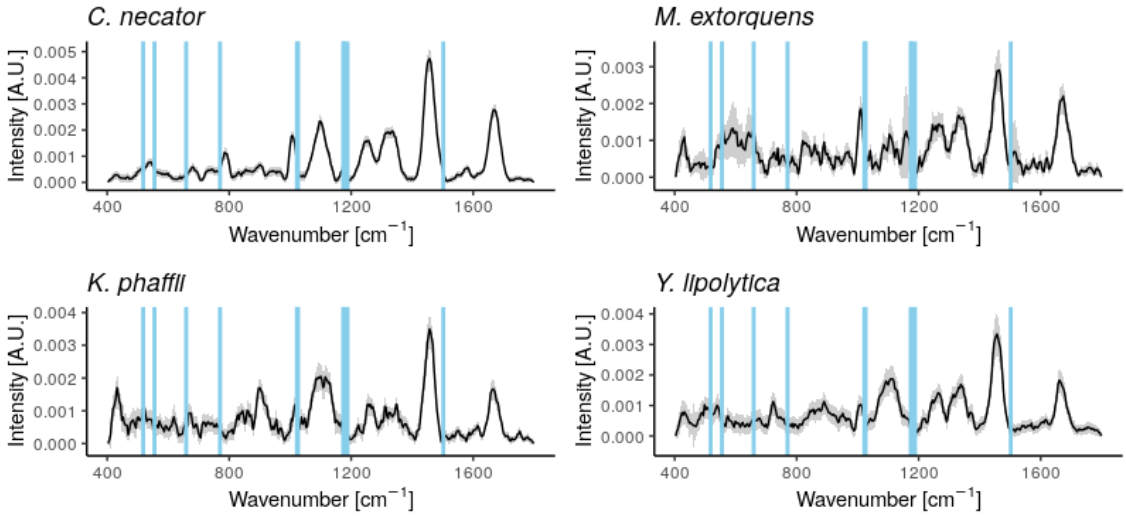


Figure 6.3: Average spectra for each culture. The standard deviation is represented as a grey smear. N=∼450. The blue regions correspond the amino acids histidine, leucine, lysine, methionine, tryptophan, phenylalanine, valine and cysteine.

6.4.1.2 Sample size

To estimate the minimum sampling size (*i.e.*, number of measured spectra) necessary to obtain stable mean Raman signal intensities, we calculated the average of the intensity obtained when adding one spectrum to the calculation. We used the large sampling dataset of four axenic cultures -*C. necator, M. extorquens, K. phaffii* and *Y. lipolytica*- where ∼450 points were measured per sample. For every organism and every amino acid, we calculated the average Raman intensity when using a single measurement (m), and when adding spectra to the calculation (m+1). We randomly selected the Raman measurements out of our dataset 1000 times. The average of the total dataset was plotted using a grey line. The 5% standard deviation was plotted using a dashed line. The results for the histidine region 555 cm^{-1} show that the number of measurements needed to reach the average is culture-dependent: while *C. necator* needs about 10 measurements, *M. extorquens* needs close to 100 (Fig. 6.4). The results for Raman regions from other amino acids (Supplementary Figure 6.11) confirm that the sampling size depends on the region that is being studied and the culture.

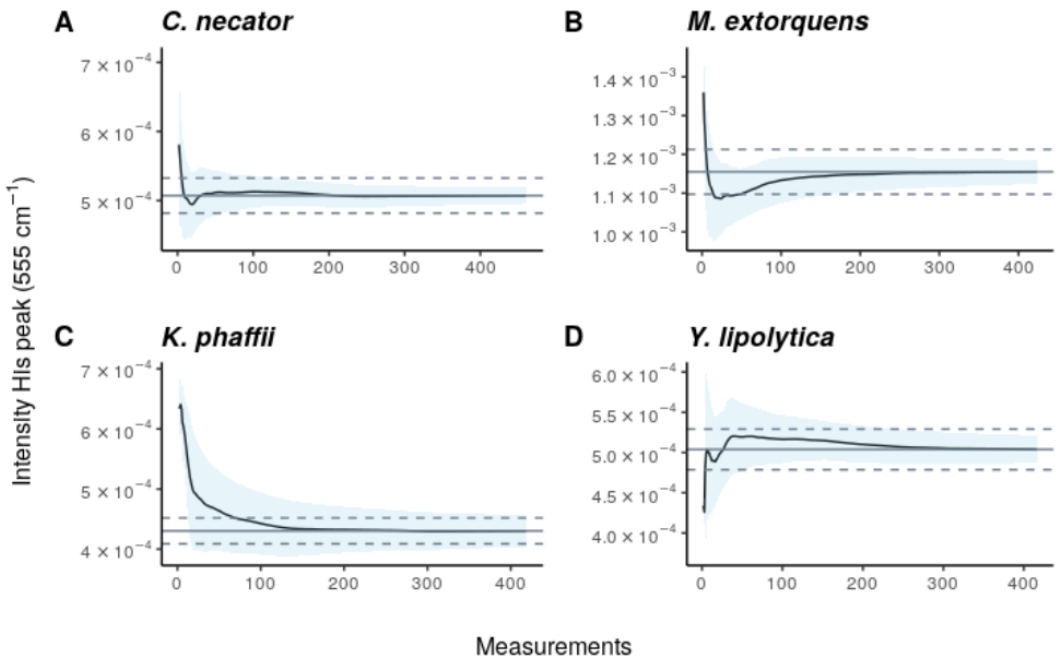


Figure 6.4: Minimum sampling size for histidine estimation in axenic cultures of *C. necator, M. extorquens, K. phaffii* and *Y. lipolytica.* The effect of adding more measurements to the average intensity was calculated. Raman measurements were selected randomly 1000 times. We represent the average of the measurements and the standard deviation in blue. In grey, the average of the total number of measurements and. The dashed line corresponds to a 5% error in the estimation. N=∼450.

6.4.2 Influence of the carbon source in the nutritional profile

We used Raman microscopy to compare the nutritional profile in two enrichment cultures. First, we inoculated a natural community in non-sterile mineral medium containing acetic acid (AA) or formic acid (FA) in triplicates. Communities were sampled at the lag, log and stationary phase (0, 8 and 13 h in the case of AA, and 0, 23 and 38 h in the case of FA) as calculated by OD_{600}. Then, fixed samples were measured using Raman spectroscopy, and 20 cells were measured per sample (a total of 60 cells were measured per timepoint).

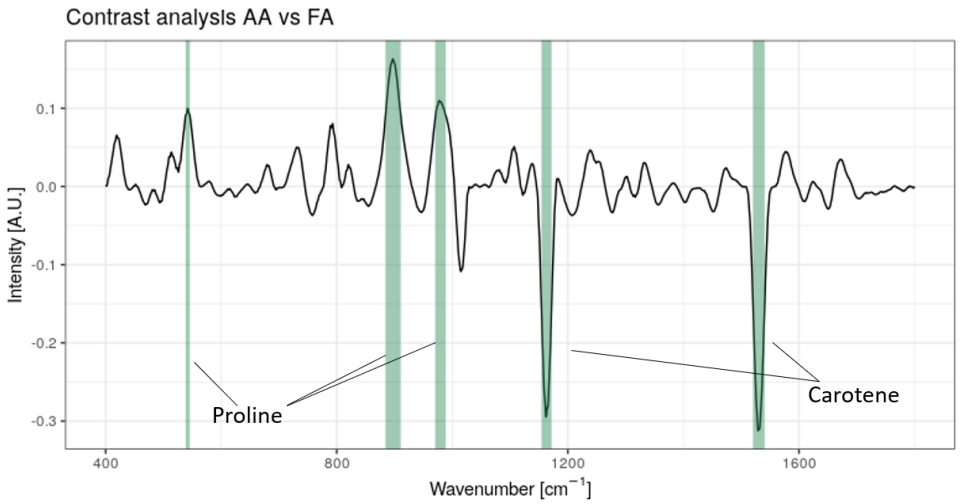


Figure 6.5: Contrast of the Raman spectra of the communities grown in acetic acid (AA) or formic acid (FA). The regions with positive intensity are more present in the AA enrichment, and those in negative to the FA community. The most prominent region are highlighted in green. Three replicates of the culture were made. In every culture, 20 cells were measured.

To understand the main metabolic components that make the communities when growing on these two carbon sources -AA and FA- different, we plotted the contrast of their Raman spectra (Fig. 6.5). We found that the AA enrichment cultures produce more proline (544, 895 and 902 cm^{-1}) while the community grown in FA has more carotenoids (1164 and 1531 cm^{-1}).

To follow the evolution of the metabolism of these communities over time, we estimated their lipid, protein and nucleic acid content in the lag, log and stationary phase. To estimate the amino acid and total protein content, we used the closest regions to those identified in the 'Calibration' section (Fig. 6.6). The other biomolecules were measured following the regions described in Teng *et al.* (2016) and Jehlička *et al.* (2014). First, we represented the average intensity of the Raman regions per cell to make estimations of the amount of biomolecules present in single cells. To convert the single cell estimation to an estimation of the amount of compound per mL of culture, we multiplied the single cell values with the number of cells in each growth stage (Fig. 6.7), measured with flow cytometry (Supplementary Figure 6.12). This way, the growth rates of the cells and how they affect production is taken into consideration. In general, the most notable differences

at the single-cell level are the highest content of carotenoids, phenylalanine and methionine in the FA-enriched cultures; and the highest content of histidine and cysteine in the AA cultures (Fig. 6.7A). When accounting for the cell density, the FA enrichment cultures, that grew more than the AA enrichment cultures (Supplementary Figure 6.13) have higher nutrient content than the AA enrichment (Fig. 6.7B).

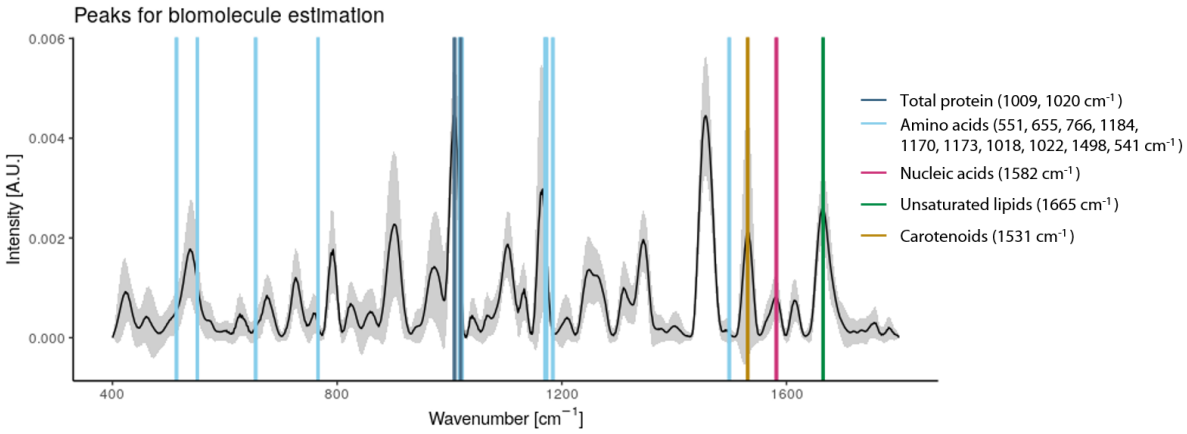


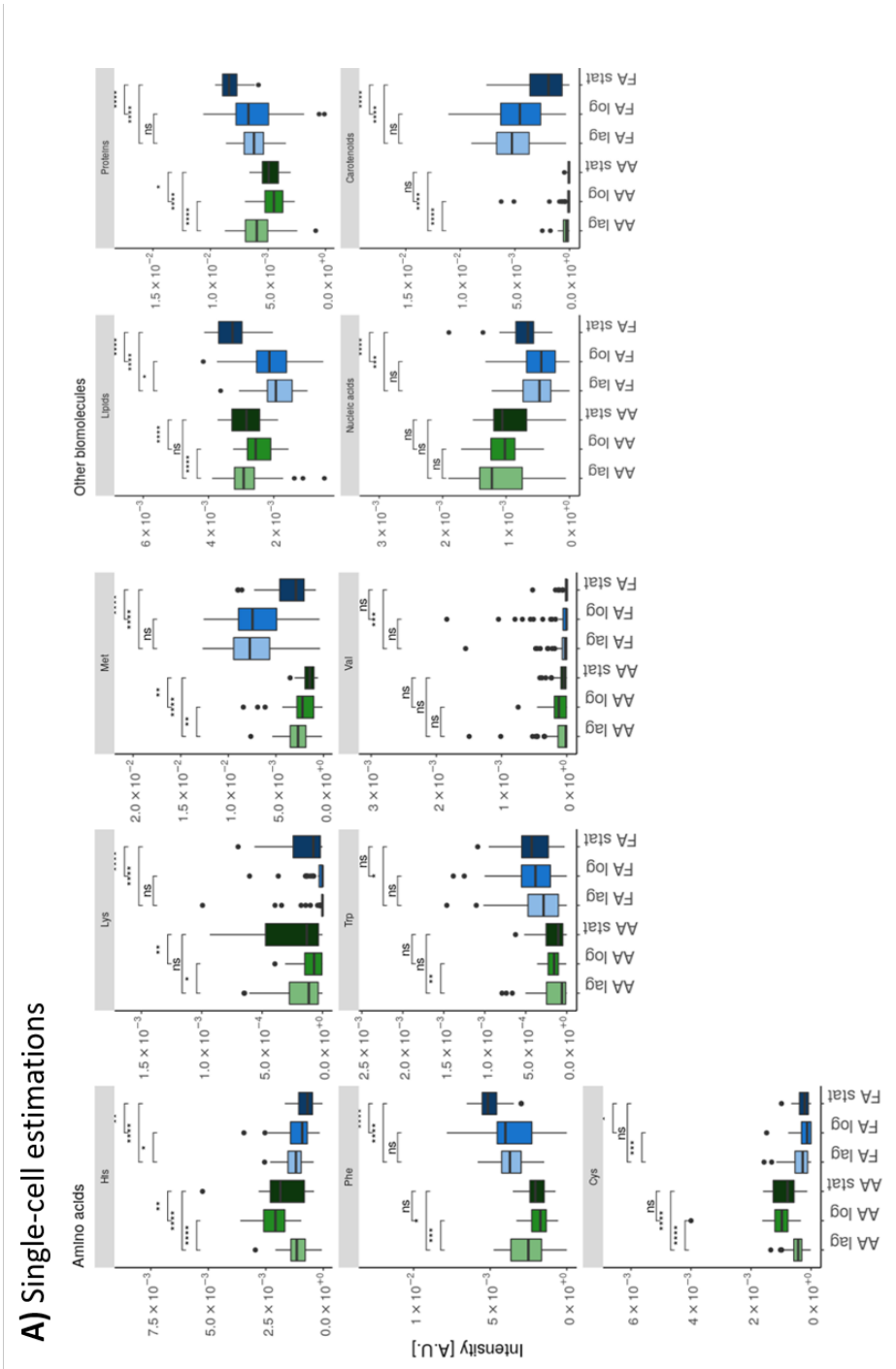
Figure 6.6: Spectral regions used for the biomolecules estimation. The average of all the cells is represented with a black line, and the standard deviation in grey. The Raman regions used to estimate the nutritional profile are shown in colors. The areas used to estimate the amino acids regions were selected in the 'Calibration' section. The regions for the other molecules are based on the literature.

6.4.3 Influence of cocultivation on the nutritional profile

We used Raman microscopy to compare the nutritional profiles of coculturing the HOBs *Xanthobacter agilis* (HOB1) or *Pinisolibacter sp.* (HOB2) with the heterotrophs *Sphingopyxis terrae* (het1), *Chryseobacterium sp.* (het2) and *Microbacterium hominis* (het3). We studied their Raman fingerprint and amino acid profile when cultured alone or cocultured. Three replicates of the cell culture were made for each condition.

Just like in the previous section, we estimated the nutritional profile of the axenic cultures and cocultures using the closest regions to those identified in the 'Calibration' section and those described in Teng *et al.* (2016) (Fig. 6.8). First, we estimated the average con-

tent of these molecules in single cells (Fig. 6.9A), to understand the metabolism of each culture. In the case of *Xantobacter agilis* (HOB1), its combination with *Microbacterium hominis* (het3) has the most interesting nutritional profile, with a higher amount of proteins and carotenoids. Also, it had a relatively high content of histidine, phenylalanine, tryptophan and cysteine. When *Pinisolibacter sp.* (HOB2) is cultured with *Chryseobacterium sp.* (het2), there is a high carotenoid content. The coculures with *Sphingopyxis terrae* (het1) and *Microbacterium hominis* (het3) have a similar protein content, but the amino acid profile of the coculture with het1 has higher histidine, valine and cysteine.


A) Single-cell estimations
Amino acids
Other biomolecules
His
Lys
Met
Val
Phe
Trp
Cys
Lipids
Nucleic acids
Proteins
Carotenoids
Intensity [A.U.]
AA lag
AA log
AA stat
FA lag
FA log
FA stat
ns

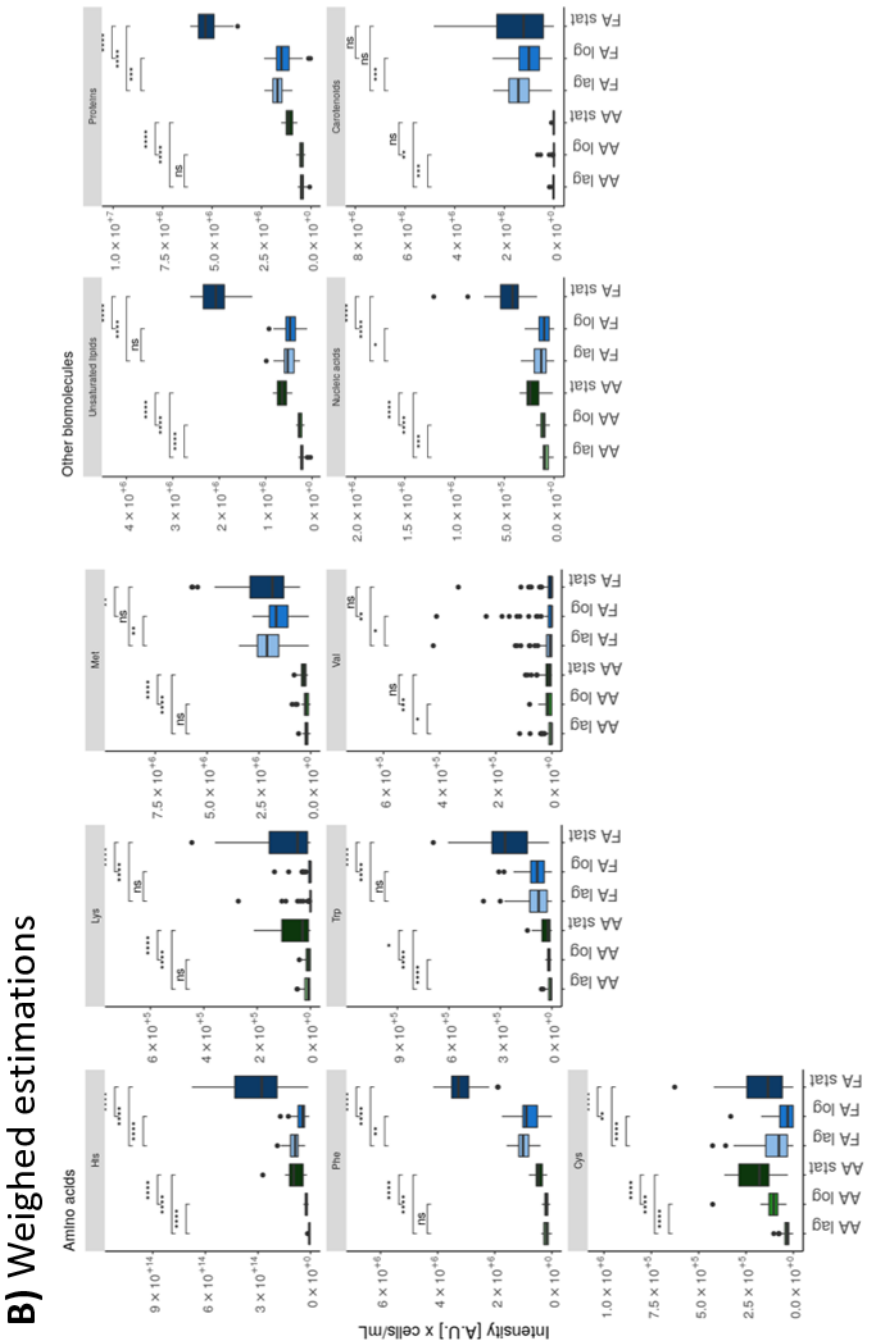


Figure 6.7: A) Single-cell and B) weighed estimations of the nutrional profile in the acetic acid (AA) and formic acid (FA) enrichment cultures in the lag, log and stationary phase. The amino acid regions regions were chosen based on the model predictions from the 'Calibration' section, summarized in Table 6.3. The codes correspond to histidine, lysine, methionine, phenylalanine, tryptophan, valine and cysteine. Three replicates of the culture were made. In every culture, 20 cells were measured. The statistical significance between the growth phases of each enrichment culture was calculated using Kruskal-Wallis (ns = $p>0.05$; * = $p\leq 0.05$; ** = $p\leq 0.01$; *** = $p\leq 0.001$; **** = $p\leq 0.0001$).

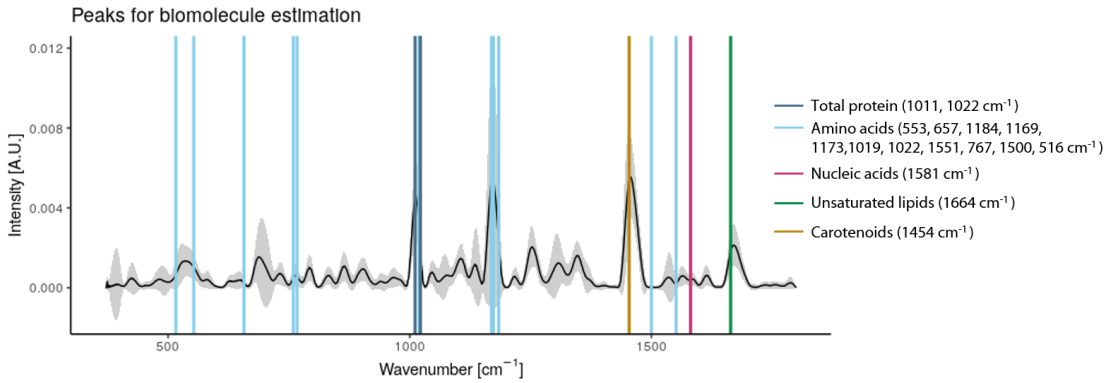
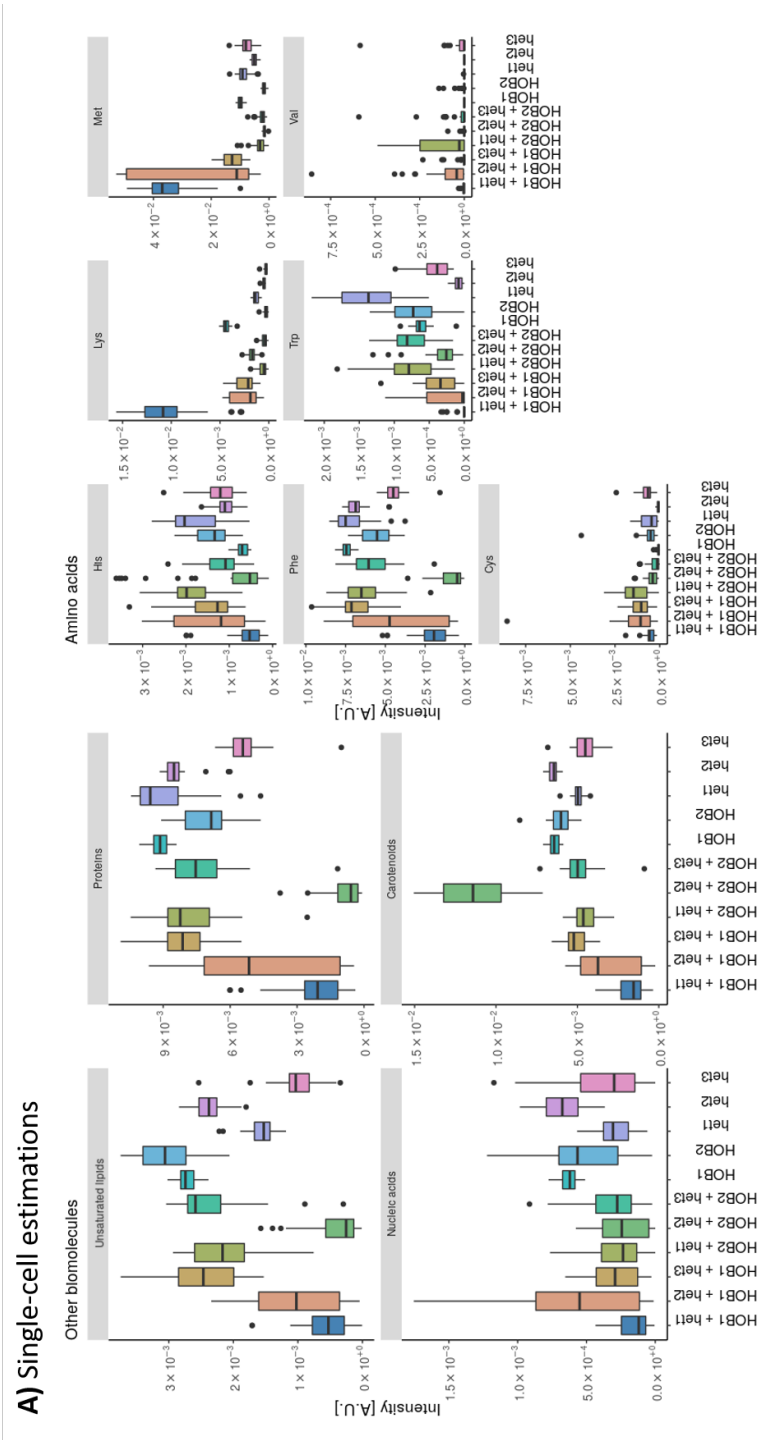


Figure 6.8: Spectral regions used for the biomolecules estimation. The average of all the cells is represented with a black line, and the standard deviation in grey. The Raman regions used to estimate the nutritional profile are shown in colors. The areas used to estimate the amino acids regions were selected in the 'Calibration' section. The regions for the other molecules are based on literature.

