Investigating the Evolutionary Origins and Cell Biology of Negativicutes

by

Richard Francis Brown

Thesis submitted for the degree of Doctor of Philosophy in Medical Sciences

University of Warwick, Warwick Medical School

October 2016
# Table of contents

List of figures ................................................................. VII
List of tables ................................................................. XI
Acknowledgements ........................................................... XIII
Author’s declaration ......................................................... XIV
Abstract ................................................................. XV
List of abbreviations ......................................................... XVI

Chapter 1  Introduction.......................................................... 1

1.1. Bacterial taxonomy .......................................................... 1

1.1.1. Phenotypic classification ........................................... 1

1.1.2. Molecular phylogenetic analysis .................................. 3

1.2. The Negativicutes .......................................................... 5

1.2.1. An enigma .............................................................. 5

1.2.2. Negativicutes in the history of microbiology ...................... 7

1.2.3. Phenotypic characteristics of the Negativicutes .................. 8

1.2.4. Genomic data .......................................................... II

1.2.5. The current taxonomic status of the Negativicutes ............... II

1.3. Physiology of the Negativicutes ........................................ 12

1.3.1. Envelope proteins .................................................... 12

1.3.2. Lipopolysaccharide .................................................. 13

1.3.3. Teichoic acids ....................................................... 14

1.3.4. Flagella .............................................................. 14

1.3.5. Sporulation ............................................................ 16

1.4. The diversity and abundance of Negativicutes ....................... 18
1.4.1. Environmental Negativicutes ........................................... 18
1.4.2. Negativicutes associated with Humans and Animals .......... 20
1.4.3. Diseases associated with the Negativicutes ....................... 20
1.4.4. Other Gram-negatives in the Firmicutes .......................... 21
1.5. The origins and evolutionary significance of the Negativicutes .... 22
1.6. Aims and objectives .......................................................... 25

Chapter 2  Materials and Methods ........................................... 26

2.1. Cultivation and Bacterial Strains ........................................ 26
2.2. Bacterial Growth Media ................................................... 27
2.2.1. DSMZ 3llc .............................................................. 27
2.2.2. DSMZ 3li .............................................................. 28
2.2.3. DSMZ 136 ............................................................. 28
2.2.4. DSMZ 14l: vitamin solution ........................................ 28
2.2.5. DSMZ 320: trace element solution SL-l0: ....................... 29
2.2.6. DSMZ 385: selenite and tungstate solution ....................... 29
2.2.7. SPM84 ................................................................. 29
2.2.8. BHI-YC ................................................................... 30
2.3. DNA preparation, extraction and sequencing .......................... 30
2.4. RNA preparation, extraction and sequencing .......................... 31
2.4.1. Viable cell counts ....................................................... 31
2.4.2. RNA sequencing ......................................................... 31
2.5. Bioinformatics methods ..................................................... 33
2.5.1. Genomes used for analysis ............................................ 33
2.5.2. Custom RefSeq BLAST database ................................... 34
2.5.3. Calculation of Orthologue groups using OrthoMCL .................. 34
2.5.4. Delineation of the core proteome ........................................ 35
2.5.5. Single and multiple next best BLAST hit(s)..............................35
2.5.6. Reciprocal best BLAST hit ....................................................35
2.5.7. Visualisation of synteny between Negativicutes ..........................36
2.5.8. Phylogenetic trees .................................................................37
2.5.9. Whole genome assembly and annotation of Sporomusa sphaeroides .................................................................. 37
2.5.10. Variant detection in 16S rRNA sequences .................................37
2.5.11. RNA-seq analysis .................................................................. 38
2.5.12. Normalisation of RNA-seq data .................................................38
2.5.13. Identification of transcriptionally active regions using toRNAdo and calculation of operons ................................................. 39
2.5.14. Visualisation of the transcriptional landscape using Biopython and genome diagram ......................................................... 39
2.5.15. Identification of putative non-coding RNA sequences in other Negativicutes ................................................................. 40
2.5.16. Figures and statistical analysis ....................................................40

Chapter 3  Evolutionary origins of the Gram-negative and Gram-positive toolkits of the Negativicutes ................................................. 41

3.1. Introduction ..................................................................................41
3.1.1. Deciphering the origins of the Negativicutes ...............................41
3.2. Results .........................................................................................44
3.2.1. Orthologue groups and core proteome of the Negativicutes ....44
3.2.2. Core-proteome phylogeny of the Negativicutes ...........................46
Chapter 3  Assignment of proteins to the Gram-positive or Gram-negative toolkit using a single-best-BLAST-hit approach ........................................ 47

3.2.4. Adding BamA to the core proteome ........................................ 49

3.2.5. Taxonomic assignments on protein products of the pan-genome using multiple best BLAST hits ........................................ 51

3.2.6. The Gram-negative toolkit is linked to the gamma-Proteobacteria ......................................................................................... 53

3.2.7. Visualising the V. parvula chromosome using the Gram Value 55

3.2.8. Gene function within the major Gram-negative locus .......... 59

3.2.9. Gene function within the minor Gram-negative locus ........... 61

3.2.10. Atypical Gram-negative features of the Negativicutes ....... 62

3.2.11. Synteny in the major Gram-negative locus .......................... 64

3.2.12. The origin of the Negativicutes ........................................... 73

  3.2.12.1. Gram-negative signature proteins ................................... 73

3.3. Discussion .................................................................................. 79

  3.3.1. A more sophisticated Negativicutes phylogeny .................... 79

  3.3.2. Defining the Gram-negative and Gram-positive toolkits ...... 80

  3.3.3. The origin of the Negativicutes ............................................. 81

  3.3.4. Halanaerobiales and Fusobacteria ....................................... 83

  3.3.5. Future work ........................................................................ 84

Chapter 4  Genome analysis of the anaerobic endospore forming Gram-negative Sporomusa sphaeroides ............................................. 85

4.1. Introduction ................................................................................ 85

  4.1.1. Sporomusa sphaeroides ....................................................... 85

  4.1.2. Rationale for genome-sequencing Sporomusa sphaeroides ...... 86
4.1.3. Nextera sequencing ................................................................. 86
4.1.4. Sequence assembly and annotation ........................................... 86
4.1.5. Genome analysis ..................................................................... 87
4.2. Results .......................................................................................... 87
  4.2.1. First sequencing run ................................................................. 87
  4.2.2. Second sequencing run .............................................................. 89
  4.2.3. The 16S ribosomal RNA genes .................................................. 91
  4.2.4. Phylogenetic analyses ............................................................... 91
  4.2.5. The Gram-negative toolkit of *S. sphaeroides* .............................. 95
  4.2.6. The sporulation toolkit of *S. sphaeroides* ................................... 97
  4.2.7. The flagellum of *S. sphaeroides* ................................................. 100
4.3. Discussion ..................................................................................... 103
  4.3.1. Contamination and sporulation .................................................. 103
  4.3.2. The 16S rRNA gene and phylogeny .......................................... 104
  4.3.3. Evidence of an ancient origin ................................................... 104
  4.3.4. Key features ........................................................................... 105
  4.3.5. Future work ............................................................................ 106

Chapter 5 The Transcriptional Landscape of the Negativicutes .... 108

5.1. Introduction ................................................................................... 108
  5.1.1. Transcriptomics ...................................................................... 109
  5.1.2. Rationale for this work .............................................................. 111
5.2. Results .......................................................................................... 112
  5.2.1. Development of a sporulation assay .......................................... 112
  5.2.2. RNA-Seq of logarithmic and stationary phases of growth ...... 118
5.2.3. Processing RNA-seq reads and quality control ..................119
5.2.4. Enhancing the genome annotation by identifying
transcriptionally active regions with toRNAdo..........................121
5.2.5. Putative non-coding RNAs in Negativicutes .....................124
5.2.6. Normalisation of RNA-Seq data........................................131
5.2.7. Analysis of differential expression.......................................136
5.2.8. Transcriptional analysis of Negativicutes.......................... 137
  5.2.8.1. Differential expression of COGs........................................137
  5.2.8.2. The conservation of both Gram-negative and Gram-
            positive operons................................................................139
  5.2.8.3. The major Gram-negative locus ......................................144
  5.2.8.4. Sporulation genes and pseudogenes.................................149
5.3. Discussion..............................................................................150
  5.3.1. Future work........................................................................152

Chapter 6  Conclusions ...............................................................153

Appendix A: Supplementary data for Chapter Three ............... 159
Appendix B: Supplementary data for Chapter Four ................ 163
Appendix C: Supplementary data for Chapter Five ................. 164
Bibliography.............................................................................182
List of figures

Figure 1.1: The Gram stain................................................................. 2

Figure 1.2: The key Gram-positive and Gram-negative features of Negativicutes................................................................. 6

Figure 1.3: Electron microscopy images of the cell wall structure of two different species of Negativicutes................................................. 9

Figure 1.4: Phase contrast images showing two species of Negativicutes..... 9

Figure 1.5: The growth in Negativicutes genomes available through NCBI...11

Figure 1.6: The structures of the Gram-positive and Gram-negative cell walls. ............................................................................... 15

Figure 1.7: Architectural differences between the Gram-negative and Gram-positive Flagella. ............................................................ 15

Figure 1.8: Timeline of the discovery of sporulating Negativicutes............ 17

Figure 1.9: Possible evolutionary scenarios for Negativicutes............... 24

Figure 3.1: Visualisation of the orthologue groups of the Negativicutes..... 45

Figure 3.2: ML phylogenetic tree of the core proteome. ........................... 46

Figure 3.3: Phyla associated with top-scoring BLAST hits from the core proteome of the Negativicutes. ................................................. 47

Figure 3.4: The major Gram-negative locus in the genome of V. parvula.... 49

Figure 3.5: ML phylogenetic tree of BamA orthologues in the Negativicutes. ......................................................................................... 51

Figure 3.6: Summary of the phyla from the multiple best BLAST hits of the Gram-negative proteins of the Negativicutes............................... 54

Figure 3.7: Assignment of multiple best BLAST hits within Proteobacteria.54
Figure 3.8: Moving average graph of the Gram Value across the genomes of C. difficile, E. coli, and V. parvula.......................... 56
Figure 3.9: The Gram Value visualised on the chromosome of V. parvula... 57
Figure 3.10: The orthologue groups of the Negativicutes relative to the V. parvula genome................................................................. 58
Figure 3.11: Detailed view of the major Gram-negative locus.................. 60
Figure 3.12: Phylogenetic location of species missing orthologue group 4... 61
Figure 3.13: Detailed view of the minor Gram-negative locus............... 63
Figure 3.14: Synteny in the major Gram-negative locus of the Negativicutes. ........................................................................................................ 64
Figure 3.15: Upstream of the major Gram-negative locus..................... 67
Figure 3.16: The core of the major Gram-negative region..................... 69
Figure 3.17: The distribution of the one-gene and two-gene configurations for orthologue group 210. ................................................................. 70
Figure 3.18: Downstream of the major Gram-negative locus of the Negativicutes.................................................................................................. 72
Figure 3.19: ML phylogenetic tree from concatenated GNSP of all bacteria identified with the same feature set.............................................. 77
Figure 3.20: The POTRA domains and partner proteins of BamA.......... 78
Figure 4.1: Coverage histogram of the first sequencing run.................. 88
Figure 4.2: The draft genome of S. sphaeroides.................................. 90
Figure 4.3: Core proteome based phylogenetic tree of Negativicutes....... 93
Figure 4.4: Comparison of the GC-content of Gram-negative and Gram-positive genes............................................................................. 93
Figure 4.5: Codon usage of Gram-negative and Gram-positive proteins.... 94
Figure 4.6: The major Gram-negative locus of synteny in Negativicutes. ... 96
Figure 4.7: Presence and absence of key sporulation proteins in Negativicutes.................................................................99
Figure 4.8: Maximum Likelihood tree of Negativicutes flagellin proteins. 101
Figure 4.9: The structure of the flagellum in S. sphaeroides.........................102
Figure 4.10: An insert in Gram-negative flagellin is absent in S. sphaeroides
.............................................................................................................103
Figure 5.1: Inhibition of sporulation with betaine. .......................................114
Figure 5.2: Growth curve of S. sphaeroides in different types of media at
37°C. ................................................................................................114
Figure 5.3: Growth of S. sphaeroides at different temperatures and brands of
BHI. ..................................................................................................115
Figure 5.4: Sporulation counts for SPM84 medium variations.......................116
Figure 5.5: Morphology of S. sphaeroides ...................................................117
Figure 5.6: PCA plots and heat maps of S. sphaeroides and V. parvula...........120
Figure 5.7: A comparison of Rockhopper and toRNAdo predictions..............122
Figure 5.8: Transcriptional regions identified by toRNAdo..........................123
Figure 5.9: Conserved putative ncRNAs in V. parvula................................129
Figure 5.10: Conserved ncRNAs in S. sphaeroides....................................130
Figure 5.11: Plot showing the distribution of counts of coding sequence
against TPM. ....................................................................................134
Figure 5.12: Relationship between the TPM cut-off and the percentage of
expressed sequences........................................................................135
Figure 5.13: The differential expression of COG at stationary vs logarithmic
phase in S. sphaeroides .....................................................................140
Figure 5.14: The differential expression of COG at stationary vs logarithmic
phase in V. parvula. ........................................................................141
Figure 5.15: Range of Gram Values from operons with either Gram-negative or Gram-positive genes. ................................................................. 143

Figure 5.16: The transcriptional landscape of the major Gram-negative locus in *V. parvula* and *S. sphaeroides*. ................................................................. 147

Figure S1: Phylogenetic tree of the 16S rRNA gene from Negativicutes. ..... 163

Figure S2: The growth curve of *V. parvula* and *S. sphaeroides*. ............. 164

Figure S3: Quality control of the extracted RNA. ....................................... 166

Figure S4: Bioanalyzer results for the 1st biological replicate in logarithmic phase. ........................................................................................................ 167
List of tables

Table 1.1: Phenotypic features of the Negativicutes.................................10
Table 1.2: Examples of the geographical locations and environments from
which species of Negativicutes have been isolated..............................19
Table 2.1: Bacterial strains.....................................................................27
Table 2.2: Indices for DNA sequencing ..................................................31
Table 2.3: Indices used for RNA sequencing ..........................................32
Table 2.4: Genomes used for analysis.....................................................33
Table 3.1: The major Gram-negative locus among genes encoding the core
proteome of the Negativicutes..............................................................48
Table 3.2: The number of proteins in each species of the Negativicutes with
entirely Gram-negative or Gram-positive BLAST hits............................53
Table 3.3: The presence and absence of the Gram-negative signature
proteins across all bacteria genera.......................................................76
Table 4.1: Summary of the sequencing data............................................89
Table 4.2: Summary of the Prokka annotation........................................89
Table 4.3: Comparison against the 16S rRNA nucleotide sequences from
NCBI......................................................................................................92
Table 4.4: The minimal machinery for sporulation in the Negativicutes....98
Table 5.1: Summary of transcriptional features......................................124
Table 5.2: Conservation of V. parvula putative ncRNAs in the class
Negativicutes and in the genus Veillonella...........................................126
Table 5.3: Conservation of S. sphaeroides putative ncRNAs in Negativicutes
and in the genus Sporomusa.................................................................127
Table 5.4: Expression of coding sequences..........................................133
Table 5.5: Differential expression between logarithmic and stationary in *V. parvula* and *S. sphaeroides*. .................................................................138

Table 5.6: Comparison of the operon structure of the major Gram-negative locus ..............................................................................................................148

Table 5.7: Expression of pseudogenes in *V. parvula* ...............................................150

Table S1: The orthologues of the GNSP for Mep45 and absence of Braun’s lipoprotein in the Negativicutes ........................................................................159

Table S2: The orthologues of the GNSP for BamA and absence of its partner proteins in the Negativicutes ........................................................................160

Table S3: The orthologues of the GNSP for lipid A biosynthesis in the Negativicutes ........................................................................................................161

Table S4: Viable cell counts of *V. parvula* and *S. sphaeroides* at logarithmic phase and stationary phase .................................................................165

Table S5: The concentration of RNA extracted from the bacterial cultures. ..............................................................................................................................166

Table S6: Table of unexpressed signature sporulation genes. ..............................168

Table S7: Predicted coding sequences with a RBS in *V. parvula* ....................170

Table S8: Predicted coding sequences with a RBS in *S. sphaeroides* ............172

Table S9: Putative ncRNA in *V. parvula*. ................................................................176

Table S10: Putative ncRNAs in *S. sphaeroides*. ..................................................178
Acknowledgements

Thank you to my supervisor Professor Mark Pallen for this opportunity, along with his expertise and advice along the way. Just as importantly, I would like to extend my thanks to my co-supervisor Dr Emma Denham. I was lucky and fortunate to have her help and guidance as I navigated my experiments and writing.

I would also like to thank Dr Chrystala Constantinidou for getting me started back in the laboratory after a hiatus from science, for her advice and for teaching me all about DNA sequencing. Also, Dr Gemma Kay for her help and for surviving three years of commuting with me between Birmingham and Coventry. A special thank you to Pavelas Sazinas as we struggled to learn Python and both shared the ups and downs of this journey.

My gratefulness to Dr Andrew Millard for his help and advice on bioinformatics as well as his suggestion I learn Python. I would also like to thank the other members of the group that made this an enjoyable journey, Emily Stoakes, Holly Hall, Dr Jacqueline Chan, Dr Martin Sergeant, Dr Nicholas Duggett, Ross Slater and everyone else in the Microbiology and Infection Unit.

I dedicate this thesis to my wife, Ivette.
Author’s declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any Degree.

The work presented (including data generated and data analysis) was carried out by the author.

Signed

Date
Abstract

The Negativicutes are a class of bacteria within the predominantly Gram-positive phylum Firmicutes that possess a diderm Gram-negative cell envelope along with lipopolysaccharides. Remarkably, some members of this class of Gram-negative bacteria can form endospores. The evolutionary origins of this enigmatic class were investigated through bioinformatics analyses, which defined and characterised their Gram-negative and Gram-positive toolkits, while analyses of Gram-negative signature proteins provided insights into the origins of the Gram-negative toolkit. The genome of Sporomusa sphaeroides, a sporulating species from the Negativicutes, was sequenced and annotated for the first time. In addition, the transcriptional landscapes of Veillonella parvula and Sporomusa sphaeroides during logarithmic and stationary phase growth were investigated using RNA-Seq, expanding the repertoire of known genome features, including non-coding RNAs within these species.
List of abbreviations

BAM  β-barrel assembly machinery
BAM  binary alignment/map
BHI  brain-heart infusion
BHI-YC brain-heart infusion with yeast and casitone
BLAST basic local alignment search tool
bp   base pair(s)
cDNA  complementary DNA
CFU   colony forming unit
COGs  clusters of orthologous groups
CSI   conserved signature insertions deletions
CSP   conserved signature proteins
DNA  deoxyribonucleic acid
DSMZ  Deutsche Sammlung von Mikroorganismen und Zellkulturen
E    expect value
GBK  GenBank file
GC-content guanine-cytosine content
GFF  general feature format file
GNSP  Gram-negative signature protein
GV   Gram value
HGT  horizontal gene transfer
IM   inner membrane
INDEL insertions deletions
kb   kilo-base-pairs
KEGG  Kyoto Encyclopedia of Genes and Genomes
LPS  lipopolysaccharide
Mb   mega-base-pairs
min  minutes
ML  maximum likelihood
mNBBH  multiple next best BLAST hit
sNBBH  single next best BLAST hit
mRNA  messenger RNA
NCBI  National Center for Biotechnology Information
ncRNA  non-coding RNA
NGS  next-generation sequencing
nm  nanometers
OD  optical density
OM  outer membrane
ORF  open reading frame
PCA  principle component analysis
PCR  polymerase chain reaction
RBBH  reciprocal best BLAST hit
RBS  ribosomal binding site
RefSeq  reference sequence database
RNA  ribonucleic acid
RNA-Seq  RNA-sequencing
RPKM  reads per kilobase per million reads
rRNA  ribosomal ribonucleic acid
RT  room temperature
SAM  sequence alignment/map
SLPs  S-layer proteins
sNBBH  single next best BLAST hit
SNP  single nucleotide polymorphism
SPM84  *Sporomusa media* 1984
TPM  transcripts per million
tRNA  transfer RNA
UTR  untranslated region
WGS  whole genome sequencing
Chapter 1
Introduction

1.1. Bacterial taxonomy

1.1.1. Phenotypic classification

The human need to impose order on nature through classification has been realised from ancient times by Aristotle in his Scala Naturae through to the 18th Century by Linnaeus with Systema Naturae. At the birth of microbiology in the 19th century, there was a similar desire to separate bacteria into groups according to characteristics such as morphology, growth and metabolic substrates.

Building on light microscopy, staining techniques were used as a tool in bacterial classification, most notably through what is now called the Gram stain, which was developed in 1884 by Danish scientist Hans Christian Gram [1]. This technique split bacteria into two groups: those that retained the purple crystal-violet/iodine stain after a decolourisation step were classified as Gram-positive and those that failed to retain the stain were classified as Gram-negative. Over a century later, the Gram stain remains a routine procedure in
microbiology, facilitating prompt diagnosis and treatment of bacterial infections [2].

The ability of the Gram stain to distinguish Gram-positive from Gram-negative bacteria depends on marked differences in the cell envelopes of these groups, particularly the much thicker peptidoglycan of the Gram-positive cell wall, which allows it to retain the dye [3]. However, another important difference between the two groups is that Gram-positive cells have a single membrane and so are sometimes described as “monoderms”, while Gram-negative cells have two membranes and so are described as “diderms” (Figure 1.1).

![Diagram of Gram-positive and Gram-negative bacteria]

**Figure 1.1: The Gram stain.**

Bacteria are fixed and stained with crystal violet, then treated with iodine, which binds the crystal violet and traps it within the cell. Ethanol is used to decolourise the Gram-negative bacteria. Gram-negative cells are counter-stained with safranin.

In the late 19th century, Ferdinand Cohn, one of the founders of microbiology, created the first bacterial taxonomy based upon cellular morphology, dividing bacteria into four groups and six genera:
Chapter One

- the sphere-shaped bacteria *Sphaerobacteria*, which included the genus *Micrococcus*.
- the rod-shaped *Microbacteria*, which included the genus *Bacterium*.
- the filamentous *Desmobacteria*, which included *Bacillus* and *Vibrio*.
- the screw-like *Spirobacteria*, which included *Spirilium* and *Spirochaeta*.

Cohn showed that the genus *Bacillus* produced heat-resistant endospores that survived the boiling that killed the more numerous vegetative cells [4, 5].

In 1923, American bacteriologist David Bergey published his manual on the collective efforts of bacterial taxonomists. This quickly became the main resource on bacterial taxonomy providing “comprehensive and authoritative descriptions of bacteria” [6].

However, the way a micro-organism was classified in the first half of the 20th Century was often based on subjective assumptions—a problem embodied in the statement “a species is whatever a competent taxonomist decides to call a species” [7]. In 1957, the British taxonomist Peter Sneath foresaw the potential of computers in bacterial taxonomy [8] and over the next decade he worked with the American bacteriologist Robert Sokal to lay the foundations of a more objective “numerical taxonomy”, in which multiple characters were given equal weighting to classify species without subjective judgments [8].

### 1.1.2. Molecular phylogenetic analysis

Friedrich Miescher discovered DNA in 1869. James Watson and Francis Crick described the double helix (using data gathered by Rosalind Franklin) in 1953 [9]. In the years that followed, elucidation of the genetic code and the central dogma of molecular biology (“DNA makes RNA makes protein”) showed how biological macromolecules can carry information within the sequences of their constituent units.
In the 1960s, Émile Zuckerkandl and Linus Pauling proposed that the rate of evolution at the molecular level was constant over time, giving rise to the idea of a “molecular clock” [10], particularly given that most of the variation at the molecular level is neutral, i.e. not due to natural selection [11]. In their landmark 1965 paper, *Molecules as documents of evolutionary history*, they recognised that, because the genetic code was triplet and degenerate, nucleic acid sequences could provide more insights into evolutionary history than amino-acid sequences [12].

In 1977, the British scientist Fred Sanger invented sequencing using "dideoxy" chain-terminators [13], which was used initially to sequence the bacteriophage λ [14] and then ultimately to sequence the human genome [15]. The ability to sequence long stretches of DNA enabled the use of this macromolecule as a powerful tool for taxonomic classification.

Also in 1977, American microbiologist Carl Woese showed that the small-subunit ribosomal RNA could be used for phylogenetic classification of microorganisms. The ubiquity of this essential molecule across all cellular life forms, twinned with its slow rate of evolution, made it ideal for reconstructing molecular phylogenies. Notably, Woese used this method to define a third domain of life, the Archaea [16].

In 1987, Naruya Saitou and Masatoshi Nei introduced the Neighbour Joining method for reconstructing phylogenetic trees. This allowed for rapid reconstruction of evolutionary history based on either nucleotide or amino-acid sequences and made it feasible to examine very large datasets [17]. Reconstructing phylogenetic trees to understand evolutionary history has become the cornerstone of phylogenetics [18].

This era of bacterial genome sequencing began in 1995, when *Haemophilus influenzae* Rd was sequenced using a whole-genome shotgun approach [19]. Although initially genome-sequencing projects were expensive and time
consuming, with the advent of “high-throughput sequencing” or “next-generation sequencing” (NGS) in the latter half of the 2000s, bacterial whole-genome sequencing (WGS) became much easier and cheaper [20-22], so that it has now become a routine technique in microbiology research laboratories.

1.2. The Negativicutes

1.2.1. An enigma

The advent of molecular phylogenetics transformed microbiology, providing a framework for classification from species all the way up to phyla. Initially, Gram-positive bacteria were lumped together in a bacterial division termed the *Firmacutes* [23], which was then subsequently renamed as a phylum, the *Firmicutes*, from the Latin for “thick skin”, due to their strong monoderm cell envelope with a thick layer of peptidoglycan. Some Gram-positive bacteria were classified into another phylum, the Actinobacteria, which also contains some genera, including *Mycobacterium* and *Actinomyces*, that are Gram-indeterminate or Gram-variable [24-26].

One of the central tenets of modern phylogenetic classification is that named taxonomic groups should be monophyletic, i.e. should contain all the descendants of a most-recent common ancestor, but no additional organisms that are not descended from that ancestor. Given that the Firmicutes was defined as the home for the majority of Gram-positive bacterial taxa, it might seem reasonable to assume that ancestor of all bacteria within this phylum possessed a typical Gram-positive monoderm cell envelope and also that all the descendent bacteria within this phylum would possess this simple cell wall. However, here morphological and molecular approaches to taxonomy clash, because according to molecular phylogenies, several bacterial genera that have a Gram-negative cell envelope actually belong in the phylum Firmicutes, and have recently been classified within their own class, the Negativicutes [27].
On first hearing this, most experienced bacteriologists think that there must have been some mistake. However, electron microscopy has proven that Negativicutes have a diderm cell wall structure and are therefore true “Gram-negatives” [28, 29]. Yet, phylogenetic analyses of 16S rRNA gene sequences show that the Negativicutes undeniably belong in the phylum Firmicutes where they form a sister group to the class Clostridia [27].

These bacteria present us with a conundrum as their genomes encode features of both Gram-positive and Gram-negative bacteria. While genetic analyses show they unquestionably sit within the “Gram-positive” phylum Firmicutes due to their 16S rRNA gene [27], they stain Gram-negative with a diderm cell envelope [30-32] and an outer membrane (OM) containing lipopolysaccharide (LPS). Astonishingly, despite being diderm Gram-negatives, some members of this group such as *Acetonema longum* [28] and *Pelosinus fermentans* [33] possess the ability—otherwise confined to Gram-positive bacteria—to form endospores [34] (Figure 1.2)!

---

**Figure 1.2: The key Gram-positive and Gram-negative features of Negativicutes.**
1.2.2. Negativicutes in the history of microbiology

Far from being an obscurity, the Negativicutes have placed a crucial role in the history of microbiology. In 1683, when Antonie van Leeuwenhoek first used his microscope to observe motile cells from the human mouth, he was probably looking at *Selenomonas*, which is now classified as a member of the Negativicutes [35]. This genus was more fully described in the 19th century and given a name that means “moon-shaped unit” after the organism’s distinctive cellular morphology [36, 37].

In 1898, Veillon and Zuber described a new bacterial species, which they named *Staphylococcus parvulus*, [38]. In 1933, Prévot [39] re-assigned this bacterium to the genus *Veillonella* and named it *V. parvula* that we now consider a member of the Negativicutes. Sixty-six years after its discovery *Veillonella* gained an important place in the history of microbiology when Bladen and Mergenhagen reported a correlation between the outer membrane fraction and endotoxic activity [40]. This established for the first time, in any bacterium, the relationship between lipopolysaccharide and endotoxin.

In 1984, the first spore-forming member of the Negativicutes was described, *Sporomusa*. It most closely resembled the non-spore-forming genus *Selenomonas* [41]. However, the endospore-forming bacterium *Clostridium quercicolum* [42], first isolated in 1971, was reclassified as a member of the Negativicutes and renamed *Dendrosporobacter quercicolum* in 2000 [43]. At the time of writing (2016), 34 genera have been classified within the Negativicutes, with *Negativicoccus* the most recent addition in 2010 [27].

The first genome of a Negativicutes was sequenced in 2007 [44] and since then there has been an explosion of genomic data. The most recent change to the Negativicutes taxonomic groupings were by Marchandin et al. [27] in 2010. The name Negativicutes was coined and this group of bacteria was promoted from an order within Clostridia to the rank of class within the Firmicutes. The
debate surrounding this decision and the most accurate way to interpret their taxonomy continues and will be discussed later.

1.2.3. Phenotypic characteristics of the Negativicutes

The Negativicutes are obligate anaerobes, i.e. they cannot grow in the presence of oxygen and generally depend upon fermentation as a means of ATP production— for example *Veillonella* is known for its ability to ferment lactate [45]. Electron microscopy has removed any doubt surrounding the debate that Negativicutes are Gram-negative (Figure 1.3). Negativicutes occur in different shapes from cocci to bacilli with variations in-between. All the sporulating species are bacilli, which are motile and have flagella. They form terminal endospores and spherical spores. The non-spore-forming Negativicutes also vary in shape (Table 1.1). Negativicutes also vary in size, as can be seen when *V. parvula* and *Sporomusa sphaeroides* are viewed under the same magnification (Figure 1.4).
Figure 1.3: Electron microscopy images of the cell wall structure of two different species of Negativicutes.

Left *A. longum* [28]. Right *Centipeda periodontii* [29]. Both show a diderm Gram-negative cell wall structure. (Outer membrane (OM), Peptidoglycan layer (PD), inner membrane (IM)).

Figure 1.4: Phase contrast images showing two species of Negativicutes.

*V. parvula* DSM2008 (left) which are cocci and *S. sphaeroides* (right) which are bacilli also show distinct differences in size. Images captured using a Leica DMi8 at 100x with oil immersion.
Table 1.1: Phenotypic features of the Negativicutes.

The presence (+) or absence (-) of the features (m) motility, (f) flagella and (s) spore-forming are shown along with the ideal growth (t) temperature in °C. A blank cell indicates it was not described in the literature.

| NCBI Taxonomy          | Ref       | Morphology                                                                 |
|------------------------|-----------|-----------------------------|-----------------|-----------------|-----------------|-----------------|-----------|-----------------|
| Selenomonadales        | 27        |                              | M   | F   | S   | T   |           |         |
| Acidaminococcaceae     |           |                              | -   | -   |    | 37  |           |         |
| Acidaminococcus        | [46-48]   | Single cells, oval or kidney-shaped diplococci                           | -   | -   |    | 37  |           |         |
| Phascolarctobacterium  | [49]      | Pleomorphic rods                                  | -   | -   |    | 37  |           |         |
| Succinilasticum        | [50]      | Short rods                                           | -   | -   |    | 39  |           |         |
| Succinispira           | [51]      | Curved rods                                      +  | +   |    | 37  |           |         |
| Veillonellaceae        | [27, 52]  |                              | M   | F   | S   | T   |           |         |
| Acidaminococcus        |           | Straight rods with rounded ends                        | +   | +   |    | 30  |           |         |
| Allisonella            | [54]      | Cells are ovoid shaped and occur in diploid or chain form | -   | -   |    | 39  |           |         |
| Anaeroarcus            | [55]      | Curved or spiral-shaped rods                       | +   | +   |    | 35  |           |         |
| Anaeroglobus           | [56]      | Coccoid to ellipsoidal, usually in pairs or, occasionally, in short chains | -   | -   |    | 37  |           |         |
| Anaeromusa             | [57]      | Curved rods                                      +  | +   |    | 38  |           |         |
| Anaerosinus            | [55]      | Curved rods or spirals                           | +   | +   |    | 37  |           |         |
| Anaerospira            | [58]      | Straight or slightly curved Rods                  | +   | +   |    | 37  |           |         |
| Anaerovibrio           | [55]      | Slightly curved rods                             | +   |    |    | 39  |           |         |
| Centipeda              | [29, 30]  | Rod-shaped with three or more curves                | +   | +   |    | 35  |           |         |
| Dendrosporobacter      | [42, 43]  | Straight Rods                                    | +   | +   |    | 30  |           |         |
| Desulfosporomusa       | [59]      | Curved rods with tapered ends                     +  | +   |    | 28  |           |         |
| Dialister              | [60-63]   | Cocacobacilli                                     | -   | -   |    | 30-37 |        |         |
| Megamonas              | [64-66]   | Rod                                           | -   | -   |    | 37  |           |         |
| Megasphaera            | [52, 67]  | Cocci                                          | -   | -   |    | 30-38 |        |         |
| Mitsuokella            | [68, 69]  | Regular or ovoid rods                            | -   | -   |    | 37-42 |        |         |
| Negativicoccus         | [27]      | Cocci                                          | -   | -   |    | 37  |           |         |
| Pectinatus             | [70-72]   | Slightly curved to helical rods & rounded ends. Single, pairs or rarely in short chains | +   | +   |    | 20 to 30 |        |         |
| Pelosinus              | [33, 73]  | Slightly curved rods                              | +   | +   |    | 30  |           |         |
| Propionispora          | [74]      | Curved to helical rod                             | +   | +   |    | 30-35 |        |         |
| Propionispora          | [75, 76]  | Cells are curved or spiral-shaped rods             | +   | +   |    | 37  |           |         |
| Psychrosinus           | [77]      | Curved rods                                    +  | +   |    | 15  |           |         |
| Quinella               | [78]      | Oval                                           | +   | +   |    | 37-39 |        |         |
| Schwartzia             | [79]      | Curved Rods                                    +  | +   |    | 39  |           |         |
| Selenomonas            | [80]      | Curved, crescent shaped                          | +   | +   |    | 37-45 |        |         |
| Sporolitus             | [81]      | Curved Rods                                    +  | +   |    | 55  |           |         |
| Sporomusa              | [41, 82, 83] | Curved rods with tapered ends                 | +   | +   |    | 30-37 |        |         |
| Sporotalea             | [84]      | Long Rods                                      +  | +   |    | 25-30 |        |         |
| Thermostinus           | [85]      | Curved Rods                                    +  | +   |    | 60  |           |         |
| Veillonella            | [39, 86-89]| Cocci in pairs, masses or short chains          | -   | -   |    | 37  |           |         |
| Zymophilus             | [90]      | Slightly curved to helical rods, occurring singly, in pairs, or in short chains. | +   | -   |    | 30  |           |         |
1.2.4. Genomic data

Analysis of 16S rRNA gene sequences provided the first clue that the Negativicutes belonged in a largely Gram-positive phylum [91]. However, most species of Negativicutes have now been fully genome sequenced, providing a wealth of data to investigate their evolutionary origins. The first published draft genome of a Negativicutes was that of the extremophile *Thermosinus carboxydivorans* Norl in 2007 [44], while the genome of *V. parvula* DSM 2008 became available in 2009. Slowly, with the availability of more genomic data the gates have started to open to a more thorough understanding and complete analysis of this class of bacteria. Eleven genome sequences have been completely finished and 86 draft genomes are available according to the National Center for Biotechnology Information (NCBI) (Figure 1.5).

![Figure 1.5: The growth in Negativicutes genomes available through NCBI.](image)

Includes both complete and draft genome sequences, correct as of May 2016.

1.2.5. The current taxonomic status of the Negativicutes

The current classification of Negativicutes came from the proposal in 2010 by Marchandin *et al.* to elevate the Veillonellaceae to the rank of a class in the Firmicutes [27] following 16S rRNA gene analysis of three unknown clinical
isolates from the Hôpital Arnaud de Villeneuve in France. Marchandin argued that “[p]hylogenetic analyses based on the 16S rRNA gene sequences of members of the phylum Firmicutes and other phyla indicated that the family Veillonellaceae forms a robust lineage clearly separated from those of the classes ‘Bacilli’, ‘Clostridia’, ‘Thermolithobacteria’ and ‘Erysipelotrichi’ in the phylum Firmicutes”. However, although the 16S rRNA gene phylogeny showed that the group remained within the Firmicutes, Marchandin et al. [27] argued the conserved Gram-negative cell envelope confirmed the coherence of the proposed class Negativicutes.

Since the inception of the Negativicutes as a class, genome sequencing has fuelled arguments for and against this formulation. A combination of 16S rRNA gene and protein-based phylogenetic trees failed to support the classification, while also splitting the sporulating Negativicutes into a separate family [92]. However, support came from a comparative genomic analysis that found 14 Conserved Signature Insertions Deletions and 48 Conserved Signature Proteins unique to Negativicutes [93]. The distributions of these within Negativicutes both supported and refined the class Negativicutes.

1.3. Physiology of the Negativicutes

1.3.1. Envelope proteins

Electron microscope images of Negativicutes show that the cell envelope ultrastructure has a three-layer organisation typical of diderms, with a cytoplasmic membrane, peptidoglycan layer and a peripheral outer membrane[76, 94] (Figure 1.3).

Braun’s lipoprotein is a membrane protein abundant in the cell walls of typical Gram-negative bacteria such as E. coli [95, 96], where it links peptidoglycan to the outer membrane, providing structural integrity [97]. However, the Negativicutes do not use Braun’s lipoprotein — instead the outer membrane is anchored to peptidoglycan using cadaverine [98].
Some Gram-negative and Gram-positive bacteria produce S-layer proteins (SLPs) as a major, metabolically costly feature of the cell envelope. SLPs self-assemble into 2-dimensional crystalline arrays that completely cover the bacterial cell surface, acting as a protective coat and a molecular sieve [99, 100]. Among the Negativicutes, *Selenomonas* produces a major membrane protein homologous to SLPs and deploys an ordered array of protein covering its surface [101]. Other typical Gram-negative outer membrane proteins have also been identified—for example, proteins from *Mitsuokella multacida* [102] and *S. ruminantium* [103] have a secondary structure that hints at the C-terminal domains forming a β-barrel formation.