Then, we multiplied these by the number of cells present in each culture as estimated by flow cytometry (Supplementary Figure 6.14) to obtain weighed estimations of every culture (Fig. 6.9A). This correction allows taking into account the cell density reached by every batch. The fact that cocultures have a higher cell density than the axenic cultures is reflected in the weighed estimations. It is specially notable how the coculture of HOB2 and het3 has a much higher protein content than the others, and high histidine, phenylalanine and tryptophan content.


A) Single-cell estimations
Other biomolecules
Unsaturated lipids
Proteins
Nucleic acids
Carotenoids
Amino acids
Met
Val
Lys
Trp
His
Phe
Cys
Intensity [A.U.]
HOB1 + het1
HOB1 + het2
HOB1 + het3
HOB2 + het1
HOB2 + het2
HOB2 + het3
HOB1
HOB2
het1
het2
het3

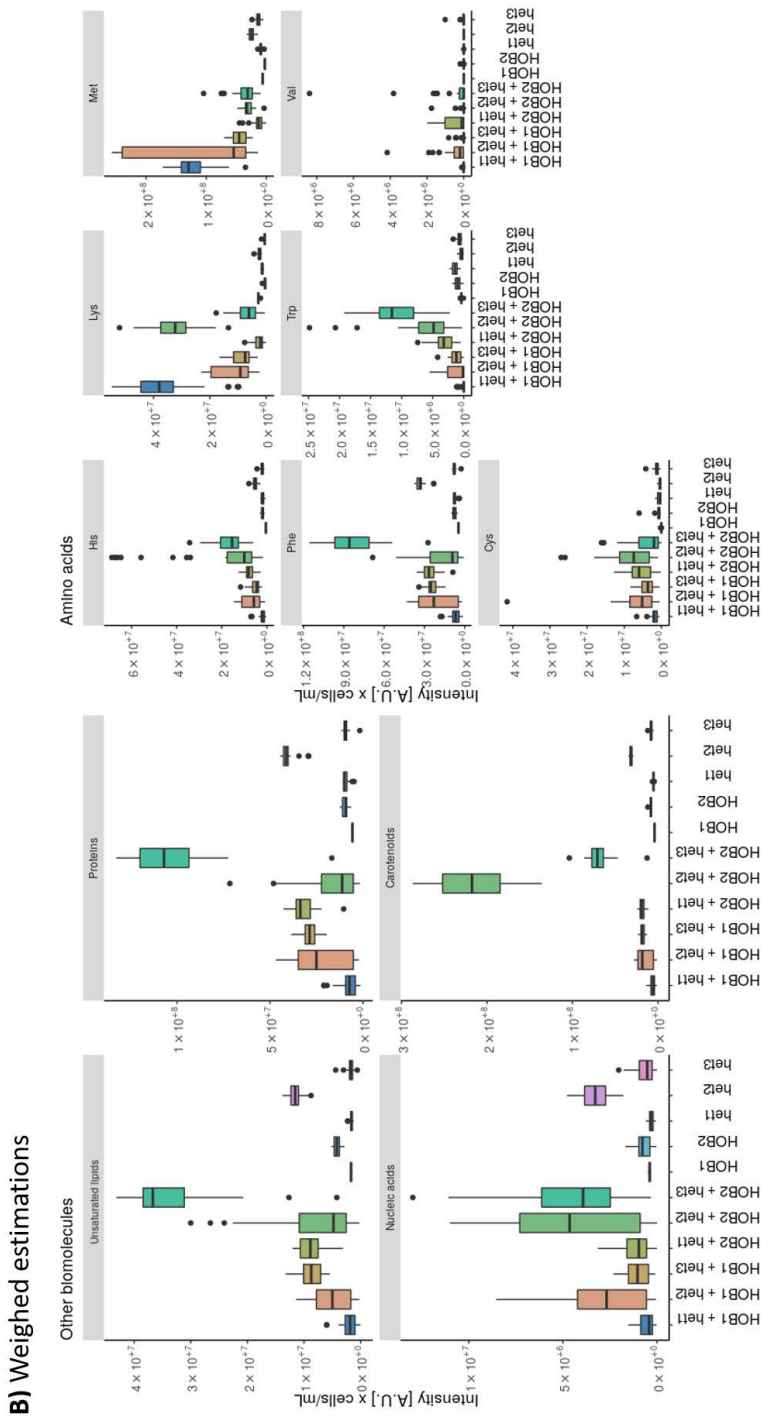


Figure 6.9: A) Single-cell estimation or B) weighed estimations of nutritional profile of axenic cultures and cocultures. The amino acid regions regions were chosen based on the model predictions from the 'Calibration' section, summarized in Table 5. The codes correspond to histidine, lysine, methionine, phenylalanine, tryptophan, valine and cysteine. The strains that correspond to each code can be found in Table 6.2. Triplicates of the cultures were made. In every culture, 30 cells were measured. The statistical significance between the cocultures of each HOB was calculated using Kruskal-Wallis (ns = $p>0.05$; * = $p\leq 0.05$; ** = $p\leq 0.01$; *** = $p\leq 0.001$; **** = $p\leq 0.0001$).

We observed if the fingerprints of the bacteria changed when grown alone or in a coculture. While the cocultures with het3, have a similar fingerprint to the HOBs' (Fig. 6.10A - 3,6), in the other combinations the fingerprint of the coculture differed to that of the axenic cultures (Fig. 6.10A - 1,2,4,5). The contrast analysis shows the metabolic changes that happen when the three heterothrophs are cultured with HOB1 or HOB2. When heterotroph 1 is cultured with HOB1, there is more tryptophan and tyrosine in the culture (regions 700 and 1173 cm^{-1}). When *Chryseobacterium sp.* (heterotroph 2) is cultured with HOB1, there is an increase in tryptophan and tyrosine, while coculturing it with HOB2 produces an increase of carotene (1448 cm^{-1}). Finally, cocultures of *Microbacterium hominis* (heterotroph 3) and HOB1 produce more tyrosine (Fig. 6.10B).

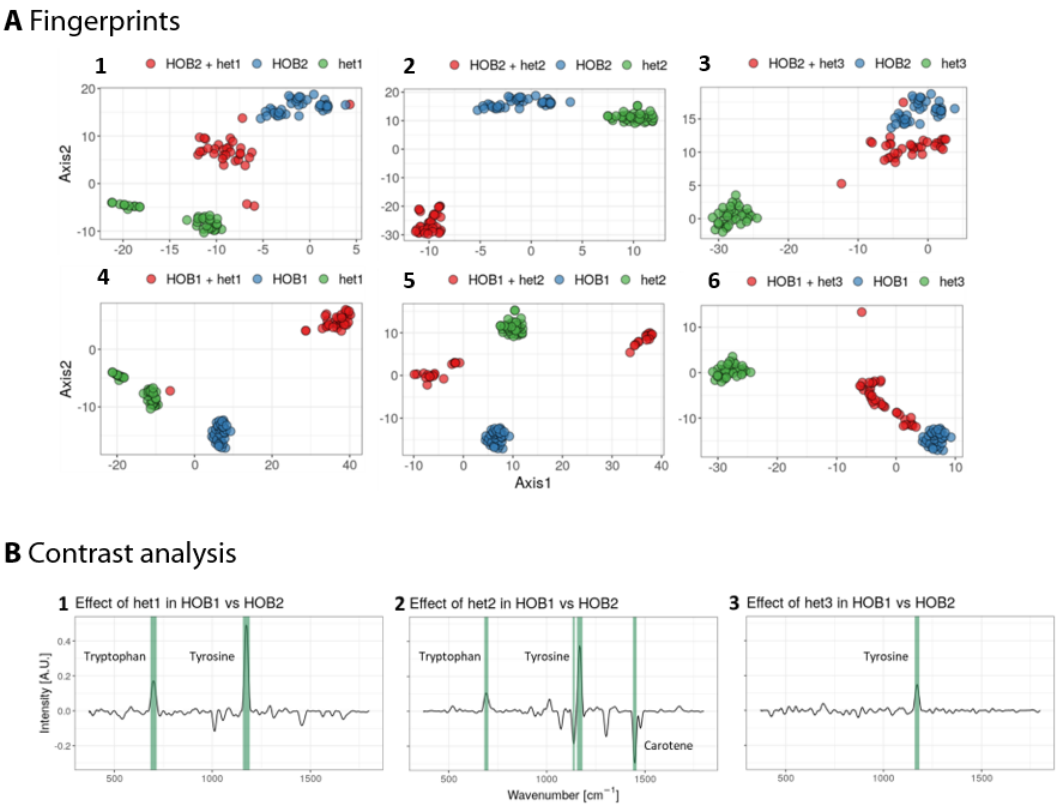


Figure 6.10: A) Representation of the fingerprint of axenic cultures and cocultures using t-SNE. A total of 30 single cells were measured per sample. B) Contrast analysis of the Raman spectra of cocultures of heterotrophs with the HOB strains. Most prominent regions are shown in green. Next to them, the molecules their tentative assignment based on the literature. Triplicates of the cultures were made. In every culture, 30 cells were measured.

6.5 Discussion

This chapter presents Raman microscopy as a tool for estimating of nutritionally valuable compounds in microbial protein production. We first used a large database of 4 strains and more than 450 cells measured per sample to benchmark the amino acid and total protein estimation. Then, we used our findings to study two set-ups: (1) enrichments with AA and FA as a carbon source and (2) cultures of hydrogen-oxidizing bacteria (HOBs) and heterotrophs, cultured alone or in combination.

In the 'Calibration' section, we show it is possible to identify characteristic regions of most of the amino acids with nutritional importance, histidine, leucine, lysine, methionine, tryptophan, phenylalanine, valine and cysteine (Fig. 6.2). We did not find unique Raman regions for leucine and tyrosine. Also, the calibration for threonine had an R^2 = 0.74, and isoleucine of 0.78. For these four amino acids, other estimation methods should be considered. Total protein content could be estimated with high accuracy ($R^2 \geq 0.93$) using two peaks that correspond to aromatic rings (1015 and 1026 cm^{-1}) (Fig. 6.2B) present in phenylalanine, tryptophan, tyrosine or histidine. The variability of the Raman signal is quite high (Fig. 6.2); however, this is not due to outliers but to single-cell heterogeneity, as shown in (Fig. 6.3). The number of cells that need to be measured to reach a stable mean amino acid signal was highly dependent on the amino acid and the strain sampled (Fig. 6.4). These different results can be explained by the heterogeneous expression of these amino acids that is found in *C. necator, M. extorquens, K. phaffii* and *Y. lipolytica*.

Our analysis focused on the amino acid requirements recommended by the World Health Organisation (WHO, 2017), which include the essential amino acids plus tyrosine and cysteine, and the total protein content. We also measured unsaturated fatty acids and carotenoids because of their nutritional relevance using peaks found in the literature. Nucleic acids were considered because the consumption of protein with high nucleic acid concentration (18–25/100 g protein dry weight) can increase the uric acid in the blood causing health disorders such as gout and kidney stone (Nasseri *et al.*, 2011). Removal of nucleic acid is necessary for the safe consumption of microbial protein .

We used the Raman regions we found in the 'Calibration' section to have significant correlation with the amino acid and total protein content to study two cases. First, to

compare microbial protein production in an AA and FA enrichment. We did a contrast analysis to identify the main metabolites that differentiate both communities (Fig. 6.5), as we expected that different communities would grow in each carbon source. We found that the AA enrichment had more proline (544, 895 and 902 cm^{-1}) while the community grown in FA has more carotenoids (1164 and 1531 cm^{-1}), probably metabolized by *Paracoccus* and *Pseudomonas*, the two most abundant genera in this enrichment (Supplementary Figure 6.15). When looking at the nutritional profile of the enrichment cultures over time, both conditions show a different single-cell profile (Fig. 6.7A); however, considering the FA enrichments in the stationary phase have the highest cell density, and express carotenoids, when looking at the weighed estimations they have the most interesting nutritional profile Fig. 6.7B). Calculating single-cell versus weighed estimations allows discerning how the single-cell metabolism changes when grown in different carbon sources, and how their growth rate affects the final metabolite content in the batch.

We used the same framework to study axenic cultures of HOBs and heterotrophic bacteria to determine if there was an improvement in their nutritional profile when grown together. An overview of the main metabolic differences when culturing the heterotrophs are with *Xanthobacter agilis* (HOB1) or *Pinisolibacter sp.* (HOB2) shows that when *Sphingopyxis terrae* (heterotroph 1) is cultured with HOB1, there is more tryptophan and tyrosine in the culture (regions 700 and 1173 cm^{-1} are more intense). When *Chryseobacterium sp.* (heterotroph 2) is cultured with *Xanthobacter agilis* (HOB1), there is an increase in tryptophan and tyrosine, while coculturing with *Pinisolibacter sp.* (HOB2) results in an increase of carotene (1448 cm^{-1}) (Fig. 6.9B). This was expected as *Chryseobacterium sp.* is known for producing carotenoids (Vila *et al.*, 2019). The single cell-estimations of the nutritional profile allows following the metabolic changes that occur between the axenic cultured and cocultres, and the effect of growing the HOBs with different heterotrophs in detail (Fig. 6.9A). Coculturing can change the metabolism of the cells growing together, as well as their growth rate. For instance, it could be that different strains are can cross-feeding or inhibiting each other. The weighed estimations shown in Figure 6.9B correct for the cell density to take this effect into account.

We confirmed the change of metabolic profile in these cultures when grown in axenic cultures or cocultures by plotting their Raman fingerprint using t-SNE (Fig. 6.10A). This is a known phenomenon previously described by Heyse *et al.* (2019). Notably, *Sphingopyxis*

terrae (heterotroph 1) presents two phenotypes when cultured alone. This could be due to cells being in a different cellular state (for example, growth stage) or having a distinct functionality. The combinations with *Microbacterium hominis* (heterotroph 3) show a Raman fingerprint that (mostly) resembles that of the axenic HOB culture (Fig. 6.10A – 3,6). It is possible that by chance we have only measured the HOBs in the mixture, and/or that growing in a coculture heterotroph 3 grows less than the HOBs. We measured a total of 30 cells per sample, which is standard when characterizing microbes with Raman spectroscopy (usually between 1 and 20 cells are measured). However, if more cells were to be measured the diversity of the sample would be better represented. Also, these are results are from single replicates. The addition of replicates would have certainly introduced more variability in the Raman fingerprint as shown previously (García-Timermans *et al.*, 2019; Teng *et al.*, 2016).

6.5.1 Methodological limitations

Raman spectroscopy presents limitations when identifying biomolecules. These include instrumental shifts and the complexity of the sample, the Raman intensity of the compounds and the choice of database.

There can be **shifts from one instrument to another** when measuring the same spectra. For instance, the 1009 cm^{-1} region from phenylalanine has been reported by De Gelder *et al.* (2007) in 1004 cm^{-1} and by Zhu *et al.* (2011) in 1005 cm^{-1}. It is important to take this into account in the experimental setup, analysing a reference spectrum and aligning the spectra if necessary, in the data processing. A study by Sjöberg *et al.* (2014) observed how free amino acids and those conforming proteins could show a shift in their Raman spectra. We took this into account when choosing the regions from Zhu *et al.* (2011), as well as the Raman shift of our instrument (see materials and methods). In our study, we also see shifts in the bands, but not those found by Sjoberg. For instance, they report that the phenylalanine 1009 cm^{-1} band was displaced to 1011 cm^{-1} when present in lysozyme. They argue that this is due to the 1015 cm^{-1} signal from tryptophan. In our case, the free phenylalanine band was present at 1012 cm^{-1}, and when measured in the microbial protein, displaced to 1022 cm^{-1}. This greater shift could be due to the presence of other amino acids, or other biomolecules. In the rest of the amino acids described by

Sjöberg *et al.* (2014), we do not find the same regions to be relevant for identification. As an example, they claim methionine is found at 1453 cm^{-1}, while in our study is best estimated using 1174 and 1177 cm^{-1}. **The complexity of our sample** and that of Sjöberg *et al.* (2014) and colleagues is not the same: while they use tripeptides and the proteins bovine serum albumin, lysozyme and b-lactoglobulin, we are measuring microbes that contain not only proteins, but also nucleic acids, lipids and carbohydrates. Meaning we expect to find different shifts and regions in these samples. Finally, it is worthy to note that Sjöberg *et al.* (2014) measured the tripeptides in H_2O, which could have sufficed to shift the Raman spectra as shown by Zhu *et al.* (2011). Zu *et al.* (2014) studied the correlation of amino acid content comparing the Raman spectra of *E. coli* with the results from ultra performance liquid chromatography. The regions that they found to be correlated are not the same that we have found in Table 6.3. In this study, we use 4 different organisms to select the Raman regions, and we compare the Raman spectra with the results of an accredited laboratory (the protocol ISO 13903:2005 (EU 152/2009 (F)) was used to measure cysteine and methionine, EU 152/2009 for tryptophan and ISO 13903:2005 (EU 152/2009 (F)) for the rest). This could explain the different results.

Some compounds that have a greater Raman intensity and are over represented in the spectra (for example, aromatic rings), while others do not show up. Therefore, although Raman spectroscopy is quantitative, this capacity can only be used to compare the same peak(s) amongst samples. **Many databases** with information on which regions are more relevant to retrieve amino acids or other biomolecules exist, and they differ from one another, making it difficult to choose from. In this manuscript, we chose the database of Zhu *et al.* (2011) for being extensive and consistent to those found in De Gelder *et al.* (2007).

For Raman spectroscopy to find a place in the biotechnology industry, standardization needs to play a pivotal role. Firstly, to minimize the impact of external factors that can affect the spectra, such as the instrument, laser power or other elements discussed in **chapter 3**. Secondly, in every setup there needs to be a study and validation -via a second established method, such as ultra-performance liquid chromatography- of the spectral regions used for the identification of biomolecules.

6.5.2 General overview

Raman spectroscopy can be used as a tool for a rapid estimation of nutritionally valuable compounds when exploring substrates or strains for microbial protein production. There had been previous efforts in this direction. For instance, Teng *et al.* (2016) had used the Raman band 1002 cm^{-1} to track the protein content in *E. coli*. Also, Schulmerich *et al.* (2012) showed how Raman spectroscopy can be used to quantify protein content in soybeans. Zu *et al.* (2014) showed the correlation of the amino acid content of *E. coli* as measured by ultra-performance liquid chromatography and with Raman spectroscopy, based on regions found on the literature.

Our work goes one step further to estimate the nutritional value of microbial protein using Raman spectroscopy, including not only total protein content, nucleic acids and unsaturated fatty acids, but also amino acids and carotenoids. First, it is important to determine what regions are relevant for the identification of the desired compounds, and how many cells should be measured to have a robust result. Here we show a calibration for the estimation protein and most nutritionally relevant amino acids (histidine, lysine, methionine, phenylalanine, tryptphan valine and cysteine). To study other amino acids, alternative methods would have to be used. In theory, it should be possible to tag amino acids or other molecules of interest with a probe that can be detected by Raman spectroscopy.

Raman spectroscopy requires little to no sample preparation, having potential to estimate the amount of nutritionally valuable compounds when different conditions are used for microbial protein production, as well as as an online monitoring tool. Changes in the desired nutritional value, or the presence of contaminants (*e.g.*, unwanted substances or foreign microorganisms) could be detected online, and be followed by a more detailed analysis using traditional tools.

6.5.3 Conclusions

- The bulk quantification of amino acids in microbial protein remains slow and time-consuming.
- Raman microscopy is presented as a single-cell alternative to quantify total protein and the content of the indispensable amino acids histidine, leucine, lysine, methionine, tryptophan, phenylalanine, valine and cysteine, with high accuracy ($R^2 \geq$ 0.93).
- Raman spectroscopy can also quantify unsaturated fatty acids, nucleic acids, vitamins and carotenoids, making it a powerful quality control tool.
- The analysis can be done at the single-cell level, useful to understand how the metabolism of the individual cells changes over time, either due to cocultivation or other factors. Batch estimations can be made by combining Raman spectroscopy with another cell counting technique, such as flow cytometry.
- This method requires little to no sample preparation, and can be used to monitor real time microbial protein production methods. It can also help researchers to choose which microbial community or growing conditions (substrate, pH, temperature) are most adequate.

6.6 Appendix

6.6.1 Acknowledgements

This chapter was written by **Cristina García-Timermans**, Myrsini Sakarika, Xiaona Hu, Korneel Rabaey, Ruben Props and Nico Boon.

The authors thank the funding that made this research possible. CGT is funded by the Flemish Fund for Scientific Research (FWO G020119N) and by the Geconcerteerde Onderzoeksacties (GOA) research grant from Ghent University (BOF15/GOA/006). MS is supported by the Catalisti cluster SBO project CO2PERATE ("All renewable CCU based on formic acid integrated in an industrial microgrid"), with the financial support of VLAIO, Belgium (Flemish Agency for Innovation and Entrepeneurship). XH is supported by the China Scholarship Council (201606260046) and Special Research Fund (BOF) of Ghent University (BOF01/SC4/518). RP is funded by the Flemish Fund for Scientific Research (FWO).

We would like to thank Yuting Guo for critically reading and reviewing the manuscript.

The authors gratefully acknowledge ValProMic for kindly providing the initial inoculum used in the experiments.

6.6.2 Conflicts of interest

The authors declare no competing interests.

6.6.3 Data availability

The raw data can be found in the repository https://github.com/CMET-UGent/RamanMP

6.6.4 Author contributions

Cristina García-Timermans, Myrsini Sakarika, Xiaona Hu, Korneel Rabaey and Nico Boon designed the experiment. MS grew the enrichment cultures and collected the samples with CGT. XH grew the mixed cultures and collected the samples with CGT. CGT acquired and analyzed the Raman spectra, and wrote this chapter. All the authors accepted the final version of this chapter.

6.6.5 Supplementary information

Table 6.4: Raman signals that correlate best with amino acid identification. EDTA = Ethylenediaminetetraacetic acid.

Compound	Unit	Value
Carbon source		
Sodium acetate	mM	12.5
Sodium formate	mM	50
Salts		
$MgSO_4 \times 7\ H_2O$	g L-1	1.0
NH_4Cl	g L-1	0.5
$CaCl_2 \times 2\ H_2O$	g L-1	0.15
FeNaEDTA*	mg L-1	0.05
Trace elements		
$Na_2EDTA^* \times 2\ H_2O$	mg L-1	0.5
$FeSO_4 \times 7\ H_2O$	mg L-1	0.2
H_3BO_3	mg L-1	0.03
$CoCl_2 \times 6\ H_2O$	mg L-1	0.02
$ZnSO_4 \times 7\ H_2O$	mg L-1	0.01
$MnCl_2 \times 4\ H_2O$	mg L-1	0.003
$Na_2MoO_4 \times 2\ H_2O$	mg L-1	0.003
$NiCl_2 \times 6\ H_2O$	mg L-1	0.002
$CuSO_4 \times 5\ H_2O$	mg L-1	2.5
Phosphate buffer		
$Na_2HPO_4 \times 12\ H_2O$	g L-1	0.717
KH_2PO_4	g L-1	0.272
Vitamins		
Riboflavin	mg L-1	0.005
Thiamine-HCl $\times 2\ H_2O$	mg L-1	0.025
Nicotinic acid	mg L-1	0.025
Pyridoxine-HCl	mg L-1	0.025
Ca-pantothenate	mg L-1	0.025
Biotin	µg L-1	0.05
Folic acid	µg L-1	0.1
B12	µg L-1	0.5

Table 6.5: Description of the solutions to make the medium used to culture the cocultures of hydrogenotrophs and heterotrophs.

Component	Concentration (/L)
Solution A	
KH_2PO_4	2.3 g
Na_2HPO_4 x 2 H_2O	2.9 g
Distilled water	50 mL
Solution B	
$MgSO_4$ x 7 H_2O	0.5 g
$CaCl_2$ x 2 H_2O	0.01 g
$MnCl_2$ x 4 H_2O	0.005 g
$NaVO_3$ x H_2O	0.005 g
Trace element solution SL-6	5 ml
Distilled water	915 mL
Solution C	
Na_2EDTA x 2 H_2O	0.06 g
$FeSO_4$ x 7 H_2O	0.05 g
Distilled water	20 mL
Solution D	
5% $NaHCO_3$	10 mL
Solution E	
Standard vitamin solution	5 mL

Table 6.6: Medium used to culture the cocultures of hydrogenotrophs and heterotrophs.