The Omp85/BamA family of proteins are present in all bacteria that have an outer membrane [104]. Through evolution, BamA has acquired additional partner subunits, which form a complex termed the β-barrel assembly machinery (BAM). This complex is responsible for the folding and insertion of proteins into the outer membrane [105]. BamA is present in Negativicutes, but little is known about its evolutionary and functional context [106].

### 1.3.2. Lipopolysaccharide

LPS is a large polymer found on the outer membrane of Gram-negative bacteria (Figure 1.6). LPS consists of a lipid A that anchors the molecule to the outer membrane, an inner and outer core and a surface-exposed O-antigen polysaccharide. LPS contributes to the structural integrity of the outer membrane—in particular, its negative charge stabilises the structure of the outer membrane. In animals, LPS acts as endotoxin, eliciting a strong response from the innate immune system. Although now recognised as a universal feature of the Gram-negative cell, the link between LPS and the outer membrane was first demonstrated in a member of the Negativicutes: in 1964 when Bladen and Mergenhagen reported an association between outer membrane fractions and endotoxic activity in *Veillonella* [40]. The O-antigen and its antigenic properties are well documented, as it is used for the
serotyping of *Salmonella enterica* and *Escherichia coli* [107]. Within a discrete bacterial group the structure of LPS shows the least diversity in lipid A and most diversity in the O-specific chains [31]. The diversity of the exposed O-antigen has been shown to be, in part, the result of horizontal gene transfer (HGT) [107-109]. The LPS of the Negativicutes also show considerable diversity. For example, the lipid A from *Centipeda periodontii* is distinct from the lipid A from *E. coli* [30]. Members of the Negativicutes have all the typical genes essential for lipid A biosynthesis, but exploit an alternative route for UDP-diacylglucosamine hydrolysis [110], using LpxI rather than LpxH [106, 111]. According to a review by Helander *et al.* in 2004, within the genus *Pectinatus*, bacteria produce smooth and rough types of LPS molecules in the same strain. The core region of the LPS is mostly conserved with a large non-repetitive saccharide that may replace the O-specific chain, which has been described as a novel architectural principle [31].

### 1.3.3. Teichoic acids

Teichoic acids are a unique feature of Gram-positive bacteria that are present in the cell wall and provide them with rigidity (Figure 1.6). They are anchored to the lipid membrane or the peptidoglycan [112]. However, there is no evidence from the scientific literature that these have been found in the Negativicutes.

### 1.3.4. Flagella

The flagellum is the chief organelle of motility in Gram-negative and Gram-positive bacteria [113]. Most of the components of the flagellum are present in both Gram-negative and Gram-positive bacteria. However, only Gram-negative bacteria have L and P rings in the outer membrane, LPS and peptidoglycan respectively (Figure 1.7) [114]. A flagellum is driven by the Mot complex, which is a rotary engine powered by the proton motive force [115].
Figure I.6: The structures of the Gram-positive and Gram-negative cell walls.

The Gram-positive cell wall contains a thick peptidoglycan layer and a single membrane. The Gram-negative cell wall has both an inner and outer membrane with a thin peptidoglycan layer. The outer membrane of the Gram-negative bacteria has LPS.

Figure I.7: Architectural differences between the Gram-negative and Gram-positive Flagella.

Gram-negative bacteria have the LPS associated L ring and peptidoglycan associated P ring. Both of which are absent in Gram-positive bacteria.
1.3.5. Sporulation

Ferdinand Cohn discovered sporulation—a key feature of some Gram-positive species of bacteria—through the detection of heat resistant endospores [4, 5]. By dividing asymmetrically, a sporulating bacterium can produce a single endospore in times of stress. This spore can persist in the environment for great lengths of time and survive extreme conditions, until conditions become more favourable and it germinates to resume vegetative growth [116].

Typically—and according to most textbooks—Gram-negative bacteria do not form endospores. The Negativicutes provide several exceptions. The first spore-forming member of the class was described in 1984 and named Sporomusa [41]. Subsequently, eight additional genera of spore-forming Negativicutes have been described in the literature (Figure 1.8).

Abecasis et al. recently described a genomic signature to predict whether a bacterium can sporulate [117]. In this context, it is worth noting that although Thermosinus carboxydivorans has not been reported to sporulate [85], it possesses many of the genes required for this process [92]. Sporulation has been characterised in detail in only a single species within the Negativicutes: A. longum [28], where cryo-tomographic imaging has shown that the prespore is engulfed by the inner membrane of the mother cell during sporulation and that, upon germination, the inner membrane forms the outer membrane of the new daughter cell.
Figure 1.8: Timeline of the discovery of sporulating Negativicutes.

Timeline and environments the different sporulating Negativicutes were discovered in. The genera are Acetonema [53], D. quercicolum [43] that was renamed and reclassified from C. quercicolum [42], Propionispora [75], Desulfosporomusa [59], Anaerospora [58], Pelosinus [33], Sporotalea [84] and Sporolituus [81].
1.4. The diversity and abundance of Negativicutes

Bacteria in the class Negativicutes are found in locations around the planet and in diverse environments, from the termite gut to the human mouth, and even in spoiled beer (Table 1.2).

1.4.1. Environmental Negativicutes

Negativicutes have been found in a diverse range of contexts from subsurface kaolin lenses in Russia [33] to rice fields in West Africa. While the majority are mesophiles, some inhabit extreme environments. For example, *Psychrosinus fermentans* [58] is a psychrophile that was isolated from the near-freezing and permanently ice-covered Lake Fryxell. At the opposite end of the spectrum are the thermophiles. For example, *Sporolituus thermophilus*, which grows optimally at 55°C, was isolated from geothermal waters of the Great Artesian Basin of Australia [81]. Another extremophile, *Thermosinus carboxydivorans* grows best at 60°C and was isolated from a hot pool in Yellowstone National Park [85].

Two genera from the Negativicutes—*Pectinatus* and *Megasphaera*—have been isolated from beer and associated with the rise of beer spoilage [118], which peaked in the early nineties but has subsided in recent years with advances in technology. Some species of Negativicutes have noteworthy metabolic features, such as *Sporomusa*, with a novel methanol degradation pathway [119] and *Veillonella* and its ability to ferment lactate [45].
Table 1.2: Examples of the geographical locations and environments from which species of Negativicutes have been isolated.

<table>
<thead>
<tr>
<th>NCBI Taxonomy</th>
<th>Ref</th>
<th>Environments of origin</th>
<th>Geographic Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selenomonadales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidaminococaceae</td>
<td>[27]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidaminococcus</td>
<td>[46-48]</td>
<td>pig, human (peritoneal fluid)</td>
<td>France</td>
</tr>
<tr>
<td>Phascolarctobacterium</td>
<td>[49]</td>
<td>faeces (human &amp; koala)</td>
<td>Australia, Japan</td>
</tr>
<tr>
<td>Succiniclasticum</td>
<td>[50]</td>
<td>cow rumen</td>
<td>Sweden</td>
</tr>
<tr>
<td>Succinispira</td>
<td>[51]</td>
<td>pond receiving effluent from an anaerobic solids digester</td>
<td>New Zealand</td>
</tr>
<tr>
<td>Veillonellaceae</td>
<td>[27, 52]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetonema</td>
<td>[28, 53]</td>
<td>termite gut</td>
<td>USA</td>
</tr>
<tr>
<td>Allisonella</td>
<td>[54]</td>
<td>cow rumen</td>
<td>USA</td>
</tr>
<tr>
<td>Anaeroaracus</td>
<td>[55]</td>
<td>rice field soil</td>
<td>West Africa</td>
</tr>
<tr>
<td>Anaeroglobus</td>
<td>[56]</td>
<td>post-operative fluid</td>
<td>France</td>
</tr>
<tr>
<td>Anaeromusa</td>
<td>[57]</td>
<td>anaerobic purification plant</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Anaerosinus</td>
<td>[55]</td>
<td>freshwater mud</td>
<td>Germany</td>
</tr>
<tr>
<td>Anaerospora</td>
<td>[58]</td>
<td>human blood culture</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Anaerovibrio</td>
<td>[55]</td>
<td>sheep rumen</td>
<td></td>
</tr>
<tr>
<td>Centipeda</td>
<td>[29, 30]</td>
<td>human periodental lesions</td>
<td>USA</td>
</tr>
<tr>
<td>Dendrosporobacter</td>
<td>[42, 43]</td>
<td>discoloured tissue in living oak tree</td>
<td>USA</td>
</tr>
<tr>
<td>Desulfosporomusa</td>
<td>[59]</td>
<td>oligotrophic lake</td>
<td>Germany</td>
</tr>
<tr>
<td>Dialister</td>
<td>[60-63]</td>
<td>human (amniotic fluid, root canal, nasal passages, faeces)</td>
<td>France, UK</td>
</tr>
<tr>
<td>Megamonas</td>
<td>[64-66]</td>
<td>human faeces, caecum (chicken &amp; duck)</td>
<td>France, Japan</td>
</tr>
<tr>
<td>Megasphaera</td>
<td>[52, 67]</td>
<td>spoiled beer, sheep rumen, human liver abscess</td>
<td>France, Italy, Sweden</td>
</tr>
<tr>
<td>Mitsuokella</td>
<td>[68, 69]</td>
<td>cow rumen, human faeces, human root canal</td>
<td>Malaysia</td>
</tr>
<tr>
<td>Negativicoccus</td>
<td>[27]</td>
<td>human toe wound</td>
<td>France</td>
</tr>
<tr>
<td>Pectinatus</td>
<td>[70-72]</td>
<td>spoiled beer</td>
<td>China, Germany, Finland</td>
</tr>
<tr>
<td>Pelosinus</td>
<td>[33, 73]</td>
<td>kaolin deposit, termite gut</td>
<td>Congo, Russia</td>
</tr>
<tr>
<td>Propionispira</td>
<td>[74]</td>
<td>cotton wood tree</td>
<td>USA</td>
</tr>
<tr>
<td>Propionispora</td>
<td>[75, 76]</td>
<td>compost, sewage treatment plant</td>
<td>Germany</td>
</tr>
<tr>
<td>Psychrosinus</td>
<td>[77]</td>
<td>frozen lake</td>
<td>Antarctica</td>
</tr>
<tr>
<td>Quinella</td>
<td>[78]</td>
<td>sheep rumen</td>
<td></td>
</tr>
<tr>
<td>Schwartzia</td>
<td>[79]</td>
<td>cow rumen</td>
<td>Sweden</td>
</tr>
<tr>
<td>Selenomonas</td>
<td>von Prowazek 1993</td>
<td>gingival crevice, yak rumen</td>
<td>China, Germany, USA</td>
</tr>
<tr>
<td>Sporolitus</td>
<td>[81]</td>
<td>geothermal waters</td>
<td>Australia</td>
</tr>
<tr>
<td>Sporomusa</td>
<td>[41, 82, 83]</td>
<td>freshwater mud, termite gut, alcohol distillery</td>
<td>France, Germany, Jamaica, USA</td>
</tr>
<tr>
<td>Sporotalea</td>
<td>[84]</td>
<td>soil samples</td>
<td></td>
</tr>
<tr>
<td>Thermosinus</td>
<td>[85]</td>
<td>hot spring</td>
<td>USA</td>
</tr>
<tr>
<td>Veillonella</td>
<td>[39, 86-89]</td>
<td>mouth (guinea pig, hamster, human, rat), human intestines</td>
<td>Australia, France, Germany, UK</td>
</tr>
<tr>
<td>Zymophilus</td>
<td>[90]</td>
<td>pitching yeast</td>
<td>Germany</td>
</tr>
</tbody>
</table>
1.4.2. Negativicutes associated with Humans and Animals

Negativicutes is present in a wide variety of invertebrate and vertebrate microbiomes. Species isolated from the intestines of invertebrates include *A. longum* [53] and *Pelosinus propionicus* [73]. Others have been isolated from ungulates such as cow [50], sheep [120] and pigs [121]. *Megamonas* is an abundant representative of the Negativicutes found in the chicken microbiome [122]. The canine oral microbiome has also shown the presence of Negativicutes with many of the same taxa present in humans [123]. The Negativicutes feature prominently in the human microbiome—for example, *Dialister* [62], *Selenomonas* [124] and *Veillonella* [125] have been found throughout the human digestive tract. The Human Microbiome Project found the Negativicutes, and in particular *Veillonella*, as abundant taxa [126], whilst another project found *Dialister invisus* highly abundant in U.S children [127].

1.4.3. Diseases associated with the Negativicutes

The Negativicutes are typically considered non-pathogenic or commensal organisms. However, it was the isolation of *Negativicoccus* from soft tissue and skin of hospitalised patients that prompted Marchandin *et al.* to define Negativicutes as a novel class. [27]. Several examples in the literature report fatal *Selenomonas sputigena* septicaemia, *Selenomonas* [128] and *Negativicoccus succinicivorans* [129] causing bacteraemia and *Veillonella* [130] causing meningitis. *Centipeda periodontii, Selenomonas* and *Mitsuokella* have been isolated from patients suffering periodontitis, but are absent or far less abundant in healthy individuals [29, 124]. Changes in the relative abundance of some genera of Negativicutes have been reported as a risk for asthma and during the treatment of Crohn's disease [131, 132].
1.4.4. Other Gram-negatives in the Firmicutes

The class Negativicutes does not encompass all the Gram-negative bacteria currently assigned to the Firmicutes. The Halanaerobiales represent a family of bacteria within the class Clostridia that like the Negativicutes are obligate fermentative anaerobes that stain Gram-negative. They were reorganised as a highly related, diverse, group of bacteria in 1995 [133]. Unlike the Negativicutes, they are obligately halophilic. They are slightly curved or flexibly rod-shaped and the majority have flagella. Species within this family include:

- *Acetohalobium arabaticum* discovered in East Crimea from lagoons of the Arapat spit [134].
- *Halanaerobium praevalens* found in the Great Salt Lake [135].
- *Halthermothrix orenii* identified from Tunisian salt lake sediment [136].
- *Orenia salinaria* discovered in Salin-de-Giraud in France from commercial salters [137].
- *Sporohalobacter lortetii* from the Dead Sea.

In addition to a Gram-negative cell envelope, some Halanaerobiales share with some Negativicutes the capability to sporulate. The relationship of the Halanaerobiales to the Negativicutes is unclear, although a concatenation of 21 ribosomal protein sequences placed them close together on a phylogenetic tree [138]. Genetic analysis of the anaerobic thermohalophilic bacterium *H. orenii* revealed that it possesses the pathways for lipid A biosynthesis. Proteins from this pathway showed 45-65% amino acid similarity to the homologues from the Negativicate *T. carboxydivorans* and phylogenetic analysis placed these two bacteria into a monophyletic group. The pathway encoded all the proteins required for lipid A biosynthesis, except LpxH—instead it uses the same LpxI pathway as Negativicutes [136]. These provide a strong argument for a monophyletic origin of the Negativicutes and Halanaerobiales. There are
several other bacterial species that are classified within the Firmicutes but appear to be Gram-negative, including Thermohalobacter berrensis [139], Caloranaerobacter azorensis [140] and Caminicella sporogenes [141]. However, their relationship to the Negativicutes and Halanaerobiales remains unresolved.

1.5. The origins and evolutionary significance of the Negativicutes

How do we explain the presence of a class of Gram-negative bacteria within a Gram-positive phylum? One approach is to piece together the origin of the Negativicutes from analyses of gene sequences. However, bacteria regularly exchange genetic material in HGT and this affects a proportion of genes in most prokaryotes [142]—there is even evidence for HGT of 16S rRNA genes [143]. If phylogenetic classification is based on the assumption that genetic material is inherited vertically, HGT creates uncertainties in the branching patterns of bacterial phyla.

Worse still, the origins of the diderm cell envelope are not yet fully understood, with several competing theories as to the origin of the cell membrane and the nature of the ancestral state: whether monoderm or diderm. Recently proposed explanations for the evolution of the diderm cell from a monoderm ancestor including antibiotic selection [144], symbiosis [145] and sporulation [28, 146]. By contrast, Cavalier-Smith has proposed that a diderm structure was the ancestral state for bacteria resulting from a key event at the dawn of cellular life in which a phospholipid vesicle engulfed a protocyttoplasm to create a cell with a double-membrane [147].

Could additional analysis of the Negativicutes help in resolving these theories? The two alternative scenarios in which the monoderm or diderm state was the original ancestral state of bacteria prompt two far-reaching evolutionary explanations for the origin of the Negativicutes. The first proposal in Figure
1.9A shows the situation if the ancestor of all present-day bacteria were diderm and monoderms are a subsequent evolutionary step. Under this scenario the Negativicutes would be an early branch of monoderm bacteria from before the outer membrane was lost. If this were the case, the Negativicutes would be the deepest branch of the Firmicutes— but there is no evidence for this in phylogenetic trees.

An alternative is that the last common bacterial ancestor was monoderm, which requires that all diderms to have evolved from a monoderm bacterium (Figure 1.9B). In this scenario the Negativicutes are an early branch of the diderm clade of bacteria. But in this situation, all diderms would nest within a sister group of the Negativicutes, which is not the case.

This leads us to conclude that the diderm nature of Negativicutes must be a derived state, resulting either from a single genome fusion event (Figure 1.9C) or from a series of HGT events (Figure 1.9D). In both scenarios, genetic material is acquired from outside the Firmicutes, but how much material would be needed to convert a Gram-positive into a Gram-negative bacterium? At a minimum, the entire repertoire of outer membrane genes would be needed for the diderm membrane. However, how could we distinguish between this minimalist scenario and a more extensive genome fusion between two different kinds of bacteria? One attractive approach would be to look for phylogenetic and compositional signals present in outer-membrane-associated genes and determine whether these point to a single phylogeny for all such genes or whether different outer membrane genes provide different signals. This approach could then be expanded to determine whether outer-membrane genes share signatures with genes associated with other functions, providing evidence for or against a single extensive genome fusion event.
Original bacterium had two membranes.
- Outer of these two membranes was lost.
- Negativicutes were formed prior to this loss.

Original bacterium had one membrane.
- Outer membrane was gained.
- Negativicutes are an early form of double membrane bacteria.

Horizontal Gene Transfer.
- The transfer included the necessary genes for outer membrane formation.

Two bacteria underwent genome fusion.
- More than just membrane genes would have been transferred.

**Figure 1.9: Possible evolutionary scenarios for Negativicutes.**

(A) The ancestral state of bacteria is diderm (pink) and Negativicutes are from just before the outer membrane is lost. (B) The ancestral state is monoderm (violet) and Negativicutes is an early diderm. (C) and (D) The event that created Negativicutes was either a Genome Fusion or a single or multiple HGT(s).
1.6. Aims and objectives

In this thesis, I aimed to address the origin and evolution of the class Negativicutes, while also shedding light on the Gram-negative and Gram-positive aspects of the group’s cell biology.

My first objective was to use sequence-based analyses to define and characterise the gene sets that encode the Gram-negative and Gram-positive toolkits present in all the Negativicutes, comparing these with similar gene sets in typical Gram-negative and Gram-positive bacteria with a view to defining the evolutionary origins of individual genes and of the Negativicutes as a whole.

The second aim was to characterise sporulation within the Negativicutes through whole-genome sequencing of *S. sphaeroides* DSM2875 (plus relevant bioinformatics analyses), twinned with the development of a reproducible experimental framework for the investigation of the transcriptional programme of events associated with sporulation.

My third aim was to investigate the transcriptomes of *V. parvula* DSM 2008 and *S. sphaeroides* DSM2875 using RNA-sequencing to better understand the contributions of the Gram-negative and Gram-positive toolkits to the cell biology of the Negativicutes. The transcriptional landscape will be characterised in the conditions of logarithmic and stationary phase to provide a dynamic view of the transcriptome.
Chapter 2
Materials and Methods

2.1. Cultivation and Bacterial Strains

All strains of Negativicutes (Table 2.1) were purchased from the DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen). Inoculated plates were incubated under the appropriate conditions (Table 2.1) in an anaerobic cabinet (Don Whitley DG250) until colonies had formed (2 to 4 days). A single colony was picked, the broth inoculated and left to grow under anaerobic conditions with shaking (IKA VXR basic Vibrax at 150 rpm) in a 50 ml plastic flask in an anaerobic cabinet. The starter culture was used to inoculate the appropriate medium to an optical density of 0.05 at a wavelength of 600 nm (OD600). This culture was grown with shaking (150 rpm) in a 50 ml plastic flask in an anaerobic cabinet.
### Table 2.1: Bacterial strains

<table>
<thead>
<tr>
<th>Species:</th>
<th>Pelosinus fermentans</th>
<th>Sporomusa sphaeroides</th>
<th>Veillonella parvula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain Number:</td>
<td>DSM17108</td>
<td>DSM2875</td>
<td>DSM2008</td>
</tr>
<tr>
<td>Notes:</td>
<td>Type strain</td>
<td>Type Strain</td>
<td>Type strain</td>
</tr>
<tr>
<td>Conditions:</td>
<td>Anaerobic</td>
<td>Anaerobic</td>
<td>Anaerobic</td>
</tr>
<tr>
<td>Temperature:</td>
<td>20°C</td>
<td>35°C</td>
<td>37°C</td>
</tr>
<tr>
<td>Incubation Time:</td>
<td>1 to 2 days</td>
<td>1 to 2 days</td>
<td>1 to 2 days</td>
</tr>
<tr>
<td>DSMZ Medium:</td>
<td>311c, 141, 320, 385</td>
<td>311, 141, 320, 385, SPM84</td>
<td>136</td>
</tr>
</tbody>
</table>

### 2.2. Bacterial Growth Media

The recipes for all media prefixed with ‘DSMZ’ were sourced from the DSMZ website [148]. Each type of media was prepared using deionised and autoclaved H₂O.

#### 2.2.1. DSMZ 311c

Autoclaved part: Yeast extract (2.00 g) and casitone (2.00 g) were dissolved into H₂O (500 ml) and autoclaved at 121°C for 15 minutes.

Filter sterilised part: K₂HPO₄ (0.35 g), KH₂PO₄ (0.23 g), NH₄Cl (0.50 g), NaCl (2.25 g), MgSO₄ x 7H₂O (0.50 g), CaCl₂ x 2H₂O (0.25 g), FeSO₄ x 7H₂O (2.00 mg), Glucose (2.00 g), NaHCO₃ (4.00 g), Resazurin (0.50 mg), Dithiothreitol* (1 mmol/L), Vitamin solution/DSMZ medium 141 (10.00 ml), Trace element solution SL-10/DSMZ medium 320 (1.00 ml) and Selenite-tungstate solution/DSMZ medium 385 (1.00 ml) were dissolved in H₂O. This was made up to a total volume of 500 ml and was filter sterilised through a filter with a pore size of 0.2 μm.

The autoclaved and filter sterilised parts were combined to make the final medium with a total volume of 1 L. If solid medium was required 1.2% Agar was added to the autoclaved part.
* Dithiothreitol was made as a 1 M stock solution and stored in the dark at -20°C.

2.2.2. **DSMZ 311**

Autoclaved part: Yeast extract (2.00 g) and casitone (2.00 g) were added to 500 ml of H₂O and autoclaved at 121°C for 15 minutes.

Filter sterilised part: K₂HPO₄ (0.35 g), KH₂PO₄ (0.23 g), NH₄Cl (0.50 g), NaCl (2.25 g), MgSO₄ x 7 H₂O (0.50 g), CaCl₂ x 2 H₂O (0.25 g), Betaine (6.70 g), FeSO₄ x 7 H₂O (2.00 mg), NaHCO₃ (4.00 g), Cysteine-HCl x H₂O (0.30 g), Sulfide (0.30 g), Vitamin solution/DSMZ medium 141 (10.00 ml), Trace element solution SL-10/DSMZ medium 320 (1.00 ml) and Selenite-tungstate solution/medium 385 (1.00 ml) were dissolved in H₂O. This was made up to a total volume of 500 ml of H₂O and was sterilised through a 0.20 µm pore filter.

The autoclaved and filter sterilised parts were combined to make the final medium with a total volume of 1 L. If solid medium was required 1.2% Agar was added to the autoclaved part.

2.2.3. **DSMZ 136**

Sodium lactate (7.5 g), Pancreatic digest of casein (5.0 g), Yeast extract (3.0 g), Tween 80 (1.0 g), Glucose (1.0 g), Sodium thioglycolate (0.75 g), Putrescine (3.0 mg) and Resazurin (1.0 mg) were dissolved into 1 L of H₂O. The solution was adjusted to pH 7.5 with solid K₂CO₃ and autoclaved at 121°C for 15 minutes.

2.2.4. **DSMZ 141: vitamin solution**

Biotin (2.00 mg), Folic acid (2.00 mg), Pyridoxine-HCl (10.00 mg), Thiamine-HCl x 2H₂O (5.00 mg), Riboflavin (5.00 mg), Nicotinic acid (5.00 mg), D-Ca-
pantothenate (5.00 mg), Vitamin B₁₂ (0.10 mg), p-Aminobenzoic acid (5.00 mg), Lipoic acid (5.00 mg) into 1 L of H₂O. The media was stirred for several hours and sterilised through a 0.20 μm pore filter.

### 2.2.5. DSMZ 320: trace element solution SL-10:

FeCl₂ x 4 H₂O (1.50 g) was dissolved into 10 ml HCl (25%; 7.7 M) and then diluted with 500 ml H₂O. The following salts were then added: ZnCl₂ (70.00 mg), MnCl₂ x 4 H₂O (100.00 mg), H₃BO₃ (6.00 mg), CoCl₂ x 6 H₂O (190.00 mg), CuCl₂ x 2 H₂O (2.00 mg), NiCl₂ x 6 H₂O (24.00 mg), Na₂MoO₄ x 2 H₂O (36.00 mg) and stirred until dissolved. Finally, the solution was made up to 1 L and sterilised through a 0.20 μm pore filter.

### 2.2.6. DSMZ 385: selenite and tungstate solution

This solution was prepared by adding NaOH (0.5 g), Na₂SeO₃ x 5H₂O (3 mg) and Na₂WO₄ x 2H₂O (4 mg) into 500 ml H₂O. The solution was sterilised through a 0.20 μm pore filter.

### 2.2.7. SPM84

The recipe for the medium to grow *S. sphaeroides* was from the first publication describing the species [41] and prepared in two parts, autoclaved and filter sterilised.

Autoclaved part: Yeast extract (2.00 g) and casitone (2.00 g) were dissolved into H₂O (500 ml) and autoclaved at 121°C for 15 minutes.

Filter sterilised part: K₂HPO₄ (0.348 g), KH₂PO₄ (0.227 g), NH₄Cl (0.5 g), MgSO₄ x 7 H₂O (0.5 g), CaCl₂ x 2H₂O (25 mg), NaCl (2.25 g), FeSO₄ x 7H₂O (2 mg), Na₂SeO₃ x 5H₂O (15 μg), NaHCO₃ (4 g), Cysteine (0.3 g) and Vitamin solution/DSMZ medium 14l (20 ml). This was made up to a total volume of 500 ml and was filter sterilised through a filter with a pore size of 0.2 μm.
The autoclaved and filter sterilised parts were combined to make the final medium with a total volume of 1 L. If solid medium was required 1.2% Agar was added to the autoclaved part.

### 2.2.8. BHI-YC

Brain-heart infusion medium (BHI) purchased from Oxoid (CMII35) was prepared according to the manufacturer’s instructions. This was supplemented with Yeast (2 g/litre) and Casitone (2 g/litre) (YC) autoclaved at 121°C for 15 minutes.

### 2.3. DNA preparation, extraction and sequencing

All protocols were carried out as per manufacturer’s instructions, unless otherwise stated. Bacterial cells were harvested ($2 \times 10^8$ cells) and the DNA extracted using the Gram-negative bacteria protocol (DNeasy Blood and Tissue protocol DNA Extraction Protocol - QIAGEN). The final elution step with EB (100 μl) was repeated, using the first elute, to increase the genomic DNA concentration. The DNA was quantitated (Qubit - Thermo Fisher Scientific), diluted to 0.2 ng/μl and 5 μl (1 ng) was used to prepare sequencing libraries (Nextera XT DNA Library Preparation Kit - Illumina). The samples were uniquely indexed with two primers Index 1 ($i7/N7XX$) and Index 2 ($i5/S5XX$) as described by the manufacturer (Table 2.2). The DNA fragment sizes and their quality were assessed (Bioanalyzer 2100 - Agilent). Instead of the Nextera XT bead normalisation method, the libraries were normalised using the standard protocol (Denature and Dilute Libraries Guide – Illumina). Each sample was diluted to 4 nm based on their fragment size, combined and denatured using NaOH (0.2 N) to prepare the final library. The final DNA library was prepared at 10 pM and sequenced using paired end technology on a version-2 500-cycle kit on the MiSeq (Illumina).
Table 2.2: Indices for DNA sequencing

<table>
<thead>
<tr>
<th>Sample</th>
<th>Index 1</th>
<th>Index 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>V. parvula</em></td>
<td>N705</td>
<td>S506</td>
</tr>
<tr>
<td><em>S. sphaeroides</em> (1st run)</td>
<td>N706</td>
<td>S501</td>
</tr>
<tr>
<td><em>S. sphaeroides</em> (2nd run)</td>
<td>N709</td>
<td>S504</td>
</tr>
</tbody>
</table>

2.4. RNA preparation, extraction and sequencing

2.4.1. Viable cell counts

All protocols were carried out as per manufacturer’s instructions, unless otherwise stated. Three biological replicates were inoculated to an OD600 of 0.05 from a starter culture. They were grown until the appropriate condition was met, either mid-logarithmic phase or stationary phase. At the required condition, the culture was serially diluted and plated onto solid media in triplicate, resulting in three biological replicates with three technical replicates per dilution. Plates were left until colonies had formed and the number of CFUs was counted to determine the viable cell count for each condition.

2.4.2. RNA sequencing

The bacterial culture was grown to the required condition and the predetermined number of cells collected (2.4.1). The RNA extraction was carried out as per the manufacturer’s protocol from the RNAprotect Bacteria Reagent Handbook (QIAGEN - December 2005, pages 41 to 43) and the miRNeasy Mini Handbook (QIAGEN - March 2014, pages 18 to 21 and 34 to 35). The RNA was extracted and treated with DNase, a PCR with 16S rRNA gene primer pairs 27f (5’AGAGTTTGATCMTGGCTCAG’) and 1492r (5’TACGGYTACCTTGTACGACTT3’) was used to check for presence of DNA:

<table>
<thead>
<tr>
<th>Reaction volumes</th>
<th>PCR Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5 μl GoTag Green (Promega)</td>
<td>95°C for 5 minutes</td>
</tr>
<tr>
<td>1 μl 16S rRNA Primer 1 – 27f</td>
<td>Then 30 cycles of:</td>
</tr>
<tr>
<td>1 μl 16S rRNA Primer 2 – 1492r</td>
<td>95°C for 30 seconds</td>
</tr>
<tr>
<td>2 μl RNA</td>
<td>51°C for 30 seconds</td>
</tr>
<tr>
<td>8.5 μl H₂O</td>
<td>72°C for 2 minutes</td>
</tr>
<tr>
<td>25 μl Total Volume</td>
<td>72°C for 10 minutes</td>
</tr>
<tr>
<td></td>
<td>4°C hold</td>
</tr>
</tbody>
</table>
The quality of the RNA and its profile with distinct 16S and 23S rRNA peaks was confirmed (Bioanalyzer - Agilent). A total of 5.0 µg of extracted RNA was depleted using RiboZERO (Illumina). The manufacturer’s instructions were followed and in the final step, the RNA was purified by ethanol precipitation for optimum recovery of small and large RNAs. The RiboZERO step resulted in a volume of 7 µl of rRNA depleted RNA. At this point the quality of the RNA was again assessed along with the depletion of rRNA (Bioanalyzer 2100) and then cDNA libraries were prepared using 5 µl of rRNA depleted RNA (TruSEQ LT - Illumina). The individual samples were identified using unique RNA Adapter Indices (Table 2.3). The cDNA fragment sizes and quality was checked (Bioanalyzer 2100) along with the cDNA concentration (Qubit). Each sample was diluted to 4 nM, then combined, denatured with NaOH (0.2 N) and diluted to 12 pM (Denature and Dilute Libraries Guide – Illumina). The library was paired end and sequenced using a version-3 150-cycle kit on the Illumina MiSeq (Illumina).

### Table 2.3: Indices used for RNA sequencing

<table>
<thead>
<tr>
<th>Sample</th>
<th>Replicates</th>
<th>Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>V. parvula</em></td>
<td>Log:1/2/3:</td>
<td>1:AR002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2:AR004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3:AR006</td>
</tr>
<tr>
<td></td>
<td>Stationary:1/2/3:</td>
<td>1:AR007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2:AR012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3:AR016</td>
</tr>
<tr>
<td><em>S. sphaeroides</em></td>
<td>Log:1/2/3/4:</td>
<td>1:AR002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2:AR004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3:AR007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4:AR0016</td>
</tr>
<tr>
<td></td>
<td>Stationary:1/2/3/4:</td>
<td>1:AR005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2:AR006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3:AR012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4:AR019</td>
</tr>
</tbody>
</table>
2.5. Bioinformatics methods

2.5.1. Genomes used for analysis

All the genomes used in this work were downloaded from the online archive of the National Center for Biotechnology Information (NCBI). Assembled and annotated genomes were used if available, otherwise unannotated contigs were annotated using Prokka (Version 1.11) [149] (Table 2.4).

Table 2.4: Genomes used for analysis.

Genomes were from the NCBI. Where completed or draft genomes were available and annotated, these were used. Otherwise, contigs were annotated using Prokka [149].

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession number</th>
<th>Contigs</th>
<th>CDS</th>
<th>Annotation</th>
<th>Genome (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaeromusa acidaminophila DSM 3853</td>
<td>NZ_ARGA00000000.1</td>
<td>64</td>
<td>3262</td>
<td>Prokka</td>
<td>3.47</td>
</tr>
<tr>
<td>Anaerococcus burkinensis DSM 6283</td>
<td>NZ_AUMI00000000.1</td>
<td>21</td>
<td>3115</td>
<td>Prokka</td>
<td>3.34</td>
</tr>
<tr>
<td>Anaerovibrio lipolyticus LB2005</td>
<td>NZ_JHYA00000000</td>
<td>32</td>
<td>2454</td>
<td>Prokka</td>
<td>2.68</td>
</tr>
<tr>
<td>Negativicoccus succinivorans DORA_17.25</td>
<td>AZMC00000000</td>
<td>356</td>
<td>1679</td>
<td>NCBI</td>
<td>1.51</td>
</tr>
<tr>
<td>Veillonella parvula DSM 2008</td>
<td>NC_013520</td>
<td>1</td>
<td>1844</td>
<td>NCBI</td>
<td>2.13</td>
</tr>
<tr>
<td>Acidaminococcus fermentans DSM 20731</td>
<td>NC_013740</td>
<td>1</td>
<td>2026</td>
<td>NCBI</td>
<td>2.33</td>
</tr>
<tr>
<td>Megasphaera elsdenii DSM 20460</td>
<td>NC_015873</td>
<td>1</td>
<td>2219</td>
<td>NCBI</td>
<td>2.47</td>
</tr>
<tr>
<td>Selenomonas ruminantium subsp. lactilytica TAM6421</td>
<td>NC_017068</td>
<td>1</td>
<td>2826</td>
<td>NCBI</td>
<td>3.00</td>
</tr>
<tr>
<td>Thermosinus carboxydivorans Norl</td>
<td>AZ_AAWL00000000</td>
<td>49</td>
<td>2750</td>
<td>NCBI</td>
<td>2.89</td>
</tr>
<tr>
<td>Megamonas funiformis YIT II815</td>
<td>NC_ADMB00000000</td>
<td>116</td>
<td>2529</td>
<td>NCBI</td>
<td>2.53</td>
</tr>
<tr>
<td>Dialister microaerophilus UPHI 345-E</td>
<td>NZ_AENT00000000</td>
<td>32</td>
<td>1310</td>
<td>NCBI</td>
<td>1.40</td>
</tr>
<tr>
<td>Phascolarctobacterium succinatutens YIT 12067</td>
<td>NZ_AENV00000000</td>
<td>118</td>
<td>2150</td>
<td>NCBI</td>
<td>2.12</td>
</tr>
<tr>
<td>Acetonema longum DSM 6540</td>
<td>NZ_AFGF00000000</td>
<td>296</td>
<td>4284</td>
<td>NCBI</td>
<td>4.32</td>
</tr>
<tr>
<td>Centipeda periodontii DSM 2778</td>
<td>NZ_AFHQ00000000</td>
<td>72</td>
<td>2631</td>
<td>NCBI</td>
<td>2.71</td>
</tr>
<tr>
<td>Anaeroglobus geminatus F0357</td>
<td>NZ_AGCH00000000</td>
<td>103</td>
<td>2148</td>
<td>NCBI</td>
<td>1.79</td>
</tr>
<tr>
<td>Pelosinus fermentans DSM 17108</td>
<td>NZ_AKVO00000000</td>
<td>65</td>
<td>4593</td>
<td>NCBI</td>
<td>4.93</td>
</tr>
<tr>
<td>Mitsukella sp. oral taxon B3I str. W9106</td>
<td>NZ_AWVT00000000</td>
<td>34</td>
<td>2309</td>
<td>NCBI</td>
<td>2.31</td>
</tr>
<tr>
<td>Succinispira mobilis DSM 6222</td>
<td>NZ_AQZZ00000000.1</td>
<td>3</td>
<td>1914</td>
<td>Prokka</td>
<td>2.09</td>
</tr>
<tr>
<td>Zymophilus raffinosivorans DSM 20765</td>
<td>NZ_ARLE00000000.1</td>
<td>66</td>
<td>3800</td>
<td>Prokka</td>
<td>4.13</td>
</tr>
</tbody>
</table>
2.5.2. Custom RefSeq BLAST database

Only genomes from RefSeq were used, because the pre-made non-redundant (nr) BLAST database included environmental samples, uncultured bacteria and microbiome projects. These are often annotated with wrongly assigned taxonomy and this metadata would impact upon the results of any analyses. The RefSeq online archive for bacteria is available at:


Within this path is “assembly_summary.txt”, a comma separated variable (CSV) from NCBI that contains the status of every bacterial RefSeq genome. This file was parsed using Python for the values found in the following columns:

| Column 8: Organism name |
| Column 11: Version status |
| Column 12: Assembly level |
| Column 20: Path to genome files |

A one species per genus BLAST database was used to prevent bias from over represented genera and greatly reduce computation time. To create a one genome per genus database, column 8 of “assembly_summary.txt” was used to randomly select a single species from each genus. This was limited to the latest version of the assembly (column 11) and a completed genome (column 12). The NCBI command line tools were then used to create the BLAST database from the files that were downloaded [150].