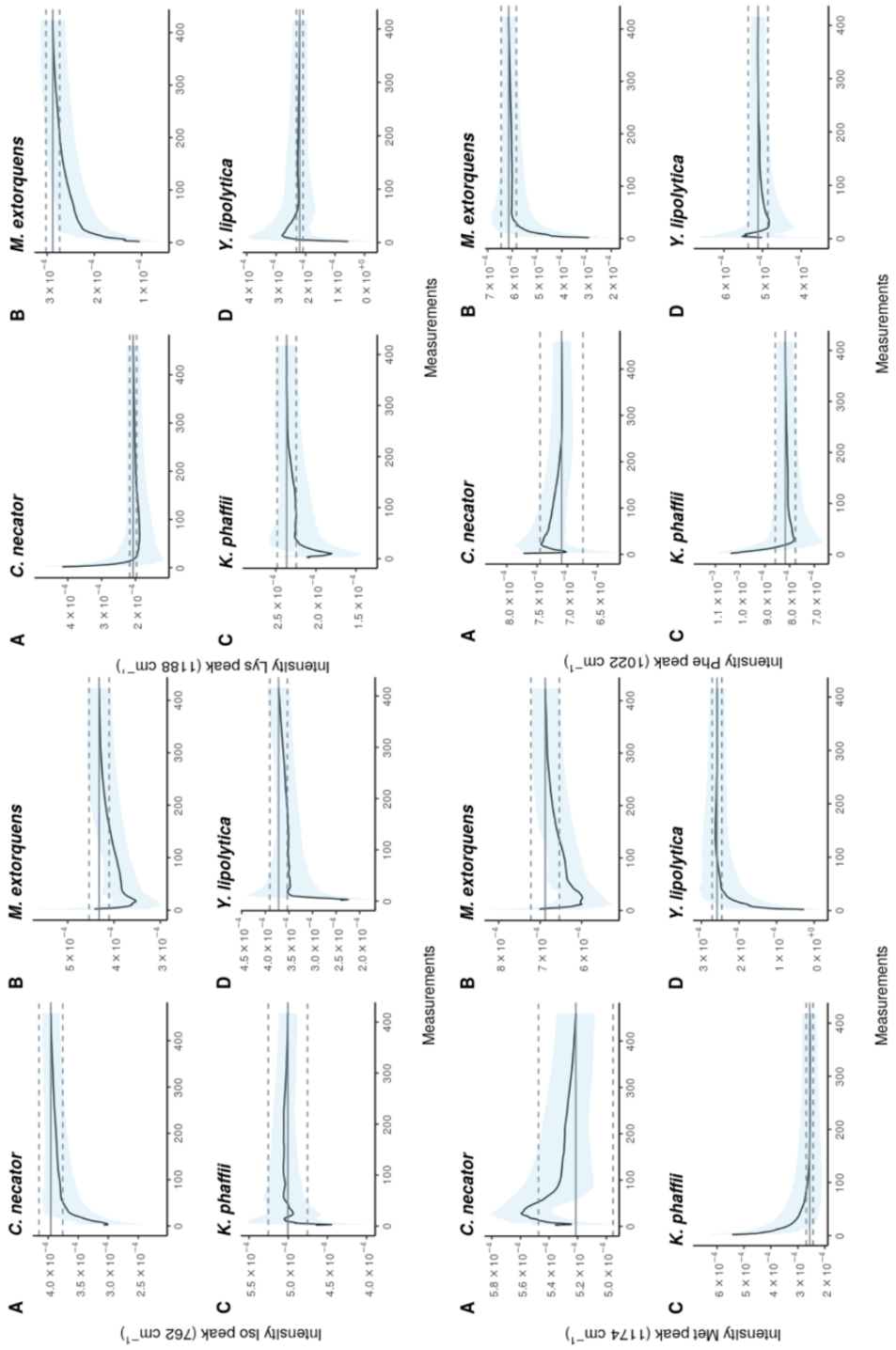
Solution	Final volume
Solution A	5 mL
Solution B	92 mL
Solution C	2 mL
Solution D	1 mL
Solution E	0.5 mL
Vatamin E	0.5 mL

Table 6.7: Amino acid results from Eurofins Denmark A/S, Denmark.

	Cupriavidus necator		***Methylobacterium extorquens***		***Yarrowia lipolytica***		***Komagataella phaffii***	
	value	**sd**	**value**	**sd**	**value**	**sd**	**value**	**sd**
Hydroxyproline	0	-	0,252	0,05	0	-	0	-
Ornithine	0	-	0	0	0	-	0	-
Threonine	0,616	0,086	0,764	0,107	0,484	0,068	0,819	0,115
Aspartic acid	1,12	0,16	1,67	0,23	0,866	0,121	1,33	0,19
Serine	0,445	0,062	0,623	0,087	0,456	0,064	0,77	0,108
Lysine	0,921	0,129	1,62	0,23	0,81	0,113	1,24	0,17
Valine	0,774	0,108	0,82	0,115	0,497	0,07	0,811	0,114
Proline	0,574	0,08	0,639	0,089	0,495	0,069	0,621	0,087
Alanine	1,2	0,17	1,26	0,18	0,65	0,091	0,789	0,11
Phenylalanine	0,522	0,073	0,59	0,083	0,393	0,055	0,622	0,087
Isoleucine	0,484	0,068	0,709	0,099	0,402	0,056	0,707	0,099
Glycine	0,686	0,096	1,2	0,17	0,508	0,071	0,596	0,083
Tyrosine	0,396	0,055	0,499	0,07	0,315	0,044	0,495	0,069
Arginine	0,842	0,118	0,597	0,084	0,525	0,074	0,69	0,097
Leucine	0,971	0,136	0,948	0,133	0,613	0,086	1,07	0,15
Histidine	0,245	0,034	0,263	0,037	0,288	0,04	0,312	0,044
Glutamic acid	1,31	0,18	2,13	0,3	0,942	0,132	1,48	0,21
Methionine	0,299	0,042	0,339	0,047	0,155	0,022	0,188	0,026
Cystein + Cystine	0,082	0,0115	0,065	0,0091	0,099	0,0139	0,116	0,016
Tryptophan	0,216	0,022	0,131	0,013	0,14	0,014	0,208	0,021

Table 6.8: Metadata aid for Raman spectra

Experiment overview	
Hypothesis	Test the capacities of Raman spectroscopy to detect amino acids and protein content
Variable(s) tested	There are two datasets. (1) The ”carbon source dataset” tests the influence of the carbon source on the spectra. (2) In the ”cocultres dataset” we test how coculturing changes the Raman spectra.
Conclusions	Raman spectroscopy can estimate total protein content and most animo acids that we tested.
Quality control (internal/external)	Silicon piece check
Samples and sample acquisition	
Material and source	1) Enrichment in acetic acid or formic acid 2) *Xanthobacter agilis, Pinisolibacter sp., Sphingopyxis terrae, Chryseobacterium sp., Microbacterium hominis*.
Growing conditions/sampling	See description in materials and methods
Filename format	<Replicate number>_<Treatment>_<Cell number>
Label in the samples	No label used
Fixation method	Filtered 4% formaldehyde solution from PFA
Integration time	40 sec
Accumulations	1
Grid	300 g/mm
Instrument	
Laser	785 nm excitation diode laser (Toptica). 175 mW of power before the objective.
Quality control	A silicon piece sample was measured with a grating of 600 g/mm, with a 1 second of acquisition time and 10 accumulations. Laser power was also monitored to detect possible variations.
Objective used (magnification)/ Numeric aperture (NA)	100x/0.9 NA (Nikon)
Camera	-70℃ cooled CCD camera (iDus 401 BR-DD, ANDOR)
Dry/water/oil objective	Dried samples
Model of spectroscope	WITec Alpha300R+
Other specifications (chromatic/ flat field correction/other)	In the samples from the enrichment cultures, we measured 20 single cells per sample. In the samples from the cocultures experiment, we measured 30 single cells per sample.
Data analysis	
Background subtraction method (if used)	No. Measurements with cosmic rays were deleted
Normalization method (peak /min-max/ area under-curve /other)	Area under the curve (‘Total Ion Current’)
Smoothing and interpolation (if done)	Smoothing, baseline correction, normalization and alignment (per group)
Statistics/Machine learning algorithm	Wilcoxon test for pairwise comparisons between two groups.
Accessibility	https://github.com/CMET-UGent/RamanMP
Other relevant information	-


A
C. necator
B
M. extorquens
C
K. phaffii
D
Y. lipolytica
Intensity Iso peak (762 cm⁻¹)
Intensity Lys peak (1188 cm⁻¹)
Intensity Met peak (1174 cm⁻¹)
Intensity Phe peak (1022 cm⁻¹)
Measurements

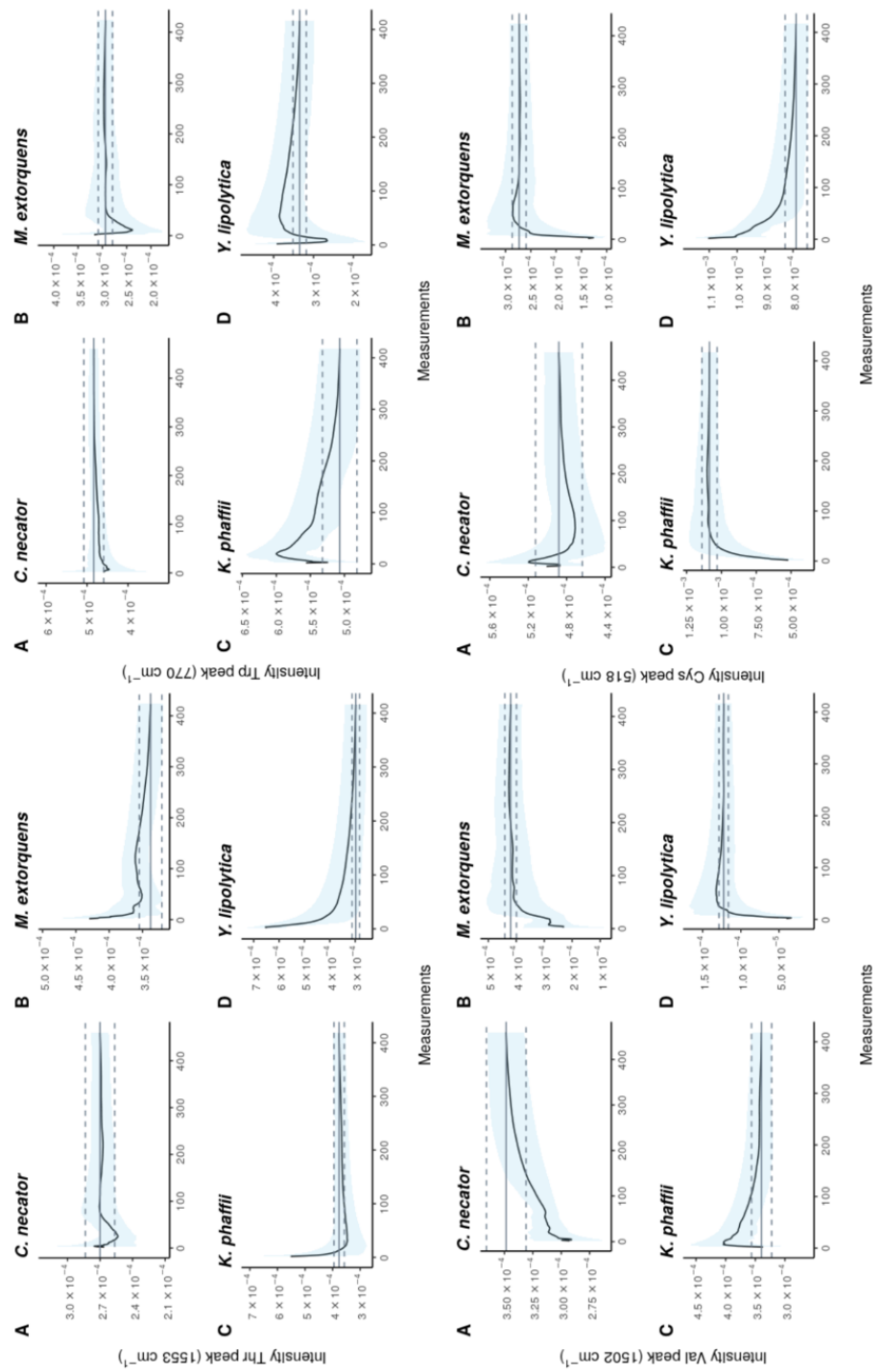


Figure 6.11: Minimum sampling size for histidine in axenic cultures of *C. necator, M. extorquens, K. phaffii* and *Y. lipolytica.* The effect of adding mor measurements to the average intensity was calculated. Raman measurements were selected randomly 1000 times. We represent the average of the measurements and the standard deviation in blue. N = $\sim$450 points. In grey, the average of the total number of measurements and. The dashed line corresponds to a 5% error in the estimation. From left to right and top dow, histidine, isoleucine, lysine; methionine, phenylalanine, threonine, tryptophan, valine and cystein.

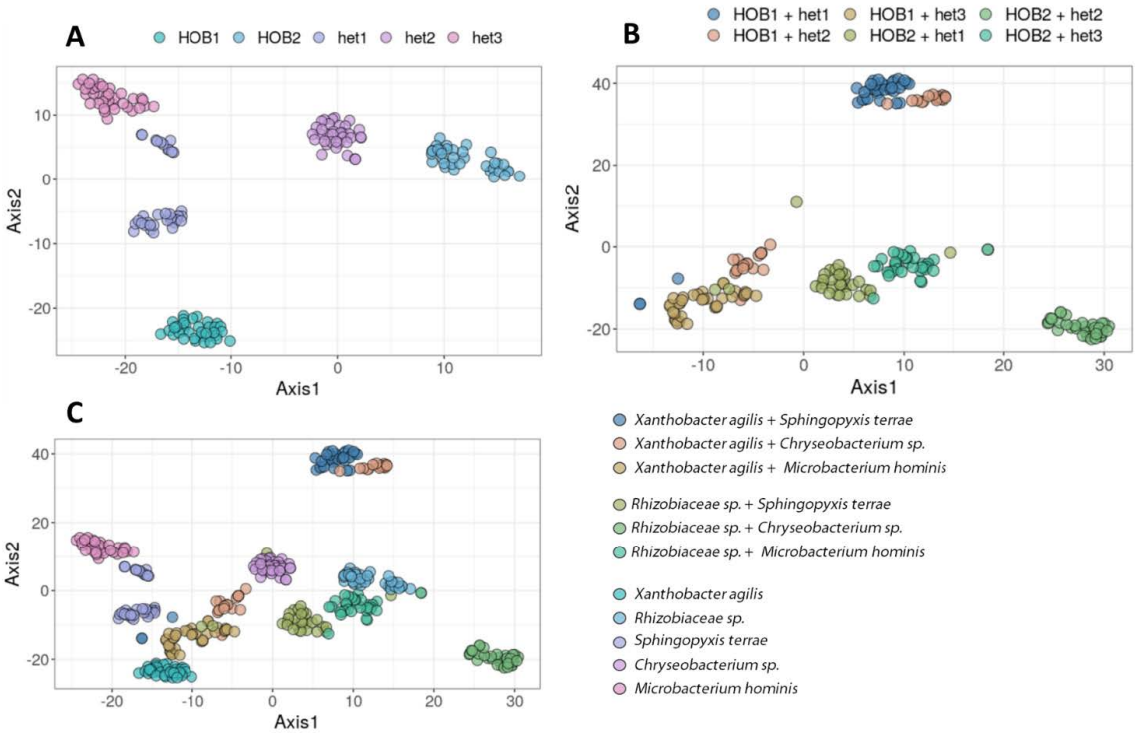


Figure 6.12: Representation of the Raman fingerprint of: A) axenic cultures, B) cocultures and C) all using t-SNE. The names of the strains can be found in Table 2. A total of 30 single cells were measured per sample.

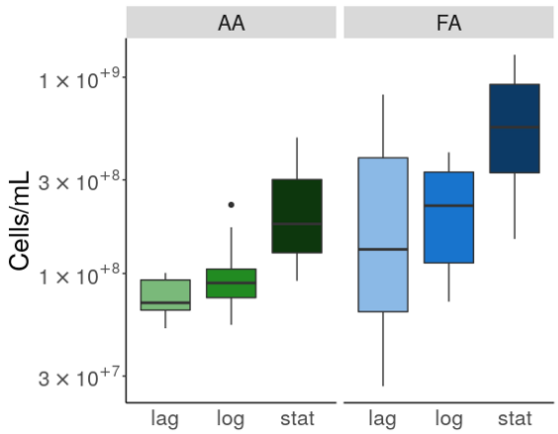


Figure 6.13: Cell concentration of the enrichment cultures grown in acetic acid (AA) and formic acid (FA), measured with flow cytometry. Triplicates of the cultures were made.

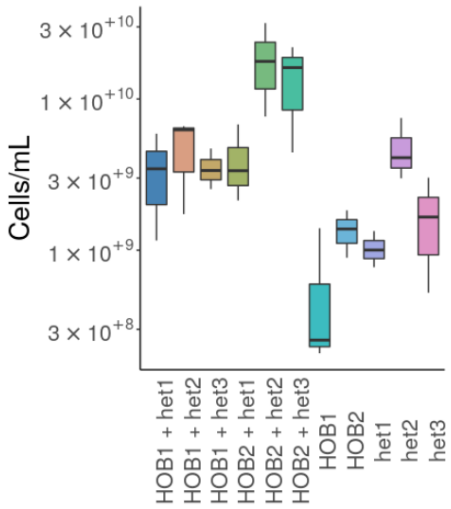


Figure 6.14: Cell concentration of the HOB and heterotrophs growing in axenic cultures or cocultures. The strains that correspond to the codes can be found in Table 6.2. Triplicates of the cultures were made.

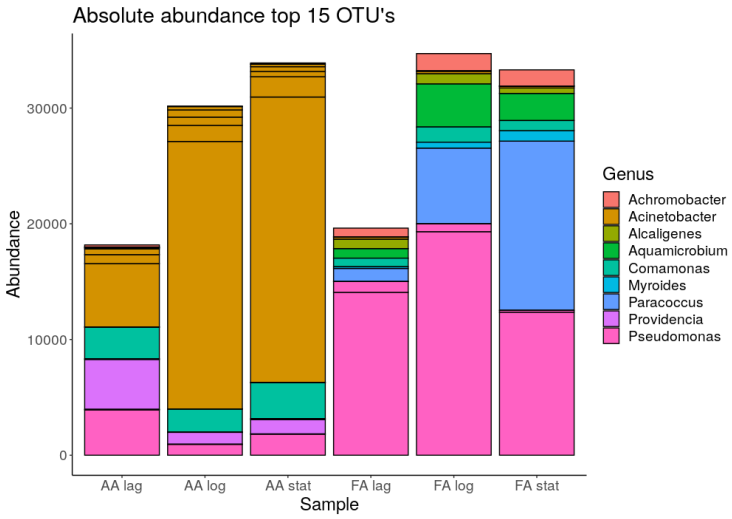


Figure 6.15: Absolute abundance of the top 15 OTUs present in the acetic acid (AA) and formic acid (FA) enrichment cultures in the lag, log and stationary phase, as estimated by 16S rRNA amplicon sequencing.

7

General discussion

7.1 Research outcomes

Studying microbial diversity is key to understanding community composition, structure, functionality and group dynamics in bacterial populations and communities. The methods most commonly used to measure microbial diversity (*i.e.*, sequencing technologies, such as 16S rRNA amplicon sequencing or whole-genome-sequencing) analyze all the cells from the sample together in bulk, and therefore give an averaged result of the composition of the population. Although informative, studying the average behaviour masks single-cell heterogeneity (Altschuler & Wu, 2010). Isogenic microbial populations present a varying degree of phenotypic heterogeneity, which allows them to distribute tasks, survive or increase fitness in a changing environment or organize the spatial structure of the population (Avery, 2006; Altschuler & Wu, 2010). Considering the importance of this phenomena in shaping microbial populations, phenotypic heterogeneity should not be

overlooked.

Single-cell technologies - such as flow cytometry, mass spectrometry, Raman spectroscopy or single cell OMICs- provide multiparametric information on individual cells, allowing to identify and study microbial phenotypes. This research explored the use of Raman spectroscopy to describe phenotypic heterogeneity and to delineate individual phenotypes in microbial populations. First, we standardized the reporting of metadata in Raman experiments (*i.e.*, experimental overview, sample(s) description, instrumental specifications and data analysis), and we proposed methodologies to identify phenotypes or quantify single-cell phenotypic heterogeneity in microbial populations. We compared the resolution of Raman spectroscopy with flow cytometry, another optical tool popular for bacterial phenotyping. Finally, we applied the methods developed to identify stressed phenotypes in bioreactors.

In this section, we list the research objectives developed in **chapter 2** and summarize their outcome.

7.1.1 Standardization of label-free Raman measurements for better reproducibility

Despite the increased use of Raman spectroscopy in microbial ecology, there was not a standardized way to report measurements. This amounted to decreased experimental reproducibility and made it difficult to share and cross compare data. In **chapter 3**, we tested the factors that influence Raman spectra in order to provide guidelines for accurate and reproducible single-cell Raman analysis. Experimental noise, intrinsic to Raman measurements, can greatly impact the outcome when very small spectral differences are driven by experimental factors, such as is often the case of microbial phenotypes. Specifically, it is known that the instrument (the type of laser, its power and the grating chosen) influences the spectra. Also, fixation with formaldehyde is the most recommended to best preserve the Raman spectra (Read & Whiteley, 2015). We proved how the sample handling, *i.e.*, the medium, storage time, extra centrifugation and resuspension steps or the drying time on the slide, can generate non-biological phenotypes.

Once the samples have been acquired, the preprocessing of the raw data needs to

be carefully considered, documented, and explained in the methodology, as discussed in **chapter 5**. The first step is the removal of cosmic rays that generate spikes in the Raman spectra. There are algorithms available to automate this task but we decided to opt for a manual elimination of the spectra that contained cosmic rays, as we considered that eliminating one or several Raman spectral wavenumbers would mean losing resolution. Then, the baseline needs to be corrected and the spectra normalized. The choice of methodology in each step will influence the outcome, and the reported intensity of the components. In this manuscript, we describe the use of certain algorithms for each of these steps, and they have provided sufficient resolution to answer to our hypothesis. However, there are many preprocessing algorithms for the different steps that we have not been tested, and that could provide a higher resolution when defining phenotypes. It is likely that depending on both the hypothesis tested and the post-processing analysis, different preprocessing steps are required.

Taking into account these factors, we developed a metadata checklist for Raman measurements. It contains questions about the experimental overview, the sample(s) description, instrumental specification and the data analysis (Fig. 7.1). We have found how often Raman spectroscopy experiments do not provide detailed information on these points, making experimental replication difficult. Also, sharing the raw spectra although encouraged by some journals is not a common trend in the field. Providing this aid as well as the raw spectra would increase data reproducibility and transparency in the field, and I hope that it may someday be included as a checklist for Raman public data repository.

7.1.2 Comparing the resolution of Raman microscopy and FCM to identify single-cell phenotypes

In **chapter 3** we explored the resolution of two optical tools for single-cell microbial fingerprinting: flow cytometry and Raman spectroscopy. Flow cytometry measures in the order of 1000 cells per second, gathering four useful parameters on a single cell (FL1, FL3, SSC and FSC). On the other hand, Raman spectroscopy gathers more information per cell, around 300 features, but it needs 30 sec to measure a single (unlabelled) cell (Fig. 7.2). We compared their performance when identifying the phenotypes of *E. coli* cells in the lag, log or stationary phase. While flow cytometry could detect population

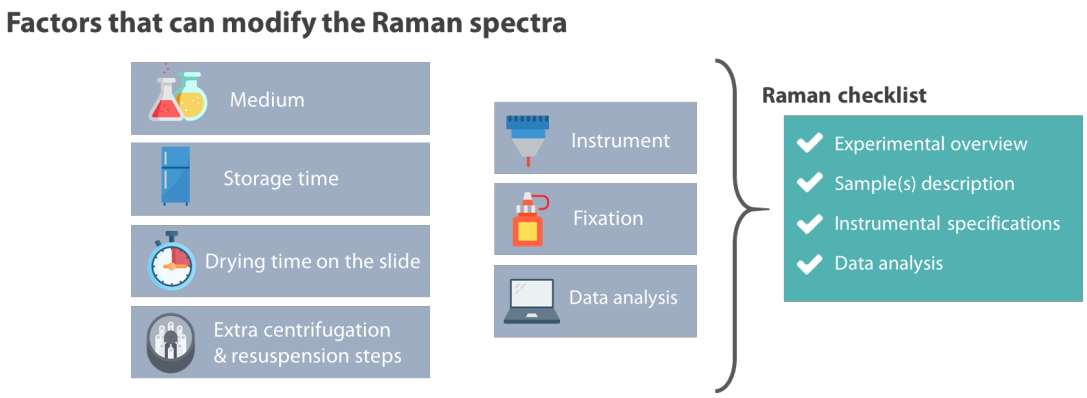


Figure 7.1: Development of Raman metadata aid. After exploring how changes in the sample manipulation (choice of medium, storage time, extra centrifugation and resuspension steps and the drying time on the slide) affect the Raman spectra, and considering known factors (instrumental variations, sample fixation and data analysis pipeline), we developed a Raman metadata aid that aims to improve reproducibility.

shifts, it was not able to clearly differentiate the phenotypes at the single cell level. Raman spectroscopy, on the other hand, could retrieve the growth stages and the replicates. More importantly, it was possible to automatically cluster cells based on their Raman spectra, using the adjusted Rand index (ARI) to optimize the t-SNE.

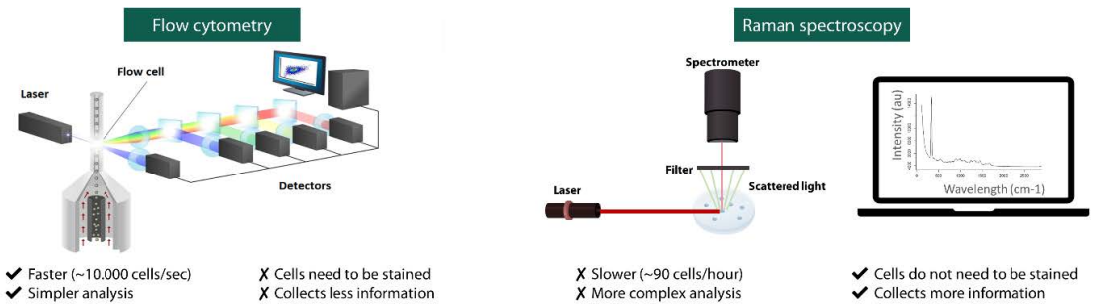


Figure 7.2: Summary of the main differences between flow cytometry and Raman spectroscopy. After sampling *E. coli* cells in a different growth stage, with a different phenotype, we determined that flow cytometry can detect population shifts, but as not enough resolution to detect changes at the single cell level. Raman spectroscopy on the other hand, was capable of distinguishing single-cell differences. We showed that using the tool PhenoGraph the phenotypes can be retrieved in a data-driven manner.