2.5.3. Calculation of Orthologue groups using OrthoMCL

Orthologous and paralogous protein sequences were delineated using OrthoMCL (Version 2.0) [151]. The first step in OrthoMCL was an all-vs-all BlastP search both between and within in all genomes in a pairwise manner.
OrthoMCL will then compute pairs within a MySQL database. The Markov Cluster Algorithm (MCL) was used to compute homologous groups. Default values were used throughout the OrthoMCL procedure. The output was a “groups.txt”, a space delimited text file. Column 1 described a group number for the homologue group and column 2 onwards in the same row were a list of homologues present in that group. Each row had a different group number and was a set of orthologues and paralogues and is described here as an ‘orthologue group’.

2.5.4. Delineation of the core proteome

The OrthoMCL output file was parsed to extract the core proteome where there is exactly one homologue of a protein per species. The rows containing these orthologue groups were identified using Python and the locus tags of the proteins present in the core proteome of each Negativicute were used to retrieve the required information from GenBank files such as the amino acid sequence.

2.5.5. Single and multiple next best BLAST hit(s)

To investigate the similarity of any protein of Negativicutes to either Gram-negative or Gram-positive a Next Best Blast Hit (NBBH) approach was used. Two Python scripts were written for either a single NBBH (sNBBH) or multiple NBBH (mNBBH). For a sNBBH the top hit was recorded and for mNBBH the first 50 BLAST hits were recorded. Both methods were restricted to an E-value cut-off of $1 \times 10^{-5}$. The locus tag of each hit was used to retrieve the taxonomy of the bacterial species of the hit. The bacterial taxonomy was also then defined as either Gram-negative or Gram-positive.

2.5.6. Reciprocal best BLAST hit

A Python script was written to calculate the Reciprocal Best BLAST Hit (RBBH) to identify orthologous proteins. Firstly, the query sequence was
searched for against a pre-defined BLAST database. The first 10 best BLAST hits from this search with an E-value cut-off of 1×10⁻⁵ were each reciprocally searched for against the original amino acid sequence. The best hit was the RBBH and classified as the putative orthologous protein.

2.5.7. Visualisation of synteny between Negativicutes

GenomeDiagram [152] in Biopython was used to draw the synteny between different species of Negativicutes. The script from the Biopython tutorial was used [153] as a starting point and extensively modified and expanded upon to visualise the synteny. There were 3 input requirements; the first was an OrthoMCL file that described the orthologue groups present in Negativicutes, secondly the GenBank files of each Negativicute species used in the analysis and finally the location of the synteny as defined by the contig, start and stop nucleotide positions identified using RBBH. If a bacterial chromosome was not fully resolved the orientation of the contig had to be considered.

The orthologue groups were divided into two types, the first type was part of the core proteome where exactly one copy of the orthologue is present in each species. The second type were those orthologue groups that are not in the core proteome. Each coding sequence was labelled with its locus tag, number of its orthologue group and a summary its orthologue group. The summary of the orthologue group for each coding sequence was abbreviated as follows:

- (P) Defined the number of species the orthologue was present in.
- (A) Defined the number of species the orthologue was absent in.
- (M) Defined the number of species with multiple copies of an orthologue.

Therefore, in an analysis of 19 genomes, (P19A0M0) would define a protein as a member of the core proteome as it is present in 19 species, absent in none and with multiple copies in none. However, (P18A1M2) would be outside the core proteome due to the protein being absent in one species and with
multiple copies in two species. If the protein was not present in any of the orthologue groups it would be defined as “Unique”.

2.5.8. Phylogenetic trees

Proteins that were from single orthologous group were aligned using MUSCLE [154]. For concatenated alignments of multiple sequences, each orthologous group was initially aligned separately as previously described and then a Python script written to concatenate all the individual alignments together. MEGA (Version 5.2.1) [155] was then used to calculate a Maximum-Likelihood (ML) tree from the aligned sequences. Both were bootstrapped with 100 replicates. Trees were visualised using FigTree [156] and additional metadata added through the use of a reference file.

2.5.9. Whole genome assembly and annotation of *Sporomusa sphaeroides*

SPAdes-3.6 [157] was used to generate a *de novo* assembly with k-mer values 21, 33, 55, 77, 99, 127, the recommended “careful” option was used to minimise mismatches. The properties of this assembly was then assessed with QUAST, part of the SPAdes software [158]. Contigs were filtered to those with a minimum coverage of at least 10x and minimum size of 200 bp. Prokka (Version 1.11) [149] was then used to annotate the draft genome. The reads were mapped back against the assembly with bowtie2 [159] and the read depth was calculated with Qualimap [160].

2.5.10. Variant detection in 16S rRNA sequences

A FASTA file of the nucleotide sequences of the 16S rRNA genes of *S. sphaeroides* were indexed using bowtie2-build [159] and the sequencing reads were mapped onto it with bowtie2 [159]. The aligned reads stored in a “SAM” file were then converted to binary “BAM” files, sorted and indexed using SAMtools [161]. The tool “mpileup” from SAMtools [161] was used to calculate a
file in the pileup format. The tool VarScan2 was then used with the “pileup file” to detect SNPs within the mapped sequencing reads using “pileup2snp” [162]. The following minimum thresholds were used: a read depth of 20, supporting reads of 15, a base quality of 30 and variant allele frequency threshold of 0.9.

2.5.11. RNA-seq analysis

The paired-end sequencing reads from RNA-seq experiments were mapped against the reference genome. The cDNA stranded preparation results in the first read and second read being in opposite orientations. The first step was to correct this, the tool seqtk [163] was used to flip the first read into the correct orientation for mapping and visualisation. The reference genome was indexed using bowtie2-build [159] and the sequencing reads were mapped onto it with bowtie2 [159]. The aligned reads stored in a “SAM” file were then converted to binary “BAM” files, sorted and indexed using SAMtools [161]. The sorted and indexed bam files were then used for further analysis and for visualisation of the RNA-seq data in tools such as Artemis [164] and a custom Python script (2.5.14).

2.5.12. Normalisation of RNA-seq data

A sorted BAM file (2.5.11) and a GFF (general feature format) file were inputted into the tool coverageBed [165], this calculates a file that contains has a count of the number of reads mapping to each feature of the genome. These “count files”, one per biological replicate, were used to manually normalise the RNA-seq data with the Transcripts per Million (TPM) method [166]. The package DESeq2 in R was used for analysis of differential gene expression with a negative binomial distribution model [167]. The same “count files”, one per biological replicate were used as an input for DESeq2.
2.5.13. Identification of transcriptionally active regions using toRNAdo and calculation of operons

A Python script and algorithm called toRNAdo [168], developed by a fellow doctoral student Pavelas Sazinas was used to identify transcriptionally active regions outside of coding sequences. The regions identified were intergenic, antisense and border non-coding RNAs (ncRNA), 5’ and 3’ Untranslated Regions (UTRs) and regions between coding sequences that have continuous expression. The border ncRNAs were those where expression crossed the borders of coding sequences and therefore contained properties of both antisense and intergenic ncRNAs.

A GFF file describing the locations of all the coding sequences in the genome was generated. RNA-seq reads were mapped to the genome and a sorted BAM file was calculated. The sorted BAM file and GFF file was inputted into coverageBed and genomeCoverageBed [165] to calculate files that described the coverage for each gene and the entire genome respectively at single nucleotide resolution for each strand. These were inputted into the Python script “toRNAdo”, which identified the different transcriptionally active regions. Each replicate was analysed individually and the presence of any transcriptional feature was compared between replicates.

2.5.14. Visualisation of the transcriptional landscape using Biopython and genome diagram

RNA-Seq data was visualised in Artemis [164]. However, for a more detailed and customisable view of the transcriptional landscape, a Python script was written. This used the drawing package GenomeDiagram [152] in Biopython and visualised the raw RNA-sequencing data, normalised data, Gram Value and features calculated by toRNAdo. The input files used to visualise the data were:

1. A genome annotation in the GenBank format.
2. The sequencing reads mapped to the genome, then processed with Genome Coverage Bed [165] to calculate text based coverage plots for each strand.

3. The transcriptionally active regions calculated by the toRNAdo Python script along with an additional calculation of overlapping genes used to determine operons.

4. Normalised data of the transcriptional landscape using expression units such as TPM and log₂ fold change calculated with DeSeq2 [167].

Two different chart types in GenomeDiagram were combined for the final visualisation. The graph plot was used to show the raw RNA-sequencing read coverage (line plot) and normalised data (bar graph). Alongside this, the track plot was used to visualise the coding sequences and prediction of features.

2.5.15. Identification of putative non-coding RNA sequences in other Negativicutes

The nucleotide sequence of the putative non-coding RNAs was searched for in the genomes of other species of Negativicutes with BLASTn. An E-value cut-off of 1x10⁻⁵ was used and the highest scoring hit was kept. The nucleotide position, E-value and bitscore of the hit was recorded. These putative ncRNAs were then visualised to show which had significant hits in different species.

2.5.16. Figures and statistical analysis

Microsoft Excel was used to create charts that included box plots, line graphs, pie charts and bar charts. It was also used for basic statistical analysis such as calculating TPM. Microsoft PowerPoint was used to create both figures and tables to present results. The tool Circos was used to visualise bacterial chromosomes with statistical data [169].
Chapter 3
Evolutionary origins of the Gram-negative and Gram-positive toolkits of the Negativicutes

3.1. Introduction

3.1.1. Deciphering the origins of the Negativicutes

The Negativicutes possess many of the features associated with Gram-negative bacteria, such as an outer membrane, beta-barrel proteins, LPS and a thin peptidoglycan layer. However, it is highly perplexing that the majority of housekeeping genes from this class are genetically closer to their counterparts in other members of the Gram-positive phylum Firmicutes [92, 170]. More puzzling still is the fact that several species from this class can sporulate, despite possessing a Gram-negative cell envelope.

Phylogenetic analyses of the Negativicutes using 16S rRNA gene sequences have shown the class to be related to the Clostridia [27, 92, 170]. Similar conclusions have stemmed from analysis of concatenated conserved protein sequences [92, 93, 170]. However, such analyses fail to take into account the
potential for different origins of the Gram-negative and Gram-positive features of these bacteria. Instead they have used concatenated sequences of proteins with no regard for the origin, so that distinctive phylogenetic signatures from sequences encoding Gram-negative and Gram-positive features are merged and lost. For a more sophisticated understanding of the evolutionary origins of the Negativicutes, we must adopt an approach to phylogenetics that distinguishes proteins with a Gram-negative or Gram-positive signature.

Much current debate focuses on whether the Negativicutes should form their own class or sit as a family within the Clostridia. A combination of 16S rRNA gene and protein-based phylogenies have been taken to undermine assignment of the Negativicutes to a class proposed by Marchandin et al. [27, 92]. However, a comparative genomic analysis found 14 Conserved Signature Indels as well as 48 Conserved Signature Proteins unique to the Negativicutes, supporting their assignment to a distinct class [93]. However, this debate has ignored other Gram-negative bacteria within the Firmicutes such as the Halanaerobiales, which have a lipid A biosynthesis pathway suggesting a typical outer membrane. It is worth noting that some of the Halanaerobiales—like some of the Negativicutes—can sporulate [136].

Because members of the Negativicutes have features of both Gram-negative and Gram-positive bacteria, we can assume that they have genetic toolkits from both. Delineating the Gram-negative or Gram-positive toolkits of the Negativicutes will allow us to identify other bacteria sharing these toolkits and will assist in elucidating the origin of the Negativicutes.

To understand the origins and correct taxonomic placement of the Negativicutes, we need to reconstruct the features of the most recent common ancestor of the entire clade, working on the assumption that features common to all Negativicutes must also have been present in the ancestor. These ancestral features can be found by identifying orthologous groups of proteins present in all the Negativicutes using the computational tool OrthoMCL.
However, when using orthologous groups of proteins in phylogenetic analyses, it is important to exclude potentially misleading paralogues, resulting from gene duplications. This means that efforts should focus on identifying the core proteome, where only one copy of each protein is encoded in the genome of each species that has been analysed.

Once we have built a core proteome for the Negativicutes, we can begin to investigate the phylogenetic affinities of each of the core proteins. Thus distinguishing between proteins that have a putative Gram-negative or a Gram-positive origin. This in turn will allow a much clearer phylogenetic analysis of the various components of the Negativicutes to counterparts in bacteria outside of this class. The first step in this process is to perform a BLAST search with each core protein of the Negativicutes, which will return hits from bacteria with either Gram-negative or Gram-positive taxonomies. These taxonomies should provide evidence of a Gram-positive or Gram-negative signature of origin for the protein.

In a pioneering study, Campbell et al. adopted a single-best-BLAST-hit approach to claim that Gram-negative proteins from the Negativicutes showed an association with the delta-Proteobacteria [106]. However, this assumes that the best BLAST hit is the closest phylogenetic relative, which has been shown to be an unsound inference [171]. In the study that follows, we reasoned that retrieving multiple best BLAST Hits would return a more robust result than using the single best hit, as it would be less prone to error from incorrect annotations and the larger sample size will add confidence and significance to the result. Additionally, retrieval of multiple hits enables reconstruction of a phylogenetic tree to determine if the protein in question groups more closely with proteins from Gram-negative or Gram-positive sources. A score can also be calculated based upon the number of BLAST Hits that were Gram-negative or Gram-positive for each protein.
Following the first round of BLAST searches, it will be important to adopt a reciprocal-best-BLAST-hits approach to confirm the phylogenetic relationships of key proteins from the Gram-negative and Gram-positive toolkits of the Negativicutes. Patterns of presence and absence of proteins in these toolkits will also prove important to the analyses, as bacteria that have the toolkits built from the same components as those found in the Negativicutes are more plausible as ancestors or donors.

The analyses that follow draw on genomes, tools and resources available in the public domain. The human oral commensal bacterium Veillonella parvula DSM2008 was used as the reference species for three reasons:

- *V. parvula* has a small genome size: 1.8 Mb, with 1,856 protein-coding sequences. This small genome size should still encode the most important Gram-negative and Gram-positive features but will be easier to handle and visualise in bioinformatics software.
- A complete genome sequence is available, with a fully closed chromosome. A completed genome is ideal for investigation of gene order and synteny.
- *V. parvula* can be grown readily under laboratory conditions and is therefore used later in this thesis for RNA-seq analysis.

### 3.2. Results

#### 3.2.1. Orthologue groups and core proteome of the Negativicutes

A single species from each genus of the Negativicutes was selected to calculate the core proteome, drawing on the ‘type species’, best or only sequenced genome for the genus. This approach provided 19 species for further analysis (Table 2.4). OrthoMCL [151] is an efficient program for identifying and grouping orthologous and paralogous protein sequences, first using an all-versus-all BLAST approach and then a Markov Cluster algorithm to group
paralogues and orthologues. This approach was applied to all protein-coding sequences from our test set of 19 species to identify groups of orthologous and paralogous proteins, designated here as an ‘orthologue group’. The smallest orthologue group had two paralogues in one species or two orthologues in two species, while the largest orthologue group had 58 orthologues and paralogues across all 19 species.

Figure 3.1 shows the distribution of these different groups when ordered by the number of species per group and then by the number of proteins per group. Over 90% of groups (5,007/5,473) are not represented in all species in the class and, of the 466 orthologue groups present in all species, only 341 had just one orthologue in each species. These 341 groups were taken to represent the core proteome of the Negativicutes.

**Figure 3.1: Visualisation of the orthologue groups of the Negativicutes.**

Orthologue groups were sorted by the number of species (red) then by the number of proteins (blue) found within each orthologue group. The area in green represents the core proteome where each group has 19 proteins in 19 species.
3.2.2. Core-proteome phylogeny of the Negativicutes

Orthologues from the core proteome were aligned with MUSCLE [154], concatenated and a phylogenetic tree constructed using MEGA [155]. The resulting tree recapitulates the topologies from Marchardin et al. [27] and Campbell et al. [93] derived from 16S and amino-acid sequences, but with stronger statistical support. The most divergent clade is the Veillonellaceae; two unnamed clades cluster together and then with the Acidaminococcaceae (Figure 3.2). Interestingly, the three spore-forming species cluster together in one of these unnamed clades.

Figure 3.2: ML phylogenetic tree of the core proteome.

ML tree based on the 341 proteins of the core proteome aligned with MUSCLE, concatenated and unrooted. The percentage bootstrap support is indicated on the nodes, which was based on the analysis of 100 replicates. The group of endospore forming Negativicutes is labelled. The groups are named as in Campbell et al.[93].
3.2.3. Assignment of proteins to the Gram-positive or Gram-positive toolkit using a single-best-BLAST-hit approach

A bespoke protein database was created from the NCBI Reference Sequence (RefSeq) Database by removing all sequences from Gram-negative members of the Firmicutes. Proteins from the core proteome of *V. parvula* were searched against this bespoke database using BLAST, with an E-value cut-off of ≤1x10⁻⁵. The top-scoring BLAST hits were assigned to a phylum by extracting taxonomic information from entries in the NCBI. Under these conditions, the phyla that provided the most hits were the Firmicutes, with 292 hits, and the Proteobacteria, with 18 hits (Figure 3.3). Other phyla accounted for 27 hits and four proteins did not have any hits matching the chosen E-value cut-off for the expect value.

Taxonomic assignments were then transferred from proteins to the genes that encoded them and mapped on to the genome of the host bacterium. This identified a major “Gram-negative” locus within the genome (Table 3.1 and Figure 3.4).

![Figure 3.3: Phyla associated with top-scoring BLAST hits from the core proteome of the Negativicutes.](image-url)
### Table 3.1: The major Gram-negative locus among genes encoding the core proteome of the Negativicutes.

Genes that encode the core proteome and are within the major Gram-negative locus are listed below. N/A indicates there were no hits for that protein.

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>E-Value</th>
<th>NBBH Phylum</th>
<th>Kegg Function</th>
<th>Product</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vpar_0522</td>
<td>2.00E-66</td>
<td>Cyanobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>hypothetical protein</td>
<td>MlaE</td>
</tr>
<tr>
<td>Vpar_0523</td>
<td>4.00E-82</td>
<td>Proteobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>ABC transporter</td>
<td>MlaF</td>
</tr>
<tr>
<td>Vpar_0526</td>
<td>4.00E-19</td>
<td>Proteobacteria</td>
<td>Intracellular trafficking, secretion, and vesicular transport</td>
<td>hypothetical protein</td>
<td>-</td>
</tr>
<tr>
<td>Vpar_0528</td>
<td>2.00E-13</td>
<td>Synergistetes</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>outer membrane chaperone Skp</td>
<td>OmpH</td>
</tr>
<tr>
<td>Vpar_0529</td>
<td>#N/A</td>
<td>#N/A</td>
<td>#N/A</td>
<td>hypothetical protein</td>
<td>-</td>
</tr>
<tr>
<td>Vpar_0530</td>
<td>#N/A</td>
<td>#N/A</td>
<td>#N/A</td>
<td>outer membrane chaperone Skp</td>
<td>OmpH</td>
</tr>
<tr>
<td>Vpar_0531</td>
<td>3.00E-90</td>
<td>Proteobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase</td>
<td>LpxD</td>
</tr>
<tr>
<td>Vpar_0533</td>
<td>2.00E-74</td>
<td>Proteobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>UDP-3-O-(3-hydroxyacetyl) glucosamine N-acyltransferase</td>
<td>LpxC</td>
</tr>
<tr>
<td>Vpar_0535</td>
<td>5.00E-91</td>
<td>Cyanobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase</td>
<td>LpxA</td>
</tr>
<tr>
<td>Vpar_0537</td>
<td>2.00E-71</td>
<td>Deferribacteres</td>
<td>Function unknown</td>
<td>hypothetical protein</td>
<td>Lpxl</td>
</tr>
<tr>
<td>Vpar_0538</td>
<td>6.00E-97</td>
<td>Proteobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>lipid-A-disaccharide synthase</td>
<td>LpxB</td>
</tr>
<tr>
<td>Vpar_0540</td>
<td>3.00E-90</td>
<td>Proteobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>3-deoxy-D-manno-octulosonic-acid transferase</td>
<td>KdtA/Waa A</td>
</tr>
<tr>
<td>Vpar_0542</td>
<td>5.00E-85</td>
<td>Verrucomicrobiota</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>3-deoxy-D-manno-octulosonate cytidyltransferase</td>
<td>KdsB</td>
</tr>
<tr>
<td>Vpar_0543</td>
<td>1.00E-124</td>
<td>Proteobacteria</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>2-dehydro-3-deoxyphospho-octonate aldolase</td>
<td>KdsA</td>
</tr>
<tr>
<td>Vpar_0544</td>
<td>1.00E-117</td>
<td>Proteobacteria</td>
<td>Carbohydrate transport and metabolism</td>
<td>KpsF/GutQ family protein</td>
<td>KdsD</td>
</tr>
<tr>
<td>Vpar_0545</td>
<td>1.00E-45</td>
<td>Deferribacteres</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>3-deoxy-D-manno-octulosonate B-phosphate phosphatase</td>
<td>KdsC</td>
</tr>
<tr>
<td>Vpar_0547</td>
<td>#N/A</td>
<td>#N/A</td>
<td>#N/A</td>
<td>hypothetical protein</td>
<td>-</td>
</tr>
<tr>
<td>Vpar_0548</td>
<td>2.00E-08</td>
<td>Bacteroidetes</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>OstA family protein</td>
<td>-</td>
</tr>
<tr>
<td>Vpar_0549</td>
<td>2.00E-104</td>
<td>Bacteroidetes</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>ABC transporter</td>
<td>LptB</td>
</tr>
<tr>
<td>Vpar_0550</td>
<td>9.00E-80</td>
<td>Synergistetes</td>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>YigP/YigQ family permease</td>
<td>LptF</td>
</tr>
<tr>
<td>Vpar_0551</td>
<td>7.00E-58</td>
<td>Synergistetes</td>
<td>Function unknown</td>
<td>hypothetical protein</td>
<td>-</td>
</tr>
<tr>
<td>Vpar_0552</td>
<td>2.00E-21</td>
<td>Cyanobacteria</td>
<td>Transcription</td>
<td>hypothetical protein</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 3.4: The major Gram-negative locus in the genome of *V. parvula*.

The major Gram-negative locus described in Table 3.1 from Vpar_0522 to Vpar_0552 is highlighted and labelled in blue on the chromosome of *V. parvula*. The locations of genes encoding other proteins of the core proteome are shown in black.

3.2.4. Adding BamA to the core proteome

One of the problems with the approach adopted above is that it excludes proteins from membership of the core proteome if >1 homologues are encoded in the same genome. This proved to be the case with BamA, a protein important in the biogenesis of the Gram-negative cell. BLAST searches with the *E. coli* BamA protein identified Vpar_0527 as an orthologue in *V. parvula*,
where it was annotated as “surface antigen (D15)”. This protein was a member of orthologue group 163 and was encoded by a gene that sat between two genes encoding components of the core proteome. Investigation of the orthologue group 163 identified that the genome of *Pelosinus fermentans* encodes two paralogues of BamA (FR7_0589 and FR7_1553), which led to this protein being excluded from the core proteome.

Three approaches were used to identify the “true” BamA orthologue in *P. fermentans*:

- **Gene order:** in *P. fermentans* FR7_1550 encodes a protein from orthologue group 418 and FR7_1555 encodes a protein from orthologue group 202. To match gene order with *V. parvula*, this suggests that FR7_1553 should be taken as the BamA orthologue.

- **Blast searches with Vpar_0123** (the *V. parvula* BamA) returned FR7_1553 as a reciprocal best hit. However, a similar search with the *E. coli* orthologue of BamA hit FR7_0589.

- **To resolve this,** an ML phylogenetic tree of BamA was built (Figure 3.5) and shows that FR7_1553 sits within the clade formed by other orthologues from the spore-forming Negativicutes, while FR7_0589 sits outside this clade. FR7_1553 was therefore adopted as the “correct” orthologue in subsequent studies.

The protein BamA was not the only protein to be excluded from the core proteome of the Negativicutes due to gene duplications: a total of 104 proteins were found to have orthologues that span the Negativicutes, but are not found in the core proteome, due to having more than one copy per genome (data not shown).
Figure 3.5: ML phylogenetic tree of BamA orthologues in the Negativicutes.

ML tree based on the amino acid sequences of BamA, the paralogue (red) and orthologue (green) of BamA in \textit{P. fermentans} are shown. Amino acid sequences were aligned with MUSCLE and a ML tree constructed in MEGA. The percentage bootstrap support is indicated on the nodes, which was based on the analysis of 100 replicates.

3.2.5. Taxonomic assignments on protein products of the pan-genome using multiple best BLAST hits

Up until now, I have made assignments based on the single top-scoring BLAST hit. However, there is published evidence to suggest that the closest BLAST hit is often not the nearest neighbour in phylogenetic terms [171]. In addition, in my initial analyses, only the core proteome was analysed and therefore important features, such as BamA were missed.

In subsequent studies, I therefore adopted a more sophisticated approach, in which taxonomic assignments were applied to protein products of the pan-genome using multiple best BLAST hits from a non-redundant database derived from RefSeq. This database was created manually and curated with a
Python script and included just one genome per genus and no genomes from Gram-negative Firmicutes. In my multiple-best-BLAST hits approach, the first 50 unique hits were recorded that had an E-value cut-off of less than 1x10^-5. This allowed assignment of a “Gram Value” to describe how much a protein was either Gram-positive or Gram-negative, in an analogue rather than binary form.

To calculate the Gram Value (GV), the sum of Gram-negative (GN) hits (H) with the dataset (mNBBH) is divided by the sum of Gram-positive (GP) and GN hits with the dataset. Hits to bacteria difficult to define as Gram-negative or Gram-positive, e.g. Mycobacterium [172], were excluded. With this equation a value of 1.0 indicated that all the hits were Gram-negative while one of 0.0 would indicate that all the hits were Gram-positive.

\[
\text{Gram Value} = \frac{\sum_{H \in \text{mNBBH}} H_{\text{GN}}}{\left( \sum_{H \in \text{mNBBH}} H_{\text{GN}} \right) + \left( \sum_{H \in \text{mNBBH}} H_{\text{GP}} \right)}
\]

A multiple-BLAST-hits search of protein sequences derived from the pan-genome confirmed that Gram-negative and Gram-positive toolkits of the genome are present throughout all the Negativicutes (Table 3.2). Relative proportions of proteins that have entirely Gram-negative (Gram Value of 1.00) or Gram-positive (Gram Value of 0.00) hits remain remarkably consistent throughout all species.
Table 3.2: The number of proteins in each species of the Negativicutes with entirely Gram-negative or Gram-positive BLAST hits.

The percentage is of the total number of coding sequences in the genome encoding proteins that have entirely Gram-negative (Gram Value of 1.00) or Gram-positive (Gram Value of 0.00) hits.

<table>
<thead>
<tr>
<th>Species</th>
<th>Total Coding Sequences</th>
<th>Gram-positive</th>
<th>%</th>
<th>Gram-negative</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. longum</em> DSM 6540</td>
<td>4284</td>
<td>285</td>
<td>6.7%</td>
<td>223</td>
<td>5.2%</td>
</tr>
<tr>
<td><em>A. fermentans</em> DSM 20731</td>
<td>2026</td>
<td>99</td>
<td>4.9%</td>
<td>81</td>
<td>4.0%</td>
</tr>
<tr>
<td><em>A. burkinensis</em> DSM 6283</td>
<td>3115</td>
<td>117</td>
<td>3.8%</td>
<td>146</td>
<td>4.7%</td>
</tr>
<tr>
<td><em>A. geminatus</em> F0357</td>
<td>2148</td>
<td>106</td>
<td>4.9%</td>
<td>109</td>
<td>5.1%</td>
</tr>
<tr>
<td><em>A. acidaminiphila</em> DSM 3853</td>
<td>3262</td>
<td>134</td>
<td>4.1%</td>
<td>173</td>
<td>5.3%</td>
</tr>
<tr>
<td><em>A. lipolyticus</em> LB2005</td>
<td>2454</td>
<td>134</td>
<td>5.5%</td>
<td>107</td>
<td>4.4%</td>
</tr>
<tr>
<td><em>C. periodontii</em> DSM 2778</td>
<td>2631</td>
<td>118</td>
<td>4.5%</td>
<td>162</td>
<td>6.2%</td>
</tr>
<tr>
<td><em>D. microaerophilus</em> UPII 345-E</td>
<td>1310</td>
<td>71</td>
<td>5.4%</td>
<td>81</td>
<td>6.2%</td>
</tr>
<tr>
<td><em>M. funiformis</em> YIT 11815</td>
<td>2529</td>
<td>119</td>
<td>4.7%</td>
<td>114</td>
<td>4.5%</td>
</tr>
<tr>
<td><em>M. elsdenii</em> DSM 20460</td>
<td>2219</td>
<td>87</td>
<td>3.9%</td>
<td>94</td>
<td>4.2%</td>
</tr>
<tr>
<td><em>Mitsuokella</em> sp. oral taxon I31 str. W9106</td>
<td>2309</td>
<td>83</td>
<td>3.6%</td>
<td>108</td>
<td>4.7%</td>
</tr>
<tr>
<td><em>N. succinicipris</em> DORA_17_25</td>
<td>1679</td>
<td>85</td>
<td>5.1%</td>
<td>70</td>
<td>4.2%</td>
</tr>
<tr>
<td><em>P. fermentans</em> DSM 17108</td>
<td>4593</td>
<td>305</td>
<td>6.6%</td>
<td>173</td>
<td>3.8%</td>
</tr>
<tr>
<td><em>P. succinatutens</em> YIT I2067</td>
<td>2150</td>
<td>114</td>
<td>5.3%</td>
<td>78</td>
<td>3.6%</td>
</tr>
<tr>
<td><em>S. ruminantium</em> subsp. lactilytica TAM6421</td>
<td>2826</td>
<td>140</td>
<td>5.0%</td>
<td>146</td>
<td>5.2%</td>
</tr>
<tr>
<td><em>S. mobilis</em> DSM 6222</td>
<td>1914</td>
<td>74</td>
<td>3.9%</td>
<td>74</td>
<td>3.9%</td>
</tr>
<tr>
<td><em>T. carboxydivorans</em> Norl</td>
<td>2750</td>
<td>217</td>
<td>7.9%</td>
<td>63</td>
<td>2.3%</td>
</tr>
<tr>
<td><em>V. parvula</em> DSM 2008</td>
<td>1844</td>
<td>83</td>
<td>4.5%</td>
<td>106</td>
<td>5.7%</td>
</tr>
<tr>
<td><em>Z. raffinosivorans</em> DSM 20765</td>
<td>3800</td>
<td>161</td>
<td>4.2%</td>
<td>171</td>
<td>4.5%</td>
</tr>
</tbody>
</table>

3.2.6. The Gram-negative toolkit is linked to the gamma-Proteobacteria

The multiple-best-BLAST-hits results from exclusively Gram-negative proteins were analysed to identify the ancestor of the Gram-negative toolkit from the Negativicutes. The Gram Value was calculated for 49,843 proteins and a Gram Value of 1.0 assigned to 2,279. These were then assigned to a phylum (Figure 3.6). The most abundant phylum is the Proteobacteria. The hits to the Proteobacteria were then assigned to classes and this revealed that gamma-Proteobacteria was the largest representative class (Figure 3.7). This stands in contrast to the assertion by Campbell *et al.* that the Gram-negative toolkit from the Negativicutes originated from the delta-Proteobacteria [106].

- 53 -
Figure 3.6: Summary of the phyla from the multiple best BLAST hits of the Gram-negative proteins of the Negativicutes.

The three largest phyla were Proteobacteria (68%), Cyanobacteria (10%) and Bacteroidetes (7%).

Figure 3.7: Assignment of multiple best BLAST hits within Proteobacteria.

The largest class was gamma-Proteobacteria (44%).
3.2.7. Visualising the *V. parvula* chromosome using the Gram Value

Next, a moving-average graph was drawn to visualise the Gram Value across the genome of *V. parvula* (Figure 3.8). Each data point on the X-axis shows the position of the protein-coding sequence on the genome. The Gram Value is mapped to the Y-axis and a rolling window averaged over 25 data points. Each of these 25 data points refers to the Gram-value of an individual gene. Therefore, a sliding window of 25 is the Gram-value averaged over 25 genes. The sliding window affects how the graph is visualised, the lower the value the more data that is shown but the rougher the graph, while the higher the value less data that is shown but the smoother the graph. A value of 25 was chosen as it delineated the genomes of *C. difficile* and *E. coli*. For comparison the same plots constructed for a Gram-negative bacterium (*E. coli*) and a Gram-positive bacterium (*C. difficile*) are shown. The *V. parvula* Gram Value occupied a similar position to that of *C. difficile*, averaging between 0.1 and 0.7. However, by contrast, the *E. coli* Gram Value occupied a position on the Y-axis considerably nearer 1.0, averaging between 0.7 and 1.0.

Three distinct areas can be identified on the *V. parvula* genome. Two loci stand out with a Gram Value of ~0.8, i.e. more Gram-negative than Gram-positive (Figure 3.8: 1 and 2). One of these, the major Gram-negative locus, is present in the core proteome. The other, the minor Gram-negative locus, is present in only 13/19 species analysed, i.e. is not part of the core proteome. A third locus shows a much lower-than-average Gram Value, i.e. is strongly Gram-positive (Figure 3.8: 3). This locus encodes the transcription and translation apparatus and its low Gram Value supports a Gram-positive origin for core intracellular functions of the Negativicutes. In addition to the moving-average graph, the Gram Values were visualised directly on the chromosome of *V. parvula* (Figure 3.9). Each gene was coloured from blue (Gram-negative) to red (Gram-positive) depending on the Gram Value. The plot had an enhanced resolution compared to the moving average graph: the moving average graph...
showed only minor and major Gram-negative loci whereas on the chromosome plot it was evident that genes encoding Gram-negative proteins are also present throughout the genome. The core proteome is shown on the inner ring of Figure 3.9. The orthologue groups of the Negativicutes were also visualised relative to the *V. parvula* genome (Figure 3.10).

Figure 3.8: Moving average graph of the Gram Value across the genomes of *C. difficile*, *E. coli*, and *V. parvula*.

Continued overleaf.
(A) The moving average graphs of the Gram-positive *C. difficile* (red) and Gram-negative *E. coli* (blue) occupy distinct ranges on the Y-axis (Gram Value). The background highlights the minimum and maximum ranges of *C. difficile* (light red) and *E. coli* (light blue). (B) The moving average graph of *V. parvula* showed two Gram-negative loci Minor (1) & Major (2) and one Gram-positive locus (3). The minimum and maximum ranges of Gram Values from *E.coli* (light blue) and *C. difficile* (light red) from (A) are highlighted. Data was averaged over 25 points.

**Figure 3.9:** The Gram Value visualised on the chromosome of *V. parvula*.

The larger outer ring shows the Gram Value for the chromosome and the inner ring shows it for the core proteome.
Figure 3.10: The orthologue groups of the Negativicutes relative to the V. parvula genome.

The minor Gram-negative locus is present in V. parvula, but only present in 13 other species of the Negativicutes. The Major Gram-negative locus is present in all 19 species of the Negativicutes. Each track shows the number of species of the Negativicutes that each orthologue was found in. If an orthologue group were found in all 19 species, it would be on the outermost track labelled 19. If the orthologue group were only found in 18 species it would be visualised on the next track inwards and so on. In each case only the number of species matters, not the number of orthologues, therefore the outer track is different to the core proteome.
3.2.8. Gene function within the major Gram-negative locus

Most proteins encoded in this locus have a Gram Value of 0.95, i.e. are most similar to proteins from Gram-negative bacteria. They are predominately members of the core proteome, present as single-copy orthologues in all Negativicutes analysed. Gene order in the locus was highly conserved across the Negativicutes. However, when MultiGeneBlast was used to detect homologues of the locus outside the Negativicutes [173], this returned no hits. Analyses using KEGG, BLAST and UniProt (Figure II) showed that proteins encoded in this locus form essential parts of the Gram-negative toolkit:

- **LPS biosynthesis**, including synthesis of lipid A and 2-keto-3-deoxymanno-octulosonic acid (KDO) [174]. One component of the lipid A biosynthesis pathway (LpxH) was missing. However, as reported for other Negativicutes [5, 12], another protein LpxI (Vpar_0537) capable of UDP-diacylglucosamine hydrolysis is present, offering an alternative route to synthesis of lipid A [11]. Also represented in the locus are orthologues of KdsA (Vpar_0543), KdsB (Vpar_0542), KdsC (Vpar_0545) and KdsD (Vpar_0544), which catalyse steps in biosynthesis of 2-keto-3-deoxymanno-octulosonic acid (KDO), the conserved core of LPS.

- **The BAM complex protein BamA** (Vpar_0527). This protein is part of the beta-barrel assembly machinery (the BAM complex), first described in the Cyanobacteria [175], and is critically involved in the assembly of these proteins into the outer membrane [176-179].

- **Maintenance of outer membrane lipid asymmetry** (MLA) [180], with orthologues of MlaD (Vpar_0524), MlaF (Vpar_0523) and MlaE (Vpar_0522). However, genes encoding orthologues of MlaB and MlaC could not be identified in the genome of *V. parvula*. MlaC typically delivers phospholipids from the outer membrane to MlaBDEF on the IM, to a currently unknown fate [180]. The absence of these two Mla proteins suggests functional differences between the Negativicutes and other better-studied Gram-negative bacteria.
**Figure 3.11: Detailed view of the major Gram-negative locus.**

Most of the orthologue groups were part of the core proteome. This is a more detailed view than what was shown from a core proteome perspective in Figure 3.4. The numbers in the table indicate the total number of orthologues/paralogues within a species.
3.2.9. Gene function within the minor Gram-negative locus

The minor Gram-negative locus is present in only 13/19 of species analysed. This is because genes encoding members of orthologue group 4 are absent from six species of the Negativicutes (Figure 3.13). These six species do not sit in a single clade in the Negativicutes phylogeny, so presumably loss of genes encoding members of orthologue group 4 has occurred in more than one lineage (Figure 3.12). Orthologue group 4 encompasses proteins annotated with adhesin-like attributes because they have either a *Yersinia* adhesin A (YadA) or haemagglutinin domain. These proteins both have a Gram Value of 1.0, indicating they are Gram-negative in origin. YadA was originally described in the outer membrane of the Gram-negative *Yersinia* [181]. The best described filamentous haemagglutinin is a major adhesin of the Gram-negative *Bordetella pertussis* [182]. It is unclear what aspects of the lifestyles of the six bacteria without these proteins mean that they do not require these Gram-negative adhesin functions.

Figure 3.12: Phylogenetic location of species missing orthologue group 4.
This locus also encodes components related to the two energy-coupled import systems, TolQ-TolR-TolA and TonB-ExbB-ExbD, which allow low-concentration substrates to be imported into the bacterial cell. The two proteins Vpar_0058 and Vpar_0059 were present in all the Negativicutes and are putatively homologues of ExbB/TolQ and ExbD/TolR respectively. However, an assignment of orthology to one system or the other is not possible from scrutiny of these proteins alone, as these pairs are structurally homologous to each other [183, 184]. However, this locus also encoded homologues of TonB (Vpar_0060 and Vpar_0061), beta-barrel proteins found on the OMs of Gram-negative bacteria involved in sensing signals from outside the bacterial cell [185]. In V. parvula and the other Negativicutes there were no significant hits for TolA, suggesting that the Negativicutes complex is most closely related to TonB-ExbB-ExbD.

3.2.10. Atypical Gram-negative features of the Negativicutes

Braun’s lipoprotein [186, 187] typically links the outer membrane to the peptidoglycan in Gram-negative bacteria such as E. coli. However, this protein is absent from the Negativicutes. They also lack the Tol-Pal system [188, 189] which in “typical” Gram-negatives plays a role in anchoring the outer membrane to peptidoglycan. Instead, in the Negativicute S. ruminantium the outer membrane is covalently anchored to peptidoglycan using cadaverine [98, 190, 191]. This diamine compound is in turn responsible for interactions between peptidoglycan and the SLH-domain protein Mep45 [103] (Vpar_0555 in V. parvula). This protein is part of the core proteome’s orthologue group 1 suggesting that all the Negativicutes use the same mechanism.
Table 3.2: Detailed view of the minor Gram-negative locus.

The six species highlighted in green are those missing orthologues of the proteins highlighted in green. The species with a start (*) are the sporulating Negativicutes. Proteins from *V. parvula* that are not found in an orthologue group are indicated by an “x”, because any orthologue group requires a minimum of 2 members. The numbers in the table indicate the total number of orthologues/paralogues within a species.
3.2.11. Synteny in the major Gram-negative locus

The major Gram-negative locus described above shows a remarkable level of synteny in most species of the Negativicutes, hinting at an important ancestral role for the locus. To investigate and visualise this synteny a Python script was written using the GenomeDiagram [152] module in Biopython. The region of synteny was first visualised using only genes encoding products from the core proteome (Figure 3.14A). Subsequently this was extended to syntenic genes encoding proteins in the same orthologue group (Figure 3.14B).

Figure 3.14: Synteny in the major Gram-negative locus of the Negativicutes.

Synteny in the major Gram-negative locus was first visualised (A) based only upon genes encoding products in the core proteome. However, further areas of synteny could be identified. These are shown in (B), where all orthologue groups were connected if in synteny. Genes in red encode products from the core proteome and were connected by blue if located together. Genes in green encode proteins outside the core proteome and are connected with green if their products are in the same orthologue group.
The region of synteny was split into three sections

- the region upstream of the core region (Figure 3.15)
- the core region (Figure 3.16)
- the region downstream of the core region (Figure 3.18).

Figure 3.16 shows a visual alignment of the core region of synteny. Synteny is disrupted at various points by a number of factors. BamA is in-between four genes (Figure 3.16-A) only present in the Negativicutes with flagella. Figure 3.16-B shows a cropped region unique to the two sporulating Negativicutes A. longum and P. fermentans, perhaps suggesting a shared role in sporulation and outer-membrane biosynthesis. Although only the first pair of genes (ALO_15707 and FR7_1561 encoding hypothetical proteins without any conserved domains) was linked, an amino acid alignment of the second unlinked pair (ALO_15702 and FR7_1562, encoding ABC transporters) showed 64.96% similarity. The first pair was part of orthologue group 2671, one of the 155 orthologue groups present only in the sporulating Negativicutes. The gene was also conserved in position in T. carboxydivorans (data not shown), although the major Gram-negative locus is split in that genome.

The cropped area in Figure 3.16-C shows two genes that in some cases have been fused into a single gene. The distribution of one-gene and two-gene arrangements in the core proteome phylogeny (Figure 3.17) can be explained only by multiple fusion or fission events. The two genes encode consecutive enzymes in lipid A biosynthesis: tetraacyldisaccharide 4’-kinase (LpxK) and 3-deoxy-D-manno-octulosonic-acid transferase (WaaA/KdtA). The effect of fusing these two genes is unknown.

Orthologue group 398 encodes KdsD, part of the KDO biosynthesis pathway. Although part of the core proteome, the synteny for this gene had not been maintained (Figure 3.16-D) as orthologues in A. burkinensis (A_bu_02109) and A. acidaminophila (A_ac_00602) were missing in this region but present on different contigs.
Figure 3.15: Upstream of the major Gram-negative locus.

This was the section upstream of the core region of synteny with the start of synteny (A), a transposase that has broken synteny (B), genes in synteny that were not calculated by orthoMCL (C) and also not calculated at (D) because the contig of A. longum ends at (E). The yellow arrow below is a reference point to compare with the core region of synteny (Figure 3.16). Genes in red were from the core proteome and were connected by blue if located together. Genes in green were outside the core proteome and connected with green if they were in the same orthologue group. The prefix of the locus tag is indicated after the species name.
Figure 3.16 continued overleaf.
**Figure 3.16: The core of the major Gram-negative region.**

This section is the core region of synteny in the major Gram-negative locus. BamA is shown in-between four genes (A) only present in the Negativicutes with flagella. Some genes are only found in the Negativicutes that form endospores (B). In one cropped region (C) two genes are fused in some species but separate in others. Finally, although they are in the core proteome some species do not have genes in synteny (D). The yellow and green arrow below is a reference point to compare the upstream (Figure 3.15) and downstream (Figure 3.18) sections respectively. Genes in red were from the core proteome and were connected by blue if located together. Genes in green were outside the core proteome and connected with green if they were in the same orthologue group. The prefix of the locus tag is indicated after the species name.
Several of the proteins downstream of the core region of synteny are putative S-layer proteins some in orthologue group 1 (Figure 3.18-A), others outside this group (Figure 3.18-B and Figure 3.18-C). A gene in this section unique to A. longum (Figure 3.18-D) and annotated as a hypothetical protein shows similarity on BLAST searching to glycosyltransferases involved in LPS synthesis[192].

It was not possible to establish synteny for this region in some genomes:

- In *T. carboxydovorans* Nor1, *D. microaerophilus* UPII 345-E, *Mitsuokella. sp. oral taxon 131 str. W9106, M. elsdenii DSM 20460* (the only one of these four species to have a fully resolved chromosome), the major Gram-negative locus is split, but synteny of some runs of genes is still maintained These species did not group together on a phylogenetic tree, indicating that loss of gene order occurred on more than one occasion.
- Synteny of the major Gram-negative locus could not be resolved in *A. geminatus, C. periodontii* and *N. succinicivorans* because of the large number of contigs in these unfinished genomes.

---

**Figure 3.17:** The distribution of the one-gene and two-gene configurations for orthologue group 210.

The species with the single large genes are shown in red and those with two smaller genes are shown in blue.
Acidaminococcus fermentans DSM 20731 (Acfer_)
Megamonas funiformis YIT 11815 (HMPREF9454_)
Veillonella parvula DSM 2008 (Vpar_)
Phascolarctobacterium succinututens YIT 12067 (HMPREF9443_)
Succinispira mobilis DSM 6222 (S_mo_)
Selenomonas ruminantium subsp. lactilytica TAM6421 (SELR_)
Zymophilus raffinosivorans DSM 20765 (Z_ra_)
Anaeroarcus burkinensis DSM 6283 (A_bu_)
Anaeromusa acidaminophila DSM 3853 (A_ac_)
Pelosinus fermentans DSM 17108 (FR7_)
Acetonema longum DSM 6540 (ALO_)
Anaerovibrio lipolyticus LB2005 (A_li_)

Figure 3.18 continued overleaf.
Figure 3.18: Downstream of the major Gram-negative locus of the Negativicutes.

This is the section downstream of the core region with S-layer proteins (A, B and C) that form the largest orthologue group in the Negativicutes and attach the outer membrane to peptidoglycan. A glycosyltransferase that is unique to *A. longum* (D), the end of synteny (E) because of two genes of unknown function present in *A. fermentans* (F). The green arrow below is a reference point to compare with the core region of synteny (Figure 3.16). Genes in red were from the core proteome and were connected by blue if located together. Genes in green were outside the core proteome and connected with green if they were in the same orthologue group. The prefix of the locus tag is indicated after the species name.
3.2.12. The origin of the Negativicutes

3.2.12.1. Gram-negative signature proteins

To investigate the origin of the Negativicutes, a set of “Gram-negative signature proteins” was used. These are proteins found in Gram-negative bacteria, absent from Gram-positive bacteria and found in all the Negativicutes. The patterns of presence or absence of such proteins can be used to identify the closest relatives of the Gram-negative toolkit of the Negativicutes. The use of signature proteins has been pioneered in the evolutionary study of Eukaryotes [193, 194].

As noted, the Negativicutes lack Braun’s lipoprotein [186, 187], which in “typical” Gram-negatives like E. coli links the outer membrane to the peptidoglycan. Instead, in S. ruminantium the outer membrane is covalently anchored to peptidoglycan using cadaverine [98, 190, 191] and the protein Mep45, which has a SLH (surface layer homology) domain [103]. Orthologues of Mep45 can be identified in all the Negativicutes (Table S1).

The BAM complex is an integral component of the outer membrane and is exclusive to diderm bacteria. BamA interacts with distinctive partner proteins in the BAM complex in different bacterial lineages [104]. In E. coli these partners include BamB, BamC, BamD and BamE [195]. A reciprocal best-hits approach was used in BLAST searches with components of the E. coli BAM to determine the presence or absence of BAM proteins in V. parvula. The BamA orthologue group I63 was present in all Negativicutes, but orthologues of the other partner proteins could not be found in the Negativicutes (Table S2).

Similarly, orthologues of all the proteins in the lipid A biosynthesis pathway were identified in all species of the Negativicutes (Table S3), with the proviso that analyses here and in the literature [106, III] have established that the
Negativicutes have an atypical pathway for LPS synthesis, using LpxI instead of LpxH.

The above observations were used to build a search strategy to identify the closest relatives, and therefore likely donors, of the Gram-negative toolkit of the Negativicutes based on the following criteria:

1. Absence of Braun’s lipoprotein but presence of Mep45, indicating an alternative method of attachment of the outer membrane to the peptidoglycan.
2. The presence of BamA, but no partner proteins from the BAM complex.

BLAST searches were performed with the relevant proteins against protein products derived from a set of 1178 genomes from RefSeq database, representing a cross section of bacterial taxa. These searches returned a group of 35 bacteria with the same pattern of presence and absence of Gram-negative signature proteins as the Negativicutes (Table 3.3). Proteins shown in this table were aligned using MUSCLE, concatenated and then used to create an unrooted phylogenetic tree using MEGA (Figure 3.19).

This tree recapitulated the topology for the Negativicutes seen in the core proteome phylogenetic tree (Figure 2). There was strong statistical support for the Halanaerobiales as the closest relative of the clade representing the Negativicutes. Within the Halanaerobiales there was strong statistical support for a division between the groups Halobacteroidaceae and Halanaerobium.

The bacterium *Ilyobacter polytropus* DSM 2926, representing the Fusobacteria, was calculated as the closest outgroup to both the Negativicutes and Halanaerobiales with good statistical support. Outside the Negativicutes-Halanaerobiales-Fusobacteria clade, sits a cluster that includes the
Synergistetes, delta-Proteobacteria and Nitrospirae. The Synergistetes have been recently classified as an independent phylum [196], with atypical Gram-negative features, but lacking Gram-positive features. The delta-Proteobacteria and Nitrospirae are more typical Gram-negative bacteria.

To refine the search for the closest relative of the Negativicutes’ Gram-negative toolkit, the domain architecture of BamA was examined. This protein contains a varying number of polypeptide transport associated (POTRA) domains [197] and can associate with various partner proteins. The POTRA domains and partner proteins stay relatively consistent within, but change between bacterial taxonomic groups [104]. The number of POTRA domains and partner proteins were assessed using InterPro [198] and reciprocal best BLAST hits respectively (Figure 3.20).

The Negativicutes, Halanaerobiales and Fusobacteria share a common pattern with the two or three POTRA domains and no partner proteins. This adds to the evidence that Fusobacteria share a common ancestor with the Negativicutes and Halanaerobiales. The lack of a third POTRA domain in some Negativicutes and Halanaerobiales may be explained either by occasional loss of the domain or a failure to identify it by InterPro.
### Table 3.3: The presence and absence of the Gram-negative signature proteins across all bacteria genera.

1178 genomes were analysed for the presence or absence of GNSP. The table highlights the different taxonomies of the resulting 35 bacteria. [X] indicated the presence and [O] the absence of a protein.

<table>
<thead>
<tr>
<th>Species</th>
<th>Taxonomy</th>
<th>Mep45</th>
<th>Lpp</th>
<th>LpaA</th>
<th>LpxB</th>
<th>LpxC</th>
<th>LpxD</th>
<th>LpxH</th>
<th>LpxI</th>
<th>BamA</th>
<th>BamB</th>
<th>BamC</th>
<th>BamD</th>
<th>BamE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetohalobium arabaticum DSM 5501</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halanaerobium hydrogeniformans</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halanaerobium praevalens DSM 2228</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halobacteroides haloides DSM 5130</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halotermophilum orosirii H 168</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetonema longum DSM 6540</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidaminococcus fermentans DSM 20731</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidaminococcus intestini RY-C-MR95</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaerococcus brevisenesis DSM 6283</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaeroglobus geminatus F0357</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaeromusa acidaminiphila DSM 3853</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaerovibrio lipolyticus LB2005</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centipeda periodontii DSM 2778</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dialister microaerophilus LPS 345-E</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Halanaerobiales</td>
<td>X O X X X X X O X X X O O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.19: ML phylogenetic tree from concatenated GNSP of all bacteria identified with the same feature set.

ML tree based on the amino acid sequences for BamA, LpxABCDI and Mep45 that were aligned with MUSCLE, concatenated and a ML tree constructed in MEGA. The percentage bootstrap support is indicated on the nodes, which was based on the analysis of 100 replicates. The Negativicutes are highlighted in blue, Halanaerobiales in red and Fusobacteria in green.
Figure 3.20: The POTRA domains and partner proteins of BamA.

BamA contains POTRA domains and partner proteins that stay relatively consistent within but change between bacterial phyla. These are visualised relative to the phylogenetic tree from Figure 3.19. Most domains conformed to Pfam profile PF07244 (POTRA/Blue). A few domains conformed to a different Pfam profile PF08479 (POTRA_2/Red).
3.3. Discussion

3.3.1. A more sophisticated Negativicutes phylogeny

Since the inception of the class Negativicutes, phylogenetic studies [92, 93, 138, 170] have sought to understand and refine their taxonomic position. Use of a core-proteome phylogeny defines the taxonomic structure of the Negativicutes better than use of the 16S rRNA gene sequences reported by Marchandin et al. [27]. However, the core proteome contains an undefined assortment of Gram-negative and Gram-positive proteins unique to the Negativicutes. This causes the phylogenetic signal from Gram-negative and Gram-positive proteins to be merged and lost. Thus, the taxonomic placement and evolutionary origins of the Negativicutes cannot be explained without taking into account both their Gram-negative and Gram-positive toolkit. Campbell et al., have also articulated a similar view, but this has not been led to an informed phylogenetic analysis that recognises the dual origins of the Negativicutes. The best available taxonomic workaround has been a compromise, placing the Gram-negative Negativicutes in a separate class within the Gram-positive Firmicutes.

Here, I have developed the Gram Value, a statistic that enabled visualisation of the Gram-negative and Gram-positive features of the Negativicutes genomes. This showed that all the Negativicutes had a major Gram-negative locus that encoded key proteins of the outer membrane. Additionally, this major Gram-negative locus was also a region of synteny in most Negativicutes species. A second minor Gram-negative locus was present in V. parvula, encoding components mostly shared with other Negativicutes. However, these regions did not account for the entire Gram-negative toolkit of the Negativicutes: the genomes were a mosaic of genes with strong Gram-negative and Gram-positive phylogenetic signatures mixed with those that were less distinctive.

It has been suggested that the Negativicutes placement as a class is incorrect [92]. However, the region of synteny unique to the Negativicutes is the
strongest possible argument that all bacteria in this class had a monophyletic origin. In addition to this region, their position as a monophyletic taxonomic group was also established using a core proteome phylogenetic tree (Figure 3.2). These findings are in agreement with a published core genome phylogenetic tree and with Conserved Signature Indels and Conserved Signature Proteins [93].

Analyses of the genomes and a review of the literature found that all the Negativicutes had the same presence-or-absence pattern for Gram-negative signature proteins. These were used to discover extant Gram-negative bacteria with the same set of Gram-negative signature proteins. Restricting the bacteria to those with the same Gram-negative features as the Negativicutes provided a robust and informed basis to undertake a phylogenetic analysis of the Gram-negative toolkit. The amino acid sequences of Gram-negative proteins from these bacteria were used to reconstruct the first Gram-negative phylogenetic tree for the Negativicutes. This showed that Halanaerobiales and the Negativicutes shared a common origin and provided robust evidence that the species *I. polytropus* from the Gram-negative phylum Fusobacteria shared a most recent common ancestor with the Gram-negative toolkit of the Negativicutes. This hypothesis is strengthened by evidence that the Negativicutes, Halanaerobiales and *I. polytropus* BAM complex share the same partner proteins and POTRA domains.

### 3.3.2. Defining the Gram-negative and Gram-positive toolkits

Before these investigations, phylogenetic analyses of the Negativicutes have focused on the Gram-positive aspects of their genomes, such as the 16S rRNA gene. However, while this research was ongoing, a comparative proteome analysis was published on the outer membrane of the Negativicute *Acidaminococcus intestini* [106], recording the phylogeny of the top BLAST hit outside the Negativicutes for each protein in the genome. This suggested that 7% of the *A. intestini* genome is most closely related to the Gram-negative
delta-Proteobacteria. However, this approach is flawed, as the closest BLAST hit is not necessarily the nearest evolutionary neighbour [171]. In this thesis, a multiple best BLAST hits approach appeared to link the Gram-negative toolkit of the Negativicutes to the gamma-Proteobacteria. However, neither of these taxa have the same BAM complex as the Negativicutes (Figure 3.20) or the same atypical Gram-negative toolkits (Table 3.3). These differences appear to rule out the Proteobacteria as the donor of a Gram-negative toolkit to the Negativicutes. Instead, it seems likely that the prevalence of Proteobacteria in the BLAST hits is a result of their over-representation in sequenced genomes. Also, using BLAST hits to determine whether a gene was Gram-negative or Gram-positive may be valid, but using this to inferring taxonomic relationships is probably not.

3.3.3. The origin of the Negativicutes

The origin of the Negativicutes must have been ancient; this is indicated by the variety of environments they have adapted to and their wide geographical distribution. Different species are present in many invertebrate and vertebrate microbiomes. They have diversified to many environmental niches, as well as evolving to survive in both extreme cold and heat.

Four possible scenarios were originally hypothesised for the origin of the Negativicutes. The first was that the Negativicutes were ancestral to all Gram-negative bacteria and therefore the first bacteria to gain an outer membrane. However, the research reported here shows that relatives of the Gram-negative toolkit of the Negativicutes are restricted to only a small subset of Gram-negative bacteria. The phylogenetic tree described here, built from this subset of Gram-negative bacteria, showed the Negativicutes nested deep within the tree, rather than sitting as the earliest branch (Figure 3.19), allowing us to discount this hypothesis.

The second hypothesis suggests the Negativicutes are ancestral to all Gram-positive bacteria. In this scenario, the outer membrane from a Gram-negative
bacterium was lost to give rise to a Gram-positive bacterium. However, this would require that all Gram-positive bacteria would sit within the same clade as the Negativicutes, and this is simply not the case [27, 92, 138, 170].

The final two scenarios implicate a transfer of genomic material

- by multiple piecemeal horizontal gene transfer events or
- by a single large event, which probably equates to a genome fusion.

These scenarios would result in a bacterium with distinct Gram-negative and Gram-positive toolkits. However, the evidence is against multiple episodes of horizontal gene transfer as the complexity of the outer membrane suggests transfer of the necessary components in a piecemeal fashion would be difficult. A more dramatic event, such as a genome fusion, is more plausible for the transfer of the outer membrane and BAM complex into the Negativicutes. This is backed up by the finding that the Gram-negative proteins are encoded throughout the genome of the Negativicutes, in addition to a single region of synteny.

Using the Gram Value, the Gram-negative and Gram-positive features of the Negativicutes have been delineated (Table 3.2 and Figure 3.9). The delineation of these two gene sets, Gram-negative and Gram-positive, proposes they have two different origins. Firstly, the HGT scenario can explain the presence the Gram-negative gene set in the genome of the Negativicutes. Secondly, a genome fusion can resolve the presence of Gram-negative genes located throughout the genome.

The unique set of Gram-negative signature proteins has also helped identify the candidate phyla that contributed these Gram-negative features. All of these candidates are Gram-negative bacteria unrelated to either Firmicutes or Clostridia. The phylogenetic tree reconstructed from the Gram-negative signature proteins identified a clade containing the Negativicutes, Halanaerobiales and I. polytropus. In addition, the Negativicutes,
Halanaerobiales and *I. polytropus* have a BAM complex with the same number of POTRA domains and no partner proteins. The similarity of domain structure of BamA supports the phylogenetic tree that and is evidence that the Gram-negative features of these three clades are related. The best scenario to explain this is that the Gram-negative features of Negativicutes arrived by HGT from an ancestor of *I. polytropus*.

The Gram-value and Gram-negative signature proteins both add strength to the argument that the origin of Negativicutes was a large scale HGT from a genome fusion. However, these results also infer a common origin to Negativicutes and Halanaerobiales which is not mirrored from phylogeny reconstructed using ribosomal proteins [138]. There is no clear way to resolve this dilemma. However, any phylogeny is dependant upon the species that are included. The phylogeny reconstructed here, from Gram-negative signature proteins, includes a clear and robust rational about the species included in the analysis.

### 3.3.4. Halanaerobiales and Fusobacteria

Many lines of evidence in this research point to the origin of the Negativicutes and Halanaerobiales as resulting from the same event. As sporulating Gram-negative bacteria are so rare, the event that created them must also be rare and probably has occurred only once. Previous work has established that the Negativicutes and Halanaerobiales should both be phylogenetically placed within Clostridia according to their Gram-positive genes [27, 92, 138, 170]. The Negativicutes and Halanaerobiales also both share the same pattern of presence and absence of Gram-negative signature proteins (Table 3.3) and components of the BAM complex (Figure 3.20). A phylogenetic tree reconstructed from these Gram-negative signature proteins confirmed a single monophyletic origin (Figure 3.19). These similarities and the phylogenetic tree support a single rare event gave rise to both the Negativicutes and Halanaerobiales.
Fusobacteria are Gram-negative obligately anaerobic non-spore-forming bacilli [199, 200]. This anaerobic phenotype matches that of the Negativicutes suggesting that donor and recipient survive in the same environments, which is a prerequisite for horizontal gene transfer. This research has also shown that a bacterium in the genus *Ilyobacter* and the family Fusobacteriaceae share the same Gram-negative signatures as the Negativicutes and Halanaerobiales. It would be incredibly unlikely these similarities occurred by chance. This supports a sophisticated phylogenetic analysis that components from these three groups of bacteria share a common Gram-negative ancestor. It has also been reported that Fusobacteria and the Negativicutes were also linked by a Conserved Signature Indel in the HSP60/GroEL protein that mediates protein folding [144]. An genomic investigation of both the genus *Ilyobacter* and the family Fusobacteriaceae would shed further light on this relationship, including how much of their genomes are related to Negativicutes outside of the currently defined Gram-negative signature proteins.

### 3.3.5. Future work

The relationship of the Negativicutes, Halanaerobiales and Fusobacteria needs to be explored further. Why was the major Gram-negative locus not found in any other bacteria using MultiGeneBLAST [173] as its presence has since been reported in Halanaerobiales [201]. Exploring these relationships will help to further understand and strengthen the arguments for the origin of the Negativicutes. Another question not yet answered is the origin of sporulation in the Negativicutes. Did the bacterium that gave rise to the Negativicutes have the ability to sporulate? Or did the sporulating Negativicutes gain this feature in a separate event? Some proteins found only in sporulating Negativicutes are encoded in the major Gram-negative region, suggesting that they are prime targets for investigating sporulation in a Gram-negative bacterium. Similarly, Gram-positive signature proteins could be used to narrow down the closest relatives and likely donors of the Negativicutes’ Gram-positive toolkit, which could shed light on whether the Gram-positive ancestor was a sporulating bacterium.
Chapter 4

Genome analysis of the anaerobic endospore forming Gram-negative
*Sporomusa sphaeroides*

4.1. Introduction

4.1.1. *Sporomusa sphaeroides*

*Sporomusa* is a genus of Gram-negative obligate anaerobes in the class
Negativicutes. Members of this group have been found in the inanimate
environment [41, 83, 202] and in insect guts [82]. The name is derived from an
ability to form endospores ‘Sporo’ and their curved shape with tapered ends
‘musa’ as in banana. The species *S. sphaeroides* was originally isolated from a
German riverbed in 1984 [41] alongside *Sporomusa ovata*, for which the
genome is already published [203]. *S. sphaeroides* has been shown to form
terminal heat-resistant endospores in the late-exponential growth phase in
betaine containing media [41]. *Sporomusa* species have previously been shown
to group with other endospore-forming Negativicutes by 16S rRNA and
concatenated protein analyses.
4.1.2. **Rationale for genome-sequencing *Sporomusa sphaeroides***

To investigate sporulation in a Gram-negative bacterium, I wished to perform RNA-Seq. However, this approach relies on the availability of a species that could be grown easily and made to sporulate in the laboratory. For this purpose, I selected *Sporomusa sphaeroides*. However, a prerequisite for a reference-based approach to RNA-Seq is the availability of a genome sequence, yet none was available for this species. I therefore set out to obtain a draft genome sequence of *Sporomusa sphaeroides*.

4.1.3. **Nextera sequencing**

The Illumina MiSeq relies on a Sequencing-by Synthesis approach. This requires the preparation of a sequencing library. Here, I used the Nextera XT sequencing kit to prepare the library. With this approach, first, a transposase is used simultaneously to fragment the DNA and add a specific adaptor sequence. Unique barcodes are added to the fragmented DNA with these adaptors, which allows multiple genomes to be sequenced together, using the barcodes to identify each genome. Once the DNA library has been loaded onto the MiSeq, the template is immobilised on a lawn of primers on the flow cell. The first step is to amplify the DNA into clusters for sequencing. As sequencing proceeds fluorescently labelled reversible terminators are added for each dNTP and then imaged. Each terminator is cleaved and then the next base is added. Paired-end sequencing allows for both ends of the fragment to be sequenced. The positional information and paired reads allows for better alignment when assembled.

4.1.4. **Sequence assembly and annotation**

After the DNA has been sequenced, the reads are delivered in a FASTQ file (FASTA with quality scores). There are two FASTQ files carrying paired forward and reverse reads. As no reference sequence exists for *S. sphaeroides*,...
the genome was assembled *de novo* with SPAdes [157], which deliver large continuous sequences or contigs. Mistakes, missing data, polymorphisms and repetitive sequences mean that the genome does not assemble into a single contig. The genome was annotated using Prokka [149] to identify features of interest such as coding sequences for proteins and ribosomal RNAs (rRNA). Standard output files in GenBank format were generated.

### 4.1.5. Genome analysis

The draft genome sequence will be a starting point to investigate the transcriptome and will offer insights into the lifestyles of a sporulating Gram-negative bacterium. The tools built and used in Chapter Three will help to interpret the genomic data beyond the Prokka annotation. Features of *S. sphaeroides* to be investigated include the genes involved in flagellar biosynthesis and sporulation, two attributes absent from *Veillonella parvula* and some other Negativicutes. The reciprocal best BLAST hits approach will be used to can detect presence/absence of key Gram-positive and Gram-negative genes, including a published list of genes defining the toolkit for sporulation [117]. As in Chapter Four, a multiple best BLAST hits approach will be used to identify phylogenetic signatures and to calculate the Gram Value for each gene. OrthoMCL will identify the core proteome of Negativicutes and the major Gram-negative locus in *S. sphaeroides* will be investigated.

### 4.2. Results

#### 4.2.1. First sequencing run

After the first sequencing run a draft genome of *S. sphaeroides* was generated using 895,720 paired-end reads (98.05% of the sequenced reads). The read lengths were between 35 bp and 251 bp, with a mean read length of 243 bp and mean insert size of 333 bp. The assembled draft genome was 9,873,628 bp with a GC content of 47.22 with 21.57-fold mean genome coverage. Coverage was lower than expected for several reasons:
• Most of the reads from the run belonged to other samples
• The apparent genome size was much larger than expected compared to other species of Sporomusa.
• The culture was contaminated. Several highly conserved proteins, such as the β subunit of bacterial RNA polymerase (RpoB), were found and BLAST search hits identified. While amino acid sequences for some proteins matched the class Negativicutes, others matched the species Paenibacillus borealis [204]. The coverage across reference (Figure 4.1) also indicated that two genomes with differing levels of coverage were sequenced.

*P. borealis* is a Gram-positive, facultative anaerobe that grows in spruce forest humus and was isolated in Finland. It is also forms endospores and many of its observable characteristics matched *S. sphaeroides*. However, as it was Gram-positive, it could be selected against using vancomycin before re-sequencing.

**Figure 4.1:** Coverage histogram of the first sequencing run.

Coverage histogram of the first sequencing run showing two distinct peaks (1 and 2). The two peaks suggested a mixed culture was sequenced.
4.2.2. Second sequencing run

After selection with vancomycin, a single colony was picked, grown and sequenced so that a draft genome of *S. sphaeroides* was generated (Table 4.1, Table 4.2 and Figure 4.2). The Gram Value predicted 317 highly Gram-positive and 206 Gram-negative proteins.

### Table 4.1: Summary of the sequencing data.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number paired-end reads</td>
<td>3,418,350</td>
</tr>
<tr>
<td>Mean insert size in bases</td>
<td>344.88</td>
</tr>
<tr>
<td>Read length in bases</td>
<td>229.79</td>
</tr>
<tr>
<td>Mean fold genome coverage</td>
<td>158.7</td>
</tr>
<tr>
<td>Largest Contig number of bases</td>
<td>485827</td>
</tr>
<tr>
<td>Smallest Contig number of bases</td>
<td>209</td>
</tr>
<tr>
<td>N50</td>
<td>155162</td>
</tr>
</tbody>
</table>

### Table 4.2: Summary of the Prokka annotation.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA, total number of bases</td>
<td>4943689</td>
</tr>
<tr>
<td>DNA coding number of bases</td>
<td>4293334 (86.8%)</td>
</tr>
<tr>
<td>DNA GC number of bases</td>
<td>2333551 (47.2%)</td>
</tr>
<tr>
<td>DNA Contigs</td>
<td>118</td>
</tr>
<tr>
<td>Genes total number</td>
<td>4694</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>4570</td>
</tr>
<tr>
<td>rRNA genes</td>
<td>12</td>
</tr>
<tr>
<td>5S rRNA gene</td>
<td>12</td>
</tr>
<tr>
<td>16S rRNA gene</td>
<td>0</td>
</tr>
<tr>
<td>23S rRNA gene</td>
<td>0</td>
</tr>
<tr>
<td>tRNA genes</td>
<td>111</td>
</tr>
<tr>
<td>other RNA genes (tmRNA)</td>
<td>1</td>
</tr>
<tr>
<td>Number of Hypothetical Proteins</td>
<td>1169</td>
</tr>
<tr>
<td>Pseudo Genes</td>
<td>0</td>
</tr>
<tr>
<td>Proteins with 1.0 Gram Value (Gram-negative like)</td>
<td>206</td>
</tr>
<tr>
<td>Proteins with 0.0 Gram Value (Gram-positive like)</td>
<td>317</td>
</tr>
</tbody>
</table>
Figure 4.2: The draft genome of *S. sphaeroides*.

Genes are coloured based upon their Gram Value; those without any BLAST hits are black. Contigs are ordered by size, the contig number is indicated on the outside. The tracks from outside to inside:

**Track 1 and Track 2: Gram Value of genes on both strands**
- These are the two coding stands, although as the genome is draft the orientation has not been resolved.

**Track 3: Proteins of the core genome with GRAM VALUE**
- The locations of the genes of the core proteome are shown on this track, their colour indicates the GRAM VALUE.

**Track 4: Proteins that had either a Gram Value of 1.0 or 0.0**
- These are gene locations of proteins that had either entirely Gram-positive (red) or Gram-negative multiple NBBH.

**Track 5: Location of the signature sporulation genes**
- The genomic locations of proteins that make up the sporulation signature are shown in Figure 4.7.

**Track 6: GC-skew**

**Track 7: GC-content**
4.2.3. The 16S ribosomal RNA genes

The 16S ribosomal RNA (rRNA) gene was not assembled by SPAdes and was absent in the Prokka genome annotation of *S. sphaeroides*. The lack of 16S rRNA genes in the assembly was checked manually by repeating the RNAmmer [205] step that Prokka uses to detect rRNA: only 5S rRNA sequences were identified. There were 17 nucleotide sequences for the 16S rRNA gene of *S. sphaeroides* in the NCBI database. Any differences between the sequenced gene and copies of the gene present in the NCBI nucleotide archive were checked. The mpileup command of SamTools was used along with VarScan to call any single nucleotide polymorphisms (SNPs). The parameters were a minimum read depth of 20, a minimum of 15 supporting reads, a minimum average base quality of 30 and 90% of reads supporting the call. Seven of the 16S rRNA nucleotide sequences were identical to those in the draft genome sequence. The nucleotide sequence from the original *S. sphaeroides* paper showed two SNP differences (Table 4.3) The reads were aligned to one of the identical 16S rRNA genes to calculate a coverage of 2,292-fold. The coverage of the 16S rRNA gene was divided by the coverage of the genome (159) to predict that 14 copies of the 16S rRNA gene exist in the genome.

4.2.4. Phylogenetic analyses

The core proteome was identified using the orthologue groups calculated with OrthoMCL [151] from *S. sphaeroides* and Negativicutes species (Table 2.4). The core proteome had 317 proteins that were used to reconstruct a phylogenetic tree (Figure 4.3). Genes encoding proteins with absolute Gram-negative or Gram-positive values (Figure 4.4) were analysed by GC-content and codon usage. Median GC-content in Gram-positive genes was 48.97% compared to 45.95% for Gram-negative genes. However, no meaningful difference was noted in codon usage between proteins with Gram-positive and Gram-negative phylogenetic signatures (Figure 4.5).
Table 4.3: Comparison against the 16S rRNA nucleotide sequences from NCBI.

The number of SNPs found when the sequencing reads are mapped against each of the 17 S. sphaeroides 16S rRNA nucleotide sequences found in the NCBI nucleotide archives.

<table>
<thead>
<tr>
<th>GI</th>
<th>Accession.Version</th>
<th>Sporomusa sphaeroides</th>
<th>Number</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>6822265</td>
<td>AJ279801.1</td>
<td>16S rRNA gene, strain DSM 2875 Type</td>
<td>2</td>
<td>1165(A-&gt;G/99.89%) 1413(G-&gt;A/100%)</td>
</tr>
<tr>
<td>219878278</td>
<td>NR_025417.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>2</td>
<td>1165(A-&gt;G/99.89%) 1413(G-&gt;A/100.00%)</td>
</tr>
<tr>
<td>397174440</td>
<td>HE966426.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 20</td>
<td>1</td>
<td>1303(A-&gt;C/100.00%)</td>
</tr>
<tr>
<td>397174441</td>
<td>HE966427.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 17</td>
<td>1</td>
<td>1128(T-&gt;C/100.00%)</td>
</tr>
<tr>
<td>397174442</td>
<td>HE966428.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 15</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>397174443</td>
<td>HE966429.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 13</td>
<td>1</td>
<td>41(G-&gt;A/100.00%)</td>
</tr>
<tr>
<td>397174444</td>
<td>HE966430.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 18</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>397174445</td>
<td>HE966431.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 14</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>397174446</td>
<td>HE966432.1</td>
<td>partial 16S rRNA gene, strain DSM 2875, clone 16</td>
<td>1</td>
<td>632(A-&gt;G/91.56%)</td>
</tr>
<tr>
<td>645320461</td>
<td>NR_117662.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>1</td>
<td>1303(A-&gt;C/100%)</td>
</tr>
<tr>
<td>645320462</td>
<td>NR_117663.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>1</td>
<td>1128 (T-&gt;C/100%)</td>
</tr>
<tr>
<td>645320464</td>
<td>NR_117664.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>645320465</td>
<td>NR_117665.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>1</td>
<td>41 (G-&gt;A/100.00%)</td>
</tr>
<tr>
<td>645320466</td>
<td>NR_117666.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>645320468</td>
<td>NR_117667.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>645320469</td>
<td>NR_117668.1</td>
<td>strain DSM 2875 16S ribosomal RNA gene, partial sequence</td>
<td>1</td>
<td>632 (A-&gt;G/91.56%)</td>
</tr>
<tr>
<td>930697316</td>
<td>LN880090.1</td>
<td>partial 16S rRNA gene, clone W2S14</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.3: Core proteome based phylogenetic tree of Negativicutes.

Maximum Likelihood tree including *S. sphaeroides* based on the 317 proteins of the core proteome aligned with MUSCLE and concatenated and unrooted. The percentage bootstrap support is indicated on the nodes, which was based on the analysis of 100 replicates. The group of endospore forming Negativicutes is highlighted in green and the location of *S. sphaeroides* is in blue. The groups are named as in Campbell *et al.*[93].

Figure 4.4: Comparison of the GC-content of Gram-negative and Gram-positive genes.

Box and whiskers plot comparing the GC-content of Gram-negative and Gram-positive genes. Gram-negative proteins were defined as those with an absolute Gram Value of 1.0 and Gram-positive a Gram Value of 0.0.
Figure 4.5: Codon usage of Gram-negative and Gram-positive proteins.

The codon usage of Gram-negative proteins (blue) and Gram-positive ones (red) compared. Gram-negative proteins were defined as those with an absolute Gram Value of 1.0 and Gram-positive a Gram Value of 0.0.
4.2.5. The Gram-negative toolkit of *S. sphaeroides*

Analysis of the genome identified the Beta-barrel Assembly Machinery protein BamA (SSPH_02650). However, as expected, none of the other BamBCDE proteins found in *E. coli* were present in *S. sphaeroides*.

Also as expected, the biosynthetic pathway for the lipid A was also identified, but LpxH was absent. As noted, the LpxI protein can replace the functionality of LpxH, [110]. The gene encoding LpxI is located between *lpxA* and *lpxB* on the genome in *Caulobacter crescentus*. BLAST searches show a potential orthologue for LpxI (SSPH_02637) in *S. sphaeroides* in the same genetic location between *lpxA* and *lpxB*. Like other Negativicutes *S. sphaeroides* lacks an orthologue of Braun’s lipoprotein [186, 187] but does encode an orthologue of Mep45, SSPH_02620 in line with expectations from Chapter Three.