7.1.3 Automatic identification of single cell phenotypes based on their Raman spectra

The preprocessed Raman spectra of single cells can be used for identifying microbial phenotypes using dimensionality reduction and/or clustering methods (Table 7.1). **Dimensionality reduction** tools are useful to visualize phenotypes. The most commonly used ones are ordination tools, that reduce multi-dimensional spaces into two dimensions. These include principal component analysis (PCA), principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS). PCA fits the differences that exist in multiple dimensions to a line that maximizes the average squared distance from a point to a line. PCoA follows the same procedure, but instead of using the raw points it uses a (dis)similarity matrix. NMDS is also based on a (dis)similarity matrix but will look to non-parametric relationships between the points in an iterative way (Palmer, 2008; Zeleny, 2020). The resolution of these algorithms is sometimes not enough to discriminate spectra that are similar, for example, to retrieve phenotypes in a monoclonal population. For this purpose, in **chapter 4** we proposed the use of t-distributed stochastic neighbor embedding (t-SNE). This algorithm reduces the dimensions of each point to two or three-dimensional points in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability (Van Der Maaten & Hinton, 2008). Clusters can be automatically identified by fine-tuning one of the t-SNE parameters (called *k*), that defines how much local information should be included when looking for similar objects. Ultimately, the choice of dimensionality reduction or clustering method depends on the experimental question. For example, while phenotypic differences between stressed and control *S. cerevisiae* were noticeable with PCA (the clustering method that transforms the data the least amongst the ones proposed) (Fig. 7.3), differentiating *E. coli* growth stages needed of t-SNE (**chapter 4**).

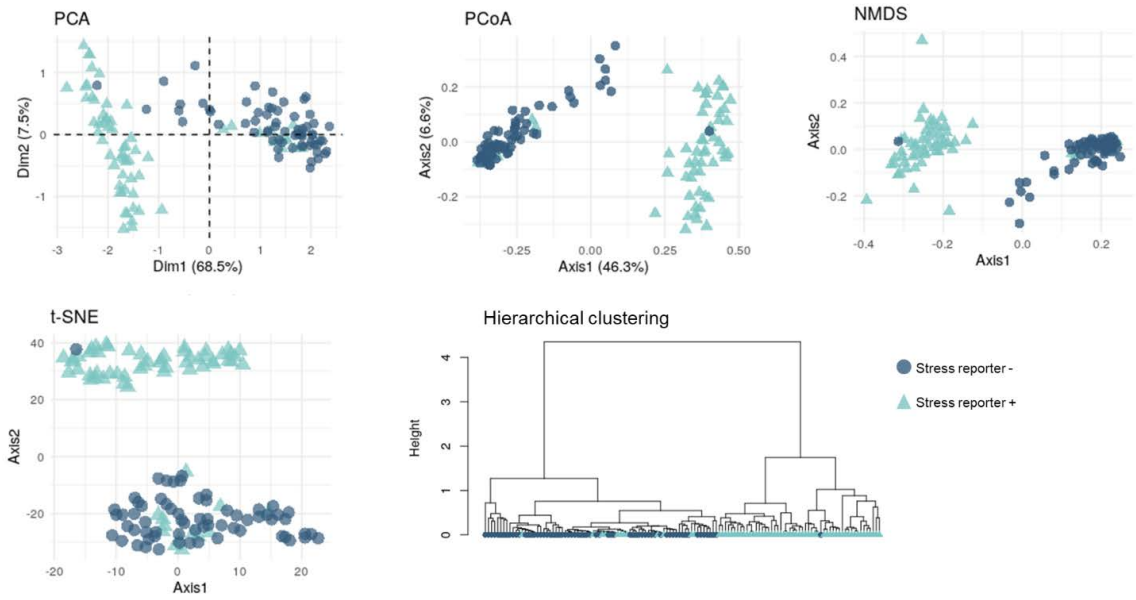


Figure 7.3: Dimensionality reduction and clustering methods for Raman spectra. *S. cerevisiae* cells with high or low expression stress reporter, classified using principal component analysis (PCA), principal coordinate analysis (PCoA) non-metric multidimensional scaling (NMDS), t-distributed stochastic neighbor embedding (t-SNE) or hierarchical clustering. N=65.

Besides these visualization tools, there are **clustering algorithms** that can help to identify to which phenotype single cells correspond to. A clustering method commonly used in Raman spectroscopy is hierarchical clustering, that aggregates spectra based on their pairwise similarity (Hedegaard *et al.*, 2011). Many agglomerative (bottom-up) algorithms can be used for this, Ward's being the most popular, because it minimizes the loss of information associated to each group (Ward, 1963). However, Kniggendorf *et al.* (2011) compared the performance of Ward algorithm and other clustering algorithms, and determined that Weighted-Average-Linkage is more reliable than Ward. Hierarchical clustering can instead use a divisive strategy (top-down), where all observations start in one cluster that is split recursively; however, this approach is less suited to generate a complete hierarchy, and therefore was not used in this manuscript (Manning *et al.*, 2008). Dissimilarity is often calculated using Bray-Curtis because -although it is not always a relevant measurement- its assumptions are widely accepted by ecologists (for example, the measure takes the value zero only when the two samples are identical and scaling

does not affect the relative values of a set) (Clarke *et al.*, 2006). In **chapter 4** we proposed the use of the spectral contrast angle for Raman spectra. This method is recommended to compare mass spectra fingerprints over other the similarity index methods (Wan *et al.*, 2002). Once that hierarchical clustering has defined to which phenotype single cells correspond to, a dendrogram can be generated. To determine where this dendrogram has to be ”cut”, the Adjusted Rand Index (ARI) can calculate the optimal number of clusters (as seen in **chapter 4**). Another clustering technique discussed in this work is partitioning around medoids (PAM), that searches for representative objects in a data set (medoids) and then assigns each object to the closest medoid in order to create clusters (Dodge, 1987). In **chapter 5** we use a reduced dataset with the principal components that explain the majority of the variance ($>$40%), calculate the optimal number of clusters using the silhouette index and then use partitioning around medoids (PAM) to identify which cluster single cells correspond to. How other ordination tools could be used with this clustering algorithm is yet to be explored. Finally, in **chapter 4** propose the use of the clustering algorithm Phenograph, that constructs a nearest-neighbor model and then divides events into communities (Levine *et al.*, 2015). First, the *hyperparameter k*, that defines how much local information is used, needs to be tuned. This hyperparameter can be optimized using ARI.

Table 7.1: Summary of the dimensionality reduction and clustering techniques used in this work.

	Technique	Input	Approach
Dimensionality reduction	Principal component analysis (PCA)	Spectral points	Linearly projects samples onto a new set of axes, such that the maximum variance is projected.
	Principal component analysis (PCoA)	(Dis)similarity matrix	Maximizes the linear correlation between the distances in the distance matrix, and the distances in a low-dimensional space.
	Non-metric multidimensional scaling (NMDS)	(Dis)similarity matrix	Finds non-parametric relationships between the dissimilarities and transforms these distances to a low-dimensional space using an iterative algorithm.
	t-distributed stochastic neighbour embedding (t-SNE)	Spectral points	Reduces the dimensions of each point to two or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.
Clustering techniques	Hierarchical clustering (bottom-up)	(Dis)similarity matrix	Aggregates points based on their similarity.
	Partitioning around medoids (PAM)	(Dis)similarity matrix	Searches for representative objects in a data set (medoids) and then assigns each object to the closest medoid in order to create clusters
	Phenograph	Spectral points	Constructs nearest-neighbor model and then divides events into communities.

7.1.4 Hill numbers to quantify single-cell diversity with Raman spectra

Diversity measurements inform about the richness and/or evenness of a microbial community. There are many metrics to calculate it, and in this work we adopt the use of the widely used Hill numbers, that are easy to interpret and represent the effective number of species. Hill numbers are typically calculated from compositional data, but have been extended to other data types as well, such as the flow cytometric fingerprint

of the population. In **chapter 5**, we extended this methodology to estimate a metric for single-cell diversity, which would refer to the metabolic diversity within each individual cell. For this, we used the single cell Raman spectra in the Hill number diversity framework. We considered every Raman signal as a component (a single or multiple molecules) and considered richness as the number of components being present in the cell, and relative contribution as their intensity (Fig. 7.4).

To calculate diversity, we chose to use the whole Raman spectrum. However, as mentioned in **chapter 5**, we did not explore how selecting only the peaks would affect the calculation: if it removed noise increasing the resolution, or on the contrary deleted relevant information. Also, the width is not taken into consideration in our calculation, when this is also a characteristic of the Raman signal of the chemical bonds. Although in this case the resolution allowed to differentiate the metabolically inactive and active populations, it would be interesting to have a larger database with different strains that are metabolically inactive and active at different degrees as measured by a second technique. Then it would be possible to explore the implications of selecting spectral peaks, and considering width or other features in the diversity calculations, and make the sc-D calculation more precise.

While other phenotypic diversity estimations (with 16S rRNA amplicon sequencing or flow cytometry data) give information about the species richness and/or abundance of a microbial community, here we present a measurement that reflects the richness and/or abundance of metabolites in single cells. With single-cell diversity estimations it is possible to find subpopulations with differential metabolic composition.

7.1.5 Applications of Raman microscopy to estimate nutritionally valuable compounds in bioproduction

Raman spectra contain (semi)quantitative information about the (bio)molecules present in the sample, as the intensity of the Raman bands is correlated to the number of molecules in each cell. For quantitative measurements, a calibration curve is needed (He *et al.*, 2017). In **chapter 5**, we used wavenumbers that have been previously associated with the lipid, protein or nucleic acid content in bacteria to determine the molecular composition of a stressed and a control *S. cerevisiae* subpopulation. We also proved that Raman

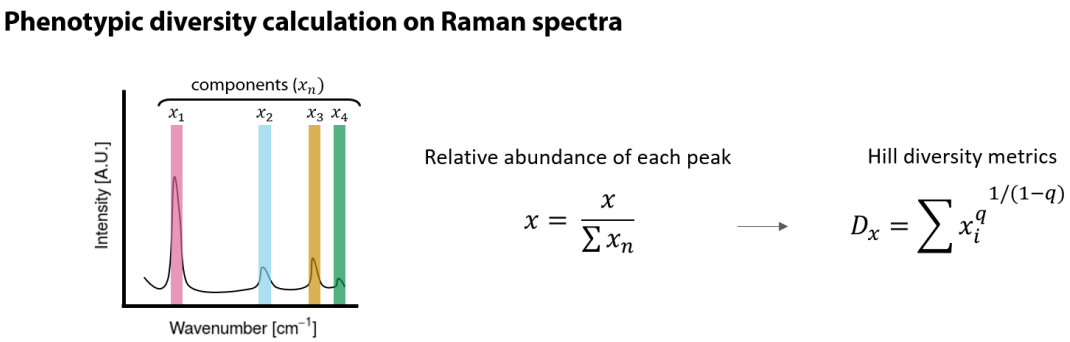


Figure 7.4: Phenotypic diversity calculation on Raman spectra. Raman peaks that correspond to one or several metabolites are considered as components. The intensity of these components (x) is used to quantify single-cell phenotypic diversity. The order of diversity (q) can be 0, 1 or 2, meaning respectively that richness, abundance or both parameters are considered in the metric. This equation considers richness and estimated abundance of metabolites in a single cell. Extracted from a figure in **chapter 5**.

spectroscopy can be used to quantify single-cell phenotypic diversity and to discriminate metabolically stressed cells using a clustering algorithm for two strains relevant for bioproduction, namely *E. coli* and *S. cerevisiae*. Stress management in bioproduction is important to maintain high production rates (Jia *et al.*, 2010).

In **capter 5**, we used Raman spectroscopy to do a multi-point calibration between the spectra of four bacterial cultures and the amino acid content as determined by an external lab. We found the peaks that best correlated with the amino acid content, and used them to estimate the nutritional profile semi-quantitively in two microbial production setups, to determine how the choice of carbon source or choice of organism(s) can influence the nutritional profile of the final product. We show how Raman spectroscopy can detect most indispensable amino acids, and the content of protein other molecules of interest, such as lipids, nucleic acids or carotenoids.

Estimating biomolecules with Raman spectroscopy has several limitations, as we explained later on this chapter (see section 'The relevance of microbial diversity'). For instance, the lack of a unified database that indicated what different peaks correspond to makes it difficult for users to compare results . Although there are peaks that are common throughout the databases, others are not. Also, there are molecules with a strong Raman

signals that can mask the weaker signals. Conversely, certain molecules can have a weak or no signal, and there are compounds that can have the same Raman spectra. Therefore, I call for a public repository for biological Raman spectra, similarly to those that exist for the Raman spectra of inorganic components -*e.g.*, FT Raman Reference Spectra of Inorganics (Geoffrey Dent, Avecia) or Raman Open Database (SOLSA)-, or for flow cytometry or sequecing data.

7.2 Raman based microbial diversity assessment

7.2.1 The relevance of microbial diversity

While there exists a consensus on the importance of biodiversity in plants and animals, and how loss in biodiversity impacts the function of an ecosystem (Willig, 2011), this is still a controverted statement in microbial ecology. Microbes have a short generation time compared to plants or animals, that should allow them to evolve and adapt more quickly. Also, it is argued that bacterial communities are diverse, and thus functionally redundant, although there is contradicting evidence about this (Roger *et al.*, 2016). Microorganisms are essential in natural biogeochemical processes (*e.g.*, carbon, nitrogen and phosphorous cycles), as well as in engineered systems (*e.g.*, food or pharmaceutical industry and wastewater treatment). It is worrisome that considering microbes inextinguishable will remove them from the biodiversity-conservation agenda, not giving them any consideration when discussing, preventing and alleviating anthropogenic disturbance (Bodelier, 2011). We discuss here some aspects that could improve diversity calculations, and other factors that should be considered alongside it to understand microbial communities.

Diversity calculations reflect the richness and/or abundance of species in a community. To take into account not only taxonomic differences, but also environmental differences, ecotypes can be considered. Ecotypes divide microorganisms according to their ecologically distinct roles, taking into account their evolution and environment: they are defined by a series of mutations that allow them to invade a certain niche. Koeppel *et al.* (2008) argued that ecotypes represent the fundamental units for bacterial diversity. However, the assumption that the environment defines the phenotype, fails to acknowledge cell-to cell

variability. The deterministic vision of ecotypes points the environment as responsible for generating diversity (via processes such as community assembly, selection or dispersal), but does not acknowledge stochasticity as a relevant force. Both environmental pressure and stochasticity seem to play a role in diversification. There are four fundamental ecological processes that generate and maintain diversity: selection, dispersal, diversification and drift (Fig. 7.5). Selection includes the changes in the community structure caused by deterministic fitness differences, such as pH, nutrient availability, salinity, oxygen, bacterial structure or other processes. Dispersal refers to the movements across space, due mostly to external forces such as wind, water or macrobes. Diversification is the genetic variation due to gene transfer, and processes of speciation and extinction amongst others. Drift corresponds to the genetic changes that result from birth death and other stochastic processes. While drift is a stochastic process generated by random processes of cell birth death and reproduction, niche selection is a deterministic process. In other processes, there is a varying degree of stochasticity (Zhou & Ning, 2017; Hanson *et al.*, 2012).

Functionality also plays a crucial role in structuring communities, as it affects not only their activity and performance, but also their resilience, resistance and structure. It can be studied at the genetic level through the ”pangenome”, which refers to all the genes present in a given species, across all isolates (Brockhurst *et al.*, 2019). The pangenome is composed by ”core genes”, that are shared by all the members of a species, and ”accessory genes”, present in some members of a species (Vernikos *et al.*, 2015). It is thought that species with large accessory genomes occupy more varied niches and more complex communities (that they are ”niche generalists”) than those with a smaller accessory genome (the ”niche specialists”) (Brockhurst *et al.*, 2019). The evolution of the pangenome is shaped by gene acquisition through horizontal gene transfer and gene loss (Sela *et al.*, 2020), and these processes are affected by the ecological processes that generate and maintain diversity discussed in the previous section (Fig. 7.5).

On the other hand, functionality is not only shaped by the gene pool but also by the differential expression of the genes, known as the phenotype. For instance, a large study of $>$30.000 marine microorganisms found that environmental conditions generated functional niches, while they only weakly influenced taxonomic composition (Louca *et al.*, 2016). As previously mentioned, there is contradicting evidence on the extent to which diversity and functional redundancy correlate. For instance, Roger *et al.* (2016) found that

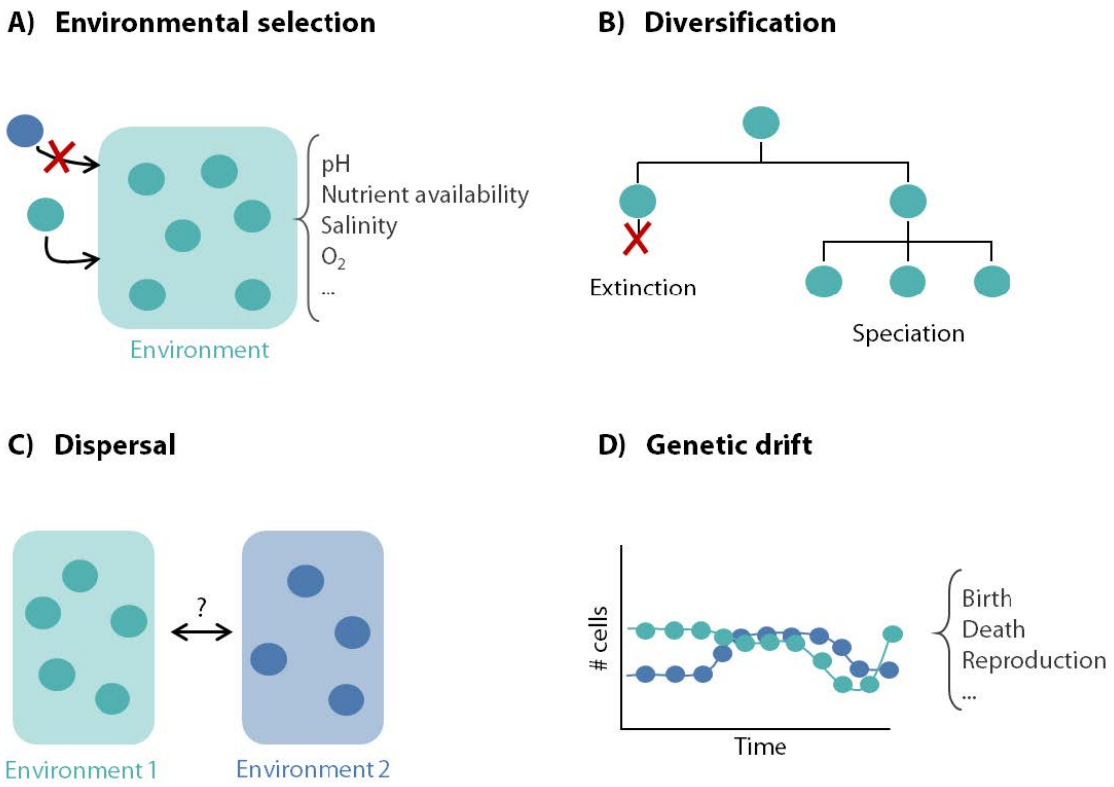


Figure 7.5: Ecological processes that generate and maintain diversity at the genotypic level: selection, dispersal, diversification and genetic drift. A) Environmental selection refers to the habitats and its conditions (pH, nutrient availability, salinity, oxygen, bacterial structure, etc). B) Diversification is the process of generating variation, and includes dormancy, gene transfer, and processes of speciation and extinction. C) Dispersal refers to the movements of organisms across space, for example via wind, water and macrobes. D) Genetic drift is a stochastic change that results from birth, death, reproduction or other random processes.

out of 24 dilution-to-extinction studies of the soil or aquatic communities, the relationship between diversity and functionality was positive in 29% of them, and negative in 10%. On the other hand, a large study by Delgado-Baquerizo *et al.* (2016) demonstrated a positive correlation between multifunctionality (measured as nutrient cycling, primary production, litter decomposition and climate regulation) and microbial diversity in the soil ecosystem, after sampling 78 global drylands and from 179 locations across Scotland. Thus, taxonomic diversity should not be considered as an indication of other community processes, such as functionality, activity, resilience, redundancy or resistance, that need

to be studied separately. Considering this, taxonomic diversity diversity should only be used in a comparative context, to understand how different environments or treatments affect community structure and/or processes (Shade, 2016; Willig, 2011).

So far we have consider the drivers of diversity at the genetic level, neglecting the importance of intra-species phenotypic expression and heterogeneity. Phenotypic expression could be, for example, the change of size in a population due to a change on the growth medium (Yao *et al.*, 2012). Phenotypic heterogeneity refers to the variance amongst single cells of isogenic populations, and can be caused by periodic oscillators (*e.g.*, biological rythms), cell ageing, mitochondrial activity (in the case of eukaryotes), cell-to-cell interactions, epigenetic modifications and stochasticity (Avery, 2006; Ackermann, 2015) (Fig. 7.6), where these mechanisms can be interdependent. Periodic oscillations occur in *Synechococcus elongatus*, where the circadian variation in light influences transcription. Cellular ageing is relevant in rod-shaped organisms, or in budding yeasts. For instance, the rod-shaped bacteria *Caulobacter crescentus* has transcriptional cascades that are

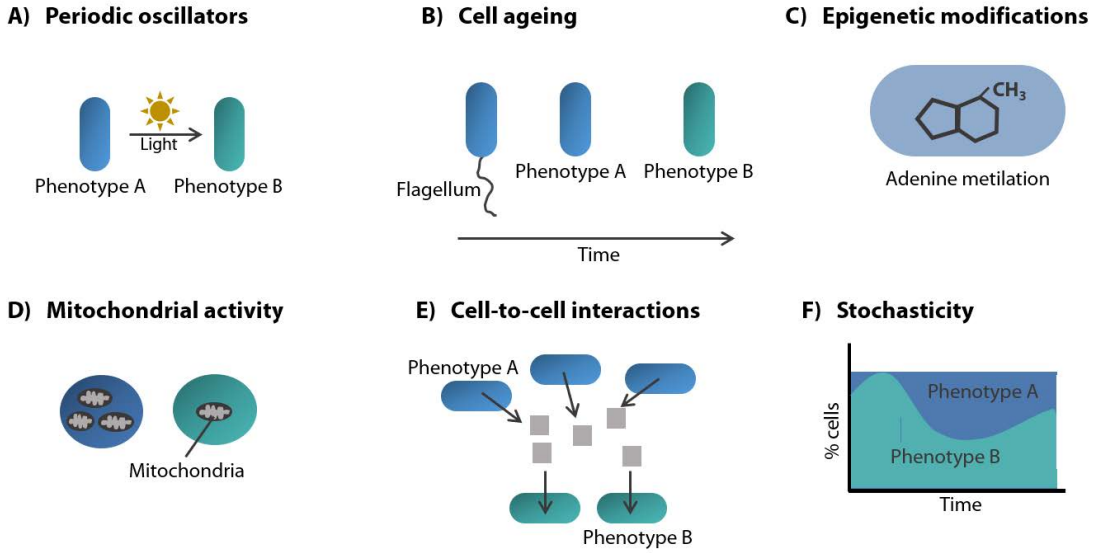


Figure 7.6: Ecological processes that generate diversity at the phenotypic level. A) Periodic oscillators, such as circadian cycles. B) Cell ageing or cell cycles C) Epigenetic modifications, such as DNA adenine methylation. D) Mitochondrial activity or number (in the case of eukaryotes), results in a different tolerance to stress. E) Cell-to-cell interactions, such as *quorum sensing* mechanism. F) Phenotypes can arise stochastically without prior knowledge of the environment.

dependent of the phases of the cell cycle (Lenz & Søgaard-Andersen, 2011). Also, in eukaryotic cells, mitochondrial number and/or activity can result in a differential resistance to stressors (Sumner & Avery, 2002). An example of cell-to-cell interactions is *quorum sensing*, which allows bacteria to signal each other so that when there is a certain cell number, they can orchestrate a response (*e.g.* virulence or biofilm formation) (Bettenworth *et al.*, 2019). The activation of *quorum sensing*-related genes can be heterogeneous, generating phenotypic differences amongst cells (Grote *et al.*, 2014). Epigenetic mechanisms regulate gene transcription or post-translational processes (*e.g.*, feedback loops and DNA adenine methylation) (Murrell *et al.*, 2005). Adam *et al.* (2008) found inheritable epigenetic patterns in *E. coli* that increase phenotypic diversity under low antibiotic concentrations. All these mechanisms that generate phenotypic heterogeneity can be deterministic or have a varying degree of stochasticity.

Stochasticity is the most prominent explanation for bacterial phenotypic heterogeneity in literature, that can arise in combination with other factors or alone (Bettenworth *et al.*, 2019). Stochastic phenotype switching seems to play an important role in the division of labour of the community, and in bet-hedging processes that allow to rapidly adapt to a sudden change in the environment (Tadrowski *et al.*, 2018). For example, persister cells (*i.e.*, cells that are resistant to antibiotics) exist in populations before exposure to an antibiotic (Dhar & McKinney, 2007). Stochasticity can be inherent to the biochemical process of gene expression (intrinsic noise) or originate from other factors that influence gene expression (extrinsic noise) (Smits *et al.*, 2006).