The region of synteny found in other Negativicutes in Chapter Three was found intact on contig 13 between SSPH_02621 and SSPH_02657 in *S. sphaeroides*. When the synteny of this locus is first compared to the other endospore-forming Negativicutes, *S. sphaeroides* appears to be missing the *lpxK* gene from orthologue group 1503, which is present in *A. longum* (ALO_15662) and *P. fermentans* (FR7_1570) (Figure 4.6). However, *lpxK* is not in fact missing from *S. sphaeroides*: instead, it is fused with the gene *waaA* (SSPH_02634), which is directly upstream. The gene *waaA* is responsible for previous step in lipid A biosynthesis pathway. However, the consequences of the fusion of these two genes (*waaA* and *lpxK*) remain unknown.
Figure 4.6: The major Gram-negative locus of synteny in Negativicutes.

The major Gram-negative locus was identified in *S. sphaeroides* between SSPH_02657 and SSPH_02621 on contig 13. Synteny is shown between the sporulating Negativicutes *S. sphaeroides*, *A. longum* and *P. fermentans*. The single difference is indicated with an arrow where *S. sphaeroides* has the two genes of other Negativicutes fused together.
4.2.6. The sporulation toolkit of *S. sphaeroides*

The original description of *S. sphaeroides* included an ability to sporulate [41]. A genomic signature of sporulation, comprising III proteins from *B. subtilis* sporulation toolkit which represent the minimal machinery for sporulation [117] was used to assess whether *S. sphaeroides* could sporulate. BLAST searches were performed of representative genomes of Negativicutes including *S. sphaeroides*. A reciprocal best blast hit approach was used; a positive result was a hit with an E-value of less than $1 \times 10^{-10}$ and negative if there were no hits.

Four genera of Negativicutes are predicted to sporulate based upon the III proteins forming the signature of sporulation (Figure 4.7). The percentage of proteins present is summarised in Table 4.4. Three of these bacteria *A. longum, P. fermentans* and *S. sphaeroides* have been previously shown to sporulate under laboratory conditions [33, 53]. Curiously, *T. carboxydovorans* is predicted to sporulate by this analysis, but this was not originally described in the literature [85]. However, the presence of sporulation genes has since been noted [92].

The genome of *S. sphaeroides* included 95% of the key sporulation genes [117], which correlated with its ability to form endospores. Three of the proteins absent in *S. sphaeroides* (DivIB, Nfo and YqxM) were also absent in the other endospore forming Negativicutes (Figure 4.7). DivIB is required in *B. subtilis* for correct formation of a polar septum [206]. Nfo (End4) is an endonuclease that has been shown to protect the spores of *B. subtilis* from accumulation of DNA damage during dormancy [207]. The absence of these two proteins in *Sporomusa* and other endospore forming Negativicutes suggest either a different mechanism or they are not required in *Sporomusa*. The final protein, YqxM is involved in the biofilm matrix of *B. subtilis* [208], but there are no reports so far that the Negativicutes can form biofilms.
About half of the minimal machinery for sporulation was also identified in non-spore formers such as *E. coli* and *V. parvula*. This was also observed in the original publication and allowed for a core genomic signature of about 50 proteins that could be used to identify sporulating bacteria. These refined signature proteins are found in 90% of spore formers but no more than 10% of non-spore formers. [117]. The non-core proteins while required for sporulation must presumably have other functions within both spore formers and non-spore formers.

**Table 4.4: The minimal machinery for sporulation in the Negativicutes.**

The 111 proteins that make up the sporulation toolkit[117] were searched for in Negativicutes. The number of these proteins present in each species (count) and percentage is shown from Figure 4.7. The endospore forming Negativicutes are highlighted in green. *E. coli* and *B. subtilis* are included as negative and positive controls respectively.

<table>
<thead>
<tr>
<th>Species</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. subtilis</em> DSM 6540</td>
<td>111</td>
<td>100%</td>
</tr>
<tr>
<td><em>A. longum</em> DSM 17108</td>
<td>107</td>
<td>96%</td>
</tr>
<tr>
<td><em>P. fermentans</em> DSM 2875</td>
<td>106</td>
<td>95%</td>
</tr>
<tr>
<td><em>T. carboxydvorans</em> Nor1</td>
<td>104</td>
<td>94%</td>
</tr>
<tr>
<td><em>E. coli</em> K-12 (negative control)</td>
<td>56</td>
<td>50%</td>
</tr>
<tr>
<td><em>A. acidaminophila</em> DSM 3853</td>
<td>70</td>
<td>63%</td>
</tr>
<tr>
<td><em>A. burkinensis</em> DSM 6283</td>
<td>70</td>
<td>63%</td>
</tr>
<tr>
<td><em>A. fermentans</em> DSM 20731</td>
<td>60</td>
<td>54%</td>
</tr>
<tr>
<td><em>A. geminatus</em> F0357</td>
<td>62</td>
<td>56%</td>
</tr>
<tr>
<td><em>A. lipolyticus</em> LB2005</td>
<td>66</td>
<td>59%</td>
</tr>
<tr>
<td><em>C. periodontii</em> DSM 2778</td>
<td>62</td>
<td>56%</td>
</tr>
<tr>
<td><em>D. microaerophilus</em> UPII 345-E</td>
<td>53</td>
<td>48%</td>
</tr>
<tr>
<td><em>M. elsdieni</em> DSM 20460</td>
<td>66</td>
<td>59%</td>
</tr>
<tr>
<td><em>M. funiformis</em> YIT II815</td>
<td>59</td>
<td>53%</td>
</tr>
<tr>
<td><em>M. sp. oral taxon</em> 131 str. W9106</td>
<td>61</td>
<td>55%</td>
</tr>
<tr>
<td><em>N. succiniciproducens</em> DORA_17_25</td>
<td>58</td>
<td>52%</td>
</tr>
<tr>
<td><em>P. succinatutens</em> YIT 12067</td>
<td>59</td>
<td>53%</td>
</tr>
<tr>
<td><em>S. mobilis</em> DSM 6222</td>
<td>60</td>
<td>54%</td>
</tr>
<tr>
<td><em>S. ruminantium subsp. lactilytica</em> TAM6421</td>
<td>65</td>
<td>59%</td>
</tr>
<tr>
<td><em>V. parvula</em> DSM 2008</td>
<td>60</td>
<td>54%</td>
</tr>
<tr>
<td><em>Z. raffinosivorans</em> DSM 20765</td>
<td>70</td>
<td>63%</td>
</tr>
</tbody>
</table>
Figure 4.7: Presence and absence of key sporulation proteins in Negativicutes.

The Negativicutes predicted to form endospores are highlighted in orange. A green box is shown if an orthologue was detected and a white box indicates an orthologue was not detected for each protein.
4.2.7. The flagellum of *S. sphaeroides*

The bacterial flagellum, the chief organelle of motility in bacteria, occurs in Gram-negative and Gram-positive bacteria [113]. The external helical flagellar filament is constructed from the protein flagellin and ends with a capping protein. Just outside the cell wall, the filament attaches to the hook. A rod is connected to the hook and forms part of the basal body that runs through the bacterial membrane and includes protein rings, which act as bearings. The S, M and C rings project into the cytoplasm from the inner membrane and attach to the peptidoglycan. These rings are present in both Gram-negative and Gram-positive bacteria and generate the force that rotates flagella. In addition to these rings, Gram-negative bacteria also have L and P rings spanning the lipopolysaccharide and peptidoglycan layers respectively [114]. The flagellum is driven by the Mot complex, which is a rotary engine powered by the proton motive force [115].

The genes encoding the flagellum of *S. sphaeroides* were identified using reciprocal best BLAST hits and BlastKOALA [209] which annotates using the KEGG database [210] (Figure 4.9). The genes that encode the proteins for the outer-membrane L-ring (FlgH; SSPH_02664) and the peptidoglycan associated P-ring (Flgl; SSPH_02665) were located clustered with other flagellar genes including those for the proximal (FlgF; SSPH_02667) and distal (FlgG; SSPH_02668) rods. The proteins FlgH, Flgl and FlgF had entirely Gram-negative multiple best BLAST hits, apart from a single mislabelled hit from a diderm Firmicute. Genes encoding components of the flagellar motor MotA (SSPH_04009) and MotB (SSPH_04010) were identified [211, 212]. An orthologue of the flagellar brake protein YcgR could not be identified. A second cluster of genes was located between SSPH_00263 (FliA) and SSPH_00297 (FlgB). Within this locus were genes encoding the basal body, motor complex, MS-ring, C-ring and hook. Two orthologues of the flagellin protein FliC (SSPH_00878/ SSPH_02559) were more closely related to those from Gram-positive rather than Gram-negative bacteria (Figure 4.8). This was
confirmed by a Conserved Signature Indel found in Gram-negative flagellins, but absent from Gram-positive flagellins (Figure 4.10). BlastKOALA also identified SSPH_04620 as a flagellin (Figure 4.8). The filament cap gene \textit{fliD} was identified (SSPH_00876) alongside \textit{flgL} (SSPH_00875) responsible for the hook-filament junction. However, the \textit{flgK} gene (SSPH_02567) was located on a different contig.

No orthologues of the flagellar master regulator FlhDC from \textit{E. coli} and \textit{Salmonella} could be found [213, 214]. However orthologues were found of the master regulators from \textit{Campylobacter jejuni} FlgR (SSPH_02908; 8e-97) and FlgS (SSPH_02991; 8e-19) [215], along with a homologue of FlhF (SSPH_00272; 5e-40), a putative GTPase necessary in flagellar organelle development in bacteria where flagella are polar [216].

\begin{figure}
\centering
\includegraphics[width=\textwidth]{flagellin_tree}
\caption{Maximum Likelihood tree of Negativicutes flagellin proteins.}
\end{figure}

The percentage bootstrap support is indicated on the nodes, which was based on the analysis of 100 replicates. The flagellin of \textit{S. sphaeroides} groups with Gram-positive bacteria (red) rather than Gram-negative bacteria (blue).
Figure 4.9: The structure of the flagellum in *S. sphaeroides*.

Proteins were identified using either a combination of RBBH, KEGG and BlastKOALA. The locus tags of the orthologues from *S. sphaeroides* are shown along with the Gram Value in brackets. Orthologues assigned by reciprocal best BLAST hits are indicated in blue.
Crop from the amino acid sequence alignment of the flagellin protein shows a 32-33 amino acid insertion in the Gram-negative flagellin compared to the Gram-positive flagellin. \textit{S. sphaeroides} does not have this insertion.

4.3. Discussion

\textit{S. sphaeroides} is one of a small group of endospore forming Negativicutes including \textit{A. longum} [217], \textit{P. fermentans}, \textit{S. ovata} [203] and \textit{Thermosinus carboxydivorans} [218]. The features described here provide a genomic context for the phenotypes described in the original paper [41]. Although genes were detected with phylogenetic signatures from both Gram-negative and Gram-positive bacteria, neither GC-content or codon usage could delineate these two classes of gene. The major Gram-positive feature of \textit{S. sphaeroides} is its ability to sporulate. Analysis showed it encodes 95% of the key proteins required for sporulation, removing any doubt that it was a sporulating Negativicute. Genes for the outer membrane were prominent in a major Gram-negative locus, which is also present in other Negativicutes. The flagella showed a divided phylogenetic background, with a Gram-negative signature for L-ring and P-ring contrasted with the flagellin proteins that had a Gram-positive phylogeny.

4.3.1. Contamination and sporulation

Initially there was doubt about the ability of \textit{Sporomusa} to make spores due to the contamination of another endospore forming bacterium. There were three likely points of contamination. The first was after arrival, the second was the
original isolation of the bacterium at DSMZ or during subsequent culturing at the centre for purchase. It would seem unlikely that it was after arrival as the contaminating bacterium was not one cultured at any point in the laboratory before or during this study. The original description of *S. sphaeroides* [41] had the nucleotide sequence of the 16S rRNA gene, although a fortunate single colony pick could have missed any contamination. It was therefore most likely to have occurred during routine culture of the bacterium at DSMZ, which holds stocks of both *S. sphaeroides* and the contaminating *P. borealis*.

### 4.3.2. The 16S rRNA gene and phylogeny

The genome was assembled *de novo*. The 16S rRNA gene contains multiple variable and highly conserved regions, which disrupt the *de novo* assembly of the gene from short reads. Mapping the reads to the 16S rRNA gene sequences found in the GenBank nucleotide archive showed 7 sequences that were exact matches. The phylogenetic trees (Figure S1 and Figure 4.3) were calculated from the 16S rRNA gene and core proteome respectively, both showed the same structure. The statistical support of all the branches in the phylogenetic tree of the core proteome were substantially higher, but as the proteins used are unique to Negativicutes so it was not possible to include an outgroup in this tree. The grouping and structure corresponded to published phylogenetic trees [27, 93, 106, 170]. *S. sphaeroides* was located alongside other sporulating Negativicutes within the Selenomonadales *incertae sedis*. These are the Negativicutes that do not fit into the two families Veillonellaceae and Acidaminococcaceae, but are within the order Selenomonadales [93].

### 4.3.3. Evidence of an ancient origin

The Gram Value was a first step to gain a global view of the genome with respect to genes encoding Gram-negative and Gram-positive features. If two different bacteria gave rise to Negativicutes, they may have had different GC-content and codon usage in their genomes. This difference would then be
reflected in the GC-content or codon usage of a gene. However, analysis showed the GC-content or codon usage could not determine whether a gene encoded a Gram-negative or Gram-positive feature. This suggests that an ancient event that gave rise to Negativicutes and *Sporomusa* and over time these differences in GC-content and codon usage have faded away.

4.3.4. **Key features**

Two lines of evidence from the genome sequence confirmed that *S. sphaeroides* is an endospore-forming Negativicute.

- the 16S rRNA gene and core proteome grouped phylogenetically with the other Negativicutes that can sporulate.
- *S. sphaeroides* also had 95% of the key sporulation proteins found in typical Gram-positive sporulators such as *B. subtilis*.

Like other Negativicutes, *S. sphaeroides* has a major Gram-negative locus and its structure resembles the locus found in the other endospore forming Negativicutes. Its presence is evidence that *S. sphaeroides* shares a common Gram-negative ancestor with all Negativicutes, but is closely related to sporulating Negativicutes. This locus is more complex in the sporulating Negativicutes compared to non-sporulating. However, there was no direct link found between the Gram-negative locus and sporulation. Within this locus was the gene for BamA, the only component of BAM complex in *S. sphaeroides*.

*S. sphaeroides* showed the same atypical biosynthetic pathway for lipid A biosynthesis as the other Negativicutes. The pathway is shared with the bacterium *C. crescentus* [110]. Another key feature of Negativicutes is that they do not use the Braun’s lipoprotein and the gene *lpp* is absent. However, it includes the protein Mep045 with a S-layer domain that links to the peptidoglycan with cadaverine, originally described in *Selenomonas*.
*ruminantium* [94, 219, 220]. This suggests the outer membrane of the spore-forming *S. sphaeroides* is similar to that in its non-spore-forming relatives.

*S. sphaeroides* was originally described as having up to 15 flagella on the concave side [41]. Not surprisingly, the genome encodes all of the genes required for a Gram-negative flagellum, the distribution of flagella correlates with the presence of FlhF, found in bacteria with polar flagella [216]. Analysis of the phylogenetic background of the flagella provided a mixed verdict. The L-ring and P-ring had a Gram-negative origin, in contrast with the flagellin proteins that had a Gram-positive origin. The presence of Gram-negative proteins has been reported in Negativicutes [106], but the presence of Gram-negative structures built with Gram-positive proteins has not. The best hypothesis to account for this divided phylogeny is if the genes originated from both Gram-positive and Gram-negative bacteria. The presence of flagella with Gram-positive and Gram-negative parts suggests that the bacterium that gave rise to Negativicutes had flagella. Gram-negative and Gram-positive genes would have combined in this bacterium and the flagella are evidence of this event. Interestingly, the regulation of the flagella in Negativicutes is different to gamma-Proteobacteria such as *E. coli* and *Salmonella* as the genome does not encode the regulators FlhDC. Instead homologues of the FlgSR two-component regulatory system of *C. jejuni*, an epsilon-Proteobacteria, were found.

### 4.3.5. Future work

The next step for the genome assembly would be to close the gaps; this would help with discovering other potential regions of synteny that may exist within the sporulating Negativicutes and *S. sphaeroides*. Long-read sequencing technologies such as the Oxford Nanopore MinION and Pacific Biosciences sequencing would help with this task. However, for the envisaged transcriptomics study on this bacterium a completed genome was not a requirement. Another important goal of future research would be to
determine whether *S. sphaeroides* follows the *C. jejuni* paradigm of sigma-54 and sigma-28 dependant expression to regulate its flagellar genes [221]. In addition, eliminating taxonomic groups using FlhDC and retaining those with FlgSR might help narrow down the bacteria that share a Gram-negative ancestor with Negativicutes.
Chapter 5
The Transcriptional Landscape of the Negativicutes

5.1. Introduction

The ability to form endospores is a trait normally associated with Gram-positive bacilli from the phylum Firmicutes. However, surprisingly, this trait is also seen in some Gram-negative species from the Negativicutes (Figure 4.3). This existence of Gram-negative spores challenges many existing assumptions in microbiology.

In the chapter that follows, I aim to compare the transcriptional landscapes of two species from the Negativicutes that can both be grown under laboratory conditions; one that does not form spores and one that does:

- *Veillonella parvula* is an non-flagellated Gram-negative coccus that does not sporulate [222] and is often isolated from the human intestinal tract and oral cavity.
• *Sporomusa sphaeroides* is a flagellated curved Gram-negative rod that can form endospores [34]. Species of *Sporomusa* have been isolated from the environment [41, 83, 202] and the insect gut [82]. *S. sphaeroides* was originally isolated from a German riverbed in 1984 [41] and its genome sequence is described in Chapter 4. *S. sphaeroides* forms terminal thermoresistant endospores in late-logarithmic growth in medium containing betaine [41].

### 5.1.1. Transcriptomics

The genomes of *V. parvula* and *S. sphaeroides* have been analysed computationally to predict protein-coding genes. However, prediction of coding sequences offers only a partial glimpse into the cell biology of these organisms. For a more rounded view, we must turn to transcriptomics.

In 1958, Francis Crick first articulated the central dogma of molecular biology: that genetic information flows from DNA to RNA to Protein [223]. Transcriptomics interrogates the flow of genetic information through the RNA that codes for proteins (messenger RNA or mRNA) However, in recent years, it has become clear that many RNA sequences that do not code for proteins still have important functions in gene regulation and in other cellular functions [224, 225]. These RNAs that do not code for proteins are known as non-coding RNAs (ncRNA). In recent years, high-throughput sequencing has shown a high prevalence of non-coding RNAs in all species where they have been looked for.

Even if all cells within a bacterial population share an identical set of genes, the pattern of expression of these genes depends upon the temporal and environmental state of the bacterial cells, with dynamic changes in the transcriptional landscape as organisms adapt, survive and succeed in diverse environments. Transcriptomics experiments that investigate
bacterial gene expression can be used to show how bacteria function under different conditions, including in response to perturbations in the environment or after genetic manipulation. At its simplest, transcriptomics can be used to understand how bacteria are regulating metabolism and other cellular functions under logarithmic and stationary phases of growth.

Technologies for analysing transcriptomes have progressed over recent years:

- *expression sequencing tags* (EST), where a partial cDNA is sequenced with the Sanger method.
- *microarrays* using specific DNA sequences (probes) attached to a solid surface. RNA is reverse-transcribed to create complementary DNA (cDNA), which then hybridises with complementary sequences on the array. Any sequences binding with non-specific low affinity are washed off. The fluorescence of the target sequence bound to the probe can then be analysed.
- *tiling arrays* are built from probes representing contiguous or partially overlapping sequences from a region or genome. These arrays allow for analysis beyond simple gene expression profiling—in particular, as probes are not restricted to annotated features, tiling arrays can characterise regions genomes where function is still unknown, delivering comprehensive bacterial transcriptomes [225].
- *RNA-seq*, which relies on bulk sequencing of cDNA molecules that have been reverse-transcribed from RNA harvested from bacterial cells. Drawing on the ease-of-use and cost-effectiveness of high-throughput sequencing, RNA-Seq delivers a snapshot of the entire transcriptome at any point in time.

RNA-Seq has recently become the method of choice for transcriptome studies [226, 227]. Advantages over previous methods include:
• no knowledge of the reference sequence or coding sequences required.
• resolution to the single nucleotide.
• low background noise.
• identification of short or poorly expressed transcripts.
• identification of the transcriptional start sites.
• identification of RNA processing, which assists in the correct mapping of RNA to repetitive sequences [228].

However, there are some drawbacks when compared to whole-genome sequences. In contrast to a genome sequence, a transcriptome is constantly changing depending on time and conditions. The high dynamic range of gene expression means sequencing has to be performed at considerable depth. In addition, biological replicates are required to evaluate the reproducibility of the conditions being assessed. All these factors lead to a substantially higher cost for RNA-Seq when compared to genome sequencing.

5.1.2. Rationale for this work

Here, I set out to map the transcriptional landscape of the Negativicutes in two species that could be readily cultured in the laboratory. Initially, I hoped to shed light on the transcriptional landscape during sporulation, which prompted me to establish a sporulation assay. However, as these efforts proved unsuccessful, I instead compared transcriptional landscapes in logarithmic and stationary phases in the two species.
5.2. Results

5.2.1. Development of a sporulation assay

The first step in developing a sporulation assay for *S. sphaeroides* was to replicate the conditions described in the first paper that documented spore formation in this species [41]. *S. sphaeroides* was grown in the defined medium described in that paper, which I have termed ‘SPM84’. In addition, cells were grown in 136, a medium recommended by the DSMZ who supplied our strain of *S. sphaeroides*. Spores were seen when *S. sphaeroides* was grown in SPM84, but not when grown in 136. In contrast to previously published observations [41], inclusion of betaine in the 136 medium inhibited sporulation (Figure 5.1).

The growth rate of *S. sphaeroides* in SPM84 was also slow, taking up to 72 hours to reach stationary phase. Therefore, the growth of *S. sphaeroides* was evaluated in other media (Figure 5.2):

- Brain heart infusion (BHI).
- BHI with yeast extract and casitone (BHI-YC).
- Müller-Hinton.

Growth was fastest and most abundant in BHI-YC. However, growth of *S. sphaeroides* was inconsistent between different brands of BHI (Figure 5.3). Only BHI purchased from Oxoid resulted in consistent and rapid growth. Although the recommended growth temperature for *S. sphaeroides* is 35°C, 37°C was used initially, as a 35°C incubator was not available. As expected, growth at 35°C was superior to growth at 37°C (Figure 5.3).

Bacterial colonies formed only on solid medium made from SPM84. The presence of spores in each type of medium was confirmed by microscopy. No spores were seen in cultures grown in BHI or BHI-YC. This resulted in
adoption for further work of a non-sporulation-inducing medium (BHI-YC) and a sporulation-inducing medium (SPM84).

In an attempt to set up a sporulation assay, *S. sphaeroides* was grown in BHI-YC to stationary phase (11 hours), harvested and then transferred to SPM84. Transfer at stationary phase was chosen to ensure that the state of the bacterial cells would remain consistent, facilitating the calculation of the percentage that sporulated. After 24 hours in SPM84, the presence of spores was confirmed by microscopy. An aliquot of the culture was taken, heat inactivated at 80°C for 10 minutes [41] to kill vegetative cells. Spores were then germinated on SPM84 agar and counted. *S. sphaeroides* would not germinate on BHI, BHI-YC or Müller-Hinton.

After 24 hours in SPM84 only 0.49% of cells had sporulated (3.81 x 10^5 spores per ml). To improve the sporulation rate, 13 different variants of the SPM84 medium were made, each variant missing a single component (Figure 5.4). The removal of sodium selenite increased sporulation to 1.92x10^5 spores per ml. However, this was only 1.54% of the cells present at stationary phase. Removal of multiple components were tested, especially those that improved the sporulation rate above SPM84 (Figure 5.4). However, the results were too inconsistent to discover a trend and so I reluctantly abandoned attempts to characterise the transcriptional landscape of sporulation in the Negativicutes and instead turned my hand to a simpler more tractable goal of assaying transcriptomes at different growth phases.
Figure 5.1: Inhibition of sporulation with betaine.

The addition of betaine to the growth medium of *S. sphaeroides* at both 0.1% and 1.0% inhibited sporulation.

Figure 5.2: Growth curve of *S. sphaeroides* in different types of media at 37°C.

The growth of *S. sphaeroides* was evaluated in other media, it was fastest and most abundant in BHI-YC (Logarithmic rate of 0.160 OD600 per hour and a maximum OD600 of 0.450), compared to Muller-Hinton (Logarithmic rate of 0.019 OD600 per hour and a maximum OD600 of 0.299), BHI (Logarithmic rate of 0.076 OD600 per hour and a maximum OD600 of 0.381) and Sporomusa ’84 (Logarithmic rate of 0.041 OD600 per hour and a maximum OD600 of 0.317). A logarithmic scale is used for the Y-axis.
BHI purchased from Oxoid resulted in consistent and rapid growth. As expected, growth at the recommended 35°C (Logarithmic rate of 0.169 OD600 per hour and a maximum OD600 of 0.561) was superior to growth at 37°C (Logarithmic rate of 0.164 OD600 per hour and a maximum OD600 of 0.433). BHI from Fluka at 37°C showed the lowest growth rate (Logarithmic rate of 0.104 OD600 per hour and a maximum OD600 of 0.331). A logarithmic scale is used for the Y-axis.

Interestingly, during these early experiments, large and small colonies formed on the solid medium (Figure 5.5A). The two colony types were observed by microscopy (Figure 5.5B and C). The larger colonies (Figure 5.5B) showed cells with morphology typical of *S. sphaeroides*, while the smaller colonies did not, and appeared coccoid (Figure 5.5C). The presence of smaller colonies on solid medium was inconsistent across technical replicates that used the same inoculant and biological replicates. Contamination was also ruled out as a cause because solid media that was uninoculated and negative controls with uninoculated medium did not exhibit small colonies. The morphology of *S. sphaeroides* with phase-bright terminal endospores (Figure 5.5D and Figure 5.5E) was as originally
described with curved shaped and tapered ends. Very rarely, long spiral forms of *S. sphaeroides* were observed (Figure 5.5F) [41].

![Figure 5.4: Sporulation counts for SPM84 medium variations.](image)

The removal of sodium selenite, dipotassium phosphate and calcium chloride dihydrate each increased rates of sporulation over SPM84.
Figure 5.5: Morphology of *S. sphaeroides*

(A) Large and small colonies formed after heat inactivation and germination, microscopy showed the morphology of cells from the large (B) and small (C) colonies. (D and E) The morphology of *S. sphaeroides* with phase-bright terminal endospores. (F) A rarely seen spiral form of *S. sphaeroides*.
5.2.2. RNA-Seq of logarithmic and stationary phases of growth

The conditions of logarithmic growth and stationary phase are sufficiently different to provide a dynamic view of the transcriptome of both \textit{V. parvula} and \textit{S. sphaeroides}. Logarithmic phase is characterised by rapid cell growth, when metabolic activity is highest. By contrast, in stationary phase restrictions on nutrients and the build up of waste products create an unfavourable growth environment leading to reduced metabolic activity.

Growth curves of \textit{V. parvula} and \textit{S. sphaeroides} were assessed and the points of logarithmic and stationary phase identified (Figure S2). As reads from RNA-Seq should ideally be mapped onto the genome sequence of the strain under analysis, I genome-sequenced the two strains under investigation. The genome of \textit{S. sphaeroides} is described in Chapter 4. The genome of \textit{V. parvula} DSM2008 was re-sequenced on a MiSeq (Illumina) and compared to the completed reference NCBI strain. The sequenced reads were mapped against the reference genome of \textit{V. parvula} DSM 2008 (GenBank accession no. NC_013520.1). The genome sequence of the strain used here was confirmed to be identical to the GenBank reference genome using BreSEQ [229].

\textit{V. parvula} was grown in DSMZ medium 136 at 37°C and the points of logarithmic at 3 hours and stationary phase at 11 hours were chosen for analysis (Figure S2A). Different medium types had been assessed for the growth of \textit{S. sphaeroides} (Figure 5.2 and Figure 5.3) and BHI-YC was chosen to define the transcriptional landscape, under conditions where no sporulation occurred. Growth in BHI-YC at 35°C was rapid compared to the other medium types and consistent between biological replicates. For logarithmic and stationary phase for \textit{S. sphaeroides}, time points at 6 hours and 11 hours were chosen. The number of bacterial cells at the logarithmic
and stationary time points were confirmed with viable cell counts (Table S4).

RNA was extracted from bacteria harvested at the selected time points (Table S5) and assayed for contamination with DNA (Figure S3). As most bacterial RNA consists of ribosomal RNA (mainly 16S and 23S rRNA), Ribo-Zero (Illumina) was used for the depletion of rRNA (Figure S4A and B), leading to substantial increase in the number of reads available for sequencing from mRNA. I followed published advice that for an rRNA-depleted sample to detect low-abundance transcripts one needs 1-2x the genome size in number of reads [230]. Maintaining strand specificity for the transcripts allows correct annotation of novel genes or transcripts antisense to coding regions that may have important regulatory roles. The technique of choice has become the deoxy-UTP (dUTP) method [231], which is incorporated during synthesis of the second strand. The cDNA library was created using the Illumina TruSeq Stranded total RNA kit, which uses this methodology. The prepared cDNA libraries (Figure S4C) were sequenced on the Illumina MiSeq at 12 pM.

5.2.3. Processing RNA-seq reads and quality control

The sequencing reads from the transcriptomes of V. parvula and S. sphaeroides were processed (Methods:2.6.12). This was repeated for each of the biological replicates and the transcriptional data was visualised in Artemis [164]. The reads that mapped to the annotated coding sequences were counted using coverageBed [165] and analysed using the tool DESeq2 [167]. The reproducibility and clustering of the biological replicates was assessed from this by calculating sample-to-sample distances to construct a dendrogram (Figure 5.6A). The replicates and conditions clustered well and this was visualised using a principle component analysis (PCA) plot (Figure 5.6B).
(A) Heatmaps: *V. parvula* (left) and *S. sphaeroides* (right)

(B) PCA plots: *V. parvula* (left) and *S. sphaeroides* (right)

**Figure 5.6: PCA plots and heat maps of *S. sphaeroides* and *V. parvula*.**

Heat maps (A) of Euclidian distances between samples and PCA plots (B) showing sample-to-sample distances for replicates in logarithmic and stationary phases of growth. The heatmap and PCA-plot were calculated using R and DESeq2.
5.2.4. Enhancing the genome annotation by identifying transcriptionally active regions with toRNAdo

Firstly, a more thorough analysis of the genome beyond computational predictions of protein-coding sequences (CDSs) from ORFs was undertaken. The annotation was expanded to identify transcriptionally active regions outside predicted CDSs using toRNAdo [168]. This is an important first step because subsequent attempts at normalisation (see below) depend on a proxy calculation of genome size, which draws on number and length of coding sequences.

The genome and transcriptionally data was then inputted into the script toRNAdo, that looks for transcriptionally active loci, defined by an increase in transcription levels, a peak in transcription levels and then a decrease in transcription levels. The location of this transcriptional activity varies in relation to CDSs—the activity could be antisense to a coding feature, intergenic or at the border of an untranslated region (UTR) and a coding sequence. Expression in unannotated areas of the genome was identified under each condition and in each replicate and novel features were called when found in at least three replicates and after a visual check in Artemis. 5’ and 3’ UTRs were identified based upon the rise and fall of expression before the start or after the end of single or multiple CDSs, which showed uninterrupted expression (Figure 5.8). Areas of increased expression at the borders of genes, antisense to a gene(s) and in intergenic regions were identified as putative non-coding RNAs (ncRNAs).

Additionally, to investigate the fidelity of toRNAdo, the 2.1 mb genome of V. parvula was scanned visually to identify and annotate regions of expression outside of predicted coding sequences. Regions that were found visually were then compared to those identified by toRNAdo and Rockhopper [232].
The script toRNAdo identified more transcriptional features, including putative ncRNAs, than Rockhopper (Figure 5.7). Additionally, toRNAdo also found all of those putative ncRNAs identified visually (data not shown). Therefore, the script toRNAdo was considered the better option to annotate transcriptional features.

Figure 5.7: A comparison of Rockhopper and toRNAdo predictions.

In region (A) toRNAdo and Rockhopper performed similarly, but in region (B) toRNAdo predictions were superior, as Rockhopper identified nothing. No regions were observed where the toRNAdo prediction was worse than Rockhopper. The Rockhopper track shows non-stranded data of putative ncRNAs (grey), operons (orange) and UTRs (red). The toRNAdo track shows stranded data of putative ncRNAs (grey), operons (orange) and 5'UTRs (blue) and 3'UTRs (green). The bottom two tracks show the raw reads (log$_2$ transformed) of stationary phase (blue) and logarithmic phase (red) as well as annotated features. The colour of the feature indicates the Gram Value from Gram-positive (red) to Gram-negative (blue) or no Gram Value (black).
To calculate operons, two types of gene-pairs were used. The first pair was calculated based upon the toRNAdo output, two genes either side of an intergenic gap of uninterrupted expression (CDS_3 and CDS_4 in Figure 5.8). The second gene-pair was calculated based upon overlapping genes on the same strand, indicating they are both a part of the same operon. (CDS_2 and CDS_3 in Figure 5.8). The collected gene-pairs were assembled into an operon based upon matching genes between pairs (CDS_3 is the matching gene in Figure 5.8). The transcriptional features identified (Table 5.1) are discussed in the next section, where all transcriptionally active regions identified outside of coding sequences will now be described as ‘putative ncRNAs’.

Figure 5.8: Transcriptional regions identified by toRNAdo

The green line shows an example of expression data from three different replicates. Highlighted in blue are the different transcriptional regions that are predicted with toRNAdo. Border, intergenic and antisense ncRNAs are identified based upon transcriptional peaks outside or on the border of known genes. The 5’ UTR and 3’ UTR is based on the rise and fall of expression outside a continuous region of expression across several genes. Operons are predicted based upon multiple genes on the same strand that are either overlapping or with uninterrupted expression between them.
Table 5.1: Summary of transcriptional features

Summary of the different transcriptional features found in the genomes of *V. parvula* and *S. sphaeroides* using toRNAdo.

<table>
<thead>
<tr>
<th></th>
<th><em>V. parvula</em></th>
<th><em>S. sphaeroides</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size</td>
<td>2,132,142 bp</td>
<td>4,943,689 bp</td>
</tr>
<tr>
<td>No. of coding sequences</td>
<td>1923</td>
<td>4566</td>
</tr>
<tr>
<td>No. of contigs</td>
<td>1</td>
<td>118</td>
</tr>
<tr>
<td>Total no. of putative ncRNAs</td>
<td>63</td>
<td>176</td>
</tr>
<tr>
<td>No. of putative border ncRNAs</td>
<td>16 (1 also called as intergenic)</td>
<td>43 (9 also called as intergenic)</td>
</tr>
<tr>
<td>No. of putative intergenic ncRNAs</td>
<td>46 (1 also called as border)</td>
<td>132 (9 also called as border)</td>
</tr>
<tr>
<td>No. of putative antisense ncRNAs</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>No. of ncRNAs called as multiple types</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>No. of 3' UTRs</td>
<td>579</td>
<td>1138</td>
</tr>
<tr>
<td>No. of 5' UTRs</td>
<td>612</td>
<td>1233</td>
</tr>
<tr>
<td>No. of operons</td>
<td>378</td>
<td>820</td>
</tr>
</tbody>
</table>

5.2.5. Putative non-coding RNAs in Negativicutes

The presence of open reading frames (ORFs) within each putative ncRNA was investigated using the tool getORF [233], using the NCBI bacterial translation table II and a minimum length of 20 amino acids [234, 235]. In *V. parvula* and *S. sphaeroides* 228 and 634 ORFs were identified respectively. Then, for each ORF the presence of a ribosomal binding site (RBS) was investigated using the tool RBS-Finder [236]. 58 and 157 coding
sequences with a RBS were predicted in *V. parvula* and *S. sphaeroides* respectively. (Table S7 and Table S8). Some putative ncRNAs had multiple potential coding sequences. None of the amino acid sequences from these putative coding sequences had any protein domains when searched with CD search tool [237]. Therefore, most newly identified transcriptional units are likely to represent ncRNAs rather than unannotated protein coding sequences.

Each of the putative ncRNA nucleotide coding sequences was searched for within their genus and class with BLASTn (Table 5.2 and Table 5.3). Hits with similar sequences were more likely to be found within the same genus rather than in other genera of Negativicutes that have a more distant phylogenetic relationship. The top three most conserved of the putative ncRNAs were intergenic or antisense RNAs. Interestingly, intergenic RNAs were also amongst some of the highest regions of expression in both *V. parvula* (Figure 5.9) and *S. sphaeroides* (Figure 5.10). A region of a putative ncRNA (SSPH_NCR001869_43) in *S. sphaeroides* showed homology to tRNA-Lys genes in the *V. parvula* genome (Vpar_R0051 and Vpar_R0042), suggesting that this is an unannotated tRNA gene. The identification of a cluster of putative ncRNAs (Table 5.3) unique to the Selenomonadaceae *incertae sedis* Cluster 2 strengthens the argument for a monophyletic origin of this clade (Figure 4.3). Examples of the ncRNAs that were well conserved within the genus and within the class were visualised (Figure 5.9 and Figure 5.10). Putative ncRNAs that were antisense to a gene (Figure 5.9B: Vpar_NCRI360993) were unsurprisingly conserved. One putative ncRNA, antisense to *rnpB*, was annotated as a border ncRNA, as the expression crossed the border of the CDS into the intergenic region.
### Table 5.2: Conservation of *V. parvula* putative ncRNAs in the class Negativicutes and in the genus *Veillonella.*

The nucleotide sequences of the putative ncRNAs from *V. parvula* were searched using BLASTn against the class Negativicutes (spore formers highlighted in grey) and other *Veillonella* (columns highlighted in blue). The percentage identity of the top hit is indicated. Only putative ncRNAs with at least one hit are shown with an E-value cut-off of 1e-5.

<table>
<thead>
<tr>
<th>Type</th>
<th>Locus tag</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>antisense</td>
<td>Vpar_NCR122743</td>
<td>86</td>
</tr>
<tr>
<td>antisense</td>
<td>Vpar_NCR1243444</td>
<td>85</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR045191</td>
<td>78</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0151984</td>
<td>82</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR019827</td>
<td>83</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR039182</td>
<td>81</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0393524</td>
<td>81</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0425316</td>
<td>85</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0360993</td>
<td>88</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0454546</td>
<td>81</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0143449</td>
<td>81</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0158476</td>
<td>85</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0196538</td>
<td>81</td>
</tr>
<tr>
<td>border</td>
<td>Vpar_NCR0204738</td>
<td>80</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0103468</td>
<td>86</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0105317</td>
<td>81</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0193319</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0150481</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0224561</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0240394</td>
<td>82</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0284139</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0313103</td>
<td>81</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0372755</td>
<td>81</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0446572</td>
<td>80</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0578777</td>
<td>10</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0584588</td>
<td>88</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0722478</td>
<td>86</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0734512</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0727866</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0732279</td>
<td>87</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0846971</td>
<td>88</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR1268129</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR1323877</td>
<td>86</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR1343574</td>
<td>87</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0159945</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0167303</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0168481</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR01573816</td>
<td>86</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0176897</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0189536</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0393522</td>
<td>87</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR2093922</td>
<td>79</td>
</tr>
<tr>
<td>intergenic</td>
<td>Vpar_NCR0548403</td>
<td>81</td>
</tr>
</tbody>
</table>
Table 5.3: Conservation of *S. sphaeroides* putative ncRNAs in Negativicutes and in the genus *Sporomusa*.