Phenotypic variation is a relevant, although largely ignored phenomenon, that can shape population-level functions. It allows for microbial populations to divide their labour, develop bet-hedging strategies, for altruistic protection and to regulate their output (Fig. 7.7). For instance, bacterial subpopulations can perform different functions that contribute to the public good, dividing their labour and allowing for a structural organisation of the population. Bet-hedging strategies, where populations exhibit different phenotypes at the same time, allow them to survive a sudden environmental change. Also, there are non-contributing cells (cheaters) that benefit from the labour of other cells. Finally, populations can regulate the number of cells displaying a producing or non-producing phenotype to modulate their output (Martins & Locke, 2015; Bettenworth *et al.*, 2019). These processes and strategies are often intertwined. For instance, *Bacillus subtilis* in

biofilms divide their labour depending on their spatial organisation. They can be motile, produce extracellular polymeric substances (EPS) or sporulate (Vlamakis *et al.*, 2008). EPS production is metabolically costly, and thus EPS non-producers are considered cheaters that benefit from the metabolism of others (Martin *et al.*, 2020). In *P. aeruginosa*, the number of EPS producers or cheaters is modulated depending on resource availability (Zhao *et al.*, 2019).

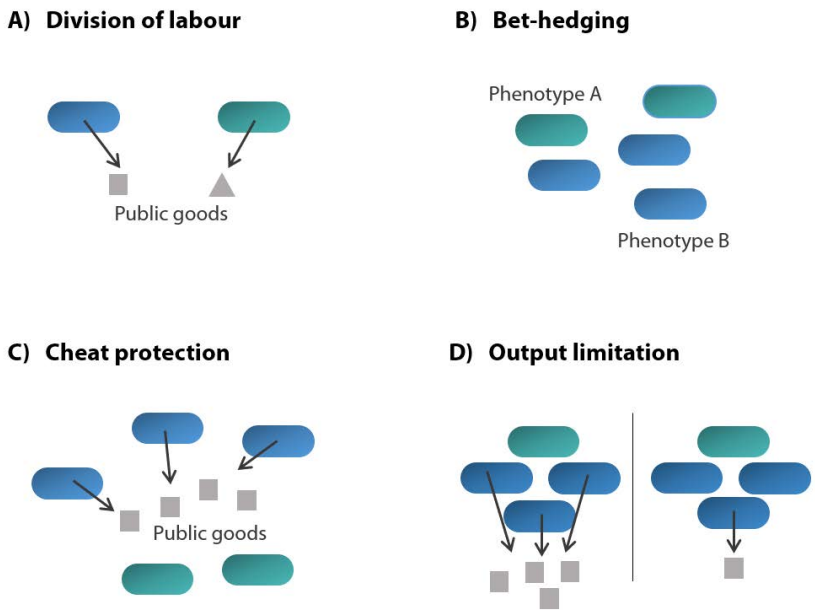


Figure 7.7: Phenotypic heterogeneity allows microbial populations to develop bet-hedging strategies, divide labour and develop strategies or cheat protection and output limitation. A) Dividing labour allows for the distribution of tasks the and spatial organisation of populations. B) Bet-hedging produces a diverse population, which helps to rapidly adapt to an environmental disruption. C) Cheat protection allows non-contributing cells to benefit from the output of the contributing cells. D) Populations can limit their output by regulating how many cells exhibit a producing or non-producing phenotype. Figure modified from Bettenworth *et al.* (2019).

In conclusion, studying diversity at the taxonomic level is insufficient, and a full understanding of microbial populations and their function requires insight on phenotypic diversity processes and their drivers. However, as I will discuss in the following section, defining

what constitutes a phenotype is a challenge by itself.

7.2.2 Defining phenotypes: how far does the rabbit hole go?

Throughout this work, we have defined 'phenotyping' as describing observable characteristics or traits amongst bacteria. There are several single-cell techniques that describe phenotypic traits of individual cells, such as single-cell (multi)-omics, fluorescent labels, imaging techniques, flow cytometry or Raman spectroscopy. However, there are practical challenges when defining phenotypes. First, because the definition of a phenotype is going to depend on the technique used to measure it; secondly because there needs to be a definition on how many differences can be contained within a single phenotype. Although we address this issue in the discussion of **chapter 4 and 5**, we expand here on our proposal, its limitations, and its implications.

When studying phenotypic diversity using Raman spectroscopy, we propose the use of operational phenotypic units (OPUs), defined as the variation of 'traits' in the functional space occupied by an 'ecological unit' (this concept was initially proposed by Carmona *et al.* (2016) to define functional diversity). The ecological unit and traits would have to be defined and justified for each context, depending on the hypothesis. For example, an ecological unit of interest could be a biofilm. However, if the spatial organization within the biofilm is important to address the hypothesis, different parts of the biofilm should be considered as ecological units. Traits could be defined in two different ways: by comparing the whole Raman spectra, or based on a functional (or multiple) Raman label(s). Unlabelled Raman spectra of bacteria can be classified as phenotypes using dimensionality reduction and/or clustering methods and similarity matrixes. Hierarchical clustering, for example, shows the dissimilarity amongst single cells generating a phenotypic tree. This requires the definition of a cut-off to delineate the different phenotypes. As previously discussed, this classification can be done in a data-driven way, using the adjusted Rand index (ARI) combined with Phenograph, or the silhouette index with partitioning around medoids (PAM) using PCA (or other (dis)similarity calculations). This proposal to define OPUs is similar to the classification system proposed by Dumolin *et al.* (2019b) for MALDI-TOF spectra, where they joined single-cell spectra that shared elements and clustered them into operational isolation units (OIUs). On the other hand, functional labels can be used

to define the phenotypical traits. For example, deuterium labelling tells about the general microbial activity, and labelled ^{13}C or ^{15}N can tell about the metabolism of single cells (Berry *et al.*, 2015; Jing *et al.*, 2018; Cui *et al.*, 2018). In this case, clustering could be based on the metabolism of individual cells.

The use of Raman spectroscopy presents several challenges, as we have pinpointed throughout this research, specifically in **chapter 6**. Microbes are complex systems and it is sometimes difficult to disentangle the Raman spectra and define what compound(s) peaks correspond to. Different molecules can have confounding Raman signals, and some can have a weak (or no) Raman signal. These limitations to study community composition affect diversity calculations and OPU definitions. Also, there are multiple databases that describe different Raman wavelengths to identify the same molecules. When picking a database to assign the peaks in Raman spectra, it is important to consider the shift that can exist between the database and the instrument used to measure the sample due to the instrumental variation. This phenomenon is complicated considering that there can be non-linear drifts in the Raman shifts. There are computational models that can automatically account for this shift between the instruments and correct it, such as for example the ”moving window fast Fourier transform cross-correlation”, that uses a standard spectrum to evaluate the shift of each spectral point (Chen *et al.*, 2018).

A second challenge is how to integrate single-cell phenotypic information into the tree of life. I propose that phenotypic units - OPUs in the case of Raman spectroscopy- are used as a ‘singe-cell layer’ of information complementary to taxonomic measurements. To explain this, let us suppose that there are two environments as shown in Fig. 7.8. Environment 1 has tree ecotypes, and environment 2 has two ecotypes. Phylogenetic inference can be made on the ancestor(s) of these ecotypes. Raman spectroscopy can study the single-cell variations of these ecotypes, by clustering them into OPUs. In this example, we find that OPU A is present in both ecotypes, while OPU B and C are exclusive of environment 1 and 2 respectively. This model could be adapted to any other single cell technique. In fact, a similar approach has been proposed by Van Rossum *et al.* (2020) to integrate metagenomic and taxonomic information although they do not consider single-cell differences.

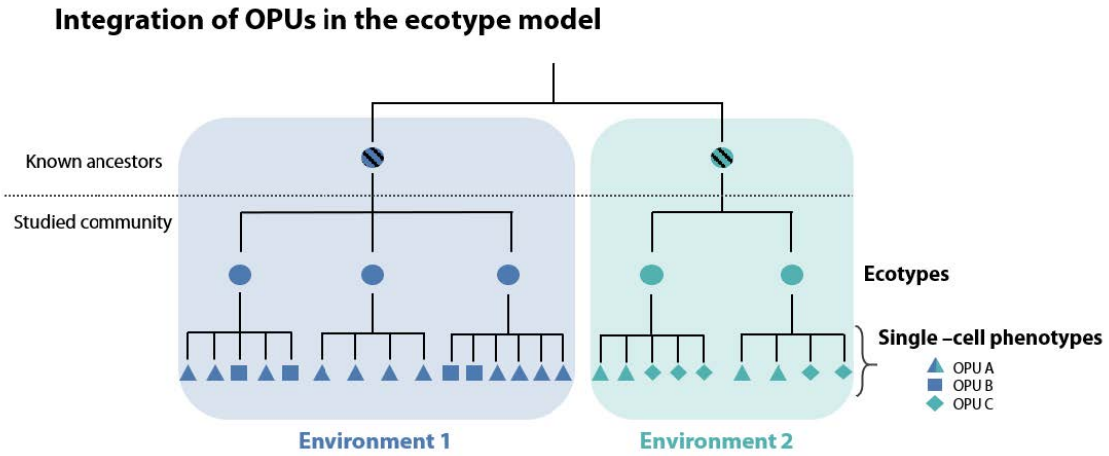


Figure 7.8: Example on the integration of OPUs into the ecotype model. The ecotypes are identified according to taxonomic studies and their environment. The ancestry line can be drawn if relevant. Single-cell information on each ecotype can also be drawn for the populations that are being studied. In the case of Raman spectroscopy studies, operational phenotypic units (OPUs) can be defined according to spectral similarities and the hypothesis tested.

7.3 Raman spectroscopy applications in natural and engineered microbial ecosystems

Raman spectroscopy is an attractive a tool for microbial ecologists. It is a single cell technique that is relatively fast, non-destructive and does not require cell labelling. It allows for the (semi)quantification of biomolecules in the sample, and phenotype or strain identification. Its speed has improved with the use of metallic particles, either in suspension, on a surface using SERS, or on the tip of a scanning probe or on the sample using TERS. Also, CARS allows to increase the signal-to-noise ratio; and isotopes can be used to follow certain metabolic pathways or the metabolic rate (*e.g.*, ^{13}C or deuterium). Finally, the microfluidics chips allow to rapidly isolate single cells. Cell sorting allows to do further analysis (*e.g.*, cultivation or sequencing) on individual microorganisms or subpopulations. In the following section, we discuss the opportunities that Raman spectroscopy offers, and

how it can help in the understanding and management of microbial communities.

Microbial resource management refers to an optimal management of microbes with the aim to develop new products and to improve existing bioprocesses, and requires answering who is in the community, what they are doing and with whom they are interacting (Verstraete *et al.*, 2007). Raman spectroscopy offers many opportunities to address these three questions in both natural and engineered communities. It can detect the strains present in a sample (who is there) (Maquelin *et al.*, 2003; Harz *et al.*, 2005; Green *et al.*, 2009) and study their functionality (what they are doing) (Berry *et al.*, 2015; Muhamadali *et al.*, 2015) at a single-cell level. Observing interrelationships (who is doing what with whom) ideally requires studying communities in their environment with minimal disruption, something that Raman spectroscopy allows. Although synthetic communities are popular for hypothesis testing in a controlled and less complex environment (De Roy *et al.*, 2014), it is known that they do not always accurately represent the dynamics observed in natural communities (Yu *et al.*, 2016). Observing microorganisms in their natural environment could be specially interesting for studying biofilms, as cell behaviour -as in motility, antibiotic tolerance or metabolism- in planktonic cultures and biofilms is not the same (Stewart & Franklin, 2008). Also, studying directly in natural communities allows to detect unculturable or viable but non-culturable organisms. It is estimated that most organisms are unculturable (or “yet to be cultured”), what could be explained by the low prevalence of certain bacteria or their slows growth, the limits of the molecular techniques in distinguishing closely related species or the inability of scientists to recreate demanding conditions necessary for growing some of these organisms (Vartoukian *et al.*, 2010; Kaeberlein *et al.*, 2002). It is also possible that certain bacteria need to be cocultured with other microorganisms to be able to grow (Wade, 2002).

Viable but non-culturable (VBNC) organisms are those who lost (temporarily or not) their ability to grow on media, although they can remain metabolically active . Many pathogenic strains have a VBNC form, and thus being able to study these cells is of great interest for the medical sciences. It is thought that persister cells, which are subpopulations that do not grow and develop antibiotic resistance, might be a type of VBNC cells, although there is discussion around this classification (Li *et al.*, 2014). Raman spectroscopy has been used to study the appearance of VBNC bacteria in water samples and lab strains by exposing them to UV and following their metabolic activity (*i.e.*, the incorporation of D_2O)

over time. In this study, activity was also assessed using 5-cyano-2,3-ditolyl tetrazolium chloride combination flow cytometry (CTC-FCM), to measure of respiration intensity (Guo *et al.*, 2019). It would be interesting to use the Raman spectra to look for metabolic changes underlying the transformation from culturable to the VBNC state.

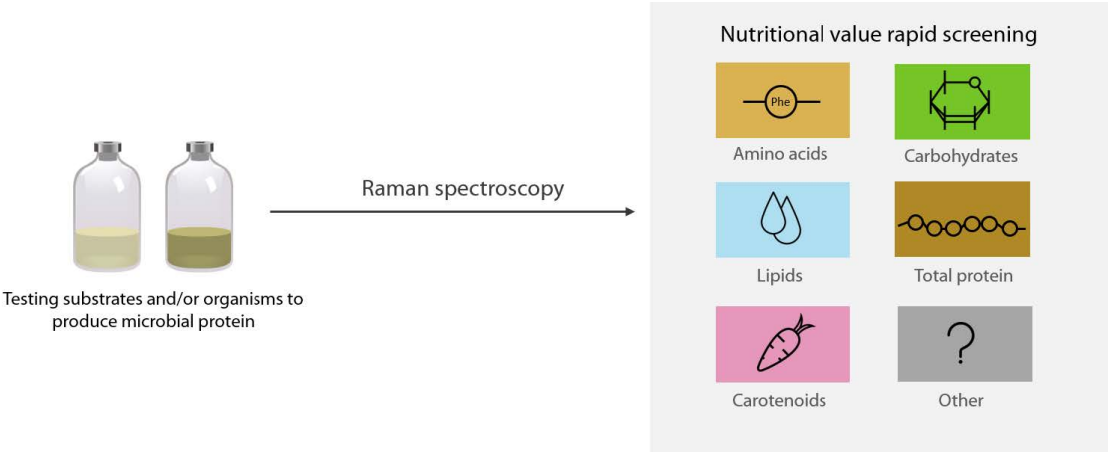


Figure 7.9: Raman spectroscopy can serve as a tool to rapidly estimate nutritionally valuable compounds, and help understand the conditions (*e.g.*, substrate, organisms) that produce a higher quality product.

The non-destructive nature of Raman analysis makes it an interesting candidate for evolutionary ecology studies, where ethical considerations call to preserve the host organism (*e.g.*, when studying animal microbiome or the microbiome of endangered plant species), or where the destruction of the sample can be avoided (*e.g.*, rocks). Keeping the sample intact reduces the risk of contamination and allows to follow up the dynamic of the system and/or further characterisation using other techniques. This is also an advantage when considering the spatial (dynamic) organisation in ecosystems. The interactions of microbial populations affect each individual's survival in a positive, negative or neutral way (for example, in mutualism, competition, parasitism or commensalism) (Faust & Raes, 2012), and these interactions can be largely modelled by the spatial organisation of the given ecosystem. A clear example of this is biofilms, where different bacteria may cooperate to build a structure that will bring protection, resistance to antimicrobials or colonisation, and is also a metabolic exchange platform (Santos *et al.*, 2018). It is possible for biofilms to create "artificial" spaces for bacteria that would otherwise not survive the environment (*e.g.*, anaerobic conditions in an aerobic environment). Another consequence

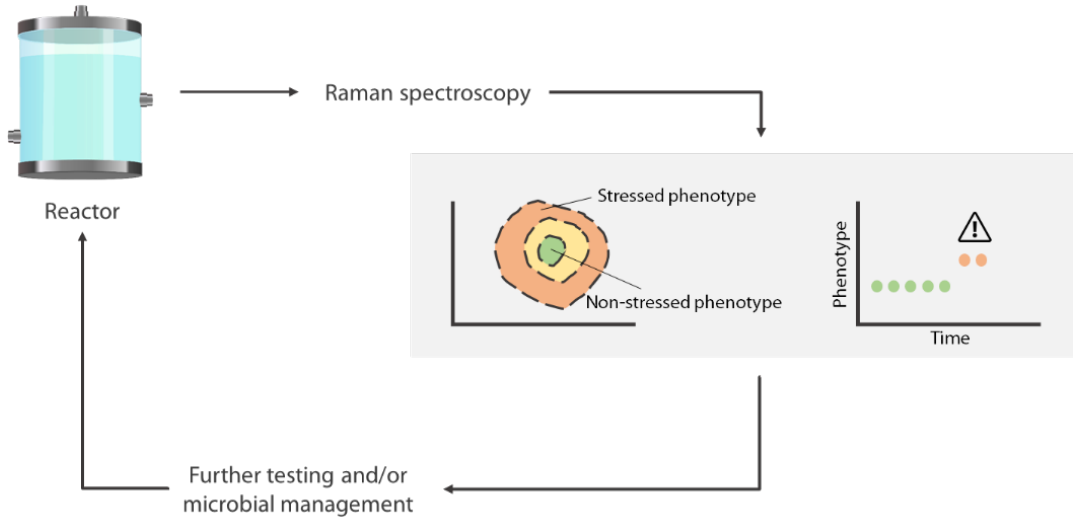


Figure 7.10: Management of microbial populations using Raman spectroscopy. Raman spectroscopy can discriminate stressed and non-stressed phenotypes. An alert system could warn the operator when the optimal phenotype is deviating from the normal. Then, the operator could further test the reactor, or steer the conditions.

of spatial organisation is phenotypic diversification. The existence of (micro)environments where some variable(s) differ can drive a differential phenotypic expression (Gough *et al.*, 2017). Raman spectroscopy allows to explore biofilm formation taking into account spatial organization and informs not only about the microorganisms and their phenotypes but also about other components of the biofilms, such as EPS (Ivleva *et al.*, 2017).

Raman spectroscopy is used in the processing of drugs and other chemicals (De Beer *et al.*, 2011), and Raman probes have been used in mammalian cell cultures to measure parameters such as glutamine, glutamate, glucose, lactate, ammonium, viable cell density, and total cell density (Abu-Absi *et al.*, 2011; Whelan *et al.*, 2012). Although the use of probes in microbial cultures might be challenging – mammalian cells have $\sim$1000 times the volume of bacterial cells (Milo *et al.*, 2009) – microfluidic chips or spectral flow cytometry-could be used. We argue that Raman spectroscopy might have a future in bioproduction using its capacity to (1) quantify (absolutely or relatively) the amount of certain compounds and to (2) fingerprint. For instance, it could be used in microbial protein production as a tool to estimate the nutritionally valuable compounds in a culture, to rapidly assess the culture conditions that help achieve the best nutritional profile (Fig. 7.9). Its fingerprinting

capacity can be used to define the fingerprint of a stressed and non-stressed culture, and to alert the operator when the population in the bioreactor is steering away to a stressed phenotype. At this point, the operator could further investigate this change, and/or steer the conditions of the reactor to achieve a more optimal production (Fig. 7.10).

Other infrared spectroscopy techniques, such as Fourier transform mid infrared (FT-MIR), near infrared red (NIR), and Fourier transform (FT)-Raman spectroscopy have been used to monitor biomass, glucose, and lactic acid lactic acid fermentation, but only FT-MIR could discriminate these compounds (Sivakesava *et al.*, 2001). However, FT-Raman has differences over dispersive Raman: it uses a 1064 cm^{-1} laser, that cannot be used in aqueous solutions, and it is considered to have less resolution (Scientific, 1996-2020), therefore, the possibilities and limitations of Raman for online monitoring of bacterial cultures remain largely unexplored.

There are certain limitations when using Raman spectroscopy in microbial communities and populations. Apart from those mentioned in the previous section, relative to the nature of Raman spectra and instrumental limitations, there are other challenges. This is because when bacteria grow together, they influence each other's phenotype (Heyse *et al.*, 2019). Meaning that if one has a database for axenic cultures, it will be difficult to use it in a coculture. If cocultures want to be followed in a dynamic study, it is best to either physically separate the strains (Heyse *et al.*, 2019), or confirm their identity using a FISH when different strains are growing together. It is important to note that the first approach might influence the behaviour of the culture as the cells will not come to proximity, even if the physical separation allows for the exchange of metabolites. Finally, the growth phase of the cells also needs to be taken into account, as can influence their phenotype.

7.3.1 Conclusion

Microbial ecology studies the relationships of microorganisms and their environment to describe, explain, predict and control microbial communities (Konopka, 2009). Understanding and measuring phenotypic diversity in these communities is key for their management. Raman spectroscopy is a rapid, label-free technique that allows to measure cells without contact, making it interesting for studying natural and engineered systems. It is able to

discriminate phenotypes and extract (semi)quantitative information on single cells. Also, Raman spectroscopy can be used to measure single-cell phenotypic heterogeneity. There are certain limitations when using Raman spectroscopy for measuring microbial populations. Some are relative to the nature of Raman spectra and instrumental limitations, but there are also challenges for the dynamic monitoring of communities.

Spatial and temporal organization, microbial diversity and functionality as well as single-cell heterogeneity are important for a better understanding of the structure of microbial communities. They allow to move beyond a static insight and to interrogate its dynamics and response to disturbances. In this research work, I discuss how Raman spectroscopy can be used in microbial ecology to describe phenotypes and phenotypic diversity, and how to integrate single-cell information to taxonomic studies. A more in-depth insight on single-cell heterogeneity and how it shapes microbial populations is key for understanding and managing community composition, structure, functionality and group dynamics.

General conclusions

- **The Raman spectra of unlabelled bacterial cells can be modified due to the sample preparation and collection.** The delays between cell fixation and taking the measurement, the time the sample stays on the slide and the number of centrifugations done during the sample preparation greatly impact single-cell classification when using supervised (*i.e.*, random forest) and unsupervised (*i.e.*, hierarchical clustering) algorithms. This can lead to the confusion of irrelevant noise as a phenotype. Hence, metadata collection is important for better experimental interpretation and reproducibility.

- **Raman microscopy can be used to identify single-cell phenotypes in isogenic populations, while flow cytometry can detect changes at the population level.** Flow cytometry is a more high-throughput technology than label-free Raman spectroscopy; however, Raman can extract many more variables per cell, without the need for staining. When we compared their resolution when detecting phenotypes in an isogenic population, we found that flow cytometry can detect phenotypic changes at the population level, whereas Raman microscopy has sufficient resolving power to identify separate phenotypes at the single-cell level. Also, Raman microscopy provides the possibility to infer which metabolic properties define different phenotypic populations and can potentially exploit this information for bioprocess monitoring.

- **Microbial phenotypes can be automatically identified based on their Raman spectra.** In this manuscript, we discuss how different dimensionality reduction techniques can be used to visualize the bacterial phenotypes. We also propose the use of different clustering techniques, such as hierarchical clustering, partitioning around medoids or Phenograph, to automatically identify the phenotye of individual

cells.

The definition of phenotypic populations using these clustering algorithms is highly dependent on the similarity threshold that is used to delimit them. We therefore suggest that researchers include validation controls in their experimental setup, so that the detected populations are ecologically meaningful and not merely arbitrarily defined groups.

- **Microbial single-cell phenotypic diversity can be quantified using Raman microscopy.** Raman spectral points correspond to a different chemical bond (or to multiple ones) that are expressed with a certain abundance. Using this information in the Hill diversity framework, it is possible to calculate the phenotypic diversity of single cells. We tested this workflow to detect stress-driven phenotypic diversity in a prokaryotic and eukaryotic population.

- **Label-free Raman microscopy can be used as a tool to estimate nutritionally valuable compounds.** The bulk quantification of amino acids in microbial protein remains slow and time-consuming, and Raman microscopy is an alternative to quantify the total protein content and content of the indispensable amino acids histidine, leucine, lysine, methionine, tryptophan, phenylalanine, valine and cysteine, in single cells with high accuracy ($R^2 \geq 0.93$). Raman microscopy can also quantify unsaturated fatty acids, nucleic acids, vitamins and carotenoids. We propose the use of this tool in biocultures to optimize the microbial populations and/or its growing conditions (substrate, pH, temperature), as well as to follow the metabolism of individual cells over time.

Afterword

This scientific work, as many others, tries to offer technical solutions and ideas that could find an application in the field of microbial ecology. A field that, I believe, will become increasingly important as part of the circular economy model. Although I am excited about these new roads, I am increasingly worried that by our offering of technical solutions, scientists do not allow other fellow citizens to think more critically about the way we produce or consume. By increasing society's hope in a technical utopian tomorrow, we might be impeding or postponing creative solutions that could address root problems underlying current and future crisis.

Acknowledgements

This thesis is the result of a number of collaborations, and would not be possible without my colleagues from CMET. Their intelligence and good humour have helped shaping this work and have increased its writer's resilience. I would like to thank the members of the Ecology cluster, whose feedback and ideas inspired these chapters, and to the CMET ATP team, that makes everything run smoothly.

A special mention to my promoters, Nico Boon and Ruben Props, and the colleagues more closely involved in my research: Jasmine Heyse, Charlotte De Rudder, Ioanna Chatzigiannidou, Josefien Van Landuyt, Myrsini Sakarika, Benjamin Buysschaert, Peter Rubbens, Frederiek-Maarten Kerckhof and Yuting Guo. I would also like to thank the members of my examination committee, that help to significantly increase the quality of this manuscript.

I am grateful to my parents, Pilar and Gabriel, and my siblings, María and Eduardo, for their support throughout all these (many) academic years. To Leo, for his company through long writing days (despite his occasional messing with the manuscript). To my grandmother Pilar, who has always encouraged me. To my friends, for keeping me grounded.

Thank you Elie, for your love and encouragement.

Brussels, October 2020
Cristina García Timermans

References

Abu-Absi, Nicholas R., Kenty, Brian M., Cuellar, Maryann Ehly, Borys, Michael C., Sakhamuri, Sivakesava, Strachan, David J., Hausladen, Michael C., & Li, Zheng Jian. 2011. Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. *Biotechnology and Bioengineering*, **108**(5), 1215–1221.

Ackermann, Martin. 2015. A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews Microbiology*, **13**(8), 497–508.

Adam, M., Murali, B., Glenn, N. O., & Potter, S. S. 2008. Epigenetic inheritance based evolution of antibiotic resistance in bacteria. *BMC evolutionary biology*, **8**, 52.

Ali, Nairveen, Girnus, Sophie, Rösch, Petra, Popp, Jürgen, & Bocklitz, Thomas. 2018. Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example. *Analytical chemistry*, **90**(21), 12485:12492.

Almarashi, Jamal F. M., Kapel, Natalia, Wilkinson, Thomas S., & Telle, Helmut H. 2012. Raman Spectroscopy of Bacterial Species and Strains Cultivated under Reproducible Conditions. *Spectroscopy: An International Journal*, **27**(5-6), 361–365.

Altschuler, Steven J., & Wu, Lani F. 2010. Cellular Heterogeneity: Do Differences Make a Difference? *Cell*, **141**(4), 559–563.

Ambriz-Avina, Veronica, Contreras-Garduno, Jorge A., & Pedraza-Reyes, Mario. 2014. Applications of Flow Cytometry to Characterize Bacterial Physiological Responses. *BioMed Research International*, **2014**, 14.

Amir, El-ad David, Davis, Kara L, Tadmor, Michelle D, Simonds, Erin F, Levine, Jacob H, Bendall, Sean C, Shenfeld, Daniel K, Krishnaswamy, Smita, Nolan, Garry P, & Pe'er, Dana. 2013. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, **31**(6), 545–552.

Andrews, Tallulah S., & Hemberg, Martin. 2018. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, **59**(feb), 114–122.

Anetzberger, Claudia, Schell, Ursula, & Jung, Kirsten. 2012. Single cell analysis of Vibrio harveyi uncovers functional heterogeneity in response to quorum sensing signals. *BMC Microbiology*, **12**(1), 209.

Anupama, & Ravindra, P. 2000. Value-added food:: Single cell protein. *Biotechnology Advances*, **18**(6), 459 – 479.

Arrigucci, Riccardo, Bushkin, Yuri, Radford, Felix, Lakehal, Karim, Vir, Pooja, Pine, Richard, Martin, December, Sugarman, Jeffrey, Zhao, Yanlin, Yap, George S, Lardizabal, Alfred A, Tyagi, Sanjay, & Gennaro, Maria Laura. 2017. FISH-Flow, a protocol for the concurrent detection of mRNA and protein in single cells using fluorescence in situ hybridization and flow cytometry. *Nature Protocols*, **12**(6), 1245–1260.

Athamneh, A. I. M., Alajlouni, R. A., Wallace, R. S., Seleem, M. N., & Senger, R. S. 2014. Phenotypic Profiling of Antibiotic Response Signatures in Escherichia coli Using Raman Spectroscopy. *Antimicrobial Agents and Chemotherapy*, **58**(3), 1302–1314.

Avery, Simon V. 2006. Microbial cell individuality and the underlying sources of heterogeneity. *Nature Reviews Microbiology*, **4**(8), 577–587.

Barbesti, Silvia, Citterio, Sandra, Labra, Massimo, Baroni, Maurizio Davide, Neri, Maria Grazia, & Sgorbati, Sergio. 2000. Two and three-color fluorescence flow cytometric analysis of immunoidentified viable bacteria. *Cytometry*, **40**(3), 214–218.

Barton, Sinead J., & Hennelly, Bryan M. 2019. An Algorithm for the Removal of Cosmic Ray Artifacts in Spectral Data Sets. *Applied Spectroscopy*, **73**(8), 893–901. PMID: 31008665.

Beattie, J. Renwick, Glenn, Josephine V., Boulton, Michael E., Stitt, Alan W., & McGarvey, John J. 2009. Effect of signal intensity normalization on the multivariate analysis of spectral data in complex 'real-world' datasets. *Journal of Raman Spectroscopy*, **40**(4), 429–435.

Beleites, C, & Sergo, V. 2012. hyperSpec: a package to handle hyperspectral data sets in R. *Journal of Statistical Software*.

Beleites, Claudia, & Salzer, Reiner. 2008. Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry*, **390**(5), 1261–1271.

Beleites, Claudia; Sego, Valter. 2017. *hyperSpec: a package to handle hyperspectral data sets in R*.

Benomar, Saida, Ranava, David, Cárdenas, María Luz, Trably, Eric, Rafrafi, Yan, Ducret, Adrien, Hamelin, Jérôme, Lojou, Elisabeth, Steyer, Jean Philippe, & Giudici-Orticoni, Marie Thérèse. 2015. Nutritional stress induces exchange of cell material and energetic coupling between bacterial species. *Nature Communications*, **6**(1), 1–10.

Berney, Michael, Hammes, Frederik, Bosshard, Franziska, Weilenmann, Hans-Ulrich, & Egli, Thomas. 2007. Assessment and Interpretation of Bacterial Viability by Using the LIVE/DEAD BacLight Kit in Combination with Flow Cytometry. *Applied and Environmental Microbiology*, **73**(10), 3283–3290.

Berry, David, Mader, Esther, Lee, Tae Kwon, Woebken, Dagmar, Wang, Yun, Zhu, Di, Palatinszky, Marton, Schintlmeister, Arno, Schmid, Markus C., Hanson, Buck T., Shterzer, Naama,

Mizrahi, Itzhak, Rauch, Isabella, Decker, Thomas, Bocklitz, Thomas, Popp, Jürgen, Gibson, Christopher M., Fowler, Patrick W., Huang, Wei E., & Wagner, Michael. 2015. Tracking heavy water (D2O) incorporation for identifying and sorting active microbial cells. *Proceedings of the National Academy of Sciences*, **112**(2), E194–E203.

Bettenworth, Vera, Steinfeld, Benedikt, Duin, Hilke, Petersen, Katrin, Streit, Wolfgang R., Bischofs, Ilka, & Becker, Anke. 2019. Phenotypic Heterogeneity in Bacterial Quorum Sensing Systems. *Journal of Molecular Biology*, **431**(23), 4530 – 4546. Underlying Mechanisms of Bacterial Phenotypic Heterogeneity and Sociobiology.

Bhatia, Ravi, & Nangul, Agam. 2013. Microorganism : A marvelous source of single cell protein. *Journal of microbiology, biotechnology and food sciences*, **3**(08), 15–18.

Bhattacharjee, Arunima, Datta, Rupsa, Gratton, Enrico, & Hochbaum, Allon I. 2017. Metabolic fingerprinting of bacteria by fluorescence lifetime imaging microscopy. *Scientific Reports*, **7**(1), 3743.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. 2016. mlr: Machine Learning in R. *Journal of Machine Learning Research*, **17**(170), 1–5.

Bizzini, A., Durussel, C., Bille, J., Greub, G., & Prodhom, G. 2010. Performance of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Identification of Bacterial Strains Routinely Isolated in a Clinical Microbiology Laboratory. *Journal of Clinical Microbiology*, **48**(5), 1549–1554.

Blondel, Vincent D, Guillaume, Jean-Loup, Lambiotte, Renaud, & Lefebvre, Etienne. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), P10008.

Bock, Christoph, Farlik, Matthias, & Sheffield, Nathan C. 2016. *Multi-Omics of Single Cells: Strategies and Applications*.

Bodelier, Paul. 2011. Toward Understanding, Managing, and Protecting Microbial Ecosystems. *Frontiers in Microbiology*, **2**, 80.

Boulesteix, Anne-Laure, Janitza, Silke, Kruppa, Jochen, & Konig, Inke R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(6), 493–507.

Bouvier, Thierry, & del Giorgio, Paul A. 2003. Factors influencing the detection of bacterial cells using fluorescence in situ hybridization (FISH): A quantitative review of published reports. *FEMS Microbiology Ecology*, **44**(1), 3–15.

Breiman, Leo. 2001. Random Forests. *Machine Learning*, **45**, 5–32.

Brenner, Don J., Fanning, George R., Rake, Adrian V., & Johnson, Karl E. 1969. Batch procedure for thermal elution of DNA from hydroxyapatite. *Analytical Biochemistry*, **28**, 447 – 459.

Brockhurst, Michael A., Harrison, Ellie, Hall, James P.J., Richards, Thomas, McNally, Alan, & MacLean, Craig. 2019. The Ecology and Evolution of Pangenomes. *Current Biology*, **29**(20), R1094 – R1103.

Bunaciu, Andrei A., Aboul-Enein, Hassan Y., & Hoang, Vu Dang. 2015. Raman Spectroscopy for Protein Analysis. *Applied Spectroscopy Reviews*, **50**(5), 377–386.

Butler, H. J., Ashton, L., Bird, B., Cinque, G., Curtis, K., Dorney, J., Esmonde-White, K., Fullwood, N. J., Gardner, B., Martin-Hirsch, P. L., Walsh, M. J., McAinsh, M. R., Stone, N., & Martin, F. L. 2016. Using Raman spectroscopy to characterize biological materials. *Nat Protoc*, **11**(4), 664–87.

Button, D. K., & Robertson, Betsy R. 2001. Determination of DNA Content of Aquatic Bacteria by Flow Cytometry. *Applied and Environmental Microbiology*, **67**(4), 1636–1645.

Carmona, Carlos P., [de Bello], Francesco, Mason, Norman W.H., & Lepš, Jan. 2016. Traits Without Borders: Integrating Functional Diversity Across Scales. *Trends in Ecology & Evolution*, **31**(5), 382 – 394.

Cereghino, Joan Lin, & Cregg, James M. 2000. Heterologous protein expression in the methylotrophic yeast Pichia pastoris. *FEMS Microbiology Reviews*, **24**(1), 45–66.

Ceuppens, S., Timmery, S., Mahillon, J., Uyttendaele, M., & Boon, N. 2013. Small Bacillus cereus ATCC 14579 subpopulations are responsible for cytotoxin K production. *Journal of Applied Microbiology*, **114**(3), 899–906.

Chai, Yunrong, Norman, Thomas, Kolter, Roberto, & Losick, Richard. 2010. An epigenetic switch governing daughter cell separation in Bacillus subtilis. *Genes & development*, **24**(8), 754—765.

Chan, J. W., Winhold, H., Lane, S. M., & Huser, T. 2005. Optical trapping and coherent anti-Stokes Raman scattering (CARS) spectroscopy of submicron-size particles. *IEEE Journal of Selected Topics in Quantum Electronics*, **11**(4), 858–863.

Chao, Anne, Chiu, Chun-Huo, & Jost, Lou. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, **45**(1), 297–324.

Chen, Hui, Liu, Yan, Lu, Feng, Cao, Yongbing, & Zhang, Zhi-Min. 2018. Eliminating Non-linear Raman Shift Displacement Between Spectrometers via Moving Window Fast Fourier Transform Cross-Correlation. *Frontiers in Chemistry*, **6**, 515.

Chen, Nathan H., Djoko, Karrera Y., Veyrier, Frédéric J., & McEwan, Alastair G. 2016. Formaldehyde Stress Responses in Bacterial Pathogens. *Frontiers in Microbiology*, **7**, 257.

Chen, H.and Zhang, Z., Miao, L., Zhan, D., Zheng, Y., Liu, Y., Lu, F., & Liang, Y. 2014. Automatic standardization method for Raman spectrometers with applications to pharmaceuticals. *Journal of Raman Spectroscopy*, **46**(1).

Chisanga, Malama, Muhamadali, Howbeer, Ellis, David I., & Goodacre, Royston. 2018. Surface-Enhanced Raman Scattering (SERS) in Microbiology: Illumination and Enhancement of the Microbial World. *Applied Spectroscopy*, **72**(7), 987–1000. PMID: 29569946.

Chiu, Liang-da, Hullin-Matsuda, Françoise, Kobayashi, Toshihide, Torii, Hajime, & Hamaguchi, Hiro-O. 2012. On the origin of the 1602 cm-1 Raman band of yeasts; contribution of ergosterol. *Journal of biophotonics*, **5**(10), 724—728.

Clarke, K. Robert, Somerfield, Paul J., & Chapman, M. Gee. 2006. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology and Ecology*, **330**(1), 55 – 80. A Tribute to Richard M. Warwick.

Clarke, Sarah, & Goodacre, Royston. 2003. *Raman Spectroscopy for Whole Organism and Tissue Profiling*. Boston, MA: Springer US. Pages 95–110.

Cowcher, David P., Xu, Yun, & Goodacre, Royston. 2013. Portable, Quantitative Detection of *Bacillus* Bacterial Spores Using Surface-Enhanced Raman Scattering. *Analytical Chemistry*, **85**(6), 3297–3302.

Cronin, Ultan P., & Wilkinson, Martin G. 2008. Bacillus cereus endospores exhibit a heterogeneous response to heat treatment and low-temperature storage. *Food Microbiology*, **25**(2), 235–243.

Cui, Li, Yang, Kai, Li, Hong-Zhe, Zhang, Han, Su, Jian-Qiang, Paraskevaidi, Maria, Martin, Francis L., Ren, Bin, & Zhu, Yong-Guan. 2018. Functional Single-Cell Approach to Probing Nitrogen-Fixing Bacteria in Soil Communities by Resonance Raman Spectroscopy with 15N2 Labeling. *Analytical Chemistry*, **90**(8), 5082–5089. PMID: 29557648.

Daly, Aisling, Baetens, Jan, & De Baets, Bernard. 2018. Ecological Diversity: Measuring the Unmeasurable. *Mathematics*, **6**(7), 119.

Davey, Hazel M., & Kell, Douglas B. 1996. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses. *Microbiological reviews*, **60**(4), 641–96.

Davis, Kimberly M., & Isberg, Ralph R. 2016. Defining heterogeneity within bacterial populations via single cell approaches. *BioEssays*, **38**(8), 782–790.

De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J.P., & Vervaet, C. 2011. Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *International Journal of Pharmaceutics*, **417**(1), 32 – 47. Advanced characterization techniques.

De Gelder, Joke, De Gussem, Kris, Vandenabeele, Peter, & Moens, Luc. 2007. Reference database of Raman spectra of biological molecules. *Journal of Raman Spectroscopy*, **38**(9), 1133–1147.

De Gelder, Joke, De Vos, Paul, Moens, Luc, & Vandenabeele, Peter. 2008. *Chapter 3 - Raman spectroscopy: Advantages and disadvantages*.

De Mey, Marjan, Lequeux, Gaspard, Maertens, Jo, Muynck, Cassandra, Soetaert, Wim, & Vandamme, Erick. 2008. Comparison of protein quantification and extraction methods suitable for E-coli cultures. *Biologicals : journal of the International Association of Biological Standardization*, **36**(06), 198–202.

De Paepe, Kim, Kerckhof, Frederiek-Maarten, Verspreet, Joran, Courtin, Christophe M., & Van de Wiele, Tom. 2017. Inter-individual differences determine the outcome of wheat bran colonization by the human gut microbiome. *Environmental Microbiology*, **19**(8), 3251–3267.

De Roy, Karen. 2014. *Microbial resource management : introducing new tools and ecological theories*. Ph.D. thesis, Ghent University.

De Roy, Karen, Clement, Lieven, Thas, Olivier, Wang, Yingying, & Boon, Nico. 2012. Flow cytometry for fast microbial community fingerprinting. *Water Research*, **46**(3), 907–919.

De Roy, Karen, Marzorati, Massimo, Van den Abbeele, Pieter, Van de Wiele, Tom, & Boon, Nico. 2014. Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental Microbiology*, **16**(6), 1472–1481.

Degenhardt, Frauke, Seifert, Stephan, & Szymczak, Silke. 2017. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 1–12.

Delgado-Baquerizo, Manuel, Maestre, Fernando, Reich, Peter, Jeffries, Thomas, Gaitan, Juan, Encinar, Daniel, Berdugo, Miguel, Campbell, Colin, & Singh, Brajesh. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nature Communications*, **7**(01).

Delvigne, Frank, Baert, Jonathan, Gofflot, Sébastien, Lejeune, Annick, Telek, Samuel, Johanson, Ted, & Lantz, Anna Eliasson. 2015. Dynamic single-cell analysis of Saccharomyces cerevisiae under process perturbation: comparison of different methods for monitoring the intensity of population heterogeneity. *Journal of Chemical Technology & Biotechnology*, **90**(2), 314–323.

Dhar, Neeraj, & McKinney, John D. 2007. Microbial phenotypic heterogeneity and antibiotic tolerance. *Current Opinion in Microbiology*, **10**(1), 30–38.

Díaz, Mario, Herrero, Mónica, García, Luis A., & Quirós, Covadonga. 2010. Application of flow cytometry to industrial microbial bioprocesses. *Biochemical Engineering Journal*, **48**(3), 385–407.

Dieterich, Daniela C., Link, A. James, Graumann, Johannes, Tirrell, David A., & Schuman, Erin M. 2006. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proceedings of the National Academy of Sciences*, **103**(25), 9482–9487.

Dodge, Y. 1987. *Statistical Data Analysis Based on the L1-norm and Related Methods*. Springer Nature Book Archives Millennium. North-Holland.

Dumolin, Charles, Aerts, Maarten, Verheyde, Bart, Schellaert, Simon, Vandamme, Tim, Van der Jeugt, Felix, De Canck, Evelien, Cnockaert, Margo, Wieme, Anneleen D., Cleenwerck, Ilse, Peiren, Jindrich, Dawyndt, Peter, Vandamme, Peter, & Carlier, Aurélien. 2019a. Introducing

SPeDE: High-Throughput Dereplication and Accurate Determination of Microbial Diversity from Matrix-Assisted Laser Desorption–Ionization Time of Flight Mass Spectrometry Data. *mSystems*, **4**(5).

Dumolin, Charles, Aerts, Maarten, Verheyde, Bart, Schellaert, Simon, Vandamme, Tim, Van der Jeugt, Felix, De Canck, Evelien, Cnockaert, Margo, Wieme, Anneleen D., Cleenwerck, Ilse, Peiren, Jindrich, Dawyndt, Peter, Vandamme, Peter, & Carlier, Aurélien. 2019b. Introducing SPeDE: High-Throughput Dereplication and Accurate Determination of Microbial Diversity from Matrix-Assisted Laser Desorption–Ionization Time of Flight Mass Spectrometry Data. *mSystems*, **4**(5).

Endesfelder, Ulrike. 2019. From single bacterial cell imaging towards in vivo single-molecule biochemistry studies. *Essays in Biochemistry*, **63**(2), 187–196.

Evans, Conor L., & Xie, X. Sunney. 2008. Coherent Anti-Stokes Raman Scattering Microscopy: Chemical Imaging for Biology and Medicine. *Annual Review of Analytical Chemistry*, **1**(1), 883–909. PMID: 20636101.

Falkowski, Paul G., & Godfrey, Linda V. 2008. Electrons, life and the evolution of Earth's oxygen cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**(1504), 2705–2716.

FAO, Food, & agriculture organization of the United Nations. 2003. *Food energy – methods of analysis and conversion factors*.

Faust, Karoline, & Raes, Jeroen. 2012. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, **10**(8), 538–550.

Feng, Ping, Gao, Ming, Burgher, Anita, Zhou, Tian, & Pramuk, Kathryn. 2016. A nine-country study of the protein content and amino acid composition of mature human milk. *Food Nutrition Research*, **60**(08).

Fouhy, Fiona, Clooney, Adam, Stanton, Catherine, Claesson, Marcus, & Cotter, Paul. 2016. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology*, **16**(06).

Gao, Dawen, Huang, Xiaoli, & Tao, Yu. 2016. A critical review of NanoSIMS in analysis of microbial metabolic activities at single-cell level. *Critical Reviews in Biotechnology*, **36**(09), 884–890.

García-Timermans, Cristina, Rubbens, Peter, Kerckhof, Frederiek Maarten, Buysschaert, Benjamin, Khalenkow, Dmitry, Waegeman, Willem, Skirtach, Andre G., & Boon, Nico. 2018. Label-free Raman characterization of bacteria calls for standardized procedures. *Journal of Microbiological Methods*, **151**(August), 69–75.

García-Timermans, Cristina, Props, Ruben, Zacchetti, Boris, Sakarika, Myrsini, Delvigne, Frank, & Boon, Nico. 2020. Raman spectroscopy-based measurements of single-cell phenotypic diversity in microbial communities. *bioRxiv*.

García-Timermans, Cristina, Rubbens, Peter, Heyse, Jasmine, Kerckhof, Frederiek-Maarten, Props, Ruben, Skirtach, Andre G., Waegeman, Willem, & Boon, Nico. 2019. Discriminating Bacterial Phenotypes at the Population and Single-Cell Level: A Comparison of Flow Cytometry and Raman Spectroscopy Fingerprinting. *Cytometry Part A*, dec, cyto.a.23952.

Gautam, Rekha, Vanga, Sandeep, Ariese, Freek, & Umapathy, Siva. 2015. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation 2015 2:1*, **2**(1), 1–38.

Georgakoudi, Irene, Tsai, Irene, Greiner, Cherry, Wong, Cheryl, DeFelice, Jordy, & Kaplan, David. 2007. Intrinsic fluorescence changes associated with the conformational state of silk fibroin in biomaterial matrices. *Optics Express*, **15**(3), 1043.

Gibb, Sebastian, & Strimmer, Korbinian. 2012. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, **28**(17), 2270–2271.

Goddard, Gregory, Martin, John C., Naivar, Mark, Goodwin, Peter M., Graves, Steven W., Habbersett, Robb, Nolan, John P., & Jett, James H. 2006. Single particle high resolution spectral analysis flow cytometry. *Cytometry Part A*, **69A**(8), 842–851.

Goodacre, Royston, Timmins, Eadaoin M, Burton, Rebecca, Kaderbhai, Naheed, Woodward, Andrew M, Kell, Douglas B, & Rooney, Paul J. 1998. Fingerprinting and artificial neural networks. *Microbiology*, **1444**(1 998), 1157–1170.

Gough, Albert, Stern, Andrew M., Maier, John, Lezon, Timothy, Shun, Tong-Ying, Chennubhotla, Chakra, Schurdak, Mark E., Haney, Steven A., & Taylor, D. Lansing. 2017. Biologically Relevant Heterogeneity: Metrics and Practical Insights. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, **22**(3), 213–237. PMID: 28231035.

Govers, Sander K., Adam, Antoine, Blockeel, Hendrik, & Aertsen, Abram. 2017. Rapid phenotypic individualization of bacterial sister cells. *Scientific Reports*, **7**(1), 8473.

Gray, Joseph V., Petsko, Gregory A., Johnston, Gerald C., Ringe, Dagmar, Singer, Richard A., & Werner-Washburne, Margaret. 2004. "Sleeping Beauty": Quiescence in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*.

Green, G. C., Chan, A. D. C., Luo, B. S., Dan, H., & Lin, M. 2009. Identification of Listeria Species Using a Low-Cost Surface-Enhanced Raman Scattering System With Wavelet-Based Signal Processing. *IEEE Transactions on Instrumentation and Measurement*, **58**(10), 3713–3722.

Grote, Jessica, Krysciak, Dagmar, Schorn, Andrea, Dahlke, Renate I., Soonvald, Liina, Müller, Johannes, Hense, Burkhard A., Schwarzfischer, Michael, Sauter, Margret, Schmeisser, Christel, & Streit, Wolfgang R. 2014. Evidence of Autoinducer-Dependent and -Independent Heterogeneous Gene Expression in Sinorhizobium fredii NGR234. *Applied and Environmental Microbiology*, **80**(18), 5572–5582.

Guo, Lizheng, Ye, Chengsong, Cui, Li, Wan, Kun, Chen, Sheng, Zhang, Shenghua, & Yu, Xin. 2019. Population and single cell metabolic activity of UV-induced VBNC bacteria determined by CTC-FCM and D2O-labeled Raman spectroscopy. *Environment International*, **130**, 104883.

Guo, Shuxia, Heinke, Ralf, Stöckel, Stephan, Rösch, Petra, Bocklitz, Thomas, & Popp, Jürgen. 2017. Towards an improvement of model transferability for Raman spectroscopy in biological applications. *Vibrational Spectroscopy*, **91**(Supplement C), 111–118.

Hakonen, A., Andersson, P. O., Stenbaek Schmidt, M., Rindzevicius, T., & Kall, M. 2015. Explosive and chemical threat detection by surface-enhanced Raman scattering: a review. *Anal Chim Acta*, **893**, 1–13.

Hanson, China A, Fuhrman, Jed A, Horner-Devine, M Claire, & Martiny, Jennifer B H. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature reviews. Microbiology*, **10**(7), 497—506.

Hardy, Ronald W, Patro, Biswamitra, Pujol-Baxley, Catherine, Marx, Christopher J, & Feinberg, Lawrence. 2018. Partial replacement of soybean meal with Methylobacterium extorquens single-cell protein in feeds for rainbow trout (Oncorhynchus mykiss Walbaum). *Aquaculture Research*, **49**(6), 2218–2224.

Harz, M., Rösch, P., Peschke, K.-D., Ronneberger, O., Burkhardt, H., & Popp, J. 2005. Micro-Raman spectroscopic identification of bacterial cells of the genus Staphylococcus and dependence on their cultivation conditions. *Analyst*, **130**, 1543–1550.

Hatzenpichler, Roland, Connon, Stephanie A., Goudeau, Danielle, Malmstrom, Rex R., Woyke, Tanja, & Orphan, Victoria J. 2016. Visualizing in situ translational activity for identifying and sorting slow-growing archaeal-bacterial consortia. *Proceedings of the National Academy of Sciences*, **113**(28), E4069–E4078.