The nucleotide sequences of the putative ncRNAs from *S. sphaeroides* were searched using BLASTn against the class Negativicutes (spore formers highlighted in grey) and other *Sporomusa* (columns highlighted in blue). The percentage identity of the top hit is indicated. Highlighted in red are a group of putative ncRNA with sequence homology within Selenomonadales incertae sedis Cluster 2 (Figure 4.3). Only putative ncRNAs with at least one hit are shown with an E-value cut-off of 1x10^-5.

<table>
<thead>
<tr>
<th>Type</th>
<th>Locus tag</th>
<th>Acinetobacter lwoffii DSM 1701</th>
<th><em>Pseudomonas</em> <em>furiicola</em> DSM 2087</th>
<th><em>Sporomusa</em> <em>ovata</em> DSM 2662</th>
<th>Thermotoga maritima DSM 25288</th>
</tr>
</thead>
<tbody>
<tr>
<td>antisense</td>
<td>SSPH_NCR0121874_10</td>
<td>86</td>
<td>95</td>
<td>95</td>
<td>90</td>
</tr>
<tr>
<td>border</td>
<td>SSPH_NCR0000427_2</td>
<td>78</td>
<td>99</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>border</td>
<td>SSPH_NCR0121326_2</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>border</td>
<td>SSPH_NCR0000007_22</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0016390_1</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0093558_6</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0121536_2</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0037869_5</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0001477_2</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0030403_1</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0085427_2</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0013458_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0174909_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0178525_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0184047_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0184381_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0196054_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0204503_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0207510_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0207919_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0215159_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0229157_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0239240_3</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0013984_4</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0023881_4</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0157456_4</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0186828_4</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0121669_5</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
</tbody>
</table>

Table 5.3 continued overleaf.
<table>
<thead>
<tr>
<th>Type</th>
<th>Locus tag</th>
<th>Z</th>
<th>X</th>
<th>Y</th>
<th>#</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0151551_6</td>
<td>79</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0182774_6</td>
<td>97</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0134132_7</td>
<td>81</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0150045_7</td>
<td>94</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0177894_7</td>
<td>98</td>
<td>99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0181006_7</td>
<td>94</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0129869_8</td>
<td>99</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0132313_8</td>
<td>95</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0150358_8</td>
<td>89</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0152612_8</td>
<td>80</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0133781_9</td>
<td>88</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0061465_10</td>
<td>94</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0030565_11</td>
<td>93</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0126729_11</td>
<td>82</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0022710_12</td>
<td>93</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0045874_13</td>
<td>97</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0079509_13</td>
<td>85</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0090287_14</td>
<td>95</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0179306_16</td>
<td>77</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0090870_90</td>
<td>86</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0052083_17</td>
<td>88</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0022874_19</td>
<td>79</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0028885_19</td>
<td>80</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0087610_19</td>
<td>82</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0035460_20</td>
<td>87</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0030340_23</td>
<td>88</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0047783_23</td>
<td>86</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0067012_24</td>
<td>88</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0074869_24</td>
<td>94</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0049974_32</td>
<td>86</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0021429_33</td>
<td>87</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR004602_35</td>
<td>89</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0033582_37</td>
<td>80</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0016839_40</td>
<td>87</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0000566_43</td>
<td>92</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR00081869_43</td>
<td>89</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0002888_43</td>
<td>92</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0010071_48</td>
<td>95</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0016465_48</td>
<td>98</td>
<td>99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0016777_48</td>
<td>83</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0016860_48</td>
<td>86</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0001176_69</td>
<td>97</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR00243970_3</td>
<td>87</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR003652_4</td>
<td>88</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0026425_25</td>
<td>98</td>
<td>99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intergenic</td>
<td>SSPH_NCR0006960_56</td>
<td>85</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5.9: Conserved putative ncRNAs in *V. parvula*

The top 7 tracks show the location and strand which toRNAdo identified the putative ncRNA in stationary phase (blue) and logarithmic (red). The final call (grey) is based upon the maximum range from each replicate. The bottom two tracks show the raw reads (log$_2$ transformed) of stationary phase (blue) and logarithmic phase (red) as well as annotated features. The colour of the feature indicates the Gram Value from Gram-positive (red) to Gram-negative (blue) or no Gram Value (black). The mapped reads in (C) look different, this is solely an artefact of GenomeDiagram and the region being lowly expressed.
Figure 5.10: Conserved ncRNAs in *S. sphaeroides*

The top 9 tracks show the location and strand which toRNAdo identified the putative ncRNA in stationary phase (blue) and logarithmic (red). The final call (grey) is based upon the maximum range from each replicate. The bottom two tracks show the raw reads (log$_2$ transformed) of stationary phase (blue) and logarithmic phase (red) as well as annotated features. The colour of the feature indicates the Gram Value from Gram-positive (red) to Gram-negative (blue) or no Gram Value (black). In some instances putative ncRNAs such as (C) reads were observed mapping antisense to the ncRNA.
5.2.6. Normalisation of RNA-Seq data

RNA-Seq analyses involve counting the number of reads that map to features of the genome. However, data should first be normalised to remove technical biases, such as the length of RNA species and sequencing depth of a sample, so as to enable accurate comparisons of expression levels of genes and genomes between and within samples [238].

The most commonly used calculation to normalise RNA-Seq data is Reads Per Kilobase per Million reads (RPKM) [239]. To calculate RPKM, the first step is to normalise for sequencing depth. Read count per gene is divided by a scaling factor, which is the total read count divided by one million. Next the reads-per-million value for each gene is divided by the gene length in kilobases, to normalise it to RPKM. However, use of RPKM leads to inconsistency in between-sample comparisons, as it characterises the genome by the total number of read counts, which is not biologically meaningful [166].

A minor modification of RPKM, known as TPM is often used to eliminate inherent biases of RPKM associated with very highly expressed genes. TPM first normalises the gene length by dividing the read count per gene by the length of the gene in kilobases. In this approach, the scaling factor is the sum of the normalised read counts per gene divided by one million. The normalised-reads-per-kilobase value of each gene is divided by the scaling factor to calculate TPM. Within a sample, RPKM and TPM are proportional; the biases that TPM accounts for are between samples. This is because while the sum of normalised reads may differ between samples, the sum of TPM between samples remains the same.

TPM [166] values were calculated for each coding sequence and for the novel putative ncRNAs in logarithmic and stationary phase. The differences
between expression in logarithmic and stationary phases were then plotted (Figure 5.11). The plot for *V. parvula* shows a negative bionominal distribution, typical of bacterial gene expression plots [240]. The observation of this distribution is important as DeSeq [167], used in this chapter, uses this model for normalisation. However, the plot for *S. sphaeroides* does not fit quite as well into a negative binomial distribution. This would have repercussions in how DeSeq models and normalises the data.

The high power of NGS leads to the issue of high sensitivity with RNA-seq, it can be unclear whether genes with low read counts are lowly expressed or the consequence of transcriptional noise [240]. Prior transcriptional data can help to resolve this by defining a cut-off between expression and transcriptional noise. However, as there is none for *V. parvula* or *S. sphaeroides* to estimate a TPM cut-off value for whether a gene is expressed, a conservative estimate based upon literature [241] was adopted that takes into account the fact that TPM scales with genome size. A TPM cut-off of 10 has been estimated for *Salmonella enterica* [241], the genomes of *V. parvula* and *S. sphaeroides* are 2.28x and 0.98x the size of *S. enterica* resulting in TPM cut-offs of 23 and 10 respectively. How the TPM cut-off correlates with the number of coding sequences showing expression was calculated (Figure 5.12). However, any TPM cut-off without experimental data should be considered arbitrary and tentative at best as no experimental methods such as RT-PCR were used to test them.

Using these estimated TPM cut-offs, it seems that a higher percentage of the *V. parvula* genome was expressed compared to *S. sphaeroides*. There are several potential explanations for this:

- the smaller genome of *V. parvula* would probably mean that more of the genome is essential.
• as *S. sphaeroides* was grown in non-sporulating conditions, genes for sporulation remained unexpressed.
• the medium that *S. sphaeroides* was grown in (BHI) was far richer than that of *V. parvula* (Table 5.4).

### Table 5.4: Expression of coding sequences

Summary of the transcriptome when the data was normalised using TPM, not RPKM. A TPM cut-off of 23 and 10 was used for *V. parvula* and *S. sphaeroides* respectively. Total genome features included both predicted or known ORFs and transcriptional active regions outside them.

<table>
<thead>
<tr>
<th></th>
<th><em>V. parvula</em></th>
<th><em>S. sphaeroides</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>TPM cut-off</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>Total genome features</td>
<td>1986</td>
<td>4867</td>
</tr>
<tr>
<td>Expressed at log</td>
<td>1791 (90%)</td>
<td>3190 (66%)</td>
</tr>
<tr>
<td>Expressed at stat</td>
<td>1496 (75%)</td>
<td>2311 (47%)</td>
</tr>
<tr>
<td>Expressed in both</td>
<td>1476 (74%)</td>
<td>2202 (45%)</td>
</tr>
<tr>
<td>Expressed in at least one</td>
<td>1807 (91%)</td>
<td>3282 (67%)</td>
</tr>
<tr>
<td>Expressed in neither</td>
<td>79 (9%)</td>
<td>1585 (33%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><em>V. parvula</em></th>
<th><em>S. sphaeroides</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Putative ncRNAs</td>
<td>63</td>
<td>176</td>
</tr>
<tr>
<td>Expressed at log</td>
<td>60</td>
<td>170</td>
</tr>
<tr>
<td>Expressed at stat</td>
<td>51</td>
<td>131</td>
</tr>
<tr>
<td>Expressed in both</td>
<td>63</td>
<td>128</td>
</tr>
<tr>
<td>Expressed in at least one</td>
<td>48</td>
<td>172</td>
</tr>
<tr>
<td>Expressed in neither</td>
<td>15</td>
<td>4</td>
</tr>
</tbody>
</table>
Figure 5.11: Plot showing the distribution of counts of coding sequence against TPM.

In logarithmic (red) more genes show higher expression while in stationary (blue) more genes show lower levels of expression. *V. parvula* is shown in (A) and *S. sphaeroides* is shown in (B).
Figure 5.12: Relationship between the TPM cut-off and the percentage of expressed sequences.

Increasing the TPM cut-off affects the number of expressed sequences differently in *V. parvula* vs *S. sphaeroides*.
5.2.7. Analysis of differential expression

When RNA-Seq is used with different conditions, it is possible to calculate the differential expression of genes between the conditions. The widely used package DeSeq, which is implemented in the statistical package R, was used to calculate differential expression of genes [242]. The hypothesis for normalisation in DESeq is that most genes are not differentially expressed. The read count for each gene is divided by the geometric mean across all samples; the medium of this ratio in a sample is then used as the scaling factor for normalisation. The test for differential expression is then based upon the negative binomial distribution.

DESeq uses P and Q-values to define if a change in expression is statistically important. The Q-value is a false discovery rate adjustment of the P-value. A P-value of 0.05 indicates 5% of all tests are false positive, while a Q-value (P-adjusted value) of 0.05 indicates 5% of the statistically significant tests (by P-value) are false positives. The dataset calculated by DESeq can be explored in R using methods such as heat maps and principle component analysis (PCA) plots to visualise and explore the variation between the biological replicates and conditions.

DESeq2 (Version 3.1) was used to normalise and analyse the differences between logarithmic and stationary phases of V. parvula growth [167, 242]. The total number of reads mapping to each gene was determined using Bedtools (Version 2.17.0) [243]. The read counts for the logarithmic and stationary phase were compared and the log$_2$-fold change between them calculated. P-values were calculated using the three biological replicates for each condition in V. parvula and four biological replicates in S. sphaeroides and adjusted for multiple testing using the false discovery rate. Significant changes in expression were based upon a P-adjusted value of ≤ 0.05 with a log$_2$ fold change of ≥ 1 [244, 245]. Statistical analysis using DeSeq showed
differentiation between conditions and good replication within conditions (Figure 5.6).

The conditions of logarithmic and stationary phase are substantially different and therefore large proportions of the genomes of both *V. parvula* and *S. sphaeroides* were differentially expressed (Table 5.5). However, the proportion of features that underwent differential expression in the two genomes was roughly the same: around half in both cases. Both genomes also showed a similar percentage of up-regulated and down-regulated genes. Furthermore, in both species, around two thirds of the putative ncRNAs were differentially expressed (Table 5.5). In *V. parvula*, ncRNAs showed the same amount of up-regulation and down-regulation in stationary phase compared to logarithmic. *S. sphaeroides* had slightly more ncRNAs down-regulated in stationary than were up-regulated.

5.2.8. Transcriptional analysis of Negativicutes

5.2.8.1. Differential expression of COGs

The genes of *V. parvula* and *S. sphaeroides* were assigned to clusters of orthologous groups (COGs) [246] using the NCBI conserved domain search [237] with an E-value cut-off of $1 \times 10^{-5}$. Only genes, reported by DESeq2, with a p-adjusted value $\leq 0.05$ and log$_2$ fold change of $\geq 1$ were considered as significantly differentially expressed and analysed (Figure 5.13 and Figure 5.14).
Table 5.5: Differential expression between logarithmic and stationary in *V. parvula* and *S. sphaeroides*.

Summary of the significantly differently expressed genes calculated using DeSeq. A significant change in expression was based upon a P-adjusted value of \( \leq 0.05 \) with a log2 fold change of \( \geq 1 \). Total coding sequences included both predicted or known ORFs and transcriptional active regions outside them.

<table>
<thead>
<tr>
<th>Coding Sequences</th>
<th>V. parvula</th>
<th>S. sphaeroides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (Excludes putative ncRNAs)</td>
<td>1987</td>
<td>4867</td>
</tr>
<tr>
<td>Differentially expressed</td>
<td>990 (49%)</td>
<td>2594 (53%)</td>
</tr>
<tr>
<td>Up-regulated in stationary (vs logarithmic)</td>
<td>524 (26%)</td>
<td>1265 (26%)</td>
</tr>
<tr>
<td>Down-regulated in stationary (vs logarithmic)</td>
<td>466 (23%)</td>
<td>1329 (27%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Putative ncRNAs</th>
<th>V. parvula</th>
<th>S. sphaeroides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>63</td>
<td>176</td>
</tr>
<tr>
<td>Differentially expressed</td>
<td>43 (68%)</td>
<td>117 (66%)</td>
</tr>
<tr>
<td>Up-regulated in stationary (vs logarithmic)</td>
<td>22 (35%)</td>
<td>57 (32%)</td>
</tr>
<tr>
<td>Down-regulated in stationary (vs logarithmic)</td>
<td>21 (33%)</td>
<td>60 (34%)</td>
</tr>
</tbody>
</table>
The logarithmic phase is characterised by the rapid growth of the bacterial population and therefore the biogenesis of bacterial cells is more pronounced in logarithmic phase than stationary. This is indicated in the down regulation of cell envelope biogenesis [M] and biogenesis [J].

Stationary phase is associated with restriction of nutrients and build up of waste products in the medium, creating an unfavourable growth environment. This change in conditions has affected the metabolism and large numbers of genes in many COGs are up- or down-regulated. Both *V. parvula* and *S. sphaeroides* showed an up-regulation of energy production and conversion [C], suggesting a major switch in their metabolism occurred at stationary phase.

A larger number of genes in *S. sphaeroides* were classified in the COG assigned to cell motility [N] than in *V. parvula*. This can be explained by *V. parvula* not having flagella while *S. sphaeroides* does. In both species, COGs assigned to cell motility were up-regulated in stationary phases of growth. The up regulation of signal transduction mechanisms [T] in *S. sphaeroides* is potentially related to the increase in cell motility, as it included many genes encoding chemotaxis proteins.

### 5.2.8.2. The conservation of both Gram-negative and Gram-positive operons

Using transcriptional information from the script toRNAdo, operons were predicted in both *V. parvula* and *S. sphaeroides* (Table 5.1). In addition, as we now know that the genomes of Negativicutes contain both “Gram-negative” and “Gram-positive” genes (Chapters 3 and 4), investigations were carried out to determine whether operons also retained a Gram-negative or Gram-positive identity.
Figure 5.13: The differential expression of COG at stationary vs logarithmic phase in *S. sphaeroides*.

Only genes, reported by DESeq2, with a p-adjusted value ≤ 0.05 and log₂ fold change of ≥1 were considered as significantly differentially expressed. The NCBI conserved domain search was used to place genes into a COG. As some COG annotations include several groups, the sum of COG groups is greater than the sum of significantly differentially expressed genes.
Figure 5.14: The differential expression of COG at stationary vs logarithmic phase in *V. parvula*.

Only genes, reported by DESeq2, with a p-adjusted value ≤ 0.05 and log₂ fold change of ≥1 were considered as significantly differentially expressed. The NCBI conserved domain search was used to place genes into a COG. As some COG annotations include several groups, the sum of COG groups is greater than the sum of significantly differentially expressed genes.
Most operons fell neatly into two groups (Figure 5.15):

- those with genes that had a Gram Value of 1.0 (i.e. “Gram-negative”)
- those with genes that had a Gram Value of 0.0 (i.e. “Gram-positive”)

Nonetheless, there were 7 and 13 operons in *V. parvula* and *S. sphaeroides* respectively, where apparently Gram-negative and Gram-positive genes were found in the same operon. However, further investigation showed that this was probably a methodological artefact, as in almost every case when a “Gram-positive” gene and a “Gram-negative” gene were in the same operon, the Gram values for one or other or both genes would be supported by very few informative BLAST hits. In a second round of analysis, these operons were removed by applying a cut-off, requiring that the Gram Value for each gene be supported by at least 10 BLAST hits. This did not alter the overall trend (Figure 5.6), but did remove all but one of the operons that had Gram-negative and Gram-positive genes.

A single operon from *S. sphaeroides* remained with both Gram-positive and Gram-negative genes with Gram Values of 0.0 and 1.0 respectively, each supported by 50 BLAST hits. The operon encoded the following genes.

- SSPH_02747 mcp8 – chemotaxis protein (Gram-positive)
- SSPH_02748 cca_1 - tRNA adenyllyltransferase (Gram-negative)
- SSPH_02749 mcp4_3 – chemotaxis protein (Gram-positive)

Homology searches showed that the two Gram-positive proteins were methyl-accepting chemotaxis proteins (MCPs) [247]. The Gram-negative protein SSPH_02748 showed homology to a tRNA adenyllyltransferase, which repairs the CCA triplet found at the 3’ end of the tRNA [248].
Operons were divided into two groups, those with genes that had a Gram Value of 1.0 (Gram-negative) or those with genes that had a Gram Value of 0.0 (Gram-positive). The range of Gram Values from all genes in those Gram-negative or Gram-positive operons was then summarised as a box plot.

Figure 5.15: Range of Gram Values from operons with either Gram-negative or Gram-positive genes.
5.2.8.3. The major Gram-negative locus

As noted, a major Gram-negative locus was identified in *V. parvula* (Chapter 3) and *S. sphaeroides* (Chapter 4). The genes in this region included *bamA* and those encoding the lipid-A biosynthesis pathway. The transcriptional landscape of this region was compared between both species of bacteria (Figure 5.16).

Absolute synteny in the major Gram-negative locus was maintained between Vpar_0522 and Vpar_0552 in *V. parvula* (Chapter 3) and SSPH_02657 and SSPH_02621 in *S. sphaeroides* (Chapter 4). Taking into account co-expression of genes in operons, the locus in *V. parvula* extended from Vpar_0519 through to Vpar_0554, while in *S. sphaeroides* the region could be expanded from SSPH_02657 through to SSPH_02613. The overall trend in the locus showed a decrease in expression during stationary phase than compared to logarithmic stage. However, this was more marked in *V. parvula* than *S. sphaeroides*.

Expression analysis also showed lower expression of *bamA* (Vpar_0527 and SSPH_02650) during stationary phase than during logarithmic stage in both species. However, this difference was significant only in *V. parvula*. This may be because in *V. parvula*, *bamA* is the first gene in an operon that ends with Vpar_0532, whereas in *S. sphaeroides*, it is in the middle of an operon that starts at SSPH_02657 and ends at SSPH_02643.

The overall operonic structure in this locus also shows differences between the two species. In *V. parvula* the region consists of six operons while in *S. sphaeroides* there are five operons (Table 5.6). Visual inspection of the data showed that the (Vpar_525/ Vpar_526) division between operons may be present at (SSPH_2654/ SSPH_2653), but it was not predicted (Figure 5.16B).
(A) The major Gram-negative locus of *V. parvula*.
(B) The major Gram-negative locus of *S. sphaeroides*. 
Figure 5.16: The transcriptional landscape of the major Gram-negative locus in *V. parvula* and *S. sphaeroides*.

Visualisation of the transcriptional landscape of the major Gram-negative locus of *V. parvula* (A) and *S. sphaeroides* (B) using toRNAdo and a custom Python script. The tracks from top to bottom are as follows:

**Track 1: Log₂ Fold Change**
The log₂ fold change is indicated on track 1 and as a bar chart on track 2. The blue and grey bars indicate genes with a significant (>1) or not significant (<1) log₂ fold change respectively. The abbreviation not significant 'NS' indicates the p adjusted-value was > 0.05.

**Track 2: Mapped reads and differential expression**
The mapped reads (log₂ transformed) are shown for logarithmic (red) and stationary phase (blue). The normalised expression data from DeSeq shows differential expression as a bar. A significant change in expression is shown in blue (log₂ fold change of ≥ 1), while an insignificant change in expression is shown in grey (log₂ fold change of < 1).

**Track 3: Features predicted by the toRNAdo script**
Transcriptional features were calculated using toRNAdo and operons were predicted. The visualised features are 5’ UTR (blue), 3’ UTR (green), operons (orange) and ncRNAs (green).

**Track 4: Annotated features**
Genes are coloured based upon their Gram Value showing Gram-positive (red) through to Gram-negative (blue). Those without any mNBBH hits or non-protein coding genes are black as a Gram Value could not be calculated. The locus tag/annotation in purple indicates genes in the major Gram-negative region.
### Table 5.6: Comparison of the operon structure of the major Gram-negative locus.

The operon structure (orange) of the major Gram-negative locus of *V. parvula* and *S. sphaeroides* was predicted using data from tRNA
d. The alignment above is based on the core proteome calculated in Chapter 4 (blue).

<table>
<thead>
<tr>
<th>Core proteome orthologue group</th>
<th>Gene</th>
<th><em>V. parvula</em> Locus tag</th>
<th>Operon</th>
<th><em>S. sphaeroides</em> Locus tag</th>
<th>Operon</th>
</tr>
</thead>
<tbody>
<tr>
<td>428</td>
<td><em>mlAE</em></td>
<td>Vpar. 522</td>
<td></td>
<td>SSPH_2657</td>
<td></td>
</tr>
<tr>
<td>429</td>
<td><em>mlAF</em></td>
<td>Vpar. 523</td>
<td></td>
<td>SSPH_2656</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>mlAD</em></td>
<td>Vpar. 524</td>
<td></td>
<td>SSPH_2655</td>
<td></td>
</tr>
<tr>
<td>430</td>
<td><em>bamA</em></td>
<td>Vpar. 526</td>
<td></td>
<td>SSPH_2653</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vpar. 527</td>
<td></td>
<td>SSPH_2652</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vpar. 528</td>
<td></td>
<td>SSPH_2651</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vpar. 529</td>
<td></td>
<td>SSPH_2650</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vpar. 530</td>
<td></td>
<td>SSPH_2649</td>
<td></td>
</tr>
<tr>
<td>227</td>
<td><em>ompH</em></td>
<td>Vpar. 528</td>
<td></td>
<td>SSPH_2648</td>
<td></td>
</tr>
<tr>
<td>228</td>
<td><em>ompH</em></td>
<td>Vpar. 531</td>
<td></td>
<td>SSPH_2645</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>lpxD</em></td>
<td>Vpar. 532</td>
<td></td>
<td>SSPH_2644</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSPH_2643</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSPH_2642</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>230</td>
<td><em>lpxC</em></td>
<td>Vpar. 533</td>
<td></td>
<td>SSPH_2640</td>
<td></td>
</tr>
<tr>
<td>231</td>
<td><em>lpxA</em></td>
<td>Vpar. 534</td>
<td></td>
<td>SSPH_2639</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>lpxB</em></td>
<td>Vpar. 535</td>
<td></td>
<td>SSPH_2638</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vpar. 536</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>232</td>
<td><em>lpxL</em></td>
<td>Vpar. 537</td>
<td></td>
<td>SSPH_2637</td>
<td></td>
</tr>
<tr>
<td>233</td>
<td><em>lpxR</em></td>
<td>Vpar. 538</td>
<td></td>
<td>SSPH_2636</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>msbA</em></td>
<td>Vpar. 539</td>
<td></td>
<td>SSPH_2635</td>
<td></td>
</tr>
<tr>
<td>234</td>
<td><em>kdtA/waa A</em></td>
<td>Vpar. 540</td>
<td></td>
<td>SSPH_2634</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>lpxK</em></td>
<td>Vpar. 541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>235</td>
<td><em>kdsB</em></td>
<td>Vpar. 542</td>
<td></td>
<td>SSPH_2633</td>
<td></td>
</tr>
<tr>
<td>236</td>
<td><em>kdsA</em></td>
<td>Vpar. 543</td>
<td></td>
<td>SSPH_2632</td>
<td></td>
</tr>
<tr>
<td>410</td>
<td><em>kdsD</em></td>
<td>Vpar. 544</td>
<td></td>
<td>SSPH_2631</td>
<td></td>
</tr>
<tr>
<td>409</td>
<td><em>kdsC</em></td>
<td>Vpar. 545</td>
<td></td>
<td>SSPH_2630</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>lpxL</em></td>
<td>Vpar. 546</td>
<td></td>
<td>SSPH_2629</td>
<td></td>
</tr>
<tr>
<td>408</td>
<td></td>
<td>Vpar. 547</td>
<td></td>
<td>SSPH_2628</td>
<td></td>
</tr>
<tr>
<td>407</td>
<td></td>
<td>Vpar. 548</td>
<td></td>
<td>SSPH_2627</td>
<td></td>
</tr>
<tr>
<td>406</td>
<td><em>lpaB</em></td>
<td>Vpar. 549</td>
<td></td>
<td>SSPH_2626</td>
<td></td>
</tr>
<tr>
<td>405</td>
<td><em>lpaF</em></td>
<td>Vpar. 550</td>
<td></td>
<td>SSPH_2625</td>
<td></td>
</tr>
<tr>
<td>404</td>
<td></td>
<td>Vpar. 551</td>
<td></td>
<td>SSPH_2624</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSPH_2623</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSPH_2622</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>403</td>
<td></td>
<td>Vpar. 552</td>
<td></td>
<td>SSPH_2621</td>
<td></td>
</tr>
</tbody>
</table>
5.2.8.4. Sporulation genes and pseudogenes

The expression of the signature sporulation genes identified in a previous publication [117] were investigated in *S. sphaeroides*. Using reciprocal best BLAST hits, 106 of these genes were identified in *S. sphaeroides*. Sixty-two of these genes were expressed in logarithmic phase, while 36 were expressed during stationary phase. Forty-four sporulation signature genes were expressed neither in logarithmic nor in stationary phase (Table S6).

Pseudogenes are once-functional genes, where frameshifts or nonsense or missense mutations have left them unable to produce functional proteins. Prokka predicted no pseudogenes in *S. sphaeroides*, while the NCBI annotation for *V. parvula* predicted 15 pseudogenes. In *V. parvula* eight pseudogenes were shown to have sufficient reads mapping that the normalised expression was above a TPM cut-off of 23 (Table 5.7).

Two apparent pseudogenes encoded transposases or integrases (Vpar_0137 and Vpar_0718). Another pseudogene was linked to the toxin PezT (Vpar_1296), which, if active, is deadly to the cell in the absence of the PezA antitoxin, for which no significant hit was found [249]. Therefore losing the ability to encode such proteins from these pseudogenes would be favourable.
Table 5.7: Expression of pseudogenes in *V. parvula*

The pseudogenes from *V. parvula*, those highlighted in green are above the TPM cut-off of 23 that putatively indicates they are expressed. Amino acid sequences were searched for using BLAST against the REFSEQ database to investigate potential function.

<table>
<thead>
<tr>
<th>Locus tag</th>
<th>TPM (log)</th>
<th>TPM (stat)</th>
<th>Putative annotation from BLAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vpar_0137</td>
<td>0</td>
<td>0</td>
<td>Transposase/Integrase</td>
</tr>
<tr>
<td>Vpar_0187</td>
<td>85</td>
<td>14</td>
<td>Unknown</td>
</tr>
<tr>
<td>Vpar_0248</td>
<td>15</td>
<td>4</td>
<td>Unknown</td>
</tr>
<tr>
<td>Vpar_0257</td>
<td>23</td>
<td>11</td>
<td>Unknown</td>
</tr>
<tr>
<td>Vpar_0262</td>
<td>35</td>
<td>4</td>
<td>Partial match to ABC transporter</td>
</tr>
<tr>
<td>Vpar_0480</td>
<td>14</td>
<td>3</td>
<td>No significant hits</td>
</tr>
<tr>
<td>Vpar_0718</td>
<td>1</td>
<td>0</td>
<td>Transposase</td>
</tr>
<tr>
<td>Vpar_0722</td>
<td>18</td>
<td>3</td>
<td>Acetyltransferase</td>
</tr>
<tr>
<td>Vpar_1165</td>
<td>235</td>
<td>164</td>
<td>No significant hits</td>
</tr>
<tr>
<td>Vpar_1296</td>
<td>33</td>
<td>6</td>
<td>Toxin PezT</td>
</tr>
<tr>
<td>Vpar_1325</td>
<td>31</td>
<td>9</td>
<td>Partial hit of Fe-S oxidoreductase</td>
</tr>
<tr>
<td>Vpar_1411</td>
<td>9</td>
<td>2</td>
<td>Only hypothetical hits</td>
</tr>
<tr>
<td>Vpar_1458</td>
<td>106</td>
<td>5</td>
<td>Prephenate dehydrogenase</td>
</tr>
<tr>
<td>Vpar_1780</td>
<td>2</td>
<td>1</td>
<td>Cation transporter</td>
</tr>
<tr>
<td>Vpar_1825</td>
<td>75</td>
<td>102</td>
<td>CRISPR-associated protein Cas6</td>
</tr>
</tbody>
</table>

5.3. Discussion

The original aim of the work on *S. sphaeroides* was to map the transcriptome during sporulation. Although this aim was not achieved, there was some success in understanding the conditions that encourage or inhibit sporulation in *S. sphaeroides* and systematic changes to culture media did improve the sporulation rate. Plus it is now clear that the addition of betaine, prevents *S. sphaeroides* from entering the sporulation process. With further work to define a minimal medium that promotes sporulation, it might soon be feasible to develop a sporulation assay for the Negativicutes. Additional exploration of the genome might provide clues as to what induces sporulation in *S. sphaeroides*. The presence of small-colonial variants, with a cellular morphology not typical of *S. sphaeroides*,
remains unexplained. Extensive repeats of the experiment showed the small coccoid cells were not due to contamination. However, the trigger for their formation remains unknown.

It has been shown that 86% of genes in *Salmonella* [241] are expressed in multiple conditions. The genomes of both *V. parvula* and *S. sphaeroides* have both been subjected to automated annotation e.g. with Prokka [149]. The results here show that under the conditions of logarithmic and stationary phase growth, the majority of these annotated genes are expressed in at least one condition: 90% and 66% in *V. parvula* and *S. sphaeroides* respectively. However, these findings depend on an arbitrary TPM cut-off [241], defined in the absence of any prior transcriptional data for the Negativicutes. This should be taken as tentative until better experimental calibration can be performed.

Mapping the transcriptional landscape in logarithmic and stationary phase for two species of the Negativicutes has also expanded the repertoire of known genome features in both species. Both species show an array of putative new ncRNAs. In addition, operon structures together with 5’ and 3’ UTRs have been defined. Interestingly, some putative ncRNAs were among the most highly expressed features of the genome. Although the ncRNAs were not experimentally verified, a BLASTn search found many of the sequences were present in other bacteria from the same genus, *Veillonella* or *Sporomusa*. There was far less conservation outside the genus and there were no equivalent sequences found between *V. parvula* and *S. sphaeroides*, which is typical of ncRNAs [226]. Many small coding sequences with RBS were found within intergenic putative ncRNAs. This suggests that some putative ncRNAs could in fact be protein-coding genes missed by automated annotation. However, proteomics analyses would be required to determine whether the associated protein sequences are in fact expressed.
V. parvula is predicted to contain a number of pseudogenes, it was determined that 8 of them had enough reads to suggest they were expressed. Signature sporulation genes were also identified in S. sphaeroides, 44 of them were not expressed in any tested condition suggesting they may play an exclusive role in sporulation.

The Gram-negative and Gram-positive character of the transcriptional landscape was also investigated. The sets of genes in operons were overwhelmingly either Gram-negative or Gram-positive respectively. This expands upon the finding of distinct Gram-negative and Gram-positive genes in Chapter 3. Transcriptional information helped map the boundaries of the conserved Gram-negative locus.

5.3.1. Future work

This work has shown that it is feasible to control sporulation in the Negativicutes. Several key steps towards the goal of a sporulation assay have been taken. The development of a sporulation medium equivalent to that used with the model organism B. subtilis will one day provide an intimate look at how a Gram-negative bacterium can sporulate. Proteomics could complete the picture for V. parvula and S. sphaeroides, helping to defining a biologically meaningful TPM cut-off for gene expression while also determining whether intergenic putative ncRNAs are in fact protein-coding sequences.
Chapter 6
Conclusions

At first glance, the Negativicutes appear to be misplaced in the taxonomic tree, as a class of Gram-negative bacteria within the Gram-positive phylum of Firmicutes. However, their apparent misplacement is a limitation of a taxonomy that says that a bacterium with Gram-negative and Gram-positive aspects can be defined only by one of those characteristics. Understanding how a class of bacteria can be both Gram-negative and Gram-positive provides a glimpse into a potentially extremely rare event in evolution.

Originally, four scenarios were hypothesised for the origin of Negativicutes:

- The origin of bacteria was Gram-negative, with Negativicutes as the first Gram-positive.
- The origin of bacteria was Gram-positive, with Negativicutes as the first Gram-negative.
- Negativicutes are a derived state, resulting from a genome fusion or at least horizontal gene transfer between a Gram-negative and a Gram-positive bacterium.
The first two hypotheses address a major unanswered question in evolutionary biology: what was the original form of the bacterial cell envelope [144, 250-255]? These hypotheses are relevant if one assumes that the Negativicutes are the missing link between monoderm and diderm bacteria [28, 201, 255, 256]. The greater complexity of the diderm cell envelope over the monoderm and its role in resistance to antibiotics suggest that it was the later invention [144, 250]. More recently, it has been suggested that diderm outer membranes are an evolutionary consequence of sporulation in a monoderm bacterium [256]. An alternative view is that the monoderm Gram-positive state has been derived through loss of the outer membrane on multiple occasions [201]. But the most parsimonious hypothesis is that the Negativicutes are an evolutionary peculiarity—the result of a sharing/combining of genomic material.

Campbell et al. adopted a single-best-BLAST-hit approach to claim that Gram-negative proteins from the Negativicutes showed an association with the delta-Proteobacteria [106]. However, this assumes that the best BLAST hit is the closest phylogenetic relative, which has been shown to be an unsound inference [171].

Early in this project, I developed a method to define whether a gene was Gram-negative or Gram-positive with a Gram Value. The Gram Value also visualised and showed how typical Gram-positives and Gram-negatives, such as C. difficile and E. coli are distinct from each other and from Negativicutes. Defining the Gram Value allowed for genome scale analysis of the class Negativicutes.

Investigations of the similarity of Negativicutes to other bacteria was also based on Gram-negative signature proteins—the first time that the relationship between Negativicutes and other bacteria has been investigated in this way. This concept was based upon the use of eukaryotic signature
proteins [193, 194]. The Gram-negative similarities of Negativicutes to other bacteria were defined based upon similar biosynthetic pathways from Gram-negative toolkits. It would be hypothesised that any two diderms would have the same set of Gram-negative signature proteins if they were related. This methodology narrowed down the bacteria that could have supplied the Gram-negative features of Negativicutes. It also showed Negativicutes was related to only a small subset of Gram-negatives, a finding previously recognised with an insert in the Hsp60 protein [144].

With a narrowed field of Gram-negative donors, an informed phylogenetic analysis could be undertaken. The phylogenetic relationships were calculated and showed that Fusobacteria had the closest relationship to the Gram-negative aspect of Negativicutes; they also share the same Hsp60 insert [144]. It also showed that Halanaerobiales and Negativicutes shared a monophyletic origin. This upholds a genome analysis of Halothermothrix orenii from the order Haloanaerobiales, which held it was monophyletic to the Negativicute T. carboxydivorans [136].

While obvious genes such as those for the outer membrane can be defined as Gram-negative, genes that were not obviously Gram-negative or Gram-positive were also found. This included part of the flagellar toolkit that was Gram-positive. Up until now proteins that are part of the Gram-negative outer membrane have been shown to be Gram-negative. The presence of proteins with a different phylogenetic origin to others within multi-protein complexes implies a genome fusion. Only such a genome scale sharing and cataclysmic merging of genetic material could cause such a dichotomy.

Identification of a major Gram-negative locus also supported the monophyletic origin of Negativicutes, as this locus showed extensive, albeit not absolute, synteny in most of the Negativicutes. The Gram-negative signature proteins present in this region link the Negativicutes and
Halanaerobiales to a common Gram-negative ancestor. This ancestor is also linked to Fusobacteria, which possess the same repertoire of Gram-negative signature proteins. Although the region of synteny was not found in Fusobacteria, it has since been reported in Halanaerobiales [201].

Analysis of the genome of *S. sphaeroides* showed that there was no codon usage or GC content difference between the Gram-negative and Gram-positive genes. This suggests any genome fusion event was ancient and over time these differences in GC-content and codon usage have faded away. This is also corroborated by the global spread and wide range of environmental niches that the Negativicutes occupy. It would have taken an extremely long time for the different species to adapt to these environments.