He, Yuehui, Zhang, Peng, Huang, Shi, Wang, Tingting, Ji, Yuetong, & Xu, Jian. 2017. Label-free, simultaneous quantification of starch, protein and triacylglycerol in single microalgal cells. *Biotechnology for Biofuels*, **10**(12).

Hedegaard, Martin A.B., Matthaüs, Christian, Hassing, Søren, Krafft, Christoph, Diem, Max, & Popp, Jürgen. 2011. Spectral Unmixing and Clustering Algorithms for assessment of single cells by Raman Microscopic Imaging. *Theoretical Chemistry Accounts*, **130**(4-6), 1249–1260.

Heyse, Jasmine, Buysschaert, Benjamin, Props, Ruben, Rubbens, Peter, Skirtach, Andre G., Waegeman, Willem, & Boon, Nico. 2019. Coculturing Bacteria Leads to Reduced Phenotypic Heterogeneities. *Applied and Environmental Microbiology*, **85**(8).

Hill, M. O. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, **54**(2), 427–432.

Holyst, Herb, & Rogers, Wade. version 4.0. *FlowFP*.

Hu, Xiaona. 2020. *Microbial protein production by autotrophic nitrogen-fixing hydrogen-oxidizing bacteria.*

Huang, W. E., Li, M., Jarvis, R. M., Goodacre, R., & Banwart, S. A. 2010. Shining light on the microbial world the application of Raman microspectroscopy. *Adv Appl Microbiol*, **70**(10), 153–86.

Huang, Wei E., Stoecker, Kilian, Griffiths, Robert, Newbold, Lyndsay, Daims, Holger, Whiteley, Andrew S., & Wagner, Michael. 2007. Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. *Environmental Microbiology*, **9**(8), 1878–1889.

Huber, D., Voith von Voithenberg, L., & Kaigala, G.V. 2018. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*, **1**, 15 – 24.

Hubert, Lawrence, & Arabie, Phipps. 1985. The trasition from bargaining to a competitive market. *Journal of Classification*, **2**, 193:218.

Hug, Laura A., Baker, Brett J., Anantharaman, Karthik, Brown, Christopher T., Probst, Alexander J., Castelle, Cindy J., Butterfield, Cristina N., Hernsdorf, Alex W., Amano, Yuki, Ise, Kotaro, & et al. 2016. A new view of the tree of life. *Nature News*, Apr.

Hutsebaut, D., Vandenabeele, P., & Moens, L. 2005. Evaluation of an accurate calibration and spectral standardization procedure for Raman spectroscopy. *Analyst*, **130**(8), 1204–14.

Ivleva, Natalia P, Kubryk, Patrick, & Niessner, Reinhard. 2017. Raman microspectroscopy, surface-enhanced Raman scattering microspectroscopy, and stable-isotope Raman microspectroscopy for biofilm characterization. *Analytical and bioanalytical chemistry*, **409**(18), 4353—4375.

Jahn, Michael, Seifert, Jana, [von Bergen], Martin, Schmid, Andreas, Bühler, Bruno, & Müller, Susann. 2013. Subpopulation-proteomics in prokaryotic populations. *Current Opinion in Biotechnology*, **24**(1), 79 – 87. Analytical biotechnology.

Jarvis, Roger M., & Goodacre, Royston. 2004. Discrimination of Bacteria Using Surface-Enhanced Raman Spectroscopy. *Analytical Chemistry*, **76**(1), 40–47.

Jehlička, Jan, Edwards, Howell G. M., & Oren, Aharon. 2014. Raman Spectroscopy of Microbial Pigments. *Applied and Environmental Microbiology*, **80**(11), 3286–3295.

Jia, Kaizhi, Zhang, Yanping, & Li, Yin. 2010. Systematic engineering of microorganisms to improve alcohol tolerance. *Engineering in Life Sciences*, **10**(5), 422–429.

Jing, Xiaoyan, Gou, Honglei, Gong, Yanhai, Su, Xiaolu, Xu, La, Ji, Yuetong, Song, Yizhi, Thompson, Ian P., Xu, Jian, & Huang, Wei E. 2018. Raman-activated cell sorting and metagenomic sequencing revealing carbon-fixing bacteria in the ocean. *Environmental Microbiology*, **20**(6), 2241–2255.

Kaeberlein, T., Lewis, K., & Epstein, S. S. 2002. Isolating "Uncultivable" Microorganisms in Pure Culture in a Simulated Natural Environment. *Science*, **296**(5570), 1127–1129.

Kambhampati, Shrikaar, Li, Jia, Evans, Bradley, & Allen, Douglas. 2019. Accurate and efficient amino acid analysis for protein quantification using hydrophilic interaction chromatography coupled tandem mass spectrometry. *Plant Methods*, **15**(12).

Kampe, B., Klob, S., Bocklitz, T, Rosch, P., & Popp, J. 2017. Recursive feature elimination in Raman spectra with support vector machines. *Frontiers of Optoelectronics*, **10**(3), 272–279.

Kassambara, Alboukadel, & Mundt, Fabian. 2017. Factoextra: extract and visualize the results of multivariate data analyses. *R package version*.

Kearns, Hayleigh, Goodacre, Royston, Jamieson, Lauren E., Graham, Duncan, & Faulds, Karen. 2017. SERS Detection of Multiple Antimicrobial-Resistant Pathogens Using Nanosensors. *Analytical Chemistry*, **89**(23), 12666–12673. PMID: 28985467.

Kerchkof, Frederiek Maarten, Buysschaert, Benjamin, Khalenkow, Dmitry, & Garcia-Timermans, Cristina. 2017. MicroRaman. *https://github.ugent.be/RamanCluster/MicroRaman_package*.

Kim, Mincheol, Oh, Hyun-Seok, Park, Sang-Cheol, & Chun, Jongsik. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, **64**($\mathrm{Pt}_2), 346 - -351.$

Kniggendorf, Ann-Kathrin, Gaul, Tobias William, & Meinhardt-Wollweber, Merve. 2011. Hierarchical Cluster Analysis (HCA) of Microorganisms: An Assessment of Algorithms for Resonance Raman Spectra. *Applied Spectroscopy*, **65**(2), 165–173.

Koch, Christin, & Müller, Susann. 2018. Personalized microbiome dynamics: Cytometric fingerprints for routine diagnostics. *Molecular Aspects of Medicine*, **59**(feb), 123–134.

Koch, Christin, Harnisch, Falk, Schröder, Uwe, & Müller, Susann. 2014. Cytometric fingerprints: evaluation of new tools for analyzing microbial community dynamics. *Frontiers in microbiology*, **5**, 273.

Koeppel, Alexander, Perry, Elizabeth B., Sikorski, Johannes, Krizanc, Danny, Warner, Andrew, Ward, David M., Rooney, Alejandro P., Brambilla, Evelyne, Connor, Nora, Ratcliff, Rodney M., Nevo, Eviatar, & Cohan, Frederick M. 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences*, **105**(7), 2504–2509.

Konopka, Allan. 2009. What is microbial community ecology? *The ISME journal*, **3**(09), 1223–30.

Kopp, C., Wisztorski, M., Revel, J., Mehiri, M., Dani, V., Capron, L., Carette, D., Fournier, I., Massi, L., Mouajjah, D., Pagnotta, S., Priouzeau, F., Salzet, M., Meibom, A., & Sabourault, C. 2015. MALDI-MS and NanoSIMS imaging techniques to study cnidarian–dinoflagellate symbioses. *Zoology*, **118**(2), 125 – 131. Animal evolution: early emerging animals matter.

Kunasundari, Balakrishnan, Murugaiyah, Vikneswaran, Kaur, Gurjeet, Maurer, Frans H. J., & Sudesh, Kumar. 2013. Revisiting the Single Cell Protein Application of Cupriavidus necator H16 and Recovering Bioplastic Granules Simultaneously. *PLOS ONE*, **8**(10).

Kursa, Miron B, & Rudnicki, Witold R. 2010. Feature Selection with the Boruta Package. *Journal Of Statistical Software*, **36**(11), 1–13.

Kusić, Dragana, Kampe, Bernd, Rösch, Petra, & Popp, Jürgen. 2014. Identification of water pathogens by Raman microspectroscopy. *Water Research*, **48**, 179 – 189.

Lee, Harry L.T., Boccazzi, Paolo, Gorret, Nathalie, Ram, Rajeev J., & Sinskey, Anthony J. 2004. In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy. *Vibrational Spectroscopy*, **35**(1-2), 131–137.

Lee, Henry H., Molla, Michael N., Cantor, Charles R., & Collins, James J. 2010. Bacterial charity work leads to population-wide resistance. *Nature News*.

Lee, Kang, Palatinszky, Marton, Pereira, Fátima, Nguyen, Jen, Fernandez, Vicente, Mueller, Anna, Menolascina, Filippo, Daims, Holger, Berry, David, Wagner, Michael, & Stocker, Roman. 2019. An automated Raman-based platform for the sorting of live cells by functional properties. *Nature Microbiology*, **4**(06).

Lenz, Peter, & Søgaard-Andersen, Lotte. 2011. Temporal and spatial oscillations in bacteria. *Nature reviews. Microbiology*, **9**(08), 565–77.

Léonard, Lucie, Bouarab Chibane, Lynda, Ouled Bouhedda, Balkis, Degraeve, Pascal, & Oulahal, Nadia. 2016. Recent Advances on Multi-Parameter Flow Cytometry to Characterize Antimicrobial Treatments. *Frontiers in Microbiology*, **7**(aug), 1225.

Letunic, I., & Bork, P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*, **44**(W1), W242–5.

Levine, Jacob H., Simonds, Erin F., Bendall, Sean C., Davis, Kara L., Amir, El-Ad D., Tadmor, Michelle, Litvin, Oren, Fienberg, Harris, Jager, Astraea, Zunder, Eli, Finck, Rachel, Gedman, Amanda L., Radtke, Ina, Downing, James R., Pe'er, Dana, & Nolan, Garry P. 2015. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis HHS Public Access. *Cell*, **162**(1), 184–197.

Li, Laam, Mendis, Nilmini, Trigui, Hana, Oliver, James D., & Faucher, Sebastien P. 2014. The importance of the viable but non-culturable state in human bacterial pathogens. *Frontiers in Microbiology*, **5**, 258.

Li, Tianlun, Wu, Ting-Di, Mazéas, Laurent, Toffin, Laurent, Guerquin-Kern, Jean-Luc, Leblon, Gérard, & Bouchez, Théodore. 2008. Simultaneous analysis of microbial identity and function using NanoSIMS. *Environmental Microbiology*, **10**(3), 580–588.

Liu, Chia-Ying, Han, Yin-Yi, Shih, Po-Han, Lian, Wei-Nan, Wang, Huai-Hsien, Lin, Chi-Hung, Hsueh, Po-Ren, Wang, Juen-Kai, & Wang, Yuh-Lin. 2016. Rapid bacterial antibiotic susceptibility test based on simple surface-enhanced Raman spectroscopic biomarkers. *Scientific Reports*, **6**(1), 23375.

Liu, Hai, Zhang, Zhaoli, Liu, Sanya, Yan, Luxin, Liu, Tingting, & Zhang, Tianxu. 2015. Joint Baseline-Correction and Denoising for Raman Spectra. *Applied Spectroscopy*, **69**(9), 1013–1022. PMID: 26688879.

Lombardi, John R., & Birke, Ronald L. 2009. A Unified View of Surface-Enhanced Raman Scattering. *Accounts of Chemical Research*, **42**(6), 734–742. PMID: 19361212.

Lorenz, Björn, Wichmann, Christina, Stöckel, Stephan, Rösch, Petra, & Popp, Jürgen. 2017. Cultivation-Free Raman Spectroscopic Investigations of Bacteria. *Trends in Microbiology*, **25**(5), 413–424.

Louca, Stilianos, Parfrey, Laura Wegener, & Doebeli, Michael. 2016. Decoupling function and taxonomy in the global ocean microbiome. *Science*, **353**(6305), 1272–1277.

Lowery, Nick Vallespir, McNally, Luke, Ratcliff, William C., & Brown, Sam P. 2017. Division of Labor, Bet Hedging, and the Evolution of Mixed Biofilm Investment Strategies. *mBio*, **8**(4).

Ludwig, Joachim, Höner zu Siederdissen, Christian, Liu, Zishu, Stadler, Peter, & Mueller, Susann. 2019 (06). *flowEMMi: An automated model-based clustering tool for microbial cytometric data*.

Mair, Florian, & Prlic, Martin. 2018. OMIP-044: 28-color immunophenotyping of the human dendritic cell compartment. *Cytometry Part A*, **93A**, 402–405.

Majcherczyk, Paul A., McKenna, Therese, Moreillon, Philippe, & Vaudaux, Pierre. 2006. The discriminatory power of MALDI-TOF mass spectrometry to differentiate between isogenic teicoplanin-susceptible and teicoplanin-resistant strains of methicillin-resistant Staphylococcus aureus. *FEMS Microbiology Letters*, **255**(2), 233–239.

Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Maquelin, K., Kirschner, C., Choo-Smith, L.-P., Ngo-Thi, N. A., van Vreeswijk, T., Stämmler, M., Endtz, H. P., Bruining, H. A., Naumann, D., & Puppels, G. J. 2003. Prospective Study of the Performance of Vibrational Spectroscopies for Rapid Identification of Bacterial and Fungal Pathogens Recovered from Blood Cultures. *Journal of Clinical Microbiology*, **41**(1), 324–329.

Maquelin, Kees, Dijkshoorn, Lenie, van der Reijden, Tanny J.K., & Puppels, Gerwin J. 2006. Rapid epidemiological analysis of Acinetobacter strains by Raman spectroscopy. *Journal of Microbiological Methods*, **64**(1), 126–131.

Martin, Marivic, Dragoš, Anna, Otto, Simon, Schäfer, Daniel, Brix, Susanne, Maróti, Gergely, & Kovács, Ákos. 2020. Cheaters shape the evolution of phenotypic heterogeneity in Bacillus subtilis biofilms. *The ISME Journal*, 06.

Martins, Bruno MC, & Locke, James CW. 2015. Microbial individuality: how single-cell heterogeneity enables population level strategies. *Current Opinion in Microbiology*, **24**, 104 – 112. Cell regulation.

Masters, Barry R. 2009 (06). *C.V. Raman and the Raman Effect*.

Matassa, Silvio, Boon, Nico, & Verstraete, Willy. 2015. Resource recovery from used water: The manufacturing abilities of hydrogen-oxidizing bacteria. *Water Research*, **68**, 467 – 478.

Matassa, Silvio, Boon, Nico, Pikaar, Ilje, & Verstraete, Willy. 2016. Microbial protein: future sustainable food supply route with low environmental footprint. *Microbial Biotechnology*, **9**(5), 568–575.

McFall-Ngai, Margaret, Hadfield, Michael G., Bosch, Thomas C. G., Carey, Hannah V., Domazet-Lošo, Tomislav, Douglas, Angela E., Dubilier, Nicole, Eberl, Gerard, Fukami, Tadashi, Gilbert, Scott F., Hentschel, Ute, King, Nicole, Kjelleberg, Staffan, Knoll, Andrew H., Kremer, Natacha, Mazmanian, Sarkis K., Metcalf, Jessica L., Nealson, Kenneth, Pierce, Naomi E., Rawls, John F., Reid, Ann, Ruby, Edward G., Rumpho, Mary, Sanders, Jon G., Tautz, Diethard, & Wernegreen, Jennifer J. 2013. Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences*, **110**(9), 3229–3236.

McIlvenna, David, Huang, Wei E., Davison, Paul, Glidle, Andrew, Cooper, Jon, & Yin, Huabing. 2016. Continuous cell sorting in a flow based on single cell resonance Raman spectra. *Lab on a Chip*, **16**(8), 1420–1429.

Milo, Ron, Jorgensen, Paul, Moran, Uri, Weber, Griffin, & Springer, Michael. 2009. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, **38**(suppl$_1$), $D750--D753$.

Mizrahi-Man, Orna, Davenport, Emily R., & Gilad, Yoav. 2013. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLOS ONE*, **8**(1), 1–14.

Mosier-Boss, Pamela. 2017. Review on SERS of Bacteria. *Biosensors*, **7**(4), 51.

Muhamadali, Howbeer, Chisanga, Malama, Subaihi, Abdu, & Goodacre, Royston. 2015. Combining Raman and FT-IR Spectroscopy with Quantitative Isotopic Labeling for Differentiation of E. coli Cells at Community and Single Cell Levels. *Analytical Chemistry*, **87**(8), 4578–4586. PMID: 25831066.

Murrell, Adele, Rakyan, Vardhman K., & Beck, Stephan. 2005. From genome to epigenome. *Human Molecular Genetics*, **14**(suppl$_1$), $R3--R10$.

Musat, Niculina, Musat, Florin, Weber, Peter Kilian, & Pett-Ridge, Jennifer. 2016. Tracking microbial interactions with NanoSIMS. *Current Opinion in Biotechnology*, **41**, 114 – 121. Analytical biotechnology.

Méheust, Raphaël, Burstein, David, Castelle, Cindy J., & Banfield, Jillian F. 2019. The distinction of CPR bacteria from other bacteria based on protein family content. *Nature News*, Sep.

Müllner, Daniel. 2013. fastcluster : Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, **53**(9).

Nasseri, A.T, Rasoul-Amini, Sara, Morowvat, Mohammad Hossein, & Younes, Ghasemi. 2011. Single Cell Protein: Production and Process. *American Journal of Food Technology*, **6**(02).

Naumann, Dieter, Helm, Dieter, & Labischinski, Harald. 1991. Microbiological characterizations by FT-IR spectroscopy. *Nature*, **351**(6321), 81–82.

Nebe-von Caron, G, Stephens, P.J, Hewitt, C.J, Powell, J.R, & Badley, R.A. 2000. Analysis of bacterial function by multi-colour fluorescence flow cytometry and single cell sorting. *Journal of Microbiological Methods*, **42**(1), 97–114.

Neugebauer, U., Rosch, P., & Popp, J. 2015. Raman spectroscopy towards clinical application: drug monitoring and pathogen identification. *Int J Antimicrob Agents*, **46 Suppl 1**, S35–9.

Nie, Lei, Wu, Gang, & Zhang, Weiwen. 2006. Correlation of mRNA Expression and Protein Abundance Affected by Multiple Sequence Features Related to Translational Efficiency in Desulfovibrio vulgaris: A Quantitative Analysis. *Genetics*, **174**(4), 2229–2243.

Nijkamp, Jurgen, Broek, Marcel, Datema, Erwin, Kok, Stefan, Bosman, Lizanne, Luttik, Marijke, Daran-Lapujade, Pascale, Vongsangnak, Wanwipa, Nielsen, Jens, Heijne, Wilbert, Klaassen, Paul, Paddon, Christopher, Platt, Darren, Kötter, Peter, Ham, Roeland, Reinders, Marcel, Pronk, Jack, Ridder, Dick, & Daran, Jean-Marc. 2012. De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology. *Microbial cell factories*, **11**(03), 36.

Noble, James, Knight, Alexander, Reason, AJ, Matola, A, & Bailey, M. 2007. A Comparison of Protein Quantitation Assays for Biopharmaceutical Applications. *Molecular biotechnology*, **37**(11), 99–111.

Noble, James E. 2014. Chapter Two - Quantification of Protein Concentration Using UV Absorbance and Coomassie Dyes. *Pages 17 – 26 of:* Lorsch, Jon (ed), *Laboratory Methods in Enzymology: Protein Part A*. Methods in Enzymology, vol. 536. Academic Press.

Novelli-Rousseau, A., Espagnon, I., Filiputti, D., Gal, O., Douet, A., Mallard, F., & Josso, Q. 2018. Culture-free Antibiotic-susceptibility Determination From Single-bacterium Raman Spectra. *Scientific Reports*, **8**(1), 3957.

Nuñez, Jamie, Renslow, Ryan, Cliff, John B., & Anderton, Christopher R. 2018. NanoSIMS for biological applications: Current practices and analyses. *Biointerphases*, **13**(3), 03B301.

O'Neill, Kieran, Aghaeepour, Nima, Špidlen, Josef, & Brinkman, Ryan. 2013. Flow Cytometry Bioinformatics. *PLoS Computational Biology*, **9**(12).

Pahlow, Susanne, Meisel, Susann, Cialla-May, Dana, Weber, Karina, Rösch, Petra, & Popp, Jürgen. 2015. Isolation and identification of bacteria by means of Raman spectroscopy. *Advanced Drug Delivery Reviews*, **89**(jul), 105–120.

Palmer, Michael W. 2008. *Ordination Methods - an Overview*.

Pearman, William F., & Fountain, Augustus W. 2006. Classification of Chemical and Biological Warfare Agent Simulants by Surface-Enhanced Raman Spectroscopy and Multivariate Statistical Techniques. *Applied Spectroscopy*, **60**(4), 356–365. PMID: 16613630.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, & Duchesnay, Édouard. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**(Oct), 2825–2830.

Perfetto, Stephen P., Chattopadhyay, Pratip K., & Roederer, Mario. 2004. Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, **4**(8), 648–655.

Pilot, Signorini, Durante, Orian, Bhamidipati, & Fabris. 2019. A Review on Surface-Enhanced Raman Scattering. *Biosensors*, **9**(2), 57.

Pinhal, Stéphane, Ropers, Delphine, Geiselmann, Johannes, & de Jong, Hidde. 2019. Acetate Metabolism and the Inhibition of Bacterial Growth by Acetate. *Journal of Bacteriology*, **201**(13).

Porter, J., Edwards, C., & Pickup, R.W. 1995. Rapid assessment of physiological status in *Escherichia coli* using fluorescent probes. *Journal of Applied Bacteriology*, **79**(4), 399–408.

Props, Ruben, Monsieurs, Pieter, Mysara, Mohamed, Clement, Lieven, & Boon, Nico. 2016. Measuring the biodiversity of microbial communities by flow cytometry. *Methods in Ecology and Evolution*, **7**(11), 1376–1385.

Quast, Christian, Pruesse, Elmar, Yilmaz, Pelin, Gerken, Jan, Schweer, Timmy, Yarza, Pablo, Peplies, Jörg, & Glöckner, Frank Oliver. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**(D1), D590–D596.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*.

R Core Team 3.6.2. 2019. *R: A Language and Environment for Statistical Computing*.

Raman, C. V., & Krishnan, K. S. 1928. A New Type of Secondary Radiation. *Nature*, **121**, 501–502.

Ramette, Alban. 2007. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, **62**(2), 142–160.

Read, D. S., & Whiteley, A. S. 2015. Chemical fixation methods for Raman spectroscopy-based analysis of bacteria. *J Microbiol Methods*, **109**(12), 79–83.

Ren, Yan, Ji, Yuetong, Teng, Lin, & Zhang, Heping. 2017. Using Raman spectroscopy and chemometrics to identify the growth phase of Lactobacillus casei Zhang during batch culture at the single-cell level. *Microbial Cell Factories*, **16**(1), 233.

Rim, John Hoon, Lee, Yangsoon, Hong, Sung Kuk, Park, Yongjung, Park, Yongjung, Kim, Myung-Sook, D'Souza, Roshan, Park, Eun Suk, Yong, Dongeun, & Lee, Kyungwon. 2015. Insufficient Discriminatory Power of Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry Dendrograms to Determine the Clonality of Multi-Drug-Resistant *Acinetobacter baumannii* Isolates from an Intensive Care Unit. *BioMed Research International*, **2015**, 535027.

Roberts, F. S. 2019. Measurement of Biodiversity: Richness and Evenness. *Springer, Cham*, **5**.

Rodriguez, J. D., Westenberger, B. J., Buhse, L. F., & Kauffman, J. F. 2011. Standardization of Raman spectra for transfer of spectral libraries across different instruments. *Analyst*, **136**(20), 4232–40.

Roger, Fabian, Bertilsson, Stefan, Langenheder, Silke, Osman, Omneya Ahmed, & Gamfeldt, Lars. 2016. Effects of multiple dimensions of bacterial diversity on functioning, stability and multifunctionality. *Ecology*, **97**(10), 2716–2728.

Ron, Eliora Z. 2013. Bacterial stress response. *In: The Prokaryotes: Prokaryotic Physiology and Biochemistry*. Springer.

RStudio. 2016. *RStudio: Integrated development environment for R*.

RStudio. 2019. *RStudio: Integrated Development for R*.

Rull, Fernando, & Martínez-Frías, Jesús. 2006. Raman spectroscopy goes to Mars. *Spectroscopy Europe*, **18**(01).

Rutherfurd, Shane M., Schneuwly, Audrey, & Moughan, Paul J. 2007. Analyzing Sulfur Amino Acids in Selected Feedstuffs Using Least-Squares Nonlinear Regression. *Journal of Agricultural and Food Chemistry*, **55**(20), 8019–8024. PMID: 17715935.

Ryabchykov, Oleg, Guo, Shuxia, & Bocklitz, Thomas. 2018. Analyzing Raman spectroscopic data. *Physical Sciences Reviews*, **0**(0), 1–16.

Saeys, Yvan, Gassen, Sofie Van, & Lambrecht, Bart N. 2016. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, **16**(7), 449–462.

Samek, Ota, Obruča, Stanislav, Šiler, Martin, Sedláček, Petr, Benešová, Pavla, Kučera, Dan, Márova, Ivana, Ježek, Jan, Bernatová, Silva, & Zemánek, Pavel. 2016. Quantitative Raman Spectroscopy Analysis of Polyhydroxyalkanoates Produced by Cupriavidus necator H16. *Sensors (Basel, Switzerland)*, **16**(11).

Sanchez-Romero, M. A., & Casadesus, J. 2014. Contribution of phenotypic heterogeneity to adaptive antibiotic resistance. *Proceedings of the National Academy of Sciences*, **111**(1), 355–360.