The transcriptomic landscape of two species of Negativicutes was assessed in different conditions for the first time: during logarithmic and stationary phase growth. The repertoire of putative non-coding RNAs expressed in these conditions was defined. Operon structures with 5’ and 3’ UTRs were also elucidated. Investigation of consensus sequences in UTRs could lead to further understandings of their similarities to Gram-negative or Gram-positive bacteria. Operons with Gram-negative or Gram-positive genes were found to contain other Gram-negative or Gram-positive genes respectively, with almost no mixed operons. The existence of conserved Gram-negative and Gram-positive operons suggests that they had different origins. Additionally, preservation of Gram-negative genes in intact operons has probably aided in maintaining their Gram-negative functions within a Gram-positive bacterium.

First steps towards a sporulation assay for a member of the Negativicutes have been taken. Different types of media were tested for both growth and sporulation. Media where *S. sphaeroides* grows faster and does not sporulate
were identified. I also discovered that the addition of betaine would absolutely inhibit sporulation. Defining sporulating and non-sporulating media allows transfer of *S. sphaeroides* that is not sporulating into a medium that induces sporulation. The medium that induced sporulation was improved three-fold, but this work needs to continue to develop a better-defined sporulation medium. This would allow the transcriptional landscape of Negativicutes to be assessed during sporulation.

So far, the relationship between the Negativicutes and other Gram-positives, such as the Clostridia has been assessed only by the phylogenetics of ribosomal proteins and the 16S rRNA gene. The phylogenetics of the Gram-positive toolkit that the Negativicutes share with other Gram-positives has yet to be investigated. Drawing on the example of Gram-negative signature proteins, Gram-positive signature proteins might provide an alternative line of evidence to investigate and define the relationship between the Negativicutes and the Clostridia. Negativicutes and Halanaerobiales should also share Gram-positive toolkits, especially as both groups are the only known sporulating Gram-negative bacteria.

Not all Negativicutes can sporulate: *V. parvula* and *S. sphaeroides* are examples of non-sporulating and sporulating respectively. Because not all Negativicutes can sporulate, is it unknown whether this function was gained or lost during descent from a common ancestor. The investigation of Gram-positive signature proteins could define a sporulating Gram-positive ancestor. However, this thesis has shown that Negativicutes and Halanaerobiales share a common origin and examples from both groups can sporulate. This would suggest that sporulation was a feature derived from the Gram-positive ancestor of both these groups.

If Negativicutes and Halanaerobiales were both derived from the same event, it implies that there was a very early split into these two
monophyletic groups. A genome fusion between two bacteria might lead to a bacterium with a highly disrupted, unstable genome and potentially two origins of replication. It could be hypothesised that subsequent DNA replication and division of this hybrid bacterium produced progeny bacteria with very different genomes that diverged into two monophyletic clades, which have since evolved into different environmental niches.

Overall, the arguments presented in this thesis strengthen the view that a rare genome fusion event gave rise to Negativicutes and Halanaerobiales. It is worth stressing that the Negativicutes and Halanaerobiales are related only to a small subset of Gram-negatives, which would not be the case if they were the missing link between monoderms and diderms. In this study of the Negativicutes, it has become clear when considering taxonomy, evolution or cell biology, it is essential to consider whether a Gram-negative or Gram-positive feature is being investigated. Only with this understanding, will the field of knowledge about these bacteria continue to move forward.
Appendix A: Supplementary data for Chapter Three

Table S1: The orthologues of the GNSP for Mep45 and absence of Braun’s lipoprotein in the Negativicutes.

The locus tag of the reciprocal best BLAST hit and the E-value is shown in brackets. O indicates the absence of any significant hit.

<table>
<thead>
<tr>
<th>Species</th>
<th>Braun's Lipoprotein</th>
<th>Mep45</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. acidaminophila DSM 3853</td>
<td>O</td>
<td>A_ac_01899(9e-64)</td>
</tr>
<tr>
<td>A. burkinensis DSM 6283</td>
<td>O</td>
<td>A_bu_01243(5e-65)</td>
</tr>
<tr>
<td>A. fermentans DSM 20731</td>
<td>O</td>
<td>Acfer_1670(3e-59)</td>
</tr>
<tr>
<td>A. geminatus Fo357</td>
<td>O</td>
<td>HMPREF0080_01527(4e-37)</td>
</tr>
<tr>
<td>A. lipolyticus LB2005</td>
<td>O</td>
<td>A_li_02135(3e-83)</td>
</tr>
<tr>
<td>A. longum DSM 6540</td>
<td>O</td>
<td>ALO_15582(5e-38)</td>
</tr>
<tr>
<td>C. periodontii DSM 2778</td>
<td>O</td>
<td>HMPREF9081_1152(2e-137)</td>
</tr>
<tr>
<td>D. microaerophilus UPII 345-E</td>
<td>O</td>
<td>HMPREF9220_0427(2e-51)</td>
</tr>
<tr>
<td>M. elsdenii DSM 20460</td>
<td>O</td>
<td>MELS_2084(2e-40)</td>
</tr>
<tr>
<td>M. funiformis YIT 11815</td>
<td>O</td>
<td>HMPREF9454_0194(3e-95)</td>
</tr>
<tr>
<td>M. sp. oral taxon 131 str. W906</td>
<td>O</td>
<td>HMPREF1985_01570(2e-173)</td>
</tr>
<tr>
<td>N. succinicivorans DORA_17_25</td>
<td>O</td>
<td>Q612_NSC0080G0013(2e-33)</td>
</tr>
<tr>
<td>P. fermentans DSM 17108</td>
<td>O</td>
<td>FR7_1582(3e-69)</td>
</tr>
<tr>
<td>P. succinatutens YIT 12067</td>
<td>O</td>
<td>HMPREF9443_00687(5e-72)</td>
</tr>
<tr>
<td>S. mobilis DSM 6222</td>
<td>O</td>
<td>S_mo_00108(6e-75)</td>
</tr>
<tr>
<td>S. ruminantium subsp. lactilytica TAM6421</td>
<td>O</td>
<td>SELR_01730(0.0)</td>
</tr>
<tr>
<td>T. carboxydivorans Nor1</td>
<td>O</td>
<td>TcarDRAFT_1331(4e-51)</td>
</tr>
<tr>
<td>V. parvula DSM 2008</td>
<td>O</td>
<td>Vpar_0555(3e-44)</td>
</tr>
<tr>
<td>Z. raffinosivorans DSM 20765</td>
<td>O</td>
<td>Z_ra_03647(4e-107)</td>
</tr>
</tbody>
</table>
Table S2: The orthologues of the GNSP for BamA and absence of its partner proteins in the Negativicutes.

Both the locus tag of the RBBH and the E-value is shown in brackets. O indicates the absence of any significant hit.

<table>
<thead>
<tr>
<th>Species</th>
<th>BamA</th>
<th>BamB</th>
<th>BamC</th>
<th>BamD</th>
<th>BamE</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. acidaminiphila</em> DSM 3853</td>
<td>A_ac_01873 (3e-174)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>A. burkinensis</em> DSM 6283</td>
<td>A_Bu_01217 (9e-167)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>A. fermentans</em> DSM 20731</td>
<td>Acfer_i696 (1e-160)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>A. geminatus</em> FO357</td>
<td>HMPREF00080_00467 (6e-154)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>A. lipolyticus</em> LB2005</td>
<td>A_li_02114 (7e-127)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>A. longum</em> DSM 6540</td>
<td>ALO_15747 (1e-156)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>C. periodontii</em> DSM 2778</td>
<td>HMPREF9081_0771 (5e-146)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>D. microaerophilus</em> UPII 345-E</td>
<td>HMPREF9220_II79 (2e-113)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>M. elsdenii</em> DSM 20460</td>
<td>MELS_0919 (0.0)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>M. funiformis</em> YIT I1815</td>
<td>HMPREF9454_02153 (2e-148)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>M. sp. oral taxon</em> I31 str. W9106</td>
<td>HMPREFI985_00217 (2e-138)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>N. succinicivorans</em> DORA_17_25</td>
<td>Q612_NSC00223G0008 (6e-138)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>P. fermentans</em> DSM 17108</td>
<td>FR7_1553 (5e-180)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>P. succinatutens</em> YIT I2067</td>
<td>HMPREF9443_00590 (9e-168)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>S. mobilis</em> DSM 6222</td>
<td>S_mo_0008 (2e-151)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>S. ruminantium</em> subsp. lactilytica TAM6421</td>
<td>SELR_26060 (5e-144)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>T. carboxydovorans</em> NorI</td>
<td>TcarDRAFT_1495 (1e-175)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>V. parvula</em> DSM 2008</td>
<td>Vpar_0527 (0.0)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><em>Z. raffinosivorans</em> DSM 20765</td>
<td>Z_ra_03621 (3e-156)</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Table S3: The orthologues of the GNSP for lipid A biosynthesis in the Negativicutes.

Both the locus tag of the RBBH and the E-value is shown in brackets. O indicates the absence of any significant hit.

<table>
<thead>
<tr>
<th>Species</th>
<th>LpxA</th>
<th>LpxB</th>
<th>LpxC</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. acidaminiphila</em> DSM 3853</td>
<td>A_ac_01883 (5e-142)</td>
<td>A_ac_01885 (9e-128)</td>
<td>A_ac_01881 (1e-112)</td>
</tr>
<tr>
<td><em>A. burkinensis</em> DSM 6283</td>
<td>A_bu_01227 (4e-141)</td>
<td>A_bu_01229 (2e-127)</td>
<td>A_bu_01225 (1e-111)</td>
</tr>
<tr>
<td><em>A. fermentans</em> DSM 20731</td>
<td>Acfer_1688 (2e-116)</td>
<td>Acfer_1686 (2e-118)</td>
<td>Acfer_1689 (5e-112)</td>
</tr>
<tr>
<td><em>A. geminatus</em> F0357</td>
<td>HMPREF0080_01152 (1e-144)</td>
<td>HMPREF0080_00396 (2e-152)</td>
<td>HMPREF0080_01150 (1e-119)</td>
</tr>
<tr>
<td><em>A. lipolyticus</em> LB2005</td>
<td>A_l_02103 (3e-123)</td>
<td>A_l_02101 (3e-105)</td>
<td>A_l_02105 (1e-106)</td>
</tr>
<tr>
<td><em>A. longum</em> DSM 6540</td>
<td>ALO_15687 (4e-140)</td>
<td>ALO_15677 (8e-139)</td>
<td>ALO_15697 (1e-104)</td>
</tr>
<tr>
<td><em>C. periodontii</em> DSM 2778</td>
<td>HMPREF9081_2444 (7e-125)</td>
<td>HMPREF9081_2442 (2e-100)</td>
<td>HMPREF9081_2447 (6e-83)</td>
</tr>
<tr>
<td><em>D. microaerophilus</em> UP1 345-E</td>
<td>HMPREF9220_1360 (1e-113)</td>
<td>HMPREF9220_0137 (4e-152)</td>
<td>HMPREF9220_1358 (6e-117)</td>
</tr>
<tr>
<td><em>M. elsdenii</em> DSM 20460</td>
<td>MELS_0093 (2e-146)</td>
<td>MELS_0091 (7e-152)</td>
<td>MELS_0095 (6e-122)</td>
</tr>
<tr>
<td><em>M. funiformis</em> YIT II815</td>
<td>HMPREF9454_02161 (6e-128)</td>
<td>HMPREF9454_02163 (3e-121)</td>
<td>HMPREF9454_02159 (6e-110)</td>
</tr>
<tr>
<td><em>M. sp. oral taxon</em> 13I str. W9106</td>
<td>HMPREF985_01162 (7e-131)</td>
<td>HMPREF985_01164 (1e-105)</td>
<td>HMPREF985_01160 (7e-96)</td>
</tr>
<tr>
<td><em>N. succinicivorans</em> DORA_17_25</td>
<td>Q612_NSC00162G002 (4e-133)</td>
<td>Q612_NSC00344G017 (2e-145)</td>
<td>Q612_NSC00249G001 (7e-109)</td>
</tr>
<tr>
<td><em>P. fermentans</em> DSM 17108</td>
<td>FR7_1565 (8e-137)</td>
<td>FR7_1567 (2e-145)</td>
<td>FR7_1563(4e-123)</td>
</tr>
<tr>
<td><em>P. succinatutens</em> YIT 12067</td>
<td>HMPREF9443_0060 (2e-122)</td>
<td>HMPREF9443_0060 (2e-113)</td>
<td>HMPREF9443_00598 (7e-121)</td>
</tr>
<tr>
<td><em>S. mobilis</em> DSM 6222</td>
<td>S_mo_00073 (2e-123)</td>
<td>S_mo_00071 (2e-107)</td>
<td>S_mo_00075 (2e-109)</td>
</tr>
<tr>
<td><em>S. ruminantium</em> subsp. lactilytica TAM421</td>
<td>SELR_25970 (1e-127)</td>
<td>SELR_25950 (4e-107)</td>
<td>SELR_25990 (2e-106)</td>
</tr>
<tr>
<td><em>T. carboxydovorans</em> Nor1</td>
<td>TcarDRAFT_0101 (6e-141)</td>
<td>TcarDRAFT_0103 (3e-148)</td>
<td>TcarDRAFT_0099 (1e-57)</td>
</tr>
<tr>
<td><em>V. parvula</em> DSM 2008</td>
<td>Vpar_0535 (0.0)</td>
<td>Vpar_0538 (0.0)</td>
<td>Vpar_0533 (0.0)</td>
</tr>
<tr>
<td><em>Z. raffinosivorans</em> DSM 20765</td>
<td>Z_ra_03631 (1e-130)</td>
<td>Z_ra_03633 (1e-129)</td>
<td>Z_ra_03629 (8e-105)</td>
</tr>
</tbody>
</table>

Table S3 continued overleaf.
<table>
<thead>
<tr>
<th>Species</th>
<th>LpxD</th>
<th>LpxH</th>
<th>LpxI</th>
<th>LpxL</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. acidaminophila</em> DSM 3853</td>
<td>A_ac_01878 (7e-112)</td>
<td>O</td>
<td>A_ac_01884 (1e-90)</td>
<td>A_ac_01892 (1e-101)</td>
</tr>
<tr>
<td><em>A. burkinensis</em> DSM 6283</td>
<td>A_bu_01222 (6e-112)</td>
<td>O</td>
<td>A_bu_01228 (2e-89)</td>
<td>A_bu_01236 (2e-103)</td>
</tr>
<tr>
<td><em>A. fermentans DSM 20731</em></td>
<td>Acfer_1692 (2e-109)</td>
<td>O</td>
<td>Acfer_1687 (1e-100)</td>
<td>Acfer_1679 (1e-83)</td>
</tr>
<tr>
<td><em>A. geminatus</em> F0357</td>
<td>HMPREF0080_00472 (6e-140)</td>
<td>O</td>
<td>HMPREF0080_00395 (2e-133)</td>
<td>HMPREF0080_00405 (4e-71)</td>
</tr>
<tr>
<td><em>A. longum</em> DSM 6540</td>
<td>ALO_15722 (2e-119)</td>
<td>O</td>
<td>ALO_15682 (1e-92)</td>
<td>ALO_04478 (3e-100)</td>
</tr>
<tr>
<td><em>C. periodontii</em> DSM 2778</td>
<td>HMPREF9081_779 (5e-106)</td>
<td>O</td>
<td>HMPREF9081_2443 (2e-89)</td>
<td>HMPREF9081_2434 (3e-79)</td>
</tr>
<tr>
<td><em>D. microaerophilus</em> UPII 345-E</td>
<td>HMPREF9220_1175 (3e-120)</td>
<td>O</td>
<td>HMPREF9220_0138 (7e-112)</td>
<td>HMPREF9220_0129 (1e-82)</td>
</tr>
<tr>
<td><em>M. elsdenii</em> DSM 20460</td>
<td>MELS_0924 (2e-151)</td>
<td>O</td>
<td>MELS_0092 (9e-131)</td>
<td>MELS_0083 (2e-105)</td>
</tr>
<tr>
<td><em>M. funiformis</em> YIT II815</td>
<td>HMPREF9454_02157 (6e-114)</td>
<td>O</td>
<td>HMPREF9454_02162 (1e-92)</td>
<td>HMPREF9454_02170 (3e-79)</td>
</tr>
<tr>
<td><em>M. sp. oral taxon 131</em> str. W9106</td>
<td>HMPREF985_0212 (9e-117)</td>
<td>O</td>
<td>HMPREF985_0116 (3e-93)</td>
<td>HMPREF985_0117 (5e-87)</td>
</tr>
<tr>
<td><em>N. succinivorans</em> DORA_17_25</td>
<td>Q612_NS000223 G0003 (8e-108)</td>
<td>O</td>
<td>Q612_NS00034 G00018 (4e-74)</td>
<td>Q612_NS00034 4G0008 (2e-94)</td>
</tr>
<tr>
<td><em>P. fermentans</em> DSM 17108</td>
<td>FR7_1558 (9e-125)</td>
<td>O</td>
<td>FR7_1566 (1e-95)</td>
<td>FR7_1575 (1e-101)</td>
</tr>
<tr>
<td><em>P. succinatutens</em> YIT 12067</td>
<td>HMPREF9443_0596 (8e-110)</td>
<td>O</td>
<td>HMPREF9443_00601 (2e-99)</td>
<td>HMPREF9443_00609 (5e-83)</td>
</tr>
<tr>
<td><em>S. mobilis</em> DSM 6222</td>
<td>S_mo_00077 (2e-110)</td>
<td>O</td>
<td>S_mo_00072 (6e-99)</td>
<td>S_mo_00064 (1e-101)</td>
</tr>
<tr>
<td><em>S. ruminantium</em> subsp. lactilytica TAM6421</td>
<td>SELR_26010 (4e-117)</td>
<td>O</td>
<td>SELR_25960 (3e-101)</td>
<td>SELR_25880 (8e-92)</td>
</tr>
<tr>
<td><em>T. carboxydovorans</em> Nor1</td>
<td>TcarDRAFT_0093 (1e-128)</td>
<td>O</td>
<td>TcarDRAFT_0102 (1e-101)</td>
<td>TcarDRAFT_1339 (5e-103)</td>
</tr>
<tr>
<td><em>V. parvula</em> DSM 2008</td>
<td>Vpar_0531 (0.0)</td>
<td>O</td>
<td>Vpar_0537 (0.0)</td>
<td>Vpar_0546 (0.0)</td>
</tr>
<tr>
<td><em>Z. raffinosivorans</em> DSM 20765</td>
<td>Z_ra_03626 (6e-118)</td>
<td>O</td>
<td>Z_ra_03632 (1e-90)</td>
<td>Z_ra_03640 (3e-93)</td>
</tr>
</tbody>
</table>
Appendix B: Supplementary data for Chapter Four

Figure S1: Phylogenetic tree of the 16S rRNA gene from Negativicutes.

ML phylogenetic tree of the 16S rRNA gene showing the relationships of the Negativicutes with *B. subtilis* as the outgroup. *S. sphaeroides* is in a box indicated with a star (*). The percentage bootstrap support is indicated on the nodes, which was based on the analysis of 100 replicates.
Appendix C: Supplementary data for Chapter Five

(A) *V. parvula*: Log and stationary phases were identified at 3 hours and 12 hours respectively. A logarithmic rate of 0.270 OD600 per hour and a maximum OD600 of 0.476 were observed.

(B) *S. sphaeroides*: Log and stationary phases were identified at 6 hours and 10 hours respectively. A logarithmic rate of 0.169 OD600 per hour and a maximum OD600 of 0.561 were observed.

**Figure S2: The growth curve of V. parvula and S. sphaeroides.**

The media was inoculated at an OD of 0.05 and log (L)/stationary (S) phases identified. *V. parvula* (A) The green, red and blue points indicate three biological replicates, each of those had three technical replicates. *S. sphaeroides* (B) the blue points indicate three biological replicates. A logarithmic scale is used for the Y-axis.
Table S4: Viable cell counts of *V. parvula* and *S. sphaeroides* at logarithmic phase and stationary phase.

The amounts of cells per ml were based on viable cell counts from three biological and three technical replicates. The ml of cells required for RNA extraction protocol was a conservative 500 million cells for *V. parvula*. The recommended maximum of 750 million cells was used for *S. sphaeroides*.

(A) *V. parvula*

<table>
<thead>
<tr>
<th>Cells/ml</th>
<th>Log</th>
<th>Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>225,000,000</td>
<td>826,666,666</td>
</tr>
<tr>
<td>Cells for RNA extraction</td>
<td>500,000,000</td>
<td>500,000,000</td>
</tr>
<tr>
<td>Culture for RNA extraction</td>
<td>2.23 ml</td>
<td>0.72 ml</td>
</tr>
</tbody>
</table>

(B) *S. sphaeroides*

<table>
<thead>
<tr>
<th>Cells/ml</th>
<th>Log</th>
<th>Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>53,833,333</td>
<td>77,388,889</td>
</tr>
<tr>
<td>Cells for RNA extraction</td>
<td>750,000,000</td>
<td>750,000,000</td>
</tr>
<tr>
<td>Culture for RNA extraction</td>
<td>13.93 ml</td>
<td>9.69 ml</td>
</tr>
</tbody>
</table>
Table S5: The concentration of RNA extracted from the bacterial cultures.

RNA was quantified using a Qubit (Thermo Fisher Scientific).

<table>
<thead>
<tr>
<th>Species</th>
<th>V. parvula</th>
<th>S. sphaeroides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log #1</td>
<td>272</td>
<td>675</td>
</tr>
<tr>
<td>Log #2</td>
<td>256</td>
<td>675</td>
</tr>
<tr>
<td>Log #3</td>
<td>223</td>
<td>496</td>
</tr>
<tr>
<td>Log #4</td>
<td>-</td>
<td>700</td>
</tr>
<tr>
<td>Stationary #1</td>
<td>235</td>
<td>1540</td>
</tr>
<tr>
<td>Stationary #2</td>
<td>278</td>
<td>1730</td>
</tr>
<tr>
<td>Stationary #3</td>
<td>267</td>
<td>1520</td>
</tr>
<tr>
<td>Stationary #4</td>
<td>-</td>
<td>1460</td>
</tr>
</tbody>
</table>

Figure S3: Quality control of the extracted RNA.

The extracted RNA was tested for DNA contamination using primers for the 16S rRNA gene (1482r and 27f). Genomic DNA of the species was used for a positive control and dH₂O for the negative control.
**V. parvula**

(A) Before depletion of ribosomal RNA using Ribo-Zero to remove rRNA.

**S. sphaeroides**

(B) After depletion of rRNA using Ribo-Zero shows depletion of 16S rRNA and 23S rRNA.

**V. parvula**

(C) After library preparation using the TruSeq LT cDNA kit.

**Figure S4: Bioanalyzer results for the 1st biological replicate in logarithmic phase.**

(A) shows the RNA ran before using Ribo-Zero, (B) shows the RNA after using Ribo-Zero and (C) shows the cDNA library ready for running on the MiSeq. The other biological replicates are not shown.
Table S6: Table of unexpressed signature sporulation genes.

The expression of signature sporulation genes, identified by RBBH, was calculated in *S. sphaeroides*. There were 44 genes below the cut-off of 23 TPM, indicating they may be specific to sporulation. The uniprot ID of the signature sporulation gene, the locus tag of the RBBH and the E-value of the hit in *S. sphaeroides* is shown.

<table>
<thead>
<tr>
<th>Signature Sporulation Gene</th>
<th>Locus Tag</th>
<th>E-value</th>
<th>TPM (log)</th>
<th>TPM (stat)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp</td>
<td>PI0727</td>
<td>SP2AA</td>
<td>SSPH_00156</td>
<td>3.00E-36</td>
</tr>
<tr>
<td>sp</td>
<td>PI0728</td>
<td>SP2AB</td>
<td>SSPH_00155</td>
<td>7.00E-38</td>
</tr>
<tr>
<td>sp</td>
<td>PI3801</td>
<td>SP2G</td>
<td>SSPH_00363</td>
<td>3.00E-26</td>
</tr>
<tr>
<td>sp</td>
<td>P62179</td>
<td>RP35</td>
<td>SSPH_00362</td>
<td>2.00E-118</td>
</tr>
<tr>
<td>sp</td>
<td>O34431</td>
<td>ATCL</td>
<td>SSPH_01307</td>
<td>0</td>
</tr>
<tr>
<td>sp</td>
<td>O34765</td>
<td>YLBJ</td>
<td>SSPH_00348</td>
<td>1.00E-103</td>
</tr>
<tr>
<td>sp</td>
<td>O34893</td>
<td>YNFG</td>
<td>SSPH_02901</td>
<td>2.00E-75</td>
</tr>
<tr>
<td>sp</td>
<td>O34991</td>
<td>YTVI</td>
<td>SSPH_02272</td>
<td>8.00E-22</td>
</tr>
<tr>
<td>sp</td>
<td>P07372</td>
<td>SP2D</td>
<td>SSPH_02672</td>
<td>2.00E-67</td>
</tr>
<tr>
<td>sp</td>
<td>P15281</td>
<td>SP3D</td>
<td>SSPH_02670</td>
<td>4.00E-32</td>
</tr>
<tr>
<td>sp</td>
<td>P19940</td>
<td>RP3G</td>
<td>SSPH_00361</td>
<td>2.00E-144</td>
</tr>
<tr>
<td>sp</td>
<td>P35149</td>
<td>SP4A</td>
<td>SSPH_00113</td>
<td>0</td>
</tr>
<tr>
<td>sp</td>
<td>P35157</td>
<td>SPMA</td>
<td>SSPH_00136</td>
<td>4.00E-60</td>
</tr>
<tr>
<td>sp</td>
<td>P35158</td>
<td>SPMB</td>
<td>SSPH_00135</td>
<td>2.00E-56</td>
</tr>
<tr>
<td>sp</td>
<td>P37558</td>
<td>YABP</td>
<td>SSPH_01438</td>
<td>5.00E-17</td>
</tr>
<tr>
<td>sp</td>
<td>P37968</td>
<td>SP2P</td>
<td>SSPH_01538</td>
<td>9.00E-27</td>
</tr>
<tr>
<td>sp</td>
<td>P49780</td>
<td>SP3AC</td>
<td>SSPH_02289</td>
<td>2.00E-15</td>
</tr>
<tr>
<td>sp</td>
<td>P49781</td>
<td>SP3AD</td>
<td>SSPH_02290</td>
<td>5.00E-42</td>
</tr>
<tr>
<td>sp</td>
<td>P49782</td>
<td>SP3AE</td>
<td>SSPH_02291</td>
<td>3.00E-74</td>
</tr>
<tr>
<td>sp</td>
<td>P49784</td>
<td>SP3AG</td>
<td>SSPH_02294</td>
<td>2.00E-18</td>
</tr>
<tr>
<td>sp</td>
<td>P50850</td>
<td>YLXY</td>
<td>SSPH_00234</td>
<td>5.00E-66</td>
</tr>
<tr>
<td>sp</td>
<td>P54468</td>
<td>YQFC</td>
<td>SSPH_00071</td>
<td>2.00E-23</td>
</tr>
<tr>
<td>sp</td>
<td>P94558</td>
<td>NTPA</td>
<td>SSPH_03911</td>
<td>2.00E-65</td>
</tr>
<tr>
<td>sp</td>
<td>P94559</td>
<td>YSNB</td>
<td>SSPH_03912</td>
<td>8.00E-23</td>
</tr>
<tr>
<td>sp</td>
<td>Q01367</td>
<td>SP3AA</td>
<td>SSPH_02287</td>
<td>6.00E-87</td>
</tr>
<tr>
<td>sp</td>
<td>Q01368</td>
<td>SP3AB</td>
<td>SSPH_02288</td>
<td>3.00E-27</td>
</tr>
<tr>
<td>sp</td>
<td>O34806</td>
<td>YTFJ</td>
<td>SSPH_00139</td>
<td>3.00E-39</td>
</tr>
<tr>
<td>sp</td>
<td>P17896</td>
<td>SP4B</td>
<td>SSPH_02311</td>
<td>4.00E-114</td>
</tr>
<tr>
<td>sp</td>
<td>P22322</td>
<td>GPR</td>
<td>SSPH_00091</td>
<td>2.00E-93</td>
</tr>
<tr>
<td>sp</td>
<td>P37968</td>
<td>SP2P</td>
<td>SSPH_01538</td>
<td>9.00E-27</td>
</tr>
<tr>
<td>sp</td>
<td>P38422</td>
<td>DACF</td>
<td>SSPH_00157</td>
<td>3.00E-95</td>
</tr>
<tr>
<td>sp</td>
<td>P39151</td>
<td>SP2R</td>
<td>SSPH_00360</td>
<td>6.00E-45</td>
</tr>
<tr>
<td>sp</td>
<td>P97030</td>
<td>QUEG</td>
<td>SSPH_02094</td>
<td>4.00E-19</td>
</tr>
<tr>
<td>sp</td>
<td>P7896</td>
<td>SP4B</td>
<td>SSPH_02313</td>
<td>4.00E-114</td>
</tr>
<tr>
<td>sp</td>
<td>P9940</td>
<td>RPSG</td>
<td>SSPH_00361</td>
<td>2.00E-144</td>
</tr>
<tr>
<td>sp</td>
<td>P2322</td>
<td>GPR</td>
<td>SSPH_00091</td>
<td>2.00E-93</td>
</tr>
<tr>
<td>sp</td>
<td>P7968</td>
<td>SP2P</td>
<td>SSPH_01538</td>
<td>9.00E-27</td>
</tr>
<tr>
<td>sp</td>
<td>P38422</td>
<td>DACF</td>
<td>SSPH_0157</td>
<td>3.00E-95</td>
</tr>
<tr>
<td>sp</td>
<td>P40868</td>
<td>SP5AC</td>
<td>SSPH_00679</td>
<td>5.00E-61</td>
</tr>
<tr>
<td>sp</td>
<td>P40869</td>
<td>SP5AD</td>
<td>SSPH_00680</td>
<td>1.00E-122</td>
</tr>
<tr>
<td>sp</td>
<td>P40870</td>
<td>S5AEA</td>
<td>SSPH_00147</td>
<td>2.00E-60</td>
</tr>
<tr>
<td>sp</td>
<td>P94559</td>
<td>YSNB</td>
<td>SSPH_03912</td>
<td>8.00E-23</td>
</tr>
<tr>
<td>sp</td>
<td>Q99171</td>
<td>TEPA</td>
<td>SSPH_00221</td>
<td>8.00E-114</td>
</tr>
<tr>
<td>sp</td>
<td>P37548</td>
<td>YABG</td>
<td>SSPH_01616</td>
<td>2.00E-79</td>
</tr>
</tbody>
</table>
Table S7: Predicted coding sequences with a RBS in *V. parvula*

Putative intergenic ncRNAs that were predicted from transcriptomics data using toRNAdo. Putative ORFs were predicted with getORF and then those with RBS were predicted with RBS-finder.pl. The start and stop positions of the ORF within the ncRNA are shown.

<table>
<thead>
<tr>
<th>Putative ncRNA</th>
<th>Start (bp)</th>
<th>Stop (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vpar_NCR0058043</td>
<td>58142</td>
<td>58285</td>
</tr>
<tr>
<td>Vpar_NCR0058043</td>
<td>58350</td>
<td>58457</td>
</tr>
<tr>
<td>Vpar_NCR0058043</td>
<td>58504</td>
<td>58566</td>
</tr>
<tr>
<td>Vpar_NCR0103468</td>
<td>103535</td>
<td>103615</td>
</tr>
<tr>
<td>Vpar_NCR0105317</td>
<td>105634</td>
<td>105702</td>
</tr>
<tr>
<td>Vpar_NCR0154081</td>
<td>154103</td>
<td>154258</td>
</tr>
<tr>
<td>Vpar_NCR0154081</td>
<td>154201</td>
<td>154293</td>
</tr>
<tr>
<td>Vpar_NCR0154081</td>
<td>154227</td>
<td>154313</td>
</tr>
<tr>
<td>Vpar_NCR0224561</td>
<td>224593</td>
<td>224646</td>
</tr>
<tr>
<td>Vpar_NCR0224561</td>
<td>224807</td>
<td>224938</td>
</tr>
<tr>
<td>Vpar_NCR0224561</td>
<td>224842</td>
<td>224943</td>
</tr>
<tr>
<td>Vpar_NCR0235534</td>
<td>235665</td>
<td>235721</td>
</tr>
<tr>
<td>Vpar_NCR0235534</td>
<td>235684</td>
<td>235785</td>
</tr>
<tr>
<td>Vpar_NCR0235534</td>
<td>235930</td>
<td>236028</td>
</tr>
<tr>
<td>Vpar_NCR0240394</td>
<td>240435</td>
<td>240539</td>
</tr>
<tr>
<td>Vpar_NCR0240394</td>
<td>240564</td>
<td>240725</td>
</tr>
<tr>
<td>Vpar_NCR0240394</td>
<td>240685</td>
<td>240738</td>
</tr>
<tr>
<td>Vpar_NCR0327255</td>
<td>327461</td>
<td>327559</td>
</tr>
<tr>
<td>Vpar_NCR0578777</td>
<td>578791</td>
<td>578892</td>
</tr>
<tr>
<td>Vpar_NCR0623588</td>
<td>623601</td>
<td>623657</td>
</tr>
<tr>
<td>Vpar_NCR0623588</td>
<td>623614</td>
<td>623676</td>
</tr>
<tr>
<td>Vpar_NCR0722478</td>
<td>722570</td>
<td>722623</td>
</tr>
<tr>
<td>Vpar_NCR0722478</td>
<td>722670</td>
<td>722750</td>
</tr>
<tr>
<td>Vpar_NCR0724452</td>
<td>724769</td>
<td>724837</td>
</tr>
<tr>
<td>Vpar_NCR0727866</td>
<td>727878</td>
<td>727973</td>
</tr>
<tr>
<td>Vpar_NCR0728229</td>
<td>728512</td>
<td>728414</td>
</tr>
<tr>
<td>Vpar_NCR0728229</td>
<td>728486</td>
<td>728391</td>
</tr>
<tr>
<td>Vpar_NCR0936691</td>
<td>936912</td>
<td>936974</td>
</tr>
<tr>
<td>Vpar_NCRII26819</td>
<td>1127123</td>
<td>1127058</td>
</tr>
<tr>
<td>Vpar_NCRI323877</td>
<td>1324001</td>
<td>1323912</td>
</tr>
<tr>
<td>Vpar_NCRI343574</td>
<td>1343833</td>
<td>1343780</td>
</tr>
<tr>
<td>Vpar_NCRI567302</td>
<td>1567382</td>
<td>1567305</td>
</tr>
<tr>
<td>Vpar_NCRi5738716</td>
<td>1573914</td>
<td>1573819</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Vpar_NCRi7643471</td>
<td>1764216</td>
<td>1764067</td>
</tr>
<tr>
<td>Vpar_NCRi8837990</td>
<td>1838284</td>
<td>1838177</td>
</tr>
<tr>
<td>Vpar_NCRi837990</td>
<td>1838274</td>
<td>1838158</td>
</tr>
<tr>
<td>Vpar_NCRi837990</td>
<td>1838192</td>
<td>1838088</td>
</tr>
<tr>
<td>Vpar_NCRi839592</td>
<td>1839527</td>
<td>1839613</td>
</tr>
<tr>
<td>Vpar_NCRi939592</td>
<td>1939710</td>
<td>1939594</td>
</tr>
<tr>
<td>Vpar_NCRi996427</td>
<td>1996488</td>
<td>1996429</td>
</tr>
<tr>
<td>Vpar_NCRi2067440</td>
<td>2067565</td>
<td>2067443</td>
</tr>
<tr>
<td>Vpar_NCRi2080150</td>
<td>2080826</td>
<td>2080749</td>
</tr>
<tr>
<td>Vpar_NCRi2080150</td>
<td>2080335</td>
<td>2080258</td>
</tr>
<tr>
<td>Vpar_NCRi2080150</td>
<td>2080262</td>
<td>2080188</td>
</tr>
<tr>
<td>Vpar_NCRi2080939</td>
<td>2081917</td>
<td>2081768</td>
</tr>
<tr>
<td>Vpar_NCRi2080939</td>
<td>2081752</td>
<td>2081609</td>
</tr>
<tr>
<td>Vpar_NCRi2080939</td>
<td>2081738</td>
<td>2081583</td>
</tr>
<tr>
<td>Vpar_NCRi2080939</td>
<td>2081455</td>
<td>2081243</td>
</tr>
<tr>
<td>Vpar_NCRi2080939</td>
<td>2081139</td>
<td>2080999</td>
</tr>
<tr>
<td>Vpar_NCRi2080939</td>
<td>2081116</td>
<td>2080991</td>
</tr>
<tr>
<td>Vpar_NCRi2082255</td>
<td>2082940</td>
<td>2082851</td>
</tr>
<tr>
<td>Vpar_NCRi2082255</td>
<td>2082571</td>
<td>2082494</td>
</tr>
<tr>
<td>Vpar_NCRi2082255</td>
<td>2082449</td>
<td>2082372</td>
</tr>
<tr>
<td>Vpar_NCRi2082255</td>
<td>2082359</td>
<td>2082294</td>
</tr>
<tr>
<td>Vpar_NCRi2083227</td>
<td>2083345</td>
<td>2083256</td>
</tr>
<tr>
<td>Vpar_NCRi2090905</td>
<td>2091232</td>
<td>2091167</td>
</tr>
<tr>
<td>Vpar_NCRi2090905</td>
<td>2091144</td>
<td>2091028</td>
</tr>
<tr>
<td>Vpar_NCRi2093922</td>
<td>2093995</td>
<td>2093924</td>
</tr>
</tbody>
</table>
Table S8: Predicted coding sequences with a RBS in *S. sphaeroides*

Putative intergenic ncRNAs that were predicted from transcriptomics data using toRNAdo. Putative ORFs were predicted with getORF and then those with RBS were predicted with RBS-finder.pl. The start and stop positions of the ORF within the ncRNA are shown.

<table>
<thead>
<tr>
<th>Putative ncRNA</th>
<th>Start (bp)</th>
<th>Stop (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSPH_NCR0019320_1</td>
<td>19643</td>
<td>19467</td>
</tr>
<tr>
<td>SSPH_NCR0043907_1</td>
<td>44356</td>
<td>44252</td>
</tr>
<tr>
<td>SSPH_NCR0043907_1</td>
<td>44312</td>
<td>44238</td>
</tr>
<tr>
<td>SSPH_NCR0043907_1</td>
<td>44038</td>
<td>43931</td>
</tr>
<tr>
<td>SSPH_NCR0043907_1</td>
<td>44004</td>
<td>43909</td>
</tr>
<tr>
<td>SSPH_NCR0072054_1</td>
<td>72070</td>
<td>72768</td>
</tr>
<tr>
<td>SSPH_NCR0210690_1</td>
<td>211173</td>
<td>211018</td>
</tr>
<tr>
<td>SSPH_NCR0363992_1</td>
<td>364261</td>
<td>364121</td>
</tr>
<tr>
<td>SSPH_NCR0363992_1</td>
<td>364121</td>
<td>364059</td>
</tr>
<tr>
<td>SSPH_NCR0378695_1</td>
<td>378948</td>
<td>378874</td>
</tr>
<tr>
<td>SSPH_NCR0378695_1</td>
<td>378765</td>
<td>378697</td>
</tr>
<tr>
<td>SSPH_NCR0448663_1</td>
<td>448739</td>
<td>448665</td>
</tr>
<tr>
<td>SSPH_NCR085807_2</td>
<td>185868</td>
<td>185809</td>
</tr>
<tr>
<td>SSPH_NCR038893_3</td>
<td>38974</td>
<td>39060</td>
</tr>
<tr>
<td>SSPH_NCR0155858_3</td>
<td>155871</td>
<td>155981</td>
</tr>
<tr>
<td>SSPH_NCR0155858_3</td>
<td>156188</td>
<td>156310</td>
</tr>
<tr>
<td>SSPH_NCR0174909_3</td>
<td>174994</td>
<td>175056</td>
</tr>
<tr>
<td>SSPH_NCR0178525_3</td>
<td>178666</td>
<td>178758</td>
</tr>
<tr>
<td>SSPH_NCR0184047_3</td>
<td>184133</td>
<td>184225</td>
</tr>
<tr>
<td>SSPH_NCR0184047_3</td>
<td>184225</td>
<td>184362</td>
</tr>
<tr>
<td>SSPH_NCR0184381_3</td>
<td>184398</td>
<td>184526</td>
</tr>
<tr>
<td>SSPH_NCR0184381_3</td>
<td>184575</td>
<td>184688</td>
</tr>
<tr>
<td>SSPH_NCR0196954_3</td>
<td>196994</td>
<td>197050</td>
</tr>
<tr>
<td>SSPH_NCR0196954_3</td>
<td>197087</td>
<td>197143</td>
</tr>
<tr>
<td>SSPH_NCR0204503_3</td>
<td>204515</td>
<td>204583</td>
</tr>
<tr>
<td>SSPH_NCR0204503_3</td>
<td>204708</td>
<td>204785</td>
</tr>
<tr>
<td>SSPH_NCR0204503_3</td>
<td>204761</td>
<td>204829</td>
</tr>
<tr>
<td>SSPH_NCR0207510_3</td>
<td>207621</td>
<td>207677</td>
</tr>
<tr>
<td>SSPH_NCR0207510_3</td>
<td>207584</td>
<td>207718</td>
</tr>
<tr>
<td>SSPH_NCR0207919_3</td>
<td>207955</td>
<td>208086</td>
</tr>
<tr>
<td>SSPH_NCR0221519_3</td>
<td>221534</td>
<td>221593</td>
</tr>
<tr>
<td>SSPH_NCR0221519_3</td>
<td>221566</td>
<td>221658</td>
</tr>
<tr>
<td>SSPH_NCR0229137_3</td>
<td>229162</td>
<td>229221</td>
</tr>
<tr>
<td>SSPH_NCR0239240_3</td>
<td>239338</td>
<td>239445</td>
</tr>
<tr>
<td>SSPH_NCR0239240_3</td>
<td>239551</td>
<td>239622</td>
</tr>
<tr>
<td>SSPH_NCR0015504_4</td>
<td>15619</td>
<td>15506</td>
</tr>
<tr>
<td>SSPH_NCR0157456_4</td>
<td>157628</td>
<td>157572</td>
</tr>
<tr>
<td>SSPH_NCR0157456_4</td>
<td>157725</td>
<td>157549</td>
</tr>
<tr>
<td>SSPH_NCR0170354_4</td>
<td>170502</td>
<td>170377</td>
</tr>
<tr>
<td>SSPH_NCR0186828_4</td>
<td>186897</td>
<td>186832</td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>121765</td>
<td>121863</td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>121824</td>
<td>121907</td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>121986</td>
<td>122048</td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>121972</td>
<td>122142</td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>121928</td>
<td>122197</td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>122094</td>
<td>122273</td>
</tr>
<tr>
<td>SSPH_NCR0140788_5</td>
<td>140810</td>
<td>140917</td>
</tr>
<tr>
<td>SSPH_NCR0140788_5</td>
<td>140913</td>
<td>140975</td>
</tr>
<tr>
<td>SSPH_NCR0163061_5</td>
<td>163077</td>
<td>163205</td>
</tr>
<tr>
<td>SSPH_NCR0033907_6</td>
<td>34087</td>
<td>34019</td>
</tr>
<tr>
<td>SSPH_NCR0033907_6</td>
<td>34004</td>
<td>33933</td>
</tr>
<tr>
<td>SSPH_NCR0093558_6</td>
<td>93587</td>
<td>93667</td>
</tr>
<tr>
<td>SSPH_NCR0151551_6</td>
<td>151593</td>
<td>151649</td>
</tr>
<tr>
<td>SSPH_NCR0151551_6</td>
<td>151576</td>
<td>151707</td>
</tr>
<tr>
<td>SSPH_NCR0151551_6</td>
<td>151830</td>
<td>151886</td>
</tr>
<tr>
<td>SSPH_NCR0182774_6</td>
<td>182814</td>
<td>182867</td>
</tr>
<tr>
<td>SSPH_NCR0182774_6</td>
<td>182899</td>
<td>182970</td>
</tr>
<tr>
<td>SSPH_NCR0182774_6</td>
<td>183019</td>
<td>183081</td>
</tr>
<tr>
<td>SSPH_NCR0134312_7</td>
<td>134299</td>
<td>134225</td>
</tr>
<tr>
<td>SSPH_NCR0150045_7</td>
<td>150286</td>
<td>150188</td>
</tr>
<tr>
<td>SSPH_NCR0150045_7</td>
<td>150423</td>
<td>150124</td>
</tr>
<tr>
<td>SSPH_NCR0181006_7</td>
<td>181242</td>
<td>181174</td>
</tr>
<tr>
<td>SSPH_NCR0181006_7</td>
<td>181149</td>
<td>181057</td>
</tr>
<tr>
<td>SSPH_NCR0129869_8</td>
<td>129946</td>
<td>130032</td>
</tr>
<tr>
<td>SSPH_NCR0132313_8</td>
<td>132394</td>
<td>132468</td>
</tr>
<tr>
<td>SSPH_NCR0132313_8</td>
<td>132407</td>
<td>132493</td>
</tr>
<tr>
<td>SSPH_NCR0132313_8</td>
<td>132493</td>
<td>132558</td>
</tr>
<tr>
<td>SSPH_NCR0132313_8</td>
<td>132423</td>
<td>132581</td>
</tr>
<tr>
<td>SSPH_NCR0150358_8</td>
<td>150370</td>
<td>150423</td>
</tr>
<tr>
<td>SSPH_NCR0150358_8</td>
<td>150428</td>
<td>150487</td>
</tr>
<tr>
<td>SSPH_NCR0150358_8</td>
<td>150433</td>
<td>150492</td>
</tr>
<tr>
<td>SSPH_NCR0150358_8</td>
<td>150514</td>
<td>150570</td>
</tr>
<tr>
<td>SSPH_NCR0150358_8</td>
<td>150545</td>
<td>150772</td>
</tr>
<tr>
<td>SSPH_NCR0152612_8</td>
<td>152820</td>
<td>153101</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>113736</td>
<td>113608</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>113530</td>
<td>113438</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>113204</td>
<td>113070</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>112952</td>
<td>112824</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>112762</td>
<td>112706</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>112737</td>
<td>112681</td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>112662</td>
<td>112597</td>
</tr>
<tr>
<td>SSPH_NCR0112960_9</td>
<td>113070</td>
<td>113156</td>
</tr>
<tr>
<td>SSPH_NCR0133781_9</td>
<td>133843</td>
<td>133908</td>
</tr>
<tr>
<td>SSPH_NCR0133781_9</td>
<td>133868</td>
<td>133975</td>
</tr>
<tr>
<td>SSPH_NCR0133781_9</td>
<td>133809</td>
<td>133988</td>
</tr>
<tr>
<td>Sample Code</td>
<td>DI</td>
<td>DI</td>
</tr>
<tr>
<td>-------------------</td>
<td>------</td>
<td>----------</td>
</tr>
<tr>
<td>SSPH_NCR0061465_10</td>
<td>61470</td>
<td>61529</td>
</tr>
<tr>
<td>SSPH_NCR0030565_11</td>
<td>30594</td>
<td>30809</td>
</tr>
<tr>
<td>SSPH_NCR0126879_11</td>
<td>127137</td>
<td>127051</td>
</tr>
<tr>
<td>SSPH_NCR0126729_11</td>
<td>127094</td>
<td>126954</td>
</tr>
<tr>
<td>SSPH_NCR0126729_11</td>
<td>127117</td>
<td>126908</td>
</tr>
<tr>
<td>SSPH_NCR0126729_11</td>
<td>126865</td>
<td>126731</td>
</tr>
<tr>
<td>SSPH_NCR0011717_12</td>
<td>11835</td>
<td>11764</td>
</tr>
<tr>
<td>SSPH_NCR0011717_12</td>
<td>11860</td>
<td>11738</td>
</tr>
<tr>
<td>SSPH_NCR0022710_12</td>
<td>22723</td>
<td>22791</td>
</tr>
<tr>
<td>SSPH_NCR0022710_12</td>
<td>22852</td>
<td>22920</td>
</tr>
<tr>
<td>SSPH_NCR0022710_12</td>
<td>22865</td>
<td>22948</td>
</tr>
<tr>
<td>SSPH_NCR0096526_12</td>
<td>96543</td>
<td>96593</td>
</tr>
<tr>
<td>SSPH_NCR0142700_12</td>
<td>142751</td>
<td>142897</td>
</tr>
<tr>
<td>SSPH_NCR0150082_12</td>
<td>150165</td>
<td>150365</td>
</tr>
<tr>
<td>SSPH_NCR0079509_13</td>
<td>79616</td>
<td>79512</td>
</tr>
<tr>
<td>SSPH_NCR0096135_14</td>
<td>96350</td>
<td>96261</td>
</tr>
<tr>
<td>SSPH_NCR0015580_16</td>
<td>15606</td>
<td>15719</td>
</tr>
<tr>
<td>SSPH_NCR0064454_16</td>
<td>64600</td>
<td>64457</td>
</tr>
<tr>
<td>SSPH_NCR0079036_16</td>
<td>79301</td>
<td>79158</td>
</tr>
<tr>
<td>SSPH_NCR0079036_16</td>
<td>79207</td>
<td>79079</td>
</tr>
<tr>
<td>SSPH_NCR0083772_16</td>
<td>83950</td>
<td>83774</td>
</tr>
<tr>
<td>SSPH_NCR0090870_16</td>
<td>91053</td>
<td>90970</td>
</tr>
<tr>
<td>SSPH_NCR0090870_16</td>
<td>91006</td>
<td>90905</td>
</tr>
<tr>
<td>SSPH_NCR0052083_17</td>
<td>52167</td>
<td>52217</td>
</tr>
<tr>
<td>SSPH_NCR0052083_17</td>
<td>52130</td>
<td>52255</td>
</tr>
<tr>
<td>SSPH_NCR0052456_18</td>
<td>52607</td>
<td>52714</td>
</tr>
<tr>
<td>SSPH_NCR0022874_19</td>
<td>23038</td>
<td>23188</td>
</tr>
<tr>
<td>SSPH_NCR0032782_19</td>
<td>33008</td>
<td>32952</td>
</tr>
<tr>
<td>SSPH_NCR0087610_19</td>
<td>88021</td>
<td>87941</td>
</tr>
<tr>
<td>SSPH_NCR0087610_19</td>
<td>88007</td>
<td>87921</td>
</tr>
<tr>
<td>SSPH_NCR0087610_19</td>
<td>87797</td>
<td>87654</td>
</tr>
<tr>
<td>SSPH_NCR0087610_19</td>
<td>87676</td>
<td>87611</td>
</tr>
<tr>
<td>SSPH_NCR0039209_20</td>
<td>39426</td>
<td>39482</td>
</tr>
<tr>
<td>SSPH_NCR0039209_20</td>
<td>39467</td>
<td>39610</td>
</tr>
<tr>
<td>SSPH_NCR0039209_20</td>
<td>39537</td>
<td>39611</td>
</tr>
<tr>
<td>SSPH_NCR0054606_20</td>
<td>54992</td>
<td>54813</td>
</tr>
<tr>
<td>SSPH_NCR0045073_22</td>
<td>45207</td>
<td>45121</td>
</tr>
<tr>
<td>SSPH_NCR0045073_22</td>
<td>45238</td>
<td>45095</td>
</tr>
<tr>
<td>SSPH_NCR0054692_22</td>
<td>54741</td>
<td>54950</td>
</tr>
<tr>
<td>SSPH_NCR0047783_23</td>
<td>47938</td>
<td>47786</td>
</tr>
<tr>
<td>SSPH_NCR0006001_24</td>
<td>6124</td>
<td>6180</td>
</tr>
<tr>
<td>SSPH_NCR0006001_24</td>
<td>6273</td>
<td>6380</td>
</tr>
<tr>
<td>SSPH_NCR0006001_24</td>
<td>6387</td>
<td>6452</td>
</tr>
<tr>
<td>SSPH_NCR0067012_24</td>
<td>67017</td>
<td>67079</td>
</tr>
<tr>
<td>SSPH_NCR0054471_29</td>
<td>54769</td>
<td>54647</td>
</tr>
<tr>
<td>SSPH_NCR0049974_32</td>
<td>50354</td>
<td>50178</td>
</tr>
<tr>
<td>SSPH_NCR0021429_33</td>
<td>21579</td>
<td>21514</td>
</tr>
<tr>
<td>SSPH_NCR0034153_35</td>
<td>34227</td>
<td>34156</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>SSPH_NCR0034602_35</td>
<td>34839</td>
<td>34973</td>
</tr>
<tr>
<td>SSPH_NCR0029448_37</td>
<td>29581</td>
<td>29646</td>
</tr>
<tr>
<td>SSPH_NCR0029448_37</td>
<td>29652</td>
<td>29771</td>
</tr>
<tr>
<td>SSPH_NCR0016839_40</td>
<td>17087</td>
<td>17001</td>
</tr>
<tr>
<td>SSPH_NCR0001869_43</td>
<td>1952</td>
<td>2005</td>
</tr>
<tr>
<td>SSPH_NCR0001869_43</td>
<td>1891</td>
<td>2040</td>
</tr>
<tr>
<td>SSPH_NCR0001869_43</td>
<td>2040</td>
<td>2162</td>
</tr>
<tr>
<td>SSPH_NCR0001869_43</td>
<td>2065</td>
<td>2172</td>
</tr>
<tr>
<td>SSPH_NCR0001869_43</td>
<td>2162</td>
<td>2236</td>
</tr>
<tr>
<td>SSPH_NCR0002888_43</td>
<td>3101</td>
<td>3166</td>
</tr>
<tr>
<td>SSPH_NCR0016048_44</td>
<td>16286</td>
<td>16158</td>
</tr>
<tr>
<td>SSPH_NCR0016048_44</td>
<td>16204</td>
<td>16049</td>
</tr>
<tr>
<td>SSPH_NCR0016465_48</td>
<td>16656</td>
<td>16718</td>
</tr>
<tr>
<td>SSPH_NCR0016465_48</td>
<td>16669</td>
<td>16719</td>
</tr>
<tr>
<td>SSPH_NCR0016777_48</td>
<td>17073</td>
<td>17005</td>
</tr>
<tr>
<td>SSPH_NCR0016777_48</td>
<td>16992</td>
<td>16930</td>
</tr>
<tr>
<td>SSPH_NCR0016777_48</td>
<td>16939</td>
<td>16835</td>
</tr>
<tr>
<td>SSPH_NCR0016860_48</td>
<td>16870</td>
<td>16974</td>
</tr>
<tr>
<td>SSPH_NCR0016860_48</td>
<td>16996</td>
<td>17070</td>
</tr>
<tr>
<td>SSPH_NCR0007597_53</td>
<td>7838</td>
<td>7686</td>
</tr>
<tr>
<td>SSPH_NCR0007597_53</td>
<td>7774</td>
<td>7634</td>
</tr>
<tr>
<td>SSPH_NCR0000220_69</td>
<td>229</td>
<td>342</td>
</tr>
<tr>
<td>SSPH_NCR0001176_69</td>
<td>1208</td>
<td>1276</td>
</tr>
<tr>
<td>SSPH_NCR0001176_69</td>
<td>1325</td>
<td>1387</td>
</tr>
</tbody>
</table>
Table S9: Putative ncRNA in *V. parvula*.

Putative ncRNAs were predicted using toRNAdo. The strand (S) and number of replicates in logarithmic (L) and stationary (S) are shown. The type (T) of expression shows border (b), antisense (a) and intergenic (i).

<table>
<thead>
<tr>
<th>Locus tag</th>
<th>Start(bp)</th>
<th>Stop(bp)</th>
<th>S</th>
<th>T</th>
<th>L</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vpar_NCR0041591</td>
<td>41591</td>
<td>44646</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0058043</td>
<td>58043</td>
<td>58626</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0073624</td>
<td>73624</td>
<td>73768</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Vpar_NCR0103468</td>
<td>103468</td>
<td>103633</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0105317</td>
<td>105317</td>
<td>105787</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0151984</td>
<td>151984</td>
<td>152395</td>
<td>+</td>
<td>b</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0153319</td>
<td>153319</td>
<td>153526</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0154081</td>
<td>154081</td>
<td>154315</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0224561</td>
<td>224561</td>
<td>224999</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0235534</td>
<td>235534</td>
<td>236077</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0240394</td>
<td>240394</td>
<td>240808</td>
<td>+</td>
<td>i</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0284239</td>
<td>284239</td>
<td>284382</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0313103</td>
<td>313103</td>
<td>313313</td>
<td>+</td>
<td>i</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0319827</td>
<td>319827</td>
<td>320210</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0327255</td>
<td>327255</td>
<td>327604</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0351582</td>
<td>351582</td>
<td>352448</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0393524</td>
<td>393524</td>
<td>394989</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0471562</td>
<td>471562</td>
<td>471825</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0486072</td>
<td>486072</td>
<td>486378</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0526078</td>
<td>526078</td>
<td>526814</td>
<td>+</td>
<td>b</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0578777</td>
<td>578777</td>
<td>579093</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0584388</td>
<td>584388</td>
<td>584571</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Vpar_NCR0623588</td>
<td>623588</td>
<td>623749</td>
<td>+</td>
<td>i</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0694803</td>
<td>694803</td>
<td>695049</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0722478</td>
<td>722478</td>
<td>722768</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0724452</td>
<td>724452</td>
<td>724922</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0727866</td>
<td>727866</td>
<td>728057</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR0728229</td>
<td>728229</td>
<td>728648</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0825816</td>
<td>825816</td>
<td>827580</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR0936691</td>
<td>936691</td>
<td>937002</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NRCl126819</td>
<td>1126819</td>
<td>1127163</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NRCl227143</td>
<td>1227143</td>
<td>1228673</td>
<td>+</td>
<td>a</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NRCl323877</td>
<td>1323877</td>
<td>1324144</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NRCl343574</td>
<td>1343574</td>
<td>1344020</td>
<td>-</td>
<td>i</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Vpar_NRCl360993</td>
<td>1360993</td>
<td>1361317</td>
<td>+</td>
<td>b</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR</td>
<td>ID</td>
<td>Length</td>
<td>ID</td>
<td>Length</td>
<td>ID</td>
<td>Length</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>----</td>
<td>--------</td>
</tr>
<tr>
<td>Vpar_NCR1541646</td>
<td>1541646</td>
<td>1542705</td>
<td>-</td>
<td>b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1543444</td>
<td>1543444</td>
<td>1544396</td>
<td>+</td>
<td>a</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Vpar_NCR1543449</td>
<td>1543449</td>
<td>1544429</td>
<td>-</td>
<td>b</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Vpar_NCR1559045</td>
<td>1559045</td>
<td>1559273</td>
<td>-</td>
<td>i</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Vpar_NCR1567302</td>
<td>1567302</td>
<td>1567603</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1568481</td>
<td>1568481</td>
<td>1568666</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Vpar_NCR1573816</td>
<td>1573816</td>
<td>1574069</td>
<td>-</td>
<td>i</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1763897</td>
<td>1763897</td>
<td>1764371</td>
<td>-</td>
<td>i</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1837990</td>
<td>1837990</td>
<td>1838298</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1839516</td>
<td>1839516</td>
<td>1839741</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1939522</td>
<td>1939522</td>
<td>1939785</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1973934</td>
<td>1973934</td>
<td>1974040</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1984746</td>
<td>1984746</td>
<td>1984988</td>
<td>+</td>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1985638</td>
<td>1985638</td>
<td>1986480</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR1996227</td>
<td>1996227</td>
<td>1996493</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR2022816</td>
<td>2022816</td>
<td>2023032</td>
<td>+</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2023028</td>
<td>2023028</td>
<td>2023228</td>
<td>-</td>
<td>i</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2042497</td>
<td>2042497</td>
<td>2042652</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2047638</td>
<td>2047638</td>
<td>2047966</td>
<td>+</td>
<td>b</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Vpar_NCR2049924</td>
<td>2049924</td>
<td>2050090</td>
<td>+</td>
<td>i</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2067440</td>
<td>2067440</td>
<td>2067590</td>
<td>-</td>
<td>i</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2080150</td>
<td>2080150</td>
<td>2080832</td>
<td>-</td>
<td>i</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2080939</td>
<td>2080939</td>
<td>2082055</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Vpar_NCR2082255</td>
<td>2082255</td>
<td>2082955</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2083227</td>
<td>2083227</td>
<td>2083602</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Vpar_NCR2090905</td>
<td>2090905</td>
<td>2091268</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2091364</td>
<td>2091364</td>
<td>2092968</td>
<td>-</td>
<td>b</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vpar_NCR2093922</td>
<td>2093922</td>
<td>2094522</td>
<td>-</td>
<td>i</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Table S10: Putative ncRNAs in *S. sphaeroides*.

Putative ncRNAs were predicted using toRNAdo. The strand (S) and number of replicates in logarithmic (L) and stationary (S) are shown. The type (T) of expression shows border (b), antisense (a) and intergenic (i).

<table>
<thead>
<tr>
<th>Locus tag</th>
<th>Start(bp)</th>
<th>Stop(bp)</th>
<th>S</th>
<th>T</th>
<th>L</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSPH_NCR0019320_1</td>
<td>19320</td>
<td>19730</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0027465_1</td>
<td>27465</td>
<td>27705</td>
<td>-</td>
<td>i</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0036017_1</td>
<td>36017</td>
<td>36144</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0043907_1</td>
<td>43907</td>
<td>44481</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0056723_1</td>
<td>56723</td>
<td>56951</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0072054_1</td>
<td>72054</td>
<td>72283</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0210690_1</td>
<td>210690</td>
<td>211201</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0363992_1</td>
<td>363992</td>
<td>364463</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0378695_1</td>
<td>378695</td>
<td>378963</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0448663_1</td>
<td>448663</td>
<td>448746</td>
<td>-</td>
<td>i</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>SSPH_NCR0000427_2</td>
<td>486354</td>
<td>488571</td>
<td>-</td>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>SSPH_NCR0013594_2</td>
<td>499521</td>
<td>499885</td>
<td>-</td>
<td>b</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SSPH_NCR0017470_2</td>
<td>503397</td>
<td>503759</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0041777_2</td>
<td>527704</td>
<td>527943</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>SSPH_NCR0059474_2</td>
<td>545401</td>
<td>545625</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0095431_2</td>
<td>581358</td>
<td>581532</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0102016_2</td>
<td>596143</td>
<td>598151</td>
<td>-</td>
<td>b</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0121326_2</td>
<td>607253</td>
<td>608448</td>
<td>-</td>
<td>b</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>SSPH_NCR0123727_2</td>
<td>609654</td>
<td>610304</td>
<td>+</td>
<td>ib</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>SSPH_NCR0126400_2</td>
<td>612327</td>
<td>612440</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0185807_2</td>
<td>671734</td>
<td>671829</td>
<td>-</td>
<td>i</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0258256_2</td>
<td>744183</td>
<td>744334</td>
<td>-</td>
<td>a</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>SSPH_NCR0026593_3</td>
<td>803410</td>
<td>803934</td>
<td>+</td>
<td>a</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>SSPH_NCR0038893_3</td>
<td>815710</td>
<td>815928</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0113381_3</td>
<td>890198</td>
<td>891642</td>
<td>+</td>
<td>ib</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>SSPH_NCR0138782_3</td>
<td>915599</td>
<td>915771</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0155858_3</td>
<td>932675</td>
<td>933207</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>SSPH_NCR0156577_3</td>
<td>933394</td>
<td>933576</td>
<td>+</td>
<td>i</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0174909_3</td>
<td>951726</td>
<td>952249</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0178525_3</td>
<td>955342</td>
<td>955712</td>
<td>+</td>
<td>i</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>SSPH_NCR0184047_3</td>
<td>960864</td>
<td>961198</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0184381_3</td>
<td>961198</td>
<td>961505</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0196954_3</td>
<td>973771</td>
<td>974094</td>
<td>+</td>
<td>i</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0204503_3</td>
<td>981320</td>
<td>981788</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0207510_3</td>
<td>984327</td>
<td>984609</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0207919_3</td>
<td>984736</td>
<td>985032</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0221519_3</td>
<td>998336</td>
<td>998690</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0229137_3</td>
<td>1005954</td>
<td>1006226</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0239240_3</td>
<td>1016057</td>
<td>1016485</td>
<td>+</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0243970_3</td>
<td>1020787</td>
<td>1023040</td>
<td>+</td>
<td>ib</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>SSPH_NCR0002495_4</td>
<td>1040930</td>
<td>1041230</td>
<td>-</td>
<td>i</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0003652_4</td>
<td>1042087</td>
<td>1042348</td>
<td>-</td>
<td>ib</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SSPH_NCR0013984_4</td>
<td>1052419</td>
<td>1052721</td>
<td>- i 4 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0015504_4</td>
<td>1053939</td>
<td>1054068</td>
<td>- i 0 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0023881_4</td>
<td>1062316</td>
<td>1063146</td>
<td>- i 4 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0030750_4</td>
<td>1069185</td>
<td>1072141</td>
<td>+ ib 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0055525_4</td>
<td>1093960</td>
<td>1094306</td>
<td>- b 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0157456_4</td>
<td>1195891</td>
<td>1196710</td>
<td>- i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0170354_4</td>
<td>1208789</td>
<td>1208955</td>
<td>- i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0186828_4</td>
<td>1225263</td>
<td>1225569</td>
<td>- i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0121669_5</td>
<td>1373210</td>
<td>1373846</td>
<td>+ i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0140788_5</td>
<td>1392329</td>
<td>1392555</td>
<td>+ i 4 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0163061_5</td>
<td>1414602</td>
<td>1414746</td>
<td>+ i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0202169_5</td>
<td>1453710</td>
<td>1453895</td>
<td>+ i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0033907_6</td>
<td>1487921</td>
<td>1488159</td>
<td>- i 3 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0034784_6</td>
<td>1488798</td>
<td>1489707</td>
<td>- ib 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0064676_6</td>
<td>1518690</td>
<td>1519696</td>
<td>+ b 3 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0093558_6</td>
<td>1547572</td>
<td>1547774</td>
<td>+ i 3 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0105671_6</td>
<td>1559685</td>
<td>1559942</td>
<td>+ i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0151551_6</td>
<td>1605565</td>
<td>1605902</td>
<td>+ i 2 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0154536_6</td>
<td>1608550</td>
<td>1608814</td>
<td>+ a 2 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0182774_6</td>
<td>1636788</td>
<td>1637137</td>
<td>+ i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0054770_7</td>
<td>1692247</td>
<td>1693190</td>
<td>+ b 3 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0075025_7</td>
<td>1712502</td>
<td>1712854</td>
<td>- i 3 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0102852_7</td>
<td>1740329</td>
<td>1741020</td>
<td>- b 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0123920_7</td>
<td>1761397</td>
<td>1762066</td>
<td>- b 2 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0134132_7</td>
<td>1771609</td>
<td>1771861</td>
<td>- i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0150045_7</td>
<td>1787522</td>
<td>1787929</td>
<td>- i 4 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0177894_7</td>
<td>1815371</td>
<td>1815549</td>
<td>- i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0181006_7</td>
<td>1818483</td>
<td>1818754</td>
<td>- i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0031241_8</td>
<td>1850848</td>
<td>1851056</td>
<td>- i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0129869_8</td>
<td>1949476</td>
<td>1949658</td>
<td>+ i 0 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0132313_8</td>
<td>1951920</td>
<td>1952188</td>
<td>+ i 2 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0150358_8</td>
<td>1969965</td>
<td>1970379</td>
<td>+ i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0152612_8</td>
<td>1972219</td>
<td>1972804</td>
<td>+ i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0007472_9</td>
<td>2007640</td>
<td>2007862</td>
<td>- b 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0053767_9</td>
<td>2053935</td>
<td>2055152</td>
<td>- ib 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0065255_9</td>
<td>2065423</td>
<td>2066926</td>
<td>- b 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0100289_9</td>
<td>2100457</td>
<td>2101367</td>
<td>- a 1 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0105774_9</td>
<td>2106942</td>
<td>2109534</td>
<td>+ b 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0112462_9</td>
<td>2112630</td>
<td>2113960</td>
<td>- i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0112960_9</td>
<td>2113128</td>
<td>2113325</td>
<td>+ i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0131587_9</td>
<td>213755</td>
<td>2133831</td>
<td>+ b 3 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0131972_9</td>
<td>2132140</td>
<td>2133467</td>
<td>- a 0 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0133781_9</td>
<td>2133949</td>
<td>2134220</td>
<td>+ i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0168064_9</td>
<td>2168232</td>
<td>2168731</td>
<td>+ a 0 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0061465_10</td>
<td>2237327</td>
<td>2237548</td>
<td>+ i 0 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0120140_10</td>
<td>2296002</td>
<td>2296625</td>
<td>- b 3 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0121874_10</td>
<td>2297736</td>
<td>2298040</td>
<td>+ a 2 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR000679_II</td>
<td>2333604</td>
<td>2333738</td>
<td>+ i 4 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPH_NCR0030565_II</td>
<td>2363490</td>
<td>2363802</td>
<td>+ i 4 4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix C

<table>
<thead>
<tr>
<th>SSPH_NCR0116840_11</th>
<th>2449765</th>
<th>2450434</th>
<th>-b 4 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSPH_NCR0126729_11</td>
<td>2459654</td>
<td>2460094</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR011717_12</td>
<td>2499904</td>
<td>2500059</td>
<td>-i 4 0</td>
</tr>
<tr>
<td>SSPH_NCR022710_12</td>
<td>2510897</td>
<td>2511198</td>
<td>+i 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0045874_12</td>
<td>2534061</td>
<td>2534149</td>
<td>+i 1 2</td>
</tr>
<tr>
<td>SSPH_NCR0078815_12</td>
<td>2567002</td>
<td>2568784</td>
<td>+b 0 4</td>
</tr>
<tr>
<td>SSPH_NCR0096288_13</td>
<td>2584475</td>
<td>2584624</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0096526_12</td>
<td>2584713</td>
<td>2584823</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0142700_12</td>
<td>2630887</td>
<td>2631143</td>
<td>+i 0 3</td>
</tr>
<tr>
<td>SSPH_NCR0149779_12</td>
<td>2637966</td>
<td>2638529</td>
<td>-b 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0150082_12</td>
<td>2638269</td>
<td>2638554</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0000032_13</td>
<td>2638686</td>
<td>2638998</td>
<td>+b 0 4</td>
</tr>
<tr>
<td>SSPH_NCR0020890_13</td>
<td>2659544</td>
<td>2659880</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0079509_13</td>
<td>2718163</td>
<td>2718295</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0090287_14</td>
<td>2872355</td>
<td>2872400</td>
<td>-i 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0096135_14</td>
<td>2878203</td>
<td>2878654</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0096209_14</td>
<td>2878277</td>
<td>2878540</td>
<td>+i 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0015580_16</td>
<td>3016865</td>
<td>3017006</td>
<td>+i 4 3</td>
</tr>
<tr>
<td>SSPH_NCR0015645_16</td>
<td>3016930</td>
<td>3017114</td>
<td>-i 0 4</td>
</tr>
<tr>
<td>SSPH_NCR0045632_16</td>
<td>3046917</td>
<td>3048193</td>
<td>-b 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0064454_16</td>
<td>3065739</td>
<td>3066058</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0079036_16</td>
<td>3080321</td>
<td>3080597</td>
<td>-i 0 3</td>
</tr>
<tr>
<td>SSPH_NCR0083772_16</td>
<td>3085057</td>
<td>3085259</td>
<td>-i 0 4</td>
</tr>
<tr>
<td>SSPH_NCR0090870_16</td>
<td>3092155</td>
<td>3092501</td>
<td>-i 4 1</td>
</tr>
<tr>
<td>SSPH_NCR0052083_17</td>
<td>3155914</td>
<td>3156187</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0052456_18</td>
<td>3253159</td>
<td>3253493</td>
<td>+i 4 3</td>
</tr>
<tr>
<td>SSPH_NCR00005339_19</td>
<td>3300614</td>
<td>3301695</td>
<td>-b 0 3</td>
</tr>
<tr>
<td>SSPH_NCR0022874_19</td>
<td>3318149</td>
<td>3318407</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0022885_19</td>
<td>3318160</td>
<td>3318488</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0032782_19</td>
<td>3328057</td>
<td>3328305</td>
<td>-i 3 0</td>
</tr>
<tr>
<td>SSPH_NCR0087610_19</td>
<td>3382885</td>
<td>3383389</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0039209_20</td>
<td>3426504</td>
<td>3426906</td>
<td>+i 4 1</td>
</tr>
<tr>
<td>SSPH_NCR0054606_20</td>
<td>3441901</td>
<td>3442398</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0000007_22</td>
<td>3556663</td>
<td>3557446</td>
<td>+b 2 1</td>
</tr>
<tr>
<td>SSPH_NCR0019379_22</td>
<td>3576035</td>
<td>3576971</td>
<td>-b 4 3</td>
</tr>
<tr>
<td>SSPH_NCR0023531_22</td>
<td>3580187</td>
<td>3580401</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0045073_22</td>
<td>3601729</td>
<td>3602016</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0053043_22</td>
<td>3609699</td>
<td>3610324</td>
<td>-b 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0054692_22</td>
<td>3611348</td>
<td>3611608</td>
<td>+i 4 3</td>
</tr>
<tr>
<td>SSPH_NCR0003040_23</td>
<td>3639024</td>
<td>3639235</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0048849_23</td>
<td>3677833</td>
<td>3678053</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0047783_23</td>
<td>3683767</td>
<td>3684092</td>
<td>-i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0006001_24</td>
<td>3720103</td>
<td>3720570</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0036291_24</td>
<td>3750393</td>
<td>3750782</td>
<td>-a 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0067012_24</td>
<td>3781114</td>
<td>3781363</td>
<td>+i 4 4</td>
</tr>
<tr>
<td>SSPH_NCR0074869_24</td>
<td>3788971</td>
<td>3789058</td>
<td>+i 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0026425_25</td>
<td>3816200</td>
<td>3817041</td>
<td>-ib 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0054155_25</td>
<td>3843930</td>
<td>3844417</td>
<td>-b 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0059622_25</td>
<td>3849397</td>
<td>3849482</td>
<td>-i 4 0</td>
</tr>
<tr>
<td>SSPH_NCR0063189_25</td>
<td>3852964</td>
<td>3853598</td>
<td>-</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------</td>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>SSPH_NCR0052489_26</td>
<td>3915704</td>
<td>3915953</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0032815_27</td>
<td>3967809</td>
<td>3968425</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0035312_27</td>
<td>3970306</td>
<td>3971819</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0035632_27</td>
<td>3970626</td>
<td>3971641</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0036660_27</td>
<td>3971654</td>
<td>3972651</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0006953_28</td>
<td>4011866</td>
<td>4012326</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0054471_29</td>
<td>4117326</td>
<td>4117743</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0058520_30</td>
<td>4126455</td>
<td>4126654</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0032476_32</td>
<td>4263570</td>
<td>4264842</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0049974_32</td>
<td>4281068</td>
<td>4281500</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0016392_33</td>
<td>4301406</td>
<td>4302858</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0021429_33</td>
<td>4306443</td>
<td>4306612</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0021508_33</td>
<td>4306522</td>
<td>4306928</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0024609_35</td>
<td>4408736</td>
<td>4409037</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0034153_35</td>
<td>4418280</td>
<td>4418487</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0034602_35</td>
<td>4418729</td>
<td>4419213</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0029448_37</td>
<td>4496078</td>
<td>4496402</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0033582_37</td>
<td>4500212</td>
<td>4500538</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0009637_38</td>
<td>4510624</td>
<td>4511096</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0016839_40</td>
<td>4579659</td>
<td>4580044</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0005887_42</td>
<td>4626863</td>
<td>4627098</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0005666_43</td>
<td>4648048</td>
<td>4648262</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0051869_43</td>
<td>4649351</td>
<td>4649864</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0002543_43</td>
<td>4650025</td>
<td>4650117</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0002888_43</td>
<td>4650370</td>
<td>4650743</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0011095_44</td>
<td>4684477</td>
<td>4684957</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0016048_44</td>
<td>4689430</td>
<td>4689684</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0010071_48</td>
<td>4775919</td>
<td>4776386</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0016465_48</td>
<td>4782313</td>
<td>4782567</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0016777_48</td>
<td>4782625</td>
<td>4782926</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0016860_48</td>
<td>4782708</td>
<td>4782920</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0007597_53</td>
<td>4854177</td>
<td>4854440</td>
<td>-</td>
</tr>
<tr>
<td>SSPH_NCR0006960_56</td>
<td>4881271</td>
<td>4883026</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR0000220_69</td>
<td>4915380</td>
<td>4915567</td>
<td>+</td>
</tr>
<tr>
<td>SSPH_NCR000176_69</td>
<td>4916336</td>
<td>4916549</td>
<td>+</td>
</tr>
</tbody>
</table>
Bibliography


<table>
<thead>
<tr>
<th></th>
<th>Authors</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
</table>


27. Marchandin, H., et al., Negativicoccus succinicivorans gen. nov., sp. nov., isolated from human clinical samples, emended description of the family Veillonellaceae and description of Negativicutes classis nov., Selenomonadales ord. nov. and Acidaminococcaceae fam. nov.


111. Kaneko, J., et al., Complete genome sequence of Selenomonas ruminantium subsp. lactilytica will accelerate further understanding of the nature of the class Negativicutes. FEMS Microbiol Lett, 2015. 362(9).


121. Chang Yj Fau - Pukall, R., et al., Complete genome sequence of Acidaminococcus fermentans type strain (VR4). (1944-3277 (Electronic)).


153. Biopython, GenomeDiagram: script to mimic Proux et al 2002 Figure 6. Biopython.


233. Williams, G., EMBOSS: getORF.