Santos, André Luis Souza dos, Galdino, Anna Clara Milesi, Mello, Thais Pereira de, Ramos, Livia de Souza, Branquinha, Marta Helena, Bolognese, Ana Maria, Columbano Neto, José, & Roudbary, Maryam. 2018. What are the advantages of living in a community? A microbial biofilm perspective! *Memorias do Instituto Oswaldo Cruz*, **113**(00).

Sapan, Christine V., Lundblad, Roger L., & Price, Nicholas C. 1999. Colorimetric protein assay techniques. *Biotechnology and Applied Biochemistry*, **29**(2), 99–108.

Schirawski, Jan, & Perlin, Michael. 2018. Plant–Microbe Interaction 2017—The Good, the Bad and the Diverse. *International Journal of Molecular Sciences*, **19**(5), 1374.

Schmid, Thomas, & Dariz, Petra. 2019. Raman Microspectroscopic Imaging of Binder Remnants in Historical Mortars Reveals Processing Conditions. *Heritage*, **2**(2), 1662–1683.

Schmidt, Thomas, Rodrigues, João, & von Mering, Christian. 2016. A family of interaction-adjusted indices of community similarity. *The ISME Journal*, **11**(12).

Schulmerich, Matthew V., Walsh, Michael J., Gelber, Matthew K., Kong, Rong, Kole, Matthew R., Harrison, Sandra K., McKinney, John, Thompson, Dennis, Kull, Linda S., & Bhargava, Rohit. 2012. Protein and Oil Composition Predictions of Single Soybeans by Transmission Raman Spectroscopy. *Journal of Agricultural and Food Chemistry*, **60**(33), 8097–8102. PMID: 22746340.

Scientific, Horiba. 1996-2020. *What is the difference between dispersive Raman and FT-Raman?*

Sela, Itamar, Wolf, Yuri I., & Koonin, Eugene V. 2020. Horizontal gene transfer barrier shapes the evolution of prokaryotic pangenomes. *bioRxiv*.

Seng, Piseth, Drancourt, Michel, Gouriet, Frédérique, La Scola, Bernard, Fournier, Pierre-Edouard, Rolain, Jean Marc, & Raoult, Didier. 2009. Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry. *Clinical Infectious Diseases*, **49**(4), 543–551.

Shade, Ashley. 2016. Diversity is the question, not the answer. *Peer J*, 01.

Sillman, Jani, Nygren, Lauri, Kahiluoto, Helena, Ruuskanen, Vesa, Tamminen, Anu, Bajamundi, Cyril, Nappa, Marja, Wuokko, Mikko, Lindh, Tuomo, Vainikka, Pasi, Pitkänen, Juha-Pekka, & Ahola, Jero. 2019. Bacterial protein for food and feed generated via renewable energy and direct air capture of CO2: Can it reduce land and water use? *Global Food Security*, **22**, 25 – 32.

Singh, R., & Riess, F. 2001. The 1930 Nobel Prize for Physics: A close decision? *Notes and Records of the Royal Society of London*, **55**(2), 267–283.

Singhal, Neelja, Kumar, Manish, Kanaujia, Pawan K., & Virdi, Jugsharan S. 2015. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology*, **6**, 791.

Sivakesava, Sakhamuri, Irudayaraj, Joseph, & Ali, Demirci. 2001. Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques. *Process Biochemistry*, **37**(4), 371 – 378.

Sjöberg, Béatrice, Foley, Sarah, Cardey, Bruno, & Enescu, Mironel. 2014. An experimental and theoretical study of the amino acid side chain Raman bands in proteins. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **128**(07), 300–311.

Smits, Wiep Klaas, Kuipers, Oscar, & Veening, Jan-Willem. 2006. Phenotypic variation in bacteria: The role of feedback regulation. *Nature reviews. Microbiology*, **4**(05), 259–71.

Song, Yizhi, Yin, Huabing, & Huang, Wei. 2016. Raman activated cell sorting. *Current opinion in chemical biology*, **33**(04), 1–8.

Spitzer, Matthew H., & Nolan, Garry P. 2016. Mass Cytometry: Single Cells, Many Features. *Cell*, **165**(4), 780–791.

Sprouffske, Kathleen, & Wagner, Andreas. 2016. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics*, **17**(1), 172.

Steen, Harald B., & Boye, Erik. 1980. Bacterial growth studied by flow cytometry. *Cytometry*, **1**(1), 32–36.

Stewart, Mary K, Cummings, Lisa A, Johnson, Matthew L, Berezow, Alex B, & Cookson, Brad T. 2011. Regulation of phenotypic heterogeneity permits Salmonella evasion of the host caspase-1 inflammatory response. *Proceedings of the National Academy of Sciences*, **108**(51), 20742–20747.

Stewart, Philip, & Franklin, Michael. 2008. Physiological Heterogeneity in Biofilms. *Nature reviews. Microbiology*, **6**(04), 199–210.

Strola, Samy Andrea, Baritaux, Jean-Charles, Schultz, Emmanuelle, Simon, Anne Catherine, Allier, Cédric, Espagnon, Isabelle, Jary, Dorothée, & Dinten, Jean-Marc. 2014. Single bacteria identification by Raman spectroscopy. *Journal of Biomedical Optics*, **19**(11), 111610.

Sumner, Edward R, & Avery, Simon V. 2002. Phenotypic heterogeneity: differential stress resistance among individual cells of the yeast Saccharomyces cerevisiae. *Microbiology*, **148**(2), 345–351.

Święciło, Agata. 2016. Cross-stress resistance in *Saccharomyces cerevisiae* yeast—new insight into an old phenomenon. *Cell Stress and Chaperones*.

Tadrowski, Andrew, Evans, Martin, & Waclaw, Bartlomiej. 2018. Phenotypic Switching Can Speed up Microbial Evolution. *Scientific Reports*, **8**(12).

Takors, Ralf, Kopf, Michael, Mampel, Joerg, Bluemke, Wilfried, Blombach, Bastian, Eikmanns, Bernhard, Bengelsdorf, Frank R., Weuster-Botz, Dirk, & Dürre, Peter. 2018. Using gas mixtures of CO, CO2 and H2 as microbial substrates: the do's and don'ts of successful technology transfer from laboratory to production scale. *Microbial Biotechnology*, **11**(4), 606–625.

Tang, Fuchou, Lao, Kaiqin, & Surani, M Azim. 2011. Development and applications of single-cell transcriptome analysis. *Nature Methods*, **8**(4), S6–S11.

Tang, Mingjie, McEwen, Gerald D., Wu, Yangzhe, Miller, Charles D., & Zhou, Anhong. 2013. Characterization and analysis of mycobacteria and Gram-negative bacteria and co-culture mixtures by Raman microspectroscopy, FTIR, and atomic force microscopy. *Analytical and Bioanalytical Chemistry*, **405**(5), 1577–1591.

Taniguchi, Yuichi, Choi, Paul J, Li, G.-W., Chen, Huiyi, Babu, Mohan, Hearn, Jeremy, Emili, Andrew, & Xie, X Sunney. 2010. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science*, **329**(5991), 533–538.

Tanniche, Imen, Collakova, Eva, Denbow, Cynthia, & Senger, Ryan. 2020. Characterizing metabolic stress-induced phenotypes of Synechocystis PCC6803 with Raman spectroscopy. *PeerJ*, **8**(03), e8535.

Taylor, Gordon T., Suter, Elizabeth A., Li, Zhuo Q., Chow, Stephanie, Stinton, Dallyce, Zaliznyak, Tatiana, & Beaupré, Steven R. 2017. Single-Cell Growth Rates in Photoautotrophic Populations Measured by Stable Isotope Probing and Resonance Raman Microspectrometry. *Frontiers in Microbiology*, **8**(1449).

Team, R Core. 2015. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Teng, Lin, Wang, Xian, Wang, Xiaojun, Gou, Honglei, Ren, Lihui, Wang, Tingting, Wang, Yun, Ji, Yuetong, Huang, Wei E., & Xu, Jian. 2016. Label-free, rapid and quantitative phenotyping of stress response in E. coli via ramanome. *Scientific Reports*, **6**(1), 34359.

Tian, Yao, & Burch, Kenneth S. 2016. Automatic Spike Removal Algorithm for Raman Spectra. *Appl. Spectrosc.*, **70**(11), 1861–1871.

Tomoyori, K, Hirano, Y, Kurihara, Kazuo, & Tamada, Taro. 2015. Background elimination using the SNIP algorithm for Bragg reflections from a protein crystal measured by a TOF singlecrystal neutron diffractometer. *Journal of Physics: Conference Series*, **664**(12), 072049.

Tuschel, David. 2016 (03). *Selecting an Excitation Wavelength for Raman Spectroscopy*.

Uckert, Kyle, & Michel, John. 2019. A Semi-Autonomous Method to Detect Cosmic Rays in Raman Hyperspectral Data Sets. *Applied Spectroscopy*, **73**(07), 000370281985058.

Ugalde, U.O., & Castrillo, J.I. 2002. Single cell proteins from fungi and yeasts. *Pages 123 – 149 of: Agriculture and Food Production*. Applied Mycology and Biotechnology, vol. 2. Elsevier.

van Belkum A, Welker M, Pincus D Charrier JP Girard V. 2017. Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry in Clinical Microbiology: What Are the Current Issues? *Critical Reviews in Biotechnology*, **37**(06), 475:483.

van de Vossenberg, Jack, Tervahauta, Heli, Maquelin, Kees, Blokker-Koopmans, Carola H. W., Uytewaal-Aarts, Marijan, van der Kooij, Dick, van Wezel, Annemarie P., & van der Gaag, Bram. 2013. Identification of bacteria in drinking water with Raman spectroscopy. *Anal. Methods*, **5**, 2679–2687.

Van Der Maaten, L. J. P., & Hinton, G. E. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, **9**, 2579–2605.

Van Nevel, S., Koetzsch, S., Weilenmann, H.U., Boon, N., & F., Hammes. 2013. Routine bacterial analysis with automated flow cytometry. *Journal of Microbiological Methods*, **94**(2), 73–76.

Van Rossum, Thea, Ferretti, Pamela, Maistrenko, Oleksandr, & Bork, Peer. 2020. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology*, 06, 1–16.

van Veen, S. Q., Claas, E. C. J., & Kuijper, Ed J. 2010. High-Throughput Identification of Bacteria and Yeast by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry in Conventional Medical Microbiology Laboratories. *Journal of Clinical Microbiology*, **48**(3), 900–907.

Vandamme, Peter. 2015. Taxonomy and classification of bacteria. *Pages 255–269 of:* Jorgensen, James H, Pfaller, Michael A, Carroll, Karen C, Funke, Guido, Landry, Marie Louise, Richter, Sandra S, & Warnock, David W (eds), *Manual of clinical microbiology*. ASM Press.

Vartoukian, Sonia R., Palmer, Richard M., & Wade, William G. 2010. Strategies for culture of 'unculturable' bacteria. *FEMS Microbiology Letters*, **309**(1), 1–7.

Veening, Jan-Willem, Smits, Wiep Klaas, & Kuipers, Oscar. 2008. Bistability, Epigenetics, and Bet-Hedging in Bacteria. *Annual review of microbiology*, **62**(07), 193–210.

Veraart, Annelies, Garbeva, Paolina, Beersum, F., Ho, Adrian, Hordijk, Kees, Meima-Franke, Marion, Zweers, A.j, & Bodelier, Paul. 2018. Living apart together—bacterial volatiles influence methanotrophic growth and activity. *The ISME Journal*, **12**(01).

Verduyn, Cornelis, Postma, Erik, Scheffers, W. Alexander, & Van Dijken, Johannes P. 1992. Effect of benzoic acid on metabolic fluxes in yeasts: A continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast*, **8**(7), 501–517.

Vernikos, George, Medini, Duccio, Riley, David R, & Tettelin, Hervé. 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology*, **23**, 148 – 154. Host–microbe interactions: bacteria • Genomics.

Verstraete, W., Wittebolle, L., Heylen, K., Vanparys, B., de Vos, P., van de Wiele, T., & Boon, N. 2007. Microbial Resource Management: The Road To Go for Environmental Biotechnology. *Engineering in Life Sciences*, **7**(2), 117–126.

Vila, Eugenia, Hornero-Méndez, Dámaso, Azziz, Gastón, Lareo, Claudia, & Saravia, Verónica. 2019. Carotenoids from heterotrophic bacteria isolated from Fildes Peninsula, King George Island, Antarctica. *Biotechnology Reports*, **21**, e00306.

Villanueva, Randle Aaron M., Chen, Zhuo Job, & Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis Using the Grammar of Graphics*. Springer-Verlag New York.

Virta, Marko, Lineri, Sanna, Kankaanpää, Pasi, Karp, Matti, Peltonen, Karita, Nuutila, Jari, & Lilius, Esa-Matti. 1998. Determination of Complement-Mediated Killing of Bacteria by Viability Staining and Bioluminescence. *Applied and Environmental Microbiology*, **64**(2), 515–519.

Vlamakis, H., Aguilar, C., Losick, R., & Kolter, R. 2008. Phenotypic variation in bacteria: The role of feedback regulation. *Genes Dev.*, **22**(7), 945–53.

Volova, T, & Barashkov, V. 2010. Characteristics of Proteins Synthesized by Hydrogen-Oxidizing Microorganisms. *Prikladnaia biokhimiia i mikrobiologiia*, **46**(11), 624–9.

Wade, William. 2002. Unculturable Bacteria—The Uncharacterized organisms that Cause Oral Infections. *Journal of the Royal Society of Medicine*, **95**(2), 81–83. PMID: 11823550.

Wahl, Joel, Sjödahl, Mikael, & Ramser, Kerstin. 2020. Single-Step Preprocessing of Raman Spectra Using Convolutional Neural Networks. *Applied spectroscopy*, **74**(4), 427–438.

Wan, Katty X., Vidavsky, Ilan, & Gross, Michael L. 2002. Comparing similar spectra: From similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, **13**(1), 85–88.

Wang, Daojing, & Bodovitz, Steven. 2010. Single cell analysis: the new frontier in 'Omics'. *Trends in Biotechnology*, **28**(6), 281–290.

Wang, Wan-Hui, Feng, Xiujuan, & Bao, Ming. 2018. *Transformation of CO2 to Methanol with Homogeneous Catalysts*. Springer. Page Chapter 6.

Wang, Yun, Ji, Yuetong, Wharfe, Emma S., Meadows, Roger S., March, Peter, Goodacre, Royston, Xu, Jian, & Huang, Wei E. 2013. Raman Activated Cell Ejection for Isolation of Single Cells. *Analytical Chemistry*, **85**(22), 10697–10701. PMID: 24083399.

Wang, Yun, Huang, Wei E, Cui, Li, & Wagner, Michael. 2016. Single cell stable isotope probing in microbiology using Raman microspectroscopy. *Current Opinion in Biotechnology*, **41**, 34 – 42. Analytical biotechnology.

Ward, Joe. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **58**(03), 236–244.

Watson, Dakota A., Brown, Leif O., Gaskill, Daniel F., Naivar, Mark, Graves, Steven W., Doorn, Stephen K., & Nolan, John P. 2008. A flow cytometer for the measurement of Raman spectra. *Cytometry Part A*, **73A**(2), 119–128.

Wayne, L.G., Brenner, D.J., Colwell, Rita, Grimont, Patrick, Krichevsky, Micah, Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, Erko, Starr, M.P., & Truper, H.G. 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic Bacteriology*, **37**(10).

Wehrli, Patrick M., Lindberg, Erika, Svensson, Olof, Sparén, Anders, Josefson, Mats, Dunstan, R. Hugh, Wold, Agnes E., & Gottfries, Johan. 2014. Exploring bacterial phenotypic diversity using factorial design and FTIR multivariate fingerprinting. *Journal of Chemometrics*, **28**(8), S681–S686.

Wesche, Alissa M., Gurtler, Joshua B., Marks, Bradley P., & Ryser, Elliot T. 2009. *Stress, sublethal injury, resuscitation, and virulence of bacterial foodborne pathogens*.

Whelan, Jessica, Craven, Stephen, & Glennon, Brian. 2012. In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnology Progress*, **28**(5), 1355–1362.

WHO, World Health Organization. 2017. *Protein and amino acid requirements in human nutrition Report of a joint FAO/WHO/UNU expert consultation*.

Willemse-Erix, Diana F. M., Scholtes-Timmerman, Maarten J., Jachtenberg, Jan-Willem, van Leeuwen, Willem B., Horst-Kreft, Deborah, Bakker Schut, Tom C., Deurenberg, Ruud H., Puppels, Gerwin J., van Belkum, Alex, Vos, Margreet C., & Maquelin, Kees. 2009. Optical Fingerprinting in Bacterial Epidemiology: Raman Spectroscopy as a Real-Time Typing Method. *Journal of Clinical Microbiology*, **47**(3), 652–659.

Willig, Michael R. 2011. Biodiversity and Productivity. *Science*, **333**(6050), 1709–1710.

Wu, Huawen, Volponi, Joanne V., Oliver, Ann E., Parikh, Atul N., Simmons, Blake A., & Singh, Seema. 2011. In vivo lipidomics using single-cell Raman spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(9), 3809–3814.

Yan, Jinyong, Han, Bingnan, Gui, Xiao, Guilong, Wang, Yunjun, Yan, Madzak, Catherine, Pan, Dujie, Wang, Yaofeng, Zha, Genhan, & Jiao, Liangcheng. 2018. Engineering Yarrowia lipolytica to Simultaneously Produce Lipase and Single Cell Protein from Agro-industrial Wastes for Feed. *Scientific Reports*, **8**(12).

Yao, Zhizhong, Davis, Rebecca M, Kishony, Roy, Kahne, Daniel, & Ruiz, Natividad. 2012. Regulation of cell size in response to nutrient availability by fatty acid biosynthesis in Escherichia coli. *Proceedings of the National Academy of Sciences*, **109**(38), E2561–E2568.

Yu, Zheng, Krause, Sascha M. B., Beck, David A. C., & Chistoserdova, Ludmila. 2016. A Synthetic Ecology Perspective: How Well Does Behavior of Model Organisms in the Laboratory Predict Microbial Activities in Natural Habitats? *Frontiers in Microbiology*, **7**, 946.

Yust, Maris, Pedroche, Justo, Girón-Calle, Julio, Vioque, Javier, Millán, Franciscp, & Alaiz, Manuel. 2004. Determination of tryptophan by high-performance liquid chromatography of alkaline hydrolysates with spectrophotometric detection. *Food Chemistry*, **85**(2), 317 – 320.

Zeleny, David. 2020. *Analysis of community ecology data in R*.

Zhang, Peiran, Ren, Lihui, Zhang, Xu, Shan, Yufei, Wang, Yun, Ji, Yuetong, Yin, Huabing, Huang, Wei E., Xu, Jian, & Ma, Bo. 2015. Raman-Activated Cell Sorting Based on Dielectrophoretic Single-Cell Trap and Release. *Analytical Chemistry*, **87**(4), 2282–2289.

Zhang, Yi, Gao, Jiaxin, Huang, Yanyi, & Wang, Jianbin. 2018. Recent Developments in Single-Cell RNA-Seq of Microorganisms. *Biophysical Journal*, **115**(2), 173 – 180.

Zhao, Kelei, Liu, Linjie, Chen, Xiaojie, Huang, Ting, Du, Lianming, Lin, Jiafu, Yuan, Yang, Zhou, Yingshun, Yue, Bisong, Wei, Kun, & Chu, Yiwen. 2019. Behavioral heterogeneity in quorum sensing can stabilize social cooperation in microbial populations. *BMC Biology*, **17**(12).

Zhou, Jizhong, & Ning, Daliang. 2017. Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiology and Molecular Biology Reviews*, **81**(4).

Zhu, Guangyong, Zhu, Xian, Fan, Qi, & Wan, Xueliang. 2011. Raman spectra of amino acids and their aqueous solutions. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **78**(3), 1187 – 1195.

Zid, Brian M., & O'Shea, Erin K. 2014. Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature*, **514**(7520), 117–121.

Zu, Theresah N. K., Athamneh, Ahmad I. M., Wallace, Robert S., Collakova, Eva, & Senger, Ryan S. 2014. Near-Real-Time Analysis of the Phenotypic Responses of Escherichia coli to 1-Butanol Exposure Using Raman Spectroscopy. *Journal of Bacteriology*, **196**(23), 3983–3991.

Curriculum vitae

CRISTINA GARCÍA TIMERMANS

crisgtimermans@gmail.com | Cristina-GT | cristinagarciatimermans | @CGTimermans

Education

PhD candidate in Bioscience Engineering

Ghent University, Belgium

Center for Microbial Ecology and Technology (CMET)

October 2016 - October 2020

- Title: Raman spectroscopy for single-cell microbial phenotyping
- Developed tools to study and quantify diversity and stress in microbial communities using Raman spectroscopy. These tools were applied for quality control in microbial protein production and to develop bioremediation strategies to alleviate deep-sea oil spills
- Keywords: single-cell, flow cytometry, Raman spectroscopy, microbial ecology

Master 'Interdisciplinary Approaches to Life Sciences'

September 2013 - June 2015

Centre des Recherches Interdisciplinaires (CRI)

Paris, France

Bachelor's degree in Medical Biochemistry

September 2009 - May 2013

University of Navarra

Pamplona, Spain

Experience

Graduate Researcher in Medical Sciences

Ghent University, Belgium

HIV lab - Dr Verhasselt's team

October 2015 - May 2016

- Cell culture and molecular biology techniques
- Work in laboratory of biosafety level 3

Startup co-founder

L'Open Lab, Paris, France

Eco-Smart Solutions

March 2015 - July 2015

- Green-tech that studied the microbiome of the metro of Paris and the use of probiotics for cleaning
- Management: Project set up, search for funding and budget management
- Scientific tasks: lab set-up, DNA sequencing and data analysis
- Communication: Pitch presentation in La Sorbonne, L'Open Lab and BIG BPI start-up showcast

Internship: mitochondrial dynamics in *Schizosaccharomyces pombe*

Paris, France

Dr Phong's team - Institut Curie

November 2014 - Mars 2015

- Microfluidics and imaging techniques

Internship: iGEM

Paris, France

iGEM Paris Bettencourt 2014

June 2014 - November 2014

- Participation in the iGEM (International Genetically Engineered Machine competition)
- Project: "The Smell of Us". Design of a probiotic cream using CRISPRs technology
- Winners of Best New Application and Best Arts and Design. Gold Medal award

Internship: Nicotinic addiction model

Paris, France

Group for Neural Theory - Ecole Normale Supérieure

February 2014 - March 2014

- Computer modelling and systems biology.

Languages

Spanish Native speaker
English Advanced
French Advanced

Skills

R, Matlab, LaTeX

Molecular techniques, flow cytometry, Raman spectroscopy, microbial ecology

Main publications

- **García-Timermans, C.**, Props, R., Zacchetti,B., Sakarika, M., Delvigne,F. , and Boon, N. Raman spectroscopy-based measurements of single-cell phenotypic diversity in microbial populations (in revision) BioRxiv https://doi.org/10.1101/2020.05.21.109934
- **García-Timermans, C.**, Rubbens,P.,Heyse, J.,Kerckhof, F.-M.,Props,R., Skirtach ,A.,Waegeman, W.,& Boon,N.(2019). Discriminating bacterial phenotypes at the population and single-cell level: A comparison of flow cytometry and Raman spectroscopy fingerprinting. CYTOMETRY PART A.http://dx.doi.org/https://doi.org/10.1002/cyto.a.23952
- **García-Timermans, C.**, Rubbens,P., Kerckhof,F.-M., Buysschaert,B., Khalenkow,D., Waegeman,W., Skir-tach,A.,& Boon,N.(2018).Label-free Raman characterization of bacteria calls for standardized procedures.JOURNAL OF MICROBIOLOGICAL METHODS,151,69–75.http://dx.doi.org/10.1016/j.mimet.2018.05.027
- Rubbens, P., Props, R., **García-Timermans, C.**, Boon, N.,& Waegeman, W.(2017).Stripping flow cytometry: How many detectors do we need for bacterial identification. CYTOMETRY PART A,91(12),1184–1191.http://dx.doi.org/10.1002/cyto.a.23284

Conference proceedings

- **García-Timermans, C.**, Rubbens,P., Kerckhof, F.-M., Props, R., Waegeman, W. & Boon, N. (2019). Fingerprinting microbial communities through flow cytometry and Raman spectroscopy. In RamanFest 2019, Oxford, United Kingdom. (Poster)
- **García-Timermans, C.**, Rubbens,P. ,Props,R., Kerckhof, F.-M.,Waegeman, W., & Boon,N.(2019). Fingerprinting microbial communities through flow cytometry and Raman spectroscopy, In Bageco 15: 15th symposium on bacterial genetics and ecology: Ecosystem drivers in a changing planet,Lisbon,Portugal. (Poster)
- **García-Timermans, C.**, Rubbens,P. ,Kerckhof,F.-M. ,Props,R. ,Waegeman,W. , & Boon,N.(2018).Single-cell bacterial characterization using flow cytometry and Raman spectroscopy, In Belgian society for microbiology, symposium abstracts,Brussels,Belgium. (Presentation)
- **García-Timermans, C.**, Rubbens,P. ,Props,R. ,Boon,N. ,& Waegeman,W. (2017).Identifying synthetic communities using flow cytometry and machine learning, In Microbial resource management,2nd international symposium,abstracts,Ghent,Belgium. (Poster)
- **García-Timermans, C.**, Buysschaert,B., Rubbens,P., Kerckhof,F.-M., Skirtach,A.,& Boon,N.(2017). Detecting phenotypes with Raman spectroscopy, In FT-IR spectroscopy in microbiological and medical diagnostics,11th workshop,abstracts,Berlin,Germany. (Poster)

Community

- Coorganiser Women in Science day (2019 and 2020), Faculty of Bioscience Engineering, Ghent University, Belgium.
- Coorganiser of the mini-symposium on microbial flow cytometry. (2019). Faculty of Bioscience Engineering, Ghent University, Belgium.